Introduction to Machine Learning

Basic Concepts, Supervised Learning, Unsupervised Learning

Author: Eng. Carlos Andrés Sierra, M.Sc.
cavirguezs@udistrital.edu.co

Full-time Adjunct Professor
Computer Engineering Program
School of Engineering
Universidad Distrital Francisco José de Caldas

2026-I

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Outline

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

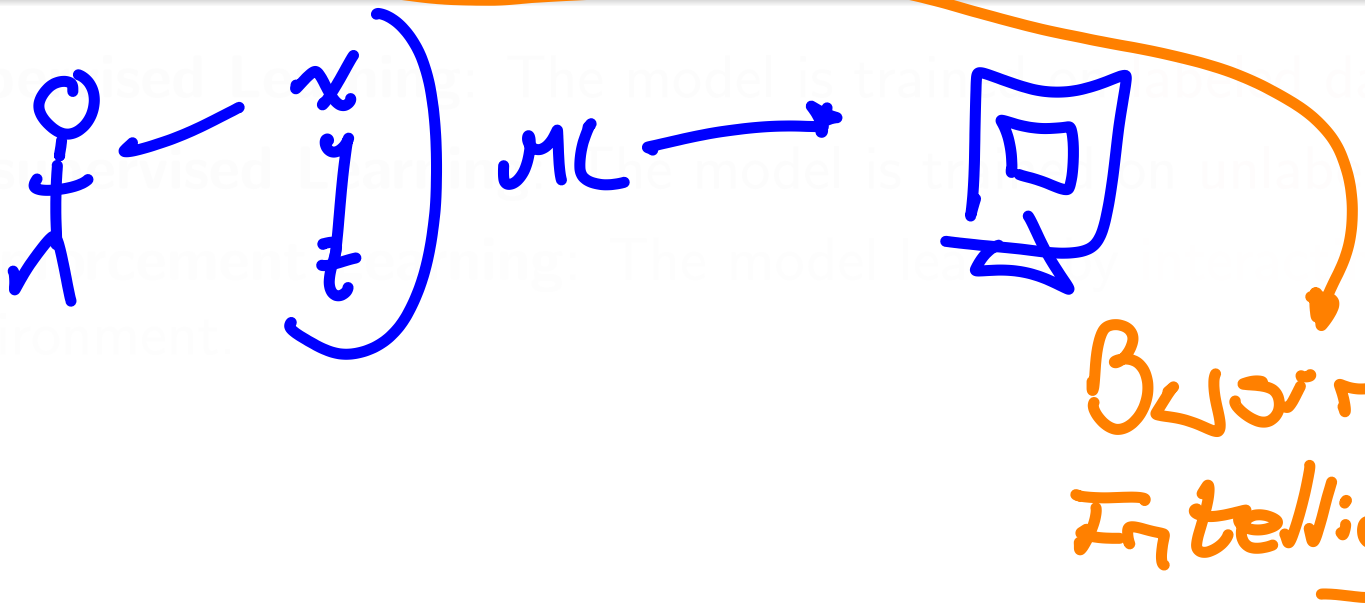
Outline

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

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`.

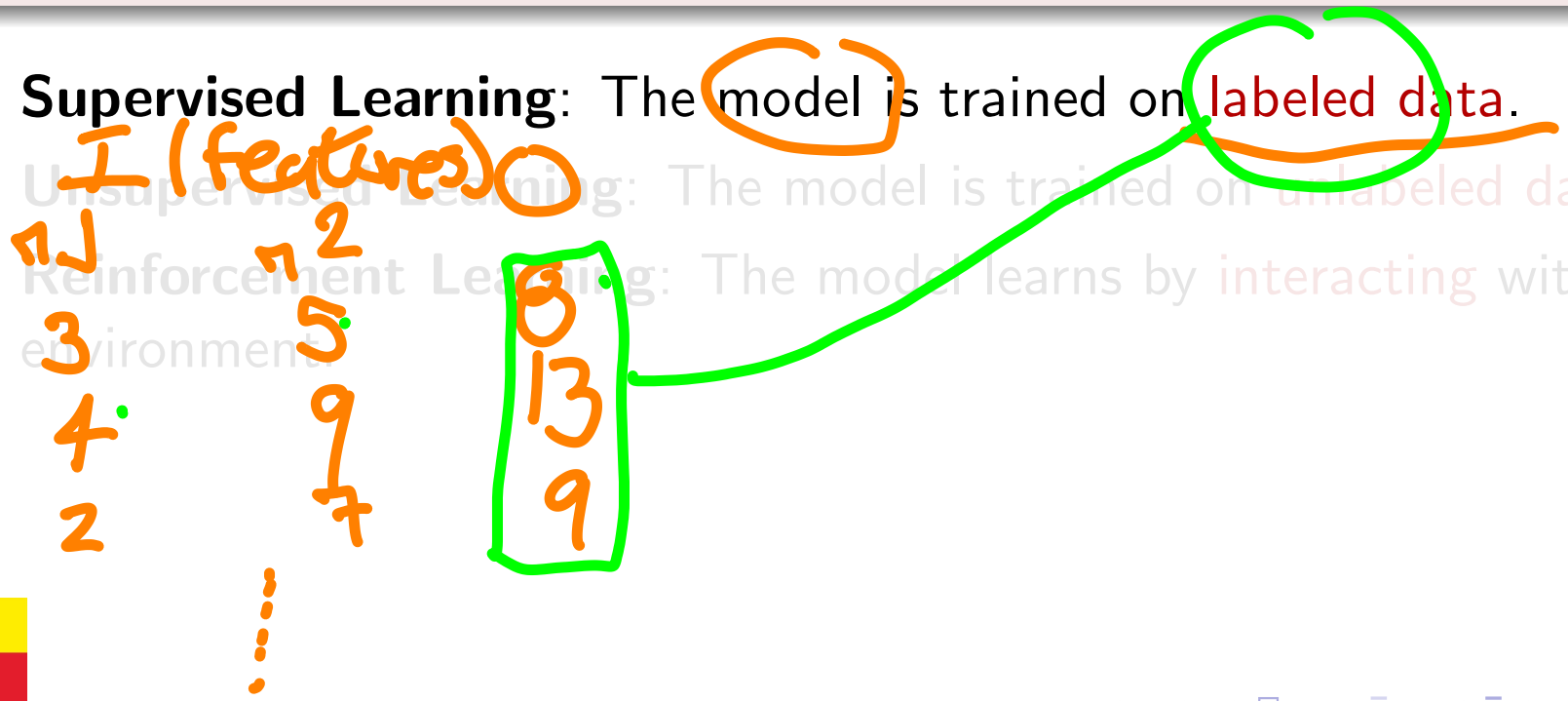


Key Concepts in Machine Learning

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.

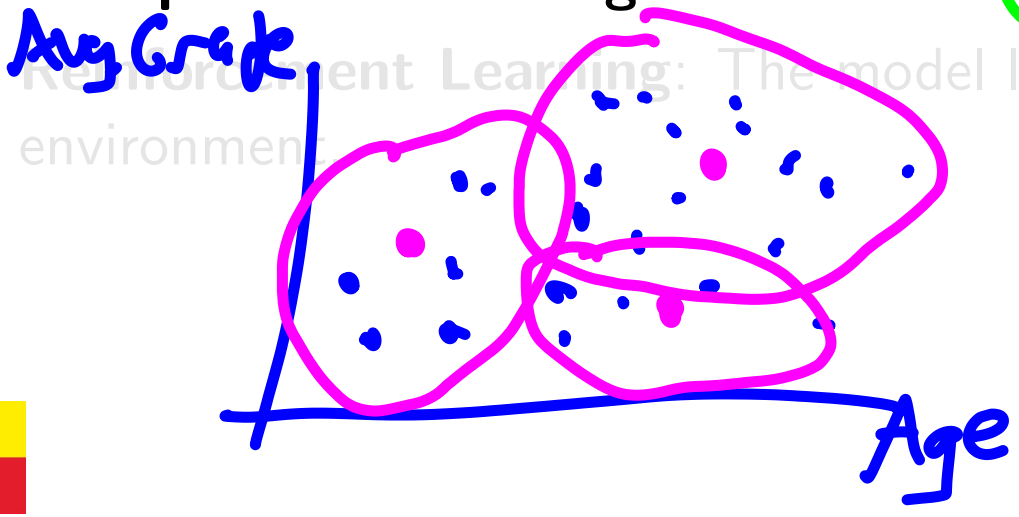


Key Concepts in Machine Learning

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.

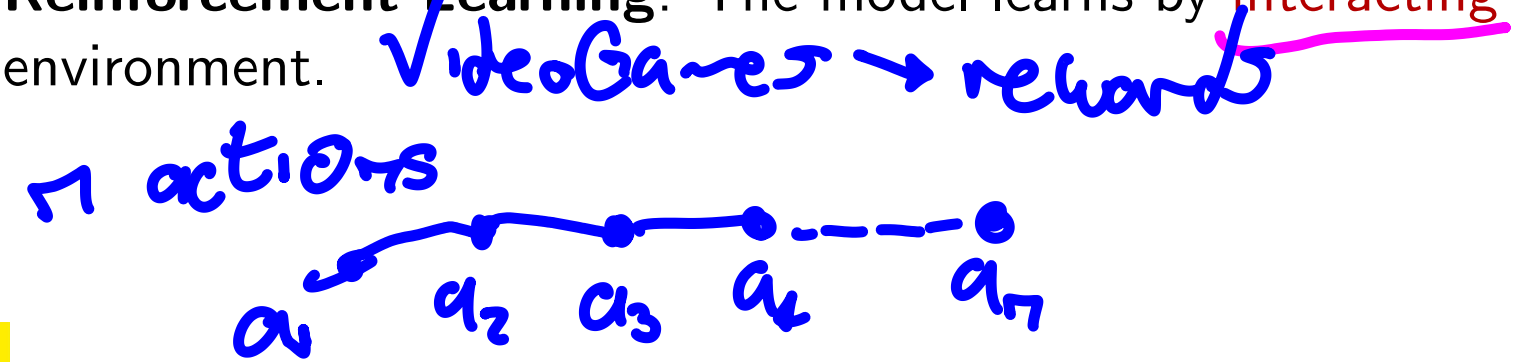


Key Concepts in Machine Learning

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`.

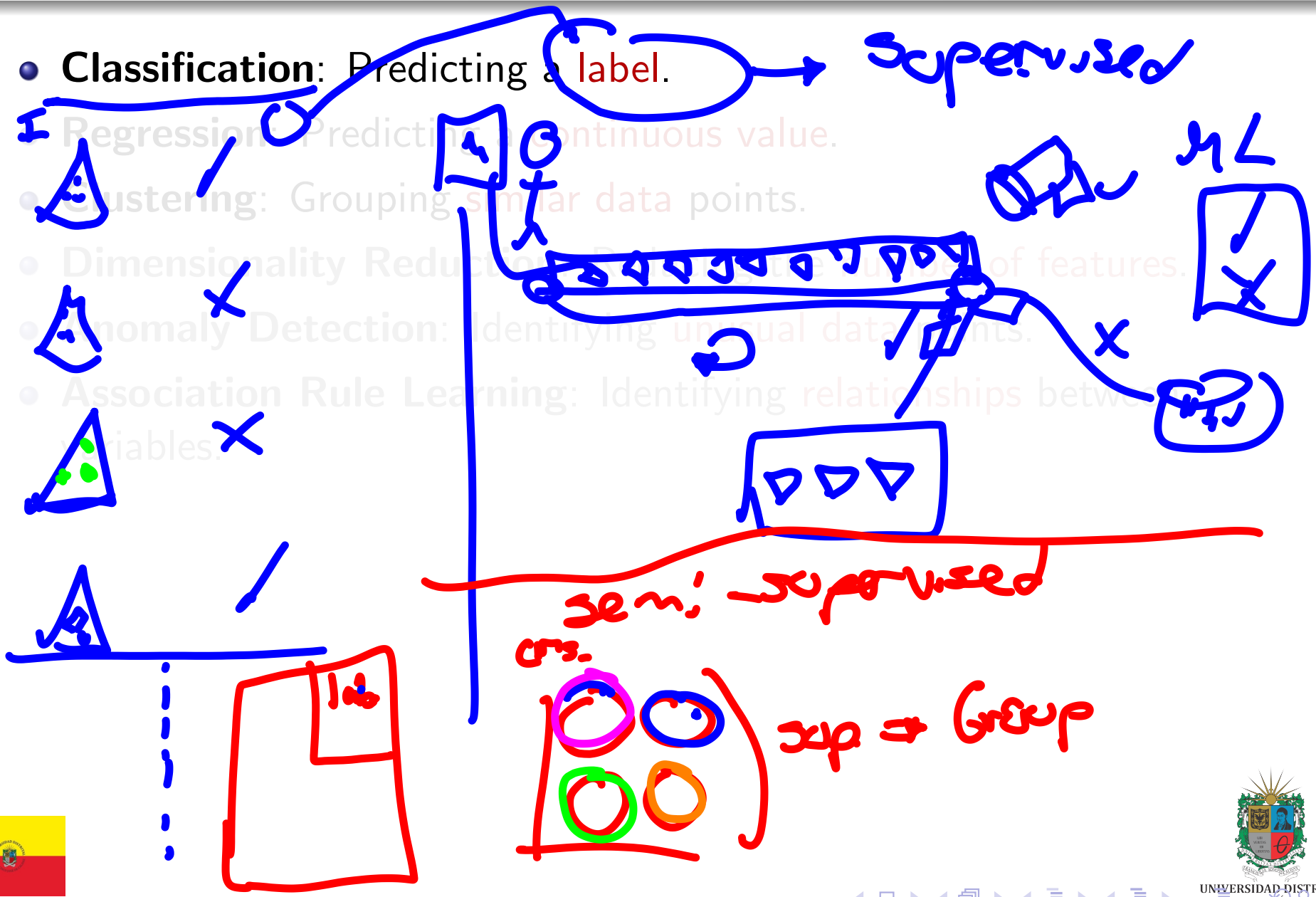
- **Supervised Learning**: The model is trained on labeled data.
- **Unsupervised Learning**: The model is trained on unlabeled data.
- **Reinforcement Learning**: The model learns by interacting with an environment.



UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

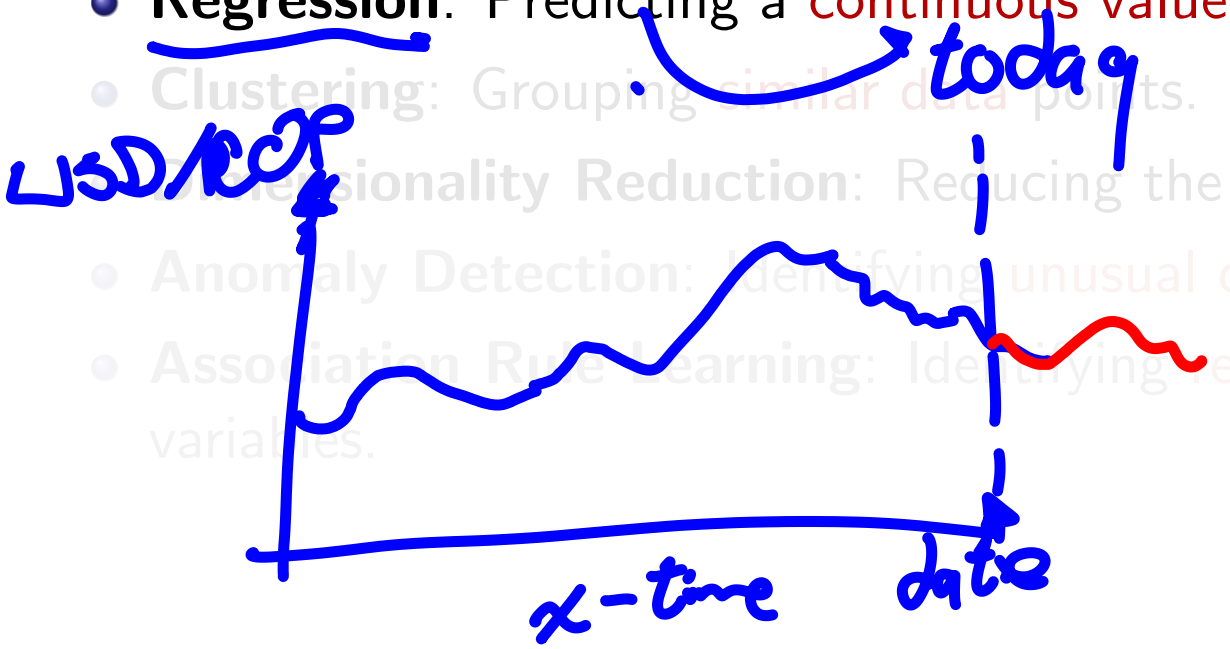
Typical Machine Learning Problems

- **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.

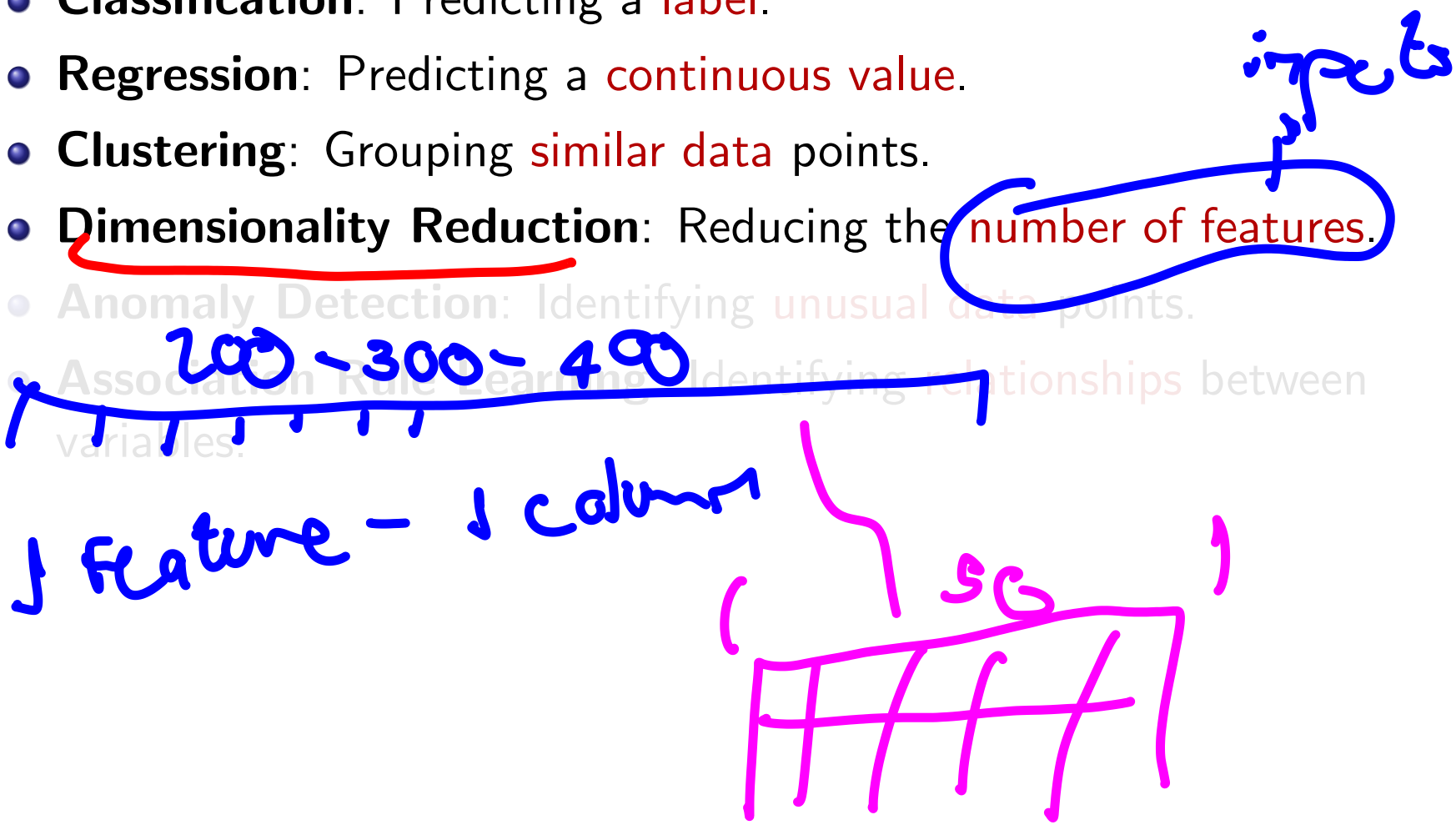


Typical Machine Learning Problems

- **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.

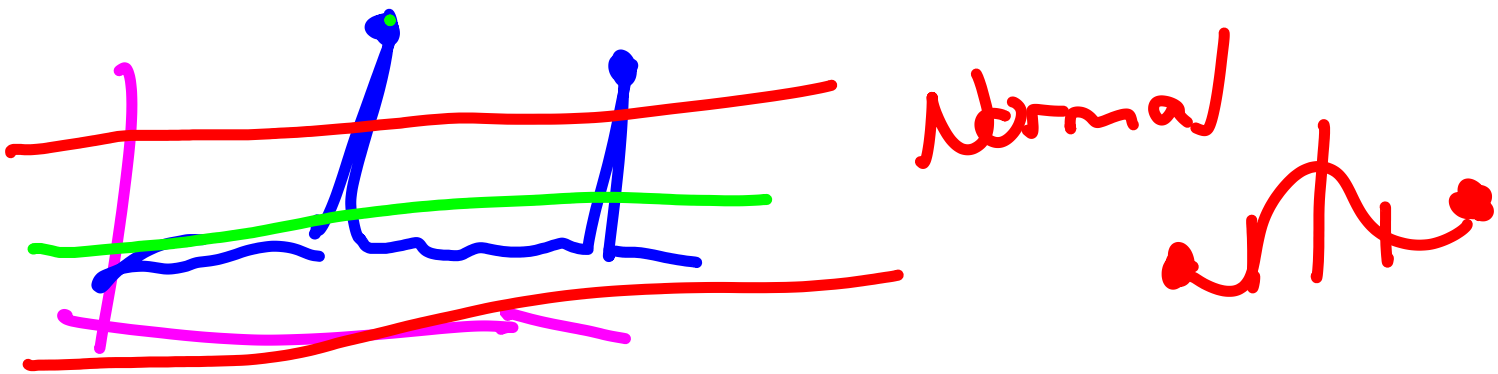
Typical Machine Learning Problems

- **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

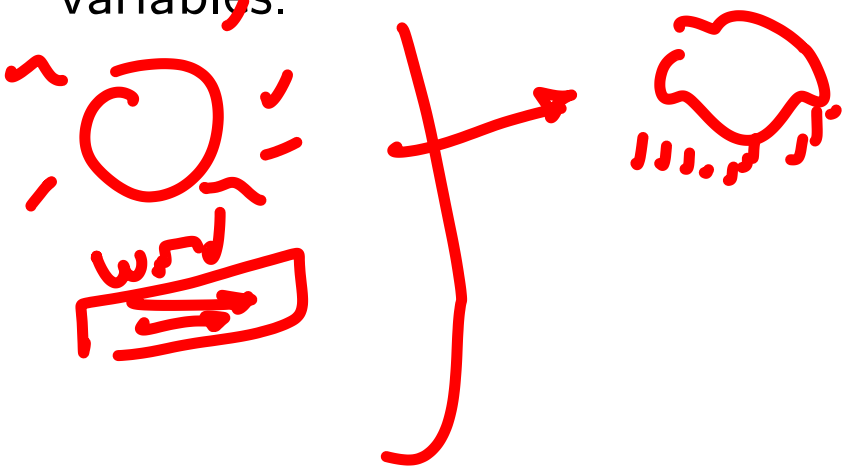
- **Classification**: Predicting a label.
- **Regression**: Predicting a continuous value.
- **Clustering**: Grouping similar data points.
- **Dimensionality Reduction**: Reducing the number of features.
- **Anomaly Detection**: Identifying unusual data points.
- **Association Rule Learning**: Identifying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



UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

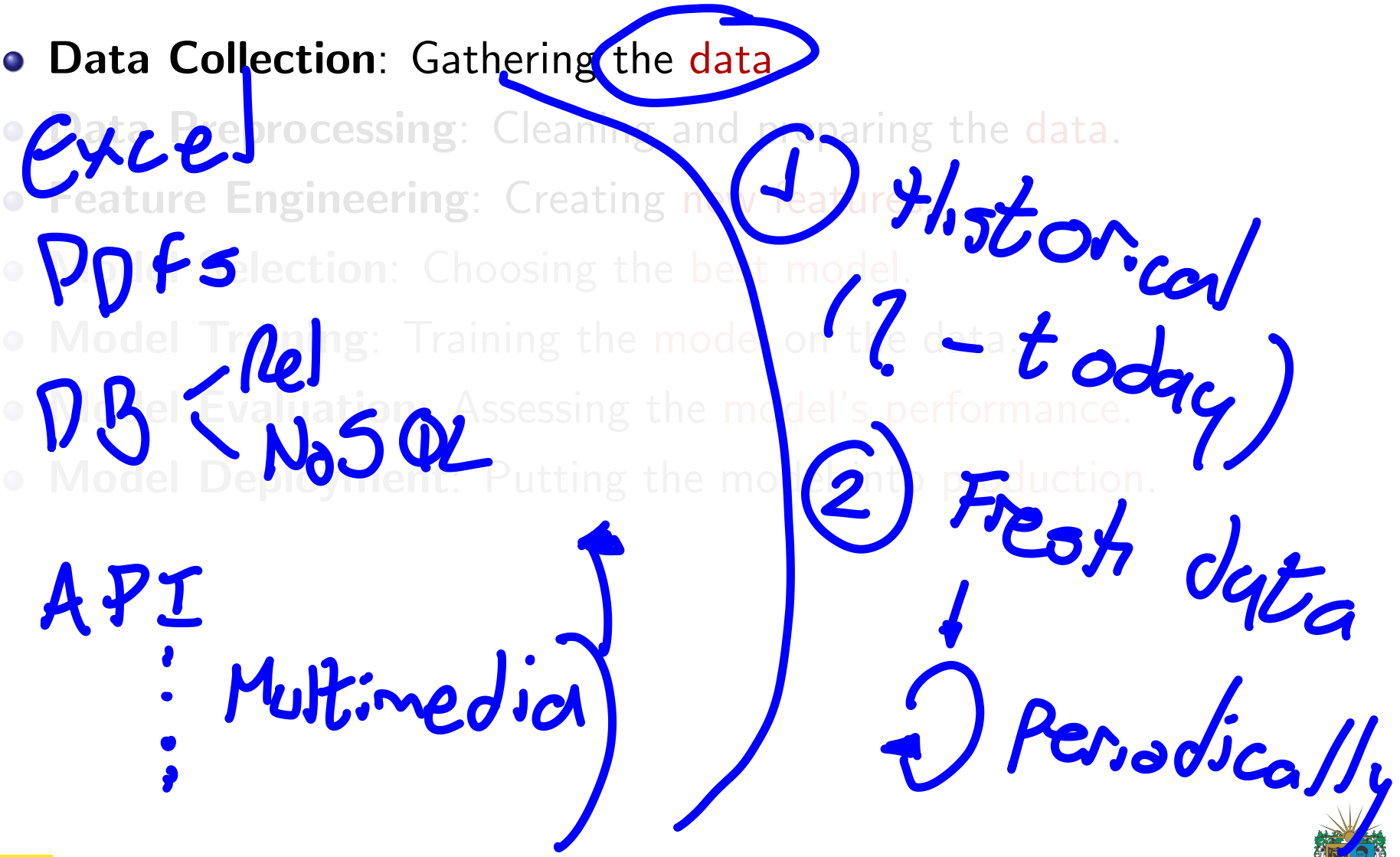
Typical Machine Learning Problems

- **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.



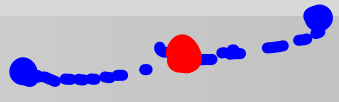
The Machine Learning Workflow

- **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.

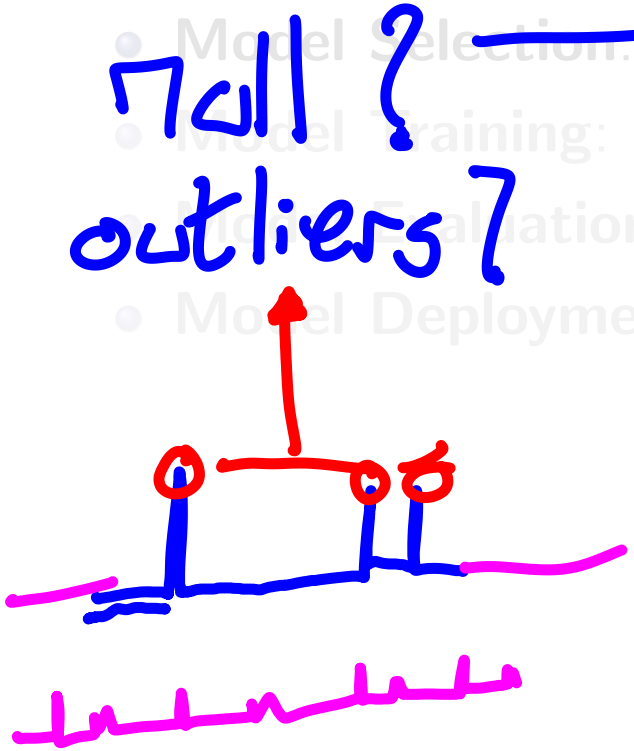


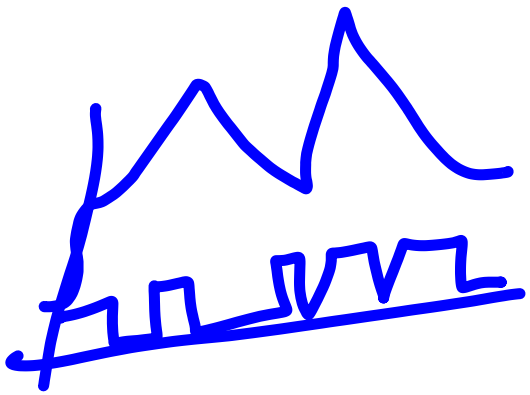
UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

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.

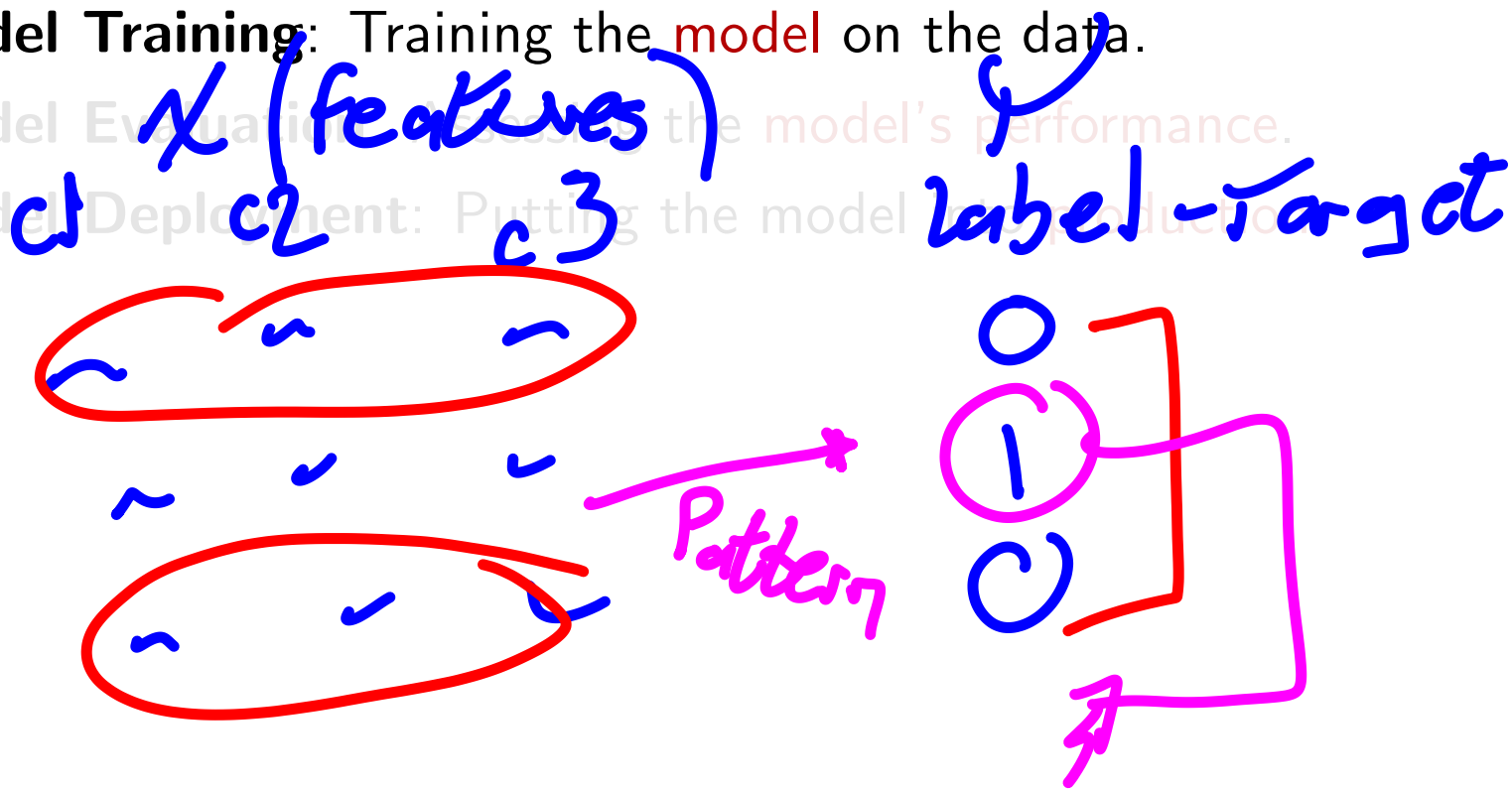
Semester	Start	Delay
8	2018-2	1
7	2022-1	1
9	2021-2	0
1	2025-1	1
2	2025-1	0

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.

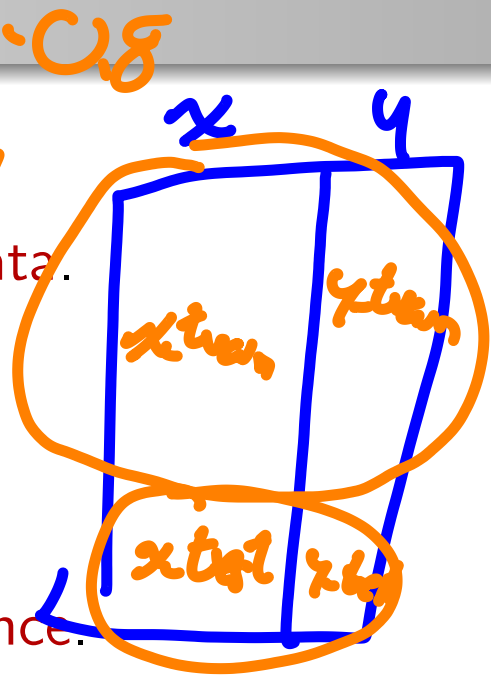
The Machine Learning Workflow

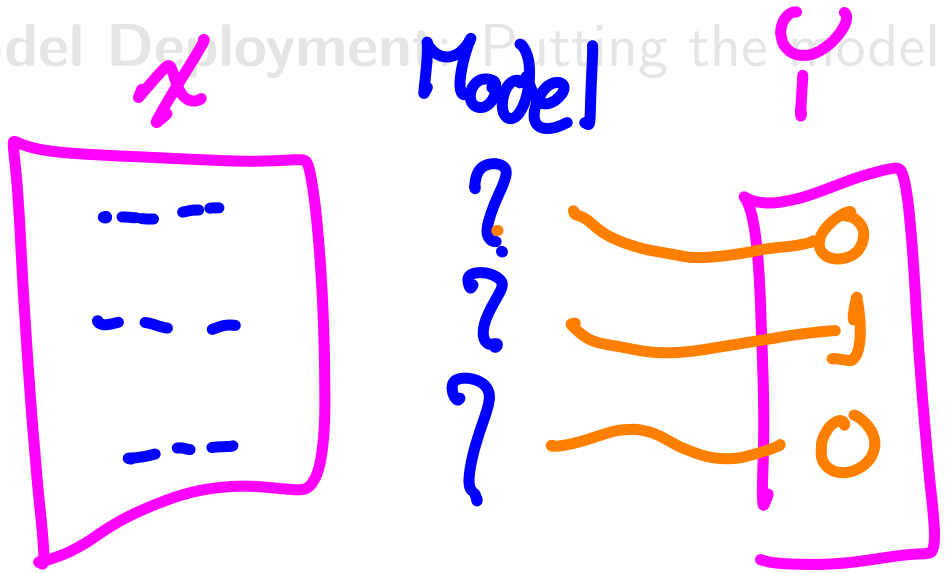
- **Data Collection**: Gathering the data.
- **Data Preprocessing**: Cleaning and preparing the data.
- **Feature Engineering**: Creating new features.
- **Model Selection**: Choosing the best model.
- **Model Training**: Training the model on the data.
- **Model Evaluation**: Assessing the model's performance.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



The Machine Learning Workflow

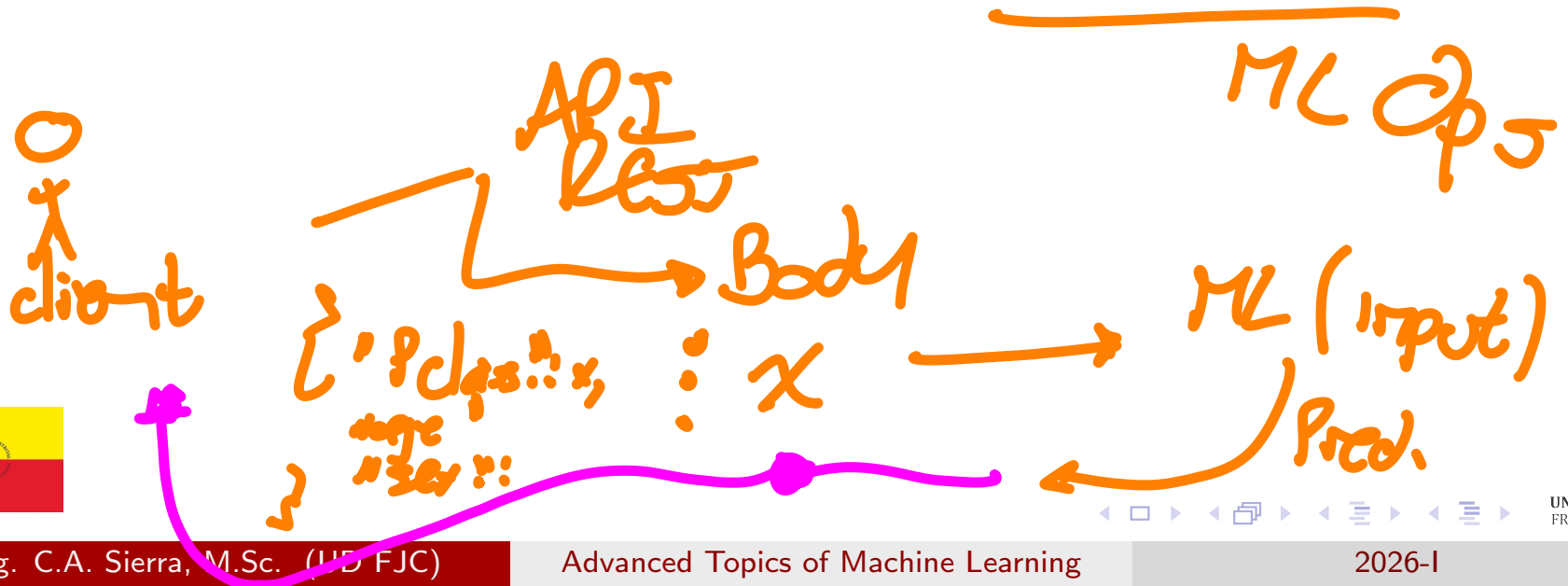
- **Data Collection**: Gathering the data.
- **Data Preprocessing**: Cleaning and preparing the data.
- **Feature Engineering**: Creating new features.
- **Model Selection**: Choosing the best model.
- **Model Training**: Training the model on the data.
- **Model Evaluation**: Assessing the model's performance.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





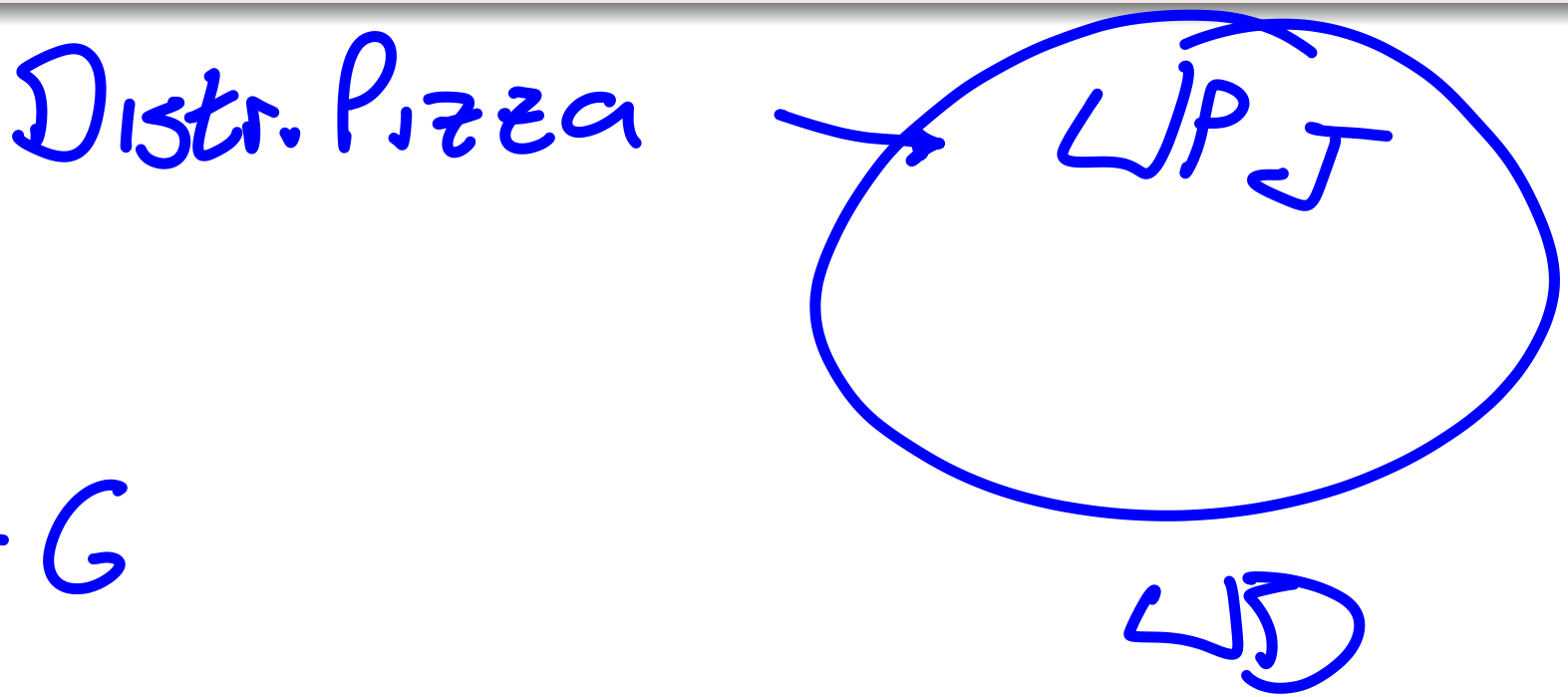
The Machine Learning Workflow

- **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.



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.



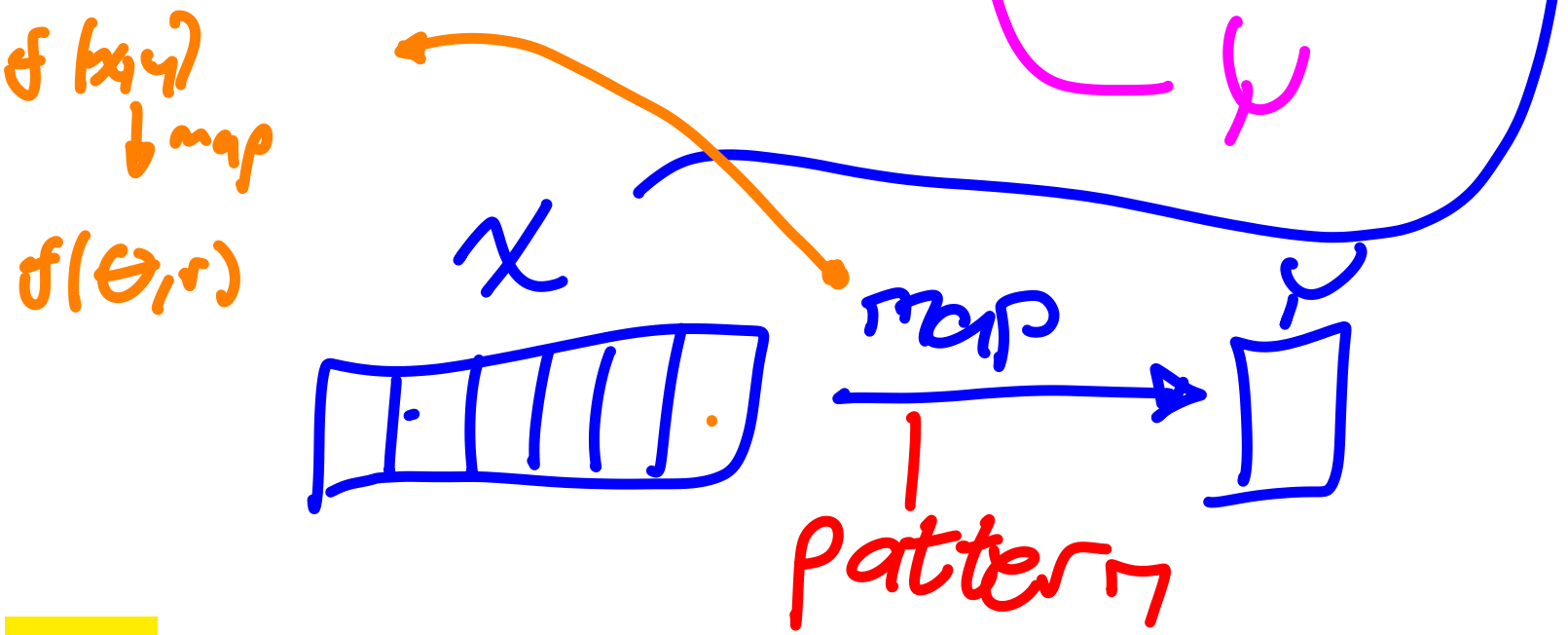
Outline

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

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

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

2+2=4 Train

7+4=11

29992+99532=? Test

Train — 80 20

3 — 2 1 Test

4 — 3 1

9 — 7 2

10 — 8 2

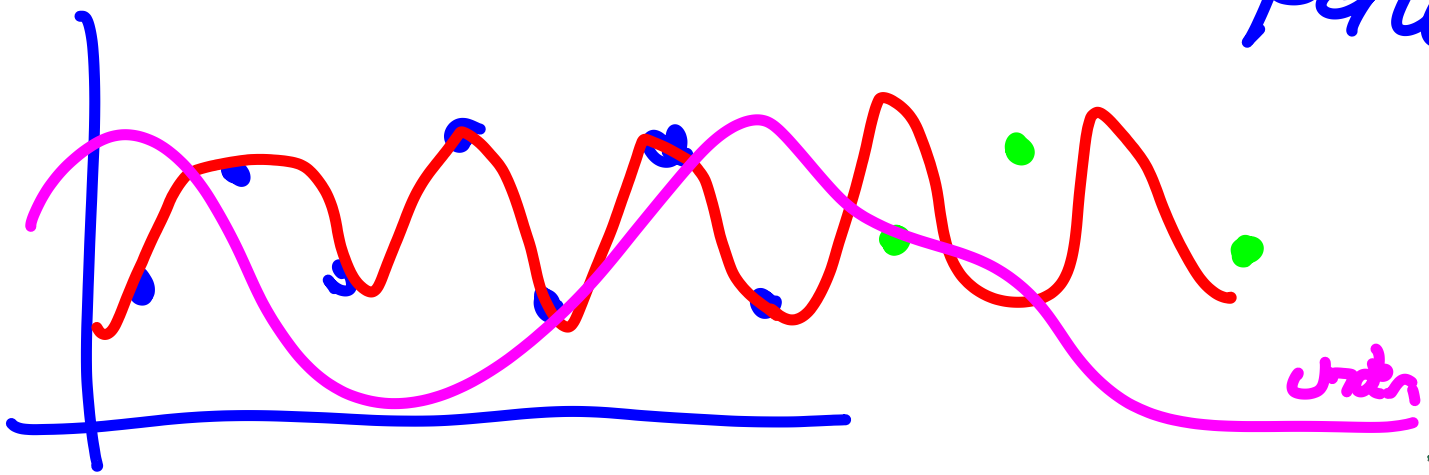
Overfitting and Underfitting

Overfitting

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

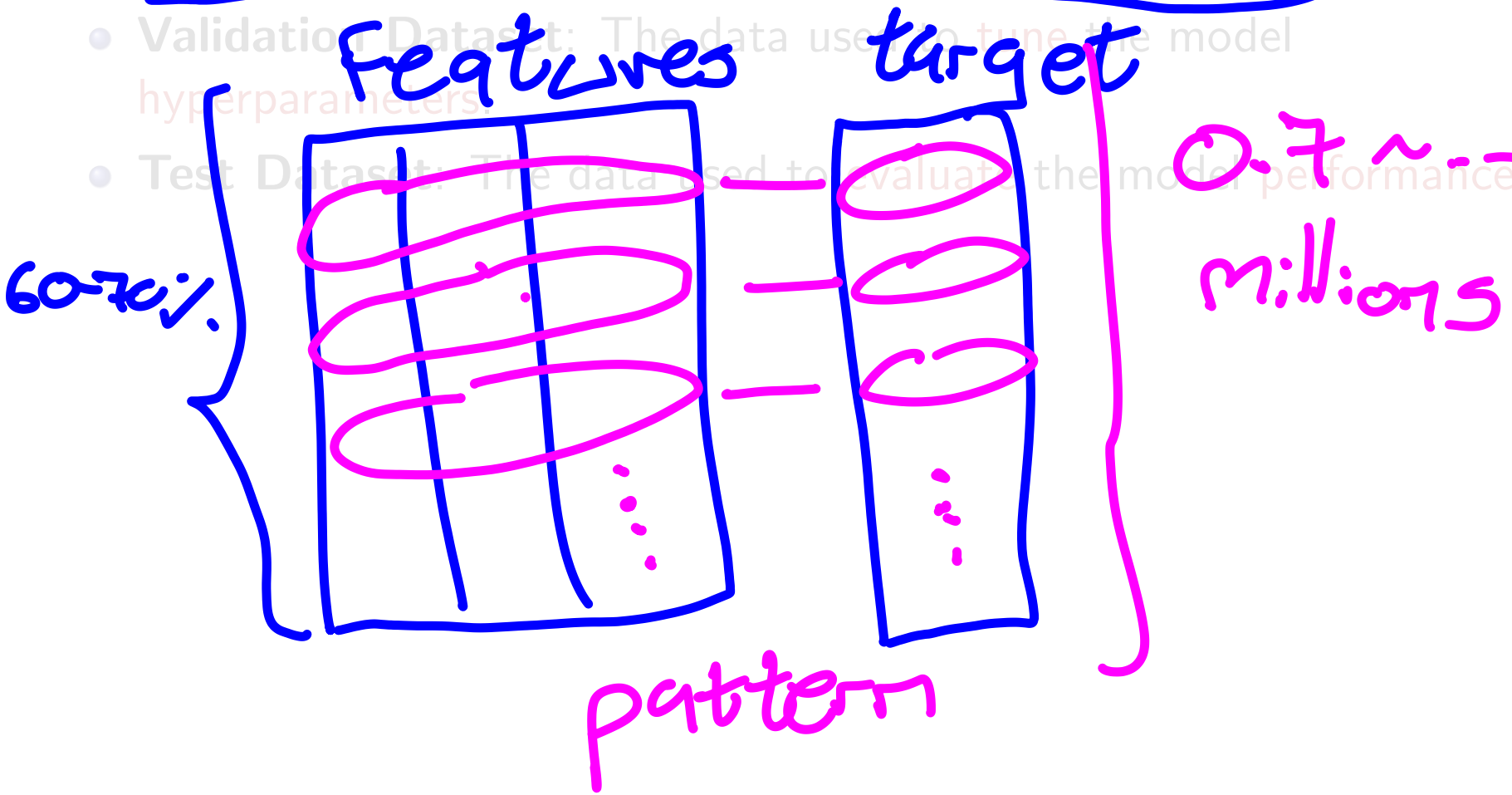
Underfitting

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



Supervised Learning Datasets

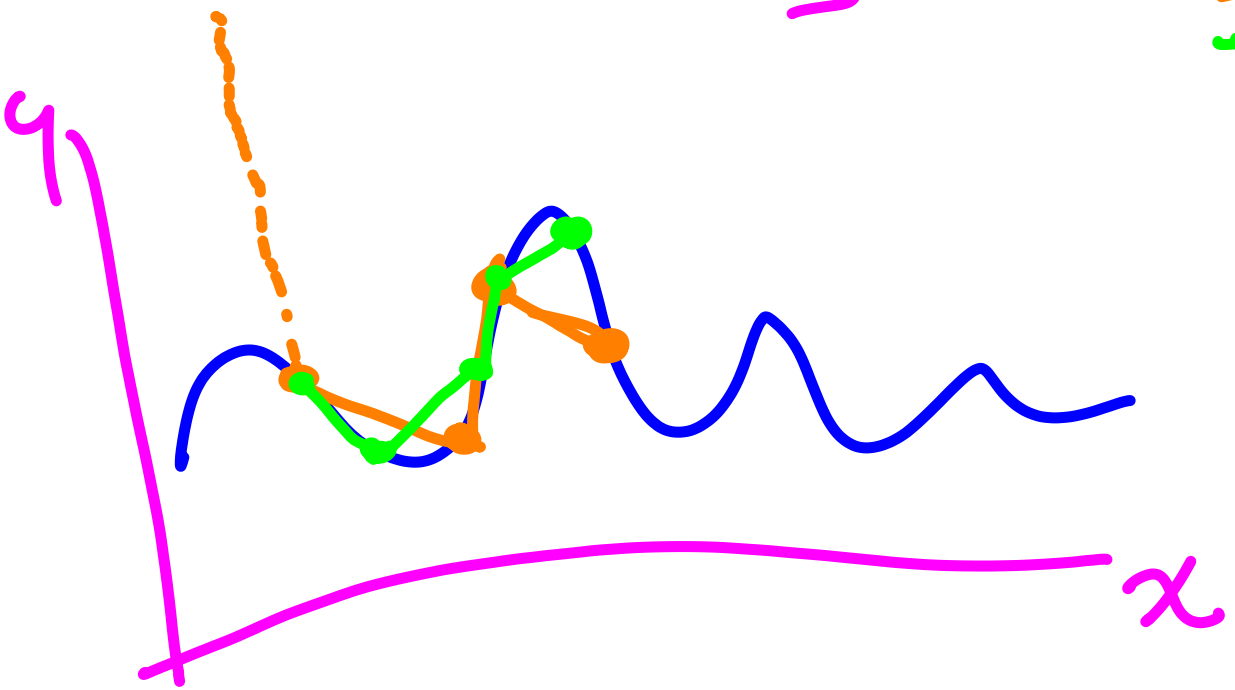
- **Training Dataset**: The data used to train the model.
- **Validation Dataset**: The data used to tune the model hyperparameters.
- **Test Dataset**: The data used to evaluate the model performance.



UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

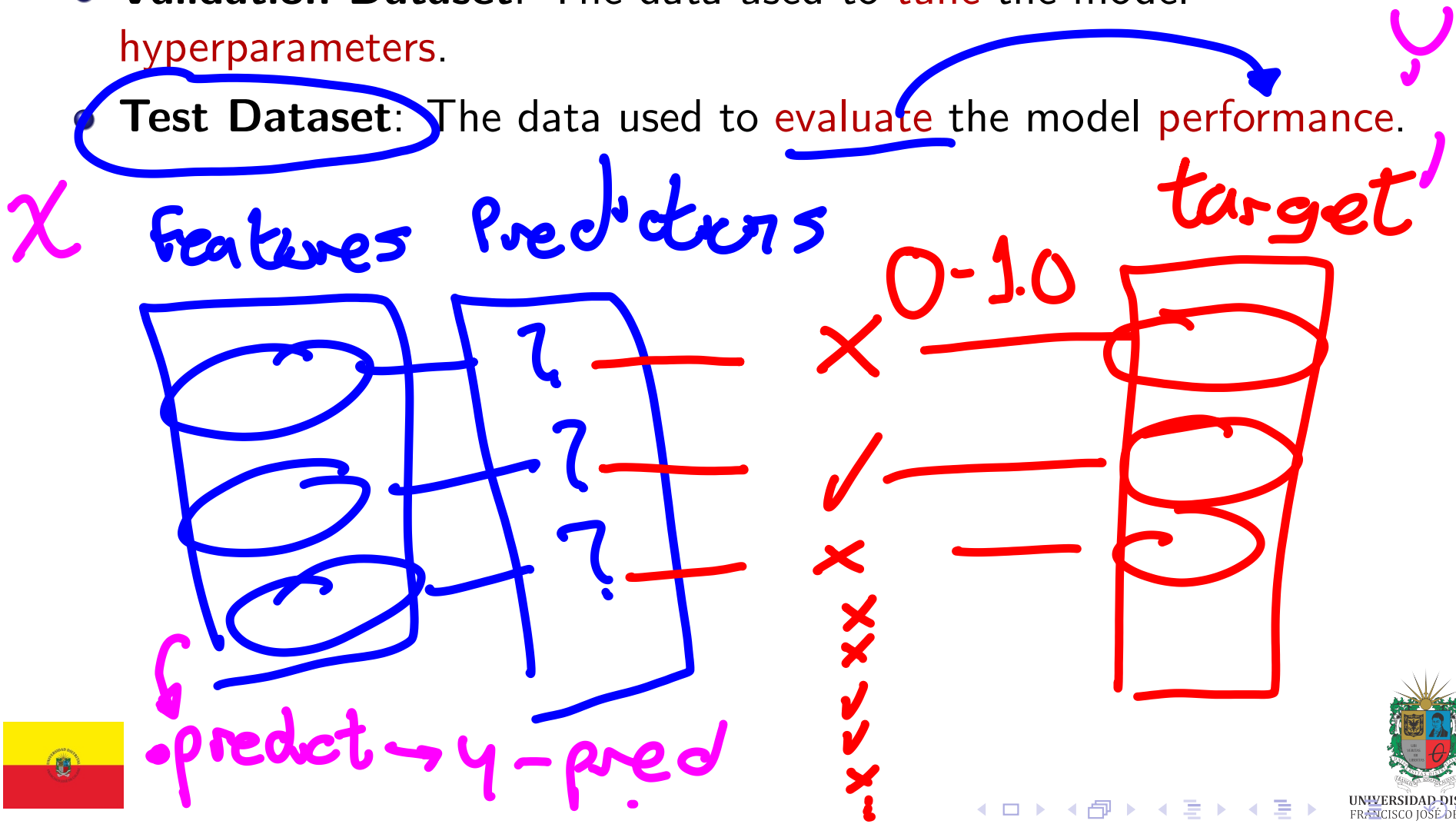
Supervised Learning Datasets

- **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.



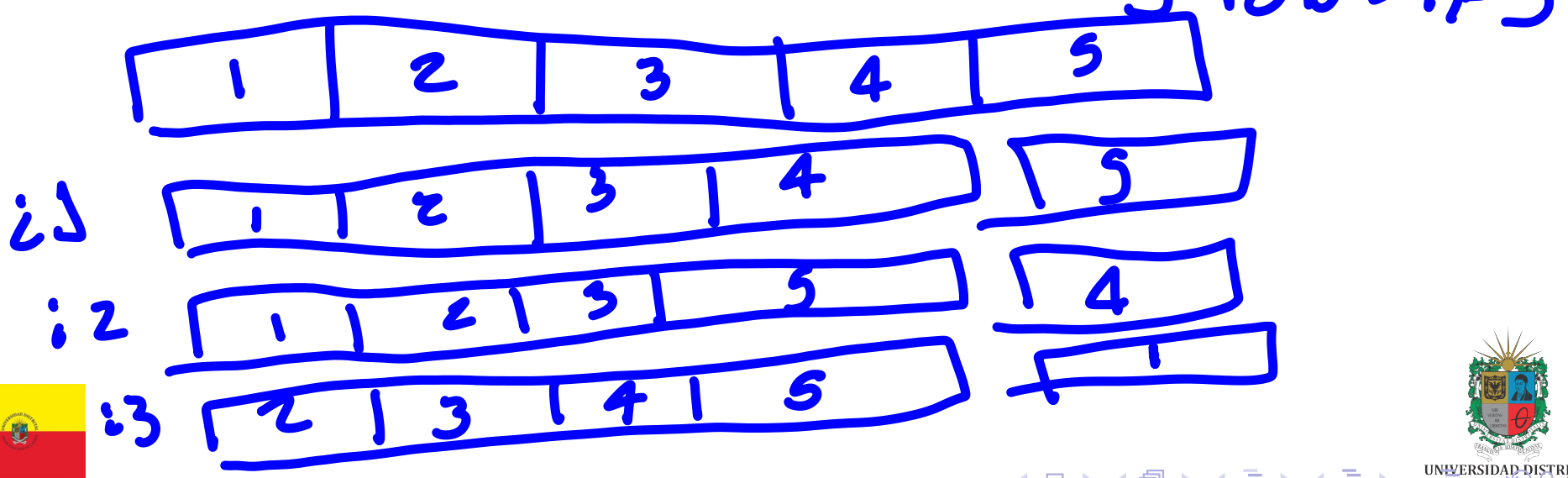
Supervised Learning Datasets

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



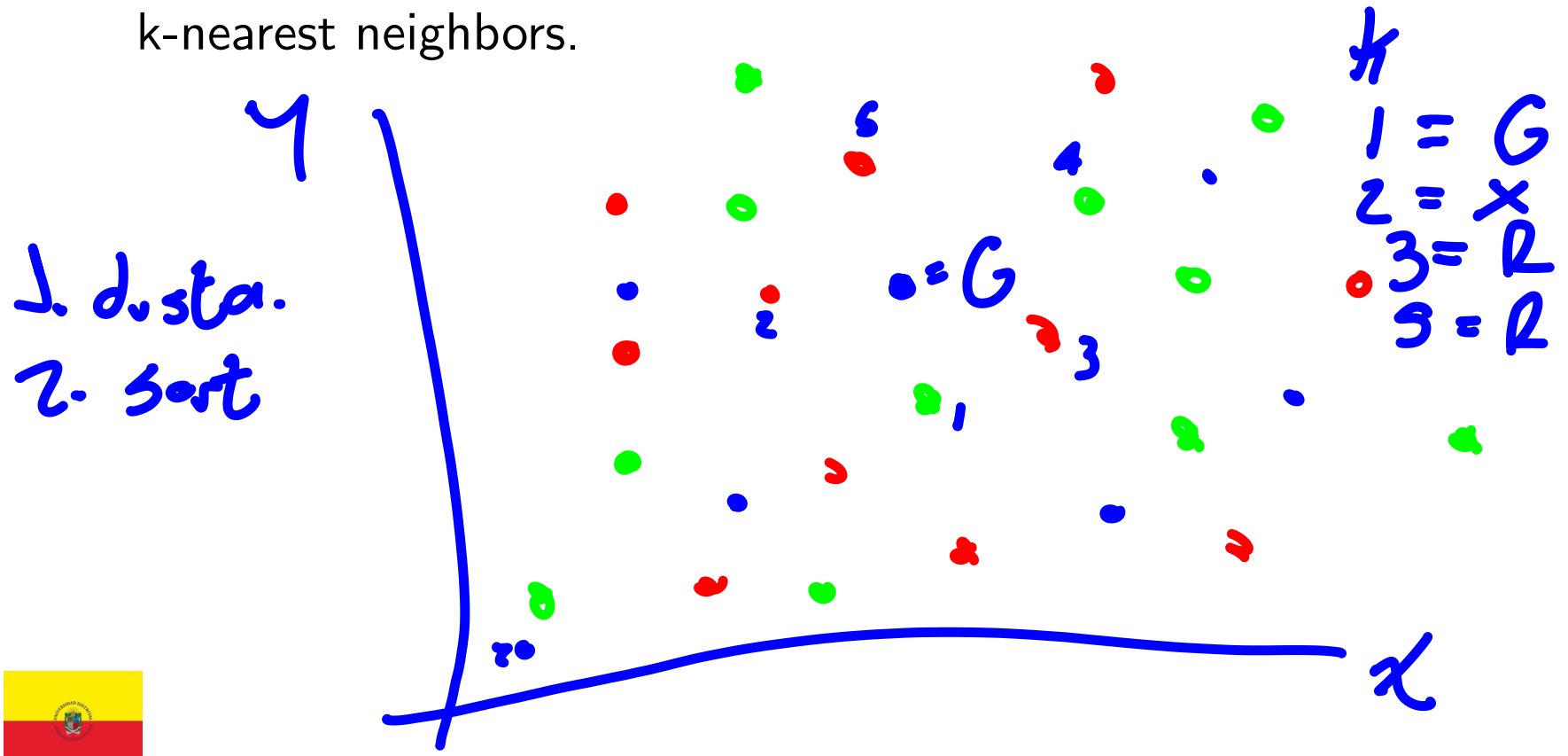
Cross-Validation

- **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.



K-Nearest Neighbors

- **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.



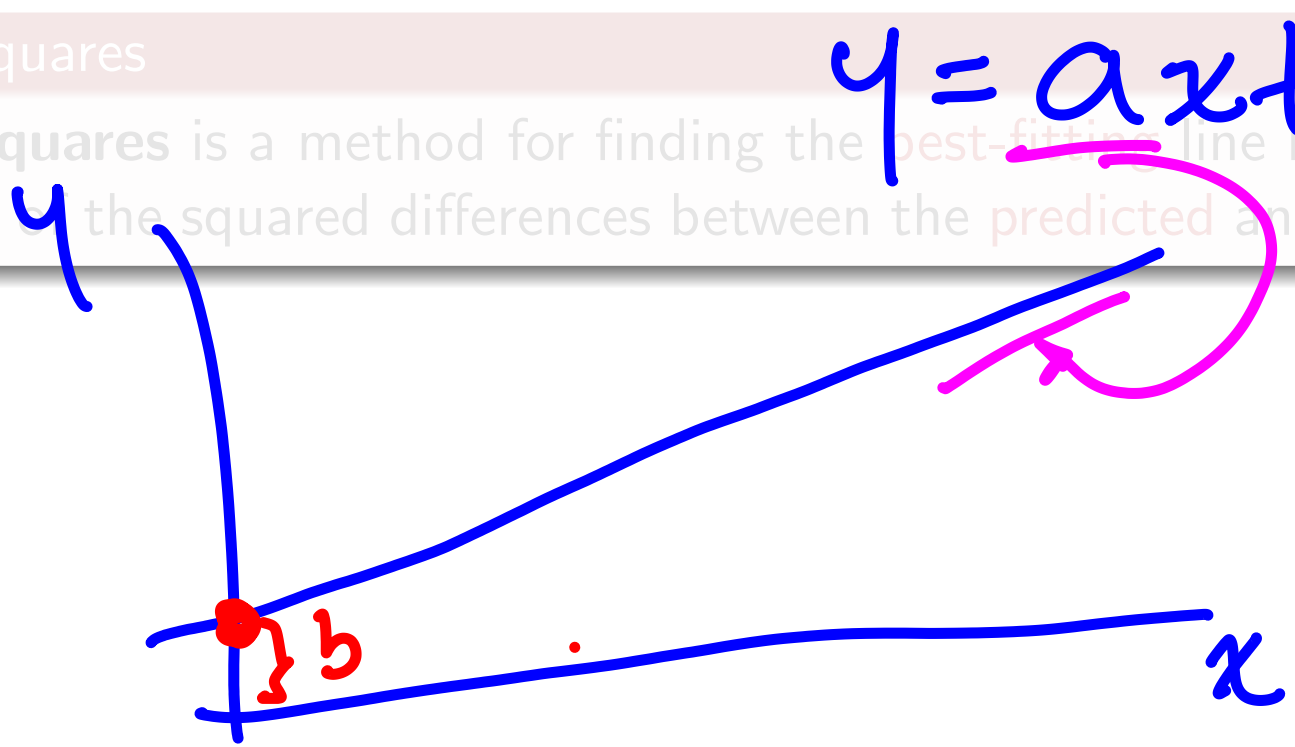
Linear Regression with Least Squares

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.



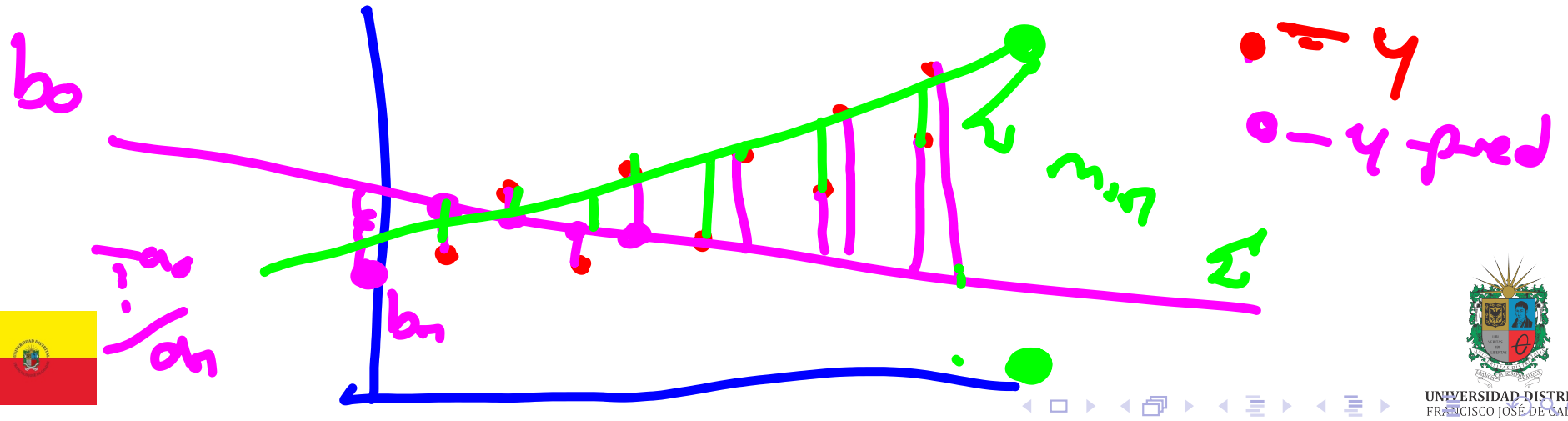
Linear Regression with Least Squares

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.

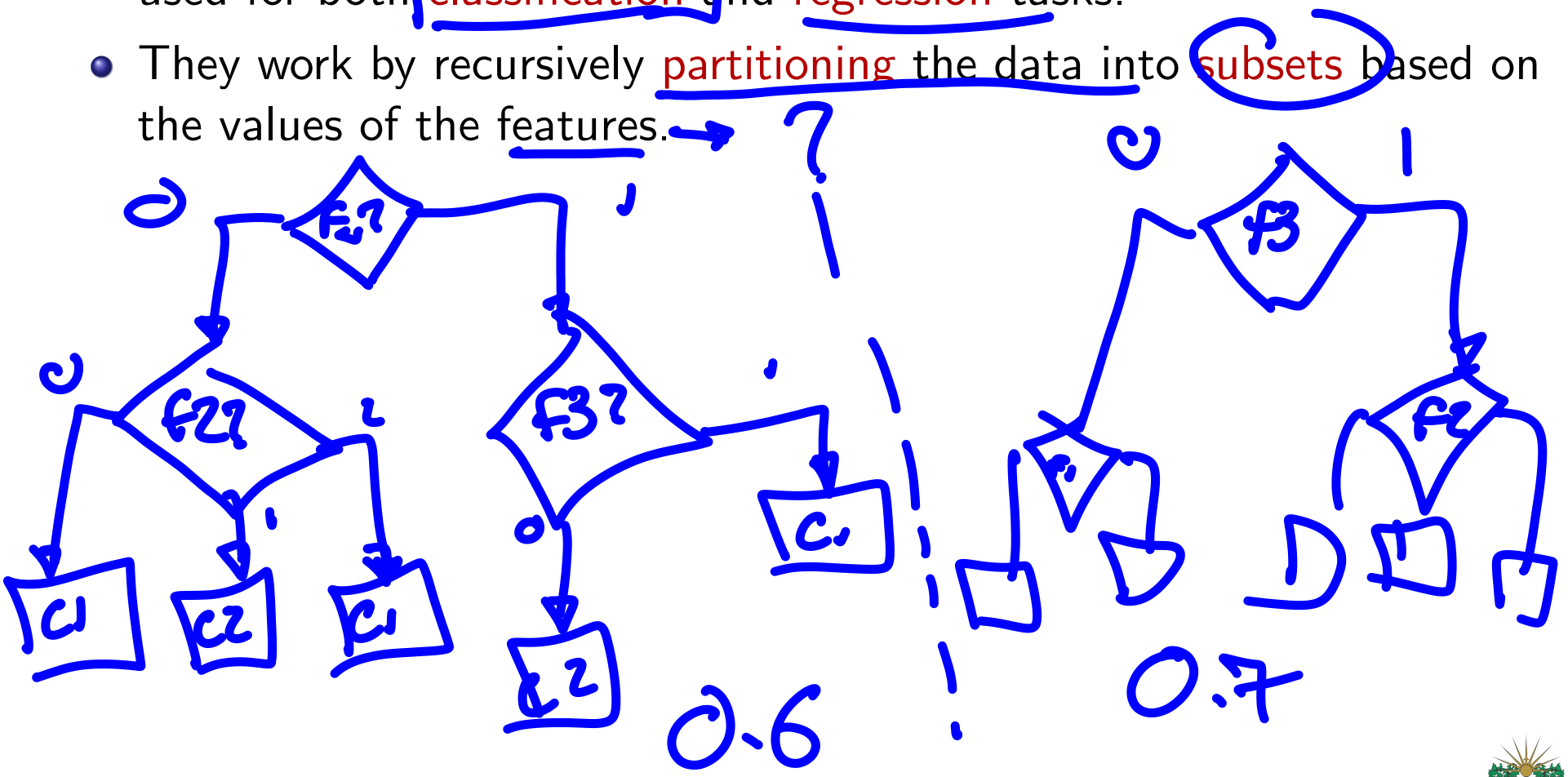
Least Squares

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.



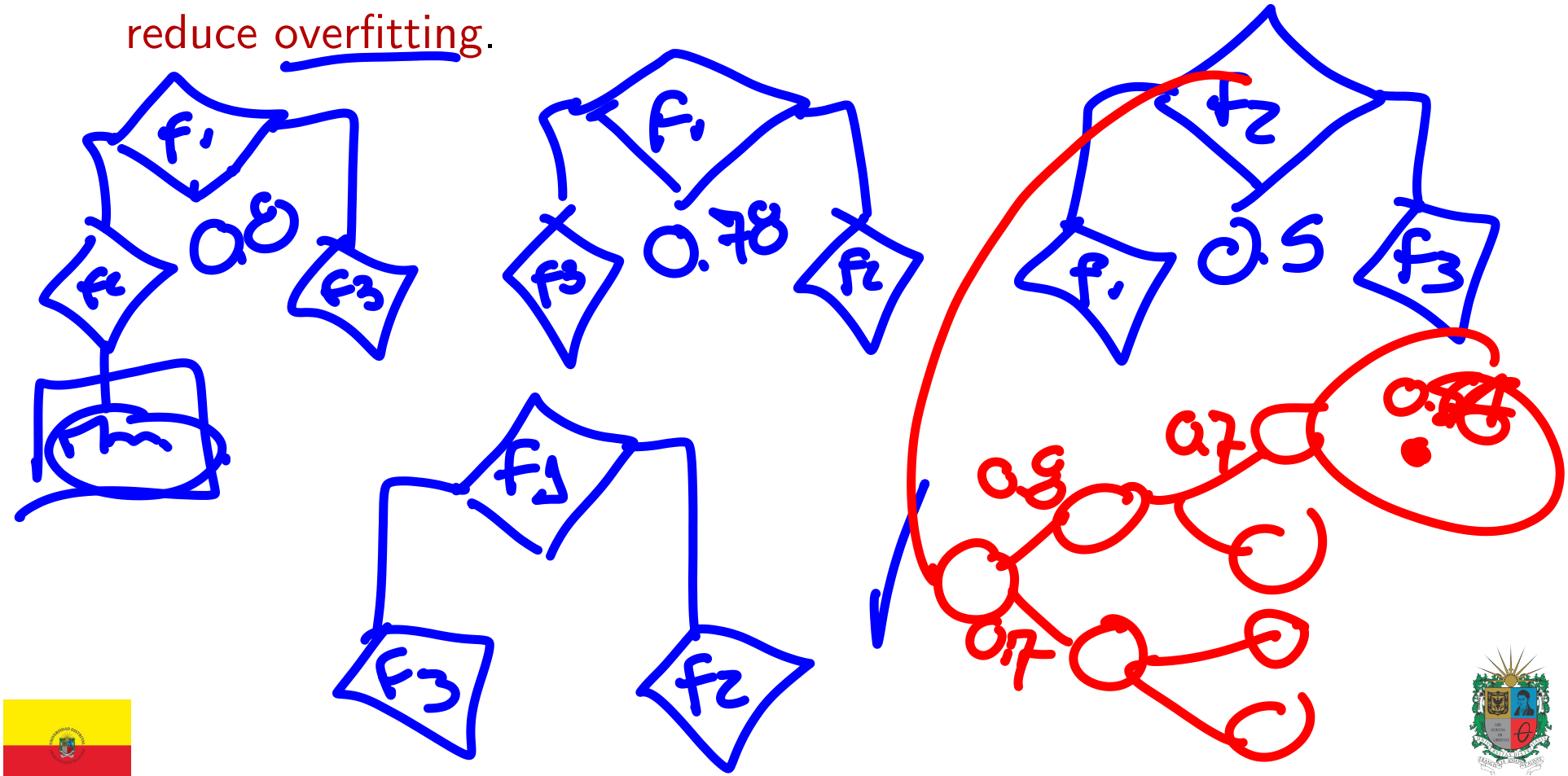
Decision Trees

- **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.



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.



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

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Introduction to Unsupervised Machine Learning

Definition

- **Unsupervised learning** is a type of machine learning where the model is trained on unlabeled data.
- It involves finding patterns and relationships in the data without any predefined labels.

Clustering

- **Clustering** is a type of unsupervised learning that involves grouping similar data points together.
- Common clustering algorithms include k-means, hierarchical clustering, and DBSCAN.

K-means

- **K-means** is a popular clustering algorithm that partitions the data into K distinct clusters based on feature similarity.
- It works by iteratively assigning data points to the nearest cluster centroid and updating the centroids based on the assigned points.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Anomaly Detection

- **Anomaly detection** is a type of unsupervised learning that involves identifying unusual data points in a dataset.
- Common anomaly detection algorithms include Isolation Forest, One-Class SVM, and Autoencoders.

Autoencoders

- **Autoencoders** are a type of neural network used for unsupervised learning.
- They work by encoding the input data into a lower-dimensional representation and then decoding it back to the original data.

Dimensionality Reduction

- **Dimensionality reduction** is a technique for reducing the number of features in a dataset while retaining as much information as possible.
- Common dimensionality reduction techniques include Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE).

Principal Component Analysis (PCA)

- **Principal Component Analysis (PCA)** is a statistical technique used for dimensionality reduction.
- It works by transforming the data into a new coordinate system where the greatest variance lies along the first principal component, the second greatest variance along the second principal component, and so on.

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Outline

Thanks!

Questions?

Repo: *https://github.com/EngAndres/ud-public/tree/main/courses/machine-learning*

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS