Academic Year	Module	Assignment Number	Assessment Type
2024/25	5CS037: Concepts and Technologies of AI	3	Report

Classification Analysis Report

Student Id	: 2408286
Student Name	: Arjabi Shrestha
Section	: L5CG15
Module Leader	: Siman Giri
Tutor	: Siman Giri
Submitted on	: 11-02-2025

Table of Contents

ABSTRACT

The goal is to predict a categorical outcome, Human Development Index (HDI) and classify countries into different levels using socio-economic factors.

The dataset used for this analysis is sourced from Kaggle which was originally developed by the United Nations Development Programme (UNDP). It includes socio-economic indicators for various countries such as GII, maternal mortality, adolescent birth rate,etc. And the target variable is Human Development Index(HDI) which is categorized into four levels(Very High, High, Medium, Low).

EDA involves handling missing values were by dropping the rows and visualization of data. Model building classification techniques includes Logistic Regression and Random Forest Classifier.

Hyperparameter were optimization using GridSearchCV for both the models.

The performance of the models was evaluated using accuracy, precision, recall, and F1-score. Loss was also observed to see how well the model generalizes

- The Logistic Regression Model shows:

  Accuracy: 76.47%, Precision: 0.76, Recall: 0.76, F1-Score: 0.76

  Train Loss: 0.4938 and Test Loss: 0.6183

- The Random Forest Classifier Model shows:

  Accuracy: 67.65%, Precision: 0.68, Recall: 0.68, F1-Score: 0.67

  Train Loss: 0.0000 andTest Loss: 0.3235

- The Hypertuned Logistic Regression Model shows:

  Accuracy: 70.59%, Precision: 0.70, Recall: 0.71, F1-Score: 0.70

  Train Loss: 0.5150 andTest Loss: 0.7245

Though all the model shows resonable results, the hypertuned Logistic Regression resulted in a slight decrease in performance than the original Logistic Regression model provided best result.

1. INTRODUCTION

1.1. Problem Statement

Classification problems involve assigning categorical data based on it's features. class based on its features. Using labeled data, a model learns to associate inputs with known classes. This trained model then classifies new, unseen data by associating it with the most likely class.

1.2. Dataset

The classification analysis this dataset was based one was sourced from Kaggle, originally developed by the United Nations Development Programme (UNDP). It contains socio-economic indicators that influence human development and gender inequality.

It contains 196 rows × 11 columns with the target variable being Human_development categorized into Very High, High, Medium, and Low. The Independent Variables include Country, GII, Male secondary education attainment (%), Female labor force participation (%), Male labor force participation (%), Maternal mortality rate (deaths per 100,000 live births), Adolescent birth rate (births per 1,000 women ages 15–19), Seats held by women in parliament (%), Female secondary education attainment (%).

This dataset aligns with SDG 5: **Gender Equality**, which aims to reduce gender discriminatio. By classifying countries based on their Human Development Index (HDI), this analysis helps identify patterns in development and gender inequality. This could be helpful for providing insights for countries to formulate targeted strategies for improving global gender equality.

1.3. Objective:

The objective is to classify countries into different Human Development Index (HDI) levels: Very High, High, Medium, and Low based on socio-economic indicators,i.e. GII, Maternal mortality rate, Adolescent birth rate, Seats held by women in parliament, Female secondary education attainment, Male secondary education attainment, Female labor force participation, and Male labor force participation. This project aims to analyze key factors influencing HDI classification and improve prediction accuracy using hyperparameter optimization, feature selection, and model tuning. The features used for classification include Country,

2. Methodology

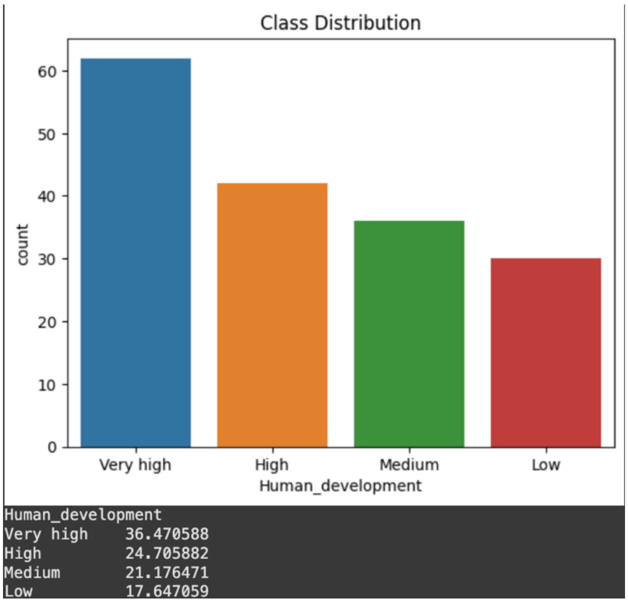
2.1 Data Preprocessing

Before building the model, the data observed and was found to be data seems fairly consistent. The dataset was cleaned by dropping rows with missing data. Additionally, scaling were performed to prepare the data for analysis, ensuring that all variables were on a comparable scale.

2.2 Exploratory Data Analysis (EDA)

Bar graphs and grouped bar graphs to better understand and visualize the relationships between variables as shown below:

Bar Graphs:



```
Human_development
Very high    36.470588
High         24.705882
Medium       21.176471
Low          17.647059
```

The "Very high" human development category represents the largest proportion (approximately 36.5%) of the observed data, nearly twice as prevalent as the "Low" category (17.6%) indicating that there are more countries that have a very high development index than there is low.

Figure 1: Class distribution of Human Development

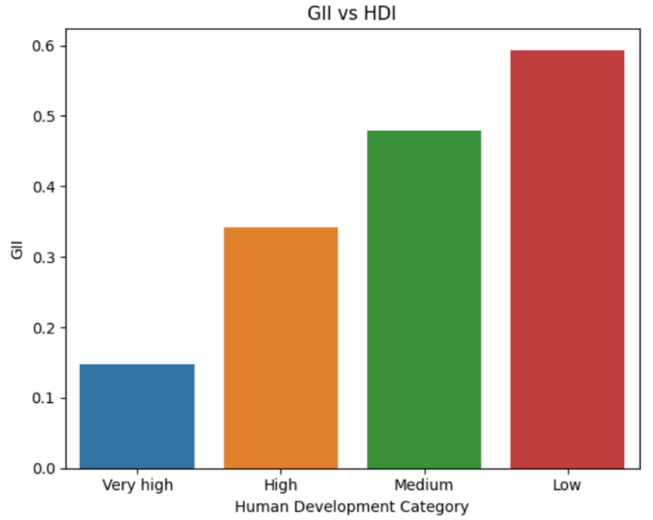


Figure 2: GII by HDI

The Gender Inequality Index (GII) shows a negative correlation as human development goes down, gender inequality goes up. This suggests that countries with lower HDI scores often face higher levels of gender disparity

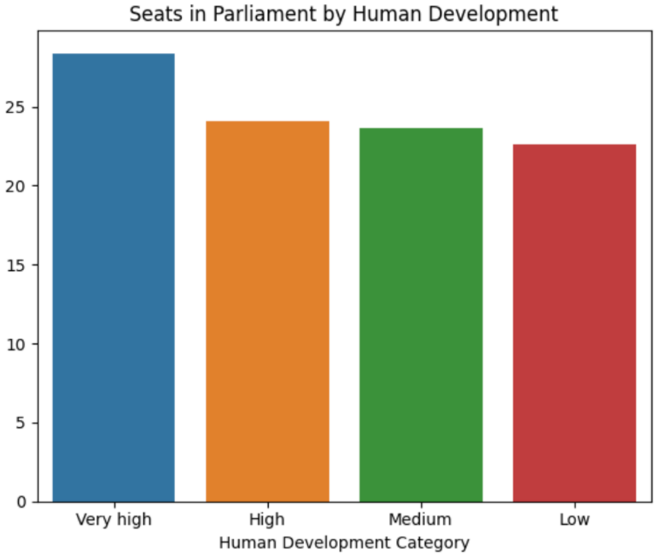


Figure 3:Seats in Parliament by Human Development

While the number of seats in parliament appears to decrease with lower. This suggests that countries with lowert HDI have poorer political representation.

Grouped Bar Graphs:

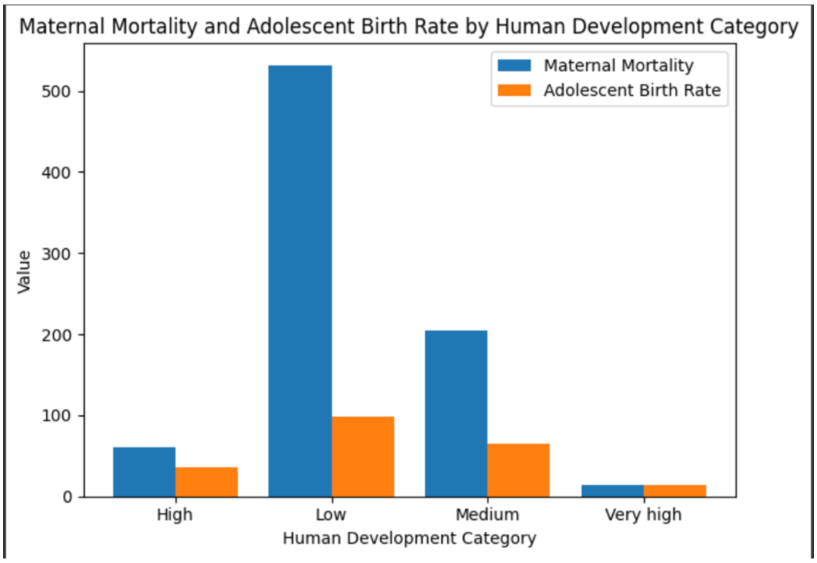


Figure 5: Maternal Mortality and Adolescent Birth Rate by Human Development Category

Both maternal mortality and adolescent birth rates exhibit a strong inverse relationship with human development categories. Suggesting that countries with lower HDI have greater maternal mortality as well as adolescent birth rate

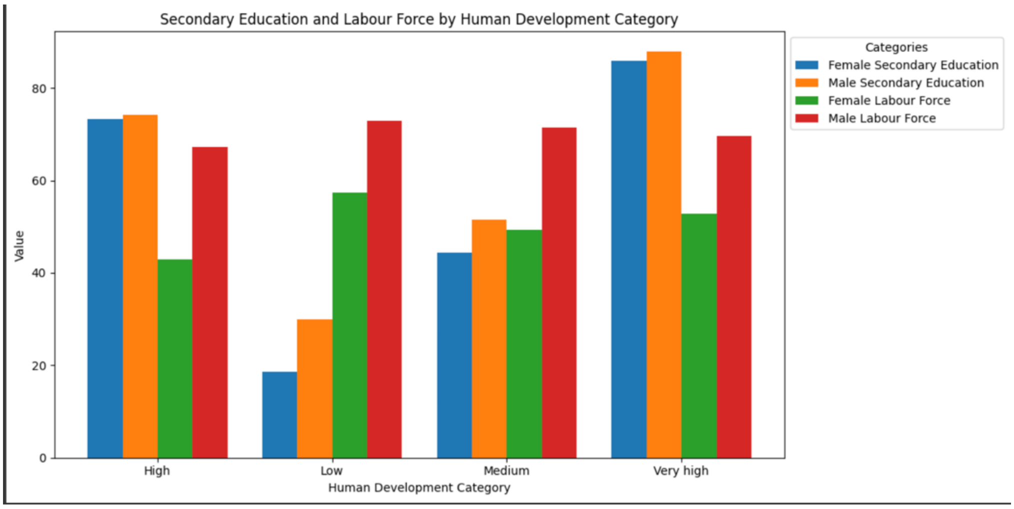


Figure 4: Secondary Education and Labour Force by Human Development Category

Male secondary education enrollment consistently exceeds female enrollment across all human development categories, reflecting a persistent gender gap in education.

Although labor force participation rises with higher human development, the gender gap remains, with male participation significantly higher than female.

2.3. Model Building

Two classification models were considered for this task:

1: Logistic Regression

2: Random Forest Classifier

The models were splitting the data into 70% training and 30% testing and the features were scaled using StandardScaler to ensure that all variables were on the same scale.

The model was trained on the scaled training data using Scikit-learn’s LogisticRegression() logistic regression and RandomForestClassifier()for random forest.

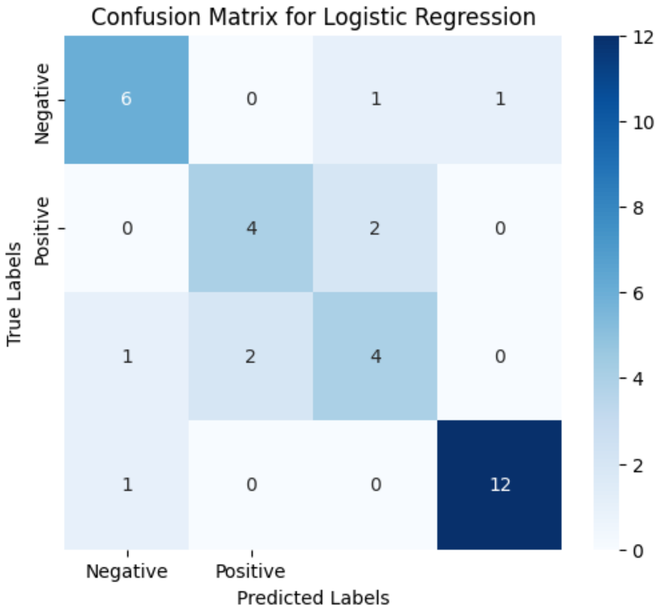
2.4. Model Evaluation

The model’s performance was evaluated using several key metrics that are essential for understanding its effectiveness, particularly when dealing with imbalanced classes. The key metrics used were: accuracy, precision,recall and f1-score .

Model 1: Logistic Regression

Accuracy: 76.47%

Confusion Matrix:



Classification Report:

```
              precision    recall  f1-score   support

        High       0.75      0.75      0.75         8
         Low       0.67      0.67      0.67         6
      Medium       0.57      0.57      0.57         7
   Very high       0.92      0.92      0.92        13

    accuracy                           0.76        34
   macro avg       0.73      0.73      0.73        34
weighted avg       0.76      0.76      0.76        34
```

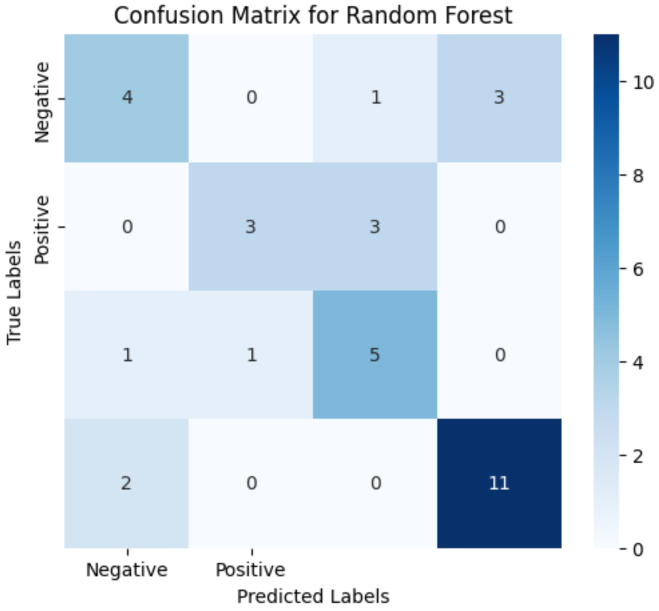
The Logistic Regression accuracy is 76.47%, with good performance in predicting the Very High class. However, the model struggled with the Medium and Low classes, showing lower precision and recall values.

Model 2: Random Forest Classifier

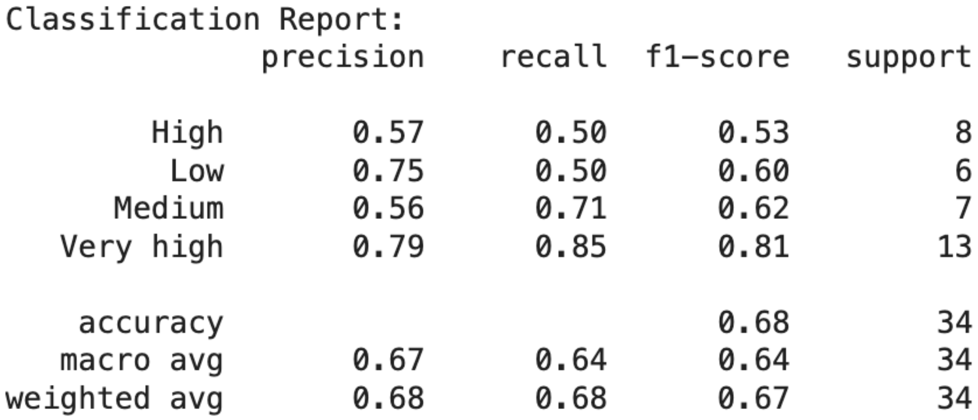
Model: Random Forest

Accuracy: 67.65%

Confusion Matrix:



Classification Report:



```
Classification Report:
              precision    recall  f1-score   support

        High       0.57      0.50      0.53         8
         Low       0.75      0.50      0.60         6
      Medium       0.56      0.71      0.62         7
   Very high       0.79      0.85      0.81        13

    accuracy                           0.68        34
   macro avg       0.67      0.64      0.64        34
weighted avg       0.68      0.68      0.67        34
```

The Random Forest accuracy is 67.65%, with the Very High class showing the best performance. The High and Low classes had lower precision and recall,while the Medium class showed decent recall but lower precision.

2.5. Hyper-parameter Optimization

GridSearchCV were used on both the models to try to enhance the model's performance.

Optimization for Model 1: Logistic Regression

For Logistic Regression, the following optimal hyperparameters were identified through GridSearchCV:

C: 10

max_iter: 100

Optimization for Model 2: Random Forest Classifier

For Random Forest Classifier, the best hyperparameters found through GridSearchCV were:

max_depth: 10

min_samples_split: 5

n_estimators: 100

2.6. Feature Selection

Feature Selection for model 1: Logistic Regression:

Recursive Feature Elimination (RFE) was used to select the most important features. The top 3 selected features were: ‘GII’, ‘Maternal_mortality’ and ‘F_secondary_educ’

Feature Selection for model 2: Random Forest:

Feature Importances were used to determine the key features in random forest. The most important features selected were: ‘F_secondary_educ’, ‘GII’, ‘Maternal_mortality’.

3. Conclusion

3.1. Key Findings

After evaluating the model’s performance on the test dataset, the results showed that the Logistic Regression model outperformed the Random Forest model, with an accuracy of 76.47% compared to 67.65%. The Logistic Regression model also showed more consistent performance across all classes, while the Random Forest model performed well in the Very High class but struggled with the Medium and Low classes. Making Logistic Regression a better choice for this dataset.

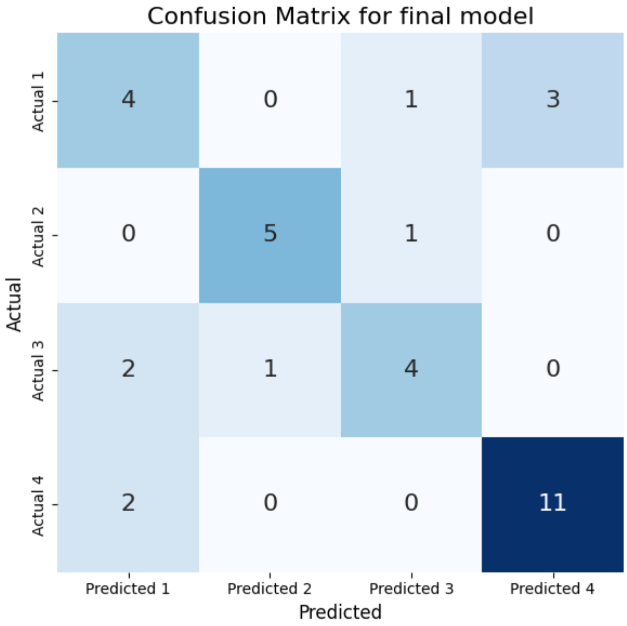
Hyper-parameter optimization through GridSearchCV and feature selection using RFE and Feature Importances helped improve the models' effectiveness by selecting the most influential features, i.e. GII, Maternal_mortality, and F_secondary_educ on both models.

3.2. Final Model

Based on Logistics regression, the final model built using the optimal hyperparameters and selected features. The key findings are as follows:

Accuracy: 70.59%

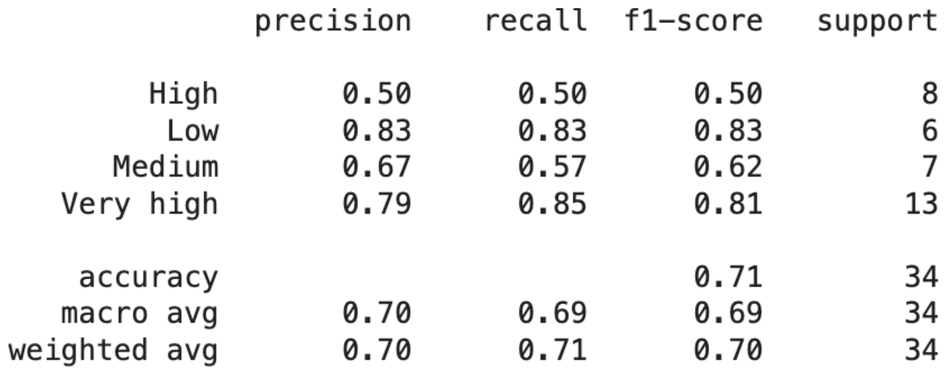
Confuion matrix:



Confusion Matrix for final model

Actual \ Predicted	Predicted 1	Predicted 2	Predicted 3	Predicted 4
Actual 1	4	0	1	3
Actual 2	0	5	1	0
Actual 3	2	1	4	0
Actual 4	2	0	0	11

Classification Report:



```
              precision    recall  f1-score   support

        High       0.50      0.50      0.50         8
         Low       0.83      0.83      0.83         6
      Medium       0.67      0.57      0.62         7
   Very high       0.79      0.85      0.81        13

    accuracy                           0.71        34
   macro avg       0.70      0.69      0.69        34
weighted avg       0.70      0.71      0.70        34
```

3.3. Challenges

The first challenge faced was that the dataset size was realtively small of only 196 rows × 11 columns. Although the data quality was good, the limited number of data was challenging in terms of model generalization and overfitting. While scikit-learn's feature selection methods were used to identify the most impactful features, and it was difficult to ensure that the selected features captured valuable information.

3.4. Future Work

For future works, searching a dataset with more data would help the model generalize better and reduce overfitting.

4. Discussion

4.1. Model Performance

The model's performance results indicate that the model performed well on the test data, with an accuracy of 70.59%. It showed good performance for the Low and Very High categories, with precision, recall, and F1-score values all above 0.70. However, the model struggled with the High and Medium categories.

4.2. Impact of Hyperparameter Tuning and Feature Selection

After applying hyperparameter tuning, the model was seen to have performance, potentiallly due to overfitting. Feature selection, on the other hand, helped identify the most important predictors, ensuring that the model focused on relevant features but in this particular dataset Feature selection had limited impact since most features were relevant, making it challenging to eliminate any without losing valuable information.

4.3. Interpretation of Result

The original logistic regression model achieved an accuracy of 76.47%, which dropped to 70.59% when it was hypertuned. The hypertuning process likely led to overfitting, resulting in poorer generalization to the test data. Additionally, the loss values increased in the tuned model indicating that the tuning process may have introduced unnecessary complexity without improving performance. The confusion matrices showed that while the tuned model improved predictions for the "Low" and "Very high" categories, its accuracy in predicting "High" and "Medium" categories decreased.

4.4. Limitation

The limited dataset size may reduce the model's ability to generalize well to new data. Additionally, while feature selection was applied, most features appeared relevant, making it difficult to confidently eliminate any without potentially losing important information.

4.5. Suggestions for Future Research

For the future, increasing the dataset size to improve model generalization and reduce overfitting would be a good idea. Additionally, experimenting with different classification algorithms, could help capture more complex patterns in the data. Further optimization of feature selection methods may also enhance model performance by ensuring that only the most relevant predictors are used.