Machine Learning

Practical File

Bachelor of Technology
Information Technology

SUBMITTED BY

Manpreet Kaur

URN- 2104528

CRN-2121070

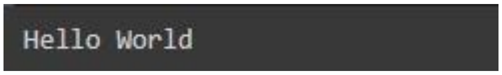
GURU NANAK DEV ENGINEERING
COLLEGE LUDHIANA-141006,INDIA

1 BASIC PYTHON PROGRAM

1.1 Print Hello World

```
Code: print( Hello World )
```

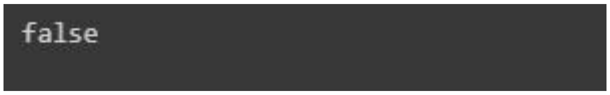
Output:



```
Hello World
```

1.2 Else If Condition

```
Code:   x=23   if   x==24:
     print("true")      else:
     print("false") Output:
```

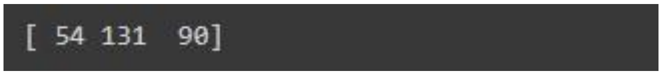


```
false
```

2 NUMPY LIBRARY

2.1 Numpy library to add two numbers

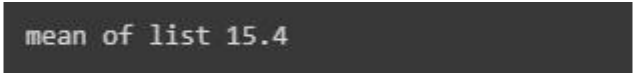
```
Code: import numpy as np a=np.array([20,97,56])
b=np.array([34]) c=np.array(a+b)
print (c) Output:
```



```
[ 54 131  90]
```

2.2 Numpy library to find mean of an array

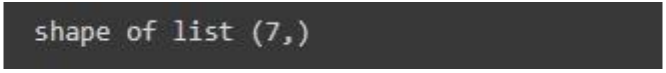
```
Code:
      import numpy as np
     marks = [16,20,18,11,12] marks_arr
= np.array(marks)
m=np.mean(marks_arr) print("mean of
list",m) Output:
```



```
mean of list 15.4
```

2.3 Numpy library to find shape of an array

```
Code: import numpy as np  s =
     [16,20,18,11,67,45,12]
arr = np.array(s) shp=arr.shape
print("shape of list",shp) Output:
```



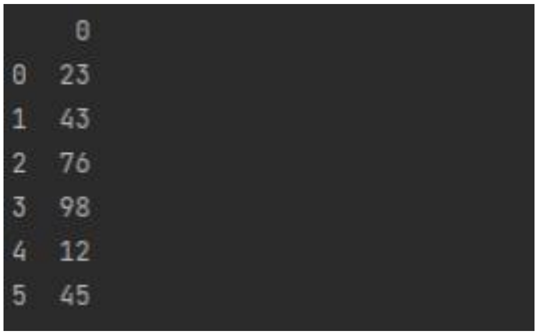
```
shape of list (7,)
```

3 PANDA LIBRARY

3.1 Using Panda for dataframing

CODE :

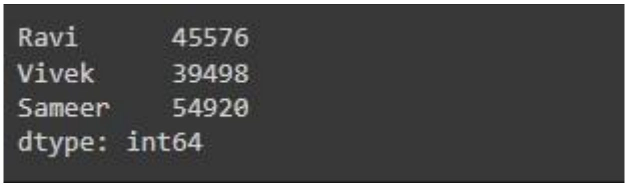
```
import    pandas    as    pd         s=pd.array([23,43,76,98,12,45],[45,76,43,98,7,1])
print(pd,DataFrame(s)) OUTPUT:
```



```
    0
0  23
1  43
2  76
3  98
4  12
5  45
```

3.2 Series in Panda Library

```
CODE    :          import   pandas   as        pd
      phonebook={
     ’Ravi’:45576,
      ’Vivek’:39498, ’Sameer’:54920 }
print(pd.Series(phonebook))
OUTPUT:
```

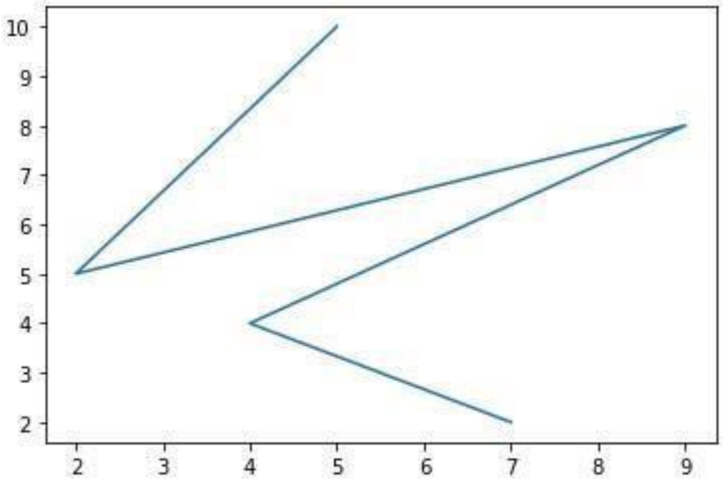


```
Ravi      45576
Vivek     39498
Sameer    54920
dtype: int64
```

4 MATPLOTLIB

CODE :

```
      from matplotlib import pyplot as plt  x=[5,2,9,4,7]
y=[10,5,8,4,2] plt.plot(x,y) plt.show() OUTPUT:
```



5 AI TOOLS

5.1 Scikit-learn

5.1.1 What is Scikit-Learn (Sklearn)?

Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of ecient tools for machine learning and statistical modeling including classication, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon NumPy, SciPy and Matplotlib.

5.1.2 Features

1. Supervised Learning algorithms = Almost all the popular supervised learning algorithms, like Linear Regression, Support Vector Machine (SVM), Decision Tree etc., are the part of scikit-learn.
2. Unsupervised Learning algorithms = On the other hand, it also has all the popular unsupervised learning algorithms from clustering, factor analysis, PCA (Principal Component Analysis) to unsupervised neural networks.
3. Clustering = This model is used for grouping unlabeled data.
4. Feature selection = It is used to identify useful attributes to create supervised models.
5. Open Source = It is open source library and also commercially usable under BSD license.

Advantages:

1. The library is distributed under the BSD license, making it free with minimum legal and licensingrestrictions.
2. It is easy to use.
3. The scikit-learn library is very versatile and handy and serves real-world purposes like the prediction ofconsumer behavior, the creation of neuroimages, etc.
4. Scikit-learn is backed and updated by numerous authors, contributors, and a vast international onlinecommunity.
5. The scikit-learn website provides elaborate API documentation for users who want to integrate thealgorithms with their platforms.

Disadvantages:

1. It is not the best choice for in-depth learning.
2. It is not optimized for graph algorithms, and it is not very good at string processing. For example,scikitlearn does not provide a built-in way to produce a simple word cloud. Scikit-learn doesn’t have a strong linear algebra library, hence scipy and numpy are used.

5.2 Pytorch

5.2.1 What is pytorch?

PyTorch is an open source machine learning library used for developing and training neural network based deep learning models. It is primarily developed by Facebook’s AI research group. PyTorch can be used with Python as well as a C++. Naturally, the Python interface is more polished. Pytorch (backed by biggies like Facebook, Microsoft, SalesForce, Uber) is immensely popular in research labs. Not yet on many production servers that are ruled by fromeworks like TensorFlow (Backed by Google) Pytorch is picking up fast. Unlike most other popular deep learning frameworks like TensorFlow, which use static computation graphs, PyTorch uses dynamic computation, which allows greater exibility in building complex architectures. Pytorch uses core Python concepts like classes, structures and conditional loops that are a lot familiar to our eyes, hence a lot more intuitive to understand.

5.2.2 Features

1. Easy Interface = PyTorch offers easy to use API; hence it is considered to be very simple to operate andruns on Python. The code execution in this framework is quite easy.

2. Python usage = This library is considered to be Pythonic which smoothly integrates with the Pythondata science stack. Thus, it can leverage all the services and functionalities offered by the Python environment.

3. Computational graphs - PyTorch provides an excellent platform which offers dynamic computationalgraphs. Thus a user can change them during runtime. This is highly useful when a developer has no idea of how much memory is required for creating a neural network model.

PyTorch is known for having three levels of abstraction as given below Tensor -

Imperative n-dimensional array which runs on GPU.

Variable - Node in computational graph. This stores data and gradient.

Module - Neural network layer which will store state or learnable weights.

Advantages

1. It is easy to debug and understand the code.
2. It includes many layers as Torch.
3. It includes lot of loss functions.
4. It can be considered as NumPy extension to GPUs.
5. It allows building networks whose structure is dependent on computation itself.

Disadvantages

1. .It has been released in 2016, so it’s new compared to others and has fewer users, and is not widelyknown.
2. Absence of monitoring and visualization tools like a tensor board.
3. The developer community is small compared to other frameworks.

5.3 TensorFlow

5.3.1 What is TensorFlow?

TensorFlow is an open-source end-to-end platform for creating Machine Learning applications. It is a symbolic math library that uses dataflow and differentiable programming to perform various tasks focused on training and inference of deep neural networks. It allows developers to create machine learning applications using various tools, libraries, and community resources. Currently, the most famous deep learning library in the world is Google’s TensorFlow. Google product uses machine learning in all of its products to improve the search engine, translation, image captioning or recommendations.

5.3.2 Features:

1. Open-source Library It is an open-source library that allows rapid and easier calculations in machine learning. It eases the switching of algorithms from one tool to another TensorFlow tool.

2. Easy to run We can execute TensorFlow applications on various platforms such as Android, Cloud,IOS and various architectures such as CPUs and GPUs. This allows it to be executed on various embedded platforms.

3. Fast Debugging: It allows you to reflect each node, i.e., operation individually concerning its evaluation.It provides computational graphing methods that support an easy to execute paradigm.

4. Effective It works with multi-dimensional arrays with the help of data structure tensor which representsthe edges in the flow graph. Tensor identifies each structure using three criteria: rank, type, shape.

Advantages

1. Open-source platform: It is an open-source platform that makes it available to all the users around andready for the development of any system on it.

2. Data visualization: TensorFlow provides a better way of visualizing data with its graphical approach. Italso allows easy debugging of nodes with the help of TensorBoard

3. Parallelism: TensorFlow finds its use as a hardware acceleration library due to the parallelism of workmodels. It uses different distribution strategies in GPU and CPU systems.

4. A user can choose to run its code on either of the architecture based on the modeling rule. A systemchooses a GPU if not specified. This process reduces the memory allocation to an extent.

Disadvantages

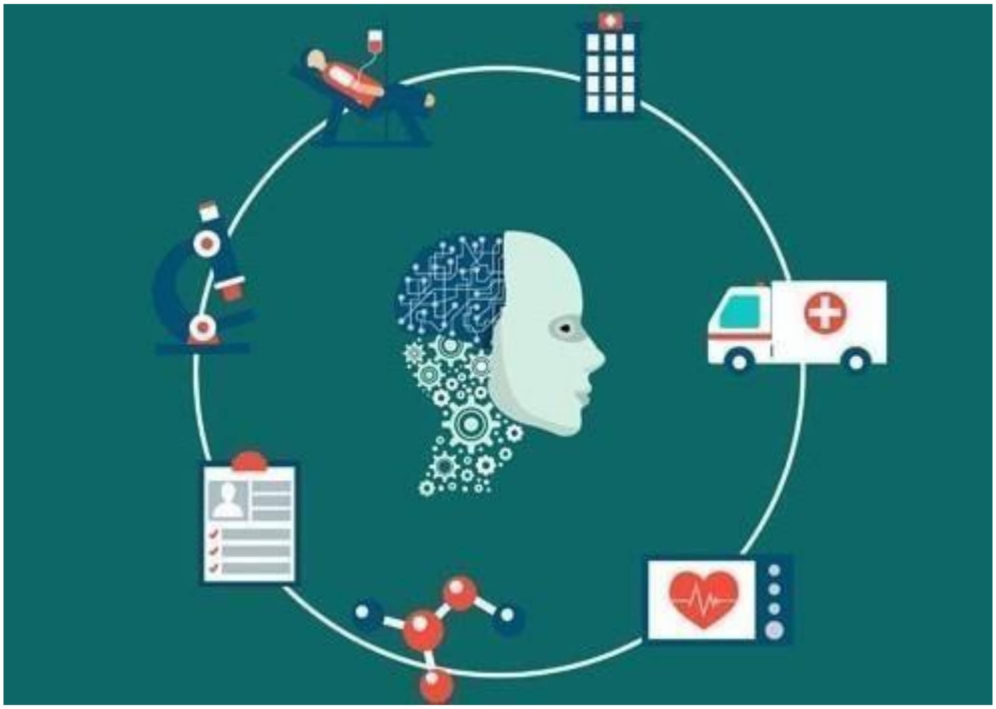
1. Frequent updates: TensorFlow releases different updates every 2-3 month, increasing the overhead for auser to install it and bind it with the existing system.

2. Inconsistent: TensorFlow provides homonyms that share similar names but different implementations,which makes it confusing to remember and use.

3. Architectural limitation: TensorFlow's architecture TPU only allows the execution of a model not to trainit.Every code needs to be executed using any platform for its support which increases the dependency for the execution.

4. Symbolic loops: TensorFlow lags at providing the symbolic loops for indefinite sequences. Hence it isreferred to as a low-level API.

6 AI in Healthcare



6.1 How AI works in healthcare

AI is able to analyze large amounts of data stored by healthcare organizations in the form of images, clinical research trials and medical claims, and can identify patterns and insights often undetectable by manual human skill sets.

6.2 Uses of AI in healthcare

1. AI supports medical imaging analysis. It supports a clinician reviewing images and scans. This enablesradiologists or cardiologists to identify essential insights for prioritizing critical cases, to avoid potential errors in reading electronic health records (EHRs) and to establish more precise diagnoses.

2. AI supports health equity. The AI and ML industry has the responsibility to design healthcare systemsand tools that ensure fairness and equality are met, both in data science and in clinical studies, in order to deliver the best possible health outcomes.

3. AI provides valuable assistance to emergency medical staff.. AI can analyze both verbal and nonverbalclues in order to establish a diagnostic from a distance. Corti is an AI tool that assists emergency medicine staff.

4. AI builds complex and consolidated platforms for drug discovery. AI algorithms are able to identify newdrug applications, tracing their toxic potential as well as their mechanisms of action. This technology led to the foundation of a drug discovery platform that enables the company to repurpose existing drugs and bioactive compounds.

5. AI can decrease the cost to develop medicines. Supercomputers have been used to predict from databasesof molecular structures which potential medicines would and would not be effective for various diseases.

6. AI analyzes unstructured data. Clinicians often struggle to stay updated with the latest medical advanceswhile providing quality patient-centered care due to huge amounts of health data and medical records. EHRs and biomedical data curated by medical units and medical professionals can be quickly scanned by ML technologies to provide prompt, reliable answers to clinicians.

7 Supervised Learning

7.1 What is supervised learning?

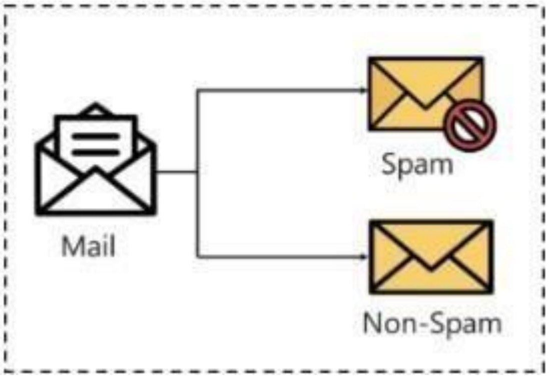
Supervised learning, also known as supervised machine learning, is a subcategory of machine learning and artificial intelligence. It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of the cross validation process. Supervised learning helps organizations solve for a variety of real-world problems at scale, such as classifying spam in a separate folder from your inbox.

How supervised learning works Supervised learning uses a training set to teach models to yield the desired output. This training dataset includes inputs and correct outputs, which allow the model to learn over time. The algorithm measures its accuracy through the loss function, adjusting until the error has been sufficiently minimized.

Supervised learning can be separated into two types of problems when data mining classification and regression:

7.2 Classification

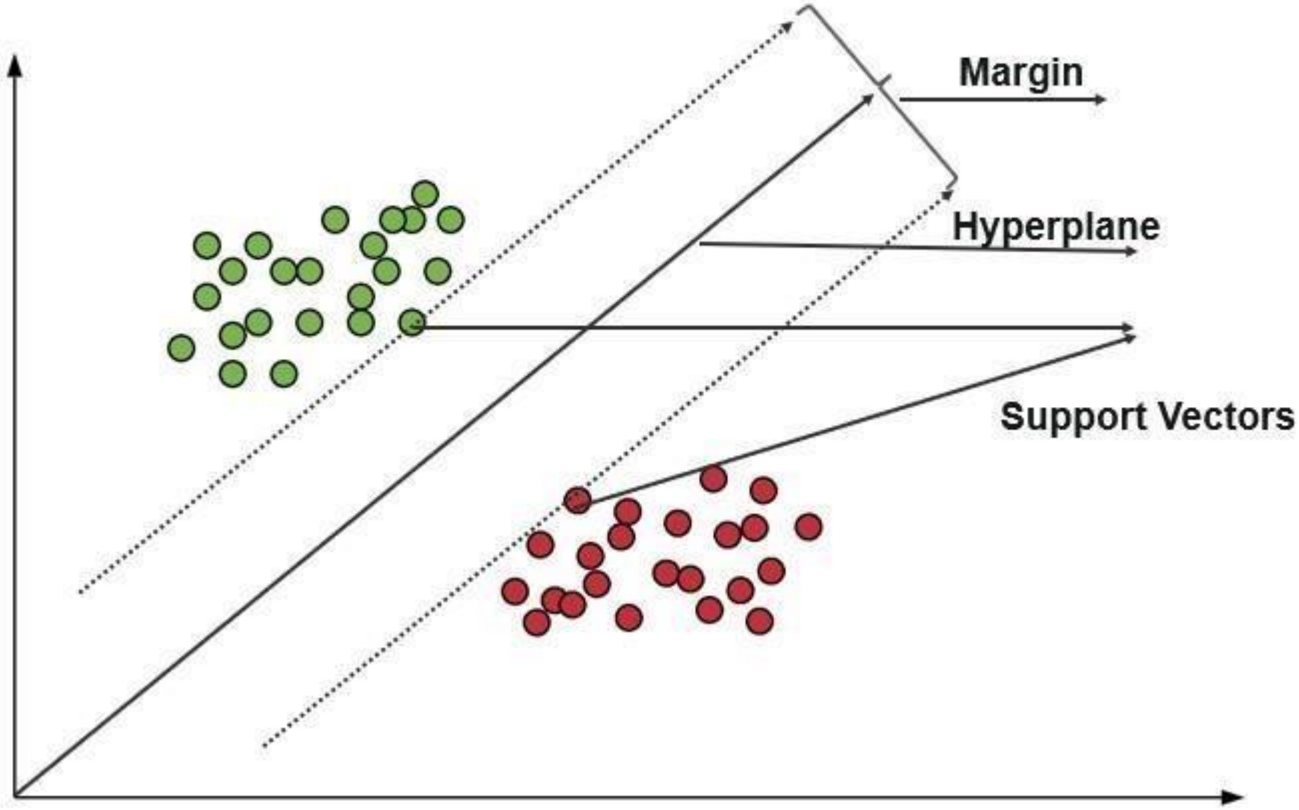
Classification is a process of categorizing a given set of data into classes, It can be performed on both structured or unstructured data. The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories. The classification predictive modeling is the task of approximating the mapping function from input variables to discrete output variables. The main goal is to identify which class/category the new data will fall into.



Common classification algorithms are linear classifiers, support vector machines (SVM), decision trees, k-nearest neighbor, and random forest, which are described in more detail below.

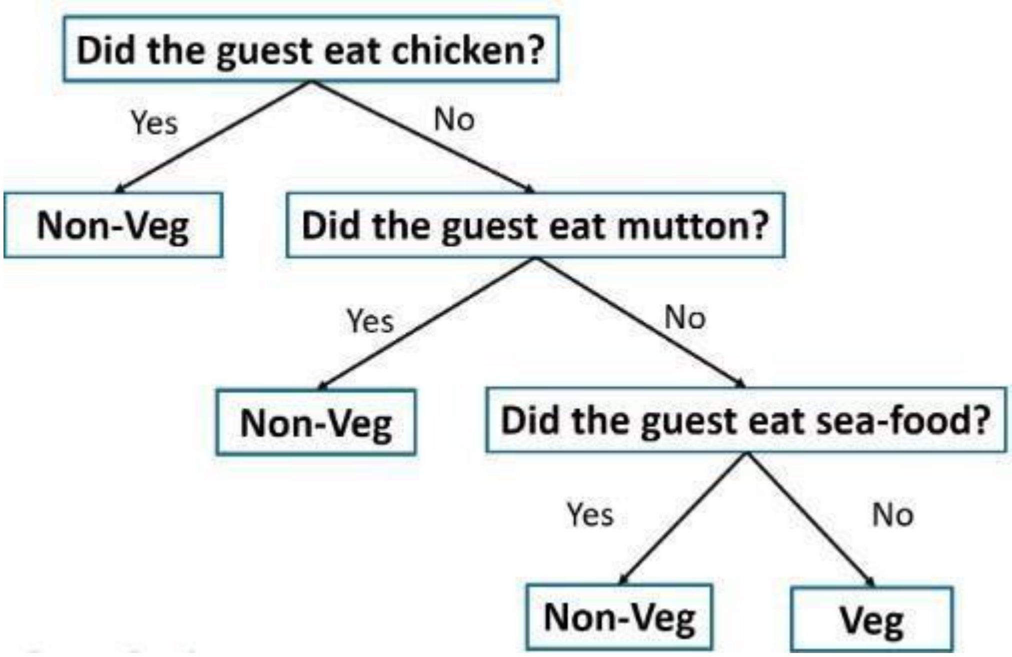
7.2.1 Support Vector Machine

The support vector machine is a classifier that represents the training data as points in space separated into categories by a gap as wide as possible. New points are then added to space by predicting which category they fall into and which space they will belong to.It uses a subset of training points in the decision function which makes it memory efficient and is highly effective in high dimensional spaces.



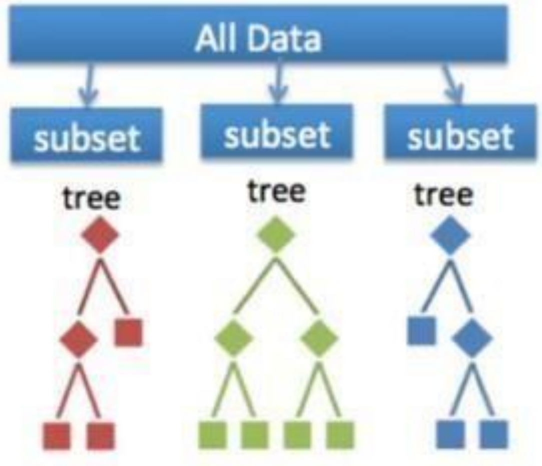
7.2.2 Decision Tree

The decision tree algorithm builds the classification model in the form of a tree structure. It utilizes the if-then rules which are equally exhaustive and mutually exclusive in classification. The process goes on with breaking down the data into smaller structures and eventually associating it with an incremental decision tree. The final structure looks like a tree with nodes and leaves. The rules are learned sequentially using the training data one at a time. Each time a rule is learned, the tuples covering the rules are removed. The process continues on the training set until the termination point is met.


Did the guest eat chicken?
Yes
No
Non-Veg
Did the guest eat mutton?
Yes
No
Non-Veg
Did the guest eat sea-food?
Yes
No
Non-Veg
Veg

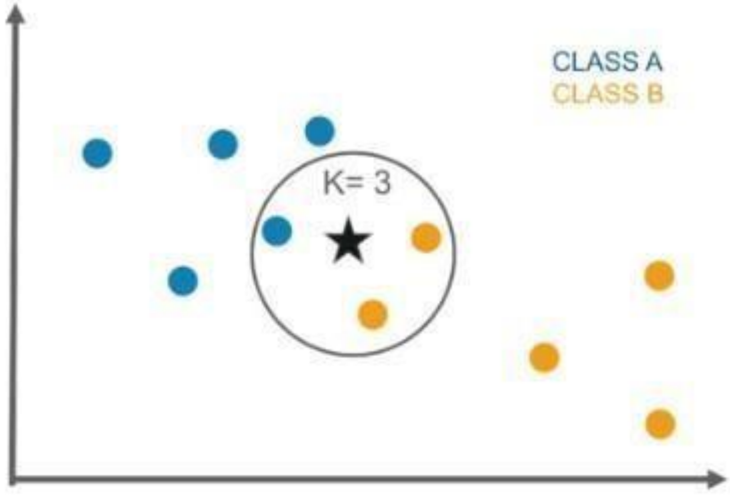
7.2.3 Random Forest

random forest are an ensemble learning method for classification, regression, etc. It operates by constructing a multitude of decision trees at training time and outputs the class that is the mode of the classes or classification or mean prediction(regression) of the individual trees.A random forest is a meta-estimator that fits a number of trees on various subsamples of data sets and then uses an average to improve the accuracy in the model's predictive nature. The sub-sample size is always the same as that of the original input size but the samples are often drawn with replacements.



7.2.4 K-Nearest Neighbor

It is a lazy learning algorithm that stores all instances corresponding to training data in n-dimensional space. It is a lazy learning algorithm as it does not focus on constructing a general internal model, instead, it works on storing instances of training data. Classification is computed from a simple majority vote of the k nearest neighbors of each point. It is supervised and takes a bunch of labeled points and uses them to label other points. To label a new point, it looks at the labeled points closest to that new point also known as its nearest neighbors. It has those neighbors vote, so whichever label most of the neighbors have is the label for the new point. The k is the number of neighbors it checks.



7.3 Regression

Regression is used to understand the relationship between dependent and independent variables. The main goal of regression is the construction of an e cient model to predict the dependent attributes from a bunch of attribute variables. A regression problem is when the output variable is either real or a continuous value i.e salary, weight, area, etc. We can also de ne regression as a statistical means that is used in applications like housing, investing, etc. It is used to predict the relationship between a dependent variable and a bunch of independent variables. Let us take a look at various types of regression techniques.Linear regression, logistical regression, and polynomial regression are popular regression algorithms.

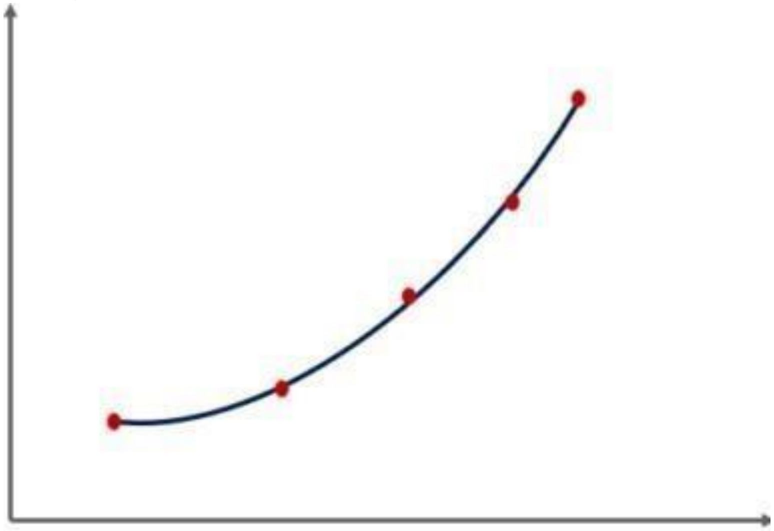
7.3.1 Simple Linear Regression

In this, we predict the outcome of a dependent variable based on the independent variables, the relationship between the variables is linear.

Simple linear regression is a regression technique in which the independent variable has a linear relationship with the dependent variable. The straight line in the diagram is the best t line. The main goal of the simple linear regression is to consider the given data points and plot the best t line to t the model in the best way possible.

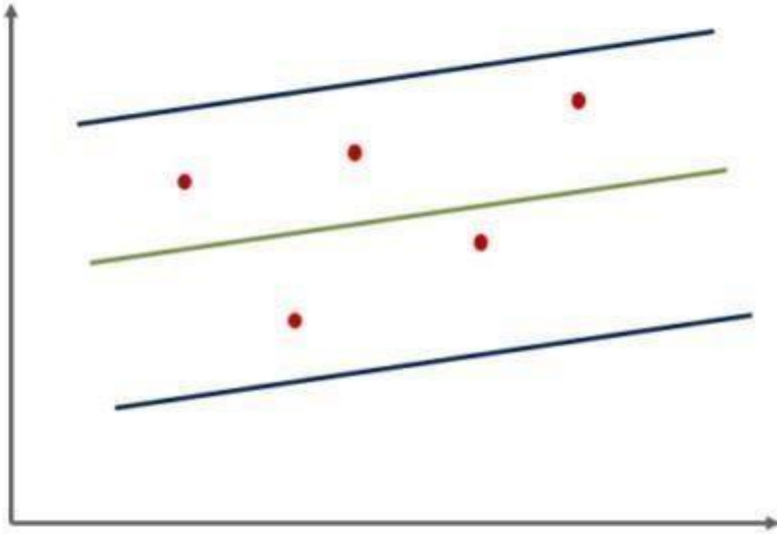
7.3.2 Polynomial Regression

In this regression technique, we transform the original features into polynomial features of a given degree and then perform regression on it.



7.3.3 Support Vector Regression

For support vector machine regression or SVR, we identify a hyperplane with maximum margin such that the maximum number of data points are within those margins. It is quite similar to the support vector machine classification algorithm. Instead of minimizing the error rate as in simple linear regression, we try to fit the error within a certain threshold. Our objective in SVR is to basically consider the points that are within the margin. Our best fit line is the hyperplane that has the maximum number of points.



7.3.4 Decision Tree Regression

A decision tree can be used for both regression and classification. In the case of regression, we use the ID3 algorithm(Iterative Dichotomiser 3) to identify the splitting node by reducing the standard deviation. A decision tree is built by partitioning the data into subsets containing instances with similar values (homogenous). Standard deviation is used to calculate the homogeneity of a numerical sample. If the numerical sample is completely homogeneous, its standard deviation is zero.

7.3.5 Random Forest Regression

In random forest regression, we ensemble the predictions of several decision tree regressions.Random forest is an ensemble approach where we take into account the predictions of several decision regression trees.

1. Select K random points
2. Identify n where n is the number of decision tree regressors to be created. Repeat steps 1 and 2 to createseveral regression trees.
3. The average of each branch is assigned to the leaf node in each decision tree.
4. To predict output for a variable, the average of all the predictions of all decision trees are taken intoconsideration.

8 Unsupervised learning

8.1 What is unsupervised learning?

Unsupervised learning, also known as unsupervised machine learning, uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. Its ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis, cross-selling strategies, customer segmentation, and image recognition. Unsupervised Learning algorithms work on datasets that are unlabelled and find patterns which would previously not be known to us. These patterns obtained are helpful if we need to categorize the elements or find an association between them

8.2 Clustering

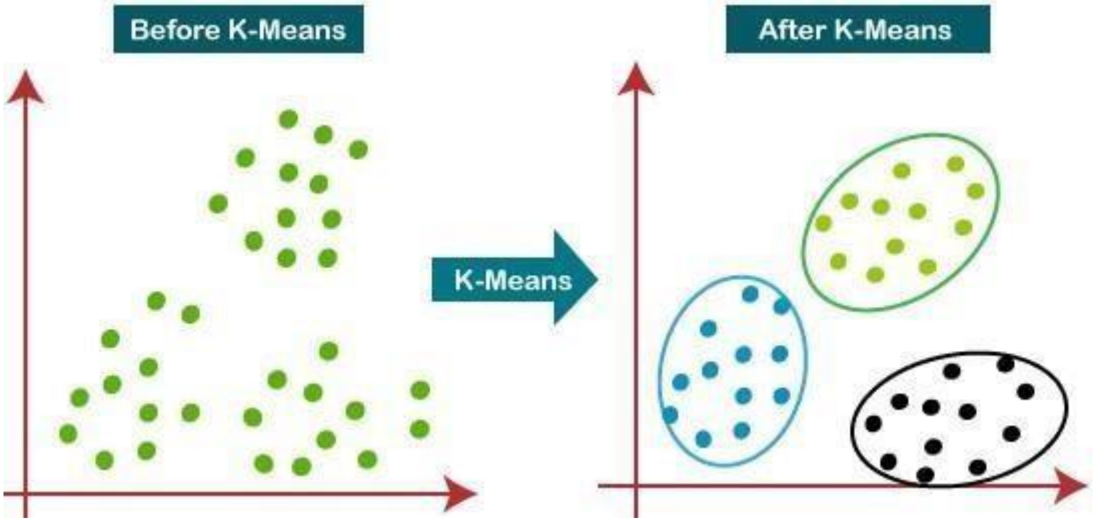
Clustering is the type of Unsupervised Learning where you find patterns in the data that you are working on. It may be the shape, size, colour etc. which can be used to group data items or create clusters. Some popular algorithms in Clustering are discussed below:

8.2.1 Hierarchical Clustering

This algorithm builds clusters based on the similarity between different data points in the dataset. It goes over the various features of the data points and looks for the similarity between them. If the data points are found to be similar, they are grouped together. This continues until the dataset has been grouped which creates a hierarchy for each of these clusters.

8.2.2 K-Means Clustering

This algorithm works step-by-step where the main goal is to achieve clusters which have labels to identify them. The algorithm creates clusters of different data points which are as homogenous as possible by calculating the centroid of the cluster and making sure that the distance between this centroid and the new data point is as less as possible. The smallest distance between the data point and the centroid determines which cluster it belongs to while making sure the clusters do not interlay with each other. The centroid acts like the heart of the cluster. This ultimately gives us the cluster which can be labelled as needed.



8.2.3 K-NN Clustering

This is probably the most simple of the Machine Learning algorithms as the algorithm does not really learn but rather classifies the new data point based on the datasets that have been stored by it. This algorithm is also called as a lazy learner because it learns only when the algorithm is given a new data point. It works well with smaller datasets as huge datasets take time to learn..