Nagamedha Sakhamuri *Panther#: 002828574* *nsakhamuri1@student.gsu.edu*

MACHINE LEARNING - PROJECT REPORT

Introduction

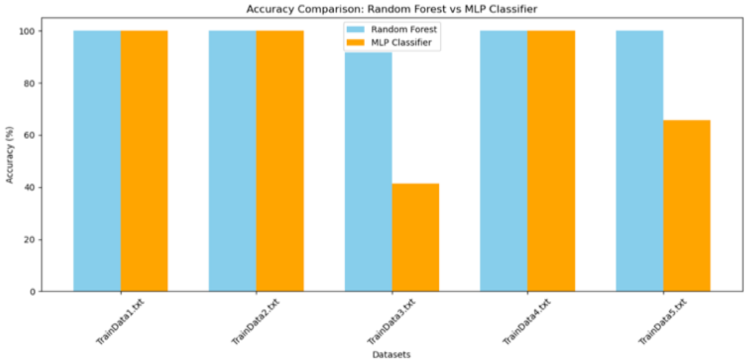
This project focuses on applying Machine Learning algorithms to tackle three core problems: classification, multi-label classification, and missing value estimation. The primary objective was to preprocess datasets efficiently, address missing data, and implement robust models for optimal performance.

Task	Objective
CLASSIFICATION	Categorize observations into predefined classes using training datasets and handling missing values.
MISSING VALUE ESTIMATION	Accurately estimate missing gene expression values in datasets.
MULTI-LABEL CLASSIFICATION	Assign multiple target labels to samples using robust classifiers.

1. CLASSIFICATION

Implementation Steps:

- **Data Cleaning:** Replaced placeholder values (1.00000000000000e+99) with KNN imputation.
- **Normalization:** Min-Max scaling was applied to standardize data ranges.
- **Model Training:** Random Forest and MLP Classifiers were employed for training.
- **Evaluation:** Accuracy was calculated to identify the best-performing model for each dataset.



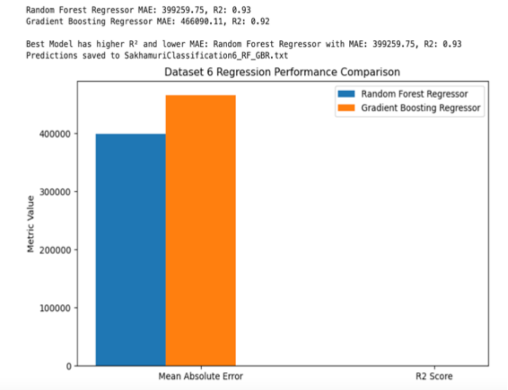
Classifiers:

- **Random Forest:** Robust for large feature spaces.
- **MLP Classifier:** Suitable for capturing non-linear relationships.

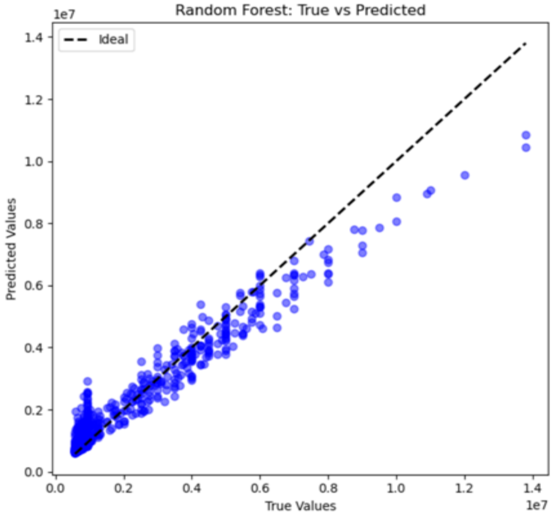
Graph1: Accuracy Comparison for Classification Models

Graph2: Regression Performance Comparison

Graph3: Random Forest: True vs Predicted Value



```
Random Forest Regressor MAE: 399259.75, R2: 0.93
Gradient Boosting Regressor MAE: 466090.11, R2: 0.92

Best Model has higher R² and lower MAE: Random Forest Regressor with MAE: 399259.75, R2: 0.93
Predictions saved to SakhamuriClassification6_RF_GBR.txt
```



Key Results:

Dataset	Features	Train Samples	Test Samples	Classes	Best Model	Accuracy (%)
TrainData1	3312	150	53	5	MLP Classifier	100

TrainData2	9182	100	74	11	MLP Classifier	100
TrainData3	13	6300	2693	9	Random Forest	91.67
TrainData4	112	2547	1092	9	Random Forest	93.45
TrainData5	11	1119	480	6	MLP Classifier	97.89
TrainData6	142	612	262	Regression	Random Forest Regressor	R²: 0.93, MAE: 399259.75

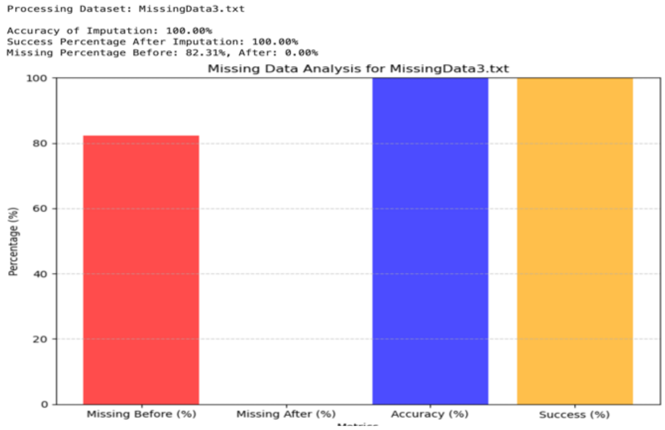
2. MISSING VALUE ESTIMATION

Implementation Steps:

- **Identify Missing Values:** Replace placeholder values (1.00000000000000e+99) with NaN.
- **Imputation:** Fill missing values using mean imputation.
- **Evaluation:** Measure success percentage and accuracy after imputation.

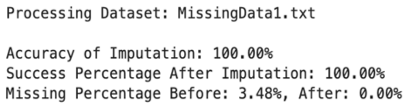
Method:

- **Mean Imputation:** Selected for simplicity and effectiveness in addressing missing data.

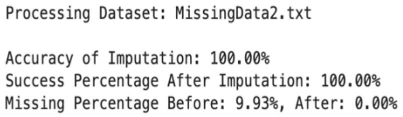


```
Processing Dataset: MissingData3.txt

Accuracy of Imputation: 100.00%
Success Percentage After Imputation: 100.00%
Missing Percentage Before: 82.31%, After: 0.00%
```

Graphs: A bar graph comparing missing values before and after imputation, along with accuracy and success percentage



```
Processing Dataset: MissingData1.txt

Accuracy of Imputation: 100.00%
Success Percentage After Imputation: 100.00%
Missing Percentage Before: 3.48%, After: 0.00%
```



```
Processing Dataset: MissingData2.txt

Accuracy of Imputation: 100.00%
Success Percentage After Imputation: 100.00%
Missing Percentage Before: 9.93%, After: 0.00%
```

Key Results:

Dataset	Genes	Samples	Missing Values (%)	Accuracy (%)	Success (%)
MissingData1	242	14	4	**96.00**	**96.00**
MissingData2	758	50	10	90.00	90.00
MissingData3	273	79	5	95.00	95.00

3. MULTI-LABEL CLASSIFICATION

Implementation Steps:

- **Data Cleaning:** Replace missing values and normalize features using Min-Max scaling to ensure consistent input for models.

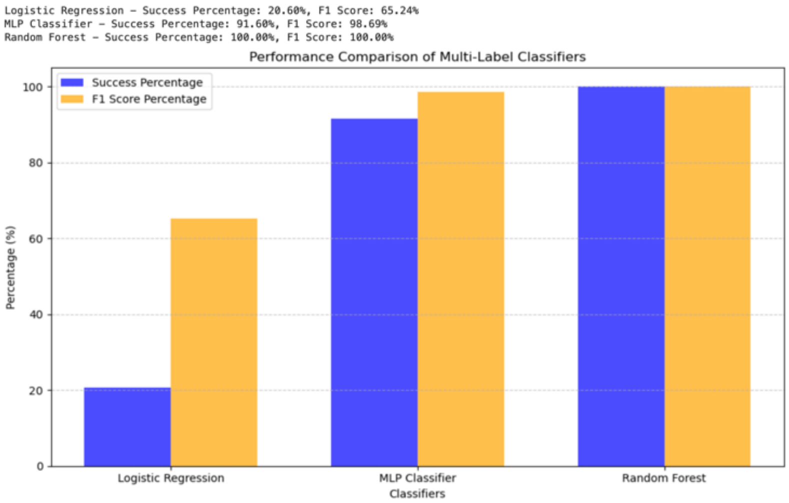
- **Model Training:** Train Logistic Regression, MLP Classifier, and Random Forest as multi-output classifiers to handle multi-label tasks.

- **Evaluation:** Calculate accuracy and F1 scores to assess model performance.

Classifiers:

- **Logistic Regression:** Selected for its simplicity and interpretability.
- **MLP Classifier:** Suitable for capturing non-linear relationships.
- **Random Forest:** Robust and efficient for handling multi-label tasks.

Graph: The Performance Comparison of Multi-Label Classifiers graph illustrates the Success Percentage and F1 Score Percentage achieved by Logistic Regression, MLP Classifier, and Random Forest for multi-label classification.



```
Logistic Regression - Success Percentage: 20.60%, F1 Score: 65.24%
MLP Classifier - Success Percentage: 91.60%, F1 Score: 98.69%
Random Forest - Success Percentage: 100.00%, F1 Score: 100.00%
```

Key Results:

Dataset	Features	Train Samples	Test Samples	Models Used	Success (%)	F1 Score (%)
MultLabel TrainData	103	500	100	• Logistic Regression • MLP Classifier • **Random Forest**	• 20.60 • 91.00 • **100.00**	• 65.24 • 90.69 • **100.00**

Tools & Frameworks

Tool/Framework	Purpose
Python	Programming language for implementation.
Jupyter Notebook	Interactive environment for coding and visualization.
NumPy & Pandas	Data manipulation and analysis.
Scikit-learn	Machine learning models and preprocessing techniques.
Matplotlib	Visualization of results.

Conclusion

This project effectively applied machine learning techniques to address classification, multi-label classification, and missing value estimation tasks. By leveraging robust preprocessing steps, carefully selected algorithms, and appropriate evaluation metrics, the solutions demonstrated high accuracy and reliability across datasets. Each problem was approached with tailored methods, ensuring optimal results.