Academic Year: 2024-25

LABORATORY MANUAL

Name of the Student:		
Class: BE	**Division: B**	**Roll No.:**
Subject: ML **(2019 Course) [410250]**		**Exam Seat No.:**
Department of Artificial Intelligence and Data Science		

Program Outcomes (PO's):

POs are statements that describe what students are expected to know and be able to do upon graduating from the program. These relate to the skills, knowledge, analytical ability attitude and behavior that students acquire through the program.

- **PO1: Engineering Knowledge:**
  Graduates will be able to apply the Knowledge of the mathematics, science and engineering fundamentals for the solution of engineering problems related to IT.
- **PO2: Problem Analysis:**
  Graduates will be able to carry out identification and formulation of the problem statement by requirement engineering and literature survey.
- **PO3: Design/Development of Solutions:**
  Graduates will be able to design a system, its components and/or processes to meet the required needs with consideration for public safety and social considerations.
- **PO4: Conduct Investigations of Complex Problems:**
  Graduates will be able to investigate the problems, categorize the problem according to their complexity using modern computational concepts and tools.
- **PO5: Modern Tool Usage:**
  Graduates will be able to use the techniques, skills, modern IT engineering tools necessary for engineering practice.
- **PO6: The Engineer and Society:**
  Graduates will be able to apply reasoning and knowledge to assess global and societal issues
- **PO7: Environment and Sustainability:**
  Graduates will be able to recognize the implications of engineering IT solution with respect to society and environment.
- **PO8: Ethics:**

Graduates will be able to understand the professional and ethical responsibility.

- **PO9: Individual and Team Work:**
  Graduates will be able to function effectively as an individual member, team member or leader in multi -disciplinary teams.
- **PO10: Communication:**
  Graduates will be able to communicate effectively and make effective documentations and presentations.
- **PO11: Project Management and Finance:**
  Graduates will be able to apply and demonstrateengineering and management principles in project management as a member or leader.
- **PO12: Life-long Learning:**
  Graduates will be able to recognize the need for continuous learning and to engage in life-long learning.

Course Objectives and Course Outcomes (COs)

Course Objectives:

- Apply regression, classification and clustering algorithms for creation of ML models
- Introduce and integrate models in the form of advanced ensembles.
- Conceptualized representation of Data objects.
- Create associations between different data objects, and the rules.
- Organized data description, data semantics, and consistency constraints of data

Course Outcomes:

On completion of the course, students will be able to–

CO1: Implement regression, classification and clustering models

CO2: Integrate multiple machine learning algorithms in the form of ensemble learning.

CO3: Apply reinforcement learning and its algorithms for real world applications.

CO4: Analyze the characteristics, requirements of data and select an appropriate data model.

CO5: Apply data analysis and visualization techniques in the field of exploratory data science

CO6: Evaluate time series data.

Dr. D. Y. Patil Educational Federation's
Dr. D. Y. PATIL COLLEGE OF ENGINEERING & INNOVATION
Department of Artificial Intelligence and Data Science
Academic Year 2024-25

CERTIFICATE

This is to certify that Mr. /Ms. _____

of Class BE - AI-DS, Roll No.____Examination Seat No. _____

has completed all the practical work in the Computer Laboratory - I [417525] satisfactorily, as prescribed by Savitribai Phule Pune University, Pune in the academic year 2024-25 (Term-I).

Place:

Date:

Course In-charge	**HOD**	**Principal**
Department of AI-DS	**Department of AI-DS**	**DYPCOEI, Varale**

INDEX

Department of Artificial Intelligence and Data Science **Class:** B.E.

Sr. No.	Name of the Experiment	Date of Conduction	Date of Checking	Page No.	Sign	Remark
1	**Assignment on PCA:** To use PCA Algorithm for dimensionality reduction. You have a dataset that includes measurements for different variables on wine (alcohol, ash, magnesium, and so on). Apply PCA algorithm & transform this data so that most variations in the measurements of the variables are captured by a small number of principal components so that it is easier to distinguish between red and white wine by inspecting these principal components.					
2	**Assignment on Predict the Price of the Uber Ride:** 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 ridge, Lasso regression models. 5. Evaluate the models and compare their respective scores like R2, RMSE, etc.					

3	**Assignment on SVM:** Implementation of Support Vector Machines (SVM) for classifying images of handwritten digits into their respective numerical classes (0 to 9).					
4	**Assignment on K-Means Clustering:** Implement K-Means clustering on Iris.csv dataset. Determine the number of clusters using the elbow method					
5	**Assignment on Random Forest Classifier:** Implement Random Forest Classifier model to predict the safety of the car.					
6	**Assignment on K-Means Clustering:** Build a Tic-Tac-Toe game using reinforcement learning in Python by using following tasks. a. Setting up the environment b. defining the Tic-Tac-Toe game c. Building the reinforcement learning model d. Training the model e. Testing the model					

Name & Signature of Course In-charge

Experiment No: 1

Feature Transformation PCA Algorithm

Name of the Student: _____

Class: BE **Batch: B1**

Date:

Mark: /10

Signature of the Course In-charge: _____

Signature of the HOD: _____

EXPERIMENT NO.1

Practical Title: Study PCA dimensionality reduction Technique

Aim: Apply PCA Algorithm on Wine Dataset and Distinguish between Red & White Wine.

Objective:

- To learn dimensionality reduction technique PCA and implement in Python.

Software Requirements:

- Anaconda with Python 3.7

Hardware Requirement:

- PIV, 2GB RAM, 500 GB HDD.

Learning Objectives:

To Learn PCA technique on given dataset.

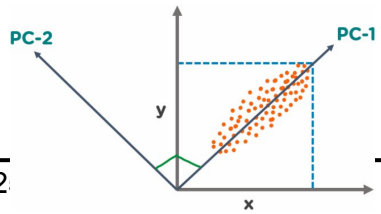
Outcome:

After completion of this assignment students are able to understand how is dimensionality reduction technique PCA work and how it will transform data from higher dimensions to lower dimensions and visualize using Matplotlib.

Theory:

Principal Component Analysis

The Principal Component Analysis is a popular unsupervised learning technique for reducing the dimensionality of data. It increases interpretability yet, at the same time, it minimizes information loss. It helps to find the most significant features in a dataset and makes the data easy for plotting in 2D and 3D. PCA helps in finding a sequence of linear combinations of variables.



In the above figure, we have several points plotted on a 2-D plane. There are two principal components. PC1 is the primary principal component that explains the maximum variance in the data. PC2 is another principal component that is orthogonal to PC1.

What is PCA

The Principal Components are a straight line that captures most of the variance of the data. They have a direction and magnitude. Principal components are orthogonal projections (perpendicular) of data onto lower- dimensional space. Now that you have understood the basics of PCA, let's look at the next topic on PCA in Machine Learning.

Dimensionality

The term "dimensionality" describes the quantity of features or variables used in the research. It can be difficult to visualize and interpret the relationships between variables when dealing with high-dimensional data, such as datasets with numerous variables. While reducing the number of variables in the dataset, dimensionality reduction methods like PCA are used to preserve the most crucial data. The original variables are converted into a new set of variables called principal components, which are linear combinations of the original variables, by PCA in order to accomplish this. The dataset's reduced dimensionality depends on how many principal components are used in the study. The objective of PCA is to select fewer principal components that account for the data's most important variation. PCA can help to streamline data analysis, enhance visualization, and make it simpler to spot trends and relationships between factors by reducing the dimensionality of the dataset.

The mathematical representation of dimensionality reduction in the context of PCA is as follows:

Given a dataset with n observations and p variables represented by the n x p data matrix X, the goal of PCA is to transform the original variables into a new set of k variables called principal components that capture the most significant variation in the data. The principal components are defined as linear combinations of the original variables given by:

$$PC\_1 = a\_11 * x\_1 + a\_12 * x\_2 + ... + a\_1p * x\_p$$

$$PC\_2 = a\_21 * x\_1 + a\_22 * x\_2 + ... + a\_2p * x\_p$$

...

$$PC\_k = a\_k1 * x\_1 + a\_k2 * x\_2 + ... + a\_kp * x\_p$$

where a_ij is the loading or weight of variable x_j on principal component PC_i, and x_j is the jth variable in the data matrix X. The principal components are ordered such that the first component PC_1 captures the most significant variation in the data, the second component PC_2 captures the second most significant variation, and so on. The number of principal components used in the analysis, k, determines the reduced dimensionality of the dataset.

Correlation

A statistical measure known as correlation expresses the direction and strength of the linear connection between two variables. The covariance matrix, a square matrix that displays the pairwise correlations between all pairs of variables in the dataset, is calculated in the setting of PCA using correlation. The covariance matrix's diagonal elements stand for each variable's variance, while the off-diagonal elements indicate the covariances between different pairs of variables. The strength and direction of the linear connection between two variables can be determined using the correlation coefficient, a standardized measure of correlation with a range of -1 to 1.

A correlation coefficient of 0 denotes no linear connection between the two variables, while correlation coefficients of 1 and -1 denote the perfect positive and negative correlations, respectively. The principal components in PCA are linear combinations of the initial variables that maximize the variance explained by the data. Principal components are calculated using the correlation matrix.
In the framework of PCA, correlation is mathematically represented as follows:

The correlation matrix C is a nxn symmetric matrix with the following components given a dataset with n variables (x_1, x_2,..., x_n):

$$C_{ij} = (sd(x_i) * sd(x_j)) / cov(x_i, x_j)$$

where sd(x i) is the standard deviation of variable x_i and sd(x_j) is the standard deviation of variable xj, and cov(x i, x j) is the correlation between variables x_i and x_j.

The correlation matrix C can also be written as follows in matrix notation:

C = X^T X / (n-1) (n-1)

Orthogonal

The term "orthogonality" alludes to the principal components' construction as being orthogonal to one another in the context of the PCA algorithm. This indicates that there is no redundant information among the main components and that they are not correlated with one another.

Orthogonality in PCA is mathematically expressed as follows: each principal component is built to maximize the variance explained by it while adhering to the requirement that it be orthogonal to all other principal components. The principal components are computed as linear combinations of the original variables. Thus, each principal component is guaranteed to capture a unique and non-redundant part of the variation in the data.

The orthogonality constraint is expressed as:

$$a\_i_1 * a\_j_1 + a\_i_2 * a\_j_2 + ... + a\_i_p * a\_j_p = 0$$

for all i and j such that i ≠ j. This means that the dot product between any two loading vectors for different principal components is zero, indicating that the principal components are orthogonal to each other.

Eigen Vectors

The main components of the data are calculated using the eigenvectors. The ways in which the data vary most are represented by the eigenvectors of the data's covariance matrix. The new coordinate system in which the data is represented is then defined using these coordinates.

The covariance matrix C in mathematics is represented by the letters v_1, v_2,..., v_p, and the associated eigenvalues are represented by _1, _2,..., _p. The eigenvectors are calculated in such a way that the equation shown below holds:

$$C\ v\_i = \lambda\_i\ v\_i$$

This means that the eigenvector v_i produces the associated eigenvalue λ_i as a scalar multiple of itself when multiplied by the covariance matrix C.

Covariance Matrix

The covariance matrix is crucial to the PCA algorithm's computation of the data's main components. The pairwise covariances between the factors in the data are measured by the covariance matrix, which is a p x p matrix.

The correlation matrix C is defined as follows given a data matrix X of n observations of p variables:

$$C = (1/n) * X^T X$$

where X^T represents X's transposition. The covariances between the variables are represented by the off- diagonal elements of C, whereas the variances of the variables are represented by the diagonal elements of C.

Algorithm:

1. Import the Required Packages
2. Read Given Dataset
3. Import the Principal Component Analysis
4. Define input & output
5. Initialize the model
6. Fit the dataset
7. Draw Scatter Plot

Conclusion:

Thus we learn how to apply PCA algorithm on Wine dataset and visualize the classes by plotting on 2D graph.

Viva Questions:

1. What is Machine Learning?
2. What is PCA Algorithm?
3. What are Different Types of ML?

Coding Efficiency	Viva	Timely Completion	Total	Dated Sign of Course In-charge
5	3	2	10	

Program:

Experiment No: 2

Assignment based on Linear Regression

Name of the Student: _____

Class: BE **Batch: B1**

Date:	Mark: /10

Signature of the Course In-charge: _____

Signature of the HOD: _____

EXPERIMENT NO.2

Practical Title: Assignment based on Linear Regression

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.

Prerequisite:

1. Basic knowledge of Python
2. Concept of preprocessing data
3. Basic knowledge of Data Science and Big Data Analytics.

Software Requirements:

- Anaconda with Python 3.7

Hardware Requirement:

- PIV, 2GB RAM, 500 GB HDD

Learning Objectives:

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.

Outcome:

Theory:

Data Preprocessing:

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

When creating a machine learning 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 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

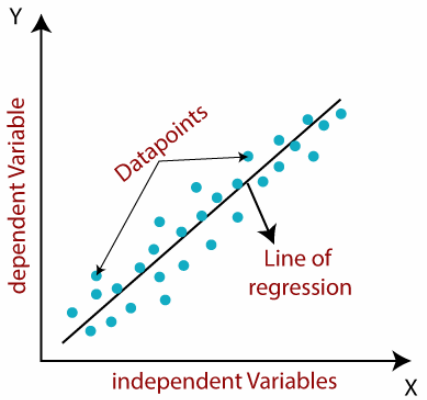
Linear Regression:

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:

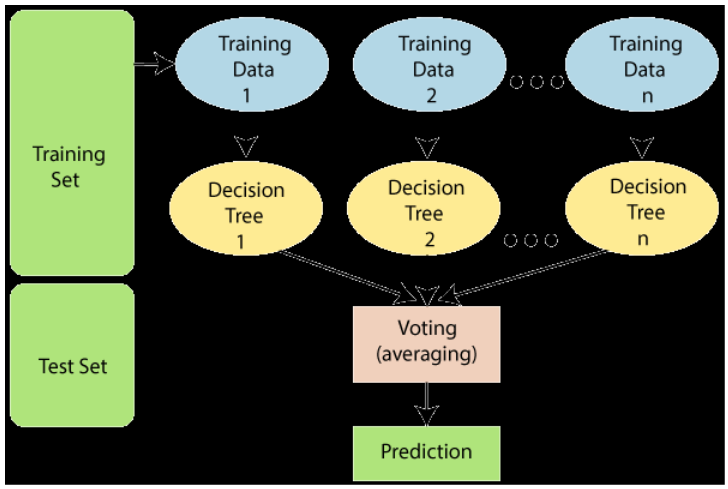


Random Forest Regression Models:

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.

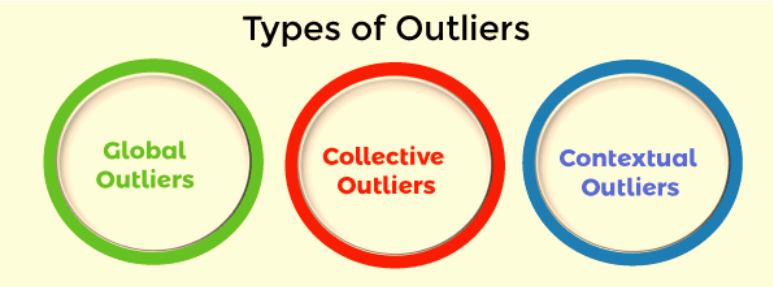
As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. "Instead of relying on one decision tree, the random forest takes the prediction from each tree 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.



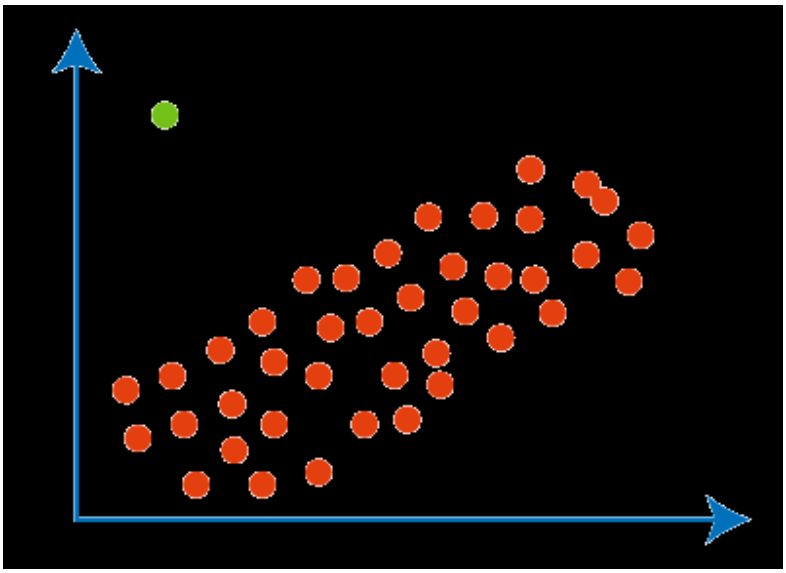
Outliers:

As the name suggests, "outliers" refer to the data points that exist outside of what is to be expected. The major thing about the outliers is what you do with them. If you are going to analyze any task to analyze data sets, you will always have some assumptions based on how this data is generated. If you find some data points that are likely to contain some form of error, then these are definitely outliers, and depending on the context, you want to overcome those errors. The data mining process involves the analysis and prediction of data that the data holds. In 1969, Grubbs introduced the first definition of outliers.



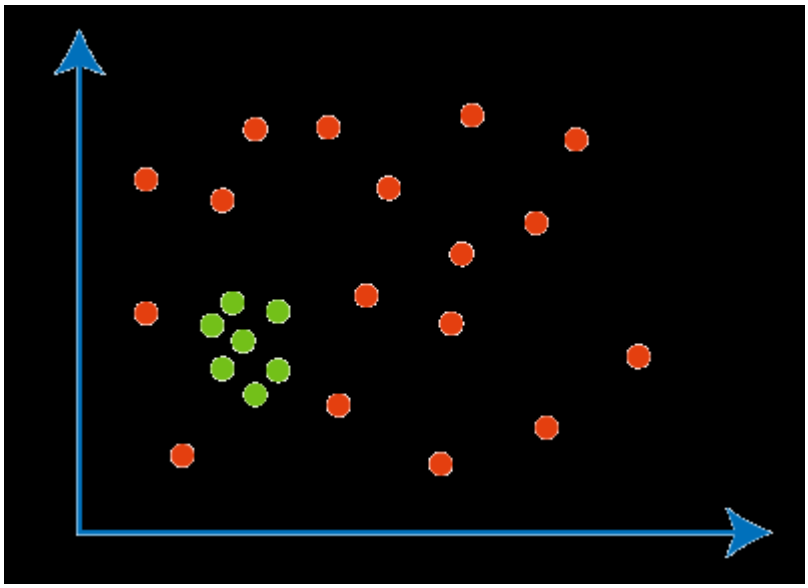
Global Outliers

Global outliers are also called point outliers. Global outliers are taken as the simplest form of outliers. When data points deviate from all the rest of the data points in a given data set, it is known as the global outlier. In most cases, all the outlier detection procedures are targeted to determine the global outliers. The green data point is the global outlier.



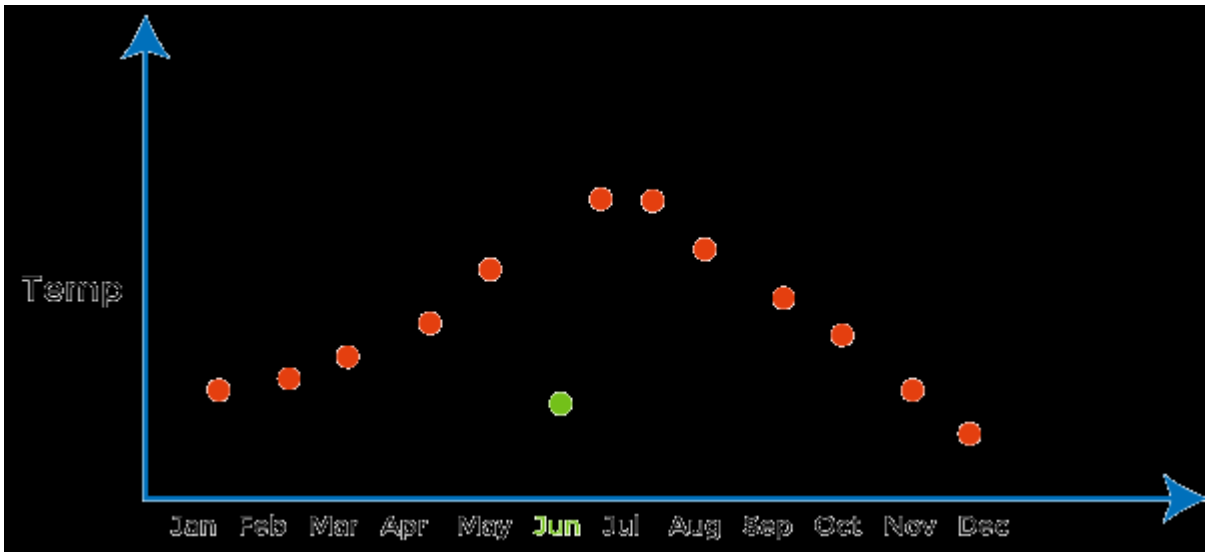
Collective Outliers

In a given set of data, when a group of data points deviates from the rest of the data set is called collective outliers. Here, the particular set of data objects may not be outliers, but when you consider the data objects as a whole, they may behave as outliers. To identify the types of different outliers, you need to go through background information about the relationship between the behavior of outliers shown by different data objects. For example, in an Intrusion Detection System, the DOS package from one system to another is taken as normal behavior. Therefore, if this happens with the various computer simultaneously, it is considered abnormal behavior, and as a whole, they are called collective outliers. The green data points as a whole represents the collective outlier.



Contextual Outliers

As the name suggests, "Contextual" means this outlier introduced within a context. For example, in the speech recognition technique, the single background noise. Contextual outliers are also known as Conditional outliers. These types of outliers happen if a data object deviates from the other data points because of any specific condition in a given data set. As we know, there are two types of attributes of objects of data: contextual attributes and behavioral attributes. Contextual outlier analysis enables the users to examine outliers in different contexts and conditions, which can be useful in various applications. For example, A temperature reading of 45 degrees Celsius may behave as an outlier in a rainy season. Still, it will behave like a normal data point in the context of a summer season. In the given diagram, a green dot representing the low-temperature value in June is a contextual outlier since the same value in December is not an outlier.



Conclusion:

In this way we learn that to how to create Decision Tree based on given decision, Find the Root Node of the tree using Decision tree Classifier.

Viva Questions:

Program:-

Experiment No: 3

Assignment based on k-NN Classification

Name of the Student: _____

Class: _____ **Batch:** _____

Date of Performance: _____

Signature of the Staff: _____

Experiment No: 4

Assignment on K-Means Clustering:

Name of the Student: _____

Class: **BE** **Batch:** _____

Date of Performance: _____

Signature of the Staff: _____