دليل علم البيانات وتعلم الآلة

Handbook of Data Science & Machine Learning

جامعة حمدان بن محمد الذكية
Hamdan Bin Mohammed Smart University

OpenAI
ChatGPT

نتاج علمي مشترك ما بين البشر ومنصة **ChatGPT**

بسم الله الرحمن الرحيم

يعتبر هذا العمل باكورة جهد وخلاصة ما بين مجموعة من المختصين في علم البيانات وتعلم الآلة ومجموعة من المتدربين الذين أنهوا دبلوم مهارات علم البيانات والذكاء الاصطناعي في رحاب جامعة حمدان الذكية - دبي - 2023/2022م ومنصة ChatGPT، يعتبر من أوائل الكتب المختصرة الناتجة من هذا التعاون. كتيب علوم البيانات وتعلم الآلة خاص بالمتخصصين بعلم البيانات وهندسة تعلم الآلة. هذا الكتيب هو دليل شامل لمجال علم البيانات وهندسة تعلم الآلة، ويغطي المفاهيم والأدوات والتقنيات الأساسية التي يستخدمها علماء البيانات لتحليل وتفسير البيانات، وخارطة الطريق لتنفيذ أي مشروع خاص بعلم البيانات وهندسة تعلم الآلة.

تم تصميم كتيب علوم البيانات وهندسة تعلم الآلة ليكون دليلاً لتنفيذ المشاريع المتخصصة بمجال علم البيانات وهندسة تعلم الآلة. ويساعدك على تطوير المهارات والمعرفة التي تحتاجها للنجاح كعالم بيانات. فإن هذا الكتيب يعد مورداً لا يقدر بثمن.

في هذا الكتيب سوف نتسلسل بالعرض اعتمادا على احتياجات مشاريع علم البيانات وتعلم الآلة، حيث البداية هي البيانات ويليها مرحلة فهم البيانات ومن ثم إعدادها وننتقل بعد ذلك نمذجة تقليل الأبعاد ومن ثم نماذج التقسيم وبعدها للعمود الفقري في مشاريع علم البيانات وتعلم الآلة وهي نماذج التنبؤ وبعد ذلك لنماذج المشاركة وأخيراً لأدوات تقييم النماذج.

فريق العمل & ChatGPT

2023

فريق العمل

فريق الخبراء والمدربين في جامعة حمدان الذكية - دبي

خبير علم البيانات والذكاء الاصطناعي أحمد العم

خبير علم البيانات والذكاء الاصطناعي حمزة أمين

فريق المتدربين في دبلوم مهارات علم البيانات والذكاء الاصطناعي في جامعة حمدان الذكية – دبي (مرتبين هجائياً)

د. آمنة الحارثي

أ. خليفة علي الخالدي

أ. صفية محمد السناني

أ. عايدة احمد مصلح

أ. علي حميد الشويهي

أ. عمر جعفر الهاشمي

أ. مريم عبدالقادر العطيشي

أ. مريم عبدالله المشجري

د. نورة عبدك البلوشي

منصة Open AI

ChatGPT

مقدمة

كتيب علوم البيانات وتعلم الآلة خاص بالمتخصصين بعلم البيانات وهندسة تعلم الآلة. هذا الكتيب هو دليل شامل لمجال علم البيانات وهندسة تعلم الآلة، ويغطي المفاهيم والأدوات والتقنيات الأساسية التي يستخدمها علماء البيانات لتحليل وتفسير البيانات، وخارطة الطريق لتنفيذ أي مشروع خاص بعلم البيانات وهندسة تعلم الآلة.

علم البيانات وهندسة تعلم الآلة هما مجالين متعددين التخصصات يجمع بين عناصر علوم الكمبيوتر والإحصاء وخبرة المجال لاستخراج الأفكار والمعرفة من البيانات المنظمة وغير المنظمة. إنه مجال سريع النمو يعمل على تحويل الصناعات ودفع الابتكار في مجموعة واسعة من المجالات، بما في ذلك الرعاية الصحية والتمويل وتجارة التجزئة والحكومة.

بصفتك عالم بيانات أو مهندس تعلم الآلة، ستكون مسؤولاً عن جمع البيانات وتنظيفها وتحليلها، فضلاً عن تطوير وتنفيذ النماذج الإحصائية وخوارزميات التعلم الآلي لاستخراج الرؤى واتخاذ قرارات تعتمد على البيانات. ستكون مسؤولاً أيضًا عن توصيل نتائجك إلى أصحاب القرار وتقديم نتائجك بشكل فعال.

تم تصميم كتيب علوم البيانات وهندسة تعلم الآلة ليكون دليلاً لتنفيذ المشاريع المتخصصة بمجال علم البيانات وهندسة تعلم الآلة. ويساعدك على تطوير المهارات والمعرفة التي تحتاجها للنجاح كعالم بيانات. فإن هذا الكتيب يعد مورداً لا يقدر بثمن.

في هذا الكتيب سوف نتسلسل بالعرض اعتمادا على احتياجات مشاريع علم البيانات وتعلم الآلة، حيث البداية هي البيانات ويليها مرحلة فهم البيانات ومن ثم إعدادها وننتقل بعد ذلك نمذجة تقليل الأبعاد ومن ثم نماذج التقسيم وبعدها للعمود الفقري في مشاريع علم البيانات وتعلم الآلة وهي نماذج التنبؤ وبعد ذلك لنماذج المشاركة وأخيراً لأدوات تقييم النماذج.

البيانات

البيانات هي شريان الحياة لعلم البيانات. إنها المادة الخام التي يستخدمها علماء البيانات لبناء نماذج إحصائية، وتدريب خوارزميات التعلم الآلي، واتخاذ قرارات تعتمد على البيانات. بدون البيانات، لن يكون هناك علم بيانات. تأتي البيانات في العديد من الأشكال والصيغ المختلفة، بما في ذلك البيانات المنظمة المخزنة في قواعد البيانات، والبيانات غير المهيكلة مثل النصوص والصور، وتدفق البيانات من أجهزة الاستشعار والمصادر الأخرى. يجب أن يكون علماء البيانات بارعين في العمل مع مجموعة واسعة من أنواع البيانات ومصادرها، ويجب أن يكونوا قادرين على استخراج رؤى ومعرفة ذات مغزى من مجموعات بيانات معقدة ومتنوعة.

تتمثل إحدى التحديات الرئيسية لعلم البيانات في الحاجة إلى العمل مع مجموعات بيانات كبيرة وغالبًا ما تكون فوضوية. يجب أن يكون علماء البيانات ماهرين في مجادلة البيانات، وعملية التنظيف، والتحضير، وتحويل البيانات إلى نموذج مناسب للتحليل. يجب أن يكونوا بارعين أيضًا في مجموعة متنوعة من الأدوات والتقنيات لتحليل البيانات وتفسيرها، بما في ذلك خوارزميات التعلم الآلي، والتحليل الإحصائي، وتصور البيانات.

كتيب علوم البيانات هو دليل شامل لمجال علم البيانات، ويغطي المفاهيم والأدوات والتقنيات الأساسية التي يستخدمها علماء البيانات لتحليل البيانات وتفسيرها. إنه مصمم لتقديم نظرة عامة واسعة على المجال ومساعدة القراء على تطوير المهارات والمعرفة التي يحتاجونها ليصبحوا علماء بيانات ناجحين.

وهناك العديد من الخصائص التي تجعل البيانات مناسبة لمشروعات علوم البيانات وتعلم الآلة:

- **الملاءمة**: يجب أن تكون البيانات ذات صلة بالمشكلة أو السؤال الذي تتم معالجته.
- **الدقة**: يجب أن تكون البيانات دقيقة وموثوقة.
- **الاكتمال**: يجب أن تكون البيانات كاملة، مع عدم وجود سجلات مفقودة أو غير كاملة.
- **الصلاحية**: يجب أن تكون البيانات صحيحة وخالية من الأخطاء أو التناقضات.
- **الاتساق**: يجب أن تكون البيانات متسقة، مع عدم وجود تناقضات أو معلومات متضاربة.
- **حسن التوقيت**: يجب أن تكون البيانات محدثة، لأن البيانات القديمة أو القديمة قد لا تكون مفيدة للمشكلة المطروحة.

وغالبا ما يتعين على علماء البيانات العمل مع مجموعات البيانات الكبيرة والمعقدة، وقد يحتاجون إلى استخدام مجموعة متنوعة من الأدوات والتقنيات لاستخراج الأفكار والمعرفة من البيانات. قد يشمل ذلك خوارزميات التعلم الآلي والتحليل الإحصائي وتصور البيانات والمزيد. وفيما يلي توضيح للأدوات والخوارزميات ذات الأهمية في ذلك، حيث نوضح في هذا الكتيب الحالات التي تستخدم بها كل حالة وخوارزمية مقسمة حسب خطوات المشروع.

1. فهم البيانات "Data Understanding"

مرحلة فهم البيانات هي عملية التعرف على البيانات التي تعمل بها. وإنها خطوة مهمة في عملية علم البيانات لأنها تساعدك على تحديد المشاكل المحتملة في البيانات، وفهم العلاقات بين المتغيرات المختلفة، وتحديد الخوارزميات والتقنيات التي من المحتمل أن تكون فعالة في تحليلك. تتضمن بعض المهام المحددة التي يتم تضمينها في فهم البيانات ما يلي:

- ✓ الحصول على البيانات واستيرادها.
- ✓ استكشاف البيانات وتحديد الأنماط.
- ✓ تصور البيانات لفهمها بشكل أفضل.
- ✓ تحديد أي نقاط بيانات مفقودة أو غير صالحة
- ✓ فهم السياق الذي تم فيه جمع البيانات.
- ✓ تحديد المنهجيات والخوارزميات المناسبة لاستخدامها في تحليلك.

ومن المهم أن تأخذ الوقت الكافي لفهم بياناتك تماما قبل الانتقال إلى الخطوات التالية في عملية علم البيانات، حيث سيساعدك ذلك على اتخاذ قرارات أفضل وتفسير نتائجك بدقة أكبر.

وعملية فهم البيانات هي عملية تكرارية: قد تحتاج إلى العودة وتكرار بعض الخطوات عدة مرات قبل أن يكون لديك فهم شامل للبيانات. والتعاون هو المفتاح؛ يمكن طلب المساعدة من الآخرين، مثل خبراء المجال أو علماء البيانات الآخرين، لتحسين فهم البيانات بشكل أفضل وتحديد الأنماط والاتجاهات المهمة. والتوثيق مهم أيضاً وخاصة أثناء تقدمك في عملية فهم البيانات، من المهم أن تتابع نتائجك وأي قرارات تتخذها. ستكون هذه الوثائق مفيدة لك وللآخرين أثناء الانتقال إلى الخطوات التالية في عملية علم البيانات.

بشكل عام، الهدف من فهم البيانات هو الحصول على فهم شامل للبيانات التي تعمل بها، بحيث يمكنك اتخاذ قرارات مستنيرة وتفسير نتائجك بدقة.

وفيما يلي عرض لأهم **أدوات فهم البيانات**، موضحاً غاية الاستخدام لكل منها؛

الإجراء

I. عرض البيانات "Display Data"

الأداة

1. Table

غاية الاستخدام

أداة **لعرض البيانات على طبيعتها أو عرض جزء منها اعتماداً على شرط ما**. وتعتبر أداة عرض للمخرجات الناتجة من أي عملية معالجة للبيانات. حيث تستخدم في أي موقع في دفق البيانات (Data Stream) لمشاريع علم البيانات.

الإجراء

II. تدقيق البيانات "Data Audit"

الأداة

1. Data Audit

غاية الاستخدام

أداة تستخدم **لأخذ نظرة أولية شاملة على البيانات**، من خلال عرض مصفوفة تربط كل متغير بمجموعة من الخصائص والمقاييس الإحصائية الوصفية لقيمه مع عرض قيمه برسم بياني.

الأداة

2. Descriptive Statistics

غاية الاستخدام

أداة تستخدم لتلخيص **البيانات الكمية** من خلال مقايس الإحصاء الوصفي لها، ويمكن أن تستخدم لقياس قوة واتجاه الارتباط الخطي بين المتغيرات.

الأداة

3. Hypothesis tests

غاية الاستخدام

أداة تستخدم **لدراسة وجود أو عدم وجود أثر معنوي (ذو دلالة إحصائية) بين المتغيرات** المستقلة والمتغير الهدف، أي على سبيل المثال هل النوع (ذكر/أنثى) كمتغير مستقل له أثر معنوي (أي ذو معنى) على مقدار الدخل المتنبئ به (المتغير الهدف) ويطلق على هذا الاختبار اسم (اختبارات فروض الفرق بين متوسطي مجتمعين مستقلين أو أكثر)، أو هل النوع (ذكر/أنثى) كمتغير مستقل له أثر معنوي (أي ذو معنى) على تعثر أو عدم تعثر العميل لسداد القرض (المتغير الهدف) ويطلق على هذا الاختبار اسم (اختبار الاستقلالية بين متغيرين وصفيين). وتطبق أيضا لدراسة وجود أو عدم وجود أثر لإجراء ما على أحد الخصائص (متغير مستقل)، على سبيل المثال دراسة أثر اتباع حمية ما على مستوى السكر في الدم، في هذه الحالة ندرس أثر الحمية من خلال متغيرين مستقلين الأول مستوى السكر في الدم لمجموعة من المرضى قبل اتباع الحمية والمتغير الثاني مستوى السكر في الدم لنفس المرضى بعد اتباع الحمية ويطلق على هذا الاختبار اسم (اختبارات فروض الفرق بين متوسطي مجتمعين غير مستقلين (مرتبطين).

الأداة

4. Correlation

غاية الاستخدام

أداة تستخدم **لقياس قوة واتجاه الارتباط الخطي بين المتغيرات** المستقلة مع الهدف والمستقلة مع بعضها البعض.

الإجراء

III. تمثيل البيانات "Visualization"

الأداة

1. Histogram

غاية الاستخدام

رسم مشابه للمدرج التكراري ما بين **متغير كمي وتكراره**، وتبرز أهميته من خلال القدرة على إدخال متغير وصفي أو أكثر على الرسمة ومشاهدة التغيرات التي تحدث عل القيم الكمية. حيث المتغيرات

الوصفية في هذا النوع يعبر عنها animation أو الألوان أو تقسيم الرسمة لعدة أجزاء.

الأداة

2. Collection

غاية الاستخدام

رسم مشابه للمدرج التكراري ما بين **متغيرين كميين**، وتبرز أهميته من خلال القدرة على إدخال متغير وصفي أو أكثر على الرسمة ومشاهدة التغيرات التي تحدث عل القيم الكمية. حيث المتغيرات الوصفية في هذا النوع يعبر عنها animation أو الألوان أو تقسيم الرسمة لعدة أجزاء.

الأداة

3. Distribution

غاية الاستخدام

رسم مشابه للتمثيل بالأعمدة ما بين **متغير وصفي وتكراره**، وتبرز أهميته من خلال القدرة على إدخال متغير وصفي أخر على الرسمة ويعبر عنه بالألوان. ويستفاد من هذا الرسم في مقارنة النسب والتكرار.

الأداة

4. Plot

غاية الاستخدام

رسم مشابه للمنحنى التكراري، حيث يمثل علاقة بين **متغيرين من أي نوع**، مع إمكانية إدخال متغير وصفي أو أكثر أو كمي على الرسمة ومشاهدة التغيرات التي تحدث عل الرسمة. حيث المتغيرات الوصفية والكمية في هذا النوع يعبر عنها animation أو تقسيم الرسمة لعدة أجزاء، أو الألوان، أو الأشكال، أو الأحجام، أو درجة اللون.

الأداة

5. Multi Plot

غاية الاستخدام

رسم مشابه للمنحنى التكراري، حيث يمثل علاقة **متغير من أي نوع على محور x** مع عدد من **المتغيرات الكمية على محور y**، مع إمكانية إدخال متغير وصفي أو اثنين على الرسمة ومشاهدة التغيرات التي تحدث عل الرسم. حيث المتغيرات الوصفية في هذا النوع يعبر عنها animation أو تقسيم الرسمة لعدة أجزاء.

الأداة

6. Time Plot

غاية الاستخدام

تستخدم هذه العقدة لرسم السلاسل الزمنية (**متغيرات كمية** تغيرها مرتبط **بالزمن**).

الأداة

7. Web

غاية الاستخدام

تستخدم لتمثيل **قوة الارتباط** بين **متغيرين وصفيين**، معتمدة في عملية الربط على عدة عوامل مثل النسبة أو تكرار الارتباط، وتعتبر الرسومات الناتجة من هذه العقدة من أهم المؤشرات الابتدائية لدراسة الارتباط والعلاقات بين المتغيرات الوصفية. وهذه العقدة تتيح خيارين للتمثيل، الأول (Web) ويستخدم في حالة دراسة **الارتباط** بين **جميع المتغيرات الوصفية** المحددة (حيث يحدد ارتباط كل متغير وصفي مع جميع المتغيرات الوصفية الأخرى). أما الخيار الثاني (Directed web) ويستخدم في حالة دراسة **الارتباط** بين **متغير وصفي واحد مع بقية المتغيرات الوصفي.**

الأداة

8. Graph board

غاية الاستخدام

أداة تمكننا من الاختيار من بين العديد من نواتج الرسوم البيانية المختلفة (bar charts, pie charts, histograms ,etc , ... ,scatterplots) في عقدة واحدة. حيث يقوم محلل البيانات باختيار الحقول التي يريد أن يرسم العلاقة بينها وطبقا لعدد الحقول ونوعها ترشح البرمجية الرسومات المتاحة لهذه البيانات ومن ثم يختار المحلل الرسمة الأكثر تناغماً مع البيانات، وتتيح العقدة اختيار تفاصيل أكثر للرسمة.

2. إعداد البيانات "Data Preparation "

تعد مرحلة إعداد البيانات خطوة مهمة في عملية علم البيانات. يتضمن تنظيف البيانات وتحويلها وتنظيمها بطريقة تجعلها مناسبة للتحليل والنمذجة. يمكن أن يكون إعداد البيانات عملية شاقة ومستهلكة للوقت، ولكنها ضرورية لنجاح أي مشروع لعلوم البيانات.

هناك العديد من المهام التي يتم تضمينها عادةً في إعداد البيانات:

- ✓ تنظيف البيانات: يتضمن تحديد وتصحيح الأخطاء والتناقضات في البيانات.
- ✓ تحويل البيانات: يتضمن ذلك تحويل البيانات إلى تنسيق أكثر ملاءمة للتحليل والنمذجة، مثل قياس البيانات أو تطبيعها.
- ✓ تكامل البيانات: يتضمن ذلك دمج البيانات من مصادر متعددة، مثل قواعد البيانات والملفات الثابتة.
- ✓ تقليل البيانات: يتضمن ذلك اختيار مجموعة فرعية من البيانات ذات الصلة بالمشكلة المطروحة، لتقليل التعقيد وتحسين الكفاءة.
- ✓ إثراء البيانات: يتضمن ذلك إضافة معلومات أو سياق إضافي للبيانات.

يعد إعداد البيانات عملية تكرارية، وقد يضطر علماء البيانات إلى أداء هذه المهام عدة مرات من أجل الحصول على البيانات في شكل مناسب للتحليل. من المهم أن يكون علماء البيانات على دراية بأدوات وتقنيات إعداد البيانات، بالإضافة إلى القيود والمزالق المحتملة لهذه التقنيات.

وفيما يلي عرض لأهم **أدوات إعداد البيانات**، موضحاً غاية الاستخدام لكل منها؛

الإجراء

I. اختيار البيانات "Selecting Data"

الأداة

1. Sample

غاية الاستخدام

أداة تستخدم هذه العقدة لاختيار أو تجاهل عينة من مصدر البيانات، **وغالباً لغايات استخدام هذه العينة في فحص النموذج الذي يتم بناءه.**

الأداة

2. Select

غاية الاستخدام

أداة تستخدم **لتضمين أو تجاهل مجموعة فرعية من السجلات في مصدر البيانات**، بناءً على حالة معينة أو شرط معين، مثلاً تضمين البيانات التي تكون قيمة المتغير Age > 30، ويستفاد منها لبناء سناريوهات مختلفة للنموذج.

الأداة

3. Filter

غاية الاستخدام

أداة تستخدم **لإجراء فلترة (إبطال أثر بدون حذف) على متغير** من حيث إهماله في عملية التحليل (أي عدم تأثيره على أي عمل مستقبلي على البيانات.

الأداة

4. Anonymize

غاية الاستخدام

أداة **تستخدم لإخفاء قيم متغير ما أو أكثر من خلال استبدال القيم الحقيقية بقيم بديلة، والهدف من ذلك المحافظة على سرية بعض البيانات** وخاصة في حالة استخدام النموذج من عدة أشخاص مثل إخفاء أسماء العملاء أرقام هواتفهم أو العناوين. مع الملاحظ أن المتغير الذي يتم إخفاء قيمة يصبح تأثيره في أي عملية مستقبلية للبيانات بقيمة الجديدة.

الأداة

5. Partition

غاية الاستخدام

أداة تستخدم هذه العقدة لتقسيم البيانات إلى عينات منفصلة لأغراض التدريب والاختبار وأغراض التحقق (اختيارية). **ويستفاد منها لفحص دقة نموذج التنبؤ بعد بناءه.**

الإجراء

II. تنظيف البيانات "Cleaning Data"

الأداة

1. Data Quality Report

غاية الاستخدام

أداة تستخدم **لتحليل دقيق لجودة للبيانات المتاحة قبل النمذجة**. والذي يساعد في كشف مشاكل البيانات ومن أشهرها؛

• تتضمن البيانات المفقودة قيمًا فارغة أو مشفرة على أنها عدم استجابة.

• أخطاء البيانات عادة ما تكون أخطاء مطبعية يتم إجراؤها عند إدخال البيانات.

• تتضمن أخطاء القياس البيانات التي تم إدخالها بشكل صحيح ولكنها تستند إلى نظام قياس غير صحيح.

• التناقضات في الترميز تشتمل عادة على وحدات القياس غير القياسية أو عدم تناسق القيمة،

• بيانات التعريف السيئة تتضمن عدم التطابق بين المعنى الظاهر للحقل والمعنى الموضح في اسم الحقل أو التعريف.

الأداة

2. Balance

غاية الاستخدام

تستخدم لتصحيح الاختلالات في نسب مجموعات البيانات. على سبيل المثال، لنفرض أن مجموعة البيانات تحتوي على ثلاث مناطق لتقديم الخدمة (شمال، وسط، جنوب)، وكانت منطقة الشمال تمثل 75٪ تقريباً من قيم متغير المنطقة، والوسط تقريباً 13%، والجنوب تشكل حوالي 12%. لذلك تعمل هذه الأداة على معالجة هذا الخلل، من خلال تصحيح نسب المجموعات بشكل متساوي تقريباً من خلال تكرار قيم مجموعات معينة أو حذف مشاهدات مجموعة ما، لبناء نماذج غير متحيزة.

الأداة

3. Filler

غاية الاستخدام

أداة تستخدم لاستبدال مجموعة من قيم المتغير أو القيم المفقودة في المتغير بقيم أخرى وفق شرط معين أو استبدال القيم المفقودة بقيم معينة.

الإجراء

III. بناء بيانات جديدة "Constructing Data "

الأداة

1. Derive

غاية الاستخدام

أداة تستخدم **لإضافة متغير (حقل) على سجل البيانات الأصلي تحت شروط معينة.**

الأداة

2. Binning

غاية الاستخدام

أداة تستخدم **لتحويل قيم متغير كمي (إنشاء متغير جديد) الى قيم اسمية (Nominal) بناء على مجموعة من العمليات الإحصائية**. أي تستخدم لتوليد متغير (حقل) جديد من النوع Nominal من خلال إعادة تصنيف قيم متغير متصل. ويستفاد منها في تحويل المتغير الكمي المتصل الى متغير اسمي (فئوي)، يستفاد منها في معالجة البيانات لتصبح قابلة لاستخدامها بخوارزميات أخرى لا تقبل إلا متغيرات اسمية مثل عقد الانحدار اللوجستي وبيز وغيرها.

الإجراء

IV. تنسيق البيانات "Formatting Data"

الأداة

1. Type

غاية الاستخدام

أداة تستخدم **لإجراء بعض التعديلات على خصائص المتغيرات أثناء بناء النموذج**، مثل مستوى القياس للمتغير أو هل هو متغير مستقل أو متغير هدف.

الأداة

2. Reorder Field

غاية الاستخدام

أداة تستخدم إعادة **ترتيب المتغيرات في ملف البيانات**، معتمدين في الترتيب على نوع المتغيرات أو أسمائها.

الأداة

3. Aggregate

غاية الاستخدام

أداة تستخدم **لتجميع البيانات الواقعة في ملف واحد وتمثل نفس الحالة (أي لها نفس قيمة متغير التعريف) وإظهار ملخص لقيم المتغيرات المقابلة لها**

الأداة

4. RFM (Recency, Frequency, Monetary) aggregate

غاية الاستخدام

أداة تستخدم في **التعاملات المالية والتسويقية، حيث تعمل على إظهار مجوعة من المعلومات المالية الخاصة بالعملاء. والمعلومات التي يتم تجميعها لكل عميل أو موظف أو وكيل هي**؛

Recency (الحداثة): وتدل على عدد الأيام التي مرت على أخر تعامل.

Frequency (التكرار): ويدل على عدد مرات التعامل.

Monetary (المالية): وتدل على مجموع المعاملات المالية التي أجريت.

الأداة

5. RFM Analysis

غاية الاستخدام

أداة تستخدم **لتقييم لنتائج عقدة RFM aggregate**، يستفاد منها لتصنيف العملاء من حيث الأفضلية اعتماداً على قيم (Recency, Frequency, Monetary)؛ حيث تعمل على إعطاء درجة لكل خاصية (Recency, Monetary ,Frequency) لكل عميل من حيث الأفضلية، وإعطاء أيضا درجة عامة لكل عميل تسمى RFM Score.

الأداة

6. Transform

غاية الاستخدام

أداة تستخدم **لإجراء مجموعة من التحويلات الرياضية على البيانات**، لوصولها لشكل التوزيع الطبيعي، لتصبح متوافقة مع مجموعة من الخوارزميات.

الإجراء
V. دمج البيانات "Integrating Data"
الأداة
1. Merge
غاية الاستخدام
أداة **تستخدم لدمج مجموعتين أو أكثر من البيانات مكونه من سجلات متماثلة ومتغيرات (حقول) مختلفة**، أو مكونه من سجلات مختلفة ومتغيرات (حقول) متماثلة، أو العمليتين معاً. ويمكن لهذه العقدة دمج ملفين أو أكثر مكونه من سجلات مختلفة ومتغيرات (حقول) مختلفة
الأداة
2. Append
غاية الاستخدام
أداة تستخدم لتجميع قواعد البيانات ذات الهياكل المماثلة، أي تجميع ملفين أو أكثر من الملفات ذات متغيرات (حقول) متماثلة وسجلات مختلفة. وفي حالة احتواء أحد الملفات على متغير أو أكثر غير موجودة في الملفات الأخرى، يمكن إظهاره والقيم المفقودة يتم تعريفها $null$.

3. طرق ونماذج تقليل الأبعاد
" Dimensional Reduction Methods & Models"

يتميز علم البيانات بالقدرة للتعامل مع البيانات الضخمة (Big Data)التي تتصف غالباً بتعدد المتغيرات كماً ونوعاً، مما يجعلها معقدة جداً لتطبيق تقنيات النمذجة عليها. فمثلاً لو تم التعامل من خلال علم البيانات مع مجموعة بيانات مكونة من ألف متغير (خاصية)، سوف ندرك صعوبة المهمة، وخاصة إذا لم يكن لدينا مشكلة محددة نرغب في حلها. إن وجود عدد كبير من المتغيرات هو نعمة ونقمة في نفس الوقت. من الرائع أن يكون لدينا الكثير من المتغيرات لبناء النماذج، ولكن هذا الأمر يمثل تحديًا بسبب الحجم.

ليس من المجدي إجراء تحليل ذو معنى لكل متغير بشكل منفرد. قد يحتاج هذا الأمر وقت طويل وتكلفة مالية عالية وتقنيات حاسوبية متطورة جداً، مما يؤثر سلباً على عجلة العمل وتطورها في الشركة أو المصنع أو البنك وغيرها. لذلك نحتاج إلى طريقة أفضل للتعامل مع البيانات ذات الأبعاد العالية (تحتوي عدد كبير من المتغيرات) أو (تحتوي عدد كبير من القيم المختلفة للمتغير الواحد) حتى نتمكن من استخلاص النماذج والأفكار من ذلك بشكل سريع. ولا يمكن تحقيق ذلك إلى من خلال تقنية النمذجة المسماة بتقليل الأبعاد Dimension Reduction.حيث أن نمذجة تقليل الأبعاد يمكن تطبيقها على المتغيرات المتعددة كماً ونوعاً أو على بيانات المتغير الواحد ذات العدد الكبير من القيم المختلفة، من خلال الاحتفاظ فقط بالمتغيرات الأكثر ملاءمة من مجموعة البيانات الأصلية من خلال إيجاد مجموعة أقل من المتغيرات المستقلة، بشرط عدم فقدان كثير من المعلومات والمحافظة على جودة النموذج المراد بناءه. ومن فوائد تطبيق تقليل الأبعاد على مجموعة بيانات؛

- ✓ تقليل المساحة المطلوبة لتخزين البيانات.
- ✓ تقليل وقت الحساب والتدريب لبناء النموذج وتطبيقه.
- ✓ الحد من الأبعاد حتى تقوم الخوارزمية بعملها بشكل مثالي. وخاصة في بعض الخوارزميات التي لا تعمل بشكل جيد عندما يكون لدينا أبعاد كبيرة.
- ✓ المساعدة في جعل عملية تمثيل البيانات بيانياً أكثر سهولة ومراقبة الأنماط بشكل أكثر وضوحاً.

وفيما يلي عرض لأهم **طرق ونماذج تقليل الأبعاد،** موضحاً غاية الاستخدام لكل منها؛

الإجراء

I. طرق تقليل الأبعاد

"Dimensions Reduction Methods"

الأداة

1. Missing Value Ratio

غاية الاستخدام

طريقة لنمذجة تقليل الأبعاد **في حالة لدينا نسبة كبيرة من القيم المفقودة لأحد أو مجموعة من المتغيرات المستقلة.**

الأداة

2. Low Variance Filter

غاية الاستخدام

طريقة لنمذجة تقليل الأبعاد **في حالة إذا كان لدينا متغير مستقل أو أكثر ذو تباين منخفض جداً أو صفري.**

الأداة

3. High Correlation Filter

غاية الاستخدام

طريقة لنمذجة تقليل الأبعاد **في حالة إذا كان لدينا متغيرات مستقلة ذات ارتباط عالي.**

الأداة

4. Mean differences were not significant Filter

غاية الاستخدام

طريقة لنمذجة تقليل الأبعاد **في حالة إذا كان لدينا متغير مستقل أو متغيرات لا يوجد لها اثر معنوي على المتغير الهدف.**

الأداة

5. Independence Filter

غاية الاستخدام

طريقة لنمذجة تقليل الأبعاد **في حالة لدينا إذا كان لدينا متغيرات وصفية مستقلة لا يوجد ارتباط بينها وبين المتغير الهدف من النوع الوصفي.**

الأداة

6. Backward Feature Elimination

غاية الاستخدام

طريقة لنمذجة تقليل الأبعاد **على المتغيرات المستقلة المتعددة كماً ونوعاً من خلال إجراء عمليات تقييم متسلسلة لأداء النموذج بعد بناءه.**

الأداة

7. Forward Feature Selection

غاية الاستخدام

طريقة لنمذجة تقليل الأبعاد **معاكسة للنمذجة السابقة الترشيح من خلال التغذية الراجعة Backward Feature Elimination. بدلاً من إزالة المتغيرات المستقلة، نحاول العثور على أفضل الميزات التي تعمل على تحسين أداء النموذج**

الإجراء

II. نماذج تقليل الأبعاد

"Dimensions reduction Models"

الأداة

1. Factor Analysis

غاية الاستخدام

خوارزمية لنمذجة تقليل الأبعاد على المتغيرات الكمية المستقلة المتعددة كماً ونوعاً، من خلال تقسيم المتغيرات الكمية المرتبطة الى مجموعات، حيث المتغيرات في كل مجموعة يكون لها ارتباط قوي فيما بينها، وارتباط منخفض مع متغيرات المجموعة (المجموعات) الأخرى.

وبعد ذلك يتم تعريف كل مجموعة كعامل (متغير). تتميز هذه العوامل بقلة عددها مقارنة بالمتغيرات الأصلية للبيانات

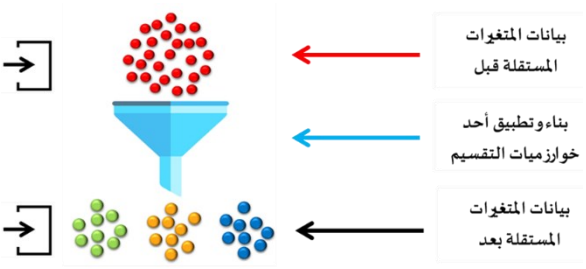
الأداة
2. Principal Component Analysis
غاية الاستخدام
خوارزمية لنمذجة تقليل الأبعاد على المتغيرات المستقلة المتعددة كماً ونوعاً، من خلال التغلب على تكرار المتغيرات (الميزات) في مجموعة البيانات. ويتم ذلك من خلال استخراج مجموعة جديدة من المتغيرات غير المرتبطة مع بعضها البعض تسمى (بالمكونات الأساسية) من مجموعة كبيرة من المتغيرات الكمية المدخلة الحالية. حيث تسعى هذه النمذجة لاستيعاب أكبر قدر ممكن من المعلومات الموجودة في البيانات وتضمينها في المكونات الأساسية.

الأداة
3. Feature Selection
غاية الاستخدام
خوارزمية لنمذجة تقليل الأبعاد على المتغيرات المستقلة المتعددة كماً ونوعاً، من خلال الاعتماد على مجموعة من الخصائص؛ إزالة المدخلات والسجلات الغير هامة والمثيرة للمشاكل، أو حقول الإدخال التي تحتوي على عدد كبير جدا من القيم المفقودة أو وجود اختلاف (Variation) كبير جدا أو صغير جدا.

4. نماذج التقسيم "Segmentation Models"

تعتبر نمذجة التقسيم (Segmentation) نمذجة تعلم الآلة غير الخاضع للإشراف(Unsupervised Learning) ، أي تنفذ على بالبيانات فقط (أي قيم المتغيرات المستقلة فقط دون قيم المتغير الهدف).

وتعمل نمذجة التقسيم (Segmentation) من خلال تقسيم السجلات الى مجموعات اعتماداً على خصائص متغير مستقل واحد أو أكثر، كل مجموعة لها صفات مشتركة، بحيث يتم التعبير عن كل سجل (صف) باسم أو رقم المجموعة التي تنتمي لها. وتختلف هذه المنهجية في علم البيانات عن التحليل الإحصائي، حيث عملية التقسيم في علم البيانات تنفذ من خلال عدد من الخوارزميات القادرة على تنفيذ عملية التقسيم اعتماداً على خليط من الخصائص المستنبطة من بيانات بعض أو كل المتغيرات المستقلة، وأيضا القدرة على التقسيم من خلال الاعتماد على خصائص لا تدرك بالطرق التقليدية. وينصح دائماً في البدء بهذه المنهجية بعد تقليل الأبعاد وخاصة عندما تكون البيانات غير مقسمة. والرسم التالي يوضح منهجية نمذجة التقسيم (Segmentation)؛



يسعى هذا النوع من النمذجة الى استخلاص التوزيع الاحتمالي (Probability Distribution) للبيانات. لغايات متعددة من أهمها؛

- الغاية الرئيسية، تقسيم البيانات الى مجموعة من النطاقات، معتمداً على التوزيع الاحتمالي التي تتبعه البيانات، حيث تتسم هذه النطاقات بخصائص

وصفات مشتركة وذات نسب معينة، ونطلق على هذا الغاية اسم التقسيم(Clustering) .

- وتستخدم أيضا في الكشف عم القيم الشاذة (Outlier)على مستوى المجموعة وليس على مستوى البيانات ككل، فمثلاً يمكن الكشف من خلالها على المسوقين المتقاعسين على مستوى المجموعة، أي من هو المسوق المتقاعس في مجموعة المسوقين الذين يبلغ متوسط مبيعاتهم اليومية $ 10,000، ومن هو المسوق المتقاعس في مجموعة المسوقين الذين يبلغ متوسط مبيعاتهم اليومية $ 5,000.
- يمكن الاستعانة بها في تقليل التباين لقيم أحد المتغيرات المستقلة، من خلال تنفيذ أحد خوارزميات التقسيم (Clustering) في تقسيم البيانات معمداً على المتغير ذو التباين المرتفع، حيث يمكن استبدال قيمه بأسماء الأقسام التي تم الحصول عليها، لأن كل قسم يمثل نطاق خاص من بيانات المتغير.
- وقد نلجأ لهذه النمذجة لتقليل أبعاد Dimensions) (Reductionالمتغيرات المستقلة، حيث أن تنفيذ أحد خوارزميات التعلم غير الخاضع للرقابة في تقسيم البيانات معتمداً على مجموعة من المتغيرات المستقلة، نحصل على أقسام كل قسم يعبر عن مجموعة مشتركة من نطاقات المتغيرات المستقلة. بحيث يمكن الاستغناء عن مجموعة المتغيرات المستقلة بأسماء الأقسام.
- ولخوارزميات التقسيم (Clustering) دور مهم ومهم جداً في إنشاء متغير تابع (الهدف) لمجموعة المتغيرات التي لا تحتوي عليه، ويظهر هذا الدور عند التعامل مع بيانات التعداد. فمثلاً عند التعامل مع بيانات التعداد السكاني لدولة ما، يمكن تقسيم البيانات التي تمثل قيم المتغيرات (الدخل، عدد أفراد الأسرة، مكان السكن، نوع السكن، قطاع الوظيفة، عدد الإصابات في الأسرة بأمراض مزمنة، ...)

الى ثلاث مجموعات من خلال أحد خوارزميات التقسيم (Clustering)، ومن خلال الخبرة بالوضع الاقتصادي وخط الفقر ومستوى الإنفاق والدخل وغيرها عند الخبير، تمكنه من أن يحكم على عناصر المجموعة الأولى أنهم فقراء والمجموعة الثانية متوسطين الحال والثالثة الأغنياء، وتسمى هذه الخبرة بالخبرة المعرفية، وفي هذه الهالة تم إنشاء متغير تابع (هدف) وقيمة الخبرة المعرفية (فقير / متوسط / غني).

وفيما يلي عرض لأهم **نماذج التقسيم**، موضحاً غاية الاستخدام لكل منها؛

الإجراء

I. التقسيم "Clustering"

الأداة

1. K-Mean

غاية الاستخدام

خوارزمية من خوارزميات التعلم الآلي الغير خاضع للإشراف، **تُستخدم لتقسيم مجموعة البيانات إلى عدد محدد مسبقًا من المجموعات**. تعمل الخوارزمية من خلال تسكين كل مدخل (صف) في مجموعة البيانات إلى الكتلة التي يكون مركزها (الوسط) الأقرب إليها، بناءً على بعض مقاييس المسافة. وتتصف العملية بالتكرارية، حيث تقوم الخوارزمية بإعادة حساب متوسط كل مجموعة وإعادة تعيين نقاط البيانات إلى المجموعات حتى يستقر تعيين نقاط البيانات إلى المجموعات. تتمثل إحدى مزايا مجموعة الوسائل K في أنها سريعة وسهلة التنفيذ، مما يجعلها خيارًا شائعا لتجميع مجموعات البيانات الكبيرة. ومع ذلك، يمكن أن تكون الخوارزمية حساسة للاختيار الأولي لمراكز الكتلة، وقد تختلف النتائج اعتمادا على التكوين الأولي. لذلك، من الشائع تشغيل الخوارزمية عدة مرات بتهيئة مختلفة واختيار التكوين الذي يعطي أفضل النتائج.

الأداة

2. Kohonen Networks

غاية الاستخدام

خوارزميات شبكات Kohonen، والمعروفة أيضًا باسم الخرائط ذاتية التنظيم (SOMs)، هي نوع من الشبكات العصبية الاصطناعية التي تُستخدم للتعلم غير الخاضع للإشراف. الهدف من شبكة Kohonen هو تعلم تمثيل منخفض الأبعاد (خريطة) لمجموعة بيانات عالية الأبعاد، بحيث يتم تعيين نقاط البيانات المتشابهة إلى المواقع القريبة في الخريطة.

تحتوي شبكات Kohonen على عدد من التطبيقات، بما في ذلك تصور البيانات والتعرف على الأنماط واستخراج الميزات. إنها مفيدة بشكل خاص لاستكشاف وفهم بنية مجموعات البيانات الكبيرة والمعقدة، حيث يمكنها الكشف عن أنماط وعلاقات قد لا تكون واضحة باستخدام تقنيات أخرى.

الأداة

3. TwoStep

غاية الاستخدام

خوارزمية من خوارزميات التعلم الآلي الغير خاضع للإشراف، تستخدم نمذجة TwoStep لتقسيم البيانات من أي نوع إلى مجموعات مميزة وخاصةً عندما لا يكون للمحلل علم أو قدره لتحديد ماهية هذه المجموعات ولا عددها عند البداية. حيث يحاول TwoStep اكتشاف أنماط في متغيرات الإدخال المحددة للتقسيم بشكل آلي. يتم تقسيم السجلات بحيث تكون السجلات داخل المجموعة متشابهة مع بعضها البعض، ولكن السجلات في مجموعات مختلفة تختلف عن بعضها.

تطبق نمذجة TwoStep من خلال منهجية تعمل على تقسيم البيانات الى مجموعات متجانسة في عناصرها من خلال خطوتين أو مرحلتين؛

الخطوة الأولى في هذه الخطوة يتم تمرير البيانات الخام بشكل منفرد ويتم توزيعها الى مجموعات جزئية ذات معنى (أي أن عناصر كل مجموعة تتسم بصفات خاصة مشتركة).

الخطوة الثانية في هذه الخطوة يتم تجميع ودمج المجموعات الجزئية الى مجموعات أكبر ذات معنى (أي أن عناصر كل مجموعة تتسم بصفات عامة مشتركة).

الإجراء

II. حديد البيانات الشاذة

" Anomaly Detection"

الأداة

1. Anomaly

غاية الاستخدام

خوارزمية من خوارزميات التعلم الآلي الغير خاضع للإشراف، تستخدم نمذجة Anomaly لتحديد القيم المتطرفة(Outliers) ، أو حالات غير عادية في البيانات (من جميع الأنواع). من خلال تقسيم البيانات الى مجموعات عناصر كل منها متجانسة ومن ثم تحديد القيم الشاذة لكل مجموعة.

ويتميز نمذجة Anomaly أن لديها القدرة على الكشف عن القيم الشاذة المتطرفة التي لا تتبع أي نموذج من نماذج فحص القيم المتطرفة، خلافا لأساليب النمذجة الأخرى التي تخزن نماذج محددة لفحص القيم المتطرفة. إذاً باستخدام هذا المؤشر من الممكن تحديد القيم المتطرفة حتى لو كانت لا تتفق مع أي نمط معروف سابقاً للقيم الشاذة، وقد يكون مفيداً بشكل خاص في التطبيقات، مثل الكشف عن الغش، حيث يمكن دائما أن تظهر أنماط جديدة من الغش غير المعروف لدينا.

5. نماذج التنبؤ "Prediction Models"

التنبؤ هو استخدام البيانات والمعلومات المتوفرة لدينا لعمل تخمينات مستنيرة حول ما يمكن أن يحدث في المستقبل أو لحالات شبيهة. حيث يمكن القيام بذلك باستخدام مجموعة متنوعة من التقنيات، مثل النمذجة الإحصائية وأشجار القرار وخوارزميات التعلم الآلي والتنبؤ المستقبلي. يمكن استخدامه في مجموعة متنوعة من المجالات، مثل التمويل والرعاية الصحية والتسويق والأمن وغيرها لاتخاذ قرارات أفضل وتحسين النتائج. ويندرج تحت مفهوم التنبؤ عدة تفرعات؛

- **الانحدار "Regression"**

الانحدار هو طريقة تُستخدم لنمذجة العلاقة بين متغير تابع من النوع **الكمي** ومتغير واحد أو أكثر من المتغيرات المستقلة. الهدف من الانحدار هو فهم كيفية ارتباط التغييرات في المتغيرات المستقلة بالتغيرات في المتغير التابع من النوع الكمي، ويتم استخدامه للتنبؤ بقيمة المتغير التابع الكمي بناءً على قيم المتغيرات المستقلة. ويمكن بناء نماذج الانحدار بعدة منهجيات؛ منهجية النماذج الإحصائية، ومنهجية نماذج شجرة القرار ومنهجية نماذج تعلم الآلة.

- **التصنيف "Classification"**

التصنيف تُستخدم لنمذجة العلاقة بين متغير تابع من النوع **الوصفي** ومتغير واحد أو أكثر من المتغيرات المستقلة. الهدف من التصنيف هو فهم كيفية ارتباط التغييرات في المتغيرات المستقلة بالتغيرات في المتغير التابع من النوع الوصفي، ويتم استخدامه للتنبؤ بقيمة المتغير التابع الوصفي بناءً على قيم المتغيرات المستقلة. ويمكن بناء نماذج الانحدار بعدة منهجيات؛ منهجية النماذج الإحصائية، ومنهجية نماذج شجرة القرار ومنهجية نماذج تعلم الآلة.

- **تحليل البقاء "Survival analysis "**

تحليل البقاء هو طريقة إحصائية تُستخدم لتحليل البيانات التي يتم فيها دراسة الوقت حتى وقوع حدث مثير للاهتمام (مثل الوفاة أو الفشل). غالبا ما يستخدم في مجالات مثل الطب والهندسة وعلم الاجتماع لفهم الوقت الذي يستغرقه حدوث حدث معين والتنبؤ به.

في تحليل البقاء، يُطلق على الوقت حتى وقوع الحدث "وقت البقاء"، ويشار إلى الحدث نفسه باسم "الفشل". الهدف من تحليل البقاء هو فهم العوامل التي تؤثر على وقت البقاء واستخدام هذه المعلومات لعمل تنبؤات حول احتمالية وقوع الحدث.

- **التنبؤ المستقبلي "Forecasting "**

وهي عملية وضع توقعات للمستقبل استناداً إلى البيانات السابقة والحالية والأكثر شيوعاً. ويجب أن تكون البيانات السابقة والحالية والمرات التنبؤ بها للمستقبل تابعة لبعضها البعض أي تخص حالة أو شخص معين. على سبيل المثال إذا علمت أوزان عادل من عمر سنة الى عمر 20 سنة فيمكن من نموذج التنبؤ المستقبلي الخاص بوزن عادل التوقع بوزنة عندما يصبح عمره 21 و22 و23 سنة مثلاً.

وفيما يلي عرض لأهم **نماذج التنبؤ**، موضحاً غاية الاستخدام لكل منها؛

الإجراء

I. الانحدار "Regression"

الأداة

1. Linear Regression

غاية الاستخدام

"نموذج إحصائي"

الانحدار الخطي هو نموذج انحدار إحصائي، يعتبر من أفضل النماذج لتمثيل العلاقة ما بين **مجموعة من المتغيرات المستقلة** ومتغير هدف من **النوع الكمي في حالة أن العلاقة خطية**، لقدرته على وصف وتمثيل العلاقات الخطية بينهما بشكل دقيق. ولكن يستلزم مجموعة من الشروط على البيانات لتطبيق هذا النموذج عليها **من** أهمها أن المتغير التابع يتبع التوزيع الطبيعي.

الأداة

2. Generalized linear Regression

غاية الاستخدام

"نموذج إحصائي"

الانحدار الخطي المعمم هو نموذج انحدار إحصائي، يستخدم لنمذجة العلاقة الخطية بين متغير تابع من **النوع الكمي** ومتغير واحد أو أكثر من **المتغيرات المستقلة. وهو امتداد لنموذج الانحدار الخطي، ويتسم بانة يفترض أن المتغير التابع يتبع توزيعا من عائلة التوزيعات الأسية، والتي تشمل التوزيعات الشائعة مثل التوزيع الطبيعي والتوزيع ذي الحدين وتوزيع بواسون. أي يتيح النموذج المعمم التعامل مع نطاق أوسع من المتغيرات التابعة، بما في ذلك المتغيرات التي لا يتم توزيعها بشكل طبيعي، فمن هنا جاء أنه أعم من الانحدار الخطي.**

الأداة

3. Generalized linear Mixed Models

غاية الاستخدام

"نموذج إحصائي"

النماذج المختلطة الخطية المعممة (GLMMs) هو نموذج انحدار إحصائي، يستخدم لنمذجة العلاقة الخطية بين متغير تابع من **أي نوع** ومتغير واحد أو أكثر من **المتغيرات المستقلة. وهو امتداد للنماذج الخطية المعممة (GLMs) ، والتي تُستخدم لنمذجة العلاقة بين متغير تابع ومتغير واحد أو أكثر من المتغيرات المستقلة. ويتسم بأنه لديه القدرة نمذجة العلاقة بين متغير تابع ومتغير واحد أو أكثر من المتغيرات المستقلة التي تحتوي على مزيج من التأثيرات الثابتة والعشوائية.**

في GLMM، يُفترض أن المتغير التابع يتبع توزيعا من عائلة التوزيعات الأسية، كما هو الحال في GLM. ومع ذلك، في GLMM، يتم التعامل مع بعض المتغيرات المستقلة (التأثيرات الثابتة) على أنها ثابتة، بينما يتم التعامل مع المتغيرات الأخرى (التأثيرات العشوائية) على أنها متغيرات عشوائية تختلف عبر الملاحظات. **يسمح هذا للنموذج بتفسير حقيقة أن تأثير بعض المتغيرات قد يختلف من ملاحظة إلى أخرى.**

الأداة

4. Logistic Regression

غاية الاستخدام

"نموذج إحصائي"

الانحدار اللوجستي هو نموذج انحدار إحصائي، **يستخدم للتنبؤ باحتمالية ظهور حدث ما**، فيجب النظر هنا أن المتغيرات المستقلة هي متغيرات **من أي نوع** والمتغير التابع **الحقيقي احتمالية ظهور حدث ما**، والمتغير التابع الوهمي هو **المتغير الذي أحد قيمة القيمة المراد التنبؤ باحتماليتها**. على سبيل المثال يرغب بنك في تحديد قيمة التأمين على القرض المقدم

للعميل اعتماداً على احتمالية تعثره، ففي هذه الحالة الهدف هو تحديد احتمالية تعثر أي عميل يرغب بأخذ قرض.

الأداة

5. Regression & Classification Tree

غاية الاستخدام

"نموذج أشجار القرار"

شجرة التنبؤ والانحدار هي أحد نماذج تعلم الآلة للانحدار، تعتمد على عملية التقسيم الغصني الثنائي. **تستخدم للتنبؤ بقيم متغير كمي، عندما تكون المتغيرات المستقلة ذات تسلسل هرمي واضح للميزات أو الفئات. واذا كان حجم البيانات طبيعي (ليس ضخم)، ودقة الفروقات بين قيم المتغيرات المستقلة ليست ذو أهمية.**

الأداة

6. Random Trees

غاية الاستخدام

"نموذج أشجار القرار"

الأشجار العشوائية هي أحد نماذج تعلم الآلة للانحدار، تعتمد على عملية التقسيم الساقي (عدد من الأشجار). **تستخدم للتنبؤ بقيم متغير كمي، عندما يكون لديك عدد كبير من المتغيرات المستقلة (المميزات) وعدد صغير نسبيًا من البيانات (الصفوف)، وعندما تريد إنشاء نموذج قوي في حالة وجود بيانات صاخبة "Noisy Data" (البيانات الصاخبة هي بيانات تحتوي على كمية كبيرة من المعلومات الإضافية التي لا معنى لها والتي تسمى الضوضاء) والقيم المتطرفة في البيانات. كما أنها سهلة التنفيذ نسبيًا ويمكنها التعامل مع البيانات عالية الأبعاد والقيم المفقودة جيدا.**

الأداة

7. Chi-squared Automatic Interaction Detection (CHAID)

غاية الاستخدام

"نموذج أشجار القرار"

شجرة مربع كاي للكشف عن التفاعل بشكل تلقائي هي أحد نماذج تعلم الآلة للانحدار، تعتمد على عملية التقسيم الغصني المتعدد. **تستخدم للتنبؤ بقيم متغير كمي، عندما تكون المتغيرات المستقلة خليط ما بين المتغيرات الكمية والوصفية، فضلاً عن البيانات المفقودة. يمكن للخوارزمية التعامل مع أي عدد من المتغيرات المستقلة، ولكنها تكون أكثر فاعلية عندما يكون هناك عدد صغير نسبياً من المتغيرات المستقلة ذات عدد كبير من القيم الكمية والوصفية.**

الأداة

8. XGBoost (eXtreme Gradient Boosting)

غاية الاستخدام

"نموذج أشجار القرار"

XGBoost هو تطبيق شائع وفعال لخوارزمية تعزيز التدرج، وهي تقنية تعلم آلي لبناء نماذج تنبؤية. تم تصميم XGBoost خصيصًا ليكون سريعًا وقابلًا للتطوير.

تعمل الخوارزمية من خلال بناء سلسلة من أشجار القرار، حيث تتعلم كل شجرة من أخطاء الشجرة السابقة. تسمى هذه العملية التعزيز، وهي تساعد على تحسين الأداء العام للنموذج.

تستخدم للتنبؤ بقيم متغير كمي، وخاصة في حالة وجود قيم مفقودة وكانت البيانات ضخمة واسعة النطاق منظمة أو غير منظمة. وهي مناسبة بشكل خاص للمهام التي تكون فيها الدقة مهمة ومجموعات البيانات كبيرة.

الأداة

9. K Nearest Neighbors

غاية الاستخدام

"نموذج تعلم الآلة"

خوارزمية أقرب K جوار KNN هي أحد نماذج تعلم الآلة للانحدار، **تستخدم للتنبؤ بقيم متغير كمي، حيث** تم تطوير نمذجة تحليل الجوار الأقرب KNN كأحد خوارزميات تعلم الآلة للتعرف على أنماط البيانات دون الحاجة إلى تطابق تام مع أي أنماط أو حالات مخزنة، أنما اعتماداً على حالات قريبة منها في الخصائص، أي نحكم على تماثل الحالات بقربها من بعضها البعض ونحكم على اختلاف الحالات ببعدها عن بعضها البعض. وبالتالي، فإن المسافة بين حالتين هي مقياس التماثل والاختلاف بينهما. وهذا يدل على أن نمذجة تحليل الجوار الأقرب هي أقل دقة في عمليات التصنيف والتنبؤ من النماذج الأخرى، **ولكن هي أكثر تفاعلية مع أي تراكيب معقدة من البيانات أو الحالات، والأكثر استجابة مع أي حالات جديدة أو طفرات في الحالات، ويفضل استخدام خوارزمية KNN إذا كان يوجد اختلاف ما بين بيانات التدريب والبيانات الجديدة المراد تطبيق النموذج عليها.**

الأداة

10. Support Vector Machine SVM

غاية الاستخدام

"نموذج تعلم الآلة"

خوارزمية آلة المتجهات الداعمة SVM هي أحد نماذج تعلم الآلة للانحدار، **تستخدم للتنبؤ بقيم متغير كمي،** وخاصة **في حالة أن أعداد حقول التنبؤ (المتغيرات المستقلة) كبير قد تصل للآلاف، وحجم بيانات (عدد صفوف قليل)**، وذلك لأنها تعتمد فقط على مجموعة فرعية من نقاط البيانات (تسمى متجهات الدعم) لعمل تنبؤات. **وأيضاً في الحالات التي لا تكون فيها البيانات قابلة للفصل خطياً، حيث يمكنها استخدام التحويلات غير الخطية للعثور على مستوى للفصل بين البيانات.**

ومع ذلك، فإن خوارزمية SVM لها أيضًا بعض المشاكل، حيث يمكن أن تكون خوارزمية SVM مكلفة من الناحية الحسابية عند التعامل مع مجموعات البيانات الكبيرة، لأنها تتطلب حل مشكلة تحسين تربيعية. ولا توفر خوارزمية SVM تقديرات احتمالية مباشرة، لذلك قد يكون من الصعب تقييم عدم اليقين في التنبؤات.

الأداة

11. Neural Networks (NN)

غاية الاستخدام

"نموذج تعلم الآلة"

خوارزمية الشبكات العصبية NN هي أحد نماذج تعلم الآلة للانحدار، **تستخدم للتنبؤ بقيم متغير كمي،** ولكن يجب أن تكون الخيار الأخير في حالة فشل النماذج الخوارزميات الأخرى، **وعند استخدامها هناك العديد من العوامل التي يجب مراعاتها عند اتخاذ قرار بشأن استخدام الشبكة العصبية لبناء نموذج انحدار على البيانات المهيكلة**:

حجم البيانات وتعقيدها: يمكن للشبكات العصبية التعامل مع مجموعات البيانات الكبيرة والمعقدة، ولكن يمكن أن يكون تدريبها وتشغيلها باهظ التكلفة من الناحية الحسابية. إذا كانت لديك مجموعة بيانات كبيرة بها العديد من الميزات، فقد تكون الشبكة العصبية خيارا جيدا.

جودة البيانات: تتطلب الشبكات العصبية كمية كبيرة من البيانات المصنفة من أجل التعلم بشكل فعال. إذا كانت لديك مجموعة بيانات كبيرة ذات فئات عالية الجودة، فقد تكون الشبكة العصبية خيارا جيدا.

بشكل عام، يمكن أن تكون الشبكات العصبية أداة قوية لبناء نماذج الانحدار على البيانات المهيكلة، لكنها ليست دائما الخيار الأفضل لكل مشكلة.

الإجراء

II. التصنيف "Classification"

الأداة

1. Generalized linear Regression

غاية الاستخدام

"نموذج إحصائي"

الانحدار الخطي المعمم هو نموذج تصنيف إحصائي، يستخدم لنمذجة العلاقة بين متغير تابع من النوع الثنائي أو **الترتيبي** ومتغير واحد أو أكثر من **المتغيرات المستقلة**. **ويتسم بانة يفترض أن المتغير التابع يتبع توزيعا من عائلة التوزيعات الأسية، والتي تشمل التوزيعات الشائعة مثل التوزيع ذي الحدين وتوزيع بواسون. أي يتيح النموذج المعمم التعامل مع نطاق أوسع من المتغيرات التابعة.**

الأداة

2. Generalized linear Mixed Models

غاية الاستخدام

"نموذج إحصائي"

النماذج المختلطة الخطية المعممة (GLMMs) هي نماذج تصنيف إحصائية، يستخدم لنمذجة العلاقة بين متغير تابع من **الوصفي** ومتغير واحد أو أكثر من **المتغيرات المستقلة. وهو امتداد للنماذج الخطية المعممة (GLMs)، والتي تُستخدم لنمذجة العلاقة بين متغير تابع ومتغير واحد أو أكثر من المتغيرات المستقلة. ويتسم بأنه لديه القدرة نمذجة العلاقة بين متغير تابع ومتغير واحد أو أكثر من المتغيرات المستقلة التي تحتوي على مزيج من التأثيرات الثابتة والعشوائية.**

في GLMM، يُفترض أن المتغير التابع يتبع توزيعًا من عائلة التوزيعات الأسية، كما هو الحال في GLM. ومع ذلك، في GLMM، يتم التعامل مع بعض المتغيرات المستقلة (التأثيرات الثابتة) على أنها ثابتة، بينما يتم التعامل مع المتغيرات الأخرى (التأثيرات العشوائية) على أنها متغيرات عشوائية تختلف عبر الملاحظات. يسمح هذا للنموذج بتفسير حقيقة أن تأثير بعض المتغيرات قد يختلف من ملاحظة إلى أخرى.

الأداة

3. Discriminant analysis

غاية الاستخدام

"نموذج إحصائي"

التحليل التميزي (Discriminant analysis) هو نموذج تصنيف إحصائي، يعرف **على انه من أساليب التحليل متعدد المتغيرات، لدية القدرة على تطوير معالم تمييزية جديدة عبارة عن تراكيب خطية يتم بنائها من المتغيرات المستقلة، لديها القدرة على التميز بين فئات المتغير التابع بطريقة مثالية، في حالة أن المتغيرات المدخلة نرغب في التعامل معها كوحدة واحدة، أي نريد دراسة أثرها مجتمعة في عملية التنبؤ**. وفي علم الإحصاء يلاحظ أن التحليل التميزي (analysis Discriminant) هو الاتجاه المعاكس لتحليل التباين المتعدد (MANOVA) في اتجاه واحد (الذي يسعى للتنبؤ بمجموعة من المتغيرات التابعة من متغير مستقل واحد فقط)، أي يمكن استخدام التحليل التميزي (Discriminant analysis) في حالة وجود متغير مستقل واحد (أسمي) وأكثر من متغير تابع (كمي)، حيث يتم تبديل المتغير المستقل ليصبح تابع والمتغيرات التابعة لتصبح مستقلة، ومن ثم نطبق التحليل التميزي فيتم دمج المتغيرات التابعة بمتغير واحد، ومن ثم نعتبر المتغير الجديد هو التابع والمتغير المستقل الأصلي هو المستقل ونطبق أي نمذجة تنبؤ. ويمكن استخدامه في نمذجة تقليل الأبعاد والتعبير عن المتغيرات بمتغير خطي واحد.

الأداة

4. Regression & Classification Tree

غاية الاستخدام

"نموذج أشجار القرار"

شجرة التنبؤ والانحدار هي أحد نماذج تعلم الآلة للانحدار، تعتمد على عملية التقسيم الغصني الثنائي. **تستخدم للتنبؤ بقيم متغير وصفي، عندما تكون المتغيرات المستقلة ذات تسلسل هرمي واضح للميزات أو الفئات. وإذا كان حجم البيانات طبيعي (ليس ضخم)، ودقة الفروقات بين قيم المتغيرات المستقلة ليست ذو أهمية.**

الأداة

5. Random Trees

غاية الاستخدام

"نموذج أشجار القرار"

الأشجار العشوائية هي أحد نماذج تعلم الآلة للانحدار، تعتمد على عملية التقسيم الساقي (عدد من الأشجار). **تستخدم للتنبؤ بقيم متغير وصفي، عندما يكون لديك عدد كبير من المتغيرات المستقلة (المميزات) وعدد صغير نسبيًا من البيانات (الصفوف)، وعندما تريد إنشاء نموذج قوي في حالة وجود بيانات صاخبة "Noisy Data" (البيانات الصاخبة هي بيانات تحتوي على كمية كبيرة من المعلومات الإضافية التي لا معنى لها والتي تسمى الضوضاء) والقيم المتطرفة في البيانات. كما أنها سهلة التنفيذ نسبيًا ويمكنها التعامل مع البيانات عالية الأبعاد والقيم المفقودة جيدا.**

الأداة

6. Chi-squared Automatic Interaction Detection (CHAID)

غاية الاستخدام

"نموذج أشجار القرار"

شجرة مربع كاي للكشف عن التفاعل بشكل تلقائي هي أحد نماذج تعلم الآلة للانحدار، تعتمد على عملية التقسيم الغصني المتعدد. **تستخدم للتنبؤ بقيم متغير وصفي، عندما تكون المتغيرات المستقلة خليط ما بين المتغيرات الكمية والوصفية، فضلاً عن البيانات المفقودة. يمكن للخوارزمية التعامل مع أي عدد من المتغيرات المستقلة، ولكنها تكون أكثر فاعلية عندما يكون هناك عدد صغير نسبياً من المتغيرات المستقلة ذات عدد كبير من القيم الكمية والوصفية.**

الأداة

7. XGBoost (eXtreme Gradient Boosting)

غاية الاستخدام

"نموذج أشجار القرار"

XGBoost هو تطبيق شائع وفعال لخوارزمية تعزيز التدرج، وهي تقنية تعلم آلي لبناء نماذج تنبؤية. تم تصميم XGBoost خصيصًا ليكون سريعًا وقابلًا للتطوير.

تعمل الخوارزمية من خلال بناء سلسلة من أشجار القرار، حيث تتعلم كل شجرة من أخطاء الشجرة السابقة. تسمى هذه العملية التعزيز، وهي تساعد على تحسين الأداء العام للنموذج.

تستخدم للتنبؤ بقيم متغير وصفي، وخاصة في حالة وجود قيم مفقودة وكانت البيانات ضخمة واسعة النطاق منظمة أو غير منظمة. وهي مناسبة بشكل خاص للمهام التي تكون فيها الدقة مهمة ومجموعات البيانات كبيرة.

الأداة
8. Quick, Unbiased, Efficient Statistical Tree (QUEST)
غاية الاستخدام
"نموذج أشجار القرار" شجرة QUEST هي أحد نماذج تعلم الآلة للتصنيف، تعتمد على عملية التقسيم الغصني الثنائي. **تستخدم للتنبؤ بقيم متغير وصفي**، تحتاج وقت أقل من وقت تحليلات C & R Tree سواء مع في حالة عدد المتغيرات الكبير أو في حالة وجود متغيرات ذات قيم متعددة. والهدف الثاني هو تقليل الميل الموجود في طرق التصنيف لتفضيل المدخلات التي تسمح بتقسيمات أكثر، أي الحقول ذات المدخلات المستمرة (مجموعة رقمية) أو تلك ذات الفئات العديدة. **تستخدم أذا كان حجم البيانات كبير وليست Big Data، وعدد قيم المتغير الهدف قليل.**

الأداة
9. C 5.0
غاية الاستخدام
"نموذج أشجار القرار" شجرة C5.0 هي أحد نماذج تعلم الآلة للتصنيف، تعتمد على عملية التقسيم الغصني المتعدد. تمتاز بقدرتها على تمثيل الشجرة بجمل شرطية تستخدم **للتنبؤ بقيم متغير وصفي**، تعتبر نماذج C5.0 قوية جدًا **في وجود مشكلات مثل البيانات المفقودة وعدد كبير من حقول الإدخال. وعادة لا تتطلب أوقات تدريب طويلة لتقديرها. بالإضافة إلى ذلك، تميل نماذج C5.0 إلى أن تكون أسهل في الفهم من بعض أنواع النماذج الأخرى، حيث أن القواعد المشتقة من النموذج لها تفسير مباشر للغاية. كما تقدم C5.0 طريقة قوية من التعليم المعزز لزيادة دقة التصنيف**

الأداة
10. Self-Learning Response
غاية الاستخدام
"نموذج تعلم الآلة" نموذج التعلم الآلي للاستجابة SLM هي أحد نماذج تعلم الآلة للتصنيف، **تستخدم للتنبؤ بالاستجابات لمجموعة من العروض مرتبة تنازلياً من الأعلى احتمالية استجابة الى الأقل**. وبمعنى أخر تعمل خوارزمية (SLR) على إنشاء نموذج يسمح بالتنبؤ بالعروض الأكثر ملاءمة للعملاء واحتمالية قبول العروض من قبلهم. أي في هذه النمذجة يوجد متغيرين هدف الأول يسمى Target field وهو يدل على العرض المقدم، والثاني يسمى Target response field وهو يدل على نتيجة الاستجابة للعرض من قبل العميل (قبول أو رفض). على سبيل المثال يرغب أحد البنوك إطلاق حملة ترويجية لبرامج القروض وهي (القرض الشخصي، قرض السكن، قرض شراء سيارة، قرض التعلم، قرض للسياحة)، ويرغب البنك بأن تكون الحملة الترويجية موجهة لعملائه، حيث يتم تقديم العرض الأمثل لكل عميل، أي الذي يُحتمل أن يتم قبوله على الأرجح. في هذه الحالة يمكنك استخدام نموذج التعلم الذاتي SLRM لتحديد خصائص العملاء الذين من المرجح أن يستجيبوا بشكل إيجابي بناءً على العروض الترويجية السابقة. **ويستفاد منها أيضا لتحديد الإجراء التالي الأمثل في حالة فشل الإجراء الأعلى احتمالية استجابة وهكذا**. وتتميز هذه الخوارزمية بالتعلم المستمر كلما أضيف بيانات على البيانات الأصلية دون الحاجة لإعادة بناء النموذج من البداية.

الأداة

11. K Nearest Neighbors

غاية الاستخدام

"نموذج تعلم الآلة"

خوارزمية أقرب K جوار KNN هي أحد نماذج تعلم الآلة للانحدار، **تستخدم للتنبؤ بقيم متغير وصفي، حيث** تم تطوير نمذجة تحليل الجوار الأقرب KNN كأحد خوارزميات تعلم الآلة للتعرف على أنماط البيانات دون الحاجة إلى تطابق تام مع أي أنماط أو حالات مخزنة، أنما اعتماداً على حالات قريبة منها في الخصائص، أي نحكم على تماثل الحالات بقربها من بعضها البعض ونحكم على اختلاف الحالات ببعدها عن بعضها البعض. وبالتالي، فإن المسافة بين حالتين هي مقياس التماثل والاختلاف بينهما. وهذا يدل على أن نمذجة تحليل الجوار الأقرب هي أقل دقة في عمليات التصنيف والتنبؤ من النماذج الأخرى، **ولكن هي أكثر تفاعلية مع أي تراكيب معقدة من البيانات أو الحالات، والأكثر استجابة مع أي حالات جديدة أو طفرات في الحالات، ويفضل استخدام خوارزمية KNN إذا كان يوجد اختلاف ما بين بيانات التدريب والبيانات الجديدة المراد تطبيق النموذج عليها.**

الأداة

12. Decision List

غاية الاستخدام

"نموذج تعلم الآلة"

نموذج قائمة القرارات هي أحد نماذج تعلم الآلة للتصنيف، **تستخدم للتنبؤ بالاستجابات لحدث ما.** وهي عبارة عن نموذج خليط ما بين نموذج الاستجابة الذاتية (SLRM) من حيث الغاية (تحديد احتمالية الاستجابة لحدث ما)، ونماذج شجرة القرارات Decisions Tree من حيث الهيكلية تقريباً. حيث يعمل نموذج قائمة القرارات على تقسيم المجتمع الى شرائح Segments، ومن ثم تحديد مجموعة من القرارات (الشروط) لكل شريحة، في حالة تحققها يستدل على الاستجابة للحدث. ويبرز اختلاف نموذج قائمة القرارات عن نموذج الاستجابة الذاتية (SLRM) ونماذج شجرة القرارات Decisions Tree فيما يلي؛

1. في نموذج الاستجابة الذاتية (SLRM) يتم بناء نموذج عام لكل مصدر البيانات، في نموذج قائمة القرارات Decision List يتم وضع قرارات (شروط) لكل شريحة من شرائح المجتمع.

2. نموذج الاستجابة الذاتية (SLRM) غير محدد المعالم، نموذج قائمة القرارات Decision List node يدرج قائمة بالقرارات واضحة المعالم لكل شريحة من شرائح المجتمع وقابلة للتعديل والحذف والإضافة ويستفاد منها في مرحلة التحليل التشخيصي (Diagnostic Analytics). أي يمكن الاستفادة من هذا النموذج ليس فقط في التصنيف، أنما أيضا في تقسيم المجتمع لشرائح ذات مواصفات محددة لغايات التعامل مع كل شريحة على حدة.

3. في نماذج شجرة القرارات Decisions Tree يتم بناء القرارات بشكل متسلسل؛

المتغير الهدف ← المعيار الأول ← المعيار الثاني ← ... ← القرار

حيث يطبق هذا التسلسل على جميع سجلات البيانات، أما في نموذج قائمة القرارات Decision List يتم بناء تسلسل خاص بكل شريحة وليس شرطاً أن يحتوي كل تسلسل جميع المتغيرات المستقلة المحددة لبناء النموذج.

الأداة

13. Bayes Network

غاية الاستخدام

"نموذج تعلم الآلة"

خوارزمية شبكة بيز هي أحد نماذج تعلم الآلة للتصنيف، **تستخدم للتنبؤ بقيم متغير وصفي، وخاصة في حالة وجود عدم اليقين (Uncertainty) في الاعتماد على المتغيرات المستقلة في التنبؤ**، ويطلق مصطلح عدم اليقين على الحالات التي لها نفس قيم البيانات (قيم المتغيرات المستقلة) وتعطي معلومات (قيم المتغير الهدف) مختلفة، على سبيل المثال "تشخيص الطبيب لنوع المرض المصاب به المريض من خلال الفحص السريري، علماً بأن الأعراض التي يعاني منها المريض هي أعراض مشتركة بين أكثر من مرض. فيتعرض الطبيب في هذه الحالة لحالة عدم اليقين في تشخيص المرض." ويمكن قياس عدم اليقين من خلال حساب الخطأ المعياري للمتوسط (Standard Error of Mean) للمتغيرات، حيث كلما زادة قيمة الخطأ المعياري للمتوسط يزداد مقدار عدم اليقين. مع العلم زيادة حجم البيانات يعمل على تقليل مقدار عدم اليقين غالباً. **وتستخدم أيضاً في حالة وجود ارتباطات معقدة ما بين المتغيرات المستقلة**. واكتسبت خوارزمية شبكة بيز هذه القدرات المتقدمة لاعتمادها عبى مفاهيم إحصائية عميقة في حساب الاحتمالات الشرطية ونظرية بيز، حيث تعتمد في التنبؤ بقيم المتغير الهدف من خلال قيم الاحتمالات لقيم المتغيرات المستقلة وأثرها على بعضها البعض

الأداة

14. Support Vector Machine SVM

غاية الاستخدام

"نموذج تعلم الآلة"

خوارزمية آلة المتجهات الداعمة SVM هي أحد نماذج تعلم الآلة للانحدار، **تستخدم للتنبؤ بقيم متغير وصفي،** وخاصة **في حالة أن أعداد حقول التنبؤ (المتغيرات المستقلة) كبير قد تصل للآلاف، وحجم بيانات (عدد صفوف قليل)**، وذلك لأنها تعتمد فقط على مجموعة فرعية من نقاط البيانات (تسمى متجهات الدعم) لعمل تنبؤات. **وأيضاً في الحالات التي لا تكون فيها البيانات قابلة للفصل خطياً، حيث يمكنها استخدام التحويلات غير الخطية للعثور على مستوى للفصل بين البيانات.**

ومع ذلك، فإن خوارزمية SVM لها أيضًا بعض المشاكل، حيث يمكن أن تكون خوارزمية SVM مكلفة من الناحية الحسابية عند التعامل مع مجموعات البيانات الكبيرة، لأنها تتطلب حل مشكلة تحسين تربيعية. ولا توفر خوارزمية SVM تقديرات احتمالية مباشرة، لذلك قد يكون من الصعب تقييم عدم اليقين في التنبؤات.

الأداة

15. Neural Networks (NN)

غاية الاستخدام

"نموذج تعلم الآلة"

خوارزمية الشبكات العصبية NN هي أحد نماذج تعلم الآلة للانحدار، **تستخدم للتنبؤ بقيم متغير وصفي،** ولكن يجب أن تكون الخيار الأخير في حالة فشل النماذج الخوارزميات الأخرى، **وعند استخدامها هناك العديد من العوامل التي يجب مراعاتها عند اتخاذ قرار بشأن استخدام الشبكة العصبية لبناء نموذج انحدار على البيانات المهيكلة**:

حجم البيانات وتعقيدها: يمكن للشبكات العصبية التعامل مع مجموعات البيانات الكبيرة والمعقدة، ولكن يمكن أن يكون تدريبها وتشغيلها باهظ التكلفة من الناحية الحسابية. إذا كانت لديك مجموعة بيانات كبيرة بها العديد من الميزات، فقد تكون الشبكة العصبية خيارا جيدا.

جودة البيانات: تتطلب الشبكات العصبية كمية كبيرة من البيانات المصنفة من أجل التعلم بشكل فعال. إذا كانت لديك مجموعة بيانات كبيرة ذات فئات عالية الجودة، فقد تكون الشبكة العصبية خيارا جيدا.

بشكل عام، يمكن أن تكون الشبكات العصبية أداة قوية لبناء نماذج الانحدار على البيانات المهيكلة، لكنها ليست دائما الخيار الأفضل لكل مشكلة.

الإجراء

III. تحليل البقاء "Survival analysis"

الأداة

1. Cox Regression

غاية الاستخدام

"نموذج إحصائي"

انحدار كوكس من أحد نماذج تحليل البقاء، هو أسلوب انحدار إحصائي خاص، ي**ستخدم لتحليل البيانات التي تكون فيها النتيجة ذات الاهتمام هي الوقت الذي يستغرقه حدوث حدث مثير للاهتمام**. يستخدم بشكل شائع في البحث الطبي لدراسة الوقت الذي يستغرقه المريض لتجربة حدث ما، مثل الوفاة أو ظهور المرض أو الشفاء، ويستخدم أيضًا في مجالات أخرى، مثل التمويل والهندسة، لدراسة الوقت يستغرق لحدث ما، مثل فشل الجهاز أو التخلف عن سداد قرض.

يعد انحدار كوكس مفيدا لدراسة العوامل التي تؤثر على الوقت الذي يستغرقه حدوث الحدث، ولإجراء تنبؤات حول احتمالية وقوع الحدث في وقت معين. غالبا ما يستخدم في البحث الطبي لتقييم فعالية العلاجات أو التدخلات على نتائج المرضى، وفي مجالات أخرى لتقييم موثوقية أو متانة المنتجات أو الأنظمة.

الإجراء

IV. التنبؤ المستقبلي "Forecasting "

الأداة

1. Exponential Smoothing

غاية الاستخدام

"نموذج إحصائي"

التنعيم الأسي هو طريقة للتنبؤ ببيانات السلاسل الزمنية (التنبؤ المستقبلي)، والتي تتضمن استخدام البيانات السابقة للتنبؤ بالقيم المستقبلية (البيانات "الصفوف" غير مستقلة عن بعضها البعض). ويستند إلى فكرة أن أحدث نقاط البيانات هي الأكثر ملاءمة للتنبؤ بالقيم المستقبلية، وأن أهمية نقاط البيانات الأقدم تتناقص بشكل كبير بمرور الوقت. ويتم تم تطبيقها غالبة إذا كانت السلسلة لا تحتوي مركبة اتجاه عام (صاعد أو هابط)، وعندما تكون البيانات قليلة التردد (التقلب) وعندما تأخذ السلسلة نمط أسي في الزيادة أو النقصان.

الأداة

2. ARIMA

غاية الاستخدام

"نموذج إحصائي"

ARIMA هي طريقة إحصائية تستخدم للتنبؤ ببيانات السلاسل الزمنية. إنه نوع من النماذج الخطية المستخدمة لالتقاط الارتباط التلقائي في البيانات، بالإضافة إلى أي اتجاه عام وموسمية.

تتكون نماذج ARIMA من ثلاث مكونات: مكون الانحدار الذاتي (AR)، الذي يقوم بنمذجة الارتباط التلقائي في البيانات؛ المكون المتكامل (I)، الذي يصوغ الفرق بين البيانات والتنبؤ السابق؛ ومكوّن المتوسط المتحرك (MA)، الذي يصوغ المخلفات (الأخطاء) للتنبؤ السابق.

بمجرد ملاءمة نموذج ARIMA للبيانات، يمكنك استخدامه لإنشاء تنبؤات للفترات الزمنية المستقبلية. يتم إنشاء التنبؤ باستخدام النموذج لتقدير القيمة المتوقعة للسلسلة الزمنية في الفترة الزمنية المستقبلية، بناءً على البيانات السابقة ومعلمات النموذج.

تعد ARIMA طريقة قوية للتنبؤ ببيانات السلاسل الزمنية، خاصةً عندما تظهر البيانات الارتباط التلقائي والاتجاه العام والموسمية

6. قواعد المشاركة "Association"

هي تقنية للكشف عن كيفية ارتباط العناصر ببعضها البعض.

فمثلاً عندما نذهب للتسوق من السوبرماركت، لدينا في كثير من الأحيان قائمة رئيسية من الأشياء لشرائها. قد تشتري ربة منزل المعكرونة والصلصة، في حين أن شاب جامعي قد يشتري المشروب الغازي والخبز المحسن. قد يساعد فهم أنماط الشراء المختلفة للزبائن على زيادة المبيعات بعدة طرق. فمثلاً إذا كان هناك زوج من العناصر X و Y، يتم شراؤها كثيراً معاً. فيمكن عمل عروض ذكية حيث اذا شريت المنتج Xتحصل خصم على المنتج Z، وبهذه الحالة تحصل زيادة في المبيعات، حيث من الممكن أن يشتري الزبون المنتج X و Y لارتباطهم وان يشري أيضا المنتج Z ليكسب العرض.

إذا قواعد المشاركة Association تسعى لتحديد الارتباطات الخفية بين العناصر المختلفة داخل قواعد البيانات الضخمة، لكي تساعد متخذ القرار بوضع المسار الأمثل في اتخاذ القرارات المختلفة.

وقد يكون لقواعد المشاركة Association الدور في رسم المسار الأمثل من الإجراءات لتحقيق هدف ما. فمثلاً يمكن تحديد مسار السلوك التعليمي الأفضل لكل طالب للوصول لأفضل النتائج.

وفيما يلي عرض لأهم **نماذج قواعد المشاركة**، موضحاً غاية الاستخدام لكل منها؛

الإجراء

I. نماذج المشاركة "Association "

الأداة

1. Sequence

غاية الاستخدام

خوارزمية Sequence **تستخدم للكشف عن الارتباطات بين العناصر** (الارتباط ما بين قيم المتغيرات وليس ارتباط المتغيرات مع بعضها البعض)، **عندما يوجد اهتمام بعملية التوالي في ظهور العناصر (العنصر الأول ومن ثم الثاني ...)**، ويكون الزمن محدد ومتماثل لكل عنصر من عناصر الدراسة. أي تستخدم للتنبؤ بوقوع الحالة التالية لوقوع حالة أو أكثر.

الأداة

2. Carma

غاية الاستخدام

خوارزمية Carma **تستخدم للكشف عن الارتباطات بين العناصر** (الارتباط ما بين قيم المتغيرات وليس ارتباط المتغيرات مع بعضها البعض)، **عندما يوجد اهتمام بظهور العناصر دون الاهتمام بأولوية الوقوع (أي جميعها بنفس المستوى).**

الأداة

3. Apriori

غاية الاستخدام

خوارزمية Apriori **تستخدم للكشف عن الارتباطات بين العناصر** (الارتباط ما بين قيم المتغيرات وليس ارتباط المتغيرات مع بعضها البعض)، **عندما يكون الاهتمام بعملية التوالي دون**

الاهتمام بزمن وقوع الحالات. أي ليس شرط أن يكون الزمن محدد ومتماثل لكل عناصر الدراسة. أي تستخدم للتنبؤ بوقوع الحالة التالية لوقوع حالة أو أكثر دون الاهتمام بزمن الوقوع. وتتميز هذه الطريقة بما يلي: أنها تتعامل مع البيانات الضخمة بكفاءة وسرعة عالية، ليس له عدد محدد على القواعد التي يمكن تتبعها، لا يتعامل إلا مع المتغيرات الفئوية. يحدد عدد التنبؤات آلياً. ويتم تحديد الحالات القبلية (Antecedents) وحالات النتيجة (Consequents).

الأداة

4. Association rules

غاية الاستخدام

خوارزمية قواعد المشاركة Association rules **تستخدم للكشف عن الارتباطات بين العناصر** (الارتباط ما بين قيم المتغيرات وليس ارتباط المتغيرات مع بعضها البعض)، **عندما يكون الاهتمام بعملية التوالي دون الاهتمام بزمن وقوع الحالات.** أي ليس شرط أن يكون الزمن محدد ومتماثل لكل عناصر الدراسة أو لا يوجد. أي تستخدم للتنبؤ بوقوع الحالة التالية لوقوع حالة أو أكثر دون الاهتمام بزمن الوقوع. وتتميز هذه الطريقة بما يلي: نلاحظ التشابه الواضح بين هذه النمذجة ونمذجة Apriori، مع وجود بعض الاختلافات. حيث تعتمد نمذجة قواعد المشاركة على بناء الجمل الشرطية ذات قيمة (Entropy) عالية، حيث يدل مقدار Entropy على مقدار المعلومات (Information) التي تقدمها الجملة الشرطية؛

if condition(s) **then** prediction(s)

على سبيل المثال، "إذا اشترى عميل معكرونة وبعد ذلك لحمة، فسيشتري هذا العميل معجون الطماطم بثقة 95٪" (مثال بسيط للتوضيح) لأن هدفا أعمق من ذلك وهو الوصول للرؤى الخفية (Hidden Insights). حيث تستخرج عقدة Association Rules مجموعة من القواعد (جمل شرطية) من مجموعة البيانات.

7. التقييم " Evaluation"

يعد تقييم النموذج خطوة مهمة في عملية تطوير نماذج علم البيانات والتعلم الآلي. حيث يساعد على تحديد مدى قدرة النموذج على التنبؤ بالبيانات غير المرئية.

يجب أن تكون عملية التقييم للنموذج عملية مستمرة، وليس مجرد حدث لمرة واحدة. من المهم إجراء تقييم مستمر لأداء النموذج عند توفر بيانات جديدة، حيث يمكن أن يساعد ذلك في تحديد الوقت الذي يتوقف فيه النموذج عن الأداء الجيد وقد يحتاج إلى التحديث أو الاستبدال.

ومن المهم تقييم النماذج باستخدام مجموعة متنوعة من المقاييس المختلفة، حيث يمكن للمقاييس المختلفة إبراز جوانب مختلفة من أداء النموذج. على سبيل المثال، تعد الدقة مقياسًا شائعًا لمهام التصنيف، ولكنها قد لا تكون أفضل مقياس لتقييم النموذج المستخدم لتقديم التوصيات، حيث قد تكون الحساسية (Sensitivity) والنوعية (Specificity) أكثر صلة.

ومن المهم أيضاً مراعاة السياق الذي سيُستخدم فيه النموذج عند تقييم أدائه. على سبيل المثال، قد يحتاج النموذج المستخدم لتشخيص الحالات الطبية إلى معدل دقة أعلى من النموذج المستخدم في التوصية بالمنتجات للعملاء.

وأيضاً من المهم مقارنة أداء النماذج المختلفة عند اختيار أفضل نموذج لمشكلة معينة. يمكن القيام بذلك عن طريق تدريب وتقييم نماذج متعددة باستخدام نفس البيانات والمقاييس، ثم اختيار النموذج الأفضل أداءً.

من المهم أن تكون على دراية بالقيود المفروضة على أي نموذج، وأن تبلغ هذه القيود لأصحاب المصلحة الذين سيستخدمون النموذج. يمكن أن يساعد ذلك في ضمان استخدام النموذج بشكل مناسب وعدم المبالغة في تفسير النتائج التي ينتجها.

وفيما يلي عرض لأهم **أدوات التقييم للنماذج،** موضحاً غاية الاستخدام لكل منها؛

الإجراء
I. **تقييم النموذج "Models Evaluation "**
الأداة
1. Classification Models Evaluation
غاية الاستخدام

لتقييم نماذج التصنيف يجب قياس مجموعة من المؤشرات، وهي؛

1. بناء مصفوفة الارتباك (Confusion Matrix)
2. الدقة العامة (Accuracy)
3. الخطأ (Error)
4. الدقة الحقيقية الإيجابية (Precision true positive)
5. الدقة الحقيقية السلبية (Precision true negative)
6. الحساسية (معدل الإيجابية الحقيقي) (Sensitivity (true positive rate))
7. النوعية (معدل السلبية الحقيقي) (Specificity (true negative rate))
8. معدل الخطأ الإيجابي (False Positive Rate)
9. معدل الخطأ السلبي (False Negative Rate)

الأداة
2. Regression Models Evaluation
غاية الاستخدام

لتقييم نماذج التصنيف يجب قياس مجموعة من المؤشرات، وهي؛

1. الحد الأدنى للخطأ (Minimum Error)
2. الحد الأعلى للخطأ (Maximum Error)
3. معدل الخطأ (Mean Error)
4. معدل الخطأ المطلق (Mean Absolute Error)
5. الانحراف المعيار للخطأ (Standard Deviation Error)
6. الارتباط الخطي (Linear Correlation)

حيث يعتبر تطابق البيانات الحقيقية (Actual) للمتغير الهدف مع البيانات المتنبئ بها (Predicted) للمتغير الهدف من أهم المؤشرات الدالة على جودة النموذج. حيث يتم قياس هذه المؤشر من خلال دراسة الارتباط الخطي (r) بين البيانات الحقيقية (Actual) للمتغير الهدف مع البيانات المتنبئ بها (Predicted) للمتغير الهدف. حيث كلما اقتراب قيمة الارتباط الخطي الى 1+ دل على زيادة جودة النموذج.

دليل خوارزميات علم البيانات وهندسة تعلم الآلة

فيما يلي عرض ملخص على شكل جدول لخوارزميات علم البيانات وهندسة تعلم الآلة، حيث يتسلسل الجدول بعرض الخوارزميات حسب أولوية الإجراءات، علماً كثير من الإجراءات يمكن أن تتكرر في أكثر من موقع في المشروع. مع العلم تم اعتماد المصطلحات باللغة الإنجليزية خوفا من الاختلاط في المصطلحات وخاصة مع ضعف تعريبها. وفيما يلي عرض للملخص؛

Data Understanding "فهم البيانات"		
Action	**Tool or Algorithm**	**The Aim**
Display data "عرض البيانات"	**Table**	أداة **لعرض البيانات على طبيعتها أو عرض جزء منها اعتماداً على شرط ما**. وتعتبر أداة عرض للمخرجات الناتجة من أي عملية معالجة للبيانات. حيث تستخدم في أي موقع في دفق البيانات (Data Stream) لمشاريع علم البيانات.
Data Audit "تدقيق البيانات"	**Data Audit**	أداة تستخدم **لأخذ نظرة أولية شاملة على البيانات**، من خلال عرض مصفوفة تربط كل متغير بمجموعة من الخصائص والمقاييس الإحصائية الوصفية لقيمه مع عرض قيمه برسم بياني
	Descriptive statistics	أداة تستخدم لتلخيص **البيانات الكمية** من خلال مقايس الإحصاء الوصفي لها، ويمكن أن تستخدم لقياس قوة واتجاه الارتباط الخطي بين المتغيرات.
	Hypothesis tests	أداة تستخدم **لدراسة وجود أو عدم وجود أثر معنوي (ذو دلالة إحصائية) بين المتغيرات** المستقلة والمتغير الهدف، أي على سبيل المثال هل النوع (ذكر/أنثى) كمتغير مستقل له أثر معنوي (أي ذو معنى) على مقدار الدخل المتنبئ به (المتغير الهدف) ويطلق على هذا الاختبار اسم (اختبارات فروض الفرق بين متوسطي مجتمعين مستقلين أو أكثر)، أو هل النوع (ذكر/أنثى) كمتغير مستقل له أثر معنوي (أي ذو معنى) على تعثر أو عدم تعثر العميل لسداد القرض (المتغير الهدف) ويطلق على هذا الاختبار اسم (اختبار الاستقلالية بين متغيرين وصفيين). وتطبق أيضا لدراسة وجود أو عدم وجود أثر لإجراء ما على أحد الخصائص (متغير مستقل)، على سبيل المثال دراسة أثر اتباع حمية ما على مستوى السكر في الدم، في هذه الحالة ندرس أثر الحمية من خلال متغيرين مستقلين الأول مستوى السكر في الدم لمجموعة من المرضى قبل اتباع الحمية والمتغير الثاني مستوى السكر في الدم لنفس المرضى بعد اتباع الحمية ويطلق على هذا الاختبار اسم (اختبارات فروض الفرق بين متوسطي مجتمعين غير مستقلين (مرتبطين).

	Correlation	أداة تستخدم **لقياس قوة واتجاه الارتباط الخطي بين المتغيرات** المستقلة مع الهدف والمستقلة مع بعضها البعض.
Visualization "تمثيل البيانات"	**Histogram**	رسم مشابه للمدرج التكراري ما بين **متغير كمي وتكراره**، وتبرز أهميته من خلال القدرة على إدخال متغير وصفي أو أكثر على الرسمة ومشاهدة التغيرات التي تحدث عل القيم الكمية. حيث المتغيرات الوصفية في هذا النوع يعبر عنها animation أو الألوان أو تقسيم الرسمة لعدة أجزاء.
	Collection	رسم مشابه للمدرج التكراري ما بين **متغيرين كميين**، وتبرز أهميته من خلال القدرة على إدخال متغير وصفي أو أكثر على الرسمة ومشاهدة التغيرات التي تحدث عل القيم الكمية. حيث المتغيرات الوصفية في هذا النوع يعبر عنها animation أو الألوان أو تقسيم الرسمة لعدة أجزاء.
	Distribution	رسم مشابه للتمثيل بالأعمدة ما بين **متغير وصفي وتكراره**، وتبرز أهميته من خلال القدرة على إدخال متغير وصفي أخر على الرسمة ويعبر عنه بالألوان. ويستفاد من هذا الرسم في مقارنة النسب والتكرار.
	Plot	رسم مشابه للمنحنى التكراري، حيث يمثل علاقة بين **متغيرين من أي نوع**، مع إمكانية إدخال متغير وصفي أو أكثر أو كمي على الرسمة ومشاهدة التغيرات التي تحدث عل الرسمة. حيث المتغيرات الوصفية والكمية في هذا النوع يعبر عنها animation أو تقسيم الرسمة لعدة أجزاء، أو الألوان، أو الأشكال، أو الأحجام، أو درجة اللون.
	Multi Plot	رسم مشابه للمنحنى التكراري، حيث يمثل علاقة **متغير من أي نوع على محور x** مع عدد من **المتغيرات الكمية على محور y**، مع إمكانية إدخال متغير وصفي أو اثنين على الرسمة ومشاهدة التغيرات التي تحدث عل الرسم. حيث المتغيرات الوصفية في هذا النوع يعبر عنها animation أو تقسيم الرسمة لعدة أجزاء.
	Time Plot	تستخدم هذه العقدة لرسم السلاسل الزمنية (**متغيرات كمية** تغيرها مرتبط **بالزمن**).
	Web	تستخدم لتمثيل **قوة الارتباط** بين **متغيرين وصفيين**، معتمدة في عملية الربط على عدة عوامل مثل النسبة أو تكرار الارتباط، وتعتبر الرسومات الناتجة من هذه العقدة من أهم المؤشرات الابتدائية لدراسة الارتباط والعلاقات بين المتغيرات الوصفية. وهذه العقدة تتيح خيارين للتمثيل، الأول (Web) ويستخدم في حالة دراسة **الارتباط** بين **جميع المتغيرات الوصفية** المحددة (حيث يحدد ارتباط كل متغير وصفي مع جميع المتغيرات الوصفية الأخرى). أما الخيار الثاني (Directed web) ويستخدم في حالة دراسة **الارتباط** بين **متغير وصفي واحد مع بقية المتغيرات الوصفي.**

	Graph board	أداة تمكننا من الاختيار من بين العديد من نواتج الرسوم البيانية المختلفة (scatterplots, … , etc ,bar charts, pie charts, histograms) في عقدة واحدة. حيث يقوم محلل البيانات باختيار الحقول التي يريد أن يرسم العلاقة بينها وطبقا لعدد الحقول ونوعها ترشح البرمجية الرسومات المتاحة لهذه البيانات ومن ثم يختار المحلل الرسمة الأكثر تناغماً مع البيانات، وتتيح العقدة اختيار تفاصيل أكثر للرسمة.

Data Preparation "إعداد البيانات"		
Action	**Tool or Algorithm**	**The Aim**
Selecting Data "اختيار بيانات"	**Sample**	أداة تستخدم هذه العقدة لاختيار أو تجاهل عينة من مصدر البيانات، **وغالباً لغايات استخدام هذه العينة في فحص النموذج الذي يتم بناءه.**
	Select	أداة تستخدم **لتضمين أو تجاهل مجموعة فرعية من السجلات في مصدر البيانات**، بناءً على حالة معينة أو شرط معين، مثلاً تضمين البيانات التي تكون قيمة المتغير Age > 30، ويستفاد منها لبناء سناريوهات مختلفة للنموذج.
	Filter	أداة تستخدم **لإجراء فلترة (إبطال أثر بدون حذف) على متغير** من حيث إهماله في عملية التحليل (أي عدم تأثيره على أي عمل مستقبلي على البيانات.
	Anonymize	أداة **تستخدم لإخفاء قيم متغير ما أو أكثر من خلال استبدال القيم الحقيقية بقيم بديلة، والهدف من ذلك المحافظة على سرية بعض البيانات** وخاصة في حالة استخدام النموذج من عدة أشخاص مثل إخفاء أسماء العملاء أرقام هواتفهم أو العناوين. مع الملاحظ أن المتغير الذي يتم إخفاء قيمة يصبح تأثيره في أي عملية مستقبلية للبيانات بقيمة الجديدة.
	Partition	أداة تستخدم هذه العقدة لتقسيم البيانات إلى عينات منفصلة لأغراض التدريب والاختبار وأغراض التحقق (اختيارية). **ويستفاد منها لفحص دقة نموذج التنبؤ بعد بناءه.**
Cleaning Data "تنظيف البيانات"	**Data Quality Report**	أداة تستخدم **لتحليل دقيق لجودة للبيانات المتاحة قبل النمذجة**. والذي يساعد في كشف مشاكل البيانات ومن أشهرها • تتضمن البيانات المفقودة قيمًا فارغة أو مشفرة على أنها عدم استجابة.

		• أخطاء البيانات عادة ما تكون أخطاء مطبعية يتم إجراؤها عند إدخال البيانات. • تتضمن أخطاء القياس البيانات التي تم إدخالها بشكل صحيح ولكنها تستند إلى نظام قياس غير صحيح. • التناقضات في الترميز تشتمل عادة على وحدات القياس غير القياسية أو عدم تناسق القيمة، مثل استخدام كل من M و Male للدلالة على النوع. • بيانات التعريف السيئة تتضمن عدم التطابق بين المعنى الظاهر للحقل والمعنى الموضح في اسم الحقل أو التعريف.
	Balance	تستخدم لتصحيح الاختلالات في نسب مجموعات البيانات. على سبيل المثال، لنفرض أن مجموعة البيانات تحتوي على ثلاث مناطق لتقديم الخدمة (شمال، وسط، جنوب)، وكانت منطقة الشمال تمثل 75٪ تقريباً من قيم متغير المنطقة، والوسط تقريباً 13%، والجنوب تشكل حوالي 12%. لذلك تعمل هذه الأداة على معالجة هذا الخلل، من خلال تصحيح نسب المجموعات بشكل متساوي تقريباً من خلال تكرار قيم مجموعات معينة أو حذف مشاهدات مجموعة ما، لبناء نماذج غير متحيزة.
	Filler	أداة تستخدم لاستبدال **مجموعة من قيم المتغير أو القيم المفقودة في المتغير بقيم أخرى وفق شرط معين أو استبدال القيم المفقودة بقيم معينة.**
Constructing New Data "بناء بيانات جديدة"	**Derive**	أداة تستخدم **لإضافة متغير (حقل) على سجل البيانات الأصلي تحت شروط معينة.**
	Binning	أداة تستخدم **لتحويل قيم متغير كمي (إنشاء متغير جديد) الى قيم اسمية (Nominal) بناء على مجموعة من العمليات الإحصائية**. أي تستخدم لتوليد متغير (حقل) جديد من النوع Nominal من خلال إعادة تصنيف قيم متغير متصل. ويستفاد منها في تحويل المتغير الكمي المتصل الى متغير اسمي (فئوي)، يستفاد منها في معالجة البيانات لتصبح قابلة لاستخدامها بخوارزميات أخرى لا تقبل إلا متغيرات اسمية مثل عقد الانحدار اللوجستي وبيز وغيرها.
Formatting Data "تنسيق البيانات"	**Type**	أداة تستخدم **لإجراء بعض التعديلات على خصائص المتغيرات أثناء بناء النموذج**، مثل مستوى القياس للمتغير أو هل هو متغير مستقل أو متغير هدف.
	Reorder Field	أداة تستخدم إعادة **ترتيب المتغيرات في ملف البيانات**، معتمدين في الترتيب على نوع المتغيرات أو أسمائها.

	Aggregate	أداة تستخدم **لتجميع البيانات الواقعة في ملف واحد وتمثل نفس الحالة (أي لها نفس قيمة متغير التعريف) وإظهار ملخص لقيم المتغيرات المقابلة لها**
	RFM, (Recency, Frequency, Monetary) aggregate	أداة تستخدم في **التعاملات المالية والتسويقية، حيث تعمل على إظهار مجموعة من المعلومات المالية الخاصة بالعملاء. والمعلومات التي يتم تجميعها لكل عميل أو موظف أو وكيل هي؛** **Recency** (الحداثة): وتدل على عدد الأيام التي مرت على أخر تعامل. **Frequency** (التكرار): ويدل على عدد مرات التعامل. **Monetary** (المالية): وتدل على مجموع المعاملات المالية التي أجريت.
	RFM Analysis	أداة تستخدم **لتقييم لنتائج عقدة RFM aggregate**، يستفاد منها لتصنيف العملاء من حيث الأفضلية اعتماداً على قيم (Recency, Frequency, Monetary)؛ حيث تعمل على إعطاء درجة لكل خاصية (Recency, Frequency, Monetary) لكل عميل من حيث الأفضلية، وإعطاء أيضا درجة عامة لكل عميل تسمى RFM Score.
	Transform	أداة تستخدم **لإجراء مجموعة من التحويلات الرياضية على البيانات**، لوصولها لشكل التوزيع الطبيعي، لتصبح متوافقة مع مجموعة من الخوارزميات
Integrating Data "دمج البيانات"	**Merge**	أداة **تستخدم لدمج مجموعتين أو أكثر من البيانات مكونه من سجلات متماثلة ومتغيرات (حقول) مختلفة**، أو مكونه من سجلات مختلفة ومتغيرات (حقول) متماثلة، أو العمليتين معاً. ويمكن لهذه العقدة دمج ملفين أو أكثر مكونه من سجلات مختلفة ومتغيرات (حقول) مختلفة
	Append	أداة تستخدم **لتجميع قواعد البيانات ذات الهياكل المماثلة، أي تجميع ملفين أو أكثر من الملفات ذات متغيرات (حقول) متماثلة وسجلات مختلفة**. وفي حالة احتواء أحد الملفات على متغير أو أكثر غير موجودة في الملفات الأخرى، يمكن إظهاره والقيم المفقودة يتم تعريفها $null$.

Dimensional Reduction Methods & Models "طرق ونماذج تقليل الأبعاد"		
Action	**Tool or Algorithm**	**The Aim**
Dimensions reduction methods	**Missing Value Ratio**	طريقة لنمذجة تقليل الأبعاد **في حالة لدينا نسبة كبيرة من القيم المفقودة لأحد أو مجموعة من المتغيرات المستقلة.**
	Low Variance Filter	طريقة لنمذجة تقليل الأبعاد **في حالة إذا كان لدينا متغير مستقل أو أكثر ذو تباين منخفض جداً أو صفري.**

"طرق تقليل الأبعاد"	**High Correlation Filter**	طريقة لنمذجة تقليل الأبعاد **في حالة إذا كان لدينا متغيرات مستقلة ذات ارتباط عالي.**
	Mean differences were not significant Filter	طريقة لنمذجة تقليل الأبعاد **في حالة إذا كان لدينا متغير مستقل أو متغيرات لا يوجد لها اثر معنوي على المتغير الهدف.**
	Independence Filter	طريقة لنمذجة تقليل الأبعاد **في حالة لدينا إذا كان لدينا متغيرات وصفية مستقلة لا يوجد ارتباط بينها وبين المتغير الهدف من النوع الوصفي.**
	Backward Feature Elimination	طريقة لنمذجة تقليل الأبعاد **على المتغيرات المستقلة المتعددة كماً ونوعاً من خلال إجراء عمليات تقييم متسلسلة لأداء النموذج بعد بناءه.**
	Forward Feature Selection	طريقة لنمذجة تقليل الأبعاد **معاكسة للنمذجة السابقة الترشيح من خلال التغذية الراجعة Backward Feature Elimination. بدلاً من إزالة المتغيرات المستقلة، نحاول العثور على أفضل الميزات التي تعمل على تحسين أداء النموذج**
Dimensional Reduction Models "نماذج تقليل الأبعاد"	**Factor Analysis**	**خوارزمية لنمذجة تقليل الأبعاد على المتغيرات الكمية المستقلة المتعددة كماً ونوعاً**، من خلال تقسيم المتغيرات الكمية المرتبطة الى مجموعات، حيث المتغيرات في كل مجموعة يكون لها ارتباط قوي فيما بينها، وارتباط منخفض مع متغيرات المجموعة (المجموعات) الأخرى. وبعد ذلك يتم تعريف كل مجموعة كعامل (متغير). تتميز هذه العوامل بقلة عددها مقارنة بالمتغيرات الأصلية للبيانات
	Principal Component Analysis	**خوارزمية لنمذجة تقليل الأبعاد على المتغيرات المستقلة المتعددة كماً ونوعاً، من خلال التغلب على تكرار المتغيرات (الميزات) في مجموعة البيانات**. ويتم ذلك من خلال استخراج مجموعة جديدة من المتغيرات غير المرتبطة مع بعضها البعض تسمى (بالمكونات الأساسية) من مجموعة كبيرة من المتغيرات الكمية المدخلة الحالية. حيث تسعى هذه النمذجة لاستيعاب أكبر قدر ممكن من المعلومات الموجودة في البيانات وتضمينها في المكونات الأساسية
	Feature Selection	**خوارزمية لنمذجة تقليل الأبعاد على المتغيرات المستقلة المتعددة كماً ونوعاً، من خلال الاعتماد على مجموعة من الخصائص**؛ إزالة المدخلات والسجلات الغير هامة والمثيرة للمشاكل، أو حقول الإدخال التي تحتوي على عدد كبير جدًا من القيم المفقودة أو وجود اختلاف (Variation) كبير جدًا أو صغير جدًا.

Segmentation Models "نماذج التقسيم"		
Action	**Tool or Algorithm**	**The Aim**
Clustering "التقسيم"	**K-Mean**	خوارزمية من خوارزميات التعلم الآلي الغير خاضع للإشراف، **تُستخدم لتقسيم مجموعة البيانات إلى عدد محدد مسبقًا من المجموعات**. تعمل الخوارزمية من خلال تسكين كل مدخل (صف) في مجموعة البيانات إلى الكتلة التي يكون مركزها (الوسط) الأقرب إليها، بناءً على بعض مقايس المسافة. وتتصف العملية بالتكرارية، حيث تقوم الخوارزمية بإعادة حساب متوسط كل مجموعة وإعادة تعيين نقاط البيانات إلى المجموعات حتى يستقر تعيين نقاط البيانات إلى المجموعات. تتمثل إحدى مزايا مجموعة الوسائل K في أنها سريعة وسهلة التنفيذ، مما يجعلها خيارًا شائعا لتجميع مجموعات البيانات الكبيرة. ومع ذلك، يمكن أن تكون الخوارزمية حساسة للاختيار الأولي لمراكز الكتلة، وقد تختلف النتائج اعتمادا على التكوين الأولي. لذلك، من الشائع تشغيل الخوارزمية عدة مرات بتهيئة مختلفة واختيار التكوين الذي يعطي أفضل النتائج.
	Kohonen Networks	خوارزميات شبكات Kohonen، والمعروفة أيضًا باسم الخرائط ذاتية التنظيم (SOMs)، هي نوع من الشبكات العصبية الاصطناعية التي تُستخدم للتعلم غير الخاضع للإشراف. الهدف من شبكة Kohonen هو تعلم تمثيل منخفض الأبعاد (خريطة) لمجموعة بيانات عالية الأبعاد، بحيث يتم تعيين نقاط البيانات المتشابهة إلى المواقع القريبة في الخريطة. تحتوي شبكات Kohonen على عدد من التطبيقات، بما في ذلك تصور البيانات والتعرف على الأنماط واستخراج الميزات. إنها مفيدة بشكل خاص لاستكشاف وفهم بنية مجموعات البيانات الكبيرة والمعقدة، حيث يمكنها الكشف عن أنماط وعلاقات قد لا تكون واضحة باستخدام تقنيات أخرى.
	TwoStep	خوارزمية من خوارزميات التعلم الآلي الغير خاضع للإشراف، تستخدم نمذجة TwoStep لتقسيم البيانات من أي نوع إلى مجموعات مميزة وخاصةً عندما لا يكون للمحلل علم أو قدره لتحديد ماهية هذه المجموعات ولا عددها عند البداية. حيث يحاول TwoStep اكتشاف أنماط في متغيرات الإدخال المحددة للتقسيم بشكل آلي. يتم تقسيم السجلات بحيث تكون السجلات داخل المجموعة متشابهة مع بعضها البعض، ولكن السجلات في مجموعات مختلفة تختلف عن بعضها.

		تطبق نمذجة TwoStep من خلال منهجية تعمل على تقسيم البيانات الى مجموعات متجانسة في عناصرها من خلال خطوتين أو مرحلتين؛ الخطوة الأولى في هذه الخطوة يتم تمرير البيانات الخام بشكل منفرد ويتم توزيعها الى مجموعات جزئية ذات معنى (أي أن عناصر كل مجموعة تتسم بصفات خاصة مشتركة). الخطوة الثانية في هذه الخطوة يتم تجميع ودمج المجموعات الجزئية الى مجموعات أكبر ذات معنى (أي أن عناصر كل مجموعة تتسم بصفات عامة مشتركة).
Anomaly Detection "تحديد البيانات الشاذة"	**Anomaly**	خوارزمية من خوارزميات التعلم الآلي الغير خاضع للإشراف، تستخدم نمذجة Anomaly لتحديد القيم المتطرفة (Outliers)، أو حالات غير عادية في البيانات (من جميع الأنواع). من خلال تقسيم البيانات الى مجموعات عناصر كل منها متجانسة ومن ثم تحديد القيم الشاذة لكل مجموعة. ويتميز نمذجة Anomaly أن لديها القدرة على الكشف عن القيم الشاذة المتطرفة التي لا تتبع أي نموذج من نماذج فحص القيم المتطرفة، خلافا لأساليب النمذجة الأخرى التي تخزن نماذج محددة لفحص القيم المتطرفة. إذاً باستخدام هذا المؤشر من الممكن تحديد القيم المتطرفة حتى لو كانت لا تتفق مع أي نمط معروف سابقاً للقيم الشاذة، وقد يكون مفيداً بشكل خاص في التطبيقات، مثل الكشف عن الغش، حيث يمكن دائما أن تظهر أنماط جديدة من الغش غير المعروف لدينا.

Prediction Models "نماذج التنبؤ"			
Action	**Tool or Algorithm**		**The Aim**
Regression "الانحدار"	**Statistical Models**	**Linear Regression**	الانحدار الخطي هو نموذج انحدار إحصائي، يعتبر من أفضل النماذج لتمثيل العلاقة ما بين **مجموعة من المتغيرات المستقلة** ومتغير هدف من **النوع الكمي في حالة أن العلاقة خطية**، لقدرته على وصف وتمثيل العلاقات الخطية بينهما بشكل دقيق. ولكن يستلزم مجموعة من الشروط على البيانات لتطبيق هذا النموذج عليها **من** أهمها أن المتغير التابع يتبع التوزيع الطبيعي.
		Generalized linear Regression	الانحدار الخطي المعمم هو نموذج انحدار إحصائي، يستخدم لنمذجة العلاقة الخطية بين متغير تابع من **النوع الكمي** ومتغير واحد أو أكثر من **المتغيرات المستقلة. وهو امتداد لنموذج الانحدار الخطي، ويتسم بانة يفترض أن المتغير التابع يتبع توزيعا من عائلة التوزيعات الأسية، والتي تشمل التوزيعات الشائعة مثل التوزيع الطبيعي والتوزيع ذي**

		الحدين وتوزيع بواسون. أي يتيح النموذج المعمم التعامل مع نطاق أوسع من المتغيرات التابعة، بما في ذلك المتغيرات التي لا يتم توزيعها بشكل طبيعي، فمن هنا جاء أنه أعم من الانحدار الخطي.
	Generalized linear Mixed Models	النماذج المختلطة الخطية المعممة (GLMMs) هو نموذج انحدار إحصائي، يستخدم لنمذجة العلاقة الخطية بين متغير تابع من **أي نوع** ومتغير واحد أو أكثر من **المتغيرات المستقلة**. **وهو امتداد للنماذج الخطية المعممة (GLMs) ، والتي تُستخدم لنمذجة العلاقة بين متغير تابع ومتغير واحد أو أكثر من المتغيرات المستقلة. ويتسم بأنه لديه القدرة نمذجة العلاقة بين متغير تابع ومتغير واحد أو أكثر من المتغيرات المستقلة التي تحتوي على مزيج من التأثيرات الثابتة والعشوائية.** في GLMM، يُفترض أن المتغير التابع يتبع توزيعا من عائلة التوزيعات الأسية، كما هو الحال في GLM. ومع ذلك، في GLMM، يتم التعامل مع بعض المتغيرات المستقلة (التأثيرات الثابتة) على أنها ثابتة، بينما يتم التعامل مع المتغيرات الأخرى (التأثيرات العشوائية) على أنها متغيرات عشوائية تختلف عبر الملاحظات. **يسمح هذا للنموذج بتفسير حقيقة أن تأثير بعض المتغيرات قد يختلف من ملاحظة إلى أخرى.**
	Logistic Regression	الانحدار اللوجستي هو نموذج انحدار إحصائي، **يستخدم للتنبؤ باحتمالية ظهور حدث ما**، فيجب النظر هنا أن المتغيرات المستقلة هي متغيرات **من أي نوع** والمتغير التابع **الحقيقي احتمالية ظهور حدث ما**، والمتغير التابع الوهمي هو **المتغير الذي أحد قيمة القيمة المراد التنبؤ باحتماليتها**. على سبيل المثال يرغب بنك في تحديد قيمة التأمين على القرض المقدم للعميل اعتماداً على احتمالية تعثره، ففي هذه الحالة الهدف هو تحديد احتمالية تعثر أي عميل يرغب بأخذ قرض.
Decision Trees	**Regression & Classification Tree**	شجرة التنبؤ والانحدار هي أحد نماذج تعلم الآلة للانحدار، تعتمد على عملية التقسيم الغصني الثنائي. **تستخدم للتنبؤ بقيم متغير كمي، عندما تكون المتغيرات المستقلة ذات تسلسل هرمي واضح للميزات أو الفئات. واذا كان حجم البيانات طبيعي (ليس ضخم)، ودقة الفروقات بين قيم المتغيرات المستقلة ليست ذو أهمية.**
	Random Trees	الأشجار العشوائية هي أحد نماذج تعلم الآلة للانحدار، تعتمد على عملية التقسيم الساقي (عدد من الأشجار). **تستخدم للتنبؤ بقيم متغير كمي، عندما يكون لديك عدد كبير من المتغيرات المستقلة (المميزات) وعدد صغير نسبيًا من البيانات (الصفوف) ، وعندما تريد إنشاء نموذج قوي في حالة وجود بيانات صاخبة "Noisy Data " (البيانات الصاخبة هي بيانات تحتوي على كمية كبيرة من المعلومات الإضافية التي لا معنى**

			لها والتي تسمى الضوضاء) والقيم المتطرفة في البيانات. كما أنها سهلة التنفيذ نسبيًا ويمكنها التعامل مع البيانات عالية الأبعاد والقيم المفقودة جيدًا.
		Chi-squared Automatic Interaction Detection (CHAID)	شجرة مربع كاي للكشف عن التفاعل بشكل تلقائي هي أحد نماذج تعلم الآلة للانحدار، تعتمد على عملية التقسيم الغصني المتعدد. **تستخدم للتنبؤ بقيم متغير كمي، عندما تكون المتغيرات المستقلة خليط ما بين المتغيرات الكمية والوصفية، فضلاً عن البيانات المفقودة. يمكن للخوارزمية التعامل مع أي عدد من المتغيرات المستقلة، ولكنها تكون أكثر فاعلية عندما يكون هناك عدد صغير نسبياً من المتغيرات المستقلة ذات عدد كبير من القيم الكمية والوصفية.**
		XGBoost (eXtreme Gradient Boosting)	XGBoost هو تطبيق شائع وفعال لخوارزمية تعزيز التدرج، وهي تقنية تعلم آلي لبناء نماذج تنبؤية. تم تصميم XGBoost خصيصًا ليكون سريعًا وقابلًا للتطوير. تعمل الخوارزمية من خلال بناء سلسلة من أشجار القرار، حيث تتعلم كل شجرة من أخطاء الشجرة السابقة. تسمى هذه العملية التعزيز، وهي تساعد على تحسين الأداء العام للنموذج. **تستخدم للتنبؤ بقيم متغير كمي، وخاصة في حالة وجود قيم مفقودة وكانت البيانات ضخمة واسعة النطاق منظمة أو غير منظمة. وهي مناسبة بشكل خاص للمهام التي تكون فيها الدقة مهمة ومجموعات البيانات كبيرة.**
	Machine Learning	**K Nearest Neighbors**	خوارزمية أقرب K جوار KNN هي أحد نماذج تعلم الآلة للانحدار، **تستخدم للتنبؤ بقيم متغير كمي، حيث** تم تطوير نمذجة تحليل الجوار الأقرب KNN كأحد خوارزميات تعلم الآلة للتعرف على أنماط البيانات دون الحاجة إلى تطابق تام مع أي أنماط أو حالات مخزنة، أنما اعتماداً على حالات قريبة منها في الخصائص، أي نحكم على تماثل الحالات بقربها من بعضها البعض ونحكم على اختلاف الحالات ببعدها عن بعضها البعض. وبالتالي، فإن المسافة بين حالتين هي مقياس التماثل والاختلاف بينهما. وهذا يدل على أن نمذجة تحليل الجوار الأقرب هي أقل دقة في عمليات التصنيف والتنبؤ من النماذج الأخرى، **ولكن هي أكثر تفاعلية مع أي تراكيب معقدة من البيانات أو الحالات، والأكثر استجابة مع أي حالات جديدة أو طفرات في الحالات، ويفضل استخدام خوارزمية KNN إذا كان يوجد اختلاف ما بين بيانات التدريب والبيانات الجديدة المراد تطبيق النموذج عليها.**

		Support Vector Machine SVM	خوارزمية آلة المتجهات الداعمة SVM هي أحد نماذج تعلم الآلة للانحدار، **تستخدم للتنبؤ بقيم متغير كمي**، وخاصة **في حالة أن أعداد حقول التنبؤ (المتغيرات المستقلة) كبير قد تصل للآلاف، وحجم بيانات (عدد صفوف قليل)**، وذلك لأنها تعتمد فقط على مجموعة فرعية من نقاط البيانات (تسمى متجهات الدعم) لعمل تنبؤات. **وأيضاً في الحالات التي لا تكون فيها البيانات قابلة للفصل خطياً، حيث يمكنها استخدام التحويلات غير الخطية للعثور على مستوى للفصل بين البيانات.** ومع ذلك ، فإن خوارزمية SVM لها أيضًا بعض المشاكل، حيث يمكن أن تكون خوارزمية SVM مكلفة من الناحية الحسابية عند التعامل مع مجموعات البيانات الكبيرة ، لأنها تتطلب حل مشكلة تحسين تربيعية. ولا توفر خوارزمية SVM تقديرات احتمالية مباشرة، لذلك قد يكون من الصعب تقييم عدم اليقين في التنبؤات.
		Neural Networks (NN)	خوارزمية الشبكات العصبية NN هي أحد نماذج تعلم الآلة للانحدار، **تستخدم للتنبؤ بقيم متغير كمي**، ولكن يجب أن تكون الخيار الأخير في حالة فشل النماذج الخوارزميات الأخرى، **وعند استخدامها هناك العديد من العوامل التي يجب مراعاتها عند اتخاذ قرار بشأن استخدام الشبكة العصبية لبناء نموذج انحدار على البيانات المهيكلة**: **حجم البيانات وتعقيدها**: يمكن للشبكات العصبية التعامل مع مجموعات البيانات الكبيرة والمعقدة، ولكن يمكن أن يكون تدريبها وتشغيلها باهظ التكلفة من الناحية الحسابية. إذا كانت لديك مجموعة بيانات كبيرة بها العديد من الميزات، فقد تكون الشبكة العصبية خيارا جيدا. **جودة البيانات**: تتطلب الشبكات العصبية كمية كبيرة من البيانات المصنفة من أجل التعلم بشكل فعال. إذا كانت لديك مجموعة بيانات كبيرة ذات فئات عالية الجودة، فقد تكون الشبكة العصبية خيارا جيدا. **بشكل عام، يمكن أن تكون الشبكات العصبية أداة قوية لبناء نماذج الانحدار على البيانات المهيكلة، لكنها ليست دائما الخيار الأفضل لكل مشكلة.**
Classification "التصنيف"	**Statistical Models**	**Generalized linear Regression**	الانحدار الخطي المعمم هو نموذج تصنيف إحصائي، يستخدم لنمذجة العلاقة بين متغير تابع من النوع الثنائي أو **الترتيبي** ومتغير واحد أو أكثر من **المتغيرات المستقلة**. **ويتسم بانة يفترض أن المتغير التابع يتبع توزيعا من عائلة التوزيعات الأسية، والتي تشمل التوزيعات الشائعة مثل التوزيع ذي الحدين وتوزيع بواسون. أي يتيح النموذج المعمم التعامل مع نطاق أوسع من المتغيرات التابعة.**

		Generalized linear Mixed Models	النماذج المختلطة الخطية المعممة (GLMMs) هي نماذج تصنيف إحصائية، يستخدم لنمذجة العلاقة بين متغير تابع من **الوصفي** ومتغير واحد أو أكثر من **المتغيرات المستقلة. وهو امتداد للنماذج الخطية المعممة (GLMs)، والتي تُستخدم لنمذجة العلاقة بين متغير تابع ومتغير واحد أو أكثر من المتغيرات المستقلة. ويتسم بأنه لديه القدرة نمذجة العلاقة بين متغير تابع ومتغير واحد أو أكثر من المتغيرات المستقلة التي تحتوي على مزيج من التأثيرات الثابتة والعشوائية.** في GLMM، يُفترض أن المتغير التابع يتبع توزيعًا من عائلة التوزيعات الأسية، كما هو الحال في GLM. ومع ذلك، في GLMM، يتم التعامل مع بعض المتغيرات المستقلة (التأثيرات الثابتة) على أنها ثابتة، بينما يتم التعامل مع المتغيرات الأخرى (التأثيرات العشوائية) على أنها متغيرات عشوائية تختلف عبر الملاحظات. يسمح هذا للنموذج بتفسير حقيقة أن تأثير بعض المتغيرات قد يختلف من ملاحظة إلى أخرى.
		Discriminant analysis	التحليل التميزي (Discriminant analysis) هو نموذج تصنيف إحصائي، يعرف **على انه من أساليب التحليل متعدد المتغيرات، لدية القدرة على تطوير معالم تمييزية جديدة عبارة عن تراكيب خطية يتم بنائها من المتغيرات المستقلة، لديها القدرة على التميز بين فئات المتغير التابع بطريقة مثالية، في حالة أن المتغيرات المدخلة نرغب في التعامل معها كوحدة واحدة، أي نريد دراسة أثرها مجتمعة في عملية التنبؤ**. وفي علم الإحصاء يلاحظ أن التحليل التميزي (analysis Discriminant) هو الاتجاه المعاكس لتحليل التباين المتعدد (MANOVA) في اتجاه واحد (الذي يسعى للتنبؤ بمجموعة من المتغيرات التابعة من متغير مستقل واحد فقط)، أي يمكن استخدام التحليل التميزي (Discriminant analysis) في حالة وجود متغير مستقل واحد (أسمي) وأكثر من متغير تابع (كمي)، حيث يتم تبديل المتغير المستقل ليصبح تابع والمتغيرات التابعة لتصبح مستقلة، ومن ثم نطبق التحليل التميزي فيتم دمج المتغيرات التابعة بمتغير واحد، ومن ثم نعتبر المتغير الجديد هو التابع والمتغير المستقل الأصلي هو المستقل ونطبق أي نمذجة تنبؤ. ويمكن استخدامه في نمذجة تقليل الأبعاد والتعبير عن المتغيرات بمتغير خطي واحد.
	Decision Trees	Regression &Classification Tree	شجرة التنبؤ والانحدار هي أحد نماذج تعلم الآلة للتصنيف، تعتمد على عملية التقسيم الغصني الثنائي. **تستخدم للتنبؤ بقيم متغير وصفي، عندما تكون المتغيرات المستقلة ذات تسلسل هرمي واضح للميزات أو الفئات. واذا كان حجم البيانات طبيعي (ليس ضخم)، ودقة الفروقات بين قيم المتغيرات المستقلة ليست ذو أهمية**

Random Trees	الأشجار العشوائية هي أحد نماذج تعلم الآلة للتصنيف، تعتمد على عملية التقسيم الساقي (عدد من الأشجار). **تستخدم للتنبؤ بقيم متغير وصفيي، عندما يكون لديك عدد كبير من المتغيرات المستقلة (المميزات) وعدد صغير نسبيًا من البيانات (الصفوف) ، وعندما تريد إنشاء نموذج قوي حالة وجود بيانات صاخبة "Noisy Data " (البيانات الصاخبة هي بيانات تحتوي على كمية كبيرة من المعلومات الإضافية التي لا معنى لها والتي تسمى الضوضاء) والقيم المتطرفة في البيانات. كما أنها سهلة التنفيذ نسبيًا ويمكنها التعامل مع البيانات عالية الأبعاد والقيم المفقودة جيدًا.**	
XGBoost	XGBoost هو تطبيق شائع وفعال لخوارزمية تعزيز التدرج، وهي تقنية تعلم آلي لبناء نماذج تصنيفية. تم تصميم XGBoost خصيصًا ليكون سريعًا وقابلًا للتطوير. تعمل الخوارزمية من خلال بناء سلسلة من أشجار القرار، حيث تتعلم كل شجرة من أخطاء الشجرة السابقة. تسمى هذه العملية التعزيز، وهي تساعد على تحسين الأداء العام للنموذج. **تستخدم للتنبؤ بقيم متغير وصفي، وخاصة في حالة وجود قيم مفقودة وكانت البيانات ضخمة واسعة النطاق منظمة أو غير منظمة. وهي مناسبة بشكل خاص للمهام التي تكون فيها الدقة مهمة ومجموعات البيانات كبيرة.**	
Quick, Unbiased, Efficient Statistical Tree (QUEST)	شجرة QUEST هي أحد نماذج تعلم الآلة للتصنيف، تعتمد على عملية التقسيم الغصني الثنائي. **تستخدم للتنبؤ بقيم متغير وصفي**، تحتاج وقت أقل من وقت تحليلات R Tree & C سواء مع في حالة عدد المتغيرات الكبير أو في حالة وجود متغيرات ذات قيم متعددة. والهدف الثاني هو تقليل الميل الموجود في طرق التصنيف لتفضيل المدخلات التي تسمح بتقسيمات أكثر، أي الحقول ذات المدخلات المستمرة (مجموعة رقمية) أو تلك ذات الفئات العديدة. **تستخدم أذا كان حجم البيانات كبير وليست Big Data، وعدد قيم المتغير الهدف قليل.**	
Chi-squared Automatic Interaction Detection (CHAID)	شجرة مربع كاي للكشف عن التفاعل بشكل تلقائي هي أحد نماذج تعلم الآلة للتصنيف، تعتمد على عملية التقسيم الغصني المتعدد. **تستخدم للتنبؤ بقيم متغير كمي، عندما تكون المتغيرات المستقلة خليط ما بين المتغيرات الكمية والوصفية، فضلاً عن البيانات المفقودة. يمكن للخوارزمية التعامل مع أي عدد من المتغيرات المستقلة، ولكنها تكون**	

			أكثر فاعلية عندما يكون هناك عدد صغير نسبياً من المتغيرات المستقلة ذات عدد كبير من القيم الكمية والوصفية.
		C5.0	شجرة C5.0 هي أحد نماذج تعلم الآلة للتصنيف، تعتمد على عملية التقسيم الغصني المتعدد. تمتاز بقدرتها على تمثيل الشجرة بجمل شرطية تستخدم **للتنبؤ بقيم متغير وصفي**، تعتبر نماذج C5.0 قوية جدًا **في وجود مشكلات مثل البيانات المفقودة وعدد كبير من حقول الإدخال. وعادة لا تتطلب أوقات تدريب طويلة لتقديرها. بالإضافة إلى ذلك، تميل نماذج C5.0 إلى أن تكون أسهل في الفهم من بعض أنواع النماذج الأخرى، حيث أن القواعد المشتقة من النموذج لها تفسير مباشر للغاية. كما تقدم C5.0 طريقة قوية من التعليم المعزز لزيادة دقة التصنيف**
Machine Learning		Self-Learning Response	نموذج التعلم الآلي للاستجابة SLM هي أحد نماذج تعلم الآلة للتصنيف، **تستخدم للتنبؤ بالاستجابات لمجموعة من العروض مرتبة تنازلياً من الأعلى احتمالية استجابة الى الأقل**. وبمعنى أخر تعمل خوارزمية (SLR) على إنشاء نموذج يسمح بالتنبؤ بالعروض الأكثر ملاءمة للعملاء واحتمالية قبول العروض من قبلهم. أي في هذه النمذجة يوجد متغيرين هدف الأول يسمى Target field وهو يدل على العرض المقدم، والثاني يسمى Target response field وهو يدل على نتيجة الاستجابة للعرض من قبل العميل (قبول أو رفض). على سبيل المثال يرغب أحد البنوك إطلاق حملة ترويجية لبرامج القروض وهي (القرض الشخصي، قرض السكن، قرض شراء سيارة، قرض التعلم، قرض للسياحة)، ويرغب البنك بأن تكون الحملة الترويجية موجهة لعملائه، حيث يتم تقديم العرض الأمثل لكل عميل، أي الذي يُحتمل أن يتم قبوله على الأرجح. في هذه الحالة يمكنك استخدام نموذج التعلم الذاتي SLRM لتحديد خصائص العملاء الذين من المرجح أن يستجيبوا بشكل إيجابي بناءً على العروض الترويجية السابقة. **ويستفاد منها أيضا لتحديد الإجراء التالي الأمثل في حالة فشل الإجراء الأعلى احتمالية استجابة وهكذا**. وتتميز هذه الخوارزمية بالتعلم المستمر كلما أضيف بيانات على البيانات الأصلية دون الحاجة لإعادة بناء النموذج من البداية.
		K Nearest Neighbors (KNN)	خوارزمية أقرب K جوار KNN هي أحد نماذج تعلم الآلة للتصنيف، **تستخدم للتنبؤ بقيم متغير وصفي**، حيث تم تطوير نمذجة تحليل الجوار الأقرب KNN كأحد خوارزميات تعلم الآلة للتعرف على أنماط البيانات دون الحاجة إلى تطابق تام مع أي أنماط أو حالات مخزنة، أنما اعتماداً على حالات قريبة منها في الخصائص، أي نحكم على تماثل الحالات بقربها من

بعضها البعض ونحكم على اختلاف الحالات ببعدها عن بعضها البعض. وبالتالي، فإن المسافة بين حالتين هي مقياس التماثل والاختلاف بينهما. وهذا يدل على أن نمذجة تحليل الجوار الأقرب هي أقل دقة في عمليات التصنيف والتنبؤ من النماذج الأخرى، **ولكن هي أكثر تفاعلية مع أي تراكيب معقدة من البيانات أو الحالات، والأكثر استجابة مع أي حالات جديدة أو طفرات في الحالات، ويفضل استخدام خوارزمية KNN إذا كان يوجد اختلاف ما بين بيانات التدريب والبيانات الجديدة المراد تطبيق النموذج عليها**.			
نموذج قائمة القرارات هي أحد نماذج تعلم الآلة للتصنيف، **تستخدم للتنبؤ بالاستجابات لحدث ما**. وهي عبارة عن نموذج خليط ما بين نموذج الاستجابة الذاتية (SLRM) من حيث الغاية (تحديد احتمالية الاستجابة لحدث ما)، ونماذج شجرة القرارات Decisions Tree من حيث الهيكلية تقريباً. حيث يعمل نموذج قائمة القرارات على تقسيم المجتمع الى شرائح Segments، ومن ثم تحديد مجموعة من القرارات (الشروط) لكل شريحة، في حالة تحققها يستدل على الاستجابة للحدث. ويبرز اختلاف نموذج قائمة القرارات عن نموذج الاستجابة الذاتية (SLRM) ونماذج شجرة القرارات Decisions Tree فيما يلي؛ **1.** في نموذج الاستجابة الذاتية (SLRM) يتم بناء نموذج عام لكل مصدر البيانات، في نموذج قائمة القرارات Decision List يتم وضع قرارات (شروط) لكل شريحة من شرائح المجتمع. **2.** نموذج الاستجابة الذاتية (SLRM) غير محدد المعالم، نموذج قائمة القرارات Decision List node يدرج قائمة بالقرارات واضحة المعالم لكل شريحة من شرائح المجتمع وقابلة للتعديل والحذف والإضافة ويستفاد منها في مرحلة التحليل التشخيصي (Diagnostic Analytics). أي يمكن الاستفادة من هذا النموذج ليس فقط في التصنيف، أنما أيضا في تقسيم المجتمع لشرائح ذات مواصفات محددة لغايات التعامل مع كل شريحة على حدة. **3.** في نماذج شجرة القرارات Decisions Tree يتم بناء القرارات بشكل متسلسل؛ المتغير الهدف ← المعيار الأول ← المعيار الثاني ← ... ← القرار حيث يطبق هذا التسلسل على جميع سجلات البيانات، أما في نموذج قائمة القرارات Decision List يتم بناء تسلسل خاص بكل شريحة وليس شرطاً أن يحتوي كل تسلسل جميع المتغيرات المستقلة المحددة لبناء النموذج.	**Decision List**		

خوارزمية شبكة بيز هي أحد نماذج تعلم الآلة للتصنيف، **تستخدم للتنبؤ بقيم متغير وصفي، وخاصة في حالة وجود عدم اليقين (Uncertainty) في الاعتماد على المتغيرات المستقلة في التنبؤ**، ويطلق مصطلح عدم اليقين على الحالات التي لها نفس قيم البيانات (قيم المتغيرات المستقلة) وتعطي معلومات (قيم المتغير الهدف) مختلفة، على سبيل المثال "تشخيص الطبيب لنوع المرض المصاب به المريض من خلال الفحص السريري، علماً بأن الأعراض التي يعاني منها المريض هي أعراض مشتركة بين أكثر من مرض. فيتعرض الطبيب في هذه الحالة لحالة عدم اليقين في تشخيص المرض." ويمكن قياس عدم اليقين من خلال حساب الخطأ المعياري للمتوسط (Standard Error of Mean) للمتغيرات، حيث كلما زادة قيمة الخطأ المعياري للمتوسط يزداد مقدار عدم اليقين. مع العلم زيادة حجم البيانات يعمل على تقليل مقدار عدم اليقين غالباً. **وتستخدم أيضاً في حالة وجود ارتباطات معقدة ما بين المتغيرات المستقلة**. واكتسبت خوارزمية شبكة بيز هذه القدرات المتقدمة لاعتمادها عبى مفاهيم إحصائية عميقة في حساب الاحتمالات الشرطية ونظرية بيز، حيث تعتمد في التنبؤ بقيم المتغير الهدف من خلال قيم الاحتمالات لقيم المتغيرات المستقلة وأثرها على بعضها البعض	**Bayes Network**		
خوارزمية آلة المتجهات الداعمة SVM هي أحد نماذج تعلم الآلة للتصنيف، تستخدم للتنبؤ بقيم **متغير وصفي**، وخاصة **في حالة أن أعداد حقول التنبؤ (المتغيرات المستقلة) كبير قد تصل للآلاف، وحجم بيانات (عدد صفوف قليل)**، وذلك لأنها تعتمد فقط على مجموعة فرعية من نقاط البيانات (تسمى متجهات الدعم) لعمل تنبؤات. وأيضا في الحالات التي لا تكون فيها البيانات قابلة للفصل خطياً، حيث يمكنها استخدام التحويلات غير الخطية للعثور على مستوى للفصل بين البيانات. ومع ذلك، فإن خوارزمية SVM لها أيضا بعض المشاكل، حيث يمكن أن تكون خوارزمية SVM مكلفة من الناحية الحسابية عند التعامل مع مجموعات البيانات الكبيرة، لأنها تتطلب حل مشكلة تحسين تربيعية. ولا توفر خوارزمية SVM تقديرات احتمالية مباشرة، لذلك قد يكون من الصعب تقييم عدم اليقين في التنبؤات.	**Support Vector Machine (SVM)**		
خوارزمية الشبكات العصبية NN هي أحد نماذج تعلم الآلة للتصنيف، **تستخدم للتنبؤ بقيم متغير وصفي**، ولكن يجب أن تكون الخيار الأخير في حالة فشل النماذج الخوارزميات الأخرى، **وعند استخدامها هناك العديد من العوامل التي يجب مراعاتها عند اتخاذ قرار بشأن استخدام الشبكة العصبية لبناء نموذج تصنيفي على البيانات المهيكلة**:	**Neural Networks (NN)**		

			حجم البيانات وتعقيدها: يمكن للشبكات العصبية التعامل مع مجموعات البيانات الكبيرة والمعقدة، ولكن يمكن أن يكون تدريبها وتشغيلها باهظ التكلفة من الناحية الحسابية. إذا كانت لديك مجموعة بيانات كبيرة بها العديد من الميزات، فقد تكون الشبكة العصبية خيارا جيدا. **جودة البيانات**: تتطلب الشبكات العصبية كمية كبيرة من البيانات المصنفة من أجل التعلم بشكل فعال. إذا كانت لديك مجموعة بيانات كبيرة ذات فئات عالية الجودة، فقد تكون الشبكة العصبية خيارا جيدا. **بشكل عام، يمكن أن تكون الشبكات العصبية أداة قوية لبناء نماذج التصنيف على البيانات المهيكلة، لكنها ليست دائما الخيار الأفضل لكل مشكلة.**
Survival analysis "تحليل البقاء"	**Cox Regression**		انحدار كوكس من أحد نماذج تحليل البقاء، هو أسلوب انحدار إحصائي خاص، ي**ستخدم لتحليل البيانات التي تكون فيها النتيجة ذات الاهتمام هي الوقت الذي يستغرقه حدوث حدث مثير للاهتمام**. يستخدم بشكل شائع في البحث الطبي لدراسة الوقت الذي يستغرقه المريض لتجربة حدث ما، مثل الوفاة أو ظهور المرض أو الشفاء، ويستخدم أيضًا في مجالات أخرى، مثل التمويل والهندسة، لدراسة الوقت يستغرق لحدث ما، مثل فشل الجهاز أو التخلف عن سداد قرض. **يعد انحدار كوكس مفيدا لدراسة العوامل التي تؤثر على الوقت الذي يستغرقه حدوث الحدث، ولإجراء تنبؤات حول احتمالية وقوع الحدث في وقت معين**. غالبا ما يستخدم في البحث الطبي لتقييم فعالية العلاجات أو التدخلات على نتائج المرضى، وفي مجالات أخرى لتقييم موثوقية أو متانة المنتجات أو الأنظمة.
Forecasting "التنبؤ المستقبلي"	**Exponential Smoothing**		التنعيم الأسي هو طريقة للتنبؤ ببيانات السلاسل الزمنية (التنبؤ المستقبلي)، والتي تتضمن استخدام البيانات السابقة للتنبؤ بالقيم المستقبلية (البيانات "الصفوف" غير مستقلة عن بعضها البعض). ويستند إلى فكرة أن أحدث نقاط البيانات هي الأكثر ملاءمة للتنبؤ بالقيم المستقبلية، وأن أهمية نقاط البيانات الأقدم تتناقص بشكل كبير بمرور الوقت. ويتم تم تطبيقها غالبة إذا كانت السلسلة لا تحتوي مركبة اتجاه عام (صاعد أو هابط)، وعندما تكون البيانات قليلة التردد (التقلب) وعندما تأخذ السلسلة نمط أسي في الزيادة أو النقصان.
	ARIMA		ARIMA هي طريقة إحصائية تستخدم للتنبؤ ببيانات السلاسل الزمنية. إنه نوع من النماذج الخطية المستخدمة لالتقاط الارتباط التلقائي في البيانات، بالإضافة إلى أي اتجاه عام وموسمية.

		تتكون نماذج ARIMA من ثلاثة مكونات: مكون الانحدار الذاتي (AR) ، الذي يقوم بنمذجة الارتباط التلقائي في البيانات؛ المكون المتكامل (I)، الذي يصوغ الفرق بين البيانات والتنبؤ السابق ؛ ومكوِّن المتوسط المتحرك (MA) ، الذي يصوغ المخلفات (الأخطاء) للتنبؤ السابق. بمجرد ملاءمة نموذج ARIMA للبيانات، يمكنك استخدامه لإنشاء تنبؤات للفترات الزمنية المستقبلية. يتم إنشاء التنبؤ باستخدام النموذج لتقدير القيمة المتوقعة للسلسلة الزمنية في الفترة الزمنية المستقبلية، بناءً على البيانات السابقة ومعلمات النموذج. تعد ARIMA طريقة قوية للتنبؤ ببيانات السلاسل الزمنية، خاصةً عندما تظهر البيانات الارتباط التلقائي والاتجاه العام والموسمية

Association Models "نماذج الارتباط"		
Action	**Tool or Algorithm**	**The Aim**
Association "نماذج المشاركة"	**Sequence**	خوارزمية Sequence **تستخدم للكشف عن الارتباطات بين العناصر** (الارتباط ما بين قيم المتغيرات وليس ارتباط المتغيرات مع بعضها البعض)، **عندما يوجد اهتمام بعملية التوالي في ظهور العناصر (العنصر الأول ومن ثم الثاني ...)**، ويكون الزمن محدد ومتماثل لكل عنصر من عناصر الدراسة. أي تستخدم للتنبؤ بوقوع الحالة التالية لوقوع حالة أو أكثر.
	Carma	خوارزمية Carma **تستخدم للكشف عن الارتباطات بين العناصر** (الارتباط ما بين قيم المتغيرات وليس ارتباط المتغيرات مع بعضها البعض)، **عندما يوجد اهتمام بظهور العناصر دون الاهتمام بأولوية الوقوع (أي جميعها بنفس المستوى).**
	Apriori	خوارزمية Apriori **تستخدم للكشف عن الارتباطات بين العناصر** (الارتباط ما بين قيم المتغيرات وليس ارتباط المتغيرات مع بعضها البعض)، **عندما يكون الاهتمام بعملية التوالي دون الاهتمام بزمن وقوع الحالات.** أي ليس شرط ان يكون الزمن محدد ومتماثل لكل عناصر الدراسة. أي تستخدم للتنبؤ بوقوع الحالة التالية لوقوع حالة أو أكثر دون الاهتمام بزمن الوقوع. وتتميز هذه الطريقة بما يلي: أنها تتعامل مع البيانات الضخمة بكفاءة وسرعة عالية، ليس له عدد محدد على القواعد التي يمكن تتبعها، لا يتعامل إلا مع المتغيرات الفئوية. يحدد عدد التنبؤات آلياً. ويتم تحديد الحالات القبلية (Antecedents) وحالات النتيجة (Consequents).

خوارزمية قواعد المشاركة Association rules **تستخدم للكشف عن الارتباطات بين العناصر** (الارتباط ما بين قيم المتغيرات وليس ارتباط المتغيرات مع بعضها البعض)، **عندما يكون الاهتمام بعملية التوالي دون الاهتمام بزمن وقوع الحالات.** أي ليس شرط أن يكون الزمن محدد ومتماثل لكل عناصر الدراسة أو لا يوجد. أي تستخدم للتنبؤ بوقوع الحالة التالية لوقوع حالة أو أكثر دون الاهتمام بزمن الوقوع. وتتميز هذه الطريقة بما يلي: نلاحظ التشابه الواضح بين هذه النمذجة ونمذجة Apriori، مع وجود بعض الاختلافات. حيث تعتمد نمذجة قواعد المشاركة على بناء الجمل الشرطية ذات قيمة (Entropy) عالية، حيث يدل مقدار Entropy على مقدار المعلومات (Information) التي تقدمها الجملة الشرطية؛ **if** condition(s) **then** prediction(s) على سبيل المثال، "إذا اشترى عميل معكرونة وبعد ذلك لحمة، فسيشتري هذا العميل معجون الطماطم بثقة 95٪" (مثال بسيط للتوضيح) لأن هدفا أعمق من ذلك وهو الوصول للرؤى الخفية (Hidden Insights). حيث تستخرج عقدة Association Rules مجموعة من القواعد (جمل شرطية) من مجموعة البيانات.	**Association rules**	

Evaluation "التقييم"		
The Aim	**Tool or Algorithm**	**Action**
لتقييم نماذج التصنيف يجب قياس مجموعة من المؤشرات، وهي؛ **10. بناء مصفوفة الارتباك (Confusion Matrix).** **11. الدقة العامة (Accuracy)** **12. الخطأ (Error)** **13. الدقة الحقيقية الإيجابية (Precision true positive)** **14. الدقة الحقيقية السلبية (Precision true negative)** **15. الحساسية (معدل الإيجابية الحقيقي) Sensitivity ((rate (true positive** **16. النوعية (معدل السلبية الحقيقي) Specificity (true (negative rate)** **17. معدل الخطأ الإيجابي (False Positive Rate)** **18. معدل الخطأ السلبي (False Negative Rate)**	**Classification Models Evaluation**	**Models Evaluation** "تقييم النموذج"

لتقييم نماذج التصنيف يجب قياس مجموعة من المؤشرات، وهي؛ 1. **الحد الأدنى للخطأ (Minimum Error)** 2. **الحد الأعلى للخطأ (Maximum Error)** 3. **معدل الخطأ (Mean Error)** 4. **معدل الخطأ المطلق (Mean Absolute Error)** 5. **الانحراف المعيار للخطأ (Standard Deviation Error)** 6. **الارتباط الخطي (Linear Correlation)** يعتبر تطابق البيانات الحقيقية (Actual) للمتغير الهدف مع البيانات المتنبئ بها (Predicted) للمتغير الهدف من أهم المؤشرات الدالة على جودة النموذج. حيث يتم قياس هذه المؤشر من خلال دراسة الارتباط الخطي (r) بين البيانات الحقيقية (Actual) للمتغير الهدف مع البيانات المتنبئ بها (Predicted) للمتغير الهدف. حيث كلما اقتراب قيمة الارتباط الخطي الى 1+ دل على زيادة جودة النموذج.	**Regression Models Evaluation**	

الخاتمة

من المحتمل أن يكون الدليل الشامل لعلوم البيانات وهندسة تعلم الآلة شمل مجموعة واسعة من الموضوعات، بما في ذلك التحليل الإحصائي والتعلم الآلي وتصور البيانات والمزيد. في ختام هذا الدليل، سيكون من المهم تلخيص النقاط الرئيسية التي تمت تغطيتها في الفصول السابقة والتفكير في الأهمية العامة والملاءمة للمعلومات المقدمة.

بالإضافة إلى ذلك، قد يقدم دليل علم البيانات اقتراحات لمزيد من القراءة أو الدراسة، أو يوفر إرشادات حول كيفية تطبيق المفاهيم والتقنيات الموضحة في الكتاب في مواقف العالم الحقيقي.

تم بحمد الله