Title: Predicting Presence of Heart Disease in Patients using Machine Learning models

By: Allie Schneider, Francis Villamater, Stephan Dupoux

Abstract: According to the CDC, heart disease (HD) is one of the leading causes of death for people of most races in the United States and 47% of all Americans have at least 1 of 3 key risk factors for HD: high blood pressure, high cholesterol, and smoking. In this study, we analyze 2020 health status survey data from across the country and implement machine learning methods to detect "patterns" which can predict a patient's condition. Using Pyspark, we've built and compared a random forest model utilizing various scalers to classify development of HD. Our models were evaluated using standard metrics (accuracy, precision, recall, F-score, False Positive Rate and False Negative Rate) and ROC/Recall curves.

INTRODUCTION:

Our dataset includes key indicators of heart disease (HD) and reflects data from the 2020 annual Center for Disease Control and Prevention (CDC) survey data of 400,000 adults related to their health status. According to the CDC, HD is one of the leading causes of death for people of most races in the United States. About 47% of Americans have at least 1 of 3 key risk factors for HD: high blood pressure, high cholesterol, and smoking. Other key indicators include diabetic status, obesity (high BMI), not getting enough physical activity or drinking too much alcohol. Detecting and preventing the factors that have the greatest impact on HD is very important in healthcare.

DATASET:

The dataset used originates from Kaggle and was adapted from a dataset by the CDC from the Behavioral Risk Factor Surveillance System (BRFSS), an annual telephone survey gathering data on health status of U.S. residents. Typically, the BRFSS completes over 400,000 adult interviews each year, with over 250 variables generated. The dataset published on Kaggle has 319,795 instances and has been reduced to 18 variables:

1. `HeartDisease` (dichotomous): Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI). [TARGET VARIABLE]
2. `BMI` (continuous): Body Mass Index (BMI)
3. `Smoking` (dichotomous): Has the respondent smoked at least 100 cigarettes in their life?
4. `AlcoholDrinking` (dichotomous): Yes, if adult men having >14 drinks/week and adult women >7 drinks/week, no, otherwise.
5. `Stroke` (dichotomous): Has the respondent ever been told they had a stroke?
6. `PhysicalHealth` (continuous): Including physical illness and injury, for how many days during the past 30 days was the respondent's physical health not good? (0-30)
7. `MentalHealth` (continuous): For how many days during the past 30 days was the respondent's mental health not good? (0-30)
8. `DiffWalking` (dichotomous): Does the respondent have serious difficulty walking or climbing stairs?
9. `Sex` (dichotomous): Male or female
10. `AgeCategory` (categorical, 13 levels): Age category
11. `Race` (categorical, 6 levels): Race/ethnicity
12. `Diabetic` (categorical, 4 levels): Has the respondent ever been told they have diabetes?

13. `PhysicalActivity` (dichotomous): Adults who reported engaging in physical activity/exercise over the past 30 days outside of their regular job
14. `GenHealth` (categorical, 5 levels): How would the respondent rate their overall health?
15. `SleepTime` (continuous): On average, how much sleep does the respondent get in a 24-hour period?
16. `Asthma` (dichotomous): Has the respondent ever been told they have asthma?
17. `KidneyDisease` (dichotomous): Excluding kidney stones, bladder infection, or incontinence, has the respondent ever been told they have kidney disease?
18. `SkinCancer` (dichotomous): Has the respondent ever been told they have skin cancer?

EXPLORATORY DATA ANALYSIS:

Fortunately, this dataset is complete and contains no missing data. However, the dataset is highly imbalanced. `HeartDisease` contained 292,422 Negative for HD and 27,373 positive for HD and `Race` was over 75% "White". While it is possible to upsample on `HeartDisease`, this would skew the prevalence of HD in the survey population to be 50/50. Instead, we elected to balance on `Race` considering that the percentage of deaths from HD amongst most races are relatively equal. Each race was resampled to equal 'Black', about 22,000 instances per `Race`.

Continuous variables were checked for normality, which none exhibited. `BMI`, `PhysicalHealth`, `MentalHealth`, and `SleepTime` all demonstrate a heavy right skew. These will be transformed with a log transformation, necessary for our model which assumes a normal distribution with numerical attributes.

METHODOLOGY:

An initial correlation map yielded non-significant results with the highest correlation value of 0.43 between `DiffWalking` and `PhysicalHealth`, which is expected. However, despite the former, we elected to utilize all 18 variables in our models as we deem them all necessary features for predicting a patient's HD status.

Pyspark pipelines were assembled for a random forest model. One hot encoding was utilized, and dummy variables were created. Data was then split 80/20 for training and testing, respectively. This training set was used to train a random forest model and evaluated on the test set.

Another concept that we implemented into our modeling is the concept of stratified sampling. Stratified Sampling allows for the training and test datasets to be equally balanced in terms of the predicted variables. The goal is to create an algorithm that can predict if a person will be diagnosed with heart disease or not. Having a model that uses a dataset that is not reflective of the whole population could cause a lot of problems with the number of false positives and false negatives. Stratified Sampling for this problem would be the best way for outputting the best model possible.

The next part of our modeling methodology is to test the viability and predictive power of the model. We use the test set to see how well the model performed on data it has never seen before and create metrics on the predictive power. Metrics that we have utilized include, but are not limited to: the F1 score, Precision, Recall, and False Positive Rate.

After testing the model, we want to see if we can create a model that can outperform the out of the box model. We implement hyperparameter tuning to create the best models by testing out new systems. This is the final part of the project as this is step is very time consuming, and computationally expensive. Overall, our methodology for this project falls in line with commonly used ML modeling data projects.