Concepts and Technologies of AI

Final Portfolio Project
An End - to - End Machine Learning Report
on Regression Task

Name: Monalika Tamang
Student ID: 2408566
Group: L4CG8

Table of Contents

Abstract

Purpose: The purpose of the report is to show the prediction of overall literacy rate based on various factors using regression techniques.

Approach: The regression of this dataset is done following the steps: Conducting Exploratory Data Analysis(EDA), building model from scratch (linear regression), building two primary models (Random forest regression and Gradient boosting), Evaluating with metrices, optimizing hyper-parameter, feature selection and rebuilding final model with best hyperparameter and features.

Key Results: The performance of all the models were evaluated based on the three metrices. The evluation of first model Random Forest demonstrated: MAE: 0.57, RMSE: 0.75, R^2: 0.99 and the evluation of second model Gradient Boosting demonstrated: MAE: 0.52, RMSE: 0.65, R^2: 1.00. The final model demonstrated: MAE: 0.05, RMSE: 0.06, R^2 Score: 1.00.

Conclusion: The performance of Final model was the best among all the other models for training data. The key insight includes the analysis of Overall Literacy Rate with the factors present in dataset.

1. Introduction:

1.1 Problem Statement:
The main aim of this project is to predict the target variable i.e. Overall Literacy Score based on the factors like education level, location and factors present on the dataset.

1.2 Dataset:
The datset used for the regression task is Digital Literacy Education Dataset obtained from Kaggle. The dataset aims to predict the final literacy score based on educational level, laction type, internet usuage score, etc. The datset aligns with SDG 4: Quality Education.

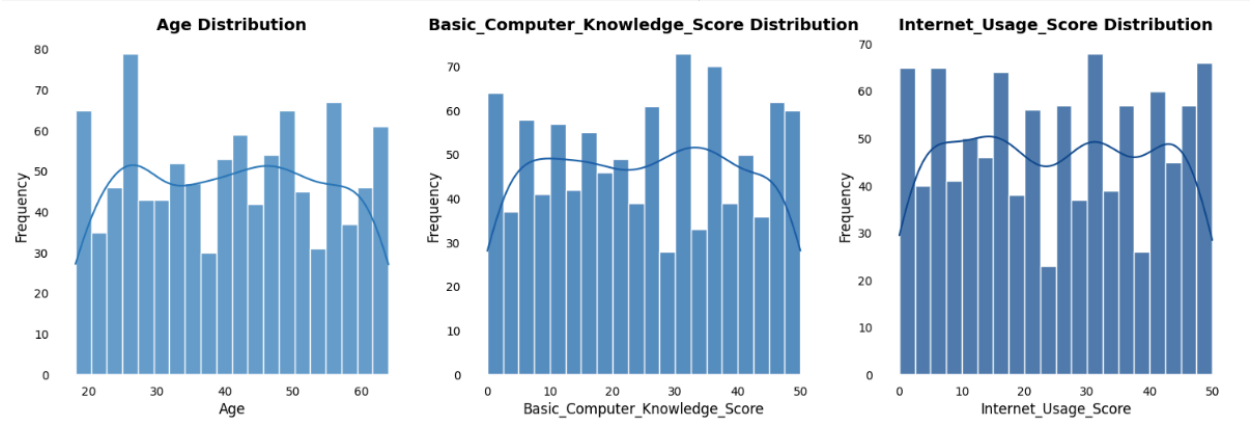
1.3 Objective:
The main objective of this task is to perform regression task with the help of models and predicting target variable of the dataset.

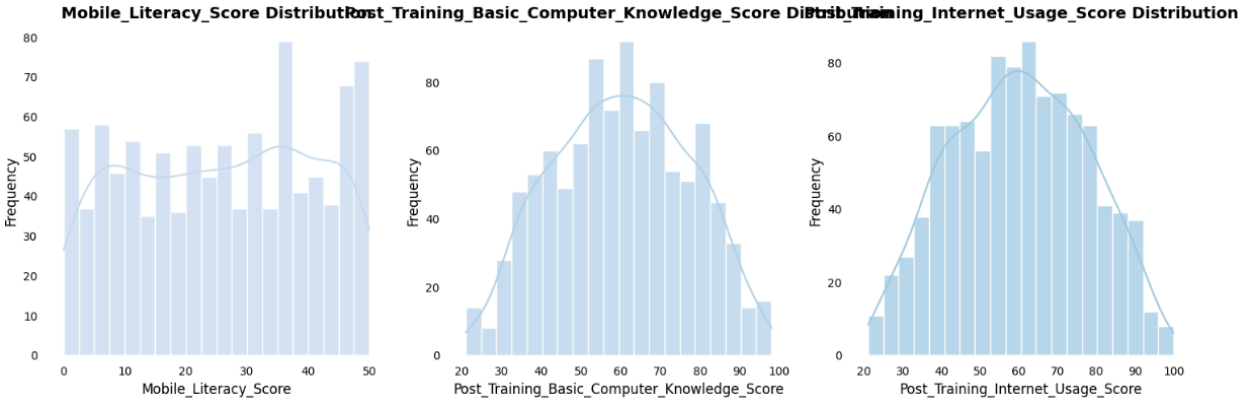
2. Methodology

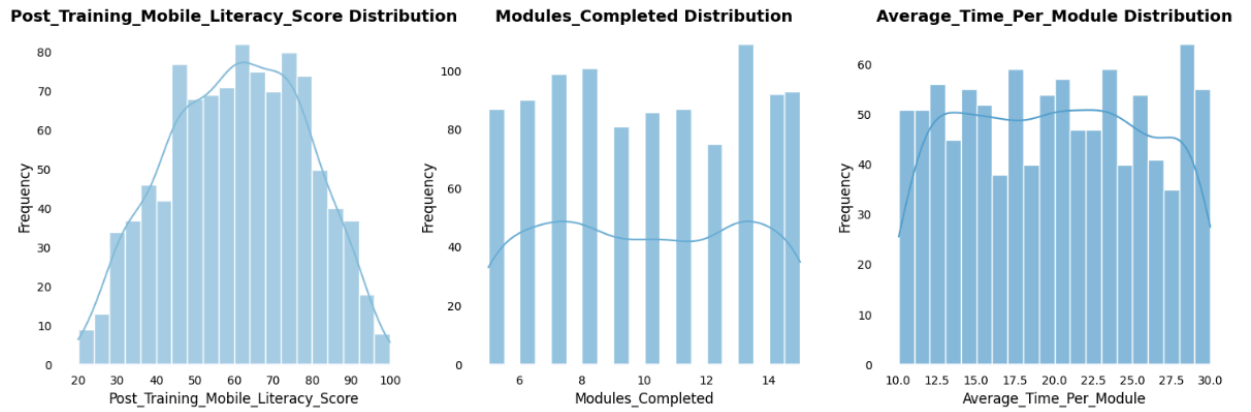
2.1 Data Processing:
First, the dataset is loaded and cleaned by handling missing values. The missing values are filled with mode.

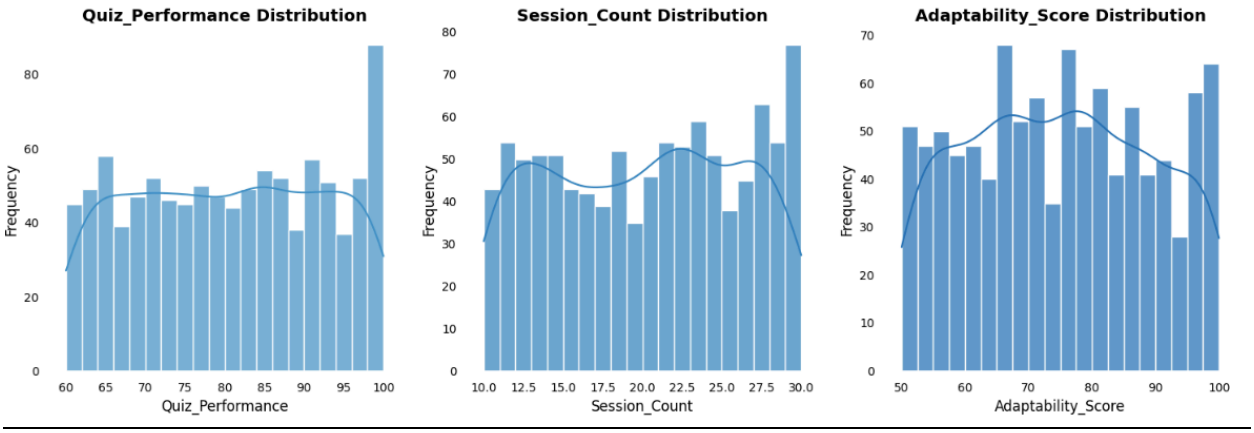
2.2 Exploratory Data Analysis (EDA):
After the data processing, it was visualized in the form of histogram, heatmap, boxplot etc. The visualization of data helps to understand and explore the information much more.

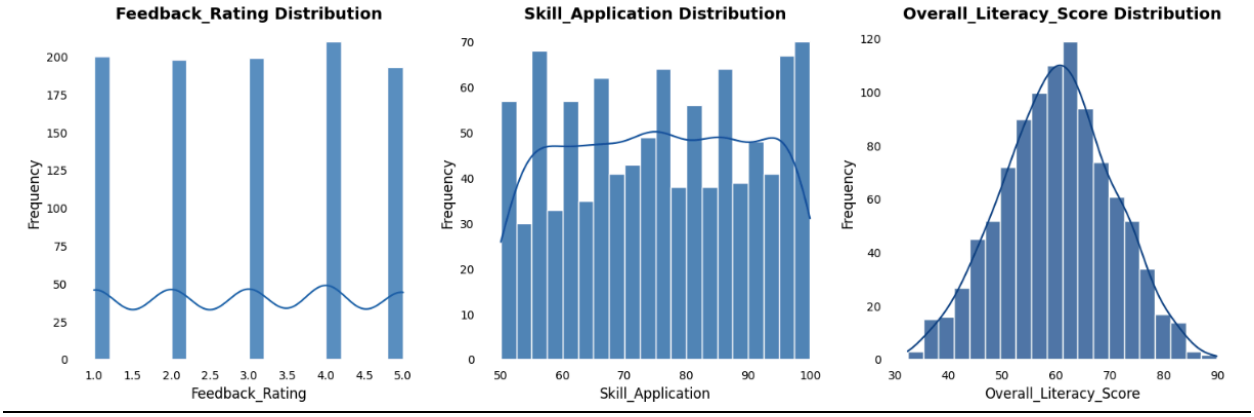
The histogram of numerical columns was plotted illiustrating the distribution of each columns with their frequencies.


Age Distribution
Frequency
Age
Basic_Computer_Knowledge_Score Distribution
Frequency
Basic_Computer_Knowledge_Score
Internet_Usage_Score Distribution
Frequency
Internet_Usage_Score

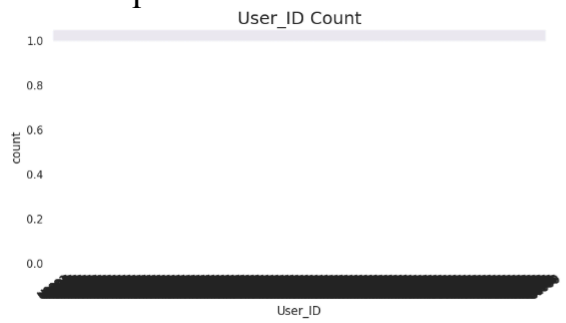

Mobile_Literacy_Score Distribution
Frequency
Mobile_Literacy_Score
Post_Training_Basic_Computer_Knowledge_Score Distribution
Frequency
Post_Training_Basic_Computer_Knowledge_Score
Post_Training_Internet_Usage_Score Distribution
Frequency
Post_Training_Internet_Usage_Score

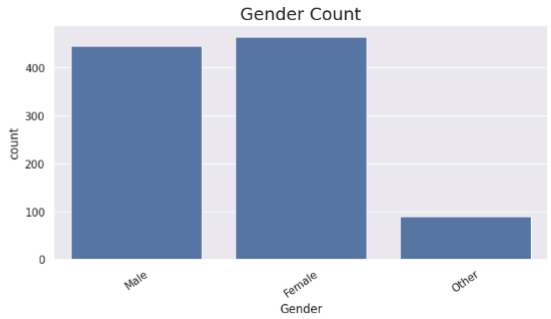

Post_Training_Mobile_Literacy_Score Distribution
Frequency
Post_Training_Mobile_Literacy_Score
Modules_Completed Distribution
Frequency
Modules_Completed
Average_Time_Per_Module Distribution
Frequency
Average_Time_Per_Module

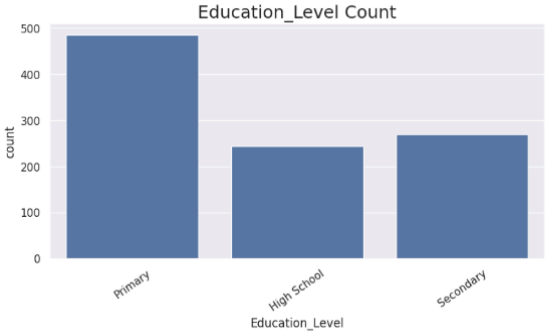


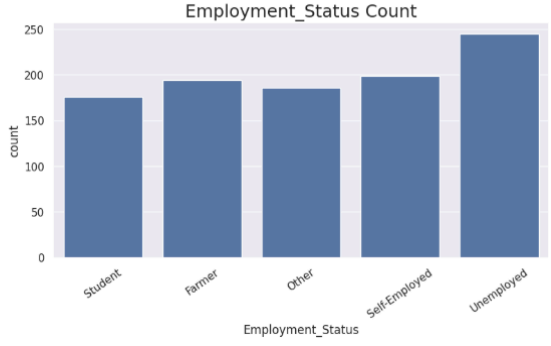


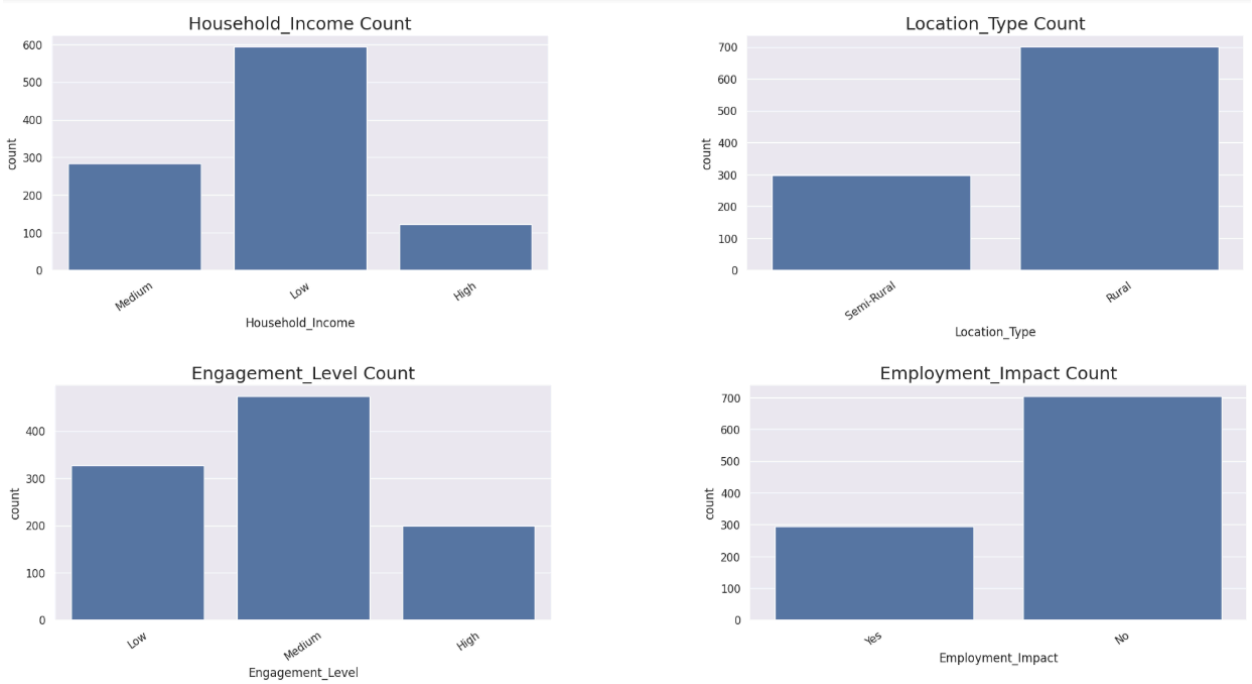
Then, many bar plots were plotted to show the categorial columns and count of their unique values.



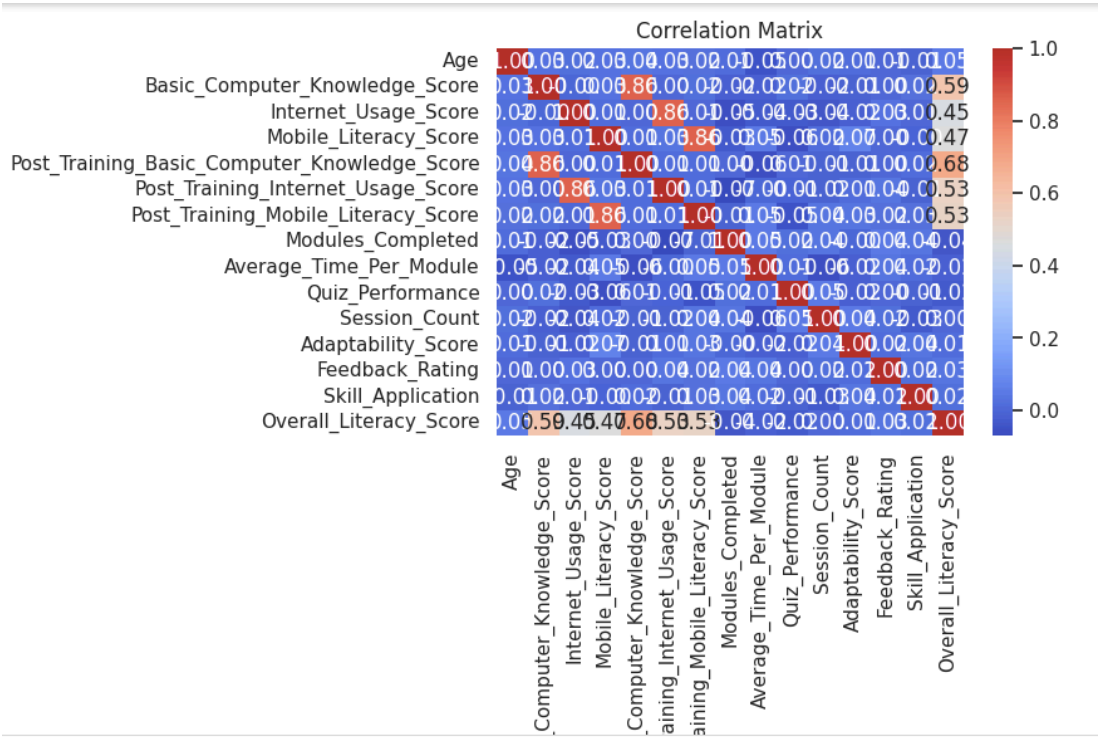




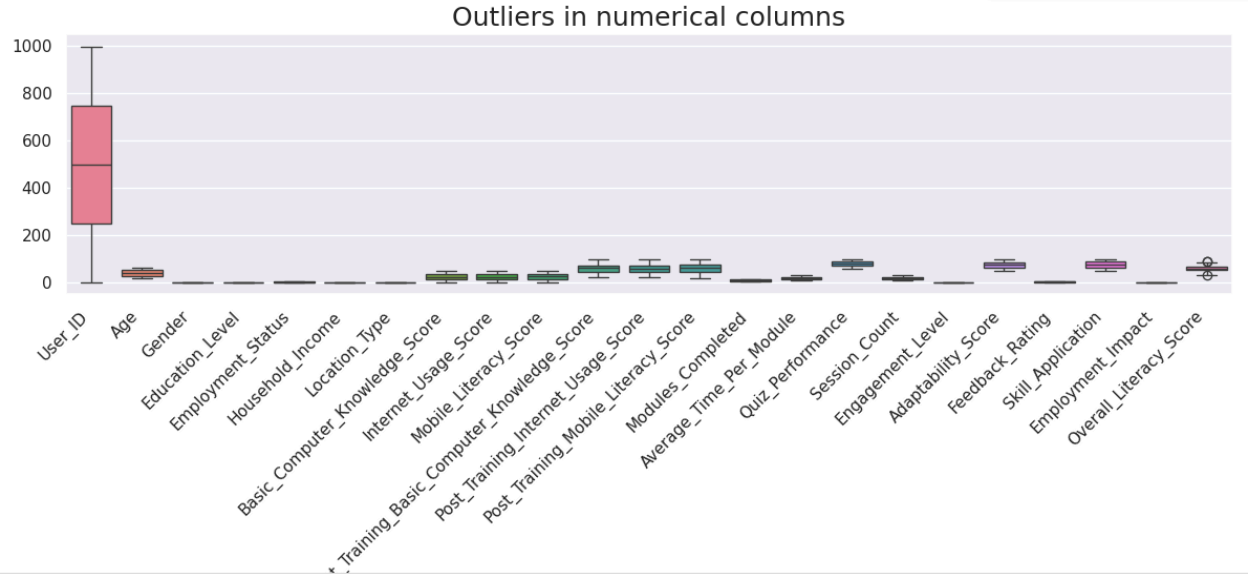




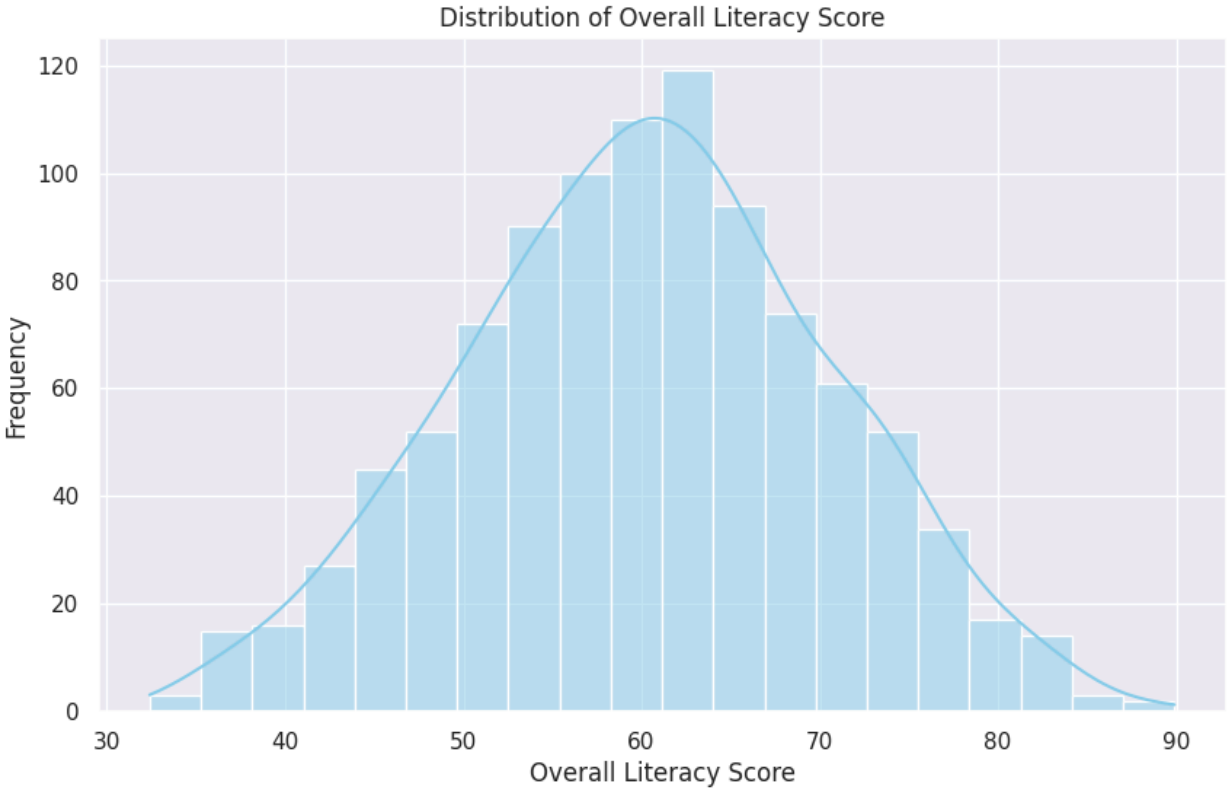
A heatmap is also plotted to represent the correlation coefficients between numerical variables of the dataset.



Similarly, box plots of each numerical column is also plotted to detect the outliers. There is outliers in Overall_Literacy_Score.



The key feature of the dataset is Overall_Literacy_Score and it is visualized in histogram.



The key insight of EDA includes the visualization of features of dataset and the key feature of dataset.

2.3 Model Building:

The next step is to build models and the models present in this task are:

1. Logistic Regression: The model is built from sratch which is then trained on sliptted train and test datas.
2. Random Forest Regression: The another is built with the help of sklearn and trained on training and test datas.
3. Gradient Boosting: The last model is also built with the help of sklearn and similarly trained.

2.4 Model Evaluation:

As for the regression task, the model evaluation is performed using metrices like Mean Absolute Error (MAE), R-Squared (r2) and Root-Mean Squared Error (RMSE). Both the models are evaluated using these three metrices and compared to find which model performs better on the basis of evaluation.

2.5 Hyper-parameter Optimization:

To find the best hyperparameter of both the models, GridSearchCV from sklearn was used.

- The best hyperparameter for Random Forest Regressor: {'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}.
- The best hyperparameter for Gradient Boostingis: {'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 5, 'n_estimators': 50}.

2.6 Feature Selection:

The feature selection of models were coducted using SelectKBest from sklearn which gave the best 5 features of each models.

- The feature indices of Random Forest Regression is : [10 11 12 9 8]
- The feature indices of Gradient Boosting is : [9 10 11 7 8]

3. Conclusion

3.1 Key findings:

The evaluation of model's performance was condusted by the three metrices: MAE, r2 and RMSE.

By comparing the metrices, the Gradient Boosting model performed better than Random Forest regression.

- The evluation of first model Random Forest demonstrated for training data: MAE: 0.57, RMSE: 0.75, R^2: 0.99.
- The evluation of second model Gradient Boosting demonstrated for training data: MAE: 0.52, RMSE: 0.65, R^2: 1.00.

3.2 Final Model:

The Gradient Boosting was retrained again using the selected features as its evaluation was better and its evaluation was done again. Similarly, the final model was built with Gradient Boosting using optimal hyperparameters and selected features. The evaluation of final model using three metrices for training data are: MAE: 0.05, RMSE: 0.06, R^2 Score: 1.00. The performance of final model was much better than improved from the other two models for the training datas.

3.3 Challenges:

There are were numerous challenges that were encountered while performing the task. Optimization of hyperparameters and selection of features were one of the main challenges.

3.4 Future Work:

Although, the performance of both the models are good, there is still room for improvement which could be helpful in future work. The use of advanced regression techniques or any other model like Random Forest could help in improvement of the model.

4. Discussion

4.1 Model's Performance:

The models used in this task Random Forest Regression and Gradient Boosting performed good based on their evaluation. After comparing their evaluation, the Gradient Boosting performed slightly better as its r2 value was greater.Likewise, the final model performed much better than the Gradient Boosting with improvement in the evaluation of metrices.

4.2 Impact of Hyperparameter Tuning and Feature Selection:

The process of hyperparameter tuning and feature selection helped to know the best hyperparameter values and best features of each model which resulted in building a final model with best hyperparameter and selected features resulting in improved output.

4.3 Interpretation of Results:

The task comes to the conclusion that the Overall_Literacy_Score can be predicted with the factors present on the dataset. The area with low literacy rate should focus on those factors where its score is less.

4.4 Limitations:

The model is too tightly fit to the testing data which might led the model to not generalize well to new data.

4.5 Suggestion for Future Research:

The future tasks and researches could perform much better by exploring any other algortithms of regression techniques.