علم البيانات

فريق العمل:

وعد المهاوش
نورة الطريف
بسمة الحسين
أسماء عواد

مها القحطاني
وضحى الخالدي
اثير عبدالمنعم
نادين العم،دي

بقيادة: مشاعل القاضي

علم البيانات من العلوم الرائجة في وقتنا الحالي والتي لها أُثرها العلمي والاقتصادي على مختلف المستويات.

ما هو علم البيانات؟

هو علم يقوم بتوظيف مختلف الطرق العلمية، مثل: العمليات الرياضية، الإحصائية، الخوارزميات، الأتمتة، والنمذجة، لاستخراج المعرفة من البيانات، بمختلف أنواعها.

سواء للمنظمة التي تمتلك هيكلة، أو تلك التي لا تملك هيكلا، أو تنظيماً واضحا. وتعتبر البيانات، الصيغة الخام، التي يتم تحويلها عبر علم البيانات، إلى حقائق أو معرفة، وبناء عليها يتم اتخاذ القرارات، واستكشاف خبايا البيانات.

ويجمع علم البيانات، بين علوم مختلفة، مثل: تحليل المشكلات، الإحصاء، علم الآلة، التمثيل الرسومي، والبرمجة. مما يسمح برؤية البيانات بشكل فعّال ومختلف وبشكل مثمر.

ما هو الفرق بين علم البيانات والبيانات الضخمه وتحليلات البيانات؟

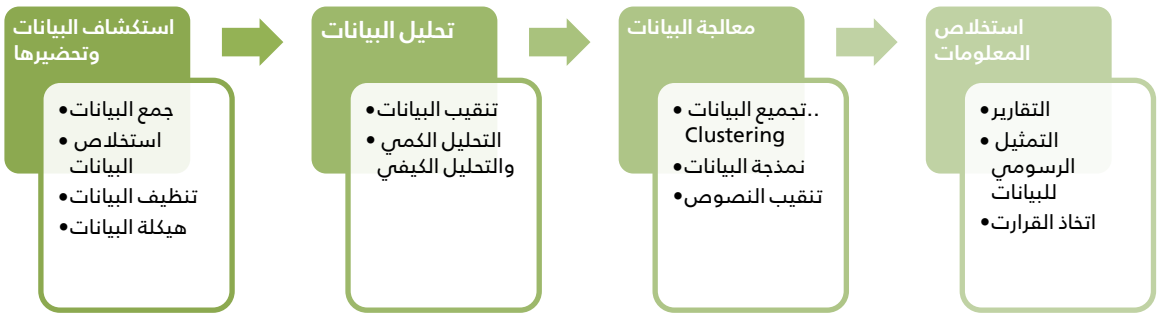
علم البيانات (Data Science) هو العلم المختص بكل ما هو متصل باستكشاف البيانات، تنظيفها، إعدادها وتحليلها، نمذجتها وتمثيلها رسوميا.

بينما **البيانات الضخمة (Big Data)** فهو علم يستخدم لتحليل البيانات ذات الأحجام الهائلة والتي ستعتمد عليها قرارات تقود المؤسسات في خططها الإستراتيجية وتكون ذات تأثير مباشر على أدائها. غالبا ما تاتي هذه البيانات من عدة مصادر وتنمو بشكل سريع وكبير مما يجعل معالجتها بالوسائل والأدوات التقليدية ليس بالشي السهل.

تحليلات البيانات (Data Analytics) وتتضمن أتمتة عمليات تجميع وإستعلام البيانات للحصول على معلومة من البيانات الخام والتي هي في شكلها الأولي.

المراحل التي يمر بها علم البيانات:

يمر علم البيانات بعدة مراحل، يمكن تلخيصها في أربعة مراحل أساسية وهي:



1. مرحلة تحليل البيانات الاستكشافي Exploratory Data Analysis (EDA)

وهي المرحلة الأولى والتي تتضمن معرفة بياناتك وهدفك الذي تسعى للوصول إليه من خلالها أو ما هو السؤال الذي تسعى للحصول على إجابة له؟

في هذه المرحلة يتم التعرف على مجموعة البيانات (dataset) وهي مجموعة من عناصر البيانات المرتبطة ببعض، وتُمثّل على هيئة جداول مكوّنه من صفوف وأعمدة وكل صف يطلق عليه سجل. ومثال ذلك، السجل الخاص بطالب معيّن يتكون من عدة حقول (اسم الطالب، الجنس، تاريخ ميلاده، تخصصه، المستوى الدراسي) وغير ذلك من بيانات الطلاب .

هي أول مرحلة من مراحل تحليل البيانات حيث يتم فيها فهم، وتلخيص وتحليل محتوى مجموعة البيانات، للحصول على فكرة أولية عما تحتويه، بعبارة أخرى مرحلة تحليل البيانات الاستكشافي تتمحور حول الوصول وتنظيم وهيكلة البيانات الخام لتهيئتها لمرحلة تحليل البيانات .

2. إعداد البيانات Data Preparation

تعتبر هذه المرحلة إمتداد للمرحلة السابقة وجزء منها وتتضمن عدة عمليات منها: إزالة العناصر المضلّلة أو العناصر غير المقبولة من مجموعة البيانات ومعالجة البيانات المفقودة وتحديد العناصر الشاذة بالإضافة إلى تصحيح البيانات غير المتوافقة وغيرها من العمليات الأخرى.

تحديد المتغيرات تعتبر أيضا عملية مهمة بمرحلة إعداد البيانات، وتتضمن تحديد أنواع المتغيرات وتحديد نوع البيانات المخزنة لكل متغير (نصي، رقمي، منطقي)، بالإضافة إلى تصنيف المتغير.

نستطيع أن نقول أن المتغيرات هي اسم الحقل في سجل الطالب، ومثال على ذلك حقل (جنس الطالب) والقيمة التي يحتوي عليها الحقل يُطلق عليها قيمة المتغير، ومثال على ذلك (ذكر أو أنثى).

3. مرحلة تحليل البيانات Data Analysis

هي المرحلة التي تتطلب استخدام التفكير المنطقي ومهارات التحليل لإيجاد نمط أو استنتاج أو حقائق من البيانات التي تم تجميعها وتهيئتها في المرحلة السابقة وهي مرحلة إعداد البيانات. تحليل البيانات يتضمن تطبيق خوارزميات وعمليات آلية على مجموعة من البيانات لاستخلاص علاقات داخل هذه البيانات. كما أن تحليل البيانات يتضمن تنقيب البيانات، تحليل النصوص، وذكاء الأعمال .

من خلال عملية تحويل البيانات قد تجد في هذه المرحلة البيانات التي تبحث عنها، ولكن قد تحتاج أحياناً إلى مراجعة سؤال البحث أو تجميع بيانات أكثر. التحليل المبدئي للأنماط والعلاقات، وانتشار البيانات، والقيم المتطرفة يساعدك في توجيه بياناتك لإجابة سؤال البحث بشكل أفضل.

في هذه العملية، يتم استخدام العديد من الأدوات وبرامج الكمبيوتر المتاحة لتسهيل هذه المهمة. ومنها مايكروسوفت فيزيو، ماتلاب، وستاتا المستخدمين تحديداً في عمليات الإحصاء المتقدمة لتحليل البيانات. بالرغم من توفر هذه البرامج المتطورة، ويعد الاكسل أيضا من البرامج الفعّالة وواسعة الإنتشار والمستخدمة لتحليل البيانات.

4. مرحلة نمذجة البيانات Modelling

هي تمثيل مبسط للبيانات، يتم انشاءه لهدف محدد. تتضمن هذه المرحلة اختيار خوارزميات وتقنيات النمذجة، توليد تصميم لنموذج البيانات، إنشاء النموذج، وتقسيم البيانات إلى بيانات التدريب وبيانات الاختبار وذلك لأهداف تقييم النموذج واختبار معاملات أخرى للنموذج الذي تم اعتماده.

4.1. إختيار تقنيات النمذجة: في هذه المرحلة يتم اختيار تقنية او أكثر من تقنيات النمذجة مثل:

4.1.1. **تقنيات التجميع:** هي عملية وضع البيانات في تجمعات متشابهة. تقسم خوارزمية التجميع مجموعة البيانات إلى عدة تجمعات حسب درجة التشابه، مثلا "هل العملاء يمكن تقسيمهم إلى مجموعات مختلفة حسب تفضيلات الشراء؟". هناك العديد من الخوارزميات المستخدمة في عملية تجميع البيانات أشهرها خوارزمية K-means clustring.

4.1.2. **تقنيات التصنيف:** هي عملية تصنيف البيانات إلى مجموعات مختلفة على حسب الهدف من تحليل البيانات، مثلا "هل من الممكن إيجاد مجموعات من العملاء لديهم احتمالية عالية لإلغاء خدماتهم بعد انتهاء العقد؟"

4.2. إنشاء تصميم لاختبار النموذج:

بعد اختيار تقنيات النمذجة، يقوم محلل البيانات باختبار جودة النموذج من خلال اختبارات تجريبية وفحص النتائج لمعرفة نقاط القوة والضعف للنموذج. يتم الاختبار عادة من خلال تقسيم البيانات إلى بيانات تدريب وبيانات اختبار، وبناء النموذج باستخدام بيانات التدريب وتقييم الجودة باستخدام بيانات الاختبار. في هذه المرحلة يمكن لمحلل البيانات قياس قوة النموذج بتوقع البيانات المعلومة مسبقا وذلك لمعرفة مدى قوته في توقع بيانات ونتائج مستقبلية.

4.3. بناء النموذج:

بعد الاختبار، يطبق محلل البيانات النموذج على البيانات التي تم تجهيزها في مرحلة "تجهيز البيانات"، ويتم حينها اتخاذ القرار للحصول على النمذجة المثاليّة.

4.4. تقييم النموذج:

في هذه المرحله يقرر محلل البيانات مدى نجاح تطبيق النمذجة المطروح، ويقوم بتحديد كفاءة التقنيات الموجودة. في الغالب يتم القرار من خلال العمل مع محلل أعمال وخبراء من ذات المجال للعمل على النتائج من منظور عملي. على سبيل المثال، نموذج لاختبار العوامل التي تؤثر على أرصدة الحسابات البنكية الخاصه قد يواجه مشاكلاً في حال رصد مجموعتين من البيانات في توقيتين مختلفين من الشهر، وسيلاحظ محلل أعمال على دراية بعمليات البنك وجود مثل هذا التعارض. بالتالي يحاول محلل البيانات أن يعيد تقييم البرنامج بناءً على المعايير المحددة من محلل الأعمال وآخذًا بعين الاعتبار أهداف النموذج.

5. مرحلة تمثيل البيانات Data Visualization

يساهم تمثيل البيانات في شكل صوري في عرض الأنماط وأبرز الاتجاهات المستلهمة من البيانات، حيث أن عدد من الأنماط قد لا يظهر بشكل واضح إذا كانت البيانات في هيئة نصية. كما أنه هنالك العديد من التطبيقات التي تسمح للمستخدم بتصفية البيانات تبعاً لمتطلباته. العديد من الخيارات التقليدية لتمثيل البيانات مثل الجداول، الرسوم البيانية، والمخططات العمودية، الخرائط الحرارية وغيرها ومؤخراً بدء استخدام التصوير ثلاثي الأبعاد لتمثيل البيانات. أدوات تمثيل البيانات تساعد في عرض البيانات في شكل متكامل ومقروء حتى لغير المختصين.

وتعد أهم الخصائص لتمثيل البيانات في صورة فعالة هي:

5.1. **النصوص:** تجنب إضافة الكثير من النصوص والحرص على إضافة المعلومات المهمة فقط كعنوان الرسم، عنوان المحور السيني والصادي. كما يجب مراعاة اتجاه وحجم النص حيث يستحسن استخدام الاتجاه الأفقي للقراءة بسهولة.

5.2. **الألوان:** مراعاة تناسب الألوان مع الموضوع، فمثلاً ارتباط اللون الوردي مع الإناث والأزرق مع الذكور بالإضافة الى اختيار لون مناسب للخلفية حيث من الرائج اختيار لون معاكس لما هو في النص. كما أن الألوان تعد عامل هام لجذب انتباه المشاهد للأجزاء والنقاط الجوهرية.

5.3. **النُسق:** التنسيقات قد تسبب في إيصال فكرة خاطئة للمشاهد لذلك يجب الحرص على اختيار الأحجام والأطوال بما يتناسب مع البيانات المُمثلة.

ووفقاً لمبادئ وليام كليفلاند فإن التمثيل الصوري للبيانات يجب أن يعرض أكبر قدر مستطاع من المعلومات بدون أن يتسبب ذلك في إجهاد معرفي للمتلقي، والوضوح هو أهم عنصر لذلك تجنب الأشكال والنصوص المضللة كما أن التصوير البياني لابد أن يكون حلول وأجوبة لتساؤلات في نطاق البيانات المُمثلة.

وبناء على تحليل البيانات وتمثيلها تتم عملية إتخاذ القرار أو استخلاص وجود رابط بين العوامل Correlation والتأكد رياضيا فيما إذا كان يعني ذلك وجود سببية Causation أم لا. وتوجد لغات برمجية كثيرة تستخدم في علم البيانات ومنها لغة بايثون Python ولغة R وأيضا توجد برامج متخصصة توفر على المختص في علم البيانات الكثير من الجهد مقارنة بالعمل الذي قد يبذله في صنع برامج من الصفر مثل برنامج Tableau.

المصادر:

- https://www.datasciencecentral.com/
- https://towardsdatascience.com/data-preparation-and-exploration-5e09b92cf00e
- https://www.dezyre.com/article/why-data-preparation-is-an-important-part-of-data-science/242
- https://whatis.techtarget.com/definition/data-set
- https://livebook.manning.com/#!/book/exploring-data-science/chapter-1/69
- https://www.fingent.com/blog/data-visualization-vs-data-analytics-whats-difference
- https://depictdatastudio.com/checklist/