Machine Learning Algorithms Laboratory

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

SSN College of Engineering
Name: Spinola theres N
`Register Number: 3122237001051`

August 5, 2025

Aim

To explore Python libraries (NumPy, Pandas, SciPy, Scikit-learn, Matplotlib) and implement machine learning workflows on real-world datasets to identify appropriate ML tasks and apply suitable models.

Part 1: Python Library Exploration

1. NumPy

NumPy is used for numerical computing and array manipulations.
Common functions:

- `np.array()`, `reshape()`, `mean()`, `dot()`, etc.

2. Pandas

Pandas is ideal for structured data manipulation.

- `read_csv()`, `groupby()`, `fillna()`, `drop()`, etc.

3. SciPy

Used for scientific computations and statistical analysis.

- `scipy.stats`, `optimize.minimize()`, etc.

4. Scikit-learn

Library for ML models and preprocessing.
Key components:

- `train_test_split()`, `StandardScaler()`, `LogisticRegression()`, etc.

5. Matplotlib

Used for data visualization.

- `plot()`, `scatter()`, `hist()`, `boxplot()`, etc.

Part 2: Identifying ML Models for Public Datasets

Public repositories such as the **UCI Machine Learning Repository** and **Kaggle Datasets** were explored to identify suitable datasets. The following datasets were downloaded, and appropriate machine learning models were proposed based on the nature of the data and problem statement.

Dataset-wise Analysis and Model Identification

1. **Loan Amount Prediction:** Dataset includes applicant details (e.g., income, credit history, marital status). The goal is to predict the **loan amount**, which is a numeric value. → **ML Type: Supervised Learning (Regression)**

2. **Handwritten Character Recognition (MNIST):** The dataset consists of grayscale images of handwritten digits (0–9) as input and digit labels as output. → **ML Type: Supervised Learning (Classification)**

3. **Email Spam Classification:** Input includes email text features. Output is a binary label (spam or not). → **ML Type: Supervised Learning (Classification)**

4. **Predicting Diabetes (Pima Indian Diabetes Dataset):** Predicts diabetes presence based on health metrics. → **ML Type: Supervised Learning (Classification)**

5. **Iris Dataset:** Dataset contains flower measurements (features) and the species type (target). → **ML Type: Supervised Learning (Classification)**

Summary Table of Datasets and ML Model Types

Dataset	Type of ML Learning	ML Task Type
Loan Amount Prediction	Supervised Learning	Regression
Handwritten Character Recognition (MNIST)	Supervised Learning	Classification
Email Spam Classification	Supervised Learning	Classification
Predicting Diabetes	Supervised Learning	Classification
Iris Dataset	Supervised Learning	Classification

Table 1: Appropriate Machine Learning Models Identified for Each Dataset

Part 3: ML Workflow Steps

The general steps followed for all datasets:

1. Loading dataset using Pandas or Scikit-learn.
2. Exploratory Data Analysis (EDA) and Visualization.
3. Data Preprocessing (null values, encoding, scaling).
4. Feature Selection using SelectKBest, Chi-square, ANOVA.
5. Data Splitting (train/test).
6. Model Building and Performance Evaluation.

Code Sample: Iris Dataset Classification

```
from sklearn.datasets import load_iris
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report

iris = load_iris()
X, y = iris.data, iris.target
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)
```

```
model = LogisticRegression()
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
print(classification_report(y_test, y_pred))
```

Inference Table

Dataset	**ML Task**	**Feature Selection Technique**	**Suitable Algorithm(s)**
Iris Dataset	Classification	ANOVA (f_classif)	Logistic Regression, SVM
Loan Amount Prediction	Regression	Correlation Matrix	Linear Regression, Decision Tree
Predicting Diabetes	Classification	SelectKBest, Chi-square	Random Forest, KNN
Email Spam Classification	Classification	SelectKBest, PCA	Naive Bayes, Logistic Regression
MNIST Digit Recognition	Classification	PCA (dimensionality reduction)	CNN, SVM, KNN

Learning Outcomes and Reflections

- Developed practical skills in data loading, preprocessing, and visualization.
- Understood the importance of selecting the correct ML algorithm.
- Learned how Python libraries work together to build ML pipelines.
- Gained insight into real-world datasets and ML task categorization.