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	UCS2612 – Machine Learning Algorithms Laboratory		
Academic Year	2025-2026 (Even)	Batch: 2023-2027	**Due Date: 23/12/25**

Experiment 1: Exploratory Data Analysis and Implementation of ML Packages

(Numpy, Scipy, Scikit-Learn, Matplotlib, Seaborn, TensorFlow, PyTorch)

Name: Rahul V S
Reg. No: 3122235001104
Section: CSE B

---

1. Aim:
To utilize Python's machine learning ecosystem to perform comprehensive Exploratory Data Analysis (EDA) on standard public datasets. The objective is to understand data distributions, identify correlations, perform preprocessing (imputation, scaling), and select appropriate machine learning algorithms for tasks ranging from regression to image classification.

2. Libraries & Environment Setup:
The following libraries were imported and utilized for data manipulation, visualization, and model building:

- **Core Data Science:** `Pandas` (Dataframes), `Numpy` (Numerical computing), `Scipy` (Scientific computing).
- **Visualization:** `Matplotlib.pyplot`, `Seaborn` (Statistical plotting).
- **Machine Learning (Scikit-Learn):**
  - *Preprocessing:* `StandardScaler`, `MinMaxScaler`, `LabelEncoder`, `SimpleImputer`.
  - *Model Selection:* `train_test_split`, `cross_val_score`.
  - *Algorithms:* `RandomForestClassifier`, `LogisticRegression`.
  - *Metrics:* `accuracy_score`, `confusion_matrix`, `classification_report`.
- **Deep Learning & Boosting:** `TensorFlow` (Keras Sequential API), `PyTorch` (`torch.nn`, `torch.optim`), `XGBoost`.

3. Methodology and Concepts:
The experimental workflow followed a structured approach to Data Science:

1. **Data Ingestion & Inspection:** Datasets were loaded using Pandas. Initial inspection involved checking dimensions (`.shape`), data types (`.info()`), and statistical summaries (`.describe()`). Missing values were identified using `.isna().sum()`.

2. **Univariate Analysis (Distribution):** Histograms and Kernel Density Estimation (KDE) plots were generated for numerical features to identify skewness and the underlying probability distribution of the data.

3. **Bivariate/Multivariate Analysis (Correlation):** Pearson correlation matrices were computed to analyze the linear relationship between features:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$$

   Heatmaps were used to visualize these correlations to detect multicollinearity.

4. **Data Preprocessing:**

   - **Imputation:** Missing values were handled using `SimpleImputer`.
   - **Encoding:** Categorical variables were converted to numeric forms using `LabelEncoder`.
   - **Scaling:** Features were normalized using `StandardScaler` and `MinMaxScaler` to ensure varying scales did not bias gradient-based models.

5. **Outlier Detection:** Boxplots were utilized to visualize the Interquartile Range (IQR) and identify anomalous data points.

4. Results and Discussions:

0.1 Dataset 1: Loan Data Analysis

The Loan dataset consists of **45,000 entries** and **14 columns**, containing information such as `person_age`, `person_income`, `loan_amnt`, and `loan_status`.

- **Data Structure:** Numerical columns with greater than 10 unique values (e.g., `loan_int_rate`, `person_emp_exp`) were isolated for distribution analysis.
- **Distribution Findings:** Histograms revealed that `person_income` and `loan_amnt` are right-skewed, suggesting that most applicants apply for smaller loans and have lower-to-middle income, with a few high-net-worth outliers.
- **Correlation:** `loan_percent_income` showed a significant relationship with loan status, indicating debt-to-income ratio is a key driver for default risk.

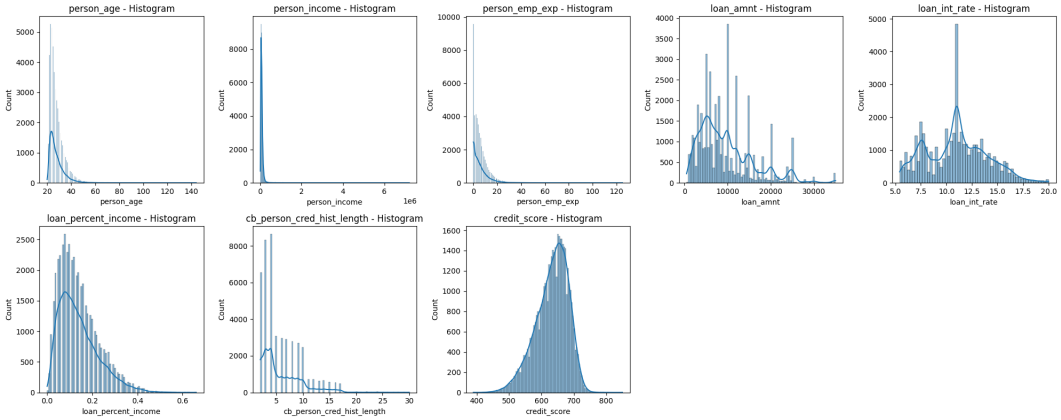
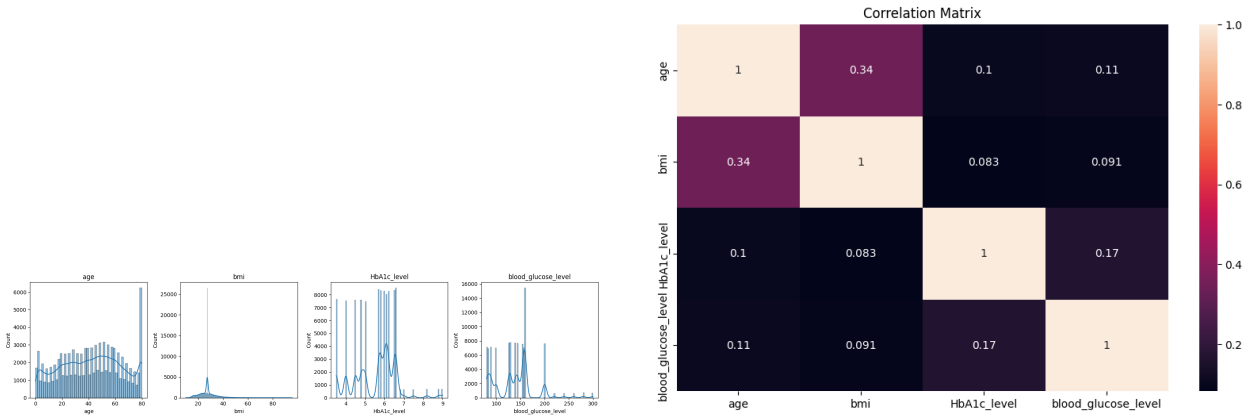


Figure 1: Distribution of Numerical Features (Age, Income, Loan Amount) in Loan Dataset

0.2 Dataset 2: Diabetes Prediction Analysis

The analysis focused on health indicators such as Glucose, BMI, and Age to predict the outcome (Diabetic/Non-Diabetic).

- **Feature Analysis:** The distributions of Glucose and BMI were analyzed. While Glucose followed a roughly normal distribution, BMI showed several outliers in the higher range.
- **Correlation:** A correlation heatmap highlighted that `Glucose` levels have the strongest positive correlation with the target variable, followed by `BMI` and `Age`.



(a) Feature Distributions

(b) Correlation Matrix

Figure 2: EDA for Diabetes Dataset

0.3 Dataset 3: Email Spam Classification

Natural Language Processing (NLP) techniques were applied to analyze text data for distinguishing 'Spam' vs. 'Ham'.

- **Class Imbalance:** The dataset showed a distinct imbalance, with legitimate emails ('Ham') significantly outnumbering 'Spam'.

- **Feature Engineering:** Word count distributions indicated that spam emails tend to have a higher word count on average compared to non-spam emails.

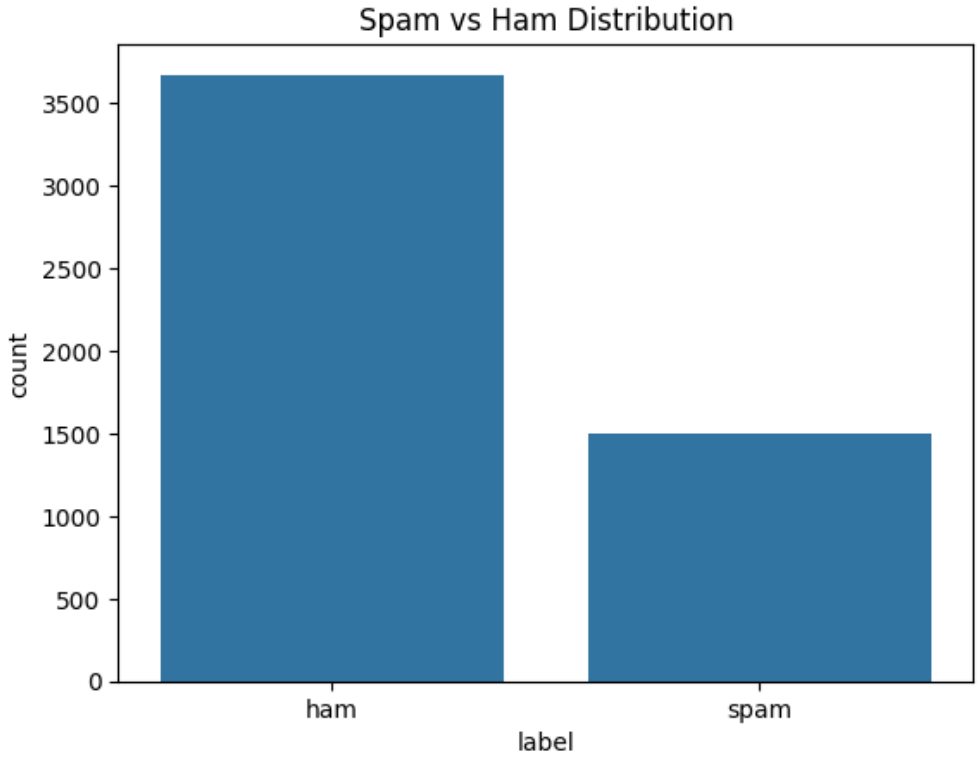


Figure 3: Class Balance and Word Count Distribution

0.4 Dataset 4: Handwritten Character Recognition (MNIST)

The MNIST dataset, consisting of 28x28 pixel grayscale images, was analyzed for image classification tasks.

- **Pixel Intensity:** Analysis of pixel values showed a bimodal distribution (background vs. digit strokes).
- **Class Distribution:** The dataset is well-balanced across digits 0-9, making it suitable for training Convolutional Neural Networks (CNNs) using TensorFlow/Keras.

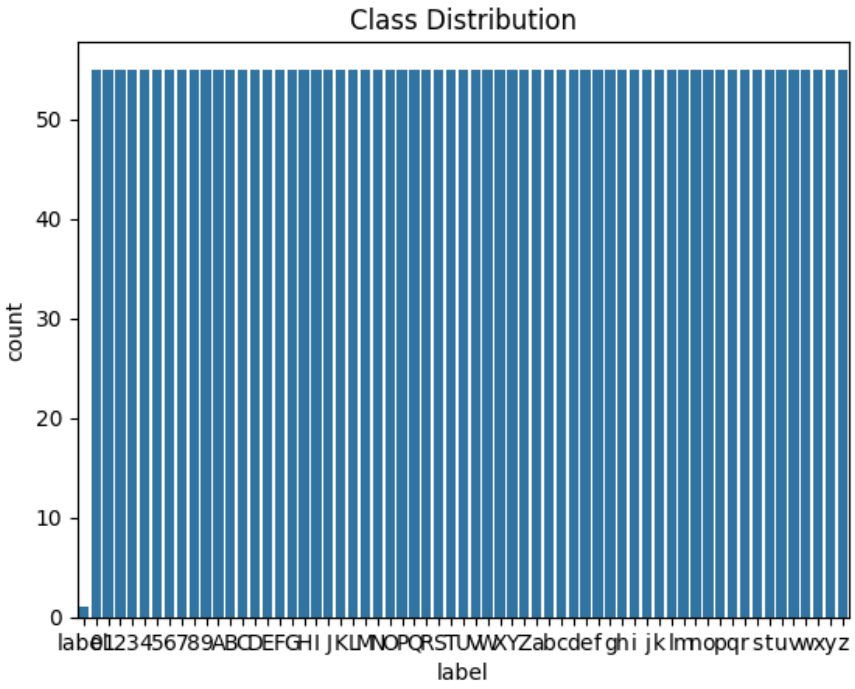


Figure 4: Class Distribution of MNIST Digits

0.5 Dataset 5: Iris Species Classification

The Iris dataset was used to study multi-class classification based on sepal and petal dimensions.

- **Separability:** Pairplots revealed that *Iris setosa* is linearly separable from the other two species based on petal length and width.
- **Correlation:** High positive correlation (> 0.9) was observed between Petal Length and Petal Width.

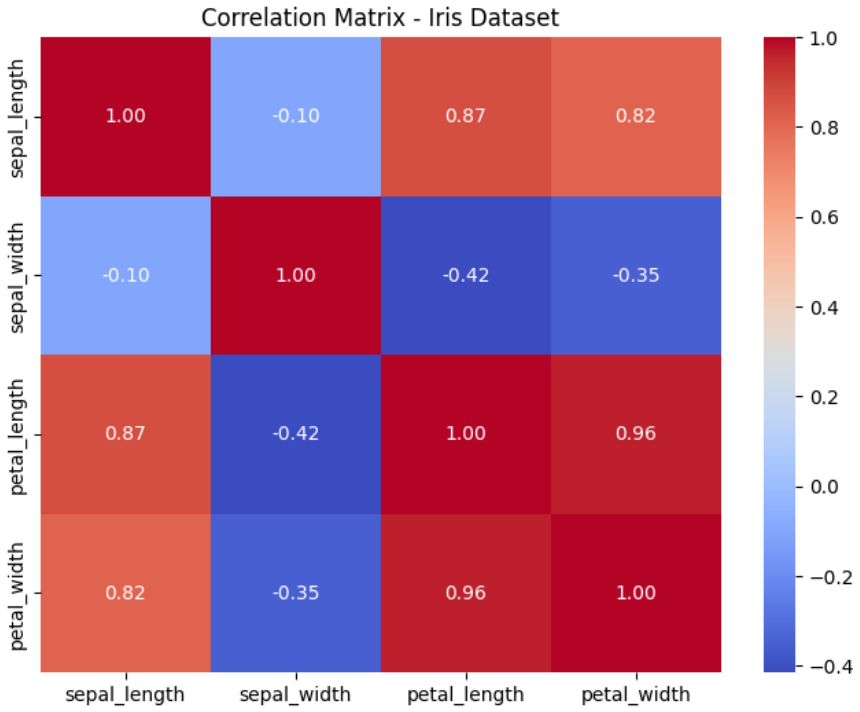


Figure 5: Correlation Heatmap of Iris Features

5. Inference & Model Selection:

Based on the EDA performed, the following Machine Learning tasks and algorithms were identified as suitable for the respective datasets:

Dataset	**Task Type**	**Key Feature Strategy**	**Suitable Algorithms**
Loan Amount	Binary Classification	Outlier removal, Scaling	XGBoost, Random Forest
Diabetes	Binary Classification	Standardization (StandardScaler)	Logistic Regression, SVM
Email Spam	NLP Classification	TF-IDF Vectorization	Naive Bayes, RNN (PyTorch)
MNIST	Image Classification	Normalization (0-1 Scaling)	CNN (TensorFlow), SVM
Iris Dataset	Multi-class	Dimensionality Reduction (PCA)	k-NN, Decision Trees

Table 1: Inference on Model Selection based on EDA

6. Conclusion:

The exploratory data analysis successfully provided insights into the structure and quality of five distinct datasets. Key takeaways include:

- **Loan Data:** Identified skewness in income data requiring log-transformation and missing values requiring imputation.
- **Preprocessing Needs:** The presence of categorical variables in Loan and Spam datasets necessitates Label Encoding or One-Hot Encoding.
- **Model Suitability:** The complexity of the MNIST dataset justifies the import of Deep Learning libraries like TensorFlow and PyTorch, while structured tabular data (Loan, Diabetes) is best suited for ensemble methods like Random Forest and XGBoost.