Assessing Predictive Models for Cardiovascular Disease: A Data-Driven Approach

Ty P. Johnson & Dr. Robert Kelley

Bellarmine University Data Science Program

Cardiovascular disease is the leading cause of death worldwide and in our country. This project will explore the disease group and seek to understand the main factors that contribute to them. Several predictive analytic models will be created and assessed to determine how well the models can accurately determine the presence of a cardiovascular disease in a patient, employing various machine learning algorithms. In addition to model building, exploratory data analysis techniques will be utilized to answer several key questions related to the factors causing cardiovascular disease. The initial k-nearest neighbor model achieved 72% accuracy, indicating that although the model isn't the least effective, there remains significant scope for enhancement. Further refinement and the incorporation of other models will look to improve predictive accuracy.

INTRODUCTION

Cardiovascular disease encompasses a wide range of disorders affecting the heart and blood vessels.

- It was responsible for 32% of 2019 global deaths
- Claims a life every 33 seconds in the United States

Everyone is at risk for cardiovascular disease, and that risk continues to grow as one gets older, but preventive measures can significantly reduce this risk. Adopting healthier eating habits, increasing physical activity, and reducing alcohol and tobacco consumption are effective strategies for reduced risk. Leveraging predictive models to enhance awareness and identify individuals at risk early paves the way for more effective implementation strategies.

OBJECTIVES

1.To accurately predict the presence of cardiovascular disease in patients using machine learning algorithms

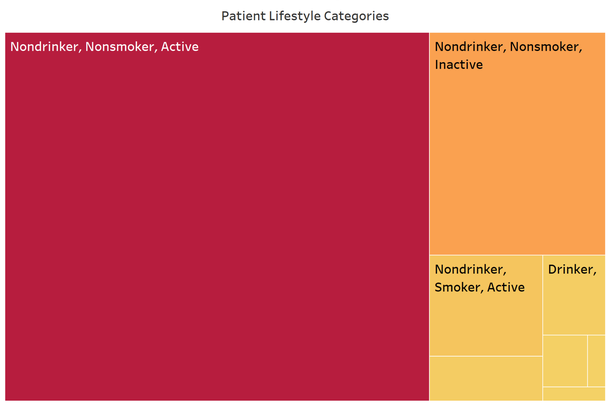
2.To identify and analyze the main factors contributing to cardiovascular disease through exploratory data analysis

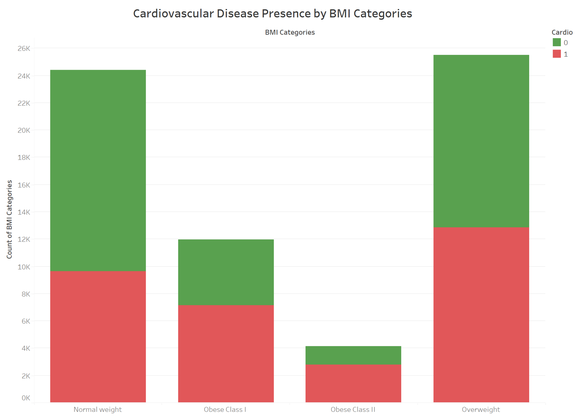
3.To evaluate the effectiveness of various predictive models in accurately predicting the presence of cardiovascular disease in patients

4.To explore the potential of data-driven approaches

MATERIALS & METHODS

The project began with an exploratory data analysis to gain insights into the dataset, which comprises 70,000 entries, including lifestyle, demographic, biometric data, and importantly, the presence of cardiovascular disease in patients. Using Python and the pandas library, the dataset was cleaned and prepared, setting the stage for addressing essential questions and facilitating model development. To enhance understanding and illustrate key findings, visualizations were crafted using both the seaborn library and Tableau, providing a clear depiction of the patient information contained in the dataset. Below is a visualization derived from the patient data mentioned.





Top 3 Lifestyle Categories
•Active, nonsmoker, nondrinker
•Inactive, nonsmoker, nondrinker
•Active, smoker, nondrinker

Overweight and obese groups have a higher proportion of cardiovascular disease present than those of normal weight.

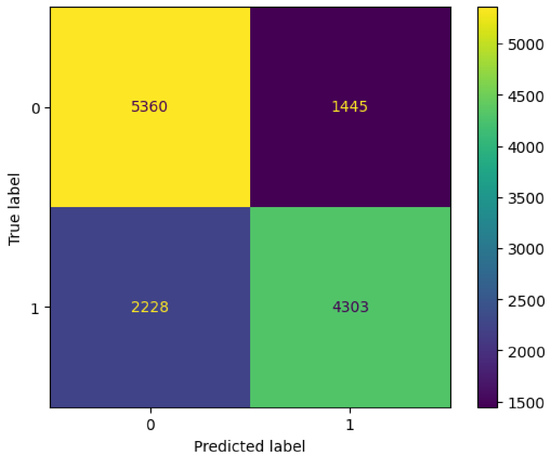
MATERIALS & METHODS (Cont.)

Then, model building began. The sci-kit learn library was used due to its extensive array of techniques and customizable options. Initially, a logistic regression model was created to serve as a foundational benchmark for future models. Then, the k-nearest neighbor model was developed with the following:

- K-Value: 257
- Train/Test Split: 80/20

RESULTS

The initial evaluation of the k-nearest neighbor model revealed an accuracy of 73%, which lags behind the baseline logistic regression model by 1%. Notably, the recall for detecting the presence of cardiovascular disease in individuals stands at 66%, raising concerns due to a 34% false negative rate. This rate is particularly alarming as it indicates a portion of missed disease diagnoses.



Overall, the performance metrics such as the F1 score, recall, and precision, of the k-neighbor model are on par with the baseline logistic regression model. However, in a medical setting, the high incidence of false negatives cannot be overlooked. It is critically important to minimize such errors to ensure early detection of the disease, for the maximization of patient health and well-being.

A notable challenge in developing the model was determining the optimal k-value amidst a large dataset, yet adjustments to the parameter scarcely impacted the model's performance. Additionally, an imbalance in critical factors associated with cardiovascular disease may be lowering overall performance.

CONCLUSION & FUTURE WORK

There are a few takeaways when it comes to the model building and medical diagnoses:

- 72% accuracy and high presence of false negatives show weakness in the KNN model
- Improvements with current KNN model likely only would be minimal
- Answer: Using a different machine learning algorithm to capture the complexity of the disorders

The dataset could also track more complex data:

- Bloodwork
- Tracking score for activity, smoking, alcohol consumption occurrences
- Examination of patients over time

In the future, the decision tree and random forest machine learning algorithms will be utilized and assessed to see if they work with the data better and provide more accurate predictions in seeing whether an individual has cardiovascular disease.

REFERENCES

World Health Organization. (n.d.). Cardiovascular diseases (CVDs). https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

Centers for Disease Control and Prevention. (2023, May 15). Heart disease facts. https://www.cdc.gov/heartdisease/facts.htm

American Heart Association. (2024, January 10). What is cardiovascular disease? https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease

Dempsey, Kuzak (2021). Exploring risk factors for cardiovascular disease [Data set]. Kaggle. https://www.kaggle.com/datasets/thedevastator/exploring-risk-factors-for-cardiovascular-diseas

CONTACT

tjohnson@bellarmine.edu