التاريخ: / /

الموضوع:

lec 2

Data Acquisition (Data Collection)

The process of gathering data and making it useful by filtering and cleaning it as per the business requirement.

هي عملية جمع الداتا وتنظيمها وتنظيفها بحيث تكون مفيدة ومناسبة لاحتياجات المشروع أو المشكلة اللي بنحلها.

← كنا قولنا إن الداتا اللي بشتغل عليها بالـ data science هي big data وبالتالي فيها تنوع كبير

variety

structure, unstructure, semi-structure

① Structured data. لها شكل معين للتخزين

This is data that is orginized in a clear and consistent format. such as tables in a database - excel files.

it is stored in rows and colums to make it easy to search and manipulate.

② un structured data. ملهاش شكل محدد

data that has no specific format or clear way of organizing it.

ex. image, vedio, ...

التاريخ: / /

الموضوع:

③ Semi-Structured data.

This is an intemediate type between structure and unstructure data,

meaning it contains some structured ~~element~~ element s, but not as fully as traditional DB.

نوع وسطي بين البيانات المنظمة وغير منظمة

زي ال JSON، سجلات (logs)

Data prepro cessing

raw data

The data we collect is often unclean. it may contain errors, missing data, duplicates, or be in the wrong format, so we need to go through a data preprocessing stage.

البيانات اللي بنجمعها عادة بتكون مش نظيفة، وعشان كده لازم نعدي بمرحلة ال preprocessing

البيانات كل ما كانت أنظف وأدق، كل ما كان النموذج اللي هنعمله أفضل وأدق في التوقع.

التاريخ: / /

الموضوع:

basic data preprocessing techniques

data preprocessing is done using four main techniques

(1) Data cleaning تنظيف

(2) Data integration دمج

(3) Data transformation تحويل

(4) Data Reduction. تقليل

① Data cleaning:-

دي عملية مهمة جدا عشان نحسن جودة البيانات قبل ما نستخدمها في التحليل أو النمذجة

Data cleaning step ← ممكن نقول عليها

A) Remove duplicate. or irrelevant observations.

أحيانا أثناء تجميع البيانات، ممكن يحصل تكرار للقيم أو إدخال بيانات غير مفيدة.

مثلاً: تدخل اسم دولة مرتين بشكل مختلف دا يعتبر تكرار

⇒ NewZeland → New Zeland

B) Fix structural errors.

دي الأخطاء اللي بتحصل بسبب الكتابة الغلط أو اختلاف طريقة الإدخال.

زي مثلا كتابة التواريخ بـ format مختلفة.

لازم نوحد الصيغة اللي هندخل بيها الداتا.

الموضوع: _____ التاريخ: / /

(c) filter unwanted outliers

. ال outliers هي القيم اللي بتلاقيها بعيدة عن باقي القيم (الداتا)
ودي ممكن تأثر على التحليل.
مش كل ال outliers غلط، لكن لو كانت غير منطقية
بيقا يفضل تشيلها
مثلاً شخص عمره 500 سنة ← مش منطقي

(D) handle missing data. →

. During analysis of data, it is important to understanding why the missing value exist.
في بعض البيانات، ممكن تلاقي cells فاضية وده له أسباب مختلفة
والحل بيعتمد على سبب نقص البيانات:- ٣ حالات
① in case, the values were not recorded,
it becomes essential to fillup the values by guessing based on the other values in that columns and rows
This is called imputation. التقدير
① يبقى لو القيم مفقودة بالصدفة (مثلاً حد نسي يسجلها)
ممكن نحاول نخمنها بناءً على القيم التانية

التاريخ: / /

② if a value is missing because it doesn't exist.
↳ leave it empty or to add a third value like NA or NaN.

② لو القيمة مش موجودة اصلا لأنها مش منطقية لشخص معين
ذي مثلاً "طول الطفل الاكبر" وهو أصلاً معندوش اطفال.
وقتها ممكن نسيب الـ cell فاضية أو نحط فيها NaN.

③ in case, it is not possible to figure out the reason for the missing, then that particular value can be dropped.

هل بنمسح الصف و العمود؟

(E) Nosiy data

دهما البيانات اللي بتحتوي على أخطاء أو معلومات غير دقيقة
عشان نعالجها في طرق مختلفة ذي:-

(a) binning methods (discretization)
is used to smooth sorted data values by consulting the values around.

وده معناه إننا هنقسم الداتا لمجموعات صغيرة هنسميها (bins)
وبعدين هنستخدم طريقة من الثلاث طرق دول:

① bin means
هنحسب المتوسط لكل bins ونبدل القيم بالمتوسط.

② smoothing by bin median.

• بنحسب الوسيط لكل مجموعة ونستخدمه بدل القيم الاصلية.

③ smoothing by bin boundaries.

• بنشوف اكبر قيمة واصغر قيمة وباقي القيم بنقاردهم بالاكبر والاصغر ونشوف هما اقرب لمين ونبدل لها بيها.

bing 1, 3, 4, 9, 15, 2, 2, 16

1, 1, 1, 16, 16, 1, 1, 16

min: 1

max: 16

نشوف 3 مثلاً اقرب لـ 1 وهلا 16 اكيد 1 فبقا نبدلها بـ 1

Algorithm:-

1- Sort the datasets
2- Partition the dataset to bins. (equal bins)
3- Calculate the arithmetic mean or median or replace by boundary (min or max value)

4- Replace each data element in each bin by the calculated mean / median / boundaries.

التاريخ: / /

الموضوع:

example\ Given data: 18, 22, 6, 6, 9, 14, 20, 21, 12, 18, 18, 16
illustrate bining by Mean, median, boundary replacement. Given bin depth = 3

Step 1: sort the data

6, 6, 9, 12, 14, 16, 18, 18, 18, 20, 21, 22

Step 2: Partition the data into equal frequency bins.

• num of bins = N/d

num of element = 12 bin depth = 3 عدد الـ element في كل bin

12/3 = 4 bins

bin1: 6, 6, 9
bin2: 12, 14, 16
bin3: 18, 18, 18
bin4: 20, 21, 22

step 3: Mean = sum of observations / num of observation

bin1: (6+6+9)/3 = 7
bin2: (12+14+16)/3 = 14
bin3: (18+18+18)/3 = 18
bin4: (20+21+22)/3 = 21

∴ step 4 Replace each data element in each bin by the calculated mean.

bin1: 7, 7, 7
bin2: 14, 14, 14
bin3: 18, 18, 18
bin4: 21, 21, 21

الموضوع: _____ التاريخ: / /

- binning using median

القيمة اللى فى النص

في حالة عدد ال ellement في ال bin

فردى

bin1 : 6, 6, 6

لو عدد القيم زوجى

هنجيب الرقمين اللى فى النص ونجمعهم ونقسمهم على 2

bin2 : 14, 14, 14

bin3 : 18, 18, 18

bin4 : 21, 21, 21

- binning using boundary Values

bin1 : 6, 6, 9

bin2 : 12, 12, 16

bin3 : 18, 18, 18

bin4 : 20, 20, 22

(b) Regression الانحدار

is used to find a mathematical equation to fit the data to smooth out the noise.

- هى طريقة رياضية بنستخدمها عشان نلاقى معادلة تناسب البيانات وتساعدنا نقلل ال noise اللى فيها.

عندنا نوعين من الانحدار

1. liner regression
2. Multiple liner regression.

التاريخ: / /

الموضوع:

① linear regression

is used to find the best line that fits two variables such that one predicts the other.

ده بنستخدمه لو عندنا متغيرين (زى عدد ساعات المذاكرة ودرجات الامتحان)

وبنحاول نرسم خط مستقيم يوضح العلاقة بينهم

بحيث نقدر نتوقع قيمة متغيرة بناء على التانى.

② Multiple linear regression

. involves more than two variables.

. هنا بيكون عندنا اكتر من متغير، بنحاول نلاقى علاقة رياضية

تربط كل المتغيرات ببعضها.

(c) clustering

. This approach is useful for organizing similar data in groups or clusters.

بنقسم الداتا لمجموعات متشابهة clusters بحيث

كل مجموعة يكون فيها بيانات شبه بعض.

. The outliers may be detected by clustering.

التاريخ: / /

الموضوع:

② Data integration

data acquired from multiple sources is integrated into a single centralized location.

• هي عبارة عن دمج بيانات من مصادر مختلفة عشان تبقى أكثر شمولاً.

The most Common approaches to integrate data :-

• في كذا طريقة نقدر نستخدمها علشان ندمج البيانات من مصادر مختلفة بحيث تبقى منظمة و أسهل في الاستخدام

① Data Consolidation

② Data propagation

③ Data Virtualization

④ Data Warehousing

دول اللي يهموا الدكتور باذن الله

التفاصيل اللي تحت دول افهمها بس.

① data Consolidation

• Means to consolidate data from several seperate sources into one data store, so that it is available to all the stake holders to enable better decision making.

بنجمع الداتا من مصادر مختلفة في مكان واحد، بحيث تكون متاحة لكل الناس اللي بيحتاجوها عشان يقدروا ياخدوا قرارات أفضل.

الموضوع: _____ التاريخ: / /

- The most common data Consolidation techinques are:

A) ETL (extract, transform, load)

بيتم استخراج البيانات من مصادر مختلفة، بعدين بنعدل عشان تكون مفهومة (ترتيبها، تنظيفها، تجميعها) وبعد كدا بنخزنها في قاعدة بيانات مركزية أو في warehouse.

.is the most widely used data consolidation technique.

(1) The data is first extracted from multiple sources.

(2) then it is transformed into an understandable format by using various functions.
(aggregating, sorting, cleaning)

(3) then transfer it to a centralized store.

(B) Data ~~Vis~~ Virtualization.

- هنا البيانات بتفضل في مكانها الأصلي، لكن بيتم عرضها للمستخدمين كأنها مجمعة في مكان واحد من غير الحاجة لنقلها فعليًا.

.the data stays in the original location.

التاريخ: / / الموضوع:

(e) Data Warehousing.

- is the integration of data from multiple sources to centralized source to facilitate decision making, reporting and query handling.

(2) Data propagation.

- في الطريقة دي، البيانات بيتم نسخها من مكان لمكان تاني (من المصدر للمكان المستهدف).

- involves copying data from one location (source) to another location (target location).

- These applications usually operate online and push data to the target location.

(3) Data Virtualization

(4) Data Warehousing.

الموضوع: _____ التاريخ: / /

③ Data transformation

• بعد ما بنجمع الداتا وننظفها، ممكن تفضل مش متناسقة أو مش جاهزة للاستخدام. عشان كده، لازم نغير شكلها وتركيبها بحيث تبقا أسهل في التحليل والاستخدام.

• Data transformation may be Constructive destructive, aesthetic or structural.

• Constructive data transformation
  - involve adding, Copying, replicating data.

• Destructive
  - involve deleting fields and records.

• Aesthetic
  - involve standardizing salutions or street names.

• Structural
  - involve renaming, moving Combining Columns in data base.

---

④ Data Reduction

- it is the process of reducing the volume of data such that data integrity is preserved.

بقلل حجم الداتا من غير ما تفقد أي معلومات مهمة، يعني حتى بعد تقليلها النتائج اللي هتطلع هتفضل زي ما هي.

التاريخ: / / الموضوع:

Data mining

- هى العملية اللى بنستخدمها عشان نستخرج معلومات مفيدة من الداتا الكتير اللى متجمعة.
- الهدف الاساسى من ال data mining هى اتخاذ قرارات مدعومة

- Data mining is the process that helps in extracting information from a given data set to identify trends, patterns, and useful data.

- The objective of using data mining is to make data-supported decisions from enormous datasets.

→ Different types of data can be mined :-

1. Data stored in database
2. data ware house
3. transactional data.
4. data streams
5. engineering design.

التاريخ: / / الموضوع: _____

The most important data mining techniques.

① association الارتباط

- discover a link between two or more items.
- it uses if-then statements to show the probability of interactions between data items.

ديما هنا بنشوف ارتباط الـ items ببعض مثلا في الشراء
لو شخص هيشتري فون غالبا هيشتري جراب مثلا.

② clustering

مبيجمع الداتا المتشابهة مع بعض في مجموعات clusters
بحيث كل cluster بيضم بيانات بتشترك في نفس الصفات
ومبتعتمد على الـ unsupervised learning.

- is a data mining technique in which clusters are created of objects that share the same characteristics.
- it is based on unsupervised learning.

لـ بيجمع الحاجات اللي شبه بعض مع بعض
لكن ميعرفش اسماءهم ايه

unsupervised = clustering
supervised = classification

التاريخ: / /

الموضوع:

③ classification

- classifies item's in a data set into predefined groups.
- it is based on ~~un~~ supervised learning

لما الدكتور كتبها في الكتاب un supervised

ولكن ان شاء الله الصح

predefined groups لأننا قولنا

⇨ The most important Data mining application.

- predictions, classifications, clusters.
- it is used in fields such as:-
  1- health care
  2- Fraud detection in banking transactions.
  3- education
  4- Market basket analysis
  5- Manufacturing and Quality.
  6- Customer Relationship Managment.

التاريخ: / /

الموضوع:

Data Interpretation تفسير البيانات

بعد ما بنحلل البيانات، لازم نقدر نفهم معناها ونوصل لاستنتاجات مفيدة.

Qualitative نوعي

- Relies on at analyting non-numerical data (pictures, symobols)
- we use it when you went to understand patterns and behaviors without clear numbers.

Quantitative كمي

- Relies on analyting numerical data with numbers and statistics.
- we use mathematical methods such as
  1- mean
  2- standard deviation
  3- Regression