Predicting Cardiovascular Disease Risk Using Lifestyle and Preventative Care Factors

Ellie Landoch

elandoch@bellarmine.edu

DS450 Data Science Senior Capstone

22 February 2026

Executive Summary

Cardiovascular disease is one of the leading causes of death in the United States, although many cases are preventable through early intervention and lifestyle changes, which first comes from identification of the risk factors. An individual's risk for cardiovascular disease is influenced by many factors, including personal behaviors and access to preventative healthcare. Because of this, it is important that predictive models take these factors into consideration rather than solely focusing on clinical data.

This project will focus on developing models to estimate cardiovascular disease risk using demographic, lifestyle, clinical, and access-to-care variables. The data used for this analysis comes from a cardiometabolic-focused subset of the National Health and Nutrition Examination Survey (NHANES), which is a publicly available dataset collected by the Centers for Disease Control and Prevention. In addition to predicting cardiovascular disease risk, this project will examine how lifestyle behaviors can act as preventative healthcare and whether model performance differs across demographic groups.

Several predictive models will be developed and compared, including Random Forest, Gradient Boosting, and k-Nearest Neighbors. Logistic regression will also be used as a baseline model to provide a point of comparison. Model performance will be evaluated using standard classification metrics, and feature importance techniques will be used to better understand which variables contribute most to predictions.

Overall, the goal of this project is to explore how predictive analytics can be used to support decision-making in preventative healthcare and lifestyle choices. The project also aims to highlight potential disparities in cardiovascular risk prediction based on demographics and model types.

Project Idea

The main goal of this project is to develop predictive models that estimate an individual's risk for cardiovascular disease based on lifestyle behaviors, clinical indicators, and access-to-care factors. Cardiovascular disease is influenced by both personal behaviors, such as levels of physical activity and smoking status, and structural factors that affect access to preventative healthcare. This project aims to combine these elements into one predictive framework. In addition to making predictions, this project will explore changing lifestyle behaviors as a form of potential preventative healthcare. It will also evaluate whether cardiovascular risk and model performance differ across demographic groups. By taking this approach, the project not only focuses on prediction accuracy but also on understanding how risk varies across different populations.

Background

Although cardiovascular disease has been extensively studied, and risk prediction tools may exist, these models often rely on demographic and clinical variables, such as age, sex, blood pressure, and cholesterol levels. While these models are useful in clinical settings, they often place less emphasis on lifestyle behaviors and access-to-care factors that influence both risk and prevention.

Lifestyle factors, such as physical activity, diet, and body mass index, play a major role in cardiovascular disease prevention. However, not all people have equal access to resources that support healthy behaviors or regular preventative care. Because of this, predictive models that do not consider these factors may fail to capture important differences in a person's cardiovascular risk.

This project uses a cardiometabolic-focused subset of NHANES, which is a nationally representative dataset collected by the Centers for Disease Control and Prevention. NHANES combines survey data, physical examinations, and laboratory measurements to assess the health and nutritional status of adults in the United States. The dataset includes approximately 10,000 observations and contains demographic variables such as age and sex, lifestyle behaviors such as physical activity and smoking, clinical indicators such as blood pressure and cholesterol, and measures related to healthcare access. A binary indicator of cardiovascular risk will be used as the main outcome variable for the predictive models.

Modeling

This project will develop and compare multiple predictive models to estimate cardiovascular disease risk. Logistic regression will be used as a baseline model because it is easier to interpret and provides a clear point of comparison, but it is not considered one of the main predictive models.

The primary predictive models that will be used in this project are Random Forest, Gradient Boosting, and k-Nearest Neighbors. Random Forest models use multiple decision trees to identify patterns and relationships in the data. Gradient Boosting models improve performance over several iterations by focusing on observations that are more difficult to classify. k-Nearest Neighbors uses a similarity-based approach by making predictions based on outcomes from individuals with similar characteristics.

Model performance will be evaluated using classification metrics to compare accuracy across models. The analysis focuses on understanding which variables contribute most to cardiovascular disease risk predictions and whether model performance differs across demographics. Emphasis is placed on interpretability and meaningful insights rather than complex model tuning.

Tools

Python will be used as the primary programming language for data preprocessing, analysis, and predictive modeling. Data cleaning and manipulation will be done using pandas and NumPy. Predictive models will be made using scikit-learn.

Visualizations and associated analysis will be done using Tableau to help interpret results and show findings clearly. GitHub will be used to document and organize project code throughout the semester, supporting reproducibility and overall project organization.

Conclusion

This project will focus on estimating cardiovascular disease risk using lifestyle, clinical, and access-to-care factors. By developing and comparing multiple predictive models, this project aims to better understand which factors are most associated with cardiovascular risk and where potential disparities may exist. The findings from this project can help inform data-driven approaches to preventive healthcare and contribute to better understanding how predictive analytics can support health outcomes.

References

Centers for Disease Control and Prevention. *National Health and Nutrition Examination Survey (NHANES).* Retrieved from https://www.cdc.gov/nchs/nhanes/

Kaggle. *NHANES cardiometabolic dataset.* Retrieved from https://www.kaggle.com/

OpenAI. *ChatGPT*. Used for brainstorming potential predictive modeling approaches.