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, Matplotlib

1 Aim:

To explore the various functions and methods available in the Python libraries.

2 Libraries used:

- Numpy
- Pandas
- Matplotlib
- Scikit-Learn
- Seaborn

3 Mathematical/Theoretical Description of the Algorithm/Objectives Performed

The following summarizes the theoretical background and purpose of each technique used:

3.1 Handling Missing Values

Missing data in a dataset can lead to biased model outcomes or training failures. To address this:

- Columns with missing values were either **removed** if deemed non-essential, or
- **Imputed using the mode** for categorical columns, ensuring no distortion in label distributions.

3.2 Feature Importance via Word Frequency Comparison

In the spam email classification task:

- Each email was represented as a **bag-of-words vector**, with each feature representing the frequency of a unique word.
- To identify which words were most indicative of spam, the **relative frequency** of each word in spam vs. non-spam emails was calculated.

3.3 Correlation Analysis Between Features and Target

For datasets with numeric input features (e.g., diabetes and iris datasets):

- **Pearson correlation coefficients** were calculated between each input feature and the target label.
- For categorical targets (e.g., species in the iris dataset), the labels were first **converted into numerical format using Label Encoding**.

3.4 Standardization of Features

Input features in real-world datasets often have **different scales and units**. To address this:

- **Z-score standardization** was applied to numeric features using the formula:

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

  where x is the feature value, μ is the mean, and σ is the standard deviation.
- This transformation ensures that all features have **zero mean and unit variance**.

3.5 Label Encoding

For datasets containing **categorical target variables** (like `species` in the iris dataset):

- Labels were encoded into numeric format using **LabelEncoder**, which assigns a unique integer to each category.
- This step is essential for machine learning algorithms that **require numerical inputs** for training and evaluation.

4 Results and Discussions:

Iris Dataset: Since this is a multi-class classification problem, algorithms like **K-Nearest Neighbors (KNN)** and **Support Vector Machine (SVM)** are well-suited.

Loan Amount Prediction: This task typically involves predicting either the loan approval status (classification) or the exact loan amount (regression). In this case, it was treated as a regression problem, for which **Linear Regression** is a suitable algorithm.

Dataset	**Type of ML Task**	**Suitable ML Algorithm**
Iris Dataset	Multi-class Classification	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 Classification	CNN, SVM

Table 1: ML Task and Suitable Algorithms for Different Datasets

Predicting Diabetes: This binary classification problem uses features like glucose level, BMI, to predict the presence of diabetes. **Support Vector Machine (SVM)** is effective for such structured datasets.

Classification of Email Spam: This involves analyzing word frequencies in emails to determine if they are spam. Algorithms such as **Logistic Regression** and **SVM** are efficient due to their ability to handle high-dimensional sparse data.

Handwritten Character Recognition: Using the MNIST dataset, this task classifies grayscale images of digits (0–9). **Convolutional Neural Networks (CNNs)** are state-of-the-art for image classification, while **SVM** can also perform well with extracted features.

5 Learning Outcomes:

- **Data Cleaning:** Missing values were handled appropriately by either removing the affected records or imputing with meaningful statistics (e.g., mode for categorical features).
- **Text Analysis:** In the spam detection task, the bag-of-words model was used to quantify word importance. Words were ranked based on their frequency difference between spam and non-spam classes, while low-frequency noise was filtered using a threshold.
- **Feature Relevance:** Correlation analysis was performed to identify which features had the strongest relationship with the output label.