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: Comprehensive Machine Learning Workflow and Exploratory Data Analysis

Name: Monesh M
Reg.No: 3122235001084
Class: CSE-B

1 Aim

To implement a comprehensive machine learning pipeline on diverse datasets (Structured, Text, and Translation) using Python's data science ecosystem. The objectives include performing extensive Exploratory Data Analysis (EDA) to derive insights, executing data preprocessing steps, and training suitable models for classification and sequence-to-sequence tasks.

2 Introduction and Libraries

This experiment explores various domains of Machine Learning, including binary classification, multiclass classification, and Neural Machine Translation (NMT). The following libraries are utilized:

- **NumPy & Pandas:** For numerical operations and data manipulation.
- **Matplotlib & Seaborn:** For high-quality data visualization and statistical plots.
- **Scikit-Learn:** For standard ML algorithms, feature extraction (TF-IDF), and preprocessing.
- **TensorFlow/Keras:** Specifically for building the Deep Learning-based Seq2Seq model for translation.

3 Theoretical Background: Exploratory Data Analysis (EDA)

EDA is a critical step to understand data patterns, anomalies, and relationships before modeling.

- **Histogram Distribution:** Visualizes the frequency distribution of a variable. It helps identify skewness ($Skew = \frac{3(Mean-Median)}{StdDev}$) and the modality of the data.
- **Box Plot (Outlier Detection):** Represents the five-number summary (Min, Q1, Median, Q3, Max). Points beyond $1.5 \times IQR$ are flagged as potential outliers.

- **Correlation Heatmap:** Uses the Pearson Correlation coefficient (r) to measure linear relationships between numeric features.
- **Pair Plot:** A matrix of scatter plots showing pairwise relationships, colored by class labels to visually inspect linear separability.

4 Overall Exploratory Data Analysis (EDA)

The following sections detail the comprehensive EDA conducted across all five datasets. This phase was fundamental in verifying data quality and informing model architecture.

4.1 Iris Dataset EDA

- **Feature Distribution:** Sepal and Petal measurements show distinct ranges. Petal length and width exhibit a bimodal distribution, indicating clear separability for at least one class.
- **Correlations:** Extremely high positive correlation (> 0.9) observed between Petal length and Petal width.
- **Class Separation:** Pair plots confirm that *Iris-setosa* is linearly separable from the other two species.

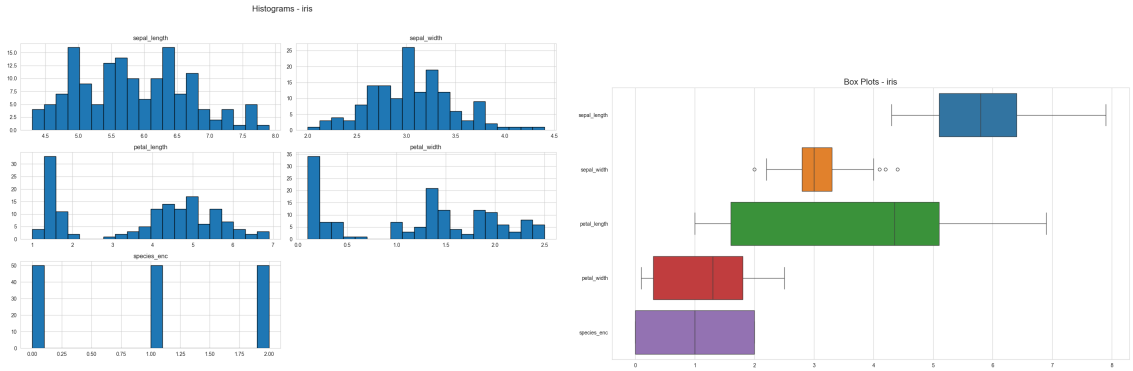

Figure 1: Iris Dataset: Histograms and Box Plots

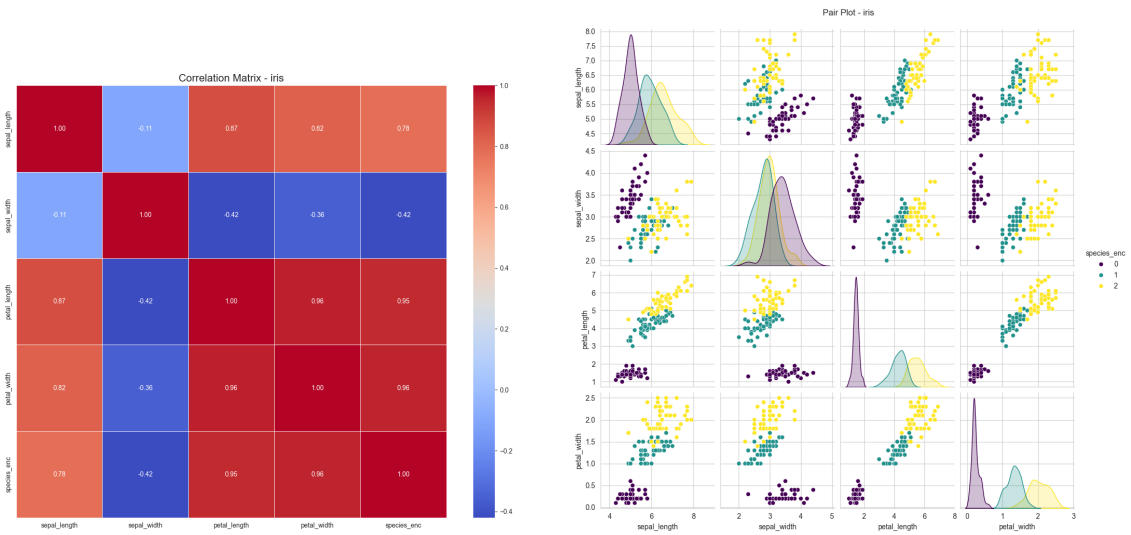

Figure 2: Iris Dataset: Correlation Heatmap and Pair Plot

4.2 Loan Eligibility EDA

- **Handling Skewness:** Applicant Income and Co-applicant Income were highly right-skewed, exhibiting long tails. Log transformations or robust scaling were considered.
- **Outlier Analysis:** Box plots identified significant outliers in the `LoanAmount` and `ApplicantIncome` columns, representing high-net-worth individuals.
- **Categorical Imbalance:** Gender and Marital Status showed moderate imbalance, which was accounted for during stratified splitting.

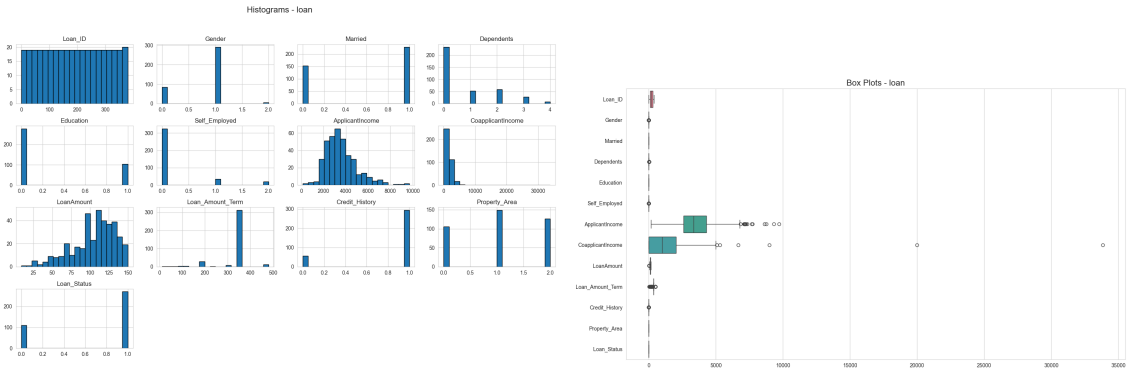

Figure 3: Loan Dataset: Histograms and Box Plots

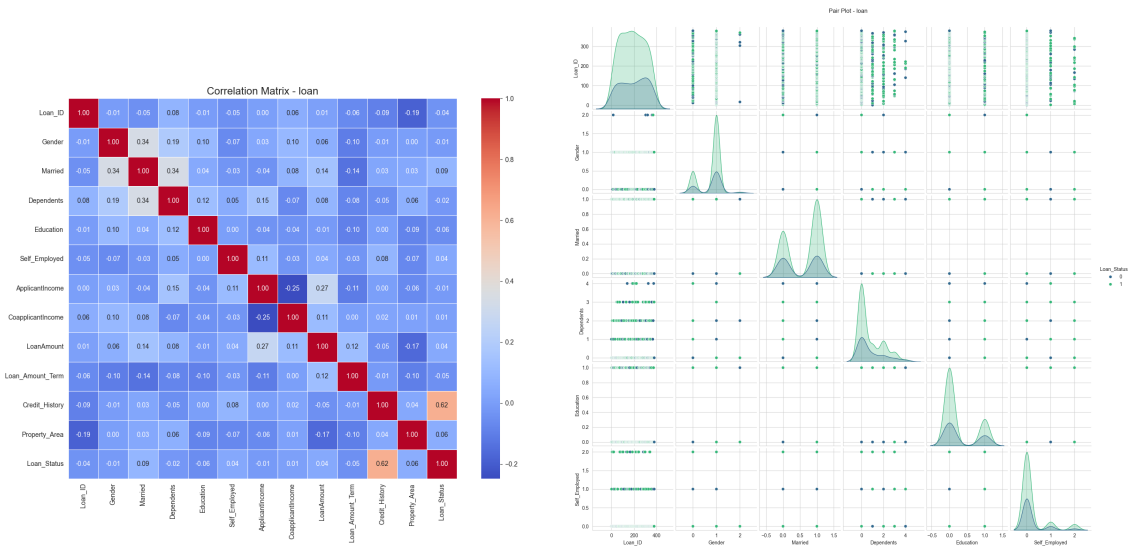

Figure 4: Loan Dataset: Correlation Heatmap and Pair Plot

4.3 Diabetes Prediction EDA

- **Bivariate Analysis:** Scatter plots of Glucose vs. BMI show a positive trend with respect to the outcome (Diabetes positive).
- **Feature Interaction:** A heatmap revealed a notable correlation between Age and pregnancies, and between BMI and skin thickness.
- **Histogram Insights:** Glucose levels follow a near-normal distribution centered around 100-120 mg/dL for non-diabetic individuals.

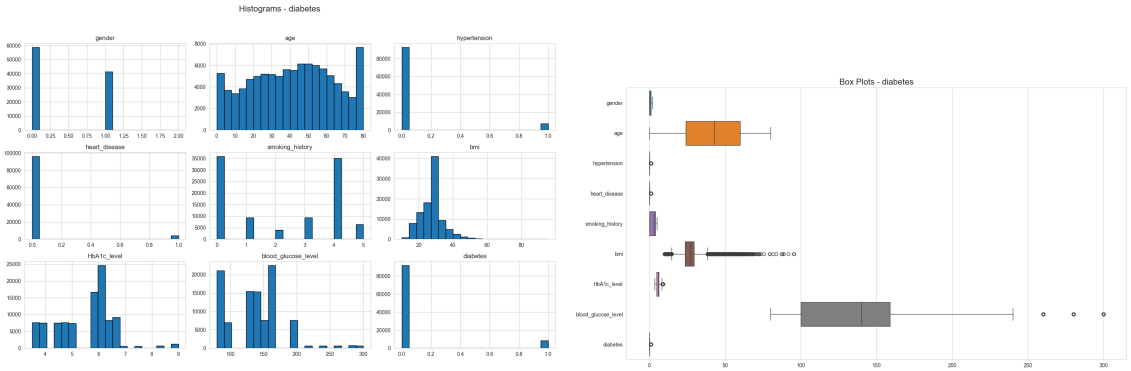

Figure 5: Diabetes Dataset: Histograms and Box Plots

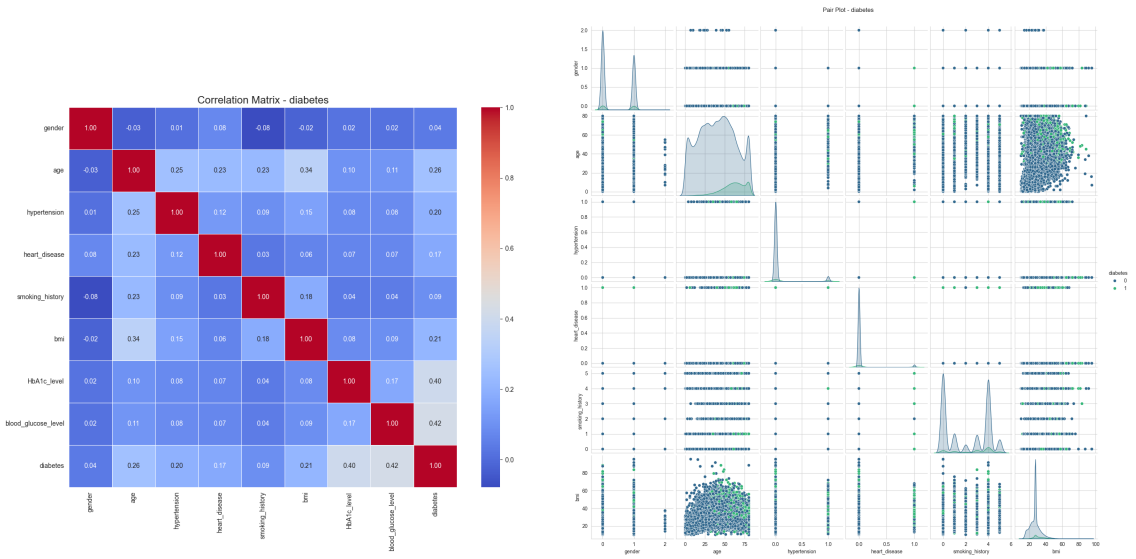

Figure 6: Diabetes Dataset: Correlation Heatmap and Pair Plot

4.4 Spam Email EDA

- **Message Length:** EDA revealed that 'Spam' messages generally have a higher character count compared to 'Ham' (normal) messages.
- **Token Frequency:** Word clouds and frequency distributions showed high occurrences of monetary and urgency-related keywords in spam.
- **Class Balance:** The dataset is imbalanced (more ham than spam), leading to the choice of the F1-score metric.

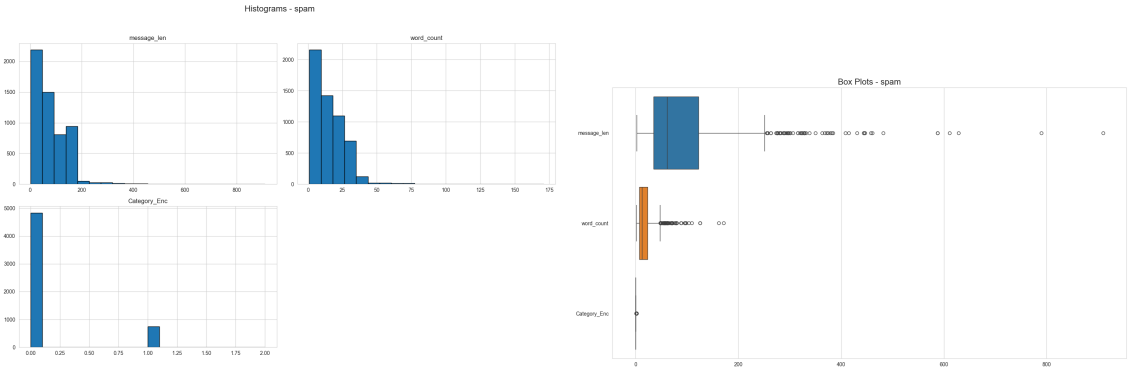

Figure 7: Spam Dataset: Histograms and Box Plots

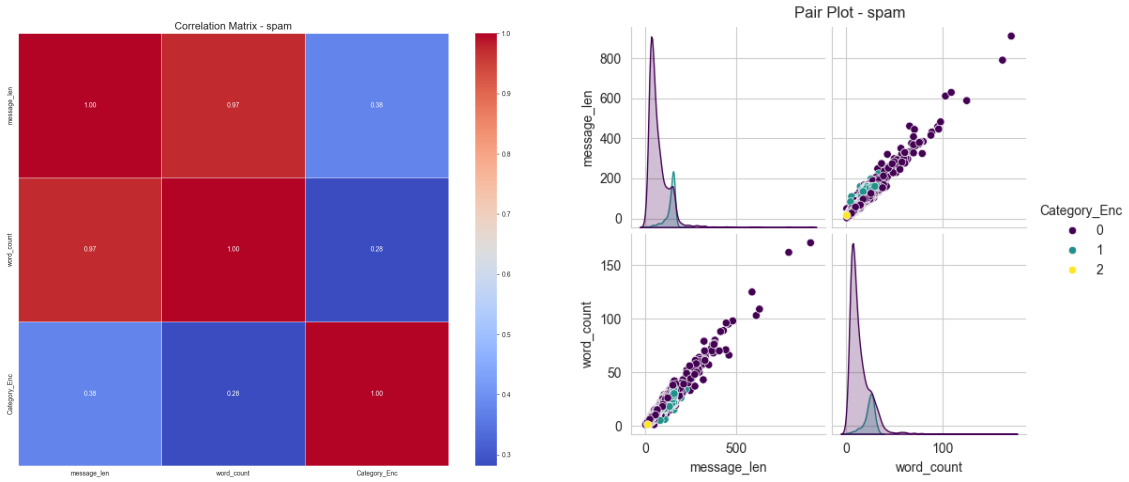

Figure 8: Spam Dataset: Correlation Heatmap and Pair Plot

4.5 MNIST / Digits Image EDA

- **Intensity Distribution:** Visualized using histograms of pixel intensities, showing most pixels are near-zero (background) with sharp peaks near 255 (stroke pixels).
- **Spatial Verification:** Visualized a 5×5 grid of digit samples to verify correct labeling and check for variations in handwriting styles.
- **Dimensionality:** Each 28×28 image was flattened, and the variance across features was assessed to ensure high information content.

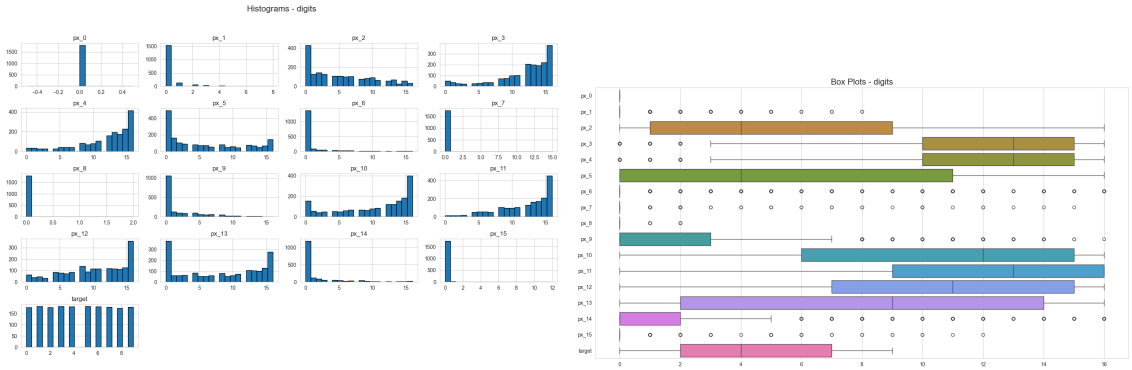

Figure 9: Digits Dataset: Histograms and Box Plots

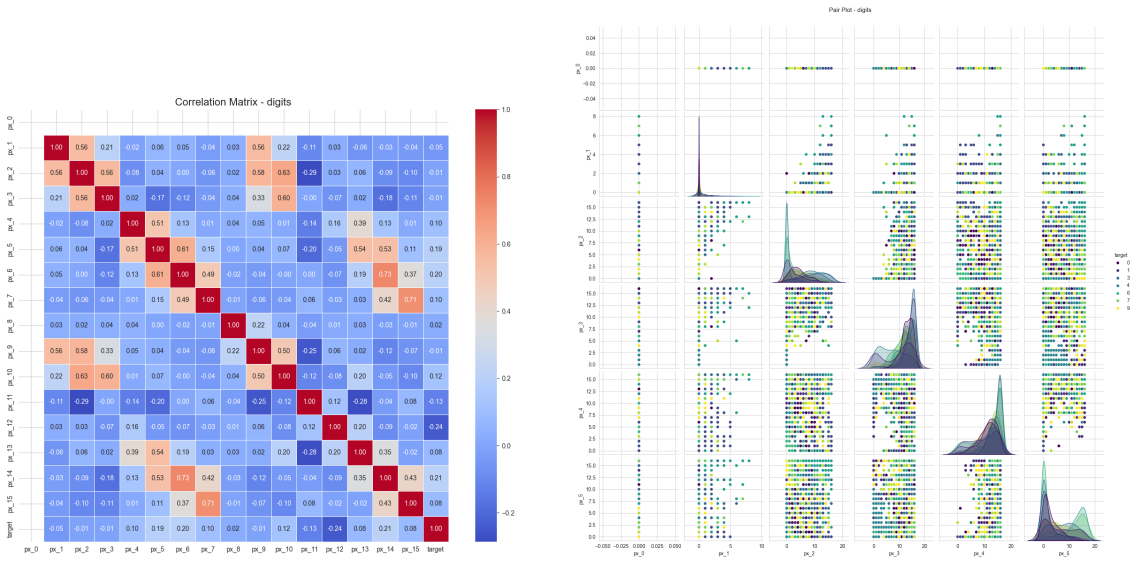

Figure 10: Digits Dataset: Correlation Heatmap and Pair Plot

5 Domain Methodology

5.1 Loan Prediction

A **Random Forest Classifier** was used to predict loan status after handling missing values and categorical mapping for features like `Married` and `Education`.

5.2 Diabetes Prediction

Implemented **Logistic Regression** to determine diabetes probability, focusing on key contributors like `Glucose` and `BMI` as identified during the correlation analysis.

5.3 Sequence-to-Sequence (NMT)

Developed an **LSTM-based Encoder-Decoder** model to translate English phrases to Tamil, leveraging a context vector to transfer linguistic meaning across time steps.

6 Conclusion

The integration of systematic EDA with advanced machine learning models provided a holistic workflow for data science projects. EDA not only highlighted the inherent structure of the five datasets but also provided the necessary justification for pre-processing techniques and model selection.