

INTRODUCTION TO MACHINE LEARNING

Introduction to Data Science

Author: Eng. Carlos Andrés Sierra, M.Sc.
carlos.andres.sierra.v@gmail.com

Lecturer
Computer Engineer
School of Engineering
Universidad Distrital Francisco José de Caldas

2024-II



Outline

- 1 Fundamentals of Machine Learning
- 2 Supervised Machine Learning
- 3 Supervised Machine Learning Algorithms
- 4 Machine Learning Models Evaluation



Outline

- 1 Fundamentals of Machine Learning
- 2 Supervised Machine Learning
- 3 Supervised Machine Learning Algorithms
- 4 Machine Learning Models Evaluation

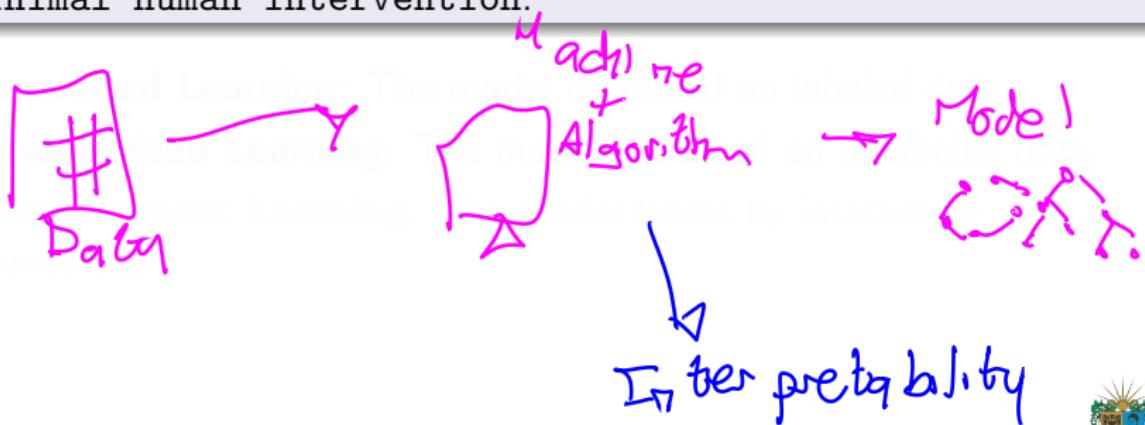


Key Concepts in Machine Learning

4

Machine Learning

- **Machine learning** is a method of data analysis that **automates** analytical model building.
- It is a **branch** of **artificial intelligence** based on the idea that systems can **learn from data**, identify patterns and **make decisions** with minimal human intervention.



Key Concepts in Machine Learning

Machine Learning

- **Machine learning** is a method of data analysis that **automates** analytical model building.
- It is a **branch** of **artificial intelligence** based on the idea that systems can **learn from data**, identify patterns and **make decisions** with minimal human intervention.
- **Supervised Learning:** The **model** is trained on **labeled data**.



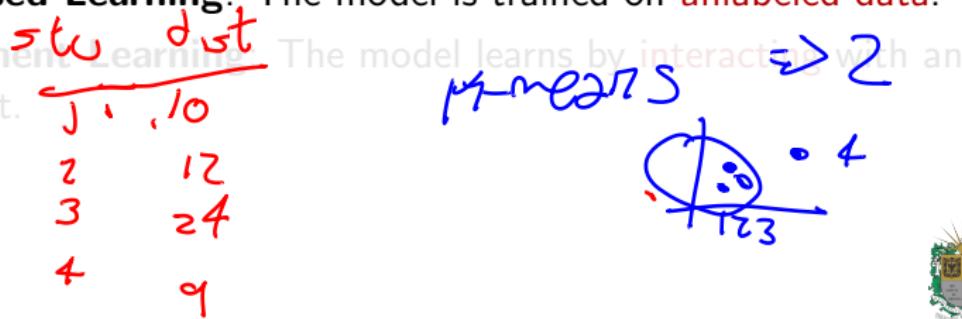
*Label
Person*



Key Concepts in Machine Learning

Machine Learning

- **Machine learning** is a method of data analysis that **automates** analytical model building.
- It is a **branch** of **artificial intelligence** based on the idea that systems can **learn from data**, **identify patterns** and **make decisions** with minimal human intervention.
- **Supervised Learning:** The model is trained on **labeled data**.
- **Unsupervised Learning:** The model is trained on **unlabeled data**.
- **Reinforcement Learning:** The model learns by interacting with an environment.



Key Concepts in Machine Learning

Machine Learning

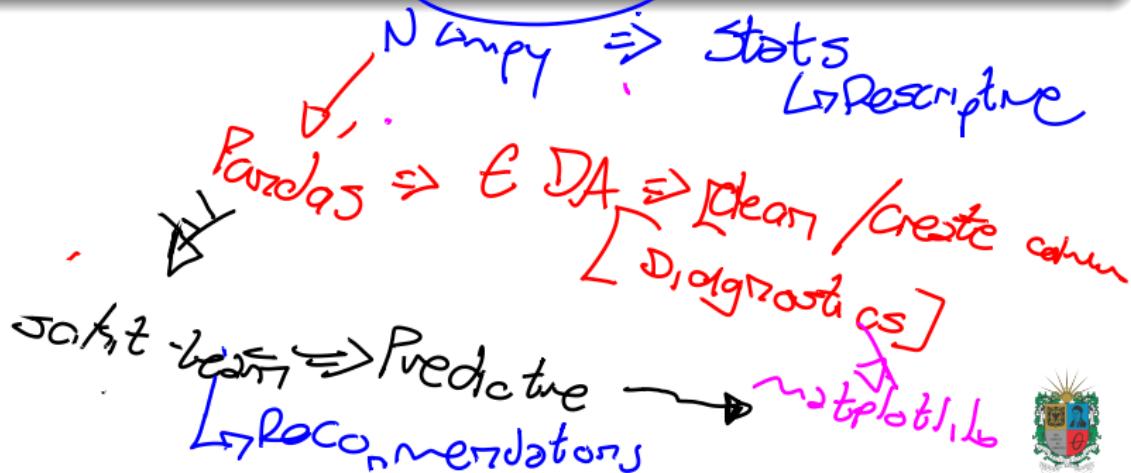
- **Machine learning** is a method of data analysis that **automates** analytical model building.
- It is a **branch** of **artificial intelligence** based on the idea that systems can **learn from data**, identify patterns and **make decisions** with minimal human intervention.
- **Supervised Learning**: The model is trained on **labeled data**.
- **Unsupervised Learning**: The model is trained on **unlabeled data**.
- **Reinforcement Learning**: The model learns by **interacting** with an environment.



Python Tools for Machine Learning

Python Tools

- **NumPy**: A library for **numerical computing**.
- **Pandas**: A library for **data manipulation** and analysis.
- **Matplotlib**: A library for **data visualization**.
- **Scikit-learn**: A library for **machine learning**.



Typical Machine Learning Problems

- **Classification:** Predicting a **label**.



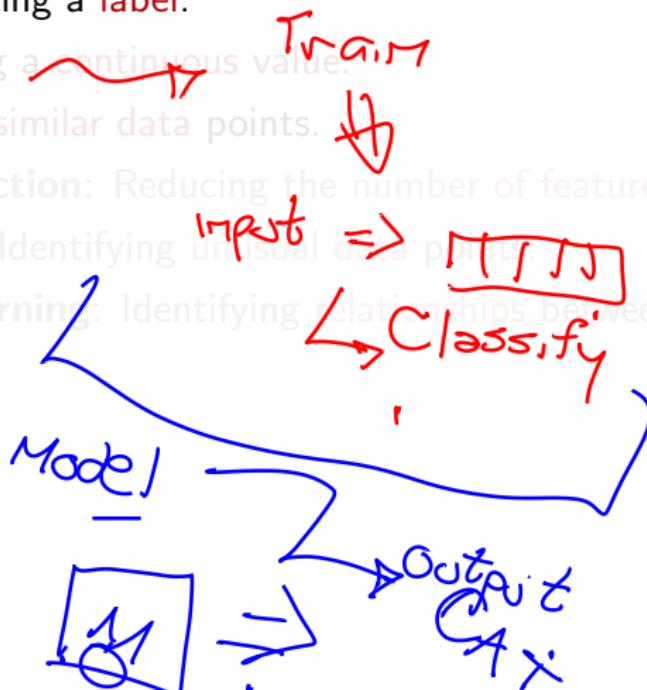
- **Regression:** Predicting a continuous value.

- **Clustering:** Grouping similar data points.

- **Dimensionality Reduction:** Reducing the number of features.

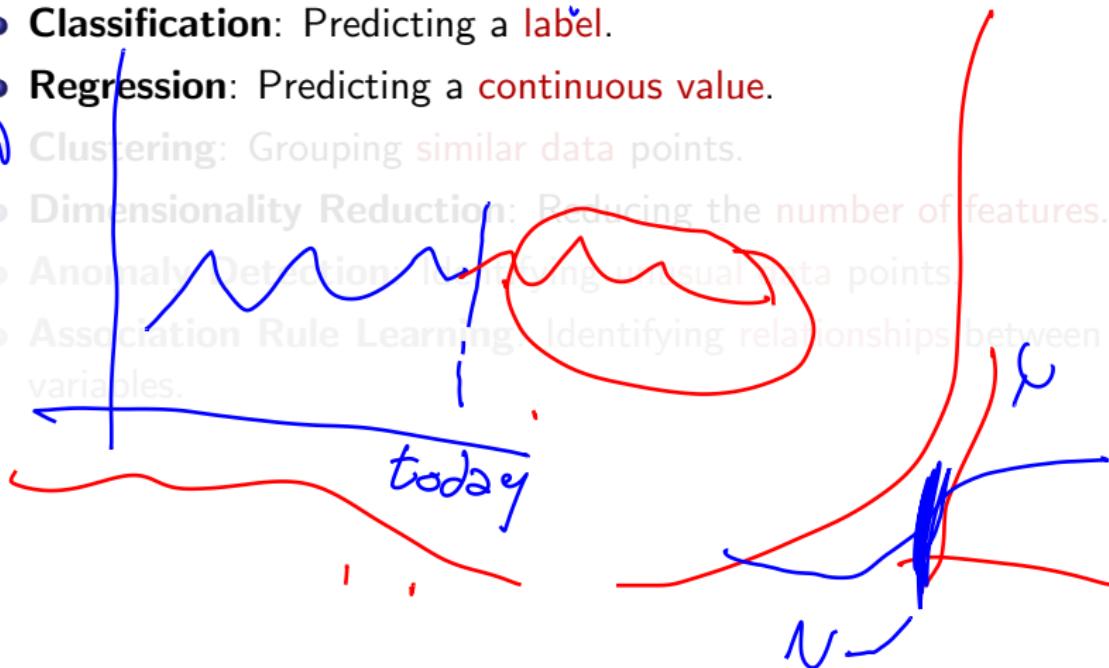
- **Anomaly Detection:** Identifying unusual data points.

- **Association Rule Learning:** Identifying relationships between variables.



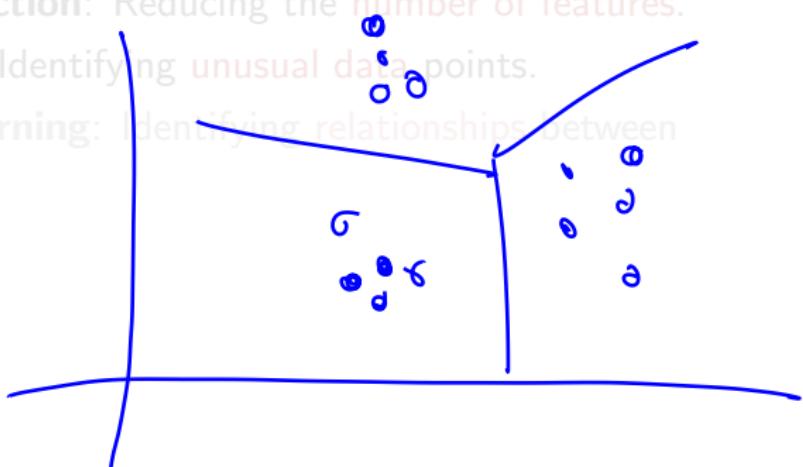
Typical Machine Learning Problems

- **Classification:** Predicting a **label**.
- **Regression:** Predicting a **continuous value**.
- **Clustering:** Grouping **similar data points**.
- **Dimensionality Reduction:** Reducing the **number of features**.
- **Anomaly Detection:** Identifying **unusual data points**.
- **Association Rule Learning:** Identifying **relationships between variables**.



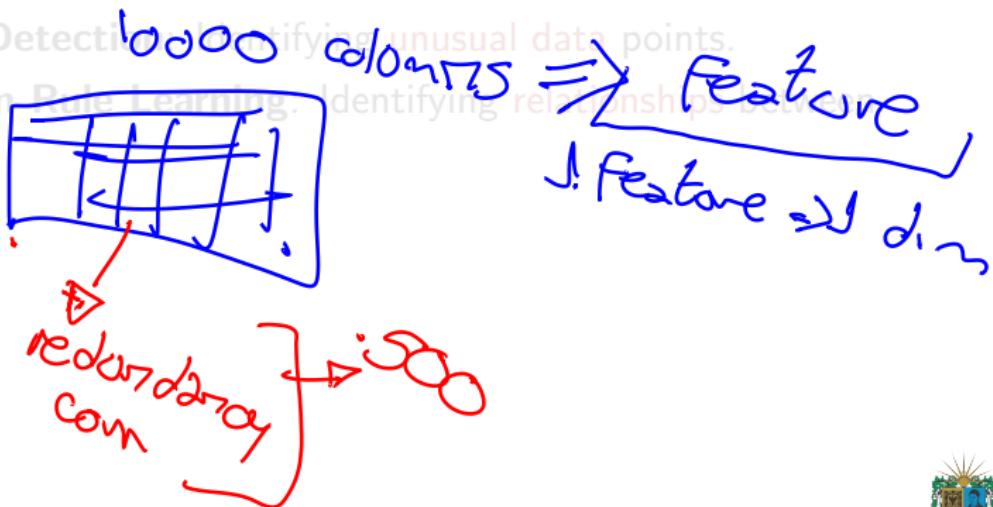
Typical Machine Learning Problems

- **Classification:** Predicting a **label**.
- **Regression:** Predicting a **continuous value**.
- **Clustering:** Grouping **similar data** points.
- **Dimensionality Reduction:** Reducing the **number of features**.
- **Anomaly Detection:** Identifying **unusual data** points.
- **Association Rule Learning:** Identifying **relationships between variables**.



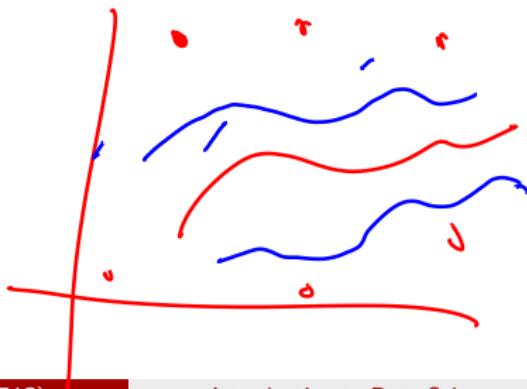
Typical Machine Learning Problems

- **Classification:** Predicting a **label**.
 - **Regression:** Predicting a **continuous value**.
 - **Clustering:** Grouping **similar data** points.
 - **Dimensionality Reduction:** Reducing the **number of features**.
 - Anomaly Detection: Identifying **unusual data** points.
 - Association Rule Learning: Identifying relationships between variables.



Typical Machine Learning Problems

- **Classification:** Predicting a **label**. ↴
- **Regression:** Predicting a **continuous value**.
- **Clustering:** Grouping **similar data** points.
- **Dimensionality Reduction:** Reducing the **number of features**.
- **Anomaly Detection:** Identifying **unusual data** points.
- **Association Rule Learning:** Identifying **relationships** between variables.



Typical Machine Learning Problems

- **Classification:** Predicting a **label**.
- **Regression:** Predicting a **continuous value**.
- **Clustering:** Grouping **similar data** points.
- **Dimensionality Reduction:** Reducing the **number of features**.
- **Anomaly Detection:** Identifying **unusual data** points.
- **Association Rule Learning:** Identifying **relationships** between
variables.

Logic Programming
Input → Rule → Outputs



The Machine Learning Workflow

- **Data Collection:** Gathering the **data**.
- Data Preprocessing: Cleaning and preparing the data.
- Feature Engineering: Creating new features.
- Model Selection: Choosing the best model.
- Model Training: Training the model on the data.
- Model Evaluation: Assessing the model's performance.
- Model Deployment: Putting the model into production.

Handwritten notes:
Data Lake
Multiple Data Sources

Handwritten note:
Data lake



The Machine Learning Workflow

- **Data Collection:** Gathering the **data**.
- **Data Preprocessing:** Cleaning and preparing the **data**.
- Feature Engineering: Creating new features *by Pandas*
- Model Selection: Choosing the **best model**.
- Model Training: Training the **model** on the **data**.
- Model Evaluation: Assessing the **model's performance**.
- Model Deployment: Putting the **model** into **production**.

*Cleaning
Imputers
Feature engineering
→ Categorical to Numerical
encoding*



The Machine Learning Workflow

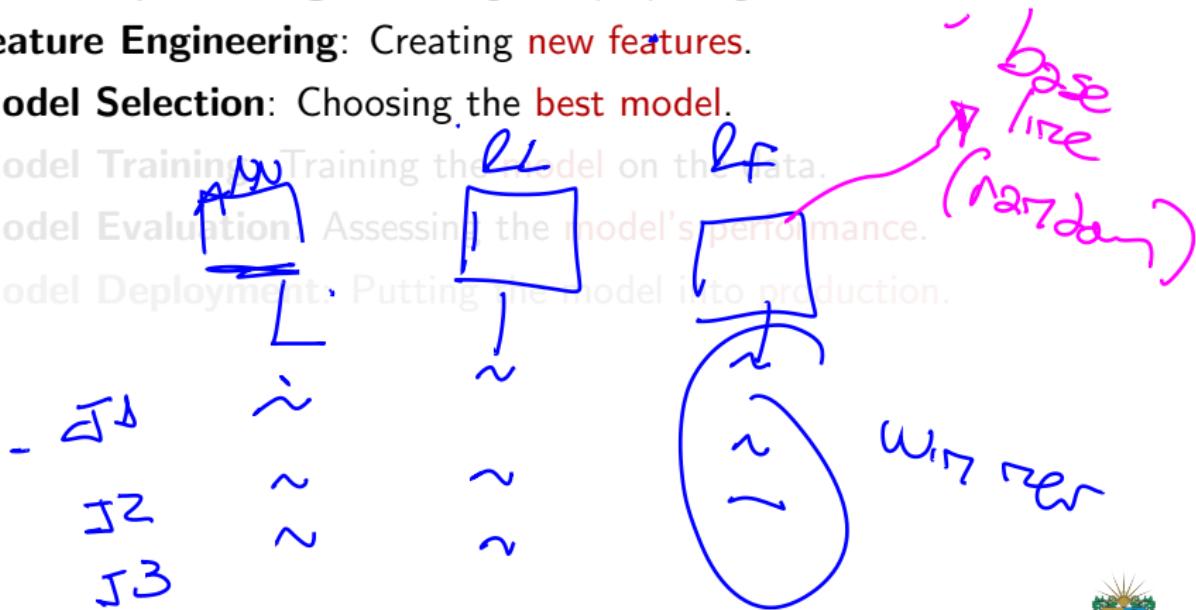
- **Data Collection:** Gathering the **data**.
- **Data Preprocessing:** Cleaning and preparing the **data**.
- **Feature Engineering:** Creating **new features**.
 - Model Selection: Choosing the best model.
 - Model Training: Training the model on the data.
↳ new columns by forming
combining to column
 - Model Evaluation: Assessing the model's performance.
 - Model Deployment: Putting the model into production.
↳ new categories \Rightarrow k-means



The Machine Learning Workflow

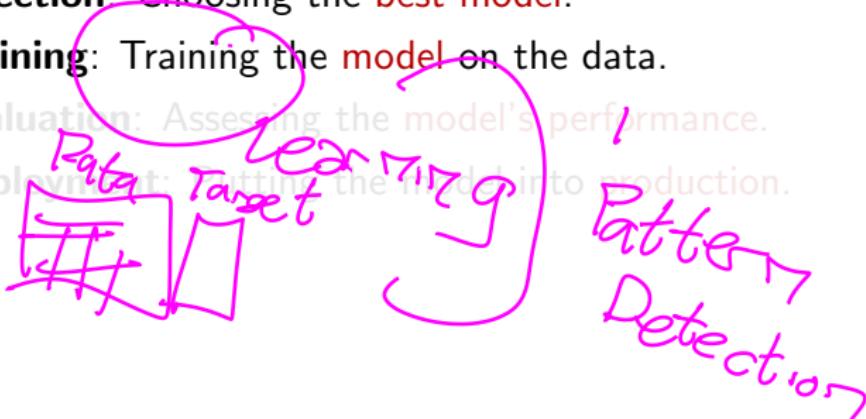
- **Data Collection:** Gathering the **data**.
 - **Data Preprocessing:** Cleaning and preparing the **data**.
 - **Feature Engineering:** Creating **new features**.
 - **Model Selection:** Choosing the **best model**.
 - **Model Training:** Training the **model** on the **data**.
 - **Model Evaluation:** Assessing the **model's performance**.
 - **Model Deployment:** Putting the **model** into **production**.

The diagram consists of four blue-outlined boxes arranged horizontally. The first box contains the handwritten word 'Data'. The second box contains a simple line graph with two points labeled 'L1' and 'L2'. The third box contains the handwritten word 'Evaluation'. The fourth box contains the handwritten word 'Deployment'. A pink wavy line starts at the top right of the 'Evaluation' box and curves down towards the bottom right, ending near the bottom right corner of the 'Deployment' box.



The Machine Learning Workflow

- **Data Collection:** Gathering the **data**.
- **Data Preprocessing:** Cleaning and preparing the **data**.
- **Feature Engineering:** Creating **new features**.
- **Model Selection:** Choosing the **best model**.
- **Model Training:** Training the **model** on the **data**.
- Model Evaluation: Assessing the **model's performance**.
- Model Deployment: Putting the **model** into **production**.



The Machine Learning Workflow

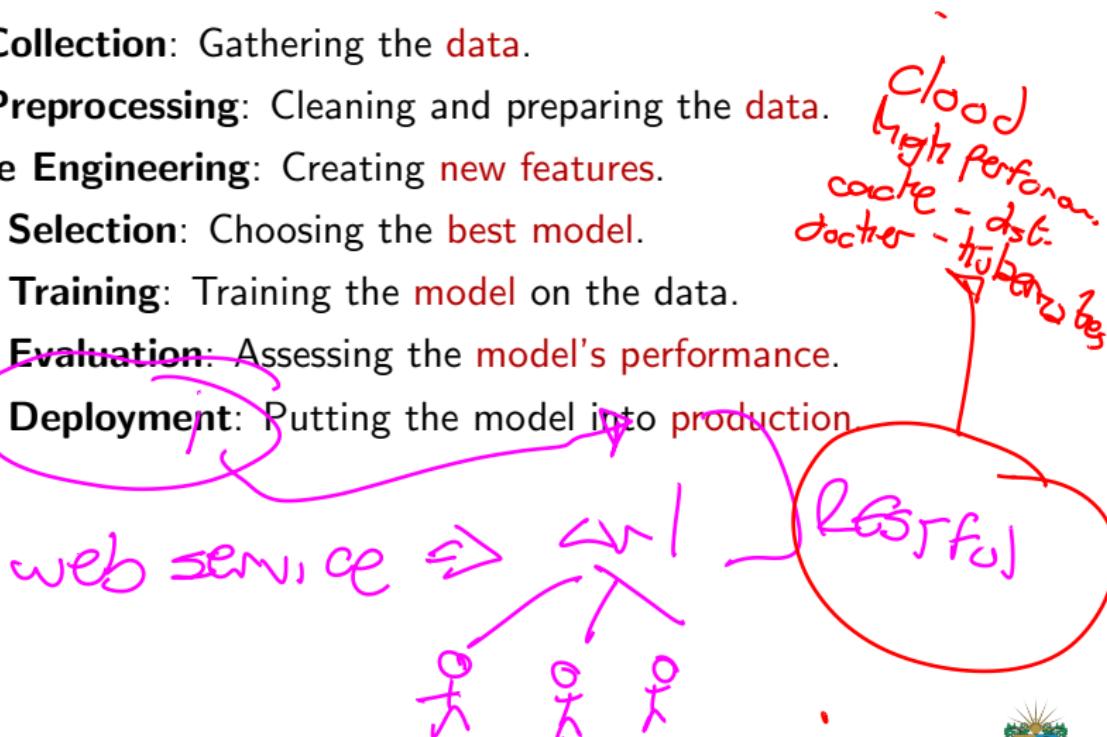
- **Data Collection:** Gathering the **data**.
- **Data Preprocessing:** Cleaning and preparing the **data**.
- **Feature Engineering:** Creating **new features**.
- **Model Selection:** Choosing the **best model**.
- **Model Training:** Training the **model** on the data.
- **Model Evaluation:** Assessing the **model's performance**.
- **Model Deployment:** Putting the **model** into **production**.

metris *CS*
trustful
error level



The Machine Learning Workflow

- **Data Collection:** Gathering the **data**.
- **Data Preprocessing:** Cleaning and preparing the **data**.
- **Feature Engineering:** Creating **new features**.
- **Model Selection:** Choosing the **best model**.
- **Model Training:** Training the **model** on the data.
- **Model Evaluation:** Assessing the **model's performance**.
- **Model Deployment:** Putting the model into **production**



Examining the Data

- **Data Exploration:** Understanding the data.

- Data Cleaning: Preparing the data.

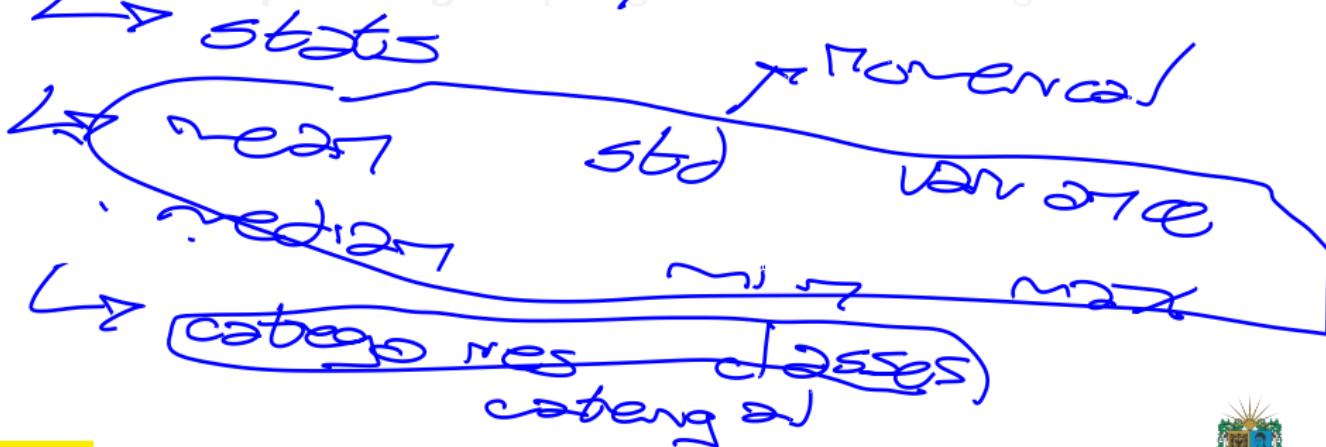


Descriptive
Analysis

- Feature Engineering: Creating new features.

- Feature Selection: Selecting the most important features.

- Data Preprocessing: Preparing the data for modeling.



Examining the Data

- **Data Exploration:** Understanding the data.
- **Data Cleaning:** Preparing the data.

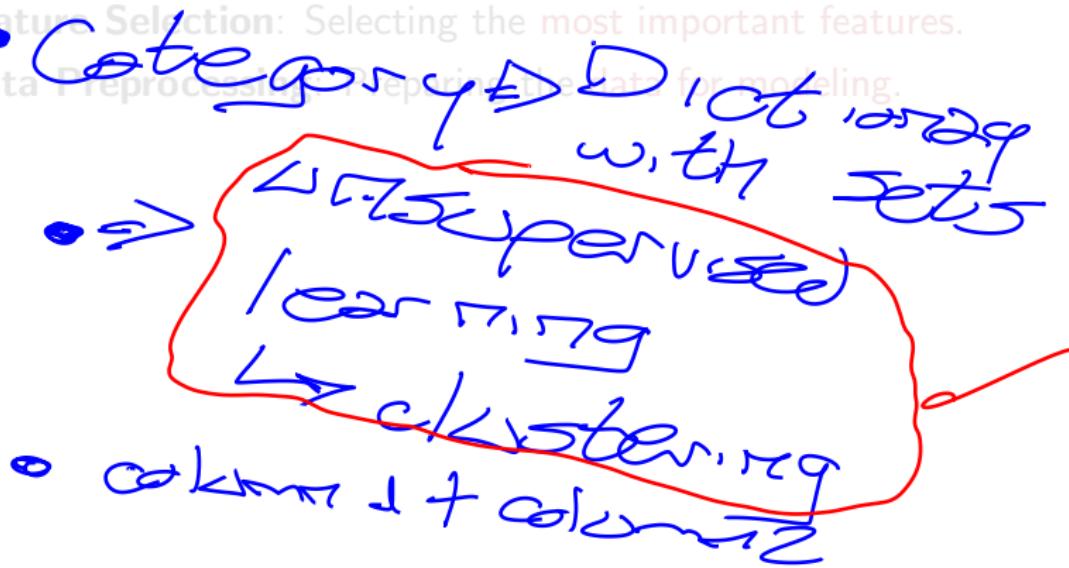
- Feature Engineering: Creating new features.
- Feature Selection: Selecting the most important features.
- Data Preprocessing: Preparing the data for modeling.

- **Drop duplicates**
↳ Bias
- **Format** ↳ string
↳ date/time



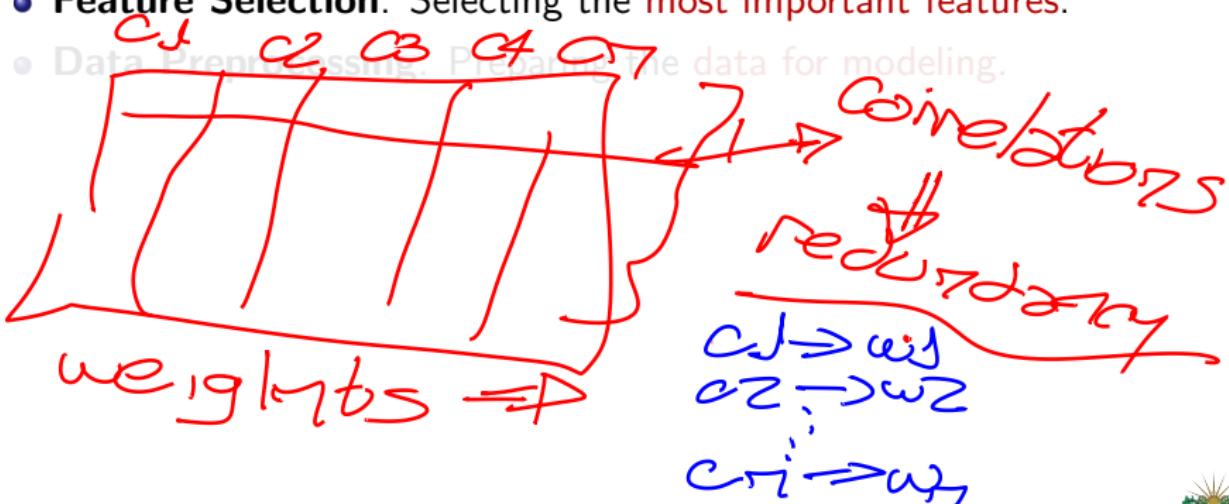
Examining the Data

- **Data Exploration:** Understanding the data.
- **Data Cleaning:** Preparing the data.
- **Feature Engineering:** Creating new features.
- Feature Selection: Selecting the most important features.
- Data Preprocessing: Preparing the data for modeling.



Examining the Data

- **Data Exploration:** Understanding the data.
- **Data Cleaning:** Preparing the data.
- **Feature Engineering:** Creating new features.
- **Feature Selection:** Selecting the most important features.
- **Data Preprocessing:** Preparing the data for modeling.



Examining the Data



- **Data Exploration:** Understanding the data.
- **Data Cleaning:** Preparing the data.
- **Feature Engineering:** Creating new features.
- **Feature Selection:** Selecting the **most important** features.
- **Data Preprocessing:** Preparing the **data** for modeling.

• Scaling \Rightarrow height \Rightarrow 140-200
 $w=0-1$

age \Rightarrow 20-80

weight \Rightarrow 50-130

$$f = \omega_1 f_1 + \omega_2 f_2 + \omega_3 f_3 \\ (\omega_1 + \omega_2 + \omega_3) = 1$$

$$\begin{aligned} & \text{A box labeled } M_{1,2,3,4,2x} \text{ with a red circle around } M_{1,2,3,4} \\ & \text{A formula: } x_{\text{pred}} = x_{\text{old}} + (x_{\text{new}} - x_{\text{old}}) \\ & \text{A red bracket under } x_{\text{new}} - x_{\text{old}} \end{aligned}$$



K-Nearest Neighbors Classification

- **K-Nearest Neighbors** is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure.
 - It is a type of **instance-based learning**, or lazy learning, where the function is only approximated locally and all computation is deferred until function evaluation.

$$d_{ec} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

distance



Algorithmic Bias

- Algorithmic bias is a **systematic error** in a model that results in **unfair outcomes**.
- It can be caused by **biased training data**, **biased algorithms** or biased decision-making.

Unbiased
pre-processing



Outline

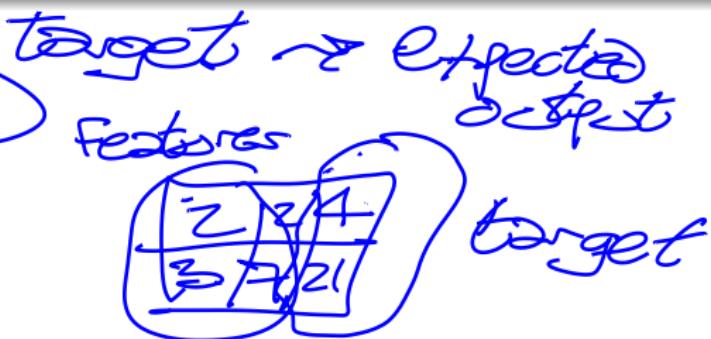
- 1 Fundamentals of Machine Learning
- 2 Supervised Machine Learning
- 3 Supervised Machine Learning Algorithms
- 4 Machine Learning Models Evaluation



Introduction to Supervised Machine Learning

Definition

- **Supervised learning** is a type of machine learning where the model is trained on labeled data.
- It involves training a model to map input data to output data based on example input-output pairs.

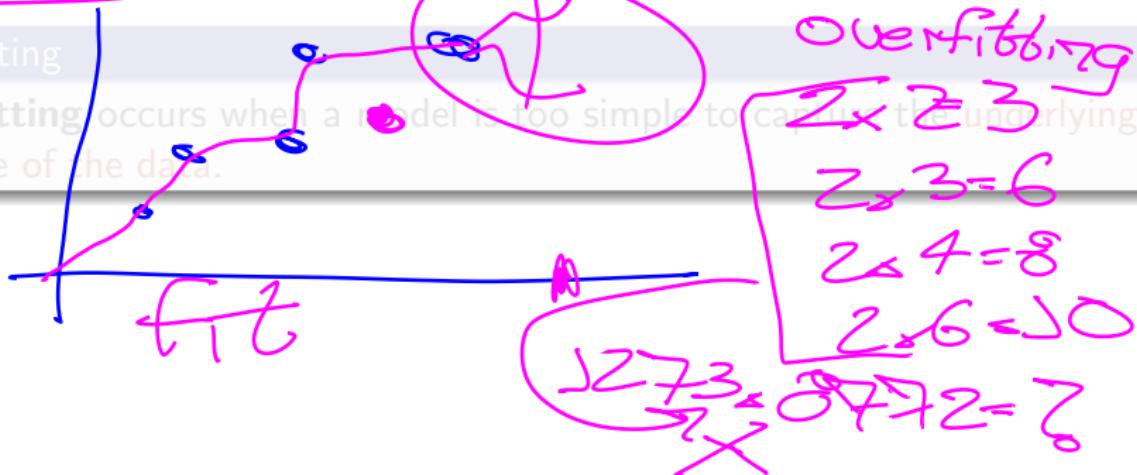


Overfitting and Underfitting

Overfitting

Overfitting occurs when a model learns the training data too well and performs poorly on new data.

Underfitting occurs when a model is too simple to capture the underlying structure of the data.



Overfitting and Underfitting

Overfitting

Overfitting occurs when a model learns the training data too well and performs poorly on new data.

Underfitting

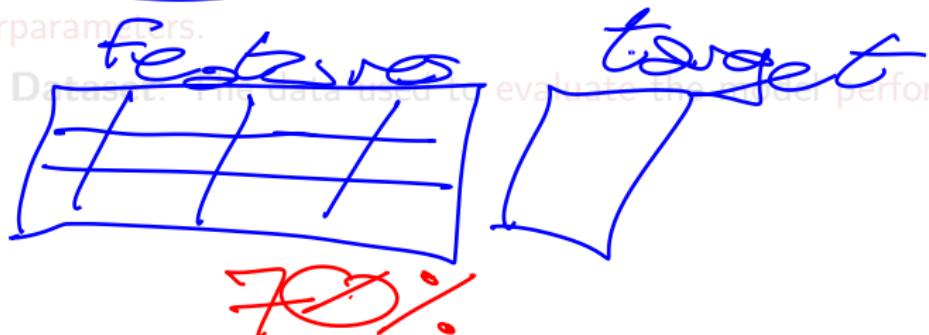
Underfitting occurs when a model is too simple to capture the underlying structure of the data.

→ No
To [better]
Random



Supervised Learning Datasets

- **Training Dataset:** The data used to **train the model**.
- **Validation Dataset:** The data used to **tune the model hyperparameters**.
- **Test Dataset:** The data used to **evaluate the model performance**.



Supervised Learning Datasets

- **Training Dataset:** The data used to train the model.
- **Validation Dataset:** The data used to tune the model

hyperparameters.

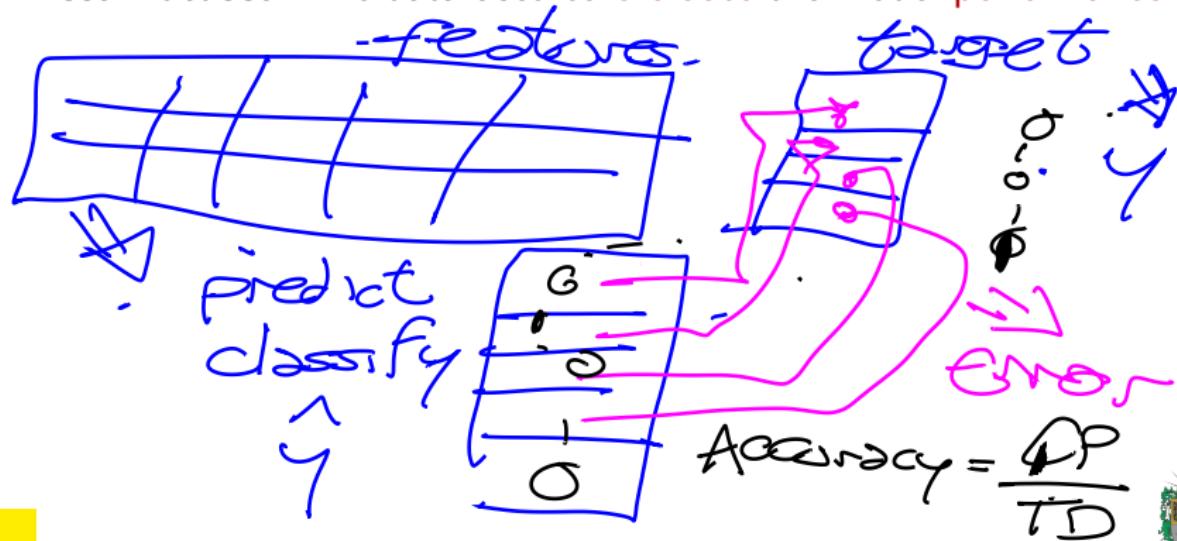
Hyperparameter Optimization

→ setup → k-fold cross-validation
Training \Rightarrow 70%
Validation \Rightarrow 30%



Supervised Learning Datasets

- **Training Dataset:** The data used to **train** the model.
 - **Validation Dataset:** The data used to **tune** the model **hyperparameters**.
 - **Test Dataset:** The data used to **evaluate** the model **performance**.



K-Nearest Neighbors: Classification and Regression

- K-Nearest Neighbors (KNN) is a simple algorithm that stores all available cases and classifies new cases based on a **similarity measure**.
- It can be used for both **classification** and **regression** tasks.
- For **classification**, the output is the **class label** of the majority of the k-nearest neighbors.
- For **regression**, the output is the **average** of the k-nearest neighbors.

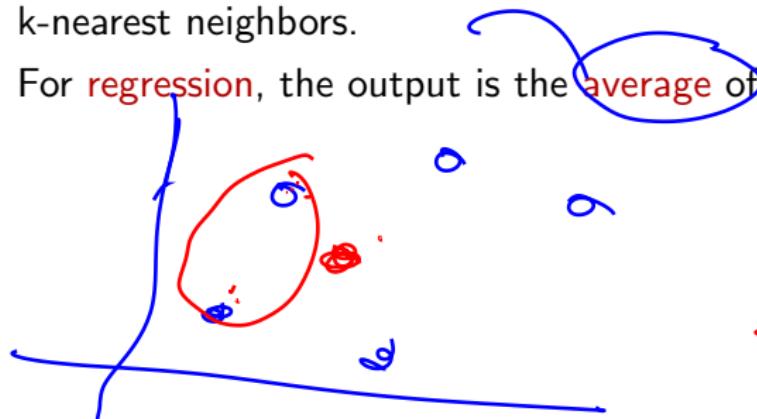
Titanic
 \rightarrow
 $k=2$
 accuracy $\Rightarrow \approx 75\%$

Survival
 $0 \rightarrow 1$
 Binary
 Classification



K-Nearest Neighbors: Classification and Regression

- **K-Nearest Neighbors (KNN)** is a simple algorithm that stores all available cases and classifies new cases based on a **similarity measure**.
- It can be used for both **classification** and **regression** tasks.
- For **classification**, the output is the **class label** of the majority of the k-nearest neighbors.
- For **regression**, the output is the **average** of the k-nearest neighbors.



$$k=2$$

$$\frac{z_1 + z_2}{2} = \underline{\underline{z}}$$



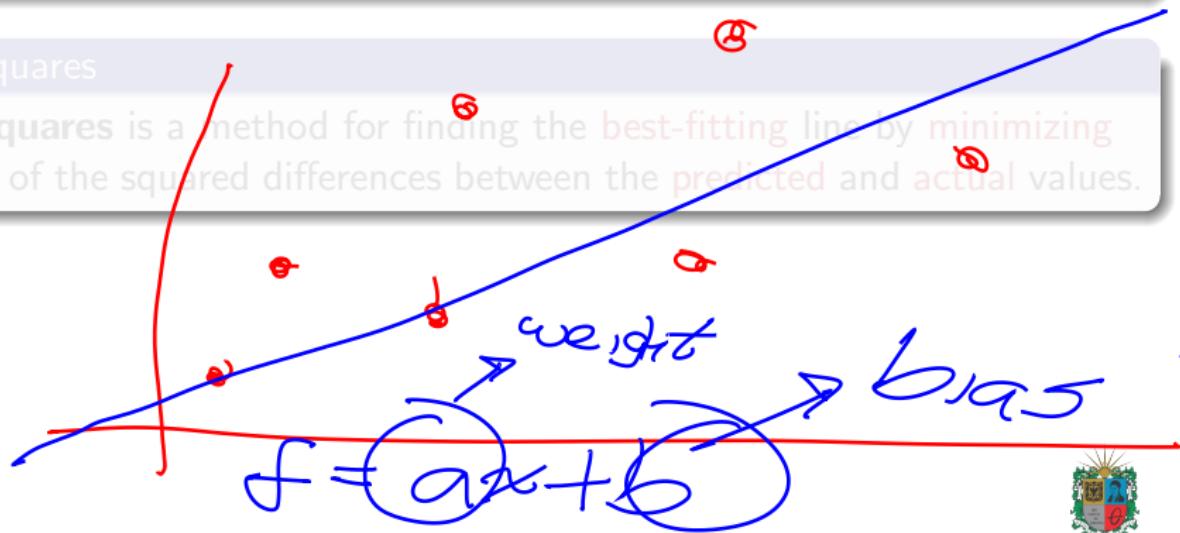
Linear Regression with Least Squares

$$f = a_0 f_0 + a_1 f_1 + \dots + a_n f_n$$

Linear Regression

- **Linear regression** is a type of regression analysis used for predicting the value of a **continuous dependent variable**.
 - It works by finding the line that best fits the data.

Least squares is a method for finding the best-fitting line by minimizing the sum of the squared differences between the predicted and actual values.



Linear Regression with Least Squares

Linear Regression

- **Linear regression** is a type of **regression analysis** used for predicting the value of a **continuous dependent variable**.
- It works by finding the **line that best fits the data**.

Least Squares

Least squares is a method for finding the **best-fitting** line by **minimizing the sum** of the squared differences between the **predicted** and **actual** values.

$$\text{función } (\beta_0, \beta_1) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

derivative



Ridge & Lasso

L2

Ridge Regression

Ridge regression is a type of **linear regression** that includes a penalty term to prevent overfitting. It works by adding a **regularization** term to the least squares objective function.

Lasso Regression

Lasso regression is a type of **linear regression** that includes a penalty term to prevent overfitting. It works by adding a **regularization** term to the least squares objective function.

W *high \rightarrow high coefficient*



Ridge & Lasso

Ridge Regression

Ridge regression is a type of **linear regression** that includes a penalty term to **prevent overfitting**. It works by adding a **regularization** term to the least squares objective function.

L1

Lasso Regression

Lasso regression is a type of **linear regression** that includes a penalty term to **prevent overfitting**. It works by adding a **regularization** term to the least squares objective function.

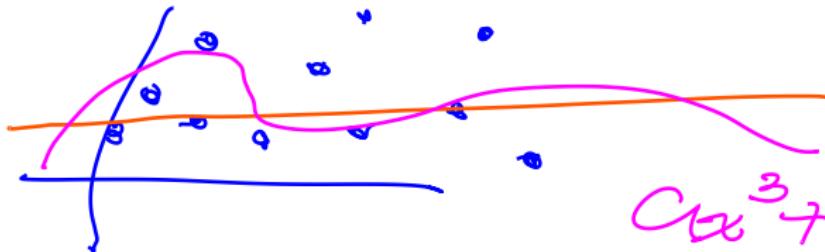
$\|w\|_1$ → No steps ↗
No. of zero coefficients ↗
 $y_{10}(\omega \text{ as } w)$



Polynomial Regression

Polynomial Regression

- **Polynomial regression** is a type of **regression analysis** that models the relationship between the independent and dependent variables as an **nth-degree polynomial**.
- It can capture **non-linear relationships** between the variables.



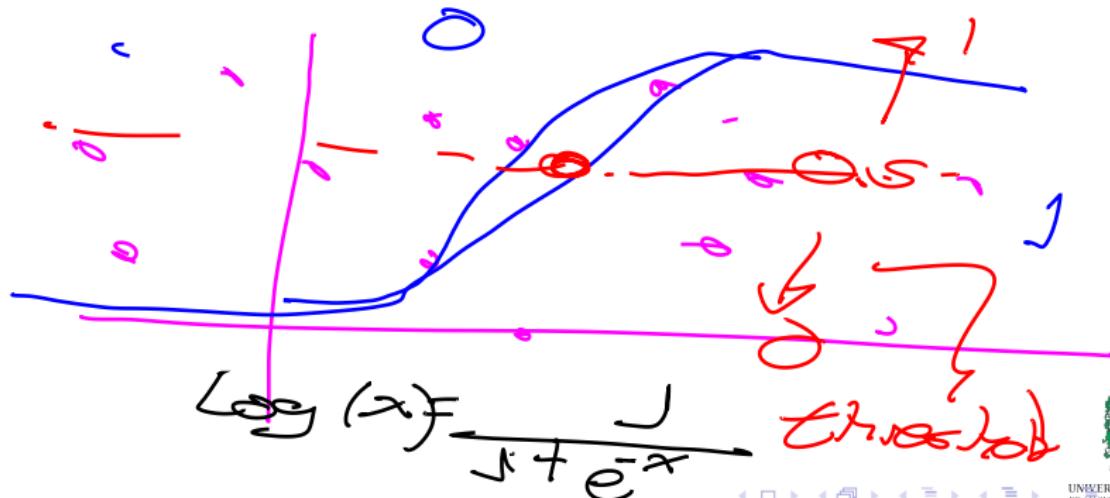
$$\alpha x^3 + \beta x^2 + \gamma x + \delta$$



Logistic Regression

Logistic Regression

- Logistic regression is a type of regression analysis used for predicting the outcome of a categorical dependent variable.
- It is used for binary classification tasks, where the output is a probability between 0 and 1.



Cross-Validation

- **Cross-validation** is a technique for **assessing the performance** of a model.
- It involves **splitting** the data into multiple subsets, training the model on some subsets, and evaluating it on others.
- Common cross-validation **techniques** include **k-fold cross-validation** and **leave-one-out cross-validation**.
- Cross-validation helps to reduce **overfitting** and provides a more accurate estimate of the model's **performance**.

A | B | C | D | E

I_E1

$T = A B C D$

$T_{test} = E$

I_E3: $T_{train} = A B$ or
 $T_{test} = C$

I_E2

$T_{train} = B C D E$
 $T_{test} = A$

I_E4: $T_{train} = A B C$
 $T_{test} = D$

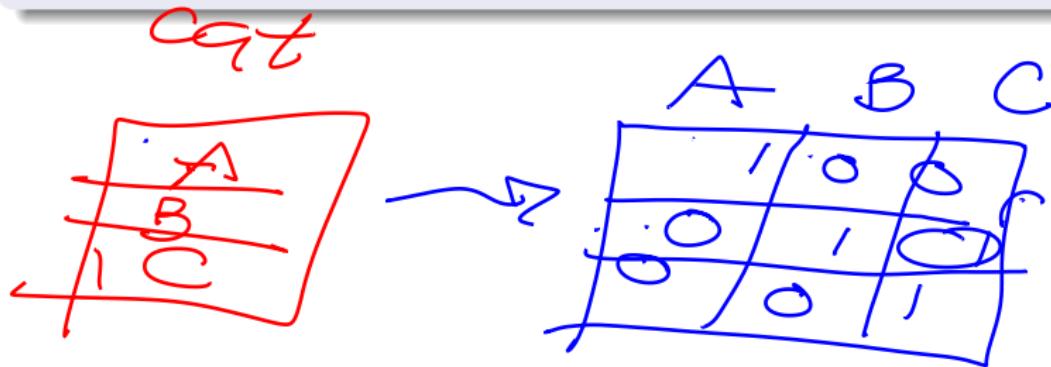
I_E5 : $T_{train} = A C D E$
 $T_{test} = B$



One-Hot Encoding

One-Hot Encoding

- One-hot encoding is a technique for converting categorical variables into numerical variables.
- It creates a binary vector for each category, with a 1 for the category and 0s for all other categories.



Data Leakage

- **Data leakage** occurs when information from the test set is inadvertently used to train the model.
- It can lead to **overfitting** and inflated performance metrics.
- Common sources of **data leakage** include **target leakage**, **train-test contamination**, and **information leakage**.
- To prevent **data leakage**, it is important to **carefully separate** the training and test data and avoid using information from the test set during training.

sklearn



Outline

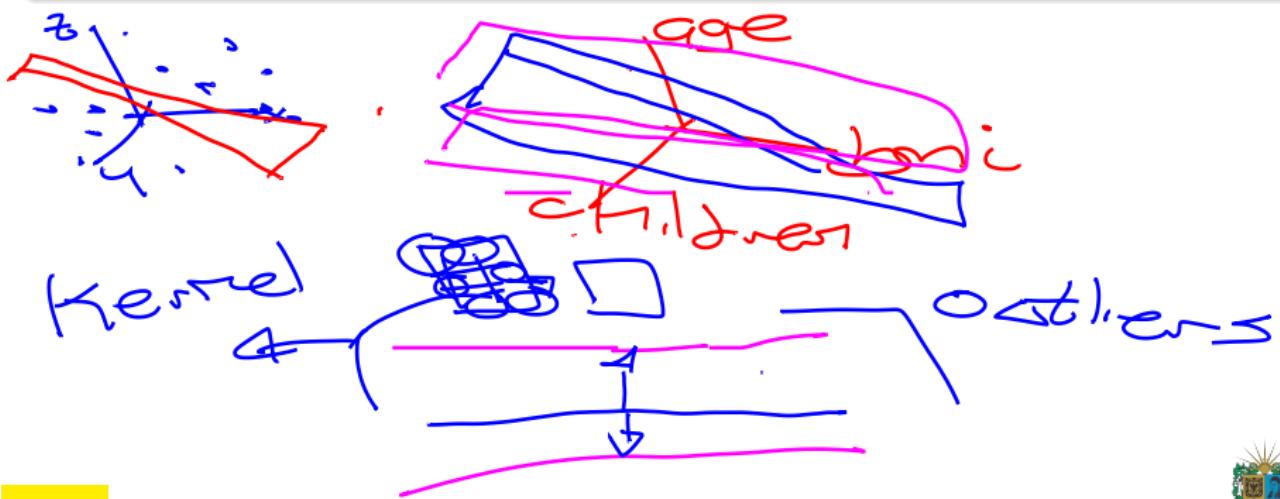
- 1 Fundamentals of Machine Learning
- 2 Supervised Machine Learning
- 3 Supervised Machine Learning Algorithms
- 4 Machine Learning Models Evaluation



Support Vector Machines

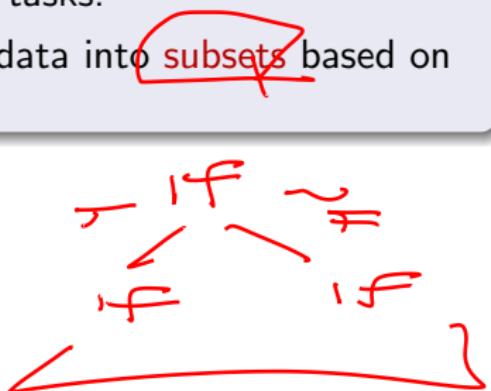
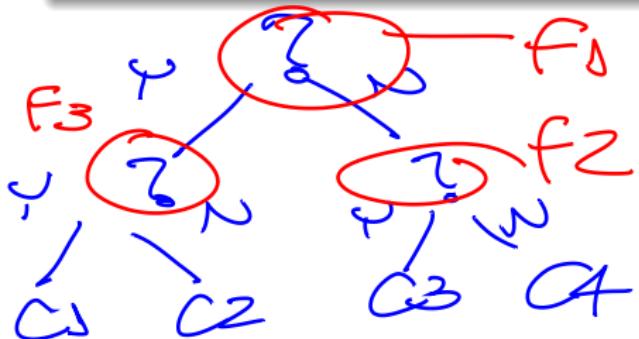
SVM

- **Support vector machines** are a type of **machine learning model** that can be used for both **classification** and **regression** tasks.
- They work by finding the **hyperplane** that best **separates** the data into different classes.



Decision Trees

- Decision trees are a type of machine learning model that can be used for both classification and regression tasks.
- They work by recursively partitioning the data into subsets based on the values of the features.

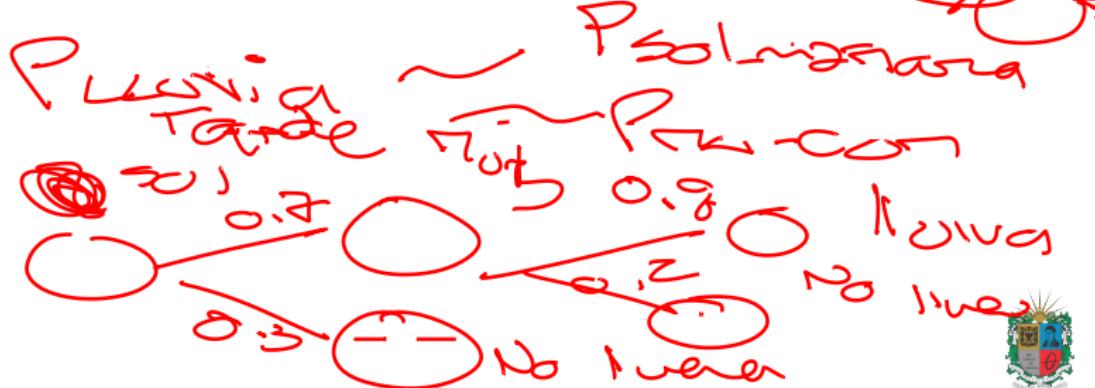


Naive Bayes Classifier

- The **naive Bayes classifier** is a simple probabilistic **classifier** based on **Bayes' theorem**.
- It assumes that the features are **conditionally independent** given the class label.

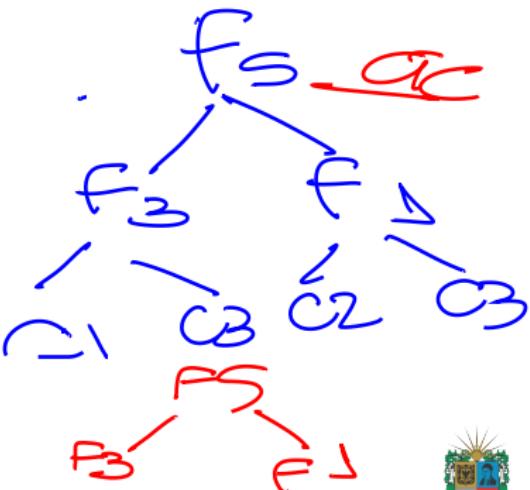
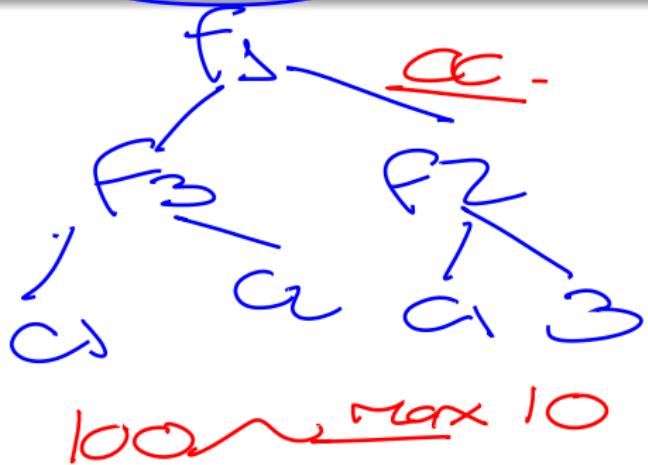
$$P(A) = 0.7$$

$$P(B|A)$$



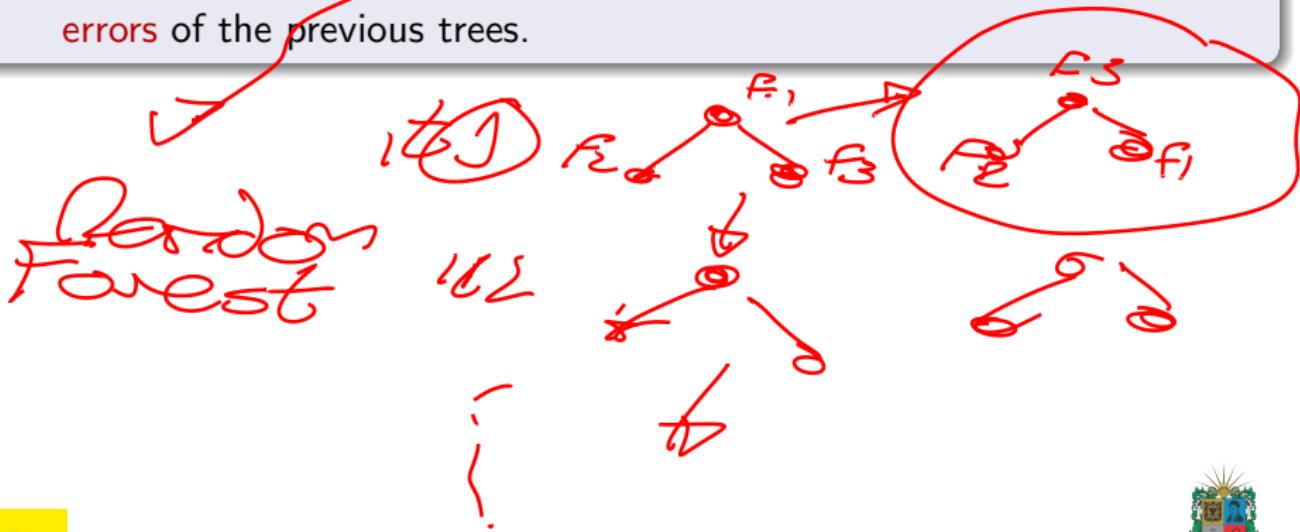
Random Forest

- **Random forest** is an **ensemble learning** method that combines **multiple decision trees** to create a strong predictive model.
- It works by building **multiple trees** and averaging their predictions to reduce **overfitting**.



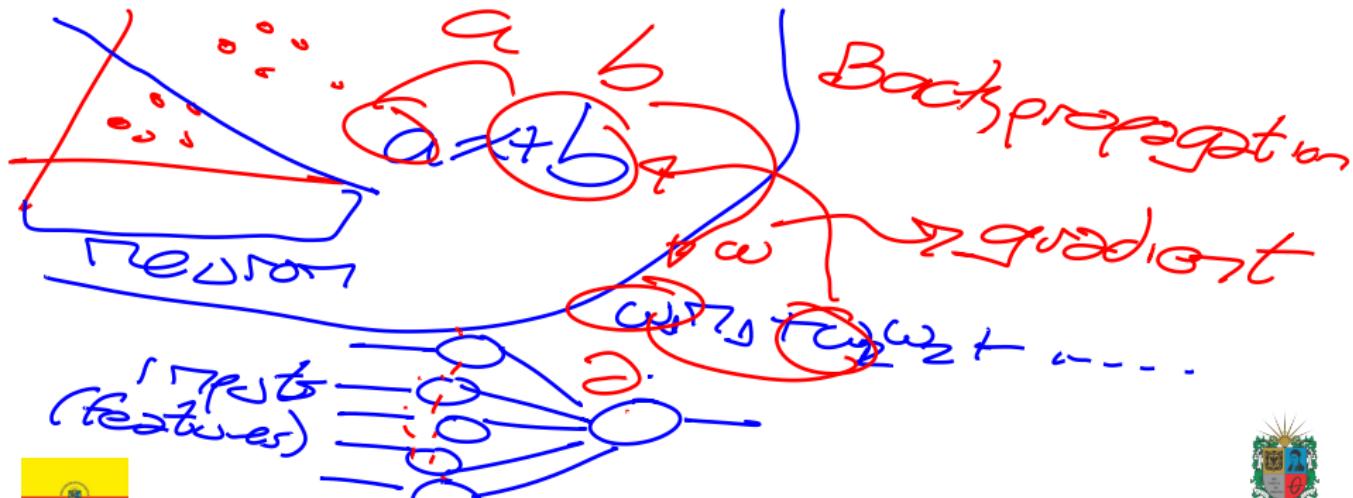
Gradient Boosted Decision Trees

- Gradient boosted decision trees are an ensemble learning method that combines multiple decision trees and gradient descent optimization to create a strong predictive model.
- They work by building trees sequentially, with each tree correcting the errors of the previous trees.



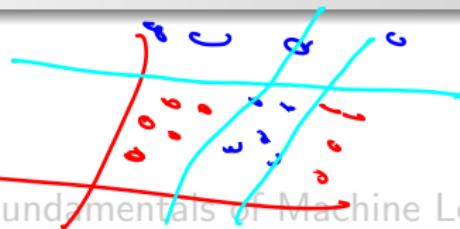
Neural Networks

- Neural networks are a type of machine learning model inspired by the **human brain**.
- They consist of **layers** of interconnected nodes that process **input data** and produce **output data**.



Outline

- 1 Fundamentals of Machine Learning



blue

- 2 Supervised Machine Learning



- 3 Supervised Machine Learning Algorithms

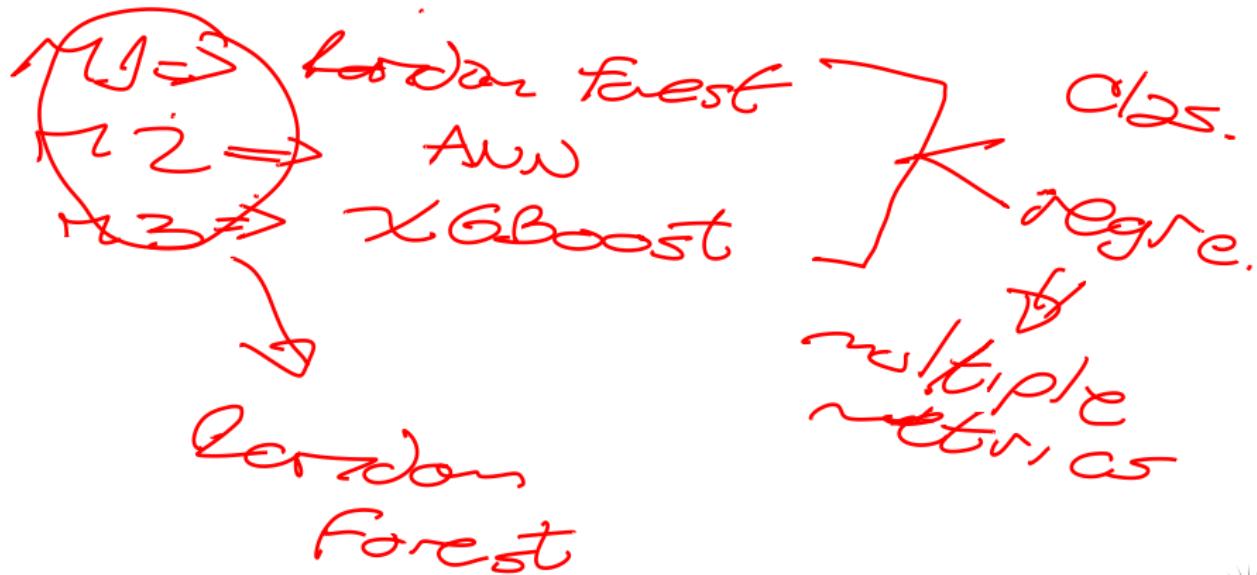
- 4 Machine Learning Models Evaluation

Deep learning



Model Evaluation & Selection

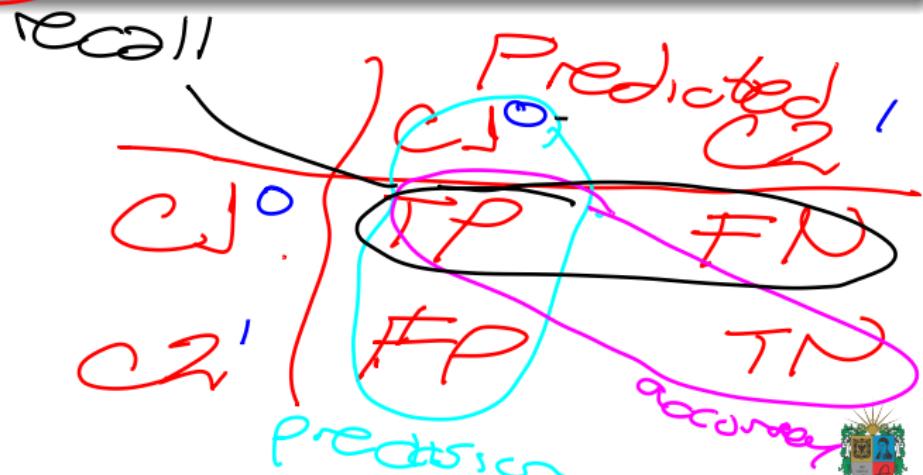
- **Model Evaluation:** Assessing the **performance** of a model.
- **Model Selection:** Choosing the **best** model for the task.



Confusion Matrices

Definition

- A **confusion matrix** is a **table** that summarizes the **performance** of a classification model.
- It shows the number of **true positives**, **true negatives**, **false positives**, and **false negatives**.



Basic Evaluation Metrics

Classification

- **Accuracy:** The proportion of **correct predictions**.
- **Precision:** The proportion of **true positives** among all positive predictions.
- **Recall:** The proportion of **true positives** among **all actual positives**.
- **F1 Score:** The harmonic mean of precision and recall.

$$AC = \frac{TP + TN}{TP + FP + FN + TN}$$



Basic Evaluation Metrics

- **Accuracy:** The proportion of **correct predictions**.
- **Precision:** The proportion of **true positives** among all positive predictions.
- Recall: The proportion of true positives among all actual positives.
- F1 Score: The harmonic mean of precision and recall.

Precision =

$$\frac{TP}{TP + FP}$$

$\approx 1 \Rightarrow$ Predicted True = True

$\approx 0 \Rightarrow$ Predicted True is False Positive



Basic Evaluation Metrics

- **Accuracy:** The proportion of **correct predictions**.
- **Precision:** The proportion of **true positives** among all positive predictions.
- **Recall:** The proportion of **true positives** among all actual positives.
- **F1 Score:** The harmonic mean of precision and recall.

Recall =

$$\frac{TP}{TP + FN}$$



→ Expected True

→ Expected True

*is Predicted True
is False*



Basic Evaluation Metrics

- **Accuracy:** The proportion of **correct predictions**.
- **Precision:** The proportion of **true positives** among all positive predictions.
- **Recall:** The proportion of **true positives** among **all actual positives**.
- **F1 Score:** The **harmonic mean** of precision and recall.

$$F_1 = \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

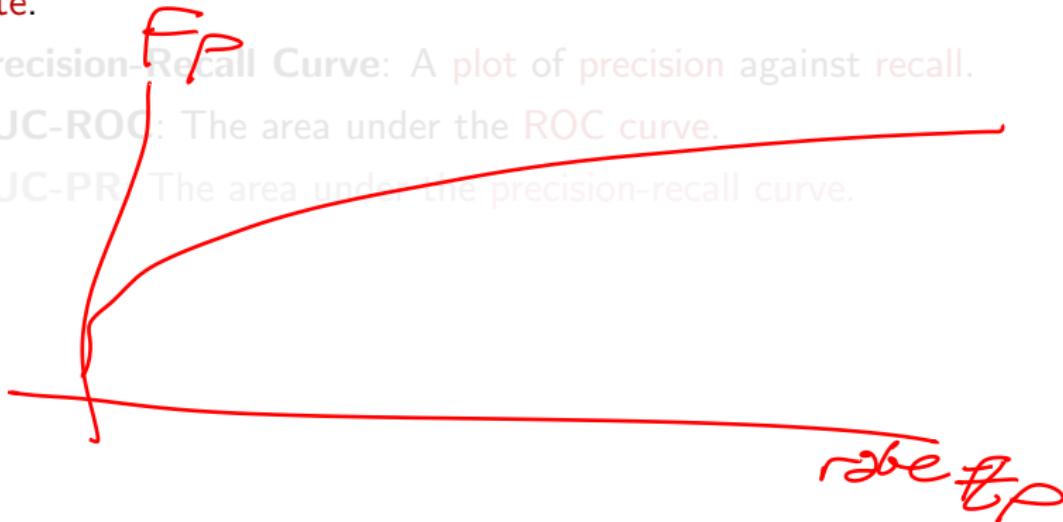
Prec.
real
real
 ≈

$$\frac{2 * 1}{2} = 1$$



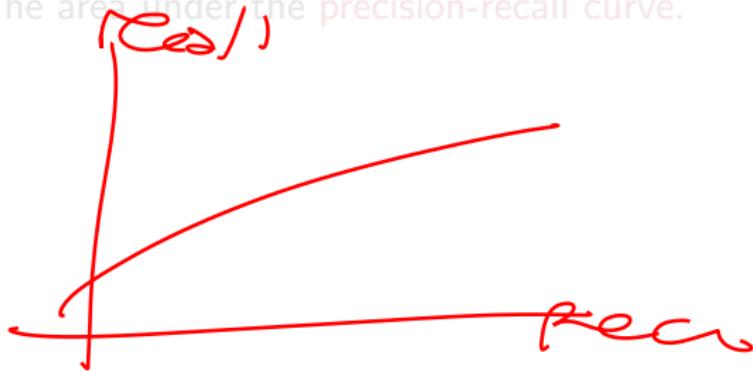
Classifier Decision Metrics

- **ROC Curve:** A plot of the true positive rate against the false positive rate.
- Precision-Recall Curve: A plot of precision against recall.
- AUC-ROC: The area under the ROC curve.
- AUC-PR: The area under the precision-recall curve.



Classifier Decision Metrics

- **ROC Curve:** A plot of the true positive rate against the false positive rate.
- **Precision-Recall Curve:** A plot of precision against recall.
- AUC-ROC: The area under the ROC curve.
- AUC-PR: The area under the precision-recall curve.



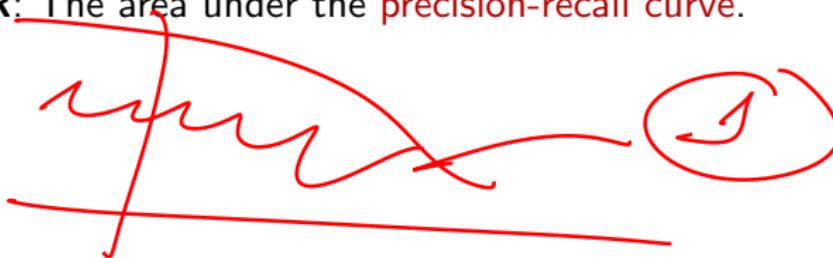
Classifier Decision Metrics

- **ROC Curve:** A plot of the true positive rate against the false positive rate.
- **Precision-Recall Curve:** A plot of precision against recall.
- **AUC-ROC:** The area under the ROC curve.
- **AUC-PR:** The area under the precision-recall curve.



Classifier Decision Metrics

- **ROC Curve:** A plot of the true positive rate against the false positive rate.
- **Precision-Recall Curve:** A plot of precision against recall.
- **AUC-ROC:** The area under the ROC curve.
- **AUC-PR:** The area under the precision-recall curve.



Multi-Class Evaluation

- **Macro-Averaging:** The **average** of the evaluation **metrics** for each class.
- **Micro-Averaging:** The evaluation metrics calculated on the aggregate confusion matrix.
- **Weighted-Averaging:** The average of the evaluation metrics **weighted** by the number of **samples** in each class.
- **One-vs-All:** A strategy for multi-class classification that trains a separate binary classifier for each class.



Multi-Class Evaluation

- **Macro-Averaging:** The **average** of the evaluation **metrics** for each class.
- **Micro-Averaging:** The evaluation metrics calculated on the **aggregate confusion matrix**.
- **Weighted-Averaging:** The average of the evaluation metrics **weighted** by the number of **samples** in each class.
- **One-vs-All:** A strategy for multi-class classification that trains a separate binary classifier for each class.



Multi-Class Evaluation

- **Macro-Averaging:** The **average** of the evaluation **metrics** for each class.
- **Micro-Averaging:** The evaluation metrics calculated on the **aggregate confusion matrix**.
- **Weighted-Averaging:** The average of the evaluation metrics **weighted** by the number of **samples** in each class.
- **One-vs-All:** A strategy for multi-class classification that trains a separate binary classifier for each class.



Multi-Class Evaluation

- **Macro-Averaging:** The **average** of the evaluation **metrics** for each class.
- **Micro-Averaging:** The evaluation metrics calculated on the **aggregate confusion matrix**.
- **Weighted-Averaging:** The average of the evaluation metrics **weighted** by the number of **samples** in each class.
- **One-vs-All:** A strategy for multi-class classification that trains a **separate binary classifier** for each class.



Regression Evaluation

- **Mean Squared Error:** The **average** of the **squared differences** between the predicted and actual values.
- **Mean Absolute Error:** The **average** of the **absolute differences** between the predicted and actual values.
- **R-Squared:** The **proportion** of the **variance** in the dependent variable that is predictable from the independent variables.
- **Adjusted R-Squared:** A modified version of R-squared that adjusts for the **number of predictors** in the model.
- **Root Mean Squared Error:** The **square root** of the **mean squared error**.



Regression Evaluation

- **Mean Squared Error:** The **average** of the **squared differences** between the predicted and actual values.
- **Mean Absolute Error:** The **average** of the **absolute differences** between the predicted and actual values.
- **R-Squared:** The **proportion** of the **variance** in the dependent variable that is predictable from the independent variables.
- **Adjusted R-Squared:** A modified version of R-squared that adjusts for the **number of predictors** in the model.
- **Root Mean Squared Error:** The **square root** of the **mean squared error**.



Regression Evaluation

- **Mean Squared Error:** The **average** of the **squared differences** between the predicted and actual values.
- **Mean Absolute Error:** The **average** of the **absolute differences** between the predicted and actual values.
- **R-Squared:** The **proportion** of the **variance** in the dependent variable that is predictable from the independent variables.
- **Adjusted R-Squared:** A modified version of R-squared that adjusts for the **number of predictors** in the model.
- **Root Mean Squared Error:** The **square root** of the mean **squared error**.



Regression Evaluation

- **Mean Squared Error:** The **average** of the **squared differences** between the predicted and actual values.
- **Mean Absolute Error:** The **average** of the **absolute differences** between the predicted and actual values.
- **R-Squared:** The **proportion** of the **variance** in the dependent variable that is predictable from the independent variables.
- **Adjusted R-Squared:** A modified version of R-squared that adjusts for the **number of predictors** in the model.
- **Root Mean Squared Error:** The **square root** of the mean **squared error**.



Regression Evaluation

- **Mean Squared Error:** The **average** of the **squared differences** between the predicted and actual values.
- **Mean Absolute Error:** The **average** of the **absolute differences** between the predicted and actual values.
- **R-Squared:** The **proportion** of the **variance** in the dependent variable that is predictable from the independent variables.
- **Adjusted R-Squared:** A modified version of R-squared that adjusts for the **number of predictors** in the model.
- **Root Mean Squared Error:** The **square root** of the mean **squared error**.



Model Calibration

- **Model calibration** is the process of adjusting the output of a model to match the **true probability distribution**.
- It is important for models that *output probabilities*, such as **logistic regression** and **support vector machines**.



Outline

- 1 Fundamentals of Machine Learning
- 2 Supervised Machine Learning
- 3 Supervised Machine Learning Algorithms
- 4 Machine Learning Models Evaluation



Thanks!

Questions?



Repo: <https://github.com/EngAndres/ud-public/tree/main/courses/data-science-introduction>

