ZEAL EDUCATION SOCIETY's

ZEAL COLLEGE OF ENGINEEIRNG AND RESEARCH, NARHE, PUNE

DEPARTMENT OF COMPUTER ENGINEERING

SEMESTER-I

[A.Y.: 2022 - 2023]

Laboratory Practice III
(410246)

LABORATORY MANUAL

Institute and Department Vision and Mission

INSTITUTE VISION	To impart value added technological education through pursuit of academic excellence, research and entrepreneurial attitude.
INSTITUTE MISSION	**M1:** To achieve academic excellence through innovative teaching and learning process. **M2:** To imbibe the research culture for addressing industry and societal needs**.** **M3:** To provide conducive environment for building the entrepreneurial skills. **M4:** To produce competent and socially responsible professionals with core human values.

DEPARTMENT VISION	To emerge as a department of repute in Computer Engineering which produces competent professionals and entrepreneurs to lead technical and betterment of mankind.
DEPARTMENT MISSION	**M1:** To strengthen the theoretical and practical aspects of the learning process by teaching applications and hands on practices using modern tools and FOSS technologies. **M2:** To endavor innovative interdisciplinary research and entrepreneurship skills to serve the needs of Industry and Society. **M3:** To enhance industry academia dialog enabling students to inculcate professional skills. **M4:** To incorporate social and ethical awareness among the students to make them conscientious professionals.

Department
Program Educational Objectives (PEOs)

PEO1: To Impart fundamentals in science, mathematics and engineering to cater the needs of society and Industries.
PEO2: Encourage graduates to involve in research, higher studies, and/or to become entrepreneurs.
PEO3: To Work effectively as individuals and as team members in a multidisciplinary environment with high ethical values for the benefit of society.

Savitribai Phule Pune University

Third Year of Computer Engineering (2019 Course)
410246: Laboratory Practice III

Teaching Scheme:	**Credit**	**Examination Scheme:**
PR: 04 Hours/Week	02	TW: 50 Marks
		PR: 50 Marks

Course Objectives:

- Learn effect of data preprocessing on the performance of machine learning algorithms
- Develop in depth understanding for implementation of the regression models
- Implement and evaluate supervised and unsupervised machine learning algorithms

Course Outcomes:

On completion of the course, student will be able to-

- **Machine Learning :**

CO1: Apply preprocessing techniques on datasets

CO2: Implement and evaluate linear regression and random forest regression models.

CO3: Apply and evaluate classification and clustering techniques

- **Design and Analysis of Algorithms:**

CO4: Analyze performance of an algorithm.

CO5: Learn how to implement algorithms that follow algorithm design strategies namely divide and conquer, greedy, dynamic programming, backtracking, branch and bound.

Block chain Technology:

CO6: Interpret the basic concepts in Block chain technology and its applications

List of Assignments

Sr. No.	TITLE
	Group B
01	Predict the price of the Uber ride from a given pickup point to 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 scores like R2, RMSE, etc. Dataset link: https://www.kaggle.com/datasets/yasserh/uber-fares-dataset
02	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 Neighbors and Support Vector Machine for classification. Analyze their performance. Dataset link: The emails.csv dataset on the Kaggle https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv
03	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 Description: 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, Tenure, Balance, etc. Link to the Kaggle project: https://www.kaggle.com/barelydedicated/bank-customer-churn-modeling Perform following steps: 1. Read the dataset. 2. Distinguish the feature and target set and divide the data set 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 (5 points).
04	Implement K-Nearest Neighbors algorithm on diabetes.csv dataset. Compute confusion matrix, accuracy, error rate, precision and recall on the given dataset. Dataset link : https://www.kaggle.com/datasets/abdallamahgoub/diabetes.
05	Implement K-Means clustering/ hierarchical clustering on sales_data_sample.csv dataset. Determine the number of clusters using the elbow method. Dataset link : https://www.kaggle.com/datasets/kyanyoga/sample-sales-data

Group B: Assignment No 1

Aim: Predict the price of the Uber ride from a given pickup point to 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 scores like R2, RMSE, etc.Dataset link:
https://www.kaggle.com/datasets/yasserh/uber-fares-dataset

Objective:

Student will learn:
1] The basic concept and implementation logic of linear regression and random forest regression model.2] Different evaluation metrics used for regression models like R2, RMSE, etc.

Outcome:

After completion of this assignment students are able to understand the How to find the outliers and correlation between to Two variable, How to Calculate the R2, RMSE.

Theory Concepts:

1. Data Pre-Processing:

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

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

It involves below steps:

- **Getting the dataset**
- **Importing libraries**
- **Importing datasets**

- **Finding Missing Data**
- **Encoding Categorical Data**
- **Splitting dataset into training and test set**
- **Feature scaling**

1) Get the Dataset:

The collected data for a particular problem in a proper format is known as the **dataset**. To use the dataset in our code, we usually put it into a CSV **file**. However, sometimes, we may also need to use an HTML or xlsx file.

2) Importing Libraries:

In order to perform data preprocessing using Python, we need to import some predefined Python libraries. These libraries are used to perform some specific jobs. There are three specific libraries that we will use for data preprocessing, which are:Numpy, Matplotlib, Pandas.

3) Importing the Datasets:

Now we need to import the datasets which we have collected for our machine learning project. But before importing a dataset, we need to set the current directory as a working directory.

4) Handling Missing data:

The next step of data preprocessing is to handle missing data in the datasets. If our dataset contains some missing data, then it may create a huge problem for our machine learning model. Hence it is necessary to handle missing values present in the dataset.

Ways to handle missing data:

By deleting the particular row: The first way is used to commonly deal with null values. In this way, we just delete the specific row or column which consists of null values.

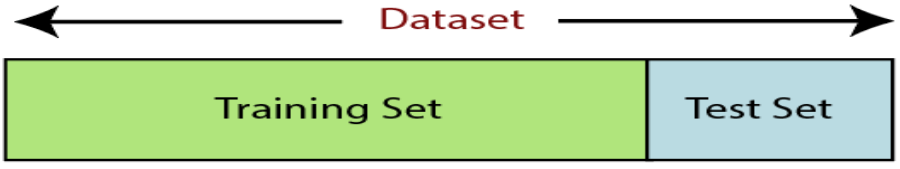
By calculating the mean: In this way, we will calculate the mean of that column or row which contains any missing value and will put it on the place of missing value. This strategy is useful for the features which have numeric data such as age, salary, year, etc. Here, we will use this approach.

5) Encoding Categorical data:

Machine learning model completely works on mathematics and numbers, but if our dataset would have a categorical variable, then it may create trouble while building the model. So it is necessary to encode these categorical variables into numbers.

6) Splitting the Dataset into the Training set and Test set:

In machine learning data preprocessing, we divide our dataset into a training set and test set. This is one of the crucial steps of data preprocessing as by doing this, we can enhance the performance of our machine learning model.



Training Set: A subset of dataset to train the machine learning model, and we already know the output.

Test set: A subset of dataset to test the machine learning model, and by using the test set, model predicts the output.

7) Feature Scaling:

Feature scaling is the final step of data preprocessing in machine learning. It is a technique to standardize the independent variables of the dataset in a specific range. In feature scaling, we put our variables in the same range and in the same scale so that no any variable dominate the other variable.

Outliers: Outlier is an observation which contains either very low value or very high value in comparison to other observed values. An outlier may hamper the result, so it should be avoided. Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distribution, and arise due to inconsistent data entry, or erroneous observations.

We can detect the outliers using Z-Score and Inter Quartile Range (IQR).

What is Linear Regression?

In a cause and effect relationship, the **independent variable** is the cause, and the **dependent variable** is the effect. **Least squares linear regression** is a method for predicting the value of a dependent variable *Y*, based on the value of an independent variable *X*.

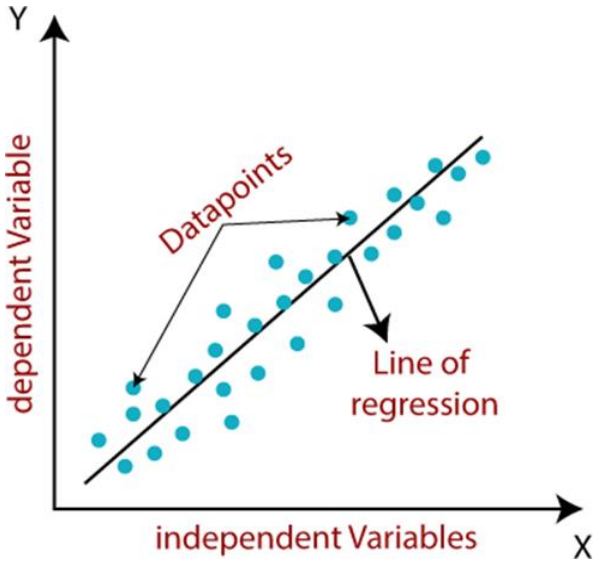
Linear Regression:

Linear regression is one of the easiest and most popular Machine Learning algorithms. It is a statistical method that is used for predictive analysis. Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

Linear regression algorithm shows a linear relationship between a dependent (y) and one or more independent

(y) variables, hence called as linear regression. Since linear regression 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.

The linear regression model provides a sloped straight line representing the relationship between the variables. Consider the below image:



Mathematically, we can represent a linear regression as: $y= a_0+a_1x+ \varepsilon$

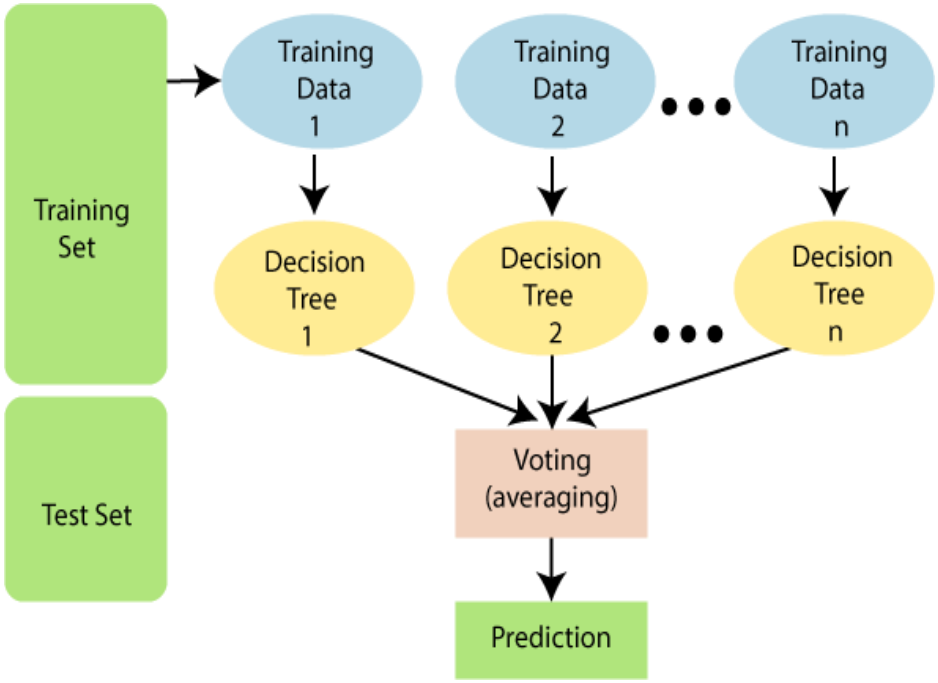
Here,
Y= Dependent Variable (Target Variable)
X= Independent Variable (predictor Variable)
a0= intercept of the line (Gives an additional degree of freedom)
a1 = Linear regression coefficient (scale factor to each
input value).ε = random error
The values for x and y variables are training datasets for Linear Regression model representation.

Random Forest:

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of *combining multiple classifiers to solve a complex problem and to improve the performance of the model.*

As the name suggests, *"Random Forest is a classifier that contains a number of decision trees on various subsetsof the given dataset and 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 and based on the majority votes of predictions, and it predicts the final output.

The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.The below diagram explains the working of the Random Forest algorithm
:



Evaluation metrics of regression models:

1] RMSE: Root Mean Squared Error is the square root of Mean Squared error. It measures the standarddeviation of residuals.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2}$$

2] R2: The coefficient of determination or R-squared represents the proportion of the variance in the dependent variable which is explained by the linear regression model. It is a scale-free score i.e. irrespective of the values being small or large, the value of R square will be less than one.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Python Packages needed

- pandas
  - Data Analytics
- numpy
  - Numerical Computing
- matplotlib.pyplot
  - Plotting graphs
- Sklearn
  - Regression Classes

Steps to establish Linear Regression

A simple example of regression is predicting weight of a person when his height is known. To do this we needto have the relationship between height and weight of a person.

The steps to create the relationship are −

- Carry out the experiment of gathering a sample of observed values of height and correspondingweight.
- Create the object of Linear Regression Class.
- Train the algorithm with dataset of X and y.
- Get a summary of the relationship model to know the average error in prediction. Also calledresiduals.
- To predict the weight of new persons, use the predict() function.

Conclusion: We have studied the Linear Regression and Random forest algorithm. Also implemented andevaluated the models using R2 and RMSE scores

Output:

- (Execute the program and attach the printout here)

Conclusion:

Thus we learn that to how to do the data preprocessing, ways for finding outliers, correlation, linear regression and random forest algorithm, finally compared the score of R2, and RMSE.

Viva Questions:

1. **What is Machine Learning?**
2. **What are the applications of machine learning?**
3. **What are the steps involved in Data preprocessing?**
4. **What is Linear Regression?**
5. **How to find the missing value?**
6. **Define Outliers. And also explain how to remove the outliers.**
7. **What is R2, RMSE, MAE,MSE.**
8. **How to calculate the Z-score? Explain with an example**

Date:	
Marks obtained:	
Sign of course coordinator:	
Name of course Coordinator :	

Group B: Assignment No 2

Aim: 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 Neighbors and Support VectorMachine for classification. Analyze their performance.
Dataset link: The emails.csv dataset on the Kaggle https://www.kaggle.com/datasets/balaka18/email-spam- classification-dataset-csv

Objective:

Student will learn:

1] The basic concept and implementation logic of K-Nearest Neighbors algorithm.
2] The basic concept and implementation logic of Support Vector Machine algorithm.
3] Different evaluation metrics used for classification models like accuracy, precision, recall, F-score, etc.

.

Outcome:

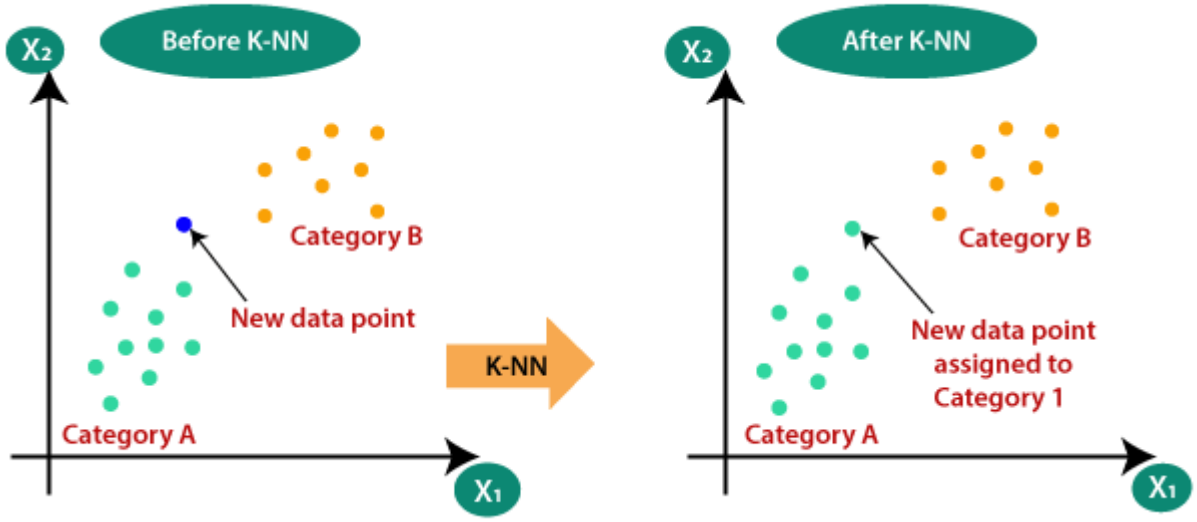
To implement the given Email is Normal State – Not Spam, b) Abnormal State – Spam.

Theory Concepts :

K-Nearest Neighbors (kNN) Algorithm-

- K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. KNN is an *non parametric lazy learning* algorithm. That is a pretty concise statement. When you say a technique is non parametric , it means that it does not make any assumptions on the underlying data distribution. It is also a lazy algorithm. What this means is that it does not use the training data points to do any *generalization*. In other words, there is *no explicit training phase* or it is very minimal. This means the training phase is pretty fast . Lack of generalization means that KNN keeps all the training data.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
- KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.
- **Example:** Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.

Why do we need a K-NN Algorithm?



How does K-NN work?

- **Step-1:** Select the number K of the neighbors
- **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
- **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbors, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.

Advantages of KNN Algorithm:

- It is simple to implement.
- It is robust to the noisy training data
- It can be more effective if the training data is large.

Disadvantages of KNN Algorithm:

- Always needs to determine the value of K which may be complex some time.
- The computation cost is high because of calculating the distance between the data points for all thetraining samples.

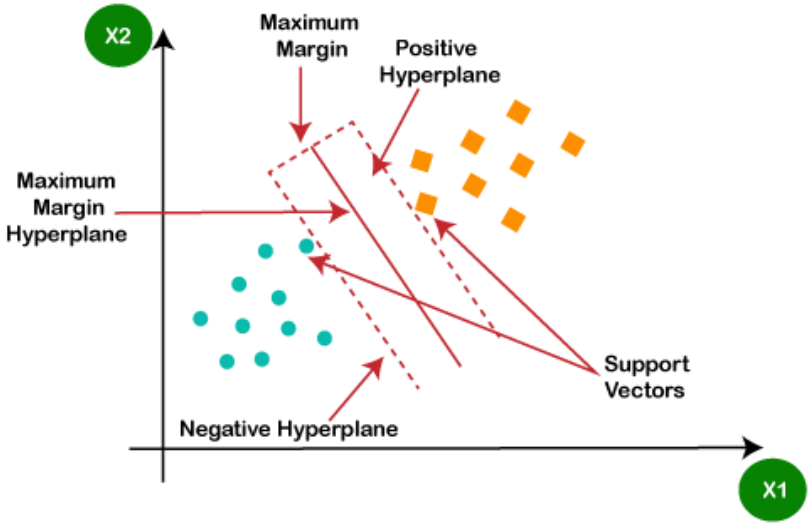
Support Vector Machine (SVM) Algorithm:

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problemsin Machine Learning.

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyper plane.



The hyper plane has classified dataset into two different classes.

Following are important concept in SVM

Support Vectors: Data points that are closed to the hyper plane is called support vectors.

Hyper plane: It is a decision plane or space which is divided between a set of objects having different classes.

Margin: It may be defined as the gap between two lines on the closed data points of different classes. It is calculated as the perpendicular distance from the line to the support vectors. Large margin is considered as a good margin and small margin as considered as bad margin.

SVM algorithm can be used for Face detection, image classification, text categorization, etc.

For analysis of performance of KNN and SVM, use different evaluation metrics like accuracy, precision,recall, F-score, etc.

Output:

(Execute the program and attach the printout here)

Conclusion:

We have studied the KNN and SVM algorithm. Also implemented and evaluated the modelsusing accuracy, precision, recall, F-score.

Viva Questions:

1. Define K-Nearest Neighbor algorithm.
2. How does KNN work?
3. Define Support Vector Machine.
4. What do you mean Hyper plane and Margin in SVM algorithm?
5. What is Confusion Matrix.
6. How to calculate the precision and recall?
7. What is F1 Score?
8. What are the advantages and dis advantages of KNN algorithm?

Date:	
Marks obtained:	
Sign of course coordinator:	
Name of course Coordinator :	

Group B: Assignment No 3

Aim: Given a bank customer, build a neural network-based classifier that can determine whether they willleave or not in the next 6 months.

Dataset Description: 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, Tenure, Balance, etc.

Link to the Kaggle project: https://www.kaggle.com/barelydedicated/bank-customer-churn-modeling Perform following steps:

1. Read the dataset.
2. Distinguish the feature and target set and divide the data set 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 (5 points).

Objective:

Student will learn:

1] The basic concept and implementation logic of normalization of data.

2] The basic concept and implementation logic of accuracy score and confusion matrix.

Theory:

Normalization in Machine Learning:

Normalization is one of the most frequently used data preparation techniques, which helps us to change the values of numeric columns in the dataset to use a common scale. 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 with the below formula:

$$Xn = (X - Xminimum) / (Xmaximum - Xminimum)$$

- Xn = Value of Normalization
- Xmaximum = Maximum value of a feature
- Xminimum = Minimum value of a feature

Accuracy Score:

Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictionsour model got right. Formally, accuracy has the following definition:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

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

Where *TP* = True Positives, *TN* = True Negatives, *FP* = False Positives, and *FN* = False Negatives.

Confusion Matrix in Machine Learning:

The confusion matrix is a matrix used to determine the performance of the classification models for a given set of test data. It can only be determined if the true values for test data are known. The matrix itself can be easily understood, but the related terminologies may be confusing. Since it shows the errors in the model performance in the form of a matrix, hence also known as an **error matrix**. Some features of Confusion matrix are given below:

- For the 2 prediction classes of classifiers, the matrix is of 2*2 table, for 3 classes, it is 3*3 table, andso on.
- The matrix is divided into two dimensions, that are **predicted values** and **actual values** along with thetotal number of predictions.

Predicted values are those values, which are predicted by the model, and actual values are the true values for the given observations

- It looks like the below table:

n = total predictions	Actual: No	Actual: Yes
Predicted: No	True Negative	False Positive
Predicted: Yes	False Negative	True Positive

Output:

(Execute the program and attach the printout here)

Conclusion: We have studied the concept of normalization of data, accuracy score and confusion matrix. Also implemented and calculated the accuracy score.

Viva Questions:

1. What is True Positive Rate(TPR) and False Positive Rate(FPR).
2. What is Normalization technique?
3. Give the formula for accuracy.
4. What do you mean by artificial neural network?
5. Define Convolutional neural network.
6. Explain Perceptron.

Date:	
Marks obtained:	
Sign of course coordinator:	
Name of course Coordinator :	

Group-B ASSIGNMENT NO:4

Aim: Implement K-Nearest Neighbors algorithm on diabetes.csv dataset. Compute confusion matrix, accuracy, error rate, precision and recall on the given dataset.
Dataset link: https://www.kaggle.com/datasets/abdallamahgoub/diabetes

Objective:

Student will learn:

1] The basic concept and implementation logic of K-Nearest Neighbors.
2] The basic concept and implementation logic of accuracy, error rate, precision and recall.

Theory Concepts*:*

K-Nearest Neighbor (KNN) Algorithm:

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suited category by using K-NN algorithm. K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data. It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

How does K-NN work?

The K-NN working can be explained on the basis of the below algorithm:

- **Step-1:** Select the number K of the neighbors
- **Step-2:** Calculate the Euclidean distance of **K number of neighbors**
- **Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.
- **Step-4:** Among these k neighbors, count the number of the data points in each category.
- **Step-5:** Assign the new data points to that category for which the number of the neighbor ismaximum.
- **Step-6:** Our model is ready.

Confusion Matrix in Machine Learning:

The confusion matrix is a matrix used to determine the performance of the classification models for a givenset of test data. It can only be determined if the true values for test data are known. The matrix itself can be easily understood, but the related terminologies may be confusing. Since it shows the errors in the model performance in the form of a matrix, hence also known as an error matrix. Some features of Confusion matrixare given below:

- For the 2 prediction classes of classifiers, the matrix is of 2*2 table, for 3 classes, it is 3*3 table, andso on.
- The matrix is divided into two dimensions, that are **predicted values** and **actual values** along with thetotal number of predictions.
- Predicted values are those values, which are predicted by the model, and actual values are the truevalues for the given observations.

o It looks like the below table:

n = total predictions	Actual: No	Actual: Yes
Predicted: No	True Negative	False Positive
Predicted: Yes	False Negative	True Positive

The above table has the following cases:

- **True Negative (TN):** Model has given prediction No, and the real or actual value was also No.
- **True Positive (TP):** The model has predicted yes, and the actual value was also true.
- **False Negative (FN):** The model has predicted no, but the actual value was Yes, it is also calledas **Type-II error**.
- **False Positive (FP):** The model has predicted Yes, but the actual value was No. It is also calleda **Type-I error.**

Accuracy - Accuracy is the most intuitive performance measure and it is simply a ratio of correctly predicted observation to the total observations. One may think that, if we have high accuracy then our model is best. Yes, accuracy is a great measure but only when you have symmetric datasets where values of false positive and false negatives are almost same. Therefore, you have to look at other parameters to evaluate theperformance of your model. For our model, we have got 0.803 which means our model is approx. 80% accurate.

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

Precision - Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. The question that this metric answer is of all passengers that labeled as survived, how many actually survived? High precision relates to the low false positive rate. We have got 0.788 precision which is pretty good.

Precision = TP/TP+FP

Recall (Sensitivity) - Recall is the ratio of correctly predicted positive observations to the all observations in actual class - yes. The question recall answers is: Of all the passengers that truly survived, how many did we label? We have got recall of 0.631 which is good for this model as it‘s above 0.5.

Recall = TP/TP+FN

Error Rate - what percentage of our prediction are wrong.

Output:
(Execute the program and attach the printout here)

Conclusion: We have studied the K-Nearest Neighbors algorithm. Also implemented and evaluated the models using accuracy, error rate, precision, and recall.

Date:	
Marks obtained:	
Sign of course coordinator:	
Name of course Coordinator :	

GROUP-B ASSIGNMENT NO. – 5

Aim: Implement K-Means clustering/ hierarchical clustering on sales_data_sample.csv dataset. Determine thenumber of clusters using the elbow method.
Dataset link : https://www.kaggle.com/datasets/kyanyoga/sample-sales-data

Objective:

Student will learn:

1] The basic concept and implementation logic of K-Means clustering/hierarchical clustering.2] The basic concept of elbow method used to determine the number of clusters.

Theory Concepts:

K-Means Clustering Algorithm:

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into differentclusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.
It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

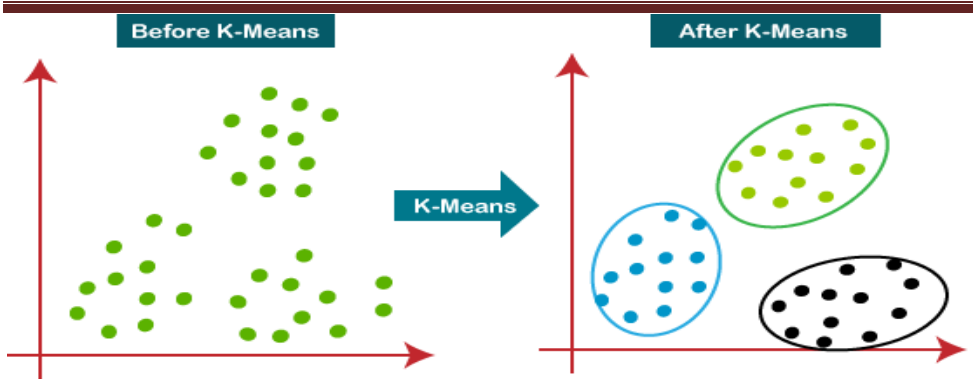
It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.
It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.
The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.

The k-means clustering algorithm mainly performs two tasks:

- Determines the best value for K center points or centroids by an iterative process.
- Assigns each data point to its closest k-center. Those data points which are near to the particular k- center, create a cluster.

Hence each cluster has datapoints with some commonalities, and it is away from other clusters.

The below diagram explains the working of the K-means Clustering Algorithm:



How does the K-Means Algorithm Work?

The working of the K-Means algorithm is explained in the below steps:

Step-1: Select the number K to decide the number of clusters.

Step-2: Select random K points or centroids. (It can be other from the input dataset).

Step-3: Assign each data point to their closest centroid, which will form the predefined K clusters.

Step-4: Calculate the variance and place a new centroid of each cluster.

Step-5: Repeat the third step, which means reassign each datapoint to the new closest centroid of each cluster.

Step-6: If any reassignment occurs, then go to step-4 else go to FINISH.

Step-7: The model is ready.

Elbow Method:

The Elbow method is one of the most popular ways to find the optimal number of clusters. This method uses the concept of WCSS value. **WCSS** stands for **Within Cluster Sum of Squares**, which defines the total variations within a cluster. The formula to calculate the value of WCSS (for 3 clusters) is given below:

$$\mathbf{WCSS= \sum_{Pi\ in\ Cluster1} distance(P_i\ C_1)^2 + \sum_{Pi\ in\ Cluster2} distance(P_i\ C_2)^2 + \sum_{Pi\ in\ CLuster3} distance(P_i\ C_3)^2}$$

In the above formula of WCSS,

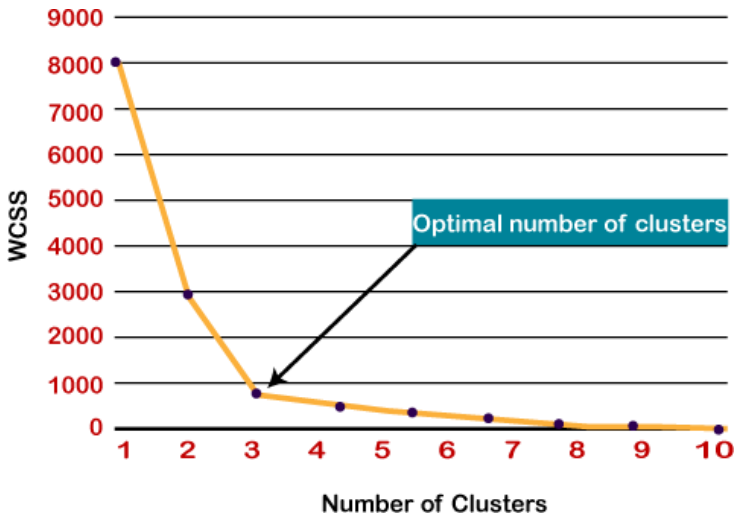
$\sum_{Pi\ in\ Cluster1} distance(P_i\ C_1)^2$: It is the sum of the square of the distances between each data point and itscentroid within a cluster1 and the same for the other two terms.

To measure the distance between data points and centroid, we can use any method such as Euclidean distanceor Manhattan distance.

To find the optimal value of clusters, the elbow method follows the below steps:

- It executes the K-means clustering on a given dataset for different K values (ranges from 1-10).
- For each value of K, calculates the WCSS value.
- Plots a curve between calculated WCSS values and the number of clusters K.
- The sharp point of bend or a point of the plot looks like an arm, then that point is considered as thebest value of K.

Since the graph shows the sharp bend, which looks like an elbow, hence it is known as the elbow method. Thegraph for the elbow method looks like the below image:



Output:
(Execute the program and attach the printout here)

Conclusion: We have studied the K-Means clustering algorithm and elbow method to find the optimalnumber of clusters. Also implemented the K-Means clustering algorithm using python language.

Viva Questions:

7. What is unsupervised learning?
8. Why do we use clustering? Give an example.
9. Explain three applications of clustering.
10. What are some of the drawbacks of k means algorithm.
11. Give the comparison k means and k mediods.
12. Define K means clustering algorithm.
13. Explain the steps involved in k means clustering.
14. Explain Elbow method.
15. What do you mean by hierarchical clustering?

Date:	
Marks obtained:	
Sign of course coordinator:	
Name of course Coordinator :	