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	Digital ID: 2310298

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

Name: R. Nithin Josh Albert
Reg.No: 3122235001090

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`)
- Standard Scaler(from sklearn.preprocessing)
- Math (Standard Library)

Mathematical and Theoretical description of the objectives performed:

- **Purpose:** Perform Exploratory Data Analysis (EDA) to understand data behavior, summarize value distributions, identify anomalies, evaluate class proportions, and uncover inter-feature relationships.
- **Summary Statistics**
  - Mean value computed as: $\mu = \frac{1}{n}\sum x_i$
  - Sample variance expressing dispersion: $s^2 = \frac{1}{n-1}\sum(x_i - \mu)^2$
  - Quantile measures (Median, $Q1$, $Q3$) used to analyze spread and construct boxplots.
- **Distribution Analysis Using Histograms**
  - Frequency-based approximation of empirical distributions through binning to reveal shape and central tendencies.

  - Kernel Density Estimation (KDE) overlaid to provide a smooth estimate of the underlying distribution:

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

    where K denotes the kernel function and h represents the bandwidth parameter.

- **Boxplot Visualization**
  - Displays median and quartiles ($Q1$, $Q3$) with the interquartile range defined as $IQR = Q3 - Q1$.
  - Whiskers extend to data points within $[Q1 - 1.5 \cdot IQR,\ Q3 + 1.5 \cdot IQR]$, while values beyond this range are flagged as outliers.
- **Feature Relationship Plots**
  - Pairwise scatter visualizations are used to detect trends, separability between classes, and nonlinear dependencies.
  - Diagonal plots present KDEs or histograms to highlight individual feature distributions.
- **Correlation Visualization**
  - Linear association between variables quantified using the Pearson correlation coefficient:

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

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

  - Coefficient values lie in the interval $[-1, 1]$, indicating both magnitude and direction of linear dependence.
- **Categorical Data Visualization**
  - Count-based bar plots are used to examine category frequencies, with optional class-wise grouping to study imbalance patterns.
- **Image Data Exploration**
  - Histograms or KDE plots of image dimensions (height and width) are used to determine consistency and preprocessing requirements.
  - Representative image samples are displayed to qualitatively assess visual characteristics.
- **Handling Missing Data**
  - Column-wise inspection of missing entries to inform strategies such as imputation or exclusion.
- **Analytical Objectives**
  - Detect skewed distributions, multiple modes, anomalous observations, correlated features, and class imbalance in order to support informed preprocessing and model selection.

Results and Discussions:

0.1 Loan Amount Prediction

	person_age	person_gender	person_education	person_income	person_emp_exp	person_home_ownership	loan_amnt	loan_in
0	22.0	female	Master	71948.0	0	RENT	35000.0	PERSC
1	21.0	female	High School	12282.0	0	OWN	1000.0	EDUCA
2	25.0	female	High School	12438.0	3	MORTGAGE	5500.0	MED
3	23.0	female	Bachelor	79753.0	0	RENT	35000.0	MED
4	24.0	male	Master	66135.0	1	RENT	35000.0	MED
...	...	...	...	...	...	...	...	
44995	27.0	male	Associate	47971.0	6	RENT	15000.0	MED
44996	37.0	female	Associate	65800.0	17	RENT	9000.0	HOMEIMPROVEM
44997	33.0	male	Associate	56942.0	7	RENT	2771.0	DEBTCONSOLIDA
44998	29.0	male	Bachelor	33164.0	4	RENT	12000.0	EDUCA
44999	24.0	male	High School	51609.0	1	RENT	6665.0	DEBTCONSOLIDA

45000 rows × 14 columns

Figure 1: Dataset Columns

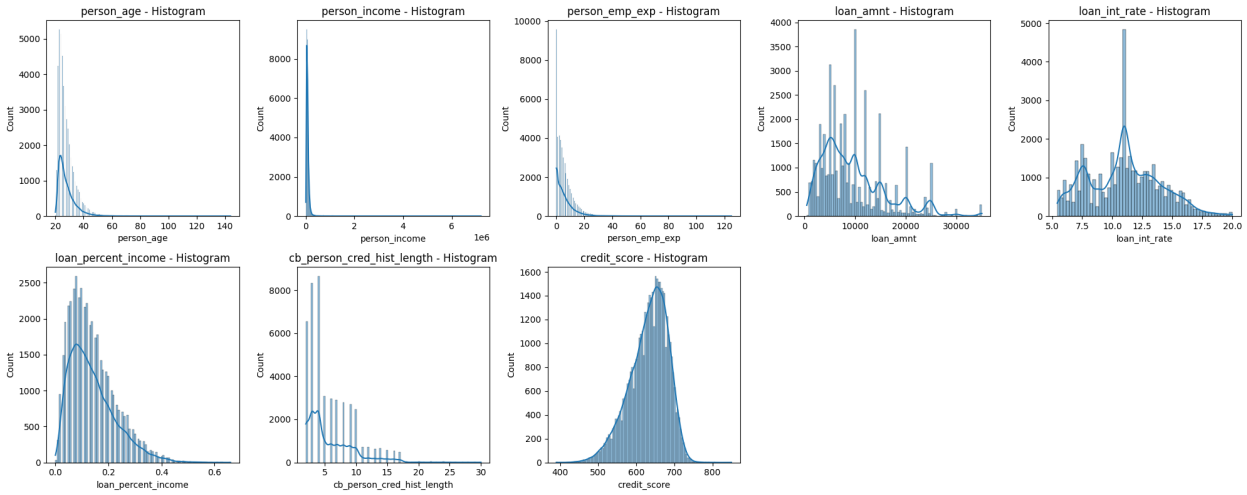

Figure 2: Loan Amount Distribution(histogram plot)

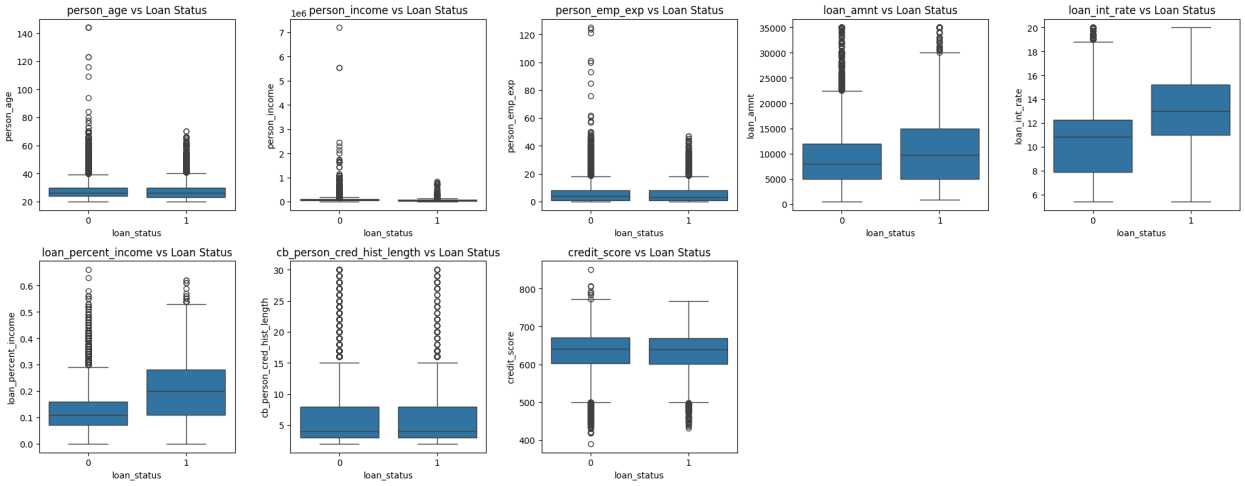

Figure 3: Boxplot Distribution

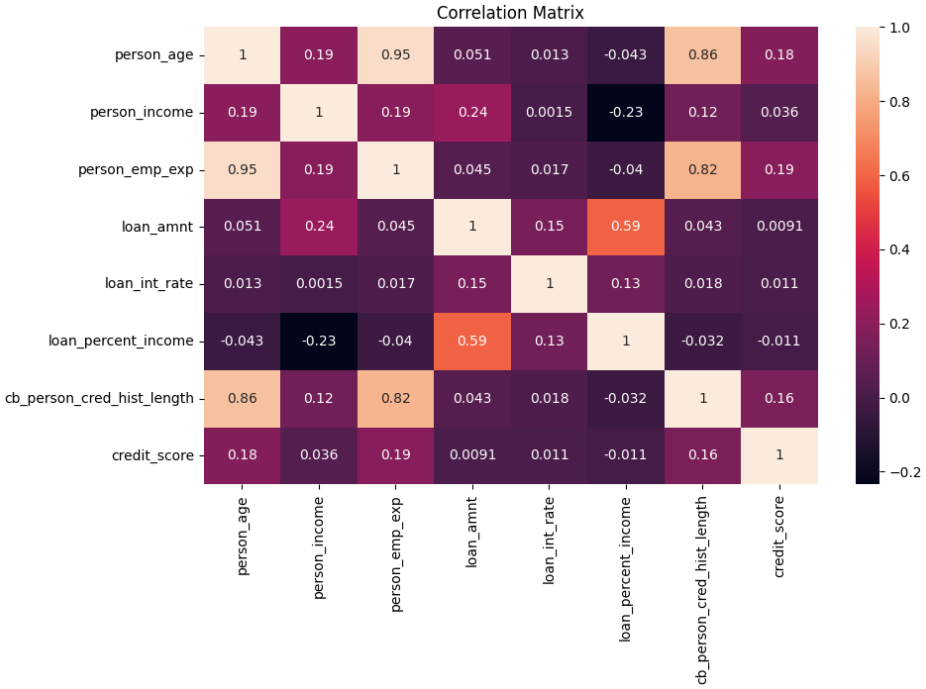

Figure 4: Correlation Matrix

0.2 Predicting Diabetes

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0
...	...	...	...	...	...	...	...	...	...
99995	Female	80.0	0	0	No Info	27.32	6.2	90	0
99996	Female	2.0	0	0	No Info	17.37	6.5	100	0
99997	Male	66.0	0	0	former	27.83	5.7	155	0
99998	Female	24.0	0	0	never	35.42	4.0	100	0
99999	Female	57.0	0	0	current	22.43	6.6	90	0

100000 rows × 9 columns

Figure 5: Dataset Columns

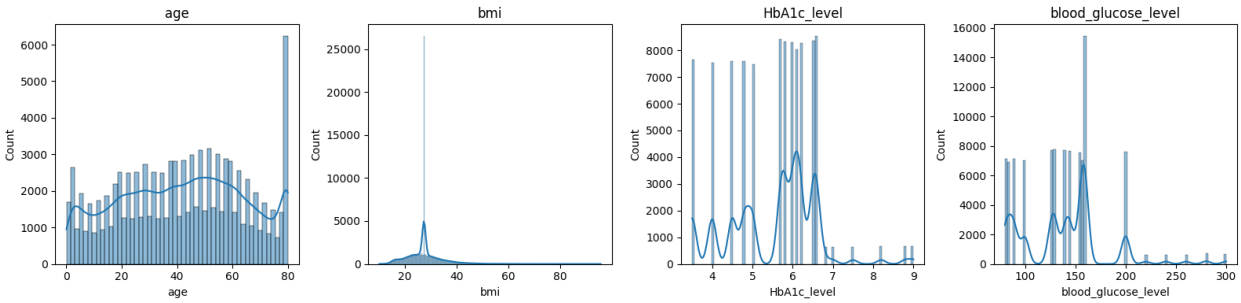

Figure 6: Histogram Distribution

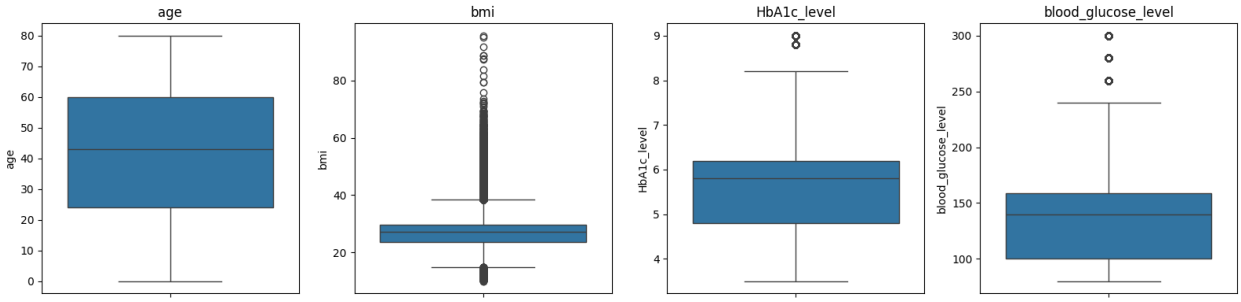

Figure 7: Boxplot distribution

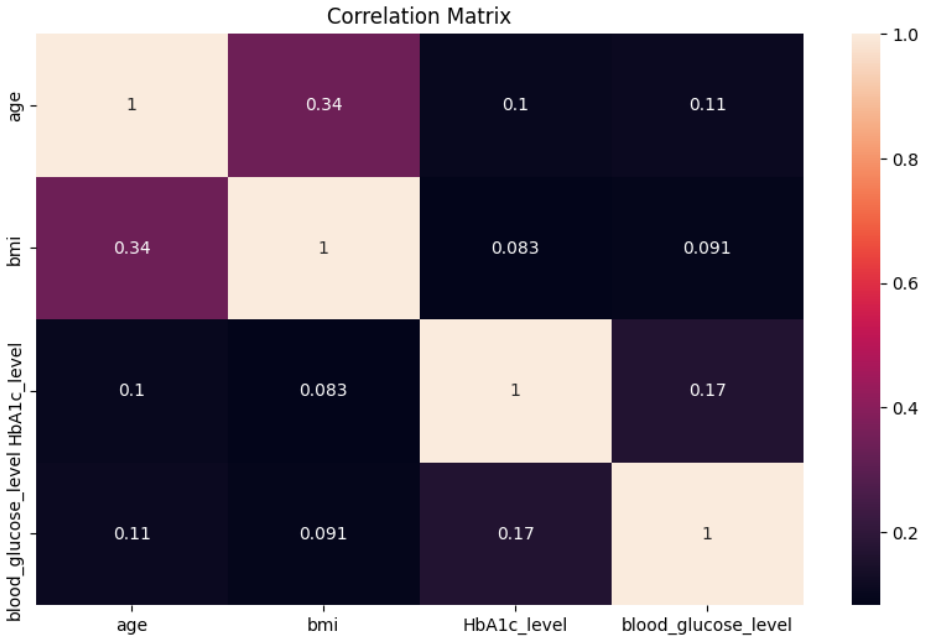

Figure 8: Correlation Matrix

0.3 Classification of Email Spam

	Unnamed: 0	label	text	label_num
0	605	ham	Subject: enron methanol ; meter # : 988291\r\n...	0
1	2349	ham	Subject: hpl nom for january 9 , 2001\r\n(see...	0
2	3624	ham	Subject: neon retreat\r\nho ho ho , we ' re ar...	0
3	4685	spam	Subject: photoshop , windows , office . cheap ...	1
4	2030	ham	Subject: re : indian springs\r\nthis deal is t...	0
...	...	...	...	...
5166	1518	ham	Subject: put the 10 on the ft\r\nthe transport...	0
5167	404	ham	Subject: 3 / 4 / 2000 and following noms\r\nhp...	0
5168	2933	ham	Subject: calpine daily gas nomination\r\n>\r\n...	0
5169	1409	ham	Subject: industrial worksheets for august 2000...	0
5170	4807	spam	Subject: important online banking alert\r\ndea...	1

Figure 9: dataset columns

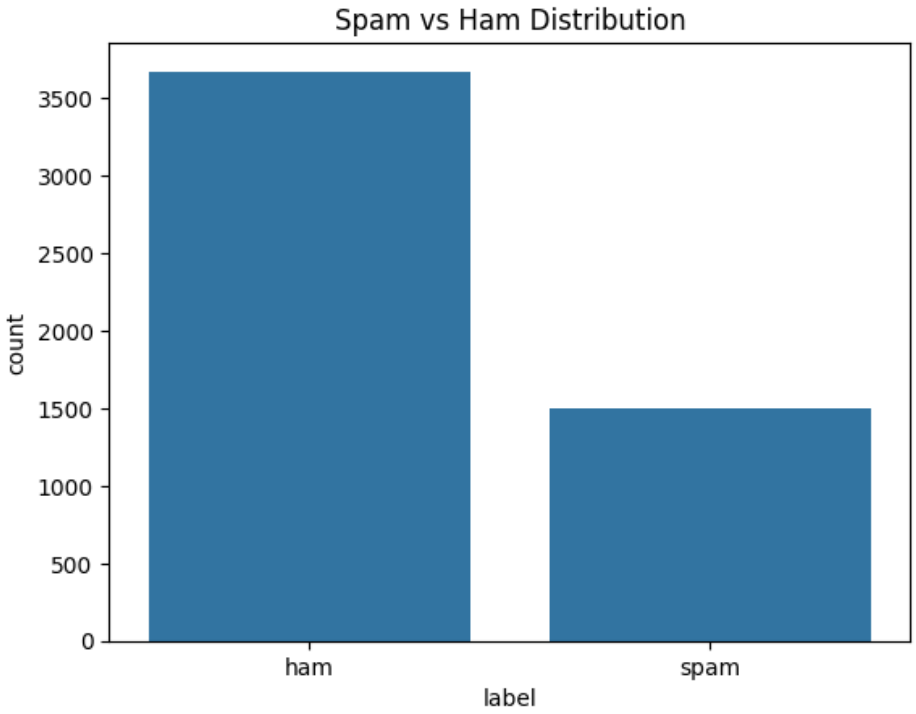

Figure 10: spam vs ham distribution

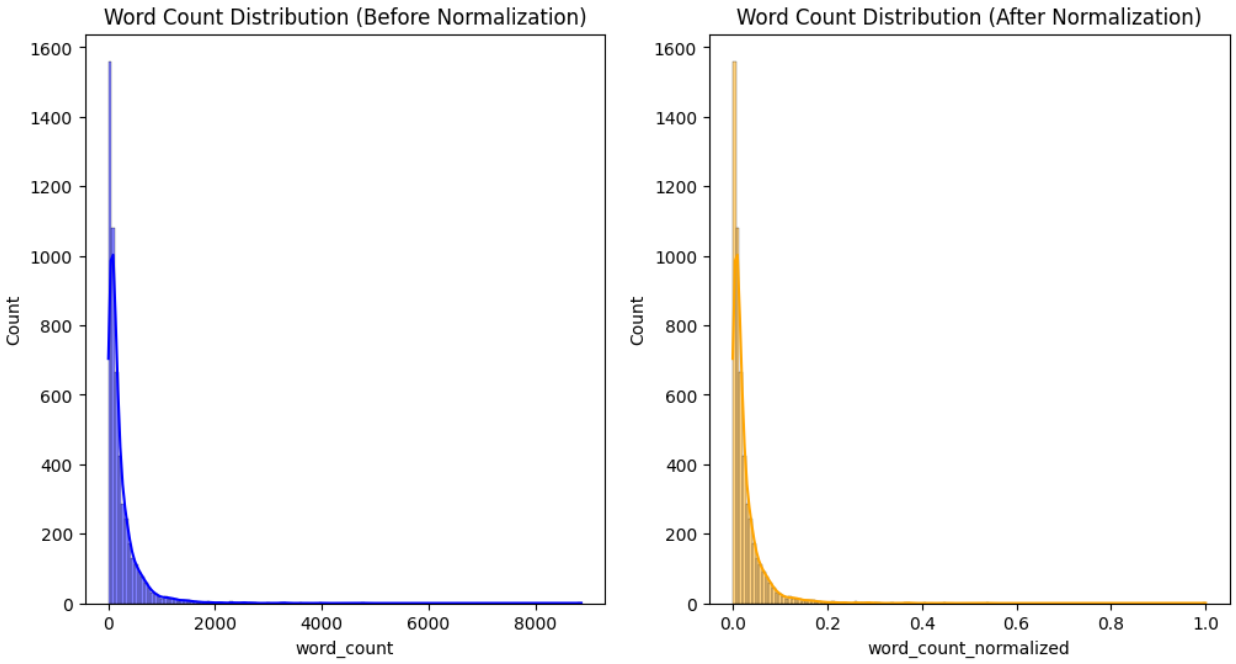

Figure 11: word count distribution(Normalisation)

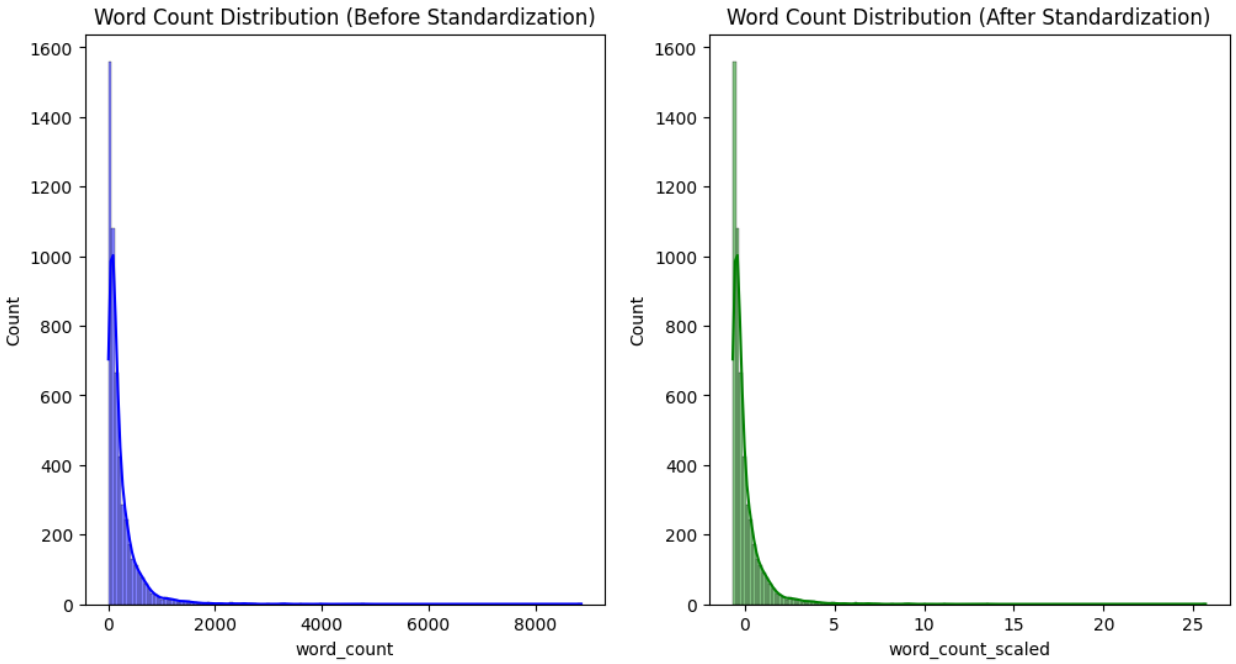

Figure 12: word count distribution(Standardisation)

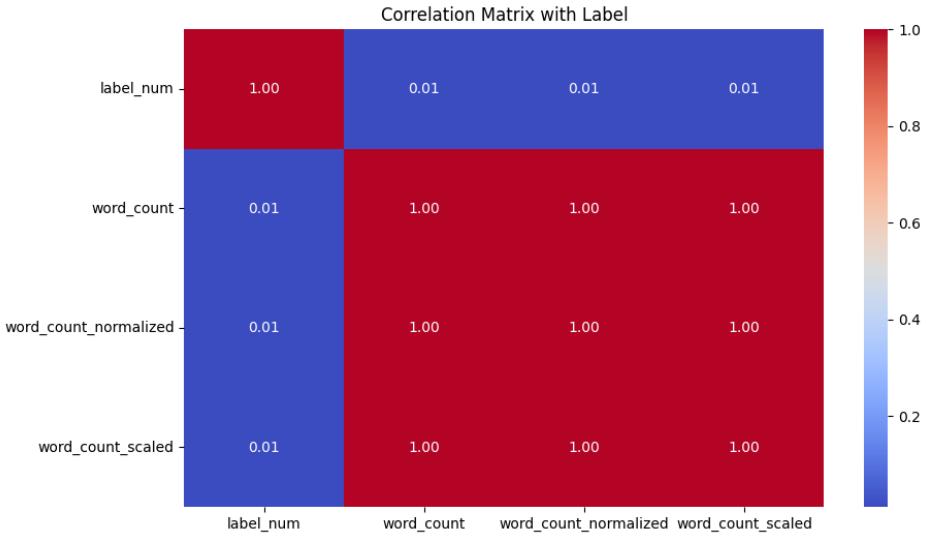

Figure 13: Correlation Matrix

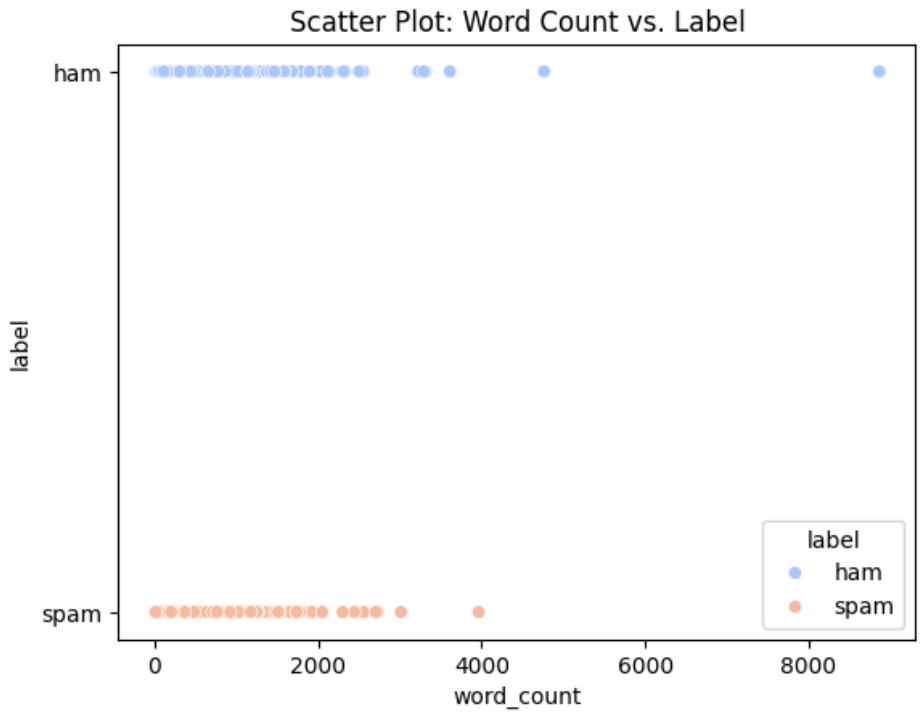

Figure 14: Correlation Matrix

0.4 Handwritten Character Recognition (MNIST)

Figure 15: sample images

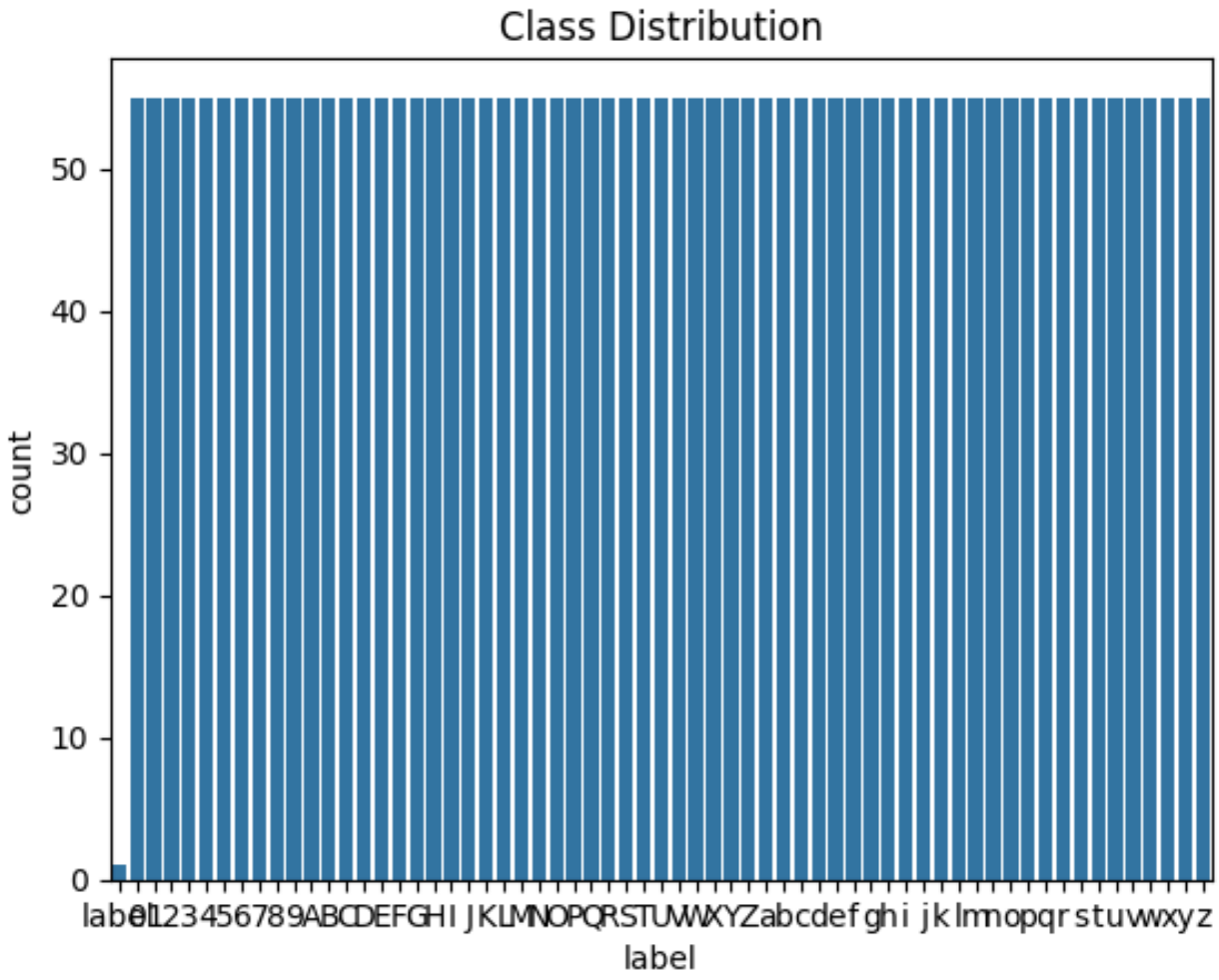

Figure 16: Digit Distribution

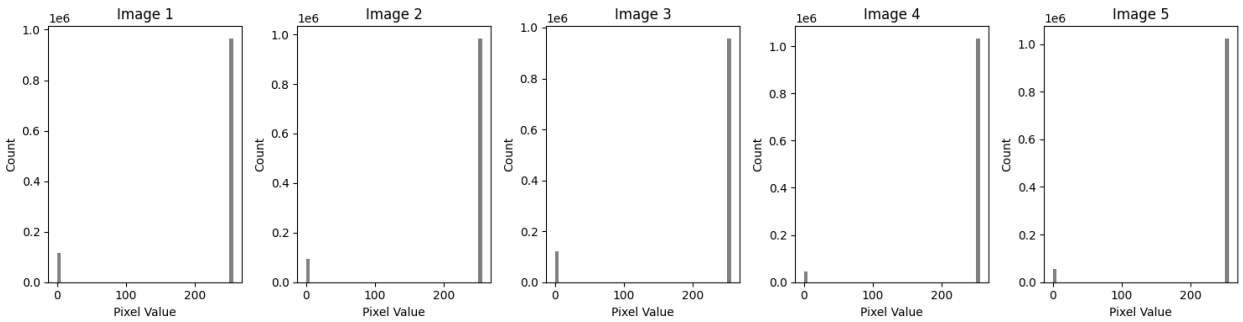

Figure 17: Pixel Intensity Analysis

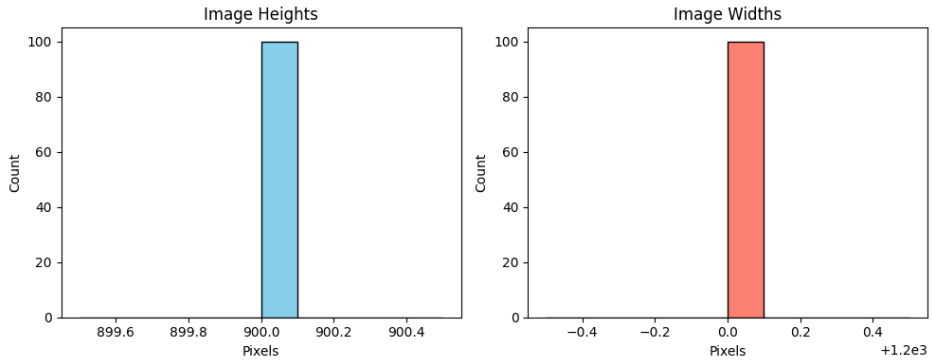

Figure 18: Height weight distribution

0.5 Iris Dataset

	sepal_length	sepal_width	petal_length	petal_width	species
0	NaN	NaN	NaN	NaN	Species
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Figure 19: Dataset Columns

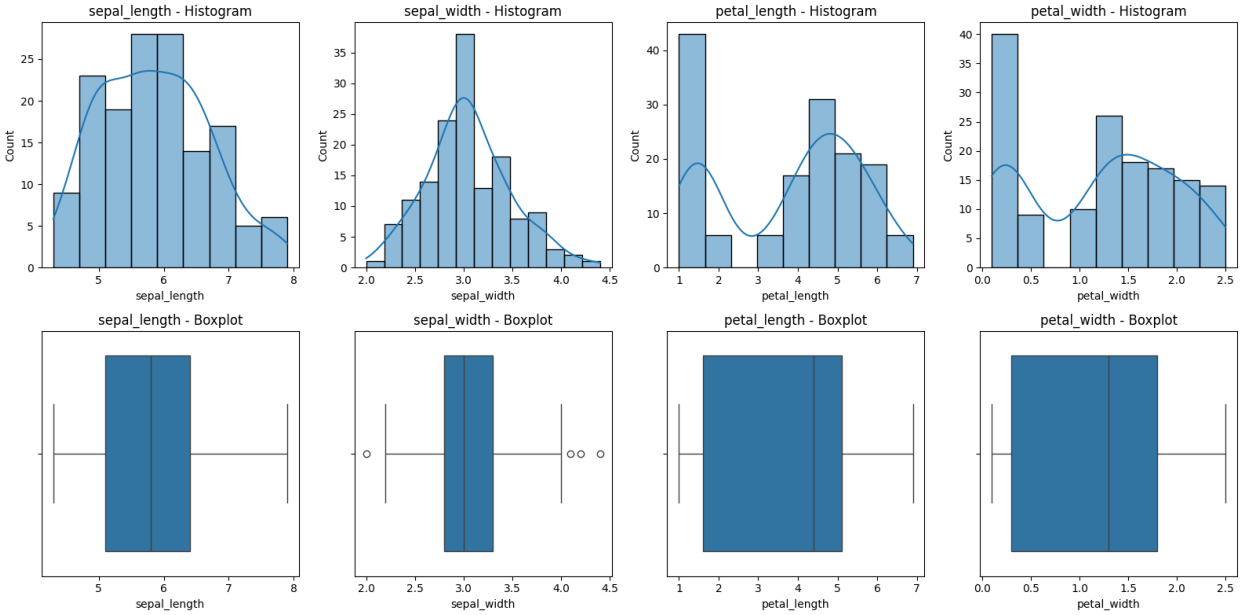

Figure 20: Histogram and Boxplot Distribution

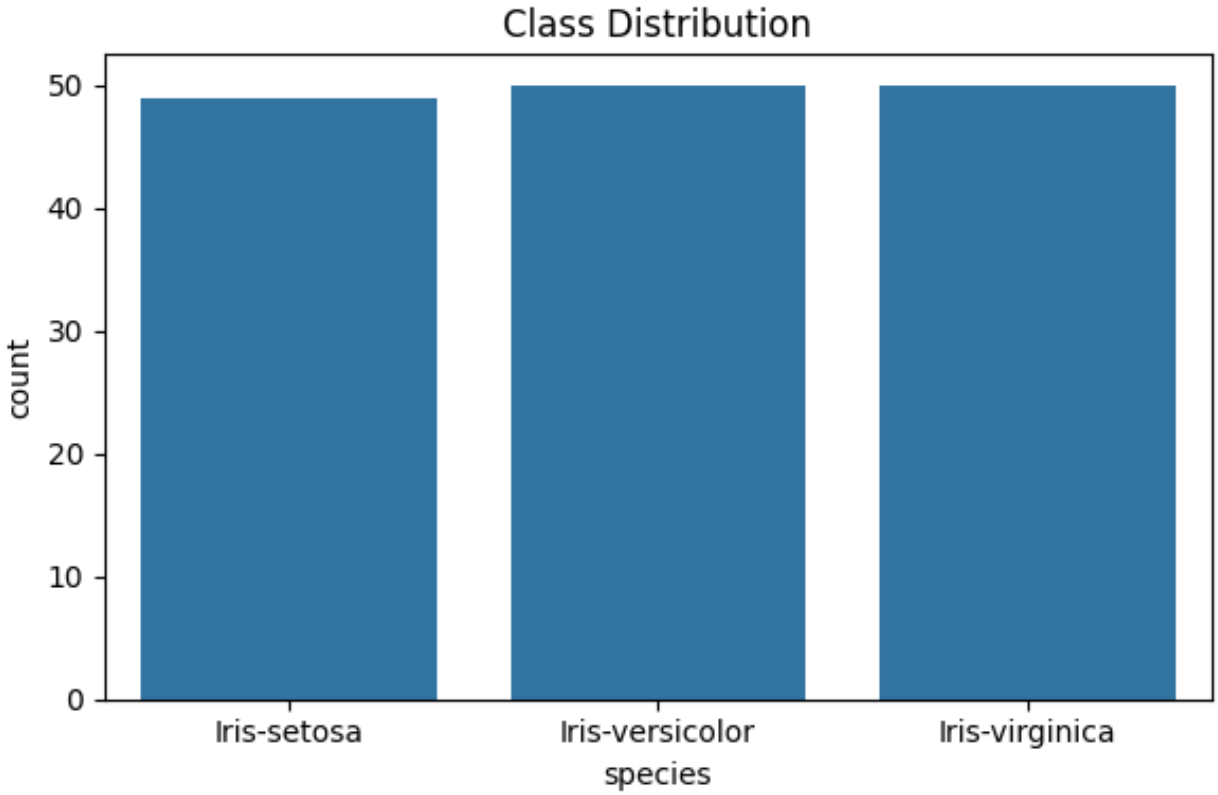

Figure 21: class distribution

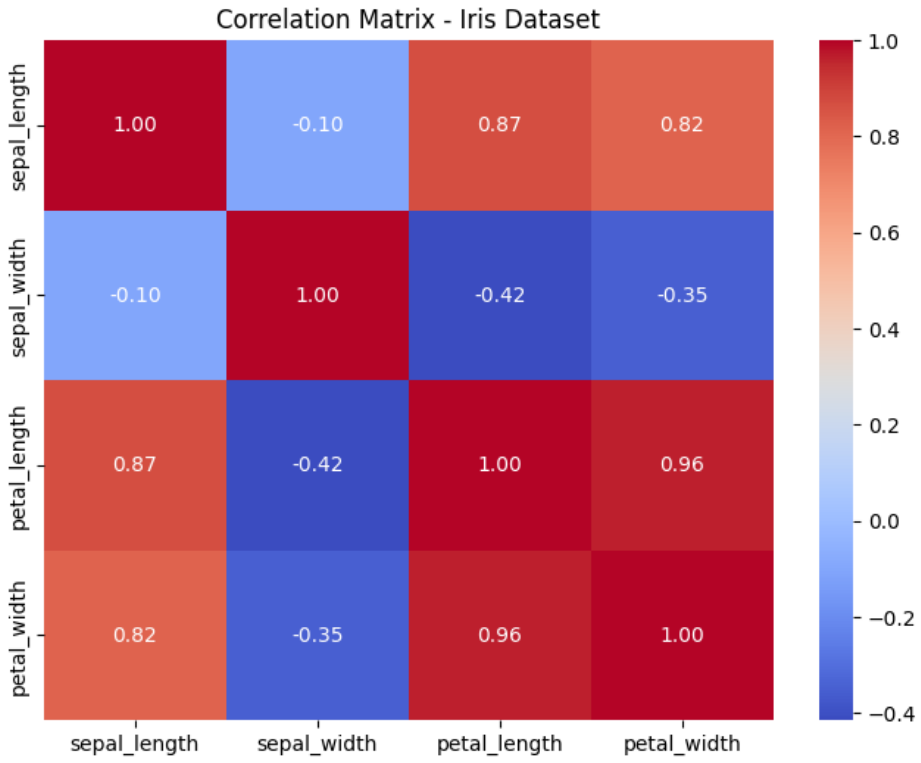

Figure 22: Correlation Matrix

Dataset	**Type of ML Task**	**Feature Selection Technique**	**Suitable ML Algorithm**
Iris Dataset	Multi-class Classification	Correlation Matrix / ANOVA F-value	k-Nearest Neighbors (k-NN), Decision Trees
Loan Amount Prediction	Regression	Recursive Feature Elimination (RFE) / Pearson Correlation	Linear Regression, Random Forest Regressor
Predicting Diabetes	Binary Classification	Chi-Square Test / SelectKBest	Logistic Regression, Support Vector Machine (SVM)
Classification of Email Spam	Binary Classification (NLP)	Information Gain / Chi-Square (on word vectors)	Naive Bayes, SVM
Handwritten Character Recognition / MNIST	Multi-class Image Classification	Principal Component Analysis (PCA)	Convolutional Neural Networks (CNN), SVM

Learning Practices:

- Examine dataset organization: Understand the structure by analyzing dimensions, data types, and missing entries.
- Explore data distributions: Develop the ability to visualize patterns using histograms, box plots, and correlation maps.
- Assess class proportions: Analyze label balance and its influence on model behavior and outcomes.
- Analyze feature interactions: Investigate relationships among variables through pair plots and correlation analysis.
- Conduct statistical evaluation: Apply techniques such as ANOVA F-tests and correlation-based feature selection.
- Utilize model-driven insights: Interpret feature importance scores obtained from Random Forest models.
- Reduce dimensional complexity: Implement PCA to support visualization and uncover latent data patterns.