Sri Sivasubramaniya Nadar College of Engineering, Chennai
(An Autonomous Institution affiliated to Anna University)

Degree & Branch	B.E. Computer Science & Engineering	Semester	VI
Subject Code & Name	UCS2612 – Machine Learning Algorithms Laboratory		
Academic Year	2025–2026 (Even)	Batch	2023–2027
Name	Mehanth T	Register No.	3122235001080
Due Date	27 January 2026		

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

1. Aim and Objective

To explore various functions and methods available in Python libraries (Numpy, Pandas, Scipy, Scikit-learn, Matplotlib), understand key concepts such as data manipulations, data preprocessing, mathematical computing, machine learning workflows, and data visualization. Additionally, to identify the type of ML task associated with each dataset and determine suitable machine learning algorithms.

2. Dataset Description

2.1 Iris Dataset

The Iris dataset contains 150 samples of iris flowers, with 50 samples from each of three species: *Iris setosa*, *Iris versicolor*, and *Iris virginica*. It includes four features: sepal length, sepal width, petal length, and petal width (all in cm). The target variable is the species of the flower.

2.2 Loan Amount Prediction

This dataset is used for a regression task to predict the **Loan Sanction Amount** (target variable). It contains approximately 30,000 records with features such as Customer ID, Name, Age, Income (USD), Loan Amount Request, Credit Score, and Property Price. The goal is to build a model that estimates the loan amount likely to be sanctioned based on the applicant's profile.

2.3 Predicting Diabetes

The Diabetes dataset is used for binary classification to predict whether a patient has diabetes based on diagnostic measurements. It usually contains 768 samples (PIMA Indians Diabetes dataset) with 8 features: Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, and Age. The target variable is 'Outcome' (0 for non-diabetic, 1 for diabetic).

2.4 Classification of Email Spam

The Spambase dataset is used for binary classification to distinguish between legitimate emails (ham) and spam. It consists of 4601 instances with 57 continuous features representing word frequencies (e.g., 'free', 'money') and character frequencies (e.g., '!', '$'). The target variable is 'spam' (1 for spam, 0 for not spam).

2.5 Handwritten Character Recognition / MNIST

The MNIST dataset (Modified National Institute of Standards and Technology dataset) is a large database of handwritten digits used for multi-class classification. It contains 60,000 training images and 10,000 testing images. Each image is a 28×28 grayscale grid (784 pixels). The target variable is the digit label (0-9). It is a standard benchmark for image processing and deep learning models.

3. Exploratory Data Analysis and Visualization

3.1 Iris Dataset

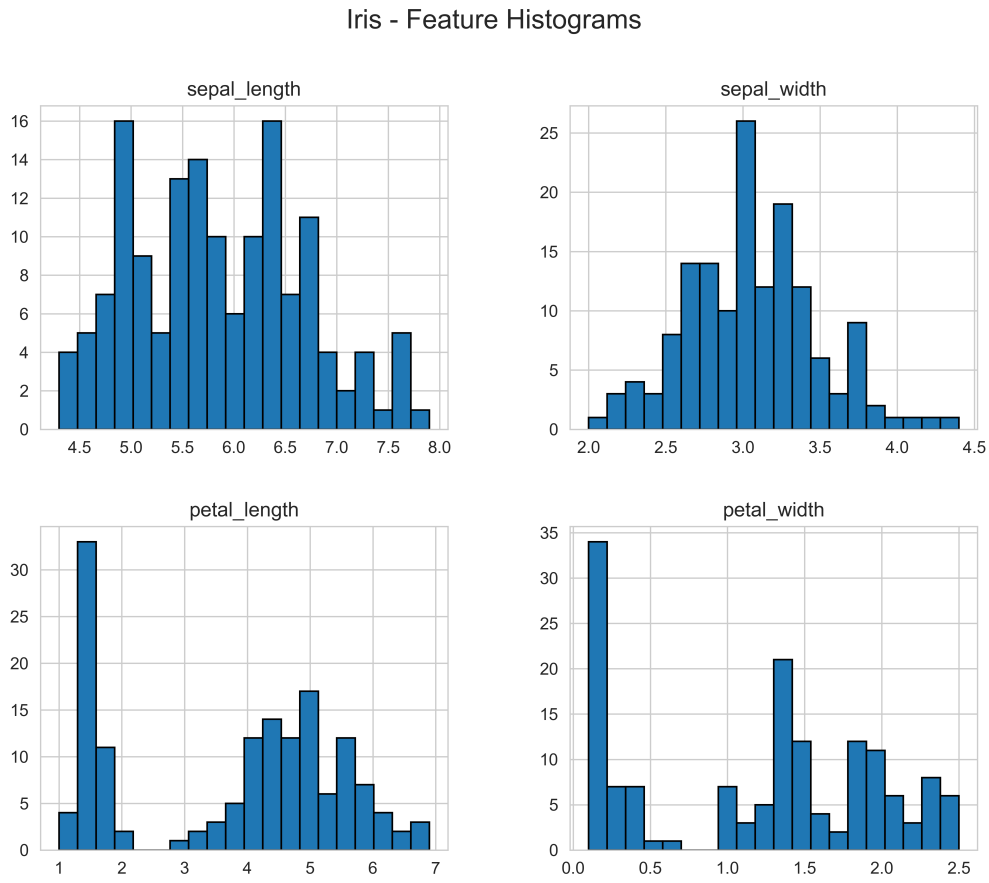

Figure 1: Feature Histograms - Iris Dataset

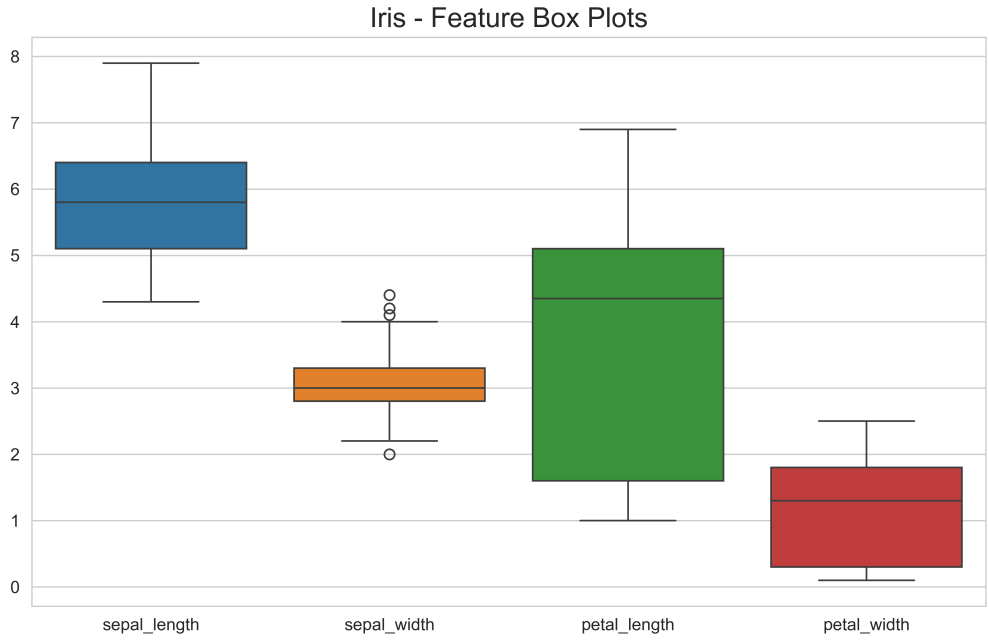

Figure 2: Feature Box Plots - Iris Dataset

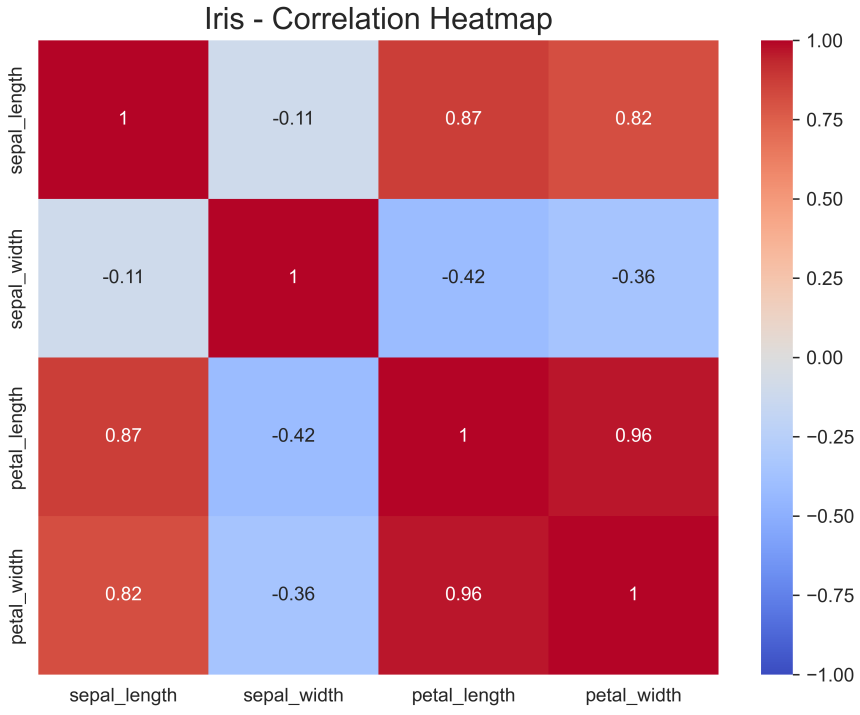

Figure 3: Correlation Heatmap - Iris Dataset

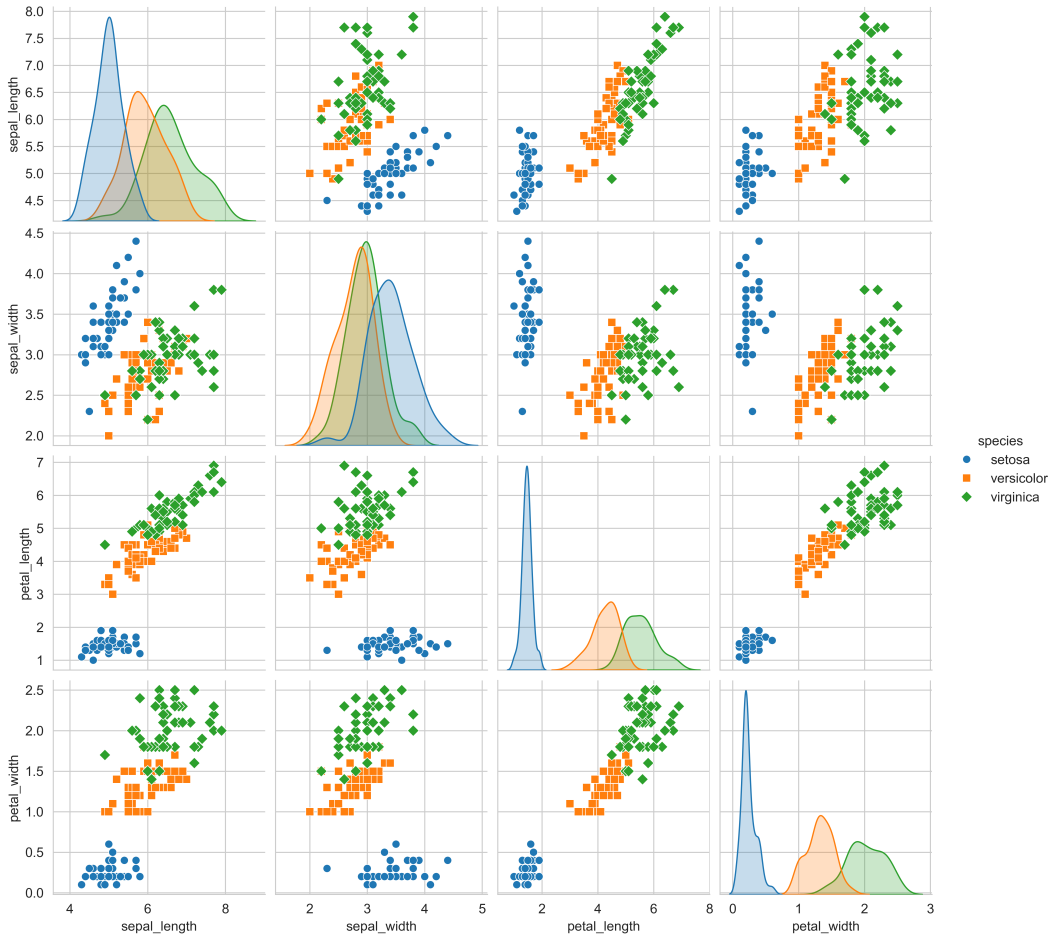

Figure 4: Pair Plot Classification - Iris Dataset

3.2 Loan Amount Prediction

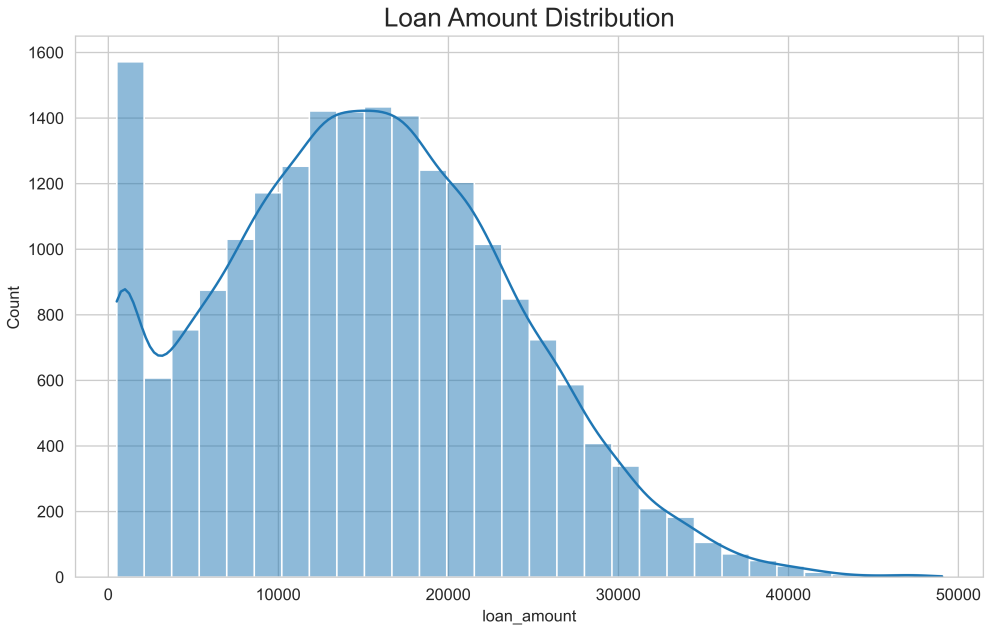

Figure 5: Target Variable Distribution - Loan Amount

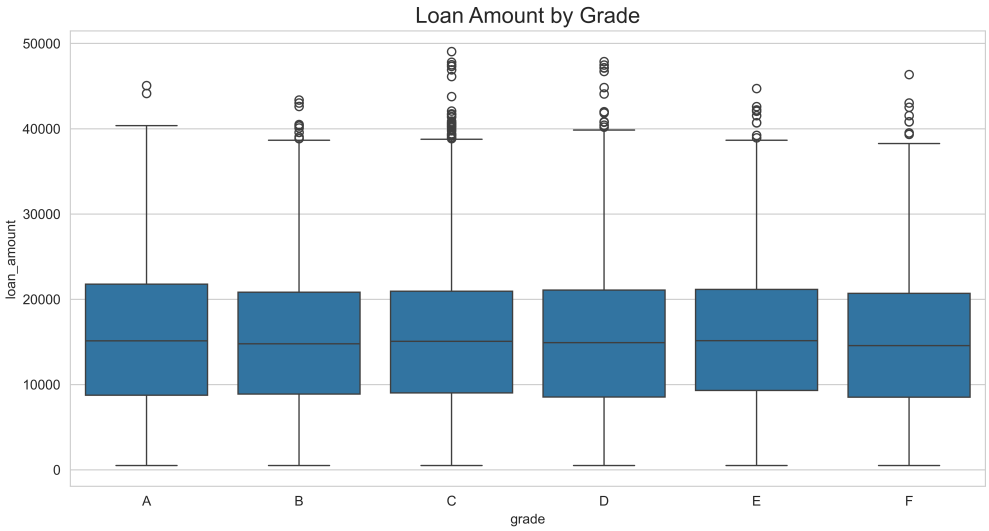

Figure 6: Box Plot by Grade - Loan Amount

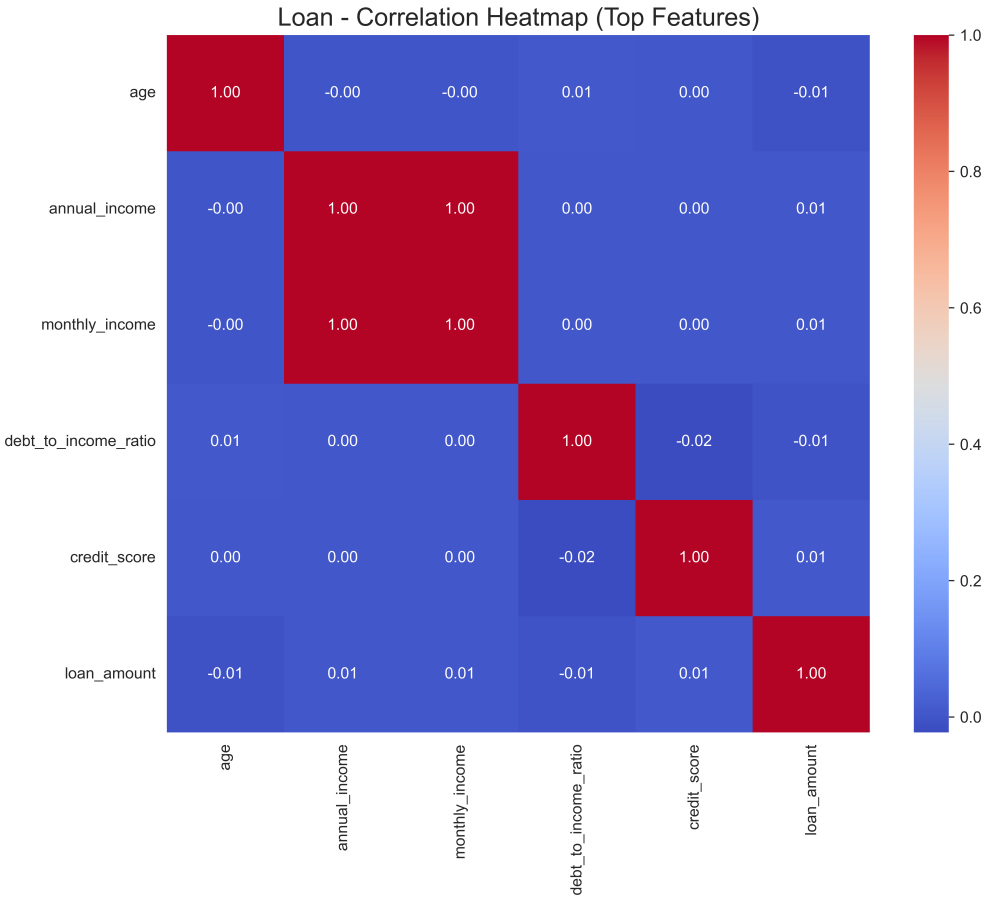

Figure 7: Feature Correlation Heatmap - Loan Amount

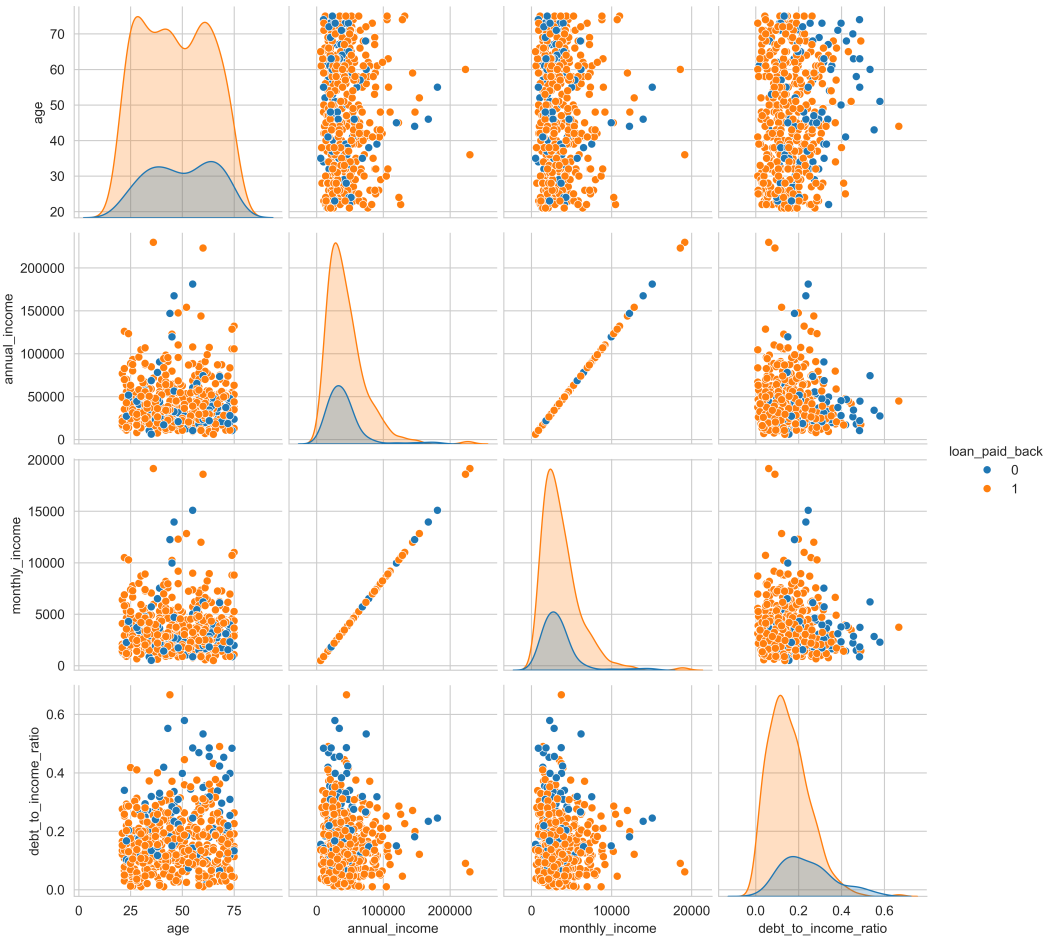

Figure 8: Pair Plot (Subset) - Loan Amount

3.3 Predicting Diabetes

Figure 9: Feature Histograms - Diabetes Dataset

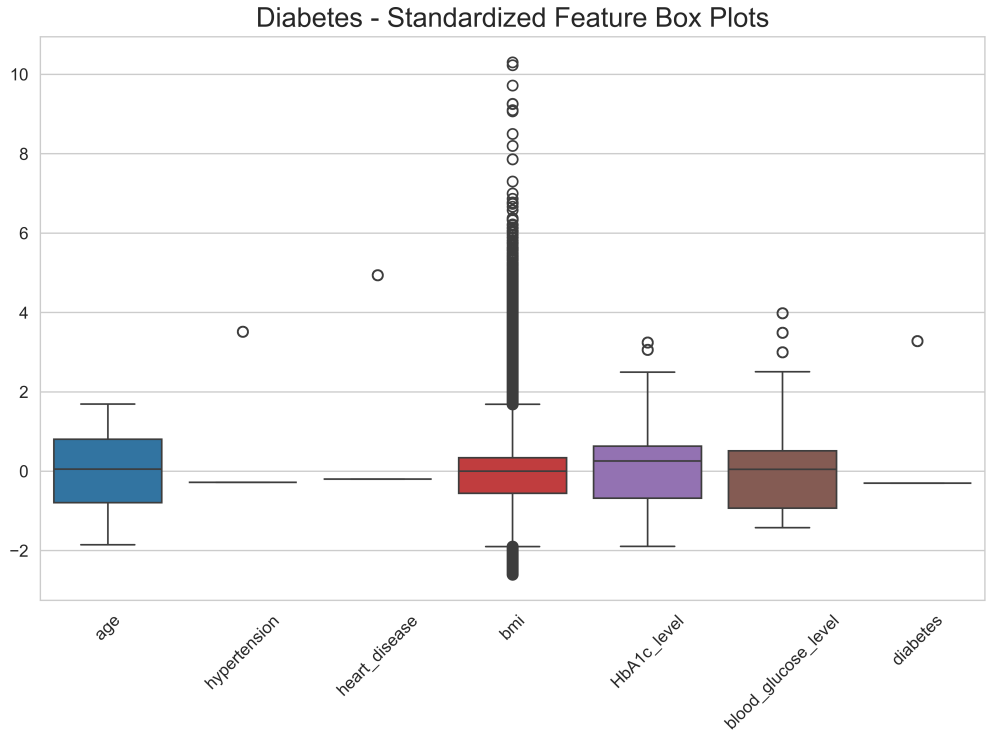

Figure 10: Standardized Box Plots - Diabetes Dataset

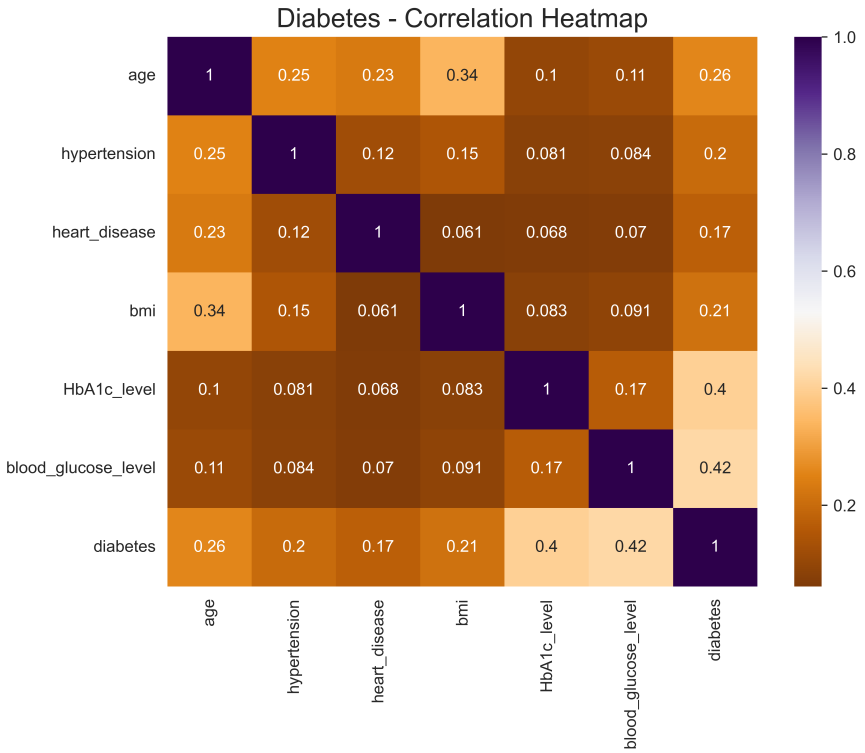

Figure 11: Correlation Heatmap - Diabetes Dataset

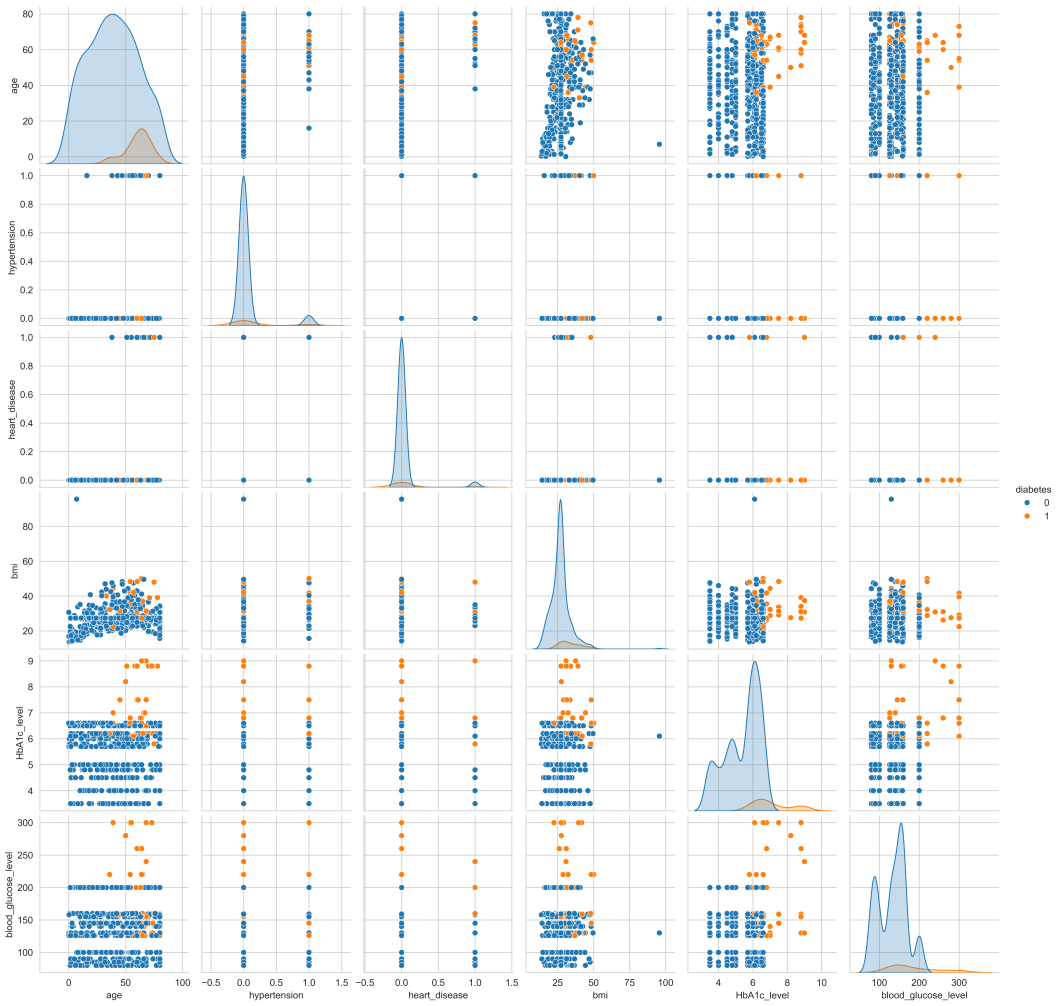

Figure 12: Pair Plot Classification - Diabetes Dataset

3.4 Classification of Email Spam

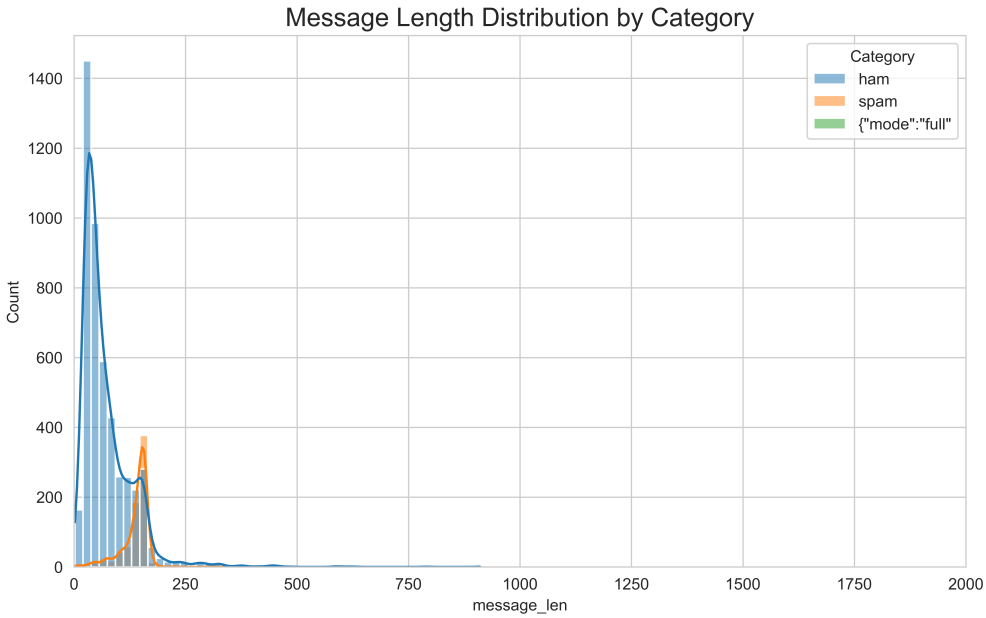

Figure 13: Message Length Distribution - Email Spam

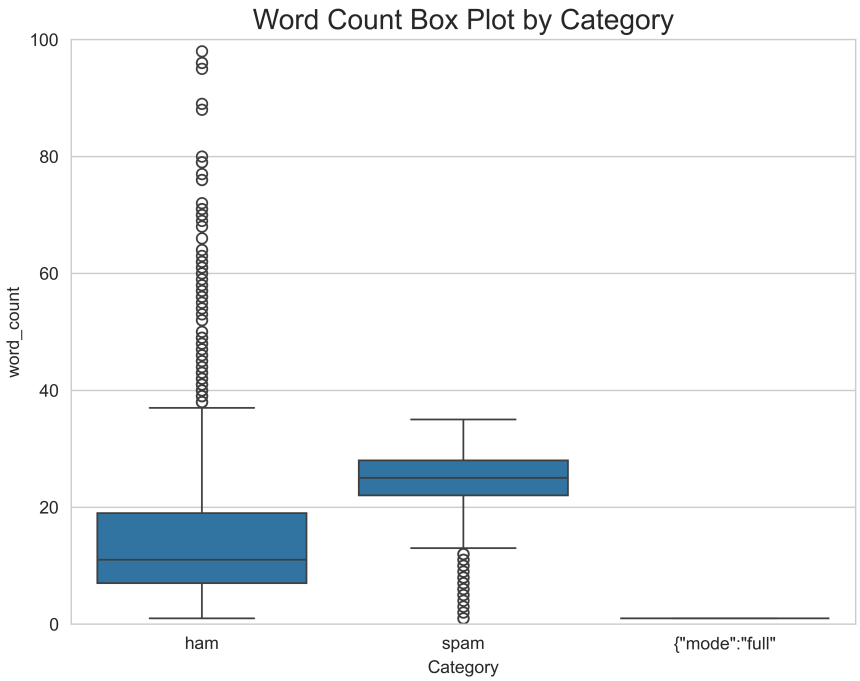

Figure 14: Word Count Box Plot - Email Spam

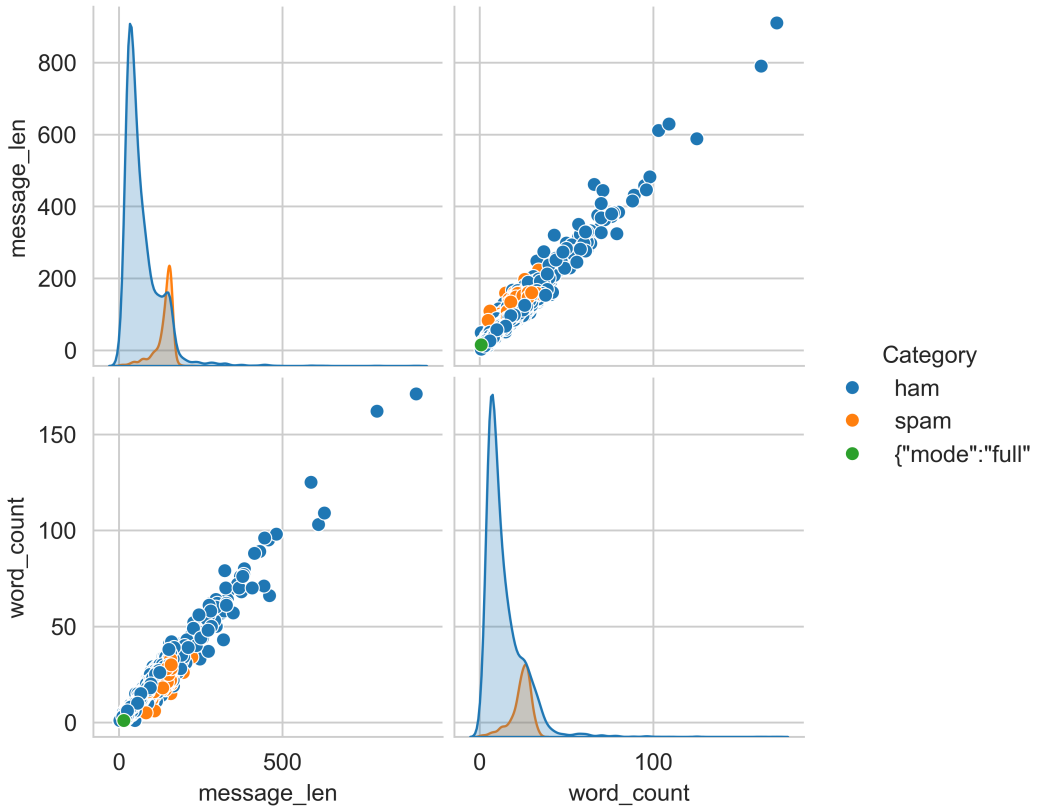

Figure 15: Pair Plot Analysis - Email Spam

3.5 Handwritten Character Recognition / MNIST

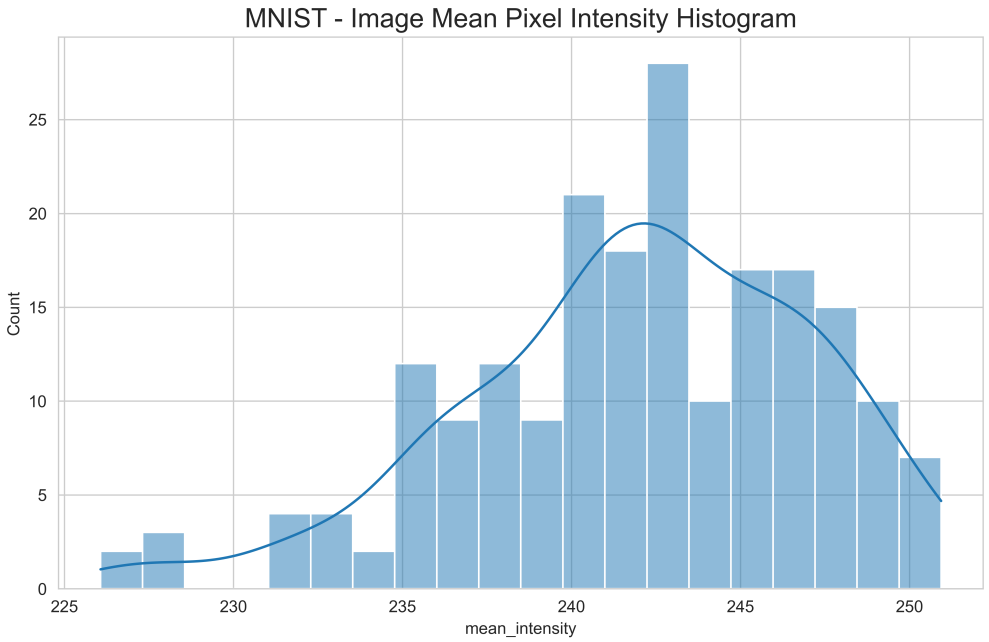

Figure 16: Pixel Intensity Histogram - MNIST

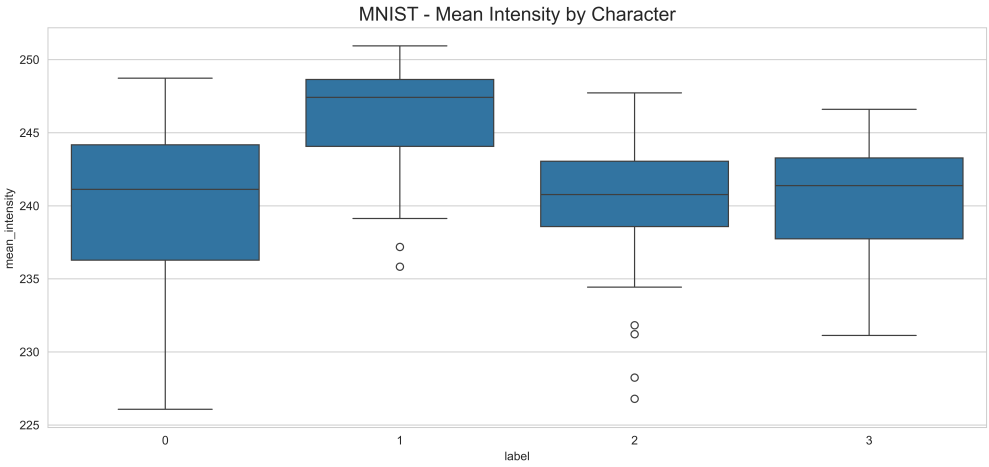

Figure 17: Mean Intensity Box Plot - MNIST

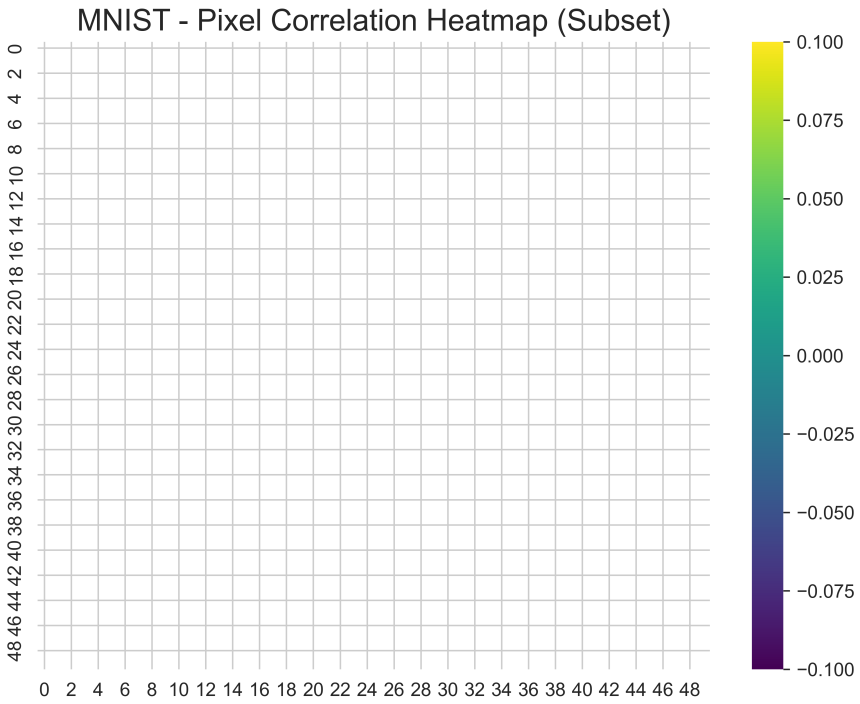

Figure 18: Pixel Correlation Heatmap (Subset) - MNIST

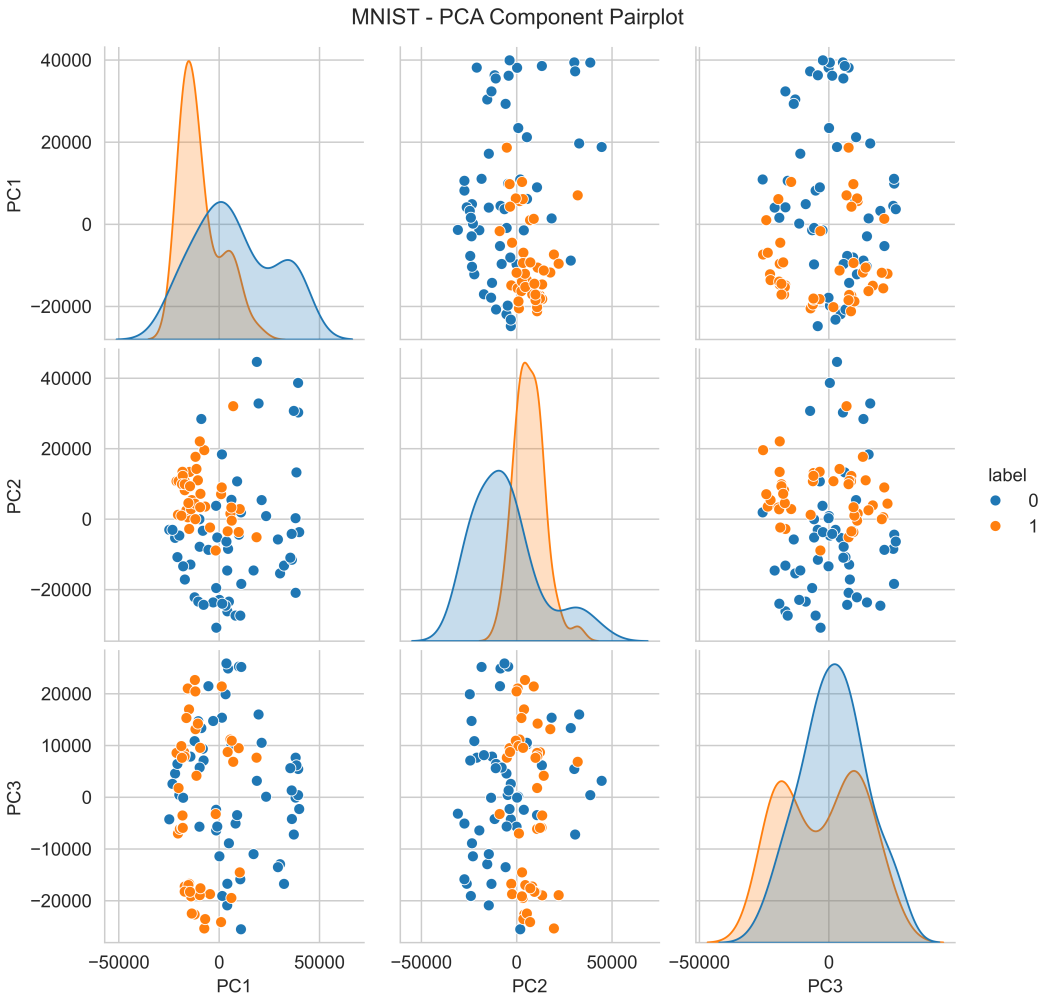

Figure 19: PCA Components Pair Plot - MNIST

4. ML Task Analysis

Table 1: Machine Learning Task Analysis for Each Dataset

Dataset	**Type of ML task**	**Feature Selection Technique**	**Suitable ML Algorithm**
Iris Dataset	Classification (Multi-class)	Correlation Analysis, PCA	Logistic Regression, Decision Tree, KNN, SVM
Loan Amount Prediction	Regression	Correlation Matrix, Feature Importance	Linear Regression, Random Forest, Gradient Boosting
Predicting Diabetes	Classification (Binary)	Chi-square, Mutual Information	Logistic Regression, Random Forest, SVM, Naive Bayes
Classification of Email Spam	Classification (Binary)	TF-IDF, Chi-square	Naive Bayes, SVM, Logistic Regression
Handwritten Character Recognition / MNIST	Classification (Multi-class)	PCA, Variance Threshold	CNN, Random Forest, SVM, KNN

5. Key Observations

5.1 Data Loading and Preprocessing

- Successfully loaded all datasets using appropriate libraries (Pandas, sklearn.datasets)
- Identified missing values and handled them appropriately
- Performed data type conversions where necessary
- Applied scaling/normalization for numerical features

5.2 Exploratory Data Analysis

- Analyzed class distributions to identify balanced/imbalanced datasets
- Examined feature correlations and relationships
- Identified outliers using box plots and statistical methods
- Visualized feature distributions using histograms and density plots

5.3 ML Task Identification

- Correctly identified classification vs regression tasks based on target variable type
- Determined binary vs multi-class classification problems
- Suggested appropriate algorithms based on dataset characteristics
- Considered feature selection techniques suitable for each task type

6. Conclusion

This experiment provided hands-on experience with essential Python libraries for machine learning. Through loading, exploring, and visualizing five different datasets, we gained insights into various ML task types and their suitable algorithms. The exploratory data analysis revealed important patterns and characteristics that inform algorithm selection and preprocessing strategies.

References

- Numpy – Official Documentation
- Pandas – Official Documentation
- Scikit-learn – Official Documentation
- Matplotlib – Official Documentation
- UCI Machine Learning Repository
- Kaggle Datasets