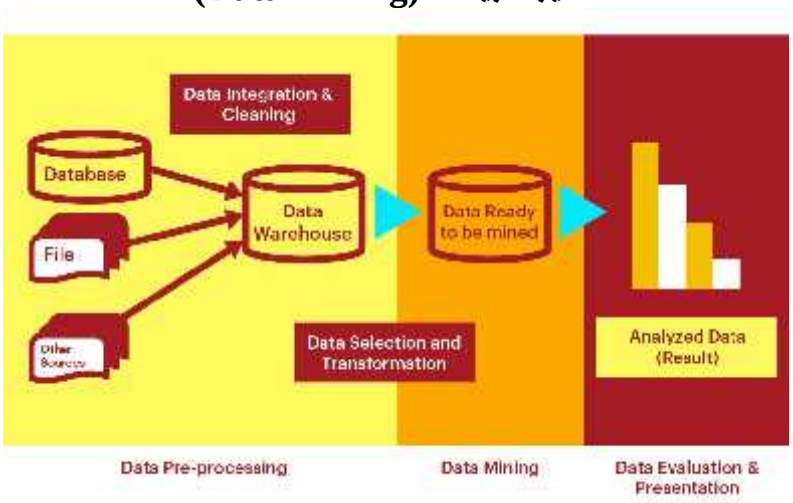
تنقيب البيانات (Data Mining)



التنقيب عن البيانات (تعدين البيانات) هو عمليَّة البحث عن مجموعة كبيرة من البيانات الخامّ وتحليلها من أجل تحديد الأنماط واستخراج المعلومات المفيدة التي تُساعد الشركات في اتِّخاذ القرار وحل المُشكلات والتنبؤ بالاتجاهات وتخفيف المخاطر وإيجاد فرص جديدة.
يتضمَّن التنقيب في البيانات أيضًا إنشاء العلاقات وإيجاد الأنماط والارتباطات بين البيانات المُختلفة لمُعالجة المشكلات، والوصول إلى معلومات قابلة للتنفيذ في هذه العمليَّة.

- **كيفية عمل التنقيب في البيانات:**

تنقيب البيانات يتضمن خطوات ممنهجة ومتسلسلة ودقيقة وهي:
اولا: معالجة البيانات وهذة المرحلة تهتم جمع البيانات من عدة قواعد بيانات وفحصها للتأكد من خلوها من الاخطاء أو النقص أو التعارض ومن ثم إعادة معالجتها وتشفيرها وتجميعها
ثانيا : تخزين البيانات في مستودع للبيانات
ثالثا: أخذ عينة من البيانات
رابعا :اختيار نوع التنقيب وصفي أو تنبئي واختيار الخوارزمية المناسبة لعمل التنقيب
خامسا :تنفيذ التنقيب لاستخراج المعارف والانماط
سادسا: تقييم المعارف المستخرجة وتحديد اين منها يعتبر مفيدا ومن ثم الاستفادة من هذه المعارف

- **المراحل الست التي تمر بها عملية التنقيب في البيانات:**

بفضل المراحل المرنة الموجودة في العملية القياسية عبر الصناعة للتنقيب في البيانات(CRISP-DM) ، يمكن لفرق البيانات التنقل ذهابًا وإيابًا بين المراحل حسب الحاجة. كما يمكن أيضًا للتقنيات البرمجية تنفيذ بعض هذه المهام أو دعمها.

1. فهم الأعمال:
يبدأ عالم البيانات أو المُنقب في البيانات بتحديد أهداف المشروع ونطاقه. ويتعاون مع أصحاب المصلحة التجاريين لتحديد معلومات معينة.

- المشكلات التي تحتاج إلى حلول

- قيود المشروع أو حدوده
- تأثير الحلول المحتملة على الأعمال

ثم يستخدم هذه المعلومات في تحديد الغرض من التنقيب في البيانات وتحديد الموارد المطلوبة لاكتساب المعرفة. وبالتالي فهم المشاكل والمسائل التي تواجهها الأعمال. وبمعنى أخر، كيف يمكن تحقيق المنفعة الأعظم من التنقيب في البيانات، مما يتطلب وجود صيغة واضحة ومحددة لأهداف الأعمال.

في هذه المرحلة، علينا تحديد:

- من أين جاءت البيانات؟
- مُن الذي جمعها وهل كان جمعها يتبع الوسائل القياسية؟
- ماذا تعني الأعمدة والصفوف المختلفة للبيانات؟
- هل هناك أي اختصارات غير معروفة أو غير واضحة؟
- وصف البيانات والتحقق من حجمها وفحص خصائصها الإجمالية.
- إمكانية الوصول وتوافر السمات. أنواع السمات، والارتباطات، والهوايات.
- فهم معنى كل سمة وقيمتها في مصطلحات الأعمال.

2. فهم البيانات:

بمجرد الوقوف على المشكلة الموجودة في العمل، يبدأ علماء البيانات في إجراء تحليل أولي للبيانات. ثم يقومون بجمع مجموعات البيانات من مصادر مختلفة، ويحصلون على حقوق إمكانية الوصول، ويقومون بإعداد تقرير حول وصف البيانات. يضم التقرير أنواع البيانات ومقدارها ومتطلبات الأجهزة والبرامج لمعالجة البيانات. وبمجرد موافقة الشركة على خطتهم، يبدؤون في استكشاف البيانات والتحقق منها. ويعالجون البيانات باستخدام الأساليب الإحصائية الأساسية، ويقيّمون جودة البيانات، ويختارون مجموعة البيانات النهائية للمرحلة التالية. حيث أن معرفة البيانات بصورة جيدة تعني مساعدة المصممين على استخدام الخوارزميات أو الأدوات المستخدمة للمسائل المحددة بدقة عالية. وهذا يقود إلى تعظيم فرص النجاح بالإضافة إلى رفع الفعالية والكفاءة لنظام اكتشاف المعرفة.

تجميع البيانات: و هي الخطوة الموجهة نحو تحديد مصدر البيانات في الدراسة بما في ذلك استخدام البيانات العامة الخارجية مثل الضرائب و غيرها.

توصيف البيانات: و هي الخطوة التي تركز على توصيف محتويات الملف الواحد من الملفات أو الجداول.

جودة البيانات و تحقيقها: هذه الخطوة تحدد ما إذا كان تقليل أو إهمال بعض البيانات غير الضرورية أو كونها رديئة الجودة و قد لا تنفع في الدراسة. لأن النموذج الجيد يحتاج إلى بيانات جيدة مما يتوجب أن تكون البيانات صحيحة و ذات مضمون دقيق.

التحليل الاسترشادي للبيانات: تستخدم الأساليب مثل الإظهار المرئي أو التصور أو عملية التحليل المباشر(OLAP) التي تؤدي إلى إجراء التحليل الأولي للبيانات. و تعتبر هذه الخطوة مهمة و ضرورية.

3. تجهيز البيانات:

يقضي المنقبين في البيانات معظم الوقت في هذه المرحلة، لأن برامج التنقيب في البيانات تتطلب بيانات عالية الجودة. وتقوم عمليات الأعمال بجمع البيانات وتخزينها لأسباب أخرى غير التنقيب، ويتوجب على المنقبين تنقيح هذه البيانات قبل استخدامها في النمذجة. ويدخل ضمن تجهيز البيانات العمليات التالية:

تنظيف البيانات: معالجة البيانات المفقودة ومعالجة أخطاء البيانات والقيم الافتراضية وتصحيحات البيانات.

الاختيار: و تعني اختيار المتغيرات المتوقعة و حجم العينة.

صياغة المتغيرات وتحويلها: حيث يجب دائما أن تصاغ المتغيرات الجديدة لبناء النماذج الفعالة.

تكامل البيانات ودمجها: حيث أن مجاميع البيانات في دراسة التنقيب عن البيانات من الممكن خزنها في قواعد بيانات متعددة الأغراض التي تكون بحاجة إلى توحيدها في قاعدة بيانية واحدة. كدمج مجموعتين مختلفتين من البيانات للحصول على مجموعة البيانات النهائية المستهدفة.

تنسيق البيانات: هنا يتم إعادة ترتيب حقول البيانات كما يتطلب في نموذج التنقيب في البيانات وتحويل أنواع البيانات أو تكوين البيانات لتقنية التنقيب المستخدمة.

4. نمذجة البيانات:

إن بناء و صياغة نموذج الحل السليم و الدقيق يتم من خلال عملية الخطأ و الصواب، حيث كثيرا ما تحتاج مثل هذه العملية إلى مساعدة المختصين في التنقيب عن البيانات بهدف اختبار و فحص مختلف البدائل للحصول على

أفضل نموذج لحل المشكلة قيد الدراسة. يتولى المنقبون في البيانات عملية إدخال البيانات التي تم تجهيزها في برامج التنقيب في البيانات ودراسة نتائجها. وللقيام بذلك، يمكنهم الاختيار من بين عدة تقنيات وأدوات خاصة بالتنقيب في البيانات. ويجب عليهم أيضًا إجراء اختبارات لتقييم جودة نتائج التنقيب في البيانات. ولأجل نمذجة البيانات، يمكن لعلماء البيانات القيام بما يلي:

- تدريب نماذج تعلّم الآلة (ML) باستخدام مجموعات بيانات أصغر ذات نتائج معروفة
- استخدام النموذج لإجراء مزيد من التحليل لمجموعات البيانات غير المعروفة
- ضبط برامج التنقيب في البيانات وإعادة تكوينها حتى تصبح النتائج مرضية

5. التقييم.

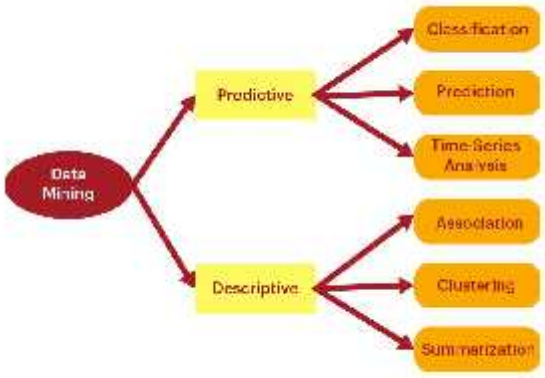
6. النشر.

➢ **فوائد التنقيب عن البيانات:**

1- تحسين عملية صُنع القرار
2- تحسين العلاقة مع العملاء
3- إدارة المخاطر
4- تقليل التكاليف
5- ميزة تنافسية للشركات

➢ **أدوات التنقيب عن البيانات:**

نماذج التنقيب في البيانات نوعان: النماذج التنبؤية (Predictives) والنماذج الوصفية.(Descriptives)
النماذج التنبؤية تهدف إلى التنبؤ بقيمة بعض الخصائص. مثل التنبؤ باحتمال الشراء للزبون.
أما **النماذج الوصفية** فتنقسم إلى: نماذج العنقدة التي تسمح بتجميع الأفراد، والأحداث، آو المنتجات في عناقيد، ونماذج الارتباط التي تسمح بتحديد العلاقات بينهم.



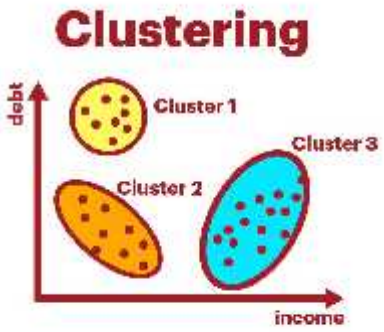
وهناك عدة أدوات للتنقيب في البيانات، من أهمها الأدوات الآتية:

1- **التلخيص:(Summarization)** يشير التلخيص إلى أساليب تفتيت كتل البيانات الكبيرة إلى مقاييس موجزة، توفر وصفا عاما للمتغيرات وعلاقاتها. ومن الأمثلة على أساليب التلخيص نذكر: المتوسطات، والمجاميع، والإحصائيات الوصفية التي تتضمن مقاييس النزعة المركزية مثل المتوسط الحسابي و الوسيط والمنوال، ومقاييس التشتت مثل الانحراف المعياري. وعلى الرغم من أن مقاييس التلخيص تعطي صورة كبيرة عن بعض التفاصيل ذات العلاقة فإنها غالبا ما تهمل تفاصيل أخرى ذات أهمية كبيرة تتعلق بسلوك المستهلك خصوصا.
2- **التصنيف: (Classification)** يتمثل التصنيف في تفسير أو التنبؤ بخاصية فرد ما من خلال خصائص أخرى. هذه الخاصية هي عموما كيفية. ويمكن انجاز التصنيف بالاعتماد على الأساليب الإحصائية القديمة مثل الانحدار والتحليل التمييزي، أو بالاعتماد على أساليب حديثة نسبيا مثل قوى الارتباط والاستنتاج المستند إلى الحالة والشبكات العصبية. وكأمثلة عن طرق التصنيف المستعملة كجزء من تطبيقات استكشاف المعرفة التي تتضمن تصنيف اتجاهات الأسواق المالية، والتحديد الآلي للأشياء المهمة في صورة كبيرة من قواعد البيانات.

3- **التنبؤ: (Prediction)** يشبه التنبؤ التصنيف أو التقدير، ما عدا أن البيانات تصنف على أساس التنبؤ بسلوكها المستقبلي أو تقدير قيمتها المستقبلية. حيث أن المتغير التابع المتنبأ به هو متغير كمي. ومن الأدوات التقليدية المستخدمة في التنبؤ نذكر على سبيل المثال: الانحدارات بأنواعها و التحليل التمييزي. أما الأساليب الجديدة فتشتمل على قواعد الارتباط و شجرة القرار و الشبكات العصبية و الخوارزميات الوراثية.

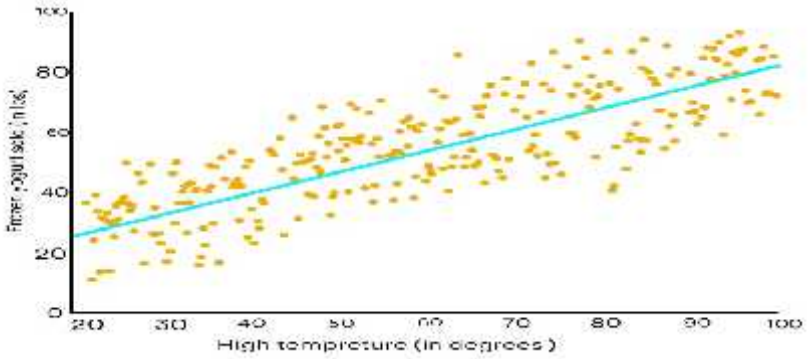
4- **العقدة أو التجزئة: (Clustering)** يتمثل التجميع العنقودي أو التجزئة إلى قطاعات في البحث عن مجموعات متجانسة في مجتمع من الأفراد. و يشير التجميع العنقودي أو التجزئة إلى قطاعات إلى عملية تشكيل مجموعات أو قطاعات مؤلفة من أفراد أو أصحاب أسر، و ذلك بالاستناد إلى معلومات متضمنة في مجاميع من المتغيرات التي تصفهم. و الغرض من التجميع العنقودي المساعدة على تطوير برامج تسويقية مصممة على مقاسات الزبائن أنفسهم، و التي بالإمكان استخدامها لاستهداف أعضاء لكل قطاع من هذه القطاعات على أمل ترغيبهم في تكرار الشراء أو التحول إلى زبائن موالين.

وتتم أساليب التجميع العنقودي غالبا بمساعدة أساليب التحليل العنقودي الإحصائية و الأساليب المستندة إلى شجرة القرار، و الشبكات العصبية و الخوارزميات الوراثية.



فيما تستخدم؟ هناك عدة طرق لاستخلاص المعرفة من التحليلات العنقودية. مثلا في التسويق لأغراض مختلفة. يستخدم تقسيم المستهلكين في التحليل العنقودي على أساس الفوائد المطلوبة من شراء المنتج.

5-**تحليل الارتباط: (Rule Analysis)** يتمثل الارتباط في البحث عن علاقات أو ارتباطات موجودة بين عدة خصائص. و يشير تحليل الارتباط إلى مجموعة من الأساليب التي تستخدم لربط أنماط الشراء عبر القطاعات المتقاطعة أو عبر الوقت. فمثلا يقوم أسلوب تحليل سلة السوق (نوع من أنواع الارتباط) باستخدام المعلومات الكامنة في السلع التي اشتراها المستهلكون فعليا للتنبؤ بالسلع المحتمل شراؤهم إياها إذا ما تم تقديم عروض خاصة لهم أو إذا تم تعريفهم بهذه السلع.



يسمى المثال أعلاه بتحليل الانحدار الخطي، والتي تعني في الأساس أنه يمكن رسم خط مستقيم لإظهار كيفية ارتباط كل متغير ببعضه البعض. في هذه الحالة، نرى أنه كلما زاد إجمالي اللبن الزبادي المجمد، ارتفعت درجة الحرارة والعكس صحيح.

إذا كان النشاط التجاري يهدف إلى إجراء تنبؤ استنادًا إلى تأثير أحد المتغيرات على الآخرين، فقد يشير إلى ما يسمى تحليل الانحدار التي تندرج أسفل تقنية التنقيب عن البيانات. يتم استخدامها عبر العديد من الصناعات لتخطيط الأعمال والتسويق، والتنبؤ المالي، والنمذجة البيئية وتحليل الاتجاهات.

الانحدار هو تقنية التنقيب عن البيانات المستخدمة للتنبؤ بمجموعة من القيم العددية(القيم المستمرة)، مع إعطاء مجموعة بيانات معينة.
على سبيل المثال، يمكن استخدام الانحدار للتنبؤ بتكلفة منتج أو خدمة ، مع الأخذ في الاعتبار المتغيرات الأخرى.

6- الكشف عن التغيرات أو الانحرافات: (Change and deviation detection)

يرتكز على استكشاف التغيرات المهمة جدا في البيانات من خلال قياسات سابقة أو قيم معيارية.

مجالات تطبيق تنقيب قواعد البيانات في منظمات الأعمال:

هناك عدة ميادين لتطبيق التنقيب في البيانات في منظمات الأعمال، من أبرزها الميادين الآتية :

1- **التسويق :**استخدمت الشبكات العصبية الاصطناعية في دراسات التسويق المستهدف بما في ذلك الحصص السوقية. و قد ساعدت هذه الأساليب التسويق على استخدام نهج تحصيص الزبائن وفقا إلى الحقائق الديمغرافية (السكانية) الأساسية مثل الجنس والعمر والمجموعات وكذلك أنماطهم الشرائية.
2- **التجزئة :**لقد استخدمت أساليب التنقيب في البيانات بصورة فعالة في التنبؤ بالمبيعات حيث أخذت العديد من المتغيرات في الدراسات مثل متغيرات السوق المتعددة، قدرات الزبائن المستندة على العادات المتبعة في الشراء. كما ساعدت أساليب مثل تحليل السلة الشرائية أو السلة السوقية كثيرا على إيجاد أي من المنتجات التي يمكن أن تشترى سوية من قبل الزبائن.
3- **البنوك** :لقد أثبتت تنبؤات الأعمال والمالية على أنها الأساليب الممتازة في تطبيقات أساليب التنقيب في البيانات. وقد استخدمت هذه الأساليب في إيجاد الأسعار المضمونة وتنبؤات السعر المستقبلية وأداء الأسهم. كما وقد حققت استخدامات مثل هذه الأساليب النجاح في تطوير أنظمة القياس الرقمية في تحديد مخاطرة القروض والاحتيالات المالية.
4- **إدارة العمليات** :حيث استخدمت الشبكات العصبية في عمليات التخطيط والجدولة وإدارة المشاريع بالإضافة إلى إدارة الجودة.

وهناك من يركز على التعامل التجاري للمؤسسة كما يلي:

1- المبيعات: يُساعد التنقيب عن البيانات في مجال المبيعات على استخدام رأس المال لدفع نمو الإيرادات بشكل أكثر ذكاءً وكفاءةً. بالإضافة إلى ذلك، يُعد استخراج البيانات أداةً قوية للشركات في صناعة المبيعات، حيث يمكن أن يساعد في تحديد الاتجاهات والأنماط والرؤى التي يُمكن استخدامها لتحسين إستراتيجيات المبيعات وزيادة الإيرادات. من خلال تقسيم العملاء بناءً على عوامل مُختلفة، وتحليل بيانات المبيعات السابقة للتنبؤ بالمبيعات المستقبلية، وتحديد المُنتجات التي يتمّ شراؤها بشكلٍ مُتكرر معًا، يُمكن للشركات استخدام علم تعدين البيانات لتحسين أداء مبيعاتها، وزيادة رضا العملاء، والحفاظ على المُنافسة في السوق.

2- التصنيع: بالنِّسبة للشركات التي تُنتج سلعتها الخاصة، يلعب استخراج البيانات دورًا أساسيًا في تحليل مقدار تكاليف كل مادة خام، وما هي المواد المُستخدمة بكفاءة أكبر، وكيف يتم قضاء الوقت على طول عملية التصنيع، وما هي الاختناقات التي تؤثّر سلبًا على العملية. بالإضافة إلى ذلك، يُساعد التنقيب في البيانات على ضمان عدم انقطاع تدفق البضائع.

3- **الموارد البشرية:** غالبًا ما يكون لدى إدارات الموارد البشرية مجموعة واسعة من البيانات المُتاحة للمُعالجة بما في ذلك بيانات عن الترقيات، ونطاقات الرواتب، ومزايا الشركة، واستطلاعات رضا الموظفين. يُمكن أن يربط علم التنقيب عن البيانات بين هذه البيانات للحصول على فهم أفضل لسبب مغادرة المُوظفين في الشركة وما الذي يُمكن فعله للحفاظ على المُوظفين المُعيَّنين حديثًا.