Machine Learning Model for Heart Disease Classification

CS 171 Final Project White Paper
Iris Feng, Alan Kashiwagi, Neeharika Kandikattu

Business Background

According to the World Health Organization (WHO), heart disease is the leading cause of death globally, taking an estimated 17.9 million lives each year, around 32% of all deaths worldwide (*Cardiovascular disease*). Just within the United States, one person dies every 34 seconds from heart disease, and it is still the leading cause of death for men, women, and people of most racial and ethnic groups. Heart disease is not only a very relevant detriment to the health of the global population but also causes immense financial strain; in the United States alone, the cost of heart disease-related healthcare services, medicines, and lost productivity due to death was $229 billion each year from 2017 to 2018, one of the largest overall contributors to healthcare spending (*Heart disease facts*).

One key to preventing and lowering the risk of heart disease is early diagnosis through analyzing a patient's health and symptoms. The presence of negative health markers such as raised blood pressure, raised glucose, high resting heart rate, and being overweight can help doctors determine if a patient has heart disease or is at risk of developing it. Having a precise, efficient way of identifying if a patient is at risk of developing or currently has heart disease would be beneficial not only for healthcare providers but also for health insurance companies alike.

Healthcare providers would benefit by being able to confidently diagnose patients early on, resulting in a higher rate of prevention prior to heart disease diagnosis. Even in those who are diagnosed with heart disease, catching signs and symptoms earlier can lead to medications being prescribed and changes made in the patient's lifestyle at an accelerated rate.

Health insurance companies would benefit by being able to categorize which customers are likely to develop heart disease. This way, they can determine what rates to charge for health insurance coverage depending on how severe the customer's symptoms are. If the customer's health improves, their rate may go down according to how much improvement their health markers show.

Objective

To address the needs discussed above, we can develop a machine learning model to classify whether or not a person will develop heart disease by using their health markers and demographic data. The health markers included in the development of the model will be tailored to exclude markers irrelevant to heart disease and ones with low correlation.

Data

a. **Data Source:** Our Heart Disease dataset was sourced from the UC Irvine Machine Learning Repository and contains data only from the Cleveland Clinic Foundation. Although the original data from the foundation contained 76 attributes, it was found that only 14 attributes contributed to detecting the presence of heart disease, so the dataset from UCI contains only those 14 attributes and over 300 rows.
b. **Features Included:** After checking the correlation of all the features in the dataset, we used all of them in our model, including age, sex, resting blood pressure, cholesterol, fasting blood sugar, etc. The dependent variable we are using to show the presence of heart disease is num. num has a value of 0 for no disease and 1-4 for.
c. **Data Cleaning:** Our first step in cleaning the data was to drop all instances that had missing values.
d. **Feature Engineering:** For the diagnosis of heart disease, we binned the "num" attribute into a dichotomous variable, assigning 0 for no disease and 1 for the presence of heart disease. We kept all the features as continuous variables. We also binned several other attributes such as sex, age, chest pain type (cp), cholesterol (chol), and more in order to create data visualizations to help us study the relationship between these variables and num.
e. **Dimension Reduction**: We decided not to use dimension reduction because with only 300 rows, the data set is very small, and we wanted to keep all the features included. Since dimension reduction is also known for causing data loss, we thought it would be best to skip it.
f. **Variable Selection:** For logistic regression, we decided to not use variable selection since we had good accuracy.
g. **Regularization:** For the Decision Tree model, we initially left all parameters to their default values, but the model performed poorly due to overfitting. Then we tried setting min_samples_leaf, min_samples_split, and max_depth to 7 and got a better fit.

Methodology Exploration

Because we are creating a model for classification, we decided on using Logistic Regression, Random Forest, SVM, KNN, and Decision Tree models to start with. All of these models have reputations for being best suited for classification, and we predicted that Logistic Regression would have the best accuracy score because it is used to predict a binary outcome, which we were seeking when we binned the dependent variable of our data into a dichotomous variable. We also favored using a Decision Tree since it does not need to make assumptions about the data, and the algorithm might fit the data better. We still fitted all the other models listed to see if there would be any wildcards that stood out that could be used for ensemble learning.

The ensemble technique we decided to use was Hard Voting, which we implemented through the EnsembleVotingClassifier.

Required Assumption
Decision tree models don't require us to make assumptions. For logistic regression, we assume that the variables aren't highly correlated with each other and that observations are independent of each other.

Model Equation
Model equations for decision tree and logistic regression credited to Géron (2019).
Cart cost function for Decision Tree:
$J(k,t_k) = m_{left}/m * G_{left} + m_{right}/m * G_{right}$
G left/right measures the impurity of the left/right subset,
m left/right is the number of instances in the left/right subset.

Logistic Regression cost function:
Equation 4-17. Logistic Regression cost function (log loss)

$$J(\boldsymbol{\theta}) = -\frac{1}{m}\sum_{i=1}^{m}\left[y^{(i)}log\left(\hat{p}^{(i)}\right) + \left(1 - y^{(i)}\right)log\left(1 - \hat{p}^{(i)}\right)\right]$$

In-Sample Validation and Performance Metric
We split the data into training and test sets and validated the performance using accuracy scores. We drew a confusion matrix for each model and found that the models tended to predict false negatives rather than false positives. Logistic Regression and Decision Tree models had the same performance at 81%, and our hard voting ensemble model that combined decision tree, random forest, support vector machine, and KNN models performed slightly better at 83%. We plan to launch using the ensemble model.

Monitoring and Maintenance Plan
We expect the model performance to change as we get more data and patient data changes over time. With the advancement of technology, we will likely have more features to work with as more correlations are made between new health markers and heart disease. With the new features and the likelihood of more data, there may come a need to implement dimension reduction or variable selection. We will monitor the performance and modify our model if it is not performing well. Other types of ensemble techniques can be considered as well as making adjustments to the individual models we used.

Conclusion

After trying different algorithms, we found that the logistic regression and decision tree models both performed well, but our ensemble model did better than expected and achieved 83% accuracy on the test data set. By launching our ensemble model, we can help the healthcare industry identify people who are likely to develop heart disease.

Our data set was small, so our model may underperform on new data. We will modify the model if it underperforms and as health markers for heart disease change. In the meantime, we can make improvements by exploring other ensemble models and new machine-learning algorithms.

In the future, we could try creating models to determine the risks of specific heart events such as heart attacks or blood clots. Models that can predict these events will be very powerful. The research on cardiovascular disease and the prevention of heart attacks will be greatly aided by these models.

Appendix

Heart disease facts. (2022, October 14). Retrieved December 9, 2022, from https://www.cdc.gov/heartdisease/facts.htm

Cardiovascular diseases. (n.d.). Retrieved December 9, 2022, from https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems* (2nd ed.). O'Reilly Media, Inc.