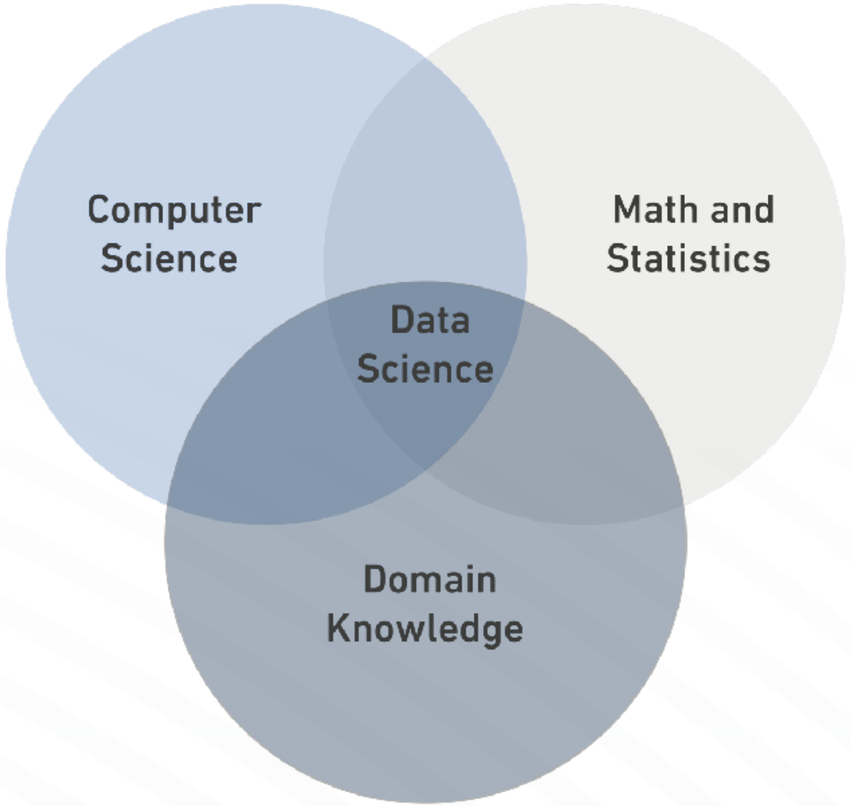
أكاديمية طويق
TUWAIQ ACADEMY

مقدمة في تعلم الآلة

ماهو علم البيانات؟

ماهو علم البيانات

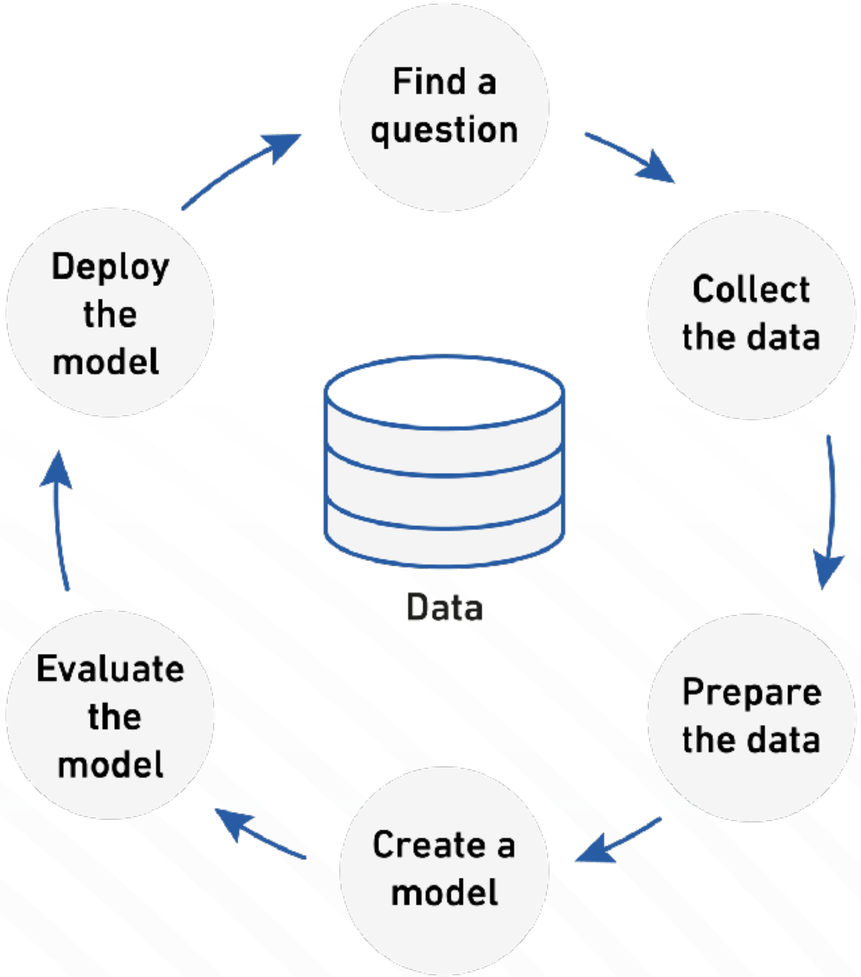


- هو علم يقوم بتطبيق الأساليب الإحصائية على البيانات بغرض الوصول لحل مشكلة معينة، حيث يتم الاستفادة من البيانات التي يتم إنتاجها بشكل يومي من عدة مصادر مثل وسائل التواصل الاجتماعي وغيرها بغرض الوصول لقرارات أو استنتاجات تقوم بحل مشكلة ما.
- ويمكن تعريفه أيضا على أنه علم قائم على التقاطع بين عدة علوم وهي علم الإحصاء وعلوم الكمبيوتر و(Domain Knowledge) والمقصود به المجال التابع للمشكلة التي نقوم بحلها وهو يختلف بحسب نوع البيانات فمثلا: بيانات المرضى تتبع المجال الصحي وبيانات الأسهم تتبع المجال المالي وهكذا.
- يهدف علم البيانات لتحويل data إلى Knowledge تساعدنا في اتخاذ القرارات.

تطبيقات علم البيانات

- التنبؤ بدرجات الحرارة للأسبوع القادم.
- التنبؤ باحتمالية إصابة شخص بمرض معين.
- معرفة آراء الناس حول حدث معين.

Data Science LifeCycle - المراحل الأساسية في علم البيانات



المراحل الأساسية في علم البيانات - Data Science LifeCycle

أي مشكلة نقوم بحلها عن طريق علم البيانات بشكل عام تمر بعدة خطوات:

أولا: طرح التساؤلات وهذا قد يكون فرضية نريد اختبارها أو قرار نريد اتخاذه أو منتج نريد إنتاجه
ثانيا: جمع البيانات المتعلقة في المشكلة التي نريد حلها وهذا يعتمد على نوع المشكلة فأحيانا لانحتاج لجمع بيانات ويمكن أن نستخدم بيانات جاهزة
ثالثا: تجهيز البيانات حتى يتم تحليلها عن طريق تنظيفها (Data Cleaning) وعمل تحويل للبيانات (Data Transforming) إلى شكل يناسب عمل التحليل.
رابعا: بناء النماذج للبيانات (Data Modeling) وهناك أشكال مختلفة للنماذج مثل: (Numerical Model) أو (Visual Model) أو (Statistical Model) أو (Machine Learning Model) ونقوم باستخدامها لإثبات فرضية معينة أو التنبؤ بنتيجة معينة.
خامسا: تقييم النموذج و في هذه المرحلة نحتاج للتأكد من النموذج هل قام بالإجابة عن التساؤلات التي طرحناها بشكل دقيق أم لا، هل ساعد في اتخاذ القرارات أو التنبؤ بنتائج معينة.
سادسا: نشر النموذج وهذه الخطوة تكون بعد التأكد من دقة النموذج، ثم بعد عملية النشر يمكن استخدام النموذج على أنواع أخرى من البيانات.

أخيرا، Data Science LifeCycle تعتبر iterative Process بحيث نقوم بتكرار هذه العملية في كل مره نطرح سؤال معين أو نحاول اتخاذ قرار لحل مشكلة ما وفي كل مره نحسن العملية حتى نصل لنتائج دقيقة، أيضا هي تعتبر غير متسلسلة Non-Sequential حيث نقوم بالتقدم لخطوات forward أو التراجع لخطوات backward بناء على النتائج التي نحصل عليها.

أشهر المكتبات في علم البيانات

مكتبة pandas

يتم استخدامها لتحليل البيانات (Data Analysis) وهيكلتها (Data Structures).

مكتبة NumPy

تقوم بتوفير إمكانية التعامل مع المصفوفات (Multidimensional Arrays) والعمليات الرياضية (Linear Algebra Functions)

مكتبة matplotlib

يتم استخدامها لتمثيل أو عرض البيانات (Data Visualization) بشكل رسومات بيانية.

أشهر المكتبات في علم البيانات

مكتبة Seaborn

هي مكتبة بُنيت فوق مكتبة matplotlip لتمثيل البيانات بطريقة متقدمة على شكل رسومات بيانية تفاعلية .

مكتبة sikit-learn

تعتبر أحد المكتبات الخاصة بتنفيذ خوارزميات تعلم الآلة (Machine Learning Algorithms) .

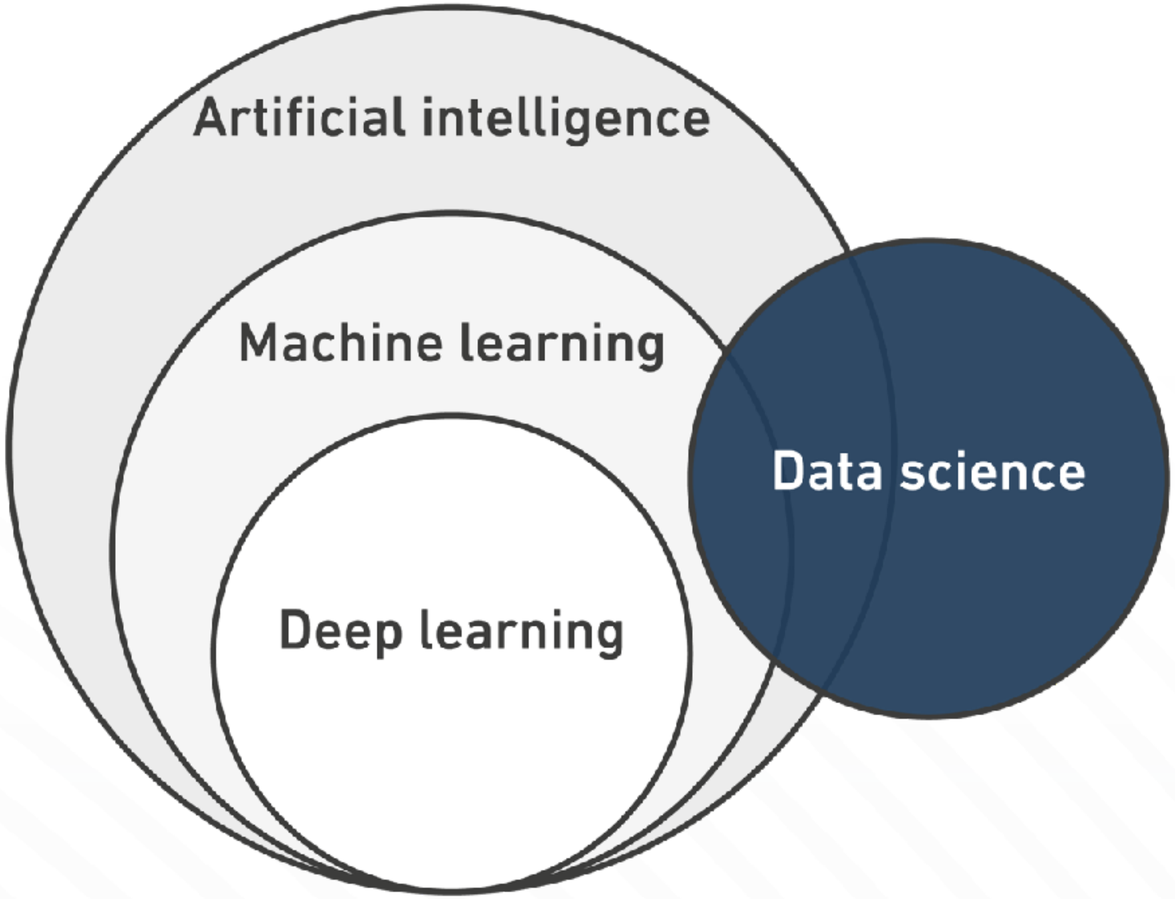
مكتبة PyTorch

تعتبر أحد المكتبات الخاصة بتنفيذ خوارزميات تعلم الآلة وتستخدم في (Natural Language Processing) و (Computer Vision) .

مكتبة TensorFlow

تعتبر أحد المكتبات الخاصة بتنفيذ خوارزميات الذكاء الاصطناعي و تعلم الآلة.

الفرق بين علم البيانات (Data Science) و الذكاء الاصطناعي (AI) و تعلم الآلة (Machine Learning) و التعلم العميق (Deep Learning)



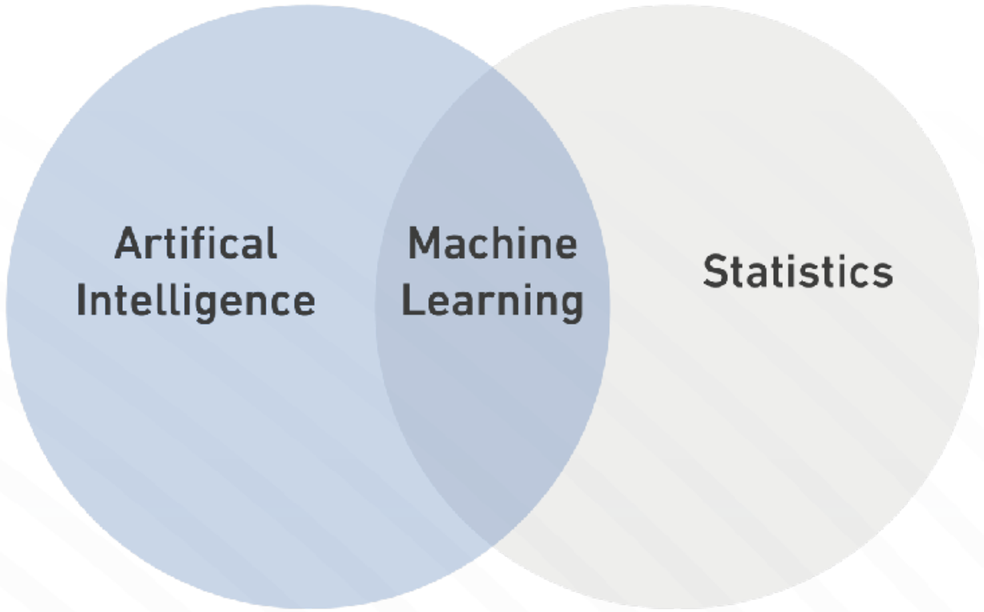
الفرق بين علم البيانات (Data Science) و الذكاء الاصطناعي (AI) و تعلم الآلة (Machine Learning) و التعلم العميق (Deep Learning)

الكثير من الأشخاص وخصوصا المبتدئين يجدون صعوبة في التفريق بين علم البيانات والذكاء الاصطناعي وتعلم الآلة و التعلم العميق وربما يعتقدون أن جميع هذه المصطلحات تشير لنفس المفهوم ولكن في الحقيقة يوجد بعض الاختلافات حيث أن:

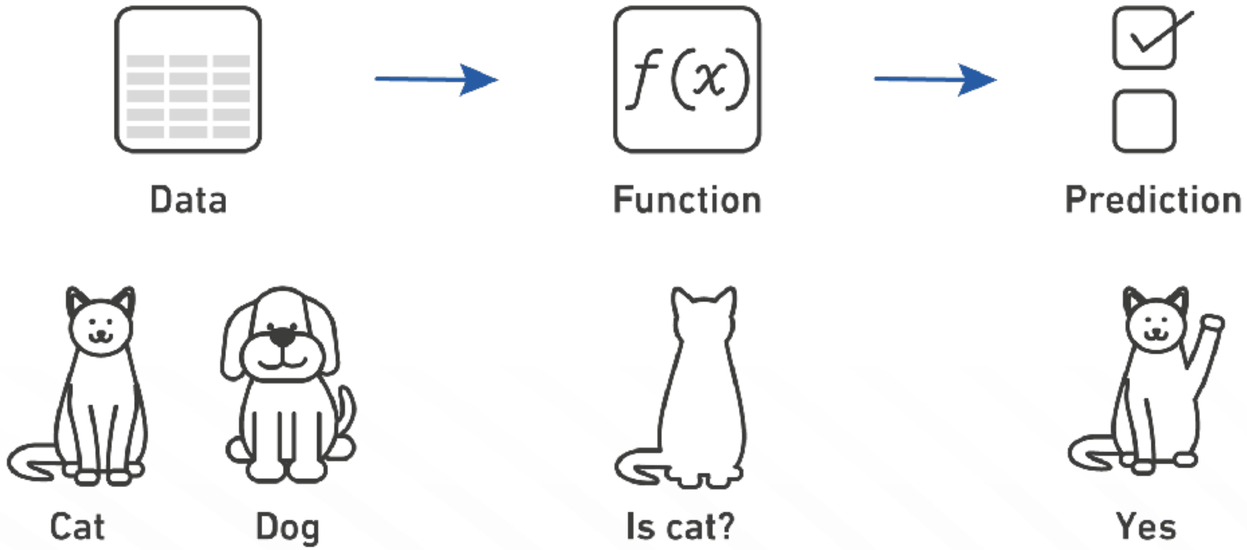
- علم البيانات يقوم باستخدام تقنيات (AI) و (Machine Learning) و (Deep Learning) وتطبيقها على البيانات بغرض الوصول لقرارات أو استنتاجات معينة.
- أما التعلم العميق (Deep Learning) فهو جزء فرعي من تعلم الآلة (Machine Learning) الذي هو أيضا جزء فرعي من الذكاء الاصطناعي (AI) وجميعها تشير للتقنيات التي تساعد الكمبيوتر على التعلم من البيانات لحل المشاكل المعقدة.

تعلم الآلة (Machine Learning)

- هو عملية تعليم الآلة إنجاز مهمة معينة دون كتابة كود صريح أو أوامر صريحة لتنفيذ هذا الأمر، ويمكن التعبير عنه بأنه جزء الذكاء الاصطناعي الذي يحتوي على احصائيات.
- أثناء عملية تعليم الآلة نقوم بإنشاء مايسمى بالنموذج (Model) الذي نقوم بتزويده بمجموعة البيانات و الخوارزمية (algorithm) للتعلم من البيانات.



تعلم الآلة (Machine Learning)



- نقوم باستعمال البيانات لتعلم بناء دالة تكون قادرة على التنبؤ بنتيجة البيانات الجديدة على سبيل المثال لنقل أننا نريد بناء دالة تقوم بتحديد هل الصورة تحتوى على قطة أم لا؟
- في البداية سوف نقوم بإنشاء بيانات تحتوي صور للقطط وصور لاتحتوي ذلك .
- ثم نقوم بتطبيق خوارزميات تعلم الآلة على مجموعة البيانات.
- تقوم هذه الخوارزميات بتعلم الدالة التي تتنبأ بالصور هل الصورة تحتوى على قطة أم لا؟

عوائل النماذج (Model Families)

الانحدار Continuous

تعتبر أنه توجد أنماط بين الخواص

* Logistic Regression
* Linear Regression
* Neural Networks

المسافة Distance

تعتبر وجود مسافة بين الخواص

* K-Means Clustering
* SVM
* DBScan

المتسلسلات الزمنية Time Series

تعتمد البيانات اللاحقة على البيانات السابقة

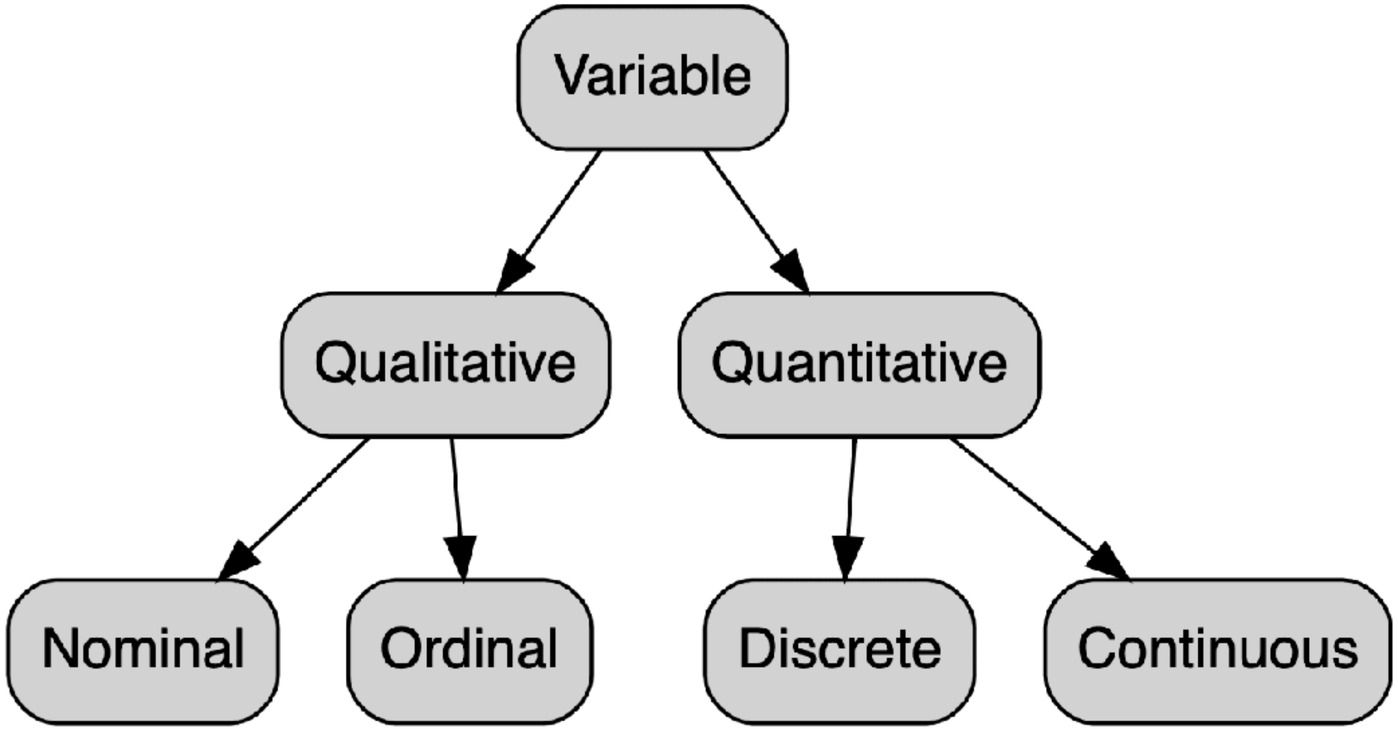
* ARIMA
* Prophet
* Markov

مصنِّفة Categorical

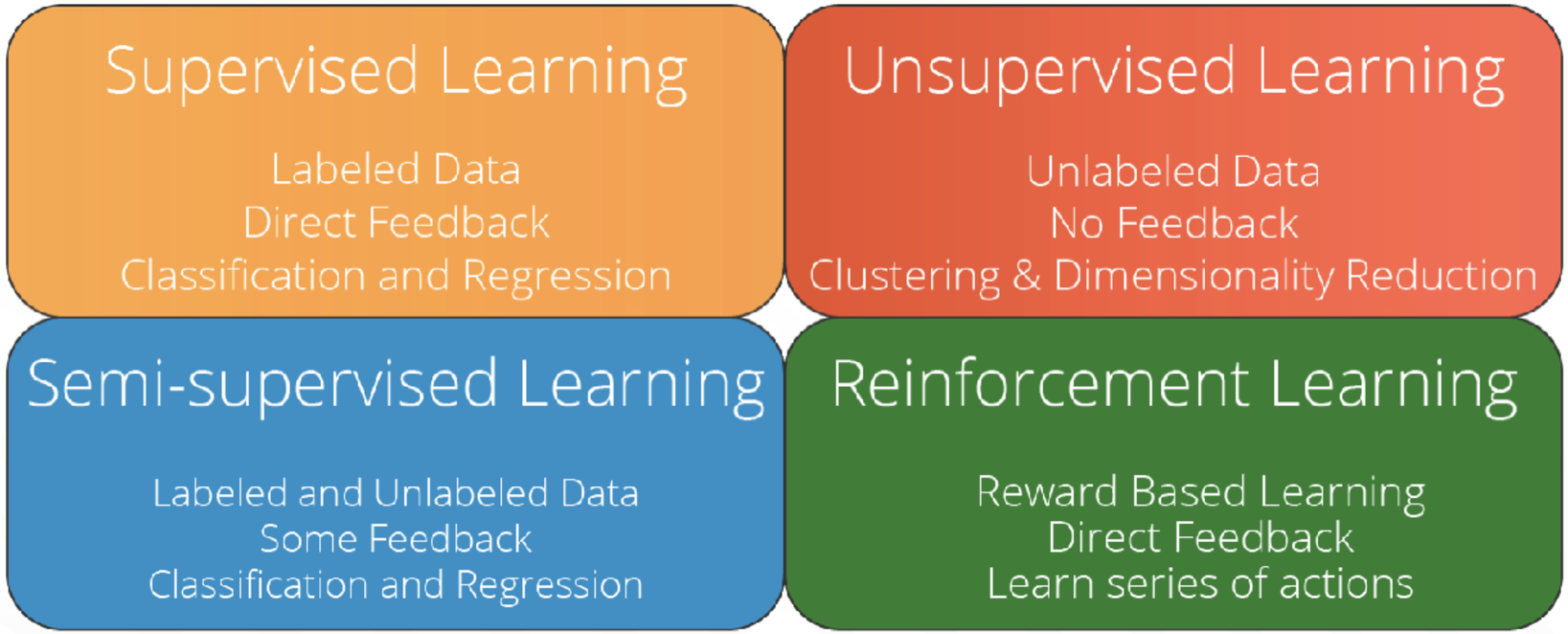
خواص تحتوي تصنيفات غير قابلة للترتيب (if statements)

* Naïve Bayes
* Decision Trees
* Random Forest

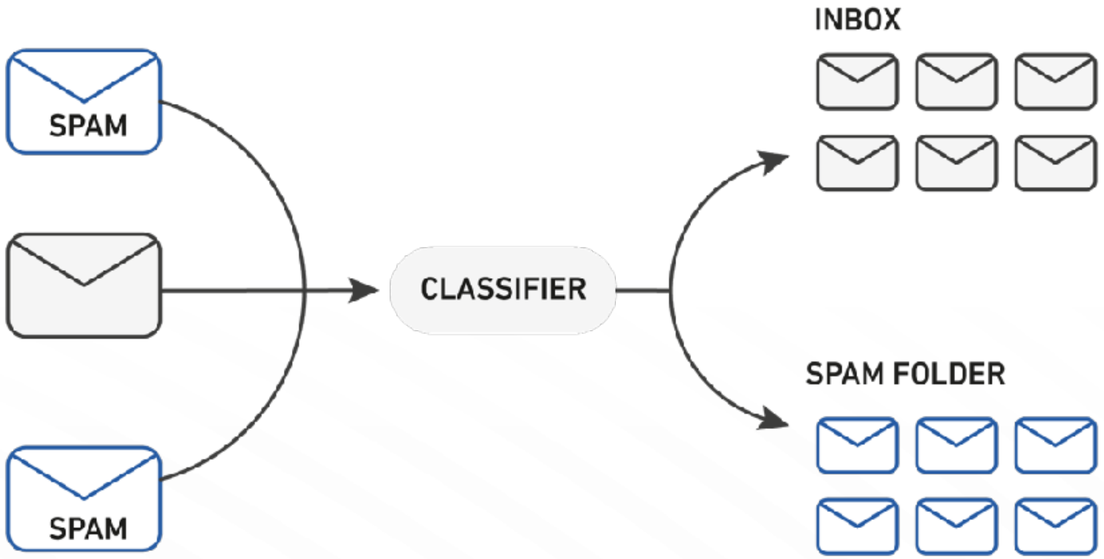
أنواع البيانات



خوارزميات تعلم الآلة (Machine Learning Algorithms)



خوارزميات تعلم الآلة (Machine Learning Algorithms)



النوع الأول (Supervised Learning)

يشير هذا النوع إلى العملية التي تتعلم فيها الآلات من مجموعات البيانات المصنفة أو المعروفة (labeled data) وتنتج نموذجًا دقيقًا قادرًا على التنبؤ بملصقات البيانات غير المرئية.

مثال: عندما نقوم بتدريب الآلة على تصنيف البريد إلى بريد مزعج (spam) أو غير مزعج (spam) خلال هذه العملية تتعلم فيها الآلة من مجموعة من البيانات المصنفة إلى بريد مزعج و غير مزعج بهدف إنشاء نموذج (Model) قادر على تصنيف أي بريد جديد إلى بريد مزعج وغير مزعج.

خوارزميات تعلم الآلة (Machine Learning Algorithms)

عادة ما تكون الخوارزميات المستخدمة في هذا النوع مصنفة لنوعين:

خوارزميات التصنيف (classification) مثل SVMs

خوارزميات الانحدار (Regression) مثل الانحدار الخطي (Linear Regression).

خوارزميات تعلم الآلة (Machine Learning Algorithms)

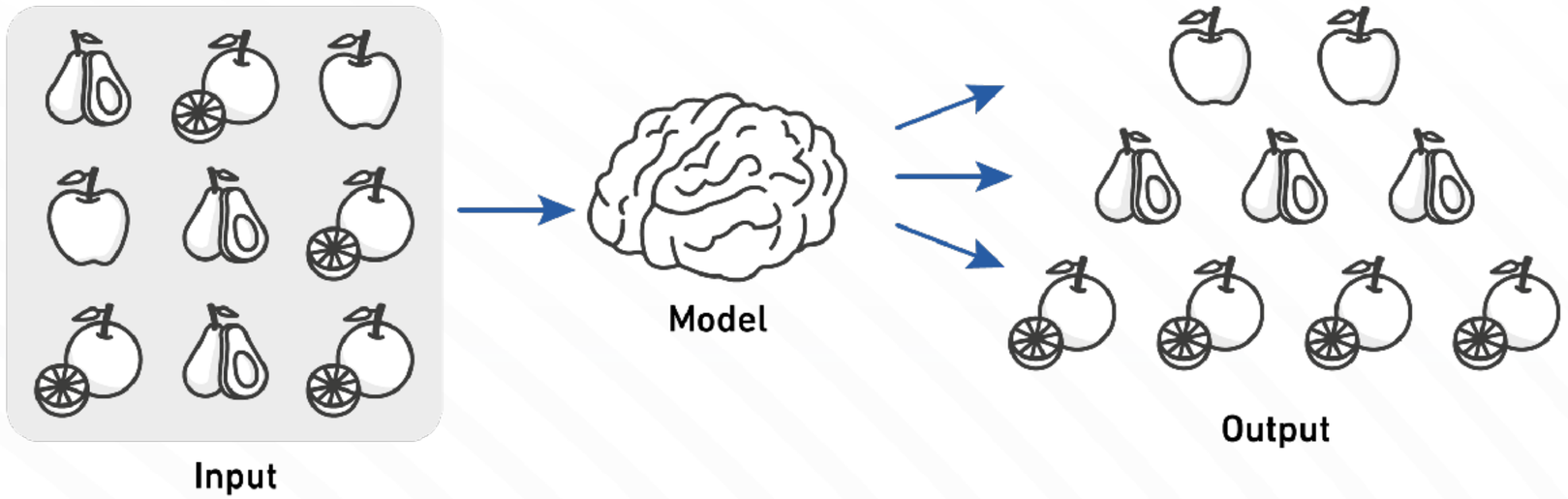
النوع الثاني (Unsupervised Learning)

يشير هذا النوع إلى العملية التي تتعلم فيها الآلات من مجموعات البيانات غير المصنفة بناءً على التشابه بين مجموعة البيانات.

مثال: عندما نقوم بتدريب الآلة على تصنيف الفواكه، في هذه الحالة لا نخبر الآلة عن اسم الفاكهة، إذا كيف تقوم الآلة بالتصنيف؟

تقوم الآلة بالتصنيف بناء على خصائص الفواكهه مثلا: اللون والحجم والشكل فمثلا: إذا كان اللون أحمر يتم تصنيف الفاكهة إلى تفاح وهكذا.

ومن أشهر الخوازميات على هذا النوع: خوارزمية k-Means.



خوارزميات تعلم الآلة (Machine Learning Algorithms)

النوع الثالث (Semi-Supervised Learning)

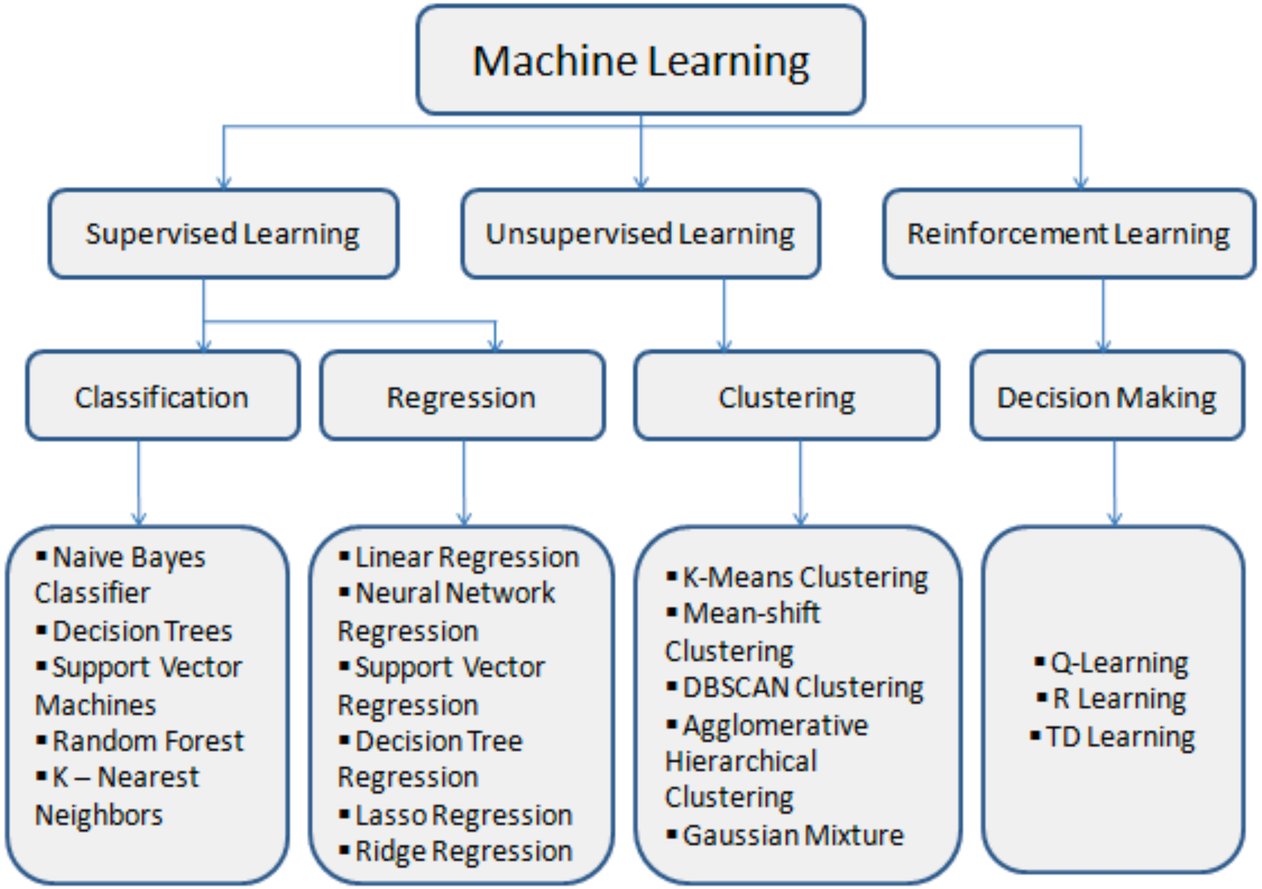
هذا النوع من التعلم نحتاجه بالغالب عندما يكون لدينا مجموعات البيانات الكبيرة ولكن عدد قليل من هذه البيانات قد تم تصنيفة، لذا تقوم الآلة بالتعلم من كلا النوعين البيانات المصنفة والغير مصنفة.

النوع الرابع (Reinforcement Learning)

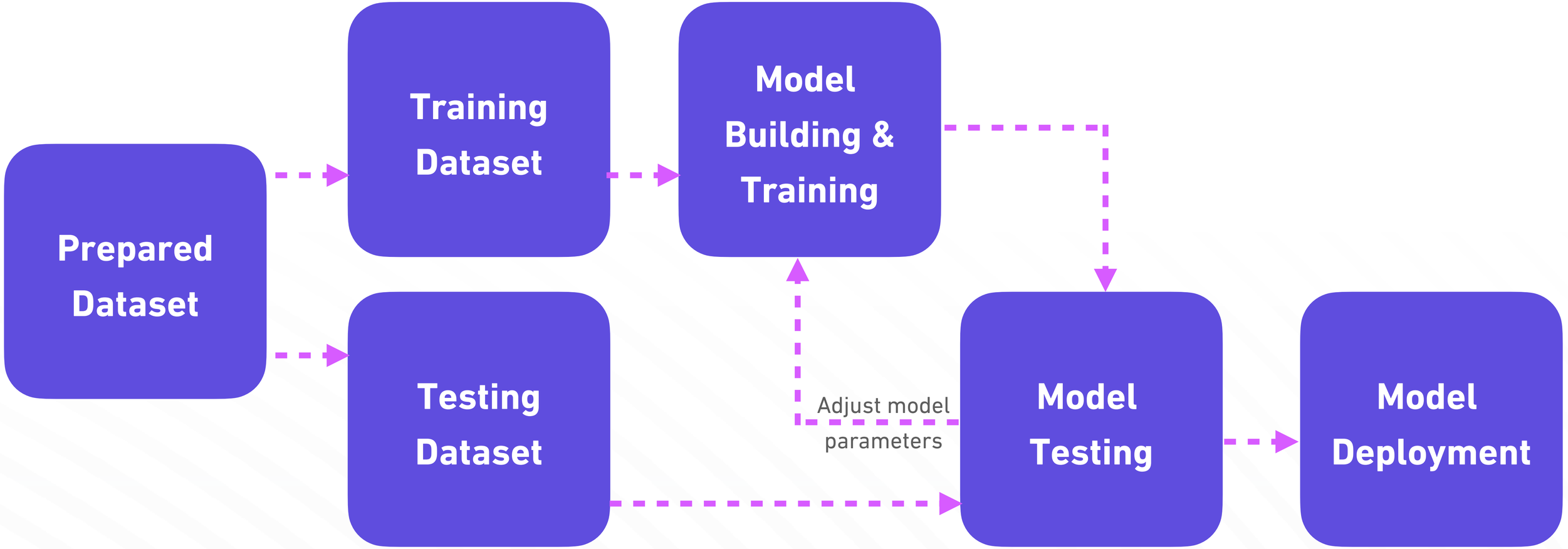
يشير هذا النوع إلى العملية التي تتعلم فيها الآلات اتخاذ قرارات (actions) بناءً على البيئة الخارجية ثم مكافأة الآلة حتى نصل لتحقيق أقصى قدر من هدف معين.

مثال: بناء الرجل الآلي.

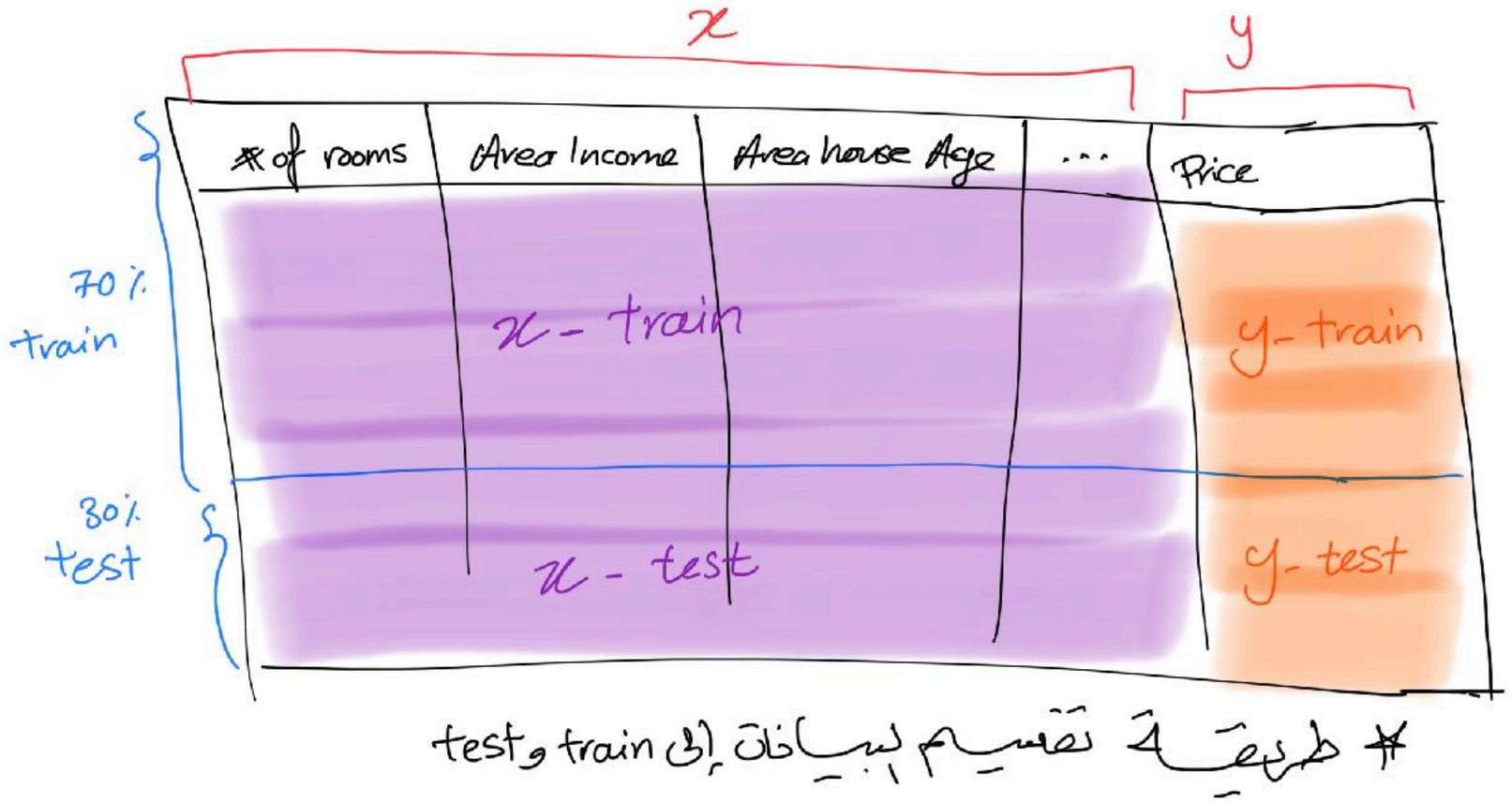
خوارزميات تعلم الآلة (Machine Learning Algorithms)



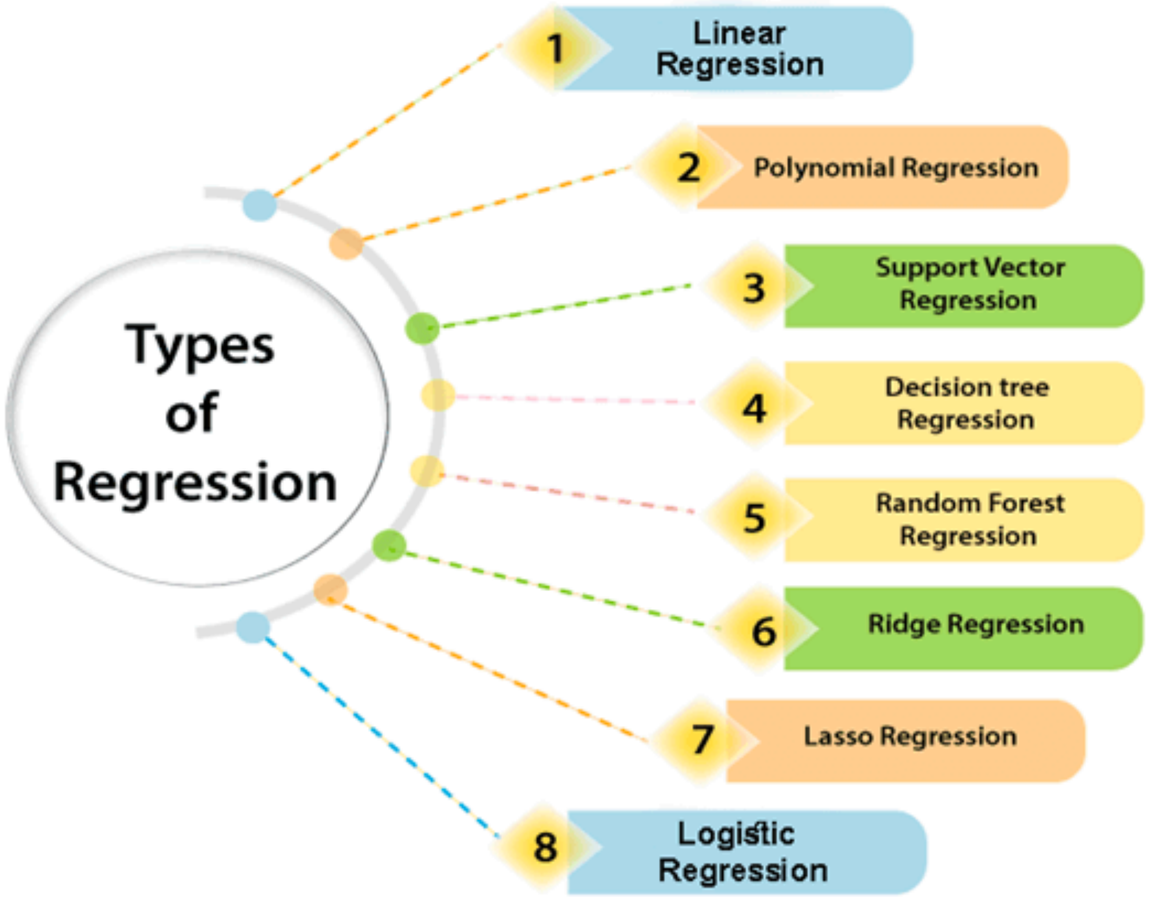
مراحل بناء نماذج تعلم الآلة (Machine Learning Models)



تقسيم البيانات



خوارزميات الانحدار (Regression Algorithms)



خوارزميات الانحدار (Regression Algorithms)

مفهوم Regression Analysis

كيفية عمل تنبؤات حول الكميات الرقمية (quantities)

أمثلة

- كيف يتغير حجم المبيعات عند تغير الأسعار؟
- كيف يتأثر حجم المبيعات بالطقس؟
- كيف يؤثر عنوان الكتاب على مبيعاته؟

خوارزميات الانحدار (Regression Algorithms)

مفهوم Regression Analysis

في جميع الأمثلة السابقة، نقوم بالبحث عن إجابة (**response**) يمكن التعبير عنها بمتغير أو عدة متغيرات مستقلة (**independent variables**) وتسمى (**covariates or predictors**)

$$\mathbf{y} = a_0 + a_1 \mathbf{x}_1$$

intercept

parameters of the model (coefficients)

خوارزميات الانحدار (Regression Algorithms)

مفهوم Regression Analysis

يمكن القول أن Regression Analysis يشير للطريقة التي نبني بها نموذج (model) للتعبير عن العلاقة بين الإجابة (response) وهي تمثل y و مجموعة المتغيرات المستقلة (independent variables) وهي تمثل xi

[Regression Role]

Build a model that can predict response from variables.

خوارزميات الانحدار (Regression Algorithms)

أنواع Regression Analysis

Simple Linear Regression

$$y = b_0 + b_1 x_1$$

Multiple Linear Regression

$$y = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n$$

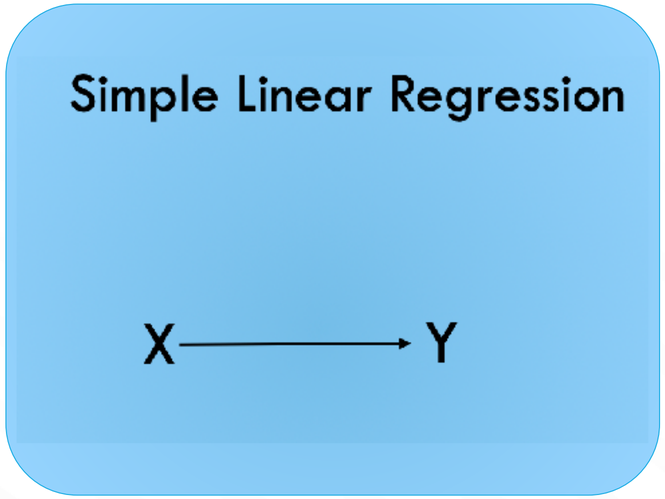
Polynomial Linear Regression

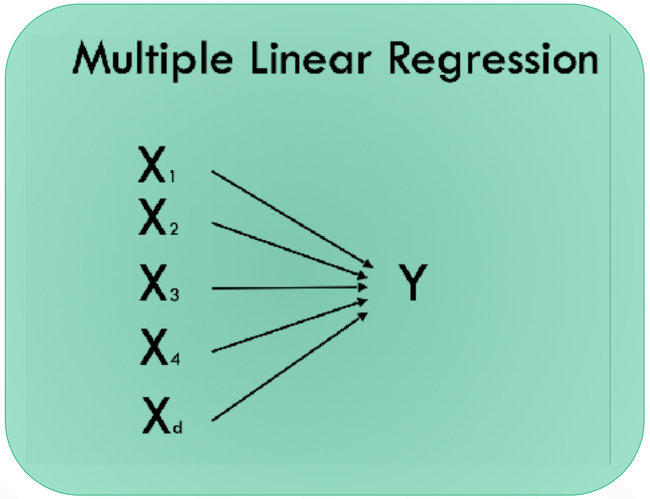
$$y = b_0 + b_1 x_1 + b_2 x_1^2 + \ldots + b_n x_1^n$$

خوارزميات الانحدار (Regression Algorithms)

أنواع Regression Analysis

- يعتبر Multiple Linear Regression امتداد لخوارزمية simple linear regression





$$\mathbf{y} = a_0 + a_1\mathbf{x}_1$$

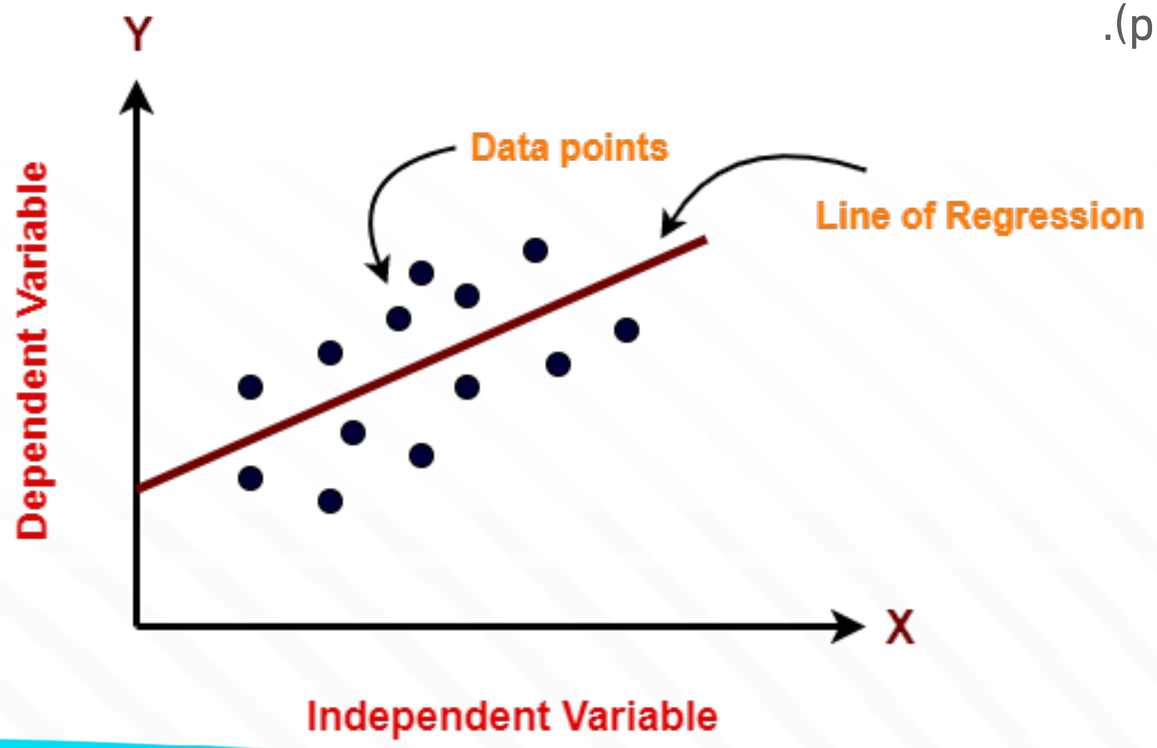
$$\mathbf{y} = a_1\mathbf{x}_1 + \cdots + a_m\mathbf{x}_m$$

خوارزميات الانحدار (Regression Algorithms)

- **لإيجاد (simple linear regression) بين مجموعة البيانات نقوم باستخدام:**

Ordinary least squares (OLS) method

وهي تعتمد على اختيار المتغيرات (a's) التي تقلل من مربع المسافة بين القيم الحقيقية (actual values) والقيم التي يتنبأ بها النموذج (predicted values).



تقييم نماذج الانحدار (Regression Evaluation Metrics)

- النوع الأول: Mean Squared Error (MSE)
- كلما انخفضت قيمة MSE كلما زادت دقة التنبؤ

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

MSE = mean squared error

n = number of data points

Y_i = observed values

$\hat{Y}_i$ = predicted values

تقييم نماذج الانحدار (Regression Evaluation Metrics)

- النوع الثاني: Mean Absolute Error (MAE)

$$\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- النوع الثالث: Root Mean Squared Error (RMSE)

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

تقييم نماذج الانحدار (Regression Evaluation Metrics)

- أفضل قيمة تساوي 1.0
- كلما قلت قيمة R2 يعتبر النموذج غير دقيق
- يمكن أن تكون قيمة R2 سالبة

$$R^2 = 1 - \frac{\sum_{1=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{1=1}^{N} (y_i - \bar{y})^2}$$

Linear Regression

مميزاتها:

- قابلة للتعميم
- مستقرة نتائجها غالبًا
- سهلة التفسير

سلبياتها:

- بطيئة في التكيف مع التغيرات
- تضمن جميع البيانات في نمط واحد
- تعاني من كثرة أبعاد الخواص
- حساسة جدًّا للقيم الشاذة

مثال دارج لتطبيقاتها: توقع الأسعار

Resources

- Practical data science with python [https://learning.oreilly.com/library/view/practical-data-science/9781801071970/].
- Data Science: The Big Picture [https://app.pluralsight.com/library/courses/data-science-big-picture/table-of-contents].
- Big data fundamentals: concepts, drivers & techniques [https://learning.oreilly.com/library/view/big-data-fundamentals/9780134291185/].
- Introduction to Data Science [https://link.springer.com/book/10.1007/978-3-319-50017-1]