BE-46

NAME : PATHAK SEJAL SUDHIR
ROLL NO: BE 46
SUBJECT : LP III - MACHINE LEARNING

GROUP B.
ASSIGNMENT NO :- 01.

TITLE :- Predict the price of the Uber ride from a given pickup point of the agreed drop-off location. Perform following tasks :-
1. Pre-process the dataset
2. Identify outliers
3. Check the correlation.
4. Implement linear regression and random Forest regression models.
5. Evaluate the models and compare their respective score like R2, RMSE, etc.

DATASET :- The project is about on world's largest taxi company Uber Inc. Now it becomes really important to manage their data properly to come up with new business ideas to get best result.

OBJECTIVE :- Students should be able to preprocess dataset and identify outliers to check correlation and implement linear regression and random Forest regression models. Evaluate them with respective scores like R2, RMSE, etc.

PREREQUISITE :- 1. Basic knowledge of Python
2. Concept of preprocessing data
3. Basic knowledge of Data Science & Big Data analytics.

BE-46

THEORY :-

DATA PREPROCESSING: It is a process of preparing the raw data & making it suitable for machine learning model. It is crusial step while creating a ML model. When creating a ML project, it is not always a case that we come across the clean & formatted data. And while doing any operation with data, it is mandatory to clean it & put in a formatted way. so for this, we use data preprocessing tast.

Why do we need data pre-processing?

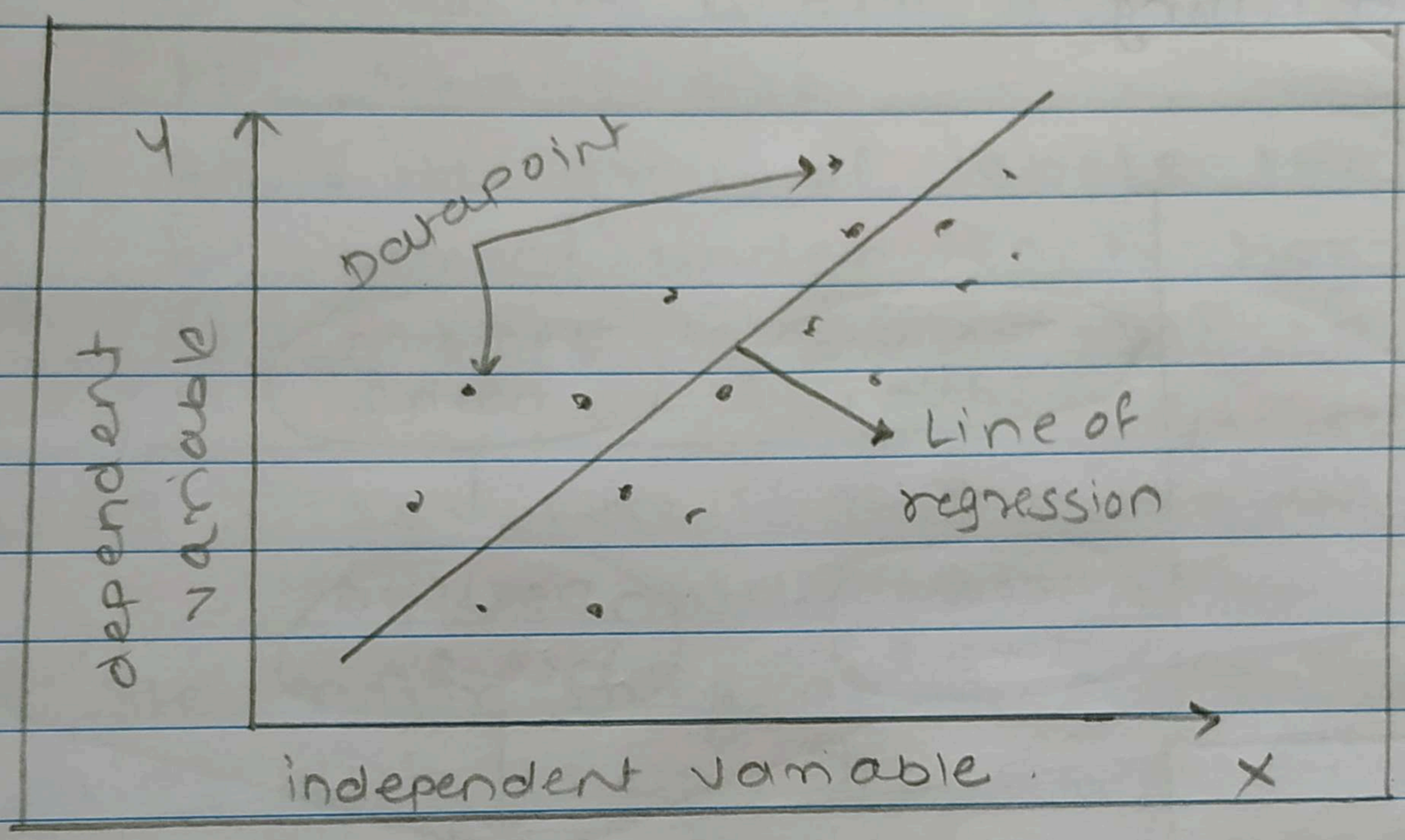
A real-world data generally contains noises, missing values and maybe in an unusable format which cannot be directly used for ML model. Data preprocessing is required tasks for cleaning the data and making it suitable for a ML model which increase the accuracy and efficiency of a ML model. It involves below steps

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing data
- Encoding categorial data
- Feature Scaling.

Linear Regression :-

① Linear regression is one of the easiest & most popular ML algorithms. It is a statistical method i.e. used for predictive analysis. Linear regression makes predictions for continuous/rea

or numericalue such as salary, age., sales.
Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independat (y) variables, hence called as linear regression.
Since LR shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the independent variable.

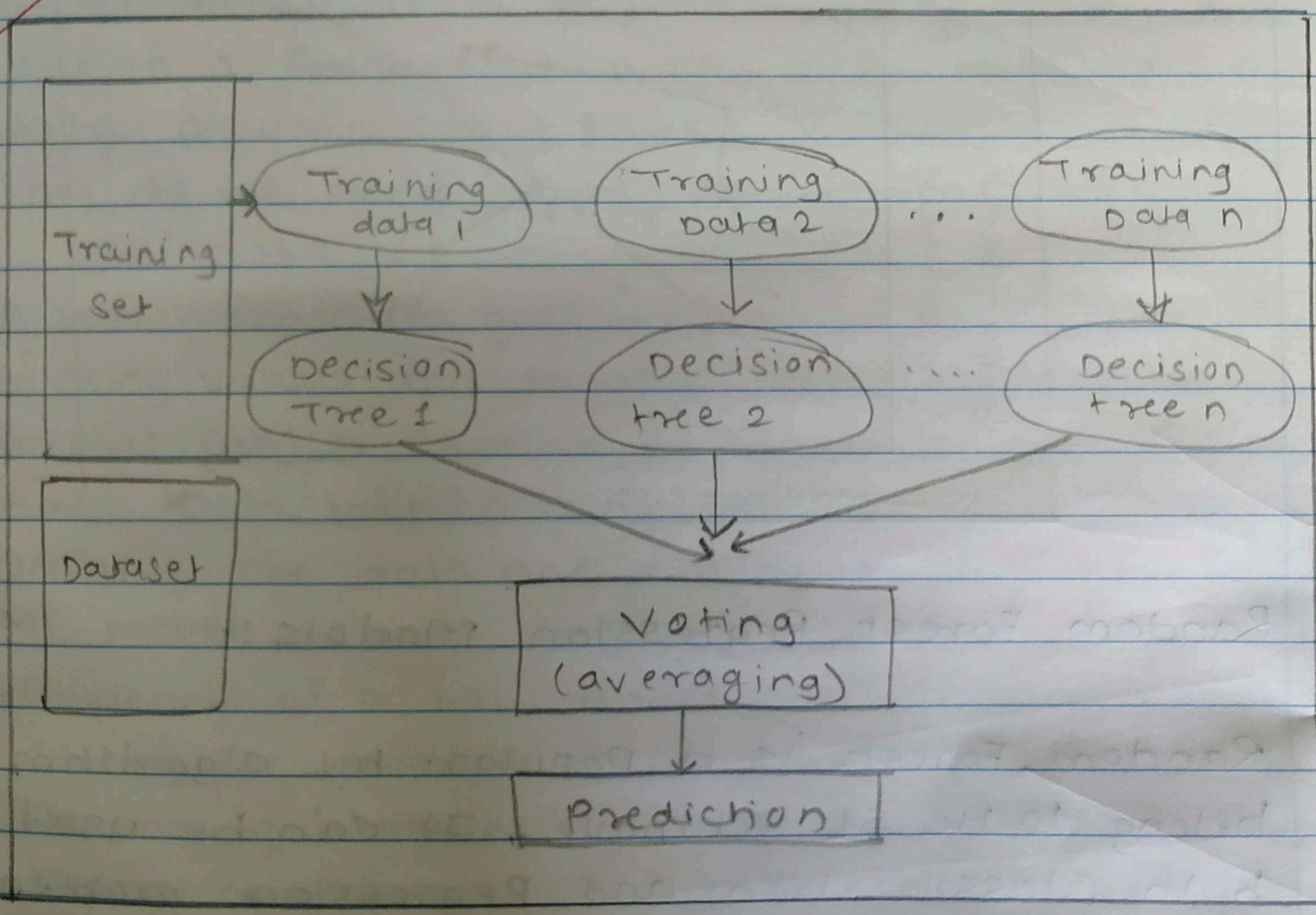


Random Forest Regression Models :-

Random Forest is a popular ML algorithm that belong to the SL technique. It can be used for both Classification and Regression problem in ML. It is based on the concept of ensemble learning, which is process of combining multiple classifier to solve a complex problem and to improve the performance of the model.
As the name suggest " Random Forest is a classifier that contains a no. of decision tree on various subse of the given dataset & takes the average to improve

the predictive accuracy of that dataset."
Instead of relying on one decision tree, the random forest takes the prediction from each tree & based on the majority votes of predictions, and it predicts the final output.

- The greater no of trees in the forest leads to higher accuracy and prevents the problem of over fitting.

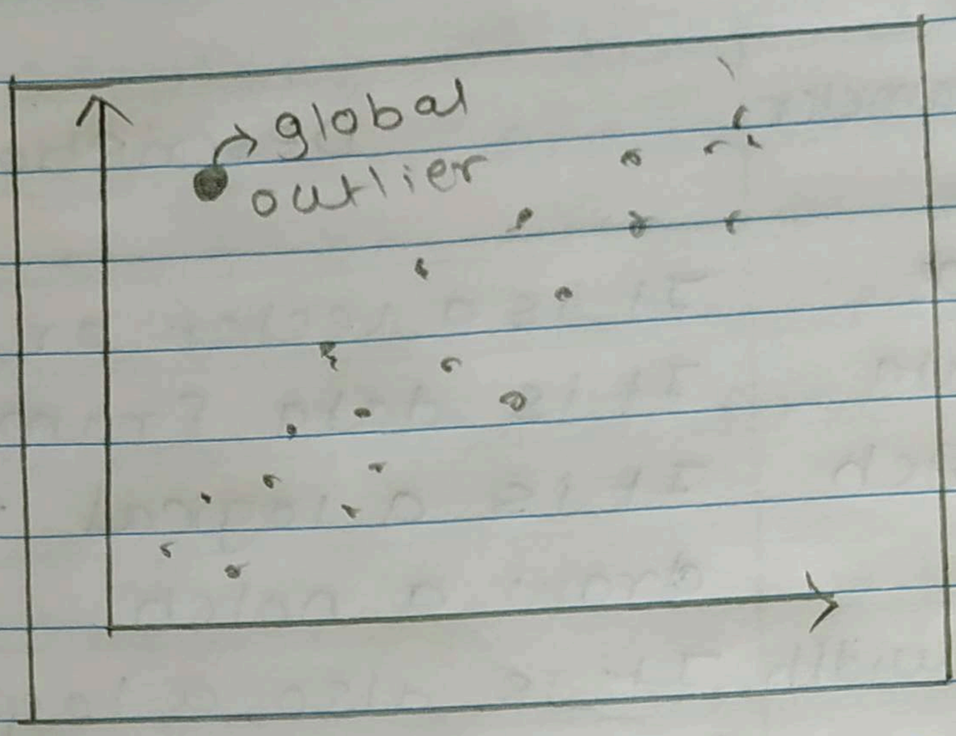


BOXPLOT :- Boxplot are a measure of how well data is distributed across a dataset
This divides the dataset into 3 quartiles. This graph represent the minimum, maximum, average, 1st quartile & 3rd quartile in dataset.
There is following Syntax of boxplot() Functio
boxplot(x, data, notch, varwidth, names, main)

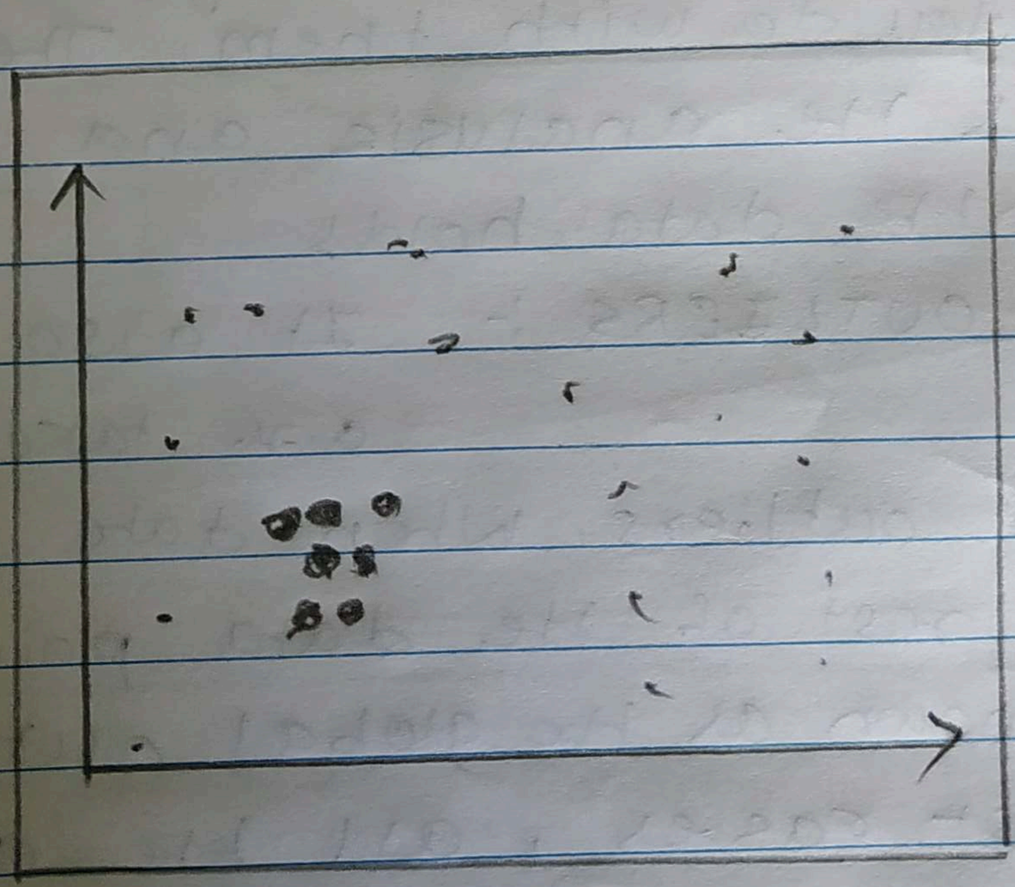
NO	Parameter	Description.
1.	x	It is a vector or a formula.
2.	data	It is data frame
3.	notch	It is a logical value. set as true to draw a notch
4.	varwidth	It is also a logical value set as true to draw the width of box same as the sample size.
5.	names	It is the group of labels that will be printed under each boxplot.
6	main	It is used to give a title to the graph.

OUTLIERS :- It refers to the data points that exists outside of what is to be expected. The major thing about the outliers is what you do with them. The data mining process involves the analysis and prediction of data that the data holds.

A) GLOBAL OUTLIERS :- It also point outliers. They are taken as the simplest as form of outliers. When data points deviate from all the rest of the data points in a given dataset, it is known as the global outliers
In most cases, all the outliers detection procedures are targeted to determine the global outliers.
The green data point is the global outlier.

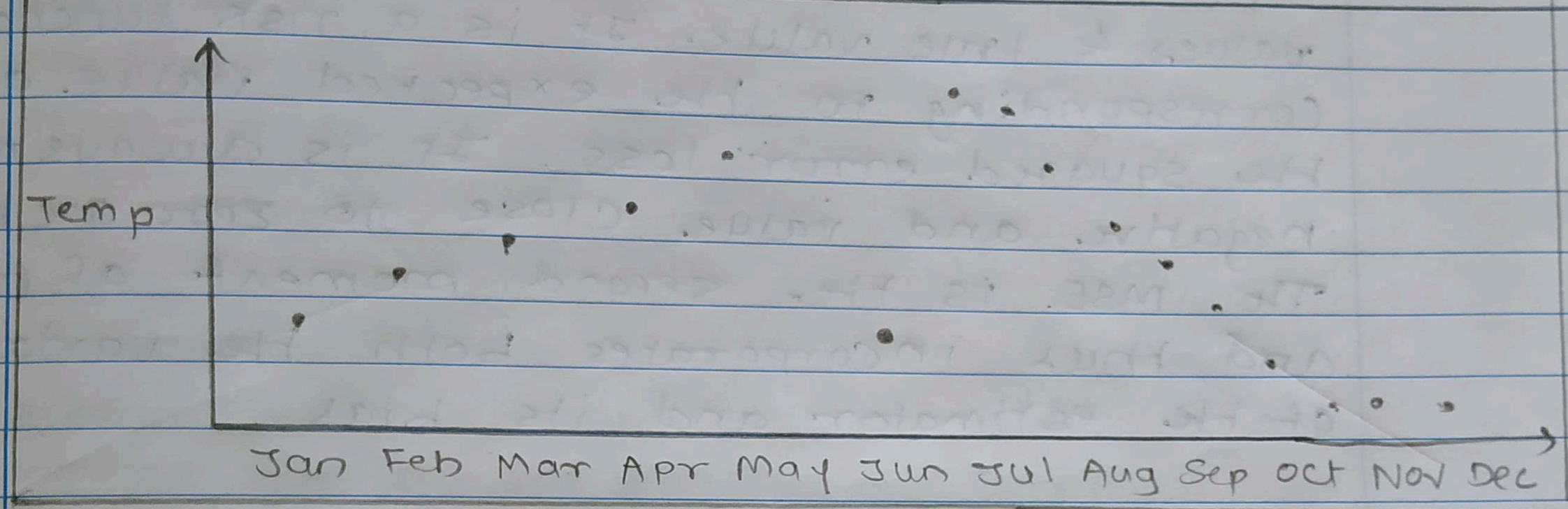


COLLECTIVE OUTLIERS :- When a group of data points deviates from the rest of the data set is called collective outliers. For example, in an Intrusion detection system, the DOS package from one system to another is taken as normal behaviour. Therefore, if this happens with the various computer, they are called collective outliers.



CONTEXUAL OUTLIERS :- "Contexual" means this Outlier introdu within a context. There are 2 types of attributes :- Contexual attributes behaviour attributes

contexual outlier analysis enables the users to examine outliers in different context & conditions, which can be useful in various applications.



Haversine :- The Haversine formula calculates the shortest distance between two points on a sphere using their latitudes. and logitudes measured along the surface. It is important for use in navigation.

Matplotlib :- It is an amazing visualization library in Python for 2D plots of arrays. It is a multi-platform data visualization library built on Numpy arrays and designed to work with the broader Scipy stack, It was introduced by John Hunter in the year 2002. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. It consists of several plot like line, bar, scatter, histogram, etc.

Mean Square Error :- The MSE or mean squared deviation (MSD) of an estimator measure the average of error squares i.e. the avg squared difference between the estimated values & true value. It is a risk function corresponding to the expected value of the squared error loss. It is always non negative and value close to zero are bett The MSE is the second moment of erro and thus incorporates both the variance of the estimator and its bias.

CONCLUSION :-

In this way we have explored concept correlation and implement linear regression and random Forest regression models.

Signature

BE-46

TSSM
Page No. :

NAME: PATHAK SEJAL SUDHIR
ROLL NO: BE 46
SUBJECT: LP III - MACHINE LEARNING
GROUP B

ASSIGNMENT NO :- 02

TITLE :- Classify the email using the binary classification method. Email spam detection has two states :-
a) Normal State - Not Spam.
b) Abnormal State - Spam.
Use K-Nearest Neighbour & Support Vector Machine For classification. Analyze Heir performance.

Dataset :- The CSV file contains 5172 rows, each row for each email. There are 3002 columns. The remaining 3000 columns are the 3000 most common words in all the emails, after excluding the non-alphabetical characters/words For information regarding all 5172 emails are stored in a compact dataframe rather than assepareate text files.

OBJECTIVE :- Student cable to classify email using the binary classification and implement email spam detection technique by using K-Nearest Neighbours & support vector machine Algorithm.

Data Preprocessing :- It is a process of preparing the raw data and making it suitable for a machine learning model. It is crusial step while creating a ML model.

When creating a ML project, it is not always a case that we come across the clean and formatted data. And while doing any operation with data, it is mandatory to clean it and put in a formatted way. so For this, we use data preprocessing task.

Why do we need Data Preprocessing?

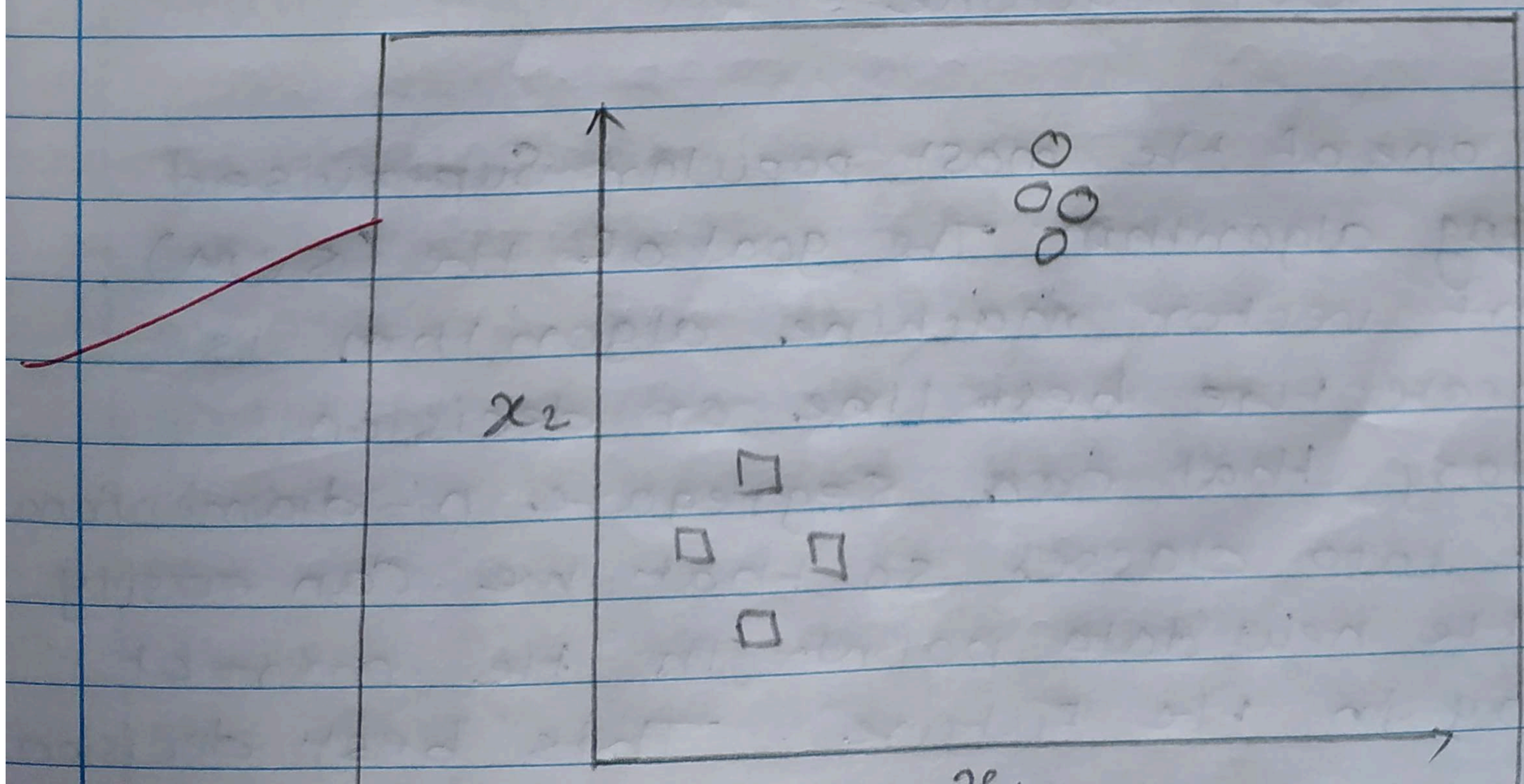
A real world data generally contains noises, missing values, and maybe in an unusable format which cannot be directly used for ML models. It is required tasks for cleaning the data & making it suitable for ML model which also increases the accuracy and efficiency of a ML model.

- Encoding categorial Data
- Splitting dataset into training & test set
- Feature Scaling

BINARY CLASSIFICATION:-

① In machine learning, binary classification is a supervised learning algorithm that categorized new observations into one of two classes.

② when we have to categorize given data into 2 distinct classes. Example - ~~On~~ on basis of health condition of a person, we have to determine whether the person has a certain disease or not.



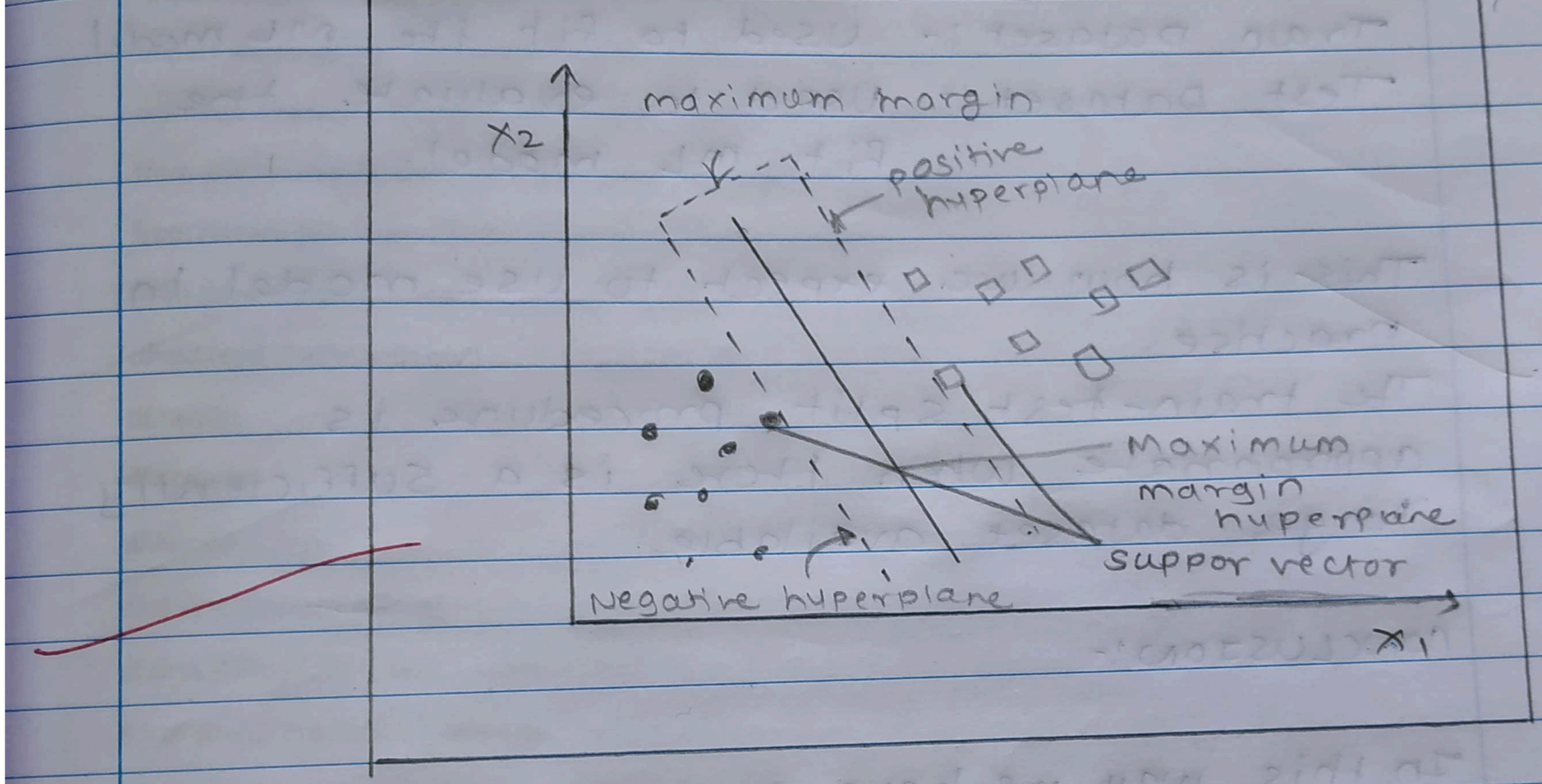
distance between the test data and all the training points.

Why do we need a KNN algorithm?

Suppose there are 2 categories i.e. category A and category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem we need a KNN algorithm.
With the help of KNN, we can easily identify the category or class of a particular dataset.

Support Vector Machine :-

It is one of the most popular Supervised learning algorithm. The goal of the (SVM) Support vector machine algorithm is to create the best line or decision boundary that can segregate n-diamentional space into classes so that we can easily put the new data point in the correct category in the future. This best decision



Train - Test Split Procedure :-

The train-test split is a technique for evaluating the performance of a machine learning program algorithm. It can be used for classification or regression problems & can be used for any supervised learning program.
The procedure involves taking a dataset & dividing it into two subsets.
The first subset is used to fit the model

predictions
values. The second dataset is referred to as the test dataset.

Train Dataset :- Used to fit the ML model

Test Dataset :- Used to evaluate the fit ML model.

This is how we expect to use model in practice.

The train-test split procedure is appropriate when there is a sufficiently large dataset available.

CONCLUSION :-

In this way we have explore the concepts of classification & implemented some methods of classification such as KNN, train-test split, Binary classification, SVM.

© Gitmati

BE-46

TSSM
Page No. :

NAME: PATHAK SEJAL SUDHIR
ROLL NO: BE 46
SUBJECT: LP III - MACHINE LEARNING
GROUP B
ASSIGNMENT NO :- 03.

TITLE :- Given a Bank customer, build a neural network based classifier that can determine whether they will leave or not in the next 6 months.

DATASET :- The case study is from an open-source dataset from Kaggle. The dataset contains 10,000 sample points with 14 distinct features such as CustomerId, CreditScore, Geography, Gender, Age, Balance, etc.

Perform the following step :-

1. Read the dataset.
2. Distinguish the feature and target set and divide the dataset into training and test sets.
3. Normalize the train and test data.
4. Initialize and build the model. Identify the points of improvement and implement the same.
5. Print the accuracy score and confusion matrix.

OBJECTIVE :- Student should be able to distinguish the feature & target set & divide the dataset into training & test sets and normalize them and students should build the model on the basis of that.

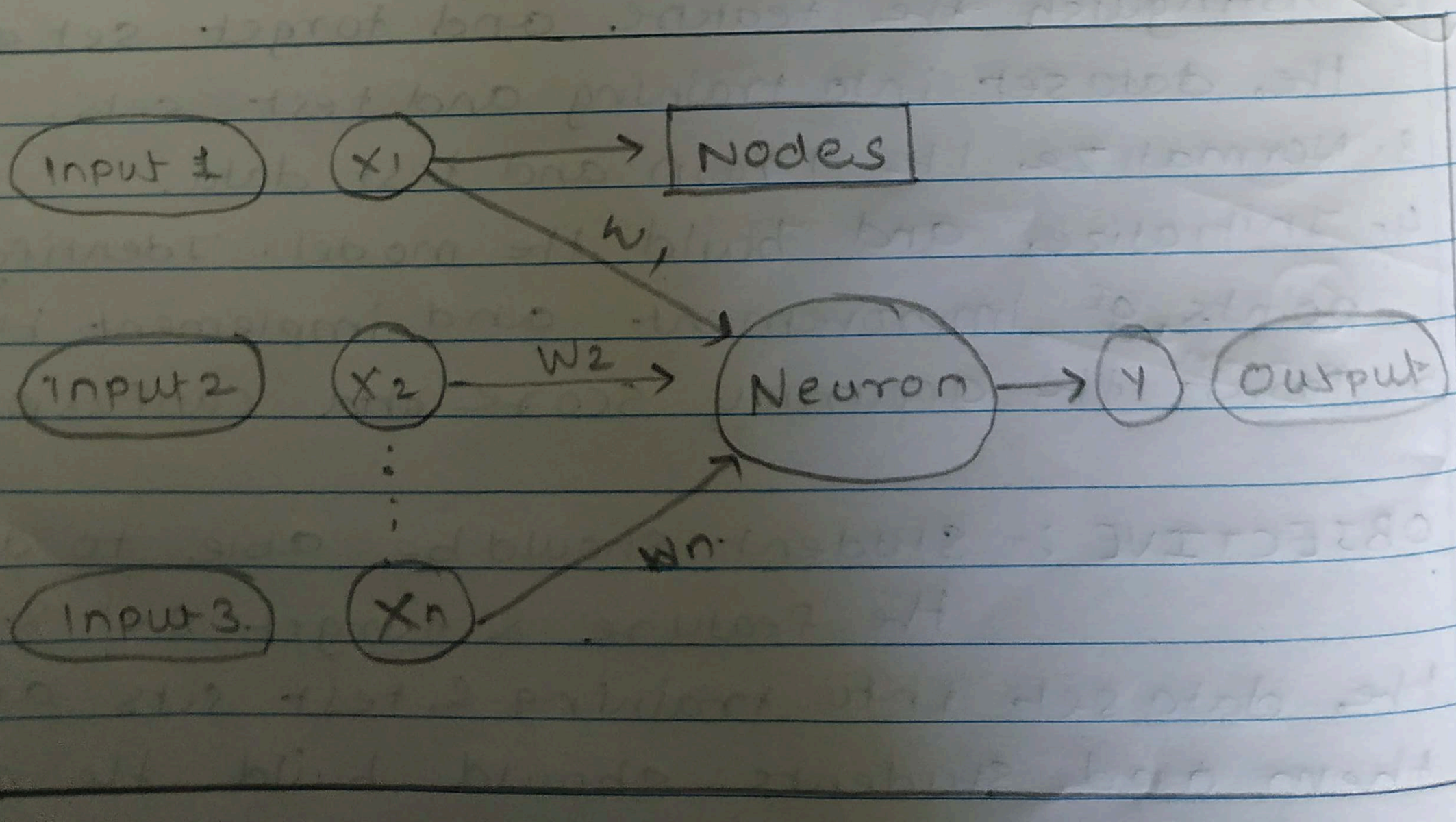
TSSM
Page No. :

BE-46

Prerequisite :- 1. Basic knowledge of Python
2. Concept of confusion Matrix.

Content of theory :-
1. Artificial Neural Network
2. keras
3. tensarflow
4. Normalization
5. Confusion Matrix.

Artificial Neural Network :- It is derived from Biological neural networks that develop the structure of a human brain. Similar to the human brain that has neurons interconnected to one other artificial neural networks also have neurons that are interconnected to one another in various layers of the networks.
The typical Artificial Neural Network looks like.



TSSM
Page No.

The architecture of an artificial neural network:

To understand the concept of architecture of an artificial neural network, we have to understand what a neural network consists of. In order to denote neural networks that consists of a larger number of artificial neurons, which are termed units arranged in a sequence of layers. Artificial Neural Network primarily consists of 3 layers:-

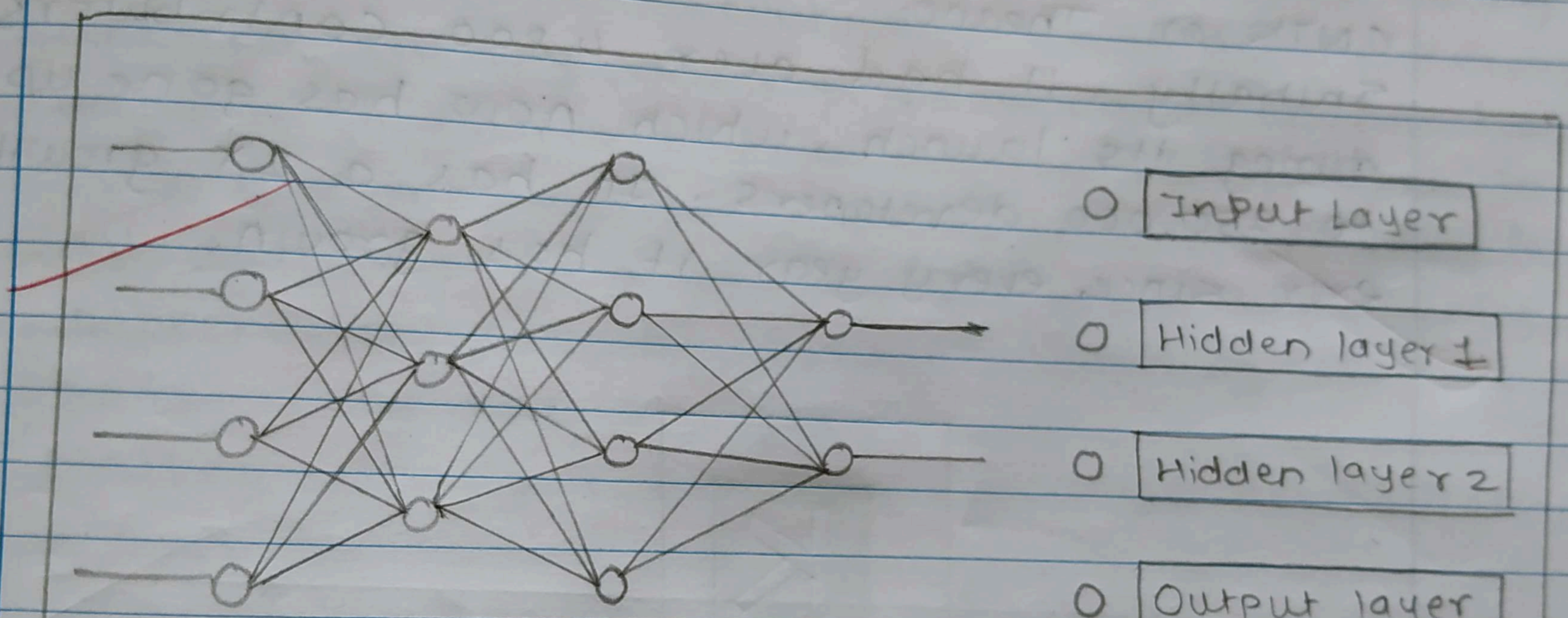


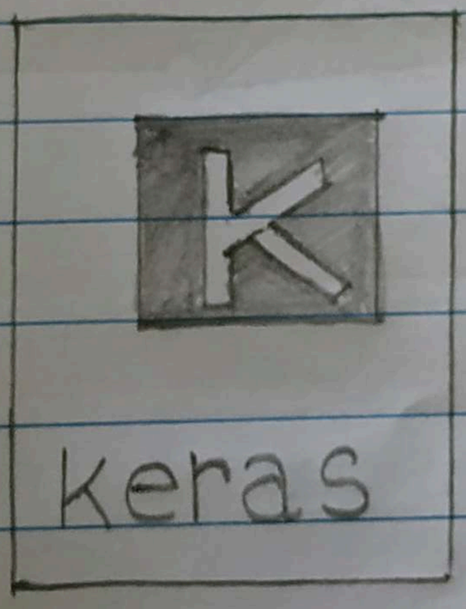
fig → Layers of Artificial Neural Network.

Input layer :- It accepts input in several different formats provided by the programmer.

Hidden layers :- It presents in-between input and output layers. It performs all calculation & hidden features, patterns

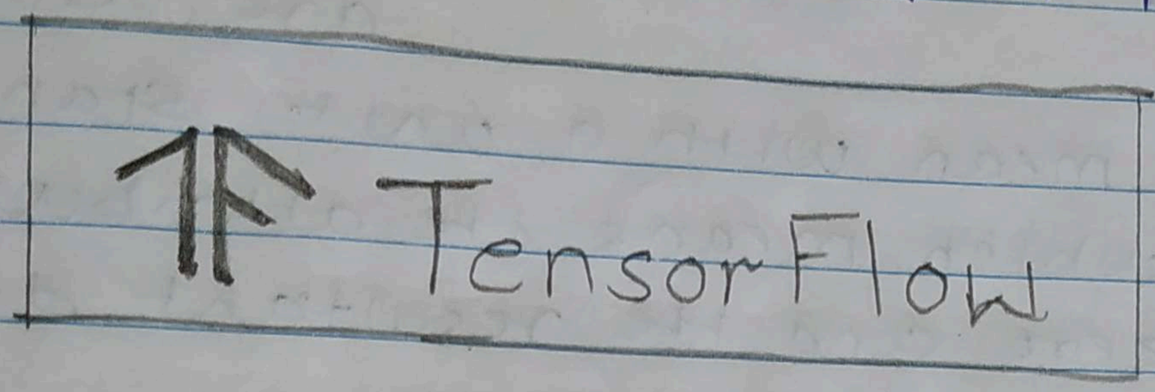
Output layers :- The input goes through series of transformations using the hidden layers, which results in output that is conveyed using this layer.

KERAS :- Keras is an open source high level Neural Network library, which is written in Python is capable enough to run on Theano, TensorFlow or CNTK.
It cannot handle low-level computations, so it makes use of Backend library to resolve it. The backend library act as a high-level API wrapper for the low-level API, which lets it run on TensorFlow, CNTK or Theano.
Initially, it had over 4800 contributors during its launch, which now has gone up to 250,000 developers. It has a 2X growth ever since every year it has grown.



TENSORFLOW :- It is Google product, which is one of the most famous deep learning tools widely used in the research area of machine learning & deep neural network. It came into market on 9th November 2015. It is built in such a way that it can easily run on multiple CPUs and GPUs

as well as on mobile operating systems. It consists of various wrappers in distinct language such Java, C++ or Python.



NORMALIZATION :- Normalization is a scaling technique in Machine learning applied during data preparation to change the values of numeric columns in the dataset to use a common scale. It is not necessary for all datasets in a model. It is required only when features of machine learning models have different ranges.

Mathematically, we can calculate normalization,

$$x_n = \frac{(X - X_{minimum})}{(X_{maximum} - X_{minimum})}$$

where,

x_n = value of Normalization
$X_{maximum}$ = Maximum value of feature
$X_{minimum}$ = Minimum value of a feature.

Normalization Techniques in Machine learning :-

1] Min-Max Scaling :- This technique help the dataset to shift &

rescale the values of their attributes, so they end up ranging between 0 and 1.

2] Standardization scaling :- In this values are centered around the mean with a unit standard deviation, which means the attribute becomes zero and the resultant distribution has unit standard deviation.

Standandization can be expressed as follow

$$x' = \frac{x - \mu}{\sigma}$$

Here,

μ = mean of Feature value

σ = standard deviation of Feature values.

CONFUSION MATRIX :- It is used to determine the performance of the classification models for a given set of test data. The matrix itself can be easily understood, but the related terminologies are confusing.

- For 2 prediction classes of classifier, the matrix is of 2*2 table.
- The matrix is divided into two dimensions, predicted value, actual value.
- Predicted are by model, actual value are by observations.

Predicted value	Actual values: Positive(1)	Actual values: Negative(0)
Positive(1)	TP	FP
Negative(0)	FN	TN

True Negative :- Model has given prediction NO, real or actual value was also NO.

True Positive :- Mode predicted Yes. real or actual value also Yes.

False Negative :- Model predicted No, actual value was 'Yes'. It is called Type II error

False Positive :- Model predicted Yes, actual value was No, It is called Type I error.

Calculation Using Confusion matrix.

1] Classification Accuracy :- To determine accuracy of classification problems. It calculated as the ratio of the number of correct predictions made by the classifier to all no of prediction made by classifier, Formula is as -

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

2] Misclassification rate / Error rate :-

It denote how often the model gives the wrong predictions. The value of error rate can be calculated as the no. of incorrect predictions to all number of predictions made by the classifier.

E

$$\text{Error rate} = \frac{FP + FN}{TP + FP + FN + TN}.$$

3] Precision : It can be denoted as the no of correct outputs provided by the model or out of all positive classes that have predicted correctly by the model, how many of them were actually true.

$$\text{Precision} = \frac{TP}{TP + FP}.$$

4] Recall : It denote as the out of total positive classes, how our model predicted correctly. the recall must be as high as possible

$$\text{Recall} = \frac{TP}{TP + FN}$$

5] F-measure : If two models have low precision and high recall or vice-versa, it is difficult to

compare these models. So, for this purpose, we can use F-score. This score helps us to evaluate the recall and precision at the same time.

$$\text{F-measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

Other important terms used in confusion matrix:

- Null Error rate: It denotes how often our model would be incorrect if it always predicted the majority class. As per the accuracy paradox, it is said that "the best classifier has a higher error rate than the null error rate.
- ROC Curve: It display classifier's performance for all possible thresholds. The graph is plotted between the true positive rate and the false positive rate.

CONCLUSION:-

In this way we build a neural network based classifier that can determine whether they will leave or not in the next 6 months.

TSSM
Page No. :

NAME : PATHAK SEJAL SUDHIR
ROLL NO : BE 46
SUBJECT : LP III - MACHINE LEARNING
Group B
ASSIGNMENT NO :- 05

TITLE :- Implement k-Nearest Neighbours algorithm on diabetes.csv dataset compute confusion matrix, accuracy, error rate, precision and recall on the given dataset

Dataset :- The datasets consists of several medical predictor variables and 1 target variables, Outcome.

OBJECTIVE :- Student should be able to preprocess dataset & identify outliers, to check correlation & implement KNN algorithm and random Forest classification models. Evaluate them with respective scores like confusion matrix, accuracy-score, mean-squared-error, r^2-score - roc-auc-score, etc.

Prerequisite :- 1. Basic knowledge of Python
2. Concept of confusion matrix
3. Concept of roc-auc curve
4. Concept of random forest and KNN
5.

BE-46

TSSM
Page No.:

THEORY:- KNN is supervised machine learning model. KNN model works by taking datapoint and looking 'k' closest labelled datapoint. The data point is then assigned to label of majority of 'k' closest point.

To start, we will use pandas to read data. After taking diabetes.csv into pandas framework. Next, call 'shape' function to see no of rows & column in the data.
row indicate no. of patience.
column indicate no. of Features.

Split up dataset into input & target. where input (x) is every coloumn expect diabetes. as diabetes will be our target.
Will use pandas 'drop' function to drop column diabetes from dataframe and stored it in variable (x) this will be our inputs.
We will insert diabetes column of our dataset into target variable (y).

Split the dataset into train and test data.
Now we will split dataset into training data & testing data. Training data is

the data that the model will learn from. The testing data is the data we will use to see how well the model performs on unseen data.

Scikit learn has a function 'train-test_split' that makes it easy for us to split our data into training & testing data.

'train_test_split' takes in 5 parameters. First 2 parameters are i/p and target data split up earlier. 'test_size' = 0.2 means 20% of all data will use for testing, which leaves 80% of data as training for model to learn from. 'random_state' = 1 ensure that we get same split each time so we reproduce our result. Stratify = y make train_split represent proportion of each value in y variable.

```
train_test_split (X, y, test_size = 0.2, random_state = 1,
                  Stratify = y).
```

Building & traing Model.

We create new KNN classifier and set 'n_neighbour' to 3, a new datapoint is labelled with by majority from the 3 nearest point.

Next, we need to train model. we use 'fit' function and passed training data as parameter to fit our model to training data

Testing the model.

once model is trained we can use predict function on model to prediction on our train data.

Now let's see how accurate model on the full test set. To do this 'score' function & pass in our test i/p and target data to see how well our model predict match up to actual result.

K-Fold cross validation :- Cross-validation is when the dataset is randomly splitup into k group one of the group is use as test and rest are used as training set. model is train on training set & scored on test set.

The train-test-split method is called 'holdout' cross validation is better using holdout method because holdout method score is dependent on how the data is split into train and test set.

Cross validation gives the mode an opportunity to test on multiple split so we can get better idea on how model is performed on unseen data.

Using cross validation our mean score is about 71.36%. This is more accurate representation of how our model will perform on unseen data. then our earlier testing using holdout method.

Hypertuning Model parameters using GridSearchCV

When built our initial KNN model we set the parameters n-neighbour to 3 as a starting

with no real logic behind the choice.
Hupertuning parameter is when you go through a process to find optimal parameters for your model to improve accuracy.
GridSearchCV works by training model multiple time on range of parameter that we specify. We can test our model with each parameter and figure out the optimal value to get best accuracy results.
For our model, we will specify range of value for 'n-neighbours' in order to see which value work best for our model to do this we will create dictionary, setting 'n-neighbours' as the key & using numpy to create an array of values from 1 to 24.

By using grid search find optimal parameter for our model, we have improved our model accuracy by over 4%.

CONCLUSION:

In this way we build a neural network-based classifier that can determine whether will leave or not in the next 6 month.

(c)

GROUP B
ASSIGNMENT NO :-

TITLE :- Implement K-means clustering / hierarchical clustering on sale_data_sample.csv dataset Determine the number of cluster using elbow method.

Data Description :- The data includes the Following features

1. Customer Id 2. Customer gender 3. Customer Age.
4. Annual Income of customer (in thousand dollars)
5. Spending source of customer (based on the customer behaviour and spending nature).

OBJECTIVE :- Student should able to understand how to use unsupervised learning segment different - different clusters or groups and used to them to train your model to predict future things.

Prerequisite :- 1. Knowledge of python
2. Unsupervised learning
3. Clustering
4. Elbow method.

Clustering algorithm try to find natural cluster in data the various aspects of how algorithm to cluster data can be turned and modified.

on principle that items within the same cluster must be similar to each other. The data is grouped in such a way that related elements are close to each other

Use of clustering :-

1] Marketing :- To identify various customer groups with existing customer data based on that customer can be provided with discount, offer, etc.

2] Real Estate :- to understand & divide various property location based on value and importance. It can process through data and identify various groups of property on basis of probable price.

3] Book store and library management :- libraries & Bookstores can use clustering to better manage the book database with proper book ordering better option can be implemented.

4] Document analysis :- often, we need to group together various reserch text and document according to similarity and in such case, we don't have any label. manually labelling large amounts of data is also not possible. Using clustering, algorithm can process the text and group into different themes.

K-means clustering :- It is unsupervised machine learning algorithm that

cluster. that need to be
It is centroid based algorithm in which each cluster is associated with a centroid. The main idea is to reduce to the distance between data point and their respective cluster centroid.
The algorithm takes raw unlabelled data as an input and divide the dataset into cluster and the process is repeated until the best cluster are found.
K. means is very easy and simple to implement. It is highly scalable can be applied to both small and large dataset. there is however problem which choosing the number of cluster or k. also, with the increase in dimentions stability decreases but overall K-means is a simple and robust algorithm that makes clustering very easy.

Data distribution:-

- Annual income falls between 50K to 85K
- There are cutomers of a wide variety of ages
- Maximum spending score is in range 40 to 60
- More female are customer than male

Now as anual income & spending score hold some patterns. we calculate sum of squared Errors for different K values. Next we choose the k for which WSS first starts to dimish. This value of K gives us the best number of clusters to make from the raw data. This result in elbow plot graph, the x axis being the no. of clusters, It is taken at the elbow joint point. This point is where value of WCSSS suddenly stop. Here 5 no. of clusters so we used K-means clustering to understand customer data. K-means is good clustering algorithm. Almost all the cluster have similar density. It is also fast and efficient in term of computational cost.

CONCLUSION:-

In this way we built K-means clustering on sales_data_sample.csv dataset and determine the no. of cluster using elbow method.