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: 29/7/25**

Experiment 1: Working with Python packages - Numpy, Scipy, Scikit-learn, Matplotlib

1 Aim:

To understand and explore essential Python libraries used in machine learning, identify the type of machine learning task for different datasets and explore the complete machine learning workflow from loading to evaluation.

2 Libraries used:

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

3 Mathematical/theoretical description of the algorithm/objective performed:

i) **Loading the Dataset:**

This step involves importing datasets (e.g., CSV files) into the Python environment using libraries such as `pandas`. Mathematically, the dataset is treated as a matrix $X \in R^{m \times n}$, where m is the number of samples and n is the number of features.

ii) **Exploratory Data Analysis (EDA) and Visualization:** EDA helps summarize the data's main characteristics using statistical and visual methods:

- **Histograms:** Show feature distributions.
- **Box plots:** Detect outliers using interquartile range (IQR).

- **Scatter plots:** Examine relationships between numeric features.
- **Heatmaps:** Visualize correlation matrices (Pearson/Spearman).

iii) **Data Preprocessing:**

- **Missing value handling:** Drop or impute missing values using mean, median, or mode.
- **Categorical encoding:** Convert categories to numerical using one-hot or label encoding.
- **Normalization/Standardization:**
  - Min-Max: $x' = \frac{x - \min(x)}{\max(x) - \min(x)}$
  - Z-score: $x' = \frac{x - \mu}{\sigma}$

iv) **Feature Selection:** Select important features to improve model efficiency and performance:

- **SelectKBest:** Selects top k features using scoring functions.
- **Chi-Square Test:** Measures independence between categorical variables: $\chi^2 = \sum \frac{(O-E)^2}{E}$
- **ANOVA (F-test):** Compares means between groups using: $F = \frac{\text{Between-group variance}}{\text{Within-group variance}}$

v) **Data Splitting:** The dataset is divided into:

- Training set (e.g., 70%)
- Validation set (e.g., 15%)
- Test set (e.g., 15%)

This helps ensure model generalization and avoids overfitting.

vi) **Performance Evaluation:** Appropriate metrics are chosen based on the task:

- **Classification:** Accuracy, Precision, Recall, F1-score, ROC-AUC
- **Regression:** MSE, MAE, R^2 Score

Visual tools like confusion matrices, ROC curves, and residual plots are used to interpret results.

Dataset	Type of ML Task	Suitable ML Algorithm
Iris Dataset	Supervised Classification	KNN, SVM
Loan Amount Prediction	Supervised Regression	Linear Regression, XGBoost
Predicting Diabetes	Supervised Classification	Logistic Regression, Random Forest
Classification of Email Spam	Supervised Classification	Naive Bayes, SVM
Handwritten Character Recognition / MNIST	Supervised Classification	CNN, SVM

Table 1: ML Task and Suitable Algorithms for Different Datasets

4 Results and Discussions:

4.1 Iris Dataset

The Iris dataset is a classification dataset used to predict the species of iris flowers based on sepal and petal dimensions; **K-Nearest Neighbors (KNN)** and **Support Vector Machine (SVM)** are employed due to their proven effectiveness on low-dimensional, well-separated datasets.

4.2 Loan Amount Prediction

Loan amount prediction involves estimating the loan amount an individual is eligible for based on financial and demographic attributes; **Linear Regression** is employed as it efficiently models the linear correlation between these input features and the continuous loan amount target.

4.3 Predicting Diabetes

The diabetes prediction task is a binary classification problem using medical diagnostic data. **Logistic Regression** is used for its simplicity and interpretability, while **Random Forest** is chosen for its robustness and ability to handle feature interactions.

4.4 Classification of Email Spam

The email spam classification task involves identifying whether an email is spam or not based on its content. **Naive Bayes** is chosen due to its effectiveness with high-dimensional text data, while **SVM** is used for its strong performance in binary classification with clear margins.

4.5 Handwritten Character Recognition / MNIST

The MNIST handwritten digit recognition task involves classifying grayscale images of digits (0–9). **Convolutional Neural Networks (CNNs)** are chosen for their ability to capture spatial

hierarchies in image data, while **SVM** serves as a strong traditional baseline for high-dimensional image classification.

EDA-Guided Model Choices

Type of Model	**Model(s)**	**When to Use**
Linear Models	Linear Regression, LDA	Data shows linear relationships and follows normality assumptions.
Nonlinear Models	Decision Trees, Random Forests	Data is highly skewed or involves complex feature interactions.
Distance-Based Models	KNN, SVM (RBF Kernel)	Features are normalized and geometric proximity is important.
Probabilistic Models	Naive Bayes	Feature independence assumption is reasonable.

Table 2: Model Selection Based on Data Characteristics

5 Learning Practices:

- Explored the use of standard datasets (Iris, MNIST, etc.) for classification and regression tasks.
- Understood supervised learning models such as KNN, SVM, Logistic Regression, Linear Regression, Random Forest, Naive Bayes, and CNN.
- Evaluated model performance using appropriate metrics like accuracy, precision, recall, and confusion matrix.
- Gained experience in selecting suitable algorithms based on data type, problem type, and dimensionality.
- Practiced preprocessing techniques including normalization, encoding, and handling missing values.
- Improved hands-on skills using Python libraries like Scikit-learn, Pandas, Matplotlib, and TensorFlow/Keras.

GitHub Repository:
https://github.com/Bloomberg890/ML$_{L}aboratory$