PAPER NAME

2407749_SubashBikramTamang_Classification_Report-4.pdf

AUTHOR

-

WORD COUNT

850 Words

CHARACTER COUNT

4766 Characters

PAGE COUNT

6 Pages

FILE SIZE

163.2KB

SUBMISSION DATE

Feb 11, 2025 5:51 PM GMT+5:45

REPORT DATE

Feb 11, 2025 5:51 PM GMT+5:45

13% 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
- 11% Submitted Works database

Excluded from Similarity Report

- Manually excluded sources

UNIVERSITY PARTNER

UNIVERSITY OF WOLVERHAMPTON

HERALD COLLEGE KATHMANDU

Concepts and Technologies of AI

AI: Classification Analysis Report

Student ID: 2407749

Student Name: Subash Bikram Tamang

Course Code: 5CS037

Abstract

The purpose of this project is to predict a category variable using some classification approaches. Materials and Methods: A medical dataset has to be used for the prediction of outcomes of patients. Thus, a healthcare dataset will be selected for this research study. The procedures include EDA, feature selection, hyperparameter tweaking, model construction using decision trees and logistic regression. The performance metrics that both models share is the following: accuracy, precision, recall, and F1 score. Performances for the two models differ; where on most fronts, the decision tree outperforms logistic regression. Tuning of hyperparameters.

1. Introduction

1.1 Problem Statement

This present study used classification algorithms in the determination of the health outcomes of the patient based on the information provided. The classification will be helpful in identifying patients who are at danger, which will aid in the medical decision-making process.

1.2 Dataset

The analysis will make use of the Kaggle Healthcare Dataset, which consists of 55,500 rows and 15 columns of features on the health status of the patients. This data responds to UNSDG in ensuring a healthy life and well-being.

1.3 Aim

This paper aims to design a predictive regression model that would be able to calculate Thus, the aim through this analysis will be to develop from it a predictive classification model which would determine probabilities of certain medical diseases based on patient data.

2. Methodology

2.1 Data Cleaning and Preprocessing

The data has been cleaned in preparation for analysis by mainly managing missing values and outliers, among other similar transformations such as normalization.

2.2 Exploratory Data Analysis

EDA was conducted to get more insights from the data through the use of visualizations like correlation matrices, bar charts, and histograms. Some of the key findings of the analysis are:

- There is a high correlation between the target variable and either one or just a few features the class distribution was highly unbalanced, thus the resampling approaches had to be applied.

2.3 Building the Models

The two models to be compared in classification are the following:

- Decision Trees
- Logistic Regression

The split of the dataset into an 80-20 split was done, and these models were trained on the training set.

2.4 Model Evaluation

The selected classifier performance is evaluated based on the metrics listed here:
• Accuracy: It is the ratio of total forecasts to correct guesses.
• Precision: This gives the percentage of the positive predictions by the algorithms that actually turn out to be true.
• Recall: This tells the ratio of real positives being predicted as positive correctly.
The F1-score is the average of recall and precision.

2.5 Hyperparameter Optimization

This model can be further improved by using techniques for hyperparameter optimisation like GridSearchCV. The best parameters for the Decision Tree are as follows:

- Maximum Depth - 5
- Minutes Samples Divided - 10

2.6 Feature Selection

Feature selection by means of Recursive Feature Elimination was carried out to go into further detail which features were more informative with regards to explaining variation in the target variable. Among the features that were chosen:

- Carats
- Slice
- Colour
- Clarity

3. Conclusion

3.1 Key Findings

Furthermore, the model performance, based on accuracy, precision, recall, and F1-score, was measured on the test dataset. In the non-linear relationship, Decision Trees outperformed Logistic Regression.

3.2 The Final Model

The final Decision Tree-based model has achieved remarkable efficacy up to 85% regarding the forecast of patient health outcomes.

3.3 Challenges

The major issues faced in the project are listed below:

- Model performance was dependent upon unbalanced data.
- It required a huge amount of time and effort for hyperparameter tuning to get the best set.

3.4 Future Work

Some suggestions regarding improvement in this respect are given below:

- Inclusion of Random Forests and other state-of-the-art classification techniques
- Synthesis of data for overcoming imbalance among classes.

4. Discussion

4.1 Evaluation of Model Performance

Assessment Recall, accuracy, precision, and F1-score of the model were evaluated. Based on the results of this evaluation, it can be observed that Decision Trees outperformed Logistic Regression in identifying the nonlinear relationship of the data.

4.2 Hyperparameter Tuning and Feature Selection

Feature selection and hyperparameter optimisation significantly enhanced the performance of the model. The model accuracy increased by 10% using these techniques.

4.3 Result Interpretation

The chosen model and characteristics did not disappoint. Some of the key results from the analysis show that these markers of health bear major significance toward the patient's outcomes.

4.4 Limitations

- Even with effective modelling, generalisability is violated due to the approach's constraints.
- The data set is small and might be biassed because it lacks demographic characteristics.

4.5 Limitations of Future Research

Following are some of the questions that were not addressed and can be considered in future studies:

- Different experiments by using ensemble classification algorithms.
- Big datasets with a lot of different patient records.
- Employ models with confidence distributions that allow for deep learning classification modeling for higher precision.

13% Overall Similarity

Top sources found in the following databases:

- 3% Internet database
- Crossref database
- 11% 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 2025-02-03 Submitted works	6%
2	University of Wolverhampton on 2025-02-09 Submitted works	3%
3	The University of Wolverhampton on 2024-02-14 Submitted works	1%
4	University of Wolverhampton on 2025-02-03 Submitted works	1%
5	ecgi.global Internet	1%
6	hdl.handle.net Internet	<1%

Excluded from Similarity Report

- Manually excluded sources

EXCLUDED OVERLAPPING SOURCES

Submitted to University of Wolverhampton on 2025-02-11 Submitted works	29%
Submitted to University of Wolverhampton on 2025-02-11 Submitted works	29%
Submitted to University of Wolverhampton on 2025-02-11 Submitted works	28%
Submitted to University of Wolverhampton on 2025-02-11 Submitted works	27%