احمد جمال الجسار

علم البيانات والبيانات الضخمة

من النظرية الى التطبيق

9 789922 682051

علم البيانات والبيانات الضخمة
من النظرية الى التطبيق

احمد جمال الجسار

٢٠٢٣

اسم الكتاب: علم البيانات والبيانات الضخمة من النظرية الى التطبيق

اسم المؤلف: احمد جمال الجسار

رقم الإيداع الدولي للكتاب: ردمك ISBN 9789922682051

الناشر: دار الدكتور للعلوم الإدارية والاقتصادية – العراق – بغداد –شارع المتنبي

سنة النشر: ٢٠٢٣

للتواصل مع المؤلف

www.ahmed-aljassar.com

E-MAIL: info@ahmed-aljassar.com

المحتويات

مقدمة الكتاب

أحدث علم البيانات والبيانات الضخمة ثورة في طريقة تحليلنا واستخدامنا للبيانات. مع ظهور التكنولوجيا، كان هناك انفجار في البيانات المتاحة للتحليل، وأصبحت الأدوات والتقنيات لتحليلها أكثر تعقيدًا. يهدف كتاب "علم البيانات والبيانات الضخمة: من النظرية الى التطبيق" إلى تقديم نظرة عامة شاملة عن مجال علم البيانات والبيانات الضخمة وطرق البحث المستخدمة في هذا المجال.

يستخدم الكتاب منهجية علمية لاستكشاف الموضوعات المختلفة في علم البيانات والبيانات الضخمة لاكتساب المعرفة. يتضمن مراقبة البيانات وافتراضها واختبارها وتحليلها لتوليد معرفة جديدة. يتبع الكتاب هذه المنهجية لاستكشاف الجوانب المختلفة لعلم البيانات والبيانات الضخمة.

تبدأ الطريقة العلمية بملاحظة ظاهرة أو مشكلة. في حالة علم البيانات، تكون الملاحظة هي النمو المتسارع للبيانات والحاجة إلى أدوات وتقنيات لفهمها. ثم يطور الكتاب فرضيات لشرح الظاهرة أو لحل المشكلة. على سبيل المثال، يمكن أن تتمثل إحدى الفرضيات في أنه يمكن استخدام خوارزميات التعلم الآلي لتحليل مجموعات البيانات الكبيرة وتحديد الأنماط والاتجاهات. ثم يختبر الكتاب هذه الفرضيات من خلال التجريب وتحليل البيانات. وأخيراً يستخلص الكتاب استنتاجات مبنية على نتائج التجارب والتحليل.

يغطي الكتاب في فصوله العشرون القصيرة والمركزة مجموعة واسعة من الموضوعات المتعلقة بعلوم البيانات والبيانات الضخمة. وتشمل: **مقدمة في علم البيانات والبيانات الضخمة**: يبدأ الكتاب بمقدمة في مجال علم البيانات والبيانات الضخمة. يناقش التطبيقات المختلفة لعلوم البيانات والتحديات والفرص التي تقدمها البيانات الضخمة.

التنقيب في البيانات والتعلم الآلي: يغطي هذا القسم أساسيات التنقيب عن البيانات والتعلم الآلي. يناقش التقنيات المختلفة المستخدمة في هذه المجالات وتطبيقاتها في علم البيانات. **تخزين البيانات ومعالجتها**: يغطي الكتاب تقنيات تخزين البيانات ومعالجتها المختلفة المستخدمة في علم البيانات، مثل Hadoop و Spark.**الحوسبة السحابية:** يناقش هذا القسم أساسيات الحوسبة السحابية وتطبيقاتها في علم البيانات والبيانات الضخمة. **لوائح خصوصية البيانات:** يغطي الكتاب مختلف

لوائح خصوصية البيانات التي تؤثر على علم البيانات والبيانات الضخمة، مثل القانون العام لحماية البيانات (GDPR) وقانون حماية حقوق الملكية الفكرية (HIPAA).**تحليلات العملاء:** يغطي هذا القسم الأدوات والتقنيات المختلفة المستخدمة في تحليلات العملاء، مثل التجزئة والنمذجة التنبؤية. **تحليلات الوسائط الاجتماعية:** يناقش هذا القسم الأدوات والتقنيات المستخدمة في تحليلات الوسائط الاجتماعية، مثل تحليل المشاعر وتحليل الشبكة. **كشف الاحتيال:** يغطي الكتاب التقنيات المختلفة المستخدمة في اكتشاف الاحتيال، مثل اكتشاف الأخطاء والتعلم الآلي. **الاتجاهات الناشئة في علم البيانات والبيانات الضخمة:** يختتم الكتاب بمناقشة الاتجاهات الناشئة في علم البيانات والبيانات الضخمة، مثل الذكاء الاصطناعي وإنترنت الأشياء.

في الختام، يقدم كتاب "علم البيانات والبيانات الضخمة: من النظرية الى التطبيق" نظرة عامة شاملة عن مجال علم البيانات والبيانات الضخمة. يغطي مجموعة واسعة من الموضوعات المتعلقة بعلوم البيانات والبيانات الضخمة، بما في ذلك التنقيب عن البيانات، والتعلم الآلي، والحوسبة السحابية، وتحليلات العملاء. يستخدم الكتاب منهجًا علميًا لاستكشاف هذه الموضوعات، بدءًا من الملاحظة وتطوير الفرضيات، والاختبار من خلال التجريب والتحليل، وانتهاءً باستخلاص النتائج. يعد الكتاب مصدرًا قيمًا لأي شخص مهتم بعلوم البيانات والبيانات الضخمة، من المبتدئين إلى المحترفين ذوي الخبرة.

الفصل الأول :مقدمة في علم البيانات والبيانات الضخمة

الفصل الأول :مقدمة في علم البيانات والبيانات الضخمة

أ- تعريف ومفهوم علم البيانات والبيانات الضخمة

علم البيانات هو مجال متعدد التخصصات يتضمن استخدام الأساليب الإحصائية والحسابية لاستخراج الأفكار والمعرفة من البيانات .وهي تشمل مجموعة واسعة من التقنيات والأدوات، بما في ذلك الإحصائيات والتعلم الآلي وتصور البيانات وإدارة قواعد البيانات.

تشير البيانات الضخمة إلى مجموعات بيانات كبيرة ومعقدة للغاية لا يمكن معالجتها أو تحليلها بسهولة باستخدام أدوات وتقنيات معالجة البيانات التقليدية .تتميز مجموعات البيانات هذه عادةً بحجمها وسرعتها وتنوعها وصحتها .غالبًا ما يتم إنشاء البيانات الضخمة عن طريق إنترنت الأشياء، ووسائل التواصل الاجتماعي، ومصادر أخرى، وتتطلب منصات وأدوات متخصصة لإدارتها وتحليلها.

ظهر مفهوم علم البيانات والبيانات الضخمة نتيجة انفجار البيانات في العصر الرقمي .مع استمرار نمو كمية البيانات التي ينتجها الأفراد والمؤسسات والحكومات، هناك حاجة متزايدة للمهنيين الذين يمكنهم إدارة هذه البيانات وتحليلها وتفسيرها بشكل فعال لاكتساب رؤى وتوجيه عملية صنع القرار.

تُستخدم علوم البيانات والبيانات الضخمة في مجموعة متنوعة من المجالات، بما في ذلك الأعمال التجارية والرعاية الصحية والحكومة والأوساط الأكاديمية لتحسين الكفاءة ودفع الابتكار وحل المشكلات المعقدة . من أجل تحقيق النجاح في مجال علم البيانات، يجب أن يكون لدى المحترفين أساس قوي في الإحصاء والبرمجة وتحليل البيانات، بالإضافة إلى فهم أحدث الأدوات والتقنيات لإدارة البيانات الضخمة وتحليلها.

ب-السياق التاريخي لعلوم البيانات والبيانات الضخمة

يمكن إرجاع مفهوم علم البيانات والبيانات الضخمة إلى الأيام الأولى للحوسبة وإدارة البيانات .في الستينيات والسبعينيات من القرن الماضي، بدأت الشركات والمؤسسات في استخدام أجهزة الكمبيوتر المركزية لتخزين ومعالجة كميات كبيرة من البيانات .ومع ذلك، كانت هذه الأنظمة المبكرة محدودة في قدراتها وتتطلب معرفة متخصصة لاستخدامها بفعالية.

في الثمانينيات والتسعينيات، سمح تطوير قواعد البيانات العلائقية وتوحيد(SQL) لغة الاستعلام الهيكلية بإدارة بيانات أكثر كفاءة ومرونة .أدى ذلك إلى ظهور تخزين البيانات، والذي تضمن دمج البيانات من مصادر متعددة في مستودع واحد للتحليل وإعداد التقارير.

أدى ظهور الإنترنت في التسعينيات وأوائل القرن الحادي والعشرين إلى انتشار مصادر البيانات، بما في ذلك وسائل التواصل الاجتماعي والتجارة الإلكترونية ومحركات البحث .خلق هذا تحديًا جديدًا للشركات والمؤسسات، حيث لم تكن تقنيات إدارة البيانات التقليدية قادرة على التعامل مع الحجم الهائل ومجموعة متنوعة من البيانات التي يتم إنشاؤها.

تمت صياغة مصطلح" البيانات الضخمة "لأول مرة في أوائل العقد الأول من القرن الحادي والعشرين لوصف هذه الظاهرة الجديدة .في عام 2008 ، اكتسب المصطلح اعترافًا سائدًا بنشر تقرير من شركة الاستشارات McKinsey & Company ، والذي حدد البيانات الضخمة كمحرك رئيسي لنمو الأعمال والابتكار.

في نفس الوقت تقريبًا، بدأت التطورات في التعلم الآلي وتحليلات البيانات في تمكين رؤى واكتشافات جديدة من مجموعات البيانات الكبيرة .أدى ذلك إلى تطوير علم البيانات كمجال للدراسة، والذي يجمع بين الأساليب الإحصائية والحسابية والخبرة الخاصة بالمجال لاستخراج الأفكار والمعرفة من البيانات.

اليوم، أصبح علم البيانات والبيانات الضخمة مكونات مهمة للعديد من الصناعات، بما في ذلك التمويل والرعاية الصحية والحكومة .مع استمرار تزايد حجم البيانات وتعقيدها، من المرجح أن تستمر أهمية علم البيانات والبيانات الضخمة في الزيادة، مما يؤدي إلى الابتكار ويشكل مستقبل الأعمال والمجتمع.

ج-تأثير علم البيانات والبيانات الضخمة على الأعمال والمجتمع

كان تأثير علم البيانات والبيانات الضخمة على الشركات والمجتمع كبيرًا وبعيد المدى .في عالم الأعمال، أحدث علم البيانات ثورة في طريقة عمل الشركات، مما مكنها من اتخاذ قرارات تعتمد على البيانات واكتساب ميزة تنافسية .من خلال الاستفادة من البيانات الضخمة والتحليلات المتقدمة، يمكن للشركات تحسين عملياتها وتحسين مشاركة العملاء وتطوير منتجات وخدمات جديدة.

أحد المجالات الرئيسية التي كان لعلم البيانات فيها تأثير كبير هو مجال التسويق والإعلان .مع وفرة البيانات المتاحة من مصادر مثل وسائل التواصل الاجتماعي وتحليلات الويب، يمكن للشركات اكتساب رؤى حول سلوك المستهلك وتفضيلاته، مما يسمح لهم بتطوير حملات تسويقية مستهدفة وشخصية.

في صناعة الرعاية الصحية، يتمتع علم البيانات بالقدرة على تحسين نتائج المرضى وخفض التكاليف من خلال تمكين تشخيصات أكثر دقة وعلاجات مخصصة .من خلال تحليل مجموعات البيانات الكبيرة من السجلات الصحية الإلكترونية والمصادر الأخرى، يمكن لعلماء البيانات تحديد الأنماط والرؤى التي يمكن أن تساعد في اتخاذ القرارات السريرية وتؤدي إلى نتائج صحية أفضل.

خارج عالم الأعمال، تعمل علوم البيانات والبيانات الضخمة أيضًا على تحويل المجتمع ككل .في مجال السياسة العامة، تستخدم الحكومات البيانات لتوجيه عملية صنع القرار وتطوير السياسات القائمة على الأدلة .على سبيل المثال، تم استخدام تحليل البيانات لمعالجة قضايا مثل الفقر والجريمة والاستدامة البيئية.

ومع ذلك، هناك أيضًا مخاوف بشأن تأثير علم البيانات والبيانات الضخمة على الخصوصية والأمان .مع تزايد حجم البيانات التي يتم إنشاؤها وتنوعها، هناك خطر يتمثل في إساءة استخدام المعلومات الشخصية أو اختراقها .لذلك من المهم وضع أطر أخلاقية وتنظيمية واضحة لاستخدام علم البيانات والبيانات الضخمة، لضمان موازنة فوائدها مع المخاطر المحتملة.

بشكل عام، يعد تأثير علم البيانات والبيانات الضخمة على الشركات والمجتمع معقدًا ومتعدد الأوجه .في حين أن هناك بالتأكيد تحديات ومخاطر مرتبطة بهذه التقنيات، لا يمكن تجاهل قدرتها على دفع الابتكار وتحسين الكفاءة وحل المشكلات المعقدة .على هذا النحو، من المهم للأفراد والمنظمات الاستمرار في استكشاف إمكانيات علم البيانات والبيانات الضخمة، مع الانتباه أيضًا لآثارها الاجتماعية والأخلاقية الأوسع.

الفصل الثاني :جمع البيانات والمعالجة المسبقة

الفصل الثاني: جمع البيانات والمعالجة المسبقة

أ- مصادر البيانات وأنواعها

في مجال علم البيانات والبيانات الضخمة، تعد مصادر البيانات وأنواعها مكونات حاسمة تشكل الأساس لجميع أنشطة التحليل والنمذجة .تشير مصادر البيانات إلى الأماكن المختلفة حيث يمكن الحصول على البيانات، بينما تشير أنواع البيانات إلى الفئات أو التصنيفات المختلفة للبيانات الموجودة.

يمكن تصنيف مصادر البيانات على نطاق واسع إلى مصادر داخلية وخارجية .تشير المصادر الداخلية إلى البيانات التي تم إنشاؤها بواسطة مؤسسة أو عمل تجاري معين، مثل بيانات المعاملات وبيانات العملاء وبيانات الموظف .من ناحية أخرى، تشير المصادر الخارجية إلى البيانات التي تم الحصول عليها من مصادر خارج المنظمة، مثل وسائل التواصل الاجتماعي وتقارير أبحاث السوق والبيانات الحكومية.

يمكن تصنيف أنواع البيانات على نطاق واسع إلى بيانات منظمة وشبه منظمة وغير منظمة .تشير البيانات المهيكلة إلى البيانات التي تم تنظيمها بتنسيق محدد مسبقًا وموحد، مثل البيانات المخزنة في قاعدة بيانات علائقية أو جدول بيانات .تشير البيانات شبه المنظمة إلى البيانات المنظمة جزئيًا ولكنها لا تتوافق مع بنية صارمة، مثل البيانات المخزنة في تنسيقات XML أو JSON تشير البيانات غير المنظمة إلى البيانات غير المنظمة بأي طريقة محددة، مثل البيانات النصية من الوسائط الاجتماعية أو ملفات الصوت والفيديو.

تشمل الأنواع الأخرى من البيانات بيانات السلاسل الزمنية، والتي تشير إلى البيانات التي يتم جمعها بمرور الوقت وغالبًا ما تُستخدم للتنبؤ والنمذجة التنبؤية، والبيانات الجغرافية المكانية، التي تشير إلى البيانات المرتبطة بموقع جغرافي محدد، مثل إحداثيات GPS أو صور القمر الصناعي.

بالإضافة إلى هذه الفئات، يمكن أيضًا تصنيف البيانات بناءً على مستوى قياسها .تتضمن مستويات القياس الشائعة الاسمي والترتيبي والفاصل الزمني والنسبة .تشير البيانات الاسمية إلى البيانات غير الرقمية ولا يمكن ترتيبها، مثل البيانات الفئوية .تشير البيانات الترتيبية إلى البيانات التي يمكن طلبها ولكن ليس لها مقياس قياس ثابت، مثل مقياس ليكرت .تشير بيانات الفاصل الزمني إلى البيانات التي تحتوي على مقياس قياس ثابت ولكن لا توجد نقطة صفر حقيقية، مثل قياسات درجة الحرارة بالدرجة المئوية أو فهرنهايت . تشير بيانات النسبة إلى البيانات التي تحتوي على مقياس قياس ثابت ونقطة صفر حقيقية، مثل قياسات الوزن أو الارتفاع.

بشكل عام، يعد فهم مصادر وأنواع البيانات المختلفة أمرًا بالغ الأهمية لعلماء ومحللي البيانات، لأنه يمكنهم من اختيار الأساليب والأدوات المناسبة لمعالجة البيانات وتحليلها وتفسيرها.

ب- تقنيات المعالجة المسبقة للبيانات

تعد المعالجة المسبقة للبيانات خطوة أساسية في علم البيانات وتحليل البيانات الضخمة، لأنها تتضمن تحويل البيانات الأولية إلى تنسيق مناسب للتحليل .تشير تقنيات المعالجة المسبقة للبيانات إلى الطرق المختلفة المستخدمة لإعداد وتنظيف البيانات للتحليل، والتأكد من أنها دقيقة وكاملة ومتسقة.

يعد تنظيف البيانات أحد الأساليب الأساسية للمعالجة المسبقة للبيانات، والذي يتضمن تحديد وتصحيح الأخطاء أو التناقضات في البيانات .يمكن أن يتضمن ذلك إزالة نقاط البيانات المكررة، وتصحيح الأخطاء الإملائية أو المطبعية، ومعالجة القيم المفقودة .يمكن معالجة القيم المفقودة باستخدام تقنيات مثل التضمين، حيث يتم استبدال القيم المفقودة بقيم مقدرة بناءً على البيانات المتاحة.

هناك أسلوب آخر مهم للمعالجة المسبقة للبيانات وهو تكامل البيانات، والذي يتضمن دمج البيانات من مصادر متعددة في مجموعة بيانات واحدة .يمكن أن تكون هذه عملية معقدة، حيث قد تستخدم البيانات من مصادر مختلفة تنسيقات أو هياكل أو وحدات قياس مختلفة .يمكن أن يتضمن تكامل البيانات تقنيات مثل دمج البيانات، حيث يتم دمج البيانات من مصادر مختلفة لإنشاء مجموعة بيانات أكثر اكتمالاً ودقة.

يعد تحويل البيانات تقنية أخرى مهمة للمعالجة المسبقة للبيانات، والتي تتضمن تحويل البيانات إلى تنسيق أكثر ملاءمة للتحليل .يمكن أن يتضمن ذلك تقنيات مثل التسوية، والتي تقيس البيانات إلى نطاق مشترك، أو هندسة الميزات، والتي تتضمن إنشاء ميزات جديدة بناءً على البيانات الموجودة.

يعد تقليل البيانات أيضًا أسلوبًا مهمًا للمعالجة المسبقة للبيانات، حيث إنه ينطوي على تقليل كمية البيانات المراد تحليلها مع الاحتفاظ بالمعلومات المهمة .يمكن أن يتضمن ذلك تقنيات مثل تقليل الأبعاد، مما يقلل من عدد المتغيرات في مجموعة البيانات، أو أخذ العينات، الذي يحدد مجموعة فرعية من البيانات للتحليل.

أخيرًا، يعتبر تحديد البيانات تقنية تتضمن تحويل البيانات المستمرة إلى فئات أو صناديق منفصلة .يمكن أن يكون هذا مفيدًا لتبسيط تحليل البيانات وجعله أكثر قابلية للتفسير، بالإضافة إلى تقليل التعقيد الحسابي للخوارزميات.

بشكل عام، تعد تقنيات المعالجة المسبقة للبيانات ضرورية لضمان دقة البيانات واكتمالها واتساقها، ولإعدادها للتحليل .من خلال اختيار تقنيات المعالجة المسبقة للبيانات المناسبة، يمكن لعلماء ومحللي البيانات تحسين جودة وموثوقية تحليلاتهم، مما يؤدي إلى رؤى أكثر دقة وفائدة.

ج-تنظيف البيانات وتحويلها

يعد تنظيف البيانات وتحويلها جانبين مهمين للمعالجة المسبقة للبيانات في علم البيانات وتحليل البيانات الضخمة .يتضمن تنظيف البيانات تحديد وتصحيح الأخطاء أو التناقضات في البيانات، بينما يتضمن تحويل البيانات تحويل البيانات إلى تنسيق مناسب للتحليل.

يعد تنظيف البيانات خطوة حاسمة في خط أنابيب المعالجة المسبقة للبيانات، حيث إنه يضمن أن البيانات دقيقة وكاملة ومتسقة .يمكن أن يتضمن تنظيف البيانات تقنيات مثل إزالة نقاط البيانات المكررة، وتصحيح الأخطاء الإملائية أو المطبعية، ومعالجة القيم المفقودة .يمكن معالجة القيم المفقودة باستخدام تقنيات مثل التضمين، حيث يتم استبدال القيم المفقودة بقيم مقدرة بناءً على البيانات المتاحة .تتضمن تقنيات تنظيف البيانات الأخرى الكشف الخارجي، حيث يتم تحديد نقاط البيانات التي تختلف اختلافًا كبيرًا عن بقية مجموعة البيانات وإما إزالتها أو تصحيحها.

يتضمن تحويل البيانات تحويل البيانات إلى تنسيق مناسب للتحليل .يمكن أن يتضمن ذلك تقنيات مثل التسوية، والتي تقيس البيانات إلى نطاق مشترك، أو هندسة الميزات، والتي تتضمن إنشاء ميزات جديدة بناءً على البيانات الموجودة .التسوية هي تقنية تُستخدم غالبًا لإحضار البيانات إلى مقياس مشترك، وهو أمر مفيد لمقارنة نقاط البيانات التي تستخدم وحدات قياس مختلفة .تتضمن هندسة الميزات إنشاء ميزات جديدة مشتقة من البيانات الموجودة، مثل حساب المتوسط أو الحد الأقصى لمجموعة من القيم.

جانب آخر مهم لتحويل البيانات هو تشفير البيانات، والذي يتضمن تحويل البيانات الفئوية إلى تنسيق رقمي يمكن تحليله بواسطة خوارزميات التعلم الآلي .يمكن أن يتضمن ذلك تقنيات مثل التشفير، حيث يتم تحويل كل قيمة محتملة لمتغير فئوي إلى متغير ثنائي منفصل.

بشكل عام، يعد تنظيف البيانات وتحويلها خطوات أساسية في المعالجة المسبقة للبيانات، لأنها تضمن أن البيانات دقيقة وكاملة ومتسقة، وأنها في تنسيق مناسب للتحليل .من خلال اختيار تقنيات تنظيف وتحويل البيانات المناسبة، يمكن لعلماء ومحللي البيانات تحسين جودة وموثوقية تحليلاتهم، مما يؤدي إلى رؤى أكثر دقة وفائدة.

الفصل الثالث :تصور البيانات واستكشافها

الفصل الثالث: تصور البيانات واستكشافها

أ-تقنيات تصور البيانات

يعد تصور البيانات جانبًا أساسيًا في علم البيانات وتحليل البيانات الضخمة، حيث يتيح للمحللين استكشاف البيانات المعقدة ونقلها بتنسيق أكثر سهولة ويسهل الوصول إليه .تشير تقنيات تصور البيانات إلى الطرق المختلفة المستخدمة لتمثيل البيانات بيانياً، باستخدام المخططات والرسوم البيانية والتصورات الأخرى.

واحدة من أكثر تقنيات تصور البيانات شيوعًا هي استخدام المخططات الشريطية والرسوم البيانية لتمثيل البيانات الرقمية .تعد المخططات الشريطية مفيدة لمقارنة القيم عبر فئات مختلفة، بينما تُستخدم الرسوم البيانية لتصور توزيع متغير واحد.

أسلوب آخر شائع لتصور البيانات هو استخدام المخططات الخطية لتمثيل الاتجاهات بمرور الوقت .تعد المخططات الخطية مفيدة لتصور كيفية تغير متغير خلال فترة مستمرة، مثل البيانات اليومية أو الأسبوعية أو الشهرية.

المخططات المبعثرة هي تقنية أخرى مهمة لتصور البيانات، تُستخدم لتصور العلاقة بين متغيرين .تعد المخططات المبعثرة مفيدة في تحديد الأنماط أو الارتباطات بين المتغيرات، مثل ما إذا كانت هناك علاقة موجبة أو سلبية بين متغيرين.

الخرائط الحرارية هي تقنية أخرى قوية لتصور البيانات، والتي تستخدم لتمثيل البيانات كمصفوفة ثنائية الأبعاد من الخلايا الملونة .تعد خرائط الحرارة مفيدة لتصور مجموعات البيانات الكبيرة أو العلاقات المعقدة بين المتغيرات.

تتضمن تقنيات تصور البيانات الأخرى المخططات الدائرية، والتي تُستخدم لتمثيل نسبة الفئات المختلفة في مجموعة البيانات، والخرائط، والتي تُستخدم لتصور البيانات المكانية.

بشكل عام، يعد تصور البيانات جانبًا مهمًا لعلوم البيانات وتحليل البيانات الضخمة، حيث يتيح للمحللين استكشاف البيانات المعقدة ونقلها بتنسيق أكثر سهولة ويسهل الوصول إليه .من خلال اختيار تقنيات تصور البيانات المناسبة، يمكن للمحللين تحسين وضوح وتأثير تحليلاتهم، ونقل نتائجهم بشكل فعال إلى جمهور عريض.

ب-تحليل البيانات الاستكشافية

يعد تحليل البيانات الاستكشافية (EDA) جانبًا مهمًا في علم البيانات وتحليل البيانات الضخمة، حيث يتضمن استخدام تقنيات إحصائية وتصور لاستكشاف مجموعة بيانات وفهمها .يتم تنفيذ EDA عادةً في المراحل الأولى من خط أنابيب تحليل البيانات، لتحديد الأنماط والاتجاهات والعلاقات في البيانات، ولإبلاغ تطوير الفرضيات أو النماذج.

الهدف الأساسي من تحليل البيانات الاستكشافية هو اكتساب فهم لبنية وخصائص البيانات، وتحديد أي قيم متطرفة أو قيم مفقودة أو غيرها من الحالات الشاذة التي قد تؤثر على التحليل .يتضمن EDA عادةً استخدام الإحصائيات الوصفية، مثل المتوسط والمتوسط والانحراف المعياري، لتلخيص البيانات والتصورات مثل الرسوم البيانية ومخططات الصندوق ومخططات التشتت، لاستكشاف العلاقات بين المتغيرات.

يتمثل أحد الأساليب الشائعة في تحليل البيانات الاستكشافية استخدام سلسلة من التصورات والاختبارات الإحصائية لتحسين فهمنا للبيانات تدريجيًا .على سبيل المثال، قد نبدأ بتصور توزيع متغير واحد باستخدام الرسم البياني أو مخطط الكثافة، ثم ننتقل إلى تصور العلاقة بين متغيرين باستخدام مخطط مبعثر أو مصفوفة الارتباط .يمكننا بعد ذلك استخدام الاختبارات الإحصائية، مثل اختبار الفرضيات أو تحليل الانحدار، لتحديد قوة وأهمية هذه العلاقات.

يتمثل أحد الجوانب الرئيسية الأخرى لتحليل البيانات الاستكشافية في تحديد التحيزات المحتملة أو العوامل المربكة التي قد تؤثر على التحليل .على سبيل المثال، قد نتحرى ما إذا كانت هناك اختلافات كبيرة بين المجموعات أو المجموعات الفرعية في البيانات، وما إذا كان يمكن تفسير هذه الاختلافات من خلال متغيرات أخرى.

بشكل عام، يعد تحليل البيانات الاستكشافية خطوة حاسمة في خط أنابيب تحليل البيانات، لأنها تتيح للمحللين اكتساب فهم عميق للبيانات، وتحديد الأنماط والعلاقات التي قد تساعد في تطوير الفرضيات أو النماذج . باستخدام مجموعة من التصورات والاختبارات الإحصائية، يمكن للمحللين استكشاف مجموعات البيانات المعقدة وتطوير فهم دقيق للعلاقات بين المتغيرات، مما يؤدي إلى رؤى أكثر دقة ومفيدة.

ج-سرد قصص البيانات

يعد سرد قصص البيانات مجالًا ناشئًا في علم البيانات وتحليل البيانات الضخمة، والذي يتضمن استخدام السرد والتصورات وتقنيات الاتصال الأخرى لنقل الرؤى والنتائج من البيانات إلى جمهور عريض .يهدف سرد قصص البيانات إلى جعل البيانات أكثر سهولة في الوصول والمشاركة، ومساعدة أصحاب المصلحة على فهم الأفكار المستمدة من البيانات والتصرف بناءً عليها.

تشمل العناصر الأساسية لسرد البيانات تحديد الجمهور واحتياجاتهم، واختيار البيانات والتصورات المناسبة لنقل الرسالة، وإنشاء سرد مقنع يوجه الجمهور من خلال البيانات .يتضمن سرد القصص الفعال للبيانات أيضًا استخدام لغة واضحة وموجزة، وتجنب المصطلحات الفنية أو التصورات شديدة التعقيد التي قد تربك الجمهور أو تطغى عليه.

تتمثل إحدى طرق سرد القصص في البيانات في استخدام هيكل سردي لتوجيه الجمهور من خلال البيانات . على سبيل المثال، قد نبدأ بتقديم المشكلة أو السؤال الذي تهدف البيانات إلى معالجته، ثم نستخدم التصورات والتقنيات الأخرى لاستكشاف جوانب مختلفة من البيانات، مثل الاتجاهات أو الأنماط أو القيم المتطرفة . يمكننا بعد ذلك استخدام هذه الأفكار لتطوير استنتاج أو توصية، وإيصال ذلك بطريقة واضحة وقابلة للتنفيذ.

جانب رئيسي آخر لسرد البيانات هو استخدام التصورات وتقنيات الاتصال الأخرى لتعزيز تأثير البيانات . على سبيل المثال، قد نستخدم الرسوم البيانية أو لوحات المعلومات التفاعلية أو مقاطع الفيديو لنقل الأفكار من البيانات ولإشراك الجمهور بطريقة أكثر تفاعلية وغامرة.

بشكل عام، يعد سرد قصص البيانات جانبًا مهمًا من علم البيانات وتحليل البيانات الضخمة، حيث يتيح للمحللين توصيل البيانات المعقدة بطريقة أكثر سهولة وجاذبية، ولمساعدة أصحاب المصلحة على فهم الأفكار المستمدة من البيانات والتصرف بناءً عليها .باستخدام مجموعة من الروايات والتصورات وتقنيات الاتصال الأخرى، يمكن للمحللين إنشاء قصص بيانات مقنعة ومؤثرة تدفع العمل والتأثير.

الفصل الرابع :التحليل الاحصائي

الفصل الرابع :التحليل الإحصائي

أ-أساسيات الاحتمالية والإحصاء

الاحتمالية والإحصاء هي مفاهيم أساسية في علم البيانات وتحليل البيانات الضخمة .الاحتمال هو فرع الرياضيات الذي يتعامل مع دراسة الأحداث العشوائية واحتمالية حدوثها، بينما الإحصاء هو علم جمع البيانات وتحليلها وتفسيرها.

في نظرية الاحتمالات، يتم تعريف الأحداث على أنها مجموعات من النتائج المحتملة للتجربة .احتمال وقوع حدث هو رقم بين 0 و 1 يمثل احتمال وقوع هذا الحدث .يمكن استخدام الاحتمالية لنمذجة الأحداث غير المؤكدة، مثل تقلبات العملات أو أنماط الطقس، ولتكوين تنبؤات حول احتمالية النتائج المستقبلية.

الإحصاء، من ناحية أخرى، يهتم بجمع البيانات وتحليلها وتفسيرها .يمكن جمع البيانات من خلال المسوحات أو التجارب أو الملاحظات، ويمكن تحليلها باستخدام مجموعة متنوعة من الأساليب الإحصائية . تُستخدم الإحصائيات الوصفية لتلخيص ووصف السمات الرئيسية لمجموعة البيانات، مثل المتوسط والوسيط والانحراف المعياري .من ناحية أخرى، تُستخدم الإحصائيات الاستدلالية لعمل تنبؤات أو استخلاص استنتاجات حول مجتمع بناءً على عينة من البيانات.

تتضمن بعض المفاهيم الأساسية في الاحتمالية والإحصاء توزيعات الاحتمالات واختبار الفرضيات وتحليل الانحدار .التوزيعات الاحتمالية هي وظائف رياضية تصف احتمالية النتائج المختلفة للتجربة .اختبار الفرضيات هو أسلوب إحصائي يستخدم لتحديد ما إذا كان من المحتمل أن تكون فرضية حول مجتمع ما صحيحة بناءً على عينة من البيانات .تحليل الانحدار هو أسلوب إحصائي يستخدم لنمذجة العلاقة بين واحد أو أكثر من المتغيرات المستقلة والمتغير التابع.

في علم البيانات وتحليل البيانات الضخمة، يتم استخدام الاحتمالات والإحصاءات لفهم وتحليل مجموعات البيانات الكبيرة، ولإجراء التنبؤات والقرارات بناءً على تلك البيانات .من خلال تطبيق الاحتمالات والإحصاءات على البيانات، يمكن للمحللين اكتساب رؤى حول الأنماط والاتجاهات في البيانات، واستخدام هذه المعلومات لاتخاذ قرارات مستنيرة ودفع نتائج الأعمال.

ب-الاستدلال الإحصائي

الاستدلال الإحصائي هو عملية استخدام بيانات العينة لعمل استنتاجات أو تنبؤات حول مجموعة سكانية . إنه جانب رئيسي من جوانب تحليل البيانات ويلعب دورًا مهمًا في علم البيانات والبيانات الضخمة.

في الاستدلال الإحصائي، نبدأ بفرضية حول المجتمع ونستخدم بيانات العينة لاختبار ما إذا كان من المحتمل أن تكون هذه الفرضية صحيحة .تتضمن العملية وضع افتراضات حول التوزيع الاحتمالي للبيانات، واستخدام الاختبارات الإحصائية لتحديد احتمالية مراقبة بيانات العينة إذا كانت الفرضية صحيحة.

هناك نوعان رئيسيان من الاستدلال الإحصائي :الاستدلال البارامترى والاستدلال غير البارامترى . يفترض الاستدلال البارامتي أن البيانات تتبع توزيع احتمالي محدد، مثل التوزيع الطبيعي، وتستخدم الاختبارات الإحصائية بناءً على هذا الافتراض .من ناحية أخرى، لا يقدم الاستدلال غير المعياري أي افتراضات حول التوزيع الاحتمالي للبيانات ويستخدم الاختبارات الإحصائية التي لا تعتمد على تلك الافتراضات.

تتضمن بعض تقنيات الاستدلال الإحصائي الشائعة المستخدمة في علم البيانات وتحليل البيانات الضخمة اختبار الفرضيات وفترات الثقة والقيم .p يتضمن اختبار الفرضية إعداد فرضية العدم، والتي تمثل الوضع الراهن أو الافتراض الذي يتم اختباره، وفرضية بديلة، والتي تمثل إمكانية وجود شيء جديد أو مختلف . ثم تُستخدم الاختبارات الإحصائية لتحديد ما إذا كانت البيانات المرصودة توفر أدلة كافية لرفض فرضية العدم لصالح الفرضية البديلة.

فترات الثقة هي مجموعة من القيم التي من المحتمل أن تحتوي على معلمة السكان الحقيقية بمستوى معين من الثقة .على سبيل المثال، فإن حد الثقة95 ٪ لوسط المجتمع يعني أنه إذا كنا سنكرر عملية أخذ العينات عدة مرات، فإن95 ٪ من فترات الثقة التي تم إنشاؤها ستحتوي على الوسط الحقيقي للمحتوى.

قيم P هي مقياس لقوة الدليل ضد فرضية العدم .أنها تمثل احتمال مراقبة بيانات العينة، أو البيانات أكثر تطرفًا من بيانات العينة، إذا كانت الفرضية الصفرية صحيحة .تشير قيمة p الصغيرة إلى أنه من غير المحتمل أن تكون البيانات المرصودة قد حدثت بالصدفة وحدها، وتقدم دليلًا لصالح الفرضية البديلة.

في الختام، يعد الاستدلال الإحصائي أداة قوية تسمح لنا بعمل تنبؤات واستخلاص استنتاجات حول السكان بناءً على بيانات العينة .إنه مكون أساسي في علم البيانات وتحليل البيانات الضخمة، ويتم استخدامه لاكتساب نظرة ثاقبة لأنماط واتجاهات البيانات، ولاتخاذ قرارات مستنيرة بناءً على تلك البيانات.

ج-اختبار الفرضيات وفترات الثقة

اختبار الفرضيات وفترات الثقة هما طريقتان إحصائيتان شائعتان تستخدمان في علم البيانات وتحليل البيانات الضخمة .كلاهما يعتمد على مبادئ الاستدلال الإحصائي، والذي يتضمن استخدام بيانات العينة لاستخلاص استنتاجات حول السكان.

اختبار الفرضيات هو أسلوب إحصائي يتضمن تقديم مطالبة، تسمى فرضية، حول معلمة مجتمع، مثل المتوسط أو الانحراف المعياري .الفرضية الصفرية هي بيان أنه لا يوجد فرق بين معلمة المجتمع وقيمة محددة، في حين أن الفرضية البديلة هي بيان بأن هناك فرقًا بين معلمة المجتمع وقيمة محددة.

لاختبار الفرضية، نستخدم بيانات نموذجية واختبارات إحصائية لتحديد ما إذا كان هناك دليل كاف لرفض الفرضية الصفرية لصالح الفرضية البديلة .مستوى الأهمية، أو مستوى ألفا، هو احتمال رفض الفرضية الصفرية عندما تكون صحيحة بالفعل .القيمة p هي احتمال ملاحظة عينة إحصائية متطرفة مثل أو أكثر تطرفًا من تلك الملاحظة، بافتراض صحة الفرضية الصفرية .إذا كانت القيمة p أقل من مستوى ألفا، فإننا نرفض الفرضية الصفرية لصالح الفرضية البديلة.

من ناحية أخرى، فإن فترات الثقة هي مجموعة من القيم التي من المحتمل أن تحتوي على معلمة السكان الحقيقية بمستوى معين من الثقة .يتم إنشاء فترات الثقة باستخدام بيانات العينة والحسابات الإحصائية . مستوى الثقة هو النسبة المئوية لمرات احتواء الفاصل الزمني على معلمة السكان الحقيقية إذا أردنا تكرار عملية أخذ العينات عدة مرات.

على سبيل المثال، لنفترض أننا نريد تقدير متوسط الراتب لجميع الموظفين في الشركة .نأخذ عينة عشوائية من الموظفين ونحسب متوسط العينة والانحراف المعياري .يمكننا بعد ذلك إنشاء فاصل ثقة لمتوسط المجتمع باستخدام بيانات العينة ومستوى معين من الثقة، مثل95 ٪ .يزودنا فاصل الثقة بمجموعة من القيم التي يمكننا أن نكون واثقين من أنها تحتوي على متوسط المحتوى الحقيقي.

في الختام، يعد اختبار الفرضيات وفترات الثقة طريقتين إحصائيتين مهمتين تستخدمان في علم البيانات وتحليل البيانات الضخمة .يتيح لنا اختبار الفرضيات اتخاذ قرارات بشأن معلمات المجتمع بناءً على بيانات العينة والاختبارات الإحصائية، بينما تزودنا فترات الثقة بمجموعة من القيم التي من المحتمل أن تحتوي

على معلمة السكان الحقيقية بمستوى معين من الثقة .تتيح لنا هذه التقنيات استخلاص رؤى ذات مغزى واتخاذ قرارات مستنيرة بناءً على البيانات.

الفصل الخامس: اساسيات تعلم الالة

الفصل الخامس :أساسيات تعلم الآلة

أ-مقدمة في التعلم الآلي

التعلم الآلي هو حقل فرعي من الذكاء الاصطناعي يتضمن استخدام التقنيات الإحصائية لتمكين أنظمة الكمبيوتر من التعلم من البيانات وتحسين أدائها في مهمة محددة بمرور الوقت .يمكن استخدام خوارزميات التعلم الآلي لتحليل مجموعات البيانات الكبيرة، وتحديد الأنماط، وعمل التنبؤات أو القرارات بناءً على تلك الأنماط.

الهدف من التعلم الآلي هو تطوير نماذج يمكنها التعلم تلقائيًا من البيانات دون أن تتم برمجتها بشكل صريح . يتم تحقيق ذلك من خلال تدريب النموذج على مجموعة بيانات، والتي تتكون من متغيرات الإدخال، تسمى الميزات، ومتغير الإخراج، يسمى المتغير المستهدف .يستخدم النموذج بعد ذلك بيانات التدريب هذه لمعرفة العلاقات بين متغيرات الإدخال والمتغير المستهدف، ويمكنه بعد ذلك إجراء تنبؤات أو قرارات بشأن بيانات جديدة غير مرئية.

هناك ثلاثة أنواع رئيسية من التعلم الآلي :التعلم تحت الإشراف والتعلم غير الخاضع للإشراف والتعلم المعزز .في التعلم الخاضع للإشراف، تتضمن بيانات التدريب كلاً من ميزات الإدخال والمتغير المستهدف، والهدف هو معرفة التعيين بين متغيرات الإدخال والإخراج .تتضمن خوارزميات التعلم الخاضع للإشراف الشائعة الانحدار الخطي والانحدار اللوجستي وأشجار القرار والشبكات العصبية.

في التعلم غير الخاضع للإشراف، تتضمن بيانات التدريب فقط ميزات الإدخال، والهدف هو اكتشاف الأنماط أو العلاقات في البيانات .تتضمن خوارزميات التعلم غير الخاضعة للإشراف التجميع، وتقليل الأبعاد، والتنقيب في قواعد الارتباط.

التعلم المعزز هو نوع من التعلم الآلي الذي يتضمن تدريب وكيل على اتخاذ إجراءات في بيئة من أجل تعظيم إشارة المكافأة .يتعلم الوكيل من عواقب أفعاله ويقوم بتعديل سلوكه وفقًا لذلك.

للتعلم الآلي العديد من التطبيقات في مجموعة متنوعة من المجالات، بما في ذلك التمويل والرعاية الصحية والنقل والتسويق .يمكن استخدامه لمهام مثل التنبؤ بضغوط العملاء، واكتشاف الاحتيال، وتشخيص الحالات الطبية، والتوصية بمنتجات أو خدمات.

باختصار، يعد التعلم الآلي أداة قوية للتحليل والتنبؤ بناءً على البيانات .وهي تتضمن نماذج تدريب على مجموعات البيانات لتعلم الأنماط والعلاقات، ويمكن استخدامها في مجموعة متنوعة من التطبيقات لتحسين عملية صنع القرار وأتمتة المهام.

١-التعلم الخاضع للإشراف

التعلم الخاضع للإشراف هو نوع من التعلم الآلي حيث تتضمن بيانات التدريب ميزات الإدخال ومتغير الإخراج، أو المتغير المستهدف الذي يحاول النموذج التنبؤ به .الهدف من التعلم الخاضع للإشراف هو معرفة رسم الخرائط بين ميزات الإدخال والمتغير المستهدف بحيث يمكن للنموذج إجراء تنبؤات دقيقة بشأن البيانات الجديدة غير المرئية.

هناك نوعان رئيسيان من التعلم الخاضع للإشراف :الانحدار والتصنيف .في الانحدار، يكون المتغير المستهدف مستمرًا والهدف هو التنبؤ بقيمة عددية .تتضمن خوارزميات الانحدار الشائعة الانحدار الخطي والانحدار متعدد الحدود ودعم الانحدار المتجه.

في التصنيف، يكون المتغير المستهدف فئويًا والهدف هو التنبؤ بالفئة أو الفئة التي تنتمي إليها بيانات الإدخال .تتضمن خوارزميات التصنيف الشائعة الانحدار اللوجستي وأشجار القرار وجيران k الأقرب والشبكات العصبية.

يتضمن التعلم الخاضع للإشراف عدة خطوات، بما في ذلك إعداد البيانات، والتدريب النموذجي، وتقييم النموذج .تتمثل الخطوة الأولى في إعداد البيانات عن طريق التنظيف والمعالجة المسبقة وتحويلها إلى تنسيق يمكن استخدامه للتدريب .يتضمن ذلك إزالة القيم المفقودة وقياس البيانات وترميز المتغيرات الفئوية.

بعد ذلك، يتم تدريب النموذج على بيانات التدريب باستخدام خوارزمية تحسين تقلل من دالة التكلفة .تقيس دالة التكلفة الفرق بين المخرجات المتوقعة للنموذج والمخرجات الفعلية.بمجرد تدريب النموذج، يتم تقييمه على مجموعة بيانات تحقق أو اختبار منفصلة لقياس أدائه .تشمل مقاييس التقييم الشائعة للتعلم الخاضع للإشراف الدقة والدقة والتذكر ودرجة F1 وتحليل منحنى.ROC

يحتوي التعلم الخاضع للإشراف على العديد من التطبيقات في مجموعة متنوعة من المجالات، بما في ذلك التعرف على الصور والكلام ومعالجة اللغة الطبيعية والتنبؤ المالي .يمكن استخدامه لحل مجموعة واسعة من المشاكل، مثل التنبؤ بأسعار الأسهم، وتصنيف رسائل البريد الإلكتروني العشوائية، وتحديد المعاملات الاحتيالية.

باختصار، يعد التعلم الخاضع للإشراف أسلوبًا قويًا للتنبؤ بمتغيرات المخرجات بناءً على ميزات الإدخال . يتضمن تدريب نموذج على البيانات المصنفة وتقييم أدائه على بيانات جديدة غير مرئية .يحتوي التعلم الخاضع للإشراف على مجموعة واسعة من التطبيقات وهو أداة أساسية في مجال التعلم الآلي.

٢-التعلم غير الخاضع للإشراف

التعلم غير الخاضع للإشراف هو نوع من التعلم الآلي لا تتضمن فيه بيانات التدريب متغيرات الإخراج أو الملصقات .بدلاً من ذلك، فإن الهدف من التعلم غير الخاضع للإشراف هو اكتشاف الأنماط أو الهياكل أو العلاقات في البيانات دون معرفة مسبقة بما تمثله البيانات.

هناك نوعان رئيسيان من التعلم غير الخاضع للإشراف :التجميع وتقليل الأبعاد .في التجميع، الهدف هو تجميع نقاط البيانات المتشابهة معًا بناءً على ميزاتها أو خصائصها .تتضمن خوارزميات التجميع الشائعة k-mean، والتكتل الهرمي ، والتجمع القائم على الكثافة.

في تقليل الأبعاد، الهدف هو تقليل عدد الميزات في البيانات مع الحفاظ على المعلومات الأكثر أهمية .يمكن أن يكون هذا مفيدًا لتصور البيانات عالية الأبعاد أو تقليل التعقيد الحسابي لنموذج التعلم الآلي .تتضمن تقنيات تقليل الأبعاد الشائعة تحليل المكونات الرئيسية (PCA) ، و t-SNE ، وأجهزة التشفير التلقائية.

يتضمن التعلم غير الخاضع للإشراف عدة خطوات، بما في ذلك إعداد البيانات والتدريب النموذجي وتقييم النموذج .تتمثل الخطوة الأولى في إعداد البيانات عن طريق التنظيف والمعالجة المسبقة وتحويلها إلى تنسيق يمكن استخدامه للتدريب.

بعد ذلك، يتم تدريب النموذج على البيانات باستخدام خوارزمية التحسين التي تزيد من مقياس التشابه أو التماسك بين نقاط البيانات .يمكن القيام بذلك باستخدام خوارزميات التجميع، والتي تجمع نقاط البيانات المتشابهة معًا، أو تقنيات تقليل الأبعاد، والتي تحدد أهم الميزات في البيانات.

بمجرد تدريب النموذج، يتم تقييمه باستخدام مقاييس مثل درجة الصورة الظلية أو مؤشر-Calinski Harabaszأو مؤشر Davies-Bouldin للتجميع وخطأ إعادة الإعمار أو التباين الموضح لتقليل الأبعاد . ومع ذلك، فإن تقييم نماذج التعلم غير الخاضعة للإشراف يمثل تحديًا بشكل عام أكثر من تقييم نماذج التعلم الخاضع للإشراف، حيث لا توجد متغيرات أو تسميات مخرجات محددة بوضوح لمقارنة التنبؤات بها.

التعلم غير الخاضع للإشراف له العديد من التطبيقات في مجموعة متنوعة من المجالات، مثل تقسيم العملاء، واكتشاف العيوب، وضغط الصور .يمكن استخدامه لاكتشاف الأنماط المخفية في البيانات التي لا تظهر على الفور، واكتساب نظرة ثاقبة حول بنية الأنظمة المعقدة.

باختصار، يعد التعلم غير الخاضع للإشراف أسلوبًا قويًا لاكتشاف الأنماط أو البنية أو العلاقات في البيانات دون معرفة مسبقة بما تمثله البيانات .يتضمن تدريب نموذج على البيانات غير المسماة وتقييم أدائه باستخدام مقاييس التشابه أو التماسك بين نقاط البيانات .يحتوي التعلم غير الخاضع للإشراف على مجموعة واسعة من التطبيقات وهو أداة أساسية في مجال التعلم الآلي.

٣-التعليم المعزز

التعليم المحسن هو مجال يسعى إلى استخدام علم البيانات والبيانات الضخمة لتحسين جودة التعليم وتعزيز تجربة التعلم للطلاب .من خلال الاستفادة من البيانات والتحليلات، يهدف التعليم المحسن إلى تحديد الأنماط والاتجاهات في سلوك الطلاب وأدائهم، واستخدام هذه المعلومات لتحسين البرامج التعليمية وتحسين نتائج الطلاب.

يعد التعلم المخصص أحد مجالات التركيز الرئيسية في التعليم المعزز .من خلال جمع البيانات حول أداء الطالب والتفضيلات وأنماط التعلم، يمكن للمعلمين إنشاء تجارب تعليمية مخصصة تكون أكثر تفاعلاً وفعالية لكل طالب على حدة .يمكن أن يشمل ذلك استخدام تقنيات التعلم التكيفي، والتي تستخدم خوارزميات التعلم الآلي لضبط المحتوى ووتيرة التعلم ديناميكيًا بناءً على أداء الطالب.

مجال آخر للتركيز هو التحليلات التنبؤية، والتي تستخدم البيانات لتحديد الطلاب المعرضين لخطر التسرب أو التخلف عن الركب .من خلال تحديد هؤلاء الطلاب في وقت مبكر، يمكن للمعلمين التدخل من خلال الدعم والموارد المستهدفة لمساعدتهم على البقاء على المسار الصحيح وتحقيق أهدافهم.

يتم استخدام علم البيانات والبيانات الضخمة أيضًا لتحسين البحث والتقييم التربوي .من خلال تحليل مجموعات البيانات الكبيرة، يمكن للباحثين اكتساب رؤى حول فعالية البرامج والتدخلات التعليمية المختلفة، واستخدام هذه المعلومات لتحسين البرامج والسياسات المستقبلية.

ومع ذلك، هناك أيضًا تحديات واعتبارات أخلاقية يجب مراعاتها في مجال التعليم المعزز .أحد المخاوف هو الخصوصية وأمن البيانات، حيث أن بيانات الطلاب حساسة ويجب حمايتها من الوصول غير المصرح به أو سوء الاستخدام .التحدي الآخر هو ضمان أن يكون استخدام البيانات والتحليلات في التعليم شفافًا وعادلاً، ولا يديم التحيز أو التمييز.

باختصار، التعليم المحسن هو مجال يسعى إلى استخدام علم البيانات والبيانات الضخمة لتحسين جودة التعليم وتعزيز تجربة التعلم للطلاب .من خلال الاستفادة من البيانات والتحليلات، يمكن للمعلمين إنشاء تجارب تعليمية مخصصة، وتحديد الطلاب المعرضين للخطر، وتحسين البحث التعليمي والتقييم .ومع ذلك، يجب مراعاة الاعتبارات الأخلاقية ومخاوف الخصوصية بعناية عند استخدام البيانات في التعليم.

الفصل السادس: تحليل الانحدار

الفصل السادس :تحليل الانحدار

أ-الانحدار الخطي

الانحدار الخطي هو طريقة إحصائية تستخدم لنمذجة العلاقة بين متغير تابع ومتغير واحد أو أكثر من المتغيرات المستقلة .الهدف من الانحدار الخطي هو العثور على أفضل خط مناسب يمكنه شرح العلاقة بين المتغيرات وعمل تنبؤات حول القيم المستقبلية للمتغير التابع.

في الانحدار الخطي، يُفترض أن يكون المتغير التابع مستمرًا، ويفترض أن تكون العلاقة بين المتغير التابع والمتغيرات المستقلة خطية .تتضمن الطريقة تقدير معاملات المعادلة الخطية، والتي تحدد ميل وتقاطع الخط الأنسب.

هناك نوعان رئيسيان من الانحدار الخطي :الانحدار الخطي البسيط والانحدار الخطي المتعدد .يتضمن الانحدار الخطي البسيط متغيرًا مستقلاً واحدًا ومتغيرًا تابعًا واحدًا، بينما يتضمن الانحدار الخطي المتعدد أكثر من متغير مستقل ومتغير تابع واحد.

يستخدم الانحدار الخطي على نطاق واسع في مختلف المجالات، بما في ذلك الاقتصاد والتمويل والعلوم الاجتماعية والهندسة .غالبًا ما يستخدم لتحليل الاتجاهات وإجراء تنبؤات، مثل التنبؤ بالمبيعات أو التنبؤ بأسعار الأسهم أو نمذجة تأثير الحملات التسويقية.

تتمثل إحدى المزايا الرئيسية للانحدار الخطي في بساطته وقابليته للتفسير .توفر الطريقة فهمًا واضحًا للعلاقة بين المتغيرات، ويمكن بسهولة تفسير المعاملات من حيث التغيير في المتغير التابع لتغيير الوحدة في المتغير المستقل.

ومع ذلك، فإن الانحدار الخطي له أيضًا بعض القيود .أحد القيود هو أنه يفترض وجود علاقة خطية بين المتغيرات، والتي قد لا تكون كذلك دائمًا .بالإضافة إلى ذلك، يمكن أن يكون الانحدار الخطي حساسًا للقيم المتطرفة والملاحظات المؤثرة، والتي يمكن أن تؤثر على دقة النموذج.

بشكل عام، يعد الانحدار الخطي طريقة قوية ومستخدمة على نطاق واسع لنمذجة العلاقة بين المتغيرات والتنبؤ .إن بساطته وقابليته للتفسير تجعله أداة قيمة لتحليل البيانات في مختلف المجالات، ولكن يجب استخدامه بحذر ويجب مراعاة حدوده بعناية.

ب- الانحدار المتعدد

الانحدار المتعدد هو امتداد للانحدار الخطي البسيط الذي يتضمن أكثر من متغير مستقل واحد .إنها طريقة إحصائية تُستخدم لنمذجة العلاقة بين متغير تابع ومتغيرين مستقلين أو أكثر .الهدف من الانحدار المتعدد هو العثور على أفضل خط ملائم يمكنه شرح العلاقة بين المتغيرات وعمل تنبؤات حول القيم المستقبلية للمتغير التابع.

في الانحدار المتعدد، يُفترض أن يكون المتغير التابع مستمرًا، ويفترض أن تكون العلاقة بين المتغير التابع والمتغيرات المستقلة خطية .تتضمن الطريقة تقدير معاملات المعادلة الخطية، والتي تحدد ميل وتقاطع الخط الأنسب.

الميزة الرئيسية للانحدار المتعدد على الانحدار الخطي البسيط هي أنه يسمح بتحليل تأثير المتغيرات المستقلة المتعددة على المتغير التابع، مع التحكم في تأثيرات المتغيرات الأخرى .يمكن أن يوفر هذا فهمًا أكثر دقة ودقة للعلاقة بين المتغيرات.

يستخدم الانحدار المتعدد على نطاق واسع في مختلف المجالات، بما في ذلك الاقتصاد والتمويل والعلوم الاجتماعية والهندسة .غالبًا ما يستخدم لتحليل تأثير العوامل المختلفة على النتيجة، مثل التنبؤ بالمبيعات بناءً على الإنفاق التسويقي، أو نمذجة تأثير العوامل الديموغرافية على النتائج الصحية.

ومع ذلك، فإن الانحدار المتعدد له أيضًا بعض القيود .أحد القيود هو أنه يفترض وجود علاقة خطية بين المتغيرات، والتي قد لا تكون كذلك دائمًا .بالإضافة إلى ذلك، يمكن أن يكون الانحدار المتعدد حساسًا للقيم المتطرفة والملاحظات المؤثرة، والتي يمكن أن تؤثر على دقة النموذج .علاوة على ذلك، يمكن أن يؤدي إدراج عدد كبير جدًا من المتغيرات المستقلة إلى التخصيص الزائد، مما قد يقلل من قابلية تعميم النموذج.

بشكل عام، يعد الانحدار المتعدد طريقة قوية ومستخدمة على نطاق واسع لنمذجة العلاقة بين المتغيرات وعمل التنبؤات .إن قدرتها على تحليل تأثير المتغيرات المستقلة المتعددة على المتغير التابع تجعلها أداة قيمة لتحليل البيانات المعقدة في مختلف المجالات، ولكن يجب استخدامها بحذر ويجب مراعاة قيودها بعناية.

ج-الانحدار اللوجستي

الانحدار اللوجستي هو طريقة إحصائية تُستخدم لنمذجة العلاقة بين متغير ثنائي تابع ومتغير واحد أو أكثر من المتغيرات المستقلة .يتم استخدامه بشكل شائع في المواقف التي يمثل فيها المتغير التابع نتيجة ثنائية، مثل ما إذا كان العميل سيجري عملية شراء أم لا، أو ما إذا كان المريض سيستجيب للعلاج أم لا.

يستخدم نموذج الانحدار اللوجستي وظيفة لوجستية لنمذجة احتمالية المتغير التابع الذي يأخذ قيمة معينة عادةً 0 أو 1 بناءً على قيم المتغيرات المستقلة .تنتج الوظيفة اللوجستية منحنى على شكل حرفS ، والذي يمثل احتمال أن يأخذ المتغير التابع قيمة معينة كدالة للمتغيرات المستقلة.

تتمثل إحدى مزايا الانحدار اللوجستي في قدرته على التعامل مع العلاقات غير الخطية بين المتغيرات المستقلة والمتغير التابع .يتم تحقيق ذلك عن طريق تحويل المتغيرات المستقلة باستخدام دالة لوغاريتمية.

يستخدم الانحدار اللوجستي على نطاق واسع في مختلف المجالات، بما في ذلك التسويق والرعاية الصحية والعلوم الاجتماعية .غالبًا ما يستخدم للتنبؤ باحتمالية وقوع حدث ما، مثل التنبؤ باحتمالية شراء العميل لمنتج أو احتمال استجابة المريض للعلاج.

ومع ذلك، فإن الانحدار اللوجستي له أيضًا بعض القيود .أحد القيود هو أنه يفترض وجود علاقة خطية بين المتغيرات المستقلة ولوغاريتم نسبة الأرجحية، والتي قد لا تكون كذلك دائمًا .بالإضافة إلى ذلك، يمكن أن يكون الانحدار اللوجستي حساسًا للقيم المتطرفة والملاحظات المؤثرة، والتي يمكن أن تؤثر على دقة النموذج .علاوة على ذلك، يمكن أن يؤدي إدراج عدد كبير جدًا من المتغيرات المستقلة إلى التخصيص الزائد، مما قد يقلل من قابلية تعميم النموذج.

بشكل عام، يعد الانحدار اللوجستي أداة قيمة لنمذجة العلاقة بين متغير ثنائي تابع ومتغير واحد أو أكثر من المتغيرات المستقلة .إن قدرتها على التعامل مع العلاقات غير الخطية والتنبؤ باحتمالية وقوع حدث يجعلها طريقة مفيدة في مختلف المجالات، ولكن يجب استخدامها بحذر ويجب مراعاة قيودها بعناية.

الفصل السابع: التجميع

الفصل السابع :التجميع

أ-التجميع

K-mean clusteringهو نوع من خوارزمية التعلم غير الخاضعة للإشراف المستخدمة في التعلم الآلي وعلوم البيانات .إنها تقنية تقسم مجموعة بيانات معينة إلى مجموعات، حيث تمثل كل مجموعة مجموعة من نقاط البيانات التي تشترك في خصائص متشابهة .تعمل الخوارزمية عن طريق التخصيص المتكرر لنقاط البيانات إلى المجموعات بناءً على قربها من مركز الكتلة، المعروف باسم النقطة الوسطى .تهدف الخوارزمية إلى تقليل مجموع المسافات المربعة بين نقاط البيانات والنقطة الوسطى المخصصة لها، والمعروفة باسم الوظيفة الموضوعية.

تُستخدم خوارزمية K-mean على نطاق واسع في مختلف المجالات، بما في ذلك التسويق وعلم الأحياء وعلوم الكمبيوتر .يتم استخدامه بشكل شائع لتجزئة العملاء، وتجزئة الصور، واكتشاف الشذوذ، والتعرف على الأنماط .تتمثل إحدى المزايا الرئيسية لتجميع الوسائل K في بساطتها وسهولة تنفيذها .كما أنه فعال من الناحية الحسابية ويمكنه التعامل مع مجموعات البيانات الكبيرة.

تحتوي خوارزمية K-mean على العديد من المعلمات التي يجب ضبطها، بما في ذلك عدد المجموعات (K)وطريقة التهيئة وقياس المسافة المستخدمة لقياس التشابه بين نقاط البيانات .يمكن أن يؤثر اختيار هذه المعلمات على جودة نتائج التجميع، ويمكن استخدام طرق مختلفة لتحديد القيم المثلى لهذه المعلمات.

على الرغم من مزاياها، فإن خوارزمية K-mean لديها بعض القيود .يفترض أن المجموعات كروية ولها تباينات متساوية، والتي قد لا تكون موجودة دائمًا في مجموعات بيانات العالم الحقيقي .كما أنه حساس للتكوين الأولي للنقط الوسطى وقد يتقارب مع حل دون المستوى الأمثل .لمعالجة هذه القيود، تم اقتراح العديد من المتغيرات لخوارزمية K-mean ، مثل التجميع الهرمي ، والتجميع الضبابي ، والتكتل الطيفي.

ب- المجموعات الهرمية

التجميع الهرمي هو نوع من خوارزمية التجميع المستخدمة في علوم البيانات والتعلم الآلي .على عكس خوارزمية K-mean ، التي تتطلب تحديد عدد المجموعات مسبقًا ، لا يتطلب التجميع الهرمي عددًا محددًا مسبقًا من المجموعات .بدلاً من ذلك، يقوم ببناء تسلسل هرمي من المجموعات عن طريق الدمج المتكرر أو تقسيم المجموعات بناءً على التشابه بين نقاط البيانات.

يمكن تصنيف خوارزمية التجميع الهرمي إلى نوعين رئيسيين :التجميع التكتلي والتقسيمي .في التجميع التكتلي، تشكل كل نقطة بيانات في البداية كتلة، ثم يتم دمج أزواج المجموعات على التوالي بناءً على تشابهها حتى يتم تكوين مجموعة واحدة تحتوي على جميع نقاط البيانات .في التجميع التقسيمي، تشكل جميع نقاط البيانات في البداية كتلة واحدة، ثم يتم تقسيم الكتلة على التوالي إلى مجموعات أصغر حتى تحتوي كل مجموعة على نقطة بيانات واحدة.

يمكن تمثيل المجموعات الهرمية باستخدام مخطط الشجرة، وهو مخطط يشبه الشجرة يعرض التسلسل الهرمي للمجموعات .يمثل ارتفاع كل عقدة في مخطط الأسنان الاختلاف بين المجموعات أو نقاط البيانات، وتمثل الفروع دمج أو تقسيم المجموعات.

تتمثل إحدى مزايا التجميع الهرمي في قدرته على التعامل مع الأشكال والأحجام المختلفة للكتل .يمكن أن يوفر أيضًا تمثيلًا مرئيًا لنتائج التجميع، والذي يمكن أن يكون مفيدًا في فهم بنية البيانات .بالإضافة إلى ذلك، يمكن قطع مخطط الأسنان على ارتفاع معين للحصول على عدد محدد من العناقيد.

ومع ذلك، يمكن أن يكون التجميع الهرمي مكلفًا من الناحية الحسابية وقد لا يكون مناسبًا لمجموعات البيانات الكبيرة .يمكن أن يكون أيضًا حساسًا لاختيار مقياس المسافة ومعيار الارتباط المستخدم لقياس التشابه بين نقاط البيانات والمجموعات.

تم اقتراح العديد من المتغيرات للتكتل الهرمي، مثل التجميع الانقسام القائم على تحليل المكون الرئيسي (PCA)، والتكتل مع معايير الارتباط المختلفة ، مثل الارتباط الكامل ، والرابط الفردي ، ومتوسط الارتباط .يمكن أن تعالج هذه المتغيرات بعض قيود التجميع الهرمي وتوفر المزيد من المرونة والتحكم في عملية التجميع.

ج-التجميع على أساس الكثافة

التجميع القائم على الكثافة هو أسلوب تعلم آلي شائع غير خاضع للإشراف يُستخدم في علم البيانات لتحديد المجموعات في مجموعة بيانات .إنه مفيد بشكل خاص عند التعامل مع مجموعات البيانات التي تحتوي على مستويات عالية من الضوضاء والقيم المتطرفة، مما يجعل من الصعب تحديد مجموعات واضحة باستخدام تقنيات التجميع الأخرى.

يعمل التجميع المستند إلى الكثافة عن طريق تحديد المناطق في مجموعة البيانات التي تحتوي على كثافة عالية من نقاط البيانات ثم تجميع تلك المناطق معًا كمجموعات .أكثر خوارزمية التجميع المستندة إلى الكثافة شيوعًا هي خوارزمية التجميع المكاني المستند إلى الكثافة للتطبيقات ذات الضوضاء(DBSCAN)

تعمل خوارزمية DBSCAN عن طريق تحديد حي حول كل نقطة بيانات في مجموعة البيانات ثم تحديد المجموعات بناءً على كثافة النقاط داخل تلك الأحياء .يتم تصنيف النقاط الموجودة في مناطق ذات كثافة منخفضة على أنها ضوضاء ويتم استبعادها من أي مجموعات.

تتمثل إحدى المزايا الرئيسية للتجميع المستند إلى الكثافة في قدرته على التعامل مع العلاقات غير الخطية بين المتغيرات في مجموعة البيانات .وهذا يجعلها مفيدة بشكل خاص في مجالات مثل تحليل الصور، حيث يمكن استخدامها لتحديد مجموعات من ميزات الصورة المتشابهة.

ومع ذلك، فإن التجميع القائم على الكثافة له أيضًا بعض القيود .يتمثل أحد التحديات الرئيسية في تحديد القيم المناسبة للمعلمات الفائقة للخوارزمية، مثل حجم الحي وعتبة الكثافة .إذا لم يتم اختيار هذه القيم بشكل مناسب، فقد لا تتمكن الخوارزمية من تحديد مجموعات ذات معنى في البيانات.

بشكل عام، يعد التجميع المستند إلى الكثافة أداة قوية في مجموعة أدوات عالم البيانات، وقدرته على تحديد المجموعات في مجموعات البيانات المزعجة تجعله مفيدًا بشكل خاص في مجموعة واسعة من التطبيقات.

الفصل الثامن :التصنيف

الفصل الثامن :التصنيف

أ-أشجار القرار

أشجار القرار هي أداة قوية للتعلم الآلي تُستخدم لحل مشاكل الانحدار والتصنيف .يتم إنشاؤها من خلال سلسلة من القرارات الثنائية التي تقسم البيانات إلى مجموعات فرعية أصغر بناءً على معايير معينة حتى يتم الوصول إلى معيار التوقف .يتم اتخاذ كل قرار أو تقسيم في الشجرة بناءً على الميزة التي توفر أكبر قدر من المعلومات، أي الخيار الذي يفصل بين الفئات بشكل أفضل أو يشرح التباين الأكبر في البيانات.

يمكن تفسير الشجرة الناتجة على أنها مجموعة من قواعد الشرط التي يمكن استخدامها للتنبؤ بالبيانات الجديدة .من السهل تفسير وتصور أشجار القرار، مما يجعلها شائعة في مجالات مثل التمويل والطب والتسويق.

ومع ذلك، يمكن أن تعاني أشجار القرار من فرط التخصيص، حيث يصبح النموذج معقدًا للغاية ويتناسب مع بيانات التدريب بشكل وثيق للغاية، مما يؤدي إلى ضعف الأداء في البيانات الجديدة .يمكن معالجة ذلك من خلال تقنيات مثل التقليم، الذي يزيل العقد ذات اكتساب المعلومات المنخفض أو يقلل من حجم الشجرة، أو باستخدام طرق التجميع مثل الغابات العشوائية أو تعزيز التدرج، والتي تجمع بين تنبؤات أشجار القرار المتعددة.بشكل عام، تعد أشجار القرار أداة قيمة في مجموعة أدوات عالم البيانات، حيث توفر طريقة بديهية وقابلة للتفسير لحل مجموعة متنوعة من مشكلات الانحدار والتصنيف.

ب-الغابات العشوائية Random Forests

تعد الغابات العشوائية طريقة تعلم جماعية شائعة تستخدم في التعلم الآلي لتحسين دقة أشجار القرار .تُنشئ الطريقة أشجار قرارات متعددة على مجموعات فرعية مختارة عشوائيًا من بيانات التدريب ثم تجمع مخرجاتها لعمل توقع نهائي.

تبدأ الخوارزمية باختيار عشوائي لمجموعة فرعية من الميزات من مجموعة بيانات التدريب وإنشاء شجرة قرار بناءً على تلك المجموعة الفرعية .يتم تكرار العملية عدة مرات، مما يؤدي إلى إنشاء أشجار قرارات متعددة .أثناء عملية بناء الشجرة، يتم تدريب كل شجرة على مجموعة فرعية مختلفة من البيانات، ويتم اختيار مجموعات فرعية مختلفة من الميزات بشكل عشوائي في كل عقدة في الشجرة.

بمجرد بناء الأشجار، تجمع خوارزمية الغابة العشوائية مخرجاتها لعمل توقع نهائي .يتم ذلك عن طريق تجميع مخرجات جميع أشجار القرار، إما من خلال التصويت بالأغلبية لمشاكل التصنيف أو المتوسط لمشاكل الانحدار.

تتميز الغابات العشوائية بالعديد من المزايا مقارنة بالأشجار ذات القرار الفردي .أولاً، هم أقل عرضة للإفراط في التجهيز، حيث يساعد الجمع بين الأشجار المتعددة في تقليل تأثير الضوضاء في البيانات .ثانيًا، تكون أكثر قوة بالنسبة للبيانات المفقودة أو الخاطئة، حيث يمكن احتساب القيم المفقودة بناءً على قيم الميزات الأخرى .أخيرًا، يمكنهم التعامل مع البيانات الفئوية والرقمية.

تتضمن بعض تطبيقات الغابات العشوائية تصنيف الصور وتحليل المشاعر والتنبؤ بتضخم العملاء .على الرغم من مزاياها العديدة، يمكن أن تكون الغابات العشوائية مكلفة من الناحية الحسابية وتتطلب ضبطًا دقيقًا للمعلمات الفائقة لتحقيق الأداء الأمثل.

ج-آلات المتجهات الداعمة (SVMs)

آلات المتجهات الداعمة هي نوع من خوارزمية التعلم الخاضع للإشراف المستخدمة في التصنيف وتحليل الانحدار .تعتمد آلات المتجهات الداعمة على مفهوم العثور على المستوى الفائق في مساحة عالية الأبعاد تفصل بين نقاط البيانات من الفئات المختلفة .يتم اختيار المستوى الفائق الذي يزيد الهامش بين الفئتين كحد القرار.

في آلات المتجهات الداعمة، يتم تحويل نقاط البيانات إلى مساحة ذات أبعاد أعلى باستخدام وظيفةkernel ، والتي تسمح بالفصل غير الخطي للفئات .يمكن أن تكون وظيفة النواة دالة خطية، أو دالة متعددة الحدود، أو دالة أساس قطري، أو دالة سيني .يعتمد اختيار وظيفة kernel على خصائص البيانات والمشكلة التي يتم حلها.

تتمتع آلات المتجهات الداعمة بالعديد من المزايا مقارنة بخوارزميات التصنيف الأخرى .إنها فعالة في المساحات عالية الأبعاد، ويمكنها التعامل مع حدود القرار غير الخطية، وهي أقل عرضة للتركيب الزائد . تعد آلات المتجهات الداعمة مناسبة أيضًا لمجموعات البيانات الصغيرة والمتوسطة الحجم.

تم تطبيق آلات المتجهات الداعمة في مجالات مختلفة، مثل التعرف على الصور والمعلوماتية الحيوية والتمويل .لقد تم استخدامها لتصنيف الأورام بناءً على بيانات التعبير الجيني، والتنبؤ بتفاعلات البروتين والبروتين، والكشف عن تزوير بطاقات الائتمان .تم استخدام آلات المتجهات الداعمة أيضًا في معالجة اللغة الطبيعية لتحليل المشاعر وتصنيف النص.

الفصل التاسع :التعلم العميق

الفصل التاسع :التعلم العميق

أ-الشبكات العصبية

الشبكات العصبية هي مجموعة فرعية من التعلم الآلي تم تصميمها على غرار بنية ووظيفة الدماغ البشري . وهي تتكون من طبقات من العقد المترابطة التي تعالج المعلومات وتنقلها، مما يسمح للشبكة بالتعلم وإجراء تنبؤات بناءً على أنماط في بيانات الإدخال.

تقع الخلية العصبية في قلب الشبكة العصبية، والتي تتلقى مدخلات من الخلايا العصبية الأخرى وتولد مخرجات بناءً على مجموع مرجح لتلك المدخلات .تحدد الأوزان الأهمية النسبية لكل مدخل في التأثير على ناتج العصبون .يتم تنظيم هذه الخلايا العصبية في طبقات، حيث تتلقى طبقة الإدخال البيانات الأولية والطبقات اللاحقة التي تقوم بمعالجة البيانات وتحويلها لإنتاج مخرجات.

هناك عدة أنواع من الشبكات العصبية، بما في ذلك شبكات التعلم المتقدم والمتكرر والتلافيفي والعميق . الشبكات المغذية هي أبسط أنواعها، حيث تتدفق المعلومات في اتجاه واحد فقط من المدخلات إلى المخرجات .تحتوي الشبكات المتكررة على حلقات تسمح بتمرير المعلومات بين الخلايا العصبية بمرور الوقت، مما يجعلها مناسبة تمامًا للبيانات المتسلسلة مثل الكلام والنص .تم تصميم الشبكات التلافيفية للتعامل مع البيانات ذات البنية الشبيهة بالشبكة، مثل الصور، واستخدام المرشحات التلافيفية لاستخراج الميزات من البيانات .شبكات التعلم العميق هي شبكات عصبية ذات طبقات عديدة وقادرة على تعلم أنماط معقدة للغاية في البيانات.

أظهرت الشبكات العصبية نجاحًا في مجموعة متنوعة من التطبيقات، بما في ذلك التعرف على الصور والكلام ومعالجة اللغة الطبيعية والسيارات ذاتية القيادة .يتم استخدامها أيضًا في صناعات مثل التمويل والرعاية الصحية والتسويق لمهام مثل اكتشاف الاحتيال وتشخيص الأمراض وتجزئة العملاء .ومع ذلك، يمكن أن تكون الشبكات العصبية مكثفة من الناحية الحسابية وتتطلب كميات كبيرة من البيانات للتدريب بفعالية.

ب-الشبكات العصبية التلافيفية (CNN)

الشبكات العصبية التلافيفية (CNN) هي نوع من الشبكات العصبية المتخصصة في مهام التعرف على الصور .في شبكة عصبية نموذجية، كل خلية عصبية في طبقة واحدة متصلة بكل خلية عصبية في الطبقة التالية .ومع ذلك، في الشبكات العصبية التلافيفية، ترتبط الخلايا العصبية في طبقة واحدة فقط بمنطقة صغيرة من الخلايا العصبية في الطبقة التالية .هذا يجعل شبكات الشبكات العصبية التلافيفية أكثر كفاءة من الشبكات العصبية التقليدية، لأنها تتطلب معلمات أقل للتعلم.

اللبنات الأساسية للشبكة العصبية التلافيفية هي الطبقات التلافيفية، والتي تطبق مجموعة من المرشحات على الصورة المدخلة، كل منها يكتشف ميزة معينة في الصورة .يتم التعرف على هذه المرشحات من خلال خوارزمية backpropagation ، والتي تعدل أوزان المرشحات لتقليل الخطأ بين الإخراج المتوقع والإخراج الفعلي.

بالإضافة إلى الطبقات التلافيفية، قد تتضمن شبكات الشبكات العصبية التلافيفية أيضًا طبقات التجميع، والتي تقلل الأبعاد المكانية لخرائط المعالم عن طريق أخذ القيمة القصوى أو المتوسطة في كل منطقة فرعية، والطبقات المتصلة بالكامل، والتي تقوم بإجراء التصنيف بناءً على الميزات التي تم اكتشافها بواسطة الطبقات التلافيفية.

حققت شبكات الشبكات العصبية التلافيفية أداءً متطورًا في مجموعة متنوعة من مهام التعرف على الصور، بما في ذلك اكتشاف الكائنات وتجزئة الصور وتصنيف الصور .كما تم تطبيقها بنجاح على أنواع أخرى من البيانات، مثل التعرف على الكلام ومعالجة اللغة الطبيعية.

ج-الشبكات العصبية المتكررة (RNNs)

الشبكات العصبية المتكررة هي نوع من الشبكات العصبية التي يمكنها التعامل مع البيانات المتسلسلة، مثل النص والكلام وبيانات السلاسل الزمنية .على عكس الشبكات العصبية المغذية التي تعالج المدخلات بترتيب ثابت، يمكن لشبكات العصبية المتكررة الحفاظ على الذاكرة الداخلية، مما يسمح لها بمعالجة المدخلات بترتيب زمني.

تتكون البنية الأساسية لشبكات العصبية المتكررة من خلية عصبية متكررة واحدة تتلقى مدخلات ومخرجاتها من الخطوة الزمنية السابقة كمدخلات .يسمح هذا للشبكة بالحفاظ على" ذاكرة "للمدخلات السابقة التي شاهدتها، والتي يمكن استخدامها للتأثير على مخرجاتها الحالية .يمكن تدريب الشبكات العصبية

المتكررة باستخدام backpropagation عبر الزمن، وهو أحد أشكال الانتشار العكسي الذي يأخذ في الاعتبار الطبيعة المتسلسلة للبيانات.

أحد التطبيقات الرئيسية لشبكات العصبية المتكررة هو معالجة اللغة الطبيعي (NLP) يمكن استخدامها في مهام مثل نمذجة اللغة والترجمة الآلية وإنشاء النصوص .بالإضافة إلى ذلك، تم تطبيق RNNs أيضًا على بيانات السلاسل الزمنية، مثل أسعار الأسهم وبيانات الطقس، لمهام مثل التنبؤ واكتشاف الانحرافات.

هناك العديد من الاختلافات في الشبكات العصبية المتكررة، بما في ذلك الذاكرة طويلة المدى (LSTM) والوحدات المتكررة ذات البوابات (GRU) ، والتي تم تصميمها لمعالجة مشكلة التدرجات المتلاشية التي يمكن أن تحدث أثناء التدريب .تستخدم هذه الاختلافات آليات بوابات إضافية للتحكم في تدفق المعلومات داخل الشبكة، مما يسمح لها بالحفاظ على المعلومات لفترات زمنية أطول وتجنب مشكلة نسيان المدخلات السابقة.

الفصل العاشر: تحليل السلاسل الزمنية

الفصل العاشر :تحليل السلاسل الزمنية

أ-بيانات السلاسل الزمنية

بيانات السلاسل الزمنية هي نوع من البيانات التي يتم جمعها بمرور الوقت، عادةً على فترات منتظمة . إنها سلسلة من الملاحظات أو القياسات المأخوذة في نقاط زمنية متتالية، ويمكن استخدامها لتحليل الاتجاهات والأنماط في البيانات والتنبؤ بها .تُستخدم بيانات السلاسل الزمنية في مجموعة متنوعة من المجالات، بما في ذلك المالية والاقتصاد والهندسة والعلوم البيئية.

تتمثل إحدى السمات الرئيسية لبيانات السلاسل الزمنية في أنها تعتمد على الوقت، مما يعني أن ترتيب الملاحظات مهم .هذا يعني أنه يمكن تحليل بيانات السلاسل الزمنية باستخدام مجموعة متنوعة من الأساليب الإحصائية المصممة لمراعاة الطبيعة الزمنية للبيانات .على سبيل المثال، يمكن تحليل بيانات السلاسل الزمنية باستخدام نماذج المتوسط المتحرك المتكامل الانحدار الذاتي (ARIMA) ، والتي تُستخدم عادةً لنمذجة الاتجاهات في البيانات والتنبؤ بها.

ميزة أخرى مهمة لبيانات السلاسل الزمنية هي أنها غالبًا ما تظهر موسمية، مما يعني أن هناك أنماطًا منتظمة تحدث في أوقات محددة من السنة .على سبيل المثال، قد تزيد مبيعات التجزئة خلال موسم العطلات، أو قد يزيد استهلاك الطاقة خلال أشهر الصيف .يمكن حساب الموسمية باستخدام طرق التحلل الموسمية، مثل تحلل الاتجاه الموسمي باستخدام LOESS (STL) أو التحلل الموسمي للسلسلة الزمنية عن طريق المتوسطات المتحركة.(SEASONAL)

بشكل عام، تعد بيانات السلاسل الزمنية نوعًا مهمًا من البيانات المستخدمة لتحليل الاتجاهات والأنماط والتنبؤ بها بمرور الوقت .إنها أداة قيمة للشركات والباحثين الذين يتطلعون إلى اكتساب رؤى حول كيفية تغير المتغيرات بمرور الوقت واتخاذ قرارات مستنيرة بناءً على تلك الأفكار.

ب-نماذج المتوسط المتحرك التلقائي او الذاتي المتكامل ARIMA

تُستخدم نماذج ARIMA المتوسط المتحرك التلقائي المتكامل بشكل شائع لتحليل بيانات السلاسل الزمنية ، وهي عبارة عن سلسلة من الملاحظات المأخوذة على فترات منتظمة بمرور الوقت .نموذج ARIMA هو نوع من النماذج الإحصائية التي تستخدم القيم والأخطاء السابقة للتنبؤ بالقيم المستقبلية في سلسلة زمنية . تُستخدم نماذج ARIMA على نطاق واسع في العديد من المجالات، بما في ذلك المالية والاقتصاد والتنبؤ بالطقس والهندسة.

يرمز الاختصار ARIMA إلى ثلاثة مكونات للنموذج AR :الانحدار التلقائي و I متكامل و MA المتوسط المتحرك يمثل مكون AR الجزء الانحدار التلقائي من النموذج، مما يعني أنه يستخدم القيم السابقة للسلسلة للتنبؤ بالقيم المستقبلية .يمثل مكون MA جزء المتوسط المتحرك من النموذج، والذي يستخدم الأخطاء) الفرق بين القيم المتوقعة والفعلية لضبط التنبؤ .يمثل المكون I الجزء المتكامل من النموذج، مما يشير إلى أن البيانات قد تم تحويلها لجعلها ثابتة، مما يعني أن متوسط وتباين السلسلة لا يتغيران بمرور الوقت.

يمكن استخدام نماذج ARIMA لعمل تنبؤات للفترات الزمنية المستقبلية، ولتحديد الأنماط والاتجاهات في البيانات .يحدد ترتيب نموذجARIMA ، المشار إليه (p d q)عدد مصطلحات الانحدار التلقائي (p) ، والمتكاملة (d) ، والمتوسط المتحرك (q) المستخدمة في النموذج .يعد اختيار القيم الصحيحة لهذه المعلمات أمرًا مهمًا للنمذجة الدقيقة والتنبؤ بالسلسلة الزمنية.

من الناحية العملية، غالبًا ما تُستخدم نماذج ARIMA جنبًا إلى جنب مع تقنيات أخرى مثل المعالجة المسبقة للبيانات وهندسة الميزات واختيار النموذج لتحسين دقتها وأدائها .يتم استخدامها أيضًا جنبًا إلى جنب مع خوارزميات التعلم الآلي لبناء نماذج أكثر تعقيدًا يمكنها التقاط العلاقات غير الخطية والأنماط المعقدة في البيانات.

ج-التجانس او التمهيد الأسي

التجانس الأسي هو طريقة للتنبؤ بالسلسلة الزمنية تُستخدم لتنعيم البيانات والتنبؤ بها باستخدام اتجاه تدريجي . إنها تقنية شائعة ومستخدمة على نطاق واسع في مجال علم البيانات ولها تطبيقات في مختلف الصناعات مثل التمويل والتسويق والاقتصاد.

تعمل نماذج التجانس الأسية من خلال أخذ متوسط مرجح للملاحظات السابقة، حيث تنخفض الأوزان بشكل كبير مع تقدم الملاحظات في السن .وهذا يعني أن الملاحظات الأحدث تُعطى وزناً أكبر في عملية التسوية، بينما الملاحظات الأقدم لها تأثير أقل على التنبؤ.

هناك أنواع مختلفة من نماذج التجانس الأسية، بما في ذلك التجانس الأسي البسيط، والتجانس الأسي الخطي لهولت، والتجانس الأسي الموسمي لهولت وينترز .يتم استخدام التجانس الأسي البسيط للتنبؤ بالبيانات بدون اتجاه أو موسمية، بينما يتم استخدام التجانس الأسي الخطي لهولت للتنبؤ بالبيانات ذات الاتجاه الخطي . يتم استخدام التنعيم الأسي الموسمي لهولت وينترز للتنبؤ بالبيانات ذات الاتجاه الموسمي.

يمكن تخصيص نماذج التسوية الأسية بشكل أكبر عن طريق ضبط معلمات التنعيم وطريقة حساب خطأ التنبؤ .يتمثل أحد الأساليب الشائعة في استخدام متوسط الخطأ التربيعي (MSE) أو متوسط الخطأ المطلق (MAE)لتقييم دقة النموذج وضبط معلمات التنعيم.

باختصار، يعد التجانس الأسي أسلوبًا قويًا للتنبؤ ببيانات السلاسل الزمنية باستخدام اتجاه تدريجي .إن مرونته وسهولة استخدامه تجعله خيارًا شائعًا في مجال علم البيانات، وقد أثبت فعاليته في مجموعة واسعة من التطبيقات.

الفصل الحادي عشر: التنقيب عن النص ومعالجة اللغة الطبيعية

الفصل الحادي عشر :التنقيب عن النص ومعالجة اللغة الطبيعية

أ-مصادر البيانات النصية

تشير مصادر البيانات النصية إلى الأشكال المختلفة للبيانات النصية غير المنظمة التي يتم إنشاؤها من مصادر مختلفة مثل وسائل التواصل الاجتماعي ورسائل البريد الإلكتروني والمقالات الإخبارية والمواقع الإلكترونية والمنتديات عبر الإنترنت وغيرها .تتميز البيانات النصية بطبيعتها غير المهيكلة، مما يجعل تحليلها واستنباط رؤى منها أمرًا صعبًا.

في السنوات الأخيرة، ازداد حجم البيانات النصية التي تم إنشاؤها بشكل كبير، مما أدى إلى الحاجة إلى طرق فعالة لتحليل هذه البيانات ومعالجتها .يمكن أن توفر البيانات النصية رؤى قيمة حول تفضيلات العملاء وآرائهم وسلوكياتهم، مما يجعلها مصدرًا قيمًا للمعلومات للشركات والحكومات والمؤسسات.

يتضمن تحليل البيانات النصية تقنيات مختلفة، مثل معالجة اللغة الطبيعية (NLP) ، والتي تتضمن استخدام الخوارزميات لتحديد الأنماط في البيانات النصية ، وتحليل المشاعر ، والتي تتضمن تصنيف بيانات النص بناءً على المشاعر والآراء المعبر عنها ، ونمذجة الموضوع ، والتي يتضمن تحديد الموضوعات الأساسية في مجموعة معينة من البيانات النصية.

أصبحت مصادر البيانات النصية ذات أهمية متزايدة في مختلف المجالات، بما في ذلك التسويق والعلوم الاجتماعية والرعاية الصحية وغيرها .على سبيل المثال، يمكن للشركات استخدام البيانات النصية لمراقبة مشاعر العملاء وتحسين خدمة العملاء

في مجال الرعاية الصحية، يمكن استخدام البيانات النصية لتحديد أنماط سلوك المريض ومراقبة تفشي الأمراض وتحسين تقديم الرعاية الصحية .في العلوم الاجتماعية، يمكن استخدام البيانات النصية لدراسة السلوك البشري واستخدام اللغة والاتجاهات الثقافية.

ب-المعالجة المسبقة للنص

المعالجة المسبقة للنص هي عملية تحويل بيانات النص الخام إلى تنسيق يمكن تحليله وفهمه بسهولة بواسطة نموذج التعلم الآلي .تتضمن هذه العملية تقنيات مختلفة مثل الترميز، والاشتقاق، وإزالة الكلمات المتوقفة، والتطبيع.

- الخطوة الأولى في المعالجة المسبقة للنص هي الترميز، والذي يتضمن تقسيم النص إلى كلمات أو عبارات فردية تسمى الرموز المميزة .هذه الخطوة مهمة لأنها تسمح لنموذج التعلم الآلي بفهم بنية البيانات النصية.

- الخطوة التالية هي الاشتقاق، والتي تتضمن اختزال الكلمات إلى شكلها الجذري .هذا مهم لأنه يقلل من عدد الكلمات الفريدة في البيانات النصية، مما يسهل التحليل.

يعد إيقاف إزالة الكلمات أسلوبًا مهمًا آخر للمعالجة المسبقة، والذي يتضمن إزالة الكلمات الشائعة مثل "the"و "is" وما إلى ذلك من البيانات النصية .تضيف هذه الكلمات القليل من المعنى إلى النص ويمكن أن تسبب ضوضاء في التحليل.

تعد التسوية أيضًا تقنية مهمة في المعالجة المسبقة للنص، والتي تتضمن تحويل البيانات النصية إلى تنسيق موحد .يمكن أن يشمل ذلك تحويل كل النص إلى أحرف صغيرة، وإزالة علامات الترقيم، واستبدال الاختصارات بأشكالها الكاملة.

بشكل عام، تعد المعالجة المسبقة للنص خطوة حاسمة في تحليل البيانات النصية والتأكد من أن نماذج التعلم الآلي يمكنها فهم البيانات وتفسيرها بدقة.

ج-تقنيات تحليل النص

تشير تقنيات تحليل النص إلى الأساليب والأساليب المختلفة المستخدمة لاستخراج رؤى ومعلومات ذات مغزى من البيانات النصية .فيما يلي بعض تقنيات تحليل النص شائعة الاستخدام:

١. تحليل المشاعر :يتضمن تحديد المشاعر أو المشاعر المعبر عنها في جزء من النص، مثل الإيجابية أو السلبية أو الحيادية.
٢. نمذجة الموضوع :هذه طريقة إحصائية تُستخدم لتحديد الموضوعات أو الموضوعات في مجموعة من المستندات النصية.
٣. التعرف على الكيان المسمى :يتضمن ذلك تحديد واستخراج الكيانات المسماة مثل الأشخاص والمؤسسات والمواقع المذكورة في جزء من النص.
٤. تصنيف النص :يتضمن هذا تصنيف المستندات النصية إلى فئات محددة مسبقًا، مثل البريد العشوائي أو رسائل البريد الإلكتروني غير العشوائية، أو المقالات الإخبارية حسب الموضوع.
٥. تلخيص النص :يتضمن ذلك إنشاء ملخص لمستند نصي أطول تلقائيًا.
٦. تجميع النص :يتضمن ذلك تجميع المستندات المتشابهة معًا بناءً على محتواها.
٧. تحليل تشابه النص :يتضمن هذا مقارنة جزأين أو أكثر من النص لتحديد درجة التشابه بينهما.

يمكن تطبيق هذه الأساليب على مصادر بيانات نصية متنوعة، مثل منشورات وسائل التواصل الاجتماعي، ومراجعات العملاء، والمقالات الإخبارية، والأوراق الأكاديمية، من بين أمور أخرى .تعتبر المعالجة المسبقة الصحيحة للبيانات النصية ضرورية لضمان نتائج دقيقة وذات مغزى من تقنيات تحليل النص.

الفصل الثاني عشر :هندسة البيانات الضخمة

الفصل الثاني عشر :هندسة البيانات الضخمة

أ-منصات البيانات الضخمة

منصات البيانات الضخمة هي أطر برمجية مصممة لإدارة وتحليل مجموعات البيانات الكبيرة والمعقدة التي لا يمكن معالجتها بسهولة بواسطة أنظمة معالجة البيانات التقليدية .توفر هذه الأنظمة الأساسية أدوات لتخزين ومعالجة وتحليل كميات كبيرة من البيانات المهيكلة وغير المهيكلة، وعادةً ما تستخدم تقنيات الحوسبة الموزعة والمعالجة المتوازية.

تتضمن بعض منصات البيانات الضخمة الشائعة المستخدمة في الصناعة Apache Hadoop و Apache Spark و Apache Cassandra و Apache Flink و Amazon Web Services (AWS) Elastic MapReduce (EMR). غالبًا ما تُستخدم هذه الأنظمة الأساسية جنبًا إلى جنب مع أدوات إدارة البيانات والتحليلات الأخرى لتوفير حلول البيانات الضخمة الشاملة.

Hadoop، على سبيل المثال ، عبارة عن منصة بيانات ضخمة مفتوحة المصدر توفر تخزين الملفات الموزعة ومعالجتها باستخدام مجموعة من أجهزة الكمبيوتر .يستخدم نموذج برمجة يسمى MapReduce لتقسيم مهام معالجة البيانات الكبيرة إلى مهام أصغر يمكن معالجتها بالتوازي عبر عقد متعددة في مجموعة . من ناحية أخرى، يعد Spark محركًا لمعالجة البيانات في الذاكرة يمكنه التعامل مع كل من معالجة الدُفعات ومعالجة الدفق في الوقت الفعلي .تشتهر بسرعتها وأدائها، ويمكن استخدامها مع لغات برمجة مختلفة ، بما في ذلك Java و Python و.Scala

Cassandra هي قاعدة بيانات NoSQL موزعة يمكنها التعامل مع كميات كبيرة من البيانات المهيكلة وغير المهيكلة .يوفر توفرًا عاليًا، وتحملًا للأخطاء، وقابلية للتوسع، مما يجعله مناسبًا للتطبيقات ذات المهام الحرجة على نطاق واسع Flink .هو نظام أساسي لمعالجة البيانات الموزعة يدعم معالجة الدُفعات والدفق .إنه يوفر زمن انتقال منخفض وإنتاجية عالية وتحمل الأخطاء، مما يجعله مناسبًا لتطبيقات معالجة البيانات في الوقت الفعلي.

AWS EMR عبارة عن منصة بيانات ضخمة مُدارة توفر Hadoop و Spark وأدوات البيانات الضخمة الأخرى كخدمة .يسمح للمستخدمين بتوفير وإدارة مجموعات البيانات الضخمة على السحابة بسرعة وسهولة، دون الحاجة إلى بنية تحتية مخصصة.

لقد أحدثت منصات البيانات الضخمة ثورة في طريقة تعامل المؤسسات مع البيانات وتحليلها، مما مكنها من استخراج رؤى قيمة من مجموعات البيانات الكبيرة والمعقدة التي كان يتعذر الوصول إليها سابقًا .كما

أنها مهدت الطريق لتطوير تطبيقات وخدمات جديدة تعتمد على البيانات، مما أدى إلى ابتكارات في مختلف الصناعات.

ب- تخزين البيانات ومعالجتها

يعد تخزين البيانات ومعالجتها من الجوانب الحاسمة لإدارة البيانات الضخمة .مع تزايد حجم البيانات وسرعتها وتنوعها، لم تعد أنظمة تخزين ومعالجة البيانات التقليدية كافية .تتطلب البيانات الضخمة حلول تخزين ومعالجة موزعة يمكنها التوسع أفقيًا والتعامل مع كميات كبيرة من البيانات.

يعد Hadoop Distributed File System (HDFS) أحد أكثر أنظمة تخزين البيانات الضخمة شيوعًا . HDFSهو نظام ملفات موزع يوفر تخزينًا قابلًا للتطوير ومتسامحًا مع الأخطاء للبيانات الضخمة .إنه مصمم للتشغيل على أجهزة سلعة ويستخدم لتخزين ومعالجة مجموعات البيانات الكبيرة عبر عقد متعددة.

حل تخزين البيانات الضخمة الشائع الآخر هو قواعد بيانات .NoSQL على عكس قواعد البيانات الارتباطية التقليدية، يمكن لقواعد بيانات NoSQL التعامل مع البيانات غير المهيكلة وشبه المنظمة، مما يجعلها مثالية لتخزين البيانات الضخمة .تتضمن بعض قواعد بيانات NoSQL الشائعة MongoDB و Cassandraو.Couchbase

لمعالجة البيانات، يعد Apache Spark محركًا شائعًا لمعالجة البيانات الضخمة مفتوح المصدر يوفر طريقة سريعة وقابلة للتطوير لمعالجة مجموعات البيانات الكبيرة .يستخدم Spark المعالجة في الذاكرة ويمكن تشغيله على Hadoop و Kubernetes ومنصات البيانات الضخمة الأخرى.

تشمل تقنيات معالجة البيانات الضخمة الأخرى Apache Flink و Apache Storm و Apache .Beamتوفر هذه التقنيات إمكانات معالجة البيانات في الوقت الفعلي ويمكن استخدامها لمعالجة الدفق ومعالجة الدُفعات.

بالإضافة إلى منصات البيانات الضخمة والتقنيات هذه، توفر الحلول المستندة إلى السحابة مثلAmazon Web Services (AWS)و Microsoft Azure و Google Cloud Platform خيارات تخزين ومعالجة للبيانات الضخمة قابلة للتطوير وفعالة من حيث التكلفة .تقدم هذه الحلول المستندة إلى السحابة مجموعة متنوعة من خدمات تخزين ومعالجة البيانات الضخمة، بما في ذلك مجموعات Hadoop وقواعد بيانات NoSQL ومحركات معالجة البيانات.

باختصار ، تتطلب البيانات الضخمة حلول تخزين ومعالجة متخصصة يمكنها التعامل مع كميات كبيرة من البيانات وتوسيع نطاقها أفقيًا .تعد Hadoop وقواعد بيانات NoSQL والحلول المستندة إلى السحابة تقنيات تخزين ومعالجة البيانات الضخمة التي توفر حلولًا قابلة للتطوير وفعالة من حيث التكلفة لإدارة البيانات الضخمة.

ج- الحوسبة الموزعة

الحوسبة الموزعة هي نموذج حوسبي يسمح باستخدام أجهزة كمبيوتر متعددة للعمل معًا في مهمة أو مشكلة . يتم استخدامه على نطاق واسع في معالجة البيانات الضخمة، حيث تكون البيانات عادةً كبيرة جدًا بحيث لا يمكن معالجتها بواسطة جهاز واحد.

في الحوسبة الموزعة، يتم تقسيم عبء العمل إلى مهام أصغر، والتي يتم تعيينها بعد ذلك إلى أجهزة كمبيوتر متعددة في الشبكة .تعمل أجهزة الكمبيوتر بشكل مستقل في المهام المخصصة لها، ويتم دمج النتائج لإنتاج الناتج النهائي .تسمى هذه العملية الحوسبة المتوازية وهي أسرع بكثير من الحوسبة التسلسلية التقليدية.

أحد التحديات الأساسية في الحوسبة الموزعة هو التأكد من أن النتائج دقيقة ومتسقة .وهذا يتطلب التنسيق والتواصل بين أجهزة الكمبيوتر وآليات التعامل مع الأعطال والأخطاء .لمواجهة هذه التحديات، غالبًا ما تشتمل منصات الحوسبة الموزعة على آليات مدمجة للتسامح مع الخطأ وآليات التكرار، مثل النسخ المتماثل ونقاط التفتيش.

تم تطوير العديد من منصات الحوسبة الموزعة خصيصًا لمعالجة البيانات الضخمة، بما في ذلكApache Hadoopو .Apache Spark توفر هذه الأنظمة الأساسية إمكانات تخزين ومعالجة موزعة، بالإضافة إلى واجهات برمجة تطبيقات عالية المستوى لمعالجة البيانات وتحليلها.

أصبحت الحوسبة الموزعة مكونًا أساسيًا في معالجة البيانات الحديثة وتحليلها .إنه يمكّن من التعامل مع مجموعات البيانات واسعة النطاق والمهام الحسابية المعقدة التي قد تكون غير عملية أو من المستحيل القيام بها باستخدام موارد الحوسبة التقليدية .مع استمرار نمو البيانات الضخمة، ستستمر الحوسبة الموزعة في لعب دور حاسم في تمكين الرؤى والابتكارات القائمة على البيانات.

الفصل الثالث عشر :تقنيات البيانات الضخمة

الفصل الثالث عشر :تقنيات البيانات الضخمة

أ-النظام البيئي Hadoop

هو إطار عمل مفتوح المصدر مصمم للتخزين الموزع ومعالجة مجموعات البيانات الكبيرة على مجموعات من الأجهزة السلعية .تم إنشاؤه بواسطة دوغ كاتنج ومايك كافاريلا في عام 2005 وسمي على اسم لعبة الفيل .أصبح Hadoop أداة أساسية لمعالجة البيانات الضخمة نظرًا لقابلية التوسع والتسامح مع الأخطاء والقدرة على التعامل مع البيانات غير المنظمة وشبه المنظمة .يشتمل نظام Hadoop البيئي على العديد من الأدوات والأطر التي تتيح أنواعًا مختلفة من معالجة وتحليلات البيانات الضخمة.

المكونات الأساسية لنظام Hadoop البيئي هي نظام الملفات الموزعة (HDFS) Hadoop و MapReduce. HDFSهو نظام ملفات موزع يوفر وصولاً عالي الإنتاجية إلى بيانات التطبيق .يخزن البيانات بطريقة موزعة عبر الكتلة ويسمح بسهولة الوصول إلى البيانات بواسطة التطبيقات . MapReduceهو نموذج برمجة وإطار عمل معالجة يستخدمان لمعالجة مجموعات البيانات الكبيرة بطريقة موزعة .يقسم MapReduce البيانات إلى أجزاء أصغر ويعالجها بالتوازي على عقد مختلفة من الكتلة، ثم يقلل النتائج للحصول على الناتج النهائي.

بصرف النظر عن HDFS و MapReduce ، يشتمل نظام Hadoop البيئي أيضًا على أدوات وأطر عمل أخرى تعمل على توسيع قدرات .Hadoop على سبيل المثال، يعد Apache Pig نظامًا أساسيًا لإنشاء برامج MapReduce بلغة عالية المستوى تسمى .Pig Latin يوفر واجهة برمجة عالية المستوى لمعالجة البيانات وتحليلها، مما يسهل العمل مع Hadoop. Apache Hive هو إطار عمل آخر لتخزين البيانات يتيح الاستعلام عن مجموعات البيانات الكبيرة وإدارتها المخزنة في HDFS باستخدام لغة شبيهة بلغة SQL تسمى.HiveQL

تشمل المكونات الشائعة الأخرى لنظام Hadoop البيئي Apache HBase ، وهي قاعدة بيانات NoSQLموزعة وقابلة للتطوير يمكنها تخزين وإدارة كميات كبيرة من البيانات غير المهيكلة وشبه المنظمة ، و Apache Spark ، وهو محرك معالجة بيانات سريع ومرن يمكنه معالجة البيانات في الذاكرة و دعم معالجة البيانات المجمعة والتدفق والتفاعلية.

بشكل عام، يوفر نظام Hadoop البيئي منصة قوية ومرنة لمعالجة البيانات الضخمة وتحليلاتها .إن قدرتها على التعامل مع معالجة البيانات وتخزينها على نطاق واسع جعلها تحظى بشعبية بين الشركات والباحثين والمطورين.

ب-Spark

هو نظام حوسبة موزع مفتوح المصدر يستخدم لمعالجة البيانات الضخمة .تم تطويره في جامعة كاليفورنيا، بيركلي، وتحتفظ به الآن مؤسسة أباتشي للبرمجيات .يوفر Spark واجهة للبرمجة بعدة لغات، بما في ذلك Javaو Scala و Python ، ويتضمن مكتبات للتعلم الآلي ومعالجة الرسوم البيانية ومعالجة الدفق.

تتمثل إحدى الميزات الرئيسية لـ Spark في قدرتها على إجراء معالجة في الذاكرة، مما يسمح لها بمعالجة مجموعات البيانات الكبيرة بشكل أسرع بكثير من أنظمة معالجة الدفعات التقليدية .يحقق Spark هذا عن طريق تخزين البيانات مؤقتًا في الذاكرة عبر مجموعة من أجهزة الكمبيوتر، مما يسمح بوصول أسرع إلى البيانات عند الحاجة.

تم تصميم Spark أيضًا للعمل مع مجموعة متنوعة من أنظمة تخزين البيانات، بما في ذلك Hadoop Distributed File System (HDFS)وCassandra و .Amazon S3 تتيح هذه المرونة إمكانية استخدام Spark مع أنظمة تخزين البيانات الحالية، بدلاً من طلب إصلاح شامل للبنية التحتية.

يتضمن Spark العديد من المكونات الأساسية، بما في ذلك Spark Core و Spark SQL و Spark Streamingو .Spark MLlib يوفر Spark Core الوظائف الأساسية للنظام، بما في ذلك القدرة على إدارة البيانات الموزعة وتشغيل العمليات الحسابية على تلك البيانات .يوفر Spark SQL واجهة تشبه SQLللعمل مع البيانات المنظمة، بينما يوفر Spark Streaming طريقة لمعالجة تدفقات البيانات في الوقت الفعلي .يتضمن Spark MLlib خوارزميات التعلم الآلي لمهام مثل التجميع والتصنيف والانحدار.

بشكل عام، تعد Spark أداة قوية لمعالجة البيانات الضخمة وتحليلها، وتستخدم على نطاق واسع في الصناعة لمهام مثل التنقيب عن البيانات، والتعلم الآلي، ومعالجة البيانات في الوقت الفعلي .تجعله مرونته وقابليته للتوسع خيارًا شائعًا للمؤسسات التي تتطلع إلى معالجة مجموعات البيانات الكبيرة وتحليلها.

ج-قواعد بياناتNoSQL

قواعد بيانات NoSQL ، أو قواعد البيانات" غير العلائقية "، هي نوع من قواعد البيانات التي لا تعتمد على النموذج العلائقي التقليدي الذي تستخدمه قواعد البيانات العلائقية مثل MySQL أو .Oracle بدلاً من ذلك، تستخدم قواعد بيانات NoSQL مجموعة متنوعة من نماذج البيانات، بما في ذلك النماذج الموجهة للمستندات، والرسم البياني، والقيمة الرئيسية، ونماذج عائلة الأعمدة.

تتمثل إحدى ميزات قواعد بيانات NoSQL في قدرتها على التعامل مع كميات كبيرة من البيانات غير المهيكلة أو شبه المنظمة، مثل تلك الموجودة في خلاصات الوسائط الاجتماعية أو بيانات مستشعر إنترنت الأشياء، بشكل أكثر كفاءة من قواعد البيانات التقليدية .يمكنهم أيضًا التوسع أفقيًا، مما يتيح سهولة توزيع البيانات عبر خوادم أو مجموعات متعددة، ويمكنهم دعم التطوير السريع باستخدام تصميم مخطط مرن.

تتضمن بعض الأمثلة على قواعد بيانات NoSQL الشائعة MongoDB و Cassandra و HBase و Neo4j.كل من قواعد البيانات هذه لها نقاط القوة والضعف الخاصة بها وهي الأنسب لحالات استخدام محددة .على سبيل المثال، تشتهر MongoDB بنموذجها الموجه نحو المستندات وغالبًا ما تستخدم لتطبيقات الويب التي تتطلب بيانات ديناميكية سريعة التغير، بينما Neo4j هي قاعدة بيانات رسومية مثالية لتخزين والاستعلام عن البيانات شديدة الارتباط، مثل الشبكات الاجتماعية أو محركات التوصية.

على الرغم من مزاياها، فإن قواعد بيانات NoSQL لها أيضًا بعض العيوب .يمكن أن تكون أقل ملاءمة لأنواع معينة من البيانات والاستعلامات التحليلية من قواعد البيانات العلائقية التقليدية، وقد تتطلب معرفة أكثر تخصصًا لإعدادها وصيانتها .ومع ذلك، نظرًا لأن البيانات الضخمة ومعالجة البيانات في الوقت الفعلي أصبحت ذات أهمية متزايدة في الأعمال التجارية الحديثة، فمن المرجح أن تستمر قواعد بيانات NoSQLفي لعب دور مهم في مشهد تخزين البيانات ومعالجتها.

الفصل الرابع عشر :الحوسبة السحابية والبيانات الضخمة

الفصل الرابع عشر :الحوسبة السحابية والبيانات الضخمة

أ-أساسيات الحوسبة السحابية

الحوسبة السحابية هي تقنية تمكن المستخدمين من الوصول إلى موارد الحوسبة عبر الإنترنت، مثل الخوادم والتخزين وقواعد البيانات والتطبيقات والخدمات الأخرى .يسمح للشركات والأفراد باستخدام هذه الموارد عند الطلب، دون الحاجة إلى الاستثمار في البنية التحتية لتكنولوجيا المعلومات الخاصة بهم وإدارتها . أصبحت الحوسبة السحابية شائعة بشكل متزايد في السنوات الأخيرة بسبب قابليتها للتوسع والمرونة والفعالية من حيث التكلفة.

تتمثل إحدى الفوائد الرئيسية للحوسبة السحابية في القدرة على زيادة الموارد أو خفضها حسب الحاجة، مما يمكّن المستخدمين من الاستجابة بسرعة وسهولة للطلب المتغير .هذا ممكن بسبب استخدام المحاكاة الافتراضية، والتي تسمح للعديد من الأجهزة الافتراضية بالعمل على خادم فعلي واحد .يقدم موفرو السحابة عادةً مجموعة متنوعة من الخدمات، بما في ذلك البنية التحتية كخدمة (IaaS) والنظام الأساسي كخدمة (PaaS)والبرامج كخدمة.(SaaS)

يوفر IaaS للمستخدمين موارد حوسبة افتراضية، مثل الأجهزة الافتراضية والتخزين والشبكات، مما يسمح لهم ببناء البنية التحتية الخاصة بهم في السحابة .توفر PaaS نظامًا أساسيًا للمطورين لإنشاء التطبيقات ونشرها دون الحاجة إلى القلق بشأن البنية التحتية الأساسية .توفر SaaS تطبيقات يمكن الوصول إليها عبر الإنترنت، مثل البريد الإلكتروني وأدوات التعاون وبرامج إدارة علاقات العملاء.(CRM)

تقدم الحوسبة السحابية أيضًا مزايا من حيث التكلفة، حيث يدفع المستخدمون فقط مقابل الموارد التي يستخدمونها، بدلاً من الاضطرار إلى الاستثمار في أجهزتهم وبرامجهم الخاصة وصيانتها .بالإضافة إلى ذلك، يقدم موفرو الخدمات السحابية عادةً مستويات عالية من الموثوقية والأمان، مع تخزين البيانات في مراكز بيانات موزعة جغرافيًا ومحمية بواسطة إجراءات أمان متقدمة.

مع استمرار تطور الحوسبة السحابية، من المتوقع أن تصبح جزءًا لا يتجزأ من مشهد تكنولوجيا المعلومات، مع تطوير خدمات وتقنيات جديدة لتلبية احتياجات الشركات والمستهلكين على حد سواء.

ب-التخزين السحابي والحوسبة

التخزين السحابي والحوسبة هما جانبان أساسيان من جوانب الحوسبة السحابية .يشير التخزين السحابي إلى تخزين البيانات على الخوادم البعيدة التي يمكن الوصول إليها عبر الإنترنت، بينما تشير الحوسبة السحابية إلى تقديم خدمات الحوسبة مثل طاقة المعالجة والبرامج والتخزين عبر الإنترنت.

يوفر التخزين السحابي العديد من المزايا مقارنة بالتخزين التقليدي داخل الشركة، مثل قابلية التوسع والموثوقية والفعالية من حيث التكلفة .يقدم موفرو التخزين السحابي خيارات تخزين متنوعة، بما في ذلك تخزين الكائنات وتخزين الكتلة وتخزين الملفات .يعد تخزين الكائنات مثاليًا لتخزين البيانات غير المهيكلة مثل الصور ومقاطع الفيديو والمستندات .يتم استخدام التخزين الكتلي للتطبيقات التي تتطلب أداءً عاليًا، مثل قواعد البيانات، بينما يتم استخدام تخزين الملفات لتخزين الملفات التي تحتاج إلى الوصول إليها من قبل العديد من المستخدمين.

يتم تصنيف خدمات الحوسبة السحابية إلى ثلاثة أنواع رئيسية :البنية التحتية كخدمة (IaaS) ، والنظام الأساسي كخدمة (PaaS) ، والبرمجيات كخدمة .(SaaS) يوفر IaaS للمستخدمين إمكانية الوصول إلى موارد الحوسبة الافتراضية مثل الخوادم والتخزين والشبكات .توفر PaaS للمستخدمين نظامًا أساسيًا لتطوير التطبيقات ونشرها، بينما توفر SaaS للمستخدمين إمكانية الوصول إلى تطبيقات البرامج عبر الإنترنت.

تقدم الحوسبة السحابية العديد من المزايا مقارنة بالحوسبة التقليدية، مثل قابلية التوسع والفعالية من حيث التكلفة وإمكانية الوصول .تتيح الحوسبة السحابية للشركات توسيع نطاق مواردها الحاسوبية أو تقليصها بناءً على احتياجاتها، دون الحاجة إلى الاستثمار في أجهزة باهظة الثمن .تتيح الحوسبة السحابية أيضًا للشركات الوصول إلى بياناتها وتطبيقاتها من أي مكان، طالما كان لديها اتصال بالإنترنت.

ومع ذلك، فإن الحوسبة السحابية لها أيضًا بعض العيوب، مثل مخاوف الأمان والخصوصية .يثير تخزين البيانات على الخوادم البعيدة مخاوف بشأن أمان البيانات والخصوصية، وتحتاج الشركات إلى ضمان حماية بياناتها من الوصول غير المصرح به.

بشكل عام، يعد التخزين السحابي والحوسبة من المكونات الأساسية للحوسبة السحابية، وتتزايد اعتماد الشركات على الحوسبة السحابية للاستفادة من مزاياها.

ج-حلول البيانات الضخمة المستندة إلى السحابة

تشير حلول البيانات الضخمة المستندة إلى السحابة إلى استخدام البنية التحتية للحوسبة السحابية لتخزين ومعالجة وتحليل كميات كبيرة من البيانات .مع زيادة كمية البيانات التي يتم إنشاؤها كل يوم، تحتاج المؤسسات إلى حلول فعالة وفعالة من حيث التكلفة لإدارة هذه البيانات وتحليلها .توفر حلول البيانات الضخمة المستندة إلى السحابة للمؤسسات القدرة على تخزين البيانات ومعالجتها وتحليلها على نطاق واسع، مع تقليل الحاجة إلى البنية التحتية المحلية المكلفة.

تعتمد حلول البيانات الضخمة المستندة إلى السحابة على موارد الحوسبة والتخزين المستندة إلى السحابة، والتي يتم توفيرها بواسطة موفري الخدمات السحابية مثل Amazon Web Services (AWS) و Microsoft Azureو .Google Cloud Platform (GCP) يقدم موفرو السحابة هؤلاء مجموعة متنوعة من الخدمات التي تسمح للمؤسسات بتخزين ومعالجة كميات كبيرة من البيانات .على سبيل المثال، توفر AWS خدمة التخزين البسيط (S3) لتخزين البيانات و Elastic MapReduce (EMR) لمعالجة البيانات وتحليلها.

تتمثل إحدى ميزات استخدام حلول البيانات الضخمة المستندة إلى السحابة في أنها توفر سعة تخزين وقوة حوسبة غير محدودة تقريبًا .يمكن للمنظمات بسهولة زيادة أو تقليص موارد التخزين والحوسبة الخاصة بهم مع تغير احتياجات البيانات الخاصة بهم، دون الحاجة إلى القيام باستثمارات رأسمالية كبيرة في الأجهزة والبنية التحتية.

ميزة أخرى لحلول البيانات الضخمة المستندة إلى السحابة هي أنها متوفرة وموثوقة للغاية .يقدم مقدمو الخدمات السحابية عادةً آليات نسخ احتياطي قوية واستعادة البيانات بعد الكوارث، مما يضمن إمكانية الوصول إلى البيانات وحمايتها دائمًا.

توفر حلول البيانات الضخمة المستندة إلى السحابة أيضًا للمؤسسات القدرة على الوصول بسرعة وسهولة إلى التحليلات المتقدمة وأدوات التعلم الآلي .يقدم موفرو السحابة مجموعة من خدمات التحليلات والتعلم الآلي التي يمكن استخدامها لتحليل البيانات واكتساب الرؤى، مثل SageMaker من AWS ، وتعلم الآلة من Azure ، ومنصة الذكاء الاصطناعي الخاصة بـ.GCP

باختصار، توفر حلول البيانات الضخمة المستندة إلى السحابة للمؤسسات نهجًا مرنًا وفعالًا من حيث التكلفة وقابلًا للتطوير لإدارة وتحليل كميات كبيرة من البيانات .باستخدام موارد التخزين والحوسبة المستندة إلى مجموعة النظراء، يمكن للمؤسسات تخزين البيانات ومعالجتها على نطاق واسع، والوصول إلى التحليلات المتقدمة وأدوات التعلم الآلي، وضمان توافر وموثوقية عالية لبياناتهم.

الفصل الخامس عشر :أخلاقيات البيانات والخصوصية

الفصل الخامس عشر :أخلاقيات البيانات والخصوصية

أ-لوائح خصوصية البيانات

لوائح خصوصية البيانات هي قواعد وإرشادات تحكم كيفية قيام المؤسسات بجمع البيانات الشخصية وتخزينها واستخدامها وحمايتها .في السنوات الأخيرة، كانت هناك زيادة في انتهاكات البيانات والهجمات الإلكترونية وغيرها من الحوادث الأمنية التي أدت إلى فقدان أو سرقة البيانات الحساسة .وقد أدى ذلك إلى إدخال العديد من لوائح خصوصية البيانات حول العالم لحماية خصوصية وأمن البيانات الشخصية.

أحد أكثر لوائح خصوصية البيانات شهرة هو اللائحة العامة لحماية البيانات (GDPR) للاتحاد الأوروبي، والتي دخلت حيز التنفيذ في عام .2018 تتطلب اللائحة العامة لحماية البيانات (GDPR) من المنظمات الحصول على موافقة صريحة من الأفراد قبل جمع بياناتهم الشخصية ومعالجتها .كما يمنح الأفراد الحق في الوصول إلى بياناتهم الشخصية وتصحيحها ومسحها، فضلاً عن الحق في الاعتراض على معالجتها في ظروف معينة .تفرض اللائحة العامة لحماية البيانات (GDPR) أيضًا غرامات كبيرة في حالة عدم الامتثال، مع عقوبات تصل إلى 20 مليون يورو أو4 ٪ من الإيرادات السنوية العالمية، أيهما أعلى.

تشمل لوائح خصوصية البيانات البارزة الأخرى قانون خصوصية المستهلك في كاليفورنيا (CCPA) في الولايات المتحدة، والذي يمنح سكان كاليفورنيا حقوقًا معينة فيما يتعلق ببياناتهم الشخصية، وقانون حماية المعلومات الشخصية والوثائق الإلكترونية (PIPEDA) في كندا، والذي يحدد قواعد جمع واستخدام والكشف عن المعلومات الشخصية من قبل مؤسسات القطاع الخاص.

بالإضافة إلى هذه اللوائح، هناك أيضًا لوائح خاصة بالصناعة تحكم استخدام البيانات الشخصية في قطاعات معينة، مثل الرعاية الصحية والتمويل .قد تفرض هذه اللوائح متطلبات إضافية على المؤسسات لحماية خصوصية وأمن البيانات الشخصية.

بشكل عام، تعد لوائح خصوصية البيانات جانبًا مهمًا من إدارة البيانات وتلعب دورًا مهمًا في حماية خصوصية وأمن البيانات الشخصية .يجب على المؤسسات التأكد من امتثالها للوائح ذات الصلة في ولايتها القضائية وتنفيذ الضمانات المناسبة لحماية البيانات الشخصية من الوصول والاستخدام والإفشاء غير المصرح به.

ب-الاعتبارات الأخلاقية في علم البيانات

نظرًا لأن علم البيانات أصبح أكثر انتشارًا وقوة، تزداد أهمية الاعتبارات الأخلاقية .هناك عدد من الاعتبارات الأخلاقية الرئيسية التي يجب مراعاتها عند التعامل مع البيانات، لا سيما في سياق التعلم الآلي والذكاء الاصطناعي.

الخصوصية من أهم الاعتبارات الأخلاقية في علم البيانات .من الأهمية بمكان أن يحترم علماء البيانات خصوصية الأفراد الذين يعملون معهم بياناتهم، وأن يتخذوا خطوات لضمان الحفاظ على أمان هذه البيانات . هذا لا يعني فقط الامتثال للوائح ذات الصلة مثل اللائحة العامة لحماية البيانات، ولكن أيضًا تجاوز هذه اللوائح لضمان سيطرة الأفراد على بياناتهم وإعلامهم بكيفية استخدامها.

اعتبار أخلاقي مهم آخر هو الإنصاف .يمكن لنماذج التعلم الآلي في بعض الأحيان إعادة إنتاج أو حتى تضخيم التحيزات الموجودة في البيانات التي تم تدريبهم عليها .يمكن أن يؤدي هذا إلى نتائج غير عادلة وتمييز ضد مجموعات معينة من الناس .لذلك من الضروري أن يعمل علماء البيانات على تحديد هذه التحيزات والتخفيف منها، على سبيل المثال من خلال ضمان أن تكون بيانات التدريب ممثلة لجمهور أوسع واستخدام تقنيات مثل التدريب على الخصومة لمواجهة التحيز بفعالية.

الشفافية هي أيضا اعتبار أخلاقي مهم .من الضروري أن يتسم علماء البيانات بالشفافية بشأن الخوارزميات التي يستخدمونها والبيانات التي يعملون بها والقرارات التي يتخذونها .هذا مهم ليس فقط لبناء الثقة مع أصحاب المصلحة، ولكن أيضًا لضمان تحديد التحيزات المحتملة والقضايا الأخلاقية ومعالجتها.

أخيرًا، تعد المساءلة عنصرًا أساسيًا في علم البيانات الأخلاقي .يحتاج علماء البيانات والمؤسسات إلى تحمل المسؤولية عن القرارات والنتائج التي تنتجها نماذجهم .وهذا يشمل الاستعداد لشرح كيفية اتخاذ القرارات، والتأكد من وجود قنوات مناسبة للانتصاف إذا اعتقد الفرد أنه عومل بشكل غير عادل، وقبول المسؤولية عن الآثار الأخلاقية للنماذج التي ينتجونها.

باختصار، تعتبر الاعتبارات الأخلاقية عنصرًا أساسيًا في علم البيانات .يجب على علماء البيانات إعطاء الأولوية للخصوصية والإنصاف والشفافية والمساءلة لضمان استخدام الأدوات القوية التي يطورونها بطريقة مسؤولة وأخلاقية.

ج- إدارة البيانات والمساءلة

تشير إدارة البيانات والمساءلة إلى مجموعة السياسات والعمليات والضوابط التي تضمن التعامل السليم مع البيانات وإدارتها واستخدامها داخل المؤسسة .مع الاستخدام المتزايد للبيانات في عمليات صنع القرار، أصبحت إدارة البيانات والمساءلة جوانب مهمة لعلوم البيانات.

تتضمن الحوكمة الفعالة للبيانات وضع سياسات وإرشادات واضحة لإدارة البيانات، مثل تخزين البيانات والوصول إليها واستخدامها .ويشمل أيضًا تحديد الأدوار والمسؤوليات لإدارة البيانات داخل المؤسسة، مثل مضيفي البيانات ومالكي البيانات .تتضمن حوكمة البيانات أيضًا التأكد من أن البيانات دقيقة وكاملة وذات جودة عالية.

تشير مسؤولية البيانات إلى مسؤولية ومسؤولية الأفراد والمؤسسات عن البيانات التي يجمعونها ويخزنونها ويستخدمونها .يتضمن ضمان استخدام البيانات وفقًا للمعايير القانونية والأخلاقية وحمايتها من الوصول أو الاستخدام أو الكشف غير المصرح به.

تعد إدارة البيانات والمساءلة أمرًا ضروريًا في علم البيانات لضمان إدارة البيانات بشكل صحيح، وأن استخدامها يتوافق مع الأهداف والغايات التنظيمية .يمكن أن تساعد الإدارة الفعالة للبيانات والمساءلة المؤسسات على اتخاذ قرارات أفضل، وتقليل المخاطر المرتبطة باستخدام البيانات، وبناء الثقة مع أصحاب المصلحة، مثل العملاء والهيئات التنظيمية.

الفصل السادس عشر :تطبيقات علوم البيانات في الأعمال

الفصل السادس عشر :تطبيقات علوم البيانات في الأعمال

أ-تحليلات العملاء

تشير تحليلات العملاء إلى استخدام تقنيات تحليل البيانات لاكتساب رؤى حول سلوك العميل وتفضيلاته وأنماط الشراء .يسمح هذا النوع من التحليل للشركات بفهم عملائها بشكل أفضل واتخاذ قرارات أكثر استنارة حول كيفية تسويق وبيع منتجاتهم أو خدماتهم.

تتضمن تحليلات العملاء جمع البيانات وتحليلها من مجموعة متنوعة من المصادر، مثل تفاعلات العملاء، ووسائل التواصل الاجتماعي، وحركة مرور مواقع الويب، ومعاملات المبيعات .ثم يتم استخدام هذه البيانات لتحديد الأنماط والاتجاهات في سلوك العملاء، والتي يمكن استخدامها لعمل تنبؤات حول نشاط العميل في المستقبل.

أحد التطبيقات الشائعة لتحليلات العملاء هو تطوير نماذج تجزئة العملاء .من خلال تقسيم العملاء إلى مجموعات مختلفة بناءً على سلوكهم وخصائصهم، يمكن للشركات تصميم استراتيجيات التسويق والمبيعات الخاصة بها لتلبية احتياجات وتفضيلات كل مجموعة بشكل أفضل.

هناك تطبيق آخر لتحليلات العملاء وهو تطوير نماذج القيمة الدائمة للعميل .(CLV) هو مقياس يقدّر القيمة الإجمالية التي سيجلبها العميل إلى النشاط التجاري على مدار علاقته .من خلال تحليل البيانات المتعلقة بسلوك العملاء وأنماط الشراء، يمكن للشركات تطوير نماذج CLV أكثر دقة، والتي يمكن استخدامها لاتخاذ قرارات أفضل حول كيفية تخصيص الموارد وجهود التسويق المستهدفة.

بشكل عام، تعد تحليلات العملاء أداة قوية للشركات التي تتطلع إلى اكتساب ميزة تنافسية من خلال فهم عملائها بشكل أفضل واستخدام هذه المعرفة لدفع قرارات الأعمال.

ب-تحليلات التسويق

تحليلات التسويق هي فرع من تحليلات البيانات التي تركز على قياس وتحليل أداء التسويق وسلوك المستهلك لإبلاغ استراتيجية التسويق والتكتيكات .يتضمن جمع وتحليل وتفسير البيانات المتعلقة بتفاعلات المستهلك مع المنتجات والخدمات وقنوات التسويق.

أحد التطبيقات الرئيسية لتحليلات التسويق هو قياس الحملات التسويقية وتحسينها .يتضمن ذلك استخدام البيانات لتقييم فعالية قنوات التسويق المختلفة، مثل وسائل التواصل الاجتماعي والبريد الإلكتروني ومحركات البحث، ولتحديد المجالات التي يمكن فيها إجراء تحسينات .من خلال تتبع سلوك المستهلك عبر القنوات، يمكن للمسوقين اكتساب رؤى حول كيفية تفاعل المستهلكين مع علامتهم التجارية وتصميم رسائلهم وحملاتهم لتلبية احتياجاتهم وتفضيلاتهم بشكل أفضل.

مجال آخر مهم لتحليلات التسويق هو تقسيم العملاء .من خلال تحليل بيانات العملاء، يمكن للمسوقين تحديد مجموعات العملاء ذات الخصائص والسلوكيات المتشابهة وتطوير حملات تسويقية مستهدفة للمشاركة بشكل أفضل في هذه المجموعات والاحتفاظ بها .يمكن أن يساعد هذا النهج الشركات على زيادة رضا العملاء وولائهم مع تقليل تكاليف التسويق.

تلعب تحليلات التسويق أيضًا دورًا مهمًا في فهم قيمة عمر العميل (CLV) ، وهي القيمة المقدرة التي سيحققها العميل للأعمال التجارية على مدار علاقته .من خلال توقعCLV ، يمكن للمسوقين تطوير استراتيجيات لتعظيم قيمة كل عميل، مثل تحديد فرص زيادة المبيعات أو البيع المتبادل وتخصيص جهود الاحتفاظ للعملاء ذوي القيمة العالية.

بالإضافة إلى هذه التطبيقات، تُستخدم تحليلات التسويق أيضًا لقياس أداء العلامة التجارية وتتبع مشاعر وسائل التواصل الاجتماعي وتحليل نشاط المنافسين .من خلال الاستفادة من البيانات والتحليلات، يمكن للشركات اتخاذ قرارات مستنيرة بشأن استراتيجياتها وأساليبها التسويقية، مما يؤدي إلى زيادة مشاركة العملاء وولائهم وإيراداتهم.

ج-التحليلات المالية

التحليلات المالية هي مجال تحليل البيانات الذي يركز على استخدام النماذج الإحصائية والخوارزميات الرياضية وتقنيات التعلم الآلي لتقييم البيانات المالية واتخاذ قرارات مستنيرة .يمكن استخدام التحليلات المالية في مجموعة متنوعة من الإعدادات، مثل الخدمات المصرفية الاستثمارية وإدارة الأصول والتأمين والمحاسبة.

في الخدمات المصرفية الاستثمارية، يمكن استخدام التحليلات المالية للمساعدة في اتخاذ قرارات الاستثمار، وتقييم اتجاهات السوق، وتقييم المخاطر المرتبطة بالأدوات المالية المختلفة .يستخدم المحللون الماليون التحليلات المالية لإنشاء نماذج تتنبأ بأسعار الأسهم وتقدير احتمالية التخلف عن السداد وتوقع الإيرادات والأرباح.

في إدارة الأصول، يمكن استخدام التحليلات المالية لإنشاء محافظ استثمارية، وتقييم المخاطر، وتحسين توزيع الأصول .يستخدم المحللون الماليون التحليلات المالية لإنشاء نماذج تتنبأ بالعوائد المستقبلية، وتقدير مخاطر الاستثمارات المختلفة، وتحديد المزيج الأمثل للأصول.

في التأمين، يمكن استخدام التحليلات المالية لتقييم المخاطر المرتبطة بالسياسات المختلفة، والتنبؤ بالمطالبات المستقبلية، وتقدير حجم الاحتياطيات التي يجب وضعها جانباً لتغطية الخسائر المستقبلية . يستخدم المحللون الماليون التحليلات المالية لإنشاء نماذج تتنبأ باحتمالية وقوع أحداث مختلفة، وتقدير تكلفة المطالبات، وتحديد السعر الأمثل لبوليصة التأمين.

في المحاسبة، يمكن استخدام التحليلات المالية لتحليل البيانات المالية، وتقييم الأداء المالي، وتحديد الاتجاهات والأنماط في البيانات المالية .يستخدم المحللون الماليون التحليلات المالية لإنشاء نماذج تحدد مجالات الاحتيال المحتملة، وتقدير احتمالية الإفلاس، وتوقع الأداء المالي المستقبلي.

بشكل عام، تعد التحليلات المالية أداة قوية لاتخاذ قرارات مستنيرة في الصناعة المالية .من خلال تحليل كميات كبيرة من البيانات المالية واستخدام نماذج وخوارزميات متطورة، يمكن للمحللين الماليين اكتساب رؤى قيمة تساعدهم على اتخاذ قرارات أفضل وتقليل المخاطر وتحسين الأداء المالي.

الفصل السابع عشر :تطبيقات علوم البيانات في الرعاية الصحية

الفصل السابع عشر :تطبيقات علوم البيانات في الرعاية الصحية

أ-السجلات الصحية الإلكترونية

السجلات الصحية الإلكترونية هي نسخ رقمية من المعلومات الصحية للمرضى التي يتم جمعها وإدارتها والوصول إليها من قبل مقدمي الرعاية الصحية .تتضمن السجلات الصحية الإلكترونية عادةً مجموعة كبيرة من المعلومات، مثل التاريخ الطبي والأدوية ونتائج المختبر ودراسات التصوير وغيرها .أصبح استخدام السجلات الصحية الإلكترونية منتشرًا بشكل متزايد في السنوات الأخيرة بسبب الفوائد المحتملة التي تقدمها، مثل تحسين رعاية المرضى، وزيادة الكفاءة، وخفض التكاليف.

تم تصميم السجلات الصحية الإلكترونية ليتم مشاركتها بين مقدمي الرعاية الصحية، مما يسمح برعاية أكثر تنسيقًا وفعالية .هذا يعني أن الأطباء والممرضات وغيرهم من المتخصصين في الرعاية الصحية يمكنهم الوصول بسرعة وسهولة إلى التاريخ الطبي للمريض ونتائج الاختبارات والمعلومات الأخرى ذات الصلة .يمكن أن تساعد السجلات الصحية الإلكترونية أيضًا في تقليل الأخطاء وتحسين سلامة المرضى من خلال توفير التنبيهات والتذكيرات لمقدمي الرعاية الصحية.

في حين أن السجلات الصحية الإلكترونية لديها القدرة على تحسين الرعاية الصحية، إلا أن هناك مخاوف بشأن استخدامها .أحد الشواغل الرئيسية هو أمان وخصوصية بيانات المريض .تحتوي السجلات الصحية الإلكترونية على معلومات حساسة يجب حمايتها لضمان الحفاظ على خصوصية المريض .هناك أيضًا خطر حدوث انتهاكات للبيانات، مما قد يؤدي إلى فقدان معلومات المريض أو سرقتها.

مصدر قلق آخر هو إمكانية استخدام السجلات الصحية الإلكترونية للتمييز ضد المرضى .على سبيل المثال، يمكن لشركات التأمين أو أصحاب العمل استخدام بيانات السجل الصحي الإلكتروني لاتخاذ قرارات بشأن التغطية أو التوظيف بناءً على الحالة الصحية للفرد .هذا يمكن أن يؤدي إلى التمييز ضد الأفراد الذين يعانون من حالات طبية معينة.

بشكل عام، يعد استخدام السجلات الصحية الإلكترونية مجالًا سريع التطور للرعاية الصحية .في حين أنها توفر فوائد محتملة لرعاية المرضى، إلا أن هناك أيضًا اعتبارات مهمة حول الخصوصية والأمان والتمييز التي يجب معالجتها لضمان استخدام السجلات الصحية الإلكترونية بطريقة أخلاقية ومسؤولة.

ب-تحليل الصور الطبية

يعد تحليل الصور الطبية مجالًا مهمًا من مجالات التصوير الطبي الذي يتضمن تطبيق تقنيات الرؤية الحاسوبية والتعلم الآلي على الصور الطبية مثل الأشعة السينية والتصوير المقطعي المحوسب والتصوير بالرنين المغناطيسي وصور الموجات فوق الصوتية .الهدف من تحليل الصور الطبية هو استخراج معلومات مفيدة من الصور الطبية التي يمكن أن تساعد في التشخيص وتخطيط العلاج ومراقبة المرض.

يتضمن تحليل الصور الطبية عدة خطوات، بما في ذلك المعالجة المسبقة للصور، وتجزئة الصورة، واستخراج الميزات، والتصنيف .تتضمن المعالجة المسبقة للصور إزالة الضوضاء والتحف والمعلومات الأخرى غير المرغوب فيها من الصور .تجزئة الصورة هي عملية تقسيم الصورة إلى مناطق اهتمام متعددة .يتضمن استخراج الميزة تحديد الميزات أو الأنماط المهمة في بيانات الصورة التي يمكن استخدامها لمزيد من التحليل .يتضمن التصنيف تعيين مناطق الصورة أو الصورة إلى فئات محددة بناءً على الميزات المستخرجة.

تُستخدم تقنيات التعلم الآلي مثل التعلم الخاضع للإشراف والتعلم غير الخاضع للإشراف والتعلم العميق في تحليل الصور الطبية لتعلم الأنماط والميزات تلقائيًا من الصور الطبية .يتضمن التعلم الخاضع للإشراف تدريب نموذج التعلم الآلي على البيانات المصنفة، بينما يتضمن التعلم غير الخاضع للإشراف تدريب نموذج على البيانات غير المسماة .لقد نجحت تقنيات التعلم العميق مثل الشبكات العصبية التلافيفية بشكل خاص في تحليل الصور الطبية نظرًا لقدرتها على تعلم الميزات الهرمية من بيانات الصور الخام.

تحليل الصور الطبية له العديد من التطبيقات في مجال الرعاية الصحية، بما في ذلك الكشف عن الأمراض، وتجزئة الورم، وكشف الآفات، والتدخلات الموجهة بالصور .إن استخدام تحليل الصور الطبية لديه القدرة على تحسين دقة وكفاءة التشخيص وتخطيط العلاج ومراقبة المرض، مما يؤدي إلى نتائج أفضل للمرضى . ومع ذلك، هناك أيضًا تحديات مرتبطة بتحليل الصور الطبية، بما في ذلك الحاجة إلى كميات كبيرة من البيانات عالية الجودة، والحاجة إلى تعليقات توضيحية الخبراء وتفسير النتائج، والحاجة إلى ضمان سلامة وخصوصية بيانات المريض.

ج-الطب الشخصي

الطب الشخصي هو مجال من مجالات الطب يستخدم خصائص المريض الفردية مثل الوراثة ونمط الحياة والبيئة لتكييف العلاجات والتدخلات الطبية لمريض معين .يهدف هذا النهج إلى توفير علاجات أكثر فعالية وكفاءة يتم تخصيصها وفقًا للاحتياجات الفريدة لكل مريض على حدة.

تلعب تقنيات علم البيانات دورًا مهمًا في الطب الشخصي من خلال تحليل كميات كبيرة من البيانات لتحديد الأنماط والعلاقات بين خصائص المريض ونتائج العلاج .على سبيل المثال، يمكن استخدام خوارزميات التعلم الآلي لتحديد الطفرات الجينية التي تزيد من خطر الإصابة بأمراض معينة أو التنبؤ بكيفية استجابة المريض لدواء معين.

أحد التحديات الرئيسية في الطب الشخصي هو إدارة وتحليل كميات كبيرة من بيانات المريض، بما في ذلك السجلات الصحية الإلكترونية، والبيانات الجينومية، وبيانات التصوير .يجب وضع تدابير فعالة لإدارة البيانات والخصوصية لضمان الحفاظ على بيانات المريض آمنة وسرية.

يتمثل التحدي الآخر في تكامل البيانات من مصادر وأشكال مختلفة، حيث يعتمد الطب الشخصي على بيانات من مجالات متعددة مثل علم الجينوم والبروتيوميات والتصوير .يعد التعاون متعدد التخصصات بين الخبراء الطبيين وعلماء البيانات والمهندسين ضروريًا لتطوير مناهج فعالة للطب الشخصي.

بشكل عام، الطب الشخصي لديه القدرة على إحداث ثورة في الرعاية الصحية من خلال توفير علاجات أكثر دقة وفعالية مصممة لاحتياجات المريض الفردية .ومع ذلك، هناك حاجة إلى استمرار البحث والتطوير في علوم البيانات والتكنولوجيا الطبية لتحقيق هذه الإمكانات بشكل كامل.

الفصل الثامن عشر :تطبيقات علوم البيانات في العلوم الاجتماعية

الفصل الثامن عشر :تطبيقات علوم البيانات في العلوم الاجتماعية

أ-تحليلات وسائل التواصل الاجتماعي

تحليلات الوسائط الاجتماعية هي عملية جمع البيانات وتحليلها من مختلف منصات التواصل الاجتماعي مثل Twitter و Facebook و Instagram و LinkedIn وغيرها .الهدف من تحليلات الوسائط الاجتماعية هو استخراج رؤى ومعلومات قيمة حول المستخدمين والاتجاهات وأنماط سلوك مستخدمي الوسائط الاجتماعية .يمكن استخدام تحليلات الوسائط الاجتماعية لمجموعة متنوعة من الأغراض، بما في ذلك التسويق والعلاقات العامة وخدمة العملاء والبحث.

لإجراء تحليلات الوسائط الاجتماعية، يستخدم علماء البيانات عادةً مجموعة من التقنيات من معالجة اللغة الطبيعية والتعلم الآلي وتصور البيانات .أحد الأساليب الشائعة المستخدمة في تحليلات الوسائط الاجتماعية هو تحليل المشاعر، والذي يتضمن تحليل نغمة وعاطفة منشورات وسائل التواصل الاجتماعي لفهم شعور المستخدمين تجاه موضوع أو علامة تجارية معينة .تشمل التقنيات الأخرى تحليل الشبكة، والذي يتضمن تحليل العلاقات بين المستخدمين واتصالاتهم على منصات الوسائط الاجتماعية، ونمذجة الموضوع، والتي تتضمن تحديد الموضوعات الرئيسية والموضوعات التي تتم مناقشتها على وسائل التواصل الاجتماعي.

أصبحت تحليلات وسائل التواصل الاجتماعي ذات أهمية متزايدة في السنوات الأخيرة حيث أصبحت منصات وسائل التواصل الاجتماعي وسيلة أساسية للتواصل ومشاركة المعلومات لكثير من الناس .تستخدم الشركات تحليلات الوسائط الاجتماعية لاكتساب نظرة ثاقبة لعملائها وتفضيلاتهم، وكذلك لتتبع فعالية حملاتهم التسويقية

يستخدم الباحثون تحليلات وسائل التواصل الاجتماعي لدراسة الاتجاهات والسلوكيات الاجتماعية، بينما يستخدمها مسؤولو الصحة العامة لمراقبة انتشار الأمراض وتحديد الفاشيات المحتملة .بشكل عام، توفر تحليلات الوسائط الاجتماعية أداة قوية لفهم عالم الوسائط الاجتماعية سريع التطور والتفاعل معه.

ب-تحليل المشاعر

تحليل المشاعر هو حقل فرعي من معالجة اللغة الطبيعية (NLP) الذي يهدف إلى تحديد واستخراج المشاعر من نص معين ، مثل مراجعة أو تغريدة .إنه ينطوي على استخدام تقنيات التعلم الآلي واللغويات الحاسوبية لتحديد النغمة العاطفية الكلية لقطعة من النص تلقائيًا.

تتضمن عملية تحليل المشاعر عادةً عدة خطوات، بما في ذلك المعالجة المسبقة للنص، واستخراج الميزات، وتصنيف المشاعر .في المعالجة المسبقة للنص، يتم تنظيف النص وتحويله إلى تنسيق يمكن تحليله بواسطة خوارزميات التعلم الآلي .يتضمن هذا إزالة كلمات التوقف، أو الاشتقاق، وتقليل أبعاد البيانات.

في استخراج الميزات، يتم استخراج الميزات ذات الصلة من النص المعالج مسبقًا .يمكن أن تتضمن هذه الميزات تكرار الكلمات وعلامات جزء من الكلام وقواميس المشاعر .ثم يتم استخدام الميزات كمدخلات لخوارزمية التعلم الآلي لتصنيف المشاعر.

يمكن إجراء تصنيف المشاعر باستخدام مجموعة متنوعة من تقنيات التعلم الآلي، مثل بايز الساذج وآلات ناقلات الدعم ونماذج التعلم العميق .الهدف هو تدريب نموذج يمكنه التنبؤ بدقة بمشاعر النص الجديد غير المرئي.

يحتوي تحليل المشاعر على العديد من التطبيقات العملية، بما في ذلك مراقبة الوسائط الاجتماعية، وإدارة سمعة العلامة التجارية، وتحليل ملاحظات العملاء .يمكن أن يساعد المؤسسات على فهم كيف ينظر العملاء إلى منتجاتهم أو خدماتهم واتخاذ قرارات تستند إلى البيانات لتحسين رضا العملاء.

ج-تحليل الشبكة

يعد تحليل الشبكة أداة قوية في علم البيانات تسمح باستكشاف وفهم العلاقات المعقدة بين نقاط البيانات .في تحليلات الوسائط الاجتماعية، يمكن استخدام تحليل الشبكة لدراسة بنية الشبكات الاجتماعية وكيفية تدفق المعلومات من خلالها .يمكن أن يوفر تحليل طوبولوجيا الشبكة والمركزية والتكتل رؤى حول السلوك الاجتماعي للمستخدمين والمساعدة في تحديد المؤثرين الرئيسيين وقادة الرأي.

في تحليل الشبكة، يتم تمثيل العلاقات بين نقاط البيانات كشبكة أو رسم بياني، حيث تمثل كل عقدة نقطة بيانات وتمثل الحواف العلاقات فيما بينها .هناك مقاييس مختلفة لتحليل الشبكة تُستخدم لاستكشاف خصائص الشبكات، بما في ذلك الدرجة المركزية، وبين المركزية، ومعامل التجميع .تقيس مركزية الدرجة عدد الاتصالات التي تمتلكها العقدة، بينما تقيس المركزية البينية مدى أهمية العقدة في توصيل أجزاء مختلفة من الشبكة .يقيس معامل التجميع ميل العقد لتشكيل مجموعات أو مجموعات.

تحليل الشبكة له العديد من التطبيقات في مختلف المجالات، بما في ذلك وسائل التواصل الاجتماعي، وعلم الأحياء، والنقل، والاقتصاد .في وسائل التواصل الاجتماعي، يتم استخدام تحليل الشبكة لفهم سلوك المستخدم وتفضيلاته، والتنبؤ بمشاركة المستخدم وانتشار المحتوى، واكتشاف الأخبار المزيفة والبريد العشوائي .في علم الأحياء، يستخدم تحليل الشبكة لدراسة تفاعلات البروتين والبروتين والمسارات الأيضية .في النقل، يتم استخدام تحليل الشبكة لتحسين المسارات والجداول الزمنية، بينما في الاقتصاد، يتم استخدام تحليل الشبكة لدراسة الشبكات التجارية والمالية.

الفصل التاسع عشر :تطبيقات علوم البيانات في الحكومة

الفصل التاسع عشر :تطبيقات علوم البيانات في الحكومة

أ- كشف الاحتيال

يعد اكتشاف الاحتيال مجالًا مهمًا في علم البيانات يركز على تحديد الأنشطة الاحتيالية في مجالات مختلفة مثل التمويل والتأمين والتجارة الإلكترونية .الهدف الأساسي من اكتشاف الاحتيال هو تحديد أي أنشطة أو معاملات مشبوهة تنحرف عن أنماط السلوك العادية، مما يسمح بمنع الأنشطة الاحتيالية المحتملة في الوقت المناسب.

يتضمن اكتشاف الاحتيال استخدام تقنيات علوم البيانات المختلفة، بما في ذلك خوارزميات التعلم الآلي والنماذج الإحصائية وتقنيات التنقيب عن البيانات .تُستخدم هذه الأساليب لتحليل كميات كبيرة من البيانات من مصادر متعددة، بما في ذلك المعاملات المالية وسلوك العملاء ومصادر البيانات الأخرى ذات الصلة.

تتضمن عملية الكشف عن الاحتيال عدة خطوات، بما في ذلك الحصول على البيانات، وتنظيف البيانات، والمعالجة المسبقة للبيانات، واستخراج الميزات، وتطوير النموذج، وتقييم النموذج .يتضمن الحصول على البيانات جمع البيانات من مصادر مختلفة، بينما يتضمن تنظيف البيانات إزالة أي بيانات غير ذات صلة أو مكررة .تتضمن المعالجة المسبقة للبيانات تطبيع البيانات وتحويلها إلى تنسيق مناسب للتحليل .يتضمن استخراج الميزات تحديد الميزات ذات الصلة من البيانات التي يمكن استخدامها لتدريب نماذج التعلم الآلي.

يتضمن تطوير النموذج اختيار وتدريب خوارزميات التعلم الآلي المناسبة التي يمكنها اكتشاف الاحتيال بشكل فعال .تتضمن هذه الخوارزميات أشجار القرار، والانحدار اللوجستي، وآلات ناقلات الدعم، والشبكات العصبية .يتضمن تقييم النموذج اختبار النماذج المطورة على مجموعة جديدة من البيانات لتحديد دقتها وفعاليتها في اكتشاف الاحتيال.

بشكل عام، يعد اكتشاف الاحتيال مجالًا مهمًا في علم البيانات يساعد المؤسسات على منع الخسائر المالية وحماية عملائها من الأنشطة الاحتيالية .مع استمرار نمو حجم البيانات، ستستمر أهمية اكتشاف الاحتيال في الزيادة، مما يجعله مجالًا أساسيًا تستثمر فيه المؤسسات.

ب- منع الجريمة

منع الجريمة هو تطبيق مهم لعلم البيانات، والذي يتضمن استخدام البيانات للتنبؤ بالأنشطة الإجرامية ومنعها .يتضمن ذلك تحليل مصادر البيانات المختلفة، مثل سجلات الجرائم وتقارير الشرطة ووسائل التواصل الاجتماعي ومصادر أخرى لتحديد الأنماط والشذوذ الذي يمكن أن يشير إلى أنشطة إجرامية.

يمكن استخدام تقنيات علوم البيانات مثل التعلم الآلي والتعلم العميق وتحليل الشبكة لبناء نماذج تنبؤية يمكن أن تساعد وكالات إنفاذ القانون على توقع الأنشطة الإجرامية ومنعها .على سبيل المثال، يمكن لنماذج الشرطة التنبؤية تحليل بيانات الجريمة التاريخية لتحديد المناطق والأوقات التي ترتفع فيها معدلات الجريمة، وتخصيص موارد الشرطة وفقًا لذلك.

علاوة على ذلك، يمكن استخدام تقنيات تحليل الشبكات الاجتماعية لتحديد أنماط النشاط الإجرامي، مثل الروابط بين الأفراد المتورطين في الجريمة المنظمة أو الجماعات الإرهابية .يمكن أن يساعد هذا وكالات إنفاذ القانون على تعطيل الشبكات الإجرامية ومنع الأنشطة الإجرامية قبل حدوثها.

بشكل عام، يلعب علم البيانات دورًا مهمًا في منع الجريمة، ومن المتوقع أن يصبح تطبيقه أكثر أهمية في المستقبل حيث تصبح الأنشطة الإجرامية أكثر تعقيدًا وتعقيدًا.

ج- تحليل السياسة العامة

تحليل السياسة العامة هو مجال دراسي يتضمن استخدام تحليلات البيانات وطرق البحث الأخرى لتقييم فعالية السياسات والبرامج الحكومية .يسعى إلى فهم تأثير السياسات على المجتمع وتحديد طرق تحسينها.

يتمثل أحد الجوانب المهمة لتحليل السياسة العامة في جمع كميات كبيرة من البيانات وتحليلها .يمكن أن يشمل ذلك جمع البيانات حول المؤشرات الاجتماعية والاقتصادية، مثل معدلات الفقر ومعدلات البطالة والتحصيل التعليمي، بالإضافة إلى بيانات حول نتائج السياسة المحددة، مثل معدلات الجريمة أو النتائج الصحية أو المؤشرات البيئية.

جانب آخر مهم لتحليل السياسة العامة هو استخدام الأساليب الإحصائية والنمذجة لتحديد الأنماط والعلاقات في البيانات .يمكن أن يشمل ذلك تحليل الانحدار، وتحليل السلاسل الزمنية، وتقنيات التعلم الآلي، من بين أمور أخرى.

غالبًا ما يعمل محللو السياسة العامة بشكل وثيق مع صانعي السياسات لتطوير توصيات قائمة على الأدلة لتحسين السياسات والبرامج .قد يتعاونون أيضًا مع أصحاب المصلحة الآخرين، مثل المنظمات المجتمعية ومجموعات المناصرة، لضمان أن السياسات مصممة لتلبية احتياجات أولئك الذين يعتزمون خدمتهم.

في السنوات الأخيرة، جعلت التطورات في تحليلات البيانات وغيرها من التقنيات من الممكن تحليل مجموعات بيانات أكبر وأكثر تعقيدًا، مما يسمح بتحليل أكثر تعقيدًا للسياسة العامة .ونتيجة لذلك، يستمر مجال تحليل السياسة العامة في التطور والتوسع، مما يوفر فرصًا جديدة لفهم وتحسين تأثير السياسات الحكومية على المجتمع بشكل أفضل.

الفصل العشرون :مستقبل علم البيانات والبيانات الضخمة

الفصل العشرون :مستقبل علم البيانات والبيانات الضخمة

أ-الاتجاهات الناشئة في علم البيانات والبيانات الضخمة

تتطور مجالات علوم البيانات والبيانات الضخمة باستمرار، وتظهر اتجاهات وتقنيات جديدة بوتيرة سريعة. تشمل بعض أهم الاتجاهات الناشئة في هذه المجالات ما يلي:

١. حوسبة الحافة :تتضمن حوسبة الحافة معالجة البيانات بالقرب من المصدر بدلاً من إرسالها إلى موقع مركزي .يكتسب هذا الاتجاه شعبية حيث تسعى المؤسسات إلى تقليل زمن الوصول وتحسين سرعات المعالجة.

٢. التعلم الآلي التلقائي :التعلم الآلي التلقائي (AutoML) عبارة عن مجموعة من التقنيات التي تسمح بأتمتة الجوانب المختلفة لعملية التعلم الآلي .يكتسب AutoML شعبية لأنه يسهل على غير الخبراء تطبيق التعلم الآلي على بياناتهم.

٣. الذكاء الاصطناعي القابل للتفسير :يشير الذكاء الاصطناعي القابل للتفسير (XAI) إلى قدرة نماذج الذكاء الاصطناعي على تقديم تفسيرات لقراراتها وإجراءاتها .يكتسب هذا الاتجاه أهمية حيث تسعى المنظمات إلى زيادة الشفافية والجدارة بالثقة لأنظمة الذكاء الاصطناعي الخاصة بها.

٤. الحوسبة الكمومية :الحوسبة الكمومية هي نموذج حوسبة جديد يعتمد على مبادئ ميكانيكا الكم .من المتوقع أن يحدث ثورة في علم البيانات والبيانات الضخمة من خلال تمكين أنواع جديدة من الخوارزميات وتقنيات معالجة البيانات.

٥. التعلم الموحد :التعلم الموحد هو نهج التعلم الآلي الموزع الذي يسمح للنماذج بالتدريب على البيانات من أجهزة متعددة دون الحاجة إلى مركزية البيانات .يكتسب هذا الاتجاه شعبية لأنه يمكّن المؤسسات من تدريب النماذج على البيانات الحساسة مع الحفاظ على خصوصية البيانات.

٦. البيانات التركيبية :تشير البيانات التركيبية إلى البيانات التي تم إنشاؤها بشكل مصطنع والتي تم تصميمها لتقليد بيانات العالم الحقيقي .إنها تكتسب شعبية كطريقة للتغلب على مخاوف خصوصية البيانات وتحسين كفاءة مشاركة البيانات.

٧. نسيج البيانات :نسيج البيانات عبارة عن بنية تتيح التكامل السلس للبيانات وإدارتها عبر مصادر وتنسيقات متعددة .تكتسب أهمية حيث تسعى المنظمات إلى إدارة بياناتها بشكل أكثر كفاءة وفعالية.

بشكل عام، من المتوقع أن تشكل هذه الاتجاهات الناشئة مستقبل علم البيانات والبيانات الضخمة وتمكين المؤسسات من استخراج قيمة أكبر من بياناتها أكثر من أي وقت مضى.

ب-تأثير علم البيانات والبيانات الضخمة على المجتمع والقوى العاملة

كان لعلوم البيانات والبيانات الضخمة تأثير كبير على المجتمع والقوى العاملة في السنوات الأخيرة .أدى النمو الهائل للبيانات وتطوير الأدوات والتقنيات التحليلية المتقدمة إلى خلق فرص للمؤسسات لتحسين عمليات صنع القرار وتطوير منتجات وخدمات جديدة .ومع ذلك، فقد أثار هذا النمو أيضًا مخاوف بشأن خصوصية البيانات والأمان والأخلاق.

كان أحد أهم تأثيرات علم البيانات والبيانات الضخمة على المجتمع هو زيادة القدرة على استخلاص الرؤى من كميات هائلة من البيانات .وقد أدى ذلك إلى تحسينات كبيرة في الرعاية الصحية والنقل والتمويل وغيرها من الصناعات .على سبيل المثال، مكّن استخدام السجلات الصحية الإلكترونية الباحثين من دراسة أعداد كبيرة من المرضى وتحديد أنماط وعوامل الخطر للأمراض .في مجال النقل، أدى استخدام تحليلات البيانات إلى تطوير أنظمة نقل ذكية يمكنها تحسين تدفق حركة المرور وتقليل الازدحام.

ومع ذلك، فإن الاستخدام الواسع النطاق لعلوم البيانات والبيانات الضخمة أثار أيضًا مخاوف بشأن خصوصية البيانات وأمانها .أدى جمع البيانات الشخصية وتحليلها إلى حدوث حالات انتهاك للبيانات وإساءة استخدام المعلومات الشخصية، مما أدى إلى زيادة الدعوات إلى لوائح حماية البيانات الأكثر صرامة.

كان تأثير علم البيانات والبيانات الضخمة على القوى العاملة إيجابيًا وسلبيًا .أدى نمو علم البيانات إلى خلق فرص عمل جديدة، مثل علماء البيانات ومهندسي البيانات ومهندسي التعلم الآلي .ومع ذلك، أدى هذا النمو أيضًا إلى مخاوف بشأن إزاحة الوظائف التقليدية حيث تعتمد المنظمات بشكل متزايد على الأتمتة والتعلم الآلي لتحليل البيانات.

علاوة على ذلك، من المتوقع أن يستمر تطور تأثير علم البيانات والبيانات الضخمة على المجتمع والقوى العاملة .من المتوقع أن يؤدي تطوير تقنيات جديدة مثل الحوسبة المتطورة وشبكات الجيل الخامس وإنترنت الأشياء إلى توليد المزيد من البيانات، مما يؤدي إلى زيادة فرص التحليل والرؤى .ومع ذلك، من المتوقع أيضًا أن تخلق تحديات جديدة تتعلق بخصوصية البيانات وأمنها وأخلاقياتها، الأمر الذي يتطلب اهتمامًا وتنظيمًا مستمرين.

ج-الاعتبارات الأخلاقية والمجتمعية في مستقبل علم البيانات والبيانات الضخمة

مع استمرار تطور مجال علم البيانات والبيانات الضخمة، من المهم مراعاة الآثار الأخلاقية والمجتمعية لتأثيرها على المجتمع .مع تزايد كمية البيانات التي يتم إنشاؤها وتحليلها، هناك مخاوف بشأن خصوصية البيانات والتحيز وإساءة استخدام البيانات المحتملة.

أحد الاعتبارات الأخلاقية الرئيسية هو حماية حقوق خصوصية الأفراد .نظرًا لأنه يتم جمع المزيد من البيانات الشخصية وتحليلها، فمن المهم التأكد من أن البيانات مؤمنة بشكل صحيح واستخدامها فقط للأغراض المقصودة .بالإضافة إلى ذلك، هناك مخاوف بشأن احتمال حدوث انتهاكات للبيانات وتأثير ذلك على الأفراد إذا تم اختراق معلوماتهم الشخصية.

اعتبار آخر مهم هو احتمال التحيز في تحليل البيانات .يمكن أن تكون خوارزميات التعلم الآلي متحيزة بناءً على البيانات التي تم تدريبها عليها، مما قد يؤدي إلى نتائج غير عادلة أو تمييزية .من المهم التأكد من أن علماء البيانات على دراية بالتحيزات المحتملة واتخاذ خطوات للتخفيف منها.

يثير استخدام البيانات الضخمة أيضًا مخاوف بشأن إزاحة الوظائف وتأثيرها على القوى العاملة .نظرًا لأن الأتمتة والتعلم الآلي أصبحت أكثر انتشارًا، فقد تصبح بعض الوظائف قديمة بينما قد يتطلب البعض الآخر مهارات وتدريبًا جديدًا .من المهم للحكومات والمنظمات أن تأخذ في الاعتبار هذه الآثار وأن تتخذ خطوات لضمان تزويد العمال بالمهارات اللازمة للنجاح في سوق العمل المتغير.

بشكل عام، من المهم لعلماء البيانات والمنظمات وصانعي السياسات النظر بعناية في الآثار الأخلاقية والمجتمعية لعلوم البيانات والبيانات الضخمة .من خلال القيام بذلك، يمكننا ضمان استخدام هذه التقنيات بطرق تفيد المجتمع ككل مع تقليل الأضرار المحتملة.

الخاتمة:

أ-ملخص النقاط الرئيسية التي تناولها الكتاب

يقدم هذا الكتاب نظرة عامة شاملة على مجال علم البيانات والبيانات الضخمة .يبدأ بتعريف القارئ بأساسيات علم البيانات، بما في ذلك مصادر البيانات المختلفة وأنواع البيانات وتقنيات معالجة البيانات .ثم يناقش تقنيات تحليل البيانات المختلفة، بما في ذلك التعلم الآلي، واستخراج البيانات، والنمذجة الإحصائية . يغطي الكتاب أيضًا العديد من منصات البيانات الضخمة، مثل Hadoop و Spark ، بالإضافة إلى الحوسبة السحابية وحلول البيانات الضخمة المستندة إلى السحابة.

يتعمق الكتاب أيضًا في التطبيقات المختلفة لعلوم البيانات والبيانات الضخمة في مجالات مختلفة، مثل تحليلات العملاء، وتحليلات التسويق، والتحليلات المالية، والسجلات الصحية الإلكترونية، وتحليلات الوسائط الاجتماعية، واكتشاف الاحتيال .يسلط الضوء على الاعتبارات الأخلاقية والمجتمعية التي تأتي مع هذه التطبيقات والحاجة إلى إدارة البيانات والمساءلة.

يختتم الكتاب بمناقشة الاتجاهات الناشئة في مجال علم البيانات والبيانات الضخمة، مثل إنترنت الأشياء، وتكنولوجيا blockchain ، والحوسبة الكمومية .كما يناقش تأثير علم البيانات والبيانات الضخمة على المجتمع والقوى العاملة والاعتبارات الأخلاقية والمجتمعية التي يجب معالجتها في مستقبل علم البيانات.

بشكل عام، يوفر هذا الكتاب مقدمة شاملة وسهلة الوصول إلى مجال علم البيانات والبيانات الضخمة، تغطي المفاهيم والتقنيات والأنظمة الأساسية والتطبيقات الرئيسية .يسلط الضوء على أهمية حوكمة البيانات والمساءلة والاعتبارات الأخلاقية ويقدم لمحة عن مستقبل علم البيانات والبيانات الضخمة.

ب-الاتجاهات المستقبلية لعلوم البيانات وأبحاث وتطبيقات البيانات الضخمة.

مع استمرار نمو وتطور مجال علم البيانات والبيانات الضخمة، هناك العديد من الاتجاهات المستقبلية المحتملة للبحث والتطبيقات .يتمثل أحد مجالات التركيز المهمة في تطوير خوارزميات وتقنيات التعلم الآلي الأكثر تعقيدًا التي يمكنها التعامل مع مجموعات البيانات المعقدة والمتنوعة بشكل متزايد .سيتطلب ذلك بحثًا مستمرًا في مجالات مثل التعلم العميق والتعلم المعزز ومعالجة اللغة الطبيعية.

اتجاه مهم آخر للبحث المستقبلي هو دمج البيانات الضخمة مع التقنيات الناشئة مثل إنترنت الأشياء (IoT) و .blockchain سيسمح ذلك بتطوير أنظمة أكثر قوة وترابطًا يمكنها دعم مجموعة واسعة من التطبيقات بشكل أفضل، من الرعاية الصحية والتمويل إلى النقل والتصنيع.

بالإضافة إلى ذلك، هناك حاجة متزايدة لتحسين إدارة البيانات والاعتبارات الأخلاقية في مجال علم البيانات والبيانات الضخمة .ويشمل ذلك وضع لوائح ومعايير أقوى لخصوصية البيانات وأمنها، بالإضافة إلى زيادة الشفافية والمساءلة في عمليات صنع القرار التي تعتمد على البيانات.

أخيرًا، سيعتمد مستقبل علم البيانات والبيانات الضخمة على التعاون المستمر والبحث متعدد التخصصات بين الخبراء في مجالات مثل علوم الكمبيوتر والإحصاءات والمجالات الخاصة بالمجال مثل الرعاية الصحية والتمويل .من خلال العمل معًا لتطوير حلول وتطبيقات مبتكرة، يمكننا الاستمرار في إطلاق العنان للإمكانات الكاملة للبيانات الضخمة من أجل تحسين المجتمع.

د.المصادر

1. An Introduction to Statistical Learning with Applications in R by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. Springer, 2013. United States.
2. Applied Predictive Modeling by Max Kuhn and Kjell Johnson. Springer, 2013. United States.
3. Big Data: Principles and best practices of scalable real-time data systems by Nathan Marz and James Warren. Manning Publications, 2015. United States.
4. Data Analysis with Open Source Tools: A Hands-On Guide for Programmers and Data Scientists by Philipp K. Janert. O'Reilly Media, Inc., 2010. United States.
5. Data Mining: Practical Machine Learning Tools and Techniques by Ian H. Witten, Eibe Frank, and Mark A. Hall. Morgan Kaufmann, 2016. United States.
6. Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking by Foster Provost and Tom Fawcett. O'Reilly Media, Inc., 2013. United States.
7. Data Science from Scratch: First Principles with Python by Joel Grus. O'Reilly Media, Inc., 2015. United States.
8. Data Smart: Using Data Science to Transform Information into Insight by John W. Foreman. Wiley, 2013. United States.
9. Doing Data Science: Straight Talk from the Frontline by Cathy O'Neil and Rachel Schutt. O'Reilly Media, Inc., 2013. United States.
10. Machine Learning Yearning by Andrew Ng. Andrew Ng, 2019. United States.
11. Machine Learning: A Probabilistic Perspective by Kevin P. Murphy. The MIT Press, 2012. United States.
12. Machine Learning: An Artificial Intelligence Approach, Volume III edited by Ryszard S. Michalski, Jaime G. Carbonell, and Tom M. Mitchell. Morgan Kaufmann, 1986. United States.

13. Pattern Recognition and Machine Learning by Christopher Bishop. Springer, 2006. United States.
14. Practical Statistics for Data Scientists: 50 Essential Concepts by Peter Bruce and Andrew Bruce. O'Reilly Media, Inc., 2017. United States.
15. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython by Wes McKinney. O'Reilly Media, Inc., 2012. United States.
16. R Graphics Cookbook: Practical Recipes for Visualizing Data by Winston Chang. O'Reilly Media, Inc., 2013. United States.
17. The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling by Ralph Kimball and Margy Ross. Wiley, 2013. United States.
18. The Elements of Statistical Learning: Data Mining, Inference, and Prediction by Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Springer, 2009. United States.
19. The Signal and the Noise: Why So Many Predictions Fail-but Some Don't by Nate Silver. Penguin Books, 2015. United States.

نبذة عن المؤلف

احمد الجسار: مستشار ومدرب خبير في الاحصاء التطبيقي وتحليل البيانات معتمد من الاكاديمية الكندية للأعمال، ومحلل بيانات محترف ومعتمد من قبل شركة IBM الامريكية في الاحصاء وعلوم البيانات، واخصائي في الادارة الرقمية الحديثة أكمل دراسته العليا وحصل على شهادة الدبلوم العالي في (الاحصاء التطبيقي) من جامعة بغداد وشهادة (الماجستير) المصغر في ادارة الاعمال من معهد ادارة الاعمال الدولي في المانيا.

يعمل في مجال التدريب والاستشارات منذ اكثر من (١٢) عام , صدر له (١٣) كتاب و(٥) إصدارات رقمية في مجال الاحصاء التطبيقي وادارة الاعمال ، وله مجموعة من البحوث المنشورة في مجلات علمية محكمة, درب اكثر من (١٥,٠٠٠) شخص حول العالم حضورياً وعبر الانترنت, وحائز على عدد من الجوائز والاشادات.

يقدم خدمات الاستشارات الإحصائية للأفراد والمنظمات. والتي تشمل تحليل البيانات، والنمذجة الإحصائية، وتصميم التجارب، وتصميم المسح وتحليله، والبحث الإحصائي التطبيقي. تهدف هذه الخدمات إلى مساعدة الافراد والمنظمات في اتخاذ قرارات مستنيرة بناءً على البيانات والأدلة.


9 789922 682051