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

Name: Darshan Parthasarathy Reg No: 3122237001009

Aim

To explore and demonstrate the functionalities of key Python libraries used in data science and machine learning, focusing on array operations, preprocessing techniques, model preparation workflows, and data visualization.

Libraries Used

- **NumPy:** For numerical computations and array manipulations.
- **Pandas:** For handling structured tabular data and data preprocessing.
- **Matplotlib:** For creating visualizations and graphical analysis.
- **Scikit-Learn:** For machine learning model building and preprocessing utilities.
- **Seaborn:** For advanced statistical data visualization.

Mathematical/Theoretical Description of the Algorithm/Objectives Performed

1. Handling Missing Values

Missing values can lead to errors or biases in model training. To address this:

- Non-critical columns with excessive missing values were dropped.
- Categorical columns were imputed using the **mode** (most frequent value) to preserve class distributions.
- Numerical columns, if necessary, could be filled using mean or median.

2. Feature Importance via Word Frequency Comparison

In a spam email classification scenario:

- Each email was converted into a bag-of-words (BoW) vector.
- Frequencies of words were computed separately for spam and non-spam classes.
- The most impactful words were identified using relative frequency comparisons.
- A frequency threshold was applied to eliminate noisy or rare words.

3. Correlation Analysis Between Features and Target

To understand the strength of relationships between input features and the target label:

- **Pearson correlation coefficients** were used for continuous variables.
- Categorical target labels were **label encoded** before correlation calculations.
- This helped in identifying predictive features to prioritize for modeling.

4. Standardization of Features

To ensure all features are on the same scale:

- **Z-score normalization** was used, given by the formula:

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

  where x is the feature value, μ is the mean, and σ is the standard deviation.

- This helps machine learning models converge faster and perform better.

5. Label Encoding

Categorical data was converted into numeric format using:

- **LabelEncoder** from scikit-learn, assigning a unique integer to each category.
- Essential for algorithms that cannot interpret textual labels.

ML Task and Suitable Algorithms

Dataset	**Type of ML Task**	**Suitable ML Algorithm**
Iris Dataset	Multi-class Classification	K-Nearest Neighbors (KNN), Support Vector Machine (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	Convolutional Neural Networks (CNN), SVM

Results and Discussions

- **Iris Dataset:** Multi-class classification using petal and sepal features was successfully demonstrated using KNN and SVM.
- **Loan Amount Prediction:** Treated as a regression problem with continuous target output, implemented using Linear Regression.
- **Diabetes Prediction:** Features such as glucose, BMI, and insulin levels were used with SVM and XGBoost to predict diabetes.
- **Spam Classification:** BoW representation with Logistic Regression and SVM yielded good performance.
- **Handwritten Digit Recognition:** The MNIST dataset was used, and classification was done using CNN and SVM.

Learning Practices

- Developed proficiency in preprocessing techniques like handling missing values and encoding.
- Gained insights into statistical techniques like correlation analysis and standardization.
- Understood how to extract meaningful features from raw text using bag-of-words.
- Explored the impact of data scaling and normalization on model performance.
- Learned to map real-world datasets to appropriate machine learning tasks and models.
- Practiced using key Python libraries effectively in an integrated machine learning pipeline.