Sri Sivasubramaniya Nadar College of Engineering, Chennai
(An autonomous Institution affiliated to Anna University)

Degree & Branch	B.E. Computer Science & Engineering	Semester	V
Subject Code & Name	ICS1512 & Machine Learning Algorithms Laboratory		
Academic year	2025-2026 (Odd)	Batch:2023-2028	**Due date:**

Experiment 1 : Working with Python packages – Numpy, Scipy, Scikit-learn, Mat- plotlib

Name: Vigneshwaran S
Reg no:3122237001059

1 Aim:

To understand and explore essential Python libraries used in machine learning, such as pandas, numpy, matplotlib, seaborn, and scikit-learn. This includes learning how to handle, visualize, and preprocess datasets effectively for model building.

2 Libraries used

- Numpy
- Pandas
- Matplotlib
- Scikit-learn
- Seaborn

3 Mathetical/theoritical description of the algorithm/objective performed:

3.1 Handling Missing Values

Missing values can reduce the accuracy of machine learning models by:

- Changing the overall data statistics
- Causing errors during training
- Producing biased predictions

So, it is important to find and handle them properly before building models.
Ways to handle missing values:

- Replace them with the mean, median, or mode of the column using the fillna() method in

pandas.

- If a column has too many missing values and is not very useful for prediction, remove it to make the dataset simpler and reduce noise.

3.2 Label encoding:

Machine learning models need numeric inputs, so categorical data (like “Yes/No” or “Graduate/Not Graduate”) must be converted to numbers.

- For binary categories, we can map values directly (e.g., “Yes” → 1, “No” → 0).
- For features with more than two categories, one-hot encoding is used. It creates separate binary columns for each category, avoiding any false order between them.

3.3 Plotting:

Data visualization helps understand patterns, relationships, and outliers in a dataset.

- **Heatmap:** Shows correlations between numeric features using colors. Darker shades mean stronger relationships, helping find related or redundant features.
- **Histogram:** Shows how a numeric feature’s values are spread, helping detect skewness, multiple peaks, outliers, or missing value gaps.
- **Box plot:** Displays data spread using quartiles, showing the median and outliers, making it easy to spot extreme values and check data symmetry.

3.4 Removal of Outliers:

After detection of outliers using boxplot, we can remove the outliers by using any one of the below mentoined ways:

- This method involves **dropping the rows that contain outlier** values beyond a certain threshold (usually outside 1.5×IQR).It helps in reducing noise from the dataset, especially when outliers are errors or irrelevant. However, **excessive removal may lead to loss of valuable information** if not done carefully.

- Instead of deleting, outlier values can be **replaced with the column's mean or median to preserve data size**. Median is preferred when data is skewed, as it minimizes distortion. This method smooths the dataset without losing records, maintaining balance in feature distributions.

3.5 Standardization:

- Standardization is a feature scaling technique that **transforms data to have a mean of 0 and a standard deviation of 1**. It is especially useful when features have different units or scales, ensuring all variables contribute equally to the model. Many machine learning algorithms, like logistic regression and KNN, perform better when data is standardized.
- The formula used for standardization is:

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

- where x is the original value, μ is the mean, and σ is the standard deviation. This process centers the data around zero and makes it easier for models to converge efficiently.

The preprocessing steps involved handling missing values, encoding categorical variables, visualiz- ing data using heatmaps, histograms, and boxplots, and addressing outliers. Additionally, feature standardization was applied to bring all variables to a common scale, ensuring better model per- formance and stability.

Dataset	Type of ML Task	Suitable ML Algorithm
Iris Dataset	Multi-class Classifica - tion	KNN, SVM
Loan Amount Prediction	Regression	Linear Regression
Predicting Diabetes	Binary Classification	SVM, XGBoost
Classification of Email Spam	Binary Classification	Logistic Regression, SVM
Handwritten Character Recognition	Multi-class Classifica - tion	CNN, SVM

Table 1: ML Task and Suitable Algorithms for Different Datasets

4 Results and Discussions:

4.1 Handwritten Digit Recognition (MNIST)

This task involves identifying digits (0–9) from grayscale image data. Due to its image-based nature, Convolutional Neural Networks (CNNs) are the most suitable choice. SVM can also be considered when using pixel-intensity features after preprocessing.

4.2 Iris Flower Classification

The Iris dataset is a classic example of multi-class classification with three flower types. K-Nearest Neighbors (KNN) and Support Vector Machines (SVM) are well-suited for this problem. These algorithms handle class boundaries effectively with proper scaling.

4.3 Diabetes Prediction

This is a binary classification task to detect the likelihood of diabetes based on medical attributes. Support Vector Machines (SVM) and Logistic Regression are the most effective for such structured data. Normalization and feature selection improve model performance.

4.4 Email Spam Classification

The goal is to categorize emails as spam or not based on word frequency features. Logistic Re- gression and SVM work best for this high-dimensional, text-based dataset. TF-IDF vectorization is often used to convert text into meaningful numerical inputs.

5 Learning Practices:

- **Handling Incomplete Data:** Missing values were addressed by replacing them with sta- tistical measures such as mean or median, or by removing irrelevant features to maintain dataset integrity.
- **Encoding Categorical Variables:** Categorical attributes were converted into numerical form using techniques like label encoding and one-hot encoding to ensure compatibility with machine learning algorithms.
- **Understanding Feature Impact:** Statistical correlation and visualization methods like heatmaps and boxplots were used to evaluate feature significance and detect redundancy or outliers.
- **Model Selection Awareness:** Based on the nature of each problem (classification or re- gression), appropriate machine learning models were selected, considering factors like data type, dimensionality, and target variable.

6.Github Link,