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

Name: KESHAV K S
Reg.No: 3122235001067
Class: CSE-B

Aim:
To explore and work with Python packages like Numpy, Scikit-learn, and Matplotlib on datasets from public repositories and identify ML tasks, feature selection techniques, and suitable algorithms.

Libraries used:

- Numpy (imported as `np`)
- Pandas (imported as `pd`)
- Matplotlib.pyplot (imported as `plt`)
- Seaborn (imported as `sns`)
- OpenCV (`cv2`)
- OS (Standard Library)
- Math (Standard Library)

Mathematical/Theoretical description of the algorithm/objective performed:

- **Objective:** Exploratory Data Analysis (EDA) — summarize distributions, detect outliers, assess class balance, and reveal pairwise relationships/correlations.
- **Descriptive statistics**
  - Mean: $\mu = \frac{1}{n}\sum x_i$
  - Sample variance: $s^2 = \frac{1}{n-1}\sum(x_i - \mu)^2$
  - Quantiles (Median, $Q1$, $Q3$) used for spread and boxplot construction.
- **Histograms**
  - Empirical density approximation via bin counts; visualizes frequency/mode structure.

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where K is the kernel (e.g., Gaussian) and h is the bandwidth.

- **Boxplots**
  - Show Median, $Q1$, $Q3$; Interquartile Range $IQR = Q3 - Q1$.
  - Whiskers typically extend to min/max within $[Q1 - 1.5 \cdot IQR,\ Q3 + 1.5 \cdot IQR]$; points outside are shown as outliers.
- **Correlation Heatmap**
  - Pearson correlation coefficient:

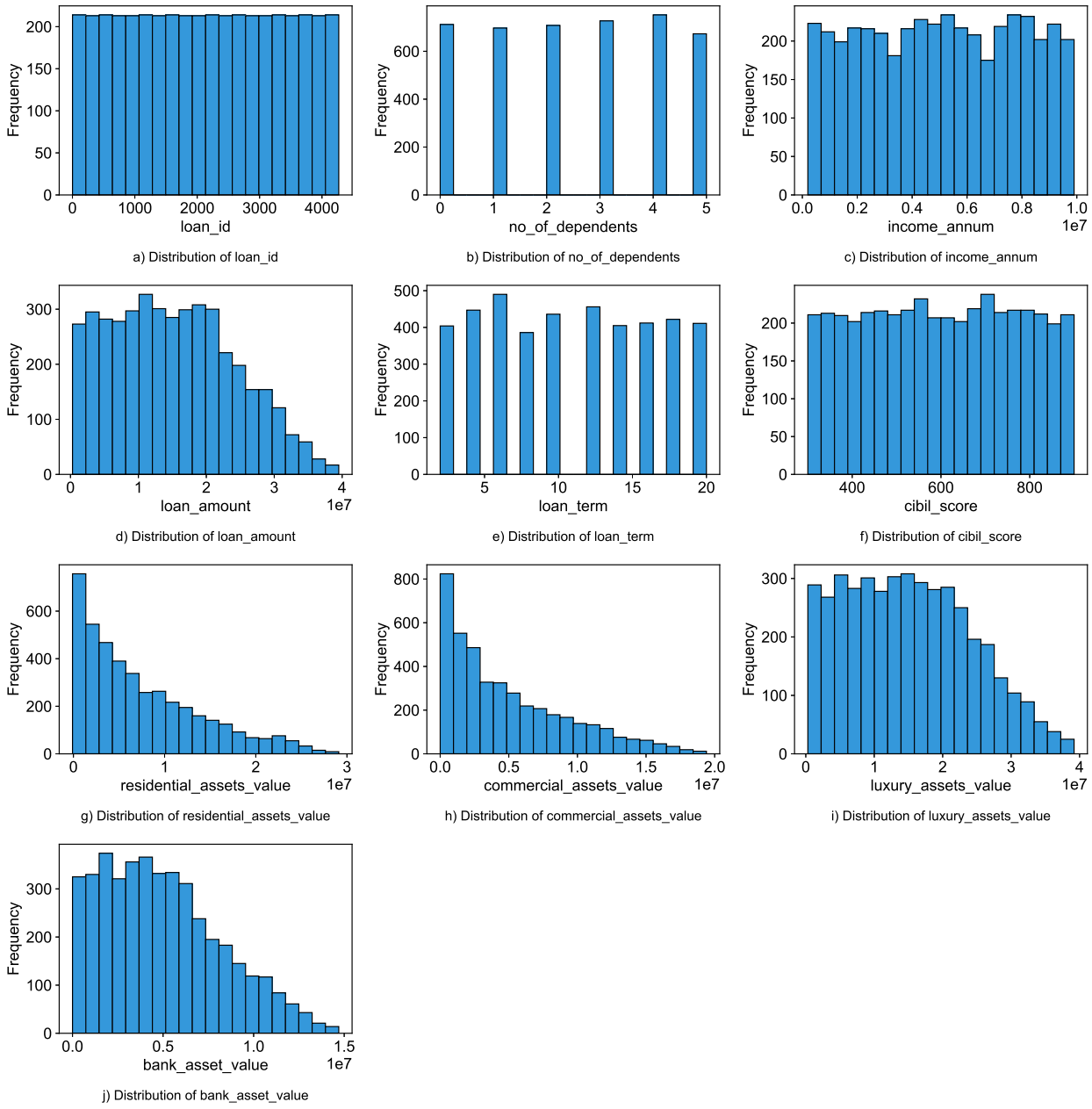
$$r_{XY} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

$$\text{cov}(X, Y) = \frac{1}{n-1} \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

  - Values in $[-1, 1]$ indicate linear association strength/direction.
- **Countplots (Categorical)**
  - Bar heights = counts per category; optionally use hue for class-conditional counts to inspect imbalance/conditional distributions.
- **Missing Values**
  - Count missing per column to decide imputation/removal.

Results and Discussions:

Loan Dataset



Histograms showing the distribution of numeric features in the dataset.

Figure 1: Distribution of numerical features using Histograms .

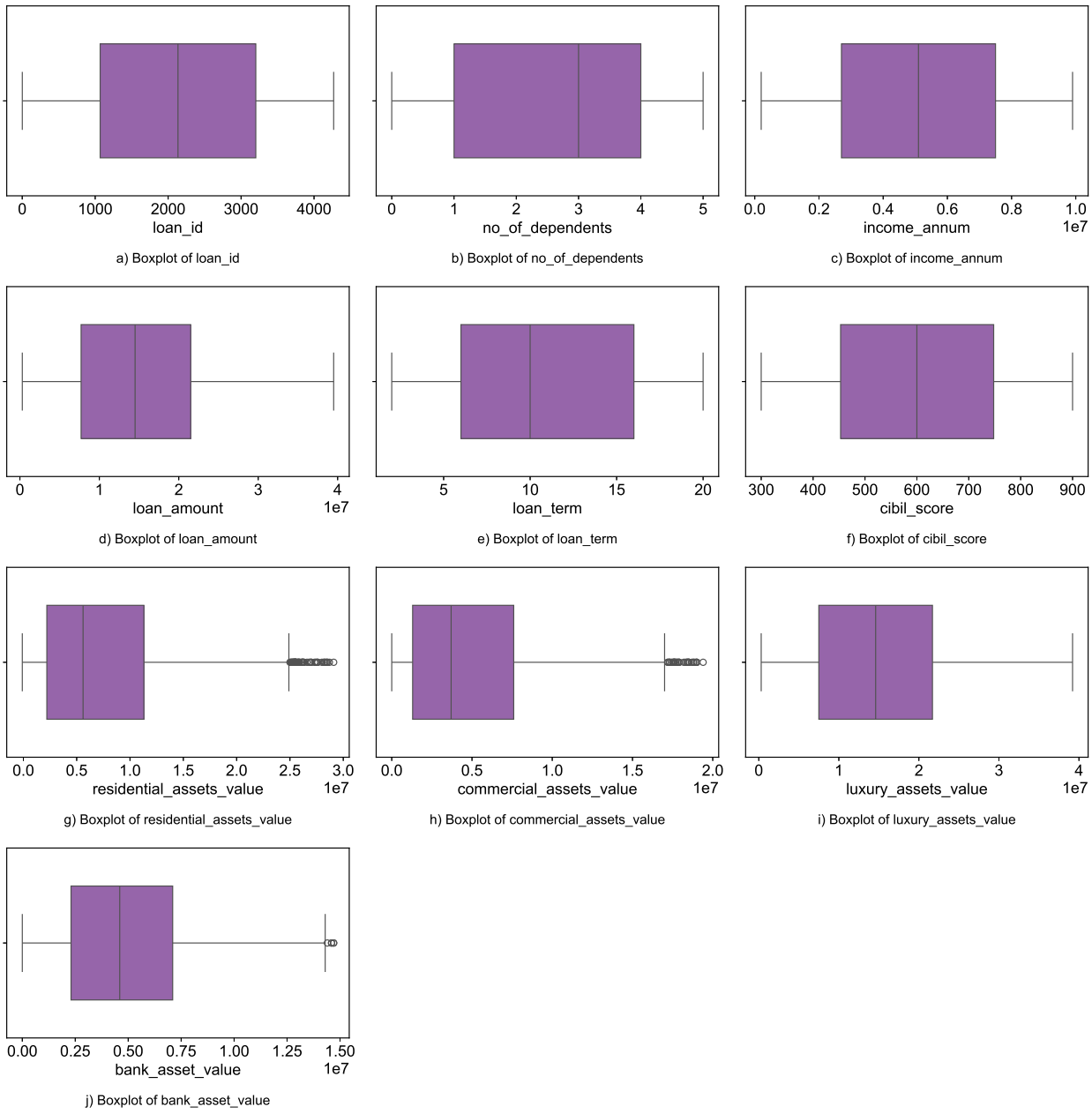


Figure: Boxplots identifying outliers and quartiles for numeric features.

Figure 2: Boxplot visualisation

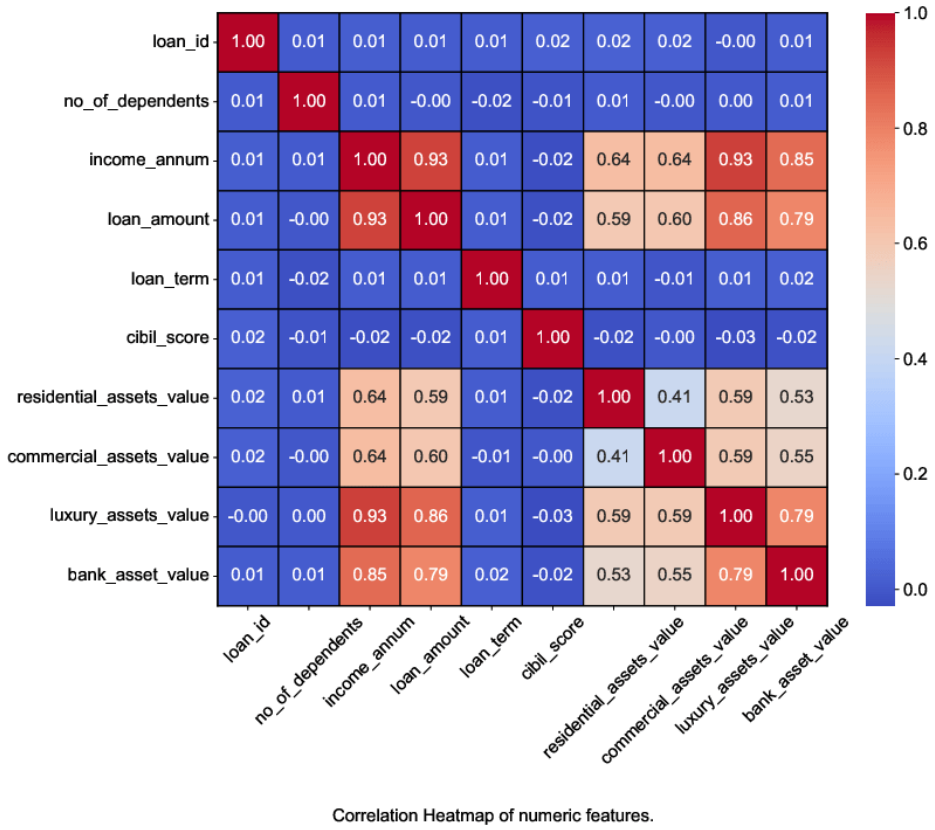


Figure 3: Correlation Heatmap of financial and asset variables.

Email Dataset

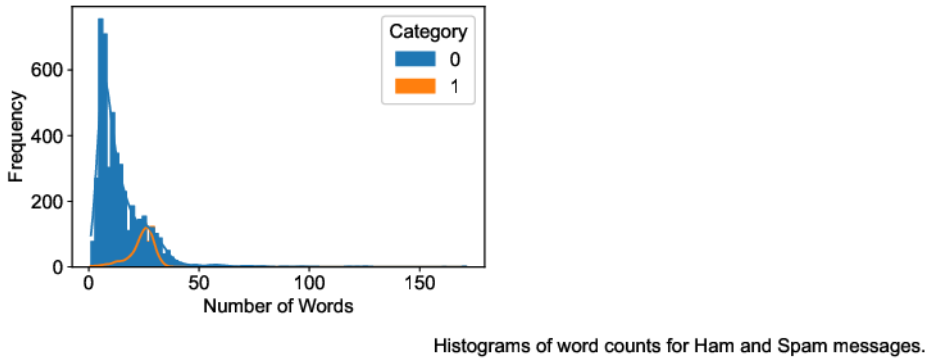


Figure 4: Word count distribution

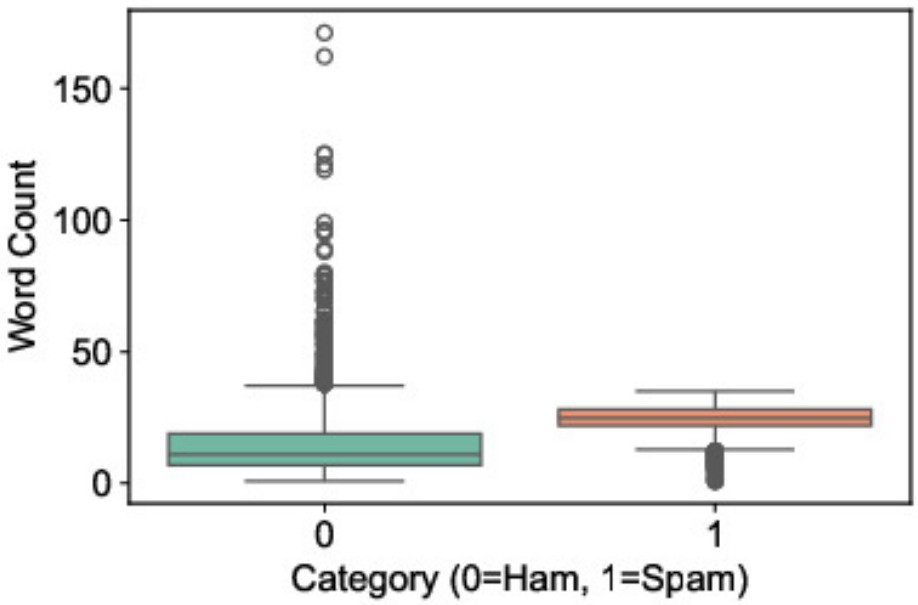


Figure 5: Category distribution

Diabetes Dataset

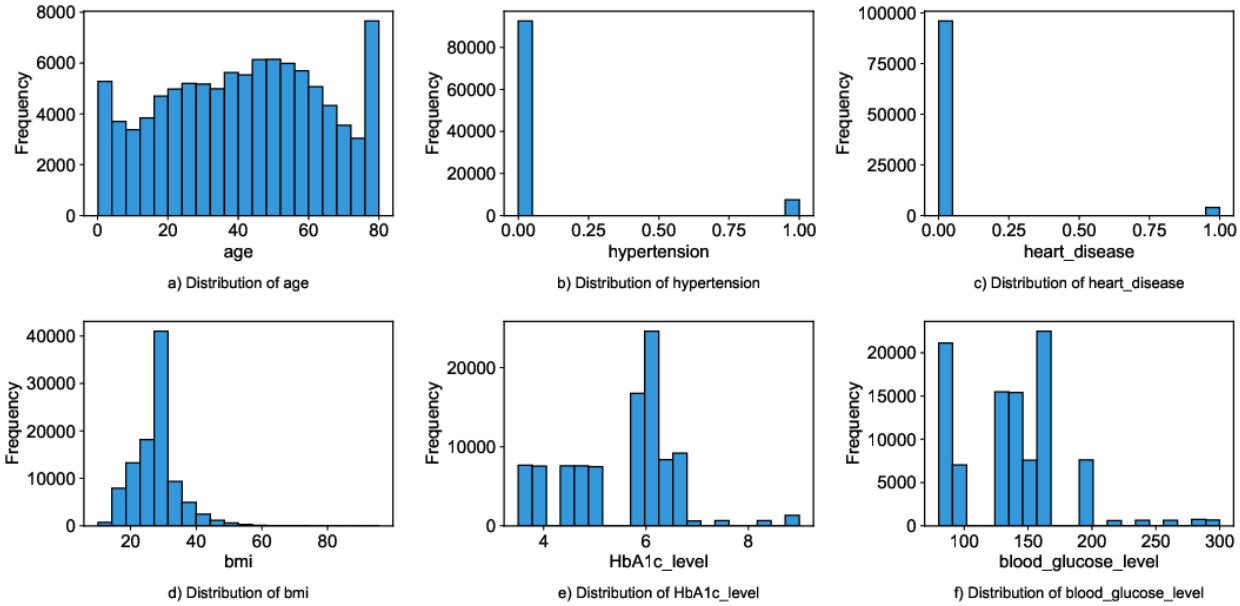


Figure: Histograms showing the distribution of numeric features in the diabetes dataset.

Figure 6: Distribution of numerical features using Histograms .

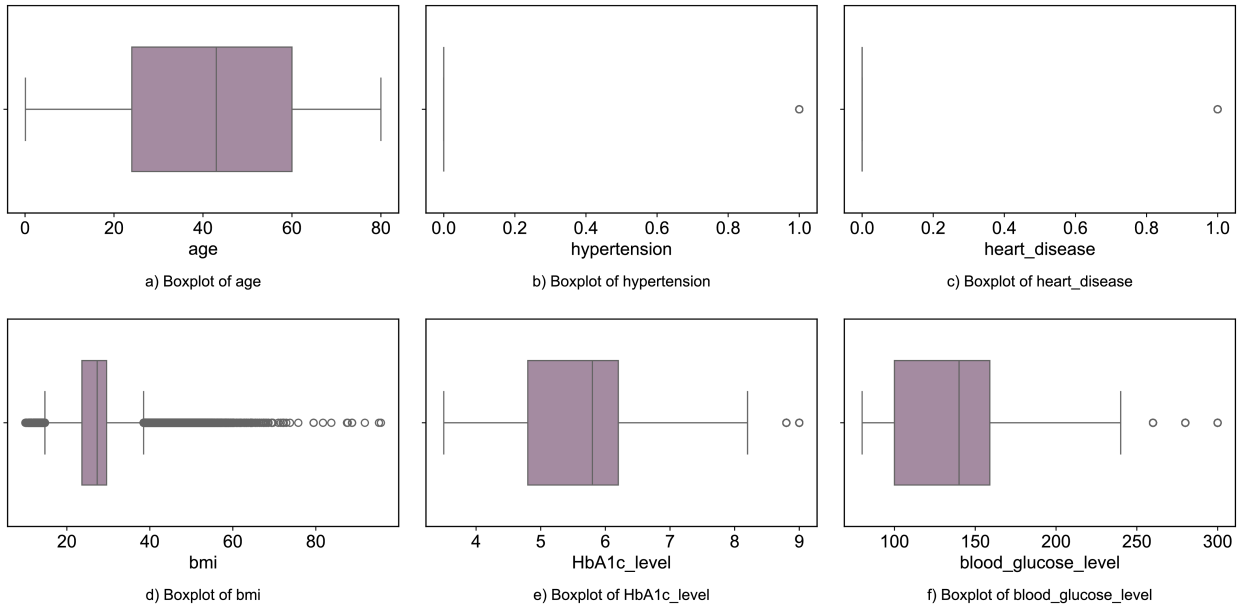


Figure: Boxplots identifying outliers and distribution of numeric health metrics.

Figure 7: Boxplot visualisation

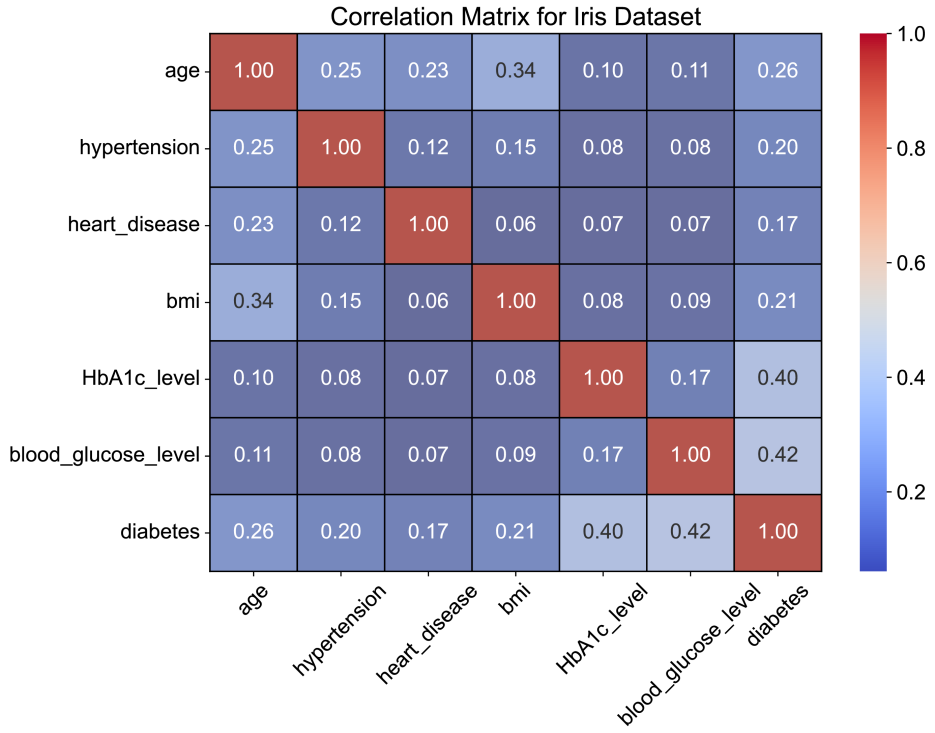


Figure: Correlation Heatmap of clinical and demographic features.

Figure 8: Correlation Heatmap

Iris Dataset

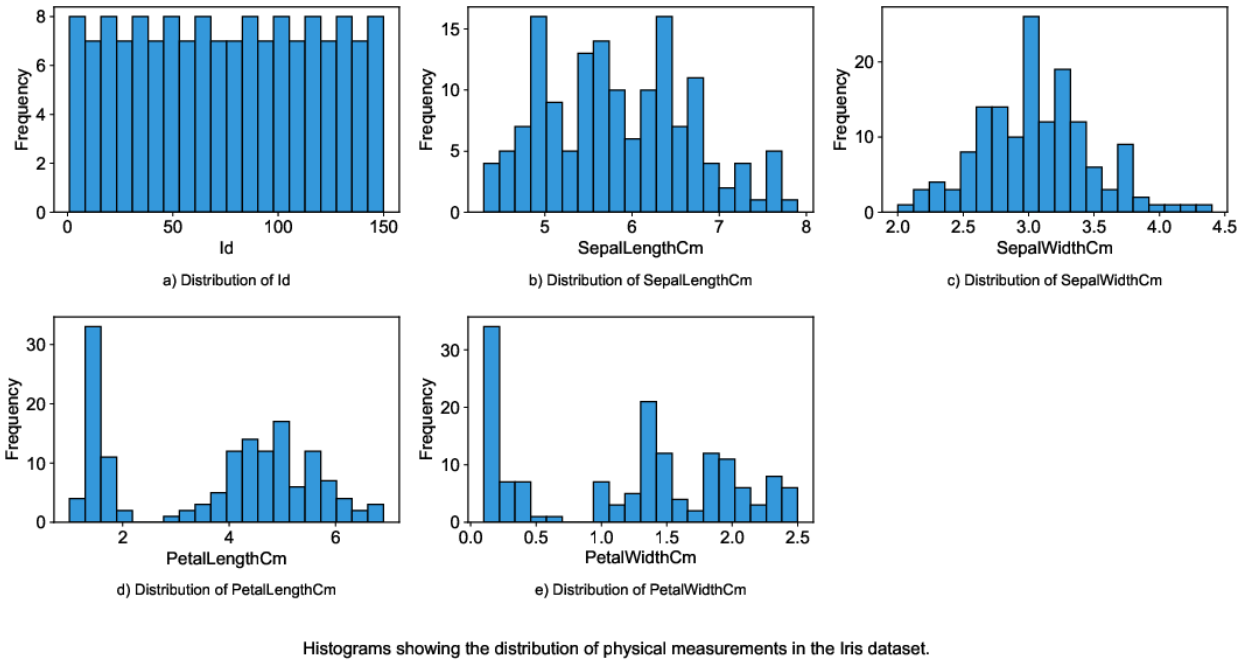


Figure 9: Distribution of numerical features using Histograms .

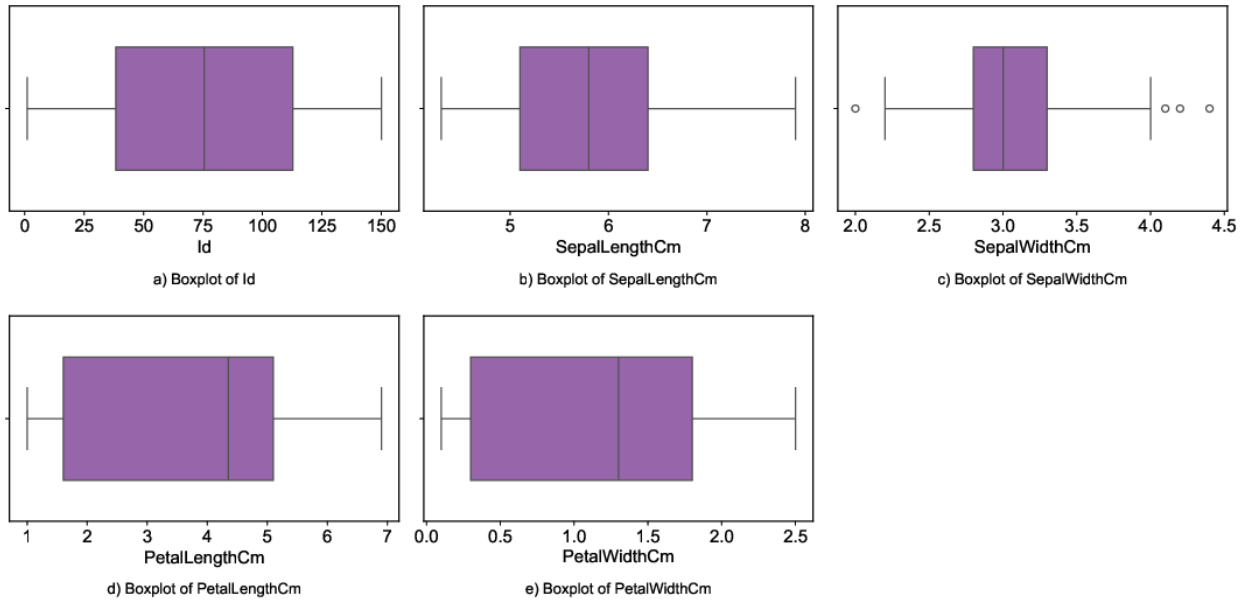


Figure 10: Boxplot visualisation

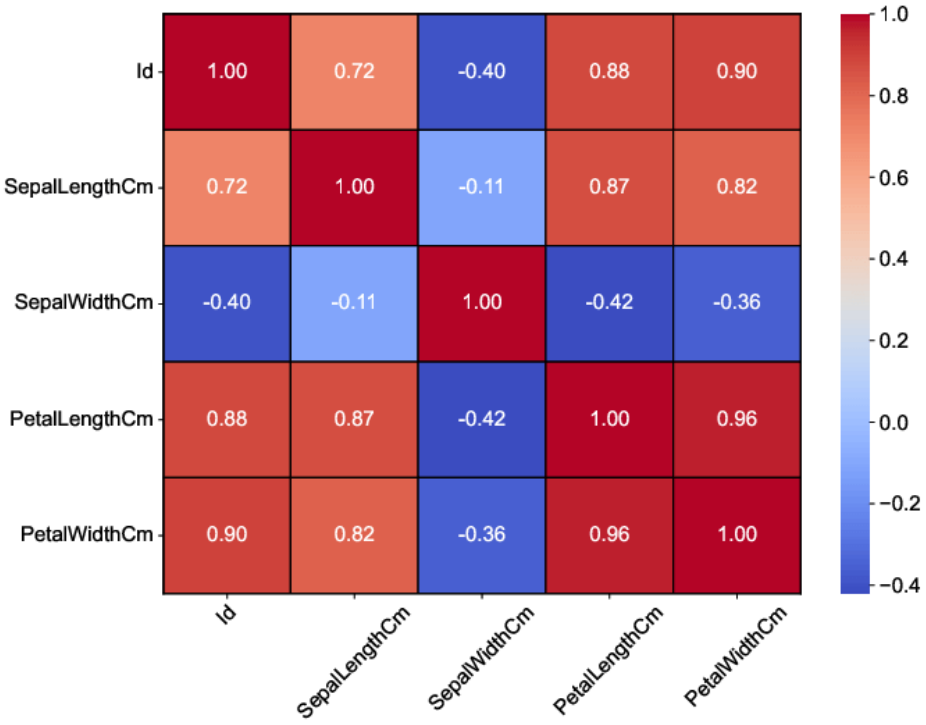


Figure 11: Correlation Heatmap

Dataset	**Type of ML Task**	**Suitable ML Algorithm**
Iris Dataset	Multi-class Classification	k-Nearest Neighbors (k-NN)
Loan Amount Prediction	Regression	Linear Regression
Predicting Diabetes	Binary Classification	Logistic Regression, Support Vector Machine (SVM)
Classification of Email Spam	Binary Classification (NLP)	Naive Bayes, SVM
Handwritten Character Recognition / MNIST	Multi-class Image Classification	Convolutional Neural Networks (CNN)

Learning Practices:

- Interpret dataset structure: Learn to inspect shape, info, and missing values.
- Visualize distributions: Gain skills in plotting histograms, boxplots, and correlation heatmaps.
- Identify class balance: Understand how label distribution affects model performance.
- Spot feature relationships: Use pairplots and correlation matrices to detect predictive variables.
- Apply statistical tests: Use ANOVA F-test and correlation filtering.
- Leverage model-based importance: Interpret Random Forest feature importances.
- Perform dimensionality reduction: Use PCA for visualization and clustering.