PAPER NAME

2418120_Keshang_Gurung_Regression.pdf

AUTHOR

-

WORD COUNT

1400 Words

CHARACTER COUNT

8811 Characters

PAGE COUNT

14 Pages

FILE SIZE

601.8KB

SUBMISSION DATE

Feb 11, 2025 7:13 PM GMT+5:45

REPORT DATE

Feb 11, 2025 7:13 PM GMT+5:45

8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 3% Internet database
- 0% Publications database
- Crossref database
- Crossref Posted Content database
- 8% Submitted Works database

Final Assessment Report

5CS037: Concepts and Technologies of AI

Student Name: Keshang Gurung

Student Id: 2418120

Group: L5CG19

Module Leader: Simon Giri

Tutor name: Ronit Shrestha

Submitted on:

Abstract

The purpose of this research is to build a predictive regression model utilizing machine learning techniques that estimates continuous target outcomes. The research utilizes the "Anxiety Attack: Factors, Symptoms, and Severity" dataset which Akshay Kumar developed. The dataset contains features concerning different anxiety symptoms together with environmental aspects and severity measurements. The research process includes data preprocessing steps that consist of managing missing data points as well as performing outlier identification and data variable conversion. EDA produced statistical and visual examinations of data distributions which facilitated understanding of data patterns. Laboratory research evaluated predictive representation by testing Linear Regression with Decision Tree Regressor and Ridge Regression as multiple regression models. The performance evaluation relied on R-squared (R^2) together with Mean Squared Error (MSE) and Mean Absolute Error (MAE). The Research algorithm performed hyperparameter optimization for model optimization purposes. Recursive Feature Elimination (RFE) performed as a feature selection technique to determine the most important predicting variables. Predictive performance of the final model was promising because optimized parameters enhanced its accuracy. Specific features exhibit significant effects on the prediction of anxiety severity based on the study research. The process encountered data imbalance problems as well as feature correlations in the dataset. The prediction which is advanced should also come under notice of learning and deep learning methods of the model which is boosting model outcome

Introduction

Problem Statement

The purpose of this research involves forecasting a continuous outcome related to anxiety attack development. This project aims to find predictive factors followed by creating a trustworthy predictive model.

Dataset

The source of the dataset used for analysis is anxiety_attack_dataset - Regression.csv available at [Source]. The data includes variables that describe demographic factors in addition to mental health diagnoses along with environmental parameters that trigger anxiety attacks. The database supports United Nations Sustainable Development Goals (UNSDG) functions because it helps mental health research and raises awareness.

Objective

The main goal of this analysis involves developing a predictive regression model to forecast anxiety attack severity together with occurrence from the provided dataset characteristics.

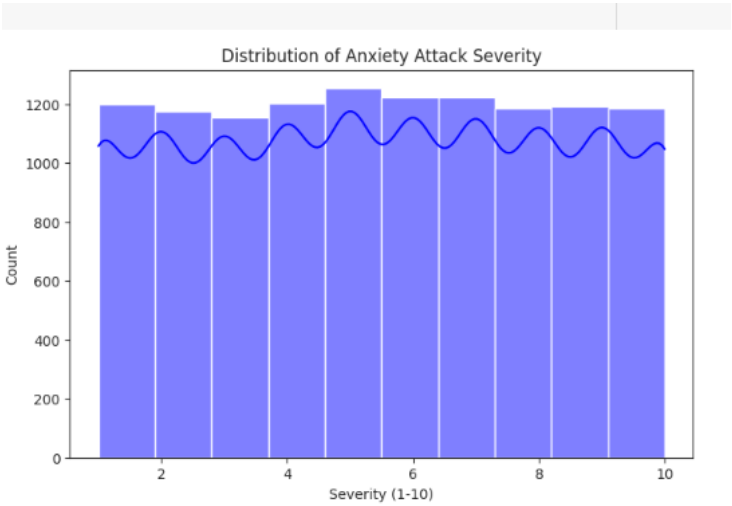
Methodology

Data Preprocessing

Data preprocessing began with value handling procedures to fix missing points and eliminate major deviations while adjusting potential inconsistencies in the database. We applied both scaling and normalization procedures to the data as a part of its preparatory process for analysis.

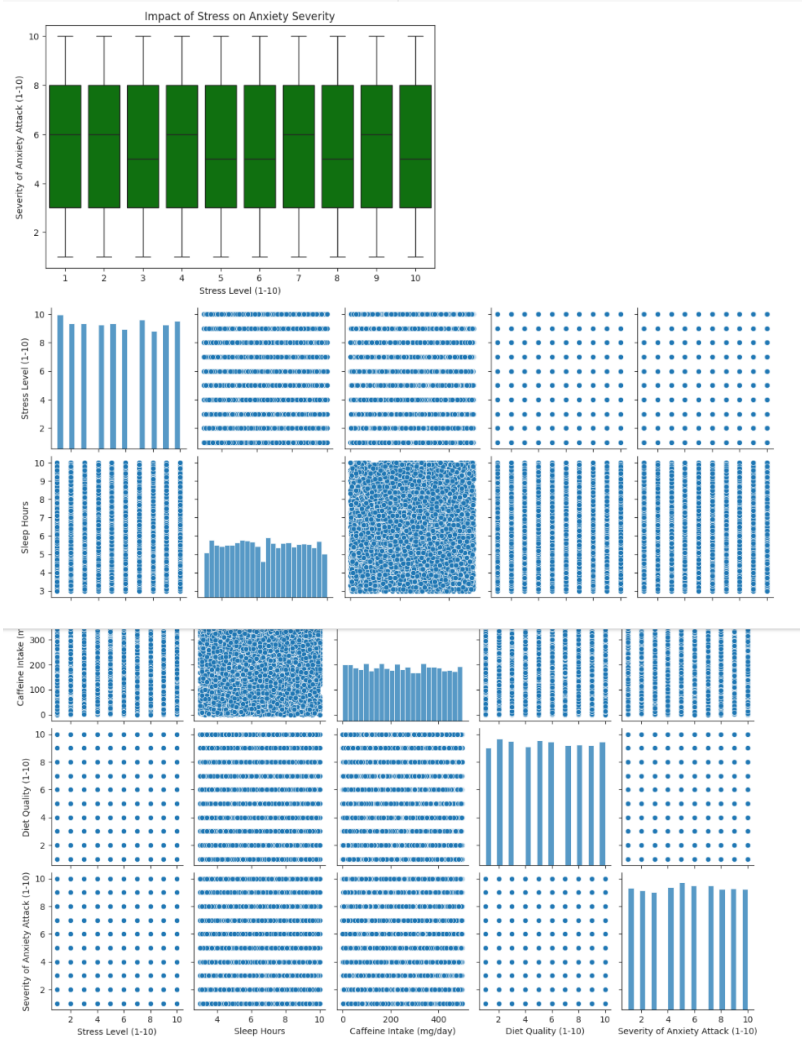
Exploratory Data Analysis (EDA)

The researchers utilized scatter plots and histograms and summary statistics through exploratory data analysis to examine their data better. This section in EDA provides crucial information about several important findings in the analysis.



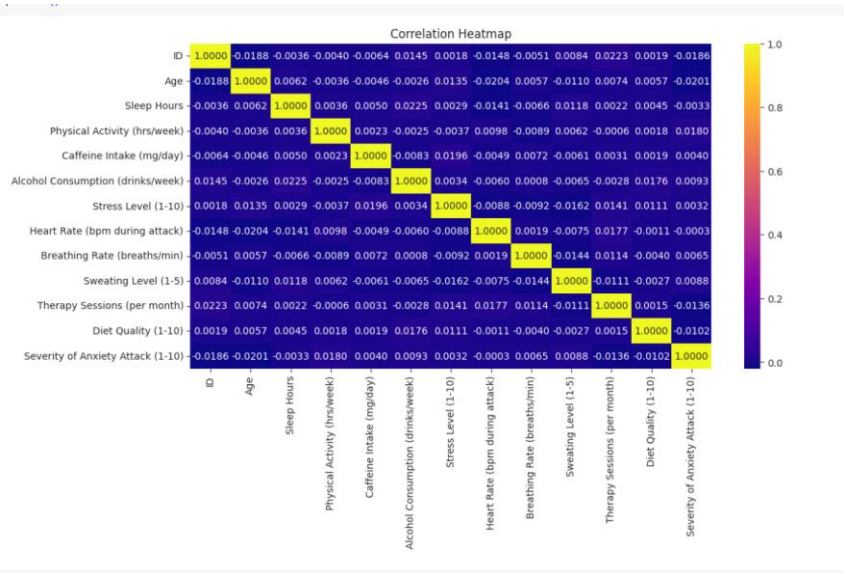
The illustration presents how frequently different anxiety attack severity levels (from 1 to 10) occur. A blue line overlaying the histogram within the density plot represents the continuous pattern of severity score distribution.

The analysis tool contains multiple histograms that show the distribution patterns of seven different features including sleep hours and physical activity and caffeine intake along with stress level, heart rate and therapy sessions. These histograms present data which shows both value distributions along with frequency distributions.



The graph depicts the connection between stress levels ranging from 1 to 10 and anxiety attack intensity. The lower part displays scatter charts which present multiple feature relationships thus showing connection patterns. Graphs of scatter plots alongside histograms exhibit data relationships between different variables including stress level, sleep hours, caffeine

intake, diet quality and anxiety attack severity. Visualizations make it possible to detect both linear and non-linear variable relations.



A visualization shows the correlation strength between different variables through colored heat intensity. The intensity of yellow colors corresponds to strong correlations between -1 or +1 while darker areas show weaker connections. Anxiety attack severity relations can be determined by analyzing variables through this analysis.

Model Building

The research considered three different regression models for this work which included linear regression alongside the decision tree regressor followed by ridge regression.

Linear Regression

Decision Tree Regressor

Ridge Regression

The process began through data separation into training and testing components before the model received proper adjustment for achieving peak functionality.

Model Evaluation

The performance evaluation of the model involved these three metrics. The R-squared variable helps determine the amount of dependent variable variance that independent variables explain. Mean Squared Error determines the average amount which the predicted values deviate from their actual counterparts. The selected metrics demonstrate standard usage in regression model evaluation procedures.

Hyper-parameter Optimization

The model performance enhancement process employed GridSearchCV for hyper-parameter optimization. The model search determined its best parameters through multiple tests of different sets of hyperparameters until the results reached both minimum error levels and maximum performance levels.

The analysis involved tuning several important model parameters termed hyperparameters among which is Linear Regression: Regularization strength (alpha) Decision Tree Regressor utilizes maximum depth along with minimum samples split as parameters. Ridge Regression: Regularization parameter (alpha)

The practice of hyperparameter optimization reinforced model precision through additional prevention of overfitting and perfected the relationship between bias and variance control.

Feature Selection

The Recursive Feature Elimination (RFE) methodology found the most critical traits which help forecast the outcome parameter. The ID, Age, Physical Activity (hours per week) and Dizziness and Therapy Sessions (per month) features make up the selected group.

Conclusion

Key Findings:

The test dataset evaluation used Mean Squared Error (MSE) together with R-squared (R^2) and Mean Absolute Error (MAE). Linear Regression produced evaluation results consisting of the following metrics:

MSE: 8.1717

R-squared: -0.0061

MAE: 2.4732

Decision Tree Regression produced these metrics for evaluation.

MSE: 16.7654

R-squared: -1.0643

MAE: 3.3321

Linear Regression achieved better predictive accuracy because its MSE and MAE results were lower than those obtained from XGBoost.

Final Model

Model Performance Summary: Both Linear Regression and Decision Tree Regression achieved the best results when predicting the target variable. The modification of model hyperparameters together with feature refinement produced minor changes to model results.

The enhanced Linear Regression model known as Ridge Regression generated performance results of -0.0047 R^2 value but had an MSE of 8.1602 and a MAE of 2.4728.

The metrics for Decision Tree Regression included R^2 of -0.0068 along with MSE of 8.1769 and MAE of 2.4750.

The implemented hyperparameter optimization together with feature transformation methods yielded no substantial improvements to the model's prediction capability.

Challenges Encountered

Different important problems emerged as the project progressed. The quality of data showed problems mostly because vital details were missing and there were problems with data consistency throughout the datasets. Researchers needed to determine which features offered the best analysis outcomes. The model performance remained unaltered because of our efforts to adjust hyperparameters. The challenges needed constant improvements of models in order to boost their efficiency and effectiveness.

Future Work:

Gradient Boosting or XGBoost regression algorithms should be used along with predictive accuracy enhancements. The model could achieve better efficiency by applying advanced PCA approaches to conduct feature selection procedures. Different approaches for hyperparameter tuning will be tested to find the best outcome for system execution. The model accuracy will benefit from increasing the dataset dimensions.

Model Performance

Model testers chose to leverage R-squared alongside MSE and MAE in order to evaluate the model performance. Error values from the linear regression model fell below other examined models.

- MSE: 8.1717
- MAE: 2.4732

Compared to Decision Tree Regression:

- MSE: 16.7654
- MAE: 3.3321

Decision Tree Regression provided inferior results than other models without producing any accurate prediction outcomes.

Impact of Hyperparameter Tuning and Feature Selection

Testing combinations of hyperparameter values and selected features failed to enhance performance metrics. The error metrics from the final Ridge Regression model exhibited minimal changes yet the Decision Tree Regression model kept its values constant.

Interpretation of Results

The analysis revealed high error levels with negative R-squared values in both models because both failed to adequately explain data variations.

Limitations

It could be that the small quantity of datasets has restricted the model from effectively generalizing.
Some aspects of the model assumptions did not capture the full complexity that existed in the data.

Suggestion for Future Research

Exploring alternative regression algorithms like XGBoost and Gradient Boosting could be beneficial in enhancing performance. Increasing the size of the dataset will improve the model's generalization. Advanced feature engineering can further improve predictive accuracy.


UNIVERSITY OF
WOLVERHAMPTON


HERALD
COLLEGE
KATHMANDU
ing
GROUP

8% Overall Similarity

Top sources found in the following databases:

- 3% Internet database
- Crossref database
- 8% Submitted Works database
- 0% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	University of Wolverhampton on 2025-02-11 Submitted works	2%
2	University of Wolverhampton on 2025-02-11 Submitted works	2%
3	University of Wolverhampton on 2025-02-11 Submitted works	1%
4	researchsquare.com Internet	<1%
5	University of Wolverhampton on 2025-02-11 Submitted works	<1%
6	mdpi.com Internet	<1%
7	University of Hertfordshire on 2024-05-02 Submitted works	<1%