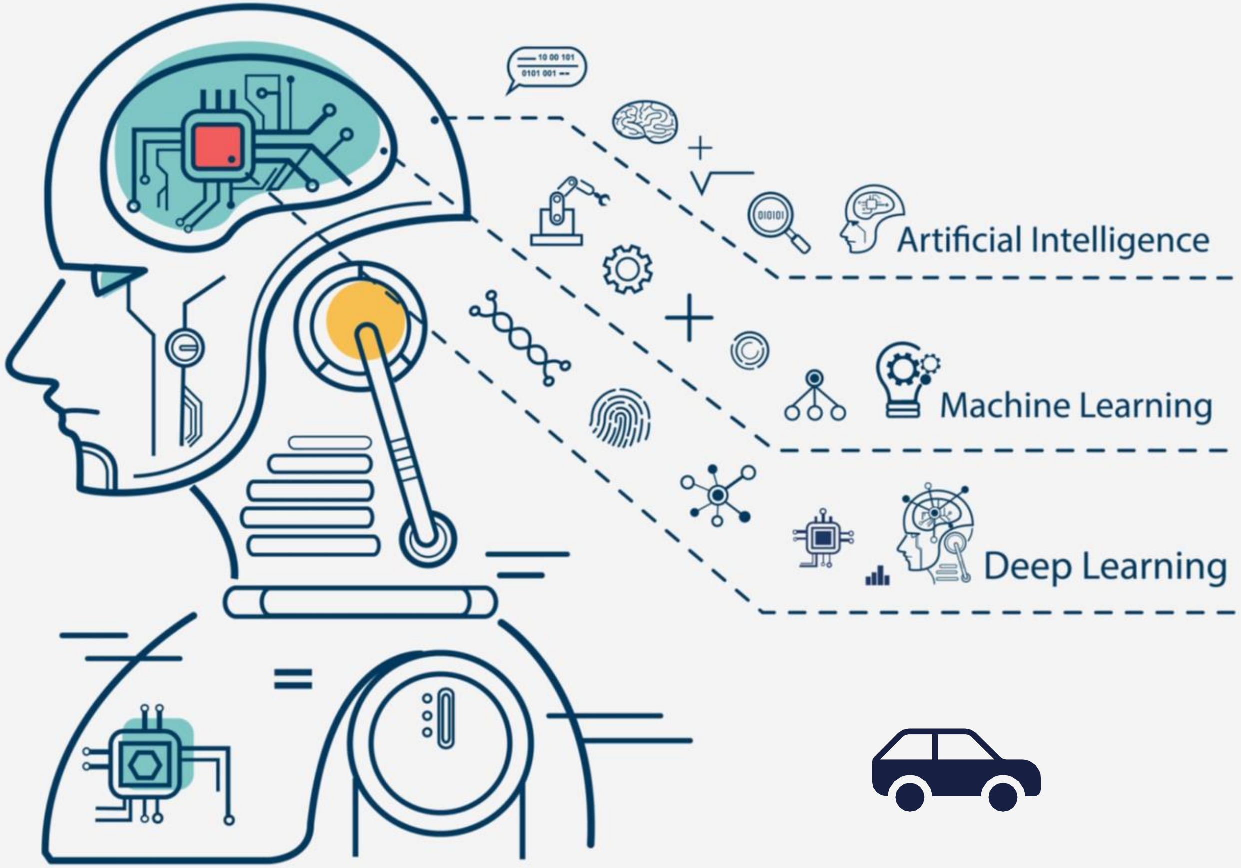
APSSDC

Andhra Pradesh State Skill Development Corporation

Skill AP
APSSDC



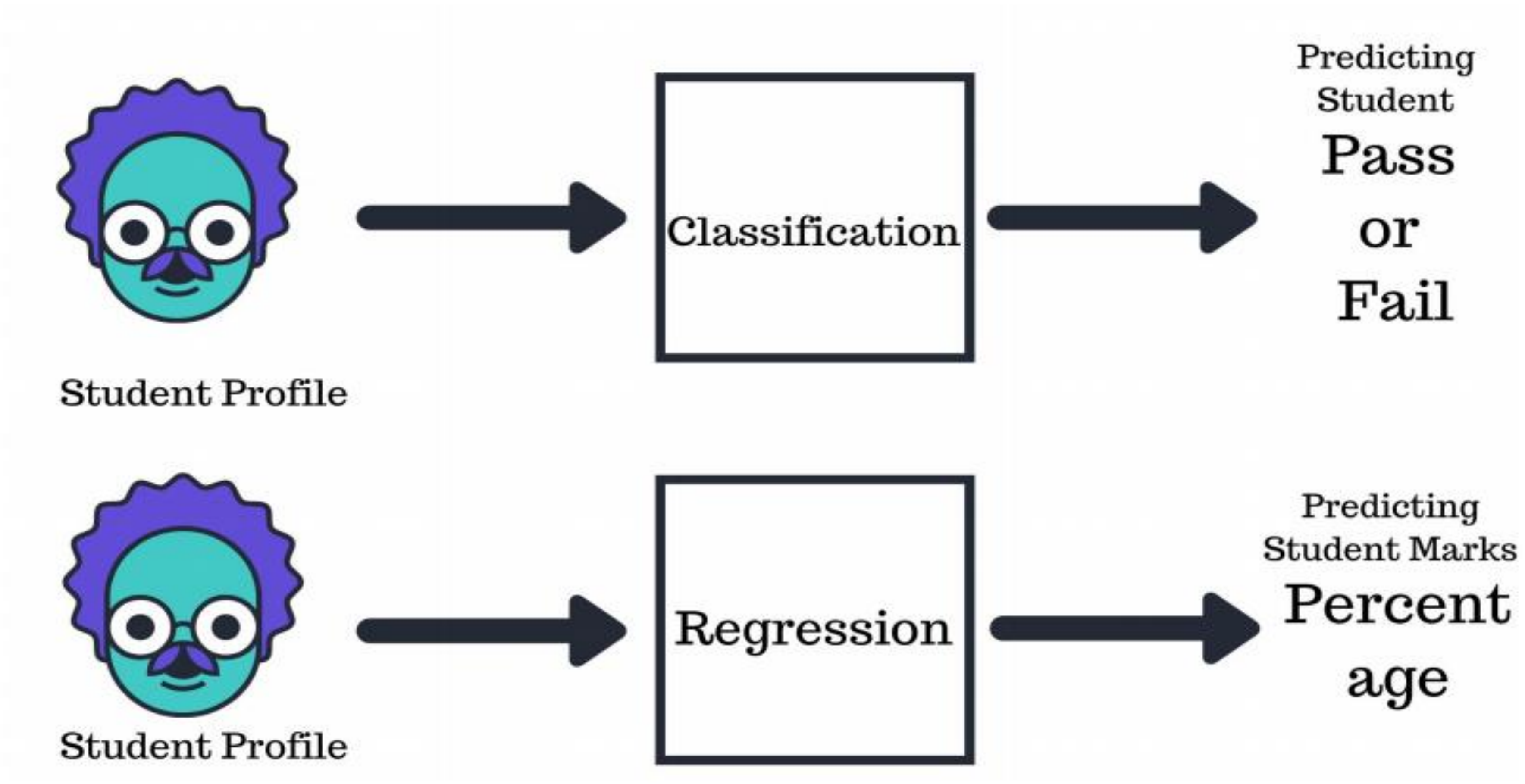
MACHINE LEARNING USING PYTHON

DAY 04 AGENDA

- What is Classification
- Types of classification and its Applications
- K-Nearest Neighbors Classifier
- Evaluation Metrics for classification Models

By Anil Kumar APSSDC

Classification Vs Regression



Student Profile → Classification → Predicting Student Pass or Fail

Student Profile → Regression → Predicting Student Marks Percent age

CLASSIFICATION

Machine Learning Classification is a process of categorizing a given set of data into classes, It can be performed on both structured or unstructured data

It is the process of predicting the class of given data points. Classes are sometimes called as targets/ labels or categories. Classification predictive modeling is the task of approximating a mapping function (f) from **input variables (X) to discrete output variables (y).**

TYPES OF CLASSIFICATION

1. Lazy learners: simply store the training data and wait until a testing data appear. When it does, classification is conducted based on the most related data in the stored training data. Compared to eager learners, lazy learners have less training time but more time in predicting.

- *Ex. **k-nearest neighbor**, Case-based reasoning*

2. Eager learners: construct a classification model based on the given training data before receiving data for classification. It must be able to commit to a single hypothesis that covers the entire instance space. Due to the model construction, eager learners take a long time for train and less time to predict.

- *Ex. Decision Tree, Naive Bayes*

TYPES OF CLASSIFICATION

There are perhaps four main types of classification tasks that you may encounter they are:

- **Binary Classification**
- **Multi-Class Classification**
- Multi-Label Classification
- Imbalanced Classification

BINARY CLASSIFICATION

Binary Classification: Classification task with two possible outcomes.
Eg: 1. Gender classification (Male/Female)
2. Email spam detection (spam or not)

binary classification tasks involve one class that is the normal state and another class that is the abnormal state.

Popular algorithms that can be used for binary classification include:

- **Logistic Regression** (It Supports only for Binary)
- k-Nearest Neighbors
- Decision Trees
- **Support Vector Machine** (It Supports only for Binary)
- Naive Bayes

By Anil Kumar APSSDC

By Anil Kumar APSSDC

MULTI-CLASS CLASSIFICATION

Multi-class classification refers to those classification tasks that have more than two class labels

Eg:1. Face classification.
2. Plant species classification.

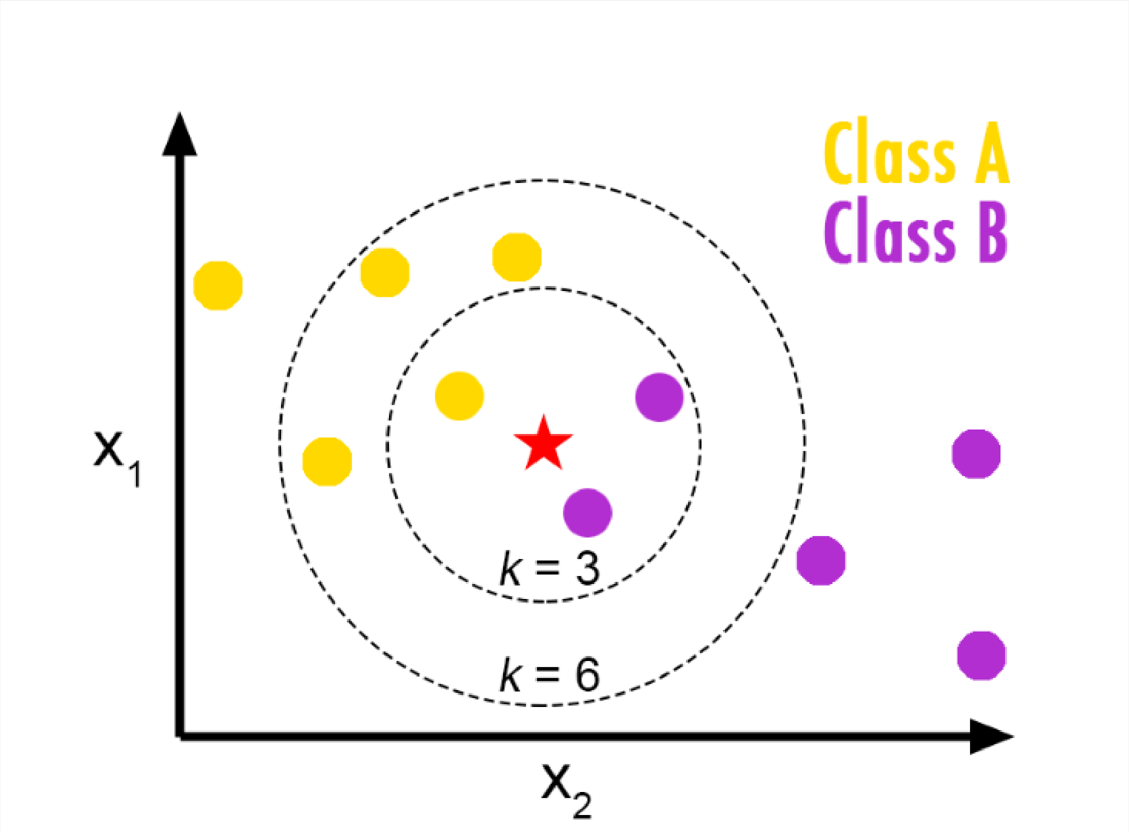
Popular algorithms that can be used for multi-class classification:

- **K-Nearest Neighbors.**
- **Decision Trees.**
- Naive Bayes.
- **Random Forest.**
- Gradient Boosting.

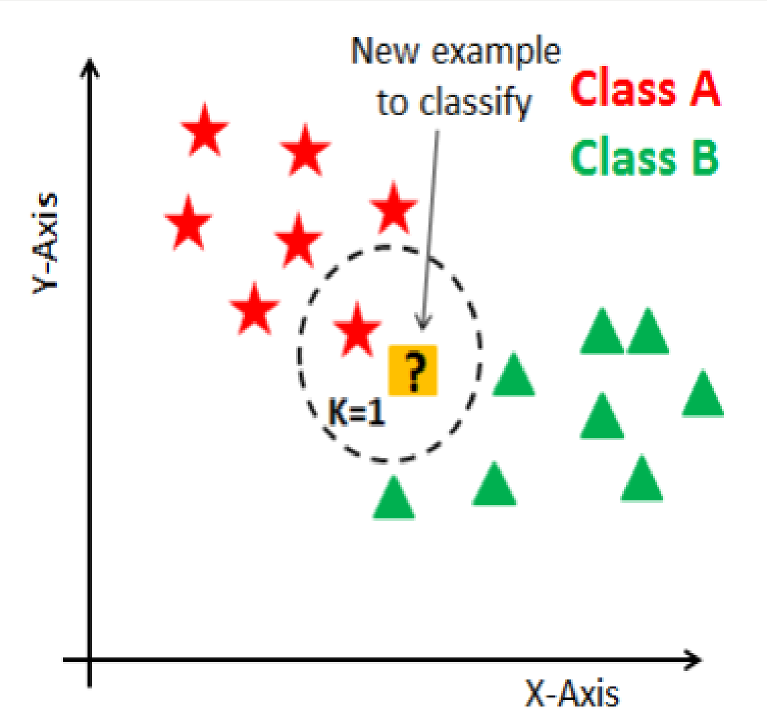
DEALING WITH NON-NUMERIC DATA

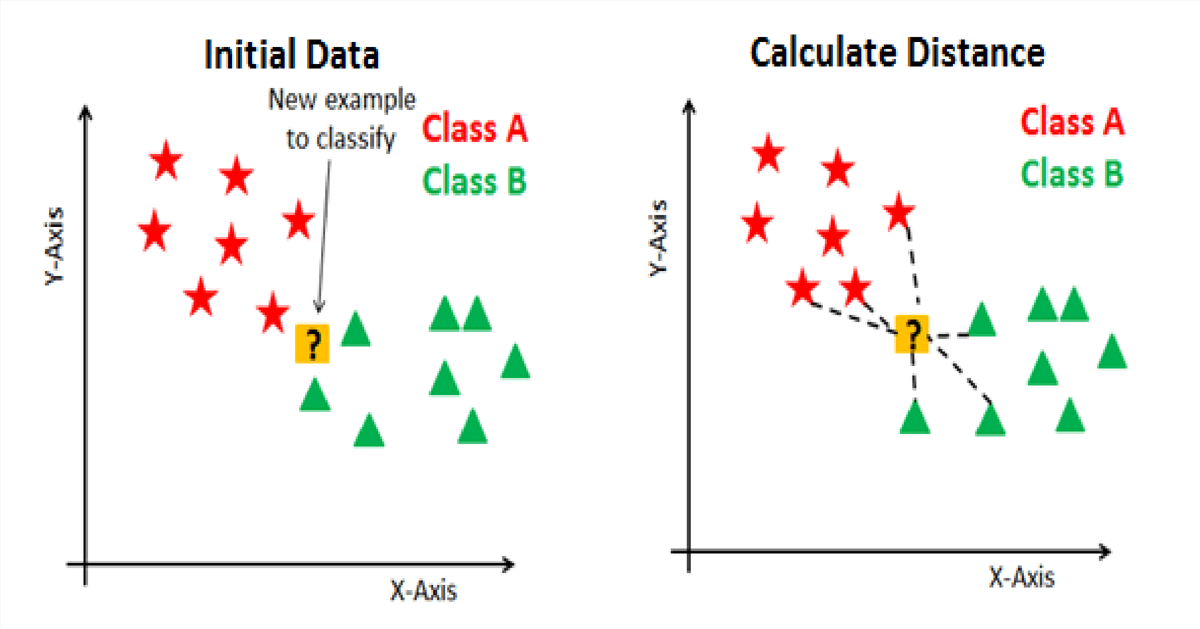
- Boolean values → convert to 0 or 1
  - Applies to yes-no/presence-absence attributes
- Non-binary characterizations
  - Use natural progression when applicable; e.g., educational attainment: HS, Inter/Dip University/Graduate, MS, PHD => 1,2,3,4,5
  - Assign arbitrary numbers but be careful about distances; e.g., color: red, yellow, blue => 1,2,3
- How about unavailable data?
  (0 value not always the answer)

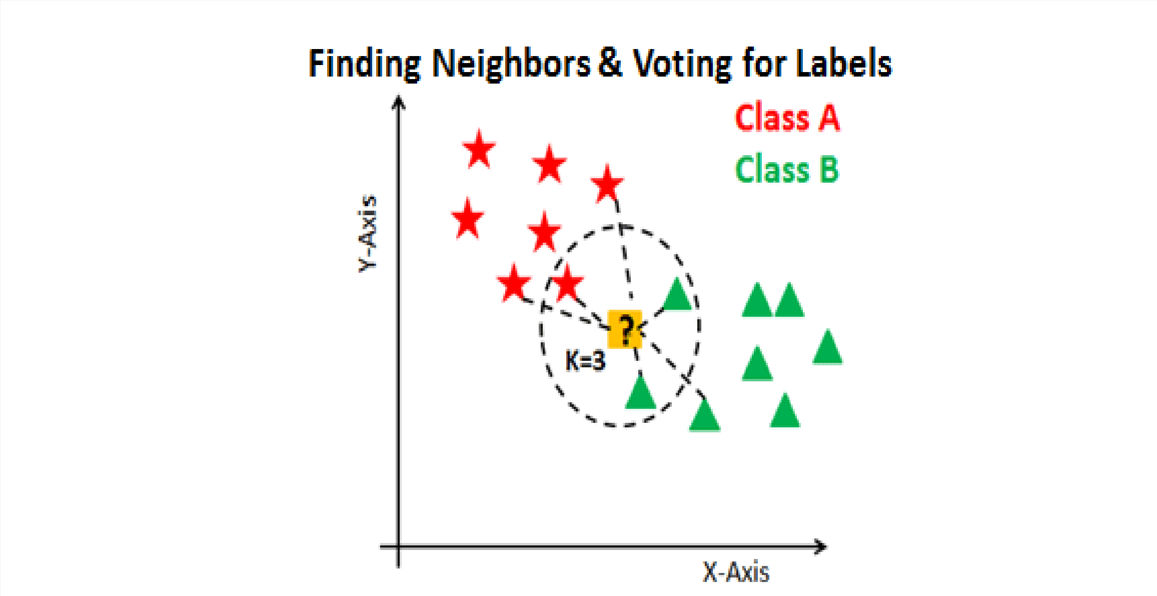
K NEAREST NEIGHBOR



By Anil Kumar APSSDC







By Anil Kumar APSSDC

By Anil Kumar APSSDC

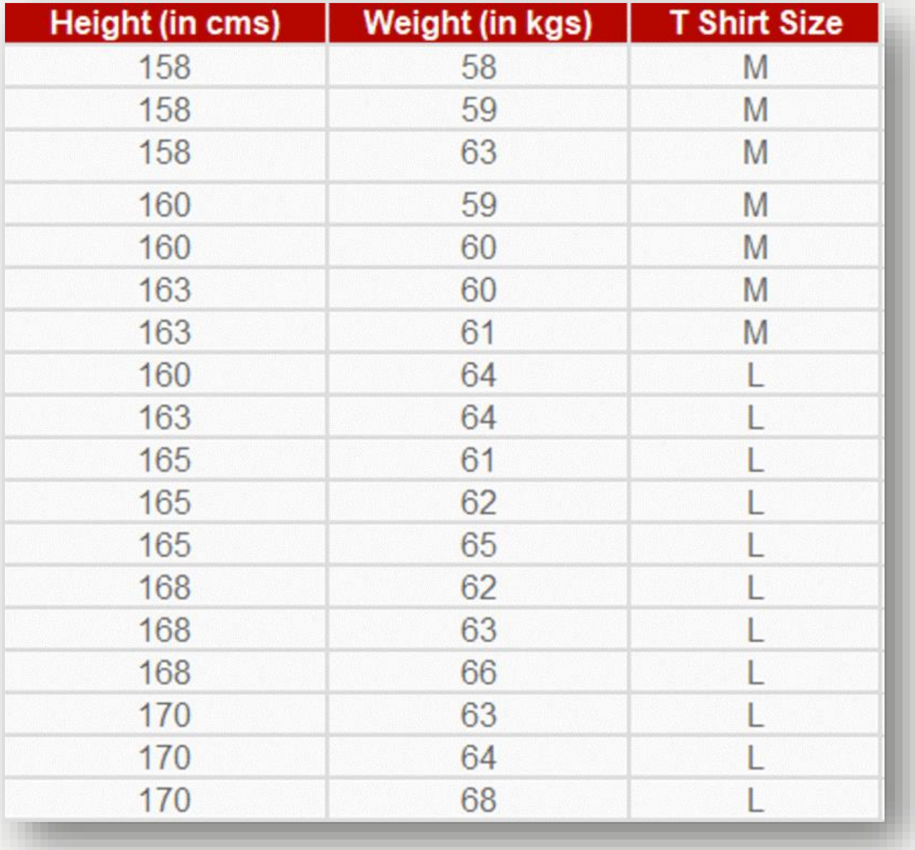
ABOUT

- The name of the algorithm, originates from the philosophy of kNN – i.e., people having similar mindset tend to stay close to each other.
- In the same way, as a part of kNN algorithm, the unknown and unlabelled data which comes for a prediction problem is judged on the basis of training data set of elements which are similar to the unknown elements.
- k in kNN algorithm represents the number of nearest neighbour points which are voting for the new test data's class.

KNN ALGORITHM MANUAL IMPLEMENTATION

- Load the data
- Initialize K to your chosen number of neighbours
- For each example in the data
    - Calculate the distance between the query example and the current example from the data.
    - Add the distance and the index of the example to an ordered collection
- Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances.
- Pick the first K entries from the sorted collection.
- Get the labels of the selected K entries, then return mode of the labels
- setting K-Nearest Neighbor algorithm forms a majority vote between the K most similar instances to a given unseen observation.

Example Dataset Contains The Height And Weight For T-shirt Size

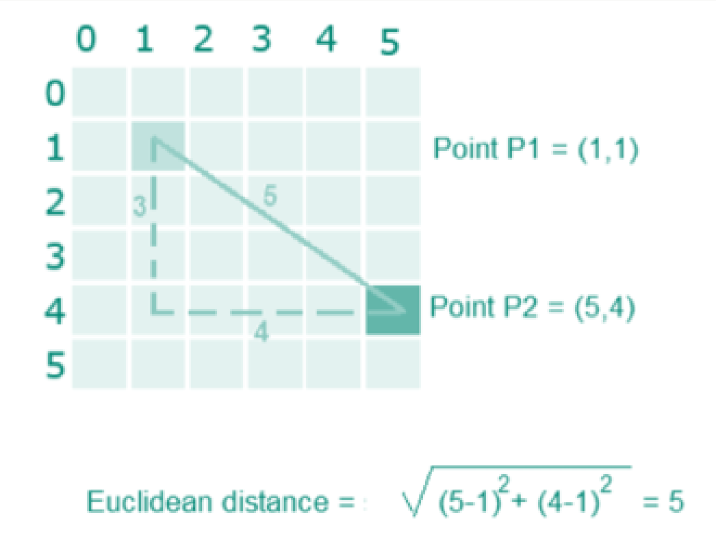


Height (in cms)	Weight (in kgs)	T Shirt Size
158	58	M
158	59	M
158	63	M
160	59	M
160	60	M
163	60	M
163	61	M
160	64	L
163	64	L
165	61	L
165	62	L
165	65	L
168	62	L
168	63	L
168	66	L
170	63	L
170	64	L
170	68	L

predict the T-shirt size of **Anna**, whose height is **161cm** and her weight is **61kg**.

Step1: calculate the Euclidean distance between the new point and the existing points

for example, Euclidean distance between point P1(1,1) and P2(5,4) is



Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2}$ = 5

Step 2: Choose the value of K and select K neighbors closet to the new point.

In this case, select the top 5 parameters having least Euclidean distance

Step 3: Count the votes of all the K neighbors / Predicting Values

By Anil Kumar APSSDC

ADVANTAGES

- ✓ The algorithm is simple and easy to implement.
- ✓ There’s no need to build a model, tune several parameters, or make additional assumptions.
- ✓ The algorithm is versatile. It can be used for classification, regression, and search.

Disadvantages

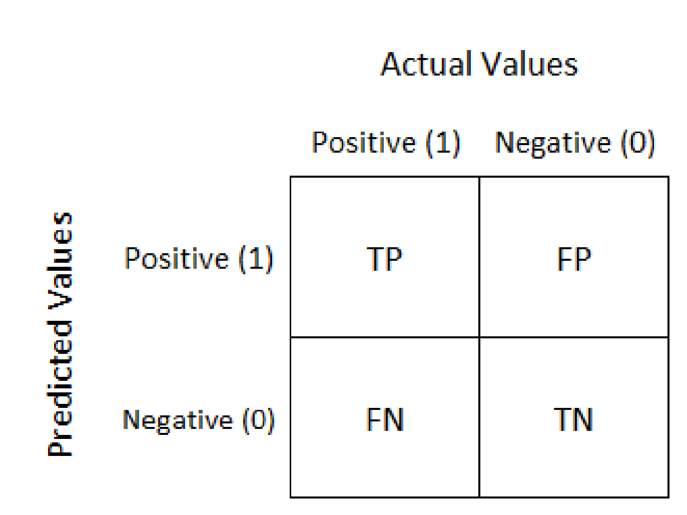
- ✓ The algorithm gets significantly slower as the number of examples and/or predictors/independent variables increase.

FURTHER READING

- This section provides more resources on the topic if you are looking to go deeper.
- Statistical classification, Wikipedia.
- Binary classification, Wikipedia.
- Multiclass classification, Wikipedia.
- Multi-label classification, Wikipedia.
- Multiclass and multilabel algorithms, scikit-learn API.

EVALUATION METRICS

- There will be four possibilities if we consider a match win/loss prediction:
  - Model Predicted win and team win – True Positive (TP)
  - Model Predicted win and team lost – False Positive (FP)
  - Model Predicted loss and team win – False Negative (FN)
  - Model Predicted loss and team loss – True Negative (TN)



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

- Confusion Matrix is the Classification Metric used to describe the performance of a classification model on a set of test data for which the true values are known.

EVALUATION METRICS

- **Model accuracy:** It is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations.

TP+TN / TP+FP+FN+TN

- **Precision:** It is the ratio of correctly predicted positive observations to the total predicted positive observations

TP / TP + FP

- **Recall(Sensitivity):** It is the ratio of correctly predicted positive observations to the all observations in actual class - yes

TP / TP + FN

- **F-Measure:** It is the weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account

2 * Precision * Recall / Precision + Recall

By Anil Kumar APSSDC