A Comparative Study of Machine Learning Workflows on Multiple Datasets

Sai Geetha M
3122235001109
Department of Computer Science
SSN College Of Engineering
Email: saigeetha2310537@ssn.edu.in

***Abstract*—This paper presents a detailed end-to-end machine learning workflow for four datasets implemented in assignments a1_iris, a1_loan, a1_diabetes, and a1_email. For each dataset, considered as Dataset 1 to Dataset 4, we describe dataset loading, exploratory data analysis (EDA), data preprocessing, feature selection, data splitting, model selection, and performance evaluation. The datasets cover both classification and regression tasks on structured tabular data and unstructured text. The work aims at understanding the complete machine learning pipeline and identifying suitable algorithms and techniques for each case, addressing Course Outcome CO1 at knowledge level K2.**

***Index Terms*—Machine learning, classification, regression, feature selection, TF-IDF, supervised learning**

I. Introduction

Machine learning workflows typically follow a sequence of steps: loading the dataset, performing exploratory data analysis (EDA), preprocessing, feature selection, splitting into training and testing sets, training models, and evaluating performance. Although this pipeline is common, the exact techniques depend on the task type (classification or regression) and the data type (tabular or text).

In this report, we treat the four assignment datasets as:

- Dataset 1: Iris Dataset (a1_iris)
- Dataset 2: Loan Amount Prediction (a1_loan)
- Dataset 3: Predicting Diabetes (a1_diabetes)
- Dataset 4: Classification of Email Spam (a1_email)

For each dataset, we describe all the required steps: (i) loading the dataset; (ii) exploratory data analysis and visualization; (iii) data preprocessing; (iv) feature selection; (v) data splitting; and (vi) performance evaluation. After that, we summarize the type of ML task, feature selection technique, and suitable algorithms in a comparative table.

II. Dataset 1: Iris Dataset

A. Loading the Dataset

The Iris dataset is a widely used benchmark dataset in machine learning for classification problems. It consists of 150 samples and 5 columns, where four columns represent numerical features and one column represents the class label (species).

The features are:

- Sepal length
- Sepal width
- Petal length
- Petal width

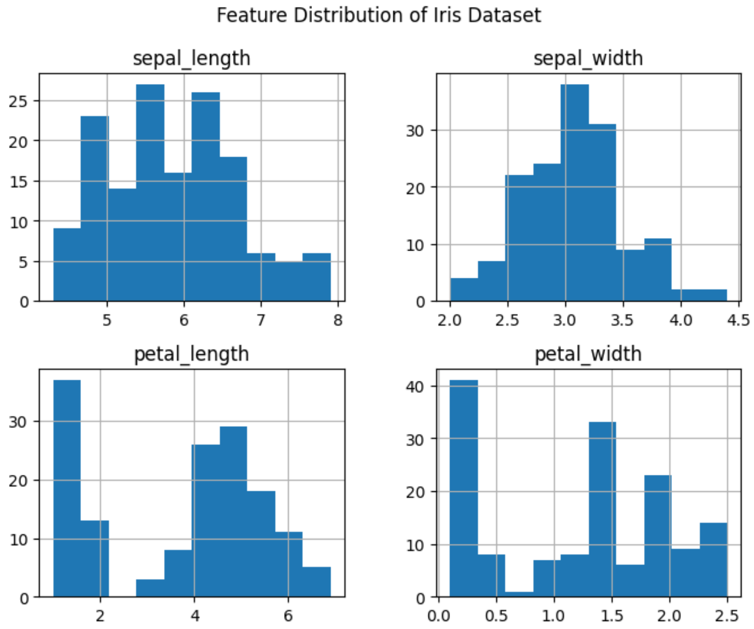

Fig. 1. Feature distribution of the Iris dataset

The target variable is *species*, which has three classes: Setosa, Versicolor, and Virginica.

The dataset is loaded from a CSV file and inspected using basic functions to understand its structure. From the dataset information, it is observed that all features are numerical and there are no missing values. Hence, the dataset is clean and ready for analysis.

B. Exploratory Data Analysis

To understand the distribution of features, histograms are plotted for all four input features. These plots show how each feature is spread across the dataset.

From the histogram plots, it is observed that the petal-related features show clearer separation between values compared to sepal features.

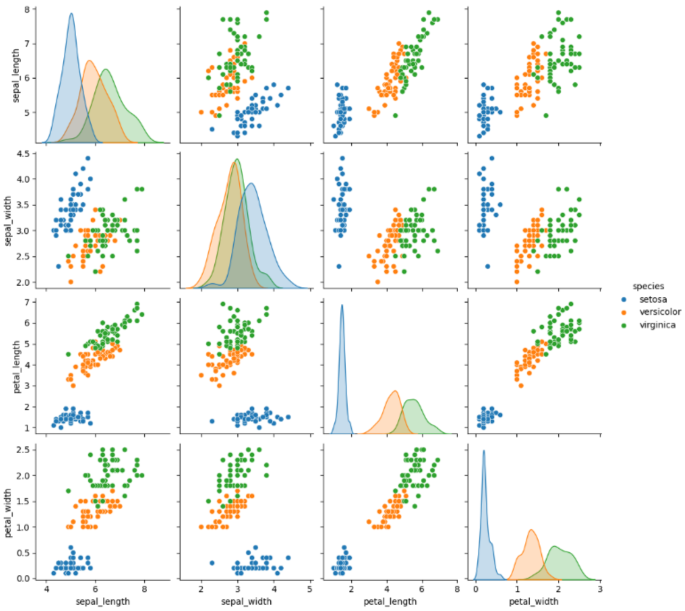

Fig. 2. Pair plot showing feature relationships and class separation

C. Pair Plot Visualization

A pair plot is generated to visualize the relationship between every pair of features, with different colors representing different species.

The pair plot shows that the Setosa class is clearly separable from the other two classes. Petal length and petal width provide strong visual separation, while Versicolor and Virginica show some overlap.

D. Label Encoding

The species column contains categorical values, which cannot be directly processed by machine learning algorithms. Therefore, label encoding is applied to convert the categorical labels into numerical form.

The encoding is as follows:

- Setosa → 0
- Versicolor → 1
- Virginica → 2

This allows the species column to be used as the target variable for classification.

E. Feature Selection

Feature selection is performed using the ANOVA F-test method. This technique measures the statistical relationship between each feature and the target variable. The top two most significant features are selected based on their scores.

It is observed that petal-related features contribute the most towards classification.

F. Data Splitting

The dataset is divided into training and testing sets using an 80:20 ratio. The training set is used to train the model, and the testing set is used to evaluate the performance of the model on unseen data.

G. Logistic Regression Model

A supervised machine learning model, Logistic Regression, is used for classification. Since the dataset contains three classes, this becomes a multiclass classification problem.

The model is trained using the training dataset and then used to predict the species labels for the test dataset.

H. Model Evaluation

The performance of the model is evaluated using classification accuracy. The accuracy obtained is found to be high (around 95–100%), indicating that Logistic Regression performs very well on the Iris dataset.

This high accuracy is due to the clean structure of the dataset and the strong separability between classes.

I. Conclusion

In this experiment, the Iris dataset was used to implement a supervised multiclass classification system using Logistic Regression. Exploratory data analysis showed that petal features play a major role in class separation. After preprocessing, feature selection, and model training, the classifier achieved high accuracy, demonstrating the effectiveness of Logistic Regression for this dataset.

III. Dataset 2: Loan Approval Dataset

A. Loading the Dataset

The Loan Approval dataset is used to perform a supervised regression task. The dataset contains 4269 records and 13 attributes related to customer and financial details.

The attributes include:

- Number of dependents
- Education
- Self-employed status
- Annual income
- Loan amount
- Loan term
- CIBIL score
- Residential asset value
- Commercial asset value
- Luxury asset value
- Bank asset value
- Loan status

The column *loan_id* is removed as it does not contribute to prediction. The dataset contains no missing values and consists of both numerical and categorical attributes.

B. Data Preprocessing

The categorical attributes such as education, self-employed, and loan status are converted into numerical values using label encoding. This transformation allows the machine learning model to process these features.

The feature column names are also cleaned by removing unnecessary spaces.

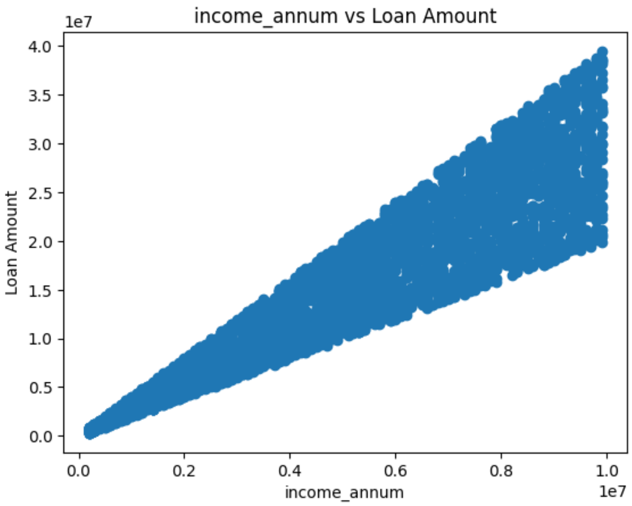

Fig. 3. Income per annum vs Loan Amount

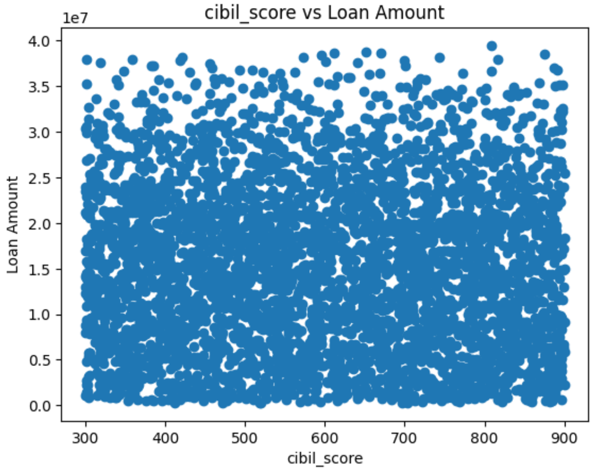

Fig. 4. CIBIL score vs Loan Amount

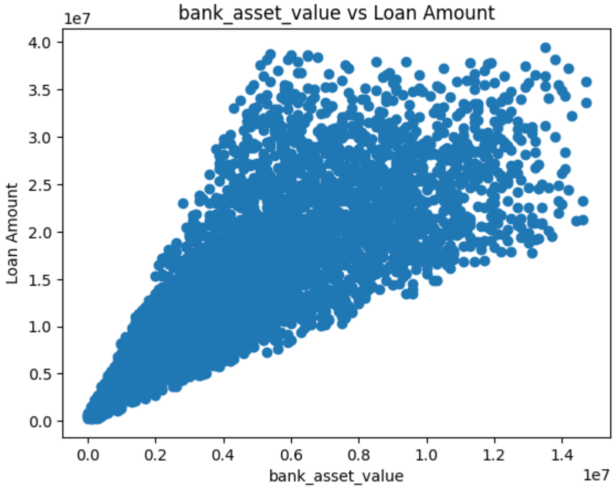

Fig. 5. Bank asset value vs Loan Amount

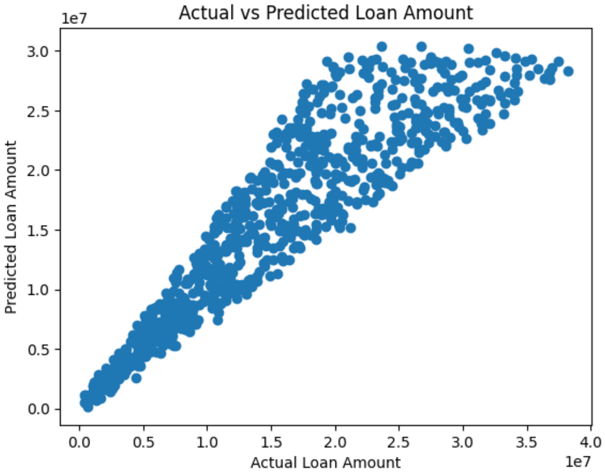

Fig. 6. Actual vs Predicted Loan Amount

C. Exploratory Data Analysis

Scatter plots are generated to visualize the relationship between selected features and the target variable (loan amount).

From the scatter plots, it is observed that income, CIBIL score, and bank asset value show a positive relationship with the loan amount.

D. Feature Scaling

Since the dataset contains features with different value ranges, standardization is applied using the StandardScaler technique. This ensures that all features are scaled to have zero mean and unit variance.

E. Data Splitting

The dataset is split into training and testing sets using an 80:20 ratio. The training set is used to train the regression model, and the testing set is used for evaluation.

F. Linear Regression Model

A supervised learning model, Linear Regression, is used to predict the loan amount. All features except the target variable are used as input for the model.

The model is trained on the training dataset and then used to predict loan amounts for the test dataset.

G. Model Evaluation

The performance of the model is evaluated using Mean Squared Error (MSE) and R-squared score (R^2).

The obtained results are:

- Mean Squared Error (MSE): 1.17×10^{13}
- R-squared Score (R^2): 0.853

The high R^2 value indicates that the model explains approximately 85% of the variance in the loan amount.

H. Actual vs Predicted Visualization

A scatter plot is generated between the actual loan amounts and predicted loan amounts.

The points are closely aligned along the diagonal line, indicating good prediction performance.

I. Conclusion

In this experiment, a supervised regression model was developed using Linear Regression to predict loan amounts. After preprocessing, feature scaling, and training, the model achieved a high R^2 score, demonstrating that Linear Regression is effective for modeling financial datasets with structured numerical features.

IV. Dataset 3: Diabetes Prediction Dataset

A. Loading the Dataset

The Diabetes Prediction dataset is used to perform a supervised binary classification task. The dataset contains 100,000 records with 9 attributes related to patient health conditions and medical history.

The attributes include:

- Gender
- Age
- Hypertension
- Heart disease
- Smoking history
- Body Mass Index (BMI)
- HbA1c level
- Blood glucose level

The target variable is *diabetes*, where:

- 0 indicates non-diabetic
- 1 indicates diabetic

The dataset contains no missing values and includes both numerical and categorical features.

B. Data Preprocessing

The categorical variables gender and smoking_history are converted into numerical values using label encoding. This enables the machine learning models to process the categorical data.

All features except the target variable are selected as input features.

C. Exploratory Data Analysis

A count plot is generated to visualize the class distribution of the diabetes variable.

The plot shows that the dataset is slightly imbalanced, with more non-diabetic samples compared to diabetic samples.

D. Correlation Analysis

A correlation heatmap is generated to understand the relationship between different features and the target variable.

From the heatmap, it is observed that blood glucose level, HbA1c level, and BMI have strong correlation with diabetes.

E. Data Splitting and Feature Scaling

The dataset is split into training and testing sets using an 80:20 ratio. Feature scaling is applied using the StandardScaler method to standardize the input features.

This step ensures that all features contribute equally during model training.

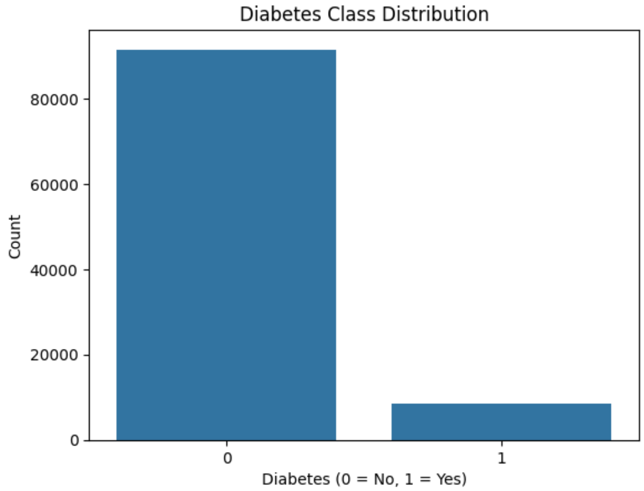

Fig. 7. Class distribution of diabetes

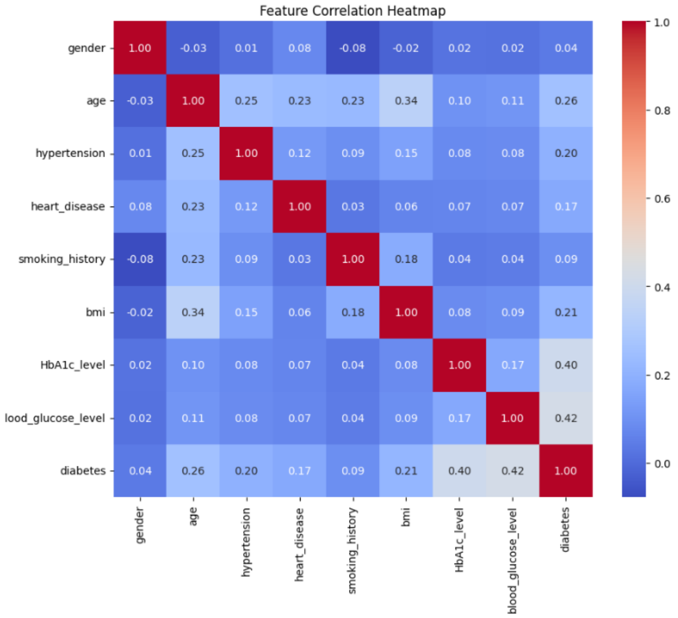

Fig. 8. Correlation heatmap of diabetes dataset

F. Model Training

Three supervised classification algorithms are trained on the dataset:

- Logistic Regression
- K-Nearest Neighbors (KNN)
- Gaussian Naive Bayes

Each model is trained using the training dataset and evaluated using the test dataset.

G. Model Evaluation

The performance of each model is evaluated using classification accuracy. The obtained results are:

- Gaussian Naive Bayes Accuracy: 90.47%

- KNN Accuracy: 96.12%
- Logistic Regression Accuracy: 95.86%

Among the three models, K-Nearest Neighbors achieves the highest accuracy.

H. Best Model Selection

Based on the evaluation results, KNN is selected as the best performing model for this dataset, as it provides the highest accuracy in predicting diabetic and non-diabetic patients.

I. Conclusion

In this experiment, a supervised binary classification system was implemented to predict diabetes using medical and lifestyle features. After preprocessing and feature scaling, three models were compared. KNN achieved the best performance with an accuracy of 96.12%, indicating that instance-based learning works effectively for this dataset.

V. Dataset 4: Email Spam Classification Dataset

A. Loading the Dataset

The Email Spam dataset is used to perform a supervised text classification task. The dataset contains 5573 email messages with two attributes: the category label and the message text.

The attributes are:

- Category (ham or spam)
- Message (email content)

The category column is renamed as *label* and the message column as *message*. The labels are encoded as:

- Ham → 0
- Spam → 1

There are no missing values in the dataset after cleaning.

B. Data Splitting

The dataset is split into training and testing sets using an 80:20 ratio. The message column is used as input and the label column is used as the target variable.

C. Text Vectorization

Since the input data is textual, TF-IDF (Term Frequency–Inverse Document Frequency) vectorization is applied to convert text into numerical feature vectors.

The TF-IDF vectorizer is configured with:

- English stop words removal
- Maximum of 3000 features

This results in a sparse feature matrix representing each email.

D. Model Training

Three supervised machine learning models are trained for spam classification:

- Multinomial Naive Bayes
- K-Nearest Neighbors (KNN)
- Logistic Regression

Naive Bayes and Logistic Regression are trained directly on TF-IDF features, while KNN is trained after converting the sparse matrix into a dense representation.

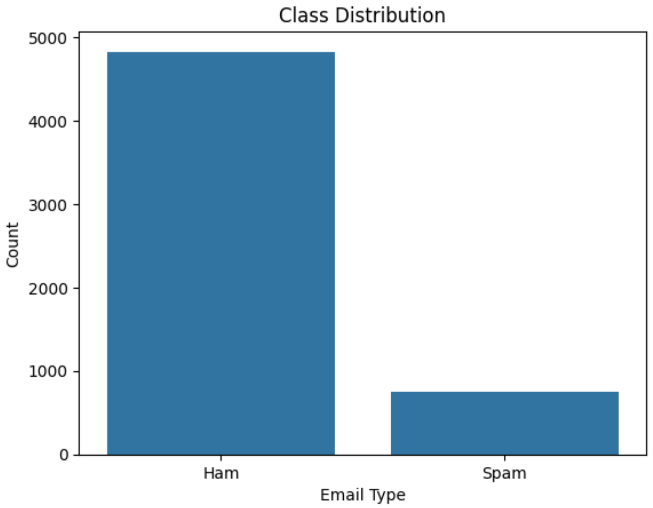

Fig. 9. Class distribution of spam and ham emails

E. Model Evaluation (Accuracy)

The models are evaluated using classification accuracy. The obtained results are:

- Naive Bayes Accuracy: 98.29%
- Logistic Regression Accuracy: 97.49%
- KNN Accuracy: 92.19%

Naive Bayes achieves the highest accuracy among the three models.

F. Class Distribution

A count plot is generated to visualize the distribution of spam and ham emails.

The dataset shows that ham emails are more frequent than spam emails.

G. Confusion Matrix Analysis

Confusion matrices are generated for all three models to analyze their prediction performance.

H. Precision, Recall and F1-score

Precision, recall, and F1-score are computed for all three models and compared using a bar chart.

I. Best Model Selection

Based on accuracy and evaluation metrics, Multinomial Naive Bayes is selected as the best model for this dataset, as it achieves the highest accuracy and strong precision-recall performance.

J. Conclusion

In this experiment, a supervised text classification system was developed to classify emails as spam or ham. After applying TF-IDF vectorization and training multiple models, Naive Bayes achieved the best performance with an accuracy of 98.29%. This demonstrates that probabilistic models work effectively for high-dimensional text data.

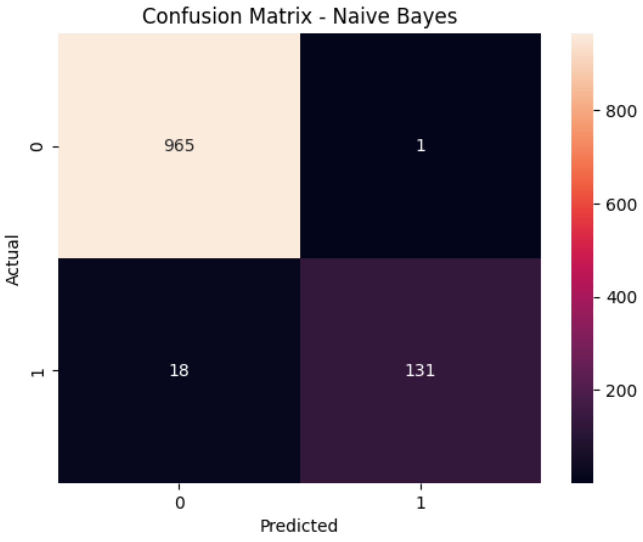

Fig. 10. Confusion Matrix - Naive Bayes

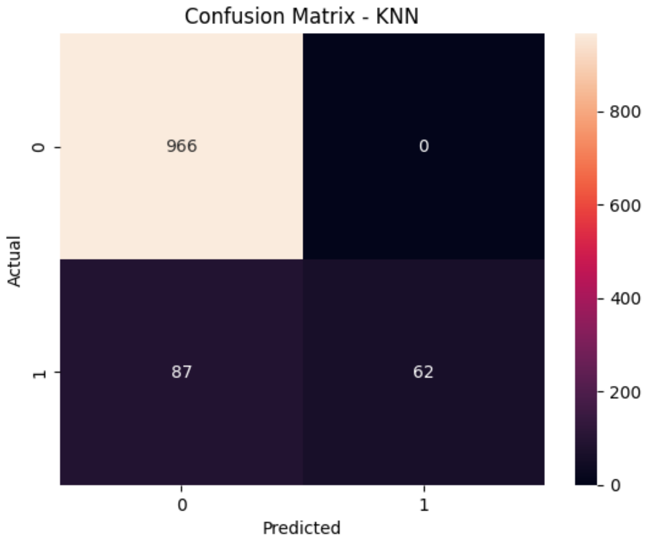

Fig. 11. Confusion Matrix - KNN

VI. Comparative Summary

Table I provides a comparative overview of the four datasets, highlighting the nature of the machine learning task, the preprocessing and feature selection strategies employed, and the corresponding algorithms used for model building.

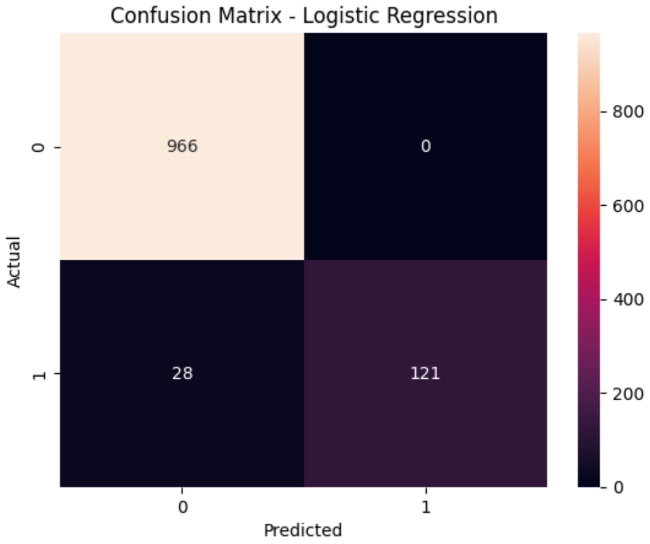

Fig. 12. Confusion Matrix - Logistic Regression

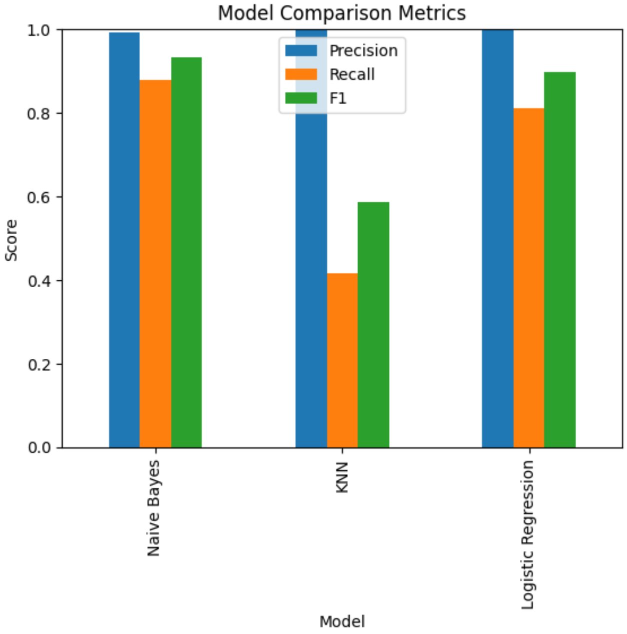

Fig. 13. Precision, Recall and F1-score comparison

TABLE I
Summary of Machine Learning Tasks and Techniques for the Four Datasets

Dataset	**ML Task**	**Feature Selection / Dimensionality**	**Algorithms**
Dataset- Iris	Multiclass classification	SelectKBest - ANOVA F-test	Logistic Regression
Dataset-Loan	Regression	Drop ID, correlation-based filtering	Linear Regression
Dataset-Diabetes	Binary classification	Correlation and ANNOVA	KNN
Dataset-Email	Binary text classification	TF-IDF with vocabulary limit	Naive Bayes, Logistic Regression

VII. Conclusion

In this report, we described the complete machine learning workflow for four different datasets: Iris, loan amount prediction, diabetes prediction, and email spam classification. For each dataset, we covered loading, EDA and visualization, preprocessing, feature selection, data splitting, model selection, and performance evaluation.

The comparison shows how the type of task (classification vs regression) and data type (tabular vs text) influence the choice of preprocessing and algorithms. This study provides a clear understanding of the supervised learning pipeline, matching the course objective CO1 at knowledge level K2. Future work may include hyperparameter tuning, cross-validation, and experimenting with more advanced models.