

Experiment 2 - (college, auto datasets)  
Aim:To study the basics of Python.

Exercise 1:

Abstract: There is a College data set, which can be found in the file College.csv. It contains a number of variables for 777 different universities and colleges in the US. The are 18 columns in the dataset.

(a) Use the pd.read\_csv() function to read the data into Python. Call the loaded data college. Make sure that you have the directory set to the correct location for the data. (b) Use the describe() method to produce a numerical summary of the variables in the data set. (c) Use the pd.plotting.scatter\_matrix() function to produce a scatterplot matrix of the first columns [Top10perc, Apps, Enroll]. (d) Use the boxplot() method of college to produce side-by-side boxplots of Outstate versus Private. (e) Create a new qualitative variable, called Elite,bybinning the Top10perc variable into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.

Exercise 2:

Abstract: This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data. (a) Which of the predictors are quantitative, and which are qualitative? (b) What is the range of each quantitative predictor? You can answer this using the min() and max() methods in numpy. (c) What is the mean and standard deviation of each quantitative predictor? (d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains? (e) Using the full data set, investigate the predictors graphically, using scatterplots . Create some plots highlighting the relationships among the predictors.Comment on your findings.

Experiment 3 - (Ice cream, HRP, 50 startups)

Aim: To implement Linear Regression

You own an ice cream business and you would like to create a model that could predict the daily revenue in dollars based on the outside air temperature (degC). You decide that a Linear Regression model might be a good candidate to solve this problem. Independent variable X: Outside Air Temperature Dependant Variable Y: Overall daily revenue generated in dollars

Exercise 1:

You work in the real estate sector and you would like to create a model that could predict the selling price of a property based on its area in square feet. You decide that a Linear Regression model might be a good candidate to solve this problem. Data set: • Independent variable X: Area of the property (in square feet) • Dependent variable Y: Selling price of the property (in dollars)

Exercise 2:

Predict the total fare amount of a Chicago taxi trip based on trip distance and duration using a Linear Regression model

Experiment 4 - (admittance, bank, smarket datasets)

Aim: To implement Logistic Regression using Python.

Admittance

Exercise 1:

You work in the loan approval department of a financial institution and would like to create a model that can help predict whether a loan application will be approved based on one of the features of the applicant. Feature selection should be done based on the results of the correlation matrix. You decide that a Logistic Regression model might be a good candidate to solve this binary classification problem.

Exercise 2:

Examine the Smarket data, which is part of the ISLP library. This data set consists of percentage returns for the S&P 500 stock index over 1,250 days, from the beginning of 2001 until the end of 2005. For each date, we have recorded the percentage returns for each of the five previous trading days, Lag1 through Lag5. We have also recorded Volume (the number of shares traded on the previous day, in billions), Today (the percentage return on the date in question) and Direction (whether the market was Up or Down on this date). Use the full data set with Direction as the response and the five lag variables plus Volume as predictors.Select which predictors can be used and perform Logistic Regression on it.

Experiment 5 (iris dataset, breast cancer, wine dataset)

Aim: To implement Decision Tree Classifier

Exercise 1: Breast cancer diagnosis is critical for timely treatment and patient survival. Using the Breast Cancer Wisconsin dataset, the goal is to develop a Decision Tree Classifier that accurately predicts whether a tumor is benign or malignant based on cell nuclei features, and evaluate its performance for reliable medical decision support.

Exercise 2: How can we utilize a Decision Tree Classifier to accurately predict the type of wine from three different cultivars based on their chemical properties, and how can we visualize the resulting tree structure to interpret the decision rules behind the classification?

Experiment 6 (iris dataset, breast cancer)

Aim: To implement Linear Support Vector Machine

Exercise 1: Apply a Linear Support Vector Machine (SVM) to the Breast Cancer dataset to classify tumors as malignant or benign based on features such as cell radius, texture, and smoothness. The problem involves training a linear SVM model to identify the optimal hyperplane that separates the two classes while maximizing the margin between them. The objective is to achieve high classification accuracy and evaluate the model's performance in distinguishing between benign and malignant tumors.

Exercise 2: Generate and side-by-side bar chart comparing accuracy, precision, recall, and F1-score between Linear vs RBF SVM and also do cross-validation comparison using Breast cancer dataset.

Experiment 7 (Palmer Penguins, iris, wine)

Aim: To implement an Ensemble model using Random Forest.

Exercise 1: Implement an ensemble model using the Random Forest algorithm on the Iris dataset. Perform the following tasks: 1. Load the Iris dataset and identify the input features and target label. 2. Split the dataset into training (80%) and testing (20%) subsets. 3. Train a Random Forest classifier and explain how ensemble learning improves over a single decision tree. 4. Evaluate the model’s performance using training and testing accuracy. 5. Visualize one of the decision trees from the ensemble and interpret the feature splits. 6. Discuss the significance of feature importance in predicting the species of Iris flowers.

Exercise 2: Develop and evaluate an ensemble learning model using the Random Forest algorithm for a real-world dataset involving multi-class classification. Your task involves selecting a dataset with multiple features (both numerical and categorical) and at least three target classes (e.g., the Wine Quality dataset). Perform necessary preprocessing such as handling missing values, encoding categorical features, and normalization if required. Train a Random Forest model with tuned hyperparameters (number of trees, max depth, min samples per split, etc.).Justify why ensemble learning (bagging + random feature selection) is preferred over a single decision tree in this scenario.

Experiment 8 (housing,iris, mall customers)

Aim: To implement k-means algorithm.

Exercise 1:

You are provided with the famous Iris dataset, which contains measurements of three different species of iris flowers. Each sample has the following attributes: ● Sepal Length (cm) ● Sepal Width (cm) ● Petal Length (cm) ● Petal Width (cm) 1. Apply K-Means clustering to group the flowers into clusters without using the species label.2. Use K = 3 (since there are 3 flower species) and fit the model. 3. Compare the clustering results with the actual species labels to check the accuracy. 4. Visualize the clusters in 2D using Petal Length vs. Petal Width. 5. Discuss whether K-Means was effective in separating the species

Exercise 2:

You are given a dataset containing the annual income (in ₹) and spending score (1–100) of customers in a shopping mall. Apply K-Means clustering on this dataset to group the customers into different clusters based on their purchasing behavior. Use the Elbow method to determine the optimal number of clusters. Visualize the clusters using a 2D scatter plot where: X-axis represents Annual Income Y-axis represents Spending Score Different clusters are shown in different colors. Interpret the characteristics of each cluster (e.g., high-income high-spending, low-income low-spending, etc.). Suggest how the shopping mall can use these insights for customer segmentation and targeted marketing strategies. Sample Dataset Columns: CustomerID Age Annual Income (₹) Spending Score (1–100)

Experiment 9 (iris, make moons, mall)  
Aim: To implement DBSCAN (Density-Based Spatial Clustering of Applications with Noise) clustering algorithm

Exercise 1: Implement the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm to identify and visualize clusters in a non-linear dataset.How can the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm be implemented to identify clusters and detect noise in a non-linear dataset such as the Make Moons dataset, and how effectively does it handle irregularly shaped clusters compared to traditional clustering methods like K-Means?

Exercise 2: How can the DBSCAN algorithm be applied to a large, real-world dataset such as the Mall Customers dataset to discover meaningful clusters of customers with similar income and spending patterns, while detecting outlier behaviors that may represent exceptional or unusual spending profiles?

Experiment 10 (iris, wine, breast cancer)

Aim: To implement PCA

Exercise 1:

You are provided with the Wine Dataset (UCI Machine Learning Repository / available in

sklearn.datasets.load\_wine).

The dataset contains 178 wine samples from three different cultivars (classes), described by 13

chemical features such as alcohol, flavanoids, color intensity, magnesium, etc.

Tasks:

1. Load the Wine dataset and display its shape and feature names.

2. Standardize the dataset so all features have zero mean and unit variance.

3. Apply PCA:

o (a) Compute the explained variance ratio for all components.

o (b) Determine how many components are required to explain at least 95% of the

variance.

4. Reduce the data to 2 principal components and plot a scatter plot with different colors for the

three wine classes.

5. Interpret the results:

o Which components explain the most variance?

o Does the 2D PCA plot show separation among the wine classes?

Exercise 2: You are provided with the Breast Cancer Wisconsin Dataset (available in sklearn.datasets.load\_breast\_cancer). The dataset contains 569 samples classified into malignant and benign tumors, described by 30 numerical features (e.g., mean radius, texture, smoothness, concavity). Tasks: 1. Load the Breast Cancer dataset and check its dimensions (number of samples × features). 2. Standardize the data before applying PCA. 3. Apply PCA to: o (a) Find the variance explained by each component. o (b) Plot the cumulative explained variance curve and determine the number of components needed to explain at least 90% of the variance. 4. Reduce the dataset to 2 principal components and create a scatter plot, coloring the points by their class (malignant vs. benign). 5. Analyze and answer:o How well do the first two components separate the two classes? o What is the minimum number of components you need to retain most of the information in the dataset?