Herald College, Kathmandu

Concepts and Technologies of AI

5CS037

An End-to-End Machine Learning Project
On Regression and Classification Task.

Report: -: Regression Analysis Report on Blood Samples Dataset

Student Name: Supriya Kunwar

Group: L5CG7

Student Id: 2408484

Module Leader: Siman Giri

Abstract

Purpose:

This is a report that projects the cholesterol level through regression techniques. The focus of this research is to develop an efficient and effective predictive model using various health-related parameters to estimate cholesterol values.

Approach:

It will contain samples for blood tests that will undergo analysis with several numerical characteristics of different medical indicators. The assembles including Exploratory Data Analysis, creating a model by using Linear Regression and Decision Tree Regression, hyper-parameter tuning using GridSearchCV, and Recursive Feature Elimination for feature selection in improving the model performances will also be included in the package.

KeyResults:

The performance of the models was assessed through different metrics: R^2, mean squared error (MSE), and root mean squared error (RMSE). Results indicate that Decision Tree Regression surpassed Linear Regression since it exhibited a higher R^2 value, indicating greater predictive power and demonstrating lower values for MSE and RMSE as measures of error.

Conclusion:

Your regression model is successful in predicting cholesterol levels. Significant findings from this study are the feature selection that enhances model performance and hyper-parameter tuning, which reduces prediction errors. Decision Tree Regression seems to have performed best in this dataset. Future endeavors with ensemble models and deep learning may provide better results.

1 Introduction

1.1 Problem Statement

This project attempts to create a regression model to predict Cholesterol from other health-related parameters found in the dataset. Various regression techniques have been analyzed with the aim of optimizing model performance in order to attain reasonable predictions.

1.2 Dataset

Blood Samples Dataset is the dataset behind this analysis, which has multiple numerical and categorical features related to the health indicators. The dataset is completely reliable and accessed from a dataset repository following the UNSDG Goal 3: Good Health and Well-being to contribute towards health diagnostics and medical predictions.

1.3 Objective

This analysis should build a regression model that is appropriately predictive of cholesterol based on existing health attributes. The project will also explore some of the different regression models; hyperparameter optimization; and finally, pick the best approach.

2 Methodology

2.1 Data Preprocessing

The pre-processing of data is the key process using which we can assure that models could be accurate and reliable. Initially, the data was checked to see if there were any values recorded missing. These missing values were either removed altogether or imputed by any statistical procedure. Any categorical variables were labelled encoded. Since the numerical values apply in different units, they affect the model performance; therefore, applying a method of feature scaling normalizes the entire dataset maintaining uniformity for all variables. Utilization of outlier detection methods has also been done with respect to the extreme values contributing towards erroneous predictions.

2.4 Model Evaluation

Key metrics were utilized to evaluate the models in terms of performance:

•R^2 Score: Measures how much variation in the target variable is explained by the model.

•Mean Absolute Error (MAE): Assesses the average absolute deviation between observed and modelled values.

•Mean Squared Error (MSE): Assess the squared difference between actual and predicted values with more penalty on larger errors.

•Root Mean Squared Error (RMSE): Presents an interpretable error measure in the same units as the target variable.

Ultimately, these assessments were compared in order to find the model with the best ability for prediction of cholesterol levels.

2.5 Hyper-parameter Optimization

Hyperparameter tuning to improve model performance was done using GridSearchCV, which tested different parameter values to zero in on the best settings for each model.

• For Linear Regression, the optimum hyperparameter was the fit_intercept setting, and this led to improved accuracy in the predictions.

• While for Decision Tree Regression, the most optimal parameters were max_depth and min_samples_split, which helped boost performance by striking between complexity and accuracy.

2.6 Feature Selection

Recursive Feature Elimination (RFE) was another technique employed in selecting the most relevant features that would improve prediction accuracy and model efficiency. This made it the more interpretable and prevented overfitting by eliminating unimportant predictors. The final subset of features retained only those most relevant to predicting levels of cholesterol.

3. Conclusion

3.1 Key Findings

While evaluating the performance of the two models, Decision Tree Regression was found to surpass Linear Regression based on the R^2 scores and the error metrics. Feature selection and hyperparameter tuning were employed to enhance the accuracy of the model. The improved model was able to predict cholesterol levels with a high degree of confidence.

3.2 Final Model

During the model comparison, Decision Tree Regression was finally selected as the model of choice based on superior predictive accuracy and being able to make fit non-linear relationships. Also, the model stood as the best compromise between complexity and performance. Therefore, it was an adequate choice for predicting cholesterol levels.

3.3 Challenges

Several challenges were encountered throughout the project:

• Feature Selection: Involved deep study to come up with the significant features.

• Complexity Balance Model: The Decision Tree model needed tuning to avoid overfitting.

3.4 Future Work

Future research can involve further steps in improving model performance, which are:

• Implementation of ensemble models such as Random Forest or Gradient Boosting and application in future predictions.

• Advanced techniques of feature engineering giving meaningful information.

• Increasing the size of the dataset helps in improving the model generalization.

4. Discussion

4.1 Model Performance

The Decision Tree model was discovered to have the best score in R^2 and possessed the least prediction error, making it the most appropriate model for cholesterol level prediction. The final model is generalized unseen test data, thus proving its reliability.

4.2 Effect of Hyperparameter Tuning and Feature Selection

Hyperparameter tuning has boosted the accuracy of the model well, while feature extraction has reduced the excess number of predictors to improve interpretability and efficiency of the model, contributing to better performance of the model in total.

4.3 Results Interpretation

The correlation analysis demonstrated that certain health indicators were correlated with cholesterol levels and thus acted as primary predictors. Hence, the final model described these relationships well, confirming its usefulness in the medical diagnostics field.

4.4 Limitations

Although regression models were effectively performed in this study, some limitations exist:

- Dataset Size - A larger dataset would increase robustness.
- Feature Unavailability - There may be some health factors that affect cholesterol levels but are not represented in the dataset.
- Model Generalization - It needs to be validated on external datasets for generalization.

4.5 Suggestions for Future Work

Further improvements can be made to extend the usefulness of this research:

•Ensemble methods may provide better accuracy by combining results from different models.

•Deep learning techniques may be required to capture complex associations between health indicators.

•Longitudinal studies on cholesterol status and trends over time.

Final Thoughts

Some high-level opportunities for further enhancement arise in real-world applications of healthcare, depending on data quality and model complexity.

The primary goal of this project was to bring to fruition a regression model predicting cholesterol levels. Decision Tree Regression emerged as the final candidate through key evaluation processes and hyperparameter tuning, and selection of important features.