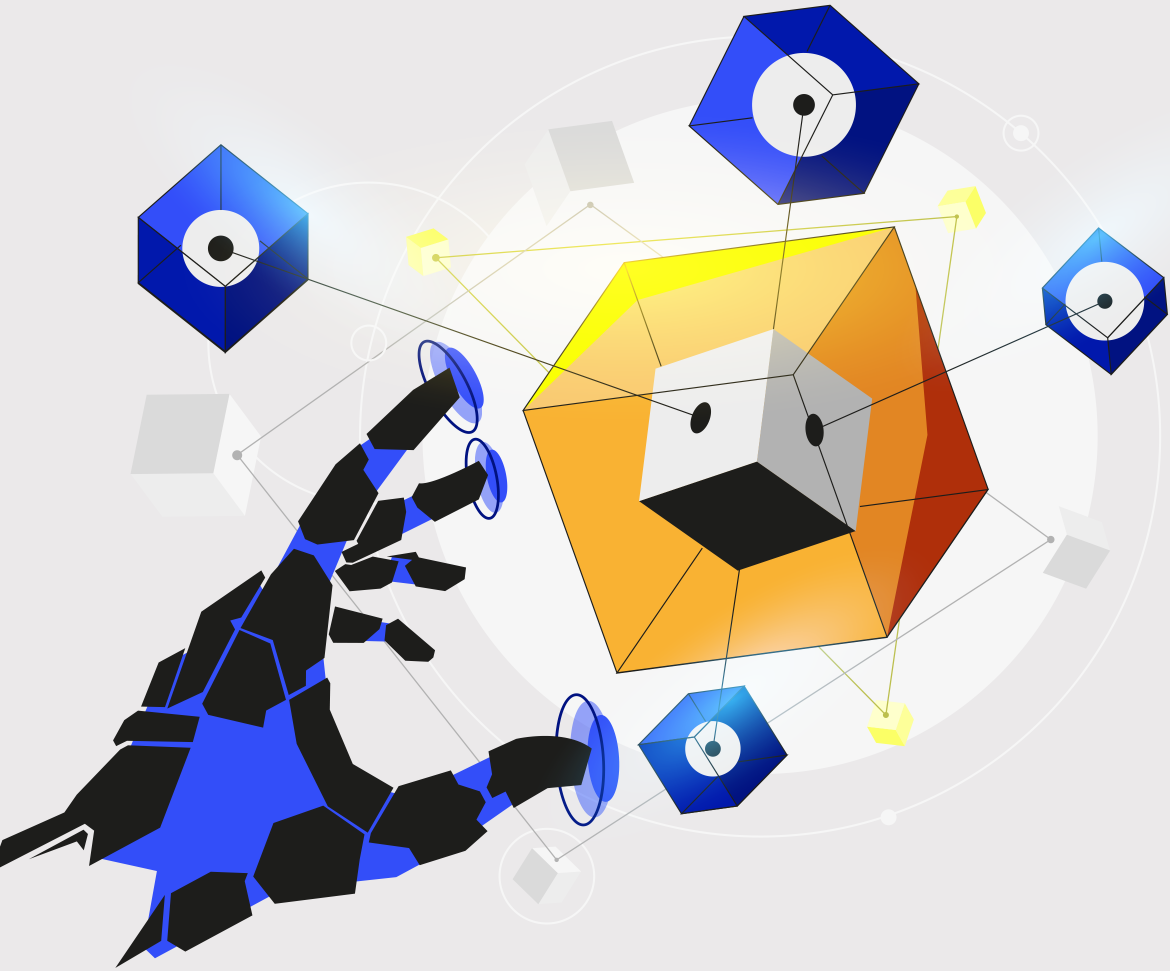
جون دي كيليهر وبريندان تيرني

علم البيانات

ترجمة رشا صلاح الدخاخني



سلسلة المعارف الأساسية

علم البيانات

تأليف
جون دي كيليهر وبريندان تيرني

ترجمة
رشا صلاح الدخاخني

مراجعة
هبة عبد العزيز غانم

هنـــداوي

علم البيانات

جون دي كيليهر
وبريندان تيرني

Data Science

John D. Kelleher
and Brendan Tierney

الناشر مؤسسة هنداوي
المشهرة برقم ١٠٥٨٥٩٧٠ بتاريخ ٢٠١٧/١/٢٦

يورك هاوس، شييت ستريت، ونديسور، SL4 1DD، المملكة المتحدة
تليفون: ٨٣٢٥٢٢ ١٧٥٣ (٠) ٤٤ +
البريد الإلكتروني: hindawi@hindawi.org
الموقع الإلكتروني: https://www.hindawi.org

إنَّ مؤسسة هنداوي غير مسئولة عن آراء المؤلف وأفكاره، وإنما يعبِّر الكتاب عن آراء مؤلفه.

تصميم الغلاف: ولاء الشاهد

الترقيم الدولي: ٤ ٣٧٧٩ ٥٢٧٣ ١ ٩٧٨

صدر الكتاب الأصلي باللغة الإنجليزية عام ٢٠١٨.
صدرت هذه الترجمة عن مؤسسة هنداوي عام ٢٠٢٥.

المحتويات

شكر وتقدير	٧
تمهيد السلسلة	٩
المقدمة	١١
١- ما علمُ البيانات؟	١٥
٢- ما المقصود بالبيانات وما المقصود بمجموعة البيانات؟	٣٩
٣- النظام البيئي لعلم البيانات	٥٧
٤- أساسيات تعلُّم الآلة	٧٥
٥- مهام علم البيانات القياسية	١١٣
٦- الخصوصية والأخلاقيات	١٣١
٧- التأثير المستقبلي لعلم البيانات ومبادئ النجاح	١٥٥
مسرد المصطلحات	١٦٧
ملاحظات	١٧٩
قراءات إضافية	١٨٥
المراجع	١٨٧

شكر وتقدير

يشكر المؤلِّفان كلًّا من بول ماكيلروي وبرايان ليهي على قراءة المسوَّدات الأولى للكتاب والتعقيب عليها بالتعليقات. كما يتوجَّهان أيضًا بالشكر إلى المراجعَين المجهولَين اللذَين قدَّما تعقيباتٍ تفصيليةً مفيدة على مسوَّدة الكتاب، ويشكران أيضًا طاقم العاملين في مؤسسة «إم آي تي بريس» على دعمهما ونصائحهما.

ويتوجَّه جون دي كيليهر بالشكر إلى أُسرته وأصدقائه على دعمهم وتشجيعهم أثناء إعداد هذا الكتاب، ويُهدي هذا الكتاب إلى والده جون بيرنارد كيليهر تقديرًا لمحبَّته وصداقته.

يتقدم بريندان تيرني بالشكر إلى جريس ودانيال وإليانور لدعمهم المستمر أثناء تأليف كتابٍ آخر (كتابه الرابع)، والموازنة بين مختلف المهام اليومية والسفر.

تمهيد السلسلة

تُقدِّم «سلسلة المعارف الأساسية» التي تنشرها مؤسسة «إم آي تي بريس» كُتبًا موجزة بلغةٍ جزلة سهلة الفهم، وشكلٍ أنيق، وحجمٍ صغير يُلائم الجيب، تُناقِش الموضوعات التي تُثير الاهتمام في الوقت الحالي. ولمَّا كانت كُتب هذه السلسلة من تأليف مفكرين بارزين، فإنها تُقدِّم آراء الخبراء بشأن موضوعاتٍ تتنوَّع بين المجالات الثقافية والتاريخية، إضافةً إلى العلمية والتقنية.

في ظلِّ ما يَشيع في هذا العصر من إشباعٍ لحظيٍّ للمعلومات، أضحى لدى الجميع القدرةُ على الوصول إلى الآراء والأفكار والشروح السطحية بسرعةٍ وسهولة، وأصبح من الصعوبة بمكانٍ أن يَحظى المرءُ بالمعرفة الأساسية التي تُيسِّر فَهمًا صادقًا للعالَم؛ وما تفعله كُتب هذه السلسلة هو أنها تُحقِّق ذلك الغرض. وكل كتابٍ من هذه الكُتب المختصَرة يُقدِّم للقارئ وسيلةً مُيسَّرة للوصول إلى الأفكار المعقَّدة، من خلال تبسيط المواد المتخصِّصة لغير المختصِّين، وشرْح الموضوعات المهمة بأبسطِ طريقةٍ ممكنة.

بروس تيدور
أستاذ الهندسة البيولوجية وعلوم الكمبيوتر
«معهد ماساتشوستس للتكنولوجيا»

المقدمة

يهدف علم البيانات إلى تحسين عملية اتخاذ القرارات من خلال الاستناد إلى الرؤى المستنيرة المستخلصة من مجموعات كبيرة من البيانات. وينطوي علم البيانات، بوصفه أحد ميادين النشاط الإنساني، على مجموعة من المبادئ وتعريفات المشكلات والخوارزميات والعمليات من أجل استخلاص الأنماط غير الواضحة، والمفيدة من المجموعات الكبيرة من البيانات. وهو علم وثيق الصلة بمجالَي التنقيب في البيانات وتعلُّم الآلة؛ لكنه أوسع نطاقًا من كليهما. اليوم، يقود علم البيانات عملية اتخاذ القرارات في جميع مناحي الحياة تقريبًا بالمجتمعات الحديثة. وتشمل بعض الطرق التي ربما يؤثر بها علم البيانات على حياتك اليومية تحديد الإعلانات التي تظهر لك عبر الإنترنت؛ والترشيحات التي تأتيك عن الأفلام والكتب ومقترحات الصداقة عبر وسائل التواصل الاجتماعي؛ ورسائل البريد الإلكتروني التي تصفى وتوضع في مجلد رسائل البريد العشوائي؛ والعروض التي تتلقَّاها عند تجديد خدمة الهاتف المحمول خاصتك؛ وتكلفة قسط التأمين الصحي الخاص بك؛ وتعاقُب إشارات المرور في منطقتك وتوقيتاتها؛ وكيفية تصميم العقاقير التي ربما تحتاج إليها؛ والأماكن التي تستهدفها الشرطة في مدينتك.

إن التوسع في استخدام علم البيانات عبر مجتمعاتنا يأتي مدفوعًا بظهور البيانات الضخمة ووسائل التواصل الاجتماعي، وزيادة القدرة الحوسبية، والانخفاض الهائل في تكلفة ذاكرة الكمبيوتر وتطوير وسائل أكثر فعالية لتحليل البيانات ونمذجتها مثل التعلُّم العميق. وتعني هذه العوامل مجتمعةً أنه صار من الأسهل على المؤسسات جمع البيانات وتخزينها ومعالجتها أكثر من أي وقتٍ مضى. وفي الوقت نفسه، تعني هذه الابتكارات التقنية والاستخدام الأوسع نطاقًا لعلم البيانات أن التحدِّيات الأخلاقية المتعلقة باستخدام البيانات وخصوصية الأفراد صارت موضوعاتٍ أكثر إلحاحًا عما كانت عليه في الماضي.

ويهدف هذا الكتاب إلى توفير مقدمةٍ إلى علم البيانات تُغطي عناصر المجال الأساسية بعُمق بحيث يقدم فهمًا مبدئيًّا للمجال.

يقدم لنا الفصل الأول مجال علم البيانات ويوفر تاريخًا موجزًا لكيفية نشأته وتطوره. كما يتناول السبب وراء اعتبار علم البيانات ذا أهميةٍ في الوقت الراهن بالإضافة إلى بعض العوامل التي تحث على اعتماده وتبنِّيه. ويُختتم الفصل باستعراض بعض الخُرافات المرتبطة بعلم البيانات وتفنيدها. أما الفصل الثاني فيقدم المفاهيم الأساسية المتعلقة بالبيانات. كما يصف المراحل القياسية لمشروع علم البيانات؛ ألا وهي فهم المشروع، وفهم البيانات، وتجهيز البيانات، والنمذجة، والتقييم، والنشر. ويركز الفصل الثالث على البنية التحتية للبيانات والتحديات التي تفرضها البيانات الضخمة ودمج البيانات المستخرَجة من مصادر متعددة. ويتمثل أحد الجوانب الخاصة بالبِنية التحتية النموذجية للبيانات، التي يمكن أن تُمثل تحديًا، في أن البيانات الموجودة في قواعد البيانات ومستودعات البيانات عادة ما تكون على وحدات خدمة مختلفة عن وحدات الخدمة المستخدَمة من أجل تحليل البيانات. وكنتيجة لذلك، عند التعامل مع مجموعات البيانات الكبيرة، يمكن قضاء وقتٍ أطول مما هو متوقَّع في نقل البيانات بين وحدات الخدمة التي توجد فيها قواعد البيانات أو مستودعات البيانات ووحدات الخدمة المستخدَمة من أجل تحليل البيانات وتعلُّم الآلة. ويبدأ الفصل الثالث بوصف بِنية علم البيانات التحتية النموذجية من أجل مؤسسةٍ ما وبعض الحلول الناشئة لتحدي نقل مجموعات البيانات الكبيرة داخل إطار البنية التحتية، التي تشمل استخدام تعلُّم الآلة المدمَج في قاعدة البيانات، واستخدام منصة هادوب لتخزين البيانات ومعالجتها، وتطوير نُظم قواعد البيانات المختلطة التي تجمع بكلِّ سلاسةٍ برامج قواعد البيانات التقليدية والحلول الشبيهة بمنصة هادوب. ويُختتم الفصل بإلقاء الضوء على بعض التحديات الخاصة بدمج البيانات عبر المؤسسة وإخراجها على هيئة شكلٍ موحَّد مناسب لتعلُّم الآلة. ويقدم الفصل الرابع مجال تعلُّم الآلة ويشرح بعضًا من أشهر الخوارزميات والنماذج الخاصة بتعلُّم الآلة، بما في ذلك الشبكات العصبية والتعلُّم العميق ونماذج الهيكل الشجري لاتخاذ القرارات (وتُعرَف أيضًا بشجرة اتخاذ القرار). ويركز الفصل الخامس على الربط بين خبرات تعلُّم الآلة ومشكلات العالم الواقعي من خلال استعراض مجموعةٍ من مشكلات المشروعات التجارية المعتادة ووصف كيفية حلِّها من خلال حلول تعلُّم الآلة. ويستعرض الفصل السادس التداعِيات الأخلاقية لعلم البيانات، وآخر المستجدات في لوائح تنظيم البيانات وبعض المناهج الحوسبية الجديدة للحفاظ على

خصوصية الأفراد في إطار العمليات المتضمنة علم البيانات. وأخيرًا، يصف الفصل السابع بعضًا من المجالات التي سيكون لعلم البيانات تأثير كبير عليها في المستقبل القريب ويتطرَّق لبعض المبادئ المهمة لتحديد ما إذا كان مشروع علم البيانات سينجح أم سيفشل.

الفصل الأول

ما علمُ البيانات؟

ينطوي علم البيانات على مجموعةٍ من المبادئ وتعريفات المشكلات والخوارزميات والعمليات التي تهدف لاستخراج الأنماط غير الواضحة والمفيدة من مجموعات البيانات الكبيرة. لقد تطورت الكثير من عناصر علم البيانات في مجالاتٍ ذات صلة مثل تعلُّم الآلة والتنقيب في البيانات. وواقع الأمر أن مصطلحات مثل: «علم البيانات» و«تعلُّم الآلة» و«التنقيب في البيانات» كثيرًا ما تُستخدَم بالتبادل بعضها مع بعض. والقاسم المشترك عبر كلٍّ هذه التخصُّصات هو التركيز على تحسين عملية اتخاذ القرار عن طريق تحليل البيانات. وعلى الرغم من أن علم البيانات يستفيد من هذين المجالَين الآخرين، فهو أوسع نطاقًا منهما. إذ يركز «تعلُّم الآلة» على تصميم الخوارزميات وتقييمها من أجل استخلاص الأنماط من البيانات المتاحة. ويتعامل «التنقيب في البيانات» بوجهٍ عامٍّ مع تحليل البيانات الهيكلية وكثيرًا ما ينطوي على التركيز على التطبيقات التجارية. أما علمُ البيانات فهو يضع كلَّ هذه الاعتبارات في الحسبان؛ ولكنه يخوض أيضًا تحدِّيات أخرى، مثل استخلاص البيانات غير الهيكلية من وسائل التواصل الاجتماعي والويب وتنقيتها ونقلها؛ واستخدام تقنيات البيانات الضخمة لتخزين مجموعات البيانات الضخمة غير الهيكلية ومعالجتها؛ هذا بالإضافة إلى المسائل المتعلقة بأخلاقيات التعامل مع البيانات واللوائح التنظيمية الخاصة بها.

ومن خلال الاستعانة بعلم البيانات، يُمكننا استخلاص أنواعٍ مختلفة من الأنماط. ربما نرغب، مثلًا، في استخلاص الأنماط التي تساعدنا في تحديد مجموعات العملاء الذين يُظهِرون سلوكياتٍ مماثلةً وأذواقًا مُتشابهة. وبالاستعانة بالمصطلحات التجارية، تُعرَف هذه المهمة بـ «تجزئة العملاء»، أما إذا استعنَّا بمصطلحات علم البيانات، فإنها تُسمَّى «التجميع». وعوضًا عن ذلك، ربما نرغب في استخلاص نمطٍ يُحدد منتجات يتكرَّر شراؤها

معًا، وهي عملية يُطلق عليها «التنقيب عن قواعد الارتباط». أو ربما نرغب في استخلاص أنماط تُحدد الأحداث الغريبة أو الشاذة، مثل مطالبات التأمين المزورة، وهي عملية تُعرف باسم «اكتشاف الشذوذ» أو «اكتشاف القِيَم الشاذة». وأخيرًا، ربما نرغب في تحديد الأنماط التي تُساعدنا على تصنيف الأشياء. على سبيل المثال، القاعدة التالية توضِّح ما قد يبدو عليه نمط التصنيف المستخلَص من مجموعة بيانات البريد الإلكتروني: «إذا اشتملت رسالة البريد الإلكتروني على عبارة «اكسب المال بسهولة»، فمن المرجح أن تكون هذه الرسالة رسالة بريد عشوائي.» والتعرف على هذه الأنواع من قواعد التصنيف يُعرف باسم «التنبؤ». وربما تبدو كلمة «تنبؤ» اختيارًا غريبًا لأن القاعدة لا تتنبأ بما سيحدُث في المستقبل: فرسالة البريد الإلكتروني إما أن تكون عشوائية أو غير عشوائية. ولذا، فمن الأفضل التفكير في أنماط التنبُّؤ على أنها تتنبَّأ بالقيمة المجهولة لسِمةٍ مُعينة بدلًا من أن نظن أنها تتنبأ بالمستقبل. في هذا المثال، نحن نتنبأ بما إذا كانت سمةُ تصنيف البريد الإلكتروني ينبغي أن تأخذ قيمة «بريد عشوائي» أم لا.

> إذا كان بإمكان أحد الخبراء أن يبتكِر نمطًا في ذهنه بسهولة، فإن هذا النمط عمومًا لا يستحقُّ الوقت والجهد اللازمَين لاستخدام علم البيانات من أجل «اكتشافه».

على الرغم من أنه بإمكاننا الاستعانة بعلم البيانات لاستخلاص شتَّى أنواع الأنماط، فإننا نرغب دومًا أن تكون هذه الأنماط غير واضحة ومفيدة على حدٍّ سواء. والمثال الذي ذكرناه في الفقرة السابقة عن قاعدة تصنيف رسائل البريد الإلكتروني هو مثالٌ بسيط وواضح جدًّا إلى حدٍّ أنه لو كانت تلك هي القاعدة الوحيدة المستخلصة من العمليات الخاصة بعلم البيانات، لأُصِبنا بخيبة الأمل والإحباط. على سبيل المثال، تُراجِع هذه القاعدة الخاصة بتصنيف رسائل البريد الإلكتروني سمةً واحدةً فقط خاصة بالبريد الإلكتروني؛ ألا وهي: هل تحتوي الرسالة على عبارة «اكسب المال بسهولة»؟ إذا كان بإمكان أحد الخبراء أن يبتكِر نمطًا في ذهنه بسهولة، فإن هذا النمط عمومًا لا يستحقُّ الوقت والجهد اللازمَين لاستخدام علم البيانات من أجل «اكتشافه». فبصفةٍ عامة، يصير علم البيانات مفيدًا عندما يكون لدينا عدد كبير من أمثلة البيانات وعندما تكون الأنماط بالِغة التعقيد بحيث يعجز البشر عن اكتشافها واستخلاصها يدويًّا. وفيما يخصُّ الحد الأدنى، يُمكننا تحديد عددٍ كبير من أمثلة البيانات على نحوٍ يفوق قدرة الخبراء على التحقُّق منه بسهولة. أما فيما

يخصُّ تعقيد الأنماط، فأُكرر أنه يمكن تحديدها في ضوء القدرات البشرية. فنحن — البشر — نُجيد بدرجة معقولة تحديد القواعد التي تتحقَّق من سِمةٍ أو سِمتَين أو ثلاثِ سمات (يُطلَق عليها في بعض الأحيان «خصائص» أو «متغيرات»)، ولكن عندما تزيد على ثلاث سمات، فقد تبدأ معاناتنا للتعامل مع التفاعلات فيما بينها. وعلى النقيض من ذلك، عادةً ما يُطبَّق علم البيانات في سياقاتٍ حيث نرغب في البحث عن أنماطٍ بين عشرات ومئات وآلاف السمات، بل وتصل إلى ملايين السمات في الحالات القصوى.

ولا تكون الأنماط التي نستنبطها باستخدام علم البيانات ذات فائدة إلا إذا وفَّرت لنا رؤيةً مستنيرة عن المشكلة بحيث تُمكننا من القيام بشيءٍ ما يُساعدنا في حل هذه المشكلة. وأحيانًا تُستخدَم عبارة «رؤية مستنيرة قابلة للتنفيذ» في هذا السياق لوصف ما نرغب أن تُوفره لنا الأنماط المستخرَجة. ويسلط مصطلح «رؤية مستنيرة» الضوء على النمط الذي ينبغي أن يُوفِّر معلومات ذات صلةٍ حول المشكلة غير الواضحة. ويُبرز مصطلح «قابلة للتنفيذ» أن الرؤية المستنيرة التي نحصل عليها ينبغي أن تكون شيئًا نتمتع بالقدرة على استغلاله بشكلٍ أو بآخر. على سبيل المثال، تَخيَّل أننا نعمل لدى شركة هواتف محمولة تحاول حل مشكلة «تسرب العملاء»؛ أي انتقال عددٍ كبير جدًّا من العملاء إلى شركات أخرى. وإحدى الطرق التي ربما يستعان بها للتعامل مع هذه المشكلة هي استخراج أنماط من البيانات المتوفرة عن العملاء السابقين تُتيح لنا تحديد العملاء الحالِيِّين المعرضين لخطر تسرُّبهم؛ ثم التواصل مع هؤلاء العملاء ومحاولة إقناعهم بالاستمرار مع شركتنا. ولا يكون النمط الذي يُمكِّننا من تحديد العملاء المحتَمل تسرُّبهم ذا فائدة بالنسبة إلينا إلا (أ) إذا كانت الأنماط تُحدد العملاء في وقتٍ مُبكر بما يكفي بحيث يكون لدينا الوقت الكافي للتواصل معهم قبل خسارتهم (ب) وإذا كانت شركتنا قادرة على تعيين فريقٍ للتواصل معهم. وهاتان الخطوتان ضروريَّتان لكي تكون الشركة قادرة على التصرف بناءً على الرؤية المستنيرة التي تُمِدُّنا بها الأنماط.

تاريخ موجز لعلم البيانات

يعود تاريخ ظهور مصطلح «علم البيانات» إلى تسعينيات القرن الماضي. إلا أن المجالات التي يعوِّل عليها هذا العلم لها تاريخ أطول من ذلك بكثير. أحد الخيوط في هذا التاريخ الأطول هو تاريخ جمع البيانات؛ والآخر هو تاريخ تحليل البيانات. في هذا القسم، نتناول التطورات الرئيسية في هذَين الخطَّين ونصِف مدى تقاربهما من مجال علم البيانات

والسبب وراء هذا التقارب. وبحكم الضرورة، يقدم هذا التناول مصطلحاتٍ جديدة أثناء وصفنا وذِكرنا للابتكارات التكنولوجية المهمة عند ظهورها. ونقدم شرحًا موجزًا لمعنى كلِّ مصطلح جديد؛ ونُعاود التطرق إلى الكثير من هذه المصطلحات في مواضع لاحقة من هذا الكتاب ونقدم تفسيرًا مفصلًا لها. سنبدأ بتاريخ جمع البيانات، ثم نستعرض تاريخ تحليل البيانات، وأخيرًا، سنتناول التطور المحرَز على صعيد علم البيانات.

تاريخ جمع البيانات

ربما يتمثل أقدم أساليب تسجيل البيانات في الثلمات المحفورة على العِصي بهدف تسجيل مرور الأيام أو الأعمدة المغروسة في الأرض لتسجيل مواقيت شروق الشمس عند حدوث الانقلاب الشمسي صيفًا وشتاءً. بيد أنه مع تطور الكتابة، زادت قدرتنا على تسجيل تجاربنا والأحداث في عالمنا من كمية البيانات التي نجمعها تزايدًا مهولًا. تطوَّر أقدم شكلٍ للكتابة في بلاد الرافدَين نحو عام ٣٢٠٠ قبل الميلاد واستُخدِم لحفظ السجلات التجارية. يلفت هذا النوع من حفظ السجلَّات الانتباه إلى ما يُعرف باسم «بيانات المعاملات التجارية». تشمل بيانات المعاملات التجارية معلوماتٍ عن حدثٍ ما مثل مبيعات خاصة بأحد الأصناف، وإصدار الفاتورة، وتسليم البضائع، والدفع ببطاقة الائتمان، والمطالبات التأمينية، وهلمَّ جرًّا. وتحظى «بيانات المعاملات غير التجارية» — مثل البيانات الديموغرافية — بتاريخٍ طويل أيضًا. إذ يرجع تاريخ أقدم إحصاء سُكاني معروف في مصر الفرعونية إلى نحو عام ٣٠٠٠ قبل الميلاد. كان السبب وراء بذْل الدول المبكرة جهدًا كبيرًا جدًّا وتسخير موارد كثيرة لعمليات جمع بيانات كبيرة هو أن هذه الدول كانت بحاجةٍ إلى زيادة الضرائب وحشد الجيوش، ممَّا يؤكد مقولة بنجامين فرانكلين الزاعمة بأن ثمة حقيقتَين فقط لا يختلف عليهما أحد في هذه الحياة؛ ألا وهما الموت والضرائب.

خلال المائة والخمسين عامًا الماضية، ساهم تطوير أجهزة الاستشعار الإلكترونية، ورقمنة البيانات، واختراع الكمبيوتر في زيادة كمية البيانات التي تُجمع وتُخزَّن زيادةً مهولةً. وكان عام ١٩٧٠ علامةً فارقة في جمع البيانات وتخزينها حين نشر «إدجار إف كود» بحثًا يشرح فيه «نموذج البيانات الارتباطية»، الذي كان في حدِّ ذاته نموذجًا ثوريًّا فيما يخصُّ تحديد كيفية تخزين البيانات (آنذاك) وفهرستها واستعادتها من قواعد البيانات. مكَّن نموذج البيانات الارتباطية المستخدِمين من استخراج البيانات من قاعدة البيانات باستخدام استعلاماتٍ بسيطة تُحدد البيانات التي يريدها المستخدِم دون إثارة القلق لدَيه

حيال الهيكل الأساسي الخاص بالبيانات أو المكان الذي خُزِّنت فيه فعليًّا. وضع بحث «كود» حجر الأساس لقواعد البيانات الحديثة وتطوير «لغة الاستعلام الهيكلية» (إس كيو إل)، وهي معيار دولي لتحديد استعلامات قواعد البيانات. تخزن قواعد البيانات الارتباطية البيانات في جداول بِبِنية تتكوَّن من صفٍّ واحد لكل مثيلٍ وعمودٍ واحد لكلِّ سمة. وهذه البِنية مثالية لتخزين البيانات لأنه من الممكن تفكيكها إلى سماتٍ بسيطة.

وتُعد قواعد البيانات هي التقنية البسيطة المستخدَمة لتخزين بيانات المعاملات التجارية أو البيانات «التشغيلية» الهيكلية (أي نوعية البيانات التي تُولِّدها العمليات التشغيلية اليومية الخاصة بمؤسسةٍ ما). ومع ذلك، نظرًا إلى أن الشركات صارت أكبر حجمًا وأكثر اعتمادًا على الأجهزة والآلات، زادت كمية البيانات التي تُنتجها الأقسام المختلفة في هذه الشركات ومدى تنوُّعها زيادة مهولة. وفي تسعينيات القرن العشرين، أدركت الشركات أنه على الرغم من أنها جمعت كمياتٍ هائلة من البيانات، فإنها واجهت صعوباتٍ مُتكررة حيال تحليل تلك البيانات. تَمثَّل جزء من المشكلة في أن البيانات كانت تُخزن عادةً في عددٍ كبير من قواعد البيانات المنفصلة بعضها عن بعض داخل الشركة الواحدة. وتمثلت صعوبة أخرى في أن قواعد البيانات كان يُحسَّن أداؤها من أجل تخزين البيانات واستعادتها، وهي الأنشطة التي تتميَّز بأعدادٍ كبيرة من العمليات البسيطة مثل «اختيار» و«إدراج» و«تحديث» و«حذف». ومن أجل تحليل بياناتها، كانت هذه الشركات بحاجةٍ إلى تقنية قادرة على تجميع البيانات والتوفيق بينها من قواعد بيانات مختلفة وهذا يَسَّر عمليات البيانات التحليلية الأكثر تعقيدًا. وقد أدى هذا التحدي إلى تطوير «مستودعات البيانات». في هذا المستودع، تُجمع البيانات من كل أقسام الشركة وتُدمج، وبالتالي تتيح للتحليل مجموعة بياناتٍ أكثر شمولًا.

وعلى مدار العقدَين الماضيَين، صارت أجهزتنا محمولةً ومتصلةً بالشبكات، ويقضي الكثيرون منَّا ساعاتٍ طويلة على شبكة الإنترنت كل يومٍ من خلال استخدام تقنيات التواصل الاجتماعي، وألعاب الكمبيوتر، والمنصَّات الإعلامية، ومحركات البحث عبر الإنترنت. وهذه التغيرات الطارئة على التكنولوجيا والطريقة التي نعيش بها لها تأثير كبير على كمية البيانات التي جُمعت. إذ تُقدر كمية البيانات التي جُمعت على مدار خمسة آلاف عامٍ منذ اختراع الكتابة وحتى عام ٢٠٠٣ بنحو ٥ إكسابايت. ومنذ عام ٢٠١٣، يُولِّد البشر هذه الكمية نفسها من البيانات «كل يوم» ويخزنونها. ومع ذلك، لم تكن كمية البيانات المجمَّعة وحدَها هي ما زاد زيادةً مهولة وإنما زاد تنوُّعها أيضًا. فقط تأمَّل

في القائمة التالية من مصادر البيانات عبر الإنترنت: رسائل البريد الإلكتروني والمدوَّنات والصور والتغريدات والإعجاب بالمنشورات والمشاركات وعمليات البحث عبر الويب وتحميل الفيديوهات وعمليات الشراء عبر الإنترنت والبودكاست. وإذا وضعنا في الاعتبار بيانات التعريف (البيانات التي تصِف بِنية البيانات الأصلية وخصائصها) لهذه الأحداث، استطعنا فهْم معنى مصطلح «البيانات الضخمة». وعادةً ما تُعرَّف البيانات الضخمة في ضوء ثلاثة عناصر: «الحجم» الضخم للبيانات، و«تنوع» نوعيَّات البيانات، و«السرعة» التي يجب أن تُعالج بها البيانات.

لقد شجع ظهور البيانات الضخمة تطور مجموعة من التقنيات الجديدة لقواعد البيانات. وكثيرًا ما يُشار إلى هذا الجيل الجديد من قواعد البيانات باسم «قواعد البيانات غير الارتباطية» (وتُعرف اختصارًا بـ NoSQL). وعادةً ما يكون لها نموذج بيانات أبسط من قواعد البيانات الارتباطية التقليدية. وتُخزِّن قاعدة البيانات غير الارتباطية البيانات على هيئة كائناتٍ ذات سِمات، باستخدام لغة ترميز كائنات مثل «جافا سكريبت أوبجكت نوتيشن» (أو جيه إس أو إن). وتكمن ميزة تمثيل البيانات على هيئة كائنات (على النقيض من النموذج القائم على الجداول الارتباطية) في أن مجموعة السِّمات الخاصة بكل كائن مُتضمنة داخله، مما يسفر عن تمثيلٍ مرِن. على سبيل المثال، ربما يحظى أحد الكائنات في قاعدة البيانات بمجموعةٍ فرعية فقط من السمات، مقارنة بالكائنات الأخرى. وعلى النقيض من ذلك، في هيكل البيانات القياسي المجدوَل والمستخدَم في قواعد البيانات الارتباطية، ينبغي أن تتمتع نقاط البيانات بالمجموعة نفسها من السمات (أي الأعمدة). وهذه المرونة في تمثيل البيانات على هيئة كائنات ذات أهمية في السياقات حيث لا يمكن تحليل البيانات إلى مجموعةٍ من السمات الهيكلية (هذا بسبب التنوُّع أو النوع). على سبيل المثال، قد يكون من الصعب تحديد مجموعة السمات التي ينبغي استخدامها لتمثيل النصِّ الحُر (مثل التغريدات) أو الصور. ومع ذلك، على الرغم من أن هذه المرونة التمثيلية تُتيح لنا تدوين البيانات وتخزينها في تنسيقاتٍ متنوعة، يجب استخراج هذه البيانات على هيئة تنسيق هيكلي قبل إجراء أي تحليل عليها.

لقد أدى ظهور البيانات الضخمة أيضًا إلى تطوير أُطرٍ جديدة لمعالجة البيانات. فعندما تتعامل مع كميات كبيرة من البيانات بسرعات عاليةٍ، قد يفيد — من المنظور الحوسبي ومن منظور السرعة — توزيع البيانات عبر وحدات خدمةٍ مُتعددة، ومعالجة الاستعلامات من خلال حساب النتائج الجزئية الخاصة بالاستعلام على كلٍّ وحدة خدمة،

ثم دمج هذه النتائج لتوليد الردِّ على هذا الاستعلام. وهذا هو النهج المتَّبع في إطار عمل «ماب رديوس» على منصة هادوب. وفي هذا الإطار، تُعيَّن البيانات والاستعلامات (أو تُوزَّع) عبر عدة وحدات خدمة، وتُحسَب النتائج الجزئية على كل وحدة خدمة، ثم تُختزل معًا (أو تُدمج معًا).

تاريخ تحليل البيانات

علم الإحصاء هو فرع من العلوم التي تتعامل مع جمع البيانات وتحليلها. ويشير مصطلح «الإحصاء» بالأساس إلى جمع بياناتٍ عن الدولة وتحليلها؛ مثل البيانات الديموغرافية أو البيانات الاقتصادية. إلا أنه مع مرور الوقت، توسَّعت نوعية البيانات التي يُستخدم فيها التحليل الإحصائي بحيث تُستخدم الإحصاءات اليوم لتحليل جميع أنواع البيانات. وأبسط شكلٍ للتحليل الإحصائي للبيانات هو تلخيص مجموعةٍ من البيانات على هيئة «إحصاءات موجزة (وصفية)» (من بينها مقاييس النزعة المركزية، مثل «الوسط الحسابي»، أو مقاييس التبايُن، مثل «المدى»). ومع ذلك، في القرنَين السابع عشر والثامن عشر، أرست أعمال أشخاصٍ مثل جيرولامو كاردانو، وبليز باسكال، وياكوب برنولي، وأبراهام دي موافر، وتوماس بايز، وريتشارد برايس أُسس نظرية الاحتمال، وخلال القرن التاسع عشر، بدأ الكثير من الإحصائيين استخدام التوزيعات الاحتمالية كأداةٍ ضمن مجموعة أدواتهم التحليلية. مكَّنَت هذه التطورات الجديدة في الرياضيات الإحصائيين من تخطِّي الإحصاءات الوصفية وبدء العمل على «التعلم الإحصائي». ويُعَد بيير سيمون دي لابلاس وكارل فريدريش جاوس اثنَين من أهم وأشهر علماء الرياضيات في القرن التاسع عشر، كلٌّ منهما قدَّم إسهاماتٍ مهمة في مجال التعلُّم الإحصائي وعلم البيانات الحديث. أخذ لابلاس أفكار توماس بايز وريتشارد برايس وطورها لتُصبح النسخة الأولى لما نُسمِّيه الآن بـ «قاعدة بايز». وطوَّر جاوس، أثناء بحثه عن الكوكب القزم المفقود سيريس، «طريقة المربَّعات الصغرى»، التي مكَّنتنا من التوصُّل إلى أفضل نموذج يلائم مجموعة البيانات بحيث يُقلل الخطأ في الملاءمة إجمالي الفروق المربعة بين نقاط البيانات في مجموعة البيانات والنموذج إلى الحد الأدنى. وفَّرت طريقة المربعات الصغرى الأساس لأساليب التعلُّم الإحصائي مثل «الانحدار الخطي» و«الانحدار اللوجستي» بالإضافة إلى تطوير نماذج «الشبكة العصبية الاصطناعية» المستخدَمة في الذكاء الاصطناعي (سنعاود التطرُّق إلى المربعات الصغرى، وتحليل الانحدار، والشبكات العصبية في الفصل الرابع).

وما بين عامَي ١٧٨٠ و١٨٢٠، في التوقيت نفسه تقريبًا الذي قدم فيه لابلاس وجاوس إسهاماتهما إلى التعلم الإحصائي، اخترع مهندس اسكتلندي يُدعى ويليام بلايفير المخططات الإحصائية وأرسى أُسس «التمثيل المرئي للبيانات» و«التحليل الاستكشافي للبيانات». ابتكر بلايفير «المخطَّط الخطِّي» و«المخطَّط المساحي» من أجل البيانات المسلسلة زمنيًّا، و«المخطَّط العمودي» لتوضيح المقارنات بين كميات الفئات المختلفة، و«المخطَّط الدائري» لتوضيح النِّسَب داخل مجموعة. ويهدف التمثيل المرئي للبيانات الكمية إلى السماح لنا باستغلال قدراتنا البصرية القوية من أجل تلخيص البيانات ومقارنتها وتفسيرها. ورغم أنه يصعُب تمثيل مجموعات البيانات الكبيرة (التي تحتوي على الكثير من نقاط البيانات) والمعقدة (التي تحتوي على الكثير من السِّمات) بشكلٍ مرئي، فإن التمثيل المرئي للبيانات لا يزال يُمثل جزءًا مهمًّا من علم البيانات. ويُعَد هذا التمثيل، على وجه التحديد، ذا فائدةٍ في مساعدة علماء البيانات في استكشاف وفَهم البيانات التي يتعاملون معها. ويمكن أن تكون التمثيلات المرئية مفيدة أيضًا في إيضاح نتائج أحد مشروعات علم البيانات. ومنذ عصر بلايفير، ازدادت مخطَّطات تمثيل البيانات زيادةً مُطردة، واليوم ثمة أبحاث متواصِلة من أجل تطوير مناهج جديدة لتمثيل مجموعات البيانات الكبيرة والمتعددة الأبعاد تمثيلًا مرئيًّا. ويتمثل أحد التطورات الحديثة في خوارزمية «تضمين الجوار العشوائي الموزع على شكل حرف T» (تي-إس إن إي)، وهي عبارة عن تقنية مفيدة لاختزال البيانات المتعددة الأبعاد إلى بُعدَين أو ثلاثة، وبالتالي تيسير التمثيل المرئي لتلك البيانات.

استمرت التطورات في نظرية الاحتمالات والإحصاء حتى القرن العشرين. إذ طوَّر كارل بيرسون اختبار الفرضية الحديث، وطور آر إيه فيشر أساليب إحصائية من أجل «التحليل المتعدد المتغيِّرات» وقدَّم فكرة «تقدير الاحتمال الأرجح» في الاستدلال الإحصائي كوسيلةٍ لاستخلاص النتائج بناءً على الاحتمالية النسبية للأحداث. وأدى عمل آلان تورينج في الحرب العالمية الثانية إلى اختراع الكمبيوتر الإلكتروني الذي كان له أثر عظيم على الإحصاء لأنه مَكَّننا من إجراء حساباتٍ إحصائية شديدة التعقيد. وخلال أربعينيات القرن العشرين والعقود التالية، طوِّر عدد من النماذج الحوسبية المهمة التي لا تزال مستخدَمة على نطاقٍ واسع في علم البيانات. وفي عام ١٩٤٣، اقترح وارن ماكولوتش ووالتر بيتس النموذج الرياضي الأول «للشبكة العصبية». وفي عام ١٩٤٨، نشر كلود شانون مقالًا بعنوان «نظرية رياضية للتواصُل»، ووضع من خلاله أساسًا لـ«نظرية المعلومات». وفي عام ١٩٥١، اقترحت

إفيلين فيكس وجوزيف هودجز نموذجًا لـ «التحليل التمييزي» (أو ما نُطلق عليه الآن مسألة «التصنيف» أو «التعرُّف على الأنماط») الذي صار أساس «نماذج أقرب الجيران» الحديثة. وبلغت هذه التطوُّرات في فترة ما بعد الحرب العالمية ذروتها في عام ١٩٥٦ مع تأسيس مجال «الذكاء الاصطناعي» في ورشة عمل بكلية دارتموث. وحتى في هذه المرحلة المبكرة من تطوير الذكاء الاصطناعي، كان قد بدأ استخدام مصطلح «تعلُّم الآلة» لوصف البرامج التي مكَّنت الكمبيوتر من التعلُّم من البيانات. وفي منتصف ستينيَّات القرن العشرين، قُدِّمت ثلاثة إسهامات مُهمة لتعلُّم الآلة. ففي عام ١٩٦٥، أوضح كتاب نيلس نيلسون بعنوان «الآلات المتعلمة» كيف يمكن استخدام الشبكات العصبية لتعلُّم النماذج الخطِّية للتصنيف. وفي العام التالي، تحديدًا في عام ١٩٦٦، طور إيرل بي هانت وجانت مارين وفيليب جيه ستون إطار نظام تعلُّم المفاهيم، الذي مثَّل الأصل الذي تنحدِر منه عائلة مهمة لخوارزميات تعلُّم الآلة التي حفَّزت ظهور نماذج شجرة اتخاذ القرار من البيانات من أعلى إلى أسفل. وفي التوقيت نفسه تقريبًا، طوَّر عدد من الباحِثين المستقلِّين النُّسخ الأولية من خوارزميات «التجميع بالمتوسطات»، التي صارت الآن الخوارزمية القياسية المستخدَمة لتجزئة البيانات (العملاء).

يُعَد تعلُّم الآلة مجالًا جوهريًّا في علم البيانات الحديث؛ ذلك لأنه يوفر الخوارزميات القادرة على تحليل مجموعات البيانات الكبيرة تحليلًا آليًّا لاستخلاص الأنماط التي من المحتمل أن تكون جاذبةً للاهتمام ومفيدةً على حدٍّ سواء. ولقد واصل هذا المجال التطوُّر والابتكار حتى يومِنا هذا. وتشمل بعض أهم التطوُّرات «النماذج التجميعية» — حيث تُجرى التنبؤات باستخدام مجموعةٍ من النماذج (أو فئة من النماذج)، ويتنبأ كلُّ نموذج بكلِّ استعلامٍ من خلال الاقتراع — و«الشبكات العصبية الخاصة بالتعلم العميق»، التي تتكوَّن من طبقات عديدة (أكثر من ثلاث طبقات) من الخلايا العصبية. وهذه الطبقات الأعمق في الشبكة قادرة على اكتشاف وتعلم تمثيلات السِّمات المعقدة (التي تتألَّف من عدة سماتٍ تفاعلية مُدخَلة جرت معالجتها بواسطة طبقاتٍ أولى)، التي تُمكِّن الشبكة بدورها من تعلُّم أنماطٍ يمكن تعميمها عبر البيانات المدخلة. ونظرًا إلى قُدرتها على تعلُّم السمات المعقدة، تتناسَب شبكات التعلم العميق على وجه الخصوص مع البيانات كثيرة الأبعاد، وبالتالي أحدثت ثورة في عدة ميادين، من بينها «رؤية الآلة» و«معالجة اللغة الطبيعية».

كما ناقشنا في معرض حديثنا عن تاريخ قواعد البيانات، شهدت أوائل السبعينيات من القرن الماضي بدايةَ تقنيةِ قواعد البيانات الحديثة مع نموذج البيانات الارتباطية الذي وضعه «إدجار إف كود» وما تبِعه من زيادة هائلة في توليد البيانات وتخزينها مما أدى إلى تطوير مستودعات البيانات في التسعينيات ولاحقًا إلى ظاهرة البيانات الضخمة. إلا أنه قبل ظهور البيانات الضخمة، وتحديدًا بحلول أواخر الثمانينيات وأوائل التسعينيات من القرن العشرين، ظهرت الحاجة إلى مجالٍ بحثي يستهدف على وجه التحديد تحليلَ هذه المجموعات الكبيرة من البيانات. وفي هذا الوقت تقريبًا بدأ استخدام مصطلح «التنقيب في البيانات» في الأوساط المستخدِمة لقواعد البيانات. وكما ناقشنا بالفعل، تمثلت إحدى الاستجابات لهذه الحاجة في تطوير مستودعات البيانات. ومع ذلك، استجاب باحثون آخرون في قواعد البيانات بالتطرُّق إلى مجالاتٍ بحثية أخرى، وفي عام ١٩٨٩، عقد جريجوري بياتيتسكي–شابيرو أول ورشة عمل عن «اكتشاف المعرفة في قواعد البيانات». ويلخص الإعلان عن هذه الورشة كيف أن الورشة ركَّزت على منهجٍ مُتعدد التخصُّصات لحل مشكلة تحليل قواعد البيانات الكبيرة؛ إذ جاء الإعلان كما يلي:

> يُثير اكتشاف المعرفة في قواعد البيانات الكثير من المسائل المهمة، خاصةً عندما تكون قواعد البيانات كبيرة الحجم. وغالبًا ما تكون هذه القواعد مصحوبةً بقدْر كبير من المعرفة بالمجال مما يُسهل عملية الاكتشاف كثيرًا. والوصول إلى قاعدة بيانات كبيرة هو أمر مكلف؛ وهنا تأتي الحاجة إلى أخذ عيناتٍ واتِّباع الأساليب الإحصائية الأخرى. وأخيرًا، يمكن أن تستفيد عملية اكتشاف المعرفة في قواعد البيانات من الكثير من الأدوات والتقنيات المتاحة من عدة مجالات مختلفة من بينها النظم الخبيرة وتعلُّم الآلة وقواعد البيانات الذكية واكتساب المعرفة والإحصاء.[1]

في الواقع، يصف المصطلحان «اكتشاف المعرفة في قواعد البيانات» و«التنقيب في البيانات» المفهوم نفسه؛ الفارق هو أن التنقيب في البيانات أكثر انتشارًا في أوساط الأعمال التجارية، أما مصطلح اكتشاف المعرفة في قواعد البيانات فهو أكثر انتشارًا في الأوساط الأكاديمية. اليوم، يُستخدم هذان المصطلحان على نحوٍ متبادل،[2] والكثير من الأماكن الأكاديمية رفيعة المستوى تستخدِم كِلا المصطلحَين. وبالطبع، يأتي المؤتمر الدولي بشأن اكتشاف المعرفة والتنقيب في البيانات على رأس أقدم المؤتمرات الأكاديمية في المجال.

ظهور علم البيانات وتطوُّره

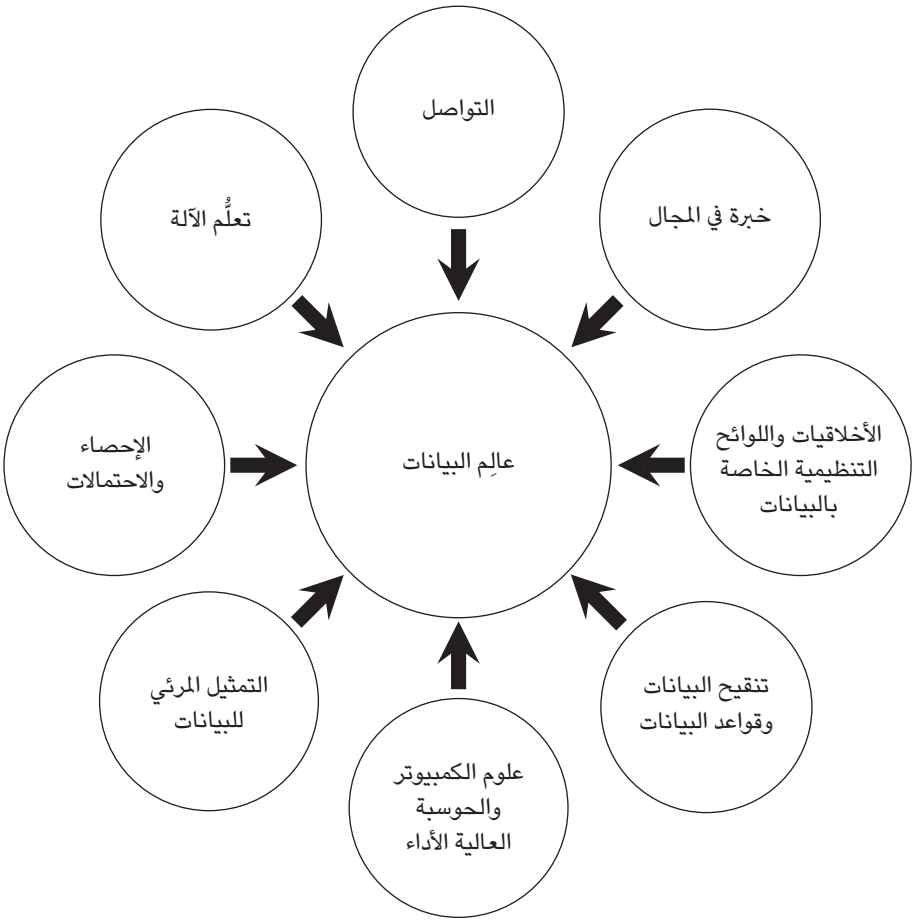
ظهر مصطلح «علم البيانات» على الساحة في أواخر تسعينيات القرن العشرين في نقاشات ذات صلة بالحاجة إلى تعاون الإحصائيين مع علماء الكمبيوتر لإدخال عنصر الدقة الرياضية إلى التحليل الحوسبي لمجموعات البيانات الكبيرة. وفي عام ١٩٩٧، سلطت المحاضرة العامة التي ألقاها «سي إف جيف وو» بعنوان: «هل يتساوى علم الإحصاء بعلم البيانات؟» الضوء على عدد من الاتجاهات الواعدة للإحصاء، من بينها توفُّر مجموعات البيانات الكبيرة/المعقدة في قواعد بيانات مهولة والاستخدام المتزايد للخوارزميات والنماذج الحوسبية. واختتمت المحاضرة بالدعوة إلى إعادة تسمية علم الإحصاء بـ «علم البيانات».

وفي عام ٢٠٠١، نشر ويليام إس كليفلاند خطة عملٍ لإنشاء قسمٍ في الجامعة مُتخصص في مجال علم البيانات (Cleveland 2001). وتؤكد الخطة على ضرورة أن يكون علم البيانات شراكةً بين الرياضيات وعلوم الكمبيوتر. كما أنها تؤكد على ضرورة فهم علم البيانات باعتباره مسعًى مُتعدِّد التخصُّصات وعلى أن يتعلم علماء البيانات كيفية العمل والتعاون مع الخبراء من المجالات المختلفة. وفي العام نفسه، نشر ليو بريمان بحثًا بعنوان «النمذجة الإحصائية: الثقافتان» (٢٠٠١). في هذا البحث، يصف بريمان النهج التقليدي في الإحصاء بأنه ثقافة نمذجة البيانات التي ترى أن الهدف الرئيسي من تحليل البيانات هو تحديد نموذج البيانات العشوائي (الخفي) (على سبيل المثال، «الانحدار الخطي») الذي يفسر كيف جرى توليد البيانات. وقارن هذه الثقافة بثقافة النمذجة الخوارزمية التي تركز على استخدام الخوارزميات الحاسوبية لابتكار نماذج تنبؤية تتَّسِم بالدقة (بدلًا من التفسير فيما يتعلق بكيفية توليد البيانات). إن تمييز بريمان بين تركيز علم الإحصاء على النماذج التي تُفسر البيانات وتركيز الخوارزميات على النماذج التي يمكن أن تتنبأ بدقة بالبيانات يُسلط الضوء على الفارق الرئيسي بين الإحصائيين والباحثين في مجال تعلُّم الآلة. ولا يزال الجدال قائمًا بين هذَين المنهجَين داخل أوساط الإحصائيين (انظر، على سبيل المثال، Shmueli 2010). وبوجهٍ عام، معظم مشروعات علم البيانات اليوم أكثر توافقًا مع منهج تعلُّم الآلة الذي يحرص على إنشاء نماذج تنبُّؤية دقيقة وأقل اهتمامًا بالتركيز الإحصائي على تفسير البيانات. وعلى الرغم من أن علم البيانات لعب دورًا بارزًا في المناقشات المتعلقة بالإحصاء ولا يزال يستعير أساليب ونماذج من علم الإحصاء، فإنه مع مرور الوقت طوَّر منهجه المميز الخاص لتحليل البيانات.

منذ عام ٢٠٠١، اتسع مفهوم علم البيانات بشكلٍ كبير ليتجاوز كونه مجرد إعادة تعريف للإحصاء. على سبيل المثال، على مدار السنوات العشر الأخيرة، كان هناك تزايد مهول في كمية البيانات المتولِّدة من الأنشطة التي تتمُّ ممارستها عبر الإنترنت (البيع بالتجزئة عبر الإنترنت، ووسائل التواصل الاجتماعي والترفيه عبر الإنترنت). لقد أسفر جمع هذه البيانات وتجهيزها لاستخدامها في مشروعات علم البيانات عن حاجة علماء البيانات لتطوير مهارات البرمجة والقرصنة لاستخراج البيانات (أحيانًا البيانات غير الهيكلية) ودمجها وتصفيتها من مصادر الويب الخارجية. كما أن ظهور البيانات الضخمة أدَّى إلى جعل علماء البيانات مُضطرِّين إلى التعامل مع تقنيات البيانات الضخمة، مثل هادوب. في الواقع، اليوم صار دور علماء البيانات موسعًا للغاية لدرجة أنه ثمة جدلٌ مُستمر حول كيفية تحديد الخبرات والمهارات اللازمة لتنفيذ هذا الدور.[3] غير أنه من الممكن سرد قائمة بالخبرات والمهارات التي قد يتفق معظم الناس على كونها ذات صلةٍ بهذا الدور، والمبينة في شكل ١-١. ومن الصعب على فردٍ واحد إتقان كل هذه المجالات، وبالطبع، يتمتع أغلب علماء البيانات بمعرفةٍ مُتعمقة وخبرة حقيقية في مجموعةٍ فرعية منها فحسب. ومع ذلك، من المهم أن نفهم ونعي إسهام كلِّ مجالٍ من هذه المجالات في مشروع علم البيانات.

يجب أن يتمتع علماء البيانات بقدْر من الخبرة في المجال. تبدأ أغلب مشروعات علم البيانات بمشكلة من الواقع الفعلي مُختصة بمجالٍ مُعين والحاجة إلى تصميم حلٍّ مستخلص من البيانات لهذه المشكلة. وكنتيجة لذلك، من المهم لعالِم البيانات أن يتمتَّع بخبرةٍ كافية في هذا المجال تُتيح له فهم المشكلة، والوقوف على سبب أهميتها، وإلى أي مدًى قد يتناسب حل المشكلة القائم على علم البيانات مع عمليات المؤسسة. وهذه الخبرة في المجال تقود اختصاصِي علم البيانات أثناء عمله نحو تحديد الحل الأمثل. كما إنها تتيح له التفاعل مع خبراء المجال الحقيقيين بطريقةٍ ملموسة لكي يتسنى له جمع المعرفة اللازمة عن المشكلة الأساسية. كما أن التحلي بقدْر من الخبرة في مجال المشروع يُتيح لعالِم البيانات الاستعانة بخبراته المكتسبة من العمل على مشروعات مشابهة في نفس المجال والمجالات ذات الصِّلة لتُساعده على تحديد نطاق تركيز المشروع.

البيانات هي محور جميع مشروعات علم البيانات. ومع ذلك، لا تعني حقيقة أن المؤسسة لها حق الوصول إلى البيانات أنه يُمكنها استغلال البيانات من الناحية القانونية أو حتى ينبغي لها ذلك من الناحية الأخلاقية. ففي أغلب الدوائر القضائية، ثمة تشريعات مناهِضة للتمييز ومؤازِرة لحماية البيانات الشخصية تُنظم عملية استخدام البيانات



شكل ١-١: مجموعة المهارات اللازمة لعالِم البيانات.

وتتحكَّم فيها. وكنتيجةٍ لذلك، يجب على عالِم البيانات أن يتفهَّم هذه التشريعات، وعلى النطاق الأوسع، يجب أيضًا أن يتمتع بفهمٍ أخلاقي لتداعيات عمله إذا كان يرغب في استخدام البيانات استخدامًا قانونيًّا ولائقًا. وسنتطرَّق إلى هذا الموضوع في الفصل السادس، حيث نُناقش اللوائح التنظيمية القانونية لاستغلال البيانات والمسائل الأخلاقية المتعلقة بعلم البيانات.

في أغلب المؤسسات، تأتي نسبة كبيرة من البيانات من قواعد البيانات الموجودة في المؤسسة. علاوة على ذلك، عند توسع هيكل البيانات الخاص بالمؤسسة، ستبدأ مشروعات علم البيانات دمج البيانات من مجموعة متنوعة من مصادر البيانات الأخرى، التي يُشار إليها عمومًا بـ «مصادر البيانات الضخمة». وقد تكون البيانات الموجودة في هذه المصادر في مجموعةٍ متنوعة من الأشكال المختلفة، أي قاعدة بيانات بشكلٍ أو آخر بصفة عامة مثل: قواعد البيانات الارتباطية أو قواعد البيانات غير الارتباطية أو هادوب. وجميع البيانات المتاحة في هذه القواعد المتنوعة ومصادر البيانات بحاجةٍ إلى دمجها وتنظيفها وتحويلها وتطبيعها، وهلمَّ جرًّا. ولهذه المهام أسماء كثيرة، مثل: «الاستخراج والتحويل والتحميل»، و«جمع البيانات»، و«تنقيح البيانات»، و«دمج البيانات»، وغيرها. وعلى غرار بيانات المصدر، البيانات المولَّدة من أنشطة علم البيانات بحاجة أيضًا إلى أن يتمَّ تخزينها وإدارتها. فقاعدة البيانات عبارة عن موقع التخزين النموذجي للبيانات المولَّدة بواسطة هذه الأنشطة لكي نتمكَّن من توزيعها بسهولةٍ ومشاركتها مع مختلف أقسام المؤسسة. ونتيجة لذلك، علماء البيانات بحاجةٍ إلى التحلِّي بالمهارات اللازمة للتفاعُل مع البيانات ومعالجتها ببراعةٍ في قواعد البيانات.

تُتيح مجموعة من مهارات علوم الكمبيوتر وأدواته لعلماء البيانات التعامل مع البيانات الضخمة ومعالجتها لتصير معلوماتٍ جديدة وذات مغزًى. وتشمل «الحوسبة العالية الأداء» حشد القدرة الحوسبية لتقديم أداءٍ أعلى مما كان مِن الممكن أن يحصل المرء عليه من جهاز كمبيوتر واحد مُستقل. الكثير من مشروعات علم البيانات تتعامل مع مجموعةٍ كبيرة جدًّا من البيانات وخوارزميات تعلُّم الآلة الباهظة التكلفة حوسبيًّا. وفي هذه المواقف، من المهم التحلِّي بالمهارات اللازمة للوصول إلى مصادر الحوسبة العالية الأداء واستخدامها. بخلاف الحوسبة العالية الأداء، لقد ذكرنا بالفعل أن علماء البيانات بحاجةٍ إلى التحلِّي بالقدرة على استخراج البيانات من مواقع الويب وتنظيفها ودمجها وكذلك التعامل مع النصوص والصور غير الهيكلية ومعالجتها. وبالإضافة إلى ذلك، قد ينتهي المطاف أيضًا بعالِم البيانات إلى إنشاء تطبيقاتٍ داخلية لأداء مهمةٍ مُعينة أو تعديل تطبيقٍ موجود بالفعل لمواءمتِهِ مع البيانات والمجال الخاضع للمعالجة. وأخيرًا، يحتاج عالِم البيانات لمهارات علوم الكمبيوتر لكي يتمكن من فهم نماذج تعلُّم الآلة وتطويرها ودمجها في تطبيقات الإنتاجية أو التطبيقات التحليلية أو التطبيقات الخلفية في إحدى المؤسسات.

إن تمثيل البيانات في صورة رسومية يجعل من الأسهل كثيرًا رؤية وفهم ما يحدُث لهذه البيانات. وينطبق التمثيل المرئي للبيانات على جميع مراحل عملية علم البيانات. فعند

مراجعة البيانات في شكل جدول، يكون من السهل إغفال أشياء مثل القِيَم الشاذة أو أنماط التوزيع أو التغيرات الطفيفة التي تطرأ على البيانات بمرور الوقت. أما حين تتُمثَّل البيانات بالصورة البيانية الصحيحة، فسوف تظهر هذه الجوانب من البيانات بوضوح. ويُعَد التمثيل المرئي للبيانات مجالًا مُهمًّا ومُتناميًا، ونرشح هنا كتابَين باعتبارهما تقديمًا ممتازًا لمبادئ وتقنيات التمثيل المرئي الفعَّال للبيانات؛ وهما: كتاب «العرض المرئي للمعلومات الكميَّة» تأليف إدوارد توفت (٢٠٠١) وكتاب «أَرِني الأرقام: توضيح تصميم الجداول والرسوم البيانية» تأليف ستيفن فيو (٢٠١٢).

تُستخدم أساليب الإحصاء والاحتمالات في جميع مراحل العملية الخاصة بعلم البيانات، بداية من تجميع البيانات والتحقُّق منها وصولًا إلى مقارنة نتائج النماذج والتحليلات المختلفة الصادرة أثناء المشروع. وينطوي تعلُّم الآلة على استخدام مجموعةٍ متنوعة من التقنيات الإحصائية والحوسبية المتقدمة لمعالجة البيانات بهدف إيجاد الأنماط. ولا يتعيَّن على عالِم البيانات الذي يُشارك في الجوانب التطبيقية لتعلُّم الآلة أن يُنشئ نُسَخه الخاصة من خوارزميات تعلُّم الآلة. فمن خلال فهم خوارزميات تعلُّم الآلة، وفيما يمكن استخدامها، وما تعنيه النتائج التي تولِّدها وما نوعية البيانات التي يمكن تشغيل خوارزميات مُعينة عليها، يستطيع عالِم البيانات أن يستفيد من خوارزميات تعلُّم الآلة حتى وإن كان لا يعرف التفاصيل الدقيقة لما تفعله الخوارزمية. وهذا يُتيح له التركيز على الجوانب التطبيقية لعلم البيانات وتجربة خوارزميات تعلُّم الآلة المتنوعة لمعرفة أيها يتناسب أكثر مع السيناريو الذي يتعامل معه والبيانات التي لديه.

أخيرًا، أحد الجوانب الرئيسية لنجاح عالِم البيانات هو التحلِّي بالقُدرة على توصيل نتائج مشروع علم البيانات. وقد توضح هذه النتائج الرؤية التي كشف عنها تحليل البيانات أو توضح مدى ملاءمة النماذج المنشأة أثناء المشروع لعمليات المؤسسة وتأثيرها المتوقَّع على آلية عمل المؤسسة. فلا جدوى من تنفيذ مشروع علم بيانات فذٍّ ما لم تُستخدَم المخرَجات منه وتوصَّل النتائج بطريقةٍ يمكن أن يفهمها الزملاء الذين لا يتمتَّعون بخلفيةٍ تقنية ويثقون بها.

أين يُستخدم علم البيانات؟

يقود علم البيانات اتخاذ القرارات في كافة جوانب المجتمعات الحديثة تقريبًا. في هذا القسم، تَصِف ثلاثُ دراساتٍ حالةً تُوضح تأثير علم البيانات؛ ألا وهي: شركات السلع

الاستهلاكية التي تستخدم علم البيانات من أجل المبيعات والتسويق؛ الحكومات التي تستخدم علم البيانات لتحسين الخدمات الصحية وأنظمة العدالة الجنائية والتخطيط العمراني؛ والرياضات الاحترافية التي تستخدم علم البيانات في استقطاب اللاعبين.

علم البيانات في مجال المبيعات والتسويق

تتمتع شركة وول مارت بإمكانية الوصول إلى مجموعات بياناتٍ كبيرة حول تفضيلات العملاء من خلال أنظمة نقاط البيع، وتتبُّع سلوك العميل على موقعها الإلكتروني، ومتابعة التعليقات حول الشركة ومنتجاتها على وسائل التواصل الاجتماعي. وعلى مدار أكثر من عقد، استخدمت شركة وول مارت علم البيانات لرفع مستويات المخزون في المتاجر، وثمة مِثال شهير على ذلك عندما أعادت تزويد مخزون فطائر بوب تارتس بنكهة الفراولة في المتاجر من جديدٍ قبل وقوع إعصار فرانسيس في عام ٢٠٠٤ بناءً على تحليل بيانات المبيعات السابقة لإعصار تشارلي الذي وقع قبل بضعة أسابيع. ولقد استخدمت شركة وول مارت، في الآونة الأخيرة، علم البيانات لتشجيع إيراداتها من البيع بالتجزئة فيما يخصُّ تقديم منتجاتٍ جديدة بناء على تحليل الاتجاهات الرائجة على مواقع التواصل الاجتماعي وتحليل أنشطة بطاقات الائتمان لتقديم توصياتٍ ومقترحات بشأن المنتجات للعملاء وتحسين تجربة العملاء عبر الموقع الإلكتروني لشركة وول مارت وإضفاء الطابع الشخصي عليها. وتعزو شركة وول مارت زيادةً يتراوح قدرُها بين ١٠ و١٥ بالمائة من المبيعات الإلكترونية إلى التحسينات الناجمة عن استخدام علم البيانات (DeZyre 2015).

المرادف لبيع المنتجات الأفضل وبيع منتجات إضافية داخل عالم الإنترنت هو «نظام التوصيات والمقترحات». إذا كنتَ قد شاهدتَ فيلمًا على منصَّة نتفليكس أو اشتريتَ منتجًا على موقع أمازون، فستعرف أن هذه المواقع الإلكترونية تستخدِم البيانات التي يجمعونها ليقدموا لك اقتراحاتٍ بخصوص ما ينبغي لك أن تُشاهده أو تشتريه في المرة التالية. ويمكن أن تُصمَّم هذه الأنظمة لترشدك بطرقٍ مختلفة: بعضها يُرشدك نحوَ الأكثر رواجًا والأفضل مبيعًا؛ بينما يُرشدك البعض الآخر نحو منتجاتٍ مُتخصصة تناسِب ذوقك على وجه الخصوص. يذكر كتاب كريس أندرسون بعنوان «الذيل الطويل» (٢٠٠٨) أنه نظرًا إلى أن الإنتاج والتوزيع صارا أقلَّ تكلفة، تحولت الأسواق من بيع كمياتٍ كبيرة من عددٍ قليل من المنتجات الرائجة إلى بيع كمياتٍ صغيرة من عددٍ أكبر من المنتجات المتخصِّصة. تُعد هذه المبادلة بين تشجيع مبيعات المنتجات الرائجة أم المنتجات المتخصِّصة قرارًا أساسيًّا لتصميم

نظام التوصيات والمقترحات وتؤثر على خوارزميات علم البيانات المستخدَمة لتطبيق هذه الأنظمة.

استخدام علم البيانات من قبل الحكومات

لقد أدركت الحكومات مميزات الاستفادة من علم البيانات في السنوات الأخيرة. ففي عام ٢٠١٥، مثلًا، ابتكرت الحكومة الأمريكية منصبَ كبير علماء البيانات في الولايات المتحدة لأول مرةٍ وعيَّنت فيه دكتور دي جيه باتيل. وكانت إحدى كبرى المبادرات التي قادتها الحكومة الأمريكية في علم البيانات من نصيب مجال الصحة. ويأتي علم البيانات في صميم المشروع الطموح لأبحاث علاج السرطان «كانسر مونشوت»[4] ومبادرة الطب الدقيق «بريسشين ميدسين». تجمع مبادرة «بريسشين ميدسين» ما بين تسلسُل الجينوم البشري وعلم البيانات بهدف تصميم أدويةٍ خاصة لكل مريضٍ حسب حالته. ويُعد برنامج «أوول أوف أس» (مبادرة كُلُّنا) جزءًا من هذه المبادرة،[5] ويجمع بياناتٍ بيئية وحياتية وبيولوجية من أكثر من مليون شخصٍ متطوع وذلك بهدف تصميم أكبر مجموعة بياناتٍ للطب الدقيق على مستوى العالم. يُحدِث علم البيانات ثورةً في طريقة تنظيم مُدننا؛ إذ إنه يُستخدم لمتابعة أنظمة البيئة والطاقة والنقل وتحليلها والتحكُّم فيها والاسترشاد بها في التخطيط العمراني على المدى الطويل (Kitchin 2014a). وسنعود إلى موضوع الصحة والمدن الذكية في الفصل السابع عند مناقشتنا للكيفية التي ستتزايد بها أهمية علم البيانات في حياتنا خلال العقود القادمة.

وتركز «مبادرة بيانات الشرطة»[6] من جانب الحكومة الأمريكية على الاستعانة بعلم البيانات بهدف مساعدة أقسام الشرطة على استيعاب احتياجات مجتمعاتها المحلية. كما أن علم البيانات يُستخدم في التنبؤ بالبؤر الإجرامية واحتمالية العودة إلى الإجرام. ومع ذلك، انتقدت الجماعات الداعية للحُرية المدنية بعضًا من استخدامات علم البيانات في مجال العدالة الجنائية. وفي الفصل السادس، سنُناقش مسائلَ أثارَها علمُ البيانات متعلقة بالخصوصية والأخلاقيات، وأحد العوامل المثيرة للاهتمام في هذه المناقشة هو أن آراء الناس فيما يتعلق بالخصوصية الشخصية وعلم البيانات تختلف من مجالٍ لآخر. ولدى الكثير من الناس المرحِّبين باستخدام بياناتهم الشخصية في الأبحاث الطبية الممولة من القطاع العام آراء مختلفة جدًّا عندما يتعلق الأمر باستخدام بياناتهم الشخصية في حفظ النظام

والعدالة الجنائية. وفي الفصل السادس، سنُناقش أيضًا استخدام البيانات الشخصية وعلم البيانات في تحديد أقساط التأمين فيما يخصُّ الحياة والصحة والسيارة والمنزل والسفر.

استخدام علم البيانات في الرياضات الاحترافية

يعرض فيلم «كرة المال» (ماني بول) (إخراج بينيت ميلر، ٢٠١١)، بطولة النجم براد بيت، الاستخدام المتزايد لعلم البيانات في مجال الرياضات الحديثة. ويحكي الفيلم المستوحى من كتابٍ يحمل العنوان نفسه (Lewis 2004) القصة الحقيقية لكيف استخدم فريق البيسبول أوكلاند أثلتيكس علم البيانات لتحسين استراتيجية استقطاب اللاعبين الجدد. أثبتت إدارة الفريق أن إحصاءات نسبة وصول اللاعب إلى القاعدة وتسديده لضربة القاعدة الإضافية أكثر إفادةً من مؤشرات الإحصاء التقليدية المعتمدة في لعبة البيسبول، مثل متوسط ضرب الكرة، للاستدلال على نجاح استراتيجية الهجوم. مكَّنت هذه الفكرة التبصُّرية فريق أوكلاند أثلتيكس من ضم قائمةٍ من اللاعبين الجدد المبخوسة قيمتُهم الحقيقية مع الالتزام بحدود ميزانية الفريق. لقد أحدث نجاح فريق أوكلاند أثلتيكس، مُستعينًا بعلم البيانات، ثورةً في رياضة البيسبول؛ حيث إن معظم فرق البيسبول الأخرى تدمج الآن استراتيجيات مُشابهة تستعين بعلم البيانات في عمليات ضمِّ اللاعبين الجُدد إليها.

تُعَد قصة فيلم «كرة المال» مثالًا واضحًا جدًّا على كيف يمكن لعلم البيانات أن يمنح مؤسسةً ما ميزة تنافسية في السوق التنافسي. ومع ذلك، ربما يكون أهم جانبٍ في قصة «كرة المال» من منظور علم البيانات المحض هو أنها تُسلط الضوء على أن القيمة الأساسية الخاصة لهذا العلم تتمثل أحيانًا في تحديد السمات الثرية بالمعلومات المفيدة. وثمة اعتقاد شائع مفاده أن قيمة علم البيانات تكمُن في النماذج التي تُنشأ أثناء العملية. ومع ذلك، بمجرد أن نعرف السمات المهمة في مجالٍ ما، فمن السهل جدًّا إنشاء نماذج مُستوحاة من البيانات. ومفتاح النجاح هنا هو الحصول على البيانات المناسبة وإيجاد السمات المناسبة. في كتاب «الاقتصاد العجيب: اقتصادي مارق يبحث في الجانب الخفي من كل شيء»، يوضح ستيفن دي ليفيت وستيفن دوبنر أهمية هذه الملاحظة عبر طائفة كبيرة من المشاكل. كما أوضحا، مفتاح فهم الحياة الحديثة هو «معرفة ما يجب قياسه وكيفية قياسه» (٢٠٠٩، ١٤). ومن خلال الاستعانة بعلم البيانات، يُمكننا كشف النقاب عن الأنماط المهمَّة في مجموعة بيانات، ويمكن أن تكشف هذه الأنماط السِّمات المهمة في المجال. والسبب وراء استخدام علم البيانات في الكثير من المجالات هو أنه بغضِّ النظر عن المجال محل الدراسة

إذا كانت البيانات المناسبة متاحة، فإنه يمكن تحديد المشكلة بكلِّ وضوح، وبالتالي يمكن لعلم البيانات أن يُساعدنا في حلِّها.

مفتاح النجاح هنا هو الحصول على البيانات المناسبة وإيجاد السمات المناسبة.

لِمَ الآن؟

لقد أسهم عدد من العوامل في نموِّ علم البيانات مؤخرًا. وكما سبق أن ذكرنا بالفعل، كان ظهور البيانات الضخمة مدفوعًا بالسهولة النسبية التي يمكن للمؤسَّسات أن تجمع بها البيانات. تستطيع الشركات في الوقت الراهن إعداد ملفات تعريف أكثر ثراءً خاصة بالعملاء الأفراد؛ هذا من خلال سجلِّ معاملات نقاط البيع، أو عدد النقرات على المنصات الإلكترونية، أو منشورات وسائل التواصل الاجتماعي، أو التطبيقات على الهواتف الذكية، أو غيرها من القنوات التي لا تُعَد ولا تُحصى. وهناك عامل آخر وهو تحويل مخزون البيانات إلى سلعةٍ تنطبق عليها وفورات الحجم، مما يجعل تخزين البيانات أقلَّ تكلفةً من ذي قبل. كما أن هناك نموًّا هائلًا في القدرة الحاسوبية. إذ تطورت بطاقات الرسومات ووحدات معالجة الرسومات بالأساس لنقل الرسومات بسرعةٍ من أجل ألعاب الكمبيوتر. والسِّمة المميزة لوحدات معالجة الرسومات أنه يُمكنها تنفيذ عمليات ضرب المصفوفات بسرعة. غير أن هذه العمليات ليست مفيدةً من أجل نقل الرسومات وحسب وإنما مفيدة أيضًا من أجل تعلُّم الآلة. وفي السنوات الأخيرة، استُغِلت هذه الوحدات وحُسِّنت بهدف استخدامها في تعلُّم الآلة، الأمر الذي ساهم في زيادة سرعة معالجة البيانات وتدريب النماذج. لقد صارت أدوات علم البيانات السهلة الاستخدام متاحة وذُلِّلت عقبات الدخول إلى علم البيانات. تعني هذه التطورات مجتمعةً أن جمع البيانات وتخزينها ومعالجتها صار أسهلَ من ذي قبل.

كانت هناك تطورات كبيرة في مجال تعلُّم الآلة في السنوات العشر الأخيرة. لقد ظهر التعلُّم العميق، على وجه الخصوص، وأحدث ثورةً في الطريقة التي يمكن أن تُعالِج بها أجهزة الكمبيوتر اللغة وبيانات الصور. ويصف مصطلح «التعلُّم العميق» فئةً من نماذج الشبكات العصبية ذات الطبقات المتعددة من الوحدات داخل الشبكة. كانت الشبكات العصبية موجودةً منذ أربعينيات القرن العشرين؛ إلا أنها تعمل بشكلٍ أفضل مع مجموعات البيانات الكبيرة والمعقدة وتستلزم وجود عددٍ كبير من الموارد الحوسبية لتدريبها. لذا، فإن

ظهور التعلُّم العميق مرتبط بزيادة البيانات الضخمة والقدرة الحوسبية. وليس على سبيل المبالغة وصف تأثير التعلُّم العميق عبر مجموعةٍ من المجالات بأنه تأثير استثنائي للغاية.

ويُعَد برنامج «ألفا جو»[7] الخاص بشركة ديب مايند مثالًا ممتازًا على كيف غيَّر التعلُّم العميق أحد مجالات البحث العلمي تغييرًا جذريًّا. ولعبة «جو» هي لعبة لَوحية ابتُكرت في الصين قبل ثلاثة آلاف سنة. وقواعد لعبة «جو» أسهل من قواعد لعبة الشطرنج؛ إذ يأخذ اللاعبون دورهم في وضع القِطع على اللوحة إما بهدف احتجاز قطع الخَصم أو محاصرة المنطقة الخاوية. ومع ذلك، فإن بساطة القواعد وحقيقة أن لعبة «جو» تستخدِم لوحةً أكبر حجمًا يعني أنه يوجد الكثير من الترتيبات المحتملة للقِطع على لوحة اللعب أكثر من لعبة الشطرنج. في الواقع، الترتيبات المحتملة لقطع لعبة «جو» أكثر من عدد الذرات الموجودة في الكون. هذا يجعل لعبة «جو» أصعب كثيرًا من الشطرنج بالنسبة لأجهزة الكمبيوتر نظرًا إلى أنه تُوجَد مساحة أكبر كثيرًا للبحث فيها وصعوبة تقييم كل ترتيبٍ من هذه الترتيبات المحتملة للقِطع. استعان فريق شركة ديب مايند بنماذج التعلُّم العميق لتمكين برنامج «ألفا جو» من تقييم ترتيبات القطع المحتملة واختيار النقلة التالية في اللعبة. كانت النتيجة أن برنامج «ألفا جو» صار أول برنامج كمبيوتر يهزم لاعبًا محترفًا في لعبة «جو»، حيث إنه في مارس عام ٢٠١٦ هزم البرنامج ليد سيدول، الحائز على لقب بطل العالم في لعبة «جو» ثمانية عشرة مرة، في مباراةٍ شاهدها ٢٠٠ مليون شخص حول العالم. ومن أجل تقدير تأثير التعلُّم العميق على لعبة «جو» تقديرًا سليمًا، يجدُر بنا أن نذكُر أنه في عام ٢٠٠٩ جاء ترتيب أفضل برنامج «جو» في العالم في مرتبةٍ أقلَّ من لاعبٍ هاوٍ متقدم المستوى؛ ولكن بعد مرور سبع سنواتٍ هزم برنامج «ألفا جو» بطل العالم في اللعبة. وفي عام ٢٠١٦، نُشر مقال يصِف خوارزميات التعلُّم العميق المستخدمة في برنامج «ألفا جو» في أكثر مجلة علمية مرموقة على مستوى العالم، مجلة «نيتشر» (Silver, Huang, Maddison, et al. 2016).

كان للتعلُّم العميق أيضًا تأثير كبير على مجموعةٍ من التقنيات المتقدمة التي نستخدِمها يوميًّا. في الوقت الحالي، يستعين موقع فيسبوك بالتعلُّم العميق للتعرُّف على الوجوه وتحليل النصوص لعرض الإعلانات مباشرة على الأشخاص بناءً على محادثاتهم عبر الإنترنت. ويستعين كلٌّ من موقع جوجل وبايدو بالتعلُّم العميق من أجل التعرُّف على الصور والتعليقات عليها والبحث والترجمة الآلية. ويستعين المساعد الافتراضي «سيري» من ابتكار شركة أبل، و«ألكسا» من ابتكار شركة أمازون، و«كورتانا» من ابتكار شركة مايكروسوفت،

و«بيكسبي» من ابتكار شركة سامسونج بخاصية التعرُّف على الصوت القائمة على التعلُّم العميق. وحاليًّا تُطور شركة هواوي مساعدًا افتراضيًّا من أجل السوق الصينية، وسيَستخدِم أيضًا التعلُّم العميق في التعرُّف على الصوت. وسوف نتناول في الفصل الرابع الشبكات العصبية والتعلُّم العميق بمزيدٍ من التفاصيل. وعلى الرغم من أن التعلُّم العميق يُعَد تطورًا تقنيًّا مُهمًّا، ربما أهم ما فيه فيما يخصُّ نمو علم البيانات هو الوعي المتزايد بقدرات هذا العلم ومميزاته واعتماد المؤسسات عليه بشكلٍ كبير والذي أسفر عن قصص نجاحها الرفيعة المستوى.

خرافات حول علم البيانات

لعلم البيانات فوائد كثيرة بالنسبة إلى المؤسسات الحديثة؛ إلا أن هناك قدرًا كبيرًا من المبالغة حوله، ولذا يجب أن نفهم ما هي حدوده. واحدة من أكبر الخُرافات هي الاعتقاد بأن علم البيانات ينطوي على عمليةٍ مُستقلة يمكننا أن نمنحها مطلق الحرية على بياناتنا بهدف العثور على حلولٍ لمشكلاتنا. ولكن في الواقع، يستلزم علم البيانات إشرافًا بارعًا من جانب البشَر عبر مختلف مراحل العملية. ويجب على المحللين وضع إطار للمشكلة، وتصميم البيانات وتجهيزها، وتحديد أيٍّ من خوارزميات تعلُّم الآلة هي الأنسب، وتفسير نتائج التحليل تفسيرًا نقديًّا؛ والتخطيط للإجراء المناسب الذي يجب اتخاذه بناءً على الرؤية (الرؤى) التي كشف عنها التحليل. ومن دون الإشراف البارع من جانب البشر، ستُخفق مشروعات علم البيانات في تحقيق أهدافها. وتأتي أفضل النتائج الخاصة بعلم البيانات عندما تتضافر الخبرة البشرية والقدرة الحاسوبية معًا، كما يقول جوردون لينوف ومايكل بيري: «التنقيب في البيانات يُتيح لأجهزة الكمبيوتر إنجاز ما تُنجزه على أفضل وجه؛ ألا وهو التنقيب عبر بياناتٍ كثيرة. وهذا بدوره يُتيح للبشر إنجاز ما ينجزونه على أفضل وجه؛ ألا وهو تحديد المشكلة وفهم النتائج» (٢٠١١، ٣).

يعني انتشار علم البيانات واستخدامه المتزايد أن أكبر تحدٍّ أمام مؤسسات كثيرة فيما يخص هذا العلم يتمثل حاليًّا في تحديد الأشخاص المؤهَّلِين كمُحللين وتوظيفهم. فالموهبة البشرية في مجال علم البيانات مطلوبة بشدة نظرًا إلى قيمتها، والعثور على مثل هذه المواهب هو المأزق الرئيسي لتحقيق الاستفادة من علم البيانات. ولكي نضع هذا النقص في الموهبة في سياقِه الصحيح، في عام ٢٠١١ توقَّع تقرير معهد ماكينزي العالمي نقصًا في الولايات المتحدة يتراوح بين ١٤٠ ألفًا و١٩٠ ألف شخصٍ يتمتَّعون بمهارات

علم البيانات والمهارات التحليلية، ونقصًا أكبر يبلُغ ١,٥ مليون مدير قادر على فَهم علم البيانات والعمليات التحليلية بمستوًى سيُمكنهم من الاستعلام عن نتائج علم البيانات وتفسيرها على النحو الصحيح (Manyika, Chui, Brown, et al. 2011). وبعد مرور خمس سنوات، وفي تقريره الصادر عام ٢٠١٦، ظل المعهد مقتنعًا بأن علم البيانات يتمتع بإمكانياتٍ هائلة وقيمةٍ غير مُستغلَّة عبر نطاقٍ واسع من التطبيقات؛ غير أن نقص الموهبة سيظل قائمًا، مع وجود عجزٍ مُتوقَّع يُقدَّر بنحو ٢٥٠ ألف عالِم بيانات على المدى القريب (Henke, Bug-hin, Chui, et al. 2016).

يتمثَّل ثاني أكبر الخُرافات حول علم البيانات في أن كل مشروع قائم على علم البيانات بحاجةٍ إلى بيانات ضخمة وبحاجة إلى استخدام التعلُّم العميق. وبوجهٍ عام، من المفيد توفير المزيد من البيانات؛ غير أن توفير البيانات «المناسبة» هو الشرط الأهم. وكثيرًا ما تُنفَّذ مشروعات علم البيانات في المؤسسات التي تتوافر لديها موارد أقل كثيرًا من شركة جوجل أو بايدو أو مايكروسوفت على صعيد البيانات والقدرة الحوسبية. وتشمل الأمثلة على نطاق مشروعات علم البيانات الأصغر حجمًا التنبُّؤ بالمطالبات في شركة تأمين تستقبل نحو ١٠٠ مطالبة في الشهر؛ والتنبؤ بنسبةِ تسرُّب الطلاب من جامعة بها أقل من ١٠ آلاف طالب؛ وتوقُّع تسرُّب أعضاء اتحادٍ قِوامه عدة آلاف من الأعضاء. ومن ثم، ليست المؤسسة في حاجة لأن تُعالج تيرابايت من البيانات أو تمتلك موارد حوسبية هائلة تحت تصرُّفها لكي تستفيد من علم البيانات.

وثالث خرافة حول علم البيانات هي أن برامج علم البيانات الحديثة يسهل استخدامها، وبالتالي تسهل ممارسة عمليات علم البيانات. صحيح أن برامج علم البيانات صارت أسهل في استخدامها. إلا أن سهولة الاستخدام هذه قد تُخفي وراءها حقيقة أن القيام بالعمليات الخاصة بعلم البيانات على النحو الصحيح يتطلَّب معرفةً صحيحة بالمجال وخبرةً فيما يتعلق بخصائص البيانات والافتراضات التي تقوم عليها خوارزميات تعلُّم الآلة المختلفة. في الواقع، من السهل القيام بالعمليات الخاصة بعلم البيانات على نحوٍ سيئ أكثر من أي وقتٍ مضى. وكما هو الحال مع أي شيءٍ آخر في الحياة، إذا كنت لا تفهم ما تفعله أثناء القيام بالعمليات الخاصة بعلم البيانات، فإنك سترتكب أخطاءً. تكمن خطورة التعامل مع علم البيانات في أن التكنولوجيا قد تجعل البشر يتهيَّبون وبالتالي يُصدقون أي نتائج تُقدمها لهم البرامج. ومع ذلك، فإنهم قد يُخطئون في تحديد المشكلة بغير قصدٍ منهم، أو يُدخلون بيانات خاطئة، أو يستخدمون تقنيات تحليلٍ ذات افتراضات غير مناسبة.

وبالتالي، من المرجَّح أن تكون النتائج التي تُقدمها البرامج إجابةً للسؤال الخطأ أو تستند إلى بياناتٍ خاطئة أو نتيجة عمليات حسابية خاطئة.

والخرافة الأخيرة حول علم البيانات التي نودُّ أن نذكُرها هنا هي الاعتقاد بأن علم البيانات يُغطي تكلفته سريعًا. وحقيقة هذا الاعتقاد مُتوقفة على سياق العمل في المؤسسة. قد تستلزِم الاستفادة من علم البيانات استثمارًا كبيرًا فيما يخصُّ تطوير البِنية التحتية للبيانات وتعيين موظفين لديهم خبرة في مجال علم البيانات. علاوة على ذلك، لن يُحقق علم البيانات نتائج إيجابية مع كل مشروع. أحيانًا، لا تُوجَد أية معلومات قيِّمة يمكن العثور عليها في البيانات، وأحيانًا أخرى لا تكون الشركة في موضعٍ يُتيح لها التصرف بناء على المعلومات القيمة التي كشف عنها التحليل. ومع ذلك، ففي السياقات التي يُوجَد فيها مشكلة تجارية مفهومة جيدًا وتُتاح فيها البيانات المناسبة وتتوفر فيها الخبرات البشرية، كثيرًا ما يوفر علم البيانات الرؤى المستنيرة القابلة للتنفيذ والتي توفر للمؤسسة الميزة التنافسية التي تحتاج إليها لتحقيق النجاح.

الفصل الثاني

ما المقصود بالبيانات وما المقصود بمجموعة البيانات؟

يعتمد علم البيانات، كما يوحي اسمه، على البيانات بالأساس. والبيان أو المعلومة، في أبسط صورهما، عبارة عن فكرةٍ مجردة لكيانٍ ما من الواقع الفعلي (شخص أو كائن أو حدث). وعادةً ما تُستخدَم مصطلحات مثل «متغير»، و«ميزة»، و«سمة» على نحوٍ متبادل لتُشير إلى فكرةٍ فردية مجردة. ويُوصف كل كيان عادةً بعدد من السمات. على سبيل المثال، يوصف الكتاب بالسمات التالية: الكاتب والعنوان والموضوع والنوع الأدبي والناشر والسعر وتاريخ النشر وعدد الكلمات وعدد الفصول وعدد الصفحات والطبعة والرقم الدولي الموحَّد للكتب وهلمَّ جرًّا.

وتتكوَّن مجموعة البيانات من بياناتٍ ذات صلة بمجموعة من الكيانات؛ وكل كيان منها يُوصف بمجموعة من السمات. وفي أبسط صورها،[1] تُرتَّب مجموعة البيانات على هيئة مصفوفة بيانات $n*m$ تُسمَّى «سجل التحليل»، حيث n هو عدد الكيانات (صفوف) وm هو عدد السِّمات (الأعمدة). وكثيرًا ما يُستخدم مصطلح «مجموعة البيانات» و«سجل التحليل» على نحوٍ تبادُلي في علم البيانات، حيث يكون سجل التحليل تمثيلًا خاصًّا لمجموعة البيانات. يوضح جدول ٢-١ سجل التحليلات لمجموعة البيانات الخاصة بكتُب الأعمال الكلاسيكية. وكل صفٍّ في الجدول يصف كتابًا واحدًا. تُستخدَم مصطلحات «مثيل»، و«مثال»، و«كيان»، و«كائن»، و«حالة»، و«فرد»، و«سجل» في مؤلَّفات علم البيانات للإشارة إلى الصف. وهكذا تحتوي مجموعة البيانات على مجموعةٍ من المثيلات، وكل مثيل يُوصف بمجموعة من السِّمات.

إن إعداد سجل التحليل هو شرط أساسي لممارسة علم البيانات. في الواقع، تُنفق الغالبية العظمى من الوقت والجهد المبذولَين في مشروعات علم البيانات في إنشاء سجل

التحليل وتنظيفه وتحديثه. وكثيرًا ما يُنشأ سجل التحليل من خلال دمج المعلومات من العديد من المصادر المختلفة؛ إذ ربما تُستخلَص البيانات من عدة قواعد بيانات أو مخازن بيانات أو ملفَّات حاسوبية بتنسيقاتٍ مختلفة (مثل جداول البيانات أو ملفات «سي إس في» (القيم المفصولة بفاصلة)) أو من خلال جمعها بجهدٍ من شبكة الإنترنت أو وسائط مواقع التواصل الاجتماعي.

جدول ٢-١: مجموعة بيانات خاصة بالأعمال الكلاسيكية.

رقم تعريفي	عنوان الكتاب	المؤلف	العام	الغلاف	الطبعة	السعر
١	«إيما»	أوستن	١٨١٥	غلاف ورقي	العشرون	٥٫٧٥ دولارات
٢	«دراكولا»	ستوكر	١٨٩٧	غلاف مُقوى	الخامسة عشرة	١٢ دولارًا
٣	«إيفانهو»	سكوت	١٨٢٠	غلاف مُقوى	الثامنة	٢٥ دولارًا
٤	«المخطوف»	ستيفنسون	١٨٨٦	غلاف ورقي	الحادية عشرة	٥ دولارات

أُدرجت أربعة كتب في مجموعة البيانات المذكورة في جدول ٢-١. وإذا استبعدنا سمة الرقم التعريفي — وهو بكل بساطة عبارة عن تسميةٍ لكل صفٍّ وبالتالي ليس ذا فائدة في التحليل — نجد أن كل كتابٍ يوصف باستخدام ستِّ سمات؛ ألا وهي: عنوان الكتاب ومؤلفه وعام النشر ونوع الغلاف ورقم الطبعة والسعر. كان بإمكاننا أن نُدرج المزيد من السمات لكل كتاب؛ إلا أننا كنَّا بحاجةٍ إلى الاختيار من السِّمات عندما كنَّا نُصمم مجموعة البيانات كما هو معتاد مع مشروعات علم البيانات. وفي هذا المثيل، نحن مُقيدون بحجم الصفحة وعدد السمات التي كان بإمكاننا أن نُدرجها. وفي أغلب مشروعات علم البيانات، تكون القيود مُتعلقةً بأيٍّ من السِّمات يُمكننا جمعها فعليًّا وأي من السمات نُصدقها بناءً على معرفتنا بالمجال ذي الصلة بالمشكلة التي نُحاول حلَّها. إن إدراج سماتٍ إضافية في مجموعة البيانات لا يأتي بدون تكلفة. أولًا: يُبذَل المزيد من الوقت والجهد في جمع المعلومات والتأكُّد من جودتها لكلِّ مثيلٍ في مجموعة البيانات ودمج هذه البيانات في سجلِّ التحليل. ثانيًا: قد يكون لإدراج سماتٍ غير ذات صلة أو متكررة تأثيرٌ سلبي على أداء الكثير من الخوارزميات المستخدَمة في التحليل. وإدراج الكثير من السِّمات في

مجموعة البيانات يزيد من احتمالية أن تجد الخوارزمية أنماطًا غير ذات صلة أو زائفة في البيانات التي تبدو ذات أهمية إحصائية فقط بسبب عينةٍ مُعينة من المثيلات الموجودة في مجموعة البيانات. وتُعد مشكلة كيفية اختيار السمة (السمات) المناسبة تحدِّيًا أمام جميع مشروعات علم البيانات، وأحيانًا يتعلق الأمر بعملية تكرار التجارب القائمة على مبدأ التجربة والخطأ حيث يتحقَّق كل تكرار من النتائج المحرزَة باستخدام مجموعاتٍ فرعية مختلفة من السمات.

هناك الكثير من أنواع السمات المختلفة، وكل نوع من السِّمات تُناسبه أنواع مختلفة من التحليل. وبالتالي فإن فهم الأنواع المختلفة من السِّمات والتعرُّف عليها هي مهارة رئيسية بالنسبة إلى عالِم البيانات. والأنواع القياسية هي سِمات «عددية»، و«اسمية»، و«ترتيبية». تصِف السمات العددية الكميات القابلة للقياس التي تُمثل باستخدام أعدادٍ صحيحة أو قِيَم حقيقية. ويمكن قياس السِّمات العددية إما على «مقياس الفاصل» أو «مقياس النسبة». تُقاس سمات الفاصل على مقياسٍ ذي فاصلٍ ثابت ولكنه اعتباطي وأصل اعتباطي — على سبيل المثال، قياسات التاريخ والوقت. من المناسب تنفيذ عمليات الترتيب والطرح على سِمات الفاصل، إلا أن العمليات الحسابية الأخرى (مثل الضرب والقسمة) غير مناسبة. ومقاييس النسبة مشابهة لمقاييس الفاصل؛ إلا أنَّ تدرُّج القياس يحتوي على صفرٍ حقيقي. وتُشير قيمة الصفر إلى أنه لم يتمَّ قياس أية كمية. وإحدى تداعِيات وجود أصلٍ صفري حقيقي في مقياس النسبة هو أننا يُمكننا وصف قيمةٍ ما على مقياس النسبة بأنها مضاعف (أو نسبة) لقيمةٍ أخرى. وتُعَد درجة الحرارة مثالًا مفيدًا للتمييز بين مقياس الفاصل ومقياس النسبة.[2] قياس درجة الحرارة على مقياس الدرجة المئوية أو مقياس فهرنهايت هو مثال على مقياس الفاصل نظرًا إلى أن القيمة صفر على أيٍّ من هذين المِقياسَين لا تُشير إلى درجة الحرارة صفر. ولذلك على الرغم من أن بإمكاننا حساب الاختلافات في درجة الحرارة على هذَين المقياسَين ومقارنة هذه الاختلافات، لا يُمكننا القول إن درجة الحرارة ٢٠ درجة مئوية هي ضعف درجة حرارة ١٠ درجات مئوية. على النقيض من ذلك، فإن قياس درجة الحرارة بالكلفن يتم على مقياس نسبةٍ لأن صفر كلفن (الصفر المطلَق) هو درجة الحرارة التي تتوقَّف عندها الحركة الحرارية بكافة أشكالها. وتشمل الأمثلة الشائعة الأخرى لقياسات مقياس النسبة المَبالِغ المالية والوزن والطول ودرجات الاختبارات الورقية (مقياس من ٠–١٠٠). في جدول ٢-١، تعد سمة «العام» مثالًا على سمة مقياس فاصل، وسمة «السعر» مثالًا على سمة مقياس نسبة.

تستقي السمات الاسمية (المعروفة أيضًا بالسمات الفئوية) القِيَم من مجموعةٍ محدودة. وهذه القيم هي أسماء (ومنها جاءت صفة السمات «الاسمية») للفئات أو الطبقات أو الحالات. ومن الأمثلة على السِّمات الاسمية سمة الحالة الاجتماعية (أعزب، مُتزوج، مُطلَّق) وسمة نوع البيرة (المِزْر، مِزْر شاحب، جعة مُعتقة، بيرة إنجليزية، بيرة ستاوت، وهلم جرًّا). والسمة الثنائية هي حالة خاصة من السمات الاسمية حيث تكون مجموعة القِيَم المحتملة مُقتصرةً على قيمتَين فقط. على سبيل المثال، قد يكون لدَينا السمة الثنائية «بريد عشوائي»، التي تصف رسائل البريد الإلكتروني إما بأنها عشوائية (صواب) أو غير عشوائية (خطأ)، أو السمة الثنائية «مُدخِّن» والتي تصف الفرد إما بأنه مُدخن (صواب) أو غير مدخن (خطأ). ولا يمكن تنفيذ عمليات ترتيبية أو حسابية على السِّمات الاسمية. لاحظ أن السِّمات الاسمية يمكن ترتيبها أبجديًّا؛ إلا أن الترتيب الأبجدي هو عملية مختلفة عن الترتيب العددي. في جدول ٢-١، «المؤلف» و«العنوان» هما مثالان على السِّمات الاسمية.

تتشابه السمات الترتيبية مع السمات الاسمية؛ مع الفارق أنه من الممكن تطبيق ترتيبٍ تدريجي على الفئات الخاصة بالسمات الترتيبية. فعلى سبيل المثال، ربما تستقي إحدى السِّمات التي تصف الإجابة على سؤال استطلاعي قيمًا من النطاق «لا يُعجبني على الإطلاق، لا يُعجبني، مُحايد، يُعجبني، يُعجبني بشدة.» وثمة ترتيبٍ طبيعي لهذه القِيَم من «لا يُعجبني على الإطلاق» إلى «يُعجبني بشدة» (أو العكس حسب العُرف المتبع). ومع ذلك، تتمثل إحدى الميزات المهمة للبيانات الترتيبية في عدم وجود مسافاتٍ متساوية بين هذه القيم. على سبيل المثال، ربما تختلف المسافة المعرفية بين «لا يُعجبني» و«محايد» عن المسافة بين «يُعجبني» و«يُعجبني بشدة». ونتيجة لذلك، ليس من المناسب تنفيذ عمليات حسابية (مثل إيجاد المتوسط) على السمات الترتيبية. في جدول ٤-١، تُعد سمة «الطبعة» مثالًا على السمة الترتيبية. والفارق بين البيانات الاسمية والترتيبية ليس واضحًا على الدوام. على سبيل المثال، فكِّر مليًّا في سِمةٍ تصف الطقس والتي يمكن أن تأخذ القيمة «مُشمس»، أو «مُمطر»، أو «مُلبَّد بالغيوم». ربما يَعتبر أحد الأشخاص هذه السمة اسمية، في ظل غياب الترتيب الطبيعي على القيم، في حين ربما يَعتبر شخصٌ آخر السمةَ ترتيبية، في ظل التعامل مع القيمة «مُلبد بالغيوم» باعتبارها قيمةً وسطية بين «مُشمس» و«مُمطر» (Hall, Witten, and Frank 2011).

يؤثر نوع البيانات الخاص بالسِّمة (عددية، أم ترتيبية، أم اسمية) على الطرق التي يُمكننا الاستعانة بها لتحليل البيانات وفهمها، ومن بين ذلك كُلٌّ من الإحصاءات الأساسية

التي يُمكننا استخدامها لوصف توزيع القِيَم التي تأخذها سمةٌ ما والخوارزميات الأكثر تعقيدًا التي نستخدِمها لتحديد أنماط العلاقات بين السِّمات. عند أبسط مستوًى للتحليل، تُتيح السِّمات العددية تنفيذ عملياتٍ حسابية، والتحليل الإحصائي النموذجي الذي يطبق على السِّمات العددية هو تحليل النزعة المركزية (باستخدام مُتوسط القيمة الخاصة بالسمة) وتشتُّت قيم السمات (باستخدام إحصاءات التبايُن أو الانحراف المعياري). ومع ذلك، ليس من المنطقي تنفيذ العمليات الحسابية على سماتٍ اسمية أو ترتيبية. ومن ثم، يشمل التحليل الأساسي لهذه الأنواع من السِّمات إحصاء عدد المرات التي تظهر فيها كل قيمةٍ في مجموعة البيانات أو حساب نسبة ظهور كل قيمةٍ أو كِلا الشيئَين.

يؤثر نوع البيانات الخاص بالسمة (عددية، أم ترتيبية، أم اسمية) على الطرق التي يُمكننا الاستعانة بها لتحليل البيانات وفهمها.

تتولد البيانات من خلال عملية تجريد، ومن ثم فإن أية بيانات تكون ناتجة عن قرارات البشر واختياراتهم. ومن أجل القيام بأي عمليةِ تجريد، يتعيَّن على شخصٍ ما (أو مجموعة من الأشخاص) أن يختار ما سيقوم بالتجريد منه وما الفئات أو وسائل القياس التي يجِب استخدامُها في التمثيل المجرد. ومعنى ذلك هو أن البيانات لا تُمثل أبدًا وصفًا موضوعيًّا للواقع. وإنما دائمًا ما تكون مُغرضة ومُتحيزة. وكما قال ألفريد كورزيبسكي: «الخريطة في حدِّ ذاتها ليست الأرض التي تُمثلها؛ ولكنها إذا كانت مُتقَنة فإنها تحوي تضاريس مُشابهة لتضاريس الأرض، وهذا ما يجعلها ذات فائدة» (عام ١٩٩٦، ٥٨).

بعبارةٍ أخرى، البيانات التي نستخدِمها لعلم البيانات ليست تمثيلًا مثاليًّا لكيانات الواقع الفعلي والعمليات التي نحاول فهمها، ولكن إذا توخَّينا الحذَر حيال كيفية تصميم البيانات التي نستخدمها وكيفية جمعها، فإن نتائج تحليلنا ستوفر رؤًى مفيدة عن مشكلات واقعنا الفعلي. وتُعَد قصة «كرة المال» التي ذكرناها في الفصل الأول مثالًا رائعًا على كيف أنَّ العامل المحدد للنجاح في الكثير من مشروعات علم البيانات يتمثل في تحديد التجريدات (السمات) المناسبة للاستعانة بها في مجالٍ مُعين. تذكر أن مفتاح نجاح قصة «كرة المال» تمثل في أن فريق أوكلاند أثلتيكس أدرك أن نسبة وصول اللاعب إلى القاعدة ونسبة تسديد ضربة القاعدة الإضافية هما أفضل سِمتَين يمكن الاستعانة بهما لتوقُّع نجاح استراتيجية الهجوم مقارنة بإحصاءات البيسبول التقليدية مثل متوسط ضرب

الكرة. إن استخدام سماتٍ مختلفة لوصف اللاعبين وفَّر لفريق أوكلاند نموذجًا مُختلفًا وأفضل من النموذج الذي تستخدِمه الفِرَق الأخرى، مما مكَّنه من التعرُّف على اللاعبين المَبخوسة قيمتُهم الحقيقية ومكنه من المنافسة مع فِرَق أكبر حجمًا بميزانيةٍ أقل.

توضح قصة «كرة المال» أن مقولة «المدخلات الخاطئة تُعطي مخرجاتٍ خاطئة» في علوم الكمبيوتر تنطبق على علم البيانات أيضًا: فإذا كانت مدخلات عملية الحوسبة خاطئة، فإن مخرجات هذه العملية ستكون خاطئة أيضًا. وبالطبع، لا نُغالي إذا ما شدَّدنا على خاصيَّتَين تُميزان علم البيانات: (أ) يجب أن نولي قدرًا كبيرًا من الاهتمام إلى كيفية إنشاء بياناتنا (فيما يخصُّ كلًّا من الاختيارات التي نقوم بها لتصميم تجريدات البيانات وجودة البيانات المستخلَصة من عمليات التجريد) و(ب) يجب علينا «التحقُّق من دقة» نتائج عملية علم البيانات — أي يجب علينا أن نستوعب أنه لمجرد أن الكمبيوتر يُحدد نمطًا في البيانات لا يعني بالضرورة أنه يُحدد رؤيةً حقيقيةً في العمليات التي نُحاول تحليلها؛ إذ ربما يكون السبب ببساطة في تحديد هذا النمط هو تحيُّزنا في تصميم البيانات واستخلاصها.

منظورات بشأن البيانات

بخلاف نوع البيانات (عددية واسمية وترتيبية)، يمكن تحديد عددٍ من الفروق المفيدة الأخرى المتعلقة بالبيانات. أحد هذه الفروق هو الفارق بين البيانات «الهيكلية» والبيانات «غير الهيكلية». البيانات الهيكلية هي بيانات يُمكن تخزينها في جدول، ويحظى كل مثيلٍ في الجدول بالهيكل نفسه (أي مجموعة السِّمات). لنضرب مثلًا بالبيانات الديموغرافية للسكان؛ حيث يصِف كل صفٍّ في الجدول شخصًا واحدًا ويتكوَّن من مجموعة السمات الديموغرافية نفسها (الاسم، والسن، وتاريخ الميلاد، والعنوان، والنوع الاجتماعي، والمستوى التعليمي، وحالة الوظيفة ... إلخ). ويمكن بسهولةٍ تخزين البيانات الهيكلية وتنظيمها والبحث فيها وإعادة ترتيبها ودمجها مع بياناتٍ هيكلية أخرى. ومن السهل نسبيًّا تطبيق علم البيانات على البيانات الهيكلية لأنها بحُكم التعريف موجودة في نسَق يناسِب الدمج في سجل تحليلات. أما «البيانات غير الهيكلية» فهي بيانات ربما يكون لكل مثيلٍ في مجموعة البيانات هيكلُه الداخلي الخاص به، وهذا الهيكل ليس بالضرورة نفس الهيكل الخاص بالمَثيلات الأخرى. على سبيل المثال، تخيَّل مجموعة بياناتٍ خاصة بصفحات الويب، ولكل صفحة ويب هيكل، ولكن هذا الهيكل يختلف من صفحةٍ لأخرى. والبيانات غير الهيكلية

أكثر شيوعًا من البيانات الهيكلية. على سبيل المثال، يمكن اعتبار مجموعات النصوص التي كتبَها البشر (رسائل البريد الإلكتروني، التغريدات، الرسائل النصية، المنشورات، الروايات، وغيرها) بياناتٍ غير هيكلية، كما هو الحال مع مجموعات ملفات الصوت والصور والموسيقى والفيديو والوسائط المتعددة. ويعني تنوع الهيكل بين العناصر المختلفة أنه من الصعب تحليل البيانات غير الهيكلية في صورتها الأصلية. يُمكننا عادة استخلاص بيانات هيكلية من البيانات غير الهيكلية باستخدام تقنيات الذكاء الاصطناعي (مثل معالجة اللغات الطبيعية وتعلُّم الآلة)، ومعالجة الإشارات الرقمية والرؤية الحاسوبية. ورغم ذلك، فإن تنفيذ واختبار هذه العمليات لتحويل البيانات هو أمر مُكلِّف ومستنزِف للوقت وقد يُضيف نفقاتٍ ماليةً كبيرة ويتسبَّب في تأخير مشروع علم البيانات.

وأحيانًا تكون السِّمات عبارة عن تجريدات «خام» مُستقاة من حدثٍ أو كائنٍ ما — على سبيل المثال طول شخص، أو عدد الكلمات في رسالة بريد إلكتروني، أو درجة الحرارة في غرفة، أو وقت الحدث أو مكان حدوثه. بيد أنه يمكن أيضًا «اشتقاق» البيانات من أجزاء أُخرى من بيانات. تأمَّل متوسط الرواتب في إحدى الشركات أو تفاوُت درجات حرارة إحدى الغُرف على مدار فترة زمنية. في كِلا المثالَين، البيانات الناتجة مُشتقة من مجموعةٍ أصلية من البيانات من خلال تنفيذ دالة على البيانات الخام الأصلية (رواتب الأفراد أو قراءات درجات الحرارة). وكثيرًا ما تتمثَّل القيمة الحقيقية لمشروع علم البيانات في تحديد سمةٍ مُشتقة واحدة (أو أكثر) ذات أهمية لتمنحنا رؤيةً ثاقبة عن مشكلةٍ ما. تخيَّل أننا نُحاول التوصُّل إلى فهمٍ أفضل لأسباب السمنة المفرطة لدى مجموعةٍ من السكان، ونحاول فهم السمات الخاصة بالفرد الذي يُصنِّف نفسه كشخصٍ يُعاني من السمنة المفرطة. سنبدأ بفحص السمات الخام للأفراد مثل الطول والوزن؛ غير أنه بعد دراسة المشكلة لبعض الوقت قد ينتهي بنا الأمر إلى ابتكار سمةٍ مُشتقة غنية أكثر بالمعلومات مثل مؤشر كتلة الجسم. ومؤشر كتلة الجسم هو نسبة كتلة الشخص إلى طوله. إن إدراك أن «التفاعُل» بين سمتَين من السِّمات الخام وهما «الكتلة» و«الطول» يوفر المزيد من المعلومات عن السمنة المفرطة أكثر مما قد تُتيحه إحدى السِّمتَين بمعزلٍ عن الأخرى — سيساعدنا على تحديد الأشخاص المعرَّضين إلى خطر الإصابة بالسمنة المفرطة في قطاع السكان. بالتأكيد مؤشر كتلة الجسم هو مِثال بسيط نستعين به هنا لتوضيح أهمية السِّمات المشتقة. ولكن ضع في اعتبارك المواقف التي نحصل فيها على رؤيةٍ بشأن مُشكلةٍ ما من خلال عدة سماتٍ مشتقة؛ حيث تُشتَق كل سمةٍ من سمتَين

إضافيتَين (أو ربما أكثر). وفي السياقات التي تتفاعل فيها عدة سِمات بعضها مع بعض، يوفر لنا علم البيانات فوائد حقيقية لأن الخوارزميات التي نستخدمها يُمكنها في بعض الحالات أن تفرق بين السمات المشتقة والبيانات الخام.

> كثيرًا ما تتمثل القيمة الحقيقية لمشروع علم البيانات في تحديد سمةٍ مشتقة واحدة (أو أكثر) ذات أهمية لتمنحنا رؤيةً ثاقبة عن مشكلةٍ ما.

بوجهٍ عام يطلَق على «البيانات الخام» التي تُجمَّع «البيانات المستخلَصة» و«البيانات الثانوية» (Kitchin 2014a). تُجمع «البيانات المستخلَصة» من خلال القياس المباشر أو الملاحظة المباشرة المصمَّمة خصوصًا لجمع البيانات. على سبيل المثال، الغرَض الأساسي من الاستطلاعات والتجارب هو جمع بيانات مُحددة حول موضوع مُعين يحظى بالاهتمام. وعلى النقيض من ذلك، البيانات الثانوية هي مُنتج فرعي لعمليةٍ ما، الغرض الأساسي منها هو أي شيءٍ آخر بخلاف استخلاص البيانات. على سبيل المثال، الغرَض الأساسي من الكثير من تقنيات وسائط التواصل الاجتماعي هو تمكين المستخدِمين من التواصل مع الآخرين. غير أنه مع كل صورة تتم مشاركتها، أو كل مدوَّنة تُنشر، أو كل تغريدة يُعاد نشرها، أو منشور يَلقى إعجابًا، تتولَّد مجموعة من البيانات الثانوية مثل: من شارك، ومن شاهد، وما الجهاز المستخدم، وفي أي وقتٍ من اليوم، وأي جهاز استُخدِم في ذلك، وكم عدد الأشخاص الذين شاهدوا/أعجبوا/أعادوا النشر وهلمَّ جرًّا. وعلى نحوٍ مماثِل، الغرض الرئيسي من موقع أمازون هو تمكين المستخدِمين من إجراء عمليات شراءٍ من خلال الموقع الإلكتروني. إلا أن كل عملية شراء تُولِّد كمياتٍ مهولة من البيانات الثانوية: ما العناصر التي يضعها المستخدِم في سلة التسوق الخاصة به، ومدة تصفُّحه الموقع الإلكتروني، وما العناصر الأخرى التي تفقَّدها، وغير ذلك.

وأحد أكثر البيانات الثانوية شيوعًا هي «بيانات التعريف»؛ ألا وهي البيانات التي تصف بياناتٍ أخرى. عندما سرَّب إدوارد سنودن وثائق حول برنامج المراقبة «بريسم» التابع لوكالة الأمن القومي الأمريكية، كشف أن الوكالة كانت تجمع كميةً مهولة من بيانات التعريف حول المكالمات الهاتفية التي يُجريها الناس. كان هذا يعني أن الوكالة لم تكن تسجل محتوى المكالمات الهاتفية فعليًّا (لم تكن تتنصَّت على المكالمات الهاتفية) وإنما كانت تجمع بياناتٍ حول المكالمات الهاتفية، مثل متى أُجريت المكالمة، ومن الذي استقبلها،

وكم استمرَّت مُدتها، وغيرها من البيانات الأخرى (Pomerantz 2015). ربما لا يبدو أن هذا النوع من جمع البيانات يُنذر بأي سوء؛ إلا أن دراسة «ميتافون» التي أُجريت بجامعة ستانفورد أوضحت أنواعَ الرؤى ذات الطبيعة الحساسة التي قد تكشف عنها بيانات تعريف المكالمات الهاتفية لأحد الأفراد (Mayer and Mutchler 2014). وحقيقة أن الكثير من المؤسسات لها أغراض مُحددة جدًّا تجعل من السهل نوعًا ما استنتاج معلوماتٍ حسَّاسة عن شخصٍ ما بناءً على مكالماته الهاتفية مع هذه المؤسسات. على سبيل المثال، أجرى بعض الأشخاص المشاركين في دراسة «ميتافون» مكالماتٍ هاتفية مع جمعية مُدمني الكحول المَجهولين ومُحامِي قضايا الطلاق والعيادات الطبية المتخصِّصة في الأمراض المنقولة جنسيًّا. وقد تكون الأنماط المتبعة في المكالمات الهاتفية كاشفة أيضًا. إذ أظهر تحليل الأنماط المأخوذة من الدراسة كيف تكشف أنماط المكالمات الهاتفية معلوماتٍ قد تكون حسَّاسة للغاية:

> تواصَل المشارك (أ) مع عدة جماعات محلية مُتخصصة في طب الأعصاب، وصيدلية مُتخصِّصة وخدمة إدارة الحالات النادرة والخط الساخن لتوفير دواء يُستخدم لعلاج التصلُّب العصبي المتعدد ... على مدار ثلاثة أسابيع، تواصَل المشارك (د) مع متجرٍ مُتخصص في تجديد المنازل وصنَّاع أقفال وموزع معدات الزراعة المائية ومتجر مُستلزمات التدخين. (Mayer and Mutchler 2014)

يركز علم البيانات عادةً على البيانات المجمعة المستخلَصة. ومع ذلك، كما توضح دراسة «ميتافون»، يمكن أن تُستخدَم البيانات الثانوية لكشف رؤيةٍ متوارية عن مواقف مُعينة. وفي السنوات الأخيرة، تزايدت فائدة البيانات الثانوية، لا سيما في مجال مشاركة العملاء وتفاعُلهم، حيث إن الربط بين مجموعات البيانات الثانوية المختلفة ينطوي على إمكانية إمداد الشركات بملفَّات تعريفٍ أكثر ثراءً عن العملاء الأفراد؛ وبالتالي يُمكِّن الشركة من توجيه خدماتها وحملات التسويق إلى عملاء مُعيَّنين. في الواقع، اليوم يتمثل أحد العوامل المحفزة لنموِّ علم البيانات في مجال الأعمال التجارية في إدراك قيمة البيانات الثانوية وقُدرَة علم البيانات على إظهار هذه القيمة للشركات.

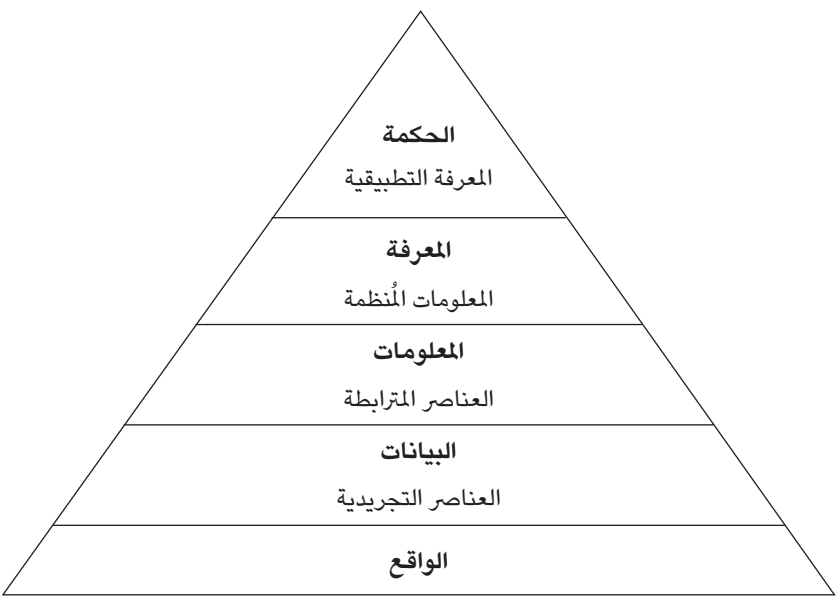
البيانات تتراكم على عكس الحكمة!

الهدف من علم البيانات هو استخدام البيانات للوصول إلى رؤية وفهم. ويَحثُّنا الكتاب المقدس على الوصول إلى الفهم من خلال السعي وراء الحكمة: «الْحِكْمَةُ هِيَ الرَّأْسُ، فَاقْتَنِ

الْحِكْمَةَ، وَبِكُلِّ مُقْتَنَاكَ اقْتَنِ الْفَهْمَ» (سفر الأمثال آية ٧:٤ [إنجيل الملك جيمس]). وهذه النصيحة في محلها؛ إلا أنها تطرح سؤالًا عن كيف ينبغي للمرء أن يبدأ السعي وراء الحكمة. الأبيات التالية من قصيدة للشاعر تي إس إليوت بعنوان: «جوقات الإنشاد» من ديوان «الصخرة» يصف فيها التسلسُل الهرمي للحكمة والمعرفة والمعلومات:

أنَّى لنا الحكمة التي أَضَعْنَاها في المعرفة؟
وأنَّى لنا المعرفة التي أَضَعْنَاها في المعلومات؟

(Eliot 1934, 96)

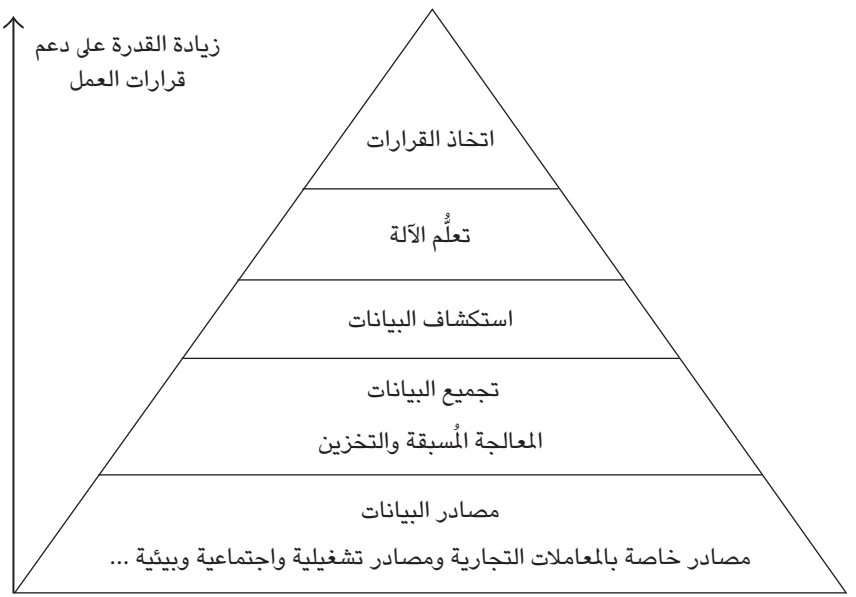


شكل ٢-١: هرم البيانات والمعلومات والمعرفة والحكمة (بتصرف من Kitchin 2014a).

يعكس التسلسل الهرمي الخاص بإليوت النموذج المعياري للعلاقات الهيكلية بين الحكمة والمعرفة والمعلومات والبيانات المعروف باسم «هرم البيانات والمعلومات والمعرفة والحكمة» (انظر شكل ٢-١). في هذا الهرم، تأتي البيانات أولًا عند سفح هذا الهرم، ثم يليها المعلومات، ثم يليها المعرفة، وتأتي الحكمة عند قمَّة الهرم. وعلى الرغم من أنه ثمة

اتفاق بوجهٍ عامٍّ على ترتيب الطبقات في هذا التسلسُل الهرَمي، فعادةً ما يكون الخلاف على الفوارق بين الطبقات والعمليات التي تتطلَّب الانتقال من طبقةٍ إلى الطبقة التالية. إلا أنه بصفة عامة:

- تنشأ البيانات من خلال التجريدات أو القياسات المأخوذة من العالم الواقعي.
- المعلومات هي بيانات جرت معالجتها، أو هيكلتها أو وضعها في سياقٍ لكي تكون ذات مغزًى بالنسبة إلى البشر.
- المعرفة هي معلومات فُسِّرت وفُهمت بواسطة البشر لكي يتمكنوا من التصرُّف وفقًا لها إذا استلزم الأمر.
- الحكمة هي التصرُّف بطريقة مناسبة بناءً على المعرفة.



شكل ٢-٢: هرم علم البيانات (بتصرُّف من Han, Kamber, and Pei 2011).

يمكن تمثيل الأنشطة في العمليات الخاصة بعلم البيانات باستخدام تسلسُلٍ هرمي مُشابه حيث يُمثل عرض الهرم كمية البيانات التي تُعالج عند كل مستوًى وكلما

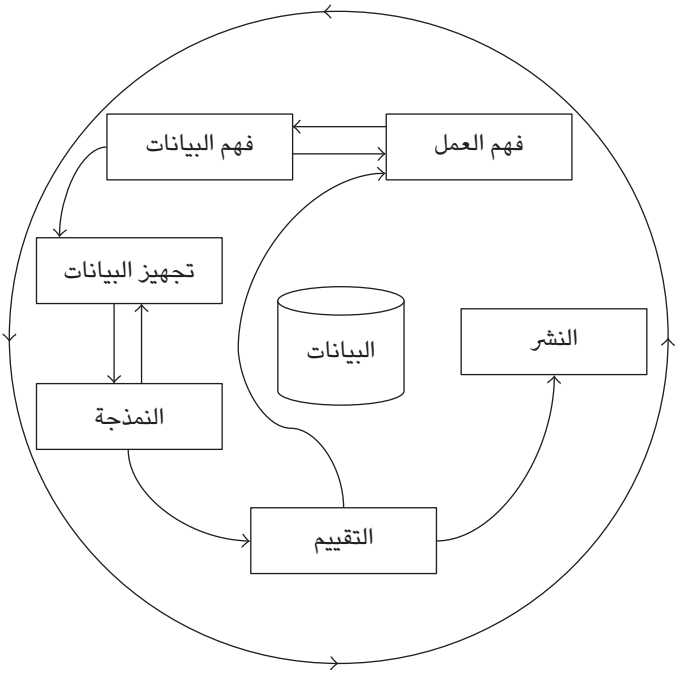
كان المستوى أعلى في الهرم، كانت نتائج الأنشطة أفيد لاتخاذ القرارات. يوضح شكل ٢-٢ التسلسل الهرمي لأنشطة علم البيانات بدايةً من استخلاص البيانات وتوليدها عبر المعالجة المسبقة والتجميع، وفهم البيانات واستكشافها، واكتشاف الأنماط، وإنشاء النماذج باستخدام تعلُّم الآلة ودعم القرارات باستخدام النماذج المستمدَّة من البيانات والمنتشرة في سياق العمل.

العملية القياسية المتعددة المجالات للتنقيب في البيانات

يتقدم الكثير من الأشخاص والشركات بانتظامٍ بمقترحات حول أفضل عمليةٍ يجب اتِّباعها لصعود هرم علم البيانات. والعملية الأشيع استخدامًا هي «العملية القياسية المتعددة المجالات للتنقيب في البيانات» (تُعرف بـ «كريسب-دي إم»)؛ والسبب الرئيسي وراء استخدامها على نطاقٍ واسع جدًّا هو أن هذه العملية مُصمَّمة لتكون مستقلةً عن أي برنامج أو مورد أو تقنية تحليل بيانات.

في البداية، طُوِّرت هذه العملية على يدِ اتحادٍ من المؤسسات يتألف من مُورِّدين روادٍ في مجال علم البيانات، ومستخدِمين نهائيين، وشركات استشارية، وباحثين. تمت رعاية مشروع «كريسب-دي إم» الأصلي جزئيًّا بواسطة المفوضية الأوروبية بموجب البرنامج الاستراتيجي الأوروبي لأنشطة البحث والتطوير في تكنولوجيا المعلومات، وقُدمت العملية لأول مرةٍ في ورشة عمل عُقدت في عام ١٩٩٩. ومنذ ذلك الحين، أُجري عدد من المحاولات لتحديث العملية، إلا أن النسخة الأصلية لا تزال شائعة الاستخدام. ولسنوات عديدة، كان هناك موقع إلكتروني مُخصص لـ «كريسب-دي إم»، إلا أنه في السنوات الأخيرة لم يعد هذا الموقع متاحًا، وفي بعض الأحيان ربما تُعيد شركة آي بي إم — التي تُعَد أحد المساهمين الأصليين في المشروع — توجيهك إلى موقع برنامج التحليل الإحصائي «إس بي إس إس». نشر الاتحاد الأصلي دليلًا تفصيليًّا للعملية خطوةً بخطوة (يتألف من ٧٦ صفحة) يسهل قراءته ومُتاح مجانًا عبر الإنترنت (انظر Chapman et al. 1999)، غير أنه يُمكن تلخيص البنية الأساسية والمهام الكبرى للعملية في بضع صفحات.

تتكوَّن عملية «كريسب-دي إم» من ستِّ مراحل: «فهم العمل التجاري، وفهم البيانات، وتجهيز البيانات، والنمذجة، والتقييم، والنشر»، كما هو مُبين في شكل ٢-٣. البيانات هي محور جميع أنشطة علم البيانات، ولهذا السبب تأتي البيانات في منتصف الرسم التوضيحي لهذه العملية. وتُشير الأسهم بين المراحل إلى الاتجاه النموذجي للعملية.



شكل ٢-٣: مراحل العملية القياسية المتعددة المجالات للتنقيب في البيانات الستة (استنادًا إلى شكل ٢-١ في Chapman, Clinton, Kerber, et al. 1999).

والعملية شِبه هيكلية، الأمر الذي يعني أن عالِم البيانات لا ينتقل دومًا عبر هذه المراحل الستة بشكلٍ خطِّي مُنتظم. استنادًا إلى النتيجة الخاصة بمرحلةٍ مُعينة، ربما يعود عالِم البيانات إلى إحدى المراحل السابقة، أو يُعيد إجراء المرحلة الحالية، أو ينتقل إلى المرحلة التالية.

في أول مرحلتَين، فهم العمل وفهم البيانات، يُحاول عالِم البيانات تحديد أهداف المشروع من خلال فهم احتياجات العمل والبيانات المتاحة. في المراحل الأولى من المشروع، غالبًا ما يتنقَّل عالِم البيانات بصورةٍ متكررة بين التركيز على فهم العمل واستكشاف البيانات المتاحة. ويشتمل هذا الانتقال عادةً تحديد مشكلة العمل ثم اكتشاف ما إذا كانت البيانات المناسبة متاحةً لتطوير حلٍّ مُستنِدٍ إلى البيانات. فإذا كانت البيانات متاحة، يمكن للمشروع المضي قدمًا؛ وإن لم تكن متاحة، سيتعيَّن على عالِم البيانات تحديد مشكلةٍ

بديلة للتعامل معها. وخلال هذه المرحلة من المشروع، سيقضي عالِم البيانات وقتًا طويلًا في الاجتماعات مع الزملاء من الأقسام التي تركز على النشاط التجاري (مثل المبيعات والتسويق والعمليات التشغيلية) لفهم مشاكلهم، ومع مديري قواعد البيانات حتى يتسنى له فهم البيانات المتاحة.

وبمجرد أن يُحدد عالِم البيانات بكل وضوحٍ مشكلة العمل ويطمئنَّ إلى أن البيانات المناسبة متوفرة، ينتقل إلى المرحلة التالية من العملية؛ ألا وهي تجهيز البيانات. ينصبُّ تركيز هذه المرحلة على إنشاء مجموعة بياناتٍ يمكن استخدامها في تحليل البيانات. وبوجهٍ عام، يشمل إنشاء هذه المجموعة من البيانات دمج مصادر البيانات من عدة قواعد بيانات. وعندما يكون لدى إحدى المؤسَّسات مخزن بيانات، ربما يكون هذا الدمج للبيانات بسيطًا إلى حدٍّ ما. وبمجرد أن أُنشئت مجموعة البيانات، يجب التحقق من جودة البيانات وتحديدها. وتشمل المشكلات النمطية لجودة البيانات القِيَم المتطرفة (الشوارد) والقيم المفقودة. والتحقق من جودة البيانات أمر مُهم للغاية، لأن وجود أخطاء في البيانات قد يكون له تأثير خطير على أداء خوارزميات تحليل البيانات.

المرحلة التالية من العملية القياسية المتعددة المجالات للتنقيب في البيانات هي مرحلة النمذجة. هذه هي المرحلة التي تُستخدَم فيها الخوارزميات الآلية لاستخراج أنماطٍ مفيدة من البيانات وإنشاء نماذج تُشفر هذه الأنماط. وتعلُّم الآلة هو مجال من علوم الكمبيوتر يُركز على تصميم هذه الخوارزميات. وفي مرحلة النمذجة، سيستخدِم عالِم البيانات عادةً عددًا من خوارزميات تعلُّم الآلة المختلفة، لتدريب عددٍ من النماذج المختلفة على مجموعة البيانات. يتدرب النموذج على مجموعة بياناتٍ من خلال تشغيل خوارزمية تعلُّم آلةٍ على مجموعة البيانات من أجل تحديد الأنماط المفيدة في البيانات وإخراج نموذجٍ يُشفر هذه الأنماط. وفي بعض الحالات، تعمل خوارزمية تعلُّم الآلة من خلال مواءمة بِنية نموذجٍ جاهز لتُناسب إحدى مجموعات البيانات، وذلك عبر ضبط معاملات النموذج الجاهز على قِيَمٍ مناسبة لمجموعة البيانات (مثل مواءمة الانحدار الخطِّي أو نموذج الشبكة العصبية ليناسب مجموعة بيانات معيَّنة). وفي حالات أخرى، تنشئ خوارزمية تعلُّم الآلة نموذجًا بالتدريج؛ جزءًا تلوَ الآخر (مثل إنشاء شجرة اتخاذ قرار، عقدةً تلوَ الأخرى، بداية من عقدة جِذر الشجرة). في معظم مشروعات علم البيانات، في النهاية يكون النموذج المولد بواسطة خوارزمية تعلُّم الآلة هو البرنامج الذي تنشره المؤسسة لمساعدتها في حل المشكلة

التي يعمل مشروع علم البيانات على حلها. وكل نموذج مُدرب بواسطة نوع مختلف من خوارزمية تعلُّم الآلة، وكل خوارزمية تبحث في البيانات عن أنواع مختلفة من الأنماط. في هذه المرحلة من المشروع، عادةً لا يعرف عالِم البيانات نوعَ الأنماط التي يجدُر به أن يبحث عنها في البيانات، ولذا، في هذا السياق، من المنطقي تجربة عددٍ من الخوارزميات المختلفة وملاحظة أي الخوارزميات تُنتِج أدقَّ النماذج عند تشغيلها على مجموعة البيانات. في الفصل الرابع، سنُقدم خوارزميات تعلُّم الآلة والنماذج بمزيدٍ من التفاصيل، ونشرح كيفية وضع خطةِ فحصٍ لتقييم دقة النموذج.

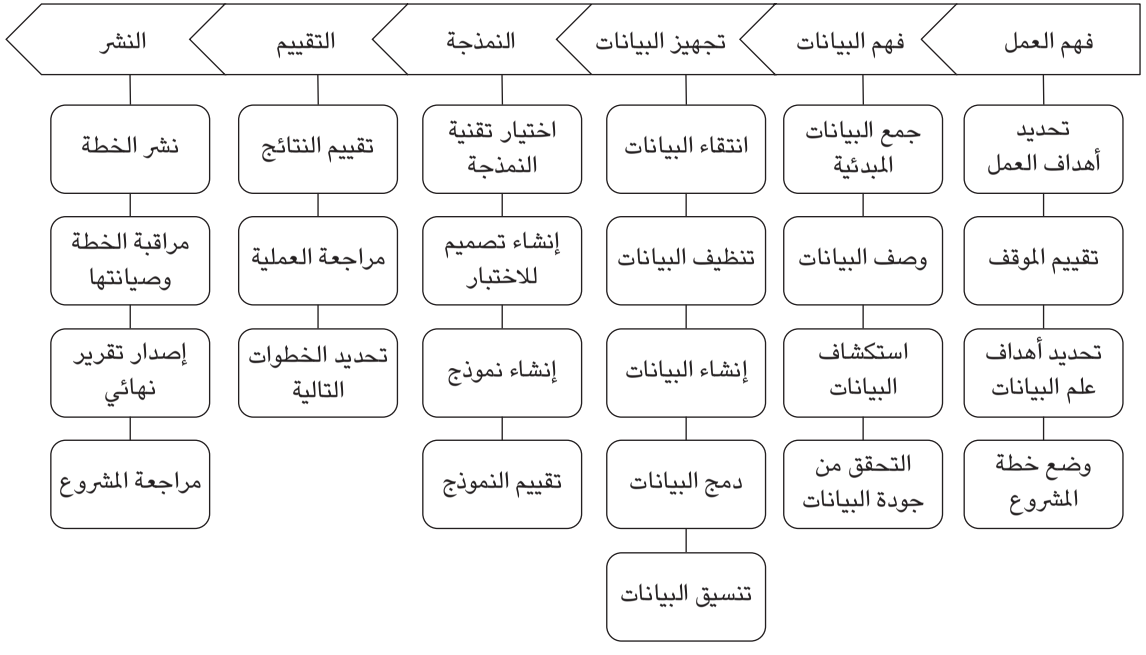
في أغلب مشروعات علم البيانات، ستكشف النتائج المبدئية لاختبار النموذج عن المشكلات الموجودة في البيانات. وأحيانًا، تظهر أخطاء البيانات عندما يُحقق عالِم البيانات في أسباب تدنِّي مستوى أداء أحد النماذج عما هو مُتوقَّع أو عندما يلاحظ أن أداء النموذج ربما يكون جيدًا على نحوٍ مثير للريبة. أو من خلال فحص بِنية النموذج، ربما يكتشف عالِم البيانات أن النموذج يعتمد على سماتٍ لم يكن يتوقعها، ونتيجة لذلك يعيد النظر في البيانات للتأكد من أن هذه السمات شُفِّرت على النحو الصحيح. ولذا، من الشائع أن يمر مشروع بهاتَين المرحلتَين في العملية عدة مرات: النمذجة، وتجهيز البيانات؛ ثم النمذجة، وتجهيز البيانات، وهكذا دواليك. على سبيل المثال، أعلن دان شتاينبرج وفريقه أنه خلال أحد مشروعات علم البيانات، أعادوا إنشاء مجموعة البيانات الخاصة بهم ١٠ مرات على مدى ستةِ أسابيع، وفي الأسبوع الخامس، وبعد تنفيذ عمليتَي تنظيف البيانات وتجهيزها عدة مرات، اكتشفوا خطأً جسيمًا في البيانات (Steinberg 2013). ولو لم يُحدَّد هذا الخطأ ويُصحَّح، لَما نجح المشروع.

تُركز المرحلتان الأخيرتان من عملية «كريسب-دي إم»، التقييم والنشر، على مدى ملاءمة النماذج للعمل وعملياته. وتركز الاختبارات التي تُجرى أثناء مرحلة النمذجة فقط على دقة النماذج بالنسبة إلى مجموعة البيانات. بينما تنطوي مرحلة التقييم على تقييم النماذج في سياقٍ أوسع تُحدِّده احتياجات العمل. فهل يحقق النموذج أهداف العمل الخاصة بالعملية؟ هل هناك أية أسباب تتعلق بالعمل وراء عدم كفاية النموذج؟ وفي هذه المرحلة من العملية، من المفيد أيضًا أن يُجري عالِم البيانات مراجعة عامة لضمان جودة أنشطة المشروع: هل هناك أي شيءٍ ناقص؟ هل يمكن تحسين أي شيء؟ وبناءً على التقييم العام للنماذج، يكون القرار الرئيسي الذي يُتخَذ أثناء مرحلة التقييم هو ما إذا كان ينبغي نشر أيٍّ من النماذج على مستوى الشركة أو ما إذا كان ينبغي تكرار عملية «كريسب-دي

إم» مرة أخرى لإنشاء نماذج أكفأ. بافتراض أن عملية التقييم اعتمدت نموذجًا أو عدة نماذج، في هذه الحالة ينتقل المشروع إلى المرحلة الأخيرة من العملية؛ ألا وهي النشر. وتشمل هذه المرحلة التحقُّق من كيفية نشر النماذج المختارة في بيئة العمل. وينطوي هذا على التخطيط لكيفية دمج النماذج في البِنية التحتية التقنية والعمليات الخاصة بالعمل. وأفضل النماذج هي النماذج التي تتلاءم بسلاسةٍ مع الممارسات الحالية للمؤسسة. وهذه النماذج لها مجموعة طبيعية من المستخدِمين الذين يواجهون مشكلةً محددة بوضوح يساعدهم النموذج في حلها. وثمة جانب آخر من النشر، ألا وهو وضع خطة لمراجعة أداء النموذج بصفةٍ دورية.

توضح الدائرة الخارجية من الرسم التوضيحي الخاص بالعملية القياسية المتعددة المجالات للتنقيب في البيانات «كريسب-دي إم» (شكل ٢-٣) إلى أيِّ مدًى تكون العملية برمَّتها متكررة. وربما تكون الطبيعة التكرارية لمشروعات علم البيانات هي الجانب الذي غالبًا ما يُتغافَل عنه في مناقشات علم البيانات. وبعد أن يُطور المشروع نموذجًا وينشره، ينبغي مراجعة النموذج بصفةٍ منتظمة للتأكُّد من أنه لا يزال يتناسب مع احتياجات العمل وأنه لم يُصبح عتيقًا. وثمة أسباب كثيرة تجعل النموذج المستنِد إلى البيانات قد يصبح عتيقًا؛ إذ ربما تكون احتياجات العمل قد تغيرت؛ أو تكون العملية التي يُحاكيها النموذج ويُقدم رؤية حولها قد تغيرت (مثل حدوث تغيرات في سلوك العملاء، وحدوث تغيرات في الرسائل العشوائية، وهكذا)؛ أو تكون تدفُّقات البيانات التي يستخدمها النموذج قد تغيرت (على سبيل المثال، ربما حُدِّث جهاز الاستشعار الذي يُغذي النموذج بالمعلومات، ومن ثم تُقدم النسخة الجديدة من جهاز الاستشعار قراءاتٍ مختلفة على نحوٍ طفيف، مما يتسبَّب في جعل النموذج أقل دقة). يعتمد تواتر هذه المراجعة على مدى سرعة تطوُّر النظام البيئي للأعمال والبيانات الذي يستخدمه النموذج. ومن الضروري إجراء مراقبةٍ مستمرة لتحديد أفضل وقتٍ لمراجعة العملية بدقةٍ مرة أخرى. وهذا ما تُمثله الدائرة الخارجية لعملية «كريسب-دي إم» المبيَّنة في شكل ٢-٣. على سبيل المثال، بناءً على البيانات، والمسألة محل الدراسة، والمجال، ربما يتعيَّن عليك مراجعة هذه العملية المتكررة بدقةٍ بصفة سنوية أو ربع سنوية أو شهرية أو أسبوعية أو حتى يومية. يُلخص شكل ٢-٤ المراحل المختلفة للعملية الخاصة بمشروع علم البيانات والمهام الرئيسية التي تنطوي عليها كل مرحلة.

من الأخطاء المتكررة التي يقع فيها الكثير من علماء البيانات المبتَدِئين تركيز جهودهم على مرحلة النمذجة في عملية «كريسب-دي إم» والتسرُّع في المراحل الأخرى.



شكل ٢-٤: مراحل العملية القياسية المتعددة المجالات للتنقيب في البيانات ومهامها (استنادًا إلى شكل ٢-٢ في Chapman, Clinton, Kerber, et al. 1999).

ولعلهم يعتقدون أن أهم ما يمكن الخروج به من أي مشروعٍ هو النموذج، وبالتالي ينبغي لعالِم البيانات أن يُخصص أغلب وقته لإنشاء النموذج وضبطه. أما علماء البيانات المخضرمون، فإنهم يقضون المزيد من الوقت في ضمان أن يُركز المشروع على هدفٍ مُحدد وأن يمتلك البيانات المناسبة. ولكي يُحقق مشروع علم البيانات نجاحًا، يجب أن يتوافر لدى عالِم البيانات فهمٌ واضح لحاجة العمل التي يُحاول المشروع أن يُلبيها. إذن فمرحلة فهم العمل هي مرحلة مهمة فعلًا من العملية. أما بخصوص الحصول على البيانات المناسبة لمشروعٍ ما، فقد وجد أحد الاستطلاعات التي أُجريت على علماء البيانات في عام ٢٠١٦ أنهم يقضون ٧٩ بالمائة من وقتهم في تجهيز البيانات. كان الوقت المستغرَق في المهام الأساسية في المشروع موزعًا كما يلي: ١٩ بالمائة مُخصص لتجميع مجموعات البيانات؛ ٦٠ بالمائة مُخصص لتنظيف البيانات وتنظيمها؛ و٣ بالمائة مُخصص لإنشاء مجموعات التدريب؛ و٩ بالمائة مخصص للتنقيب في البيانات بحثًا عن أنماط؛ و٤ بالمائة مُخصص لتحسين الخوارزميات؛ و٥ بالمائة مُخصص لأداء المهام الأخرى (Crowd-Flower 2016). وتأتي نسبة الـ ٧٩ بالمائة المخصصة لتجهيز البيانات من جمع الوقت المستغرَق في جمع البيانات وتنظيفها وتنظيمها. وظلَّت هذه النتيجة المتمثلة في أن حوالي ٨٠ في المائة من وقت المشروع ينقضي في جمع البيانات وتجهيزها، ثابتة في جميع استطلاعات الرأي التي تمَّت في مجال علم البيانات لعددٍ من السنوات. أحيانًا، تفاجئ هذه النتيجة الناس لأنهم يتخيلون أن علماء البيانات يقضون وقتهم في إنشاء النماذج المعقدة لاستخراج رؤيةٍ ثاقبة من البيانات. ولكن الحقيقة ببساطة هي أنه بغضِّ النظر عن مدى جودة تحليلك للبيانات، فإن هذا التحليل لن يُحدد الأنماط المفيدة ما لم يُجْرَ على البيانات المناسبة.

الفصل الثالث

النظام البيئي لعلم البيانات

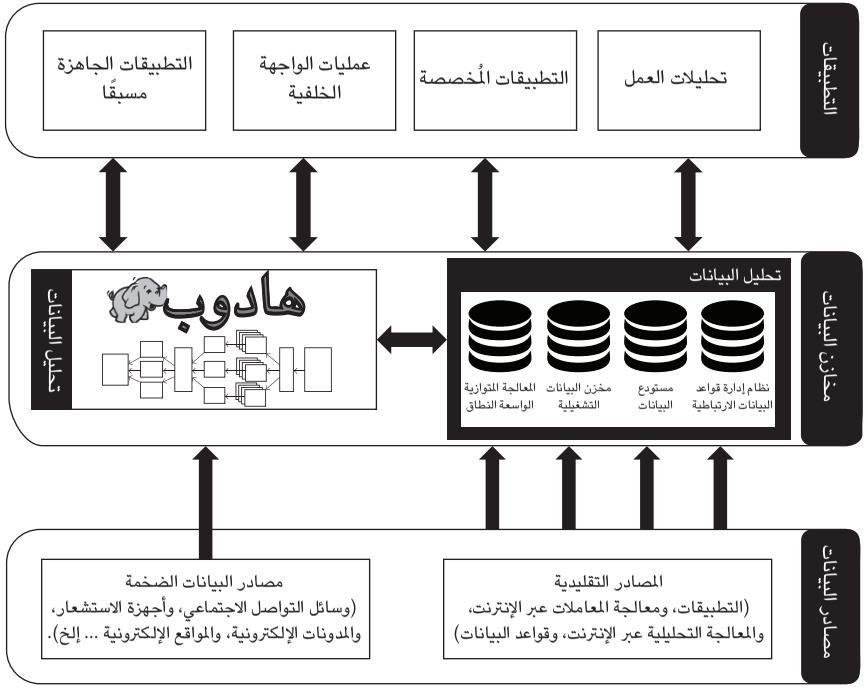
تتنوع مجموعة التقنيات المستخدمة لممارسة علم البيانات عبر مختلف المؤسسات. فكلما كانت المؤسسة أكبر أو كانت كمية البيانات التي تتم معالجتها أكثر أو كِلا الأمرَين معًا، زادت درجة تعقيد النظام البيئي التكنولوجي الداعم لأنشطة علم البيانات. وفي معظم الحالات، يحتوي هذا النظام على أدواتٍ ومكوناتٍ من عدد من مورِّدي البرامج المختلفين، مما يسفر عن معالجة البيانات بالعديد من التنسيقات المختلفة. وهناك طَيف من المناهج التي تستطيع المؤسسة أن تختار منها عند تطوير نظامها البيئي لعلم البيانات. على أحد طرفي الطيف، ربما تقرر المؤسسة الاستثمار في مجموعة أدواتٍ تجارية مدمجة. وعلى الطرف الآخر، ربما تُنشئ نظامًا بيئيًّا مخصَّصًا عن طريق دمج مجموعة من اللغات والأدوات المفتوحة المصدر. وبين هذَين النقيضَين، يوفر بعض مورِّدي البرمجيات حلولًا تتكون من مزيج من المنتجات التجارية والمنتجات المفتوحة المصدر. ومع ذلك، على الرغم من أن المزيج المحدَّد من الأدوات سيختلف من مؤسسةٍ إلى أخرى، ثمَّة قاسم مشترك فيما يخص المكونات الموجودة في معظم بِنَى علم البيانات.

يوفر شكل ٣-١ نظرة عامة رفيعة المستوى على بِنية بيانات تقليدية. وهذه البنية ليست مُخصصة لبيئات البيانات الضخمة فحسب؛ وإنما لجميع بيئات البيانات بكافة أحجامها. وفي هذا الرسم التوضيحي، تتكوَّن المساحات الثلاث الرئيسية من «مصادر البيانات»، حيث تُولَّد جميع البيانات في أي مؤسسة؛ و«مخازن البيانات»، حيث تُخزن البيانات وتُعالج؛ و«التطبيقات»، حيث تتم مشاركة البيانات مع المستخدِمين.

تمتلك جميع المؤسسات تطبيقاتٍ تولِّد وتستخلِص بياناتٍ عن العملاء والمعاملات، وبيانات تشغيلية عن كل شيءٍ له علاقة بكيفية سير العمل في المؤسسة. وتتضمن مصادر البيانات والتطبيقات إدارة العملاء، والطلبيات، والتصنيع، والتسليم، وإصدار الفواتير،

والمعاملات البنكية، والشئون المالية، وإدارة علاقات العملاء، ومركز الاتصالات، وتطبيقات تخطيط موارد المؤسسة، وما إلى ذلك. وعادةً ما يُشار إلى هذه الأنواع من التطبيقات على أنها أنظمة «معالجة المعاملات عبر الإنترنت». بالنسبة إلى الكثير من مشروعات علم البيانات، تُستخدَم البيانات المستخلَصة من هذه التطبيقات لتشكيل مجموعة البيانات الأوَّلية المدخلة لخوارزميات تعلُّم الآلة. وبمرور الوقت، يزداد حجم البيانات المستخلَصة من التطبيقات المتعددة داخل المؤسسة أكثر فأكثر وتبدأ المؤسسة في التشعُّب لاستخلاص البيانات التي جرى تجاهلها، أو التي لم استُخلِصت فيما مضى، أو التي لم تكن متاحةً من قبل. ويُشار إلى هذه البيانات الأحدث عادةً بـ «مصادر البيانات الضخمة» لأن حجم البيانات التي تُستخلَص أكبر بكثيرٍ من تطبيقات التشغيل الرئيسية الخاصة بالمؤسسة. تشمل بعض مصادر البيانات الضخمة الشائعة حركة النقل عبر الشبكة، وبيانات تسجيل الدخول من التطبيقات المتعددة، وبيانات أجهزة الاستشعار، وبيانات المدونات الإلكترونية، وبيانات وسائل التواصل الاجتماعي، وبيانات مواقع الإنترنت، وهلم جرًّا. في مصادر البيانات التقليدية، تُخزَّن البيانات عادة في قاعدة بيانات. ومع ذلك، نظرًا إلى أن التطبيقات المرتبطة بالكثير من مصادر البيانات الضخمة الأحدث ليست مصممة بالأساس لتخزين البيانات على المدى الطويل — كما هو الحال مع البيانات المتدفقة مثلًا — تختلف تنسيقات التخزين وبِنياته لهذا النوع من البيانات من تطبيقٍ إلى آخر.

ومع زيادة عدد مصادر البيانات، يزداد أيضًا التحدِّي المتمثل في القدرة على استخدام هذه البيانات لإجراء التحليلات ومشاركتها عبر المؤسسة على نطاق أوسع. وعادةً ما يُستخدم مستوى مخازن البيانات، الموضح في شكل ٣-١، للتعامل مع مشاركة البيانات وتحليلات البيانات عبر المؤسسة. وينقسم هذا المستوى إلى جزأين. يُغطي الجزء الأول برامج مشاركة البيانات المعتادة التي تستخدِمها معظم المؤسسات. والشكل الأكثر شيوعًا لبرامج دمج البيانات التقليدية وتخزينها هو نظام إدارة قواعد البيانات الارتباطية. وعادةً ما تُمثل هذه الأنظمة التقليدية حجر الأساس في حلول ذكاء الأعمال داخل أي مؤسسة. وحلول ذكاء الأعمال هي نظم سهلة الاستخدام لدعم اتخاذ القرارات وتُتيح تجميع البيانات ودمجها ونقلها وكذلك تحليلها. وبناءً على مستوى اكتمال بنية ذكاء الأعمال، يمكن أن تتألف هذه البِنية من أي شيءٍ بداية من نسخةٍ أساسية لأحد التطبيقات التشغيلية وصولًا إلى «مخزن البيانات التشغيلية» وإلى حلول قواعد بيانات المعالجة المتوازية الواسعة النطاق ومستودعات البيانات.



شكل ٣-١: بنية تقليدية للبيانات الصغيرة والضخمة من منظور علم البيانات (مُستوحًى من شكلٍ مأخوذ من نشرة «هورتونووركز»، ٢٣ أبريل، ٢٠١٣، https://hortonworks.com/blog/hadoop-and-the-data-warehouse-when-to-use-which).

عملية تخزين البيانات في مستودعات البيانات هي في الأساس عملية تجميع للبيانات وتحليلها بهدف دعم اتخاذ القرارات. ومع ذلك، ينصبُّ تركيز هذه العملية على إنشاء مستودع بيانات مركزي جيد التصميم. ومن هذا المنطلَق، يُعد مستودع البيانات موردًا مهمًّا لعلم البيانات. ومن منظور علم البيانات، إحدى المزايا الكبرى لوجود مستودع بيانات هي إنجاز المشروع في وقتٍ أقصر بكثير. تُعد البيانات المكَوِّن الأساسي لأية عملية خاصة بعلم البيانات، ولذا ليس من المستغرب أنه في الكثير من المشروعات يُستغرَق أغلب الوقت ويُبذَل أغلب الجهد في العثور على البيانات وتجميعها وتنظيفها قبل البدء في تحليلها. فإذا توفَّر مستودع بيانات بإحدى الشركات، عادةً ما يقلُّ الجهد والوقت المبذولان في تجهيز البيانات الخاصة بمشروعات علم البيانات على نحوٍ ملحوظ. ومع ذلك، من الممكن إنجاز العمليات الخاصة بعلم البيانات رغم عدم وجود مستودع بيانات مركزي. وينطوي إنشاء مستودع بيانات مركزي على أكثرَ مِن مجرد تكديس البيانات المأخوذة من عدة قواعد بيانات تشغيلية في قاعدة بيانات واحدة.

كثيرًا ما يستلزِم دمج البيانات من عدة قواعد بيانات الكثيرَ من العمل غير الآلي لحل مشكلات عدم التوافق بين قواعد البيانات المصدرية. ومصطلح «الاستخراج والتحويل والتحميل» هو المصطلح المستخدَم لوصف العمليات والأدوات التقليدية المستخدَمة لدعم تعيين البيانات ودمجِها ونقلها بين قواعد البيانات. وتختلف العمليات التقليدية التي تُنفَّذ في مستودع البيانات عن العمليات البسيطة التي تُنفذ عادة في قاعدة بيانات نموذج البيانات الارتباطية القياسية. ويُستخدم مصطلح «المعالجة التحليلية عبر الإنترنت» لوصف هذه العمليات. تركز عمليات المعالجة التحليلية عبر الإنترنت عمومًا على توليد ملخَّصات للبيانات القديمة وتتضمَّن تجميع البيانات من مصادر مُتعددة. على سبيل المثال، ربما نُقدم الطلب التالي الخاص بالمعالجة التحليلية عبر الإنترنت (سنُعبر عنه هنا باللغة العربية لتيسير القراءة): «اكتب تقريرًا عن مبيعات جميع المتاجر حسب المنطقة وحسب الفترة ربع السنوية وقارن هذه الأرقام بأرقام العام الماضي.» ما يوضحه هذا المثال هو أن نتيجة طلب المعالجة التحليلية عبر الإنترنت غالبًا ما تُشبه تقارير العمل القياسية التي نتوقَّع أن نراها. وتمكن عمليات المعالجة التحليلية عبر الإنترنت المستخدِمين من تقسيم البيانات وتجزئتها وتدويرها في المستودع للحصول على طرق عرضٍ مختلفة لهذه البيانات. وتعمل هذه العمليات على تمثيلٍ للبيانات يُسمَّى «مكعب البيانات» الذي تُنشأ فوق مستودع البيانات. ولمكعب البيانات أبعاد ثابتة مُحددة مسبقًا حيث يُمثل كل بُعدٍ خاصية معينة للبيانات.

وستكون أبعاد مكعب البيانات المطلوب في مثال طلب المعالجة التحليلية السابق على النحو التالي: «المبيعات حسب المتاجر»، و«المبيعات حسب المنطقة»، و«المبيعات حسب الفترة ربع السنوية». الميزة الأساسية وراء الاستعانة بمُكعب البيانات ذي مجموعة الأبعاد الثابتة هي أنها تسرع من زمن الاستجابة لعمليات المعالَجة التحليلية عبر الإنترنت. ونظرًا إلى أن مجموعة أبعاد مكعب البيانات مبرمجة مسبقًا في نظام المعالجة التحليلية عبر الإنترنت، يمكن أن يوفر النظام واجهاتٍ رسوميةً سهلة الاستخدام لتحديد طلبات المعالجة التحليلية عبر الإنترنت. ومع ذلك، يُقيد تمثيل مكعب البيانات أيضًا أنواع التحليلات التي يمكن إجراؤها باستخدام المعالجة التحليلية عبر الإنترنت لتقتصر على مجموعة الاستعلامات التي يمكن توليدها باستخدام أبعادٍ مُحددة مسبقًا. وبالمقارنة، تقدم لغة الاستعلام الهيكلية (إس كيو إل) واجهة استعلامٍ أكثر مرونة. أيضًا، على الرغم من أن نظم المعالجة التحليلية عبر الإنترنت مفيدة لاستكشاف البيانات وإعداد التقارير، فإنها لا تُتيح نمذجة البيانات أو الاستخراج التلقائي للأنماط من البيانات. وبمجرد تجميع البيانات من كافة أنحاء المؤسسة وتحليلها داخل نظام ذكاء الأعمال، يمكن استخدام هذا التحليل باعتباره مدخلاتٍ لمجموعة من المستهلكين عند مستوى التطبيقات الموضح في شكل ٣-١.

يتعامل الجزء الثاني من مستوى مخازن البيانات مع إدارة البيانات الناتجة عن مصادر البيانات الضخمة الخاصة بالشركة. في هذه البنية، تُستخدَم منصة هادوب لتخزين هذه البيانات الضخمة وتحليلها. وهادوب هي منصة مفتوحة المصدر طوَّرتها مؤسسة أباتشي للبرمجيات، وهي مصمَّمة خصوصًا لمعالجة البيانات الضخمة. وتستخدِم منصة هادوب نظامَ تخزينٍ ومعالجة موزعًا عبر مجموعات من وحدات الخدمة. ومن خلال استخدام نموذج برمجة «ماب رديوس»، تُسرِّع هادوب من عملية معالجة الاستعلامات في مجموعات البيانات الكبيرة. ويُنفذ نموذج «ماب رديوس» استراتيجية «التقسيم — التنفيذ — التجميع»؛ بحيث: (أ) تُقسَّم مجموعة البيانات الكبيرة إلى أجزاء منفصلة، ويُخزَّن كل جزء على عقدة (كمبيوتر) مختلفة في مجموعة الأجهزة؛ (ب) ثم يُنفذ استعلام على جميع الأجزاء بالتوازي؛ (ج) وتُحسب نتيجة الاستعلامات من خلال جمع النتائج المتولدة على الأجزاء المختلفة. غير أنه خلال العامَين الماضيَين استُخدِمت منصة هادوب أيضًا كامتداد لمستودع بيانات المؤسسات. وبالأساس، كان من شأن مستودعات البيانات أن تُخزِّن بيانات ثلاث سنوات؛ أما الآن فهي تستطيع تخزين بياناتٍ أكثرَ مِن عشر سنوات، وهذا الرقم قيد الزيادة المستمرة. ومع ذلك، عندما تزداد كمية البيانات في مستودع

البيانات، يجب أن تتزايد متطلبات التخزين والمعالجة الخاصة بقاعدة البيانات ووحدة الخدمة أيضًا. وقد يكون لهذا الشرط آثار كبيرة من حيث التكلفة. ويتمثل البديل في نقل بعضٍ من البيانات القديمة إلى مستودع بيانات لتخزينها في هادوب. على سبيل المثال، من شأن مستودع البيانات أن يُخزن أحدث البيانات، لِنَقُل بيانات ثلاثِ سنواتٍ مثلًا، التي يجب أن تكون متاحةً على نحو متكرر لتحليلها وتمثيلها بسرعة، في حين البيانات الأقدم والأقل استخدامًا تُخزن على منصة هادوب. وتحظى معظم قواعد البيانات على مستوى المؤسسة بسماتٍ تربط مستودع البيانات بمنصة هادوب، مما يُتيح لعالِم البيانات الاستعلام عن البيانات في كلا المكانين كما لو أنها موجودة جميعًا في بيئةٍ واحد، وهذا باستخدام لغة الاستعلام الهيكلية. وقد يشمل استعلامه الوصول إلى بعض البيانات في قاعدة بيانات المستودع وبعض البيانات الأخرى الموجودة على منصة هادوب. ستنقسم معالجة الاستعلام تلقائيًّا إلى جزأين منفصلين، كلٌّ منهما يعمل على نحوٍ مستقلٍّ عن الآخر، وستُجمع النتائج تلقائيًّا وتُدمَج قبل أن تظهر مرة أخرى أمام عالِم البيانات.

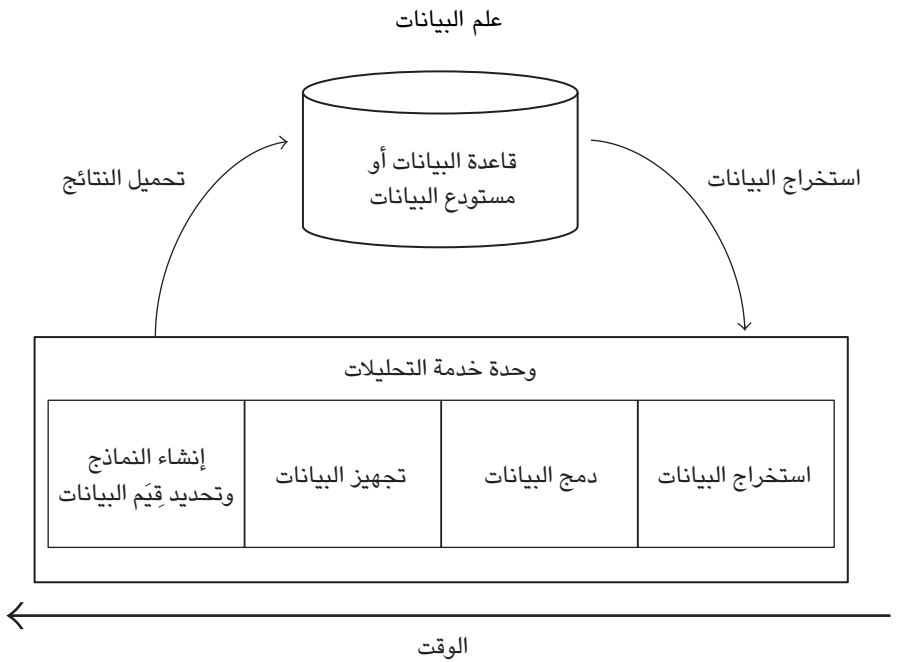
وتحليل البيانات مرتبط بكِلا جزأي مستوى مخازن البيانات الموضح في شكل ٣-١. وقد يحدث هذا التحليل للبيانات الموجودة في كل جزءٍ من مستوى البيانات، ويمكن مشاركة النتائج الخاصة بتحليل البيانات بين الجزأين في أثناء القيام بتحليلٍ إضافي للبيانات. غالبًا ما تكون البيانات المستمدَّة من المصادر التقليدية نظيفة نسبيًّا وغنية بالمعلومات مقارنة بالبيانات المستخلصة من مصادر البيانات الضخمة. ورغم ذلك، يعني حجم الكثير من مصادر البيانات الضخمة وطبيعتها الآنية أنَّ الجهد المبذول في إعداد وتحليل هذه المصادر للبيانات الضخمة ربما يكون له مردود يتمثل في الوصول إلى رؤًى إضافيةٍ لم يكن من الممكن الوصول إليها من خلال البيانات المستمَدة من المصادر التقليدية. ويمكن الاستعانة بمجموعة متنوعة من تقنيات تحليل البيانات المطورة عبر عددٍ من مجالات البحث المختلفة (من بينها معالجة اللغة الطبيعية، والرؤية الحاسوبية، وتعلُّم الآلة) لتحويل البيانات الضخمة غير الهيكلية، الشحيحة المعلومات والمنخفضة القيمة، إلى بياناتٍ ثرية بالمعلومات وعالية القيمة. ويمكن دمج هذه البيانات العالية القيمة مع بيانات أخرى عالية القيمة مُستمَدة من مصادر تقليدية بهدف إجراء المزيد من تحليل البيانات. ويُعد الوصف المذكور في هذا الفصل والموضح في شكل ٣-١ هو البنية النموذجية للنظام البيئي لعلم البيانات. ويتناسب مع أغلب المؤسسات، الصغير منها والكبير على حدٍّ سواء. غير أنه مع توسع حجم المؤسسة، تزداد أيضًا درجة التعقيد الخاصة بنظامها

البيئي لعلم البيانات. على سبيل المثال، ربما لا تحتاج المؤسسات الأصغر حجمًا إلى منصة هادوب؛ إلا أنها ستكون بالِغة الأهمية بالنسبة إلى المؤسسات الكبيرة جدًّا.

نقل الخوارزميات إلى البيانات

ينطوي المنهج التقليدي لتحليل البيانات على استخراج البيانات من مختلف قواعد البيانات، ودمجها، وتنظيفها، ووضعها في مجموعاتٍ فرعية، وإنشاء نماذج تنبُّؤ. وبمجرد أن تُنشأ نماذج تنبؤ، فإنه يمكن تنفيذها على البيانات الجديدة. ذكرنا في الفصل الأول أن نموذج التنبُّؤ يتنبَّأ بالقيمة المفقودة الخاصة بسِمةٍ ما: عامل تصفية البريد العشوائي هو نموذج تنبُّؤ يتنبَّأ بما إذا كانت سِمة التصنيف الخاصة بالبريد الإلكتروني ينبغي أن تحتوي على قيمة «عشوائي» أم لا. إن تنفيذ نماذج التنبُّؤ على المَثيلات في البيانات الجديدة من أجل توليد القِيَم المفقودة يُعرف بـ «تحديد قِيَم البيانات». بعد ذلك، ربما تُحمَّل النتائج النهائية، بعد تحديد قِيَم البيانات الجديدة، مرة أخرى على قاعدة بياناتٍ بحيث يمكن استخدام هذه البيانات الجديدة كجزءٍ من سير العمل، أو لوحة مراقبة الأداء، أو غيرها من الممارسات التقييمية للشركة. يوضح شكل ٣-٢ أن الكثير من عمليات معالجة البيانات التي تنطوي على تجهيز البيانات وتحليلها تتمُّ على وحدة خدمةٍ منفصلة عن قواعد البيانات ومستودع البيانات. وبالتالي، يمكن قضاء قدْر كبير من الوقت في مجرد نقل البيانات من قواعد البيانات وإعادة النتائج إليها مرة أخرى.

تُقدِّم تجربة أُجْرِيَتْ في معهد دبلن للتكنولوجيا بخصوص إنشاء نموذج انحدارٍ خطِّي مثالًا على الوقت المستغرق في كل مرحلةٍ من العملية. يُقضى من ٧٠ إلى ٨٠ بالمائة من الوقت تقريبًا في استخراج البيانات وتجهيزها؛ أما الوقت المتبقي فيُقضى في إنشاء النماذج. ومن أجل تحديد قيمة البيانات، يُقضى ٩٠ بالمائة من الوقت تقريبًا في استخراج البيانات وحفظ مجموعة البيانات التي حُدِّدت قيمتها مرةً أخرى في قاعدة البيانات؛ ويُقضى ١٠ بالمائة فقط من الوقت في تحديد القِيَم فعليًّا. وتستند هذه النتائج إلى مجموعات البيانات التي تتكوَّن من عددٍ يتراوح ما بين ٥٠ ألف سجلٍّ وحتى ١,٥ مليون سجل. ولقد أدرك أغلب مُقدمي خدمات قواعد البيانات للشركات الوقت الذي يتمُّ توفيره إذا لم يُقضَ الوقت في نقل البيانات ولقد حلُّوا هذه المشكلة من خلال دمج وظيفة تحليل البيانات وخوارزميات تعلُّم الآلة في مُحركات قواعد البيانات الخاصة بهم. وتستكشف الأقسام التالية من هذا الفصل كيف تُدمج خوارزميات تعلُّم الآلة في قواعد



شكل ٣-٢: العملية التقليدية لإنشاء نماذج تنبؤية وتحديد قِيَم للبيانات.

البيانات الحديثة، وكيف يعمل تخزين البيانات في عالَم البيانات الضخمة الخاص بمنصة هادوب، وكيف يُتيح الاستعانة بمزيجٍ من هذَين المنهجَين للمؤسسات العمل بسهولة ويُسر مع جميع بياناتها باستخدام لغة الاستعلام الهيكلية بوصفها لغةً مشتركة للوصول إلى البيانات والتحليل وأداء تعلُّم الآلة والتحليلات التنبُّؤية في الوقت الفعلي.

> يمكن قضاء قدْر كبير من الوقت في مجرد نقل البيانات من قواعد البيانات وإعادة النتائج إليها مرة أخرى.

قاعدة البيانات التقليدية أم قاعدة البيانات التقليدية الحديثة

يواصل مورِّدو خدمات قواعد البيانات الاستثمار في تطوير قابلية التوسُّع في قواعد بياناتهم، ومستوى أدائها، وتأمينها، وتأدية وظائفها. فقواعد البيانات الحديثة أكثر تطورًا من قواعد البيانات الارتباطية التقليدية. فهي تستطيع أن تُخزن البيانات وتستعِلم

عنها في مجموعةٍ متنوعة من التنسيقات المختلفة. فبالإضافة إلى التنسيقات الارتباطية التقليدية، من الممكن أيضًا تحديد أنواع الكائنات، وتخزين الوثائق وتخزين كائنات JSON والبيانات المكانية والاستعلام عنها، وهلمَّ جرًّا. تأتي معظم قواعد البيانات الحديثة بعددٍ كبير من الدوال الإحصائية، لدرجة أن بعضها يحتوي على عددٍ من الدوال الإحصائية مساوٍ لمعظم التطبيقات الإحصائية. على سبيل المثال، تأتي قاعدة بيانات أوراكل بأكثر من ٣٠٠ دالة إحصائية مختلفة ولغة استعلام هيكلية مُدمجة بها. وتغطي هذه الدوال الإحصائية أغلبية التحليلات الإحصائية التي تحتاجها مشروعات علم البيانات وتشمل أغلب الدوال الإحصائية — إن لم تكن كلها — المتوفرة في الأدوات واللغات الأخرى مثل لغة البرمجة آر. ربما يتيح استخدام الوظيفة الإحصائية المتوفرة في قواعد البيانات في إحدى المؤسسات أداء تحليلات البيانات بأسلوبٍ أكفأ وقابلٍ للتطوير أكثر باستخدام لغة الاستعلام الهيكلية. علاوة على ذلك، لقد دمج معظم المورِّدين الروَّاد لخدمات قواعد البيانات (من بينهم أوراكل، ومايكروسوفت، وآي بي إم، وإنتربرايز دي بي) الكثير من خوارزميات تعلُّم الآلة في قواعد بياناتهم، ويمكن تشغيل هذه القواعد باستخدام لغة الاستعلام الهيكلية. ويُعرف تعلُّم الآلة المدمج في مُحرك قواعد البيانات والذي يمكن الوصول إليه باستخدام لغة الاستعلام الهيكلية باسم «تعلُّم الآلة المدمَج في قاعدة البيانات». قد يقود هذا النوع من التعلُّم إلى إنشاءٍ أسرعَ للنماذج وانتشارٍ أسرعَ للنماذج والنتائج على حدٍّ سواء لتُستعمَل في التطبيقات ولوحات مراقبة الأداء التحليلية. وتتلخَّص الفكرة وراء تعلُّم الآلة المدمَج في قاعدة البيانات في الأمر التالي: «انقل الخوارزميات إلى البيانات بدلًا من نقل البيانات إلى الخوارزميات.»

والمزايا الرئيسية لاستخدام تعلُّم الآلة المدمَج في قاعدة البيانات هي كما يلي:

- **لا حاجة لنقل البيانات:** تستلزم بعض منتجات علم البيانات تصدير البيانات من قواعد البيانات وتحويلها إلى تنسيقٍ مُخصص لإدخالها إلى خوارزميات تعلُّم الآلة. وبالاستعانة بتعلُّم الآلة المدمَج في قاعدة البيانات، لا حاجة لنقل البيانات أو تحويلها. وهذا يجعل العملية بأكملِها أقلَّ تعقيدًا وأقل استهلاكًا للوقت وأقل عرضة للأخطاء.
- **توفير أداء أسرع:** في ظل العمليات التحليلية التي تُجرى في قاعدة البيانات وفي ظلِّ غياب نقل البيانات، من الممكن الاستفادة من قدرات الحوسبة الخاصة بوحدة خدمة قاعدة البيانات، مما يوفر أداءً أسرع حتى ١٠٠ مرة من أداء المنهج التقليدي. تتمتع

أغلب وحدات خدمة قواعد البيانات بمواصفاتٍ عالية، ذات وحدات معالجة مركزية كثيرة وقدرة على إدارة الذاكرة بكفاءة بهدف معالجة مجموعات البيانات التي تحتوي على أكثر من مليار سجل.

- **توفير أمانٍ عالٍ:** تُوفر قاعدة البيانات وصولًا إلى البيانات الموجودة في قاعدة البيانات على نحوٍ خاضع للتحكم وقابل للمراجعة والتدقيق، مما يُسرع إنتاجية عالِم البيانات مع توفير عنصر الحماية للبيانات. يتجنَّب تعلُّم الآلة المدمَج في قاعدة البيانات المخاطر الأمنية المادية الكامنة في استخلاص البيانات وتنزيلها على وحدات خدمةٍ تحليلية بديلة. وعلى النقيض من ذلك، يُسفر عن العملية التقليدية إنتاج العديد من النُّسَخ (وربما إصدارات مختلفة) من مجموعات البيانات في مستودعاتٍ منفصلة عبر المؤسسة.
- **قابلية التوسع:** يمكن أن تتوسَّع قاعدة البيانات بسهولةٍ في إجراء التحليلات مع زيادة حجم البيانات؛ هذا إذا أُدخلت خوارزميات تعلُّم الآلة إلى قاعدة البيانات. تُصمم برامج قواعد البيانات من أجل إدارة كمياتٍ كبيرة من البيانات بكفاءة، عن طريق استغلال وحدات المعالجة المركزية المتعددة والذاكرة الموجودة على وحدة الخدمة ليُتاح تشغيل خوارزميات تعلُّم الآلة بالتوازي. كما أن قواعد البيانات شديدة الفعالية في معالجة مجموعات البيانات الكبيرة التي لا تحتويها الذاكرة بسهولة. لقد تطوَّرت قواعد البيانات على مدار أكثر من ٤٠ عامًا لضمان تمكُّنها من معالجة مجموعات البيانات بسرعة.
- **نشر وبيئات الوقت الفعلي:** يمكن نشر النماذج التي تمَّ تطويرها باستخدام خوارزميات تعلُّم الآلة المدمَجة في قاعدة البيانات واستخدامها في بيئات الوقت الفعلي. ويُتيح هذا دمج النماذج في التطبيقات اليومية، مما يوفِّر تنبؤات للمستخدِمين والعملاء النهائيِّين في الوقت الفعلي.
- **النشر في بيئة الإنتاج:** قد يتطلَّب نشر النماذج المطوَّرة باستخدام برامج تعلُّم الآلة المستقلة إعادة تشفيرها بلغاتِ برمجةٍ أخرى قبل دمجِها في تطبيقات المؤسسة. لكنَّ هذا ليس هو الحال مع نماذج تعلُّم الآلة المدمَجة في قاعدة البيانات. فلُغة الاستعلام الهيكلية هي لغة قاعدة البيانات الأساسية، ويمكن استخدامُها واستدعاؤها من أية لغةِ برمجةٍ أو أداة أخرى من أدوات علم البيانات. وبالتالي يمكن دمج النماذج المدمجة في قاعدة البيانات بسهولةٍ في تطبيقات الإنتاج.

تستغل الكثير من المؤسسات مزايا تعلُّم الآلة المدمَج في قاعدة البيانات. وتتنوَّع ما بين المؤسسات الصغيرة والمتوسطة وحتى المؤسسات الكبيرة التي تستخدِم البيانات الضخمة. فيما يلي بعض الأمثلة على المؤسسات التي تستخدِم تقنيات تعلُّم الآلة المدمَج في قاعدة البيانات:

- شركة فيسيرف، وهي شركة أمريكية تُقدم الخدمات المالية وخدمات التحليل والكشف عن الاحتيال. تحولت شركة فيسيرف من الاستعانة بعدَّة مورِّدين للخدمات الخاصة بتخزين البيانات وتعلُّم الآلة إلى الاستعانة بإمكانيات تعلُّم الآلة المدمجة في قواعد بياناتها. ومن خلال الاستعانة بتعلُّم الآلة المدمَج في قاعدة البيانات، تضاءل الوقت المستغرق لإنشاء/تحديث ونشر نموذج كشف الاحتيال من أسبوع تقريبًا إلى بضع ساعاتٍ فحسب.
- شركة ٨٤.٥١° (كانت تُعرف سابقًا باسم شركة دانهامبي الولايات المتحدة الأمريكية)، وهي شركة مُتخصصة في علم بيانات العملاء. تستعين الشركة بالعديد من المنتجات التحليلية المختلفة لإنشاء نماذج العملاء المختلفة. كان من المعتاد أن تستغرق أكثر من ٣١٨ ساعة شهريًّا لنقل البيانات من قواعد بياناتها إلى أدوات تعلُّم الآلة والعكس مرةً أخرى، بالإضافة إلى ٦٧ ساعة شهريًّا لإنشاء النماذج. وعندما تحولت الشركة إلى استخدام خوارزميات تعلُّم الآلة المدمجة في قاعدة بياناتها، لم يعُد هناك حاجة إلى نقل البيانات. وبقِيَت البيانات في قاعدة البيانات. ووفرت الشركة على الفور ٣١٨ ساعة شهريًّا. ونظرًا إلى أنها كانت تستخدِم قاعدة بياناتها كمحرِّكٍ حوسبي، استطاعت أن تتوسَّع في تحليلاتها، ومن ثم تضاءل الوقت المستغرَق في إنشاء أو تحديث نماذج تعلُّم الآلة من أكثر من ٦٧ ساعة إلى ساعةٍ واحدة شهريًّا. وهذا وفَّر للشركة ستة عشر يومًا كل شهر. لقد أصبحت الآن قادرةً على الحصول على نتائج أسرع ويمكنها الآن أن تقدم لعملائها نتائج في وقتٍ أقرب بكثيرٍ بعد إجراء عملية شراء.
- شركة وورجيمينج، صاحبة ابتكار لعبة «وورلد أوف تانكس» (عالم الدبابات) وغيرها من الألعاب. تستعين الشركة بتعلُّم الآلة المدمَج في قاعدة البيانات لنمذجة وتوقع كيفية التفاعُل مع عملائها الذين يفوق عددهم ١٢٠ مليون عميل.

البنية التحتية للبيانات الضخمة

على الرغم من أن قاعدة البيانات التقليدية (الحديثة) تتَّسم بفعالية مذهلة في معالجة بيانات المعاملات التجارية، فإن الحاجة تدعو إلى وجود بِنية تحتية جديدة لإدارة جميع أشكال البيانات وتخزينها على المدى الطويل في عصر البيانات الضخمة. ويمكن لقاعدة البيانات التقليدية المعاصرة أن تتعامل مع أحجام البيانات الكبيرة والتي يصل حجمها إلى بِضع بيتابايت؛ إلا أنه ربما تُصبح حلول قواعد البيانات باهظةً على نحوٍ تعجيزي بالنسبة إلى هذا الحجم من البيانات. وعادةً ما يُشار إلى مشكلة التكلفة هذه بـ «التوسُّع العمودي». في نموذج البيانات التقليدي، كلما زادت كمية البيانات التي تُضطَر إحدى المؤسسات إلى تخزينها ومعالجتها خلال فترة زمنية معقولة، زاد حجم وحدة خدمة قاعدة البيانات اللازمة لذلك، وفي المقابل زادت التكلفة من أجل إعدادات وحدة الخدمة وترخيص قاعدة البيانات. ربما تستطيع المؤسسات استيعاب والاستعلام عن مليار سجلٍّ بصفة يومية/أسبوعية باستخدام قواعد البيانات التقليدية، غير أنها ربما تُضطَر إلى استثمار أكثر من ١٠٠ ألف دولار فقط لشراء العتاد اللازم لإجراء هذا الحجم من المعالجة.

تُعَد هادوب منصة مفتوحة المصدر طورتها وأطلقتها مؤسسة أباتشي للبرمجيات. وهي منصة مُجربة لاستيعاب وتخزين أحجام مهولة من البيانات بطريقة فعَّالة وقد تكون أقل تكلفةً بكثير من منهج قاعدة البيانات التقليدية. في منصة هادوب، تُقسَّم البيانات وتُجزَّأ بطرقٍ متنوعة، وتُوزَّع هذه الأجزاء من البيانات عبر العقد على منصة هادوب. تُعالِج أدوات التحليل المتنوعة التي تتعامل مع منصة هادوب البيانات الموجودة على كل عقدة من العُقَد (في بعض المثيلات يمكن أن تتواجد هذه البيانات على الذاكرة)، مما يُتيح معالجةً سريعة للبيانات نظرًا إلى أن عمليات التحليل تتم بالتوازي عبر العُقَد. ولا حاجة لاستخراج البيانات أو لعمليات «الاستخراج والتحويل والتحميل». تُجرى عملية تحليل البيانات حيث يتمُّ تخزينها.

وعلى الرغم من أن منصَّة هادوب هي أشهر إطارِ معالجةٍ للبيانات الضخمة، فهي ليست الوحيدة بأية حال من الأحوال. تشمل إطارات معالجة البيانات الضخمة الأخرى كلًّا من «ستورم»، و«سبارك»، و«فلينك». وكل هذه الأُطُر جزء من مشروعات مؤسَّسة أباتشي للبرمجيات. ويكمن الاختلاف بين هذه الأطر في حقيقة أن منصة هادوب مُصمَّمة أساسًا من أجل معالجة البيانات على دفعات. والمعالجة على دفعاتٍ مناسبة عندما تكون مجموعة البيانات ثابتةً بلا تغيُّر أثناء عملية المعالجة وعندما تكون نتائج المعالجة ليست مطلوبةً

فورًا (أو على الأقل عندما لا يكون عنصر الوقت حرجًا للغاية). أما نظام «ستورم» فهو مُصمَّم لمعالجة البيانات المتدفقة. وفي معالجة البيانات المتدفقة، تتم معالجة كل عنصرٍ بمجرد أن يدخل النظام، وبالتالي تُعرَّف عمليات المعالجة للعمل على كلِّ عنصرٍ فردي في البيانات المتدفقة بدلًا من العمل على مجموعة البيانات بأكملها. على سبيل المثال، ربما تُعطي المعالجة على دفعاتٍ متوسط قِيَم مجموعة من البيانات، في حين أن المعالجة المتدفقة تُعطي تسميةً فردية أو قيمة فردية لكل عنصر في البيانات المتدفقة (مثل حساب درجة التفاعل مع كل تغريدة من التغريدات المتدفقة على موقع تويتر). ونظام «ستورم» مُصمَّم من أجل معالجة البيانات في الوقت الفعلي ووفقًا لموقع ستورم الإلكتروني،[1] لقد أصبح معيارًا مرجعيًّا لمعالجة أكثر من مليون حقلٍ مترابط في الثانية الواحدة وفي كل عقدة. و«سبارك» و«فلينك» إطاران للمعالجة المختلطة (المعالجة على دفعاتٍ والمعالجة المتدفقة). ونظام «سبارك» هو بالأساس نظام معالجة بالدفعات، مُشابه لمنصة هادوب؛ إلا أنه يتمتع ببعض قدرات المعالجة المتدفقة؛ في حين أن «فلينك» هو إطار معالجة متدفِّقة ولكن يمكن استخدامه أيضا لمعالجة البيانات على دفعات. وعلى الرغم من أن هذَين الإطارَين لمعالجة البيانات الضخمة يقدمان لعلماء البيانات خيارًا من الأدوات التي تُلبي مُتطلبات البيانات الضخمة الخاصة بالمشروع، فإن الاستعانة بهذين الإطارَين قد يكون له عيب يتمثل في اضطرار عالِم البيانات الآن إلى تحليل البيانات في مكانَين مختلفَين، أي في قواعد البيانات التقليدية الحديثة ومخزن البيانات الضخمة. ويُلقي القسم التالي نظرةً على كيفية حل هذه المشكلة تحديدًا.

عالَم قواعد البيانات المختلطة

إذا كانت إحدى المؤسسات لا تمتلك بيانات بالحجم والمقدار اللازمَين للاستعانة بمنصة هادوب، سيتطلب الأمر برنامج قواعدِ بياناتٍ تقليديًّا لإدارة بياناتها. ومع ذلك، تذكر بعض المؤلَّفات أن أدوات تخزين ومعالجة البيانات المتاحة في عالم هادوب ستحلُّ محلَّ قواعد البيانات الأكثر تقليديةً. ومِن الصعب جدًّا رؤية حدوث هذا، وفي الآونة الأخيرة صار هناك الكثير من المناقشات الدائرة حول اتِّباع منهج أكثرَ توازنًا لإدارة البيانات فيما يُسمى «عالَم قواعد البيانات المختلطة». وهذا العالَم هو المكان الذي تُوجَد فيه قواعد البيانات التقليدية وعالَم هادوب معًا.

في عالَم قواعد البيانات المختلطة، ترتبط قواعد بيانات الشركة بالبيانات المخزَّنة على منصة هادوب وتعملان معًا، مما يتيح المعالجة الفعَّالة للبيانات ومشاركتها وتحليلها.

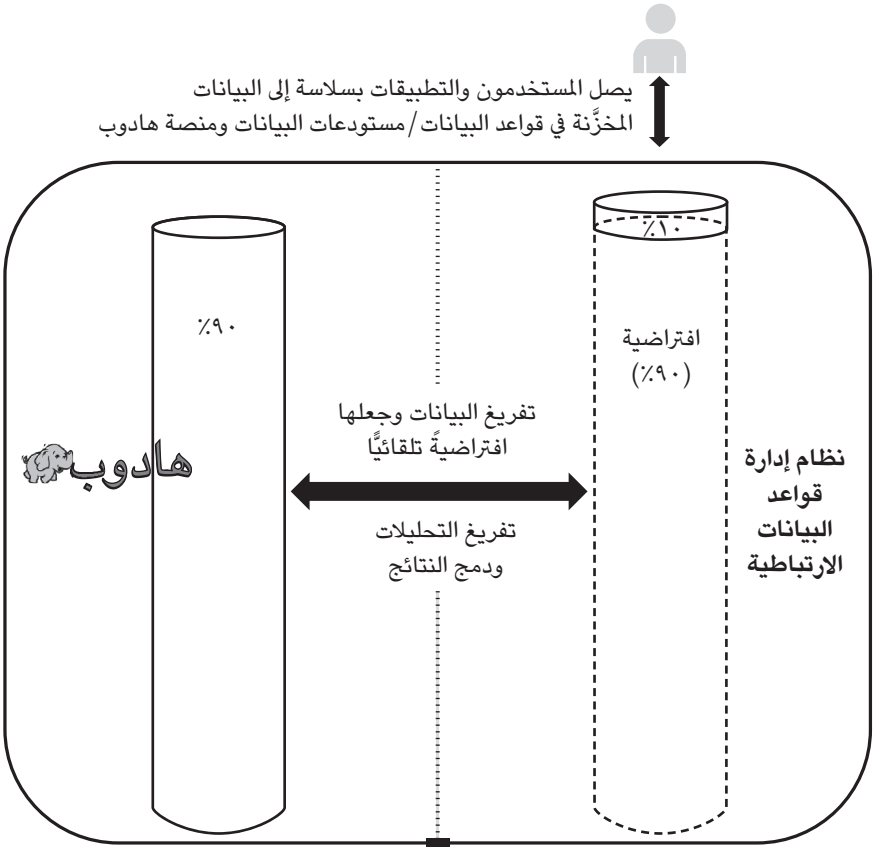
ويوضح شكل ٣-٣ مخزن بياناتٍ تقليديًّا؛ ولكن بدلًا من تخزين جميع البيانات على قاعدة البيانات أو في مُستودع البيانات، يُنقل أغلبها إلى منصة هادوب. ويُنشأ رابطٍ بين قاعدة البيانات ومنصة هادوب، ليُتيح لعالِم البيانات الاستعلام عن البيانات كما لو كانت موجودة جميعًا في مكانٍ واحد. وعالِم البيانات ليس بحاجةٍ إلى الاستعلام عن جزء البيانات الموجودة في المستودع ثم الاستعلام في خطوةٍ منفصلة عن الجزء المخزَّن على منصة هادوب. ويمكنه الاستعلام عن البيانات كما كان يفعل دائمًا، وسيُحدِّد الحل أي جزءٍ من الاستعلام سيُنفَّذ في مستودع البيانات وأي جزء سيُنفَّذ تنفيذه في منصة هادوب. وستُدمَج نتائج الاستعلام التي تم التوصل إليها في كِلا الموقعَين وتُقدَّم إلى عالِم البيانات. وبالمثل، مع زيادة حجم مُستودع البيانات، لن يُستعلَم عن بعض البيانات الأقدم بكثرة. وبالتالي ينقل حلُّ قاعدة البيانات المختلطة تلقائيًّا البيانات الأقل استخدامًا إلى منصة هادوب وينقل البيانات الأكثر استخدامًا إلى مستودع البيانات. وتوازن قاعدة البيانات المختلطة تلقائيًّا موقع البيانات بناءً على تواتر الوصول إلى البيانات ونوع عمليات البيانات التي تُجرى.

> توازن قاعدة البيانات المختلطة تلقائيًّا موقع البيانات بناءً على تواتر الوصول إلى البيانات ونوع عمليات البيانات التي تُجرى.

إحدى مزايا هذا الحل المختلط هو أن عالم البيانات سيظل يستخدِم لغة SQL للاستعلام عن البيانات. ولا يحتاج إلى تعلُّم لغة أخرى للاستعلام عن البيانات أو إلى استخدام مجموعةٍ متنوعة من الأدوات المختلفة. وبناءً على الاتجاهات الحالية، سيمتلك مورِّدو خدمات قواعد البيانات ومورِّدو حلول دمج البيانات وجميع موردي خدمات التخزين السحابي للبيانات حلولًا مُشابهة لهذا الحل المختلط في المستقبل القريب.

تجهيز البيانات ودمجها

يشمل دمج البيانات استخراج البيانات من مصادرها المختلفة ودمجها معًا لتوفر رؤية موحدة للبيانات من مختلف أقسام المؤسسة. وتُعد السجلات الطبية مثالًا جيدًا على هذا الدمج. كوضعٍ مثالي، من المفترض أن يكون لكل شخصٍ سجل صحِّي واحد، وأن يستخدم كلُّ مستشفًى أو منشأة طبية أو طبيب مُمارس عام رقم التعريف نفسه للمريض أو وحدات القياس نفسها، ونظام التصنيف نفسه، وهكذا. ولكن لسوء الحظ، يمتلك كل



شكل ٣-٣: قواعد البيانات ومستودعات البيانات ومنصة هادوب وهي تعمل معًا (مُستوحًى من شكلٍ في مستنَدٍ تقني صادرٍ عن منصة بيانات جلونت، ٢٠١٧، https://gluent.com/wp-content/uploads/2017/09/Gluent-Overview.pdf).

مستشفًى تقريبًا نظامه المستقل لإدارة شئون المرضى، وكذلك كل مُختبر من المختبرات الطبية داخل المستشفى. تأمُل التحديات الكامنة في العثور على سجل أحد المرضى وتحديد النتائج الصحيحة للمريض الصحيح. وهذه هي التحديات التي يُواجهها مستشفًى واحد فحسب. وفي السيناريوهات التي تتشارك فيها مستشفيات متعددة في بيانات المرضى، تُصبح مشكلة الدمج مشكلةً عويصة. وبسبب هذا النوع من التحديات، تستغرق المراحل الثلاث الأولى من العملية القياسية المتعددة المجالات للتنقيب في البيانات (كريسب-دي إم)

من ٧٠ إلى ٨٠ بالمائة من إجمالي زمن مشروع علم البيانات، مع تخصيص غالبية هذا الوقت لعملية دمج البيانات وحدها.

يُعَد دمج البيانات من عدة مصادر أمرًا صعبًا حتى عندما تكون البيانات هيكلية. ومع ذلك، عندما يتعلق الأمر ببعض مصادر البيانات الضخمة الأحدث، حيث تكون البيانات شِبه الهيكلية أو غير الهيكلية هي القاعدة، فإن تكلفة دمج البيانات وإدارة البِنية التحتية يمكن أن تُصبح كبيرة. وتُعد بيانات العملاء مثالًا توضيحيًّا على تحديات دمج البيانات. يمكن أن تُوجَد بيانات العملاء في العديد من التطبيقات المختلفة (وقواعد البيانات المقابلة لتلك التطبيقات). سيحتوي كل تطبيقٍ على جزء مختلف قليلًا من بيانات العملاء. على سبيل المثال، قد تحتوي مصادر البيانات الداخلية على التصنيف الائتماني للعميل، ومبيعات العميل، والمدفوعات، ومعلومات الاتصال الخاصة بمركز الاتصال، إلى آخِره. وربما تُتاح أيضًا بيانات إضافية عن العميل من مصادر البيانات الخارجية. في هذا السياق، يستلزِم تكوين رؤية متكاملة عن العميل استخراج البيانات من كلِّ مصدرٍ من هذه المصادر ودمجها معًا.

ستتضمَّن عملية دمج البيانات النموذجية عددًا من المراحل المختلفة، تتكوَّن من استخراج البيانات وتنظيفها وتوحيدها ونقلها وفي النهاية دمجها لتكوين نسخةٍ موحدة واحدة من البيانات. يمكن أن يكون استخراج البيانات من مصادر البيانات المتعددة أمرًا صعبًا لأن العديد من مصادر البيانات لا يمكن الوصول إليها إلا باستخدام واجهة معينة خاصة بذلك المصدر. ونتيجة لذلك، يجب أن يتمتع علماء البيانات بمجموعة واسعة من المهارات حتى يكونوا قادرين على التفاعل مع كل مصدر من مصادر البيانات من أجل الحصول على البيانات.

وبمجرد أن تُستخرَج البيانات من المصدر، يجب التحقُّق من جودة البيانات. وتنظيف البيانات هي عملية اكتشاف البيانات التالِفة أو غير الدقيقة، أو تنظيفها، أو استبعادها من البيانات المستخرَجة. على سبيل المثال، ربما يتعيَّن تنظيف معلومات عنوان العميل من أجل تحويلها إلى صيغةٍ موحدة. بالإضافة إلى ذلك، ربما يكون هناك بيانات مُكررة في مصادر البيانات، في تلك الحالة من الضروري تحديد سجلِّ العميل الصحيح الذي يجب استخدامه وإزالة جميع السجلات الأخرى من مجموعات البيانات. ومن المهم التأكُّد من أن القِيَم المستخدمة في مجموعة البيانات مُتماثلة. على سبيل المثال، ربما يستخدِم أحدُ تطبيقات المصدر قِيمًا عددية لتمثيل التصنيف الائتماني للعميل؛ في حين يستخدِم تطبيقٌ

آخَرُ مزيجًا من القِيَم العددية والحروف. في هذا السيناريو، يجب اتخاذ قرارٍ بشأن نوع القيم التي ستُستخدَم، وبعد ذلك تغيير القِيم التي يختلف نوعها عن النوع الذي حُدِّد لهذا العنصر. على سبيل المثال، تخيَّل أن إحدى السِّمات في مجموعة البيانات هي مقاس حذاء العميل. يمكن أن يشتري العملاء أحذيةً من مختلف المناطق حول العالم، غير أن النظام العددي المستخدَم لتحديد مقاسات الأحذية في أوروبا يختلف قليلًا عن ذلك المستخدم في الولايات المتحدة والمملكة المتحدة وغيرها من الدول. وقبل إجراء تحليل البيانات ونمذجتها، يجب توحيد قيم هذه البيانات.

وينطوي نقل البيانات على تغيير البيانات أو تجميعها من قيمة إلى أخرى. ويمكن استخدام مجموعة متنوعة من التقنيات أثناء هذه الخطوة وتشمل تسوية البيانات وتوزيعها في فئاتٍ وتطبيعها وكذلك كتابة كودٍ مُخصص لإجراء عملية نقلٍ مُعينة. ويتضح مثال شائع على نقل البيانات في عملية معالجة عُمر أحد العملاء. في الكثير من مهام علم البيانات، التمييز الدقيق بين أعمار العملاء ليس مفيدًا بشكلٍ خاص. فالفارق بين عميلٍ في عمر الثانية والأربعين وآخر في الثالثة والأربعين ليس مُهمًّا بوجهٍ عام، على الرغم من أن التمييز بين عميلٍ في الثانية والأربعين وآخر في الثانية والخمسين قد يكون مفيدًا. ونتيجة لذلك، غالبًا ما يُنقل عمر العميل من العمر الأصلي إلى فئةٍ عمرية عامة. وتُعَد هذه العملية لتحويل الأعمار إلى فئاتٍ عمرية مثالًا على تقنية نقل بيانات تُسمى «التوزيع في فئات». وعلى الرغم من أن التوزيع في فئات هي عملية مباشرة نسبيًّا من المنظور التقني، فإن التحدي هنا يتمثل في تحديد الحدود الأفضل لنطاق الفئة لتطبيقه أثناء عملية التوزيع في فئات. وقد يؤدي تطبيق الحدود الخاطئة إلى حجب فروقٍ مُهمة في البيانات. ومع ذلك، ربما يستلزم العثور على الحدود المناسبة معرفةً خاصة بالمجال أو الاعتماد على التجربة والخطأ.

وتتمثل الخطوة الأخيرة للدمج في إنشاء البيانات التي تُستخدَم كمدخلاتٍ لخوارزميات تعلُّم الآلة. وتُعرف هذه البيانات بـ «الجدول الرئيسي للتحليل».

إنشاء الجدول الرئيسي للتحليل

أهم خطوة في إنشاء الجدول الرئيسي للتحليل هي اختيار السمات التي ستُضمَّن في التحليل. يعتمد الاختيار على معرفة المجال وعلى تحليل العلاقات بين السمات. فلنضرب مثلًا بسيناريو يُركز التحليل فيه على عملاء إحدى الخدمات. في هذا السيناريو، يُعتبر من

المفاهيم الشائعة الاستخدام في المجال والتي ستجعل تصميمك واختيارك للسمات مُستنيرًا تفاصيل التعاقد مع العميل والمعلومات الديموغرافية والاستخدام والتغيرات الطارئة على الاستخدام، والاستخدام الخاص، والمرحلة الحالية في العملية القياسية المتعددة المجالات للتنقيب في البيانات، وروابط الشبكة، وما إلى ذلك. وعلاوة على ذلك، من المرجَّح أن تكون السمات التي وُجد أنها مرتبطة ارتباطًا كبيرًا بالسمات الأخرى متكررة ومن ثم ينبغي استبعاد واحدة من السمات المترابطة. وقد يسفر حذف السمات المتكررة عن نماذج أبسط يسهل فهمها، ويُقلل أيضًا من احتمالية إنتاج خوارزمية تعلُّم الآلة نموذجًا يتناسَب مع أنماطٍ زائفة في البيانات. تحدد مجموعة السمات المختارة لتضمينها ما يُعرف باسم «سجل التحليل». ويشمل سجل التحليل عادة كلًّا من السمات الخام والسمات المشتقة على حدٍّ سواء. وكل مثيل في الجدول الرئيسي للتحليل يُمثله سجل تحليل واحد، ومن ثم فإن مجموعة السمات المتضمنة في سجل التحليل تُحدد شكل المثيلات التي سيُجرى عليها التحليل.

وبعد أن صُمِّم سجل التحليل، يجب استخراج مجموعة من السجلات وتجميعها لإنشاء مجموعة بيانات مناسبة للتحليل. وعندما تُنشأ هذه السجلات وتخزَّن — في قاعدة بيانات مثلًا — يُشار عمومًا إلى مجموعة البيانات هذه بـ «الجدول الرئيسي للتحليل». وهذا الجدول عبارة عن مجموعة البيانات المستخدمة كمدخلاتٍ في خوارزميات تعلُّم الآلة. يقدم الفصل التالي مجال تعلُّم الآلة ويصف بعضًا من أشهر خوارزميات تعلُّم الآلة المستخدمة في علم البيانات.

الفصل الرابع

أساسيات تعلُّم الآلة

أفضل ما قيل عن علم البيانات هو أنه شراكة بين عالِم البيانات وجهاز الكمبيوتر. في الفصل الثاني، وصفنا العملية التي يتَّبعها عالم البيانات: مراحل العملية القياسية المتعددة المجالات للتنقيب في البيانات. وتُحدد هذه العملية القياسية سلسلةً من القرارات يتعين على عالم البيانات أن يتَّخذها والأنشطة التي ينبغي أن يشارك فيها لجعل هذه القرارات مستنيرةً ولتنفيذها. في هذه العملية، تتمثل المهام الكبرى لعالِم البيانات في تحديد المشكلة وتصميم مجموعةِ البيانات وتجهيز البيانات وتحديد نوع تحليل البيانات المراد تطبيقه، وتقييم نتائج تحليل البيانات وتفسيرها. وما يُساهم به جهاز الكمبيوتر في هذه الشراكة هو القدرة على معالجة البيانات والبحث عن أنماطٍ مُحددة في البيانات. وتعلُّم الآلة هو مجال دراسة يُطوِّر الخوارزميات التي تتَّبعها أجهزة الكمبيوتر لتحديد الأنماط واستخلاصها من البيانات. وتُطبَّق خوارزميات تعلُّم الآلة وتقنياتها بالأساس أثناء مرحلة النمذجة في العملية القياسية المتعددة المجالات للتنقيب في البيانات. وينطوي تعلُّم الآلة على عمليةٍ تتألف من خطوتَين.

أولًا: تُطبَّق خوارزمية تعلُّم الآلة على مجموعة بياناتٍ لتحديد الأنماط المفيدة الموجودة في البيانات. وهذه الأنماط يمكن تمثيلها بعدة طرق مختلفة. وفي موضعٍ لاحق من هذا الفصل، سوف نصف بعض التمثيلات الشائعة؛ ولكنها تشمل الهيكل الشجري لاتخاذ القرار، ونماذج الانحدار، والشبكات العصبية. وتُعرف هذه التمثيلات للأنماط باسم «النماذج»، وهذا هو السبب أن هذه المرحلة من مراحل العملية القياسية المتعددة المجالات للتنقيب في البيانات تُعرف باسم «مرحلة النمذجة». ببساطة، تنشئ خوارزميات تعلُّم الآلة نماذج باستخدام تمثيلٍ مُعين (شبكة عصبية أو هيكل شجري أو أي شيءٍ غيرهما).

ثانيًا: بمجرد أن يُنشأ النموذج، يُستخدم من أجل التحليل. وفي بعض الحالات، ما يُهم هي بِنية النموذج. فبِنية النموذج يمكن أن تكشف عن السمات المهمة في مجالٍ ما. على سبيل المثال، في المجال الطبي، ربما نقوم بتطبيق خوارزمية تعلُّم الآلة على مجموعة بيانات خاصة بمرضى السكتة الدماغية ونستخدِم بنية النموذج لنُحدد العوامل التي لها علاقة قوية بالسكتة الدماغية. وفي حالات أخرى، يُستخدم النموذج لوصف أمثلة جديدة أو تصنيفها. الغاية الأساسية من نموذج تصفية البريد العشوائي هو وصف رسائل البريد الإلكتروني الجديدة إما بأنها رسائل عشوائية أو غير عشوائية بدلًا من كشف السمات المحددة لرسائل البريد العشوائي.

التعلُّم الخاضع للإشراف في مقابل التعلُّم غير الخاضع للإشراف

تُصنَّف أغلبية خوارزميات تعلُّم الآلة ضمن إحدى فئتَين: «تعلُّم خاضع للإشراف» أو «تعلُّم غير خاضع للإشراف». يهدف التعلُّم الخاضع للإشراف إلى إنشاء دالَّةٍ وتعليمها كيفية تعيين قيمة السمة التي تصف مثيلًا (السمة المستهدفة) بالاستدلال بقِيَم سماتٍ أخرى لذلك المثيل. على سبيل المثال، عندما يُستخدم التعلُّم الخاضع للإشراف لتدريب أداة تصفية البريد العشوائي، تحاول الخوارزمية إنشاء دالةٍ تعيِّن قيمةً للسمة المستهدفة (عشوائي/غير عشوائي) بالاستدلال بقِيَم السمات التي تصف البريد الإلكتروني؛ وتكون الدالة التي تُنشئها الخوارزمية هي نموذج تصفية البريد العشوائي الذي تُنتجه الخوارزمية. إذن، في هذا السياق، النمط الذي تبحث عنه الخوارزمية في البيانات هو دالَّة تُعيِّن قيمة السمة المستهدفة بالاستدلال بقِيَم السمات المدخَلة، والنموذج الناتج عن الخوارزمية هو برنامج كمبيوتر يُنفذ هذه الدالة. يشمل التعلُّم الخاضع للإشراف البحث عبر الكثير من الدوالِّ المختلفة لإيجاد الدالة التي تستطيع تعيين أفضل مُخرجات ملائمة للمدخلات. ومع ذلك، بالنسبة إلى أية مجموعة بيانات ذات درجةٍ معقولة من التعقيد يُوجَد عددٌ كبير جدًّا من تكوينات المدخلات وما يقابلها من التعيينات المحتملة للمخرجات التي تعجز معها الخوارزمية أن تُجرب جميع الدوال المحتملة. ونتيجة لذلك، صُممت كل خوارزمية من خوارزميات تعلم الآلة للبحث عن أنواعٍ مُعينة من الدوال أو تفضيل تلك الأنواع بعينها أثناء بحثها. وتُعرف تلك التفضيلات بـ «التحيُّز الاستقرائي» (أو تحيُّز التعلُّم) الخاص بالخوارزمية. ويتمثل التحدي الفعلي أمام استخدام تعلُّم الآلة في العثور على الخوارزمية

التي يتناسب تحيُّزها الاستقرائي على أفضل نحوٍ مع مجموعةٍ مُعينة من البيانات. وبوجهٍ عام، تشمل هذه المهمة إجراء تجارب على عددٍ من الخوارزميات المختلفة للعثور على أفضل واحدةٍ تتماشى مع تلك المجموعة من البيانات.

> يتمثل التحدي الفعلي أمام استخدام تعلُّم الآلة في العثور على الخوارزمية التي يتناسب تحيُّزها الاستقرائي على أفضل نحوٍ مع مجموعة مُعينة من البيانات.

هذا النوع من تعلُّم الآلة «خاضع للإشراف» لأن كلَّ مثيلٍ في مجموعة البيانات يُدرِج كلًّا من قِيَم المدخلات وقيمة المخرج (المستهدف) لكلِّ مثيل. وبالتالي، خوارزمية التعلُّم يمكن أن تقود بحثها إلى أفضل دالةٍ من خلال مراجعة إلى أيِّ مدًى تتناسب كل دالة جرت تجربتُها مع مجموعة البيانات، وفي الوقت نفسه تؤدي مجموعة البيانات دور المشرف لعملية التعلُّم من خلال تقديم تقارير. ومن الواضح أنه من أجل حدوث التعلُّم الخاضع للإشراف يجب أن يُوصَف كل مثيلٍ في مجموعة البيانات بالقيمة الخاصة بالسمة المستهدفة. ومع ذلك، عادة ما يكون السبب وراء كون السمة المستهدفة مثيرةً للاهتمام هو أنها ليس من السهل تقدير قيمتِها مباشرة، وبالتالي لا يمكن إنشاء مجموعة بياناتٍ مكونة من مثيلاتٍ وُصِّفت بكلِّ سهولة. وفي هذا السيناريو، يستلزم الأمر قدرًا كبير من الوقت والجهد لإنشاء مجموعة بيانات بالقِيَم المستهدَفة قبل أن يتم تدريب النموذج باستخدام التعلُّم الخاضع للإشراف.

في التعلُّم غير الخاضع للإشراف، لا يوجد سمة مستهدفة. وكنتيجة لذلك، يمكن استخدام خوارزميات التعلُّم غير الخاضع للإشراف بدون استثمار وقتٍ وجهد في توصيف مثيلات مجموعة البيانات حسب السمة المستهدفة. ومع ذلك، عدم وجود سمة مُستهدفة يعني أيضًا أن عملية التعلُّم صارت أصعب: بدلًا من المشكلة المحددة الخاصة بالبحث عن تعيينات مُخرجات للمدخلات تُناسِب مجموعة البيانات، صار للخوارزمية مهمة أكثر عمومية تتمثل في البحث عن ثوابت في البيانات. والنوع الأكثر شيوعًا للتعلُّم الخاضع للإشراف هو «تحليل المجموعات» أو «التحليل العنقودي»، حيث تبحث الخوارزمية عن مجموعات المثيلات التي يشبِه بعضها بعضًا أكثر من تشابُهها بمثيلاتٍ أخرى في البيانات. عادة تبدأ خوارزميات التجميع بتخمين عددٍ من المجموعات أو العناقيد، ثم تحدِّث المجموعات أو العناقيد على نحوٍ مُتكرر (عن طريق حذف مثيلات من مجموعةٍ وإضافتها

إلى مجموعةٍ أخرى) لكي يزداد التشابُه داخل المجموعة الواحدة والتنوع عبر المجموعات المختلفة.

ثمة تحدٍّ مرتبط بمسألة التجميع يتمثل في معرفة كيفية قياس درجة التشابه. فإذا كانت جميع السمات في مجموعة البيانات هي سمات عددية وتتمتع بنطاقاتٍ متشابهة، ربما يكون من المنطقي على الأرجح حساب المسافة الإقليدية (المعروفة باسم «مسافة الخط المستقيم») بين المثيلات (أو الصفوف). تُعامل الصفوف القريبة بعضها من بعضٍ على المسافة الإقليدية على أنها مُتشابهة. ومع ذلك، ثمة عددٌ من العوامل قد تجعل حساب درجة التشابُه بين الصفوف أمرًا مُعقدًا. ففي بعض مجموعات البيانات، للسمات العددية المختلفة نطاقات مختلفة، مما ينتُج عنه ألا يكون التبايُن في قِيَم الصفوف في أحد السمات على نفس القدْر من أهمية التبايُن بنفس المقدار في سمةٍ أخرى. في هذه الحالات، ينبغي تطبيع السمات بحيث يكون لها جميعًا النطاق نفسه. وثمة عامل تعقيد آخر في حساب درجة التشابُه ألا وهو أنه يمكن اعتبار الأشياء متشابهةً بعدة طرق مختلفة. أحيانًا تكون بعض السِّمات أهم من سماتٍ أخرى، لذا قد يكون من المنطقي تقدير بعض السمات في ضوء حسابات المسافة الإقليدية، أو لعلَّ مجموعة البيانات تشمل بياناتٍ غير عددية. ربما تتطلَّب السيناريوهات الأكثر تعقيدًا تصميم معايير مُخصصة للتشابُه لاستخدامها بواسطة خوارزمية التجميع.

ويمكن توضيح التعلُّم غير الخاضع للإشراف عن طريق مِثال واقعي. تخيَّل أننا مُهتمون بتحليل أسباب إصابة الذكور الأمريكيين البالِغين ذوي البشرة البيضاء بمرَض السُّكَّر من النوع الثاني. سنبدأ بإنشاء قاعدة بيانات، وفيها كل صفٍّ يُمثل شخصًا واحدًا وكل عمود يُمثل سمةً نعتقد أنها ذات صلةٍ بالدراسة. ولهذا المثال، سندرج السمات التالية: طول الفرد بالمتر ووزنه بالكيلوجرام، وعدد الدقائق التي يُمارس فيها الرياضة كل أسبوع، ومقاس حذائه، واحتمالية الإصابة بمرض السُّكر مُمثلة بنسبة مئوية بناءً على عدد الاختبارات السريرية ودراسات مَسحية عن نمط الحياة. ويوضح جدول ٤-١ جزءًا من هذه المجموعة من البيانات. من الواضح أنه يمكن إدراج سماتٍ أخرى — مثل عمر الشخص — ويمكن استبعاد بعض السمات — مثل مقاس الحذاء الذي لن يكون ذا أهمية خاصة لتحديد ما إذا كان شخصٌ ما سيُصاب بمرض السُّكَّر أم لا. وكما ناقشنا في الفصل الثاني، يُعد اختيار أي السمات التي ستُضمَّن أو تُستبعَد من مجموعة البيانات هي مهمة أساسية في علم البيانات، ولكن لأغراض هذه المناقشة سنعمل على مجموعة البيانات هذه دون تغيير.

جدول ٤-١: مجموعة بيانات خاصة بدراسة الإصابة بمرض السكر.

رقم تعريفي	الطول (بالمتر)	الوزن (بالكيلوجرام)	مقاس الحذاء	التمارين الرياضية (عدد الدقائق في الأسبوع)	مرض السكَّر (احتمالية الإصابة بالنسبة المئوية)
١	١,٧٠	٧٠	٥	١٣٠	٠,٠٥
٢	١,٧٧	٨٨	٩	٨٠	٠,١١
٣	١,٨٥	١١٢	١١	٠	٠,١٨

ستبحث خوارزمية التجميع غير الخاضعة للإشراف عن مجموعات الصفوف المتشابهة معًا أكثر من تشابهها مع الصفوف الأخرى في البيانات. وتُحدد كل مجموعةٍ من هذه المجموعات ذات الصفوف المتشابهة مجموعة من المثيلات المشابهة. على سبيل المثال، تستطيع خوارزميةٌ ما أن تُحدد أسباب المرض أو الأمراض المصاحبة (الأمراض التي تظهر معًا) من خلال إلقاء نظرة على قِيَم السمات المتكررة بصورةٍ نسبية داخل مجموعةٍ ما. إن الفكرة البسيطة المتمثلة في البحث عن مجموعاتٍ من الصفوف المتشابهة هي فكرة عظيمة جدًّا ولها تطبيقات في مناحٍ كثيرة بالحياة. ويتمثل تطبيقٌ آخر لتجميع الصفوف في تقديم توصيات بمنتجات معينة إلى العملاء. إذا أُعجِب عميل بكتاب أو أغنية أو فيلم، فلعلَّه يستمتع بكتاب آخر أو أغنية أخرى أو فيلم آخر من المجموعة نفسها.

نماذج التنبؤ الخاصة بتعلُّم الآلة

التنبؤ هو مهمة تقدير قيمة السمة المستهدفة من أجل مثيلٍ مُعين بناءً على قيم السمات الأخرى (أو سمات الإدخال) لذلك المثيل. وهذه هي المشكلة التي تحلها خوارزميات تعلُّم الآلة؛ فهي تُولِّد نماذج تنبؤ. وهنا يمكننا أيضًا استخدام مثال نموذج تصفية البريد العشوائي الذي استخدمناه لتوضيح التعلُّم الخاضع للإشراف: نحن نستخدم هذا النوع من التعلُّم لتدريب نموذج تصفية البريد العشوائي، ونموذج تصفية البريد العشوائي هو نموذج تنبؤ. ويتمثل الاستخدام الشائع لنموذج التنبؤ في تقدير قيمة السِّمة المستهدفة في المثيلات الجديدة غير الموجودة في مجموعة بيانات التدريب. واستكمالًا لمثال البريد

العشوائي، نُدرب نموذج تصفية البريد العشوائي (نموذج التنبؤ) على قاعدة بياناتٍ من رسائل البريد الإلكتروني القديمة ثم نستخدِم هذا النموذج للتنبؤ بما إذا كانت الرسائل الجديدة تندرج تحت البريد العشوائي أم لا. وربما تكون مشكلات التنبؤ هي النوع الأكثر شيوعًا الذي يُستخدم من أجله تعلُّم الآلة، ولذلك يركز باقي هذا الفصل على التنبؤ باعتباره دراسة حالةٍ لتوضيح تعلُّم الآلة. وسوف نبدأ توضيح نماذج التنبؤ بتوضيح مفهومٍ من المفاهيم الأساسية في التنبؤ؛ ألا وهو «تحليل الارتباط». ثم نشرح كيف تعمل خوارزميات تعلُّم الآلة لإنشاء أنواعٍ مختلفة من نماذج التنبؤ الشائعة، بما فيها نماذج الانحدار الخطي، ونماذج الشبكة العصبية، والهياكل الشجرية الخاصة باتخاذ القرار.

العلاقات الارتباطية ليست علاقاتٍ سببية، ولكن بعضها مفيد

تصف «العلاقة الارتباطية» قوة الارتباط بين سمتَين.[1] وبوجهٍ عام، يمكن أن تصف العلاقة الارتباطية أي نوع من الارتباط بين سمتَين. ولمصطلح «العلاقة الارتباطية» معنًى إحصائي مُحدَّد، والذي يُستخدم عادةً كاختصار لـ «معامل ارتباط بيرسون». ويقيس معامل ارتباط بيرسون قوة العلاقة الخطية بين سمتَين عدديتَين. وتتراوح قيمته من −١ إلى +١. يُستخدم حرف «r» للإشارة إلى قيمة بيرسون أو معامل الارتباط بين السمتَين. ويشير معامل الارتباط $r = 0$ إلى أن السمتَين غير مرتبطتَين بعضهما ببعض. ويُشير معامل الارتباط $r = +1$ إلى أن السمتَين بينهما علاقة ارتباطية موجبة مثالية، بمعنى أن كلَّ تغيير يحدُث في إحدى السمتَين يُصاحبه تغيير مُماثل في السمة الأخرى في الاتجاه نفسه. ويُشير معامل الارتباط $r = -1$ إلى أن السمتَين تجمعهما علاقة ارتباطية سالبة مثالية، بمعنى أن كلَّ تغيير يحدُث في إحدى السمتين يصاحبه تغيير معاكس في السمة الأخرى. وتتمثل الإرشادات العامة لتفسير معاملات ارتباط بيرسون في أن قيمة $r \approx \pm 0.7$ تشير إلى علاقةٍ خطية قوية بين السمتَين؛ وتُشير $r \approx \pm 0.5$ إلى علاقة خطية متوسطة، وتُشير $r \approx \pm 0.3$ إلى علاقة ضعيفة، وتُشير $r \approx 0$ إلى عدم وجود علاقة بين السمتَين.

وفي حالة دراسة احتمالية الإصابة بمرَض السُّكَّر، من شأننا أن نتوقَّع من واقع معرفتنا بآلية عمل الجسم البشري أنه سيكون هناك علاقات بين بعض السِّمات المدرجة في جدول ٤-١. على سبيل المثال، من المعروف بوجهٍ عام أنه كلما كانت قامة الشخص أطول، كان مقاس حذائه أكبر. ومن شأننا أيضًا أن نتوقَّع أنه كلما مارس الشخص تمارين رياضيةً أكثر، أصبح أخف وزنًا، رغم أن شخصًا طويل القامة من المحتمل أن

يكون أثقل وزنًا من شخصٍ قصير القامة يُمارس رياضة بالقدْر نفسه. ومن شأننا أيضًا أن نتوقع أنه لن يكون هناك علاقة واضحة بين مقاس حذاء الشخص ومقدار ممارسته للتمارين الرياضية. يُقدم شكل ٤-١ ثلاثة مخططاتِ تشتُّتٍ توضح كيف تنعكس هذه البديهيات على البيانات. ويوضح مخطَّط التشتُّت العلوي كيف تنتشر البيانات إذا كان المخطط يعتمِد على مقاس الحذاء وطول الشخص. وثمة نمط واضح في هذا المخطط: تتحرك البيانات من الزاوية السفلية ناحية اليسار إلى الزاوية العلوية ناحية اليمين، مما يُشير إلى العلاقة التي مفادها أنه عندما يكون الأفراد أطول قامة (أو عندما نتَّجه يمينًا على المحور السيني)، فإنهم يميلون إلى ارتداء مقاسات أحذية أكبر (نتحرك إلى أعلى على المحور الصادي). وبوجهٍ عام، يُشير نمط البيانات المتجهة من الأسفل يسارًا إلى الأعلى يمينًا في مخطَّط التشتُّت على علاقةٍ ارتباطية موجبة بين هاتين السمتَين. وإذا حسبنا ارتباط بيرسون بين مقاس الحذاء وطول القامة، يكون معامل الارتباط $r = 0.898$ يشير إلى علاقة ارتباطية موجبة قوية بين هاتَين السمتَين. ويُبين مخطَّط التشتُّت الأوسط كيف تنتشر البيانات عندما نرسم بيانيًّا العلاقة بين الوزن وممارسة التمارين الرياضية. وهنا يتمثل النمط العام في الاتجاه المعاكس، من أعلى اليسار إلى أسفل اليمين، مما يُشير إلى علاقة ارتباط سالبة: كلما زادت التمرينات الرياضية التي يُمارسها الفرد، صار أخفَّ وزنًا. ويكون معامل ارتباط بيرسون لهاتَين السمتَين كما يلي $r = -0.710$، مما يُشير إلى علاقة سالبة قوية. ويوضح مخطَّط التشتُّت الأخير، بالأسفل، بيانيًّا العلاقة بين ممارسة التمارين الرياضية ومقاس الحذاء. في هذا المخطَّط، البيانات موزعة عشوائيًّا على نحوٍ نِسبي، ومعامل ارتباط بيرسون لهاتَين السمتَين هو $r = -0.272$، مما يشير إلى عدم وجود علاقة ارتباطية حقيقية بين السمتَين.

حقيقة أن تعريف ارتباط بيرسون الإحصائي على أنه ارتباط بين سمتَين تجعل استخدام هذه العلاقة الإحصائية لتحليل البيانات مقتصرًا فقط على أزواج السمات الثنائية. ولكن لحُسن الحظ يمكننا تخطي هذه المشكلة من خلال استخدام الدوالِّ على مجموعاتٍ من السمات. في الفصل الثاني، قدَّمنا مؤشر كتلة الجسم بوصفه دالة لوزن الشخص وطول قامته. والمقصود بها تحديدًا نسبة وزن الشخص (بالكيلوجرام) مقسومة على مُربع طوله (بالأمتار). ابتُكِر مؤشر كتلة الجسم في القرن التاسع عشر على يد عالِم رياضيات بلجيكي، يُدعى أدولف كوتيليه، ويُستخدم هذا المؤشر لتصنيف الأفراد إلى فئات: ناقص الوزن، أو ذي وزن طبيعي، أو زائد الوزن، أو يُعاني من السمنة. وتُستخدَم

شكل ٤-١: مخطَّطات التشتُّت الخاصة بالعلاقة الارتباطية بين مقاس الحذاء وطول القامة، والوزن والتمارين الرياضية، ومقاس الحذاء والتمارين الرياضية.

النسبة بين الوزن والطول لأن مؤشر الكتلة مُصمَّم ليكون ذا قيمة مماثلة بالنسبة إلى الأشخاص الذين يندرجون تحت الفئة نفسها (ناقص الوزن أو ذو وزن طبيعي أو زائد الوزن أو يعاني من السمنة) بغض النظر عن طول قامتهم. نحن نعرف أن ثمة علاقة ارتباطية موجبة بين الوزن والطول (بوجه عام، كلما كان الشخص أطول قامة، كان أثقل وزنًا)، إذن من خلال قسمة الوزن على الطول، نحسب تأثير الطول على الوزن. ونقسم على مربع الطول لأن الأشخاص يزيدون عرضًا كلما صاروا أطول، ولذا، تربيع الطول هي محاولة لحساب إجمالي حجم الشخص في هذه المعادلة. وثمة جانبان لمؤشر كتلة الجسم مثيران للاهتمام في مناقشتنا للعلاقة الارتباطية بين عدة سمات. أولًا: مؤشر الكتلة هو دالة تأخذ عددًا من السمات كمدخلاتٍ وتُعيِّن على أساسها قيمةً جديدة. في الواقع، يُنشئ هذا التعيين سمةً جديدة مُشتقة في البيانات (بخلاف السمات الخام). ثانيًا: نظرًا إلى أن مؤشر كتلة جسم الشخص هو قيمة عددية مفردة، يُمكننا أن نحسب العلاقة الارتباطية بينها وبين السِّمات الأخرى.

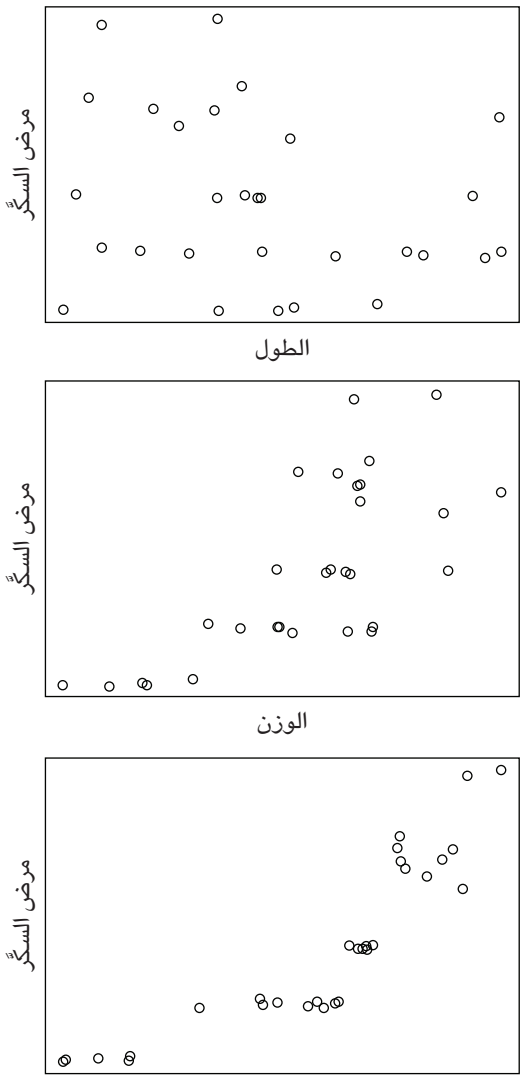
في دراسة الحالة الخاصة بأسباب إصابة الذكور الأمريكيين البالِغين ذوي البشرة البيضاء بمرض السكَّر من النوع الثاني، نحن مُهتمون بتحديد ما إذا كان أي من السمات ذا علاقةٍ ارتباطية قوية بالسمة المستهدَفة التي تصف احتمالية إصابة شخصٍ ما بمرض السكَّر. ويقدم شكل ٤-٢ ثلاثة مخطَّطات تشتُّت، يوضح كل منها بيانيًّا العلاقة الارتباطية بين السمة المستهدَفة (مرض السكَّر) وسمة أخرى: الطول والوزن ومؤشر كتلة الجسم. في مخطَّط التشتُّت الخاص بالطول ومرض السكَّر، لا يبدو أنه يوجَد نمط معيَّن في البيانات، مما يشير إلى أنه لا تُوجَد علاقة ارتباطية حقيقية بين هاتَين السمتَين (معامل ارتباط بيرسون هو $r = -0.277$). ويُبيِّن مخطَّط التشتُّت الأوسط توزيع البيانات بيانيًّا باستخدام الوزن واحتمالية الإصابة بالسكَّر. ويشير انتشار البيانات إلى وجود علاقة ارتباطية موجبة بين هاتَين السمتَين؛ بمعنى أنه كلما زاد وزن الشخص، زادت احتمالية إصابته بمرض السكَّر (معامل ارتباط بيرسون هو $r = 0.655$). ويوضح مخطَّط التشتُّت الأخير مجموعة البيانات مرسومة بيانيًّا باستخدام مؤشر كتلة الجسم والإصابة بالسكَّر. والنمط في هذا المخطَّط مُشابه لمخطَّط التشتُّت الأوسط: البيانات المنتشرة من الأسفل يسارًا إلى الأعلى يمينًا، تُشير إلى علاقة ارتباطية موجبة. غير أنه في هذا المخطَّط، المثيلات شديدة الارتباط ببعضها البعض، مما يُشير إلى أن العلاقة الارتباطية بين مؤشر كتلة الجسم ومرض السكَّر أقوى من العلاقة الارتباطية بين الوزن ومرض السكَّر. في الواقع،

معامل ارتباط بيرسون لمرض السكَّر ومؤشر كتلة الجسم لهذه المجموعة من البيانات هو $r = 0.877$.

يوضح مثال مؤشر كتلة الجسم أنه من الممكن ابتكار سِمة مُشتقة جديدة من خلال تحديد دالَّة تأخذ سمات متعددة كمدخلٍ لها. كما يُبيِّن أيضًا أنه من الممكن حساب معامل ارتباط بيرسون بين هذه السمة المشتقة وسمة أخرى في مجموعة البيانات. علاوة على ذلك، يمكن لسمة مشتقة أن تحظى بعلاقة ارتباطية مع سمة مستهدَفة على نحوٍ أوثق من العلاقة التي تربط بين أيٍّ من السمات المستخدَمة لتوليد السمة المشتقة وبين السمة المستهدَفة. وتتمثل إحدى الطرق لفهم سبب وجود علاقة ارتباطية أكثر إيجابية بين مؤشر كتلة الجسم وسمة الإصابة بمرض السكَّر مقارنة إما بالطول أو بالوزن في أن احتمالية إصابة شخصٍ ما بهذا المرض متوقفة على التأثير المتبادَل بين الطول والوزن، ويُمثل مؤشر كتلة الجسم هذا التأثير المتبادَل على نحوٍ مناسب فيما يخصُّ احتمالية الإصابة بمرض السكَّر. ويهتم المعالجون بمؤشر كتلة الجسم الخاص بالأفراد لأنه يوفر لهم المزيد من المعلومات عن احتمالية إصابة الشخص بمرض السكَّر من النوع الثاني أكثر مما يوفِّره طول الشخص أو وزنه فحسْب على نحوٍ مستقل.

ذكرنا بالفعل أن اختيار السمة هي مُهمة أساسية في علم البيانات. ويُعد تصميم السمة مهمةً أساسية أيضًا. تكمن القيمة الحقيقية لعلم البيانات غالبًا في تصميم سمةٍ مشتقة ذات علاقة ارتباطية قوية بسمةٍ ما تثير اهتمامنا. وبمجرد أن تُحدِّد السمات المناسبة لتستخدِمها لتُمثِّل البيانات، يمكنك تصميم نماذج دقيقة بسرعة نسبيًّا. ويُعد اكتشاف السمات المناسبة وتصميمها هو الجزء الصعب. وفي حالة مؤشر كتلة الجسم، صمَّم البشر هذه السمة المشتقة في القرن التاسع عشر. غير أن خوارزميات تعلُّم الآلة يمكنها فهم التأثيرات المتبادلة بين السمات وإنشاء سماتٍ مشتقة مفيدة من خلال البحث عبر توليفاتٍ مختلفة من السمات والتأكد من العلاقة الارتباطية بين هذه التوليفات والسمة المستهدفة. ولهذا السبب تعلُّم الآلة مفيد في سياقات حيث تُسهم الكثير من السمات ذات التأثير المتبادل الضعيف في العملية التي نُحاول فهمها.

من المفيد تحديد سمةٍ (خام أو مشتقة) ذات علاقة ارتباطية وثيقة بسمةٍ مستهدَفة لأن السمة المرتبطة ربما تُعطينا رؤية عن العملية التي تسببت في الظاهرة التي تُمثلها السمة المستهدفة: تشير حقيقة أن مؤشر كتلة الجسم مرتبط ارتباطًا وثيقًا باحتمالية إصابة الشخص بمرض السكَّر إلى أن الوزن في حدِّ ذاته لا يُسهم في إصابة الشخص



شكل ٤-٢: مخطَّطات التشتُّت الخاصة باحتمالية الإصابة بمرض السكَّر فيما يتعلق بطول القامة، والوزن، ومؤشر كتلة الجسم.

بالسكَّر وإنما ما يُسهم في الإصابة هي معاناة الشخص من السمنة. أيضًا إذا كانت السمة المدخَلة مرتبطة ارتباطًا وثيقًا بسمةٍ مستهدفة، فمن المرجَّح أن تكون مدخلًا مفيدًا في نموذج التنبؤ. وعلى غرار تحليل الارتباط، ينطوي التنبؤ على تحليل العلاقات بين السمات. ولكي نتمكن من تعيين السمة المستهدَفة من القِيَم الخاصة بمجموعة سماتٍ مُدخلة، يجب أن يكون هناك علاقة ارتباطية بين السمات المدخلة (أو دالة مُشتقة تطبق عليها) والسمة المستهدفة. وإذا لم تكن هذه العلاقة الارتباطية موجودة (أو لا تستطيع الخوارزمية العثور عليها)، إذن السمات المدخَلة ليست ذات صلةٍ بمسألة التنبؤ، وأفضل ما يستطيع النموذج أن يفعله هو تجاهل تلك المدخلات والتنبؤ بالاتجاه الرئيسي لتلك السمة المستهدفة[2] في مجموعة البيانات. وعلى العكس، إذا كان هناك ارتباط وثيق بين السمات المدخلة والسمة المستهدفة، من المرجَّح أن تكون خوارزمية تعلُّم الآلة قادرة على إنشاء نموذج تنبؤ دقيق للغاية.

الانحدار الخطي

عندما تتكوَّن مجموعة بيانات من سماتٍ عددية، حينئذٍ كثيرًا ما تُستخدَم نماذج التنبؤ المعتمدة على الانحدار. ويقدِّر «تحليل الانحدار» القيمة المتوقعة (أو المتوسطة) لسمةٍ عددية مستهدفة عندما تكون جميع سمات الإدخال ثابتة. والخطوة الأولى في تحليل الانحدار هي افتراض بِنية العلاقة بين السمات المدخلة والسمة المستهدفة. حينئذٍ يُحدَّد النموذج الرياضي القائم على المعاملات للعلاقة المفترَضة. يُسمَّى هذا النموذج القائم على المعاملات بـ «دالة الانحدار». يمكنك التفكير في دالة الانحدار باعتبارها آلةً تُحوِّل المدخلات إلى قيمةٍ مُخرجة والتفكير في المعاملات باعتبارها الإعدادات التي تتحكم في سلوك الآلة. وربما تحتوي دالة الانحدار على عدة مُعاملات، وينصبُّ تركيز تحليل الانحدار على إيجاد الإعدادات الصحيحة لهذه المعاملات.

من الممكن افتراض ونمذجة العديد من أنواع العلاقات المختلفة باستخدام تحليل الانحدار. ونظريًّا القيد الوحيد على بِنية العلاقة التي يمكن نمذجتها هو القدرة على تحديد دالَّة الانحدار المناسبة. وفي بعض المجالات، ربما يكون هناك أسباب نظرية قوية تفرِض نوعًا معينًا من العلاقة، ولكن في ظلِّ غياب هذا النوع من نظرية المجال فمن الأفضل البدء بافتراض أبسط شكلٍ للعلاقات — ألا وهي العلاقة الخطية — ثم المضي قدمًا لوضع نموذجٍ للعلاقات الأكثر تعقيدًا إذا لزم الأمر. وأحد الأسباب للبدء بعلاقةٍ خطية هو أن

دوال الانحدار الخطي من السهل نسبيًّا تفسيرها. والسبب الآخر هو الاعتقاد السائد بأن إبقاء الأمور بسيطةً بقدْر الإمكان هي فكرة سديدة بوجهٍ عام.

عند افتراض علاقة خطِّية، يُطلَق على تحليل الانحدار «انحدار خطي». وأبسط تطبيقٍ للانحدار الخطي هو نمذجة العلاقة بين سمتَين: سمة مُدخَلة X (س) وسمة مستهدفة (مُخرجة) Y (ص). وفي هذه المسألة البسيطة للانحدار الخطي، يكون شكل دالة الانحدار كما يلي:

$$Y = \omega_0 + \omega_1 X$$

ودالة الانحدار هذه هي مجرد معادلة خط (كثيرًا ما تُكتب على هذا الشكل: $y = mx + c$) مألوفة لأغلب من درسوا مادة الهندسة في المرحلة الثانوية.[3] ويُعَد المتغيِّران ω_0 وω_1 مُعامِلَين لدالَّة الانحدار. ويغير تعديل هذَين المعامِلَين طريقة تعيين الدالة للمُخرَج Y بناءً على المدخَل X. والمعامل ω_0 هو نقطة التقاطع مع المحور الصادي (Y) (أو الرمز C المستخدَم في مادة الهندسة بالمرحلة الثانوية) التي تُحدد نقطة تقاطع الخط مع المحور الرأسي y عندما تساوي X صفرًا. ويُحدد المعامِل ω_1 درجة انحدار الخط (أي هو المكافئ للرمز m في نسخة المرحلة الثانوية).

وفي تحليل الانحدار، تكون مُعاملات دالة الانحدار مجهولةً في البداية. وتحديد مُعاملات دالة الانحدار يكافئ البحث عن الخط الذي يتناسَب مع البيانات على أفضل وجه. وتبدأ استراتيجية تحديد هذه المعاملات بتخمين قِيَم المعاملات ثم تحديث المعاملات على نحوٍ مُتكرر لتقليل الخطأ الإجمالي للدالة على مجموعة البيانات. ويُحسب الخطأ الإجمالي في ثلاث خطوات:

(١) تُنفذ الدالة على مجموعة البيانات، وتقدر قيمة السِّمة المستهدفة لكل مثيلٍ موجود في البيانات.

(٢) يُحسب خطأ الدالة لكل مثيلٍ من خلال طرح القيمة التقديرية للسمة المستهدفة من قيمتها الحقيقية.

(٣) يتم تربيع خطأ الدالة لكلِّ مثيل، ثم تُجمَع هذه القِيَم التربيعية.

يتم تربيع خطأ الدالَّة لكل مثيلٍ في الخطوة الثالثة لتجنُّب إلغاء أثر الأخطاء المتعاكسة عند المبالغة في تقدير القيمة المستهدفة وعند التقليل منها. وتربيع الخطأ يجعل الخطأ

موجبًا في كِلتا الحالتَين. ويُعرف هذا القياس للخطأ باسم «مجموع الأخطاء التربيعية»، وتُعرف استراتيجية إعداد دالة خطِّية من خلال البحث عن المعامِلات التي تُقلل هذا المجموع إلى الحد الأدنى باسم «المربعات الصغرى». ويتحدَّد مجموع الأخطاء التربيعية بالمعادلة التالية:

$$SSE = \sum_{i=i}^{n} (target_i - prediction_i)^2$$

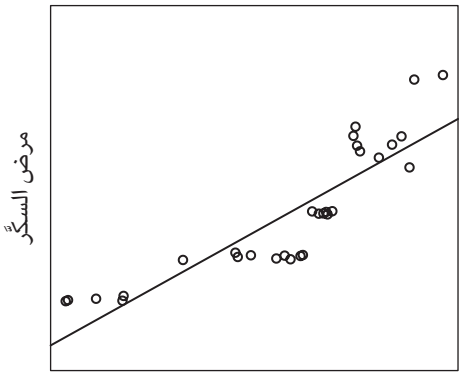
حيث تحتوي مجموعة البيانات على عدد n من المثيلات، و$target_i$ هي قيمة السمة المستهدفة للمثيل i في مجموعة البيانات، $prediction_i$ وهو القيمة التقديرية للسمة المستهدفة باستخدام دالَّة بالمثيل نفسه.

ولكي نُنشئ نموذج تنبؤٍ قائمًا على الانحدار الخطي الذي يُقدِّر احتمالية إصابة الفرد بمرض السكَّر بناءً على مؤشر كتلة جسمه، نستبدِل سمةَ مؤشر كتلة الجسم بالرمز X، وسمة الإصابة بالسكَّر بالرمز Y، ونستعين بخوارزمية المربعات الصغرى لإيجاد الخط الأكثر ملاءمةً لمجموعة بيانات الإصابة بالسكَّر. يوضح شكل ٤-٣(أ) الخط الأكثر ملاءمةً ومَوضِعَهُ بالنسبة إلى المثيلات في مجموعة البيانات. وفي شكل ٤-٣(ب)، تُظهِر الخطوط المتقطِّعة الخطأ (أو القيمة الباقية) لكل مثيلٍ لهذا الخط. وباستخدام منهج المربعات الصغرى، يكون الخط الأكثر ملاءمةً هو الخط الذي يُقلل إجمالي القيم الباقية التربيعية إلى أدنى حد. ومعادلة هذا الخط كما يلي:

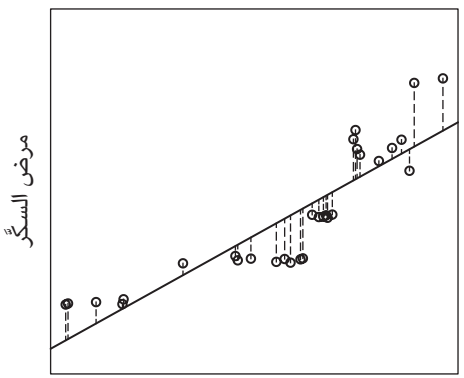
الإصابة بمرض السكر = − ٧,٣٨٤٣١ + ٠,٥٥٥٩٣ ⋆ مؤشر كتلة الجسم

تُشير قيمة معامِل الميل $\omega_1 = 0.55593$ إلى أن النموذج يزيد من الاحتمالية المقَدَّرة لإصابة الشخص بالسكَّر بنسبةٍ تزيد قليلًا عن نصف بالمائة مع كل زيادة مقدارها وحدة واحدة على مؤشر كتلة الجسم. ومن أجل التنبؤ باحتمالية إصابة الشخص بالسكَّر، نُدخِل بكل بساطة مؤشر كتلة جسم الشخص في النموذج. على سبيل المثال، إذا كان مؤشر كتلة الجسم يساوي ٢٠، يتنبَّأ النموذج باحتمالية الإصابة بالسكَّر بنسبة ٣,٧٣ بالمائة، وعندما يساوي مؤشر كتلة الجسم ٢١، يتنبَّأ النموذج باحتمالية الإصابة بنسبة ٤,٢٩.[4]

وفي باطن هذه العملية، يحسُب نموذجُ انحدار خطِّي، مُعَد باستخدام أسلوب المربعات الصغرى، فعليًّا المتوسطَ المرجَّحَ عبر المثيلات. في الواقع، تؤكد قيمة معامِل انحدار الميل $\omega_0 = -7.38431$ أن الخط الأكثر ملاءمةً يمر عبر النقطة المحددة بمتوسط قيمة مؤشر كتلة الجسم ومتوسط قيمة الإصابة بالسكَّر من واقع مجموعة البيانات. فإذا أُدخلت قيمة



(أ)



(ب)

شكل ٤-٣: (أ) خط الانحدار الأكثر ملاءمةً للنموذج هو «الإصابة بمرض السكر = −٧,٣٨٤٣١ + ٠,٥٥٥٩٣ * مؤشر كتلة الجسم». (ب) توضح الخطوط الرأسية المتقطِّعة القيمة المتبقية لكل مثيل.

متوسط مؤشر كتلة الجسم في مجموعة البيانات (مؤشر كتلة الجسم = ٢٤,٠٩٣٢)، فإن النموذج يقدِّم قيمةَ احتمالية الإصابة بالسكَّر بنسبة ٤,٢٩ في المائة، وهي القيمة المتوسطة للإصابة بمرض السكَّر وفقًا لمجموعة البيانات.

يتوقف ترجيح (تحديد وزن) المثيلات على المسافة الفاصلة بين المثيل والخط: كلما ابتعد أحد المثيلات عن الخط، زادت القيمة المتبقية لذلك المثيل، وستُرجح الخوارزمية ذلك المثيل من خلال تربيع القيمة المتبقية. وإحدى تداعِيات تحديد الوزن هي أن المَثيلات ذات القِيم المتطرفة (الشاذة) يكون لها تأثير كبير على نحوٍ غير متناسب على عملية إعداد الخط الأكثر ملاءمة، مما يُسفر عن إبعاد الخط عن المثيلات الأخرى. وبالتالي، من المهم التحقق من القِيَم الشاذة في مجموعة البيانات قبل إعداد الخط الأكثر ملاءمةً لمجموعة البيانات (أو بعبارة أخرى، تدريب دالة انحدار خطِّي على مجموعة البيانات) باستخدام خوارزمية المربعات الصغرى.

يمكن التوسع في نماذج الانحدار الخطي لاستيعاب عدة مدخلات. يُضاف معاملٌ جديد إلى النموذج من أجل كل سمةٍ مُدخلة جديدة، وتُحدَّث المعادلة الخاصة بالنموذج لتشملَ نتيجة ضرب السمة الجديدة في المعامل الجديد ضمن المجموع. على سبيل المثال، من أجل التوسع في النموذج ليشمل سِمتَي التمارين الرياضية والوزن كمدخلات، ستصير معادلة دالة الانحدار كما يلي:

الإصابة بمرض السكر = $\omega_1 + \omega_0$ مؤشر كتلة الجسم + ω_2 التمارين الرياضية + ω_3 الوزن.

في علم الإحصاء، تُعرف دالة الانحدار التي تُعيِّن مُخرجًا واحدًا من عدة مدخلات بهذه الطريقة باسم «دالة انحدار خطِّي مُتعدد». تُعد بِنية دالة الانحدار المتعدد المدخلات أساسًا لمجموعةٍ من خوارزميات تعلُّم الآلة، من بينها الشبكات العصبية.

وتُعد العلاقة الارتباطية والانحدار مفهومَين متشابهَين حيث إن كليهما تِقنيتان تُركزان على العلاقات بين السمات في مجموعة البيانات. وتركز العلاقة الارتباطية على اكتشاف ما إذا كان هناك علاقة موجودة بين سِمتَين، ويركز الانحدار على نمذجة علاقة مفترضة بين السمات بغرَض القدرة على تقدير قيمة إحدى السمات المستهدَفة بناءً على قِيَم سِمة أو أكثر من السِّمات المدخلة. في الحالات المحددة لعلاقة بيرسون الارتباطية والانحدار الخطي، تقيس علاقة بيرسون الارتباطية درجة وجود علاقةٍ خطية بين سمتَين، والانحدار الخطي المدرب باستخدام المربعات الصغرى هو عملية لإيجاد خطٍّ أكثر ملاءمةً يتنبأ بقيمة سمةٍ بمعلومية قيمة سِمة أخرى.

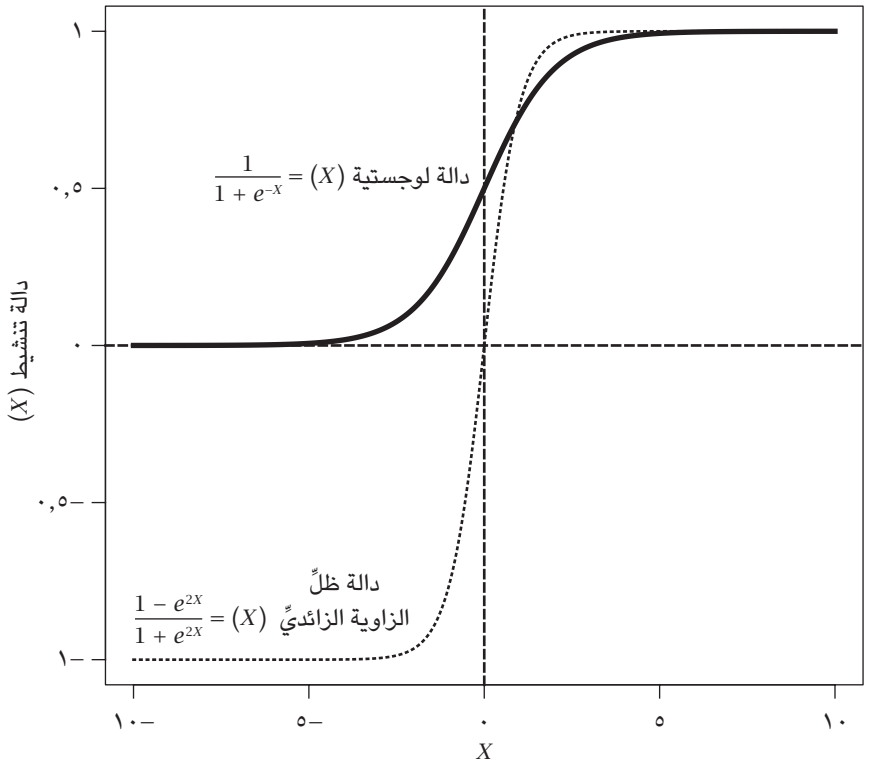
الشبكات العصبية والتعلُّم العميق

تتكون «الشبكة العصبية» من مجموعة من الخلايا العصبية (أو العصبونات). تأخذ الخلية العصبية مجموعة من القيم العددية كمُدخلٍ لها ثم تعيِّن قيمة مُخرَجة وحيدة. والخلية العصبية، في جوهرها، هي بكل بساطة دالة انحدار خطي متعدد المدخلات. الفارق الوحيد المهم بين الاثنين أنه في الخلية العصبية يُمرَّر مخرج دالة الانحدار الخطي المتعدد المدخلات عبر دالةٍ أخرى يُطلق عليها «دالة تنشيط».

تُنفِّذ دوال التنشيط هذه عملية تعيين غير خطية لمُخرَج دالة الانحدار الخطي المتعدد المدخلات. ثمة دالتان تنشيطيَّتان شائعتا الاستخدام ألا وهما الدالة اللوجستية ودالة ظلِّ الزاوية الزائديِّ (انظر شكل ٤-٤). تأخذ كلتا الدالتَين قيمةً واحدة X بصفتها مدخلًا؛ في الخلية العصبية، هذه القيمة X هي المخرج الناتج عن دالة الانحدار الخطِّي المتعدد المدخَلات التي نفذتها الخلية العصبية على مدخلاتها. وتستخدِم كلتا الدالتين عدد أويلر، e، الذي يساوي تقريبًا ٢,٧١٨٢٨١٨٢. أحيانًا يُطلق على هاتَين الدالتَين «دوال الضغط» لأنهما تأخذان أية قيمةٍ بين عددٍ لا نهائي موجب وعددٍ لا نهائي سالب ويقومان بتعيينها إلى نطاقٍ صغير مُحدد مسبقًا. ونطاق مخرجات الدالة اللوجستية يكون من ٠ إلى ١، ونطاق دالة ظلِّ الزاوية الزائدي يكون من −١ إلى ١. وكنتيجة لذلك، دائمًا ما تكون مُخرَجات الخلية العصبية التي تستعين بالدالة اللوجستية بوصفها دالَّتها التنشيطية ما بين صفر وواحد. وتتَّضح حقيقة أن كِلتا الدالتَين اللوجستية وظل الزاوية الزائدي تُنفذان عمليات تعيين غير خطية في شكل المنحنيات التي تتخذ شكل حرف S. والسبب وراء تنفيذ عمليات تعيين غير خطية في الخلية العصبية هو أن أحد أوجه قصور دالة الانحدار الخطي المتعدد المدخلات يتمثل في أن الدالة خطية، كما يتَّضح من اسمها، وإذا نفَّذت جميع الخلايا العصبية داخل الشبكة عمليات التعيين الخطية فحسب، فسوف تقتصِر الشبكة العصبية ككلٍّ على تعلُّم الدوال الخطية فقط. ومع ذلك، فإن تنفيذ دالة التنشيط غير الخطية في الخلايا العصبية الخاصة بالشبكة تُتيح للشبكة تعلُّم الدوالِّ الأكثر تعقيدًا (غير الخطية).

تجدُر الإشارة إلى أن كل خليةٍ عصبية في الشبكة تُجري مجموعة بسيطة جدًّا من العمليات:

(١) ضرب كل مُدخلٍ في وزن.
(٢) جمع نتائج عمليات الضرب معًا.
(٣) تمرير هذه النتيجة عبر دالة تنشيط.

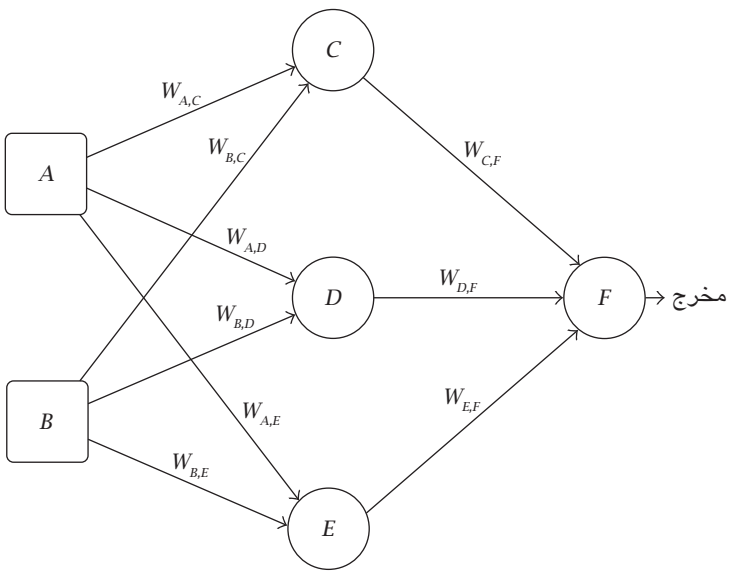


شكل ٤-٤: رسم بياني للدالة اللوجستية ودالة ظلِّ الزاوية الزائديِّ أثناء تنفيذهما على المدخل x.

تُعد العمليَّتان الأولى والثانية مجرد عملياتٍ حسابية لدالة انحدار مُتعدد المدخلات، والعملية الثالثة هي تنفيذ دالة تنشيط.

لكل الوصلات بين الخلايا العصبية في شبكةٍ ما يُوجَد اتجاه مُعين ووزن مُرتبط بها. ووزن الوصلة المتجهة إلى داخل خليةٍ عصبية هو الوزن الذي تمنحه الخلية العصبية للمُدخَل الذي تستقبِله في تلك الوصلة عند حساب دالة الانحدار المتعدد المدخلات على مدخلاتها. ويوضح شكل ٤-٥ البنية الهيكلية لشبكةٍ عصبية بسيطة. ويُمثل المربَّعان الموجودان على يسار الشكل، المكتوب عليهما A وB، مواضع في الذاكرة نستخدِمها لتقديم

البيانات المدخلة إلى الشبكة. ولا تُنفذ أية عمليات لمعالجة البيانات أو تحويلها في تلك المواضع. يمكنك أن تعتبر تلك العقد خلايا عصبية خاصة بالمدخلات أو خلايا استشعارية، حيث يُضبط تنشيط مخرجاتها حسب قيمة المدخَل.[5] وتُمثِّل الدوائر الموجودة في شكل ٤-٥ (المكتوب عليها C وD وE وF) الخلايا العصبية في الشبكة. غالبًا ما يفيد التفكير في الخلايا العصبية في الشبكة على أنها مُرتَّبة على هيئة طبقات. ولهذه الشبكة ثلاث طبقات من الخلايا العصبية: طبقة المدخلات وهي تحتوي على A وB؛ وطبقة مَخفيَّة وتحتوي على C وD وE؛ وطبقة المخرجات وتحتوي على F. ويصف مصطلح «الطبقة المخفية» حقيقة أن الخلايا العصبية في هذه الطبقة ليست موجودة في طبقة المدخلات ولا في طبقة المخرَجات؛ وإنما هي مخفيَّة عن الأنظار.



شكل ٤-٥: شبكة عصبية بسيطة.

تمثل الأسهم، التي تربط بين الخلايا العصبية في الشبكة، اتجاه تدفُّق المعلومات عبر هذه الشبكة. فمن الناحية التقنية، تُعَد هذه الشبكة بعينها شبكةً عصبية ذات تغذيةٍ أمامية لأنه لا يوجد حلقات تكرار في هذه الشبكة: تُشير جميع الوصلات إلى الأمام من

المدخلات إلى المخرجات. وهذه الشبكة متصلة بعضها ببعض بالكامل لأن كل خلية عصبية متصلة بجميع الخلايا العصبية الأخرى في الطبقة التالية من الشبكة. ومن الممكن إنشاء عدة أنواعٍ مختلفة من الشبكات العصبية من خلال تغيير عدد الطبقات، وعدد الخلايا العصبية في كل طبقة، ونوع دوالِّ التنشيط المستخدَمة، واتجاه الوصلات بين الطبقات، وغيرها من المعاملات. في الواقع، يتضمن قدْر كبير من الجهد المطلوب لتطوير شبكةٍ عصبية لأداء مهمةٍ معينة، التجريبَ، للعثور على أفضل تصميمٍ للشبكة لكي تؤدي تلك المهمة.

تمثل التسميات على كلِّ سهم الوزن الذي تمنحه العقدة الموجودة في نهاية السهم للمعلومات التي تُنقل عبر تلك الوصلة. على سبيل المثال، السهم الرابط بين C وF يُشير إلى أن المخرَج من C يُمرر كمُدخل إلى F، وسوف تمنح F الوزن WC,F للمُدخل القادم من C.

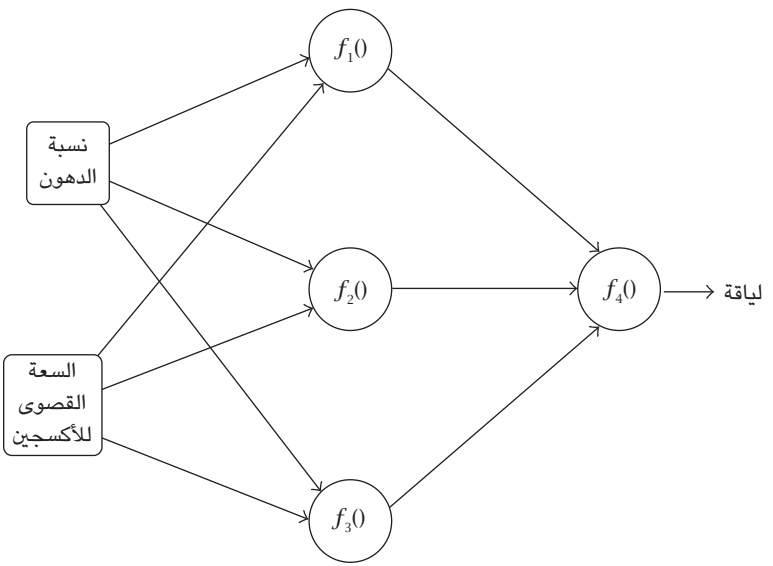
إذا افترضنا أن الخلايا العصبية في الشبكة الموضحة بشكل ٤-٥ تستخدِم دالة تنشيطٍ من نوع ظل الزاوية الزائدي، يُمكننا إذن كتابة المعادلة الحسابية التي تُجرى في الخلية العصبية F من الشبكة على النحو التالي:

المخرج = ظل الزاوية الزائدي $(\omega_{C,F}C + \omega_{D,F}D + \omega_{E,F}E)$.

يوضح التعريف الرياضي لعملية المعالجة التي تُجرى في الخلية العصبية F أن المخرج النهائي للشبكة يُحسب باستخدام تركيبةٍ من مجموعة دوال. وتعني عبارة «تركيبة من الدوال» أن المخرَج الخاص بدالة واحدة يُستخدم كمُدخلٍ لدالة أخرى. في هذه الحالة، مُخرجات الخلايا العصبية C وD وE تُستخدَم كمُدخلاتٍ للخلية العصبية F، وبالتالي تتكوَّن الدالة التي تستخدِمها خلية F من الدوال التي تنفذها الخلايا C وD وE.

يجعل شكل ٤-٦ هذا الوصف الخاص بالشبكات العصبية أكثر واقعية، موضحًا شبكة عصبية تأخذ نسبة الدهون في الجسم لشخصٍ والسعة القصوى للأكسجين (مقياس للحد الأقصى لكمية الأكسجين التي يمكن لشخصٍ استخدامها في الدقيقة) كمُدخَل وتحسب مستوى لياقة هذا الشخص.[6] تحسب كل خلية عصبية في الطبقة الوسطى من الشبكة دالةً قائمة على نسبة الدهون في الجسم والسعة القصوى للأكسجين: $f_1()$ و$f_2()$ و$f_3()$. تُظهِر كل دالةٍ التفاعل بين المدخلات بطريقةٍ مختلفة. تمثل هذه الدوال بالأساس سماتٍ جديدة مُستقاة من المدخلات الخام إلى الشبكة. وهي تُشبه سمة مؤشر كتلة الجسم المذكورة آنفًا؛ ويُحسَب مؤشر كتلة الجسم كدالةٍ للوزن والطول. وأحيانًا من الممكن تفسير

ما يُمثله المخرج الخاص بخلية عصبية في الشبكة إلى الحد الذي يمكن أن يُقدِّم وصفًا نظريًّا لما تُمثله السمة المشتقة وفهم سبب كون هذه السمة المشتقة مفيدة للشبكة. ومع ذلك، عادةً لا يكون للسمة المشتقة، التي تحسبها الخلية العصبية، معنًى رمزي بالنسبة للبشر. وبدلًا من ذلك، تُصوِّر هذه السمات التفاعلاتِ بين السمات الأخرى التي وجدتها الشبكة مفيدة. تحسب العقدة الأخيرة في الشبكة f_4 دالةً أخرى — عبر مُخرَجات $f_1()$ و$f_2()$ و$f_3()$ — تُعَد مخرجاتها هي مستوى اللياقة المتوقَّع الناتج عن الشبكة. ونُكرر مرة أخرى أن هذه الدالة ربما لا تكون ذات مغزًى بالنسبة للبشر، باستثناء حقيقة أنها تُحدد تأثيرًا متبادلًا وجدَتِ الشبكة أنه ذا علاقة ارتباطية وثيقة بالسمة المستهدفة.



شكل ٤-٦: شبكة عصبية تتنبأ بمستوى لياقة شخصٍ ما.

يشمل تدريب الشبكة العصبية إيجاد الأوزان الصحيحة للوصلات الموجودة في الشبكة. ولفهم كيفية تدريب شبكة عصبية، من المفيد البدء في التفكير في كيفية تدريب الأوزان من أجل خليةٍ عصبية وحيدة في طبقة المخرجات الخاصة بالشبكة. افترض أن لدَينا مجموعة بيانات تدريب تحتوي على مُدخلاتٍ ومُخرجات مستهدَفة لكلِّ مَثيل. افترض أيضًا

أن الوصلات الآتية إلى الخلية العصبية لها أوزان مُعينة. فإذا أخذنا مثيلًا من مجموعة البيانات وقدَّمنا قِيمًا للسمات المدخَلة لهذا المثيل في الشبكة، ستتنبَّأ الخلية العصبية بالسمة المستهدَفة على هيئة مُخرج. ومن خلال طرح هذه القيمة المتنبِّئة من القيمة المحسوبة للسمة المستهدَفة في مجموعة البيانات، يُمكننا حساب خطأ الخلية العصبية لذلك المثيل. ومن خلال الاستعانة ببعض أساسيات حساب التفاضل والتكامل، من الممكن استنباط قاعدةٍ لتحديث الأوزان الخاصة بالوصلات الآتية من الخلية العصبية بمعلومية قياس خطأ المخرَج الخاص بالخلية العصبية بهدف تقليل نسبة خطأ الخلية العصبية. وسيختلف التعريف الدقيق لهذه القاعدة باختلاف دالَّة التنشيط التي استخدمتها الخلية العصبية لأن دالة التنشيط تؤثر على السمة المشتقة المستخدَمة لاشتقاق القاعدة. ولكن يُمكننا تقديم التفسير البديهي التالي لآلية عمل قاعدة تحديث الوزن:

(١) إذا كان الخطأ يُساوي صفرًا، إذن لا ينبغي لنا تغيير الأوزان الممنوحة للمُدخلات.

(٢) إذا كان الخطأ بالموجب، سنُثقل الخطأ إذا قُمنا بزيادة مُخرَجات الخلية العصبية، إذن يجب أن نزيد أوزان جميع الوصلات التي يكون فيها المدخَل بالموجب ونُثقل أوزان الوصلات التي يكون فيها المدخَل بالسالب.

(٣) إذا كان الخطأ بالسالب، سنُثقل الخطأ إذا قلَّلنا مُخرَجات الخلية العصبية، وبالتالي يجب أن نُثقل أوزان جميع الوصلات التي يكون فيها المدخَل بالموجب ونزيد أوزان الوصلات حيث يكون المدخَل بالسالب.

تكمن الصعوبة في تدريب شبكةٍ عصبية في أن قاعدة تحديث الوزن تتطلَّب تقديرًا للخطأ الموجود في خلية عصبية، وعلى الرغم من أنه يسهل حساب الخطأ في كل خليةٍ عصبية من طبقة المخرَجات الخاصة بالشبكة، فمن الصعب حساب الخطأ الخاص بالخلايا العصبية في الطبقات الأولى. والطريقة القياسية لتدريب شبكة عصبية هو استخدام خوارزمية تُسمَّى «خوارزمية الانتشار العكسي» لحساب الخطأ لكلِّ خليةٍ عصبية في الشبكة واستخدام قاعدة تحديث الوزن لتعديل الأوزان في الشبكة.[7] وتُعد خوارزمية الانتشار العكسي خوارزمية تعلُّم آلةٍ خاضع للإشراف، ومن ثم تفترض مجموعة بيانات مدربة لها مُدخَلات ومُخرَج مستهدَف لكل مثيل. يبدأ التدريب بتعيين أوزانٍ عشوائية لكل وصلةٍ من الوصلات الموجودة في الشبكة. تحدِّث الخوارزمية بعد ذلك الأوزان في الشبكة على نحوٍ متكرِّر من خلال عرض مثيلات التدريب من مجموعة البيانات على

الشبكة وتحدِّث أوزان الشبكة إلى أن يتحسَّن أداء الشبكة كما هو متوقَّع منها. ويأتي اسم خوارزمية «الانتشار العكسي» من حقيقة أنه بعد تقديم كلِّ مثيلٍ تدريبي إلى الشبكة، تُمرِّر الخوارزمية خطأ الشبكة على نحوٍ عكسي عبر الشبكة بدايةً من طبقة المخرَجات وتحسب عند كل طبقة في الشبكة أخطاء الخلايا العصبية الموجودة في تلك الطبقة قبل مشاركة هذا الخطأ مرةً أخرى مع الخلايا العصبية الموجودة في الطبقة السابقة. وفيما يلي الخطوات الأساسية التي تقوم بها هذه الخوارزمية:

(١) حساب خطأ الخلايا العصبية الموجودة في طبقة المخرجات والاستعانة بقاعدة تحديث الوزن لتحديث الأوزان الداخلة إلى هذه الخلايا العصبية.

(٢) مشاركة الخطأ المحسوب عند إحدى الخلايا العصبية مع كلِّ خليةٍ عصبية في الطبقة السابقة المتصلة بتلك الخلية العصبية بالتناسُب مع وزن الوصلة الرابطة بين الخليتَين العصبيتَين.

(٣) بالنسبة إلى كل خليةٍ عصبية في الطبقة السابقة، حساب إجمالي أخطاء الشبكة التي تَسبَّبَت فيها الخلية العصبية عن طريق جمع الأخطاء التي انتشرت انتشارًا عكسيًّا واستخدام نتيجة مجموع هذه الأخطاء لتحديث الأوزان الخاصة بالوصلات الداخلة إلى هذه الخلية العصبية.

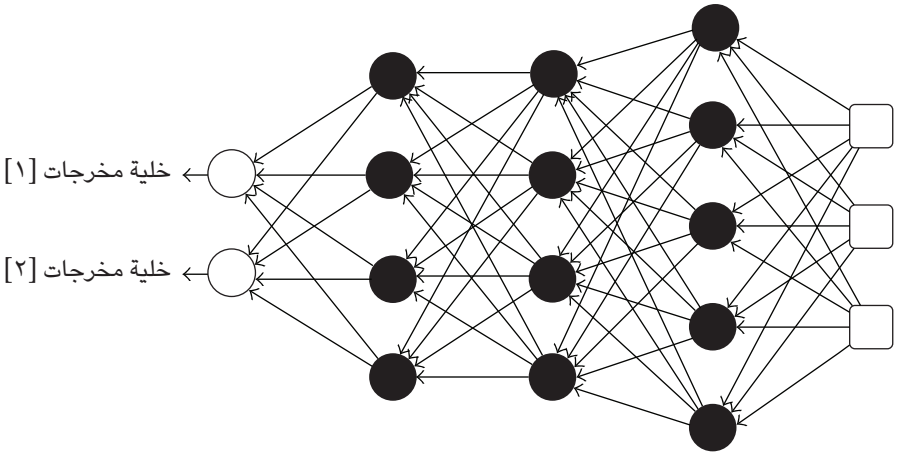
(٤) التعامل مع باقي الطبقات في الشبكة من خلال تكرار الخطوتَين الثانية والثالثة حتى تُحدَّث أوزان الوصلات ما بين خلايا المدخلات والطبقة الأولى من الخلايا العصبية المخفية.

في الانتشار العكسي، تُحسَب تحديثات الأوزان لكل خليةٍ عصبية من أجل الإقلال من أخطاء الخلية العصبية في المثيل التدريبي، لا من أجل التخلُّص نهائيًّا من الأخطاء. والسبب وراء ذلك أن الهدف وراء تدريب الشبكة هو تمكينها من التعميم على المَثيلات الجديدة غير الموجودة في بيانات التدريب بدلًا من حفظ بيانات التدريب. وبالتالي، كل مجموعة من تحديثات الأوزان تدفع الشبكة نحوَ مجموعةٍ من الأوزان التي تُناسب بوجهٍ عام مجموعة البيانات بالكامل، ومن خلال العديد من عمليات التكرار تستقر الشبكة على مجموعةٍ من الأوزان التي ترصد التوزيع العام للبيانات بدلًا من التفاصيل المخصصة لمثيلات التدريب. وفي بعض نُسَخ الانتشار العكسي، تُحدَّث الأوزان بعد تقديم عددٍ من المثيلات (أو مجموعة من المثيلات) للشبكة وليس بعد كل مثيلٍ تدريبي. التعديل

الوحيد اللازم إجراؤه على هذه النُّسَخ هو أن تستخدِم الخوارزمية متوسط خطأ الشبكة على مجموعةٍ من المثيلات باعتباره مقياس الخطأ عند طبقة المخرجات لعملية تحديث الوزن.

أحد أكثر التطوُّرات التقنية المثيرة للاهتمام خلال السنوات العشر الأخيرة هو ظهور التعلُّم العميق. وشبكات «التعلُّم العميق» هي ببساطة شبكات عصبية ذات طبقاتٍ متعددة[8] من الوحدات المخفية؛ بعبارةٍ أخرى، هي «عميقة» من حيث عدد الطبقة المخفية التي تحتويها. للشبكة العصبية الموجودة في شكل ٤-٧ خمس طبقات: طبقة مدخلات على اليسار تحتوي على ثلاث خلايا عصبية، وثلاث طبقات مخفية (الدوائر السوداء)، وطبقة مُخرَجات واحدة على اليمين تحتوي على خليتَين. توضح هذه الشبكة أنه يمكن أن يكون هناك عدد مختلف من الخلايا العصبية في كل طبقة: طبقة المدخَلات بها ثلاث خلايا عصبية؛ الطبقة الأولى المخفيَّة بها خمس؛ وكل طبقة من الطبقتَين المخفيتَين التاليتَين بها أربع؛ وطبقة المخرجات بها اثنتان. توضح هذه الشبكة أيضًا أن طبقة المخرَجات من الممكن أن تحتوي على عدة خلايا عصبية. واستخدام عدة خلايا عصبية للمخرجات مفيد إذا كانت السمات المستهدفة من نوع البيانات الاسمية أو الترتيبية التي لها مستويات مختلفة. وفي هذه السيناريوهات، تُعدَّ الشبكة بحيث يكون هناك خلية عصبية واحدة للمُخرَجات في كل مستوًى، ويتم تدريب الشبكة بحيث يكون لكل مُدخَل خلية مُخرَجات واحدة فقط تُخرج تنشيطًا عاليًا (مما يدل على المستوى المستهدف المتوقع).

كما في الشبكات السابقة التي ألقينا نظرةً عليها، الشبكة المبينة في شكل ٤-٧ متصلة بعضها ببعض بالكامل، وهي شبكة تغذية أمامية. ومع ذلك، ليست جميع الشبكات شبكةَ تغذية أمامية مُتصلة بالكامل. في الواقع، طُوِّرت أشكال مُتعددة من طوبولوجيا الشبكة. على سبيل المثال، تقدم الشبكات العصبية التكرارية الحلقات التكرارية في طوبولوجيا الشبكة: تُرجَع مُخرجات الخلية العصبية الخاصة بمُدخل مُعين إلى الخلية العصبية أثناء معالجة الإدخال التالي. تُكوِّن هذه الحلقة التكرارية ذاكرة للشبكة تُمكنها من معالجة كل مُدخَل في سياق المدخلات السابقة التي عالجتها. ونتيجة لذلك، تُعد الشبكات العصبية التكرارية مناسبةً لمعالجة البيانات المتسلسلة مثل اللغة.[9] ثمة بِنية أخرى مشهورة للشبكات العصبية العميقة ألا وهي الشبكة العصبية الالتفافية. صُمِّمت هذه الشبكات في الأصل من أجل استخدامها مع الصور (Le Cun 1989). وإحدى الخصائص المرغوبة



شكل ٤-٧: شبكة عصبية عميقة.

لشبكة التعرُّف على الصور هي أنها ينبغي أن تكون قادرة على التعرُّف على ما إذا كانت سمةٌ بصرية معينة قد ظهرت في صورةٍ ما بغضِّ النظر عن موضع حدوثها في الصورة. على سبيل المثال، إذا كانت شبكةٌ ما تجري عملية التعرُّف على الوجوه، فإنها يجب أن تكون قادرةً على التعرُّف على شكل العين إذا كانت العين موجودةً في الركن العلوي الأيسر أم في وسط الصورة. تُحقق الشبكات العصبية الالتفافية هذا لأنها تحتوي على مجموعاتٍ من الخلايا العصبية التي تتشارك في نفس مجموعة الأوزان الخاصة بمُدخلاتها. وفي هذا السياق، فلنضرب مثلًا بمجموعة أوزان المدخلات على أنها تعرف دالة بحيث تعطي نتيجة «صواب» في حال إذا ظهرت سمة بصرية مُعينة في مجموعة البكسلات التي تُمرَّر إلى هذه الدالة. هذا يعني أن كل مجموعة من الخلايا العصبية التي تتشارك في أوزانها تتعلَّم التعرُّف على سمةٍ بصرية مُعينة، وكل خلية عصبية في المجموعة تؤدي دور جهاز كشفٍ عن تلك السمة. وفي الشبكة العصبية الالتفافية، تُرتَّب الخلايا العصبية داخل كل مجموعةٍ بحيث تفحص كل خليةٍ موضعًا مختلفًا في الصورة، وتُغطي المجموعة الصورة بأكملها. ونتيجة لذلك، إذا كانت السمة البصرية التي تبحث عنها المجموعة موجودةً في أي مكانٍ بالصورة، فستتعرف عليها إحدى الخلايا العصبية في المجموعة.

تأتي قوة الشبكات العصبية العميقة من حقيقة أنها يمكن أن تتعلَّم السمات المفيدة تلقائيًّا، مثلما تفعل الخلايا الكاشفة عن سمةٍ ما في الشبكات العصبية الالتفافية. في

الواقع، أحيانًا يُعرف التعلُّم العميق باسم «التعلُّم التمثيلي» لأن هذه الشبكات العميقة تتعلم بالضرورة تمثيلًا جديدًا للبيانات المدخلة يعتبر أفضل في التنبؤ بالسمة المستهدَفة من المدخل الأساسي الخام. تعرِّف كل خليةٍ عصبية في الشبكة دالةً تعيِّن القيم المدخلة إلى الخلية العصبية إلى سمة جديدة مُخرجة. ومن ثم، ربما تتعلم خلية عصبية في الطبقة الأولى من الشبكة دالَّةً تعيِّن القيم الخام المدخَلة (مثل الوزن والطول) إلى سمة أفيدَ مِن القِيَم المدخَلة الفردية (مثل مؤشر كتلة الجسم). ومع ذلك، تُغذَّى الخلايا العصبية الموجودة في الطبقة الثانية بالمخرَجات الخاصة بهذه الخلية، بالإضافة إلى المخرجات الخاصة بالخلايا العصبية المجاورة في الطبقة الأولى، وتحاول الخلايا العصبية في الطبقة الثانية أن تتعلَّم الدوال التي تعين مُخرجات الطبقة الأولى إلى تمثيلاتٍ جديدة أكثر فائدة. وتستمر هذه العملية الخاصة بتعيين مُدخلات إلى السمات الجديدة وتغذية الدوال الجديدة بهذه السمات الجديدة كمدخلاتٍ عبر الشبكة، وبينما تزداد الشبكة عمقًا، يُمكنها أن تتعلَّم تعييناتٍ أكثر تعقيدًا من المدخلات الخام إلى تمثيلات السمة الجديدة. إن القدرة على تعلُّم التعيينات المعقدة للبيانات المدخلة تلقائيًّا إلى تمثيلات مفيدة هي ما تجعل نماذج التعلُّم العميق دقيقةً للغاية في المهام الكثيرة الأبعاد (مثل معالجة الصور والنصوص).

ومن المعروف منذ فترةٍ طويلة أن جعل الشبكات العصبية أعمقَ يُتيح للشبكة أن تتعلم تعيينات أعقد للبيانات. والسبب وراء أن التعلُّم العميق لم يُحقق نجاحًا فوريًّا إلا في السنوات القليلة الماضية هو أن المزيج المعتاد المتمثل في الاستهلال بأوزانٍ عشوائية يتبعها خوارزمية انتشار عكسي لا يؤتِي ثماره بشكلٍ جيد مع الشبكات العميقة. وتتمثل إحدى مشكلات خوارزمية الانتشار العكسي في أن الخطأ تتم مشاركته نظرًا إلى أن العملية تتم بشكلٍ عكسي عبر الطبقات، وبالتالي في الشبكة العميقة عندما تصِل الخوارزمية إلى الطبقات الأولى من الشبكة، حينئذٍ لن تكونَ تقديرات الخطأ مفيدة.[10] ونتيجة لذلك، لا تتعلم الطبقات الموجودة في الأجزاء الأولى من الشبكة عمليات التحويل المفيدة للبيانات. وفي السنوات القليلة الماضية، طوَّر الباحثون أنواعًا جديدة من الخلايا العصبية وأضافوا تعديلاتٍ على خوارزمية الانتشار العكسي التي تتعامل مع هذه المشكلة. وقد وُجِد أيضًا أن توخِّي الحذَر بشأن تحديد أوزان عشوائية للشبكة في البداية أمرٌ مفيد. وكان هناك عاملان آخران جعلا من الصعب تدريب الشبكات العميقة، ألا وهما أن تدريب شبكةٍ عصبية يتطلب قدرًا مهولًا من القدرة الحوسبية، وتؤتي الشبكات العصبية ثمارها على

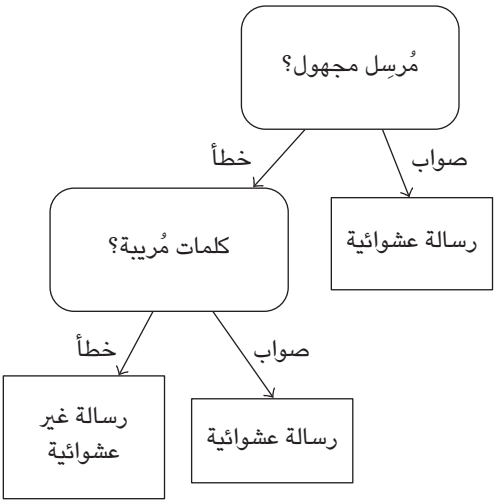
نحوٍ أفضل عندما يكون هناك قدرٌ كبير من بيانات التدريب. وكما ناقشنا بالفعل، في السنوات الأخيرة أدَّت الزيادات الكبيرة في إتاحة القدرة الحوسبية ومجموعات البيانات الكبيرة إلى جعل الشبكات العميقة أكثر جدوى.

الهياكل الشجرية لاتخاذ القرار

يعمل الانحدار الخطِّي والشبكات العصبية على أفضل نحوٍ مع المدخَلات العددية. فإذا كانت السمات المدخَلة في مجموعة البيانات سماتٍ اسميةً أو ترتيبية في الأساس، فربما تكون خوارزميات ونماذج تعلُّم الآلة الأخرى، مثل الهياكل الشجَرية لاتخاذ القرار، مناسبةً أكثر لهذه البيانات.

يشفر الهيكل الشجري لاتخاذ القرار مجموعةً من قواعد if-then-else على هيئة شجرة. ويوضح شكل ٤-٨ هيكلًا شجريًّا مستخدمًا لتحديد ما إذا كانت رسالة البريد الإلكتروني عشوائيةً أم غير عشوائية. يمثل المستطيلان مُستديرَا الزوايا اختباراتٍ تخضع لها السمات، أما المربعات فتشير إلى القرار أو التصنيف. يشفر هذا الهيكل الشجري القواعد التالية: «إذا كانت رسالة البريد الإلكتروني من مُرسِل مجهول، إذن فهي رسالة عشوائية؛ وإذا لم تكن من مُرسِل مجهول؛ ولكنها تحتوي على كلماتٍ مُريبة، إذن فهي رسالة عشوائية؛ وإذا لم تكن من مُرسِل مجهول ولا تحتوي على كلماتٍ مُريبة؛ إذن فهي ليست رسالة عشوائية.» وفي الهيكل الشجري لاتخاذ القرار، يُتخَذ القرار الخاص بمثيلٍ عن طريق البدء عند قمة الهيكل الشجري نزولًا إلى الأسفل من خلال إخضاع المثيل لسلسلةٍ من اختبارات السمات. وتُحدد كل عقدةٍ في الهيكل الشجري سمة واحدة للاختبار، وتسير العملية على طول الهيكل الشجري إلى أسفلَ، عقدةً بعقدةٍ من خلال اختيار الفرع المنحدِر من العقدة الحالية ذات المسمَّى المناسب للقيمة الخاصة بالسمة الاختبارية الخاصة بالمثيل. القرار النهائي هو تسمية العقدة الطرفية (أو الورقة) التي ينحدِر إليها المثيل.

يحدد كل مسار في الهيكل الشجري، بدايةً من الجذر وصولًا إلى الأوراق، قاعدةً تصنيفية تتألف من سلسلة من الاختبارات. والهدف من خوارزمية التعلُّم القائمة على الهياكل الشجرية هو إيجاد مجموعة من القواعد التصنيفية التي تُقسِّم مجموعة بيانات التدريب إلى مجموعاتٍ من المثيلات لها نفس قيمة السمة المستهدفة. الفكرة هي إذا كانت القاعدة التصنيفية يُمكنها أن تفصل من مجموعة البيانات مجموعةً فرعية من المثيلات



شكل ٤-٨: هيكل شجري لتحديد ما إذا كانت رسالة البريد الإلكتروني عشوائيةً أم غير عشوائية.

التي لها نفس القيمة المستهدفة، وإذا كانت هذه القاعدة التصنيفية مُتحقِّقة أو تعطي نتيجة true لمثيلٍ جديد (بمعنى أن المثيل يسري على ذلك المسار في الهيكل الشجري)، إذن فعلى الأرجح يكون التنبؤ الصحيح لهذا المثيل الجديد هو القيمة المستهدفة التي تتشاركها جميع مثيلات التدريب التي تنطبق عليها هذه القاعدة.

تُعد خوارزمية ثنائية التفرُّع التكرارية ٣ (آي دي ٣) هي المنشأ الذي تنحدِر منه أحدث خوارزميات تعلُّم الآلة القائمة على الهياكل الشجرية لاتخاذ القرار (Quinlan 1986). تنشئ خوارزمية آي دي ٣ هيكلًا شجريًّا لاتخاذ القرار بأسلوبٍ تكراري يعطي الأولوية للتعمق، مُضيفة عقدة واحدة في كل مرة، بدءًا من عقدة الجذر. وتبدأ هذه الخوارزمية باختيار سمةٍ ما عند عقدة الجذر لإخضاعها للاختبار. ينشأ فرع من الجذر لكل قيمةٍ في نطاق هذه السمة الاختبارية ويُسمى بتلك القيمة. على سبيل المثال، سينحدِر فرعان من أي عُقدةٍ ذات سمةٍ ثنائية اختبارية. بعد ذلك تُقسَّم مجموعة البيانات: يسير كلُّ مثيل في مجموعة البيانات إلى أسفل الفرع وتُعطى له تسمية فئوية تتناسَب مع قيمة السمة الاختبارية للمثيل. ثم تنمي خوارزمية آي دي ٣ كل فرع باستخدام العملية نفسها المستخدمة لإنماء عقدة الجذر: أي اختيار سمة اختبارية، وإضافة عقدة ذات فروع، وتقسيم البيانات من خلال تحويل المَثيلات إلى الفروع ذات الصِّلة. وتستمر هذه العملية

إلى أن تُصبح لجميع المثيلات على أحد الفروع القيمة نفسها للسمة المستهدَفة، وفي هذه الحالة تُضاف العقدة الختامية إلى الشجرة وتُسمى بقيمة السمة المستهدفة التي تشاركها جميع المثيلات على الفرع.[11]

تختار خوارزمية آي دي ٣ السمة التي ستُختبَر عند كل عقدة في الشجرة بحيث تُقلل عدد الاختبارات المطلوبة لإنشاء مجموعاتٍ نقية (أي مجموعات المثيلات التي لها نفس القيمة الخاصة بالسمة المستهدفة). وإحدى الطرق لقياس نقاء مجموعةٍ ما هو استخدام معيار «الإنتروبيا» لكلود شانون. والحد الأدنى الممكن للإنتروبيا لمجموعةٍ ما هو صفر، وقيمة الإنتروبيا للمجموعة النقية هي صفر. تعتمد القيمة العددية القصوى للإنتروبيا الخاصة بمجموعة بيانات على حجم المجموعة وعدد الأنواع المختلفة من العناصر التي قد تُوجَد في المجموعة. وتمتلك أي مجموعةٍ الحد الأقصى من الإنتروبيا عندما تكون جميع عناصرها مختلفة الأنواع.[12] تختار هذه الخوارزمية السِّمة التي ستُختبَر عند عقدةٍ لكي تكون السمة التي تُنتج الإنتروبيا الأقل وزنًا بعد تقسيم مجموعة البيانات عند العقدة باستخدام هذه السمة. ويحسب وزن الإنتروبيا لسمةٍ ما عن طريق: (١) تقسيم مجموعة البيانات باستخدام السمة؛ (٢) حساب الإنتروبيا الخاصة بالمجموعات الناتجة؛ (٣) تقدير وزن كل إنتروبيا حسب الجزء من البيانات الموجود في المجموعة؛ (٤) ثم تجميع النتائج.

يُدرِج جدول ٤-٢ مجموعةَ بياناتٍ خاصة برسائل البريد الإلكتروني تُوصَف فيه كل رسالةٍ عن طريق عددٍ من السمات وما إذا كانت الرسالة عشوائية أم غير عشوائية. وتأخذ سمة «مرفق» القيمة «صواب» إذا كانت رسالة البريد الإلكتروني تحتوي على ملفٍّ مرفق، أما إذا لم يكن بها ملف مرفق، فستكون قيمة هذه السمة «خطأ» (في هذه العينة من رسائل البريد الإلكتروني، لا تحتوي أيٌّ من الرسائل على مرفق).

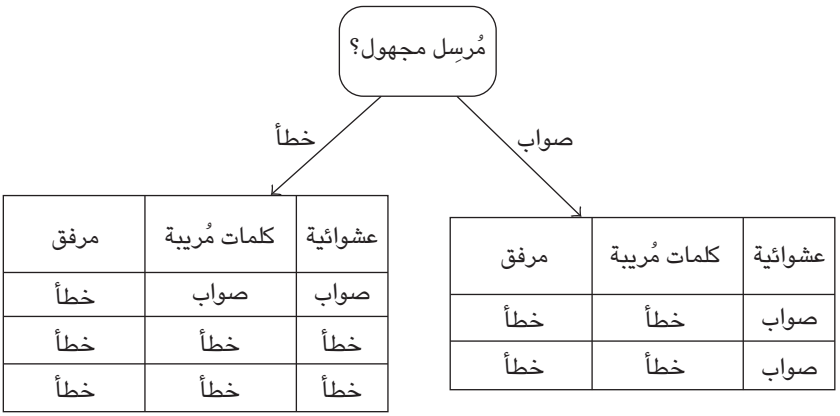
جدول ٤-٢: مجموعة بيانات خاصة برسائل البريد الإلكتروني: عشوائية أم غير عشوائية؟

مرفق	كلمات مُريبة	مُرسِل مجهول	عشوائية
خطأ	خطأ	صواب	صواب
خطأ	خطأ	صواب	صواب
خطأ	خطأ	خطأ	صواب

مرفق	كلمات مُريبة	مُرسِل مجهول	عشوائية
خطأ	خطأ	خطأ	خطأ
خطأ	خطأ	خطأ	خطأ

تأخذ سمة «كلمات مريبة» القيمة «صواب» إذا كان البريد الإلكتروني يحتوي على كلمةٍ أو أكثر من قائمةٍ مُحدَّدة مسبقًا من الكلمات المريبة. وتأخذ سمة «مُرسِل مجهول» القيمة «صواب» إذا كان مرسِل رسالة البريد الإلكتروني غير موجود في دليل جهات الاتصال الخاصة بالمتلقي. هذه هي مجموعة البيانات التي استخدمت لتدريب الهيكل الشجري لاتخاذ القرار المبيَّن في شكل ٤-٨. في هذه المجموعة للبيانات، تُعَد سِمات «مرفق»، و«كلمات مُريبة»، و«مُرسِل مجهول» هي السمات المدخلة وتُعد سمة «عشوائية» هي السمة المستهدفة. وتقسم سمة «مُرسِل مجهول» مجموعة البيانات إلى مجموعاتٍ أكثر نقاءً مقارنةً بأيٍّ من السمات الأخرى (تحتوي مجموعةٌ على مثيلاتٍ حيث «عشوائية = صواب» ومجموعة أخرى على مثيلات حيث «عشوائية = خطأ» وتضم الأخيرة معظم المثيلات). ونتيجة لذلك، توضع سمة «مُرسل مجهول» عند عقدة الجذر (انظر شكل ٤-٩). وبعد هذه التقسيمة المبدئية، تُصبح جميع المثيلات الموجودة على الفرع الأيمن لها نفس قيمة السمة المستهدفة. أما المثيلات الموجودة على الفرع الأيسر فتحتوي على قيمتَين مختلفتَين للسمة المستهدفة. وينتج عن تقسيم المثيلات على الفرع الأيسر باستخدام سِمة «كلمات مُريبة» مجموعتان نقيَّتان: الأولى حيث «عشوائية = خطأ» والثانية حيث «عشوائية = صواب». ومن ثم، تُختار سمة «كلمات مريبة» باعتبارها سمةً اختبارية للعقدة الجديدة على الفرع الأيسر (انظر شكل ٤-١٠). عند هذه النقطة، تكون مجموعة البيانات الفرعية الموجودة عند طرف كل فرعٍ نقية، وبالتالي تنتهي الخوارزمية وتنتج الهيكل الشجري لاتخاذ القرار المبين في شكل ٤-٨.

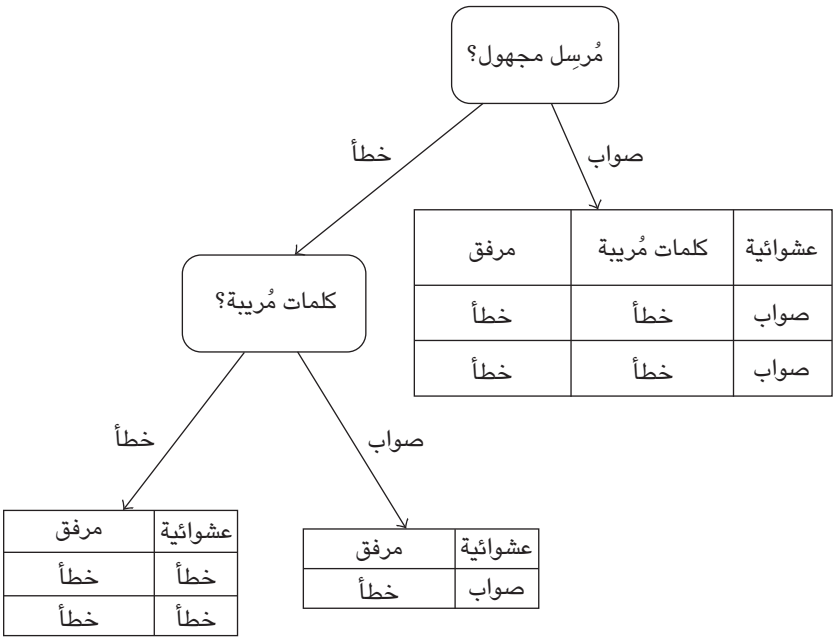
إحدى نقاط القوة التي تتمتع بها الهياكل الشجرية لاتخاذ القرار هي أنها يسهل فهمُها. كما أنه من الممكن ابتكار نماذج دقيقة للغاية استنادًا إلى هذه الهياكل. على سبيل المثال، يتألف «نموذج الغابة العشوائية» من مجموعة من الهياكل الشجرية، حيث يتم تدريب كل هيكل على عينةٍ فرعية من بيانات التدريب، ويكون التنبؤ الذي يُنتجه النموذج لاستعلامٍ فردي هو التنبؤ الأكثر شيوعًا عبر جميع أشجار الغابة. وعلى الرغم من أن



شكل ٤-٩: إنشاء عقدة الجذر في الهيكل الشجري.

الهياكل الشجرية لاتخاذ القرار تستطيع التعامل مع كلٍّ من البيانات الاسمية والترتيبية؛ فإنها تواجِه صعوبةً في التعامل مع البيانات العددية. ففي أحد الهياكل الشجرية، ينحدِر فرع مُنفصل من كل عقدة عن كل قيمةٍ في نطاق السمة الاختبارية عند العقدة. أما السِّمات العددية فلها عددٌ لا نهائي من القِيم في نطاقاتها، وذلك يعني ضمنيًّا أن الهيكل الشجري سيحتاج عددًا لا نهائيًّا من الفروع. وأحد الحلول لهذه المشكلة هو تحويل السمات العددية إلى سماتٍ ترتيبية، على الرغم من أن القيام بذلك يستلزِم تعيين الحدود المناسبة؛ وهو أمر قد يكون صعبًا أيضًا.

أخيرًا، نظرًا إلى أن خوارزمية تعلُّم الآلة القائمة على الهياكل الشجرية تُقسِّم مجموعة البيانات بصورةٍ متكررة كلما كبر الهيكل الشجري، فإنها تُصبح أكثر حساسيةً نحو التشويش (مثل المَثيلات المضللة). تصير المجموعة الفرعية للأمثلة (المَثيلات) الموجودة على كل فرعٍ أصغر فأصغر، وبالتالي تصير عينة البيانات التي تستند إليها كل قاعدةٍ تصنيفية أصغر. وكلما كانت عينة البيانات المستخدمة لتحديد القاعدة التصنيفية أصغر، صارت القاعدة أكثر حساسيةً للتشويش. ونتيجة لذلك، من الجيد الإبقاء على الهياكل الشجرية سطحية. ويتمثل أحد المناهج في إيقاف نموِّ الفرع عندما يكون عددُ المثيلات على الفرع لا يزال أقلَّ من الحَدِّ المحدد مسبقًا (على سبيل المثال، ٢٠ مثيلًا). وتسمح مناهج أخرى بنموِّ الهيكل الشجري، ثم بعد ذلك يُقلَّم. تستعين هذه المناهج عادةً باختباراتٍ إحصائية



شكل ٤-١٠: إضافة العقدة الثانية إلى الهيكل الشجري.

أو أداء النموذج على مجموعة من المثيلات المختارة بدقة لأداء هذه المهمة المتمثلة في تحديد التفريعات القريبة من الجزء السفلي من الهيكل الشجري والتي ينبغي إزالتها.

التحيُّز في علم البيانات

الهدف من تعلُّم الآلة هو إنشاء نماذج تشفِّر التعميمات الصحيحة استنادًا إلى مجموعات البيانات. وثمة عاملان مُهمان يساهمان في التعميم (أو النموذج) الذي ستولده خوارزمية تعلُّم الآلة من مجموعة البيانات. العامل الأول هو مجموعة البيانات التي تعمل عليها الخوارزمية. إذا لم تكن مجموعة البيانات تُمثل المجتمع الإحصائي، فلن يكون النموذج الذي تُولده الخوارزمية دقيقًا. على سبيل المثال، في موضعٍ سابق طوَّرنا نموذج انحدار خطِّي يتنبأ باحتمالية إصابة الفرد بمرض السكَّر من النوع الثاني استنادًا إلى مؤشر كتلة جسمه. تم توليد هذا النموذج من مجموعةِ بياناتٍ خاصة بالذكور الأمريكيين البالِغين ذوي البشرة البيضاء. ونتيجة لذلك، ليس من المرجَّح أن يكون هذا النموذج دقيقًا إذا

استُخدم للتنبؤ باحتمالية إصابة الإناث أو الذكور من عِرق مختلفٍ أو خلفيات عرقية مختلفة. يصف مصطلح «تحيُّز العينة» إلى أي مدًى يمكن أن تقدم العملية المستخدمة لاختيار مجموعة البيانات تحيزاتٍ إلى التحليل اللاحق، سواء أكان هذا التحليل إحصائيًّا أم لإنتاج نماذج تنبؤ باستخدام تعلُّم الآلة.

العامل الثاني الذي يؤثر على النموذج المتولد من مجموعة البيانات هو اختيار خوارزمية تعلُّم الآلة. هناك الكثير من هذه الخوارزميات، وكل واحدة منها تشفِّر طريقة مختلفة لاستنباط التعميمات من مجموعة البيانات. تُعرف نوعية التعميم الذي تشفره الخوارزمية بـ «التحيز الاستقرائي» الخاص بالخوارزمية (أو أحيانًا يُطلق عليه «تحيز النمذجة» أو «تحيز الاختيار»). على سبيل المثال، تشفر خوارزمية الانحدار الخطي تعميمًا خطيًّا من البيانات وبالتالي تتجاهل العلاقات غير الخطية التي ربما تتناسَب بدرجةٍ أكبر مع البيانات. عادةً ما يُفهم التحيز على أنه شيءٌ سيئ. على سبيل المثال، التحيز في اختيار العينة هو التحيز الذي سيحاول عالِم البيانات أن يتجنَّبه. ومع ذلك، في ظلِّ غياب التحيز الاستقرائي لا يمكن أن يكون هناك تعلُّم، وإنما ستكون الخوارزمية قادرةً على حفظ البيانات فقط.

ومع ذلك، نظرًا إلى أن خوارزميات تعلُّم الآلة تتحيَّز نحو البحث عن الأنواع المختلفة من الأنماط، ونظرًا إلى أنه لا يوجد تحيُّز استقرائي يصلح لجميع المواقف، فإنه لا يُوجَد ما يُعرَف بأفضل خوارزمية تعلُّم آلة. في الواقع، تنص النظرية المعروفة باسم «نظرية لا شيء مجاني»، أو «نظرية لا غداء مجاني» (Wolpert and Macready 1997) على أنه لا تُوجَد خوارزمية تعلُّم آلة أفضل تتفوَّق في متوسط أدائها على جميع الخوارزميات الأخرى عبر مجموعات البيانات المحتملة كافة. لذلك، عادةً ما تشمل مرحلة النمذجة الخاصة بالعملية القياسية المتعددة المجالات للتنقيب في البيانات إنشاء عدة نماذج باستخدام خوارزميات مختلفة، ومقارنة النماذج لتحديد أي خوارزمية تُولِّد أفضل نموذج. وتختبر هذه التجارب عمليًّا أي تحيزٍ استقرائي يُنتِج في المتوسط أفضل نماذج لمجموعة البيانات والمهمة المحددة.

تقييم النماذج: التعميم لا الحفظ

بمجرد أن يختار عالِم البيانات مجموعة من خوارزميات تعلُّم الآلة ليُجربها على مجموعة بيانات، تكون المهمة الرئيسية التالية هي وضع خطة فحصٍ لكيف ستُقيَّم النماذج التي

تم توليدها بواسطة هذه الخوارزميات. الهدف من خطة الفحص هو التأكُّد من أن التقييم يُقدِّم تقديراتٍ واقعية لأداء النموذج على البيانات التي لم يسبق رؤيتها. وليس من المرجَّح أن يبلي نموذج التنبؤ الذي يحفظ مجموعة البيانات فحسب بلاءً حسنًا في تقدير القِيَم من أجل الأمثلة الجديدة. وتتمثَّل إحدى المشكلات المرتبطة بحفظ البيانات فحسب في أن معظم مجموعات البيانات تحتوي على تشويش. وبالتالي، فإن نموذج التنبؤ الذي يحفظ البيانات فحسب يحفظ أيضًا التشويش الموجود في البيانات. وتتمثل مشكلة أخرى مرتبطة بحفظ البيانات فحسب في اختزال عملية التنبؤ على البحث في جدول؛ دون إيجاد حلٍّ لمشكلة كيفية التعميم من بيانات التدريب على أمثلةٍ جديدة لا توجَد في الجدول.

وجزء من خطة الفحص مرتبطٌ بكيفية استخدام مجموعة البيانات لتدريب النماذج واختبارها. يجب أن تُستخدَم مجموعة البيانات لغرَضَين مختلفَين. الغرض الأول هو إيجاد الخوارزمية التي تُولِّد أفضل نموذج. والغرض الثاني هو تقييم أداء التعميم الخاص بالنموذج الأفضل؛ أي إلى أي مدًى من المرجَّح أن يُجيد النموذج التعامل مع البيانات التي لم يسبق رؤيتها. والقاعدة الذهبية لتقييم النماذج هي أنه لا ينبغي أبدًا اختبار النماذج بناءً على نفس البيانات التي دُرِّبت عليها. إن الاستعانة بالبيانات نفسها لتدريب النماذج واختبارها أشبَهُ بإعطاء الطلَّاب أسئلة الاختبار في ليلة الامتحان. بالطبع، سيُبلي الطلَّاب بلاءً حسنًا في الاختبار؛ ولن تعكس درجاتهم إجادتهم الحقيقية للمادة الدراسية بوجهٍ عام. وهذا هو الوضع أيضًا مع نماذج تعلُّم الآلة: إذا قُيِّم النموذج بناءً على البيانات نفسها التي تدرَّب عليها، فستكون نتائج التقييم متفائلةً مقارنة بالأداء الحقيقي للنموذج. وتتمثل العملية المعيارية لضمان أن النماذج لا يُمكنها اختلاس النظر إلى بيانات الاختبار أثناء التدريب في تقسيم البيانات إلى ثلاثة أجزاء: مجموعة تدريب، ومجموعة تحقُّق، ومجموعة اختبار. وستختلف نِسَب هذه المجموعات باختلاف المشروعات، إلا أن تقسيم المجموعات على هذا النحو: ٥٠:٢٠:٣٠ و٤٠:٢٠:٤٠ هو التقسيم الشائع. وحجم مجموعة البيانات هو عامل رئيسي في تحديد التقسيمات: بوجهٍ عام، كلما كانت مجموعة البيانات كبيرة، كانت مجموعة الاختبار كبيرة. تُستخدَم مجموعة التدريب لتدريب مجموعةٍ مبدئية من النماذج. ثم تُستخدَم مجموعة التحقُّق لمقارنة أداء هذه النماذج على البيانات التي لم يسبق رؤيتها. تمكِّننا مقارنة أداء هذه النماذج المبدئية بمجموعة التحقُّق من تحديد أي الخوارزميات تُولِّد النموذج الأفضل. وبمجرد اختيار أفضل خوارزمية، يمكن دمج مجموعة التدريب والتحقُّق معًا لتُصبح مجموعة تدريب أكبر، وتُغذَّى أفضل

خوارزمية بهذه المجموعة لكي تُنشئ النموذج النهائي. ومن الأهمية بمكانٍ ألا تُستخدَم مجموعة الاختبار خلال عملية اختيار أفضل خوارزمية، ولا ينبغي أن تُستخدَم لتدريب هذا النموذج النهائي. وإذا اتبعت هذه التحفظات، إذن يمكن استخدام مجموعة الاختبار لتقدير أداء التعميم الخاص بهذا النموذج النهائي على البيانات التي لم يسبق رؤيتها.

> القاعدة الذهبية لتقييم النماذج هي أنه لا ينبغي أبدًا اختبار النماذج بناءً على نفس البيانات التي تدرَّبت عليها.

المكون الرئيسي الآخر لخطة الفحص هو اختيار معايير تقييمٍ مناسبة لاستخدامها أثناء التجربة. بوجهٍ عام، تُقيَّم النماذج استنادًا إلى أي مدًى تتوافق عادةً مُخرجات النموذج مع المخرجات المذكورة في مجموعة الاختبار. فإذا كانت السِّمة المستهدفة قيمةً عددية، إذن مجموع الأخطاء التربيعية هي إحدى الطرق لقياس دقة النموذج على مجموعة الاختبار. وإذا كانت السمة المستهدَفة اسميةً أو ترتيبية، إذن تكون أسهل طريقة لتقييم دقة النموذج هي حساب نسبة الأمثلة في مجموعة الاختبار التي تنبَّأ بها النموذج على النحو الصحيح. ومع ذلك، من المهم في بعض السياقات تضمين تحليل الخطأ داخل التقييم. إذا كان النموذج مستخدَمًا في سياق تشخيصٍ طبيٍّ مثلًا، يكون الأمر أكثر خطورة إذا شخَّص النموذج مريضًا على أنه شخصٌ سليم مما إذا كان الشخص سليمًا وشُخِّص على أنه مريض. ربما يسفر تشخيص شخصٍ مريض على أنه سليم عن إعادته إلى المنزل دون تلقِّي العناية الطبية المناسبة، ولكن إذا شخَّص النموذج شخصًا سليمًا على أنه مريض، فمن المرجَّح اكتشاف هذا الخطأ بواسطة الفحوصات الطبية التالية التي سيُجريها المريض. وبالتالي ينبغي أن يُعطي مقياس التقييم المستخدَم لتقييم هذه الأنواع من النماذج وزنًا أكبر لنوعٍ من الأخطاء على غيره عند تقييم أداء النموذج. وبمجرد أن تُنشأ خطة الاختبار، يستطيع عالِم البيانات أن يبدأ تدريب النماذج وتقييمها.

ملخص

استُهِلَّ هذا الفصل بقول إن علم البيانات بمثابة علاقة شراكة بين عالِم البيانات والكمبيوتر. ويوفر تعلُّم الآلة مجموعةً من الخوارزميات التي تُولِّد نماذج من مجموعةٍ

كبيرة من البيانات. ومع ذلك، ستعتمد فائدة هذه النماذج من عدمها على خبرة عالِم البيانات. ولكي ينجح مشروع علم البيانات، ينبغي أن تكون مجموعة البيانات ممثلة للمجال وينبغي أن تتضمن سماتٍ ذات صلة. ينبغي أن يُقيِّم عالِم البيانات مجموعة من خوارزميات تعلُّم الآلة لتحديد الخوارزمية التي تولد أفضل النماذج. وينبغي أن تتبع عملية تقييم النموذج القاعدة الذهبية التي تنص على أن النموذج ينبغي ألا يُقيَّم بناءً على البيانات التي تدرَّب عليها.

حاليًّا المعيار الأساسي، في أغلب مشروعات علم البيانات، لاختيار النموذج الذي سيُستخدم هو دقة النموذج. ومع ذلك، في المستقبل القريب، ربما تؤثر لوائح الخصوصية واستخدام البيانات على اختيار خوارزميات تعلُّم الآلة. على سبيل المثال، ستدخل اللائحة العامة لحماية البيانات حَيِّز التنفيذ في الاتحاد الأوروبي في ٢٥ مايو ٢٠١٨. سنناقش هذه اللوائح فيما يخصُّ استخدام البيانات في الفصل السادس، ولكن في الوقت الراهن نودُّ أن نُشير إلى أنه ربما يبدو أن بعض البنود في هذه اللائحة تفرض «الحق في التفسير» فيما يخص عمليات اتخاذ القرار الآلية.[13] ومن بين الآثار المحتملة لهذا الحق هو أنه ربما يصير استخدام النماذج، مثل الشبكات العصبية التي يصعب تفسير قراراتها المرتبطة بالأفراد، أمرًا إشكاليًّا. وفي ظل هذه الظروف، ربما تجعل الشفافية وسهولة تفسير بعض النماذج، مثل الهياكل الشجرية لاتخاذ القرار، استخدام هذه النماذج أكثر ملاءمة.

في النهاية، العالم يتغير، ولكن النماذج لا تتغير. ويكمُن في صميم عملية تعلُّم الآلة الخاصة بإنشاء مجموعة البيانات وتدريب النموذج وتقييمه افتراض أن المستقبل لن يختلف عن الماضي في شيء. ويُعرَف هذا الافتراض باسم «افتراض الثبات»: العمليات أو السلوكيات التي تُنمذَج تتَّسِم بالثبات عبر الزمن (أي أنها لا تتغير). ومجموعات البيانات في حدِّ ذاتها قديمة بمعنى أن البيانات هي تمثيلات للملاحظات التي دُوِّنت في الماضي. ولذا، في الواقع، تبحث خوارزميات تعلُّم الآلة عبر الماضي عن أنماطٍ ربما تُعمَّم على المستقبل. ومن الواضح أن هذا الافتراض لا تثبُت صحته على الدوام. يستخدم علماء البيانات مصطلح «انحراف المفاهيم» لوصف كيف قد تتغير العملية أو السلوك، أو تنحرف، مع مرور الوقت. ولهذا السبب تتقادم النماذج وتحتاج إلى إعادة تدريبها من جديد ولهذا السبب تتضمن العملية القياسية المتعددة المجالات للتنقيب في البيانات الدائرة الخارجية المبينة في شكل ٢-٣ للتأكيد على أن علم البيانات يتَّسِم بالتكرارية. يجب على العمليات أن تضمن مرحلة ما بعد نشر النموذج للتأكد من أن النموذج لم يتقادم، وعندما

يتقادم، يجب إعادة تدريبه. ومعظم هذه القرارات لا يمكن تنفيذها آليًّا، وإنما تتطلَّب رؤيةً ومعرفة بشرية. سيجيب جهاز الكمبيوتر عن الأسئلة التي تطرح عليه، ولكن ما لم يُولَ الاهتمام، فمن السهل أن يُطرَح السؤال الخطأ.

الفصل الخامس

مهام علم البيانات القياسية

واحدة من أهم المهارات التي يجب أن يتمتَّع بها عالِم البيانات هي القدرة على صياغة مشكلة واقعية على شكل مهمة قياسية خاصة بعلم البيانات. ويمكن تصنيف معظم مشروعات علم البيانات على أنها تنتمي إلى واحدة من أربع فئات عامة للمهام:

- التجميع (أو التجزئة)
- اكتشاف الشذوذ (أو القيم الشاذة)
- التنقيب عن قواعد الارتباط
- التنبؤ (بما في ذلك المسائل الفرعية الخاصة بالتصنيف والانحدار)

قد يساعد فهم المهمة التي يستهدفها المشروع في اتخاذ الكثير من القرارات المتعلقة بالمشروع نفسه. على سبيل المثال، يتطلَّب تدريب نموذج التنبؤ أن يتضمن كلٌّ مثيل من المثيلات في مجموعة البيانات قيمة السمة المستهدفة. وبالتالي، تُرشدنا معرفة أن المشروع يتنبأ (عبر المتطلبات) فيما يخصُّ تصميم مجموعة البيانات. إن فهم المهمة يساعد أيضًا في تحديد أي خوارزميَّات تعلُّم الآلة التي يجب استخدامها. وعلى الرغم من وجود عددٍ كبير من خوارزميات التعلُّم، فكل خوارزمية مُصمَّمة لمهمةٍ معينة خاصة بالتنقيب في البيانات. على سبيل المثال، خوارزميات تعلُّم الآلة التي تُولِّد نماذج الهياكل الشجرية مُصمَّمة بالأساس لمهامِّ التنبؤ. وثمة علاقة «متعدد إلى واحد» بين خوارزميات تعلُّم الآلة والمهمة، وبالتالي فإن معرفة المهمة لا تخبرك بالخوارزمية التي يجب استخدامها على وجه التحديد، إلا أنها تُحدد مجموعةً من الخوارزميات المصممة لأداء المهمة. ونظرًا إلى أن مهمة علم البيانات تؤثر على تصميم مجموعةِ البيانات واختيار خوارزمية التعلُّم، يجب اتخاذ القرار الخاص بأي مهمةٍ سيستهدفها المشروع في مرحلةٍ مبكرة من مراحل المشروع، حبَّذا أثناء مرحلة فهم طبيعة العمل من مراحل العملية القياسية المتعددة المجالات للتنقيب في

البيانات. ومن أجل توفير فهمٍ أفضل لكل مهمةٍ من هذه المهام، يوضح هذا الفصل كيفية صياغة بعض مشكلات العمل القياسية على هيئة مهام.

من هم عملاؤنا؟ (التجميع)

واحد من مجالات تطبيق علم البيانات الأكثر شيوعًا في أوساط العمل التجاري هو دعم حملات التسويق والمبيعات. يتطلَّب تصميم حملة تسويقية موجَّهة نحو عملاء بعينِهم فَهم العميل المستهدَف. ولدى معظم الشركات مجموعة متنوعة من العملاء ذوي احتياجاتٍ متنوعة، وبالتالي من المرجَّح أن يفشل استخدام منهج واحدٍ يناسِب الجميع مع شريحةٍ كبيرة من قاعدة العملاء. ثمة منهج أفضل يتمثل في تحديد عددٍ من نماذج الشخصيات للعملاء أو الملفَّات التعريفية للعملاء، بحيث يكون كلٌّ منها ذا صلةٍ بشريحة مهمة من قاعدة العملاء، وبالتالي ذا صلة بتصميم حملات التسويق الموجَّهة لكل نموذج شخصية. ويمكن إنشاء هذه النماذج باستخدام الخبرة التخصُّصية، إلا أنه من الجيد بوجهٍ عامٍّ أن تستند نماذج الشخصيات على البيانات التي تمتلكها الشركة عن عملائها. وكثيرًا ما يُغفِل الحدس البشري تجاه العملاء شرائح مهمة مُبهمة المعالم أو لا يوفر مستوى الدقة المطلوب من أجل التسويق المفصَّل. على سبيل المثال، تذكُر ميتا إس براون (٢٠١٤) كيف أن الصورة النمطية المعروفة لـ «الأم المهتمة بتمرين كرة القدم» (أي ربة المنزل التي تعيش في الضواحي وتقضي وقتًا طويلًا في توصيل أبنائها بالسيارة إلى تمرين كرة القدم أو أية رياضةٍ أخرى) لم تُصنف ضمن قاعدة عملاء في أحد مشروعات علم البيانات. غير أن الاستعانة بعملية تجميعٍ مبنية على البيانات أظهرت نماذج شخصياتٍ أكثر تحديدًا للعملاء، مثل «الأمهات العاملات بدوامٍ كامل خارج المنزل واللاتي لديهنَّ أطفال صغار يمكثون في مراكز رعاية نهارية» و«الأمهات العاملات بدوامٍ جزئي ولديهنَّ أولاد في المرحلة الثانوية» و«السيدات المهتمات بالغذاء والصحة واللاتي ليس لديهنَّ أبناء». وتُحدد هذه النماذج الخاصة بالعملاء أهدافًا أوضحَ من أجل حملات التسويق وربما تُسلط الضوء مسبقًا على شرائح غير معلومة في قاعدة العملاء.

> كثيرًا ما يُغفِل الحدس البشري تجاه العملاء شرائح مُهمةً مُبهمةَ المعالم أو لا يوفر مستوى الدقة المطلوب من أجل التسويق المفصَّل.

ويتمثل منهج علم البيانات القياسي لهذا النوع من التحليلات في صياغة المشكلة على هيئة مهمة «تجميع». وينطوي التجميع على فرز المثيلات في مجموعة البيانات إلى مجموعاتٍ فرعية تحتوي على المثيلات المتشابهة. ويتطلَّب التجميع عادة مُحللًا مُتخصصًا ليُقرر أولًا عدد المجموعات الفرعية التي يودُّ تحديدها في البيانات. وربما يكون هذا القرار معتمدًا على معرفةٍ بالمجال أو على معرفة بأهداف المشروع. بعد ذلك تُشغَّل خوارزمية التجميع على البيانات مع إدخال العددِ المرغوب من المجموعات الفرعية بصفته أحدَ مُعامِلات الخوارزمية. وعندئذٍ تُنشئ الخوارزمية هذا العدد من المجموعات الفرعية من خلال تجميع المَثيلات بناءً على تشابُهِ قِيَم سِماتها. وبمجرد أن تُنشئ الخوارزمية العناقيد (التجميعات)، يُراجعها شخصٌ خبير بالمجال لتحديد ما إذا كانت ذات مغزًى أم لا. وفي سياق تصميم حملة التسويق، تشتمل هذه المراجعة على التأكُّد مما إذا كانت المجموعات تعكس نماذج شخصيات العملاء بصورةٍ منطقية أو تُحدد النماذج الشخصية الجديدة التي لم تكن توضع في الحسبان من قبل.

تعتبر السمات التي يمكن استخدامها لوصف العملاء من أجل وضعهم في مجموعاتٍ كثيرة للغاية؛ ولكنها تضمُّ على سبيل المثال معلوماتٍ فئوية (مثل العمر، والنوع، وما إلى ذلك)، ومعلومات عن الموقع (مثل الرمز البريدي، أو العنوان في القرية أو المدينة، وما إلى ذلك)، ومعلومات خاصة بالمعاملات (مثل ما المنتجات أو الخدمات التي قاموا بشرائها)، والإيرادات التي تُحققها الشركة منهم، ومنذ متى وهم يتعاملون مع الشركة، وما إذا كانوا أعضاء في برنامج بطاقة الولاء، وما إذا كانوا قد سبق لهم إرجاع مُنتجٍ أو تقديم شكوى بشأن الخدمة، وما إلى ذلك. وكما هو الحال بالنسبة إلى جميع مشروعات علم البيانات، فإن أحد أكبر التحدِّيات التي يُواجهها التجميع (تكوين العناقيد) هو تحديد أي السِّمات يُدمَج وأيها يُستبعَد لتحقيق أفضل النتائج. وينطوي اتخاذ هذا القرار بشأن اختيار السمات على تكرارَ التجارب والتحليل البشري لنتائج كلِّ عملية تكرار.

أشهر خوارزمية من خوارزميات تعلُّم الآلة مستخدِمة للتجميع هي خوارزمية «التجميع بالمتوسطات» (أو ما يُعرف بالإنجليزية بخوارزمية k-means). ويُشير حرف k المستخدَم في التسمية الإنجليزية إلى أن الخوارزمية تبحث في البيانات عن التجميعات (العناقيد) التي عددها k. وقيمة k مُحدَّدة مسبقًا وغالبًا ما تُحدد من خلال عمليةٍ قائمة على التجربة والخطأ بقِيَمٍ مختلفة لـ k. وتفترض هذه الخوارزمية أن جميع السمات التي تصف العملاء في مجموعة البيانات هي سمات عددية. وإذا تضمنت مجموعة البيانات

سماتٍ غير عددية، إذن يجب تعيين هذه السمات إلى قِيَم عددية من أجل استخدام خوارزمية التجميع بالمتوسطات؛ وإلا يجب أن تُعدل الخوارزمية من أجل التعامل مع هذه القيم غير العددية. وتتعامل الخوارزمية مع كلِّ عميلٍ باعتباره نقطة في سحابة النقاط (أو مخطط التشتُّت)، حيث يتحدَّد موضع العميل من خلال قِيَم سماته في ملفِّه التعريفي. والهدف من الخوارزمية هو إيجاد موضع مركز كل عنقود في سحابة النقاط. وبما أن هناك عدد k من العناقيد، إذن فهناك عدد k من مراكز العناقيد (أو المتوسطات) — ومن هنا تأتي تسمية الخوارزمية.

تبدأ هذه الخوارزمية بانتقاء عدد k من المثيلات بوصفِها مراكز عناقيد أولية. وأفضل ما يمكن القيام به حاليًّا هو استخدام خوارزمية تُسمَّى «خوارزمية التجميع بالمتوسطات++» لانتقاء مراكز العناقيد الأولية. والفكرة الأساسية وراء خوارزمية التجميع بالمتوسطات++ تحديدًا هي أنه من الأفضل نشر مراكز العناقيد الأولية بقدْر الإمكان. ومِن ثَم، في خوارزمية التجميع بالمتوسطات++ يُحدَّد أول مركز عنقودٍ عن طريق التحديد العشوائي لإحدى المثيلات في مجموعة البيانات. ويُحدَّد مركز العنقود الثاني وما يَليه من مراكز عن طريق تحديد مثيلٍ من مجموعة البيانات مع احتمالية أن المثيل المحدد يتناسب مع المسافة المربعة إلى أقرب مركز عنقود موجود. وبمجرد تحديد جميع مراكز العناقيد ذات العدد k، تعمل الخوارزمية عن طريق تكرار عملية تتكوَّن من خطوتَين: أولًا: توزيع كل مثيلٍ على أقرب مركز عنقود، ثم ثانيًا: تحديث مركز العنقود ليكون في منتصف المثيلات الموزَّعة عليه. وفي أول تكرار، تُوزع المَثيلات على أقرب مركز عنقود تُنتجه خوارزمية التجميع بالمتوسطات++ ثم تُحرك مراكز العناقيد هذه بحيث توضع في وسط المثيلات الموزعة عليها. ومن المرجَّح أن يؤدي نقل مراكز العناقيد إلى وضعها على نحوٍ أقرب من بعض المثيلات وأبعد عن مثيلات أخرى (من ذلك أن تكون أبعد عن بعض المثيلات الموزعة على مركز العنقود). ثم يُعاد توزيع المثيلات مرةً أخرى على أقرب مركز عنقود مُحدَّث. وستظلُّ بعض المثيلات موزعةً على المركز نفسه، وربما يُعاد توزيع مثيلات أخرى على مركز عنقود آخر. وتستمرُّ هذه العملية الخاصة بتوزيع المثيلات وتحديث المراكز إلى أن تتوقف المثيلات عن التوزيع على مركز عنقودٍ آخر أثناء عملية التكرار. وخوارزمية التجميع بالمتوسطات ليست خوارزمية حتمية، بمعنى أنه من المرجح أن تُسفر مواضع البدء المختلفة لمراكز العناقيد عن عناقيد مختلفة. ونتيجة لذلك، تُشغَّل الخوارزمية عادةً عدة مرات، ثم تُقارَن نتائج مرات التشغيل المختلفة هذه لتحديد أي من هذه العناقيد أكثر منطقيةً في ضوء معرفة عالِم البيانات وفهمه للمجال.

وكما هو الحال بالنسبة إلى جميع مشروعات علم البيانات، فإن أحد أكبر التحديات التي يواجهها التجميع هو تحديد أي السِّمات يُدمج وأيها يُستبعَد لتحقيق أفضل النتائج.

عندما يُحكَم على مجموعة من عناقيد نماذج شخصيات العملاء بأنها مفيدة، عادةً ما تُمنح هذه العناقيد أسماءً لتعكس السمات الرئيسية الخاصة بنماذج الشخصيات. ويُحدد مركز كل عنقودٍ نموذج شخصية مختلفًا، حيث ينتج وصف نموذج الشخصية من قِيم السمات الخاصة بمركز العنقود ذي الصلة. وخوارزمية التجميع بالمتوسطات ليست مُلزمةً بإنتاج عناقيد متساوية الحجم، بل إنها من المرجَّح أن تُنتج عناقيد مختلفة الحجم. وأحجام العناقيد من الممكن أن تكون مفيدة، لأنها ربما تساعد في توجيه عملية التسويق. على سبيل المثال، قد تكشف عملية التجميع (تكوين العناقيد) عن عناقيد صغيرة مركزة من العملاء تَغفُل عنها حملات التسويق الحالية. أو ربما تركز استراتيجية بديلة على عناقيد تحتوي على عملاء يجلبون نسبةً كبيرة من الإيرادات. وأيًّا كانت استراتيجية التسويق المتبعة، يُعتبر فهم الشرائح داخل قاعدة العملاء شرطًا أساسيًّا لنجاح التسويق.

إحدى مُميزات التجميع كمنهجٍ تحليلي هو أنه يمكن تطبيقه على معظم أنواع البيانات. ونظرًا إلى تعدُّد استعمالاته، عادةً ما يُستخدم التجميع كأداة لاستكشاف البيانات أثناء مرحلة فهم البيانات في كثيرٍ من مشروعات علم البيانات. كما يُعَد التجميع مفيدًا في مجموعةٍ واسعة من المجالات الأخرى. على سبيل المثال، استخدم التجميع لتحليل الطلاب المسجلين في دورة دراسية مُعينة من أجل تحديد مجموعات الطلاب الذين يحتاجون إلى دعمٍ إضافي أو الذين يُفضلون مناهج تعليميةً مختلفة. كما أنه استُخدِم من أجل تحديد مجموعات المستندات المتشابهة في مجموعةٍ من المستندات، وفي مجال العلوم، استُخدِم في مجال المعلوماتية الحيوية لتحليل تسلسُل الجينات في تحليل الرقائق الجينية الدقيقة.

هل هذا احتيال؟ (اكتشاف الشذوذ)

يتضمن اكتشاف الشذوذ أو تحليل القيم الشاذة البحث عن مَثيلات لا تتوافق مع البيانات النمطية الواردة في مجموعة البيانات وتحديد هذه المثيلات. وكثيرًا ما يُشار إلى هذه الحالات غير المتوافقة بـ «قيم الشذوذ» أو «القيم الشاذة». وغالبًا ما يُستخدم اكتشاف الشذوذ في

تحليل المعاملات المالية من أجل رصد أنشطة الاحتيال المحتملة وبدء تحقيقاتٍ بشأنها. فعلى سبيل المثال، ربما يؤدي اكتشاف الشذوذ إلى كشف النقاب عن معاملاتٍ احتيالية لبطاقة الائتمان من خلال تحديد المعاملات التي حدثت في مكانٍ غير معتاد أو تلك التي تضمنت مبالغ كبيرةً غير معتادة مقارنةً بمعاملاتٍ أخرى مُسجَّلة على بطاقةٍ ائتمانية مُعينة.

يتمثل المنهج الأول الذي تستعين به أغلب الشركات لاكتشاف الشذوذ في تحديد عددٍ من القواعد يدويًّا بناءً على الخبرة بالمجال والتي تُساعد في تحديد الأحداث الشاذة. وعادةً ما يتمُّ تحديد هذه المجموعة من القواعد باستخدام لغة الاستعلام الهيكلية أو أية لغةٍ أخرى وتُطبَّق على البيانات الواردة في قواعد بيانات الشركة أو مخزن البيانات. لقد بدأت بعض لغات البرمجة تضمين أوامر مُحددة لتيسير عملية تشفير هذه الأنواع من القواعد. فعلى سبيل المثال، تشمل تطبيقات قواعد البيانات المكتوبة بلغة الاستعلام الهيكلية الآن دالَّة التعرُّف على الأنماط المتطابقة (أو ما تُعرف باسم داخلة MATCH_RECOGNIZE) من أجل تيسير التعرُّف على الأنماط المتطابقة في البيانات. ويتمثل نمط شائع لعمليات الاحتيال الخاصة ببطاقات الائتمان عندما تُسرق بطاقة ائتمان، ويتأكد السارق أولًا من أن البطاقة لا تزال قيدَ العمل وذلك من خلال شراء شيءٍ صغير باستخدام البطاقة، وإذا تمت تلك المعاملة بنجاح، يُتبِع السارق عملية الشراء بعمليةٍ أخرى لشيء باهظ الثمن بأسرع ما يمكن قبل أن توقَف البطاقة. تُمكِّن دالةُ التعرف على الأنماط المتطابقة بلغة الاستعلام الهيكلية مبرمجي قواعد البيانات من كتابةِ نصوص برمجة تتعرَّف على سلاسل المعاملات التي تتم على بطاقة الائتمان التي تتطابق مع هذا النمط وإما توقف البطاقة تلقائيًّا أو تُرسِل تحذيرًا إلى الشركة المصدِرة لبطاقة الائتمان. وبمرور الوقت، ومع التعرُّف على مزيدٍ من المعاملات الشاذة — على سبيل المثال من خلال العملاء الذين يُبلِّغون عن معاملاتٍ احتيالية — يُتوسَّع في مجموعة القواعد التي تُحدد المعاملات الاحتيالية من أجل التعامل مع هذه المثيلات الجديدة.

العيب الأساسي في المنهج القائم على القواعد المستخدَم لاكتشاف الشذوذ هو أن تحديد القواعد بهذه الطريقة يعني أن الأحداث الشاذة لن يُتعرَّف عليها إلا بعد وقوعها بالفعل ولفت انتباه الشركة إليها. فمن الناحية المثالية، تودُّ معظم المؤسسات أن تتمتَّع بالقدرة على تحديد القِيَم الشاذة فور ظهورها لأول مرة أو إذا ظهرت رغم عدم الإبلاغ عنها. يُعد اكتشاف الشذوذ، في بعض النواحي، نقيضًا للتجميع: الهدف من التجميع هو

تحديد مجموعات المثيلات المتشابهة، في حين أن الهدف من اكتشاف الشذوذ هو العثور على المثيلات المختلفة عن باقي البيانات في مجموعة البيانات. ومن هذا المنطلق، يمكن الاستعانة بالتجميع لتحديد القِيَم الشاذَّة تلقائيًّا. وثمة منهجان للاستعانة بالتجميع في اكتشاف الشذوذ. المنهج الأول هو أنه ستُجمَّع البيانات العادية معًا، وستكون السجلات الشاذة في عناقيد منفصلة. ستكون العناقيد التي تحتوي على السجلَّات الشاذة صغيرة، وبالتالي ستكون مختلفة بوضوح عن العناقيد الكبيرة التي تُوجد فيها الكتلة الأساسية من السجلات. والمنهج الثاني هو قياس المسافة بين كلِّ مثيلٍ ومركز العنقود. وكلما كان المثيل بعيدًا عن مركز العنقود، زاد الاحتمال أن يكون شاذًّا وبالتالي يستلزم التحقيق.

وثمة منهج آخر لاكتشاف الشذوذ وهو تدريب نموذج تنبُّؤ، مثل هيكلٍ شجري، لتصنيف المَثيلات إما شاذة أو غير ذلك. ومع ذلك، تدريب هذا النموذج يستلزم عادةً مجموعة بيانات تدريبية تحتوي على سجلَّات شاذة وأخرى عادية. ولا يكفي أن يكون لديك عددٌ قليل من المثيلات التي تحتوي على سجلَّات شاذة؛ فمن أجل تدريب نموذج تنبؤ عادي، يجب أن تحتوي مجموعة البيانات على عددٍ معقول من المثيلات من كل فئة. ومن الناحية المثالية، يجب أن تكون مجموعة البيانات متوازنة؛ في حالة النتيجة الثنائية، من شأن التوازن أن يعني تقسيم البيانات بنسبة ٥٠:٥٠. وبوجهٍ عام، لا يمكن الحصول على هذا النوع من بيانات التدريب لاكتشاف الشذوذ؛ إذ بحكم تعريفها، القيم الشاذة هي أحداث نادرة، ربما تظهر في ١ إلى ٢ بالمائة من البيانات أو أقل. وهذا القصور في البيانات يعوق استخدام نماذج التنبُّؤ العادية الجاهزة. ومع ذلك، ثمة خوارزميات تعلُّم الآلة تُعرَف باسم «مُصنفات الفئة الواحدة» مُصمَّمة للتعامل مع نوعية البيانات غير المتوازنة التي تتميز بها مجموعات بيانات اكتشاف الشذوذ.

تُعد خوارزمية «آلة المتَّجه الدَّاعم ذات الفئة الواحدة» من مصنفات الفئة الواحدة المعروفة. بصفة عامة، تفحص هذه الخوارزمية البيانات كوحدة واحدة (أي فئة واحدة) وتُحدد السمات الأساسية للمثيلات وسلوكها المتوقع. وتشير الخوارزمية بعد ذلك إلى مدى تشابُه أو عدم تشابُه كل مثيلٍ عن السمات الأساسية والسلوك المتوقع. يمكن استغلال هذه المعلومات بعد ذلك لتحديد المثيلات التي تستحق المزيد من التحقيق (أي القيم الشاذة المسجلة). وكلما زاد اختلاف المثيل، زادت احتمالية ضرورة التحقُّق منه.

وتعني حقيقة أن القيم الشاذة نادرة أنه قد يسهل عدم الانتباه لها ويصعب تحديدها. ونتيجة لذلك، عادةً ما يجمع عالِم البيانات عددًا من النماذج المختلفة لاكتشاف

القِيَم الشاذة. الفكرة هي أن النماذج المختلفة ستكتشف نوعيات مختلفة من القيم الشاذة. وبوجهٍ عام، هذه النماذج تُستخدَم لتكملة القواعد المعروفة داخل المؤسسة التي حددت الأنواع المختلفة من الأنشطة الشاذة. تُدمج النماذج المختلفة معًا في حلٍّ لإدارة القرار يُمكِّننا من الاستفادة من التنبؤات الناتجة من كل نموذج في تنوير القرار الخاص بناتج التنبؤ النهائي. على سبيل المثال، إذا صنف نموذج واحد فقط من أصل أربعة نماذج إحدى المعاملات على إنها معاملة احتيالية، فربما يقرر نظام اتخاذ القرار أنها ليست معاملة احتيالية حقيقية، وقد تُتجاهَل المعاملة. وعلى العكس من ذلك، إذا صنفت ثلاثة أو أربعة نماذج من أصل الأربعة نماذج المعاملة على أنها معاملة احتيالية محتملة، فسيتم وضع علامة بجوار المعاملة لكي يتحقق منها عالِم البيانات.

ويمكن تطبيق عملية اكتشاف الشذوذ في الكثير من المجالات الإشكالية بخلاف حالات الاحتيال في بطاقات الائتمان. وبصفة عامة، يُستخدم اكتشاف الشذوذ في غرف المقاصَّة لتحديد المعاملات المالية التي تستلزم المزيد من التحقيق لتحديد ما إذا كانت حالات احتيال مُحتملة أو غسيل أموال. ويُستخدم في تحليل مطالبات التأمين لتحديد ما لا يتوافق مع المطالبات النموذجية للشركة. وفي الأمن السيبراني، تُستخدَم لتحديد عمليات اقتحام الشبكة من خلال رصد حالات القرصنة المحتملة أو السلوك غير النمطي من قبل الموظفين. وفي المجال الطبي، قد يكون تحديد القيم الشاذة في السجلَّات الطبية مفيدًا في تشخيص الأمراض ودراسة العلاجات وآثارها على الجسم. وفي النهاية، ومع انتشار أجهزة الاستشعار والاستخدام المتزايد لتكنولوجيا إنترنت الأشياء، سيلعب اكتشاف الشذوذ دورًا مهمًّا في مراقبة البيانات وتحذيرنا عند وقوع أحداث شاذة تستلزم اتخاذ إجراء.

هل تريد بطاطس مقلية مع هذا الطلب؟ (التنقيب عن قواعد الارتباط)

يُعد البيع المتقاطع — أو الاقتراح على العملاء الذين يشترون منتجات أنهم ربما بحاجةٍ أيضًا إلى شراء منتجات تكميلية أخرى أو منتجات ذات صلة — من الاستراتيجيات القياسية في المبيعات. الفكرة هي زيادة إجمالي معدل إنفاق العملاء من خلال حثِّهم على شراء المزيد من المنتجات وفي الوقت نفسه تحسين خدمة العملاء من خلال تذكيرهم بمنتجات أرادوا شراءها على الأرجح؛ ولكنهم ربما نسوها. والمثال الكلاسيكي على البيع المتقاطع هو عندما يسأل نادل في مطعم هامبورجر زبونًا طلبَ للتو هامبورجر: «هل تريد بطاطس مقلية مع هذا الطلب؟» تعرف محلات السوبر ماركت ومتاجر البيع

بالتجزئة أن المتسوِّقين يشترون المنتجات في مجموعاتٍ ويستغلون هذه المعلومة لخلق فُرَصٍ للبيع المتقاطع. على سبيل المثال، عملاء السوبر ماركت الذين يشترون النقانق من المرجَّح أن يشتروا كاتشب وبيرة أيضًا. وبالاستعانة بهذه النوعية من المعلومات، يستطيع المتجر أن يُصمم نسقًا معينًا لتوزيع المنتجات على الأرفف. وبالتالي، فإن وضع النقانق والكاتشب والبيرة بعضها بجوار بعض على أرفف المتجر يساعد العملاء في جمع هذه المجموعة من المنتجات سريعًا وربما يؤدي أيضًا إلى زيادة المبيعات لأن العملاء الذين يشترون النقانق ربما يرون منتجَي الكاتشب والبيرة اللذين نسوا حاجتهم إليهما وبالتالي يشترونهما. إن فهم هذه النوعيات من الارتباط بين المنتجات هو أساس جميع عمليات البيع المتقاطع.

يُعتبر التنقيب عن قواعد الارتباط تقنية تحليل بيانات غير خاضعة للإشراف تهدف إلى البحث عن مجموعات العناصر التي كثيرًا ما يتكرَّر وجودها معًا. ويتمثل المثال الكلاسيكي للتنقيب عن قواعد الارتباط في «تحليل سلة التسوق»؛ حيث تحاول متاجر البيع بالتجزئة تحديد مجموعات السلع التي تُشترى معًا مثل النقانق والكاتشب والبيرة. ومن أجل إجراء هذا النوع من تحليل البيانات، يتعقَّب المتجر مجموعة السلع (أو سلة التسوق) التي يشتريها كل عميل أثناء كل زيارة إلى المتجر. ويصف كل صفٍّ في مجموعة البيانات سلة واحدة من السلع التي اشتراها عميل مُعين في زيارة معينة إلى المتجر. وهكذا تكون السمات في مجموعة البيانات هي المنتجات التي يبيعها المتجر. وبأخذ هذه البيانات في الاعتبار، تبحث عملية التنقيب عن قواعد الارتباط عن السلع التي يتكرَّر وجودها معًا داخل سلة التسوق في كل مرة. وخلافًا للتجميع واكتشاف الشذوذ، اللذين يُركزان على تحديد أوجه التشابُه أو الاختلاف بين المثيلات (أو الصفوف) في مجموعة البيانات، فإن التنقيب عن قواعد الارتباط يركز على البحث في العلاقات بين السِّمات (أو الأعمدة) في مجموعة البيانات. وبوجهٍ عام، فإنها تبحث عن علاقات الارتباط بين المنتجات التي تُشترى في نفس الوقت. وباستخدام التنقيب عن قواعد الارتباط، يستطيع المتجر البدء في الإجابة عن أسئلةٍ بخصوص سلوكيات العملاء من خلال البحث عن أنماط ربما تُوجَد في البيانات. ومن بين الأسئلة التي يمكن الاستعانة بتحليل سلة التسوق للإجابة عنها ما يلي: «هل كانت حملة التسويق مُجدية؟ هل تغيرت أنماط الشراء لدى هذا العميل؟ هل وقع حدث مهم في حياة العميل؟ هل تتأثر سلوكيات الشراء بموقع المنتج في المتجر؟ من الذي يجب أن نستهدفه بمنتجنا الجديد؟»

خوارزمية أبريوري هي الخوارزمية الأساسية المستخدمة لإنتاج قواعد الارتباط. وتحتوي على عملية من خطوتين:

(١) إيجاد جميع توليفات العناصر التي توجَد معًا في مجموعة من التعاملات بحدٍّ أدنى مُحدد من التواتر والتكرار. ويُطلق على هذه التوليفات «مجموعة العناصر المتكررة».

(٢) إنشاء قواعد تعبر عن احتمالية وجود العناصر معًا داخل مجموعة العناصر المتكررة. تحسب خوارزمية أبريوري احتمالية وجود عنصرٍ في مجموعة العناصر المتكررة بمعلومية وجود عنصرٍ آخر أو عناصر أخرى.

تُنشئ خوارزمية أبريوري قواعد ارتباط تُعبر عن وجود علاقات محتملة بين العناصر الموجودة في مجموعات العناصر المتكررة. وتتَّخذ قاعدة الارتباط الصيغة التالية: IF antecedent, THEN consequent (بمعنى إذا توافر العنصر «السابق»، فهذا يعني توافر العنصر «التالي»). تنص هذه القاعدة على أن وجود العنصر أو مجموعة العناصر «السابقة»، يعني ضمنيًّا وجود عنصر أو عناصر أخرى في سلة التسوق نفسها «العناصر التالية». على سبيل المثال، ربما تنصُّ القاعدة المستمدة من مجموعة العناصر المتكررة التي تحتوي على العناصر «أ» و«ب» و«ج» على أنه إذا توافر العنصران «أ» و«ب» معًا في معاملة ما، فمن المرجح أن تتضمن المعاملة العنصر «ج» أيضًا:

IF {hot-dogs, ketchup}, THEN {beer}.

تشير هذه القاعدة إلى أن العملاء الذين يشترون النقانق والكاتشب من المرجَّح أن يقوموا بشراء البيرة أيضًا. وثمة مثال مُتكرر على قوة التنقيب عن قواعد الارتباط يتمثل في مثال الارتباط بين «البيرة والحفاضات» الذي يصف كيف استغل أحد المتاجر الأمريكية المغمورة في الثمانينيات من القرن العشرين نظامًا حاسوبيًّا قديمًا لتحليل بيانات فواتير المشتريات الخاصة بالمتجر ووجد علاقة ارتباطية غريبة بين الحفَّاضات والبيرة في مشتريات العملاء. ووُضعت نظرية لفهم هذه القاعدة ألا وهي أن الأُسَر التي لديها أطفال صغار تستعدُّ لقضاء عطلات نهاية الأسبوع وأنها تدرك أنها ستكون بحاجة إلى حفاضات للأطفال وستقضي العطلة معًا في المنزل. وضع المتجر العنصرَين (الحفاضات والبيرة) مُتجاورَين، وبالتبعية ارتفعت المبيعات. فُنِّدت قصة وجود ارتباط بين البيرة والحفاضات باعتبارها قصة ملفَّقة، غير أنها لا تزال تُقدم مثالًا مفيدًا على الفوائد المحتملة للتنقيب عن قواعد الارتباط بالنسبة إلى متاجر البيع بالتجزئة.

ثمة قياسان إحصائيَّان أساسيَّان مرتبطان بقواعد الارتباط؛ ألا وهما: «الدعم» و«الثقة». تشير نسبة «دعم» قاعدة الارتباط — أو معدل المعاملات التي تشمل كلًّا من العناصر السابقة والعناصر التالية نسبةً إلى العدد الإجمالي للمعاملات — إلى مدى تكرار وجود العناصر الواردة في قاعدة الارتباط معًا. أما نسبة «الثقة» في قاعدة الارتباط — أو معدل عدد المعاملات التي تشمل كلًّا من العناصر السابقة والعناصر التالية بالنسبة إلى عدد المعاملات التي تشمل العناصر السابقة — فهي الاحتمال الشرطي بأن العنصر التالي سوف يتوفر بشرط وجود العنصر السابق. إذن، على سبيل المثال، تعني نسبة الثقة التي تساوي ٧٥ بالمائة في قاعدة ارتباط تربط بين عنصري «النقانق» و«الكاتشب» وعنصر «البيرة» أنه في ٧٥ بالمائة من الحالات التي يشتري فيها العملاء كلًّا من «النقانق» و«الكاتشب»، سيشترون أيضًا «البيرة». أما نسبة دعم القاعدة فتشير ببساطة إلى نسبة السلال التي تنطبق عليها القاعدة في مجموعة البيانات. على سبيل المثال، تشير نسبة الدعم التي تساوي ٥ بالمائة إلى أن ٥ بالمائة من جميع السلال في مجموعة البيانات تحتوي على العناصر الثلاثة الموجودة في قاعدة «النقانق والكاتشب والبيرة».

حتى مجموعات البيانات الصغيرة قد تسفر عن إنشاء عددٍ كبير من قواعد الارتباط. ومن أجل التحكم في درجة تعقيد تحليل هذه القواعد، من المعتاد تنقيح مجموعة القواعد المتولدة لتشمل فقط القواعد التي تتميز بنسبة دعمٍ وثقة عاليتَين. والقواعد التي لا تتمتع بنسبتَين عاليتين من الدعم والثقة ليست مثيرة للاهتمام نظرًا إلى أن القاعدة لا تُغطي سوى نسبةٍ صغيرة جدًّا من السلال (نسبة دعم منخفضة)، أو لأن العلاقة بين العناصر السابقة والعناصر التالية ضعيفة (نسبة ثقة منخفضة). وينبغي أيضًا تنقيح القواعد عديمة الأهمية أو غير القابلة للتفسير. تُمثل القواعد عديمة الأهمية علاقاتٍ ارتباطية واضحة ومعروفة جيدًا لأي شخصٍ يفهم في هذا المجال من الأعمال. وتمثل القاعدة غير القابلة للتفسير علاقاتٍ ارتباطية غريبة جدًّا لدرجةٍ يصعب معها فهم كيف يمكن تحويل القاعدة إلى إجراءٍ مفيد بالنسبة إلى الشركة. ومن المرجَّح أن تكون القاعدة غير القابلة للتفسير ناتجة عن عينة بياناتٍ غريبة (أي أن القاعدة تمثل ارتباطًا زائفًا). وبمجرد أن تُنقَّح مجموعة القواعد، يستطيع عالِم البيانات تحليل القواعد المتبقية لفهم أي المنتجات يرتبط بعضها ببعض، وتطبيق هذه المعلومة الجديدة في الشركة. وعادة ستستخدم الشركات هذه المعلومة الجديدة لتحديد نسَق توزيع المنتجات في المتجر أو لتنفيذ بعض حملات التسويق الموجَّه إلى العملاء. وقد تشمل هذه الحملات إجراء تحديثات

لمواقعها الإلكترونية لتشمل المنتجات الموصى بها، والإعلانات داخل المَتجر، ورسائل البريد الإلكتروني المباشرة، والبيع المتقاطع لمنتجاتٍ أخرى من خلال فريق التحصيل (كاشير) وهلم جرًّا.

وتصير عملية التنقيب عن قواعد الارتباط أقوى عندما ترتبط سلال العناصر ببياناتٍ ديموغرافية بخصوص العملاء. ولهذا السبب ينفذ الكثير من تجار التجزئة برامج بطاقة الولاء نظرًا إلى أن هذه البرامج تسمح لهم ليس فقط بالربط بين العميل وبين سلال التسوق المختلفة له بمرور الوقت وإنما تسمح لهم أيضًا بربط سلة التسوق بالمعلومات الديموغرافية الخاصة بالعميل. ويمكِّن دمج هذه المعلومات الديموغرافية في تحليل الارتباط من أن يكون التحليل مُركِّزًا على معلومات ديموغرافية مُعينة، والتي قد تساعد أكثر في التسويق والإعلانات الموجَّهة. على سبيل المثال، يمكن استخدام قواعد الارتباط المزوَّدة بمعلوماتٍ ديموغرافية مع العملاء الجدد الذين لا يتوافر لدى الشركة معلوماتٍ عن عاداتهم الشرائية؛ ولكن لديها معلومات ديموغرافية عنهم. وفيما يلي مثال على قاعدة ارتباط مزوَّدة بمعلومات ديموغرافية:

IF gender(male) and age(<35) and {hot-dogs, ketchup},
THEN {beer}.
[*Support = 2%, Confidence = 90%.*]

وتعني أنه إذا كان النوع الاجتماعي للعميل ذكَرًا والسنُّ أقلَّ من ٣٥ واشترى نقانق وكاتشب، فسوف يشتري بيرة).
[الدعم = ٢٪، الثقة = ٩٠٪.]

يركز نطاق التطبيق المعتاد للتنقيب عن قواعد الارتباط على ماهية المنتجات الموجودة في سلَّة التسوق وماهية المنتجات غير الموجودة في هذه السلة. يفترِض هذا أن المنتجات تُشترى في زيارةٍ واحدة إلى المتجر أو الموقع الإلكتروني. ومن المحتمَل أن ينجح هذا النوع من السيناريوهات مع معظم سيناريوهات البيع بالتجزئة وغيرها من السيناريوهات ذات الصِّلة. ومع ذلك، يكون التنقيب عن قواعد الارتباط مفيدًا أيضًا في نطاقٍ من المجالات بخلاف البيع بالتجزئة. على سبيل المثال، في مجال الاتصالات عن بُعد، يساعد تطبيق التنقيب عن قواعد الارتباط على استخدام العملاء شركات الاتصالات عن بُعد في تصميم طرُق لتجميع الخدمات المختلفة معًا في باقات. وفي مجال التأمين، يُستخدم التنقيب عن قواعد الارتباط لمعرفة ما إذا كانت هناك علاقات ارتباطية بين المنتجات والمطالبات. وفي

المجال الطبي، يُستخدم التنقيب عن قواعد الارتباط للتحقُّق مما إذا كان هناك تفاعُلات بين العلاجات والأدوية الموجودة وتلك الجديدة. وفي مجال الخدمات المصرفية والمالية، يُستخدم لمعرفة أي منتجاتٍ يمتلِكها العملاء عادة وما إذا كان من الممكن تطبيق هذه المنتجات على العملاء الجدد أو العملاء الحاليين. ويمكن الاستعانة بالتنقيب عن قواعد الارتباط لتحليل سلوكيات الشراء على مدى فترةٍ زمنية. على سبيل المثال، يميل العملاء إلى شراء المنتج «س» و«ص» اليوم، وفي غضون ثلاثة أشهر يشترون المنتج «ع». ويمكن اعتبار هذه الفترة الزمنية سلةَ تسوُّق، على الرغم من أنها فترة تمتدُّ على مدار ثلاثة أشهر. ويؤدي تنفيذ التنقيب عن قواعد الارتباط على هذا النوع من السلال المحدَّدة زمنيًّا إلى توسيع نطاقات تطبيق التنقيب عن قواعد الارتباط لتشمل جداول الصيانة واستبدال قطع الغيار والمكالمات الخدمية والمنتجات المالية وما إلى ذلك.

تسرُّب العملاء أو الاحتفاظ بهم، تلك هي المسألة (التصنيف)

إحدى مهام العمل القياسية في إدارة العلاقات مع العملاء هي تقييم احتمالية أن يتَّخذ عميل فردي إجراءً ما. يُستخدم مصطلح «نمذجة المَيل» لوصف هذه المهمة، لأن الهدف منها هو وضع نموذج لمَيل الفرد نحو القيام بشيءٍ ما. وقد يكون هذا الإجراء أيَّ شيءٍ، بداية من الاستجابة إلى حملات التسويق وصولًا إلى التعثر في سداد قرضٍ أو التوقُّف عن استخدام خدمة. إن القدرة على تحديد العملاء الذين من المرجَّح أن يتوقفوا عن استخدام خدمةٍ مُعينة هو أمر ذو أهمية بالِغة بالنسبة إلى شركات خدمات الهاتف المحمول. تتكلف هذه الشركات مبالغ طائلة لاجتذاب العملاء الجدد. وفي الواقع، تشير التقديرات بوجهٍ عام إلى أن اجتذاب عميل جديد يكلف أكثر من الاحتفاظ بعميلٍ حالي بمقدارٍ يتراوح بين خمس وستِّ مرات (Verbeke et al. 2011). ونتيجة لذلك، تحرص الكثير من الشركات أشدَّ الحرص على الاحتفاظ بعملائها الحاليِّين. ومع ذلك، تريد هذه الشركات أن تُقلل التكاليف أيضًا إلى الحد الأدنى. وعلى الرغم من أنه قد يكون من السهل الاحتفاظ بالعملاء، من خلال تقديم أسعار مُخفَّضة وتحديثات رائعة لخدمات الهاتف إلى جميع العملاء، فإن هذا لا يُعَد خيارًا واقعيًّا. وبدلًا من ذلك، ترغب هذه الشركات أن تقتصر العروض التي توفرها لعملائها على أولئك الذين من المرجَّح أن يتركوا الشركة في المستقبل القريب. فإذا استطاعت تحديد العميل الذي بصدد التوقف عن استخدام الخدمة وإقناعه بمواصلة

استخدامها، ربما من خلال تقديم تحديثٍ أو حزمة جديدة للفواتير، يُمكنها توفير الفارق بين تكلفة إغراء العميل بالبقاء وتكلفة اجتذاب عميل جديد.

يُستخدم مصطلح «تسرب العملاء» (أو خسارة العملاء) لوصف عملية تخلي العملاء عن خدمةٍ ما وانضمامهم إلى شركة خدمية أخرى. وبالتالي، تُعرَف مسألة التنبؤ بالعميل الذي من المحتمَل أن يتوقف عن استخدام الخدمة في المستقبل القريب باسم «التنبؤ بتسرُّب العملاء». وكما يُوحي الاسم، فهذه مهمة تنبؤية. وتتمثل هذه المهمة في تصنيف ما إذا كان العميل عُرضةً للتسرُّب من الخدمة أم لا. تستخدِم الكثير من الشركات هذا النوع من التحليل للتنبؤ باحتمالية تسرُّب العملاء في شركات الاتصالات وخدمات المرافق والخدمات البنكية والتأمين وغيرها من المجالات. وأحد المجالات النامية التي تركز عليها الشركات هو التنبؤ بمعدل دوران العمالة أو تسرُّب العمالة: أي العمالة التي من المرجَّح أن تترك الشركة في غضون فترةٍ زمنية محددة.

وعندما يُنتج نموذج التنبؤ تسميةً فئوية أو فئة لمُدخَلٍ ما، يُعرف النموذج باسم «نموذج التصنيف». ويتطلب تدريب نموذج التصنيف بياناتٍ قديمة، حيث يُسمى كل مثيلٍ بتسميةٍ فئوية ليشير إلى ما إذا كان الحدث المستهدَف قد وقع لذلك المثيل أم لا. على سبيل المثال، يتطلب تصنيف عملية تسرُّب العملاء مجموعة بيانات تُمنَح فيها تسمية فئوية لكل عميلٍ (صف واحد لكل عميل) بحيث تشير إلى ما إذا كان هذا العميل قد تسرَّب أم لا. وستشمل مجموعة البيانات سمة، تُعرف باسم «السمة المستهدَفة»، التي تدرج هذه التسمية الفئوية لكل عميل. وفي بعض المثيلات، يكون وضْعُ تسميةٍ فئوية إلى جوار خانة العميل، للدلالة على تسرُّبه أو عدمه، مهمةً بسيطة نسبيًّا. على سبيل المثال، ربما يتواصل العميل مع الشركة ويُلغي بكل بساطةٍ اشتراكه أو تعاقده مع الشركة. ومع ذلك، في بعض الحالات، ربما لا تُميَّز واقعة التسرب ببساطة. على سبيل المثال، ليس لدى جميع عملاء شركات خدمات الهواتف المحمولة عقودٌ شهرية. فبعضهم يمتلك عقود الدفع المسبق (أو الشحن المسبق) حيث يقومون فيها بشحن أرصدة هواتفهم على فتراتٍ غير منتظمة عند الحاجة إلى المزيد من الرصيد على الهاتف. وقد يكون من الصعب تحديد ما إذا كان هذا النوع من العملاء قد تسرَّبوا أم لا: هل خسرت الشركة العميل الذي لم يُجْرِ مكالمة هاتفية منذ أسبوعين، أم من الضروري أن يكون رصيد العميل صفرًا ولم يقم بأي نشاطٍ لمدة ثلاثة أسابيع قبل اعتباره عميلًا متسربًا؟ بمجرد تحديد حدث التسرُّب من المنظور التجاري، فمن الضروري إذن تطبيق هذا على هيئة كودٍ من أجل تعيين تسميةٍ فئوية مستهدفة لكل عميلٍ في مجموعة البيانات.

ثمة عامل تعقيد آخر مرتبط بإنشاء مجموعة بيانات مُدربة لنماذج التنبؤ بتسرُّب العملاء يتمثل في ضرورة أخذ الفجوات الزمنية في الاعتبار. فالهدف من وراء التنبؤ بتسرُّب العملاء هو عمل نموذج للمَيل (أو الاحتمالية) أن العميل سيتسرَّب في وقتٍ ما في المستقبل. ونتيجة لذلك، فإن لهذا النوع من النماذج بُعدًا زمنيًّا يجب وضعه في الاعتبار أثناء إنشاء مجموعة البيانات. ومجموعة السمات الواردة في مجموعة البيانات الخاصة بنموذج المَيل مأخوذة من فترتَين زمنيتَين منفصلتَين: فترة «المراقبة» وفترة «النتائج». وفترة المراقبة هي الفترة التي تُحسَب فيها قِيَم سمات كل مدخل. أما فترة النتائج فهي الفترة التي تحسب فيها السمة المستهدفة. والهدف التجاري من ابتكار نموذج للتنبؤ بتسرُّب العملاء هو تميكن الشركة من التدخل بشكلٍ أو بآخر قبل تسرُّب العميل؛ أو بعبارة أخرى إغراء العميل بمواصلة الاستعانة بالخدمة. وهذا يعني أنه يجب إجراء التنبؤ بتسرب العملاء في وقتٍ سابق على توقف العميل عن استخدام الخدمة فعليًّا. ومدة هذه الفترة مساوية لمدة فترة النتائج، والنتيجة التي يخرج بها نموذج التنبؤ تفيد بأن العميل سيتوقَّف عن استخدام الخدمة في غضون فترة النتائج هذه. على سبيل المثال، يمكن تدريب النموذج على التنبؤ بأن العميل سيتسرَّب في غضون شهر أو شهرين، بناءً على سرعة الشركة في إجراء عملية تدخل لإقناع العميل بالبقاء.

ويؤثر تحديد فترة النتائج على البيانات التي ينبغي استخدامها كمُدخلاتٍ للنموذج. فإذا كان النموذج مصممًا للتنبؤ بأن العميل سيتوقف عن استخدام الخدمة في غضون شهرَين من اليوم الذي يُشغَّل فيه النموذج على سجلِّ ذلك العميل، فعندما يتم تدريب هذا النموذج، ينبغي حساب السِّمات المدخَلة التي تصف العملاء القدامى الذين توقفوا عن استخدام الخدمة بالفعل باستخدام البيانات المتاحة عن هؤلاء العملاء قبل شهرين من توقفهم عن استخدام الخدمة. وبالمِثل ينبغي حساب السمات المدخَلة التي تصف العملاء النشطين حاليًّا باستخدام البيانات المتاحة عن نشاط هؤلاء العملاء منذ شهرَين. ويضمن إنشاء مجموعة البيانات بهذه الطريقة أن جميع المثيلات في مجموعة البيانات هذه — والتي تتضمَّن العملاء المتسرِّبين والعملاء النشِطين — تصف العملاء في وقت تصميم النموذج أثناء رحلتهم الفردية كعملاء للتنبؤ بما إذا كانوا سيتسرَّبون أم لا قبل شهرَين من اتخاذهم القرار.

تستخدِم جميع نماذج مَيل العملاء تقريبًا سماتٍ تصف المعلومات الديموغرافية الخاصة بالعميل كمُدخلات: السن، والنوع الاجتماعي، والوظيفة، وما إلى ذلك. وفي

السيناريوهات المتعلقة بخدمة مستمرة، من المرجَّح أيضًا أن تشتمل على سماتٍ تصف المرحلة التي يُوجَد فيها العميل في مراحل تطوُّر العملاء: «عميل مستجد»، «عميل لا يزال في منتصف فترة التعاقد»، «عميل يقترب من نهاية التعاقد». ومن المحتمَل أيضًا أن يكون هناك سماتٌ متعلقة بمجالٍ بعينه. على سبيل المثال، من السمات المعتادة في نماذج تسرب العملاء من شركات الاتصالات متوسط فاتورة العميل، والتغيرات الطارئة على مبالغ الفواتير، ومتوسط الاستخدام، والالتزام بدقائق الاستخدام التي توفرها له الخطة التي اشترك فيها أو تَخطِّيها بصفةٍ عامة، ونسبة المكالمات الموجهة لمستخدمي الشبكة إلى الموجهة لمن هم خارجها وربما نوعية الهاتف المستخدَم.[1] ورغم ذلك، تتنوع السمات المحددة المستخدمة في كل نموذجٍ من مشروع إلى آخر. وسجل جوردون لينوف ومايكل بيري (٢٠١١) أنه في أحد مشاريع التنبؤ بتسرُّب العملاء في كوريا الجنوبية، وجد الباحثون أنه من المفيد تضمين سمةٍ تصف معدل تسرُّب العملاء المرتبط بهاتف العميل (أي نسبة تسرُّب العملاء الذين يستخدمون هذا الهاتف بالذات خلال فترة المراقبة). ومع ذلك، عندما ذهبوا إلى تصميم نموذجٍ مُشابه للتنبؤ بتسرُّب العملاء في كندا، كانت سمة الهاتف المستخدم/معدل التسرب عديمة الفائدة. كان الفارق أنه في كوريا الجنوبية قدمت شركة خدمات الهاتف المحمول خصوماتٍ كبيرة على الهواتف الجديدة للعملاء الجدد، في حين أنه في كندا قُدمت نفس نسبة الخصومات إلى العملاء الحاليين والجدد على حدٍّ سواء. وكان التأثير الإجمالي أن الهواتف القديمة في كوريا الجنوبية شجَّعت تسرُّب العملاء؛ وشُجِّع الناس على ترك شركة والانضمام إلى أُخرى من أجل الاستفادة بالخصومات، ولكن في كندا لم يكن هذا الحافز موجودًا من الأساس.

بمجرد إنشاء مجموعة بيانات ذات تسمية فئوية، تكون المرحلة الكبرى في إنشاء نموذج التصنيف هي استخدام خوارزمية تعلُّم آلة لإنشاء النموذج. وأثناء النمذجة، من المفيد تجربة عددٍ من خوارزميات تعلُّم الآلة المختلفة لتحديد الخوارزمية التي تعمل بشكلٍ أفضل على مجموعة البيانات. وبمجرد اختيار النموذج النهائي، تُقدَّر الدقة المحتملة لتوقُّعات هذا النموذج على المثيلات الجديدة من خلال اختباره على مجموعةٍ فرعية من البيانات لم تُستخدَم أثناء مرحلة تدريب النموذج. وإذا اعتبر النموذج دقيقًا بالدرجة الكافية ومناسبًا لاحتياج الشركة، يُنشَر النموذج ويُطبَّق على البيانات الجديدة إما في عمليةٍ مجمعة أو في الوقت الفعلي. ومن أهم مراحل نشر النموذج التأكد من تشغيله بطريقةٍ ملائمة واستخدام الموارد المناسبة بحيث يُستغَل النموذج بفاعلية. لا فائدة تُرجى

من إنشاء نموذج للتنبؤ بتسرُّب العملاء ما لم يَنتُج عن تنبؤات النموذج اتخاذ إجراءات لاستمالة العملاء لتستطيع الشركة الاحتفاظ بهم.

بالإضافة إلى التنبؤ بالتسمية التصنيفية، تستطيع نماذج التنبؤ أن تُعطينا مقياسًا عن مدى تأكد النموذج من التنبؤ الذي وصل إليه. يُسمى هذا المقياس «احتمالية صحة التنبؤ» وله قيمة تتراوح ما بين صفر وواحد. وكلما كانت القيمة أعلى، زاد احتمال أن يكون التنبؤ صحيحًا. ويمكن استخدام قيمة مقياس «احتمالية صحة التنبؤ» لإعطاء الأولوية للعملاء الذين يجِب التركيز عليهم. على سبيل المثال، في التنبؤ بتسرُّب العملاء تريد الشركة التركيز على العملاء الأكثر عرضةً للتوقُّف عن استخدام الخدمة. ومن خلال الاستعانة بقيمة احتمالية صحَّة التنبؤ وترتيب العملاء بناءً على هذه القيمة، يمكن للشركة أن تركز على العملاء الرئيسيين (الأكثر عُرضة للتوقُّف عن استخدام الخدمة) قبل الانتقال إلى العملاء ذوي القيمة الأقل فيما يخصُّ احتمالية صحة التنبؤ.

كم ستكون تكلفة هذا؟ (الانحدار)

التنبؤ بالأسعار هي مهمة تقدير سعر المنتج في نقطة زمنية مُعينة. قد يكون هذا المنتج سيارةً أو منزلًا أو برميل نفط أو سهمًا أو إجراءً طبيًّا. ومن الواضح أن الوصول إلى تقدير حقيقيٍّ لسعر شيءٍ ما هو أمر مهم بالنسبة إلى شخصٍ يفكر في شراء هذا الشيء. وتعتمد دقة نموذج التنبؤ بالأسعار على المجال. على سبيل المثال، نظرًا إلى تقلُّبات سوق الأوراق المالية، فمن الصعب جدًّا التنبؤ بسعر سهمٍ ما غدًا. وعلى العكس من ذلك، ربما يكون من الأسهل التنبؤ بسعر منزلٍ في مزادٍ نظرًا إلى أن تغيُّر أسعار المنازل يتم بوتيرةٍ أبطأ بكثيرٍ من الأسهم.

حقيقة أن التنبؤ بالأسعار يشمل تقدير قيمة سمةٍ مستمرة تعني أنه يُتعامَل معه بوصفه «مسألة انحدار». ومسألة الانحدار تُشبه من الناحية الهيكلية مسألة التصنيف، ففي كلتا الحالتَين، يشمل الحل الذي يُقدمه علم البيانات تصميم نموذج يُمكنه التنبؤ بالقيمة المفقودة لسمةٍ معينة بمعلومية مجموعة من السمات المدخلة. الفارق الوحيد أن التصنيف ينطوي على تقدير قيمة سمةٍ فئوية، أما الانحدار فينطوي على تقدير قيمة سمةٍ مستمرة. يتطلب تحليل الانحدار مجموعة بيانات مُدرَج فيها قيمة السمة المستهدفة في كل مثيلٍ قديم. ويوضح نموذج الانحدار الخطِّي، المتعدد المدخَلات الذي قدَّمناه في الفصل

الرابع، البنية الأساسية لنموذج الانحدار، حيث إن معظم نماذج الانحدار الأخرى عبارة عن تنويعات لهذا المنهج. ولا تتغير البنية الأساسية لنموذج الانحدار الخاص بالتنبؤ بالأسعار بغضِّ النظر عن المنتج الذي يُطبق عليه النموذج؛ وكل ما يتغير هو أسماء السمات وعددها. على سبيل المثال، من أجل التنبُّؤ بسعر منزل، ستشمل المدخَلات سماتٍ مثل مساحة المنزل وعدد الغرف وعدد الطوابق ومتوسط سعر المنازل في المنطقة ومتوسط مساحة المنازل في المنطقة وما إلى ذلك. على النقيض من ذلك، من أجل التنبؤ بسعر سيارة، ستشمل السمات المدخَلة عمر السيارة وعدد الأميال التي قطعتها والمسجَّلة على عدَّاد المسافات، وحجم المحرك وماركة السيارة وعدد الأبواب وما إلى ذلك. وفي كل حالة، وبمعلومية البيانات المناسبة، تعمل خوارزمية الانحدار على تحديد إلى أي مدًى تُساهم كل سمةٍ من هذه السمات في السعر النهائي.

وكما هي الحال مع جميع الأمثلة التي ضربناها على مدار هذا الفصل، المثال التطبيقي على استخدام نموذج الانحدار للتنبؤ بالأسعار هو مثال توضيحي فقط لنوعية المشكلات التي يكون من المناسب صياغتها على شكل مهمةٍ لنمذجة الانحدار. ويمكن الاستعانة بتنبُّؤ الانحدار في مجموعةٍ واسعة النطاق من المسائل الأخرى في العالم الواقعي. وتشمل المسائل النمطية للتنبؤ باستخدام الانحدار حساب الأرباح، وقيمة المبيعات وحجمها، وحساب الحجم، والطلب، والمسافة، والجرعة.

الفصل السادس

الخصوصية والأخلاقيات

اليوم يُعَد أكبر مجهول يواجِه علم البيانات هو كيف ستختار المجتمعات الإجابة عن نسخة جديدة من السؤال القديم المتعلق بكيفية تحقيق التوازن الأمثل بين حريَّات وخصوصية الأفراد والأقليات من ناحية والحفاظ على أمن المجتمع ومصالِحه من ناحية أخرى. وفي سياق مُتصل بعلم البيانات، تعاد صياغة هذا السؤال القديم على النحو التالي: ما الذي نعتبره، كمجتمع، طرقًا معقولة لجمع واستخدام البيانات الخاصة بالأفراد في سياقاتٍ متنوعة مثل مكافحة الإرهاب، وتحسين العلاج الدوائي، ودعم أبحاث السياسات العامة، ومكافحة الجريمة، واكتشاف الاحتيال، وتقييم مخاطر الائتمان، وتقييم المخاطر التأمينية، وتوجيه الإعلانات للمجموعات المستهدفة؟

ويَعِد علم البيانات بأن يقدم طريقةً لفهم العالم من خلال البيانات. وفي عصر البيانات الضخمة الحالي، هذا الوعد مُغرٍ للغاية، وبالطبع، يمكن الاستعانة بعددٍ من الحجج لدعم تطوير التقنيات والبنية التحتية القائمة على البيانات واستخدامها. وترتبط إحدى الحجج الشائعة بتحسين الكفاءة والفعالية والتنافُسية؛ وهي حجة يدعمها بعض الأبحاث الأكاديمية في سياق العمل التجاري على الأقل. على سبيل المثال، أظهرت دراسة شملت ١٧٩ شركة كبيرة مطروحةً للتداول العام في عام ٢٠١١ أنه كلما كان اتخاذ قرارات الشركة معتمدًا على البيانات، أصبحت الشركة أكثر إنتاجية؛ إذ جاء فيها: «وجدنا أن الشركات التي تبنَّت اتخاذ القرارات بناءً على البيانات تزيد مخرجاتها وإنتاجيتها بنسبة ٥–٦ بالمائة عما هو متوقَّع في ضوء استثماراتها الأخرى واستخدام تكنولوجيا المعلومات» (Brynjolfsson, Hitt, and Kim 2011, 1).

وثمة حجة أخرى للاعتماد المتزايد على تقنيات وممارسات علم البيانات وتتعلق بإضفاء الطابع الأمني على كل شيء. لفترة طويلة، استعانت الحكومات بحجة أن المراقبة

تُحسن مستوى الأمن. ومنذ الهجمات الإرهابية التي وقعت في الحادي عشر من سبتمبر ٢٠٠١ بالولايات المتحدة، ومع كل هجمةٍ إرهابية تالية في مختلف أنحاء العالم، لاقت الحجة رواجًا شعبيًّا أكثر. بالطبع، كثيرًا ما استخدمت في النقاش العام الذي أُثير بسبب اعترافات إدوارد سنودن عن برنامج المراقبة «بريسم» التابع لوكالة الأمن القومي الأمريكية والبيانات التي تُجمع بصفةٍ دورية عن المواطنين الأمريكيِّين. ومثال صارخ على قوة هذه الحجة هو إنفاق الوكالة ١,٧ مليار دولار على مركز بيانات في مدينة بلوفديل بولاية يوتا يتمتع بقدرةٍ على تخزين كمياتٍ هائلة من المكالمات الخاضعة للمراقبة (Carroll 2013).

غير أنه في الوقت نفسه، تجاهد المجتمعات والحكومات والشركات من أجل فهم الآثار الطويلة الأمد لعلم البيانات في عالم البيانات الضخمة. وفي ضوء التطور السريع لتقنيات جمع البيانات وتخزينها وتحليلها، ليس من المستغرَب أن تحدث تغيرات سريعة أيضًا في الأطر القانونية المعمول بها وفي النقاشات الأخلاقية الأوسع نطاقًا حول البيانات، لا سيما مسألة خصوصية الأفراد. وعلى الرغم من ذلك، فإن من المهم فهم المبادئ القانونية الأساسية التي تحكُم جمع البيانات واستخدامها. بالإضافة إلى ذلك، سلَّط النقاش الأخلاقي حول استخدام البيانات والخصوصية الضوءَ على التوجهات المهمة التي ينبغي لنا أن نعيها كأفرادٍ ومواطنين.

المصالح التجارية في مقابل خصوصية الأفراد

يمكن النظر إلى علم البيانات في إطار أنه يجعل العالم مكانًا أكثر ثراءً وأمنًا للعيش فيه. إلا أنه يمكن استخدام هذه الحجج نفسها من جانب المنظمات المختلفة التي لدَيها أجندات مختلفة. على سبيل المثال، قارن دعوات الجماعات المؤيدة للحريات المدنية المنادية بأن تكون الحكومة أكثر انفتاحًا وشفافية في جمع البيانات واستخدامها وإتاحتها على أمَل تمكين المواطنين من محاسبة هذه الحكومات، بالدعوات المماثلة من جانب أوساط الشركات التجارية التي تأمُل في استخدام هذه البيانات لزيادة أرباحها (Kitchin 2014a). في الواقع، يُعَد علم البيانات سلاحًا ذا حدَّين. يمكن استخدامه لتحسين جودة حياتنا من خلال جعل الحكومة أكثر كفاءة والأدوية أكثر فعاليةً والرعاية الطبية أكثر جودة، والتأمين أقل تكلفة والمدن أكثر ذكاء من خلال تقليل معدلات الجريمة وما إلى ذلك. وفي الوقت نفسه، يمكن استخدامه أيضًا للتجسُّس على حياتنا الشخصية ولاستهدافنا بإعلاناتٍ غير مرغوبة والتحكم في سلوكياتنا سرًّا وعلانية على حدٍّ سواء (إذ يمكن للخوف من المراقبة أن يؤثر علينا بقدْر ما تؤثر علينا المراقبة نفسها).

كثيرًا ما تظهر الجوانب المتناقضة الخاصة بعلم البيانات في التطبيقات نفسها. على سبيل المثال، ترتكز الاستعانة بعلم البيانات في تعهُّد تغطية التأمين الصحي على مجموعات البيانات التسويقية الخاصة بأطرافٍ خارجية والتي تحتوي على معلوماتٍ مثل العادات الشرائية وسجلِّ البحث عبر الإنترنت، بالإضافة إلى مئات السمات الأخرى المتعلقة بنمط حياة الأفراد (Batty, Tripathi, Kroll, et al. 2010). واستغلال بيانات الأطراف الخارجية هو أمر مُزعج لأنه ربما يقيد حرية الناس؛ حيث يتجنَّب الناس أنشطةً معينة مثل زيارة مواقع الرياضات الخطيرة، خوفًا من تكبُّد أقساط تأمينية أعلى من جراء تعقب نشاط تصفُّح الفرد للمواقع الإلكترونية (Mayer-Schönberger and Cukier 2014). ومع ذلك، فإنَّ مُبرر استخدام هذه البيانات هو أنها تقوم بديلًا لمصادر المعلومات الأكثر شمولًا وتكلفة مثل فحوصات الدم، وعلى المدى الطويل ستُقلل التكاليف والأقساط وبالتالي تزيد من عدد الأشخاص المشتركين في خدمات التأمين الصحي (Batty, Tripathi, Kroll, et al. 2010).

يُعَد استخدام البيانات الشخصية في التسويق الموجَّه مثالًا واضحًا على الصدام ما بين المصالح التجارية والاعتبارات الأخلاقية في علم البيانات. فمن منظور الإعلانات التجارية، الحافز وراء استخدام البيانات الشخصية هو وجود علاقة بين التسويق والخدمات والمنتجات المصمَّمة خصوصًا من جانب، وفعالية التسويق من جانب آخر. ولقد تبيَّن أن استغلال البيانات الشخصية المتاحة على شبكات التواصل الاجتماعي — مثل تحديد العملاء الذين على صلةٍ بعملاء سابقين — يزيد كفاءة حملة التسويق عبر رسائل البريد الإلكتروني المباشرة لخدمات التواصل عن بُعد من ثلاث إلى خمس مراتٍ مقارنة بالمناهج التسويقية التقليدية (Hill, Provost, and Volinsky 2006). وثمة ادِّعاءات مُماثلة حول فعالية تخصيص التسويق عبر الإنترنت بالاستعانة بالبيانات الشخصية للعميل. على سبيل المثال، قارنَت دراسة حول تكلفة وفعالية الإعلانات الموجهة في الولايات المتحدة في عام ٢٠١٠ بين التسويق العشوائي عبر الإنترنت (عند إطلاق حملة إعلانات عبر مجموعة من المواقع الإلكترونية دون استهداف عملاء مُعينين أو مواقع مُعينة) و«الاستهداف السلوكي»[1] (Beales 2010). وجدت الدراسة أن التسويق السلوكي كان أكثر تكلفة (بـ ٢,٦٨ مرة) ولكنه أكفأ، مقرونًا بزيادة معدل التحويل (يقيس معدل التحويل مدى تحوُّل التصفُّح السلبي إلى إجراءات يتخذها العملاء على المواقع الإلكترونية) بأكثر من الضعف مقارنةً بالتسويق العشوائي عبر الإنترنت. أجرى باحثون من جامعة تورونتو ومعهد

ماساتشوستس للتكنولوجيا دراسةً شهيرة أخرى عن فعالية الإعلانات عبر الإنترنت المَبنية على البيانات (Goldfarb and Tucker 2011). واستعانوا بسنٍّ مشروع قانون حماية الخصوصية في الاتحاد الأوروبي[2] الذي حدَّ من قدرة شركات الإعلانات على تعقُّب سلوك المستخدِمين عبر الإنترنت في المقارنة بين فعالية الإعلانات عبر الإنترنت في ظل القيود الجديدة (داخل الاتحاد الأوروبي) وفعالية الإعلانات عبر الإنترنت في ظلِّ غياب القيود الجديدة (في الولايات المتحدة وغيرها من الدول غير الأعضاء بالاتحاد الأوروبي). ووجدت الدراسة أن الإعلانات عبر الإنترنت كانت أقلَّ كفاءة بدرجةٍ كبيرة في ظل القيود الجديدة، مع تسجيل انخفاضٍ بنسبة ٦٥ بالمائة في نية شراء المشاركين في الدراسة. كانت نتائج هذه الدراسة محلَّ جدال (انظر، على سبيل المثال، Mayer and Mitchell 2012)، إلا أن الدراسة استُخدِمت لدعم الحجة القائلة بأنه كلما زادت كمية البيانات المتاحة عن الشخص، زادت فعالية الإعلانات الموجهة إلى ذلك الشخص. ويرى مؤيدو التسويق الموجَّه المبني على البيانات أن هذا النوع من التسويق مُربح لشركات الدعاية والإعلان وللعملاء على حدٍّ سواء، زعمًا بأن شركات الدعاية والإعلان تُقلل تكاليف التسويق من خلال الحدِّ من الإعلانات المهدرة وتُحقق معدلاتِ تحويلٍ أعلى، وأن العملاء يتلقَّون المزيد من الإعلانات ذاتِ الصِّلة باهتماماتهم.

في أحسن الأحوال، يستند هذا المنظور المثالي لاستخدام البيانات الشخصية من أجل التسويق الموجَّه على فَهم انتقائي للمشكلة. وربما واحدةٌ من أكثر القصص المقلقة المتعلقة بالإعلانات الموجهة جاءت في صحيفة «نيويورك تايمز» في عام ٢٠١٢ وتضمَّنت متجر تارجت للبيع بالتجزئة بأسعارٍ مخفَّضة (Duhigg 2012). من المعروف جيدًا في مجال التسويق أن الحمْل والإنجاب يُمثلان إحدى الفترات في حياة المرء التي تتغير فيها عادات التسوق لدَيه تغييرًا جذريًّا. ونظرًا إلى هذا التغيير الجذري، يرى المسوِّقون أن الحمْل فرصة لتغيير عادات التسوق لدى شخصٍ وتغيير ولائه للعلامات التجارية، والكثير من متاجر البيع بالتجزئة تستعين بسجلات المواليد المتاحة للجماهير لتشجيع التسويق المصمَّم خصوصًا للآباء الجدد، من خلال إرسال عروض متعلقة بمنتجات الأطفال الرُّضَّع. ومن أجل الحصول على ميزة تنافسية، أراد متجر تارجت تحديد العميلات اللاتي في مرحلةٍ مبكرة من الحمل (مثاليٌّ أثناء الثلث الثاني من الحمل) حتى لو لم تقُل العميلة إنها حامل.[3] وهذه الفكرة مكَّنت متجر تارجت من أن يبدأ في التسويق المخصص للعميلات قبل أن تعرف متاجر البيع بالتجزئة الأخرى أن هذه العميلة تنتظِر طفلًا عما قريب.

ومن أجل تحقيق هذا الهدف، أطلق متجر تارجت مشروعًا قائمًا على علم البيانات بهدف التنبؤ بما إذا كانت العميلة حاملًا بناءً على تحليل عادات التسوق لدَيها. وكانت نقطة الانطلاق للمشروع هو تحليل عادات التسوق لدى النساء اللاتي سجلنَ في دفتر «هدايا المولود الجديد» على موقع متجر تارجت. وكشفَ التحليل أن الأمَّهات الحوامل يمِلنَ إلى شراء كمياتٍ أكبر من كريم الجلد عديم الرائحة في بداية الثلث الثاني من الحمل وكذلك مُكملات غذائية مُعينة خلال أول ٢٠ أسبوعًا من الحمل. وبناءً على هذا التحليل، أنشأ متجر تارجت نموذجًا مَبنيًّا على البيانات استعان بـ ٢٥ منتجًا ومؤشرًا وأعطى درجةً لكلِّ عميلة تتعلق «بالتنبؤ بالحمل». وكان «نجاح» هذا النموذج جليًّا جدًّا عندما ذهب رجل إلى متجر تارجت ليشتكي من حقيقة أن ابنته المراهقة الطالبة في مرحلة التعليم الثانوي استقبلت كوبونات ملابس أطفال وأسِرَّة لحديثي الولادة على بريدها الإلكتروني. واتَّهم الرجل متجر تارجت بمحاولة تشجيع ابنته على الحمل. ومع ذلك، وعلى مدار الأيام التالية، تبيَّن أن ابنة الرجل في الواقع كانت حامِلًا؛ ولكنها لم تُخبر أحدًا. هكذا، كان نموذج التنبؤ بالحمل الخاص بمتجر تارجت قادرًا على تحديد طالبةٍ في المرحلة الثانوية حامل والتصرُّف بناء على هذه المعلومة قبل أن تُخبر هي حتى أُسرتها.

التداعيات الأخلاقية لعلم البيانات: الملف التعريفي والتمييز

تُلقي قصة اكتشاف متجر تارجت حمل طالبةٍ في المرحلة الثانوية، دون رضاها عن الإعلان عن حمِلها، الضوءَ على كيف يمكن استخدام علم البيانات في إنشاء ملفٍّ تعريفٍ اجتماعي، ليس فقط للأفراد وإنما أيضًا للأقليَّات في المجتمع. وفي كتاب بعنوان «شخصيتك اليومية: كيف تُحدد صناعة الإعلانات الجديدة هويتك وجدارتك» (٢٠١٣)، يناقش جوزيف تورو كيف يستعين المسوقون بملفٍّ التعريف الرقمي لتصنيف الأشخاص إما كـ «أهداف» أو «عناصر غير مُجدية» ثم يستعينون بهذه الفئات لإضفاء الطابع الشخصي على العروض والإعلانات الدعائية الموجهة إلى العملاء الأفراد: «أولئك المصنَّفون كعناصر غير مُجدية تُتجاهَل أو تُوجَّه إلى منتجاتٍ أخرى يرى المسوقون أنها ذات صِلة أكثر بأذواقهم أو دخلهم» (١١). قد يسفر إضفاء الطابع الشخصي هذا عن معاملةٍ تفضيلية لبعض العملاء وعن تهميش آخرين. ومن الأمثلة الواضحة على هذا التمييز التسعير التفاضُلي على المواقع الإلكترونية، حيث تُفرَض تكلفةٌ على بعض العملاء أكثر منها على عملاء آخرين لنفس المنتج بناءً على ملفات التعريف الشخصية الخاصة بالعملاء (Clifford 2012).

قد يسفر إضفاء الطابع الشخصي هذا عن معاملة تفضيلية لبعض العملاء وعن تهميش آخرين.

يُنشأ هذه الملفات التعريفية بواسطة البيانات المجمَّعة من عددٍ من مصادر البيانات المشوَّشة والمتحيزة المختلفة، وبالتالي قد تكون هذه الملفات التعريفية عادةً مُضللة عن فردٍ ما. والأسوأ من ذلك هو أن هذه الملفات التعريفية التسويقية تُعامَل معاملة المنتجات وكثيرًا ما تُباع إلى الشركات الأخرى، مما يسفر عن أنَّ التقييم التسويقي السلبي للفرد يلاحِق ذلك الفرد عبر العديد من المجالات. لقد ناقشنا بالفعل استغلال مجموعة البيانات التسويقية في تعهُّد تغطية التأمين الصحي (Batty, Tripathi, Kroll, et al. 2010)، ولكن هذه الملفات التعريفية قد تشقُّ طريقها أيضًا نحو تقييمِ مَخاطر الائتمان والكثير من عمليات اتخاذ القرار الأخرى التي تؤثر على حياة الأشخاص. وثمة جانبان لهذه الملفَّات التعريفية التسويقية يجعلانها إشكاليةً على نحوٍ خاص. أولًا: إنها صندوق أسود في حدِّ ذاتها، وثانيًا إنها لا تتغيَّر. تتَّضح طبيعة الملفات التعريفية الغامضة كالصندوق الأسود عندما يفكر المرء كم أنه من الصعب على الفرد أن يعرف ماهية البيانات المسجلَّة عنه وأين ومتى سُجِّلت، وآلية عمل عمليات اتخاذ القرار التي تستخدِمها. ونتيجة لذلك، إذا انتهى المطاف بالفرد ليُصبح على قائمة الممنوعين من السفر أو على القائمة السوداء للائتمان، فـ «من الصعب تحديد أسباب التمييز والتصدِّي لها» (Kitchin 2014a, 177). علاوة على ذلك، في العالم الحديث عادةً ما تُخزَّن البيانات لوقت طويل. ومن ثم، فإن البيانات المسجلة عن حدثٍ ما في حياة الفرد تدوم فترة طويلة بعد الحدث. وكما حذَّر تورو: «إن تحويل ملفات التعريف الفردية إلى تقييماتٍ فردية هو ما يحدث عندما يصير الملف التعريفي أشبه بسمعةٍ للفرد» (٢٠١٣، ٦).

علاوة على ذلك، قد يؤدي علم البيانات في الواقع إلى استدامة التحيُّز وتعزيزه، ما لم يُستخدم بحرصٍ شديد. وأحيانًا يُقال إن علم البيانات موضوعي؛ أي أنه يستند إلى الأرقام، وبالتالي فإنه لا يُشفَّر أو يملك آراء متحيزة تؤثر على قرارات البشر. والحقيقة أن خوارزميات علم البيانات تعمل بطريقةٍ لا أخلاقية أكثر من كونها طريقة موضوعية. يستخرج علم البيانات أنماطًا في البيانات؛ ومع ذلك، إذا قامت البيانات بتشفير علاقة مُتحيزة في المجتمع، فمن المرجَّح أن تُحدِّد الخوارزمية هذا النمط وتستنِد مخرجاتها إلى النمط نفسه. وبالتأكيد، كلما كان التحيُّز أكثر ثباتًا في مجتمعٍ ما، ظهر ذلك النمط

التحيُّزي في البيانات عن ذلك المجتمع، وزادت احتمالية أن تستخرج خوارزمية علم البيانات ذلك النمط الخاص بالتحيُّز وتُكرره. على سبيل المثال، وجدَت دراسة أجراها باحثون أكاديميون على نظام الإعلانات عبر الإنترنت من جوجل أن النظام يُظهر الإعلانات المتعلقة بالوظائف ذات الأجور المرتفعة على نحوٍ أكثر تواترًا للمشاركين الذين يُبين الملفُّ التعريفي الخاص بهم على جوجل أنهم ذكور مقارنةً بأولئك الذين يُظهر ملفُّهم التعريفي أنهم إناث (Datta, Tschantz, and Datta 2015).

وحقيقة أن خوارزميات علم البيانات يمكن أن تُعزز التحيز هو أمر مُزعج على نحو خاص عند تطبيق علم البيانات في مجال حفظ الأمن والنظام العام. تُعَد برامج «التنبؤ بالجرائم»، أو ما يُعرف بـ PredPol،[4] أداة خاصة بعلم البيانات مُصمَّمة للتنبؤ بالموعد والمكان المرجَّح لحدوث جريمة. وعند نشر تطبيق هذه البرامج على مدينةٍ ما، تُولِّد تقريرًا يوميًّا يسرد قائمةً بعددٍ من البؤر الإجرامية على خريطة (مناطق صغيرة ٥٠٠ قدم في ٥٠٠ قدم) حيث يعتقد النظام أن من المرجح حدوث جرائم في هذه البؤر ويوزع دوريات شرطة عليها في الأوقات التي يعتقد النظام أن الجريمة ستقع فيها. استعانت أقسام الشرطة في كلٍّ من الولايات المتحدة والمملكة المتحدة بهذه الأداة. الفكرة وراء هذا النوع من نظام المراقبة الذكية هو أنه يمكن استغلال موارد الشرطة بفعالية. من الناحية الظاهرية، يبدو هذا تطبيقًا مهمًّا لعلم البيانات، فهو يسفر عن استهدافٍ فعَّال للجريمة وتقليل تكاليف حفظ الأمن والنظام. ومع ذلك، أُثيرت شكوك حول دقَّة برنامج «التنبؤ بالجرائم» وفعالية مبادرات حفظ النظام بواسطة برامج التنبؤ المماثلة (هانت، وسوندرز، وهوليوود ٢٠١٤؛ فريق عمل أوكلاند برايفيسي ٢٠١٥؛ هاركنيس ٢٠١٦). ولقد لوحظ أيضًا إمكانية استخدام هذه الأنواع من الأنظمة لتشفير ملفَّات تعريفٍ تنطوي على تمييزٍ عرقي أو طبقي في أعمال المراقبة وحفظ النظام (Baldridge 2015). قد يؤدي نشر عناصر الشرطة بناءً على البيانات القديمة إلى حضور أعلى للشرطة في مناطق مُعينة — عادة المناطق الأقل حظًّا من الناحية الاقتصادية — مما يؤدي بدوره إلى ارتفاع مستويات الجريمة المسجَّلة في هذه المناطق. بعبارة أخرى، يُصبح التنبؤ بالجرائم نبوءةً ذاتية التحقق. ونتيجة هذه الحلقة المفرغة أن بعض الأماكن ستكون مستهدفةً على نحوٍ غير متكافئ لمراقبة الشرطة مما يتسبَّب في انهيار الثقة بين الأشخاص الذين يعيشون في تلك المجتمعات المحلية وبين المؤسسات الأمنية (Harkness 2016).

في الواقع، قد يؤدي علم البيانات إلى استدامة التحيُّز وتعزيزه، ما لم يُستخدم بحرصٍ شديد.

مثال آخر على حفظ الأمن القائم على البيانات هو برنامج «قائمة الخاضعين للمراقبة الاستراتيجية» الذي تستعين به إدارة شرطة شيكاغو في محاولة منها للحدِّ من جرائم السلاح. أنشئت القائمة لأول مرة في عام ٢٠١٣، وفي تلك الفترة ضمت القائمة ٤٢٦ شخصًا قُدِّر أنهم مُعرضون بشدة لارتكاب أو الوقوع ضحية للعنف المسلح. وفي محاولة للتصدي للجرائم المسلحة على نحوٍ استباقي، تواصلت إدارة شرطة شيكاغو مع جميع الأشخاص الوارد ذكرهم في القائمة لتحذيرهم من أنهم خاضعون للمراقبة. فوجئ بعض الأشخاص بشدَّةٍ من وجودهم في القائمة لأنه على الرغم من أن لديهم صحيفة جنائية لارتكابهم جُنحًا؛ فإنه لم تُسجَّل في صحيفتهم الجنائية أية جرائم عُنف مسلَّح (Gorner 2013). أحد الأسئلة التي يجب طرحُها حول هذا النوع من جمع البيانات للتصدي للجرائم هو: إلى أي مدى تكون التكنولوجيا دقيقة؟ وجدت دراسة حديثة أن الأشخاص المذكورين في قائمة الخاضعين للمراقبة الاستراتيجية لعام ٢٠١٣ «لم يكن من المرجَّح وقوعهم ضحية لجريمة القتل أو تعرضهم لإطلاق نار على نحوٍ أكثر أو أقل من مجموعة الضبط» (Saunders, Hunt, and Hollywood 2016). ومع ذلك، وجدت هذه الدراسة أيضًا أن الأفراد المذكورين في القائمة كانوا أكثر عرضةً للاعتقال بتهمة ارتكاب حوادث إطلاق النار، رغم أنها أشارت إلى أن هذه الاحتمالية الأكبر قد تعود إلى حقيقة أن هؤلاء الأفراد موجودون في القائمة، الأمر الذي أدَّى إلى زيادة وعي ضبَّاط الشرطة بهؤلاء الأفراد (Saunders, Hunt, and Hollywood 2016). واستجابة إلى هذه الدراسة، صرَّحت إدارة شرطة شيكاغو أنها تحدِّث الخوارزمية المستخدَمة لإعداد قائمة الخاضعين للمراقبة الاستراتيجية بانتظامٍ وأن فعالية القائمة قد زادت منذ عام ٢٠١٣ (Rhee 2016). وثمة سؤال آخر عن قوائم مكافحة الجريمة المَبنية على البيانات وهو: كيف ينتهي الأمر بوجود الفرد على هذه القائمة؟ ويبدو أن نسخة ٢٠١٣ من قائمة الخاضعين للمراقبة الاستراتيجية قد جُمِعت باستخدام تحليل الشبكة الاجتماعية الخاصة بالفرد، بما في ذلك تاريخ حوادث اعتقال أحد معارف هذا الفرد وحوادث إطلاق النار التي تورَّطوا بها، وهذا من بين سماتٍ أخرى للفرد (Dokoupil 2013; Gorner 2013). ومن ناحية أخرى، يبدو من المنطقي استخدام تحليل الشبكة الاجتماعية، إلا أنها تفتح الباب أمام مشكلةٍ واقعية جدًّا وهي مبدأ الإدانة بالتلازُم. وإحدى المشاكل المقترنة بهذا

المبدأ هي أنه قد يكون من الصعب تحديد ما يَعنيه التلازم المقصود هنا. هل العيش في الشارع نفسه مع مجرم كافٍ ليكون هناك تلازُم؟ علاوة على ذلك، في الولايات المتحدة، حيث تكون الغالبية العظمى من السجناء ذكورًا أمريكيين من أصلٍ إفريقي أو لاتيني، من المرجَّح أن يسفر السماح لخوارزميات التنبُّؤ بالجرائم بالاستعانة بمفهوم الإدانة بالتلازُم عن تنبؤات تستهدف بالأساس الشباب ذوي البشرة الملوَّنة (Baldridge 2015).

تعني الطبيعة التنبؤية لبرامج التنبؤ بالجرائم أنه مِن المحتمَل أن تختلف معاملة الأفراد ليس بسبب ما اقترفوه وإنما بسبب الاستدلالات المَبنية على البيانات بخصوص ما قد يفعلونه. ونتيجة لذلك، ربما تعمل هذه الأنواع من الأنظمة على تعزيز الممارسات القائمة على التمييز العنصري من خلال تكرار الأنماط في البيانات القديمة وربما يؤدي هذا إلى نبوءات تتحقَّق ذاتيًّا.

التداعيات الأخلاقية لعلم البيانات: إنشاء سجن «بانوبتيكون» آخر

إذا قضيتَ وقتًا كافيًا في التعرض للترويج التجاري لعلم البيانات، فسوف ينتابُك شعور بأن أية مشكلة يمكن حلها باستخدام تكنولوجيا علم البيانات، ما دام لدَينا قدر كافٍ من المعلومات الصحيحة. وهذا التسويق لعلم البيانات يُغذي وجهة نظرٍ مفادها أن المنهج القائم على استخدام البيانات في الحوكمة هو أفضل طريقة للتعامل مع المشكلات الاجتماعية المعقدة مثل الجريمة والفقر وتدنِّي مستوى التعليم وتدنِّي مستوى الصحة العامة: وكل ما نحتاجه لحلِّ هذه المشكلات هو وضع أجهزة استشعار في مجتمعاتنا لمراقبة كل شيء، ودمج جميع البيانات وتشغيل الخوارزميات لتوليد رؤًى رئيسية تُقدم الحل.

عند قبول وجهة النظر هذه، عادة ما يُشدَّد على عمليتَين. الأولى هي أن المجتمع يصير ذا طبيعة تكنوقراطية أكثر، وتبدأ جوانب الحياة في التنظيم بواسطة أنظمةٍ قائمة على البيانات. وتوجد بالفعل أمثلة على هذا النوع من التنظيم التكنولوجي؛ على سبيل المثال، في بعض الولايات القضائية يُستخدم علم البيانات حاليًّا في جلسات الاستماع للإفراج المشروط (Berk and Bleich 2013) وجلسات إصدار الأحكام (Barry-Jester, Casselman, and Goldstein 2015). وبالنظر إلى مثالٍ من خارج النظام القضائي، تأمل كيف تُنظم تقنيات المدن الذكية حركة تدفق المرور عبر المدن بخوارزميات تُحدد بفعاليةٍ أي تدفق مرور يحصل على الأولوية عند تقاطعٍ مُعين في أوقات مختلفة من اليوم

(Kitchin 2014b). ومن بين النواتج الثانوية لهذا التنظيم التكنوقراطي انتشار أجهزة الاستشعار التي تدعم أنظمة التنظيم الآلية. والعملية الثانية هي «الزحف الرقابي»، حيث يُعاد استخدام البيانات التي جُمِعت لغرضٍ مُعين لأداء غرضٍ آخر وتُستخَدم في التنظيم بطريقةٍ أخرى (Innes 2001). على سبيل المثال، كاميرات الشوارع التي رُكِّبت في لندن بغرض تنظيم الازدحام وتحصيل رسوم الازدحام (رسوم الازدحام المروري في لندن هي رسوم يومية لقيادة المركبات داخل لندن أثناء أوقات الذروة) قد استخدمت بغرَض أداء مهامَّ أمنية (Dodge and Kitchin 2007). تشمل أمثلةً أخرى على «الزحف الرقابي» تقنية تُسمَّى «شوتسبوتر» وهي تتألف من شبكة ميكرفونات مركَّبة على نطاق المدينة ومصمَّمة للتعرُّف على أصوات الطلقات النارية والإبلاغ عن أماكنها غير أنها تسجل أيضًا المحادثات، وقد استُخدِم بعض هذه المحادثات في الإدانات الجنائية (Weissman 2015)، وأيضًا استخدام أنظمة الملاحة داخل السيارة لفرض المراقبة ورصد غراماتٍ على سائقي السيارات المستأجَرة الذين يقودون السيارات خارج الولاية (Elliott 2004; Kitchin 2014a).

ويتمثل أحد جوانب «الزحف الرقابي» في الرغبة في دمج البيانات الواردة من مختلف المصادر بهدف تقديم صورةٍ أكثرَ اكتمالًا للمجتمع وبالتالي ربما الكشف عن رؤًى أعمق للمشكلات الموجودة في النظام. غالبًا ما توجد أسباب وجيهة وراء إعادة استخدام البيانات لأغراضٍ مختلفة. وفي الواقع تُوجَّه نداءاتٌ كثيرة لدمج البيانات التي تحتفظ بها مختلف الهيئات الحكومية لأغراضٍ مشروعة، مثل دعم البحوث الصحيَّة ومن أجل مصلحة الدولة ومواطنيها. غير أنه من منطلق الحريات المدنية، هذه الاتجاهات مُثيرة للقلق جدًّا. إن تشديد المراقبة، ودمج البيانات من مصادر مُتعددة، والزحف الرقابي، والحوكمة الاستباقية (مثل برامج التنبُّؤ بالجرائم) قد تسفر عن مجتمع قد يُعامل فيه الفرد بارتيابٍ فقط لأنه قام بسلسلةٍ من التصرُّفات أو المقابلات الحسنة النية غير ذات الصِّلة والتي تتوافق مع نمَطٍ يَعُدُّه النظام الرقابي المبني على البيانات مُريبًا. والعيش في هذا النوع من المجتمعات من شأنه أن يُحيلنا من مواطنين أحرار إلى سجناء في سجن الفيلسوف الإنجليزي بنثام، أي سجن «بانوبتيكون»،[5] مما يجعلنا نُمارس ضبط النفس باستمرار على سلوكياتنا تخوفًا من الاستنتاجات التي قد تُستدَلُّ منها. والفارق بين الأفراد الذين يعتقدون أنهم غير خاضعين للمراقبة ويتصرفون وفقًا لذلك والأشخاص الذين يمارسون ضبط النفس خوفًا، حيث إنهم نزلاء في سجن «بانوبتيكون»، أشبهُ بالفارق بين مجتمعٍ حُرٍّ ودولةٍ استبدادية.

رحلة البحث عن الخصوصية المفقودة

عندما يتفاعل الأفراد مع مجتمعاتٍ حديثة من الناحية التقنية ويتنقَّلون عبرها، لا يكون أمامهم خيار سوى أن يتركوا وراءهم أثرًا للبيانات يمكن تعقُّبهم من خلاله. ففي العالم الحقيقي، يعني انتشار كاميرات المراقبة أنه يمكن جمع بياناتٍ مكانية عن الأفراد في كل مرةٍ يظهرون في الشارع أو عند متجرٍ أو عند ساحة انتظار سيارات، وانتشار الهواتف المحمولة يعني أنه يمكن تعقُّب أشخاصٍ كثيرين بواسطة هواتفهم. وتشمل الأمثلة الأخرى عن جمع بيانات العالَم الحقيقي تسجيل مُشتريات بطاقات الائتمان، واستخدام برامج بطاقات الولاء في السوبرماركت، وتتبُّع عمليات السحب من ماكينات السحب الآلي، وتتبُّع إجراء المكالمات الهاتفية. وفي عالم الإنترنت، تُجمَع البيانات عن الأفراد عندما يزورون أو يُسجلون الدخول على المواقع الإلكترونية، أو يرسِلون رسائل بريد إلكتروني، أو يُشاركون في التسوق الإلكتروني، أو يُقيمون مطعمًا أو متجرًا، أو يستخدِمون تطبيقًا لقراءة الكتب الإلكترونية، أو يُشاهدون محاضرةً في دورةٍ تدريبية مجَّانية عبر الإنترنت أو يُسجلون إعجابهم بمنشورٍ على أحد مواقع التواصل الاجتماعي أو ينشرونه. ولكي نضع كمية البيانات التي تُجمَع عن الفرد العادي في مجتمعٍ حديث من الناحية التقنية في نِصابها الصحيح، قدَّر تقرير صادر عن «هيئة حماية البيانات الهولندية» في عام ٢٠٠٩ أن المواطن الهولندي العادي يُدرَج في عددٍ من قواعد البيانات يتراوح من ٢٥٠ إلى ٥٠٠ قاعدة بيانات، مع تزايد هذا العدد وصولًا إلى ألف قاعدة بيانات بالنسبة للأشخاص الأكثر نشاطًا اجتماعيًّا (Koops 2011). تُحدِّد نقاط البيانات المتعلقة بشخصٍ ما، عند جمعها معًا، «البصمةَ الرقمية» لذلك الشخص.

ويمكن جمع البيانات الموجودة في البصمة الرقمية في سياقَين يُمثلان إشكاليةً من منظور الخصوصية. أولًا: يمكن جمع البيانات عن شخصٍ ما دون علمه أو معرفته. ثانيًا: في بعض السياقات، ربما يختار شخصٌ ما مشاركة بياناتٍ عن نفسه وآرائه؛ ولكن ربما يكون لدَيه قليل من العلم أو لا يكون لديه علم أو سيطرة على كيفية استخدام هذه البيانات أو كيفية مشاركتها مع أطرافٍ أخرى وإعادة استخدامها لأغراضٍ أخرى. ويُستخدم مصطلحا «الظل الرقمي» و«البصمة الرقمية»[6] للتفرقة بين هذين السياقَين لجمع البيانات: فالظلُّ الرقمي يشمل جمع البيانات عن الفرد دون علمِه أو رضاه أو معرفته، أما البصمة الرقمية فهي عبارة عن أجزاءٍ من البيانات التي يُتيحها الفرد علنًا وعن قصد (Koops 2011).

وبالطبع، جمع بيانات عن شخصٍ بدون معرفته أو رضاه هو أمر مُثير للقلق. غير أن قوة تقنيات علم البيانات الحديثة المستخدَمة للكشف عن الأنماط الخفية في البيانات بالإضافة إلى عملية دمج البيانات من عدة مصادر وإعادة استخدامها لأغراضٍ أخرى تعني أنه حتى البيانات التي جُمِعت بمعرفة الفرد ورضاه، في سياقٍ قد يجعلها ذات آثارٍ سلبية على ذلك الفرد، يستحيل حتى توقُّعها. اليوم، مع الاستعانة بتقنيات علم البيانات الحديثة، حتى المعلومات الشخصية جدًّا التي ربما لا نرغب في الإعلان عنها واخترْنا عدم مشاركتها لا يزال من الممكن الاستدلال عليها على نحوٍ موثوق من البيانات غير ذات الصِّلة التي ننشُرها بمحض اختيارنا على مواقع التواصل الاجتماعي. على سبيل المثال، يُعجَب الكثيرون، بمحض اختيارهم، بشيءٍ على موقع فيسبوك لأنهم يرغبون في إبداء الدعم لصديق. ومع ذلك، ومن خلال الاستعانة بكل بساطة بالأشياء التي يُعجَب بها الأفراد على موقع فيسبوك تستطيع النماذج المَبنية على البيانات أن تتوقَّع بدقة التوجُّه الجنسي الخاص بالشخص وآراءه السياسية والدينية ومستوى ذكائه وسِماته الشخصية واستخدامه للمواد المسببة للإدمان مثل المشروبات الكحولية والمخدرات والسجائر؛ بل ويمكنها أن تُحدِّد ما إذا كان والدا ذلك الشخص ظلَّا معًا حتى بلغ الشخص السنَّ القانونية (Kosinski, Stillwell, and Graepel 2013). وتُثبَت روابط غير بديهية بهذه النماذج من خلال التنبؤ بالمثلية الجنسية لفردٍ نتيجة إعجاب الفرد بحملةٍ للدفاع عن حقوق الإنسان (لكلٍّ من الذكور والإناث) والتنبؤ بعدم مَيلِه للتدخين نتيجة إعجابه بسيارات هوندا (Kosinski, Stillwell, and Graepel 2013).

المناهج الحوسبية للحفاظ على الخصوصية

في السنوات الأخيرة، كان هناك اهتمام متزايد بالمناهج الحوسبية للحفاظ على خصوصية الأفراد طوال عملية تحليل البيانات. وهناك اثنان من المناهج المعروفة هما: «الخصوصية التفاضلية» (ويُطلق عليها أيضًا الخصوصية التبايُنية) و«التعلُّم التشاركي» (ويُطلق عليه أيضًا «التعلُّم المتحد»).

الخصوصية التفاضلية هي منهج رياضي لحلِّ مشكلة معرفة معلومات مفيدة عن مجموعة سكانية، وفي الوقت نفسه عدم معرفة أي شيءٍ عن الأفراد داخل المجموعة. وتستخدِم الخصوصية التفاضلية تعريفًا خاصًّا للخصوصية: خصوصية الفرد لا يتم

المساس بها من خلال تضمين بياناته في عملية تحليل البيانات إذا كانت الاستنتاجات التي تم التوصُّل إليها من خلال التحليل لن تتأثر إذا لم تُضمَّن بيانات الفرد. ويمكن الاستعانة بعددٍ من العمليات لتطبيق الخصوصية التفاضلية. وتأتي في صلب هذه العمليات فكرة ضخ التشويش إما في عملية جمع البيانات أو في الاستجابات إلى الاستعلامات الخاصة بقاعدة البيانات. ويحمي التشويش خصوصية الأفراد ولكن يمكن إزالته من البيانات على مستوى التجميع بحيث يمكن حساب إحصائيات مفيدة على مستوى المجموعة السكانية ككل. وتُعد تقنية الاستجابة المعشَّاة مثالًا مفيدًا على إجراء ضخ التشويش في البيانات وتقدم تفسيرًا بديهيًّا لآلية عمل الخصوصية التفاضلية. وتتمثل حالة استخدام هذه التقنية في استطلاع رأيٍ يتضمن سؤالًا حساسًا إجابته بنعم أو لا (أي سؤال متعلق بانتهاك القانون، أو بالإصابة بمرض ما، وما إلى ذلك). يُوجَّه المشاركون في الاستطلاع إلى الإجابة عن السؤال الحسَّاس باستخدام الإجراء التالي:

(١) اقذف عملة معدنية في الهواء والْتقطها وأبقِ النتيجة سرًّا (هل كانت على الصورة أم الكتابة؟).

(٢) إذا كانت النتيجة كتابة، فأجب بـ «نعم».

(٣) إذا كانت النتيجة صورة، فأجب بصدق.

سيحصل نصف المشاركين في هذا الاستطلاع على «كتابة» ويجيبون بـ «نعم»؛ وسيجيب النصف الآخر بصدق. وهكذا، فإن العدد الحقيقي للمشاركين الذين كان يُفترَض أن يُجيبوا بـ «لا» في المجموعة بأكملها ضعف عدد الذين أجابوا فعليًّا بـ «لا» (تقريبًا) (فالعملة المعدنية أداة منصفة وتختار عشوائيًّا، وبالتالي ينبغي أن يعكس توزيع إجابات نعم/لا بين المشاركين الذين حصلوا على نتيجة «كتابة» عدد المشاركين الذين أجابوا بالحقيقة). وبالوضع في الاعتبار العدد الحقيقي للإجابة بـ «لا»، يُمكننا أن نحسب العدد الحقيقي للإجابة بـ «نعم». ومع ذلك، على الرغم من أننا الآن لدينا عدد دقيق لمن أجابوا على السؤال الحساس بـ «نعم» في المجموعة، فإننا لا نستطيع تحديد أي من المشاركين الذين أجابوا بنعم ينطبق عليهم بالفعل الظرف الحساس. ثمة موازنة بين كمية التشويش الذي يُضخُّ في البيانات وفائدة البيانات في تحليل البيانات. تتعامل الخصوصية التفاضلية مع هذه الموازنة من خلال تقديم تقديرات لحجم التشويش المطلوب بأن تُوضَع في الاعتبار عوامل مثل توزيع البيانات داخل قاعدة البيانات، ونوع استعلام قاعدة البيانات الذي تتم معالجته وعدد الاستعلامات التي نرغب من خلالها في ضمان خصوصية الفرد.

قدَّمت سينثيا دوورك وآرون روث (٢٠١٤) مقدمةً إلى الخصوصية التفاضلية ونظرة عامة على عدة مناهج لتطبيق الخصوصية التفاضلية. والآن، تُطبَّق تقنيات هذه الخصوصية في عددٍ من المنتجات الاستهلاكية. على سبيل المثال، تستخدم شركة أبل الخصوصية التفاضلية في نظام التشغيل «آي أو إس ١٠» (iOS 10) لحماية خصوصية المستخدِمين الأفراد وفي الوقت نفسه تتعلَّم أنماط الاستخدام لتحسين النصوص التنبُّؤية في تطبيقات الرسائل ولتحسين وظيفة البحث.

في بعض السيناريوهات، تأتي البيانات المستخدَمة في مشروع علم البيانات من مصادر متعدِّدة ومتباينة. على سبيل المثال، ربما تساهم مستشفيات عديدة في مشروعٍ بحثي واحد، أو تجمع إحدى الشركات بياناتٍ من عددٍ مهول من مستخدمي أحد تطبيقات الهاتف المحمول. وبدلًا من إضفاء طابع المركزية على هذه البيانات في مستودع بيانات واحد وإجراء التحليل على البيانات المجمعة، يتمثل المنهج البديل في تدريب نماذج مختلفة على مجموعات البيانات الفرعية في مصادر البيانات المختلفة (أي في المستشفيات كلٍّ على حدة أو على هواتف المستخدِمين كلٍّ على حدة) ثم دمج هذه النماذج المدرَّبة معًا. تستعين شركة جوجل بهذا المنهج «للتعلُّم التشاركي» لتحسين المقترحات الخاصة بالاستعلام التي تُقدمها لوحة مفاتيح جوجل المستخدَمة على نظام الأندرويد (McMahan and Ramage 2017). وفي إطار عمل التعلُّم التشاركي الخاص بجوجل، يحتوي جهاز المحمول مبدئيًّا على نسخةٍ مُحملة من التطبيق الحالي. وعندما يستخدِم المستخدِم التطبيق، تُجمَع بيانات التطبيق الخاصة بذلك المستخدِم على هاتفه وتُستخدَم بواسطة خوارزمية تعلُّم محلية موجودة على هذا الهاتف لتحديث النسخة المحلية من النموذج. ثم يُرفَع هذا التحديث المحلي للنموذج على السحابة مع باقي التحديثات المحلية للنماذج الخاصة بالمستخدِمين الآخرين، ويُنشأ نموذج بمتوسط التحديثات التي تمَّت على كل هذه النماذج. يحدَّث النموذج الأساسي بعد ذلك باستخدام هذا المتوسط. وباستخدام هذه العملية، يمكن تحسين النموذج الأساسي، ويمكن في الوقت نفسه حماية خصوصية المستخدِمين الفرديِّين إلى الحدِّ الذي لا تُشارك فيه سوى تحديثات النموذج فقط؛ لا بيانات استخدام من جانب المستخدِمين.

الحقيقة أن خوارزميات علم البيانات تعمل بطريقة لا أخلاقية أكثر من كونها طريقة موضوعية.

الأطر القانونية لتنظيم استخدام البيانات وحماية الخصوصية

ثمة تنوُّع عبر الولايات القضائية المختلفة في القوانين المتعلقة بحماية الخصوصية والاستخدام المسموح به للبيانات. ومع ذلك، ثمة ركنان أساسيَّان موجودان عبر معظم الولايات القضائية الديمقراطية في التشريعات المناهضة للتمييز وتشريعات حماية البيانات الشخصية.

ففي معظم الولايات القضائية، تحظُر التشريعات المناهضة للتمييز التمييزَ على أساسِ أيٍّ من الأسباب التالية: الإعاقة، والعمر، والجنس، والعرق، والانتماء العرقي، والجنسية، والتوجُّه الجنسي، والآراء الدينية أو السياسية. وفي الولايات المتحدة، يحظُر قانون الحقوق المدنية لعام ١٩٦٤[7] التمييز على أساس اللون، أو العرق، أو الجنس، أو الدين، أو الجنسية. ولقد وسَّعت التشريعات اللاحقة نطاق هذه القائمة؛ على سبيل المثال، وسع قانون الأمريكيين ذوي الإعاقة لعام ١٩٩٠[8] نطاق حماية الأفراد من التمييز ليشمل الحماية من التمييز القائم على الإعاقة. وثمة تشريعات مُشابهة معمول بها في العديد من الولايات القضائية الأخرى. على سبيل المثال، يحظر ميثاق الحقوق الأساسية للاتحاد الأوروبي التمييز على أساس أية أسباب بما فيها العرق، واللون، والأصل العرقي أو الاجتماعي، والسمات الجينية، والجنس، والعمر، والميلاد، والإعاقة، والتوجه الجنسي، والدين أو المعتقد، والممتلكات، والعضوية في أقليةٍ وطنية، والرأي السياسي أو أيِّ رأيٍ آخر (ميثاق ٢٠٠٠).

ويُوجَد حالة مُماثلة من التبايُن والتداخل فيما يخص تشريعات الخصوصية في مختلف الولايات القضائية. ففي الولايات المتحدة الأمريكية، وفَّرت مبادئ الممارسات العادلة للمعلومات (عام ١٩٧٣)[9] الأساس للكثير من التشريعات التالية لحماية الخصوصية في تلك الولايات القضائية. وفي الاتحاد الأوروبي، تُعد توجيهات حماية البيانات (مجلس الاتحاد الأوروبي والبرلمان الأوروبي عام ١٩٩٥) هي الأساس للكثير من تشريعات حماية الخصوصية الخاصة بتلك الولايات القضائية. تتوسَّع اللائحة العامة لحماية البيانات (مجلس الاتحاد الأوروبي والبرلمان الأوروبي عام ٢٠١٦) في نطاق مبادئ حماية البيانات المنبثقة من توجيهات حماية البيانات وتوفر لوائح حماية بيانات مُتسقة وقابلة للتنفيذ قانونيًّا في جميع الدول الأعضاء في الاتحاد الأوروبي. غير أن المبادئ الأكثر قبولًا على نطاقٍ واسع فيما يتعلق بالخصوصية الشخصية والبيانات هي المبادئ التوجيهية لحماية الخصوصية وتدفُّقات البيانات الشخصية عبر الحدود التي نشرتها

منظمة التعاون الاقتصادي والتنمية (١٩٨٠). وضمن هذه المبادئ التوجيهية، تُعرَف البيانات الشخصية على أنها سجلَّات متعلقة بفردٍ يمكن التعرف عليه، يُعرف باسم «صاحب البيانات». تُحدد المبادئ التوجيهية ثمانية مبادئ (متداخلة) مُصممة لحماية خصوصية صاحب البيانات:

(١) مبدأ تقييد جمع البيانات: لا ينبغي الحصول على البيانات الشخصية إلا على نحوٍ قانوني وبمعرفة صاحب البيانات وبموجب موافقته.

(٢) مبدأ جودة البيانات: أي بيانات شخصية تُجمَع ينبغي أن تكون ذات صلةٍ بالغرض الذي تُستخدَم من أجله؛ وينبغي أن تكون دقيقة وكاملة ومُحدَّثة.

(٣) مبدأ تحديد الغرض: في توقيت جمع تلك البيانات الشخصية أو قبلها، ينبغي أن يُخطَر صاحب البيانات بالغرض الذي ستُستخدَم من أجله. علاوة على ذلك، على الرغم من أن التغييرات الطارئة على الغرض جائزة، لا ينبغي إقحامها بصورة اعتباطية (يجب أن تكون الأغراض الجديدة متوافقة مع الغرض الأصلي) وينبغي أن يتمَّ تحديدها لصاحب البيانات.

(٤) مبدأ تقييد الاستخدام: استخدام البيانات الشخصية مُقيد بالغرض الذي أُبلغ به صاحب البيانات، وينبغي ألا تُفشى البيانات إلى أطرافٍ أخرى دون موافقة صاحب البيانات أو بموجب سلطة القانون.

(٥) مبدأ الضمانات الأمنية: ينبغي أن تكون البيانات الشخصية محميةً بضماناتٍ أمنية ضد الحذف، أو السرقة، أو الإفشاء، أو التعديل، أو الاستخدام غير المصرَّح به.

(٦) مبدأ الشفافية: يجب أن يكون لدى صاحب البيانات القدرة على الحصول على معلومات بسهولة معقولة بشأن جمع بياناته الشخصية وتخزينها واستخدامها.

(٧) مبدأ مشاركة الفرد: يحقُّ لصاحب البيانات الوصول إلى بياناته الشخصية والاعتراض عليها.

(٨) مبدأ المساءلة: يتحمل معالج البيانات مسئولية الالتزام بهذه المبادئ.

تقر الكثير من الدول، بما فيها دول الاتحاد الأوروبي والولايات المتحدة، بالمبادئ التوجيهية لمنظمة التعاون الاقتصادي والتنمية. في الواقع، يمكن إرجاع مبادئ حماية البيانات المذكورة في اللائحة العامة لحماية البيانات الخاصة بالاتحاد الأوروبي إلى حدٍّ كبير إلى المبادئ التوجيهية لمنظمة التعاون الاقتصادي والتنمية. تنطبق هذه اللوائح على

جمع وتخزين ونقل ومعالجة البيانات الشخصية المتعلقة بمواطني الاتحاد الأوروبي داخل الاتحاد الأوروبي ولها تداعِيات على تدفُّقات هذه البيانات خارج الاتحاد الأوروبي. وحاليًّا، تعمل العديد من البلدان على تطوير قوانين مُماثلة لحماية البيانات مُتَّسقة مع لوائح الاتحاد الأوروبي.

نحو علم بياناتٍ أخلاقي

من المعروف جيدًا أنه على الرغم من الأطر القانونية المعمول بها، كثيرًا ما تجمع الدول القومية البيانات الشخصية عن مواطنيها والأجانب بدون علم هؤلاء الأشخاص، وعادةً ما يكون هذا تحت مُسمَّى الأمن والاستخبارات. وتشمل الأمثلة برنامج المراقبة «بريسم» التابع لوكالة الأمن القومي الأمريكية، وبرنامج «تمبورا» التابع لمكاتب الاتصالات الحكومية البريطانية (Shubber 2013)؛ ونظام الأنشطة التحقيقية التشغيلية التابع للحكومة الروسية (Soldatov and Borogan 2012). تؤثر هذه البرامج على تصوُّر العامة للحكومات واستخدام تقنيات الاتصال الحديثة. ففي عام ٢٠١٥، أشارت نتائج استطلاع رأي بعنوان «استراتيجيات الخصوصية الأمريكية بعد فضيحة سنودن» إلى أن ٨٧ بالمائة من المشاركين كانوا على درايةٍ بمراقبة الاتصالات الهاتفية والإلكترونية، وصرح ٦١ بالمائة ممن كانوا على دراية بهذه البرامج بأنهم فقدوا الثقة بأن هذه البرامج كانت تخدم المصلحة العامة، وقال ٢٥ بالمائة إنهم غيروا طريقة استخدامهم للتقنيات كردٍّ فعلٍ على معرفتهم بهذه البرامج (Rainie and Madden 2015). لقد سُجلت نتائج مشابهة في استطلاعات رأي أوروبية، فكان أكثر من نصف الأوروبيين على دراية بجمع بيانات على نطاقٍ كبير من قبل وكالات حكومية وصرح معظم المشاركين بأن هذا النوع من المراقبة كان له أثر سلبي على ثقتهم فيما يخصُّ كيفية استخدام بياناتهم الشخصية المتاحة عبر الإنترنت (استطلاعات يوروباروميتر ٢٠١٥).

في الوقت نفسه، يتجنَّب الكثير من الشركات الخاصة اللوائح المتعلقة بالبيانات الشخصية والخصوصية من خلال ادعاء استخدام البيانات المشتقة أو المجمعة أو المجهولة المصدر. وبهذه الطريقة، تدَّعي الشركات أن البيانات لم تُعد بياناتٍ شخصية، وهو ما يسمح لها — على حدِّ قولها — بجمع البيانات دون معرفة الفرد أو موافقته ودون غرَض واضح ومباشر للبيانات؛ وبحفظ البيانات لفتراتٍ زمنية طويلة، وبإعادة استخدام البيانات لغرَض آخر أو بيعها حينما تسنح فرصة تجارية لذلك. ويزعم الكثير من أنصار

الفُرَص التجارية لعلم البيانات والبيانات الضخمة أن القيمة التجارية الحقيقية لعلم البيانات تأتي من إعادة استخدام البيانات أو «قيمتها الاختيارية» (Mayer-Schönberger and Cukier 2014). ويلقي أنصار إعادة استخدام البيانات الضوء على ابتكارين من الابتكارات التكنولوجية تجعل من جمع البيانات وتخزينها استراتيجيةً تجارية مناسبة: أولًا: اليوم يمكن جمع البيانات على نحوٍ سلبي بقليل من الجهد أو حتى دون بذل أي جهدٍ أو دون وعي من جانب الأفراد الذين يُتعقَّبون؛ وثانيًا أصبح تخزين البيانات غير مكلفٍ نسبيًّا. ومن هذا المنطلَق يكون من المنطقي على الصعيد التجاري تسجيل البيانات وتخزينها في حال وجود فرصة تجارية مستقبلية (ربما لا يمكن التنبؤ بها) تجعلها ذات قيمة.

وتتعارض الممارسات التجارية الحديثة الخاصة باحتكار البيانات وإعادة استخدامها لأغراضٍ أخرى وبيعها تمامًا مع مبدأي «تحديد الغرض» و«تقييد الاستخدام» في المبادئ التوجيهية لمنظمة التعاون الاقتصادي والتنمية. علاوة على ذلك، تُغفَل أهمية مبدأ «تقييد جمع البيانات» إذا قدمت إحدى الشركات اتفاقية خصوصية إلى العميل مُعدة بحيث تكون غير صالحة للقراءة أو بحيث تحتفظ الشركة بحق تعديل الاتفاقية دون الرجوع إلى العميل أو إخطاره بذلك أو أي من الأمرَين. وكلما حدث ذلك، تحولت عملية إخطار العميل والحصول على موافقته إلى مجرد ممارسة لا معنى لها مُتمثلةً في وضع علامة صح داخل مربع فحسب. وعلى غرار الرأي العام تجاه المراقبة الحكومية تحت مُسمى الحفاظ على الأمن، فإن الرأي العام سلبي تمامًا تجاه جمع المواقع الإلكترونية التجارية للبيانات الشخصية وإعادة استخدامها لأغراضٍ أخرى. مرة أخرى باستخدام استطلاعات الرأي الأمريكية والأوروبية كاختبارٍ حاسم للرأي العام الأوسع نطاقًا، وجدنا استطلاع رأيٍ أُجري في عام ٢٠١٢ لمستخدمي الإنترنت بأمريكا ووصل إلى أن ٦٢ بالمائة من البالغين الذين شاركوا في الاستطلاع لم يكونوا على دراية بكيفية تقييد استخدام المعلومات التي تجمعها المواقع الإلكترونية عنهم، وصرح ٦٨ بالمائة بأنهم غير راضِين عن ممارسة توجيه الإعلانات المستهدفة لأنهم غير راضين عن تتبُّع سلوكهم عبر الإنترنت وتحليله (Purcell, Brenner, and Rainie 2012). ووجد استطلاع رأي حديث للمواطنين الأوروبيين نتائج مشابهة؛ حيث شعر ٦٩ بالمائة من المشاركين بأن جمع بياناتهم ينبغي أن يستلزِم موافقتهم الصريحة على ذلك؛ وأن ١٨ بالمائة فقط من المشاركين يقرءون بنود الخصوصية بالكامل حقًّا. علاوة على ذلك، صرح ٦٧ بالمائة من المشاركين أنهم لا يقرءون بنود الخصوصية

لأنهم يجدونها طويلة على نحوٍ مبالَغ فيه، وصرح ٣٨ بالمائة أنهم وجدوها غير واضحة أو من الصعب جدًّا فهمها. ووجد استطلاع الرأي أيضًا أن ٦٩ بالمائة من المشاركين كانوا قلقِين بشأن استخدام معلوماتهم لأغراضٍ تختلف عن الأغراض التي جُمعت من أجلها، وشعر ٥٣ بالمائة من المشاركين بالانزعاج من شركات الإنترنت التي تستخدِم معلوماتهم الشخصية لتصميم الإعلانات الموجهة لأفرادٍ بعينِهم (استطلاعات يوروباروميتر ٢٠١٥).

إذن الرأي العام، في الوقت الراهن، سلبي بوجهٍ عام نحو كلٍّ من المراقبة الحكومية وجمع البيانات الشخصية وتخزينها وتحليلها من جانب شركات الإنترنت. واليوم، يتفق معظم المعلِّقين على أننا بحاجةٍ إلى تحديث تشريعاتِ خصوصية البيانات وأن التغييرات تحدث. ففي عام ٢٠١٢، نشر الاتحاد الأوروبي والولايات المتحدة مراجعات وتحديثات فيما يخصُّ حماية البيانات وسياسات الخصوصية (European Commission 2012; Federal Trade Commission 2012; Kitchin 2014a, 173). وفي عام ٢٠١٣، وُسِّع نطاق المبادئ التوجيهية لمنظمة التعاون الاقتصادي والتنمية لتشمل مزيدًا من التفاصيل فيما يخص تنفيذ مبدأ المساءلة. وعلى وجه الخصوص، تحدد المبادئ التوجيهية الجديدة مسئولياتِ معالج البيانات فيما يتعلق بوضع برنامج لإدارة الخصوصية ولتحديد ما يستلزِمه هذا البرنامج بوضوحٍ وكيف يجب تصميمه فيما يخصُّ إدارة المخاطر المتعلقة بالبيانات الشخصية (OECD 2013). وفي عام ٢٠١٤، كسب مواطن إسباني، يُدعى ماريو كوستيجا جونزاليس، قضيةً رفعها أمام محكمة العدل الأوروبية ضد شركة جوجل (C-131/12 [2014]) مؤكدًا حقه في أن تُنسى بياناته. وحكمت المحكمة بأنه بإمكان المرء، تحت شروطٍ معينة، أن يطلُب من محرك البحث عبر الإنترنت إزالة روابط صفحات الويب الناتجة عن عمليات البحث باسم هذا الشخص. وتضمَّنت أسباب هذا الطلب أن البيانات غير دقيقة أو قديمة أو أنها احتُفِظ بها لوقتٍ أطول مما هو ضروري لأغراضٍ تاريخية أو إحصائية أو علمية. ولهذا الحكم آثار كبرى على جميع مُحركات البحث عبر الإنترنت، بل إنه ربما يكون له أيضًا آثار على محتكري البيانات الضخمة. على سبيل المثال، ليس من الواضح في الوقت الحاضر ماهية آثاره على مواقع التواصل الاجتماعي مثل فيسبوك وتويتر (Marr 2015). وقد تأكَّد مفهوم الحق في أن تُنسى بيانات المرء عبر ولايات قضائية أخرى. على سبيل المثال، يؤكد قانون «المسح» بكاليفورنيا على حق الشخص القاصر في إزالة مواد نشرها عبر الإنترنت أو من خلال خدمات الهاتف المحمول بناءً على طلبه. ويحظر القانون أيضًا على شركات خدمات الإنترنت أو عبر الإنترنت أو خدمات الهاتف المحمول جمع بياناتٍ شخصية متعلقة بشخص قاصر لأغراض الإعلانات

الموجهة أو السماح لطرف آخر بالقيام بذلك.[10] وكمثالٍ أخير على التغيرات التي تحدث، في عام ٢٠١٦ وُقِّع على اتفاقية «حماية الخصوصية بالاتحاد الأوروبي والولايات المتحدة» واعتمادها (European Commission 2016). وينصبُّ تركيز هذه الاتفاقية على توحيد التزامات حماية خصوصية البيانات عبر الولايتَين القضائيتَين. والغرَض منها هو تعزيز حقوق حماية بيانات مواطني الاتحاد الأوروبي في السياق الذي تُنقل فيه البيانات خارج حدود دول الاتحاد الأوروبي. هكذا، فرضت هذه الاتفاقية التزاماتٍ أقوى على الشركات التجارية فيما يخصُّ شفافية استخدام البيانات، وآليات رقابة قوية وعقوبات مُحتملة وكذلك قيود وآليات رقابة للهيئات العامة في تسجيل البيانات الشخصية أو الوصول إليها. ومع ذلك، وفي وقت تأليف هذا الكتاب، خضعت قوة تلك الاتفاقية وفعاليتها للاختبار في قضيةٍ نُظِرت أمام المحاكم الأيرلندية. والسبب الذي جعل النظام القضائي الأيرلندي في قلب هذا النقاش هو أن الكثير من شركات الإنترنت الأمريكية متعددة الجنسيات (مثل جوجل وفيسبوك وتويتر ... إلخ) توجَد مقراتها الرئيسية لحسابات أوروبا والشرق الأوسط وأفريقيا في أيرلندا. ونتيجة لذلك، يكون المفوض المعني بحماية البيانات في أيرلندا مسئولًا عن تنفيذ لوائح الاتحاد الأوروبي فيما يخصُّ عمليات نقل البيانات عبر الحدود الوطنية التي تقوم بها هذه الشركات. ويوضح التاريخ الحديث أنه من الممكن للقضايا القانونية أن تُسفر عن تغيراتٍ كبرى وسريعة في اللوائح المعنية بكيفية التعامل مع بيانات الأشخاص. في الواقع، تُعَد اتفاقية «درع الخصوصية بين الاتحاد الأوروبي والولايات المتحدة» نتيجة مباشرة لقضية رفعها ماكس شريمز، وهو محامٍ نمساوي وناشط في مجال حماية الخصوصية، ضد فيسبوك. وكانت المحصلة النهائية لقضية شريمز في عام ٢٠١٥ إبطال اتفاقية «الملاذ الآمن» القائمة بين الاتحاد الأوروبي والولايات المتحدة بأثر فوري، ووضعت اتفاقية درع الخصوصية كاستجابةٍ طارئة لهذه النتيجة. ومقارنة باتفاقية «الملاذ الآمن» الأصلية، عزَّزت اتفاقية درع الخصوصية حقوق مواطني الاتحاد الأوروبي في حماية بياناتهم (O'Rourke and Kerr 2017)، وربما يؤدي أي إطار عملٍ جديد إلى تعزيز هذه الحقوق. على سبيل المثال، ستُتيح اللائحة العامة لحماية البيانات لمواطني الاتحاد الأوروبي حماية بياناتهم بموجب القانون اعتبارًا من مايو ٢٠١٨.

ومن منظور علم البيانات، توضح هذه الأمثلة أن اللوائح المتعلقة بخصوصية البيانات وحمايتها في حالة تغيُّر مُستمر. ومن المؤكد أن الأمثلة المذكورة هنا مأخوذة من سياقاتٍ خاصة بالولايات المتحدة والاتحاد الأوروبي؛ ولكنها تُشير إلى اتجاهات أوسع

نطاقًا فيما يتعلق بالخصوصية وتنظيم البيانات. ومن الصعب جدًّا التنبؤ بكيفية تطوُّر هذه التغييرات على المدى الطويل. ويوجَد مجموعة من أصحاب المصالح على هذا الصعيد: تأمَّل الأجندات المختلفة لكبرى شركات الإنترنت، وشركات الإعلان وشركات التأمينات، ووكالات الاستخبارات، والهيئات الأمنية، والحكومات، ومؤسسات البحث العلمي الطبي والاجتماعي وجماعات الحريات المدنية. ولكلٍّ من هذه القطاعات المختلفة في المجتمع أهداف واحتياجات مختلفة فيما يخصُّ استخدام البيانات، وبالتالي فإن لدَيها وجهات نظر مختلفة بشأن كيفية صياغة لوائح حماية خصوصية البيانات. علاوة على ذلك، من المحتمَل أن تكون لدينا كأفرادٍ وِجهاتُ نظر مُتغيرة اعتمادًا على المنظور الذي نتبنَّاه. على سبيل المثال، ربما نرحب جدًّا بمشاركة بياناتنا الشخصية وإعادة استخدامها في سياق الأبحاث الطبية. ومع ذلك، كما سجلت استطلاعات الرأي العام في أوروبا والولايات المتحدة، الكثيرون منَّا لديهم تحفُّظات بخصوص جمع البيانات وإعادة استخدامها ومشاركتها في سياق الإعلانات الموجهة. وبشكلٍ عام، ثمة وجهتا نظر في سياق الحديث حول مستقبل خصوصية البيانات. هناك وجهة نظر تُنادي بتعزيز اللوائح المتعلقة بجمع البيانات الشخصية وتذهب في بعض الحالات إلى حدِّ تمكين الأفراد من التحكُّم في كيفية جمع بياناتهم ومشاركتها واستخدامها. أما وجهة النظر الأخرى فتنادي بإلغاء فرض القيود التنظيمية على جمع البيانات؛ ولكنها تؤيد أيضًا سَنَّ قوانين أشد لمعالجة إساءة استخدام البيانات الشخصية. ومع وجود العديد من أصحاب المصالح ووجهات النظر المختلفة، لا توجد إجابات سهلة أو واضحة للأسئلة المطروحة عن الخصوصية والبيانات. من المرجح أن تتحدد الحلول النهائية التي يُجرى تطويرها بناءً على كل قطاع على حدة وتشمل حلولًا وسطًا يتفاوض عليها أصحاب المصالح المعنيين.

وفي مثل هذا السياق المائع، من الأفضل التصرف بصورةٍ متحفظة وأخلاقية. وبينما نعمل على تطوير حلول جديدة لمشاكل العمل من وجهة نظر علم البيانات، ينبغي أن نضع في اعتبارنا المسائل الأخلاقية المتعلقة بالبيانات الشخصية. وثمة أسباب عملية وجيهة للقيام بهذا. أولًا: سيضمن التصرف على نحوٍ أخلاقي وبشفافية مع البيانات الشخصية أن تحظى الشركة بعلاقات طيبة مع عملائها. فيمكن أن تتسبَّب الممارسات غير اللائقة المتعلقة بالبيانات الشخصية في إلحاق أضرارٍ جسيمة بسمعة الشركة وتدفع عملاءها إلى الانتقال إلى الشركات المنافسة (Buytendijk and Heiser 2013). ثانيًا: ثمة خطورة تتمثل في أن تكثيف عمليات دمج البيانات وإعادة استخدامها وتكثيف عمليات إنشاء

ملفاتٍ شخصية واستهداف المستخدِمين، سيؤدي إلى إثارة الرأي العام بشأن خصوصية البيانات في السنوات القادمة، مما سيسفر عنه لوائح أكثر تشددًا. إن التصرف الواعي بشفافية وعلى نحو أخلاقي هو أفضل طريقة لضمان عدم تعارُض حلول علم البيانات التي نطوِّرها مع اللوائح الحالية أو اللوائح التي قد تدخل في حيِّز التنفيذ في السنوات القادمة.

تذكر أفرا كير (٢٠١٧) قضيةً تعود لعام ٢٠١٥ توضح كيف أن عدم مراعاة الاعتبارات الأخلاقية قد يؤدي إلى عواقب وخيمة بالنسبة إلى مُطوري التكنولوجيا والموردين لخدماتها. وقد ترتب على هذه القضية قيام لجنة التجارة الفيدرالية الأمريكية بفرض غراماتٍ على مطوري وناشري ألعاب التطبيقات بموجب قانون حماية خصوصية الأطفال عبر الإنترنت. لقد دمج مطورو الألعاب إعلاناتٍ خاصةً بجهاتٍ خارجية في ألعابهم المجانية. ويُعَد دمج إعلانات جهاتٍ خارجية هي ممارسة معتادة في نموذج عمل الألعاب المجانية؛ إلا أن المشكلة نشأت من أن الألعاب كانت مُصمَّمة للأطفال دون سنِّ الثالثة عشرة. ونتيجة لذلك، من خلال مشاركة بيانات المستخدِم مع شبكات الدعاية والإعلان، كان المطوِّرون يشاركون أيضًا في الحقيقة بيانات الأطفال، وبالتالي ينتهكون قانون حماية خصوصية الأطفال عبر الإنترنت. وفي إحدى المرات، تقاعس المطورون عن إبلاغ شبكات الدعاية والإعلان أن التطبيقات كانت مُخصَّصة للأطفال. وبالتالي كان من الممكن أن يظهر للأطفال إعلانات غير لائقة، وفي هذه الحالة حكمت لجنة التجارة الفيدرالية بأن يتحمَّل ناشرو اللعبة مسئولية ضمان توفير محتوى وإعلانات مناسبة لأعمار الأطفال الذين يلعبون هذه الألعاب. لقد كان هناك عدد متزايد من هذه النوعية من الحالات في السنوات الأخيرة، ولقد دعت عددٌ من المنظمات، من بينها لجنة التجارة الفيدرالية (٢٠١٢) الشركات إلى اعتماد مبدأ «تضمين الخصوصية في التصميم» (Cavoukian 2013). طُوِّرت هذه المبادئ في التسعينيات من القرن العشرين وصار مُعترفًا بها عالميًا من أجل حماية الخصوصية. فهي تنادي بأنه ينبغي أن تكون الخصوصية هي الوضع الافتراضي لعملية تصميم التكنولوجيا وأنظمة المعلومات. ويتطلب اتِّباع هذه المبادئ من المصمم أن يُبادر بالسعي، وعن وعيٍ، إلى تضمين اعتبارات الخصوصية في تصميم التقنيات والممارسات التنظيمية وهياكل نُظم الشبكات.

وعلى الرغم من أن الحجج الداعمة للجانب الأخلاقي من علم البيانات واضحة، فليس من السهل دومًا التصرف على نحوٍ أخلاقي. وإحدى الطرق لجعل التحدي المتمثل في

الجانب الأخلاقي من علم البيانات ملموسًا أكثر هو أن تتخيل نفسك تعمل في شركةٍ ما كعالِم بيانات على مشروعٍ مُهم تجاريًّا. وفي خضم تحليل البيانات، حددت عددًا من السمات المتفاعلة التي تدلُّ معًا على عِرق مُعين (أو سمة شخصية أخرى مثل الدين والنوع الاجتماعي وما إلى ذلك). أنت تعرف أنه لا يمكنك من الناحية القانونية أن تستخدِم سمة العرق في النموذج الذي تُصمِّمه، ولكنك تؤمِن بأن هذه السمات البديلة ستُمكنك من الالتفاف حول تشريعات مناهضة التمييز. أنت تعتقد أيضًا أن تضمين هذه السِّمات في النموذج سيجعل نموذجك يؤتي ثماره، على الرغم من أنك تشعر بالقلق بطبيعة الحال من أن هذه النتيجة الناجحة قد تتحقق لأن النموذج سيتعلم تعزيز التمييز الموجود بالفعل في النظام. سلْ نفسك: «ماذا سأفعل؟»

الفصل السابع

التأثير المستقبلي لعلم البيانات ومبادئ النجاح

ثمة اتجاه واضح في المجتمعات الحديثة يتمثل في انتشار النظم التي يمكنها استشعار العالم الخارجي والتفاعل معه؛ مثل الهواتف الذكية والمنازل الذكية والسيارات الذاتية القيادة والمدن الذكية. ويُشكل هذا الانتشار للأجهزة الذكية وأجهزة الاستشعار تحديًا أمام خصوصيتنا؛ إلا أنه يُحفز نمو البيانات الضخمة وتطور نماذج التكنولوجيا الحديثة، مثل «إنترنت الأشياء». في هذا السياق، سيكون لعلم البيانات تأثير مُتزايد عبر مجالات عديدة في حياتنا. ومع ذلك، ثمة مجالان سيؤدي علم البيانات فيهما إلى تطوُّرات مهمة خلال العقد القادم؛ ألا وهما: الطب الشخصي وتطوير المدن الذكية.

علم البيانات الطبية

في السنوات الأخيرة، سعى مجال الطبِّ إلى استخدام علم البيانات والتحليلات التنبؤية. كان الأطباء في الماضي يعتمدون بشكلٍ أساسي على خبراتهم وحدسهم في تشخيص الأمراض وتحديد خطة العلاج المناسبة. وتؤكد حركة الطب القائم على الأدلة والطب الدقيق فكرة أن القرارات الطبية ينبغي أن تستند إلى البيانات، وتربط، بصورةٍ مثاليةٍ، أفضل البيانات المتاحة بحالة كل مريضٍ على حدة وتفضيلاته الشخصية. على سبيل المثال، في حالة الطب الدقيق، تُتيح تقنية التحديد السريع لتسلسل الجينوم تحليل جينومات المرضى المصابين بأمراضٍ نادرة من أجل تحديد الطفرات التي تسبَّبت في المرض، وبالتالي تصميم واختيار العلاجات المناسبة لكل فردٍ على حدة. ومن العوامل الأخرى التي تُشجع على استخدام علم البيانات في مجال الطب تكلفة الرعاية الصحية. يمكن استخدام علم البيانات، لا سيما التحليل التنبؤي، لأتمتة بعض عمليات الرعاية الصحية. على سبيل المثال، استخدمت

التحليلات التنبؤية لتحديد التوقيت الذي ينبغي فيه إعطاء المضادات الحيوية وغيرها من الأدوية إلى الأطفال والكبار على حدٍّ سواء، ومن المعروف على نطاقٍ واسع أن هذا النهج قد أنقذ العديد من الأرواح.

تُطوَّر أجهزة استشعار طبية يرتديها أو يبتلِعها المريض أو تُزرع داخله من أجل مراقبة العلامات الحيوية وسلوك المريض ووظائف أعضائه على مدار اليوم. وتُجمَع هذه البيانات باستمرارٍ وتُرسَل مرة أخرى إلى وحدة خدمة مراقبة مركزية. وفي وحدة الخدمة هذه، يمكن لمسئولي الرعاية الصحية الوصول إلى البيانات التي جُمعت من جميع المرضى، وتقييم حالاتهم، وفهم الآثار التي يُحدثها العلاج، ومقارنة نتائج كل مريض بنتائج المرضى الآخرين الذين يعانون من حالاتٍ مُماثلة لإعلامهم بما يجب أن يحدُث في الخطوة التالية من النظام العلاجي الخاص بكل مريض. يستعين علم الطب بالبيانات التي جُمعت من خلال هذه الأجهزة ويدمجها مع بياناتٍ إضافية من أجزاء مختلفة من مهنة الطب والصناعة الدوائية لتحديد آثار الأدوية الحالية والجديدة. وتُطوَّر برامج علاجية مُصممة خصوصًا بناءً على نوع المريض وحالته المرضية وكيفية استجابة جسمه للأدوية المختلفة. بالإضافة إلى ذلك، يوفر هذا النوع الجديد من علم البيانات الطبية المعلومات لأبحاثٍ جديدة عن الأدوية وتفاعلاتها، وتصميم أنظمة مراقبة أكثر كفاءة وتفصيلًا، واكتشاف رؤًى أعمق من التجارب السريرية.

المدن الذكية

تعتمد العديد من المدن حول العالم على تقنيات جديدة لتتمكن من جمع البيانات التي أنتجها مواطنوها واستخدامها من أجل إدارة مؤسسات المدينة ومرافقها وخدماتها على نحوٍ أفضل. وهناك ثلاثة عوامل تمكين أساسية لهذا الاتجاه: علم البيانات والبيانات الضخمة وإنترنت الأشياء. يشير مصطلح «إنترنت الأشياء» إلى ربط الأجهزة المادية وأجهزة الاستشعار المادية بالإنترنت بحيث تتمكن هذه الأجهزة من مشاركة المعلومات. ربما يبدو هذا الأمر عاديًّا؛ ولكن له فائدة تتمثل في أننا يُمكننا الآن التحكم في الأجهزة الذكية عن بُعد (مثل منازلنا إذا هُيِّئت بشكلٍ مناسب) ويفتح الباب أمام إمكانية أن يؤدي الاتصال الشبكي بين الآلات إلى تمكين البيئات الذكية من التنبؤ باحتياجاتنا والاستجابة لها بشكل مُستقلٍّ (على سبيل المثال، يوجد الآن ثلاجات ذكية متاحة تجاريًّا يمكنها أن تُحذرك عندما يوشك الطعام أن يتلف وتُتيح لك طلب الحليب الطازج عبر هاتفك الذكي).

تدمج مشروعات المدن الذكية البيانات اللحظية من العديد من مصادر البيانات المختلفة في مركز بياناتٍ واحد، حيث يُجرى تحليلها واستخدامها للاسترشاد بها في قرارات إدارة المدن وتخطيطها. وتتضمَّن بعض مشروعات المدن الذكية بناء مدنٍ جديدة تمامًا تتَّصف بالذكاء بالكامل من الألف إلى الياء. وتُعَد مدينة «مصدر» في الإمارات العربية المتحدة ومدينة «سونجدو» في كوريا الجنوبية من المدن الجديدة تمامًا التي بُنيت بالتكنولوجيا الذكية مع التركيز على أن تكون صديقةً للبيئة وأن تستخدِم الطاقة بفعالية. ومع ذلك، أغلب مشروعات المدن الذكية تتضمَّن تجديد مدن موجودة بالفعل وتزويدها بشبكاتٍ جديدة من أجهزة الاستشعار ومراكز معالجة البيانات. على سبيل المثال، في مشروع «سمارت سانتاندر» في إسبانيا،[1] رُكِّب أكثر من ١٢ ألف جهاز استشعار مُتصلة جميعها بشبكة عبر المدينة لقياس درجة الحرارة والضوضاء والإضاءة المحيطة ومستويات أول أكسيد الكربون وأماكن وقوف السيارات. وعادةً ما تركز مشروعات المدن الذكية على تطوير الاستخدام الفعَّال للطاقة، وتخطيط حركة المرور وتوجيهها، وتخطيط خدمات المرافق لتلبية احتياجات السكان والنمو السكاني.

تبنت اليابان مفهوم المدينة الذكية مع التركيز بشكلٍ خاص على تقليل استهلاك الطاقة. وقامت شركة طوكيو للطاقة الكهربائية بتركيب أكثر من ١٠ ملايين عدادٍ ذكي في جميع المنازل الواقعة في نطاق منطقة خدمات الشركة.[2] وفي الوقت نفسه، تعمل الشركة على تطوير وإطلاق تطبيقات هواتف ذكية تُمكن العملاء من متابعة استهلاك الكهرباء في منازلهم أثناء الوقت الفعلي للاستهلاك وتغيير بنود التعاقد على خدمات الكهرباء الخاصة بهم. وتتيح هذه التطبيقات الخاصة بالهواتف الذكية للشركة أن تُرسِل نصائح شخصية حول توفير الطاقة لكل عميلٍ حسب استخدامه. وخارج المنازل، يمكن استخدام تكنولوجيا المدن الذكية لتقليل استهلاك الطاقة من خلال الإنارة الذكية للشوارع. ويعمل مشروع «نموذج جلاسكو لمدن المستقبل» على تجريب إضاءة الشوارع التي تُشغَّل ويوقَف تشغيلها بناء على وجود الأفراد في الشارع من عدمه. ويُعَد ترشيد استهلاك الطاقة أولويةً قصوى بالنسبة إلى جميع المباني الحديثة، لا سيما المباني الحكومية المحلية والمباني التجارية الكبيرة. ويمكن تحسين كفاءة استهلاك الطاقة في هذه المباني عن طريق التحكُّم التلقائي في أدوات التحكُّم في المناخ من خلال الجمع بين تكنولوجيا الاستشعار عن بُعد والبيانات الضخمة وعلم البيانات. ومن الفوائد الإضافية لأنظمة مراقبة هذه المباني الذكية هو أنها تستطيع مراقبة مستويات التلوث وجودة الهواء ويُمكنها تفعيل الضوابط والتحذيرات الضرورية في الوقت الفعلي.

تُعد وسائل النقل والمواصلات مجالًا آخر تستخدم فيه المدن علم البيانات. نفذت الكثير من المدن أنظمة مراقبة وإدارة حركة المرور. وتستعين هذه الأنظمة ببيانات الوقت الفعلي للتحكم في تدفق حركة المرور عبر المدينة. على سبيل المثال، يُمكنها التحكم في تسلسل إشارات المرور في الوقت الفعلي، في بعض الحالات لإعطاء أولوية لوسائل النقل العام. كما تعد بيانات شبكات النقل في المدن مفيدة لتخطيط وسائل النقل العام. تفحص المدن الطرق والجداول الزمنية وإدارة وسائل النقل لضمان دعم الخدمات لأكبر عددٍ من الأشخاص وخفض التكاليف المرتبطة بتقديم خدمات النقل. بالإضافة إلى نمذجة الشبكة العامة، يُستخدم علم البيانات أيضًا لمراقبة وسائل النقل الرسمية في المدينة لضمان الاستخدام الأمثل لها. وتجمع هذه المشروعات ظروف الحركة المرورية (التي تجمع بيانات عنها بواسطة أجهزة الاستشعار الموزعة عبر شبكة الطرق، وعند إشارات المرور وما إلى ذلك)، ونوع المهمة التي تُنفَّذ، وغيرها من الظروف لتحسين تخطيط الطرق، وتُزود وسائل النقل بإعدادات الطرق المتغيرة من خلال التحديثات المباشرة والتغيرات الطارئة على مساراتها.

بجانب توفير استهلاك الطاقة وتحسين خدمات النقل، يُستخدم علم البيانات لتحسين جودة تقديم خدمات المرافق وتنفيذ التخطيط طويل الأمد لمشروعات البنية التحتية. وتخضع عملية تقديم خدمات المرافق بكفاءة للمراقبة المستمرة بناءً على الاستهلاك الحالي والاستهلاك المتوقع، وتأخذ عملية المراقبة في الاعتبار الاستهلاك السابق في ظل ظروفٍ مماثلة أيضًا. وتستعين شركات المرافق بعلم البيانات بطرقٍ عدة. إحدى هذه الطرق هي مراقبة شبكة توصيل المرافق: الإمداد وجودة الإمداد وأية مشكلات متعلقة بالشبكة والمناطق التي تتطلب استخدامًا أعلى من المتوقع وإعادة التوجيه التلقائي للإمداد وأية عيوب في الشبكة. طريقة أخرى تستخدمها شركات المرافق هي مراقبة عملائها. فهي تبحث عن الاستخدام غير العادي الذي قد يُشير إلى بعض الأنشطة الإجرامية (على سبيل المثال، وكر لبيع المخدرات)، وعن العملاء الذين ربما يُغيرون المعدات والعدادات الخاصة بالمبنى الذي يُقيمون فيه، والعملاء الذين من المرجَّح أن يتخلفوا عن سداد الأقساط الخاصة بهم. ويُستعان بعلم البيانات أيضًا لدراسة أفضل طريقةٍ لتخصيص الوحدات السكنية والخدمات المرتبطة بها في تخطيط المدن. تُصمم نماذج النمو السكاني للتنبؤ بالمستقبل، واستنادًا إلى عمليات محاكاة متنوعة، يستطيع القائمون على تخطيط المدن تقدير متى وأين تكون هناك حاجة لخدمات دعم معينة، مثل المدارس الثانوية.

مبادئ مشروعات علم البيانات: لماذا تنجح المشروعات أو تفشل؟

أحيانًا يفشل مشروع علم البيانات لأنه لا يُحقق ما كان مرجوًّا منه بسبب تورُّطه في بعض المشكلات التقنية أو السياسية، ولا يقدم نتائج مفيدة، وفي أغلب الأحيان يُشغَّل مرة (أو مرتَين) ولكن لا يُعاد تشغيله مرة أخرى. وكما هي الحال مع عائلات ليو تولستوي السعيدة [التي تحدث عنها في بداية رواية «أنَّا كارنينا» حيث قال: «جميع العائلات السعيدة تتشابه، لكن لكلِّ عائلةٍ تعيسة طريقتها الخاصة في التعاسة» وكان يعني أنه لكي تكون العائلة سعيدة، يجب أن تُحقق النجاح في عدة نواحٍ متنوعة (الحب، والوضع المالي، والحالة الصحية، والمصاهرة)، أما إذا فشلت في أي من هذه النواحي، فسيتسبب ذلك في التعاسة؛ ومن ثم فجميع العائلات السعيدة تتشابه لأنها تُحقق النجاح في جميع النواحي، أما العائلات التعيسة فقد تكون تعيسة لمجموعة من الأسباب المختلفة]،[3] يعتمد نجاح مشروع علم البيانات على عددٍ من العوامل. تحتاج مشروعات علم البيانات الناجحة إلى التركيز والبيانات عالية الجودة والأشخاص المناسبين والرغبة في تجربة نماذج متعددة والاندماج في هياكل وعمليات تكنولوجيا المعلومات الخاصة بالشركة، والتأييد من جانب الإدارة العليا، وإدراك الشركة أنه نظرًا إلى أن العالم يتغير، فإن النماذج تُصبح قديمة وتحتاج إلى إعادة إنشائها مرة أخرى. ومن المرجح أن يؤدي الفشل في أيٍّ من هذه العوامل إلى فشل المشروع برمَّته. يتناول هذا القسم بالتفصيل العوامل المشتركة التي تُحدد نجاح مشروعات علم البيانات وكذلك الأسباب النمطية لفشل هذه المشروعات.

التركيز

يبدأ كل مشروع ناجح من مشروعات علم البيانات بتحديد المشكلة التي سيُساعد المشروع في حلِّها، بكل وضوح. وهذه الخطوة بديهية من عدة نواحٍ: فمن الصعب أن ينجح مشروعٌ ما لم يكن له هدف واضح من البداية. إن وجود هدفٍ محدد جيدًا يرشد عملية اتخاذ القرار بشأن البيانات التي يجب استخدامها، وخوارزميات تعلُّم الآلة التي يجب الاستعانة بها، وكيفية تقييم النتائج، وكيف سيُستخدم التحليل والنماذج ونشرها على نطاقٍ واسع، ومتى يحين الوقت الأمثل لتنفيذ العملية مرةً أخرى من أجل تحديث التحليل والنماذج.

البيانات

ينبغي تحديد أي البيانات التي سيحتاج إليها المشروع. يساعد وجود فهم واضح للبيانات المطلوبة في توجيه المشروع إلى مكان وجودِ هذه البيانات. كما أنه يساعد في تحديد البيانات غير المتوفِّرة حاليًّا، وبالتالي يحدد بعض المشروعات الإضافية التي يمكن أن تسعى إلى جمع وتوفير هذه البيانات. ومع ذلك، من المهم ضمان أن تكون البيانات المستخدَمة عالية الجودة. فقد يكون لدى المؤسسات تطبيقات سيئة التصميم، ونموذج بيانات سيئ للغاية، وطاقم موظفين غير مُدرَّبين على نحوٍ مناسب لضمان إدخال بيانات جيدة. في الواقع، قد تؤدي عوامل لا تُعد ولا تُحصى إلى ظهور بياناتٍ ذات جودة سيئة في النظم. وبالطبع، الحاجة إلى بيانات عالية الجودة هو أمر مُهم للغاية لدرجة أن بعض المؤسسات قامت بتعيين أشخاصٍ يتفقدون البيانات باستمرار، ويُقيِّمون جودتها، ثم يكوِّنون أفكارًا عن كيفية تحسين جودة البيانات المدخَلة من خلال التطبيقات والأشخاص الذين يُدخلون البيانات. وبدون بيانات عالية الجودة، من الصعب جدًّا أن يُكتب النجاح لمشروع علم البيانات.

عند الحصول على البيانات المطلوبة، من المهم دومًا التحقُّق من البيانات التي تُجمع واستخدامها عبر المؤسسة. ولسوء الحظ، فإن النهج الذي يتَّبعه بعض مشروعات علم البيانات للحصول على البيانات هو النظر إلى البيانات المتاحة في قواعد بيانات المعاملات التجارية (أو أي مصادر أخرى للبيانات) ثم دمج هذه البيانات وتنقيتها قبل الانتقال إلى استكشاف البيانات وتحليلها. ويتجاهل هذا النهج تمامًا فريق ذكاء الأعمال وأي مستودع بيانات ربما يكون موجودًا. وفي الكثير من المؤسسات، يجمع فريق ذكاء الأعمال ومستودع البيانات بالفعل بيانات المؤسسة ويقوم بتنظيفها ونقلها ودمجها في مستودعٍ مركزي واحد. فإذا كان يُوجَد مستودع بيانات بالفعل، فهو يحتوي على الأرجح على جميع أو أغلب البيانات المطلوبة للمشروع. وبالتالي، يمكن لمستودع البيانات أن يُوفر قدرًا كبيرًا من الوقت المستغرَق في دمج البيانات وتنظيفها. كما أنه سيحتوي على بياناتٍ أكثر بكثير مما تحتويه قواعد بيانات المعاملات التجارية الحالية. وإذا تمَّت الاستعانة بمستودع البيانات، فمن الممكن العودة لعدة سنوات، وإنشاء نماذج تنبؤية باستخدام البيانات القديمة، وتطبيق هذه النماذج عبر فترات زمنية متعددة، ثم قياس مستوى الدقة التنبؤية لكل نموذج. وتُتيح هذه العملية مراقبة التغير في البيانات وكيف تؤثر على النماذج. بالإضافة إلى هذا، من الممكن مراقبة الفروق في النماذج التي تُنتجها خوارزميات

تعلُّم الآلة وكيف تتطوَّر النماذج بمرور الوقت. إن اتباع هذا النهج يُسهِّل توضيح طريقة عمل النماذج وسلوكها على مدى عدة سنوات ويساعد في بناء ثقة العملاء فيما يُنجَز وما يمكن تحقيقه. على سبيل المثال، في أحد المشروعات حيث أُتيحت في مستودع البيانات بياناتٌ قديمة خاصة بخمس سنوات سابقة، أمكن إثبات أن الشركة كان بإمكانها توفير ٤٠ مليون دولار أميركي أو أكثر خلال تلك الفترة الزمنية. ولو لم يكن مستودع البيانات متاحًا أو مستخدمًا، ما كان من الممكن التوصُّل إلى هذا الاستنتاج. أخيرًا، عندما يستعين المشروع بالبيانات الشخصية من الضروري التأكد من أن استخدام هذه البيانات يتوافق مع اللوائح المعنية بمكافحة التمييز والحفاظ على الخصوصية.

الأشخاص المناسبين

يضم مشروع علم البيانات الناجح عادة فريقًا من الأشخاص الذين يتمتَّعون بمزيج من الكفاءات والمهارات الخاصة بعلم البيانات. وفي معظم المؤسسات، يمكن لمجموعةٍ متنوعة من الأشخاص المشاركة في مشروعات علم البيانات، بل يجب عليهم أن يشاركوا فيها، ومن بين هؤلاء الأشخاص: الأشخاص الذين يعملون مع قواعد البيانات، والذين يتعاملون مع عمليات «الاستخراج والتحويل والتحميل»، والذين يتعاملون مع دمج البيانات، ومديرو المشروعات، وخبراء تحليل الأعمال، وخبراء المجال، وغيرهم. غير أن المؤسسات غالبًا ما تحتاج إلى توظيف مُتخصِّصين في علم البيانات؛ أي أشخاص يتمتعون بمهارات التعامل مع البيانات الضخمة، وتطبيق نماذج تعلُّم الآلة، وصياغة مشكلات العالم الفعلي بحيث يكون لها حلول مَبنية على البيانات. وعلماء البيانات الناجحون مُستعدون وقادرون على العمل والتواصل مع فريق الإدارة والمستخدِمين النهائيين وجميع الأشخاص المعنيين لتوضيح وشرح كيف يمكن لعلم البيانات أن يدعم عملهم. فمن الصعب أن تجد في المؤسسة أشخاصًا يتمتعون بالمهارة التقنية المطلوبة والقدرة على التواصل والعمل مع مختلف الأشخاص. ومع ذلك، فإن هذا المزيج ضروري لنجاح مشروعات علم البيانات في أغلب المؤسسات.

> يضم مشروع علم البيانات الناجح عادة فريقًا من الأشخاص الذين يتمتعون بمزيجٍ من الكفاءات والمهارات الخاصة بعلم البيانات.

النماذج

من المهم تجربة مجموعة متنوعة من خوارزميات تعلُّم الآلة لاكتشاف أيها تعمل بشكل أفضل مع مجموعات البيانات. وفي مواضع كثيرة جدًّا من الأدبيات، تُضرب أمثلة لحالات استُخدِمت فيها خوارزمية تعلُّم آلة واحدة فقط. فقد يناقش المؤلفون الخوارزمية التي نجحت معهم أو التي فضلوها. وحاليًّا هناك اهتمام كبير باستخدام الشبكات العصبية والتعلُّم العميق. ومع ذلك، يمكن استخدام الكثير من الخوارزميات الأخرى ويجب النظر في هذه البدائل واختبارها. علاوة على ذلك، بالنسبة إلى مشروعات علم البيانات القائمة في الاتحاد الأوروبي، ربما تُصبح اللائحة العامة لحماية البيانات، التي دخلت حيِّز التنفيذ في مايو ٢٠١٨، عنصرًا مهمًّا في تحديد اختيار الخوارزميات والنموذج. ومن الآثار الجانبية المحتملة لهذه اللائحة أن «حق الفرد في التفسير» فيما يخصُّ عمليات اتخاذ القرارات الآلية التي تؤثر عليه ربما يحدُّ من استخدام النماذج المعقدة التي يصعب تفسيرها وشرحها (مثل نماذج الشبكات العصبية العميقة) في بعض المجالات.

الاندماج في هياكل وعمليات تكنولوجيا المعلومات الخاصة بالشركة

عندما يُحدَّد الهدف من مشروع علم البيانات، من المهم أيضًا تحديد كيف ستُطبَّق مخرجات المشروع ونتائجه على هيكل تكنولوجيا المعلومات الخاص بالشركة وعلى عمليات العمل. ويتضمن ذلك تحديد مكان وكيفية دمج النموذج في النظم الحالية وكيفية استخدام النتائج المتولِّدة من قبل المستخدِمين النهائيين للنظام أو ما إذا كانت النتائج سيُستفاد منها في عمليةٍ أخرى. وكلما زادت أتمتة هذه العملية، تستطيع الشركة الاستجابة على نحوٍ أسرع مع الملف التعريفي المتغير الخاص بعملائها، وبالتالي خفض التكاليف وزيادة الأرباح المحتملة. على سبيل المثال، إذا أُنشئ نموذج تقييم مخاطر العملاء لعملية الإقراض في بنكٍ ما، ينبغي تضمين هذا النموذج في نظام الاستقبال الذي يتلقَّى طلب القرض الذي يُقدمه العميل. بهذه الطريقة، عندما يُدخل موظف البنك طلب القرض، يستطيع أن يحصل على تقريرٍ مباشر من جانب النموذج. يستطيع الموظف استخدام هذا التقرير المباشر للتعامل مع أية مشكلات تُثار مع العميل. مثال آخر هو كشف الاحتيال. قد يستغرق تحديد حالة الاحتيال المحتمل، التي تستلزم التحقيق، من أربعة إلى ستة أسابيع. ومن خلال الاستعانة بعلم البيانات وتضمينه في نظم مراقبة المعاملات،

تستطيع المؤسسات حاليًّا اكتشاف حالات الاحتيال المحتمَلة على الفور تقريبًا. ومن خلال أتمتة النماذج المبنية على البيانات ودمجها، يتحقَّق وقت استجابة أسرع، ويمكن اتخاذ الإجراءات في الوقت المناسب. وإذا لم تُدمج النواتج والنماذج التي أنشأها مشروع ما في عمليات العمل، فإن هذه النواتج لن تُستخدَم، وفي النهاية سيفشل المشروع.

دعم الإدارة العليا

بالنسبة إلى معظم المشروعات في معظم المؤسسات، يعتبر الدعم الذي تُقدمه الإدارة العليا أمرًا مهمًّا لنجاح الكثير من مشروعات علم البيانات. ومع ذلك، يُركز أغلب مديري تكنولوجيا المعلومات على الأولويات الملحَّة: التأكد من سيرورة العمل، والحفاظ على تشغيل الأنظمة وضمان تشغيل التطبيقات اليومية، والتأكد من وجود عمليات النسخ الاحتياطي والاسترداد (واختبارها)، وما إلى ذلك. تتم رعاية مشروعات علم البيانات الناجحة من قِبل كبار مديري الشركة (وليس مدير تكنولوجيا المعلومات) نظرًا إلى أن كبار مديري الشركة لا يُركزون على التكنولوجيا وإنما يُركزون على العمليات التي ينطوي عليها مشروع علم البيانات وكيف يمكن استخدام نواتج مشروع علم البيانات لصالح الشركة. وكلما ركز راعي المشروع على هذه العوامل، حقق المشروع نجاحًا أكبر. وعندئذٍ سيتصرف بوصفه المسئول عن إعلام باقي الشركة بالمشروع وإقناعهم به. ولكن حتى عندما يكون لمشروع علم البيانات مديرٌ كبير يدعمه، فقد تفشل استراتيجية علم البيانات على المدى الطويل إذا تم التعامل مع المشروع المبدئي لعلم البيانات باعتباره تمرينًا روتينيًّا وتأديةَ واجبٍ فحسب. ينبغي ألا تنظر الشركة إلى مشروع علم البيانات باعتباره مشروعًا لمرة واحدة ولن يتكرَّر. لكي تتمكن الشركة من جني فوائد طويلة الأجل، فإنها بحاجة إلى مضاعفة قُدرتها على تنفيذ مشروعات علم البيانات على نحوٍ مُتكرر واستخدام نواتج هذه المشروعات. يتطلب الأمر التزامًا طويل الأجل من الإدارة العُليا لاعتبار علم البيانات استراتيجية.

> لكي تتمكَّن الشركة من جني فوائد طويلة الأجل، فإنها بحاجةٍ إلى مضاعفة قُدرتها على تنفيذ مشروعات علم البيانات على نحوٍ متكرر واستخدام نواتج هذه المشروعات.

التكرار

تحتاج معظم مشروعات علم البيانات إلى التحديث والتجديد بصورةٍ شبه منتظمة. وفي كل تحديثٍ جديد أو تكرار، يمكن إضافة بيانات جديدة ويمكن إضافة تحديثات جديدة، وربما يمكن استخدام خوارزمية جديدة، وهكذا. وتختلف وتيرة عمليات التكرار من مشروعٍ إلى آخر؛ فقد تكون يومية، أو ربع سنوية، أو نصف سنوية، أو سنوية. ينبغي تضمين عمليات التحقق في مُخرجات علم البيانات المنتجة لاكتشاف متى تكون النماذج بحاجة إلى التحديث (انظر Kelleher, Mac Namee, and D'Arcy 2015 لتفسير كيف يُستخدم مؤشر الاستقرار لتحديد متى يجب تحديث النموذج).

أفكار ختامية

دائمًا يضع البشر فرضيات عن العالم الخارجي ويحاولون فهمه من خلال تحديد أنماط في تجاربهم فيه. ويُعَدُّ علم البيانات هو أحدث تجسيد لهذا السلوك الباحث عن الأنماط. ومع ذلك، على الرغم من أن لعلم البيانات تاريخًا طويلًا، فإن مدى تأثيره على الحياة الحديثة غير مسبوق. ففي المجتمعات الحديثة، عادةً تُستخدم كلمات مثل «دقيق»، «ذكي»، «موجَّه»، «مُخصَّص» للإشارة إلى مشروعات علم البيانات مثل: «الطب الدقيق»، و«الشرطة الدقيقة»، و«الزراعة الدقيقة»، و«المدن الذكية»، و«النقل الذكي»، و«الإعلانات الموجهة»، و«وسائل الترفيه المخصصة». والعامل المشترك بين كل هذه المجالات في حياة الإنسان هو أن ثمة قرارات يجب اتخاذها: ما العلاج الذي ينبغي لنا استخدامه مع هذا المريض؟ أين ينبغي لنا أن نوزع موارد الشرطة المتوفرة لدينا؟ ما كمية الأسمدة التي ينبغي لنا أن نوزعها؟ كم عدد المدارس الثانوية التي ينبغي لنا أن نبنيها في السنوات الأربع المقبلة؟ لمن يجب أن نُرسِل هذا الإعلان؟ ما الفيلم أو الكتاب الذي ينبغي أن نوصي به لهذا الشخص؟ وقدرة علم البيانات على مساعدتنا في اتخاذ القرار هي التي تُشجعنا على اعتماد استخدامه. وإذا نُفِّذ علم البيانات بشكلٍ جيد، يمكنه تقديم «رؤى عملية» تؤدي إلى قرارات أفضل ونتائج أفضل في النهاية.

يعتمد علم البيانات، في شكله الحديث، على البيانات الضخمة، والقدرة الحاسوبية، والبراعة البشرية في عدد من مجالات المساعي العلمية (بداية من التنقيب في البيانات والبحث في قواعد البيانات وصولًا إلى تعلم الآلة). لقد حاول هذا الكتاب تقديم نظرة عامة

على الأفكار والمفاهيم الأساسية اللازمة لفهم علم البيانات. تجعل مراحل العملية القياسية المتعددة المجالات للتنقيب في البيانات عملية علم البيانات واضحة وتقدم هيكلًا لمسار علم البيانات بداية من البيانات وصولًا إلى المعرفة: فهم المشكلة، وإعداد البيانات، والاستعانة بتعلم الآلة لاستخراج الأنماط وإنشاء النماذج، واستخدام النماذج للحصول على رؤًى عملية. يتطرق هذا الكتاب أيضًا إلى بعض المخاوف الأخلاقية المتعلقة بخصوصية الأفراد في عالَم علم البيانات. إن لدى الناس مخاوف حقيقية ومبررة من أن الحكومات وأصحاب المصالح ربما يستغلُّون علم البيانات في التلاعُب بسلوكنا ومراقبة أفعالنا. ونحن — كأفراد — نحتاج إلى تطوير آراء مُستنيرة حول نوع عالم البيانات الذي نرغب أن نعيش فيه، والتفكير في القوانين التي نريد أن تضعها مجتمعاتنا لتوجيه استخدام علم البيانات في الاتجاهات المناسبة. وعلى الرغم من المخاوف الأخلاقية التي قد تُساورنا بشأن علم البيانات، فإن الجنِّيَّ قد خرج من القمقم بالفعل وصار لعلم البيانات تأثيرات مُهمة على حياتنا اليومية وستستمر هذه التأثيرات في المستقبل. وعند استخدامه على نحوٍ مناسب، سوف يكون له القدرة على تحسين جودة حياتنا. ولكن إذا أردنا أن تُحقق الشركات التي نعمل بها والمجتمعات التي نعيش فيها والأسر التي نتشارك حياتنا معها الاستفادة من علم البيانات، فنحتاج إلى فهم واستكشاف ماهية علم البيانات وطريقة عملِه وما يمكنه القيام به (وما لا يمكنه القيام به). ونأمل أن يكون هذا الكتاب قد زوَّدك بالأساسيات الضرورية التي تحتاجها للانطلاق في هذه الرحلة.

مسرد المصطلحات

اكتشاف شذوذ البيانات

يُقصَد بها عملية البحث عن البيانات الشاذة أو المتطرفة في مجموعة البيانات، وتحديد أمثلة عليها. وعادةً ما يُشار إلى هذه الحالات غير المطابقة بـ «قيم الشذوذ» أو «القيم الشاذة». وغالبًا ما تُستخدَم هذه العملية في تحليل المعاملات المالية من أجل رصد أنشطة الاحتيال المحتملة وبدء تحقيقاتٍ بشأنها.

الاستخراج والتحويل والتحميل

يُستخدم هذا المصطلح لوصف العمليات والأدوات المستخدَمة عادةً للمساعدة في تعيين البيانات ودمجها ونقلها بين قواعد البيانات.

إنترنت الأشياء

يقصد به ربط الأجهزة وأجهزة الاستشعار بحيث يتسنَّى لهذه الأجهزة مشاركة المعلومات فيما بينها. ويشمل مجال الاتصال بين آلة وآلة، الذي يُطوِّر نظمًا تمكِّن الآلات ليس فقط من مشاركة المعلومات، وإنما تمكنها أيضًا من الاستجابة لهذه المعلومات واتخاذ الإجراءات اللازمة دون أي تدخُّل بشري.

انتشار عكسي

خوارزمية الانتشار العكسي هي إحدى خوارزميات تعلُّم الآلة وتُستخدَم في تدريب الشبكات العصبية. تحسب الخوارزمية مقدار ما تُسهم به كل خلية عصبية داخل الشبكة في الخطأ

الحادث في هذه الشبكة. ومن خلال عملية حساب الخطأ هذه لكلِّ خلية عصبية يمكن تحديث أوزان الأخطاء بناءً على المدخلات التي تُمرَّر إلى كل خليةٍ عصبية، ومن ثَمَّ الحد من إجمالي الأخطاء في الشبكة. سُميت خوارزمية الانتشار العكسي بهذا الاسم لأن تنفيذها يتم من خلال عمليةٍ من مرحلتَين. في المرحلة الأولى، يُمرَّر مثيل إلى الشبكة في صورة مُدخَل، وتتدفق المعلومات تدفقًا أماميًّا عبر الشبكة حتى تولِّد الشبكة تنبؤًا خاصًا لذلك المثيل. وفي المرحلة الثانية، يُحسَب خطأ الشبكة الخاص بهذا المثيل من خلال مقارنة تنبؤ الشبكة بالمخرج الصحيح لذلك المثيل (كما هو مُحدد بموجب مجموعة البيانات التدريبية) ثم تتم مشاركة هذا الخطأ مرة أخرى (أو ينشر عكسيًّا) عبر الخلايا العصبية في الشبكة، حيث يُوزَّع على طبقة تلوَ الأخرى بدءًا من طبقة المخرجات.

انحدار خطي

عندما يُفترَض وجود علاقة خطية في تحليل الانحدار، يُطلق على التحليل الانحدار الخطي. يُستخدم نوع شائع من نماذج التنبؤ لتقدير قيمة سمةٍ مستهدفة عددية بناءً على مجموعةٍ من السمات المدخلة العددية.

بيانات

البيانات، في أبسط صورها، هي معلومة مجردة عن كيانٍ قائم في الواقع الفعلي (شخص أو شيء أو حدث).

بيانات المعاملات

معلوماتٌ عن حدثٍ ما، مثل بيع سلعة معينة أو إصدار فاتورة أو تسليم بضائع أو الدفع ببطاقة الائتمان، وهلمَّ جرًّا.

بيانات تعريف

عبارة عن بياناتٍ وصفية تصف هياكل بياناتٍ أخرى وخصائصها، ومن الأمثلة على بيانات التعريف الطابع الزمني الذي يصف وقت جمع أحد البيانات. تُعَد بيانات التعريف أحد أكثر أنواع البيانات الثانوية شيوعًا.

بيانات ثانوية

هي البيانات التي تكون ناتجًا ثانويًّا لعمليةٍ ما الهدف الرئيسي منها ليس جمع البيانات. على سبيل المثال، تُنتج مجموعة من البيانات الثانوية مع كل صورة تتم مشاركتها أو إرسالها أو إعادة إرسالها أو الإعجاب بها في تويتر؛ بياناتٍ على غرار مَن شارك الصورة، ومَن شاهدها، والجهاز المستخدَم في ذلك، وفي أي توقيت من اليوم، وهكذا. يُرجى مقارنتها بمصطلح «بيانات مستخلصة».

بيانات ضخمة

عادةً ما تتحدَّد البيانات الضخمة في ضوء ثلاثة عوامل: الحجم الهائل للبيانات، واختلاف أنواع البيانات، والسرعة اللازمة لمعالجة هذه البيانات.

بيانات غير هيكلية

نوع من البيانات يمكن أن يكون فيه لكلِّ مثيلٍ في مجموعة البيانات هيكلُه الداخلي الخاص به؛ أي ليس بالضرورة أن يكون الهيكل متماثلًا في جميع المثيلات. على سبيل المثال، غالبًا ما تكون البيانات النصية بياناتٍ غير هيكلية وتستلزِم تطبيق سلسلة من العمليات عليها حتى يتسنى استخراج تمثيلٍ هيكلي لكل مثيل.

بيانات مستخلَصة

هي البيانات التي تُستخلَص من خلال عملية قياس مباشرة مُصممة خصوصًا لجمع البيانات. يُرجى مقارنتها بتعريف «بيانات ثانوية».

بيانات هيكلية

هي البيانات التي يمكن تخزينها في جدول. وتكون لكل مثيلٍ في الجدول مجموعة السمات نفسها. يُرجى مقارنتها بمصطلح «بيانات غير هيكلية».

تجميع

تحديد مجموعات المثيلات المتشابهة في مجموعة بياناتٍ ما.

تحليل الانحدار

يقدِّر القيمة المتوقعة (أو المتوسطة) لسمةٍ عددية مستهدفة عندما تكون جميع قيم السمات المدخلة ثابتة. ويفترض تحليل الانحدار نموذجًا رياضيًّا قائمًا على المعاملات للعلاقة المفترضة بين المدخَلات والمخرَجات المعروفة باسم «دالة الانحدار». ويمكن أن تحتوي دالة الانحدار على معاملاتٍ متعددة، ويركز تحليل الانحدار على إيجاد الإعدادات الصحيحة لهذه المعاملات.

تحليل البيانات

يقصد به أي عملية لاستخلاص معلوماتٍ مفيدة من البيانات. وتشمل أنواع تحليل البيانات التمثيل المرئي للبيانات، والإحصاءات الموجزة، وتحليل الارتباط، والنمذجة باستخدام تعلُّم الآلة.

تصنيف

هي مهمة يُتنبأ من خلالها بقيمة سمةٍ مستهدفة لمثيل ما بناءً على قيم مجموعة من السمات المدخلة، حيث تكون السمة المستهدفة من نوع البيانات الاسمية أو الترتيبية.

تعلُّم الآلة

مجال في أبحاث علوم الكمبيوتر يركز على إنشاء وتقييم خوارزمياتٍ يمكنها استخراج أنماطٍ مفيدة من مجموعات البيانات. وتأخذ خوارزمية تعلم الآلة مجموعة بيانات باعتبارها مدخَلات، وتنتج نموذجًا يشفر الأنماط التي استخرجتها الخوارزمية من البيانات.

تعلُّم الآلة المدمَج في قواعد البيانات

يُقصَد به استخدام خوارزميات تعلُّم الآلة المدمَجة في حلِّ قاعدة البيانات. ويفيد تعلُّم الآلة المدمج في قواعد البيانات في تقليل الوقت المستغرَق في نقل البيانات داخل قواعد البيانات وخارجها بهدف تحليلها.

تعلُّم خاضع للإشراف

شكل من أشكال تعلُّم الآلة يكون الهدف فيه هو إنشاء دالَّة وتعليمها كيفية تقدير قيمة سمةٍ مستهدفة خاصة بمثيلٍ بالاستدلال بمجموعة من قِيَم السمات المدخَلة الخاصة بهذا المثيل نفسه.

تعلُّم عميق

نموذج التعلُّم العميق هو عبارة عن شبكةٍ عصبية تتضمَّن عدة طبقاتٍ (أكثر من طبقتَين) من الوحدات المخفية (أو الخلايا العصبية). وتُوصَف الشبكات العميقة بالعمق في ضوء عدد طبقات الخلايا العصبية داخل الشبكة. ويتألف الكثير من الشبكات العميقة حاليًّا من عشرات بل من مئات الطبقات. وتنبُع قوة نماذج التعلُّم العميق من قدرة الخلايا العصبية الموجودة في الطبقات الأخيرة على تعلُّم سماتٍ مفيدة مُشتقة من السِّمات التي تعلمَتْها الخلايا العصبية في الطبقات الأولى.

تعلُّم غير خاضع للإشراف

شكل من أشكال تعلُّم الآلة يكون الهدف فيه هو تحديد أنماطٍ مُتسقة في البيانات. وقد تتضمن هذه الأنماط مجموعاتٍ من المَثيلات المتشابهة داخل البيانات أو أنماط وعلاقات بين سماتٍ مختلفة. وعلى عكس التعلُّم الخاضع للإشراف، لا تُحدَّد سمة مُستهدَفة في مجموعة البيانات في هذا الشكل من التعلُّم.

تنبُّؤ

يُقصَد بالتنبؤ في سياق علم البيانات وتعلُّم الآلة مهمة تقدير قيمة إحدى السمات المستهدفة لمثيلٍ مُعين بناءً على قِيم سماتٍ أخرى (أو السمات المدخَلة) لذلك المثيل.

تنقيب عن قواعد الارتباط

أسلوب لتحليل البيانات غير خاضع للإشراف، ويهدف إلى البحث عن مجموعات العناصر التي كثيرًا ما يتكرَّر وجودها معًا. وتتمثل حالة الاستخدام الكلاسيكية لهذا الأسلوب في تحليل سلة التسوق، حيث تحاول متاجر البيع بالتجزئة تحديد مجموعات العناصر التي تُشترى معًا، مثل شراء النقانق والكاتشب والبيرة معًا.

تنقيب في البيانات

هي عملية استخراج أنماطٍ مفيدة من مجموعة البيانات لحل مشكلةٍ محددة جيدًا. تُحدِّد العملية القياسية المتعددة المجالات للتنقيب في البيانات المراحل القياسية لمشروع التنقيب في البيانات. وعملية التنقيب في البيانات وثيقة الصلة بعلم البيانات، ولكنها بوجهٍ عامٍّ ليست على القدر نفسه من سعة النطاق.

جدول التحليل الرئيسي

جدول يحتوي فيه كل صفٍّ على البيانات المتعلقة بمثيلٍ معين، ويصف فيه كل عمود القيم الخاصة بسمةٍ معينة لكل مثيل. وهذه البيانات هي المدخل الأساسي لخوارزميات تعلُّم الآلة والتنقيب في البيانات.

حوسبة عالية الأداء

يركز مجال الحوسبة العالية الأداء على تصميم أطُر عمل وتنفيذها لربط عددٍ كبير من أجهزة الكمبيوتر معًا بحيث يمكن لمجموعة الأجهزة المرتبطة معًا تخزين كمياتٍ مهولة من البيانات ومعالجتها بكفاءة.

خلية عصبية

تستقبل الخلية العصبية عددًا من قِيَم الإدخال في صورة مُدخلات، ثم تعين من خلالها قيمة إخراج واحدة في صورة مُخرجات. وتتم هذه العملية عادةً من خلال تنفيذ دالة انحدارٍ خطي متعددة المدخلات على قِيم الإدخال هذه ثم تمرير ناتج دالَّة الانحدار عبر دالَّة تنشيط غير خطية، مثل الدالة اللوجستية أو دالة ظلِّ الزاوية الزائدي.

سمة

يُوصَف كل مثيلٍ في مجموعة البيانات بعددٍ من السمات (المعروفة أيضًا بـ «الميزات» أو «المتغيرات»). تُسجِّل السمة معلومةً مُعينة عن المثيل. وقد تكون السمة خامًا أو مشتقة.

سمة خام

معلومة مجردة عن كيانٍ ما؛ أي قياس مباشر لهذا الكيان؛ على سبيل المثال، طول شخص مُعين. يُرجى مقارنتها بمصطلح «سمة مشتقة».

سمة مستهدفة

يُقصَد بها في مهامِّ التنبؤ السمة التي تم تدريب نموذج التنبؤ من أجل تقدير قيمتها.

سمة مُشتقة

هي سمة توجَد قيمتها بتطبيق دالَّة على بياناتٍ أخرى بدلًا من استخدام أداة قياس مباشر مأخوذة من الكيان نفسه. ومن أمثلة السمات المشتقة السمة التي تصف قيمةً متوسطة في مجتمع إحصائي. يُرجى مقارنتها بمصطلح «سمة خام».

شبكة عصبية

هي أحد أنواع نماذج تعلُّم الآلة، يُطبَّق على هيئة شبكة مكوَّنة من وحدات معالجة بسيطة تُسمى الخلايا العصبية. ويمكن إنشاء مجموعة متنوعة من أنواع الشبكات العصبية المختلفة من خلال تعديل طوبولوجيا الخلايا العصبية في الشبكة. تُعد الشبكة العصبية المتصلة بالكامل ذات التغذية الأمامية أحد الأنواع الشائعة للغاية من الشبكات التي يمكن تدريبها باستخدام الانتشار العكسي.

علاقة ارتباطية

يقصد بها قوة الارتباط بين سِمتَين.

علم البيانات

مجال ناشئ يدمج مجموعة من تعريفات المشكلات والخوارزميات والعمليات التي يمكن الاستعانة بها في تحليل البيانات من أجل استخراج رؤًى عملية قابلةٍ للتنفيذ من مجموعات البيانات (الكبيرة). وعلم البيانات وثيق الصلة بمجال التنقيب في البيانات، إلا أنه يفوقه

من حيث سعة النطاق ومجالات التركيز والاهتمام. يتعامل هذا العلم مع كلٍّ من البيانات (الضخمة) الهيكلية وغير الهيكلية، ويشمل مبادئ مُستقاة من عدة مجالاتٍ، من بينها تعلُّم الآلة وعلم الإحصاء وأخلاقيات البيانات والقواعد التنظيمية للبيانات والحوسبة العالية الأداء.

عملية قياسية متعددة المجالات للتنقيب في البيانات

تُحدد هذه العملية المراحل القياسية لأيِّ مشروعٍ من مشروعات التنقيب في البيانات. وعادةً ما تمر مشروعات علم البيانات بنفس المراحل.

قاعدة البيانات

هي مستودع مركزي لتخزين البيانات. ويتمثل هيكل قاعدة البيانات الأكثر شيوعًا في قاعدة البيانات الارتباطية، التي تخزَّن من خلالها البيانات على هيئة جداول تتألف من صفٍّ واحد لكل مثيلٍ وعمود واحد لكلِّ سمة. ويُعد هذا التمثيل تمثيلًا نموذجيًّا لتخزين البيانات بهيكلٍ واضح يمكن تفكيكه إلى سماتٍ أساسية.

قاعدة بيانات المعالجة المتوازية الواسعة النطاق

في هذا النوع من قواعد البيانات، تُقسَّم البيانات عبر عدة وحدات خدمة، ويمكن لكل وحدةِ خدمة معالجة البيانات الموجودة عليها محليًّا على نحوٍ مستقل.

لغة الاستعلام الهيكلية

لغة قياسية دولية لتحديد استعلامات قاعدة البيانات.

مثيل

يحتوي كل صفٍّ في مجموعة البيانات على معلوماتٍ عن مثيلٍ واحد (يُعرف أيضًا بـ «مثال»، أو «كيان»، أو «حالة»، أو «سجل»).

مجموعة البيانات

مجموعة من البيانات ذات الصلة بمجموعة من المثيلات، حيث يُوصَف كل مثيلٍ في ضوء مجموعة من السمات. وتُنظَّم مجموعة البيانات، في أبسط صورة لها، على هيئة مصفوفة $n*m$؛ حيث n عدد المثيلات (الصفوف) وm عدد السمات (الأعمدة).

مخزن البيانات التشغيلية

يدمج نظام مخزن البيانات التشغيلية البيانات التشغيلية أو الخاصة بالمعاملات من عدة أنظمة للمساعدة في إنشاء تقارير حول العمليات التشغيلية المختلفة.

مستودع البيانات

عبارة عن مخزنٍ مركزي يتضمن بياناتٍ مُستقاة من مجموعة من المصادر عبر مؤسسةٍ ما. تتم هيكلة البيانات بأسلوب يسهل معه إنشاء تقارير موجزة من البيانات المجمَّعة. ويُستخدم مصطلح «المعالجة التحليلية عبر الإنترنت» لوصف العمليات النموذجية التي تتم على مستودع البيانات.

مدينة ذكية

تحاول مشروعات المدن الذكية بوجهٍ عام دمج البيانات الفورية القادمة من العديد من مصادر البيانات المختلفة في مركز بياناتٍ واحد، حيث تُحلَّل وتُستخدَم للاسترشاد بها في قرارات إدارة المدن وتخطيطها.

معالجة المعاملات عبر الإنترنت

هذه المعالجة مُصممة للمعاملات القصيرة على البيانات عبر الإنترنت (مثل الإدراج والحذف والتحديث وغيرها) مع التأكيد على سرعة معالجة الاستعلامات وضمان صحة البيانات في البيئات التي يمكن الوصول إليها من جهاتٍ متعددة. قارن بينها وبين «المعالجة التحليلية عبر الإنترنت»، المصممة من أجل عملياتٍ أكثر تعقيدًا على البيانات القديمة.

معالجة تحليلية عبر الإنترنت

تُنشئ عمليات المعالجة التحليلية عبر الإنترنت ملخَّصاتٍ للبيانات القديمة وتجمع البيانات من مصادر متعددة. هذه العمليات مُصمَّمة لإنشاء ملخصاتٍ شبيهة بالتقارير، وهي تُتيح للمستخدِمين تقسيم البيانات في مستودع البيانات وتجزئتها وإعادة تنظيمها في جداول محورية باستخدام مجموعةٍ من الأبعاد المحددة مسبقًا، مثل المبيعات حسب المتجر والمبيعات حسب الفترة ربع السنوية وهكذا. يُرجى مقارنتها بمصطلح «معالجة المعاملات عبر الإنترنت».

نظام إدارة قواعد البيانات الارتباطية

هو نظام إدارة قواعد بيانات يستند إلى نموذج البيانات الارتباطية الذي طوَّره إدجار فرانك كود. تُخزِّن قواعدُ البيانات الارتباطية البيانات في مجموعةٍ من الجداول، حيث يكون لكلِّ جدولٍ منها هيكل مكوَّن من صفٍّ واحد لكل مثيل وعمود واحد لكل سمة. ويمكن إنشاء روابط بين الجداول من خلال تضمين سماتٍ أساسية في الجداول المتعددة. يتناسب هذا الهيكل مع استعلامات لغة الاستعلام الهيكلية التي من شأنها تحديد العمليات التي ستُجرى على البيانات الموجودة في الجداول.

نموذج

في سياق تعلُّم الآلة، يُعَدُّ النموذج هو تمثيل أحد الأنماط المستخرجة من مجموعة بياناتٍ ما باستخدام تعلُّم الآلة. ومن ثَمَّ، يتم تدريب النماذج، أو جعلها ملائمة لمجموعة البيانات، أو إنشاؤها عن طريق تطبيق خوارزمية تعلُّم آلة على مجموعة البيانات. وتشمل التمثيلات الشائعة للنماذج الهيكل الشجري لاتخاذ القرار والشبكات العصبية. يُحدِّد نموذج التنبؤ علاقة (أو دالَّة) يُوجِد بموجبها قيمة سمةٍ مستهدفة بناءً على قيم مجموعة من السمات المدخلة. وبمجرد إنشاء النموذج، يمكن تطبيقه على أي حالاتٍ جديدة مشابهة من نفس المجال. على سبيل المثال، من أجل تدريب نموذجٍ لتصفية البريد العشوائي، نقوم بتطبيق خوارزمية تعلُّم آلة على مجموعة بياناتٍ خاصة برسائل بريد إلكتروني قديمة مُصنفة على أنها عشوائية أو غير عشوائية. وبمجرد أن يتم تدريب النموذج، يُمكن استخدامه لتصنيف (أو لتصفية) رسائل البريد الجديدة التي لم تكن موجودةً في مجموعة البيانات الأصلية.

هادوب

منصة مفتوحة المصدر طوَّرتها مؤسسة أباتشي للبرمجيات، وهي مصمَّمة خصوصًا لمعالجة البيانات الضخمة. وتستخدِم التخزين والمعالجة الموزعة عبر مجموعاتٍ من الأجهزة.

هرم البيانات والمعلومات والمعرفة والحكمة

نموذج للعلاقات الهيكلية بين البيانات، والمعلومات، والمعرفة، والحكمة. في هذا الهرم، تأتي البيانات أولًا عند سفح الهرم، تليها المعلومات، ثم المعرفة، ثم الحكمة عند قمة الهرم.

هيكل اتخاذ القرار الشجري

نوع من نماذج يتنبأ بتشفير قواعد if-then-else على هيئة هيكلٍ شجري. كل عقدة في هذا الهيكل الشجري تُحدد سمةً واحدة لاختبارها، ويُحدد المسار المتَّجه من العقدة الجذرية إلى العقدة الطرفية سلسلة من الاختبارات التي يجب أن يجتازها المَثيل حتى يمكن التنبؤ بتسمية العقدة الطرفية لذلك المثيل.

ملاحظات

الفصل الأول: ما علمُ البيانات؟

(1) Quote taken from the call for participation sent out for the KDD workshop in 1989.

(2) Some practitioners do distinguish between data mining and KDD by viewing data mining as a subfield of KDD or a particular approach to KDD.

(3) For a recent review of this debate, see *Battle of the Data Science Venn Diagrams* (Taylor 2016).

(4) For more on the Cancer Moonshot Initiative, see https://www.cancer.gov/research/key-initiatives.

(5) For more on the All of Us program in the Precision Medicine Initiative, see https://allofus.nih.gov.

(6) For more on the Police Data Initiative, see https://www.policedatainitiative.org.

(7) For more on AlphaGo, see https://deepmind.com/research/alphago.

الفصل الثاني: ما المقصود بالبيانات وما المقصود بمجموعة البيانات؟

(1) Although many data sets can be described as a flat *n* m* matrix, in some scenarios the data set is more complex: for example, if a data set describes the evolution of multiple attributes through time, then each time point in the data set will be represented by a two-dimensional flat *n* m* matrix, listing the state of the attributes at that point in time, but the overall data set will be three dimensional, where time is used to link the two-dimensional snapshots. In these contexts, the term *tensor* is sometimes used to generalize the *matrix* concept to higher dimensions.

(2) This example is inspired by an example in Han, Kamber, and Pei 2011.

الفصل الثالث: النظام البيئي لعلم البيانات

(1) See Storm website, at http://storm.apache.org.

الفصل الرابع: أساسيات تعلُّم الآلة

(1) This subheading, Correlations Are Not Causations, but Some Are Useful, is inspired by George E. P. Box's (1979) observation, "Essentially, all models are wrong, but some are useful."

(2) For a numeric target, the average is the most common measure of central tendency, and for nominal or ordinal data the mode (or most frequently occurring value is the most common measure of central tendency).

(3) We are using a more complex notation here involving ω_0 and ω_1 because a few paragraphs later we expand this function to include more than one input attribute, so the subscripted variables are useful notations when dealing with multiple inputs.

(4) A note of caution: the numeric values reported here should be taken as illustrative only and not interpreted as definitive estimates of the relationship between BMI and likelihood of diabetes.

(5) In general, neural networks work best when the inputs have similar ranges. If there are large differences in the ranges of input attributes, the attributes with the much larger values tend to dominate the processing of the network. To avoid this, it is best to normalize the input attributes so that they all have similar ranges.

(6) For the sake of simplicity, we have not included the weights on the connections in figures 14 and 15.

(7) Technically, the backpropagation algorithm uses the chain rule from calculus to calculate the derivative of the error of the network with respect to each weight for each neuron in the network, but for this discussion we will pass over this distinction between the error and the derivative of the error for the sake of clarity in explaining the essential idea behind the backpropagation algorithm.

(8) No agreed minimum number of hidden layers is required for a network to be considered "deep," but some people would argue that even two layers are enough to be deep. Many deep networks have tens of layers, but some networks can have hundreds or even thousands of layers.

(9) For an accessible introduction to RNNs and their natural-language processing, see Kelleher 2016.

(10) Technically, the decrease in error estimates is known as the *vanishing-gradient problem* because the gradient over the error surface disappears as the algorithm works back through the network.

(11) The algorithm also terminates on two corner cases: a branch ends up with no instances after the data set is split up, or all the input attributes have already been used at nodes between the root node and the branch. In

both cases, a terminating node is added and is labeled with the majority value of the target attribute at the parent node of the branch.

(12) For an introduction to entropy and its use in decision-tree algorithms, see Kelleher, Mac Namee, and D'Arcy 2015 on information-based learning.

(13) See Burt 2017 for an introduction to the debate on the "right to explanation."

الفصل الخامس: مهام علم البيانات القياسية

(1) A customer-churn case study in Kelleher, Mac Namee, and D'Arcy 2015 provides a longer discussion of the design of attributes in propensity models.

الفصل السادس: الخصوصية والأخلاقيات

(1) Behavioral targeting uses data from users' online activities—sites visited, clicks made, time spent on a site, and so on—and predictive modeling to select the ads shown to the user.

(2) The EU Privacy and Electronic Communications Directive (2002/58/EC).

(3) For example, some expectant women explicitly tell retailers that they are pregnant by registering for promotional new-mother programs at the stores.

(4) For more on PredPol, see http://www.predpol.com.

(5) A Panopticon is an eighteenth-century design by Jeremy Bentham for institutional buildings, such as prisons and psychiatric hospitals. The defining characteristic of a Panopticon was that the staff could observe the inmates without the inmates' knowledge. The underlying idea of this

design was that the inmates were forced to act as though they were being watched at all times.

(6) As distinct from digital footprint.

(7) Civil Rights Act of 1964, Pub. L. 88–352, 78 Stat. 241, at https://www.gpo.gov/fdsys/pkg/STATUTE–78/pdf/STATUTE–78–Pg241.pdf.

(8) Americans with Disabilities Act of 1990, Pub. L. 101–336, 104 Stat. 327, at https://www.gpo.gov/fdsys/pkg/STATUTE–104/pdf/STATUTE–104–Pg327.pdf.

(9) The Fair Information Practice Principles are available at https://www.dhs.gov/publication/fair–information–practice–principles–fipps.

(10) Senate of California, SB–568 Privacy: Internet: Minors, Business and Professions Code, Relating to the Internet, vol. division 8, chap. 22.1 (commencing with sec. 22580) (2013), at https://leginfo.legislature.ca.gov/faces/billNavClient.xhtml?bill_id=201320140SB568.

الفصل السابع: التأثير المستقبلي لعلم البيانات ومبادئ النجاح

(1) For more on the SmartSantander project in Spain, see http://smartsantander.eu.

(2) For more on the TEPC's projects, see http://www.tepco.co.jp/en/press/corp–com/release/2015/1254972_6844.html.

(3) Leo Tolstoy's book *Anna Karenina* (1877) begins: "All happy families are alike; each unhappy family is unhappy in its own way." Tolstoy's idea is that to be happy, a family must be successful in a range of areas (love, finance, health, in–laws), but failure in any of these areas will result in unhappiness. So all happy families are the same because they are successful in all areas, but unhappy families can be unhappy for many different combinations of reasons.

قراءات إضافية

About Data and Big Data

Davenport, Thomas H. *Big Data at Work: Dispelling the Myths, Uncovering the Opportunities*. Cambridge, MA: Harvard Business Review, 2014.

Harkness, Timandra. *Big Data: Does Size Matter?* New York: Bloomsbury Sigma, 2016.

Kitchin, Rob. *The Data Revolution: Big Data, Open Data, Data Infra–structures, and Their Consequences*. Los Angeles: Sage, 2014.

Mayer–Schönberger, Viktor, and Kenneth Cukier. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Boston: Eamon Dolan/Mariner Books, 2014.

Pomerantz, Jeffrey. *Metadata*. Cambridge, MA: MIT Press, 2015.

Rudder, Christian. *Dataclysm: Who We Are (When We Think No One's Looking)*. New York: Broadway Books, 2014.

About Data Science, Data Mining, and Machine Learning

Kelleher, John D., Brian Mac Namee, and Aoife D'Arcy. *Fundamentals of Machine Learning for Predictive Data Analytics*. Cambridge, MA: MIT Press, 2015.

Linoff, Gordon S., and Michael J. A. Berry. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Indianapolis, IN: Wiley, 2011.

Provost, Foster, and Tom Fawcett. *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. Sebastopol, CA: O'Reilly Media, 2013.

About Privacy, Ethics, and Advertising

Dwork, Cynthia, and Aaron Roth. 2014. "The Algorithmic Foundations of Differential Privacy." *Foundations and Trends® in Theoretical Computer Science* 9 (3–4): 211–407.

Nissenbaum, Helen. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford, CA: Stanford Law Books, 2009.

Solove, Daniel J. *Nothing to Hide: The False Tradeoff between Privacy and Security*. New Haven, CT: Yale University Press, 2013.

Turow, Joseph. *The Daily You: How the New Advertising Industry Is Defining Your Identity and Your Worth*. New Haven, CT: Yale University Press, 2013.

المراجع

Anderson, Chris. 2008. *The Long Tail: Why the Future of Business Is Selling Less of More*. Rev. ed. New York: Hachette Books.

Baldridge, Jason. 2015. "Machine Learning and Human Bias: An Uneasy Pair." *TechCrunch*, August 2. http://social.techcrunch.com/2015/08/02/machine-learning-and-human-bias-an-uneasy-pair.

Barry-Jester, Anna Maria, Ben Casselman, and Dana Goldstein. 2015. "Should Prison Sentences Be Based on Crimes That Haven't Been Committed Yet?" *FiveThirtyEight*, August 4. https://fivethirtyeight.com/features/prison-reform-risk-assessment.

Batty, Mike, Arun Tripathi, Alice Kroll, Peter Wu Cheng-sheng, David Moore, Chris Stehno, Lucas Lau, Jim Guszcza, and Mitch Katcher. 2010. "Predictive Modeling for Life Insurance: Ways Life Insurers Can Participate in the Business Analytics Revolution." Society of Actuaries. https://www.soa.org/files/pdf/research-pred-mod-life-batty.pdf.

Beales, Howard. 2010. "The Value of Behavioral Targeting." Network Advertising Initiative. http://www.networkadvertising.org/pdfs/Beales_NAI_Study.pdf.

Berk, Richard A., and Justin Bleich. 2013. “Statistical Procedures for Forecasting Criminal Behavior.” *Criminology & Public Policy* 12 (3): 513–544.

Box, George E. P. 1979. “Robustness in the Strategy of Scientific Model Building.” In *Robustness in Statistics*, ed. R. L. Launer and G. N. Wilkinson, 201–236. New York: Academic Press.

Breiman, Leo. 2001. “Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author).” *Statistical Science* 16 (3): 199–231. doi:10.1214/ss/1009213726.

Brown, Meta S. 2014. *Data Mining for Dummies*. New York: Wiley. http://www.wiley.com/WileyCDA/WileyTitle/productCd-1118893174,subjectCd-STB0.html.

Brynjolfsson, Erik, Lorin M. Hitt, and Heekyung Hellen Kim. 2011. “Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance?” SSRN Scholarly Paper ID 1819486. Social Science Research Network, Rochester, NY. https://papers.ssrn.com/abstract=1819486.

Burt, Andrew. 2017. “Is There a ‘Right to Explanation’ for Machine Learning in the GDPR?” https://iapp.org/news/a/is-there-a-right-to-explanation-for-machine-learning-in-the-gdpr.

Buytendijk, Frank, and Jay Heiser. 2013. “Confronting the Privacy and Ethical Risks of Big Data.” *Financial Times*, September 24. https://www.ft.com/content/105e30a4-2549-11e3-b349-00144feab7de.

Carroll, Rory. 2013. “Welcome to Utah, the NSA’s Desert Home for Eavesdropping on America.” *Guardian*, June 14. https://www.theguardian.com/world/2013/jun/14/nsa-utah-data-facility.

Cavoukian, Ann. 2013. “Privacy by Design: The 7 Foundation Principles (Primer).” Information and Privacy Commissioner, Ontario, Canada. https://www.ipc.on.ca/wp-content/uploads/2013/09/pbd-primer.pdf.

Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, and Rudiger Wirth. 1999. “CRISP–DM 1.0: Step–by–Step Data Mining Guide.” ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP–DM.pdf.

Charter of Fundamental Rights of the European Union. 2000. *Official Journal of the European Communities* C (364): 1–22.

Cleveland, William S. 2001. “Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics.” *International Statistical Review* 69 (1): 21–26. doi:10.1111/j.1751–5823.2001.tb00477.x.

Clifford, Stephanie. 2012. “Supermarkets Try Customizing Prices for Shoppers.” *New York Times*, August 9. http://www.nytimes.com/2012/08/10/business/supermarkets–try–customizing–prices–for–shoppers.html.

Council of the European Union and European Parliament. 1995. “95/46/EC of the European Parliament and of the Council of 24 October 1995 on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data.” *Official Journal of the European Community* L 281: 38–1995: 31–50.

Council of the European Union and European Parliament. 2016. “General Data Protection Regulation of the European Council and Parliament.” *Official Journal of the European Union* L 119: 1–2016. http://ec.europa.eu/justice/data–protection/reform/files/regulation_oj_en.pdf.

CrowdFlower. 2016. *2016 Data Science Report*. http://visit.crowdflower.com/rs/416–ZBE–142/images/CrowdFlower_DataScienceReport_2016.pdf.

Datta, Amit, Michael Carl Tschantz, and Anupam Datta. 2015. “Automated Experiments on Ad Privacy Settings.” *Proceedings on Privacy Enhancing Technologies* 2015 (1): 92–112.

DeZyre. 2015. “How Big Data Analysis Helped Increase Walmart’s Sales Turnover.” May 23. https://www.dezyre.com/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109.

Dodge, Martin, and Rob Kitchin. 2007. “The Automatic Management of Drivers and Driving Spaces.” *Geoforum* 38 (2): 264–275.

Dokoupil, Tony. 2013. “‘Small World of Murder’: As Homicides Drop, Chicago Police Focus on Social Networks of Gangs.” *NBC News*, December 17. http://www.nbcnews.com/news/other/small-world-murder-homicides-drop-chicago-police-focus-social-networks-f2D11758025.

Duhigg, Charles. 2012. “How Companies Learn Your Secrets.” *New York Times*, February 16. http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html.

Dwork, Cynthia, and Aaron Roth. 2014. “The Algorithmic Foundations of Differential Privacy.” *Foundations and Trends® in Theoretical Computer Science* 9 (3–4): 211–407.

Eliot, T. S. 1934 [1952]. “Choruses from ‘The Rock.’” In *T. S. Eliot: The Complete Poems and Plays—1909–1950*. San Diego: Harcourt, Brace and Co.

Elliott, Christopher. 2004. “BUSINESS TRAVEL; Some Rental Cars Are Keeping Tabs on the Drivers.” *New York Times*, January 13. http://www.nytimes.com/2004/01/13/business/business-travel-some-rental-cars-are-keeping-tabs-on-the-drivers.html.

Eurobarometer. 2015. “Data Protection.” Special Eurobarometer 431. http://ec.europa.eu/COMMFrontOffice/publicopinion/index.cfm/Survey/index#p=1&instruments=SPECIAL.

European Commission. 2012. “Commission Proposes a Comprehensive Reform of the Data Protection Rules–European Commission.” January

25. http://ec.europa.eu/justice/newsroom/data-protection/news/120125_en.htm.

European Commission. 2016. "The EU–U.S. Privacy Shield." December 7. http://ec.europa.eu/justice/data-protection/international-transfers/eu-us-privacy-shield/index_en.htm.

Federal Trade Commission. 2012. *Protecting Consumer Privacy in an Era of Rapid Change*. Washington, DC: Federal Trade Commission. https://www.ftc.gov/sites/default/files/documents/reports/federal-trade-commission-report-protecting-consumer-privacy-era-rapid-change-recommendations/120326privacyreport.pdf.

Few, Stephen. 2012. *Show Me the Numbers: Designing Tables and Graphs to Enlighten*. 2nd ed. Burlingame, CA: Analytics Press.

Goldfarb, Avi, and Catherine E. Tucker. 2011. "Online Advertising, Behavioral Targeting, and Privacy." *Communications of the ACM* 54 (5): 25–27.

Gorner, Jeremy. 2013. "Chicago Police Use Heat List as Strategy to Prevent Violence." *Chicago Tribune*, August 21. http://articles.chicagotribune.com/2013-08-21/news/ct-met-heat-list-20130821_1_chicago-police-commander-andrew-papachristos-heat-list.

Hall, Mark, Ian Witten, and Eibe Frank. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. Amsterdam: Morgan Kaufmann.

Han, Jiawei, Micheline Kamber, and Jian Pei. 2011. *Data Mining: Concepts and Techniques*. 3rd ed. Haryana, India: Morgan Kaufmann.

Harkness, Timandra. 2016. *Big Data: Does Size Matter?* New York: Bloomsbury Sigma.

Henke, Nicolaus, Jacques Bughin, Michael Chui, James Manyika, Tamim Saleh, and Bill Wiseman. 2016. *The Age of Analytics: Competing in a Data-Driven World*. Chicago: McKinsey Global Institute. http://

www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world.

Hill, Shawndra, Foster Provost, and Chris Volinsky. 2006. "Network-Based Marketing: Identifying Likely Adopters via Consumer Networks." *Statistical Science* 21 (2): 256–276. doi:10.1214/088342306000000222.

Hunt, Priscillia, Jessica Saunders, and John S. Hollywood. 2014. *Evaluation of the Shreveport Predictive Policing Experiment*. Santa Monica, CA: Rand Corporation. http://www.rand.org/pubs/research_reports/RR531.

Innes, Martin. 2001. "Control Creep." *Sociological Research Online* 6 (3). https://ideas.repec.org/a/sro/srosro/2001-45-2.html.

Kelleher, John D. 2016. "Fundamentals of Machine Learning for Neural Machine Translation." In *Proceedings of the European Translation Forum*, 1–15. Brussels: European Commission Directorate-General for Translation. https://tinyurl.com/RecurrentNeuralNetworks.

Kelleher, John D., Brian Mac Namee, and Aoife D'Arcy. 2015. *Fundamentals of Machine Learning for Predictive Data Analytics*. Cambridge, MA: MIT Press.

Kerr, Aphra. 2017. *Global Games: Production, Circulation, and Policy in the Networked Era*. New York: Routledge.

Kitchin, Rob. 2014a. *The Data Revolution: Big Data, Open Data, Data Infrastructures, and Their Consequences*. Los Angeles: Sage.

Kitchin, Rob. 2014b. "The Real-Time City? Big Data and Smart Urbanism." *GeoJournal* 79 (1): 1–14. doi:10.1007/s10708-013-9516-8.

Koops, Bert-Jaap. 2011. "Forgetting Footprints, Shunning Shadows: A Critical Analysis of the 'Right to Be Forgotten' in Big Data Practice." Tilburg Law School Legal Studies Research Paper no. 08/2012. *SCRIPTed* 8 (3): 229–56. doi:10.2139/ssrn.1986719.

Korzybski, Alfred. 1996. “On Structure.” In *Science and Sanity: An Introduction to Non-Aristotelian Systems and General Semantics*, CD-ROM, ed. Charlotte Schuchardt-Read. Englewood, NJ: Institute of General Semantics. http://esgs.free.fr/uk/art/sands.htm.

Kosinski, Michal, David Stillwell, and Thore Graepel. 2013. “Private Traits and Attributes Are Predictable from Digital Records of Human Behavior.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (15): 5802–5805. doi:10.1073/pnas.1218772110.

Le Cun, Yann. 1989. *Generalization and Network Design Strategies*. Technical Report CRG-TR-89-4. Toronto: University of Toronto Connectionist Research Group.

Levitt, Steven D., and Stephen J. Dubner. 2009. *Freakonomics: A Rogue Economist Explores the Hidden Side of Everything*. New York: William Morrow Paperbacks.

Lewis, Michael. 2004. *Moneyball: The Art of Winning an Unfair Game*. New York: Norton.

Linoff, Gordon S., and Michael J. A. Berry. 2011. *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*. Indianapolis, IN: Wiley.

Manyika, James, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. 2011. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*. Chicago: McKinsey Global Institute. http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation.

Marr, Bernard. 2015. *Big Data: Using SMART Big Data, Analytics, and Metrics to Make Better Decisions and Improve Performance*. Chichester, UK: Wiley.

Mayer, J. R., and J. C. Mitchell. 2012. “Third-Party Web Tracking: Policy and Technology.” In *2012 IEEE Symposium on Security and Privacy*, 413–27. Piscataway, NJ: IEEE. doi:10.1109/SP.2012.47.

Mayer, Jonathan, and Patrick Mutchler. 2014. “MetaPhone: The Sensitivity of Telephone Metadata.” *Web Policy*, March 12. http://webpolicy.org/2014/03/12/metaphone-the-sensitivity-of-telephone-metadata.

Mayer-Schönberger, Viktor, and Kenneth Cukier. 2014. *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Reprint. Boston: Eamon Dolan/Mariner Books.

McMahan, Brendan, and Daniel Ramage. 2017. “Federated Learning: Collaborative Machine Learning without Centralized Training Data.” *Google Research Blog*, April. https://research.googleblog.com/2017/04/federated-learning-collaborative.html.

Nilsson, Nils. 1965. *Learning Machines: Foundations of Trainable Pattern-Classifying Systems*. New York: McGraw-Hill.

Oakland Privacy Working Group. 2015. “PredPol: An Open Letter to the Oakland City Council.” June 25. https://www.indybay.org/newsitems/2015/06/25/18773987.php.

Organisation for Economic Co-operation and Development (OECD). 1980. *Guidelines on the Protection of Privacy and Transborder Flows of Personal Data.* Paris: OECD. https://www.oecd.org/sti/ieconomy/oecdguidelinesontheprotectionofprivacyandtransborderflowsofpersonaldata.htm.

Organisation for Economic Co-operation and Development (OECD). 2013. *2013 OECD Privacy Guidelines*. Paris: OECD. https://www.oecd.org/internet/ieconomy/privacy-guidelines.htm.

O’Rourke, Cristín, and Aphra Kerr. 2017. “Privacy Shield for Whom? Key Actors and Privacy Discourse on Twitter and in Newspapers.” In “Redesigning or Redefining Privacy?,” special issue of

Westminster Papers in Communication and Culture 12 (3): 21–36. doi:http://doi.org/10.16997/wpcc.264.

Pomerantz, Jeffrey. 2015. *Metadata*. Cambridge, MA: MIT Press. https://mitpress.mit.edu/books/metadata-0.

Purcell, Kristen, Joanna Brenner, and Lee Rainie. 2012. "Search Engine Use 2012." Pew Research Center, March 9. http://www.pewinternet.org/2012/03/09/main-findings-11/.

Quinlan, J. R. 1986. "Induction of Decision Trees." *Machine Learning* 1 (1): 81–106. doi:10.1023/A:1022643204877.

Rainie, Lee, and Mary Madden. 2015. "Americans' Privacy Strategies Post-Snowden." Pew Research Center, March. http://www.pewinternet.org/files/2015/03/PI_AmericansPrivacyStrategies_0316151.pdf.

Rhee, Nissa. 2016. "Study Casts Doubt on Chicago Police's Secretive 'Heat List.'" *Chicago Magazine*, August 17. http://www.chicagomag.com/city-life/August-2016/Chicago-Police-Data/.

Saunders, Jessica, Priscillia Hunt, and John S. Hollywood. 2016. "Predictions Put into Practice: A Quasi-Experimental Evaluation of Chicago's Predictive Policing Pilot." *Journal of Experimental Criminology* 12 (3): 347–371. doi:10.1007/s11292-016-9272-0.

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25 (3): 289–310. doi:10.1214/10-STS330.

Shubber, Kadhim. 2013. "A Simple Guide to GCHQ's Internet Surveillance Programme Tempora." *WIRED UK*, July 24. http://www.wired.co.uk/article/gchq-tempora-101.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, et al. 2016. "Mastering the Game of *Go* with Deep Neural Networks and Tree Search." *Nature* 529 (7587): 484–489. doi:10.1038/nature16961.

Soldatov, Andrei, and Irina Borogan. 2012. “In Ex-Soviet States, Russian Spy Tech Still Watches You.” *WIRED*, December 21. https://www.wired.com/2012/12/russias-hand.

Steinberg, Dan. 2013. “How Much Time Needs to Be Spent Preparing Data for Analysis?” http://info.salford-systems.com/blog/bid/299181/How-Much-Time-Needs-to-be-Spent-Preparing-Data-for-Analysis.

Taylor, David. 2016. “Battle of the Data Science Venn Diagrams.” *KDnuggets*, October. http://www.kdnuggets.com/2016/10/battle-data-science-venn-diagrams.html.

Tufte, Edward R. 2001. *The Visual Display of Quantitative Information*. 2nd ed. Cheshire, CT: Graphics Press.

Turow, Joseph. 2013. *The Daily You: How the New Advertising Industry Is Defining Your Identity and Your Worth*. New Haven, CT: Yale University Press.

Verbeke, Wouter, David Martens, Christophe Mues, and Bart Baesens. 2011. “Building Comprehensible Customer Churn Prediction Models with Advanced Rule Induction Techniques.” *Expert Systems with Applications* 38 (3): 2354–2364.

Weissman, Cale Gutherie. 2015. “The NYPD’s Newest Technology May Be Recording Conversations.” *Business Insider*, March 26. http://uk.businessinsider.com/the-nypds-newest-technology-may-be-recording-conversations-2015-3.

Wolpert, D. H., and W. G. Macready. 1997. “No Free Lunch Theorems for Optimization.” *IEEE Transactions on Evolutionary Computation* 1 (1): 67–82. doi:10.1109/4235.585893.


هنـــداوي