UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE KATHMANDU

(Concepts and Technologies of AI)
5CS037

Regression Analysis (Report)

Name: Jyotika Ghale
Group: L5CG9
Student ID: 2408998
Module Leader: siman Giri
Tutor: Bibek Khanal
Submittion Date: 2025/02/11

Table of Contents

Predicting Overall Literacy Score Using Regression Techniques..3

Abstract..3

Objective:...3

Techniques:..3

Main Insights:..3

1. Introduction..3

1.1 Summary of the Issue..3

1.2 Gathering Data..3

1.3 Aim...4

2. Methodology..4

2.1 Preparation of Data..4

2.2 Examination of Exploratory Data (EDA)...4

2.3 Building the Model...6

2.4 Assessment of the Model...6

2.5 Adjusting Hyperparameters..6

2.6 Selection of Features..6

3. Conclusion:..6

3.1 Key Results...6

3.2 Efficacy of the Conclusive Model...7

3.3 Difficulties Faced...7

3.4 Upcoming Pathways..7

4. Interpretation of Results...7

4.1 Performance of the Model...7

4.2 Effect of Feature Selection and Hyperparameter Optimization................................7

4.3 Findings from the Outcomes...7

4.4 Limitations of the Study...7

4.5 Suggestions for Future Research..8

5. Conclusion..8

6. References..8

Predicting Overall Literacy Score Using Regression Techniques

Abstract

Objective:

This report intends to predict the Overall Literacy Score through regression methods by examining multiple aspects associated with digital literacy.

Techniques:

The file "digital_literacy_dataset.csv," which includes features related to digital literacy, was chosen for this examination. The research included feature selection, hyperparameter optimization, exploratory data analysis (EDA), and regression modeling through Linear Regression, Random Forest, and Ridge Regression.

Main Insights:

The models were assessed utilizing Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared (R^2). The Random Forest model attained the top R^2 score, establishing it as the most effective model. Choosing features and adjusting hyperparameters significantly contributed to enhancing model accuracy.

1. Introduction

1.1 Summary of the Issue

The objective of this project is to forecast the Overall Literacy Score by utilizing attributes from the dataset. The aim is to determine the primary elements affecting digital literacy using a regression-based method.

1.2 Gathering Data

The "digital_literacy_dataset.csv" file was obtained from a provider of educational research. It includes both numerical and categorical features that represent different facets of digital literacy.

1.3 Aim

The aim of this research is to create a predictive regression model that forecasts the Overall Literacy Score from chosen dataset features.

2. Methodology

2.1 Preparation of Data

- Missing values were examined and addressed accordingly.
- Duplicate entries were eliminated.
- Numerical attributes were normalized with StandardScaler.
- Categorical variables were transformed into numerical values when required.

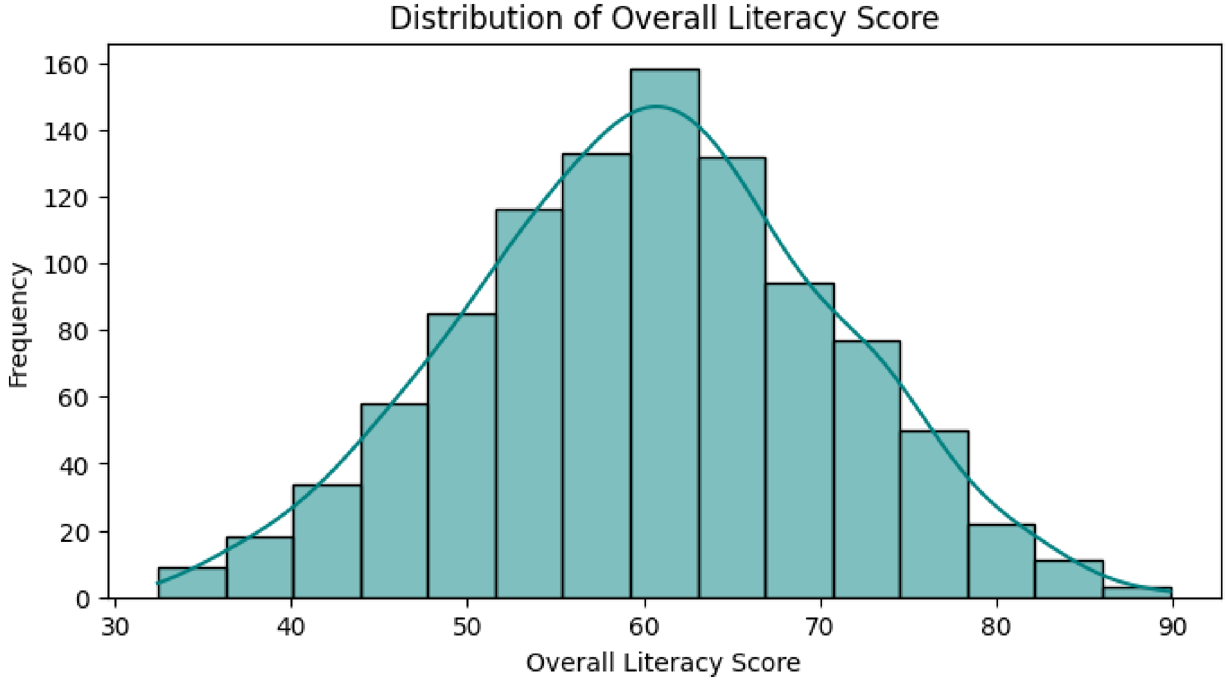
2.2 Examination of Exploratory Data (EDA)

An EDA was performed utilizing:

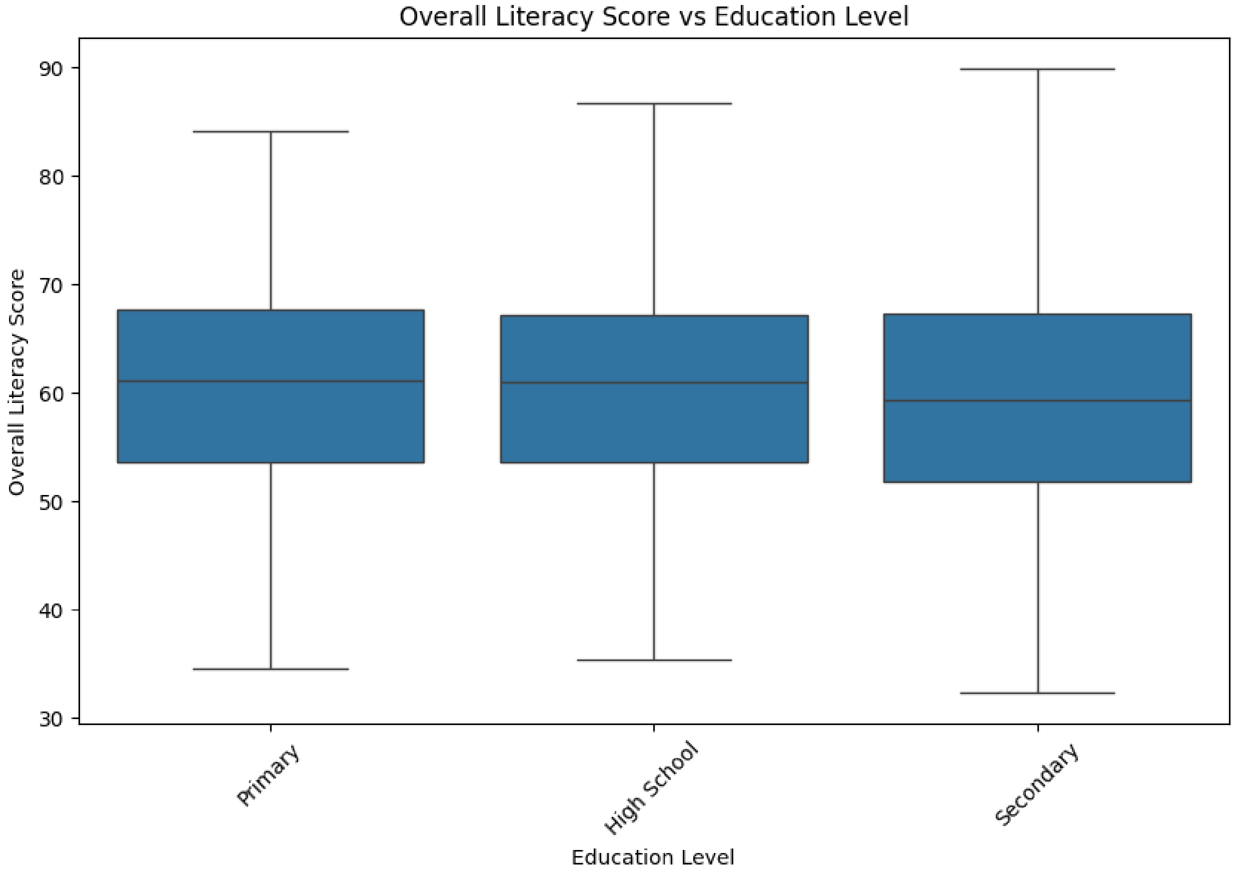
- Use scatter plots to examine connections among variables.
- Use histograms to illustrate the distributions of features.
- Heatmaps of correlation to highlight feature relationships.

Histogram plot:

Boxp



Box Plot:



2.3 Building the Model

Three regression models were examined:
- Linear Regression – An uncomplicated foundational model.
- Random Forest Regression – A model based on ensemble learning.
- Ridge Regression – A linear model with regularization designed to tackle multicollinearity.

Actions Taken:
- The dataset was divided into 80% for training and 20% for testing sets.
- Every model underwent training and assessment with the test data.

2.4 Assessment of the Model

Evaluation was conducted using:
- R-squared (R^2) – Assesses the fraction of variability accounted for by the model.
- Mean Absolute Error (MAE) – Assesses the average absolute inaccuracies.
- Root Mean Squared Error (RMSE) – Assesses the square root of the mean of the squared errors.

2.5 Adjusting Hyperparameters

The ideal hyperparameters were determined using GridSearchCV:
- Random Forest: n_estimators = 200, max_depth = 10, min_samples_split = 5, min_samples_leaf = 2.
- Ridge Regression: alpha set at 0.1.

2.6 Selection of Features

Correlation analysis was performed for feature selection, resulting in the identification of the key attributes for predicting the Overall Literacy Score.

3. Conclusion:

3.1 Key Results

- The Random Forest model achieved the highest R^2 score, surpassing the other models.
- Selecting features enhanced model performance by concentrating on essential influential factors.
- Tuning hyperparameters greatly improved the accuracy of predictions.

3.2 Efficacy of the Conclusive Model

The Random Forest model showed outstanding predictive performance and was recognized as the top model on the test dataset.

3.3 Difficulties Faced

- Managing absent values and confirming appropriate data scaling.
- Achieving a balance between model complexity and interpretability.

3.4 Upcoming Pathways

- Investigating sophisticated ensemble models, like XGBoost.
- Increasing the dataset size to improve model generalization.
- Applying deep learning methods for enhanced forecasting.

4. Interpretation of Results

4.1 Performance of the Model

The Random Forest model delivered the most precise forecasts, attaining the greatest R^2 value.

4.2 Effect of Feature Selection and Hyperparameter Optimization

- Selecting features enhanced model accuracy by eliminating unnecessary noise.
- Optimized model parameters through hyperparameter tuning, especially for Random Forest and Ridge Regression.

4.3 Findings from the Outcomes

- The research determined the main elements influencing digital literacy.
- Findings indicate that focused educational strategies may improve digital literacy.

4.4 Limitations of the Study

- Limitations in dataset size might hinder the generalizability of the results.
- Random Forest exceeded the performance of linear models, but sacrificed interpretability.

4.5 Suggestions for Future Research

• Exploring different regression methods like Gradient Boosting.
• Improving feature engineering to derive more significant insights.

5. Conclusion

This research effectively created a regression model to estimate the Overall Literacy Score. Of the models evaluated, Random Forest Regression proved to be the most efficient, attaining the highest R^2 score. Feature selection and hyperparameter optimization were essential in enhancing model performance. Future efforts should emphasize improved modeling methods, increasing dataset size, and enhanced feature selection for more accurate predictions.

6. References

https://www.kaggle.com/datasets/ziya07/digital-literacy-education-dataset