2024

DATA MINING-REPORT

Muhammad Nadeem

Mehar Hamid Ishfaq

TABLE OF CONTENTS

INTRODUCTION

Exploratory Data Analysis (EDA) on the Heart Disease dataset revealed valuable insights into various health-related parameters and their association with the likelihood of heart disease. This analysis aimed to enhance skills in data exploration and contribute to a better understanding of the complex interplay of factors influencing cardiovascular health. The dataset, initially containing 1024 rows and 14 columns, underwent thorough cleaning, including the removal of 723 duplicate entries.

DATA PRE-PROCESSING

Data preprocessing played a pivotal role in preparing the dataset for predictive modeling. Key steps included handling outliers, visualizing the distribution of numerical and categorical variables, and identifying relationships between features. Categorical variables were converted for better interpretability, and numerical variables were standardized for modeling. Additionally, the presence of outliers and correlations between variables were carefully considered.

DATA PRE-PROCESSING

In the code, several preprocessing tasks were implemented to ensure data quality and prepare the dataset for further analysis. These tasks included handling duplicate entries by identifying and removing duplicate rows using the drop_duplicates() method to maintain analysis integrity. Additionally, categorical variables were renamed to enhance readability and simplify interpretation, such as converting numeric representations of binary categories to intuitive labels ('M' for male and 'F' for female). Selected numerical variables were converted into categorical ones for better visualization and analysis, such as converting the target variable from numerical (0, 1) to categorical ('N', 'Y') for clearer communication of insights. Extensive exploratory data visualization was conducted to understand the distribution of numerical and categorical variables, identify outliers, and explore relationships between variables using various plots like histograms, density plots, box plots, bar plots, and scatter plots. Outliers were identified through visualization and analysis, indicating the presence of unusual or extreme values, which may require further investigation. Correlation analysis was performed to identify relationships between numerical variables, visualized using a correlation heatmap to identify variables with significant correlations, aiding in feature selection and building predictive models.

PREDICTIVE MODELING

Two machine learning models, Logistic Regression and Random Forest, were employed for predictive modeling. The features were preprocessed using a combination of standard scaling for numerical variables and one-hot encoding for categorical variables. The models were trained on the preprocessed data, and their accuracy was evaluated on the test set.

No. 01 — Logistics Regression

The Logistic Regression model achieved an accuracy score of 0.80.
This model is suitable for binary classification tasks and provides insights into the impact of different features on the likelihood of heart disease.

No. 02 — Random Forest

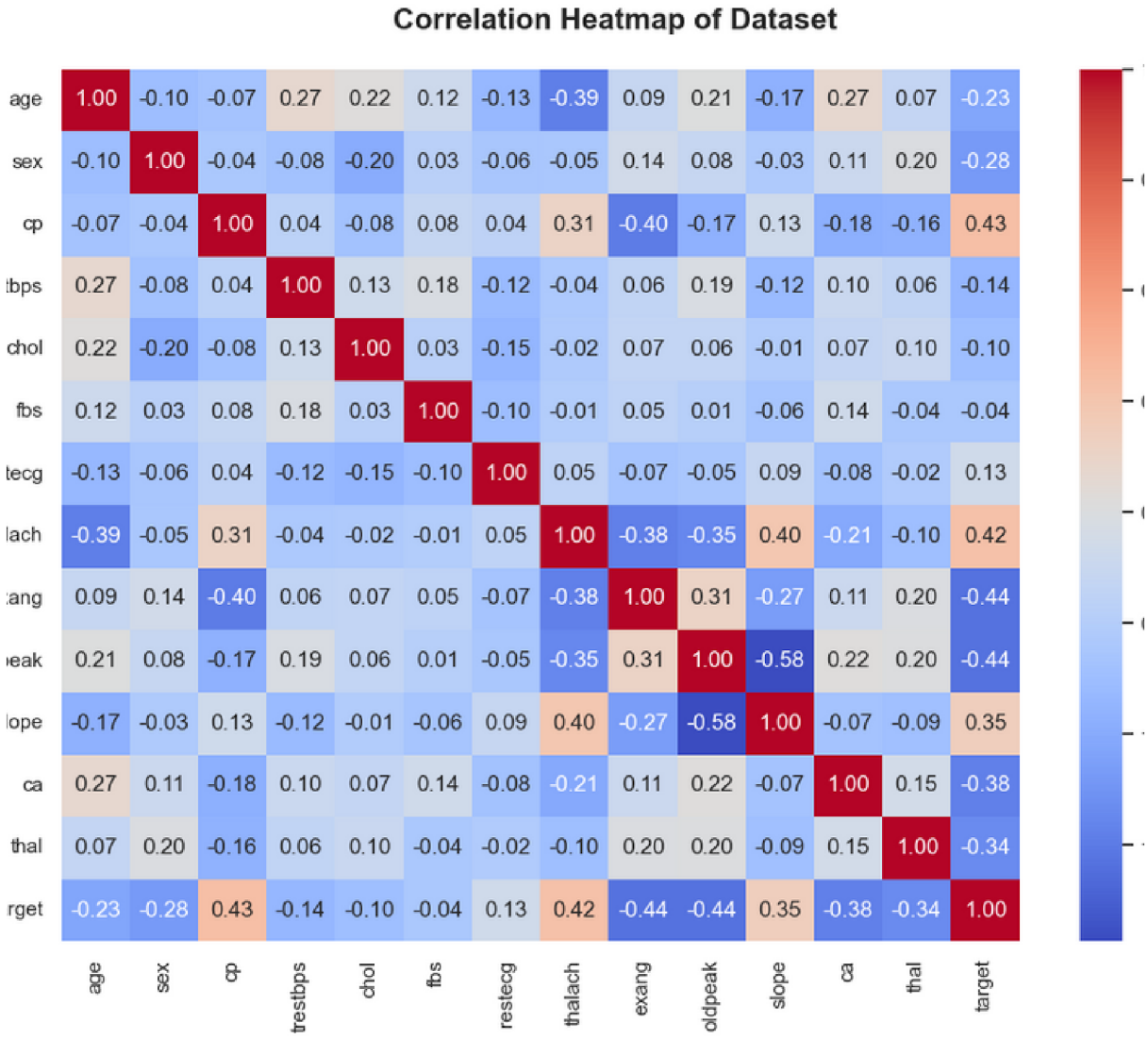
The Random Forest model demonstrated an accuracy score of 0.81.
By leveraging an ensemble of decision trees, Random Forest excels in capturing complex relationships within the data, contributing to its predictive performance.

Handling Class Imbalance

I do not have any missing value. In real-world scenarios, handling class imbalance is crucial, as imbalanced datasets can lead to biased models. Techniques such as oversampling, undersampling, or the use of advanced algorithms like SMOTE can be implemented to address this issue.

In above scenerio We have a class that is imbalanced Attribute name is “Class” and classes is “No Fraud and Fraud ” Check the number of count in this above scenerio it shows that “No Fraud” has more instances then that “Fraud” has less instances. it may affect on your Prediction When We train and Test our Model.

Relationships Between Attributes



100% *Here is above you Find the relationships Between All attributes that is in our Dataset.*

NEXT STEPS

As we move forward, our next steps involve building upon the foundation laid by our data mining efforts. Sustainability reports are not just about reflecting on past achievements but also about looking forward and outlining strategies for continuous improvement. This Report serves as a dynamic tool for tracking our impact and progress over time. Our strategy includes

No. 01 — Action

Implementing targeted interventions based on insights gained from predictive modeling to address specific health risk factors identified in the EDA.

No. 02 — Action

Collaborating with healthcare professionals and stakeholders to develop tailored interventions and preventive measures aimed at reducing the prevalence of heart disease within our community.

No. 03 — Action

Continuing to monitor and evaluate the effectiveness of implemented interventions through ongoing data collection and analysis, ensuring adaptive management and continuous improvement.

CONCLUSION

- Exploratory Data Analysis (EDA) provided valuable insights into various health-related parameters and their association with the likelihood of heart disease.
- Data preprocessing played a pivotal role in preparing the dataset for predictive modeling, including handling outliers, visualizing distributions, and identifying relationships between features.
- Logistic Regression and Random Forest models were employed for predictive modeling, achieving accuracy scores of 0.80 and 0.81, respectively.
- The presence of class imbalance was addressed through techniques such as oversampling, undersampling, or the use of advanced algorithms like SMOTE.
- Relationships between attributes were explored to understand their impact on heart disease likelihood, guiding the development of targeted interventions.
- Next steps involve implementing targeted interventions, collaborating with stakeholders, and continuing to monitor progress towards reducing the prevalence of heart disease.

- In conclusion, our data mining efforts have provided valuable insights into the complex interplay of factors influencing cardiovascular health. By leveraging data-driven approaches, we have identified key risk factors and developed predictive models to support targeted interventions and preventive measures. As we continue our journey towards achieving the Sustainable Development Goals (SDGs), we remain committed to making significant strides in improving public health and well-being. Together, with the support of our stakeholders and partners, we are confident in our ability to create positive impact and contribute to a healthier future for all.

ACKNOWLEDGEMENTS

- We extend our sincere gratitude to all those who have contributed to this project, including our dedicated team of researchers, writers, designers, and collaborators from local and partner organizations. We also thank our contributors and donors for their invaluable support in our mission to advance the Sustainable Development Goals (SDGs) and improve lives globally.

We thank you for your continued support in our efforts to contribute to this report.

Contact

Muhammad Nadeem, Mehar Hamid Ishfaq

201980050@gift.edu.pk

201980038@gift.edu.pk

linkedin.com/in/muhammad-nadeem-5a1517242

github.com/NadeemMughal?tab=repositories