Experiment 1: Working with Python Packages – Numpy, Scipy, Scikit-Learn, Matplotlib

Sreeram GM

July 29, 2025

Sri Sivasubramaniya Nadar College of Engineering, Chennai

(An autonomous institution affiliated to Anna University)

Subject Code & Name: ICS1512 – Machine Learning Algorithms Laboratory

Academic Year: 2025–2026 (Odd)

Batch: 2023–2028

Due Date: 12.07.2025

1 Aim

To familiarize with essential Python libraries and explore their functionalities for tasks such as array operations, data preprocessing, numerical computing, machine learning workflows, and data visualization.

2 Libraries Used

- NumPy
- Pandas
- Matplotlib
- Scikit-Learn
- Seaborn

3 Mathematical / Theoretical Overview

This experiment focuses on key preprocessing and analytical steps required for efficient machine learning. The major components include:

1. **Handling Missing Values:** Missing data can bias models or cause failures during training. Columns with missing values were either discarded (if non-essential) or imputed using the mode for categorical attributes to maintain label integrity.

2. **Feature Importance via Word Frequency:** In spam email classification, emails were represented as bag-of-words vectors. Feature relevance was assessed by comparing word frequency across spam and non-spam classes. Rare words were filtered out to retain significant features.

3. **Correlation Analysis:** For numeric datasets (e.g., diabetes, iris), Pearson correlation coefficients were computed between input features and the target variable. Label encoding was applied to categorical targets.

4. **Feature Standardization:** Features often differ in scale. To normalize, Z-score standardization was applied:
$$z = \frac{x - \mu}{\sigma}$$
ensuring zero mean and unit variance.

5. **Label Encoding:** For categorical outputs, classes were converted to integers using LabelEncoder, making them compatible with algorithms requiring numeric inputs.

4 Dataset and Suitable Algorithms

Table 1: Datasets and Corresponding Algorithms

Dataset	**Task Type**	**Suitable Algorithms**
Iris Dataset	Multi-class Classification	KNN, SVM
Loan Amount Prediction	Regression	Linear Regression
Diabetes Prediction	Binary Classification	SVM
Email Spam Detection	Binary Classification	Logistic Regression, SVM
Digit Recognition	Multi-class Classification	CNN, SVM

Results and Discussions

- **Iris Dataset:** Classified flowers into three species using sepal and petal dimensions. KNN and SVM were suitable choices.

- **Loan Amount Prediction:** Treated as a regression task using Linear Regression.

- **Diabetes Prediction:** Binary classification using SVM for structured numeric data.

- **Email Spam Detection:** Used word frequency features, classified with Logistic Regression and SVM.
- **Digit Recognition:** Implemented CNN for image classification; SVM works well with extracted features.

5 Learning Outcomes

- Applied data cleaning techniques for missing values.
- Performed text analysis using Bag-of-Words for spam detection.
- Conducted feature relevance checks via correlation analysis.
- Standardized numerical features for improved model performance.