Heart Watch: A Data-Driven Approach to Heart Failure Prevention

Shanthibooshan Subramanian

Bellevue University

Predictive Analytics 630

Andrew Hua

Heart Watch: A Data-Driven Approach to Heart Failure Prevention

Final Project Paper

Introduction

The venture, "Heart Watch: A Data-Driven Approach to Heart Failure Prevention," emerges from the critical need to combat cardiovascular diseases, which claim around 17 million lives annually. With a specific focus on heart failure, the project aims to leverage electronic medical records for precise analysis and prediction. By employing rigorous data analysis and advanced machine learning techniques, the goal is not only to detect heart failure early but also to guide personalized interventions, ultimately contributing to the reduction of fatalities and enhancing global heart health. The dataset chosen for this endeavor combines five independent heart-related datasets, selected for their depth of information and potential to reveal intricate patterns.

Data Selection

The dataset, selected for my project, is a meticulously curated compilation tailored for in-depth analysis of heart health. Comprising 319,795 entries and 18 columns, it stands out for its comprehensive nature, covering demographic and clinical factors and consolidating five independent heart-related datasets. It serves as a vital resource for my predictive modeling endeavor. The dataset includes key attributes like age, gender, chest pain type, blood pressure, cholesterol levels, and more. Its significance lies in being one of the largest heart disease datasets, offering me a robust foundation for developing predictive models and gaining valuable insights into factors influencing heart disease.

Modeling and Methods

The "Heart Watch" modeling phase involves a systematic approach utilizing three distinct machine learning models: Logistic Regression, Decision Trees, and Random Forest.

I will utilize the Logistic Regression model for binary classification to identify individuals at risk of heart failure. Its application of logistic functions will allow me to gain insights into how individual features impact heart failure prediction. Moreover, its interpretability will be a valuable asset in guiding my feature selection process.

The Decision Trees component will be instrumental in capturing non-linear relationships and interactions among features. This is crucial for recognizing complex patterns within the dataset. The visual representation of Decision Trees will aid me in unraveling intricate relationships in the data, providing deeper insights that can guide subsequent decision-making processes.

To enhance predictive accuracy and mitigate overfitting, I will incorporate the Random Forest model as an ensemble. By leveraging the strengths of multiple Decision Trees, this approach aims to create a robust and generalized model capable of effectively distinguishing between heart disease and non-heart disease cases.

My evaluation methodology will encompass standard performance metrics such as accuracy, precision, recall, F1-score, and ROC AUC. This multifaceted approach will ensure a nuanced understanding of my models' predictive capabilities, offering insights into both overall accuracy and sensitivity to specific aspects of heart failure prediction.

To fortify the robustness of the models, k-fold cross-validation is implemented. This technique involves partitioning the dataset into k subsets and iteratively using k-1 subsets for training and the remaining subset for validation. This process is repeated k times, ensuring that each subset serves as the validation data exactly once. The results are averaged across all iterations, providing a reliable estimate of the models' performance. The confusion Matrices are pivotal in visualizing the models' true positive, true negative, false positive, and false negative predictions. This visualization aids in understanding how well the models are making predictions and which aspects may require further refinement.

Results Interpretation

In the culmination of the "Heart Watch" project, the results derived from the Random Forest Classifier hold profound implications for heart failure prediction. The model exhibits an exceptional accuracy rate of 97%, a testament to its robustness and effectiveness. This high accuracy indicates the model's ability to discern between individuals with and without heart disease, showcasing its potential for real-world applications. Precision, recall, and F1 scores provide a more nuanced understanding of the model's performance. The Random Forest Classifier achieves a perfect precision score of 1.00 for non-heart disease cases, indicating that when the model predicts an individual as not having heart disease, it is highly accurate. This precision is a critical metric in scenarios where avoiding false positives is crucial.

For heart disease cases, the model delivers a commendable precision score of 0.96. This high precision signifies the model's proficiency in correctly identifying individuals with heart disease. The balance between precision and recall suggests that the Random Forest Classifier not only accurately classifies non-heart disease instances but also demonstrates sensitivity to the presence of heart disease.

The F1 score, which considers both precision and recall, further emphasizes the model's reliability. The Random Forest Classifier's F1 score corroborates its effectiveness in providing a harmonious balance between precision and recall, solidifying its role as a powerful tool for heart disease prediction.

Conclusion

As we conclude the "Heart Watch" project, it becomes evident that the Random Forest Classifier stands out as a promising and reliable model for heart failure prediction. Its remarkable accuracy, precision, and balanced performance across key metrics position it as a compelling solution for early detection and intervention in heart disease cases.

The success of the Random Forest Classifier underscores the potential of machine learning models in healthcare applications, particularly in predictive analytics for critical conditions like heart failure. The project's initial objectives, centered around developing accurate predictive models, have been not only met but exceeded. Moving forward, it is imperative to acknowledge the broader implications of these findings. The deployment of accurate heart failure prediction models holds the promise of revolutionizing patient care. Early detection facilitated by such models can lead to timely interventions, ultimately improving patient outcomes and quality of life.

In conclusion, the "Heart Watch" project, driven by a meticulous modeling approach and comprehensive evaluation, not only achieves its primary objective of developing accurate predictive models but also opens avenues for impactful applications in healthcare. The Random Forest Classifier's success paves the way for future endeavors in leveraging machine learning for proactive heart failure prevention, aligning with the broader goal of enhancing public health and well-being.

References

Heart Failure dataset, from Kaggle website: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records

https://ieeexplore.ieee.org/document/10064888

https://www.ndc.scot.nhs.uk/docs/Heart%20Failure%20Dataset.pdf

https://www.sciencedirect.com/science/article/pii/S2001037016300460

https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5

https://www.nhlbi.nih.gov/research/heart-failure

https://www.ahajournals.org/doi/full/10.1161/CIRCRESAHA.121.318172

https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5

https://www.nhlbi.nih.gov/research/heart-failure

https://www.ahajournals.org/doi/full/10.1161/CIRCRESAHA.121.318172