Introduction to Machine Learning Classification

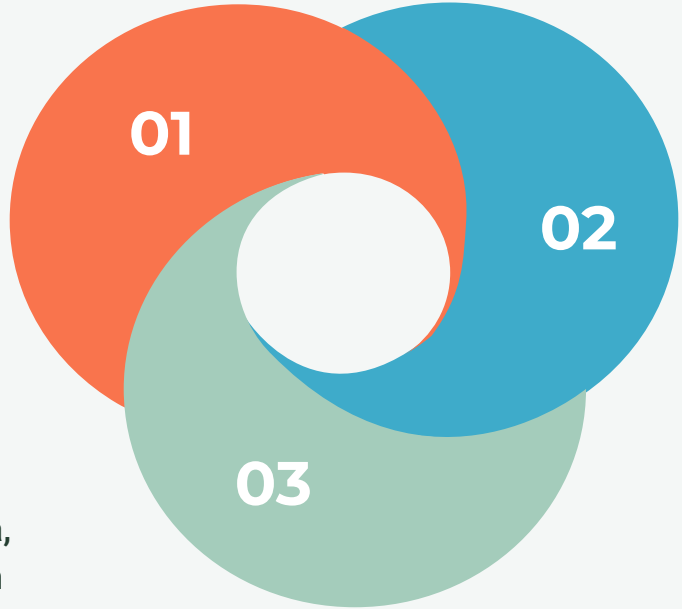
This presentation provides an overview of Machine Learning classification and its applications.

Meher Kharbachi

What is Machine Learning Classification?

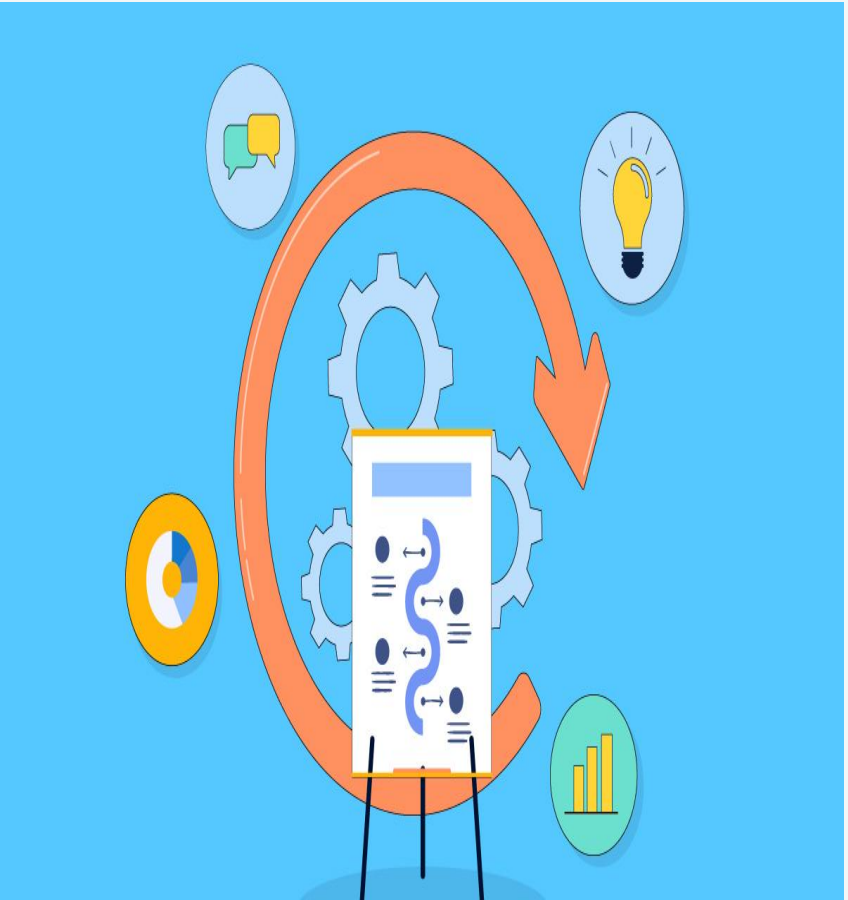
Classification is a fundamental task in machine learning that involves categorizing data into predefined classes or labels based on their features.

Classification enables the automatic categorization of data, making it a crucial component in various real-world applications.



It falls under the category of supervised learning, where the algorithm learns from labeled training data to make predictions or decisions.

Key components of classification



Features and Labels

The process involves identifying relevant features in the data and assigning appropriate labels to them.

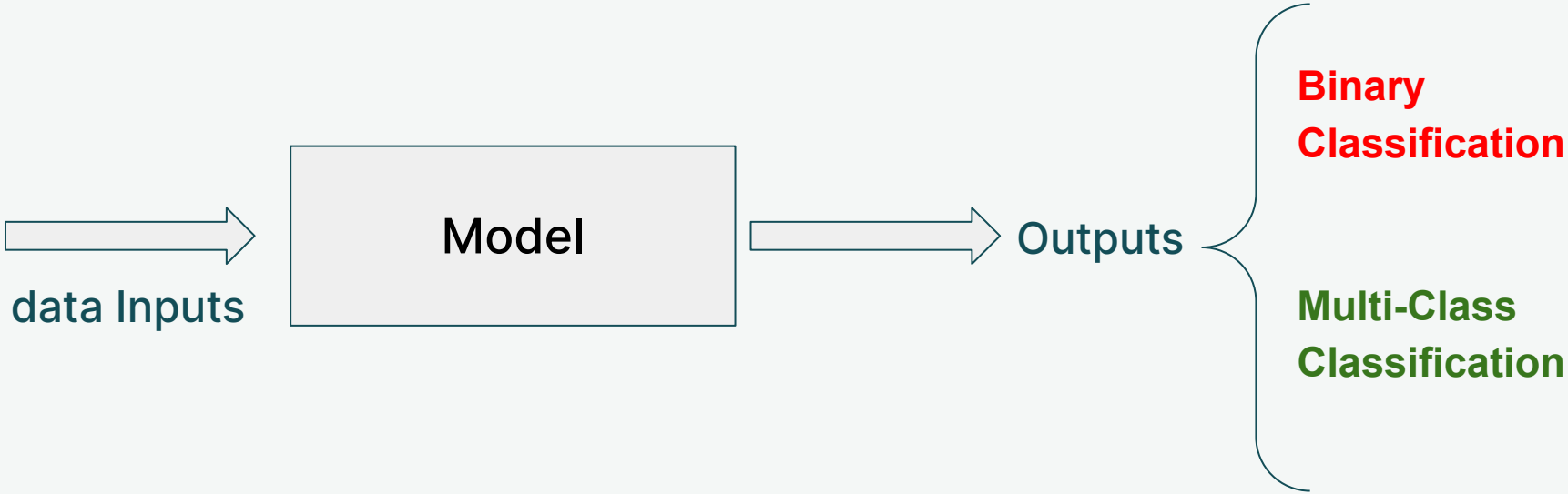
Training and testing

The model is trained on a labeled dataset and then tested on new, unseen data to evaluate its accuracy.

Decision Boundaries

Classification algorithms create decision boundaries to distinguish between different classes in the feature space.

Types of Machine Learning Classification



Binary Classification

Binary classification involves categorizing data into two distinct classes or categories.

Example:

- Spam Detection (Spam vs Ham)
- Disease diagnosis (Cancer vs not Cancer)
- Sentiment Analysis (Happy , Angry, ..)

Multi-Class Classification

Multi-class classification assigns data points to more than two classes, making it suitable for scenarios with multiple categories.

Example:

- Handwritten digit recognition
- language identification
- image classification

Challenges in Classification

➔ Overfitting and Underfitting

- Overfitting occurs when a model is too complex, fitting noise instead of the underlying pattern
- Underfitting happens when a model is too simple, failing to capture the complexity of the data

➔ Imbalanced Data

- Imbalanced classification deals with datasets where the distribution of classes is significantly unequal.
- Techniques such as resampling and ensemble methods are employed to handle imbalanced datasets.

Common Classification Algorithms

Decision Trees

decision trees are tree-like structure used for classification tasks.

They partition the data into subsets based on the most significant attribute.

Random Forest

RF is an ensemble learning method based on decision trees.

It combines multiple decision trees to improve accuracy and reduce overfitting.

Support Vector Machines (SVM)

SVM is an algo for classification and regression tasks.

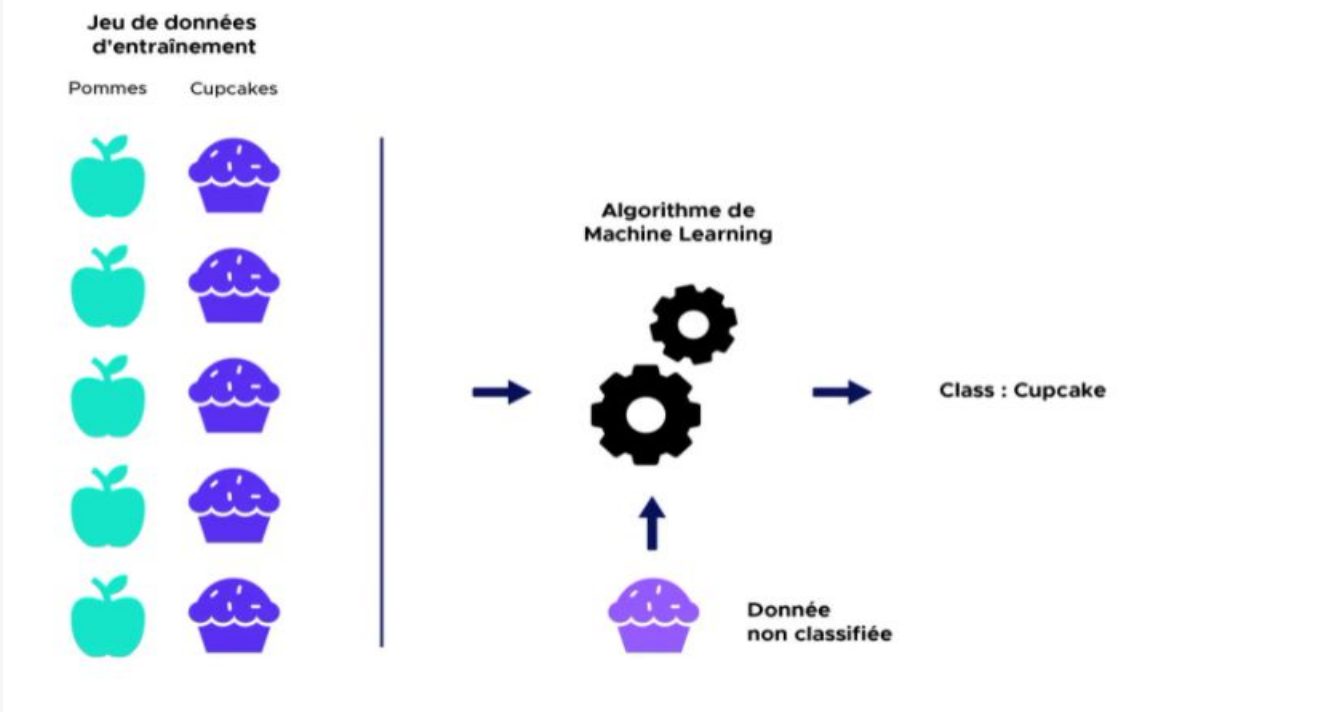
It finds the optimal hyperplane that best separates data points in a high dimensional space.

K-Nearest Neighbors (KNN)

KNN is a simple algo for classification and regression.

It classifies data points based on the majority vote of their k- nearest neighbors.

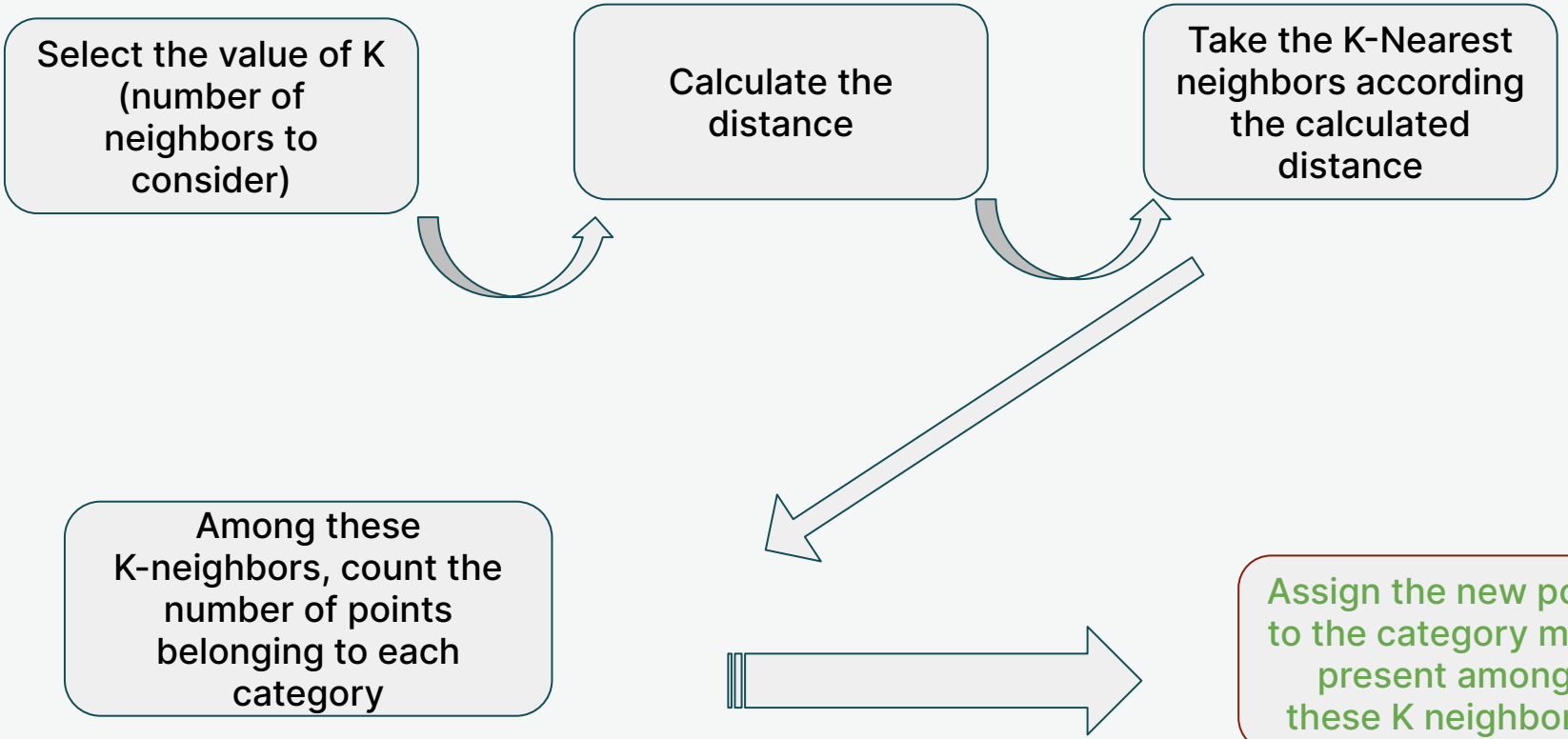
K-Nearest Neighbors (KNN)



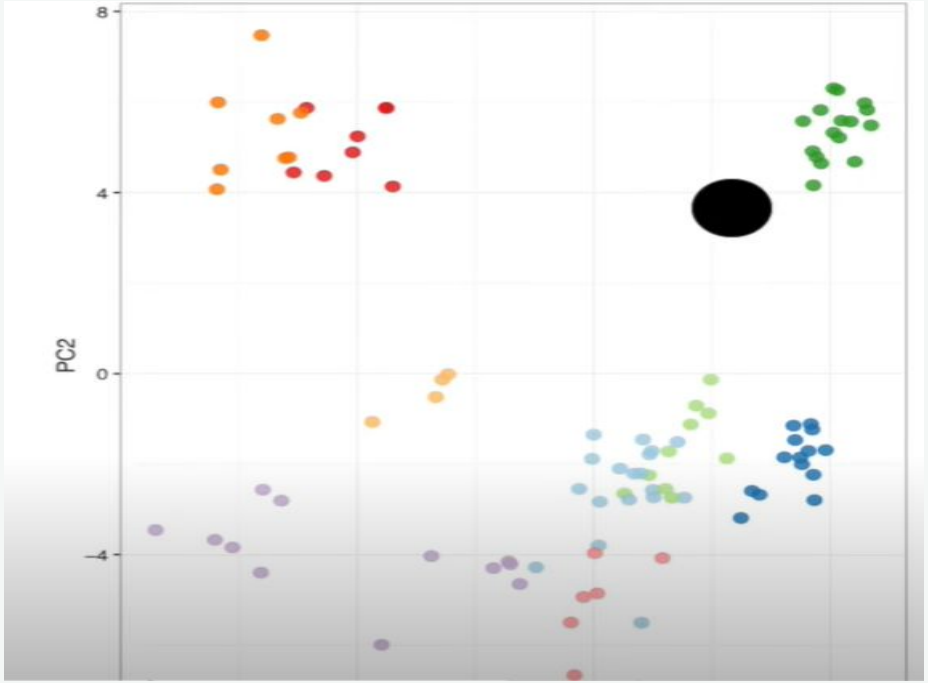
KNN is a simple algo for classification.

Belongs to the class of simple supervised learning algorithms.

K-Nearest Neighbors (KNN) Steps

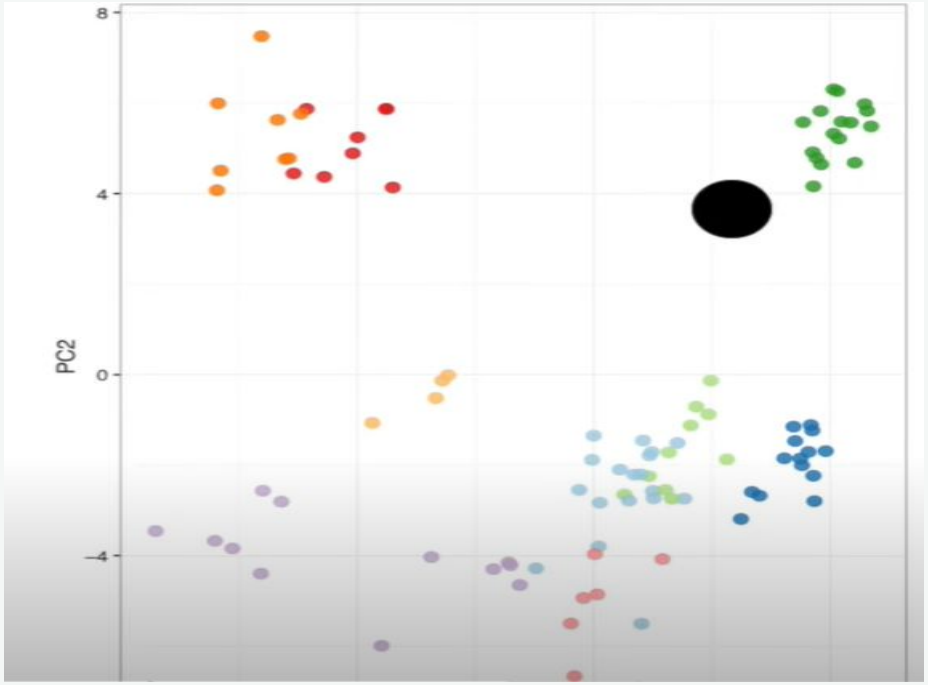


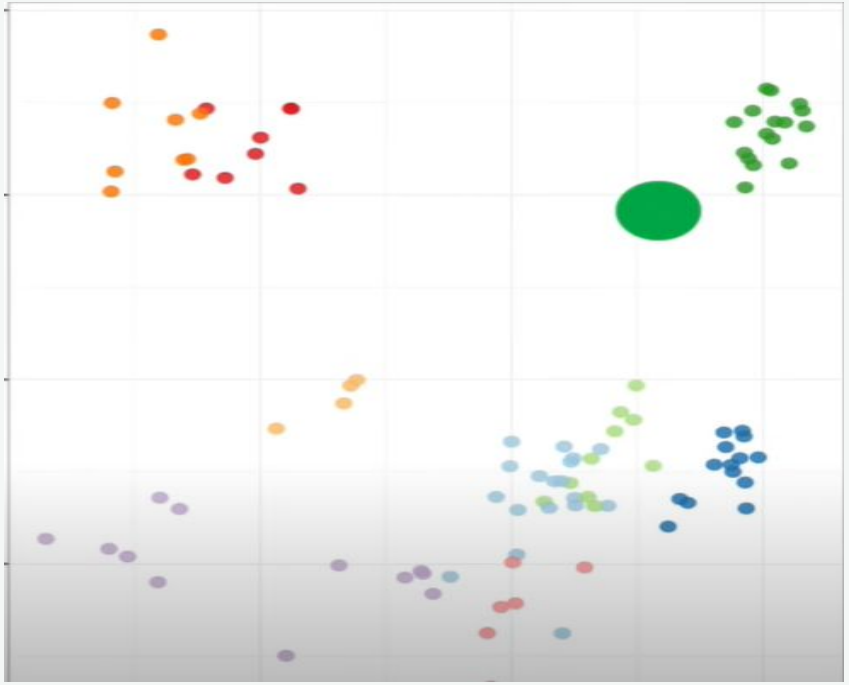
K-Nearest Neighbors (KNN) Example 1



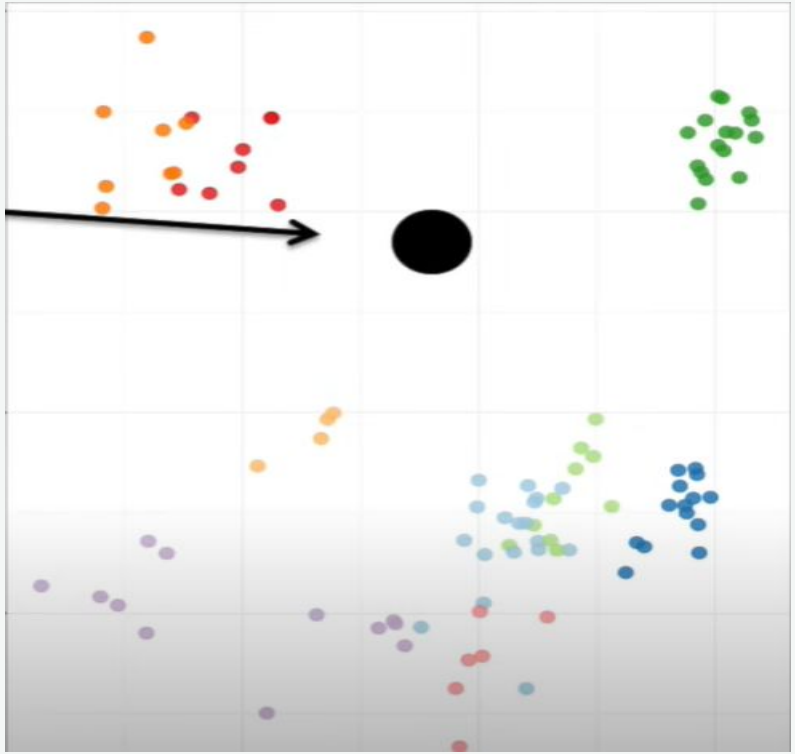
The new point should be ?????

K-Nearest Neighbors (KNN) Example 1



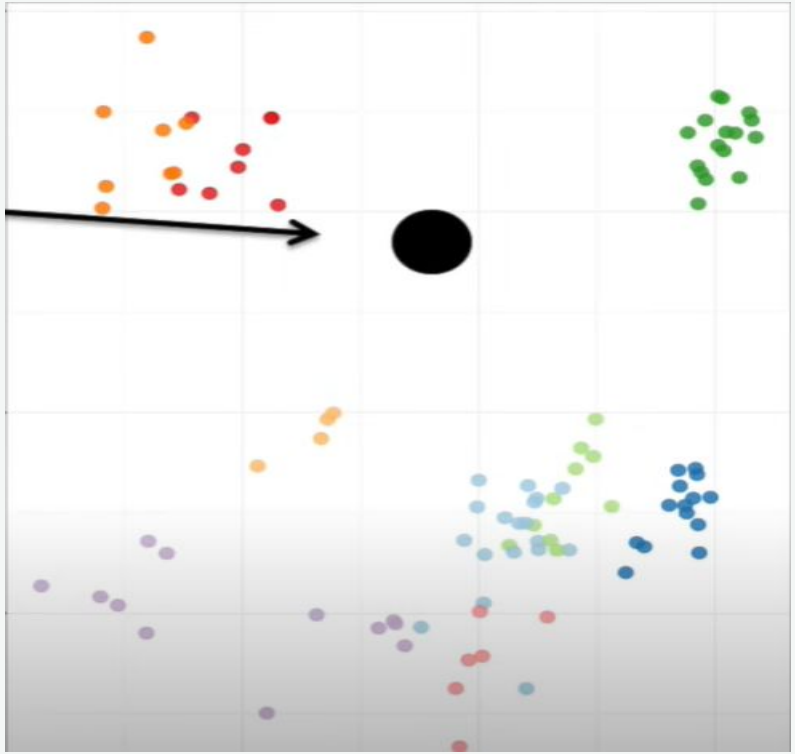


K-Nearest Neighbors (KNN) Example 2



The new point should be ?????

K-Nearest Neighbors (KNN) Example 2



K=11

7 RED

3 Orange

1 Green

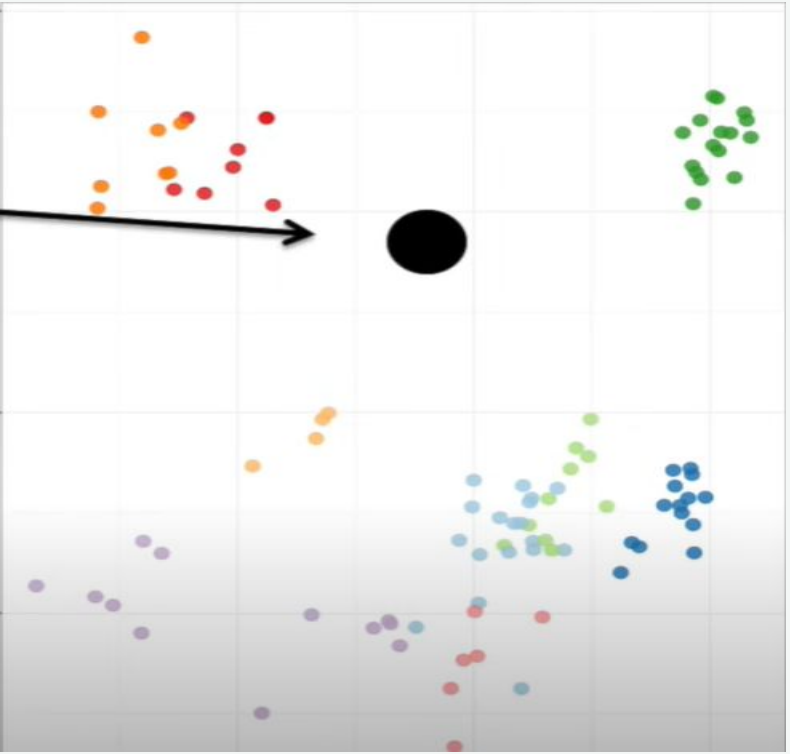
K-Nearest Neighbors (KNN) Example 2

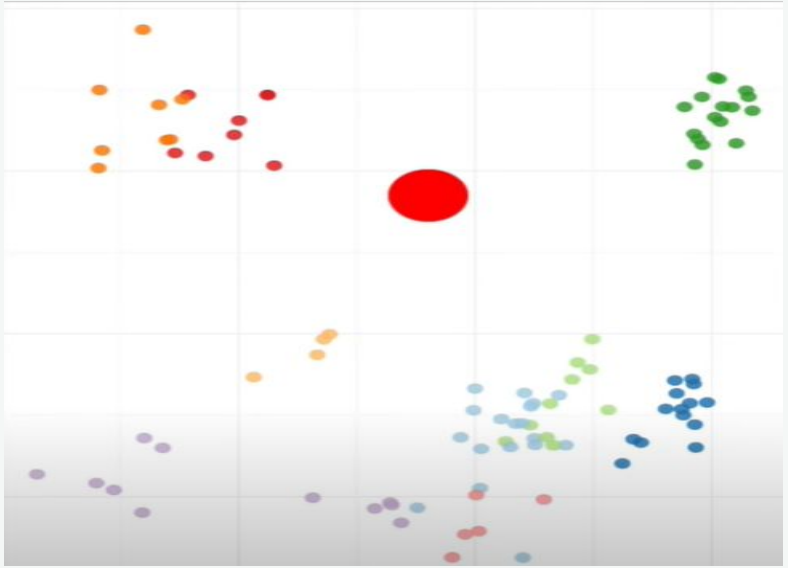
K=11

7 RED

3 Orange

1 Green





KNN Pros and Cons

The algorithm is simple and easy to implement.

- Does not work well with a large dataset.
- Does not work well with high dimensions.
- K's choice.

Evaluation

- The evaluation is performed using a test dataset

- In binary classification, we have to define: “negative” vs “positive” class

Most used metrics

- Confusion Matrix
- Accuracy Score
- Recall
- Precision
- log loss
- ...

Example:
cheap (negative) & expensive(positive)
not cancer (negative) & cancer (positive)
Ham (negative) & spam (positive)

Evaluation

- The evaluation is performed using a test dataset

- In binary classification, we have to define: “negative” vs “positive” class

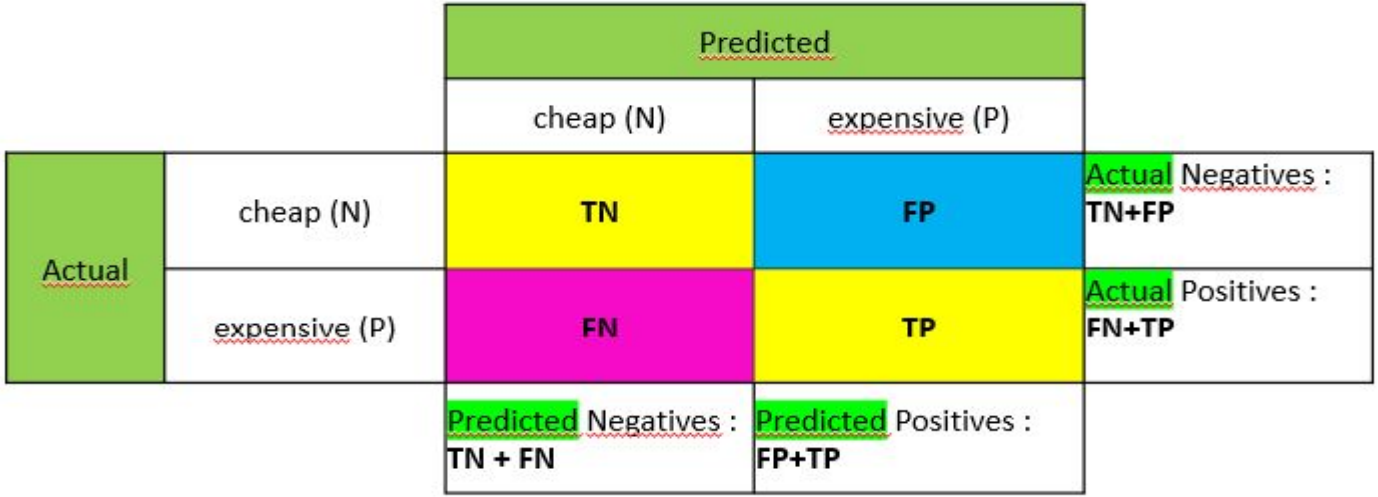
Most used metrics

- Confusion Matrix
- Accuracy Score
- Recall
- Precision
- log loss
- ...

Example:
cheap (negative) & expensive(positive)
not cancer (negative) & cancer (positive)
Ham (negative) & spam (positive)

Evaluation: Confusion Matrix

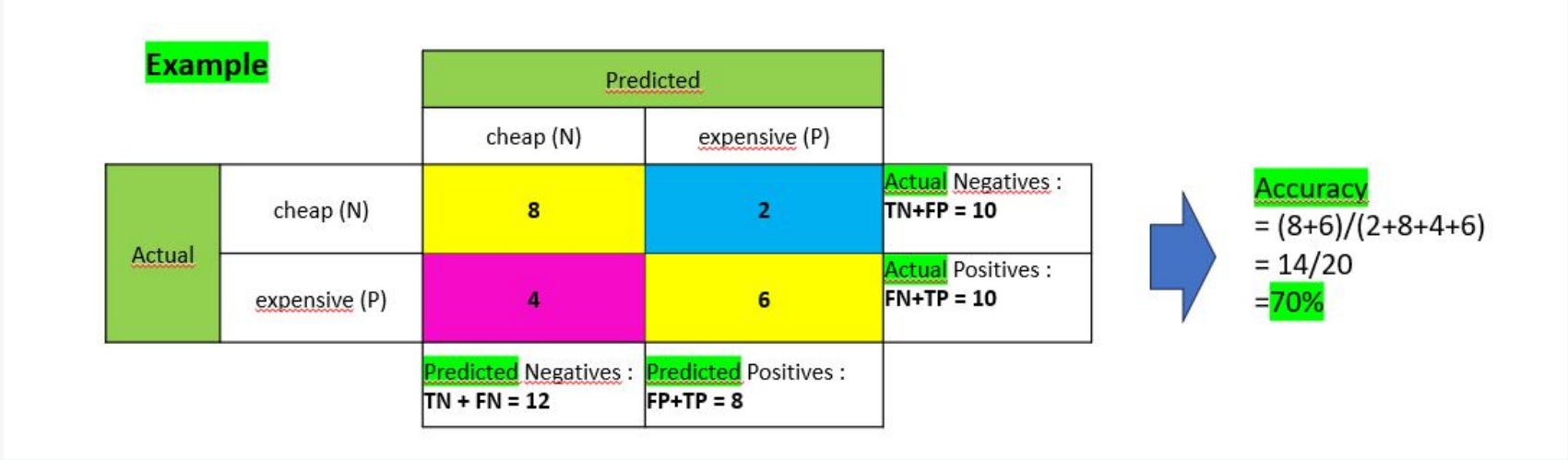
- True Negatives TN : The number of “cheap” houses correctly classified as “cheap”
- False Positives FP: The number of “cheap” houses incorrectly classified as “expensive”
- False Negatives FN: The number of “expensive” houses incorrectly classified as “cheap”
- True Positives TP: The number of “expensive” houses correctly classified as “expensive”



Evaluation: Accuracy Score

It is a measure of how often the classifier correctly predicts both “cheap” and “expensive” houses

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$



Evaluation: Sensitivity vs Specificity

- Sensitivity $= \frac{TP}{TP + FN}$ → True Positive Rate

- Specificity $= \frac{TN}{TN + FP}$ → True Negative Rate

Evaluation: Sensitivity vs Specificity

Cancer example

- Cancer (TP)
- Not caner (TN)
- Cancer but predicted not cancer (FN)
- Not cancer predicted cancer (FP)

which situation is the most serious ???

Spam Example

- Spam (TP)
- Not Spam (TN)
- Spam but predicted not Spam (FN)
- Not Spam predicted Spam (FP)

which situation is the most serious ???

Thank you for your time and attention 🙂