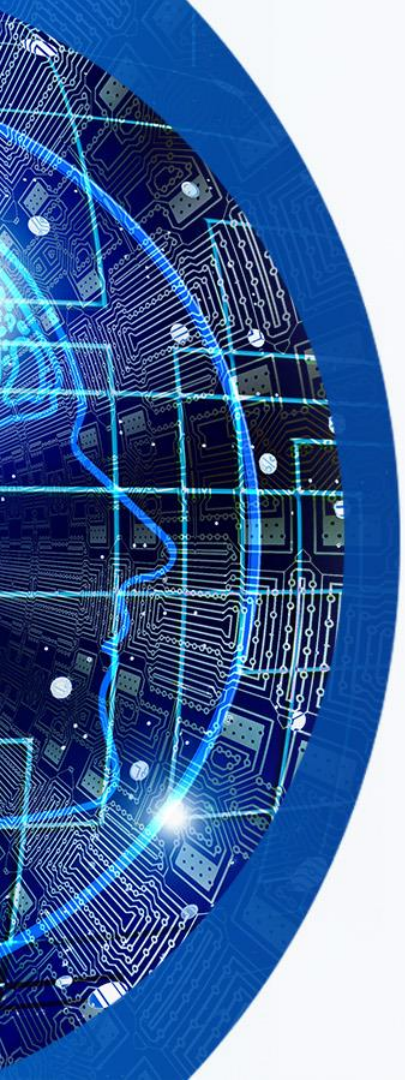
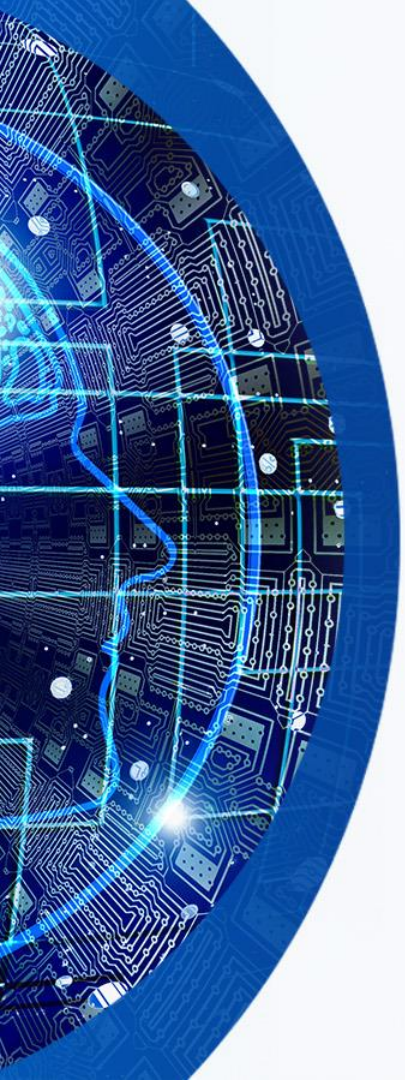
الفصل الثاني : ادارة البيانات

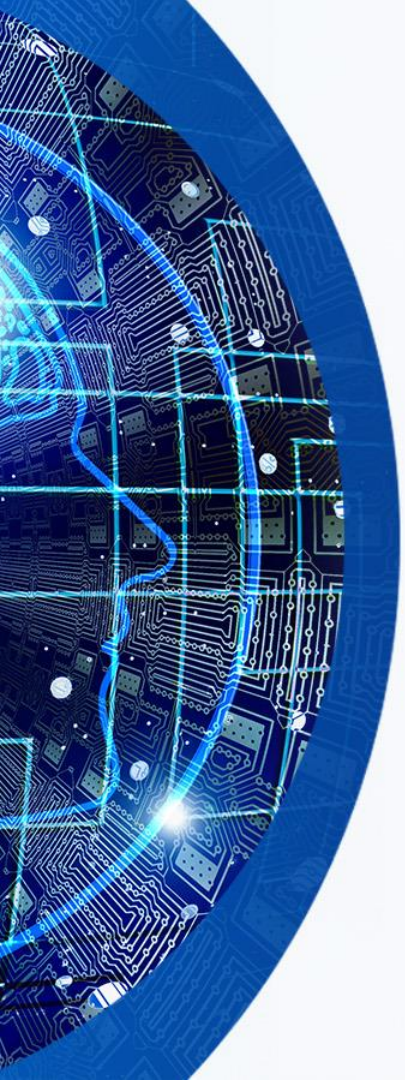


1- التعرف على أنواع البيانات.
2- معالجة و تحضير وتنقية البيانات.
3-الدراسة الاحصائية للبيانات
4-التمثيل المرئي للبيانات



مقدمة

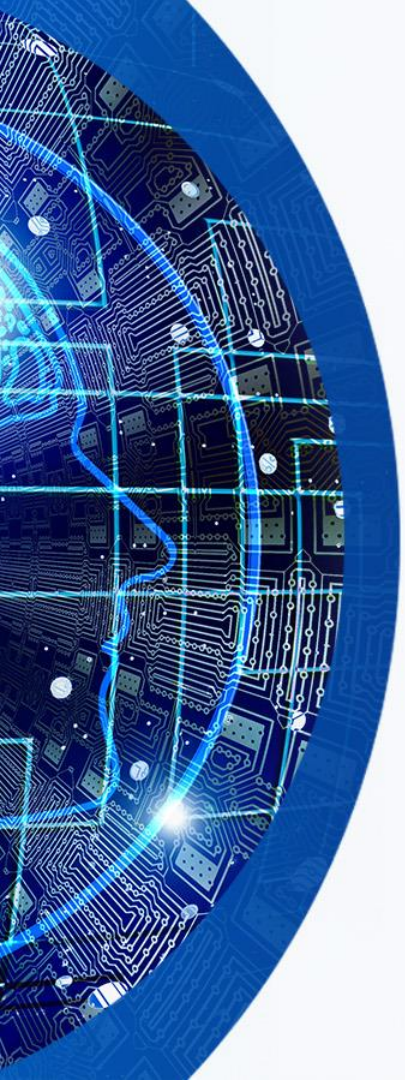
- من المهم تخزين البيانات ومعالجتها بشكل صحيح دون أخطاء . يلعب نوع البيانات دورا مهما في تحديد استراتيجية المعالجة المسبقة للحصول على النتائج المناسبة أو نوع التحليل الإحصائي الذي يجب استخدامه للحصول على أفضل النتائج. يتيح لنا فهم أنواع البيانات المختلفة اختيار نوع البيانات الذي يناسب دراستنا و وأهدافنا.



انواع البيانات

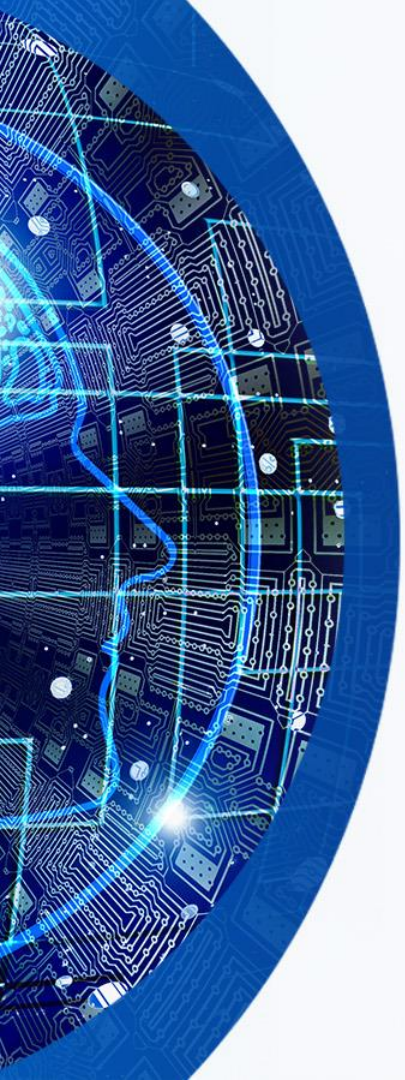
1. **البيانات النوعية** Qualitative Data

تصف البيانات النوعية ، أو بيانات الفئوية ، تكون باستخدام مجموعة محدودة من الفئات المنفصلة. وهذا يعني أن هذا النوع من البيانات لا يمكن عده أو قياسه بسهولة باستخدام الأرقام و بالتالي فهو مقسم إلى فئات. جنس الشخص (ذكر أو أنثى) يعد جنس الاشخاص وألوانهم وأعراقهم أمثلة على هذا النوع من البيانات. على سبيل المثال . يمكن تصنيف كل هذه المعلومات على أنها بيانات نوعية. هناك نوعان عامان من البيانات النوعية: البيانات الاسمية Categorical Data و المتسلسلة. Ordinal



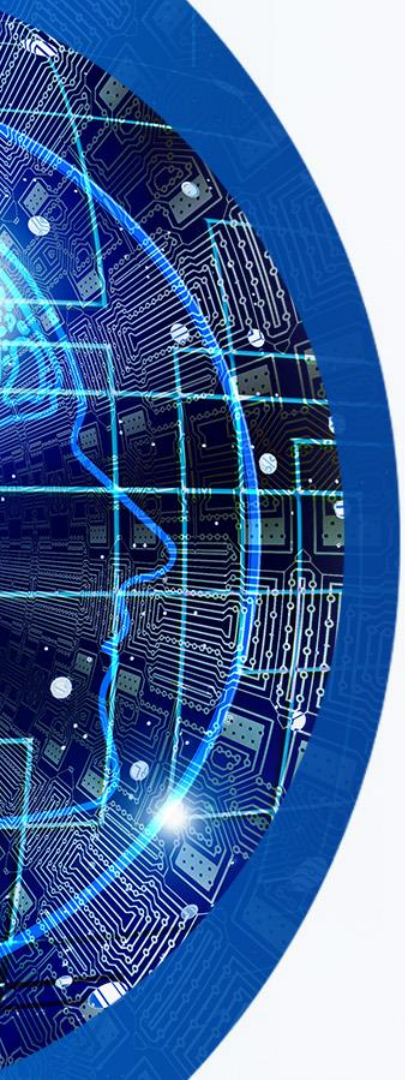
البيانات الاسمية

- يتم تعريف البيانات الاسمية على أنها بيانات تستخدم لتسمية المتغيرات أو عنونتها بدون أي كمية. عادة لا يوجد ترتيب جوهري للبيانات الاسمية. على سبيل المثال ، مثال لون العين هو متغير اسمي له عدة فئات(أزرق ، أخضر ، بني) ولا توجد طريقة لترتيب هذه الفئات من الأعلى إلى الأدنى.



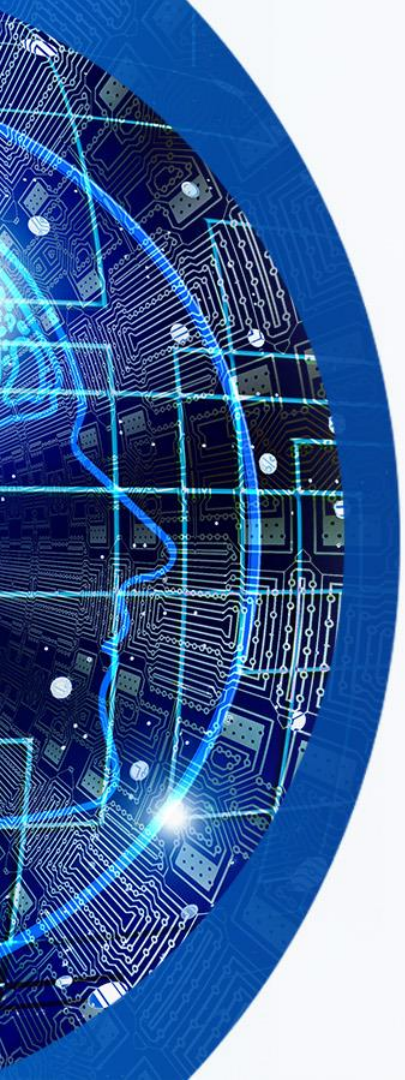
البيانات المتسلسلة

•البيانات المتسلسلة هي نوع من البيانات المصنفة و المرتبة. متغيرات البيانات المتسلسلةمدرجة بالترتيب. عادة ما يتم ترقيم المتغيرات التسلسلية للإشارة إلى ترتيب القائمة. ومع ذلك ، لا يتم قياس الأرقام أو تحديدها رياضيا ، ولكن فقط يتم تعيينها كعنوان تعليق. على سبيل المثال،إذا أخذنا في الاعتبار حجم إحدى العلامات التجارية للملابس ، فيمكننا تصنيفها بسهولة إلى صغيرة ومتوسطة وكبيرة ، على التوالي .



البيانات الكمية Quantitative Data

- البيانات الكمية هي بيانات قابلة للقياس. بمعنى آخر ، يمكن حسابها أو قياسها ويمكن الحصول على قيمة عددية لها. سعر الهاتف الذكي ، والخصم المعروض ، وتردد معالج الهاتف الذكي أوذاكرة الوصول العشوائي لهذا الهاتف ، كلها تندرج في فئة أنواع البيانات الصغيرة. و هناك عددا لا حصر له من القيم التي يمكن أن تحتوي عليها الصفة. رئيسيان من البيانات الكمية نوعان : البيانات المتقطعة و البيانات المستمرة



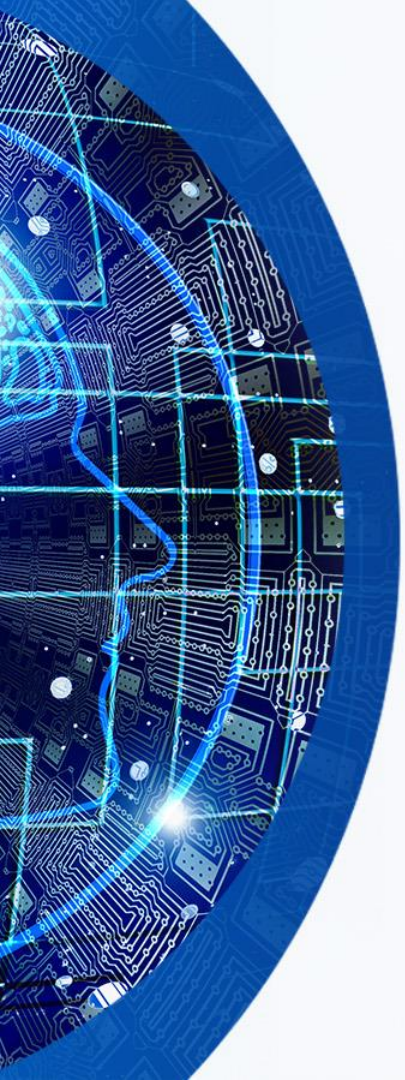
البيانات المتقطعة Discrete data

- البيانات المتقطعة قابلة للعد وتحتوي على أعداد صحيحة فقط. عدد مكبرات الصوت المحمولة،وعدد الكاميرات ، وعدد مراكز المعالج ، وعدد بطاقات SIM المدعومة كلها أمثلة على أنواعالبيانات المنفصلة.

- **البيانات المستمرة** Continuous data

- البيانات المستمرة هي البيانات التي يمكن تقسيمها بشكل كبير إلى مستويات أكثر دقة. يمكن قياسه على مقيا س scale أو بشكل مستمر continuum

- ويمكن أن يكون له أي قيمة عددية تقريبا. على سبيل المثال، يمكنك قياس طولك بمقاييس دقيقة للغاية ، مثل الأمتار ، والسنتيمتر ، والمليمترات ، وما إلى ذلك. يمكنك تسجيل البيانات المستمرة في قياسات مختلفة مثل درجة الحرارة و الوقت وما إلى ذلك.

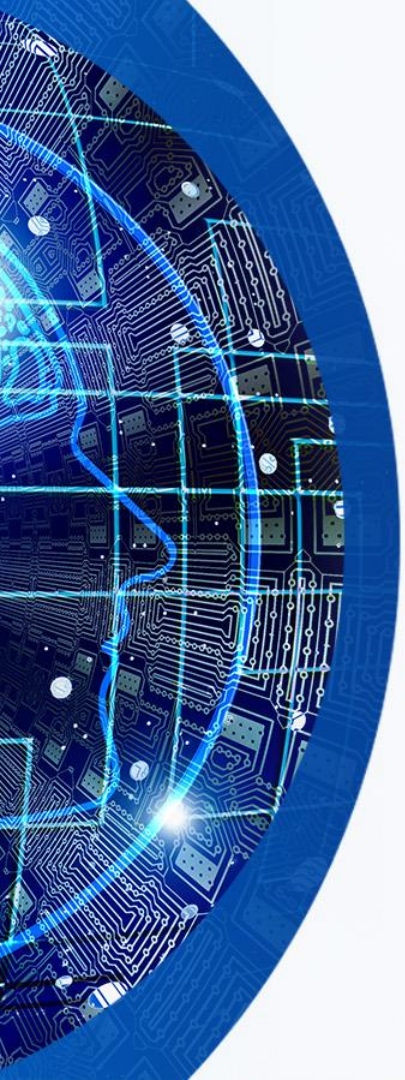


- **بيانات السلاسل الزمنية** Time Series Data

- تحليل السلاسل الزمنية هو طريقة خاصة للتحليل المتسلسل لنقاط البيانات التي تم جمعهاخلال فترة زمنية.

- سواء كنا نريد التنبؤ باتجاهات السوق المالية أو استهلاك الكهرباء ، فإن الوقت عامل مهم يجب الآن مراعاته في نماذجنا. على سبيل المثال توقع وقت ذروة استهلاك الكهرباء في اليوم. للقيام بذلك ، يكفي استخدام بيانات السلاسل الزمنية. بيانات السلاسل الزمنية هي سلسلة من الأرقام التي يتم جمعها على فترات منتظمة خلال فترة زمنية.

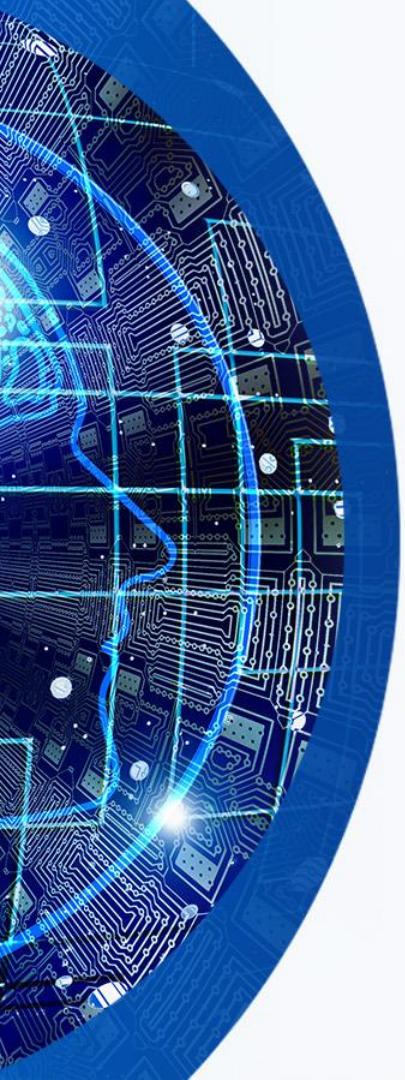
- في السلسلة الزمنية ،غالبا ما يكون الوقت متغيرا مستقلا والهدف عادة هو عمل تنبؤات للمستقبل.



هياكل البيانات Data frame

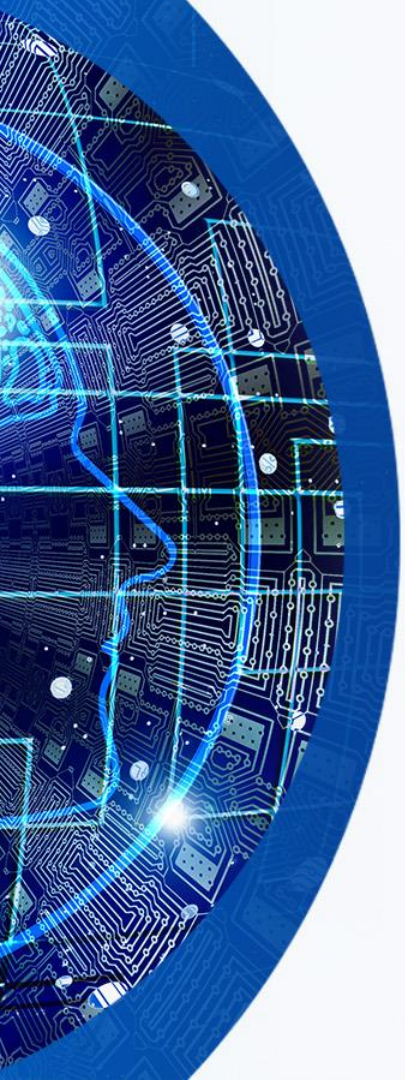
غالبا ما تكون هياكل البيانات عبارة عن ملف يكون فيه كل كائن في صف وكل عمود يتوافق مع إحدى ميزات هذه الكائنات. على سبيل المثال، يوضح الجدولل ادناه مجموعة بيانات تحتوي على معلومات الطلاب. يشير كل صف إلى طالب، وكل عمود عبارة عن ميزة تصف بعض جوانب الطالب، مثل رقم الطالب وسنة الالتحاق ومتوسط درجة النقاط ومجال الدراسة.

رقم الطالب	سنة الالتحا ق	المعد ل	مجال الدراسة
976001	2020	14.45	هندسة ميكانيكية
974120	2021	15.03	اعلام الي
990245	2022	13.95	علوم اقتصادية



الميزات العامة للهياكل البيانات

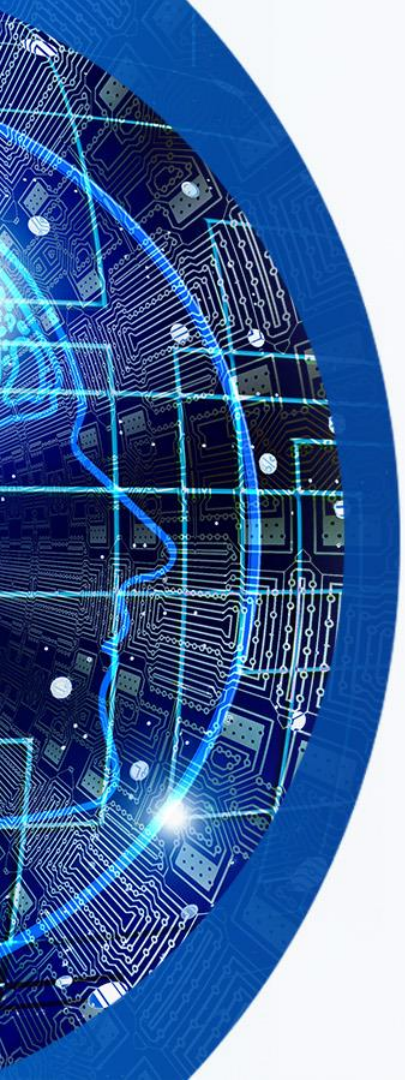
1. ثلاث خصائص عامة تستخدم عند استعمال العديد من هياكل البيانات ولها تأثير كبير على استخدام تقنيات التعلم الآلي هي:
2. الأبعاد dimensionality
3. التشتت sparsity
4. الدقة. resolution



ا لابعاد: أبعاد هياكل البيانات هي عدد الميزات التي تمتلكها الكائنات في مجموعةالبيانات. تختلف البيانات منخفضة الأبعاد نوعيا عن البيانات المتوسطة أو عالية الأبعاد. في الواقع، أحيانا تسمى المشكلات المتعلقة بتحليل البيانات عالية الأبعاد بالمشكلات متعددة الأبعاد..

التشتت: يعد التشتت ميزة ، حيث يجب تخزين وترتيب القيم غير الصفرية فقط. هذا يوفر وقتا كبيرا في الحساب والتخزين.

الدقة (دقة العرض) : تختلف خصائص البيانات باختلاف درجات الدقة.. تعتمد الأنماط في البيانات أيضا على مستوى الدقة.



أمثلة على البيانات عالية الأبعاد

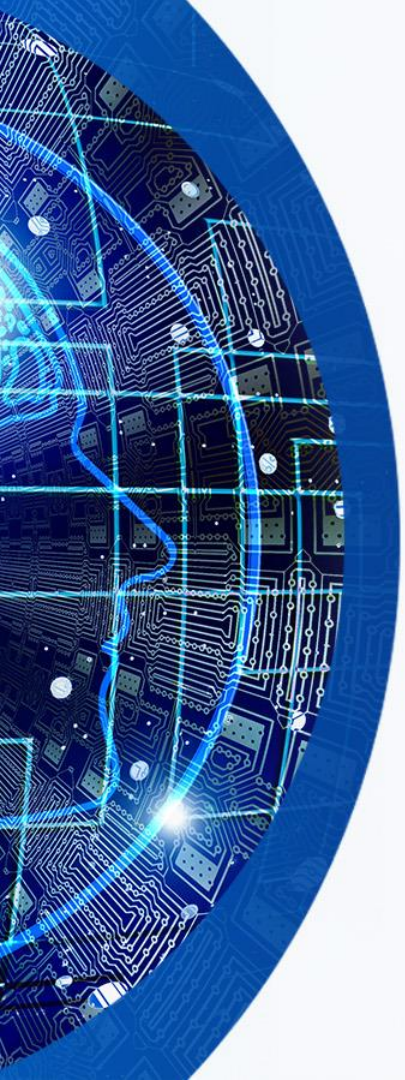
توضح الأمثلة التالية بيانات عالية الأبعاد في مجالات مختلف :

مثا ل 1: البيانات المالية

البيانات عالية الأبعاد شائعة أيضا في مجموعات البيانات المالية، حيث يمكن أن يكون عددخصائص سهم معين كبيرا جدا، على سبيل المثال، حجم التداول ونسبةالسعر إلى الربح والقيمة السوقية للسهم و ربحية السهم و معدل توزيع الأرباح وما إلى ذلك. في هذا النوع من البيانات، و قد يكون عدد السمات أكبر بكثير من عدد المشاهدات.

مثا ل 2: البيانات الصحية

البيانات عالية الأبعاد شائعة في مجموعة من البيانات الصحية أن عدد خصائص شخص معين يمكن أن يكون كبيرا جدا ،على سبيل المثال ،ضغط الدم، ومعدل ضربات القلب أثناء الراحة ، والحالة المناعية، والتاريخ الجراحي، والطول، والوزن، والحالة، إلخ. في مجموعة البيانات هذه ، من الشائع أن يتجاوز عدد السمات عدد المشاهدات.

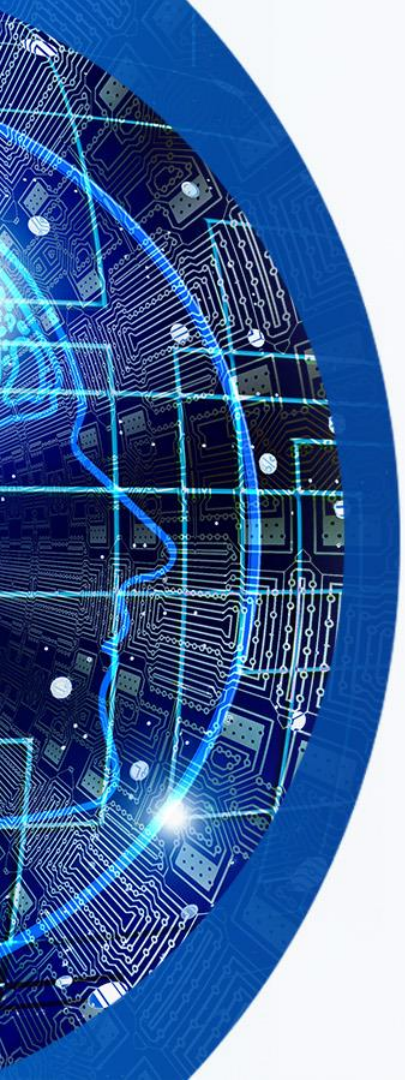


جمع البيانا ت Data collection

- يعد إنشاء مجموعة بيانات كبيرة أمرا شاقا إذا تم إجراؤه يدويا. ولكن يمكن لأساليب مثل تجريف الويب و زاحف الشبكة أتمتة عملية جمع البيانات وتسهيل إنشاء مجموعات البيانات للتحليل.

1. **تجريف ويب** web scraping

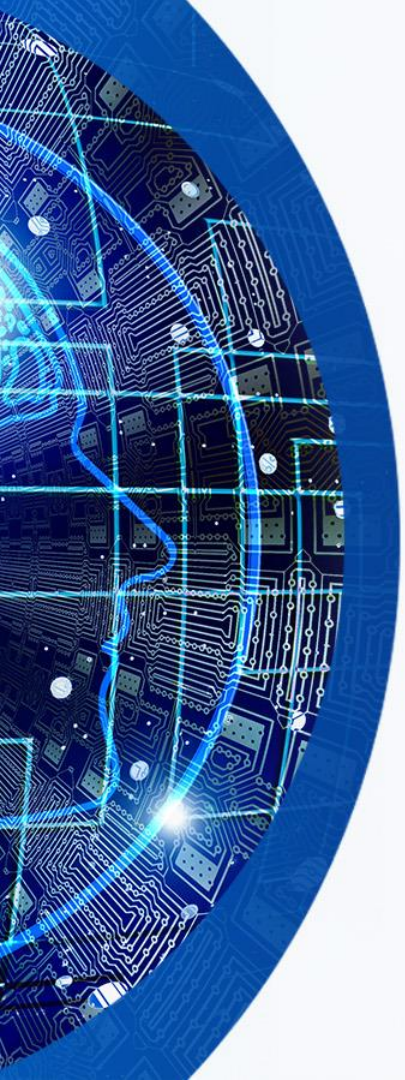
- في عالم اليوم التنافسي، يبحث الجميع عن طرق للابتكار واستخدام تقنيات جديدة. و يوفر تجريف الويب حلا لأولئك الذين يرغبون في الوصول تلقائيا إلى بيانات الويب المنظمة. يعد تجريف الويب مفيدا إذا كان موقع الويب العام الذي تريد تلقي المعلومات منه لا يحتوي على واجهة برمجة تطبيقات أو لديه وصول محدود فقط إلى البيانات.



•يمكن تقسيم عملية تجريف الويب أكملها إلى مراحل مختلفة وشرحها بإيجاز على النحو التالي :

1. **المرحلة الأولى** –جلب البيانا ت: Fetch dataفي هذه الخطوة ، يجب تحديد مواقع الويب التي يمكن من خلالها الوصول إلى البيانات. يمكن بعد ذلك إجراء الجلب باستخدام بروتوكول HTTP، وهو بروتوكول إنترنت يُستخدم لإرسال الطلبات وتلقيها من خادم الويب .

- **المرحلة الثانية** - استخراج المعلوما ت: Extracting Information بعد جلب مستندات HTML المطلوبة ، فإن الخطوة التالية هي استخراج المعلومات التي نحتاجها من موقع الويب. يمكن القيام بذلك باستخدام عدة تقنيات مثل تحليل HTML و DOM و XPath ومطابقة أنماط النص.
- **المرحل ة الثالثة** - تحويل البيانا ت Data Transformation : بعد استخراج المعلومات المطلوبة من المواقع المطلوبة URL ، ستكون البيانات غير منظمة. يمكن بعد ذلك تحويلها إلى نموذج منظم مثل CSVأو جدول بيانات أو pdf، للعرض التقديمي أو التخزين .

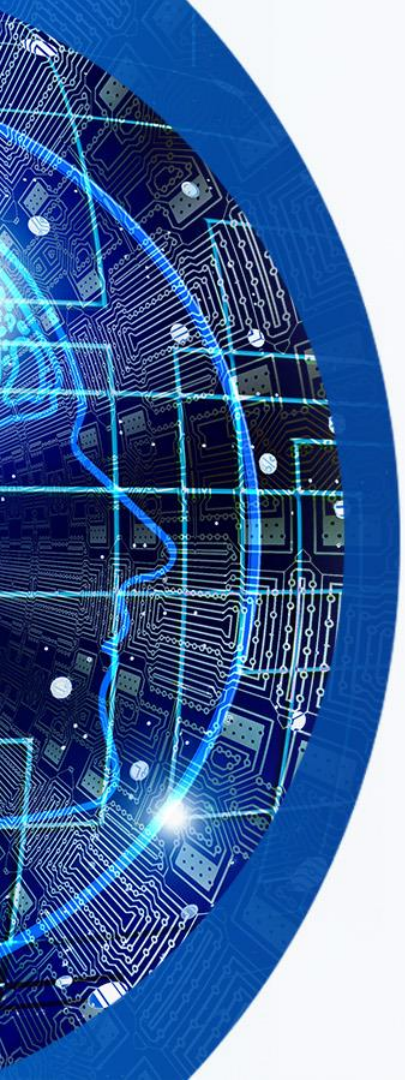


، قواعد البيانات

- بعد استخراج البيانات من الويب بنجاح، فإن الخطوة التالية هي تخزينها بتنسيق قابل للاستخدام. هناك العديد من تنسيقات الملفات المتاحة لتخزين البيانات المحذوفة من الويب، بما في ذلك CSV و Excelوقواعد البيانات. يمكن أن يعتمد اختيار التنسيق الصحيح على طبيعة البيانات وحجمها والاستخدام المقصود للبيانات.

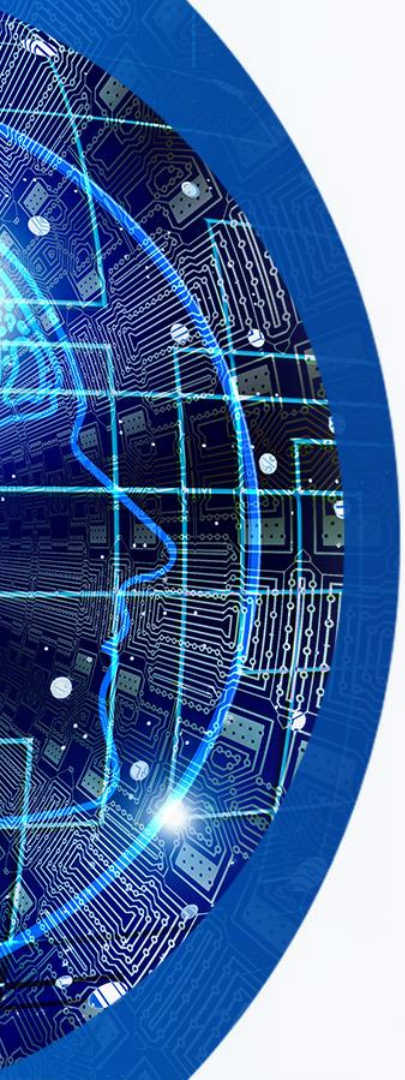
- **Comma-separated values** (CSV .

- CSV(القيم المفصولة بفواصل) هو تنسيق ملف شائع يستخدم لتخزين البيانات الجدولية. إنه تنسيق بسيط، حيث يمثل كل صف سجلاً وكل عمود يمثل حقلاً. إحدى المزايا الرئيسية لملف CSV هي أنه سهل القراءة والكتابة، ويمكن فتحه في أي محرر نصوص أو برنامج جداول بيانات. علاوة على ذلك، فإن ملفات CSV خفيفة الوزن وتستهلك مساحة أقل مقارنة بالتنسيقات الأخرى. ومع ذلك، تتمتع ملفات CSV بدعم محدود لأنواع البيانات وقد لا تكون مناسبة لتخزين هياكل البيانات المعقدة.



- .اكسل

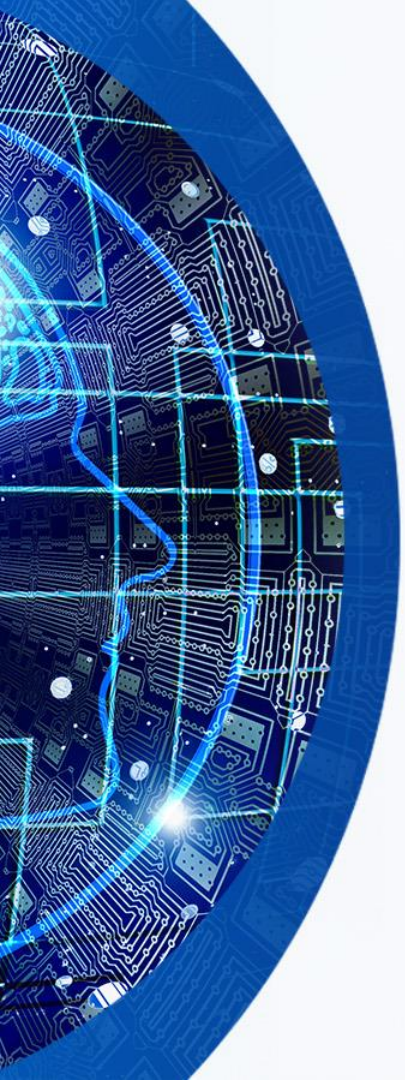
- Excelهو برنامج جداول بيانات يستخدم على نطاق واسع ويسمح للمستخدمين بتخزين البيانات ومعالجتها وتحليلها. تشبه ملفات Excel ملفاتCSV ، ولكنها تحتوي على ميزات إضافية مثل التنسيق والصيغ والمخططات. إحدى المزايا الرئيسية لبرنامج Excel هي أنه يوفر واجهة سهلة الاستخدام لتحليل البيانات وتصورها. علاوة على ذلك، يمكن لملفات Excel التعامل مع مجموعات بيانات أكبر من ملفات CSV وتوفير دعم أفضل لأنواع البيانات. ومع ذلك، يمكن أن تصبح ملفات Excel بطيئة وغير مستقرة عند التعامل مع مجموعات كبيرة من البيانات، وقد لا تكون مناسبة لتخزين البيانات التي تتطلب تحديثات متكررة.



2- معالجة البيانات و تحضيرها

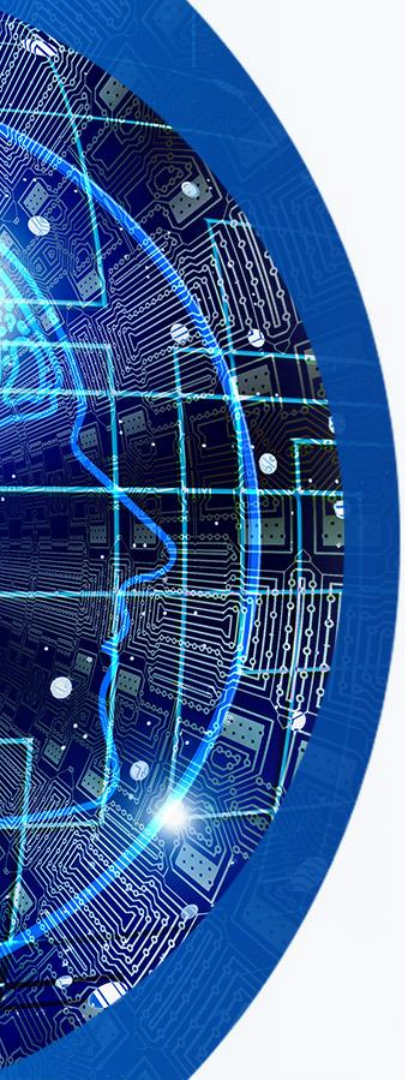
- قد تكون هناك مشاكل بسبب خطأ بشري أو عيوب في عملية جمع البيانات. او قد لا تكون بعض القيم موجودة، وفي حالات أخرى، قد توجد عناصر مزيفة أو مكررة. على سبيل المثال، قد تكون هناك حالتان مختلفتان لشخص عاش مؤخرا في عنوانين مختلفين. حتى إذا كانت جميع البيانات متوفرة وتبدو جيدة، فقد يكون هناك تناقضات، على سبيل المثال، يبلغ طول الشخص مترين، لكنه يزن 20 كلغ فقط.

مما يعني أنها تتطلب التنظيف والمعالجة المسبقة قبل أن تتمكن من استخدامها للتحليل. في هذا القسم، سنستكشف كيف يمكنك تنظيف البيانات ومعالجتها مسبقاً باستخدام R.



تنظيف البيانات Cleaning data

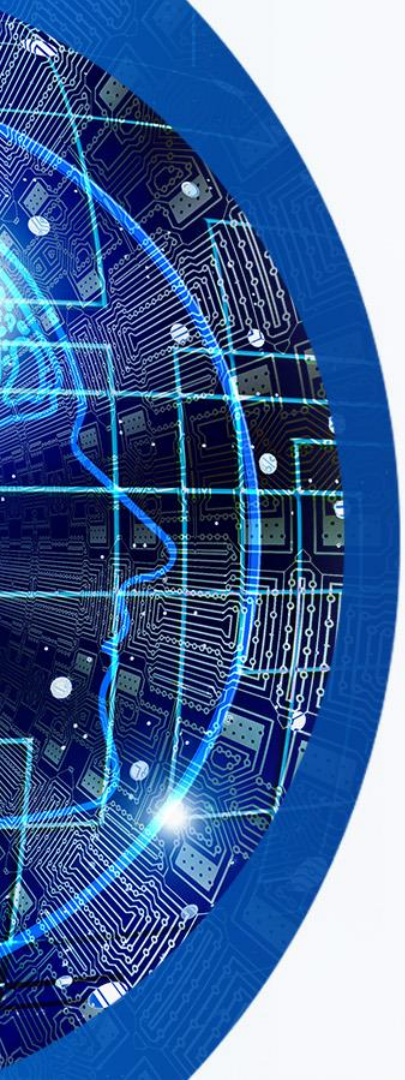
تنظيف البيانات هي عملية إعداد البيانات للتحليل عن طريق إزالة أو تعديل البيانات غير الصحيحةأو غير الكاملة أو غير الملائمة أو المكررة أو غير المناسبة. عادة ما تكون هذه البيانات غير ضروريةأو مفيدة في تحليل البيانات، لأنها قد تعطل العملية أو تقدم نتائج غير دقيقة. هناك عدة طرق لتنظيف البيانات، اعتمادا على كيفية تخزين المعلومات والاستجابات.



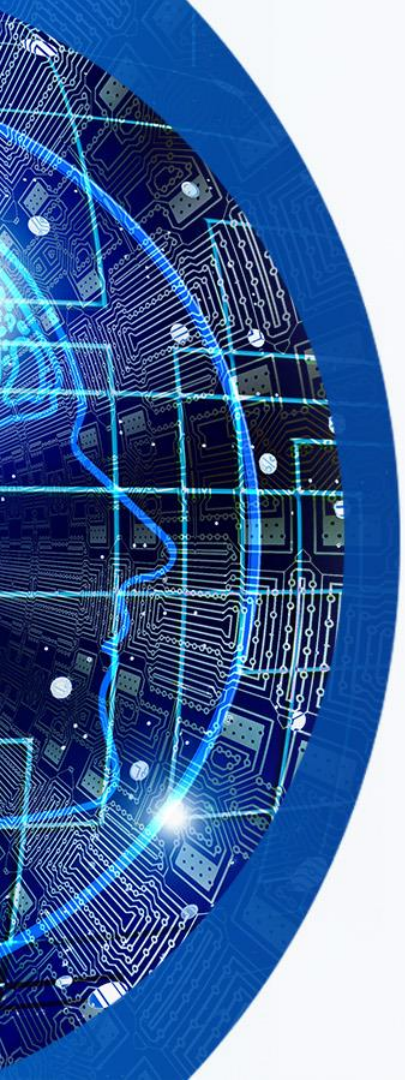
القيم المفقودة Missing Values

- في بعض الأحيان قد تكون البيانات بالتنسيق الصحيح، ولكن بعض القيم مفقودة. مثلا جدول يحتوي على معلومات العميل التي لا تتضمن بعض أرقام هواتف المنزل. قد يكون هذابسبب أن بعض الأشخاص ليس لديهم هاتف منزلي وبدلا من ذلك يستخدمون هواتفهم المحمولة كهاتف رئيسي.

- وجود قيم مفقودة في بياناتك ليس بالضرورة مشكلا ومع ذلك، يمكن للبيانات المفقودة ان تؤدي الى نتائج غير صحيحة. و هناك عدة طرق للتعامل مع هذه المشكلة، لكن لكل طريقة مزايا وعيوب.

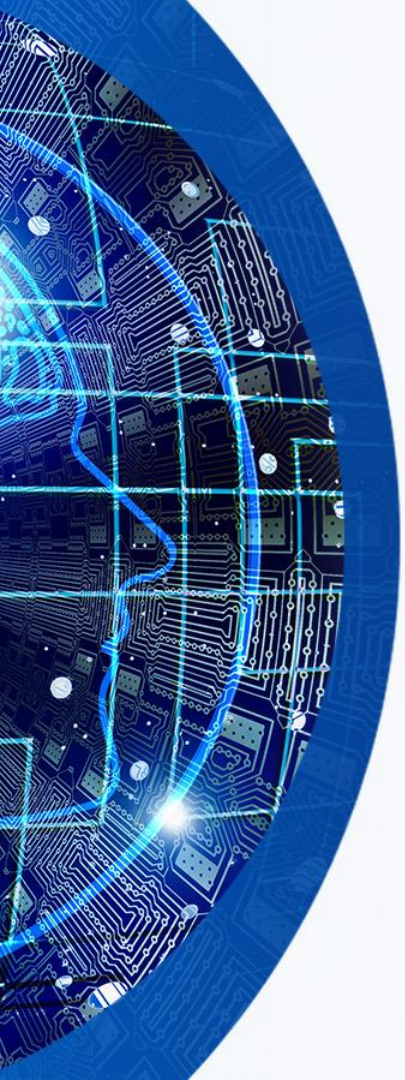


- 1- **حذف العناصر أو خصائص البيانا ت**: إستراتيجية بسيطة وفعالة لحذف العناصر ذات القيم المفقودة. ومع ذلك، حتى كائن البيانات يحتوي على بعض المعلومات، وإذا كان عددا كبيرا من العناصر يحتوي على قيم مفقودة، فقد يكون التحليل الموثوق به صعبا أو مستحيلا. ومع ذلك، إذا كان عدد قليل من العناصر في مجموعة البيانات تحتوي على قيم مفقودة، فقد يكون حذفها مفيدا. تتمثل الإستراتيجية ذات الصلة في حذف السمات التي تحتوي على قيم مفقودة. و لكن يجب أن يتم ذلك بحذر، حيث قد تكون الميزات المحذوفة ميزة مهمة في التحليل.



- 2- **تقدير القيم المفقود ة** : في بعض الأحيان يمكن تقدير البيانات المفقودة بشكل موثوق. على سبيل المثال، سلسلة زمنية تتغير منطقيا ولكن بها بعض القيم المبعثرة المفقودة. في مثل هذه الحالات، يمكن تقدير القيم المفقودة (استكمالinterpolated) باستخدام القيم المتبقية.
- 3-**البيانات المتشابهة**. في هذه الحالة، غالبا ما تُستخدم قيم النقطة القريبة من القيمةالمفقودة لتقدير القيمة المفقودة. إذا كانت السمة متصلة، فسيتم استخدام **متوسط قيمةسمة الجيران الأقرب**. إذا كانت الصفة منفصلة، فيمكن اعتماد القيمة الأكثر شيوعا للسمة.

- 4-**تجاهل القيم المفقودة** أثناء التحليل: هناك طريقة أخرى للتعامل مع البيانات المفقودة وهي تجاهل هذه القيم. على سبيل المثال، افترض أن العناصر مجمعة ويجب حساب التشابه بين أزواج عناصر البيانات. إذا كان أحد العناصر أو كلاهما يحتوي على قيم مفقودة لبعض الخصائص، فلا يمكن حساب التشابه إلا باستخدام الخصائص التي لاتحتوي على قيم مفقودة.



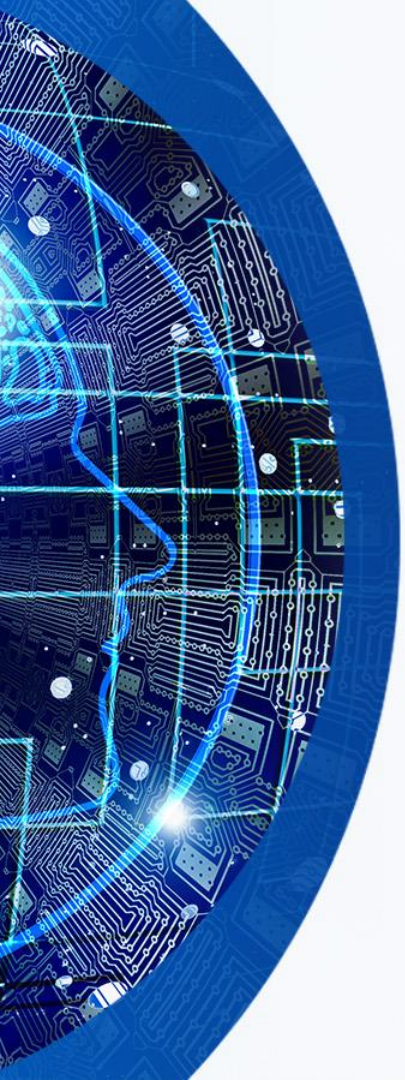
التعامل مع القيم المفقودة Handling missing values

- غالباً ما تحتوي البيانات على قيم مفقودة، والتي يمكن تمثيلها على أنها `NA` أو سلسلة فارغة.
- للتعامل مع القيم المفقودة، يمكنك استخدام الدالة `is. Na()` للتحقق مما إذا كانت هناك قيمة مفقودة، والدالة `na. Omit()` لإزالة القيم المفقودة من مجموعة بيانات. على سبيل المثال، إذا كان لديك إطار بيانات يحتوي على قيم مفقودة، فيمكنك إزالتها بهذه الطريقة:

```
Df <- data. Frame(x = c(1, 2, NA, 4), y = c("a" ،"b" ،"" ،"d"))
```

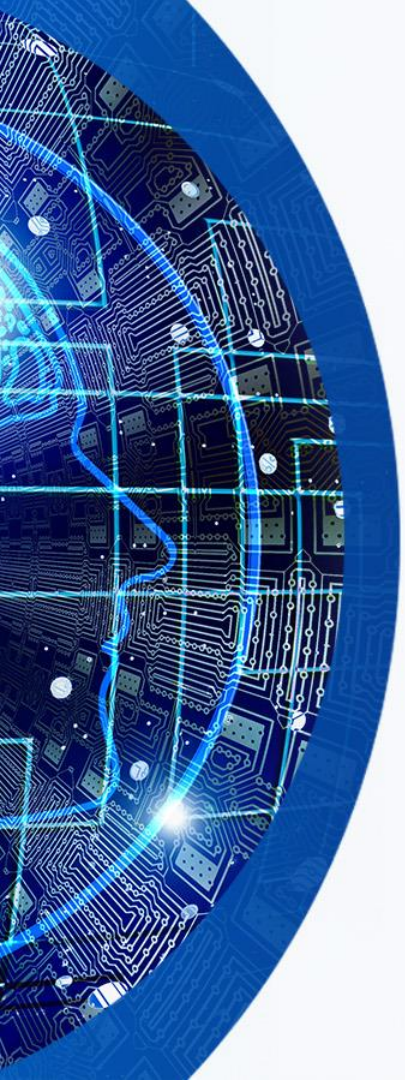
```
Clean_df <- na. Omit(df)
```

- يستخدم هذا الرمز الدالة `na. Omit()` لإزالة أي صفوف من إطار البيانات تحتوي على قيم مفقودة. سيحتوي إطار البيانات الناتج فقط على صفوف تحتوي على بيانات كاملة.



البيانات الشاذة او المتطرفة Outliers

في التعلم الآلي، لا تقل جودة البيانات أهمية عن جودة النموذج أو التصنيف التنبئي. ومع ذلك،في بعض الأحيان في مجموعة البيانات، توجد بيانات مسجلة تختلف اختلافا كبيرا عن الحالات الأخرى، وتميز نفسها في ميزة واحدة أو أكثر. هذه البيانات، المعروفة باسم البيانات المتطرفة، يمكن ان تؤدي إلى حدوث حالات شاذة في النتائج التي تم الحصول عليها من خلال الخوارزميات والأنظمة التحليلية في .النماذج الخاضعة للاشراف، يمكن أن تخدع البيانات المتطرفة عملية التدريب ، مما قد يؤدي إلى فترات تدريب أطول أو يؤدي إلى نماذج أقل دقة.



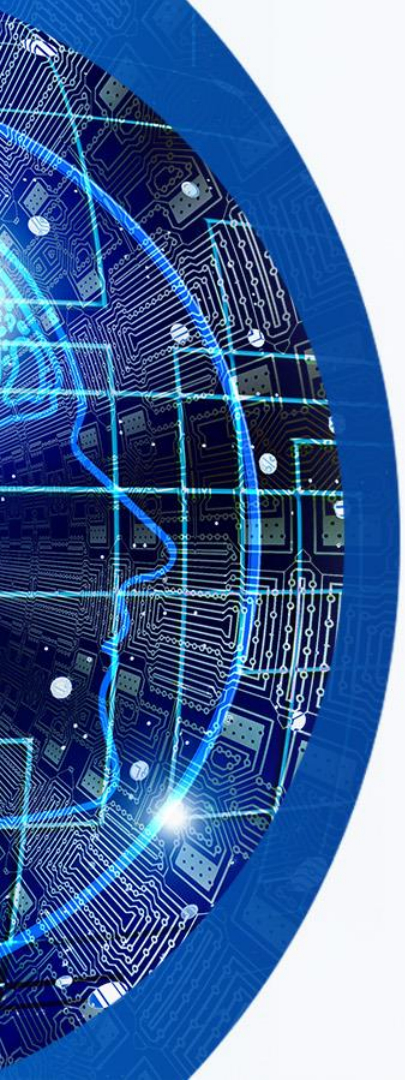
تأثير البيانات المتطرفة على التحليل

- البيانات المتطرفة لها تأثير كبير على نتيجة تحليل البيانات. فيما يلي بعض التأثيرات الأكثر شيوعا:

- قد يكون لها تأثير كبير على المتوسط والانحراف المعياري.

1. إذا لم يتم توزيع تشتت النقاط المتطرفة بشكل عشوائي ، فيمكنها تقليل الحالةالطبيعية[1] . normality

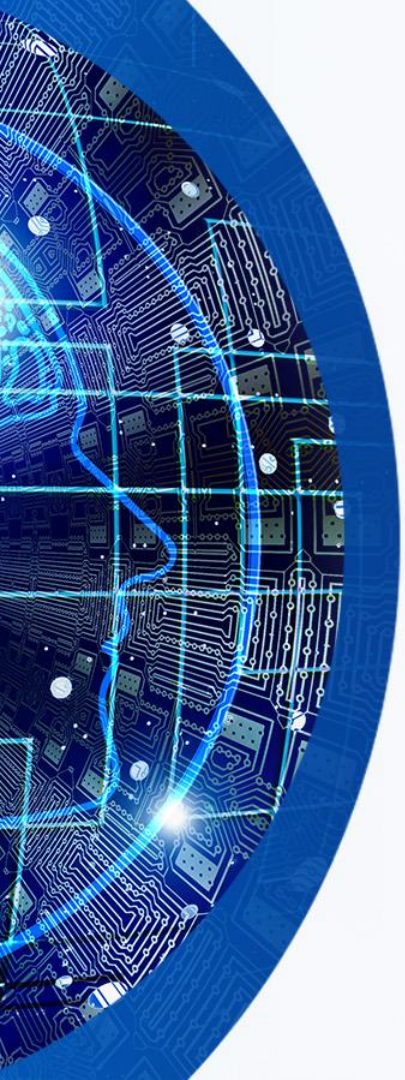
- يمكن أن تسبب التحيز(Bias)أو تؤثر على التقديرات.
- يمكن أن تؤثر على الافتراض الأساسي للتوقع(الانحدار) والنماذج الإحصائية الأخر ى .



كشف النقاط المتطرفة

•هناك عدة طرق للعثور على المواقع المتطرفة. تستخدم كل هذه الطرق أسلوبا للعثور على قيم غير معتادة مقارنة بمجموعات البيانات الأخرى. هنا قمنا بإدراج عدد قليل من هذه التقنيات:

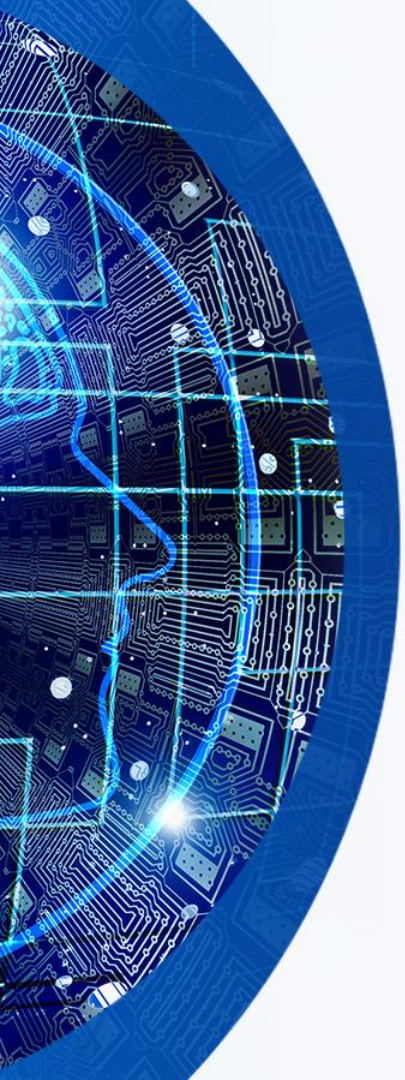
- الترتيب: الترتيب هو أبسط تقنية لتحليل البيانات المتطرفة. قم بتحميل مجموعة البيانات الخاصة بك إلى أي نوع من أدوات معالجة البيانات ، مثل جدول بيانات ، وقم بفرز القيم حسب الحجم. بعد ذلك ، تحقق من نطاق قيم نقاط البيانات المختلفة. إذا كانت كل نقطة بيانات أعلى أو أقل بشكل ملحوظ من النقاط الأخرى في مجموعة البيانات ، فيمكن اعتبارها عنصرا بعيدا. طريقة فرز البيانات على مجموعة بيانات صغيرة فعالة للغاية.
- باستخدام الرسوم البيانية .طريقة أخرى لتحليل البيانات البعيدة هي الرسوم البيانية. ارسمجميع نقاط البيانات على الرسم البياني واعرف النقاط الأبعد عن النقاط الأخرى. باستخدام طريقة الرسم التخطيطي مقارنة بطريقة الترتيب، يمكننا تصور المزيد من نقاط البيانات التيتسهل رؤية النقاط المتطرفة. يمكننا تحديد النقاط المتطرفة باستخدام المخططات الصندوقية[2] boxplotو مخططات المدرج ا لتكراري[3] histogram ومخططات التشتت[4]. scatter plot



تحويل البيانات Data transformation

غالباً ما يتم تخزين البيانات كسلاسل، حتى لو كانت تمثل قيمة رقمية أو تاريخاً. لاستخدام هذه البيانات للتحليل، قد نحتاج إلى تحويلها إلى نوع البيانات المناسب. على سبيل المثال، إذا كان هيكل بيانات و تريد تحويله الى مصفوفة او العكس

يجب تحويل البيانات إلى نظام قابل للقراءة ومتوافق. و فيما يلي بعض العمليات الهامة التي يتم استخدامها لتحويل البيانات .



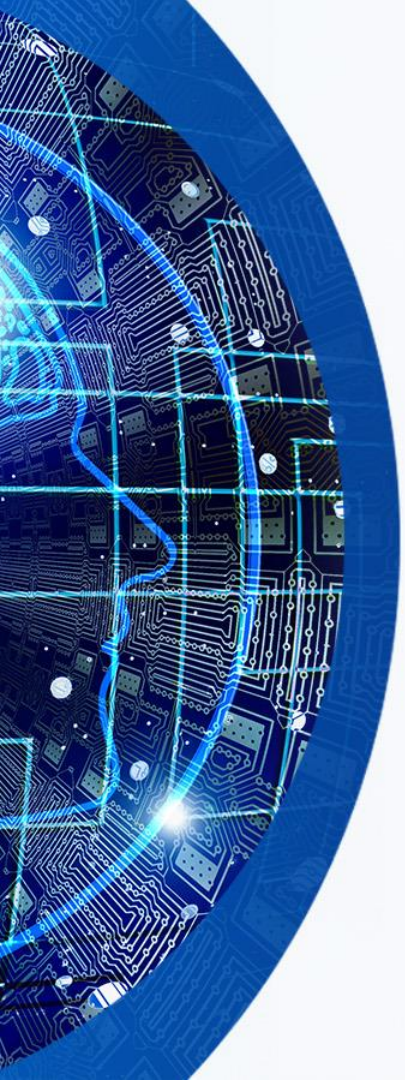
البيانات المكررة Duplicate Data

- قد تحتوي البيانات المدروسة على قيم مكررة، مما قد يؤدي إلى تحريف تحليلك. لتحديد التكرارات وإزالتها، يمكنك استخدام الدالة `duplicated()` والدالة `unique()`. على سبيل المثال، إذا كان لديك متجه سلاسل يحتوي على قيم مكررة، فيمكنك إزالتها بهذه الطريقة:

- Strings <- c("foo", "bar", "baz", "foo")

- Unique_strings <- unique(strings)

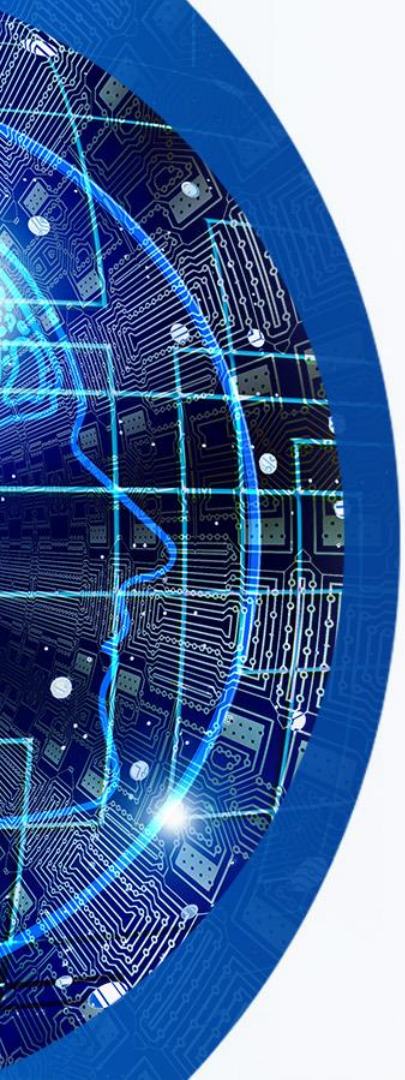
- يستخدم هذا الكود الدالة `unique()`` لإزالة أي قيم مكررة . و النتيجة تحتوي على قيم فريدة فقط.



التجميع Aggregation

•تجميع البيانات هو طريقة يتم فيها جمع البيانات الأولية واستخدامها بإيجاز للتحليل. على سبيل المثال، يمكن جمع البيانات الأولية خلال فترة زمنية لتقديم إحصائيات مثل المتوسط، والحدالأدنى، والحد الأقصى، والمجموع. بعد تجميع البيانات وكتابتها كتقرير، يمكنك تحليل البيانات المجمعة لاكتساب رؤى حول مصادر محددة. بمعنى آخر، يمكن أن يمكّن تجميع البيانات المحللين من الوصول إلى كميات كبيرة من البيانات وفحصها في فترة زمنية معقولة. يمكن أن يمثل صف من البيانات المجمعة المئات أو الآلاف أو حتى أكثر من البيانات الدقيقة. تتضمن أمثلة ال بيانات المجمعة ما يلي:

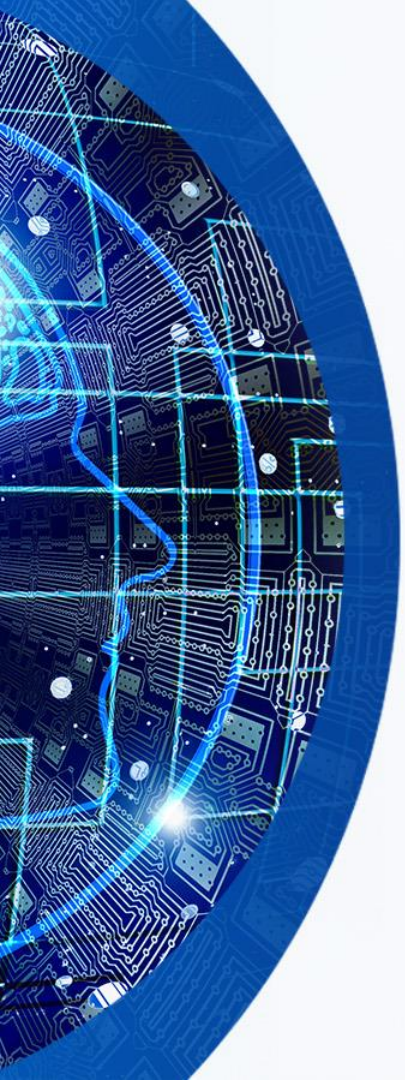
- متوسط عمر العميل حسب المنتج. لم يتم تحديد كل عميل بشكل منفصل ، ولكن بالنسبة لكل منتج ، يتم تخزين متوسط عمر العميل .
- عدد العملاء حسب الدولة. بدلا من مراجعة كل عميل ، يتم توفير عدد من العملاء من كل بلد.



التقطيع Discretization [1]

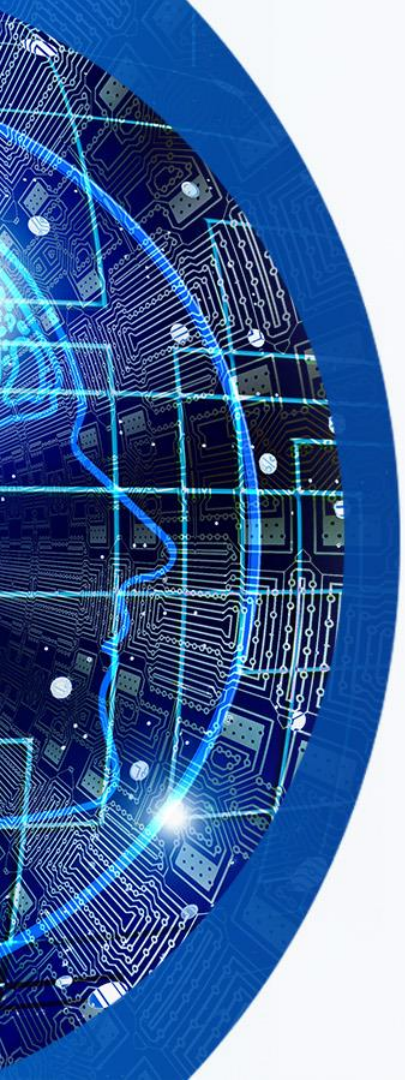
غالبا ما نواجه البيانات التي يتم جمعها من العمليات المستمرة مثل درجة الحرارة والضوءالمحيط وسعر سهم الشركة. لكن في بعض الأحيان نحتاج إلى تقسيم هذه القيم المستمرة إلى أجزاء أكثر قابلية للتحكم)لأن بعض خوارزميات التعلم الآلي، وخاصة خوارزميات التصنيف، تتطلب تقطيع البيانات لتكون سمات. مثال على ذلك عمر الافراد كما يلي:

1 ,5,9,4,7,11,41,71,31,81,91,13,33,63,24,44,64,07,47,87,77



يوضح الجدول أدناه هذه البيانات بعد التقطيع:

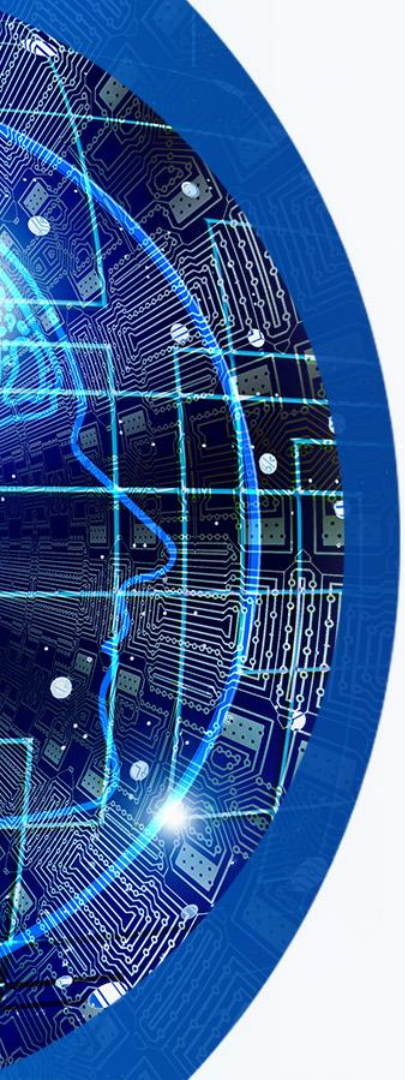
العمر	العمر	العمر	العمر	السمة
70,74,77,78	31,33,36,42,44,46	11,14,17,13,18,19	1,5,4,9,7	
المسنين	الكبار	الشباب	الاطفال	بعد التقطيع



التنعيم Smoothing [2]

•يتم تنفيذ تنعيم البيانات باستخدام خوارزميات متخصصة لإزالة الضوضاء من مجموعة البيانات.

•تسمح هذه العملية بإبراز أنماط البيانات المهمة. يمكن أن يساعد تنعيم البيانات في التنبؤ بالاتجاهات. على الرغم من أن تنعيم البيانات يمكن أن يساعد في كشف الأنماط في البيانات المعقدة، إلا أن تنعيم البيانات لا يوفر بالضرورة تفسيرا للموضوع أو الأنماط التي تساعد في التعرف عليه. في بعض الأحيان، قد يؤدي تنعيم البيانات إلى إزالة نقاط البيانات القابلة للاستخدام.إذا كانت مجموعة البيانات موسمية ولا تعكس بشكل كامل الواقع الناتج عن نقاط البيانات، فقديؤدي ذلك إلى تنبؤات غير صحيحة.



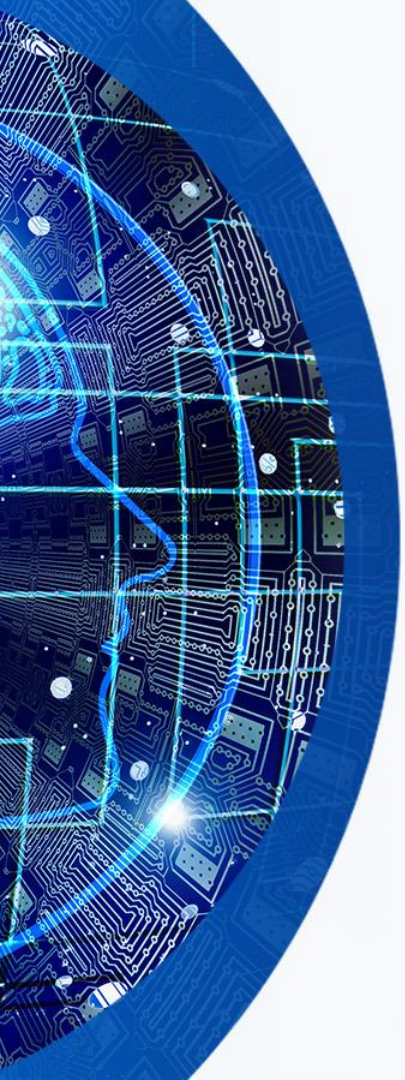
التحجيم Scaling

•في العديد من خوارزميات التعلم الآلي، لجلب جميع الميزات في موقف واحد، يتعين علينا التحجيم حتى لا نؤثر على عدد كبير من النماذج لمجرد حجمها الكبير. تعد ميزات التحجيم Feature scalingفي التعلم الآلي واحدة من أهم الخطوات في معالجة البيانات قبل إنشاء نموذج التعلم الآلي. يمكن أن يفرق التحجيم بين نموذج التعلم الآلي السيئ والنموذج الأفضل.

- أكثر تقنيات تحجيم السمات شيوعا هي التوحيد القياسي Standardization

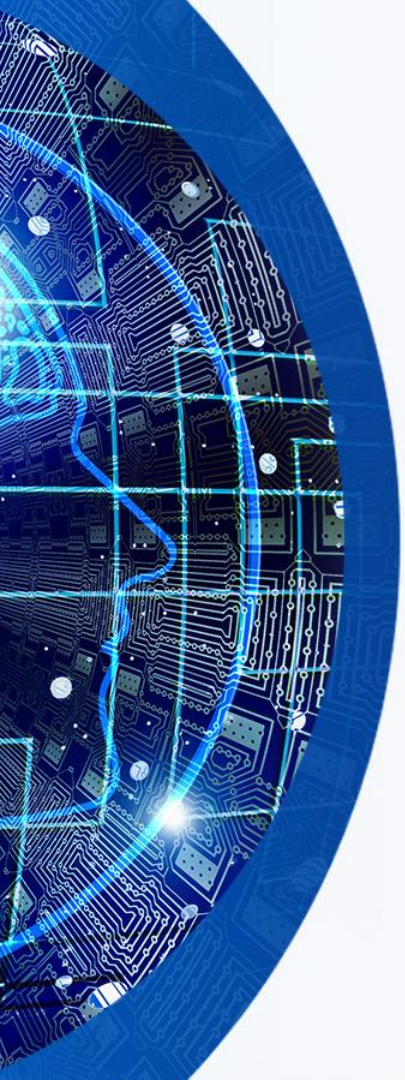
- normalization التسوية .

- يتم استخدام التوحيد عندما نريد تقييد قيمنا بين رقمين، عادة بين [0،1] أو[1،1-]. بينما يحول التسوية البيانات إلى متوسط صفر وتباين 1.



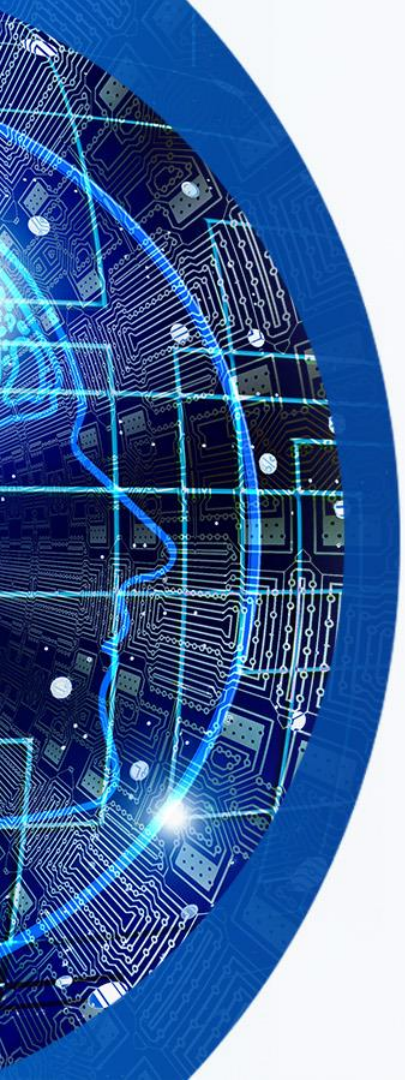
سبب تحجيم البيانات ؟

خوارزميات التعلم الآلي ترى الأرقام فقط. ومن ثم، إذا كان هناك اختلاف كبير في نطاق الأرقام، فإنهم يضعون الافتراض الأساسي بأن الأرقام في النطاق الأعلى لها مزايا اكبر. وهكذا، يبدأ هذا العددالكبير في لعب دور أكثر حسما أثناء تدريب النموذج. بالإضافة إلى ذلك، تعمل خوارزميات التعلم الآلي على الأرقام ولا تعرف ما يمثله الرقم. يبلغ وزنه 10 جرامات وسعره 10 دولارات، وهو يمثل شيئين مختلفين تماما، وهو أمر واضح للبشر، ولكن بالنسبة للنموذج،كلاهما يعتبر ميزة. لنفترض أن لدينا خاصيتين للوزن والسعر أن قيم الوزن لا يجب أن تكون أعدادا أكبر. ومن ثم، تفترض الخوارزمية أنه نظرا لأن الوزن أكبر من السعر، فإن الوزن أهم من السعر. لهذا السبب، تلعب هذه الأرقام الأكثر أهمية دورا أكثر حسما في تدريب النموذج. لذلك، فإن تحجيم السمات مطلوب لإحضار جميع السمات في موقف واحد دون أي أولية .



التوحيد القياسي Standardization [1]

التوحيد القياسي للبيانات هو أسلوب مهم يتم إجراؤه كخطوة معالجة مسبقة قبل العديد من نماذج التعلم الآلي لتوحيد نطاق ميزات مجموعة بيانات الإدخال. يحدث التوحيد القياسي عندماتختلف خصائص مجموعة بيانات الإدخال اختلافا كبيرا في نطاقها. بعبارات أبسط، عندما يتم قياس البيانات بوحدات قياس مختلفة)على سبيل المثال ،كيلوغرامات، أمتار،كيلومترات، إلخ.(تسبب هذه الاختلافات في نطاق الميزات الأساسية مشاكل في العديد من نماذج التعلم الآلي. على سبيل المثال، بالنسبة للنماذج التي تستند إلى حساب المسافة، إذا كان لإحدى الخصائص نطاق واسع من القيم، يتم ضبط المسافة بواسطة خاصية معينة. لنفترض أن لدينا مجموعة بيانات ثنائيةالأبعاد بخاصيتين للطول بالأمتار والوزن بالكيلوجرام، والتي تتراوح بين [1إلى 2] متر و[30إلى 90]كجم، على التوالي. بغض النظر عن النموذج القائم على المسافة الذي تستخدمه بناءعلى مجموعة البيانات هذه، فإن خاصية الوزن ستسود على خاصية الارتفاع وستكون لها حصة أكبر في حساب المسافة؛ فقط لأنه يحتوي على قيم أعلى مقارنة بالارتفاع .لذلك، لتجنب هذه المشكلة وحلها، من الضروري تحويل الميزات إلى مقاييس مماثلة باستخدام توحيد البيانات.



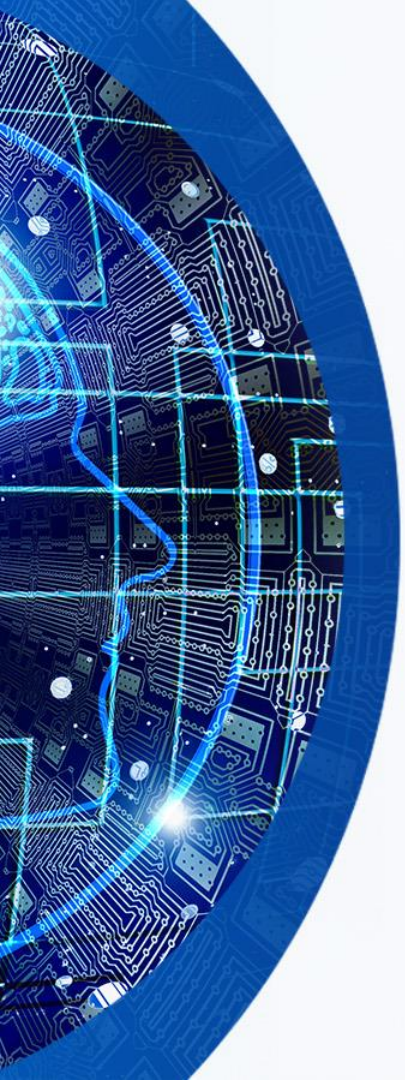
كيفية توحيد البيانات ؟

•تعتبر الدرجة Z ، والتي تسمى أيضا الدرجة القياسية standard score ، واحدة من أكثر الطرق شيوعا لتوحيدالبيانات، والتي يمكن إجراؤها عن طريق طرح المتوسط وتقسيمه على الانحراف المعياري لكلقيمة لكل خاصية. معادلتها الرياضية على النحو التالي :

$$Z=(x-\mu)/\sigma$$

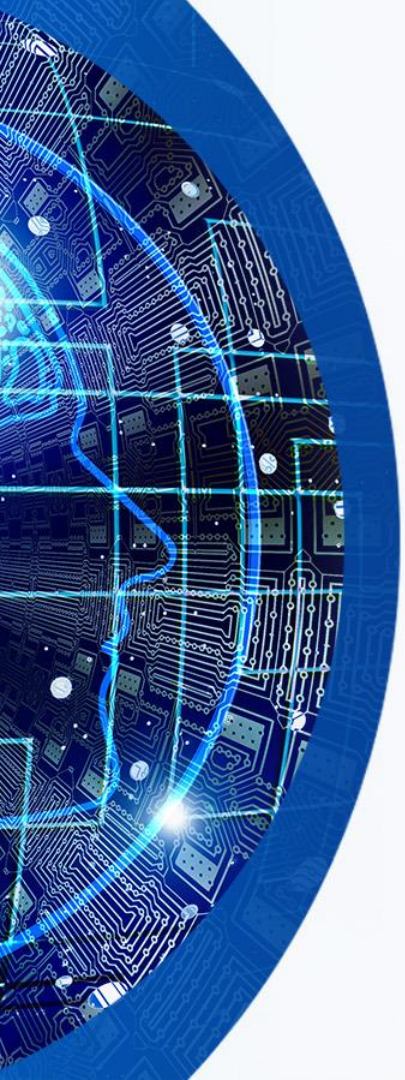
•في هذه المعادلة، x هي الدرجة الأولية ، و μ هي متوسط العينة و σ هي الانحراف المعياري للعينة.

- بمجرد اكتمال توحيد البيانات، سيكون لجميع السمات متوسط صفر، وانحراف معياري بواحد ،وبالتالي، نفس المقياس



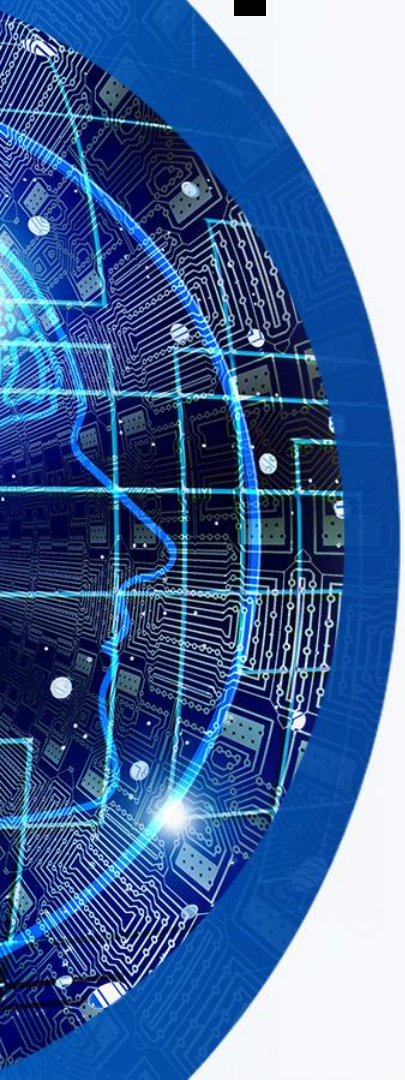
•قبل استخدام أي من نماذج وطرق التعلم الآلي، من الأفضل معرفة متى ولماذايجب علينا استخدام التوحيد القياسي للبيانات:

1. PCA: في تحليل المكونات الرئيسية PCA Principal Component Analysis ، تكتسب الميزات ذات التباين العالي / النطاق الواسع وزنا أكبر من تلك ذات التباين المنخفض ، ونتيجة لذلك ، فإنها تهيمن بشكل غير معقول على المكونات الرئيسية الأولى. يمكن أن يمنع التوحيد هذا من خلال توفير نفس الوزن لجميع الميزات.
2. قبل التجميع: نماذج التجميعClustering هي خوارزميات تعتمد على المسافة تستخدم معيارالمسافة لقياس التشابه بين الملاحظات. لذلك ، سيكون للميزات عالية النطاق تأثير أكبر على التجميع. ومن ثم ، فإن التوحيد مطلوب قبل إنشاء نموذج التجميع.
3. KNN: K-أقرب الجيرا ن k-nearest neighbors هي خوارزمية تصنيف تعتمد على المسافة تصنف الملاحظات الجديدة بناء على أوجه التشابه على سبيل المثال، معايير المسافة مع الملاحظات الموسومة من مجموعة التدريب. يسمح التوحيد لجميع المتغيرات بالمشاركة بالتساوي في قياس التشابه .
4. SVM: تحاول خوارزمية آلة المتجهات الداعمة Support Vector Machineتعظيم المسافة بين لوحة الفاصل ومتجهات الدعمة. إذا كانت الخاصية تحتوي على قيم كبيرة جدا ، فإنها تهيمن على الخصائص الأخرى عند حساب المسافة. لذلك، يعطي التوحيد جميع الميزات نفس التأثير على معيار المسافة.



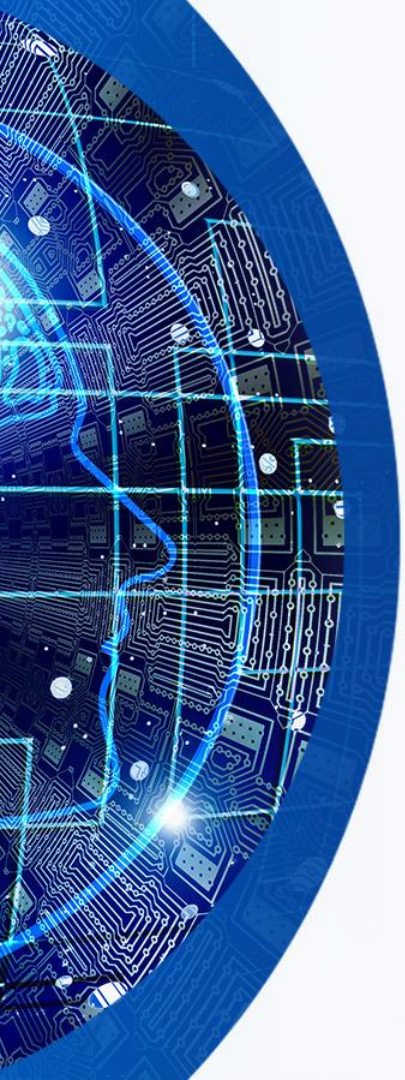
التسوية Rescaling

هو جزء من تقنيات المعالجة المسبقة وتنقية البيانات، وبشكل أكثر عمومية، نوع من تحجيم الميزات. الغرض الرئيسي من هذه التقنية هو جعل البيانات متسقة عبر جميع السجلات والحقول (دون تغيير نطاق القيم) يساعد هذا في إجراء اتصالات بين بيانات الإدخال، مما يساعد بدوره في تنظيف البيانات وتحسين جودتها. يتم استخدام هذا النوع من التحجيم عندما يكون للبيانات نطاق متنوع (للخصائص نطاقات مختلفة) ولا تفترض الخوارزميات التي يتم تدريبها عليها مسبقا توزيع البيانات (مثل الشبكات العصبية.)



تعطي التسوية وزنا / أهمية متساوية لكل متغير بحيث لا يشوه متغير واحد أداء النموذج فياتجاه واحد ؛ فقط لأنهم أكثر عددا. إن أسلوب التحجيم الأكثر شيوعا والأكثر استخداما هو التحجيم ، والمعروف أيضا باسم تسوية الحد الاقل-الحد الاكثر، والذي يتم حسابه على النحو التالي:

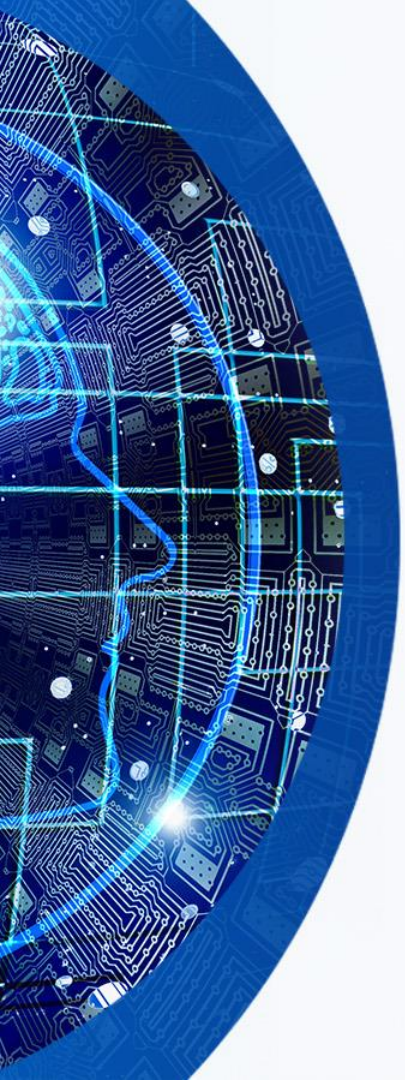
$$\acute{x} = \; x - min(x) \;\; / \;\; max(x) - min(x)$$

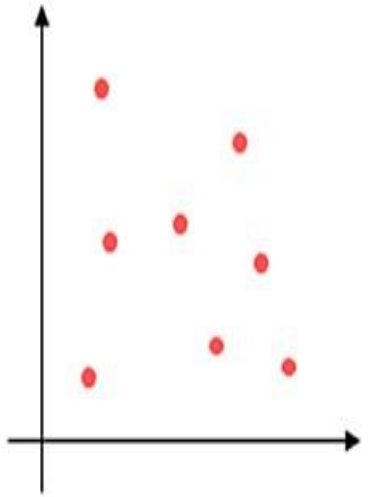


التوحيد او التسوية؟

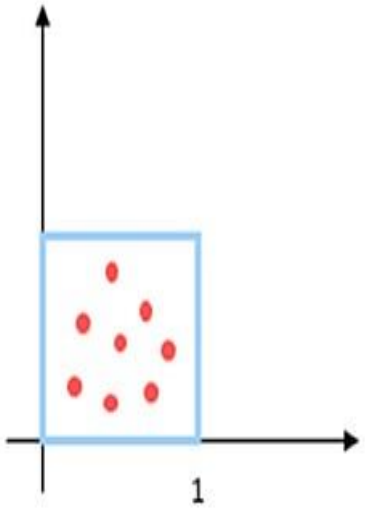
تعد التسوية مفيدة عندما تعلم أن توزيع البيانات الخاص بك لا يتبع التوزيع الغاوسي)منحنى الجرس(. يمكن أن يكون هذا مفيدا في الخوارزميات التي لا تفترض أي توزيع للبيانات، مثلKNN أو الشبكات العصبية. من ناحية أخرى ،يمكن أن يكون التوحيد مفيدا في الحالات التي تتبع فيها البيانات توزيعا غاوسيا (يفترض التوحيد أن بياناتك لها توزيع غاوسي). ومع ذلك، هذا ليس صحيحا بالضرورة، ولكن إذا كان توزيع الميزات الخاص بك هو غاوسي، فإن هذه التقنية تكون أكثر فعالية. أيضا، على عكس التوحيد، ليس للتسوية حدود. لذلك، حتى إذا كان لديك الكثير من البيانات المتطرفة في بياناتك، فلن تتأثر بالتسوية. ومع ذلك، يعتمد اختيار استخدام التوحيد أو التسوية على مشكلتك وخوارزمية التعلم الآلي الخاصة بك. لا توجد قاعدة صارمة وسريعة لإخبارك بموعد توحيد بياناتك. يمكنك دائما ملاءمة[1] نموذجك مع البيانات الخام والموحدة والمسواة ومقارنة الأداء للحصول على أفضل النتائج.

بينما يضع التوحيد القيم الأساسية ضمن نطاق معين ، تضعها التسوية في توزيع متوسطه صفر وانحرافه المعياري واحد

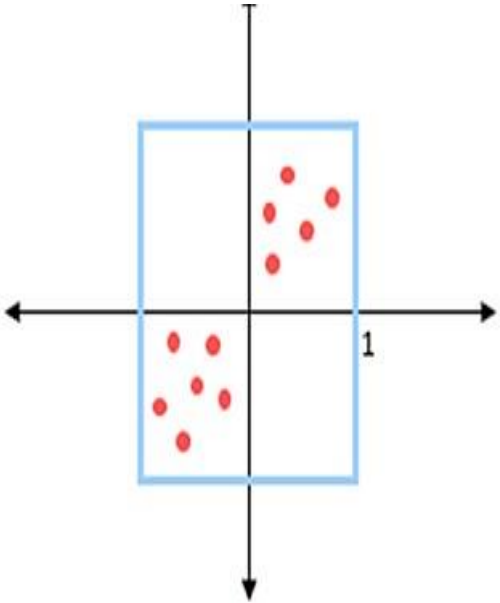




البيانات الحقيقية



بعد التسوية



بعد التوحيد

شكرا على المتابعة