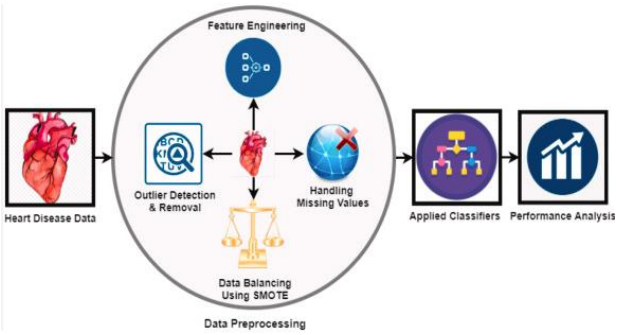

VitalFlow Guardian: Data-Powered Cardiac Risk Management

Bellevue University
DSC630 - Predictive Analytics

Submitted By:
Debabrata Mishra

Instructor:
Andrew Hua

Table of Contents

1.0 Introduction

Cardiovascular diseases result in the loss of about 17 million lives globally every year, predominantly through events like myocardial infarctions and heart failures. Among these, heart failure (HF) occurs when the heart struggles to adequately pump blood to meet the body's needs. Addressing this critical health issue involves tapping into electronic medical records, a promising resource for meticulous analysis. These records compile a vast dataset encompassing symptoms, physiological traits, and vital clinical test results. By employing biostatistical analysis, hidden patterns and correlations surface, often beyond the scope of even the most adept medical practitioners.

The emergence of machine learning has marked a transformative phase, enabling predictions of patient survival based on this extensive data repository. In the United States, heart disease remains a leading cause of mortality, demanding a holistic understanding of its multifaceted nature and interconnected contributing factors. Our pursuit involves unraveling the intricate web of variables that impact cardiac health, empowering individuals with insights crucial for informed decisions and fostering enduring heart health.

"VitalFlow Guardian," a dedicated initiative committed to crafting predictive models for cardiac risk management. Through meticulous data analysis and cutting-edge machine learning methodologies, our objective is to unearth invaluable insights. These insights won't just aid in early detection but will also pave the way for tailored interventions, ultimately saving lives and fortifying the health and longevity of individuals worldwide.

2.0 Data Selection

The dataset chosen for my project is an intricately curated compilation tailored specifically for in-depth analysis of heart health. With a total of 319,795 entries and 18 columns, it distinguishes itself with its comprehensive coverage, encompassing various demographic and clinical factors, and amalgamating five distinct heart-related datasets. This dataset serves as an indispensable resource for my predictive modeling endeavors. It encompasses crucial attributes such as age, gender, chest pain type, blood pressure, cholesterol levels, and more. Its significance lies in being one of the largest datasets on heart disease, providing me with a sturdy foundation for constructing predictive models and extracting invaluable insights into the determinants of heart disease.

There are no null values, eliminating the need to drop any rows. Most columns exhibit two unique values, Yes and No. The heart disease column, specifically, manifests a significant

imbalance with 292,422 instances of No and only 27,373 instances of Yes, indicating an unbalanced dataset. Strategies such as under-sampling or over-sampling will be necessary during the data transformation phase.

Given the dataset's imbalance, oversampling was implemented to achieve balance, a crucial adjustment to prevent model bias toward the majority class. Some categorical variables were encoded as binary (0 and 1) instead of creating dummy variables, simplifying the dataset and reducing dimensionality without compromising pertinent information. Continuous features underwent scaling to a range between 0 and 1, ensuring uniform impact across models.

3.0 Modeling and Methods

I have utilized a diverse set of machine learning models for VitalFlow Guardian: Data-Powered Cardiac Risk Management, encompassing Logistic Regression, Decision Trees, Random Forest, XGBoost, and Gradient Boosting Machine (GBM). Each model serves a distinct purpose in identifying and estimating critical factors influencing the likelihood of heart failure in individuals.

Logistic Regression: Logistic regression acts as the foundational model for binary classification, distinguishing individuals at risk of heart risks from those not at risk. This model offers a clear insight into how individual features impact heart failure prediction, aiding in feature selection and enhancing interpretability.

Decision Trees: Decision trees capture non-linear relationships and interactions among features, enabling the recognition of complex patterns. They amplify our understanding of intricate data relationships and can be visually represented, facilitating easy interpretation.

Random Forest: Random Forest, an ensemble model, amalgamates multiple decision trees to enhance predictive accuracy and mitigate overfitting. By harnessing the strengths of decision trees, it fortifies the model's resilience and generalization.

XGBoost (Extreme Gradient Boosting): Implements gradient boosting for enhanced performance, efficiency, and model accuracy. This maximizes predictive power by systematically improving weak learners, refining model predictions iteratively.

Gradient Boosting Machine (GBM): Like XGBoost, GBM employs boosting techniques to enhance model performance by sequentially improving weak learners. This enhances model accuracy through a series of iteratively refined predictions.

"VitalFlow Guardian- Cardiac Risk Management" project undergoes a thorough evaluation encompassing a diverse range of assessment methodologies. Our comprehensive strategy integrates standard performance metrics, cross-validation techniques, and model-specific evaluations.

I have employed the fundamental performance metrics such as accuracy, precision, recall, F1-score, and ROC AUC, to gain a holistic view of our model's predictive capabilities. These metrics serve not only to measure overall performance but also provide nuanced insights into different facets of heart failure prediction.

The incorporation of confusion matrices is pivotal in our evaluation toolkit. These matrices visually represent true positives, true negatives, false positives, and false negatives, offering profound insights into our models' prediction mechanisms.

In essence, evaluation approach amalgamates past learnings to forge a robust and enlightening assessment of our prediction models. This multifaceted strategy, embracing diverse techniques and metrics, aims to furnish accurate and actionable insights, thereby enhancing heart health outcomes.

4.0 Results Interpretation

In my analysis, I undertook a critical step of partitioning the dataset into two fundamental subsets: the training set and the test set, with the intent of subjecting our predictive models to rigorous evaluation. This partitioning was performed with a meticulous balance, ensuring that the class distribution of heart disease and non-heart disease cases was maintained consistently between the training and test datasets. The training dataset, representing 70% of the total data, contained 192,166 instances classified as non-heart disease and 192,072 cases with a confirmed presence of heart disease. This partitioning scheme was mirrored precisely in the test dataset, comprising 82,290 nonheart disease and 82,384 heart disease cases, allowing us to maintain an equilibrium of cases across the two classes.

Subsequently, I sought to assess the performance of five distinct machine learning models in the context of heart disease prediction. The Random Forest Classifier stood out with its exceptional performance, exhibiting an accuracy rate of 96,21% when applied to the testing dataset. This remarkable accuracy rate indicated the model's profound ability to differentiate between heart disease and non-heart disease cases with great precision. For non-heart disease cases, it achieved a perfect precision score of 1.00, highlighting its capability to accurately classify instances as non-heart disease. Furthermore, the model delivered a commendable precision score of 0.93 for heart disease cases, demonstrating

its proficiency in distinguishing these cases as well. In addition to these metrics, the Random Forest Classifier excelled in terms of recall and F1 scores, further emphasizing its reliability.

	Model	Accuracy	Cross Validation Score	ROC AUC Score
0	Logistic Regression	76.05%	83.54%	76.05%
1	Decision Tree	94.82%	94.39%	94.82%
2	Random Forest Classifier	96.22%	99.58%	96.22%
3	XGBoost (Extreme Gradient Boosting):	78.26%	86.12%	78.25%
4	Gradient Boosting Machine (GBM)	76.47%	83.95%	76.47%

Based on the feature importance values obtained from your model, it appears that the most important features for predicting heart disease are as follows:

- BMI (Body Mass Index)
- Age Category
- Sleep Time
- Physical Health
- Mental Health

These features have been ranked based on their importance in the model's decision-making process, with BMI being the most important feature, followed by Age Category and Sleep Time. It's essential to note that these importance values are relative to the specific model and dataset used, and interpretations should be made in the context of the model's performance and domain expertise.

5.0 Conclusion

As we conclude the initial phase of our project, it is evident that the Random Forest Classifier has exhibited exceptional predictive power with a 96.21% accuracy rate. This signifies its pivotal role in detecting potential heart disease cases. To build on this success and maximize the accuracy of our predictions, we propose several recommendations.

First and foremost, it is imperative to expand the dataset to include a more comprehensive and diverse set of health records. This will enable the model to capture a broader spectrum of patient characteristics, symptoms, and risk factors, enhancing its ability to detect heart disease accurately. Collaboration with healthcare institutions and organizations for data acquisition should be considered to achieve this goal.

Feature engineering is another avenue to explore, aiming to identify the most influential variables contributing to heart disease. A deeper understanding of the features with the

highest predictive power can guide healthcare practitioners in risk assessment and diagnosis.

Model maintenance is crucial to ensure long-term performance and reliability. Regular updates and retraining are recommended to adapt to evolving trends and data patterns. As heart disease research advances, incorporating the latest medical insights into the model can lead to more accurate predictions.

The healthcare industry is continuously evolving, with new diagnostic techniques and treatments emerging. Collaborating with medical professionals and experts to refine the model and align it with current clinical standards is advisable. This collaboration can facilitate the development of a tool that not only predicts heart disease but also offers valuable insights to healthcare providers for early intervention and tailored treatment strategies.

In conclusion, the Random Forest Classifier has proven to be a promising tool for heart disease prediction. By implementing these recommendations, we can enhance its capabilities, contributing to more precise and proactive healthcare interventions in the battle against heart disease.

6.0 References

1. https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease
2. https://www.kaggle.com/code/andls555/heart-disease-prediction
3. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction
4. https://archive.ics.uci.edu/dataset/45/heart+disease
5. https://towardsdatascience.com/heart-disease-uci-diagnosis-prediction-b1943ee835a7
6. https://www.analyticsvidhya.com/blog/2022/03/logistic-regression-on-uci-dataset/
7. https://www.sciencedirect.com/science/article/abs/pii/S0010482521004662
8. https://www.ndc.scot.nhs.uk/docs/Heart%20Failure%20Dataset.pdf
9. https://www.sciencedirect.com/science/article/pii/S2001037016300460
10. https://www.nhlbi.nih.gov/research/heart-failure
11. https://www.ahajournals.org/doi/full/10.1161/CIRCRESAHA.121.318172
12. https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction