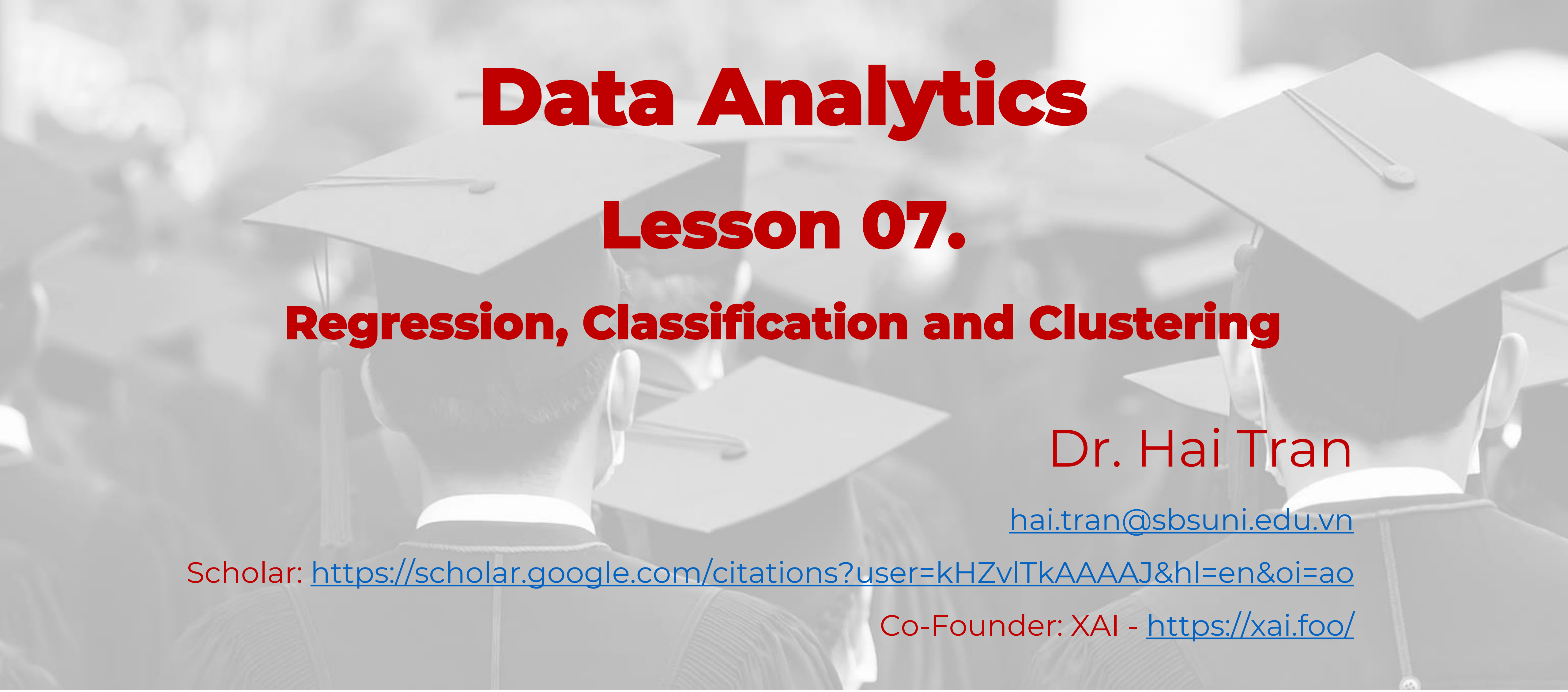


Data Analytics

Lesson 07.

Regression, Classification and Clustering

Dr. Hai Tran

hai.tran@sbsuni.edu.vn

Scholar: https://scholar.google.com/citations?user=kHZvlTkAAAAJ&hl=en&oi=ao

Co-Founder: XAI - https://xai.foo/

Saigon Business School

In partnership with

UNIVERSITY CANADA WEST

MACQUARIE University SYDNEY·AUSTRALIA

University of Northampton

AMITY GLOBAL INSTITUTE

Saigon Business School

Learning materials

- Textbook
  - Evans, J. (2016) Business Analytics. 2nd edn. Pearson.
  - Runkler, T. (2016) Data Analytics: Models and Algorithms for Intelligent Data Analysis. 2nd edn. Vieweg+Teubner Verlag.
- Online reference materials
  - archive.ics.uci.edu/ml/
  - powerbi.microsoft.com
  - https://github.com/topics/data-analysis-python
  - https://media.pearsoncmg.com/ph/esm/esm_evans_eba3e_20/tools/eba3e_analytic_solver.html
  - https://data.imf.org/

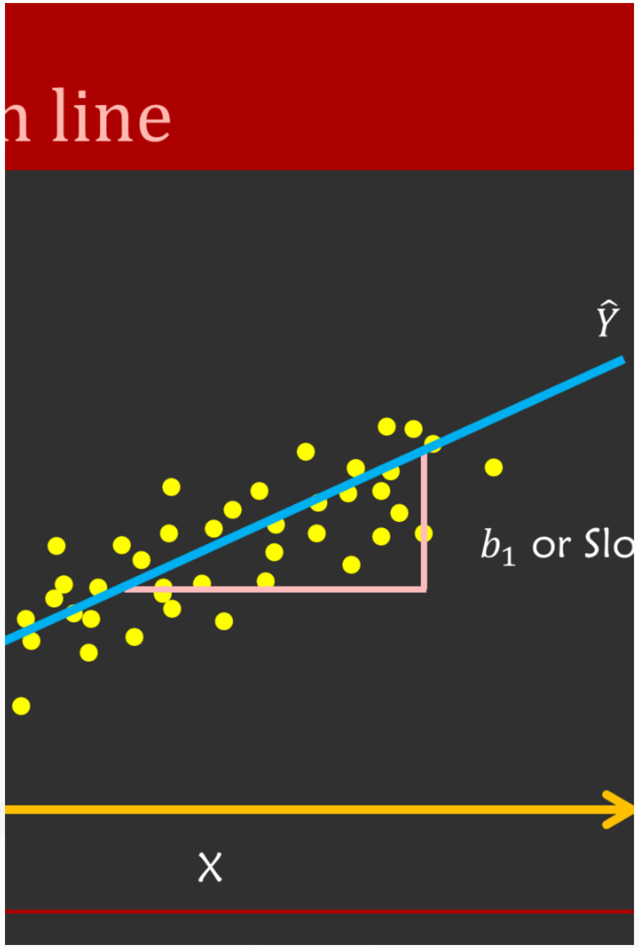
Saigon Business School

Agenda

- Lesson 1: Understanding Data Analytics Terminologies.
- Lesson 2: Foundation of Business Analytics
- Lesson 3: Visualizing and Exploring data
- Lesson 4: Applying Descriptive Analytic Techniques
- Lesson 5: Data Modeling
- Lesson 6: Predictive Analytics
- Lesson 7: Regression, Classification and Clustering
- Lesson 8: Forecasting Techniques
- Lesson 9: Investigating Predictive Analytic Techniques
- Lesson 10: Introduction to Data Mining
- Lesson 11: Demonstrating Prescriptive Analytic Methods
- Lesson 12: Recap and advanced topics

Saigon Business School

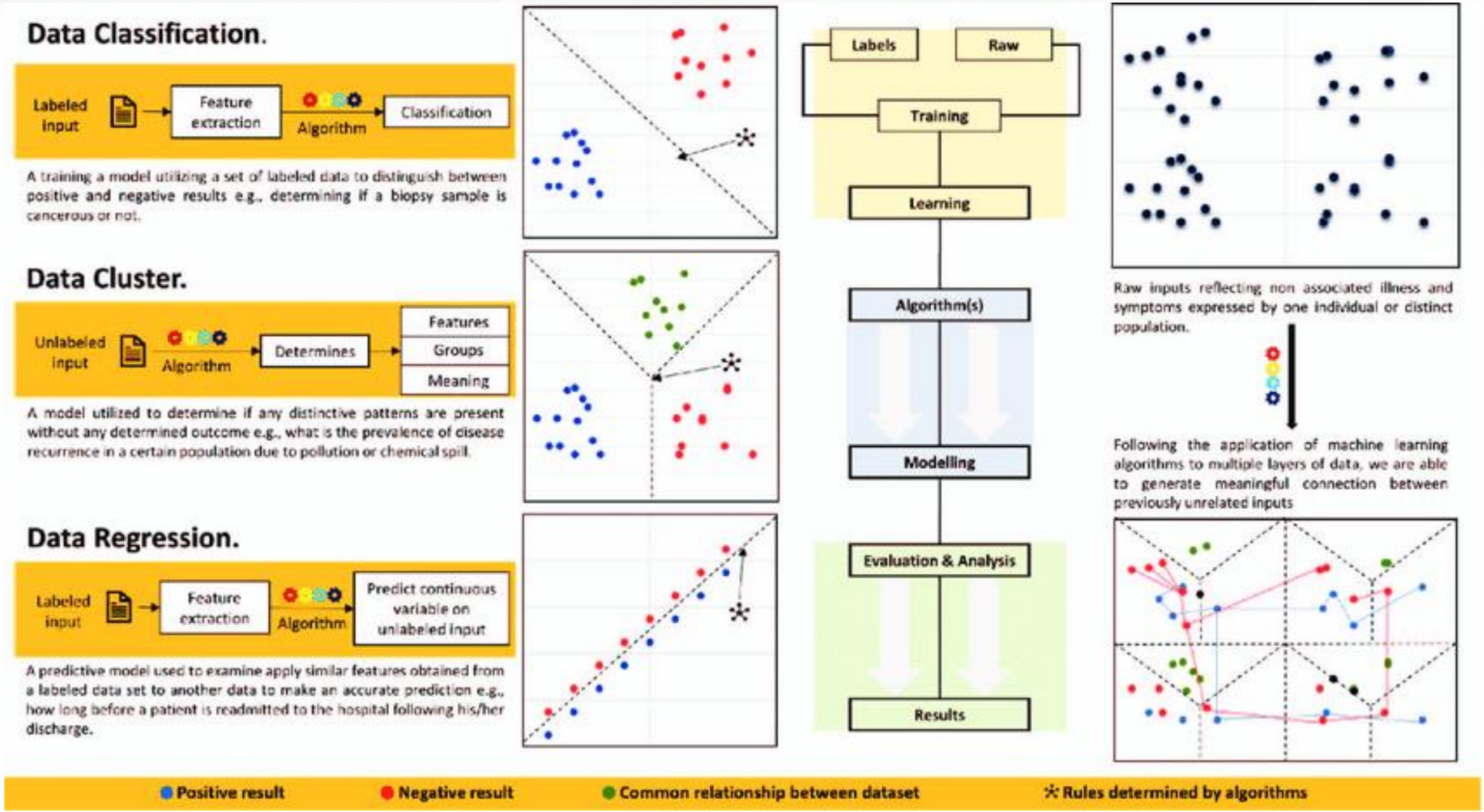
Regression, Classification and Clustering



In this presentation, we will explore the fascinating world of Regression, Classification, and Clustering. Get ready to dive into the depths of data analysis and uncover the power of these techniques!

Binary classification	Multiclass classification	Regression	Clustering
Supervised learning technique	Supervised learning technique	Supervised learning technique	Unsupervised learning technique
Target variable can take only two categorical values as only two target categories (classes) exist.	Target variable can take any one of the multiple categorical values as multiple target categories (classes) exist.	Target variable can take any one of the infinite within a range.	The output variables are not given to us. We try to cluster the given data into clusters and extract useful information out of it.
Output variable-Discrete	Output variable-discrete	Output variable-continuous	Output variable-not given
Example:-Given 1000 images of bananas and apples classified as bananas and apples respectively. Classify 10 images whether they are of a banana or apple.	Example:- Given 1000 images of bananas ,apples and potatoes classified as bananas, apples and potatoes respectively. Classify 10 images whether they are of a banana, apple or potato.	Example:-Given the weights and heights of 1000 people. Predict the weights of 20 people whose height is known.	Given images related to basic shapes like circles, rectangles, triangles etc find clusters, useful patterns amongst them.

Regression, Classification and Clustering



Saigon Business School

Regression, Classification and Clustering

Regression

1 **Predictive Modeling** 📊

Regression allows us to predict continuous outcomes based on input variables, making it invaluable for forecasting and trend analysis.

2 **Linear Relationships** 📈

We'll delve into linear regression, where we analyze the relationships between variables and fit a line to our data to make accurate predictions.

3 **Examining Residuals** 🔎

We'll also explore the concept of residuals to assess the quality of our model and identify any patterns or biases that may exist.

Saigon Business School

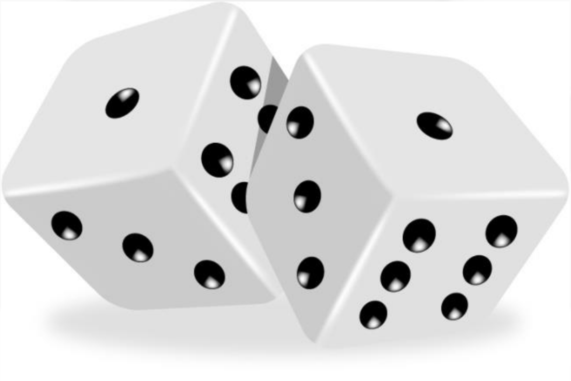
Regression, Classification and Clustering

Classification

Approach	Key Features	Applications
Supervised Learning 🎯	Classifies data based on labeled examples, making it useful for spam detection, image recognition, and sentiment analysis.	Social media monitoring, fraud detection, medical diagnosis
Unsupervised Learning 🤔	Finds patterns and similarities in data without any prior labels, enabling clustering and anomaly detection.	Market segmentation, customer profiling, recommendation systems

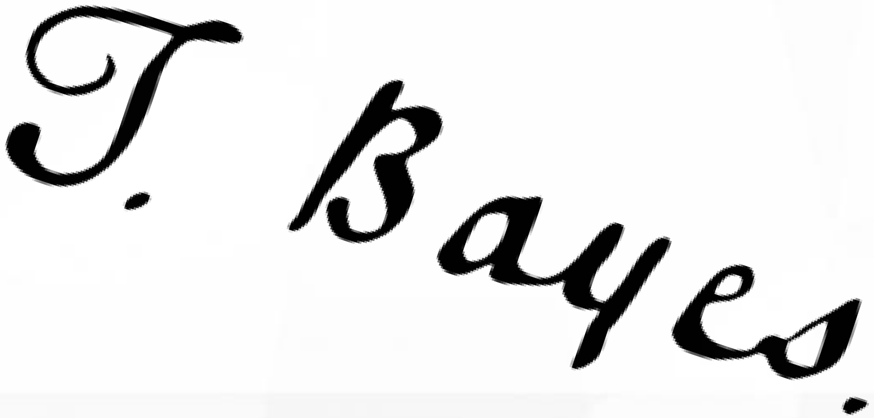
Saigon Business School

Thomas Bayes
(1702 – 1761)



Probability Bayes:

$$\boxed{P(H\,|\,X) = \frac{P(X\,|\,H)P(H)}{P(X)}}$$



Saigon Business School

Example: Consider a training dataset consisting of classified datasets as follows:

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Apply Bayes classification to predict the class of the dataset.

$$X = (age = youth,\ income = medium,\ student = yes,\ credit\_rating = fair)$$

There are 02 classes of data corresponding to buys_computer = yes and buys_computer = no.

Saigon Business School

X' =(senior, high, no, excellent)

$$P(buys\_computer = yes) = 9/14 = 0.643$$
$$P(buys\_computer = no) = 5/14 = 0.357$$

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

$$P(age = youth \mid buys\_computer = yes) = 2/9 = 0.222$$
$$P(age = youth \mid buys\_computer = no) = 3/5 = 0.600$$
$$P(income = medium \mid buys\_computer = yes) = 4/9 = 0.444$$
$$P(income = medium \mid buys\_computer = no) = 2/5 = 0.400$$
$$P(student = yes \mid buys\_computer = yes) = 6/9 = 0.667$$
$$P(student = yes \mid buys\_computer = no) = 1/5 = 0.200$$
$$P(credit\_rating = fair \mid buys\_computer = yes) = 6/9 = 0.667$$
$$P(credit\_rating = fair \mid buys\_computer = no) = 2/5 = 0.400$$

Consequently:

$$P(X \mid buys\_computer = yes) = P(age = youth \mid buys\_computer = yes) \times P(income = medium \mid buys\_computer = yes) \times P(student = yes \mid buys\_computer = yes) \times P(credit\_rating = fair \mid buys\_computer = yes) = 0.222 \times 0.444 \times 0.667 \times 0.667 = 0.044.$$

Likewise

$$P(X \mid buys\_computer = no) = 0.600 \times 0.400 \times 0.200 \times 0.400 = 0.019$$
$$P(X \mid buys\_computer = yes)P(buys\_computer = yes) = 0.044 \times 0.643 = 0.028$$
$$P(X \mid buys\_computer = no)P(buys\_computer = no) = 0.019 \times 0.357 = 0.007$$

⇒X belonging to the class of data corresponding to buys_computer = yes

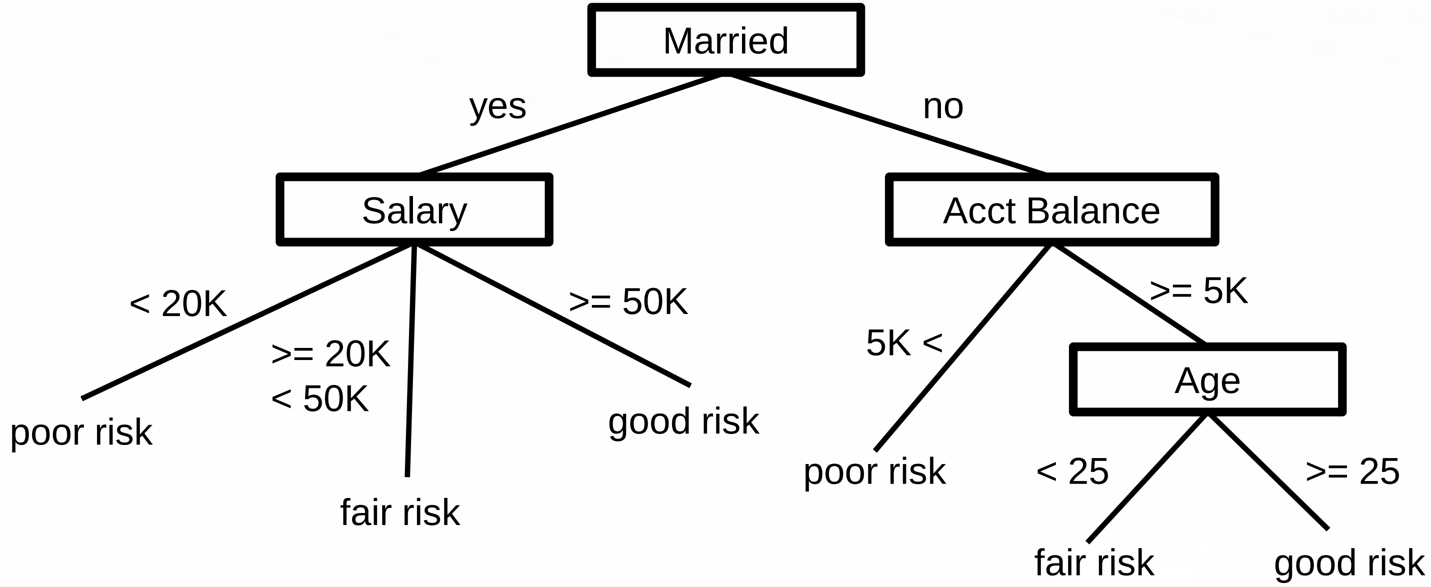
Saigon Business School

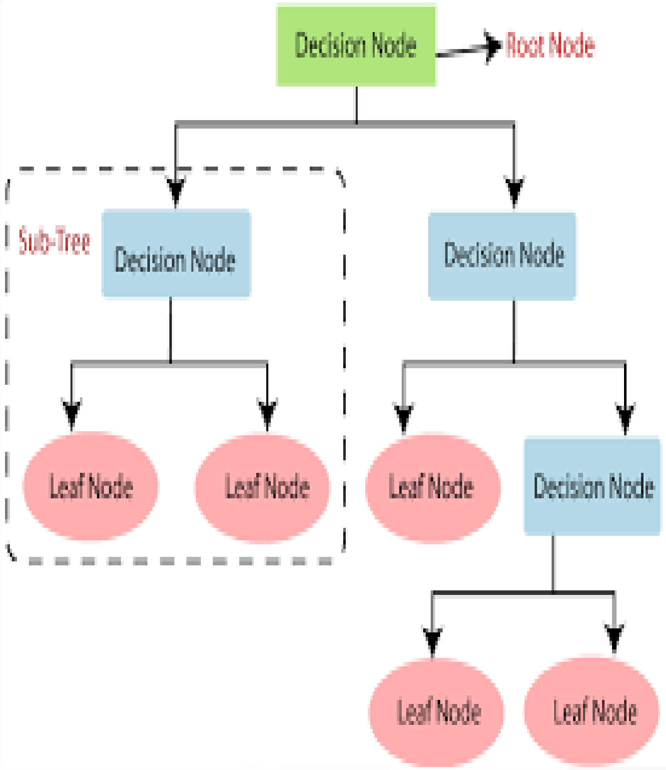
Exerciser

- X =(senior, high, no, excellent) belonging to the class of data corresponding to what?

RID	age	income	student	credit_rating	Class: buys_computer
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

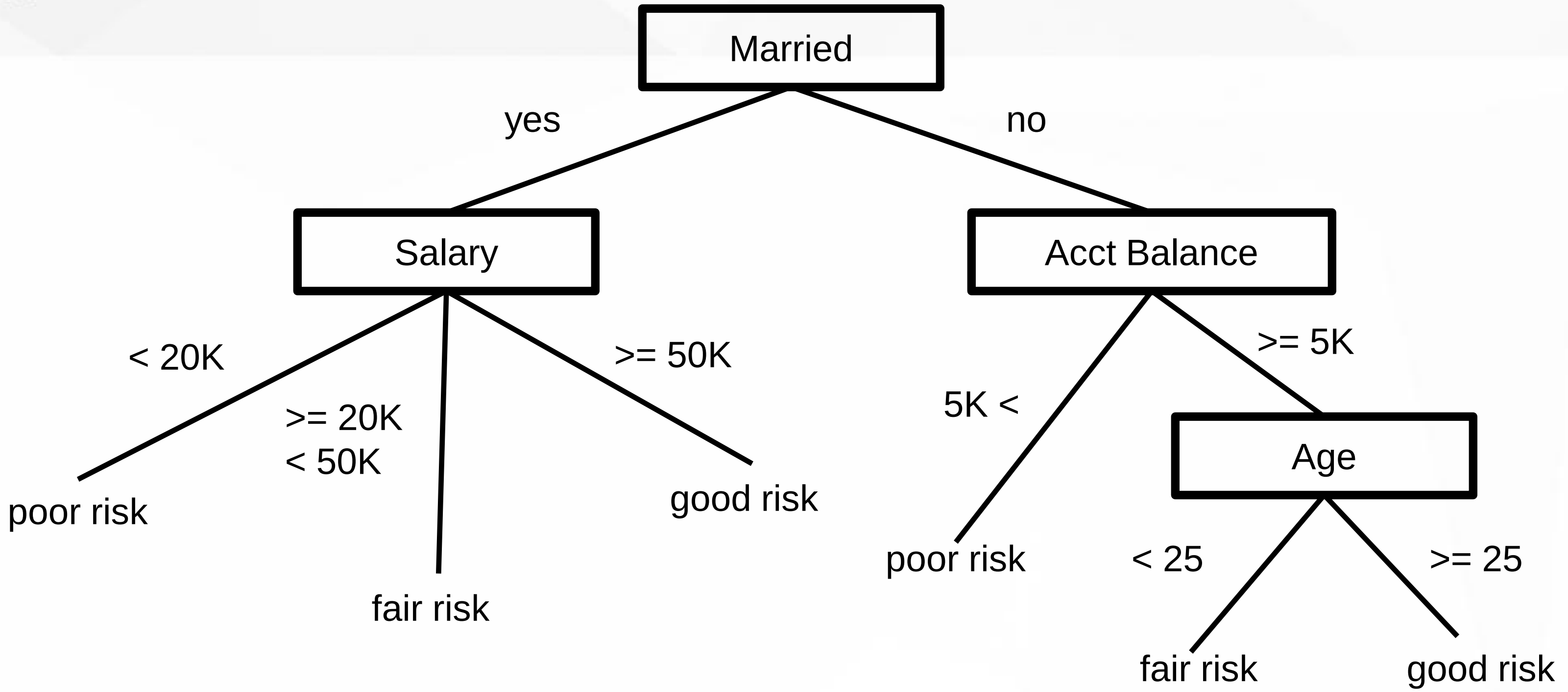
CLASSIFICATION BASED ON DECISION TREE



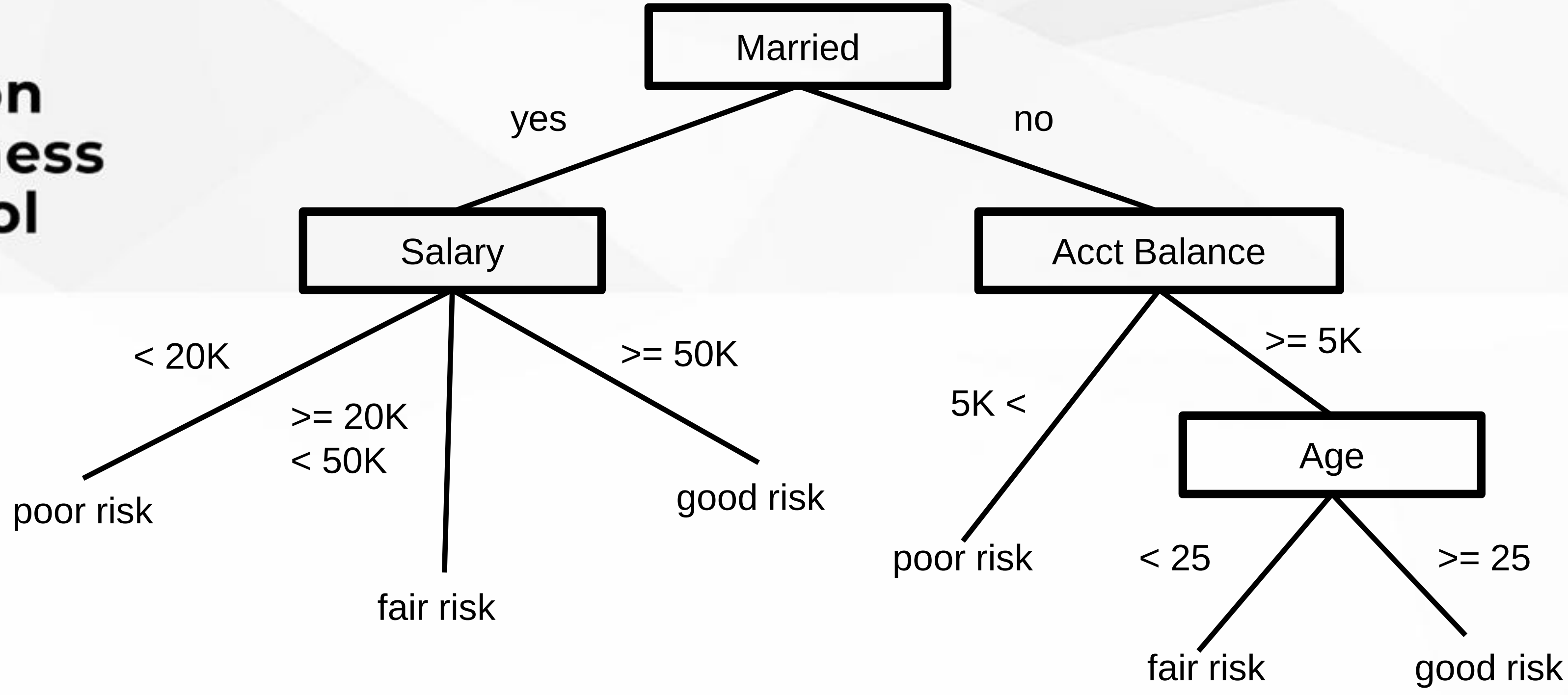


sbsedu.vn

CLASSIFICATION BASED ON DECISION TREE



1. **If** (Married = yes) And (Salary > 20K) **Then** Class = poor risk
2. **If** (Married = yes) And (50K > Salary >= 20K) **Then** Class = fair risk
3. **If** (Married = yes) And (Salary >= 50K) **Then** Class = good risk
4. **If** (Married = no) And (Acct Balance < 5K) **Then** Class = poor risk
5. **If** (Married = no) And (Acct Balance >= 5K) And (Age < 25) **Then** Class = fair risk
6. **If** (Married = no) And (Acct Balance >= 5K) And (Age >= 25) **Then** Class = good risk



Name	Age	Married	Salary	Acct Balance	Class
Alice	19	yes	30K	6K	?
Pike	28	no	60K	7K	?
Tom	35	yes	10K	10K	?
Peter	24	no	20K	8K	?
Lucas	40	no	20K	3K	?

Name	Age	Married	Salary	Acct Balance	Class
Alice	19	yes	30K	6K	fair risk
Pike	28	no	60K	7K	good risk
Tom	35	yes	10K	10K	poor risk
Peter	24	no	20K	8K	fair risk
Lucas	40	no	20K	3K	poor risk

SBS
Saigon Business School

sbsedu.vn

E	X	A	Q	C	H	O	G	Q	O	G	X	T	Z	T	A	Q	N	P	J	N	N	Z	U
R	Y	W	K	W	J	P	C	X	J	P	S	Y	Z	J	C	E	Z	I	N	A	G	R	O
O	L	S	U	Q	W	K	A	D	D	F	V	L	T	A	L	L	Y	V	N	Y	F	W	F
W	D	I	P	M	Y	Y	H	R	L	R	S	N	I	Z	O	C	Y	S	J	K	A	A	T
R	J	D	P	I	S	X	G	B	G	U	N	R	D	L	M	U	R	I	Y	O	E	Z	W
D	Q	W	Y	M	C	T	N	E	M	E	G	A	N	A	M	I	H	J	A	Y	R	X	N
L	J	J	R	P	I	T	V	A	R	R	L	C	D	A	Z	I	S	Y	A	U	K	U	E
G	A	S	K	P	G	Q	O	E	P	O	R	C	A	B	V	J	F	D	R	E	H	T	O
C	A	B	C	T	V	D	S	G	Z	L	S	Z	R	F	V	D	U	K	M	O	S	V	Z
Z	S	I	E	F	U	E	F	S	R	S	W	B	Y	I	S	G	S	Q	F	N	X	Z	V
Z	Y	B	O	L	Y	Q	W	N	J	A	B	G	Q	U	C	Y	K	W	V	I	A	O	J
Q	O	D	J	P	S	V	J	M	O	U	P	M	G	D	A	T	A	C	V	J	I	T	Y
I	X	J	T	N	T	D	B	U	F	O	P	H	I	N	F	O	R	M	A	T	I	O	N
R	B	F	Y	W	O	I	Q	S	L	R	D	U	Q	K	I	S	A	F	R	G	Y	A	F
C	G	U	T	P	S	P	O	N	O	C	A	H	N	L	E	G	Y	K	R	P	L	C	Z
H	R	V	K	P	T	G	L	N	V	J	L	F	I	N	J	F	E	V	Q	I	Q	H	O
O	I	H	M	P	C	J	O	N	S	I	Q	J	A	C	P	M	V	G	K	H	T	F	Y
I	Q	L	Z	M	B	P	H	S	N	M	I	X	B	A	R	G	R	A	P	H	I	G	N
C	V	X	R	I	W	P	Q	E	O	U	R	E	K	I	Q	G	U	U	T	D	T	U	W
E	K	E	E	B	A	A	P	X	M	Q	B	L	X	Y	I	C	S	P	Y	B	L	M	F
S	W	Z	Q	R	Q	L	E	T	L	U	I	B	C	H	A	R	T	W	Y	G	E	B	W
Z	I	T	G	L	O	C	B	J	Z	G	L	C	S	J	F	D	O	X	S	C	H	E	X
N	T	I	N	T	P	V	K	O	L	Z	M	D	P	S	D	H	F	J	P	W	B	L	B
S	O	V	X	B	H	C	Q	G	U	S	Q	U	E	S	T	I	O	N	S	N	Q	S	O

Saigon Business School

```python
# Import necessary libraries
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report
from sklearn.preprocessing import LabelEncoder
```

1

```python
# Encode categorical variables
label_encoder = LabelEncoder()
titanic_data['Sex'] = label_encoder.fit_transform(titanic_data['Sex'])
titanic_data['Embarked'] = label_encoder.fit_transform(titanic_data['Embarked'])
# Split the data into features and target variable
X = titanic_data[['Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare', 'Embarked']]
y = titanic_data['Survived']
```

2

```python
# Make predictions on the test set
predictions = model.predict(X_test)
# Evaluate the model
accuracy = accuracy_score(y_test, predictions)
print(f'Accuracy: {accuracy:.2f}')
# Display classification report
print('Classification Report:')
print(classification_report(y_test, predictions))
```

3

```python
# Load the Titanic dataset
titanic_data = pd.read_csv('titanic.csv')

# Preprocess the data
# Drop unnecessary columns or fill missing values as needed
titanic_data = titanic_data[['Pclass', 'Sex', 'Age', 'SibSp', 'Parch', 'Fare', 'Embarked', 'Survived']]
titanic_data = titanic_data.dropna()
```

4

```python
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
# Create a logistic regression model
model = LogisticRegression()
# Train the model
model.fit(X_train, y_train)
```

5

Saigon Business School

Regression, Classification and Clustering


```
# Import necessary libraries
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score,
classification_report
from sklearn.preprocessing import LabelEncoder

# Load the Titanic dataset
titanic_data = pd.read_csv('titanic.csv')

# Preprocess the data
# Drop unnecessary columns or fill missing values as
needed
titanic_data = titanic_data[['Pclass', 'Sex', 'Age',
'SibSp', 'Parch', 'Fare', 'Embarked', 'Survived']]
titanic_data = titanic_data.dropna()

# Encode categorical variables
label_encoder = LabelEncoder()
titanic_data['Sex'] =
label_encoder.fit_transform(titanic_data['Sex'])
titanic_data['Embarked'] =
label_encoder.fit_transform(titanic_data['Embarked'])
```

Saigon Business School

Regression, Classification and Clustering

Clustering

1. **K-means Clustering**

   One of the most widely used clustering algorithms, K-means partitions data into K clusters based on distance measurements.

2. **Hierarchical Clustering**

   This approach creates a tree-like structure to represent relationships between data points, making it useful for visualizing hierarchical structures.

3. **DBSCAN**

   A density-based algorithm that groups together data points based on their density in a given region, as opposed to distance.

Regression, Classification and Clustering

Differences between Regression, Classification, and Clustering

Purpose

Regression predicts continuous outcomes, classification categorizes data, and clustering identifies inherent groupings.

Input

Regression and classification rely on labeled data, while clustering does not require any pre-existing labels.

Output

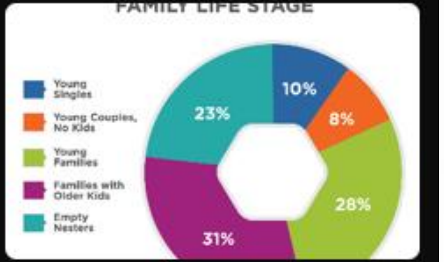
Regression and classification provide specific predictions or classes, while clustering simply groups similar instances.

Applications

Regression is used for forecasting, classification for image recognition, and clustering for customer segmentation.

Regression, Classification and Clustering

Applications of Regression, Classification, and Clustering



Predictive Maintenance 💡

Regression analysis can help anticipate when machines or infrastructure require maintenance, reducing downtime and costs.

Medical Diagnostics 🩺

Classification algorithms help doctors make accurate disease diagnoses based on symptoms, medical history, and test results.

Marketing Strategy 🎯

Clustering allows businesses to businesses to identify target target markets and tailor marketing campaigns to specific specific customer groups.

Regression, Classification and Clustering

- In this example, we use features like 'pclass', 'fare', 'survived', and 'sex' to predict the 'fare' (ticket fare) of passengers. The model is evaluated using metrics such as Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error. The scatter plot visually represents the relationship between actual and predicted fare values.

- https://colab.research.google.com/drive/13yN08T4WQZfVtVez4W1GgDB9Cz_yOdUA?usp=sharing

```
import pandas as pd
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import metrics
import matplotlib.pyplot as plt
import numpy as np  # Add this line to import NumPy

# Load the Titanic dataset
titanic = sns.load_dataset('titanic')

# Let's consider a subset of features for simplicity
features = ['pclass', 'fare', 'survived', 'sex']

# Drop rows with missing values for simplicity in this
example
titanic_subset = titanic[features].dropna()

# Convert categorical variable 'sex' to numerical (0
for male, 1 for female)
titanic_subset['sex'] =
titanic_subset['sex'].map({'male': 0, 'female': 1})
```

Google classroom task

- #MachineLearning #clustering vs classification concept **CLUSTERING vs CLASSIFICATION**
- https://www.youtube.com/watch?v=BgJewx3bC5g.
  - Watch and investigate.
  - Submit your answer:
    - YouTube Link
    - Compare classification to clustering in table format.

Regression, Classification and Clustering

1. What is the primary objective of regression analysis? A. Classification of data B. Prediction of numerical outcomes C. Grouping similar data points D. Identifying outliers in the dataset
2. In a simple linear regression, what is the role of the dependent variable? A. To be predicted based on the independent variable B. To remain constant throughout the analysis C. To be controlled by the researcher D. To represent the categorical data in the model
3. What does the term "residuals" refer to in regression analysis? A. Predicted values in the model B. The difference between observed and predicted values C. Independent variables in the model D. Outliers in the dataset
4. In multiple regression, how many independent variables are considered? A. One B. Two C. More than two D. None
5. What does the coefficient of determination (R-squared) indicate in regression analysis? A. The slope of the regression line B. The strength of the relationship between variables C. The number of independent variables D. The standard deviation of the residuals
6. What is the primary goal of classification algorithms? A. Predicting numerical values B. Grouping similar data points C. Identifying outliers in the dataset D. Estimating correlation coefficients
7. Which algorithm is commonly used for binary classification problems? A. Decision Trees B. K-Means C. Principal Component Analysis (PCA) D. Linear Regression
8. Question: What is the purpose of a confusion matrix in classification? A. Evaluating the performance of a classification model B. Identifying outliers in the dataset C. Grouping similar data points D. Predicting numerical outcomes
9. Question: What is the main objective of clustering analysis? A. Predicting numerical values B. Grouping similar data points based on similarity C. Estimating correlation coefficients D. Identifying outliers in the dataset
10. Question: Which algorithm is commonly used for hierarchical clustering? A. K-Means B. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) C. Agglomerative Hierarchical Clustering D. Support Vector Machines (SVM)

Saigon Business School

Learning Mission

30 mins

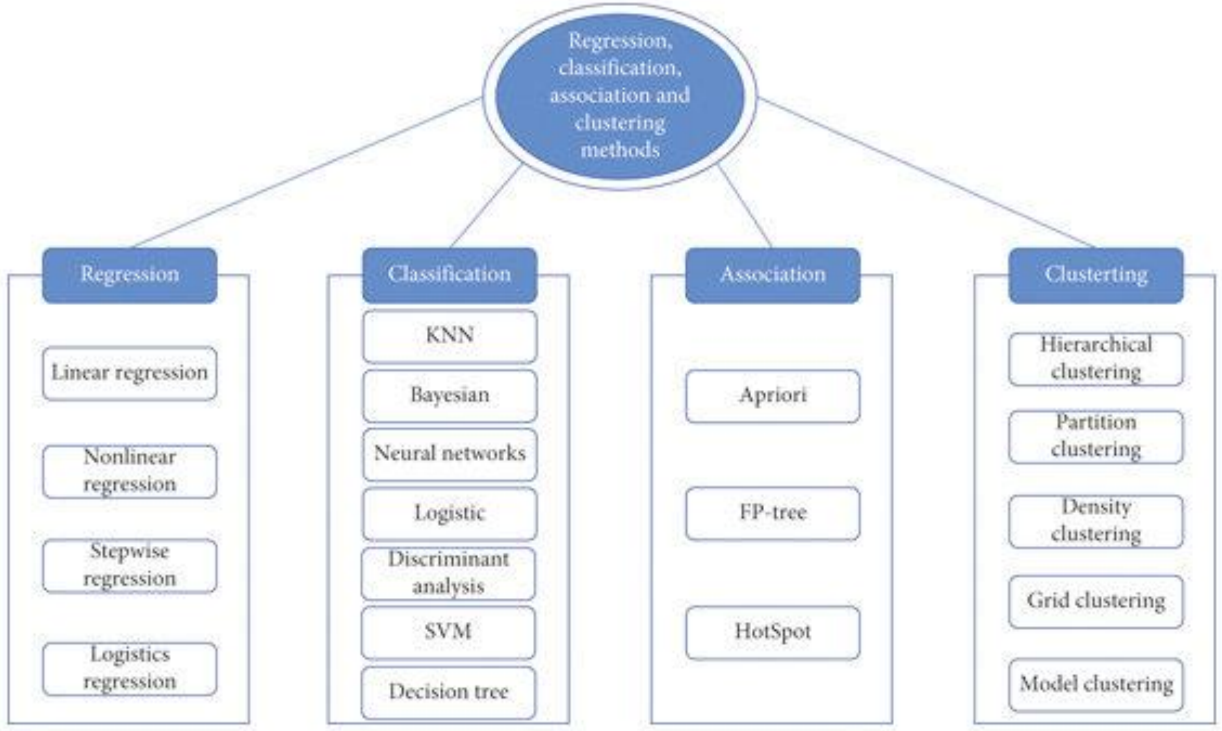
Reading paper 'Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine'Discussion and answer:

- Search this paper on https://scholar.google.com/.
- How to Regression apply in healthcare.
- How to Classification apply in healthcare.
- How to Clustering apply in healthcare.
- List down some names of tech in Data Analytics bale to apply for healthcare.

Ahmed, Z., Mohamed, K., Zeeshan, S., & Dong, X. (2020). Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine. *Database*, *2020*, baaa010.

Conclusion and Questions

- Regression, Classification, and Clustering are indispensable tools in the world of data analysis and machine learning. They offer unique ways to extract insights from data, make predictions, and enhance decision-making. Understanding their differences, applications, and limitations is crucial to leveraging their power effectively.



Saigon Business School

Thank you

sbsedu.vn

TP, FP, FN, TN

1. **True Positive (TP):**
   Definition: The model correctly predicted instances of the positive class.
2. **False Positive (FP):**
   Definition: The model incorrectly predicted instances of the positive class when they actually belong to the negative class.
3. **False Negative (FN):**
   Definition: The model incorrectly predicted instances of the negative class when they actually belong to the positive class.
4. **True Negative (TN):**
   Definition: The model correctly predicted instances of the negative class.

Saigon Business School

TP, FP, FN, TN

- These terms are often used to calculate various metrics that help assess the performance of a binary classification model, such as accuracy, precision, recall, and F1 score.

  - **Accuracy:** $\frac{TP+TN}{TP+TN+FP+FN}$
  - **Precision:** $\frac{TP}{TP+FP}$
  - **Recall (Sensitivity or True Positive Rate):** $\frac{TP}{TP+FN}$
  - **F1 Score:** $2 \times \frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$

Saigon Business School

TP, FP, FN, TN

- Suppose you have a binary classification model for detecting whether an email is spam (positive) or not spam (negative). Let's assume you have the following results based on 100 emails:
- True Positives (TP): 25 emails
- False Positives (FP): 10 emails
- False Negatives (FN): 5 emails
- True Negatives (TN): 60 emails

Accuracy:

Accuracy =

Accuracy =

Precision:

Precision =

Precision =

Recall (Sensi

Recall =

Recall =

F1 Score:

F1 Score =

F1 Score =

Scale multiple variables

Saigon Business School

- When your data has different values, and even different measurement units, it can be difficult to compare them. What is kilograms compared to meters? Or altitude compared to time?
- The answer to this problem is scaling. We can scale data into new values that are easier to compare.
- It can be difficult to compare the volume 1.0 with the weight 790, but if we scale them both into comparable values, we can easily see how much one value is compared to the other.
- There are different methods for scaling data, in this tutorial we will use a method called standardization.
- The standardization method uses this formula:
- $z = (x - u) / s$

Car	Model	Volume	Weight	CO2
Toyota	Aygo	1.0	790	99
Mitsubishi	Space Star	1.2	1160	95
Skoda	Citigo	1.0	929	95
Fiat	500	0.9	865	90
Mini	Cooper	1.5	1140	105
VW	Up!	1.0	929	105
Skoda	Fabia	1.4	1109	90
Mercedes	A-Class	1.5	1365	92
Ford	Fiesta	1.5	1112	98
Audi	A1	1.6	1150	99