Author: Alysen Casaccio

Title: Heart Disease Prediction Using 2022 CDC Survey Data

Topic

Heart Disease Prediction Using 2022 CDC Survey Data

Business Problem

Heart disease is a leading cause of death in the U.S., and early prediction can significantly improve patient outcomes by enabling timely interventions. This project aims to develop a machine-learning model to predict the likelihood of heart disease based on key health indicators and lifestyle factors.

Datasets

The project will utilize publicly available CDC survey data with 246,022 rows after removing NaNs, featuring variables such as:

- State of Residence
- Gender
- General Health
- Last Checkup
- Physical Activity
- Mental and Physical Health Ratings
- Number of Sleep Hours
- Had Heart Attack (Y/N)
- Multiple other health-related factors

Methods

The project will use the following methods:

- **Data Cleaning and Preparation**: Ensure data is accurate and formatted correctly.
- **Descriptive Statistics**: Summarize heart disease rates and examine data to select a target variable.
- **Clustering**: Group patients to identify patterns and similar profiles.
- **Predictive Modeling**: Develop and evaluate classification models (Logistic Regression, Random Forest, SVM) to predict heart disease. Techniques like hyperparameter tuning and cross-validation will be employed.
- **Comparative Analysis**: Compare predicted and actual rates to assess model accuracy.
- **Visualization**: Create visualizations to display trends and predictions.

Ethical Considerations

- **Patient Privacy**: Ensure data is anonymized and compliant with HIPAA.
- **Bias and Fairness**: Addressing potential biases in the data related to race, socioeconomic status, and geographic location to ensure fair and unbiased model predictions.

- **Accessibility**: Ensuring the model is interpretable and actionable for healthcare providers.

Potential Challenges or Risks

- **Class Imbalance:** Comments on the data source file from other data scientists indicated that the target variable had an imbalance that needs to be managed.
- **Data Quality**: Ensuring data accuracy and consistency.
- **Feature Selection**: Identifying the most relevant features from a large pool of variables.
- **Model Interpretability**: Ensuring that models are interpretable for hospital administrators.

Audience Questions

1. How did you handle the class imbalance in the dataset?
2. What feature engineering techniques did you apply?
3. How did you ensure the model is not biased towards any demographic group?
4. What were the most significant predictors of heart disease according to your model?
5. How do you interpret the model's predictions in a real-world healthcare setting?
6. What measures did you take to ensure data privacy and security?
7. How did the model's performance vary with different algorithms?
8. Can the model be applied to other diseases or health conditions?
9. How do you plan to validate the model's predictions?
10. What challenges did you face during the data preprocessing stage, and how did you overcome them?

References

Abbott, D. (2014). *Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst*. Indianapolis: Wiley.

Madhumita Pal, S. P. (2022). Risk Prediction of Cardiovascular Disease Using Machine Learning Classifiers. *Open Medicine*, 1100-1113.

Morid MA, K. K. (2018 April). Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation. *AMIA*, 1312-1321.

Nadiah A. Baghdadi, S. M. (2023, September). Advanced Machine Learning Techniques for Cardiovascular Disease Early Detection and Diagnosis. *Journal of Big Data*, 10.

Sidra Abbas, S. O. (2024). Artificial Intelligence Framework for Heart Disease Classification. *Scientific Reports*, 3123.

Umarani Nagavelli, D. S. (2022). Machine Learning Technology-Based Heart Disease Detection Models. *Journal of Healthcare Engineering*.

Yar Muhammad, M. T. (2020). Early and Accurate Detection and Diagnosis of Heart Disease Using Intelligent Computational Model. *Sci Rep*. doi:https://doi.org/10.1038/s41598-020-76635-9