Machine Learning

Report File

July 20, 2022

BACHLORS OF TECHNOLOGY

Information Technology

SUBMITTED BY

JASMEEN KAUR GILL

University Roll Number : 2104390

College Roll No: 2121138

GURU NANAK DEV ENGINEERING COLLEGE

LUDHIANA-141006, INDIA

Contents

1 Aim : Introduction to Python Libraries 1

1.1 Aim : Use of NumPy Library. 1

1.2 Aim: Use of pandas Library. 1

1.3 Aim: Use of Matplotlib Library. 2

1.4 Aim: Use of Sciket - Learn library 2

2 Aim: Introduction to concept of Data Science 3

3 Aim: Introduction to concept of Artificial Intelligence 4

4 Aim: Introduction to concept of Machine Learning 5

5 Aim: Regression Techniques used in Machine Learning 6

6 Aim: Classification Technique used in Machine Learning 9

7 Aim: Clustering techniques used in Machine Learning 12

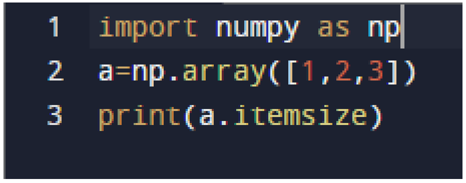
1 Aim : Introduction to Python Libraries

A Python library is a collection of related modules. It contains bundles of code that can be used repeatedly in different programs.

1.1 Aim : Use of NumPy Library.

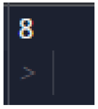
NumPy: NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

Implementation:



```
import numpy as np
a=np.array([1,2,3])
print(a.itemsize)
```

Output:

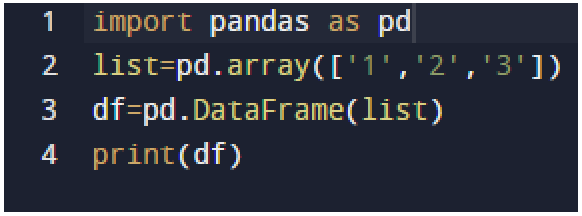


```
8
>
```

1.2 Aim: Use of pandas Library.

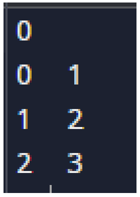
Pandas: Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

Implementation:



```
import pandas as pd
list=pd.array(['1','2','3'])
df=pd.DataFrame(list)
print(df)
```

Output:



```
0
0  1
1  2
2  3
```

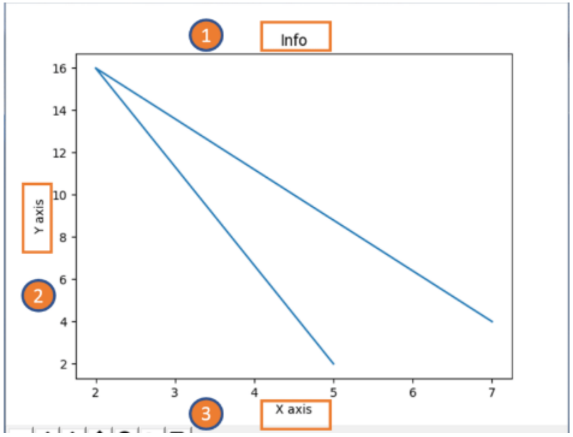
1.3 Aim: Use of Matplotlib Library.

Matplotlib: Matplotlib is a plotting library for the Python programming language. It is used to plot 2D graphics of datasets.

Implementation:

```
from matplotlib import pyplot as plt

 #Plotting to our canvas

 plt.plot([1,2,3],[4,5,1])

 #Showing what we plotted

 plt.show()
```

Output:



1.4 Aim: Use of Sciket - Learn library

Sciket - Learn: Scikit-learn is a free machine learning library for Python. It features various algorithms like support vector machine, random forests, and k-neighbours, and it also supports Python numerical and scientific libraries like NumPy and SciPy. It is a key library for the Python programming language that is typically used in machine learning projects. It is used for clean, uniform and streamlined data.

2 Aim: Introduction to concept of Data Science

Introduction to Data Science

Data Science is a combination of multiple disciplines that uses statistics, data analysis, and machine learning to analyze data and to extract knowledge and insights from it. Firstly we gather Raw Data then after organising it becomes information. Data is of two types structured and unstructured. Example of a data could be information collected for researsh paper.

Dataset: Dataset is a collection of data.

Database: It is an organised collection of data stored and accessed electronicaly

Database Table: In case of tabular data, dataset corresponds to one or more database tables, where columns and rows represents record of dataset.

Components of Data Science: Two main components of data science are visualisation and statistics.

Applications of Data Science

It is used in banking and electricity department (for keeping record).

It is used in route planning and navigation.

Also used to for see delays of flight/ship.

It is used to forecast new year revenue.

Used in stock market, e-commerce.

3 Aim: Introduction to concept of Artificial Intelligence

Introduction to Artificial Intelligence

Artificial intelligence (AI) broadly refers to any human-like behavior displayed by a machine or system. In AI's most basic form, computers are programmed to "mimic" human behavior using extensive data from past examples of similar behavior.

Strong AI: It can perform variety of functions eventually teaching itself to solve for new problems.

Weak AI: It can perform specific tasks like answering questions based on input and output.

Applications of AI

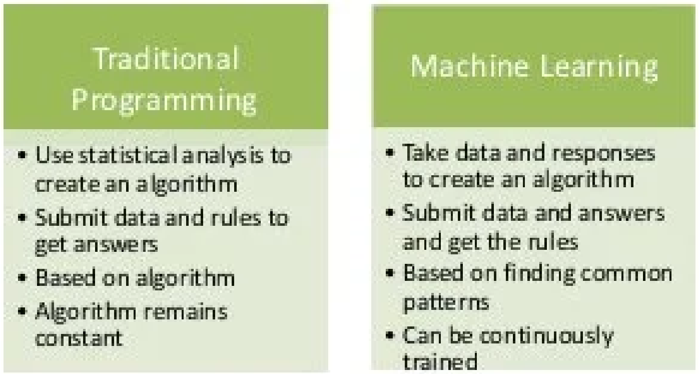
Personalized Shopping (preferences , interest)
AI-powered Assistants (natural language procesing)
Fraud Prevention and Creating Smart Content
Voice Assistants (moniter student's data thoroughly)
Personalized Learning
Autonomous Vehicles (electric cars)

4 Aim: Introduction to concept of Machine Learning

Introduction to Machine Learning

Machine learning is programming computers to optimize a performance criterion using example data or past experience . We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience. The model may be predictive to make predictions in the future. It is making machine learn things.

Machine Learning vs Traditional Learning



Traditional Programming	Machine Learning
• Use statistical analysis is to create an algorithm	• Take data and responses to create an algorithm
• Submit data and rules to get answers	• Submit data and answers and get the rules
• Based on algorithm	• Based on finding common patterns
• Algorithm remains constant	• Can be continuously trained

Types of Machine Learning

Supervised Learning: Supervised machine learning requires labelled input and output data during the training phase of the machine learning lifecycle. This training data is often labelled. Machine learn things from training data and then apply knowledge to test data.

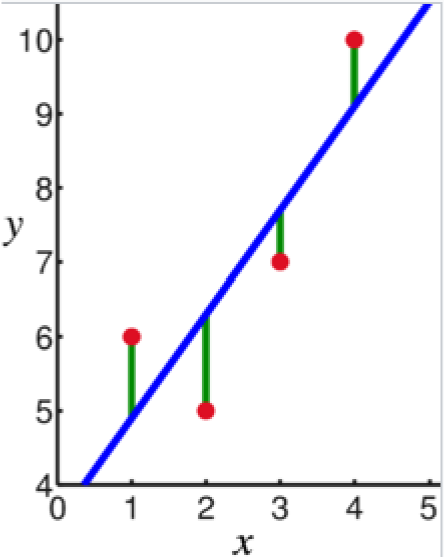
Unsupervised Learning: Unsupervised machine learning is the training of models on raw and unlabelled training data. It is often used to identify patterns and trends in raw datasets, or to cluster similar data into a specific number of groups.

Semisupervised Learning: Semi-supervised machine learning is a combination of supervised and unsupervised learning. It uses a small amount of labeled data and a large amount of unlabeled data, which provides the benefits of both unsupervised and supervised learning

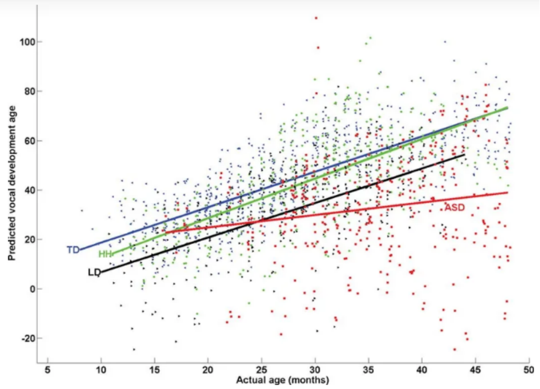
5 Aim: Regression Techniques used in Machine Learning

Simple Linear Regression: Simple linear regression is a regression model that estimates the relationship between one independent variable and one dependent variable using a straight line.

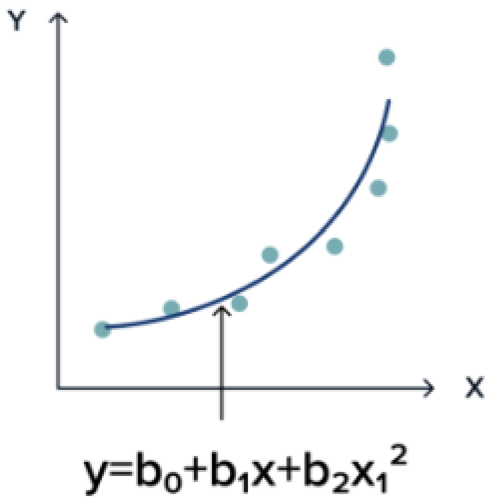
It's equation is y=ax+b, where x is input and y is output.



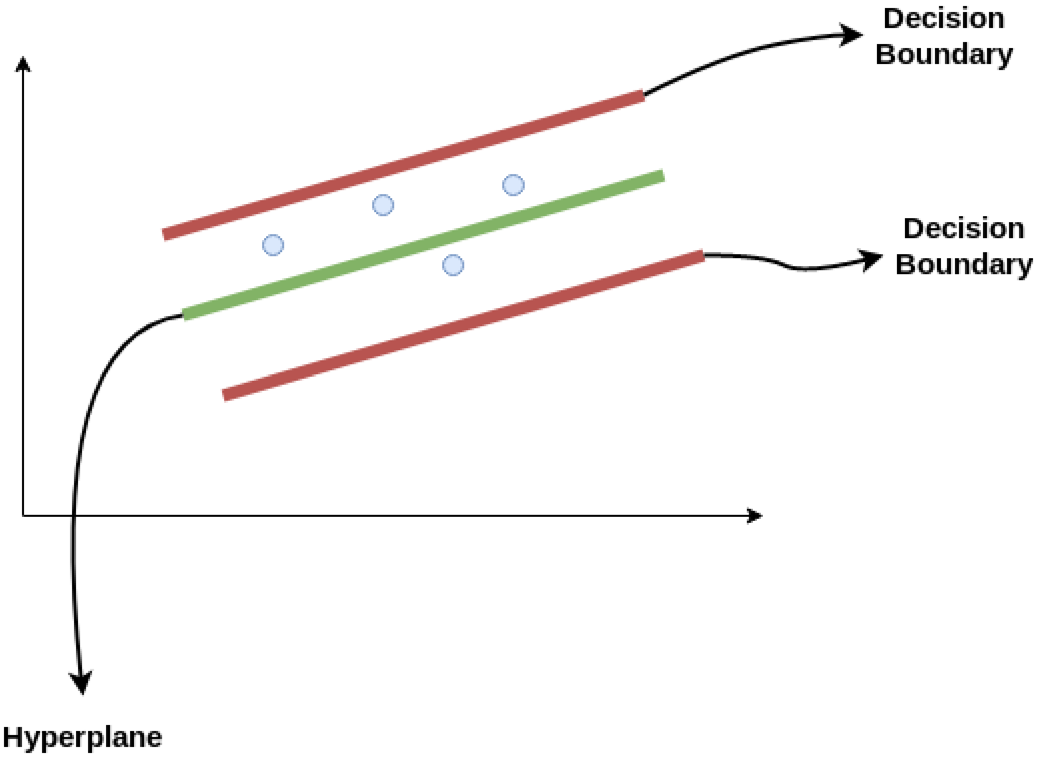
Multiple Linera Regression: Multiple linear regression refers to a statistical technique that is used to predict the outcome of a variable based on the value of two or more variables.



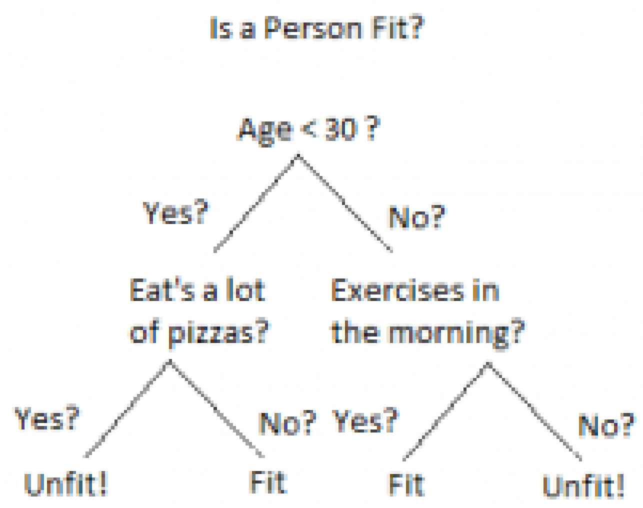
Polynomial Regression: It is a form of regression analysis in which the relationship between the independent variable x and the dependent variable y is modelled as an nth degree polynomial in x.



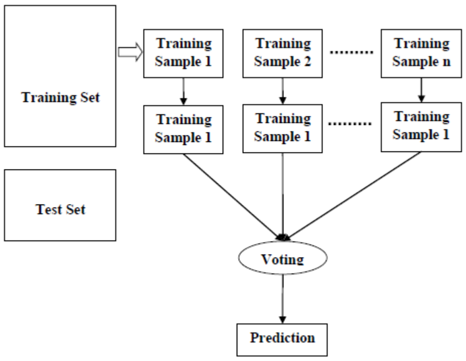
Support Vector Regression: It is an regression algorithm. Our objective, when we are moving on with SVR, is to basically consider the points that are within the decision boundary line. Our best fit line is the hyperplane that has a maximum number of points.



Decision Tree: It is a graphical representaion for getting all possible solutions to a problem decision based on given conditions. It is used to memic the human thinking ability. It is mostely used for classification problems due to accuracy problem. It has decision node which signifies decision eithe yes or no, leaf nodes which signifies output of decision and branches showing decision rules.

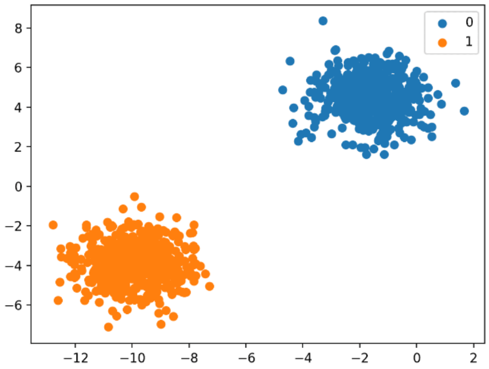


Random Forest: It is a supervised machine learning algorithm made up of decision trees. Random Forest is used for both classification and regression. It is better than decision tree as it predicts output with higher accuracy and takes less training time.



6 Aim: Classification Technique used in Machine Learning

Classification: It is a supervised learning concept which basically categorizes a set of data into classes. The most common classification problems are – speech recognition, face detection, handwriting recognition, document classification, etc. It gives output in the form of yes or no.



K-nearest neighbors (kNN): It is supervised machine learning algorithm used to solve both classification and regression tasks. It gives output as 0 (representing not) or 1 (representing yes).

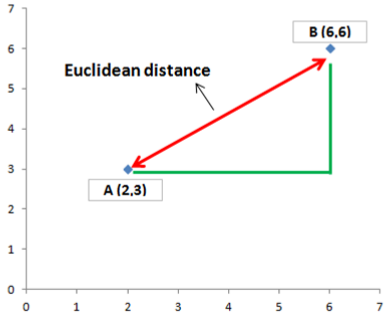
Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is between -3 to +3. It can be reversed after task is being done called reverse scaling.

Step 1: Select the value of K neighbors(say k=5)

Step 2: Find the K (5) nearest data point for our new data point based on euclidean distance(which we discuss later)

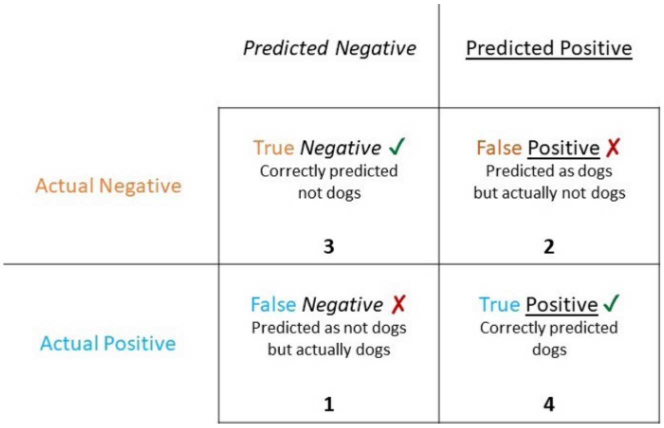
Step 3: Among these K data points count the data points in each category

Step 4: Assign the new data point to the category that has the most neighbors of the new datapoint



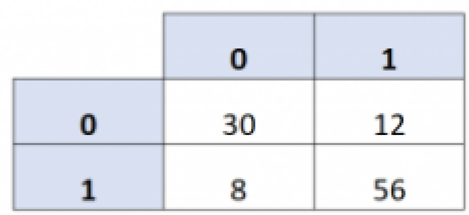
$$\boldsymbol{Euclidean\ distance}\ (\boldsymbol{a},\boldsymbol{b}) = \sqrt{(\boldsymbol{a_1}-\boldsymbol{b_1})^2 + (\boldsymbol{a_2}-\boldsymbol{b_2})^2}$$

Confusion Matrix: It is a useful machine learning method which allows you to measure precision and accuracy. It is used for classification tasks. False negative and false positive is caused due to the fault of algorithm.



	Predicted Negative	Predicted Positive
Actual Negative	True *Negative* ✓ Correctly predicted not dogs **3**	False Positive ✗ Predicted as dogs but actually not dogs **2**
Actual Positive	False *Negative* ✗ Predicted as not dogs but actually dogs **1**	True Positive ✓ Correctly predicted dogs **4**

Example:

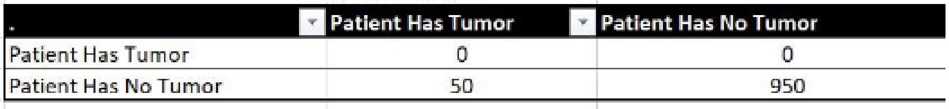


	0	**1**
0	30	12
1	8	56

Total = 106
TN = 30, TP = 56
FN = 8, FP = 56
Accuracy = (TP+TN) / (TP+TN+FP+FN)
Accuracy = 86 / 106
Accuracy (in %) =81%

Accuracy Paradox: The accuracy paradox is the paradoxical finding that accuracy is not a good metric for predictive models when classifying in predictive analytics.

Example:



.	Patient Has Tumor	Patient Has No Tumor
Patient Has Tumor	0	0
Patient Has No Tumor	50	950

Total = 1000

Accuracy = (TP + TN) / (TP + FP + FN + TN)

Accuracy = (950 + 0) / (950 + 0 + 50 + 0)

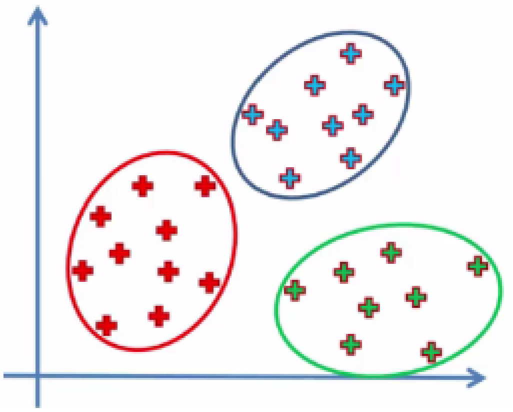
Accuracy = 950 / 1000

Accuracy (in %) = 95%

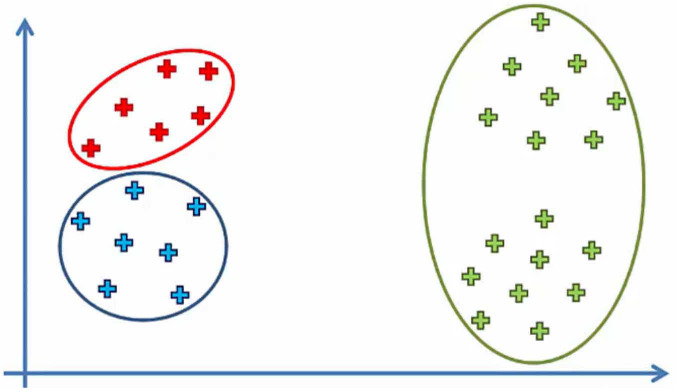
Model is clearly biased as in any case it is saying that patient has no tumor. Hence, accuracy is not a reliable metric to determine a model performance.

7 Aim: Clustering techniques used in Machine Learning

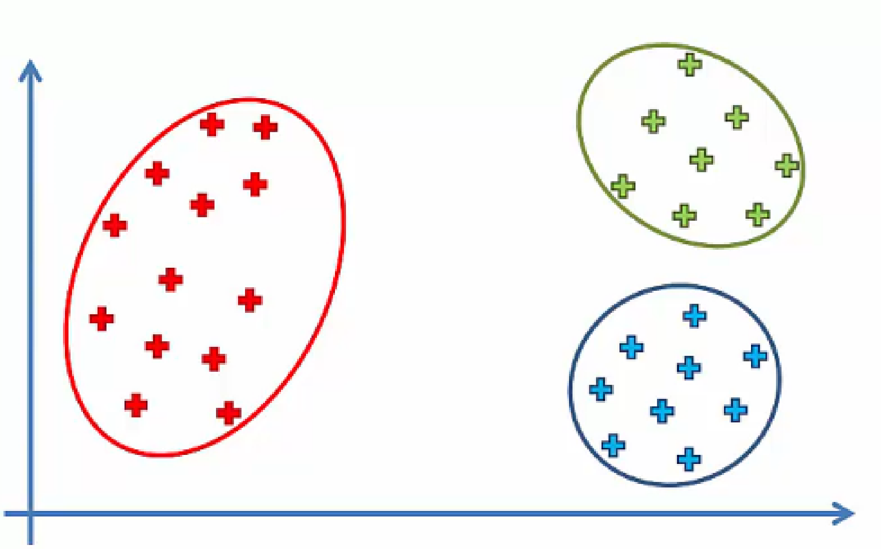
Clustering: It is type of unsupervised learning. It makes group out of available set of data. Clustering is the task of dividing the data points into a number of groups such that data points in the same groups are more similar to other data poimts.



Random Initialization Trap: It is a problem that occurs in K-means algorithm. In it when the centroids of the clusters to be generated are explicitly defined by the user then inconsistency may be created and this may sometimes lead to generating wrong clusters in the dataset as shown in the given figure. In order to remove this apply k++ algorithm.



The correct random initialisation would lead to formation of given three clusters as shown in figure below.

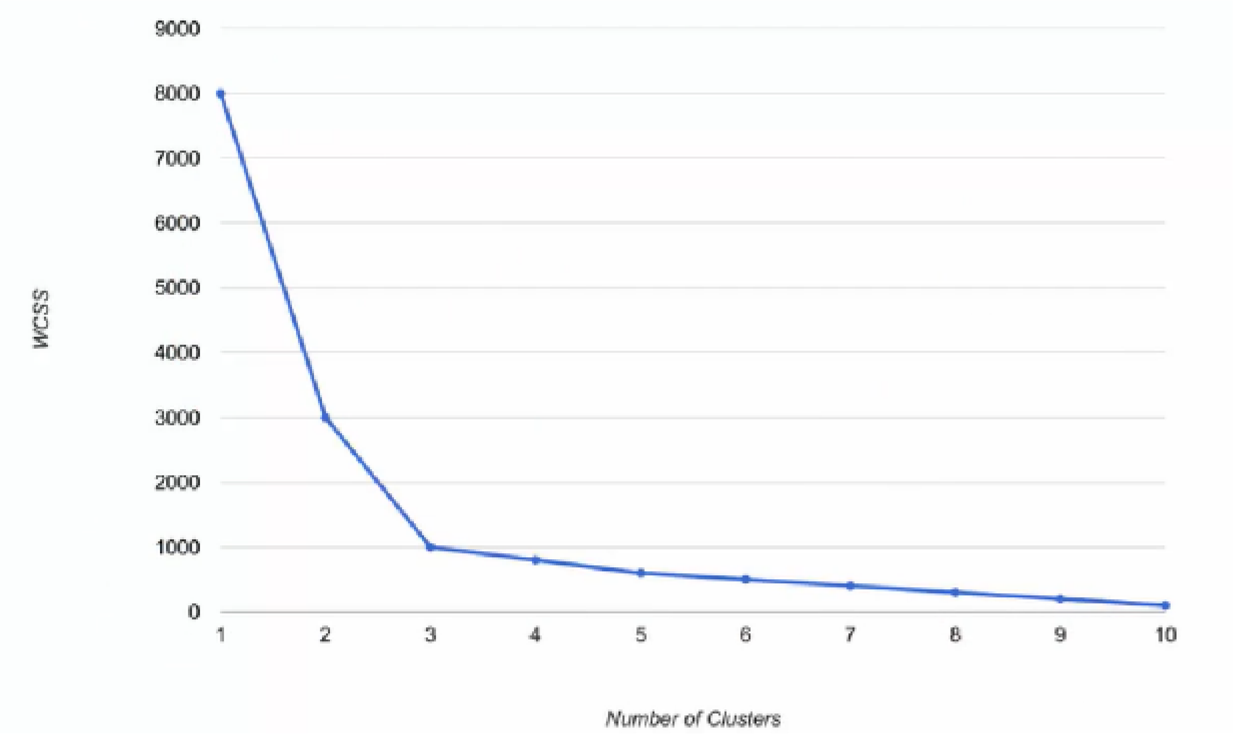


How to choose right number of clusters?

In order to choose right number of clusters we will find WCSS (within cluster sum of squares). It's maximum is one when it has only one cluster and minimum value is zero if every point is taken as cluster.

$$\mathrm{WCSS} = \sum_{\mathrm{P_i\ in\ Cluster\ 1}} \mathrm{distance}(\mathrm{P_i}, \mathrm{C_1})^2 + \sum_{\mathrm{P_i\ in\ Cluster\ 2}} \mathrm{distance}(\mathrm{P_i}, \mathrm{C_2})^2 + \sum_{\mathrm{P_i\ in\ Cluster\ 3}} \mathrm{distance}(\mathrm{P_i}, \mathrm{C_3})^2$$

Elbow Method: This method is used to know optimal number of clusters.


WCSS
Number of Clusters