Kelly Farrell - knf7vg
Karan Manwani - akp4he
Amanda Maruca - qfu2fy
Fall 2021 Semester

Prediction of Heart Disease in American Adolescents and Adults

Abstract

As a leading cause of death and disease in the USA, heart disease affects millions of people. Detection of cardiovascular conditions is therefore important, so that early interventions can alleviate some of the disease burdens faced by heart disease patients. Using data from the CDC's NHANES research project, we created a model to predict heart disease within the 2017-2018 participant cohort that incorporated blood test lab values, dietary and supplement logs, demographic information, and a medical history questionnaire. Out of the many variables measured for NHANES, these were chosen based on relevance and accessibility--a priority for our group was that the necessary variables could be collected within a variety of treatment settings to ensure ease of use. After upsampling our target heart disease group and performing a variety of transformation and feature engineering tasks, we trained logistic regression, LASSO, random forest, and support vector machine (SVM) models using 5-fold cross-validation. The models' performances were evaluated and compared on a separate testing data sample. While all of the models performed well, LASSO and random forest showed the highest levels of accuracy and lowest rate of false negative results. Due to the lower computational load required to train the LASSO model, our group would be most likely to endorse this method for implementation within healthcare settings.

Intro

Health is a state of complete physical, mental, and social well-being, which requires following the advice of medical professionals and taking preventative measures to reduce the possibility of diseases and maintain good energy levels. There are varying degrees of health and challenges presented among individuals, and across populations. Our group was interested in using a robust dataset including measures of health on the individual level as well as larger groups. This led us to the long-running NHANES research project, which characterizes health and illness factors using surveys, examination, and lab data values from a representative sample of the US population.

The National Health and Nutrition Examination Survey is a result of the National Health Survey Act that was passed in 1956. The purpose of this act was to collect statistical data on the amount, distribution, and effects of illness and disability in the United States. A branch of the US Public Health Service in the US Department of Health and Human Services conducted seven separate examination surveys that have evolved from the 1960s to present. The first three surveys were titled “National Health Examination Surveys” and focused on selected chronic diseases of adults and the growth and development of children. In 1970, nutrition was found to be increasingly more important in its relationship to health. This discovery prompted the proceeding survey names to be modified to “National Health and Nutrition Examination Survey” and to include a variety of dietary information in addition to health information of participants. The results of past years of NHANES data have been used by various government agencies to recommend public health measures and investigate drivers of health and disease across the US population.

For our project, we decided to explore datasets from NHANES based on a questionnaire from cohort year 2017-2018. We analyzed a variety of datasets including lab results, supplement data, and dietary data. Heart disease was chosen as the target variable due to the large burden it causes--millions of Americans are diagnosed with heart disease each year, which drives shorter lifespans and poorer health outcomes in those affected. Early detection of heart disease could result in earlier interventions, reducing the devastating impacts of heart disease on a large scale. Through data exploration and model creation, our goal is to identify which contributing factors are the most significant in predicting heart disease in participants.

Exploratory Data Analysis

The NHANES dataset contains hundreds of measures across many modalities. Our group wanted to include a diversity of datatypes while also limiting the amount of variables, both to prevent our model from becoming overly complex, but also in keeping with our goal of accessibility--it was important to us to use tests and questionnaires that are likely already a part of a standard doctor’s appointment so that results could be used in many patient settings. Based on these criteria, and a review of relevant literature, we chose to include some measures from a complete blood draw (CBC), fasting insulin levels (blood), high-sensitivity C-reactive protein test (blood), a medical history questionnaire, patient demographics, and survey questions relating to the use of dietary supplements and diet history. (A diagram showing the complete variable list is below.)

Lab Datasets	**Dietary Datasets (Food Log Day 1)**	**Supplement Questionnaire (Past 30 days)**	**Medical History Questionnaire**	**Demographics Survey**
Fasting Insulin	How many grams were eaten of: -Cholesterol -Carbohydrates -Fiber -Monounsaturated, polyunsaturated, and saturated fats -Sugars -Protein -Sodium	Were any dietary supplements taken? Were any antacids taken?	Diagnosis of: -gout -congestive heart failure -COPD -coronary heart disease -stroke -angina -thyroid problem -emphysema -liver problems (fatty liver, liver fibrosis, cirrhosis, viral	Gender

	-Added sodium (during food preparation)		hepatitis, autoimmune hepatitis, jaundice, other) -chronic bronchitis -gallstones	
Lymphocyte Number	How many mcg were eaten of: -Vitamins A -Vitamin B1 -Vitamin B12 -Vitamin B2 -Vitamin B6 -Vitamin C -Vitamin D(D2, D3) -Vitamin E -Vitamin K -Calcium -Folate/Folate DFE -Folic Acid -Added Vitamins B12 and E	How many total dietary supplements were taken? How many total Antacids were taken?	(If diagnosed) Do you still have asthma/thyroid problem/liver problem?	Age
Monocyte Number			(If history of cancer) -1st cancer- where was it? -2nd cancer- where was it? -3rd cancer- where was it?	Race/Hispanic origin
Eosinophils Number		How many mcg were taken of: -Vitamin B1 -Vitamin B2 -Vitamin B6 -Vitamin B12 -Vitamin C -Vitamin D -Vitamin K -Niacin -Folic acid -Folate DFE?	(If diagnosed) Have you had emergency care for asthma in the last year?	Served active duty in the US Armed Forces?
Basophils Number			Have you ever had a blood transfusion?	Have you ever served in foreign country?
Hematocrit %	How many mcg were eaten of: -Alpha-carotene -Beta-carotene -Beta-cryptoxanthin -copper -Iron -Lutein + zeaxanthin -Lycopene -Magnesium -Niacin -Phosphorous -Potassium -Retinol -Selenium -Theobromine -Total Choline -Zinc		(If diagnosed) What type of arthritis do you have?	Country of birth
Mean Cell Hemoglobin			Have you had abdominal pain in the last 12 months? If so, where was the most uncomfortable pain? Have you seen a doctor for this pain?	Citizenship status
Platelet Count		How many mcg were taken of: -Lycopene -Lutein + Zeaxanthin	Have you had treatment for anemia in the last 3 months?	Marital Status
Nucleated Red Blood Cell Count			(If diagnosed) Have you had surgery for gallstones?	Pregnancy status
High-Sensitivity C-Reactive Protein Level	How many mcg were eaten of: -Alcohol -Caffeine	How many mcg were taken of: -Calcium -Phosphorous -Magnesium -Iron -Zinc -Copper -Sodium -Potassium -Selenium -Iodine	Has a close relative had: -Diabetes? -Heart attack?	Education Level
LDL (Cholesterol)	Compare food eaten yesterday to your usual diet		Has a doctor asked you to: -reduce/control weight? -reduce fat in diet? -reduce salt in diet? -reduce calories in diet? -increase exercise?	Total # of people in household, total # of people in family
	Dietary recall status		Are you currently: -reducing/controlling weight?	Ratio of family

Triglycerides	(confidence in memory of foods eaten)		-reducing fat in diet? -reducing salt in diet? -reducing calories in diet? -increasing exercise?	income to poverty level

The graphs below show breakdowns of the demographic data of the participants' gender, race/hispanic origin, and age. The top-left graph (Figure 1) shows that the gender breakdown of the study was split almost evenly with 4557 males and 4697 females. The top-right graph (Figure 2) shows the race/hispanic origin of the participants with 1367 Mexican Americans, 820 Other Hispanic, 3150 Non-Hispanic White, 2115 Non-Hispanic Black, and 1802 Other Races including multi-racial. The bottom graph shows the breakdown of participants' ages, most of which were under 30 years of age.

Figure 1

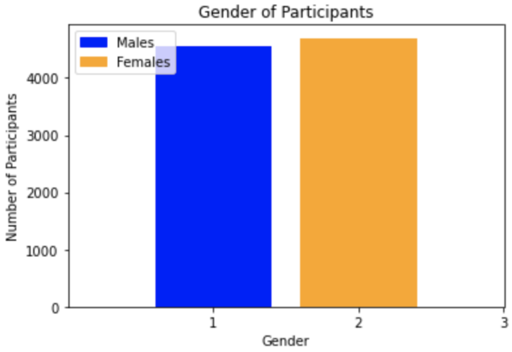
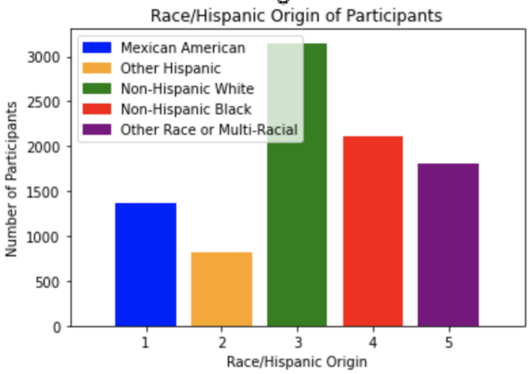


Figure 2



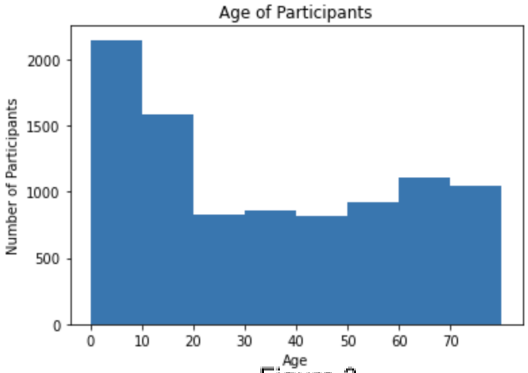


Figure 3

The graph below (Figure 4) shows how many participants have (or have a history of) heart disease. Less than 1000 participants disclosed that they have congestive heart failure, COPD, angina/angina pectoris, heart attack, or history of stroke. About 5000 participants reported that they do not have any of these conditions relating to heart disease. Since the difference is so large in participants with heart disease and participants without heart disease, we decided to upsample the heart disease group in order to improve the model's predictive abilities. Without creating more equally-sized groups, a model could simply have guessed that all participants did not have heart disease without consideration of the independent variables and achieve around 80% accuracy. Such a model would not provide any benefit to patients or their healthcare providers.

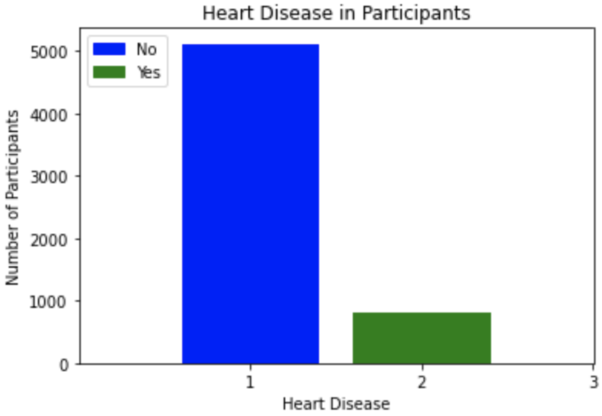


Figure 4

The boxplots below are examples of variables we compared against participants with and without heart disease. The Sugar Intake boxplot (Figure 5) shows that those with no heart disease history and those with heart disease history generally consume the same amount of sugar. The High-Sensitivity C-Reactive Protein boxplot (Figure 6) shows that participants with heart disease are more likely to report higher measures of the C-Reactive protein, which has been found to have a high correlation with heart disease in previous studies.

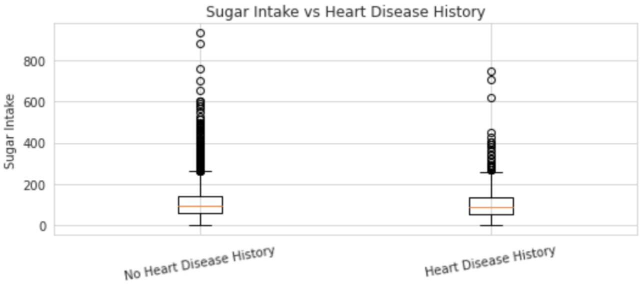


Figure 5

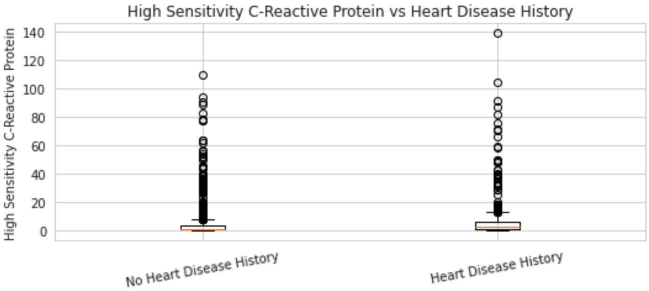


Figure 6

Methods

Data Cleaning

The 2017-2018 NHANES cohort included 9300 total participants, who all completed demographic questionnaires. Of those, 8400 filled out both a high-sensitivity C-reactive protein blood test and complete blood draw, 7500 had their total blood cholesterol measured, and 3100 had fasting insulin and triglycerides tested. 4399 values were imputed for insulin and triglycerides (methods described in section below). 9255 participants answered survey questions about dietary supplements, 8900 completed a medical history questionnaire, and 7484 submitted data about the foods they'd eaten in the last two days. Joining all of this information based on participant ID# yielded a total of 6582 lines in our experimental dataset.

Despite the high number of participants ages 0-11, we chose to focus on participants aged 12 and older, as the lab values such as cholesterol were not collected for those 11 and younger. While we chose to impute some missing lab values (this is described in more detail in the methods section), we did not feel it was methodologically sound to generalize group means from the 12+ sample to the younger pediatric population. Removing the younger participants brought the total number of lines to 5787. We also removed rows where participants answered "Unsure" or "I don't know" to various survey questions, which amounted to only 13 rows.

Preprocessing

Missing Data

While a majority of participants answered the survey measures in their entirety, only about ⅓ of the total NHANES 2017-2018 cohort (3100 participants) had fasting insulin and total blood cholesterol measured. In order to utilize this data in our model, we imputed weighted means based on age and gender for these two variables. The weighted means were calculated within a new column and then added into the dataset using a 'fillna' statement from the pyspark.sql module.

Feature Engineering

Computation of Target Variable

Our target variable was a binary indicator of heart disease. We needed to engineer this target based on the medical history questionnaire. If participants indicated that they had ever been told by a doctor that they had congestive heart failure, COPD, angina/angina pectoris, heart attack, or stroke, the value of the target variable label assigned was 1.0. Otherwise, the label value was 0.0. Target variable values were assigned using a nested 'when' statement from pyspark.sql.functions.

One-Hot Encoding, VectorAssembly, and Scaling

As many of the independent variables were categorical, we used the one-hot encoding method from the pyspark.ml.feature module to create 'dummy' variables. New columns containing sparse vectors were created, with the suffix '_one_hot' added to the original feature name. The original features were then removed from the independent variables list. Numeric variables were scaled using the StandardScaler, and all independent variables were combined into a single vector using the VectorAssembler module, both from pyspark.ml.feature.

Minority Upsampling

Only 714 participants were identified as being diagnosed with heart disease using the target variable. We upsampled the heart disease group with a custom function based on filtering for each label group (heart disease vs. controls). The function upsampled with replacement until the ratio of the two labels were about the same. The final ratio of participants was 5157 heart disease participants to 5060 controls, with a new total of 10217 lines of data. We utilized caching to optimize model training and reduce the memory load.

Test/Train Split and Cross-Validation

70% of the data was used to train the model, with the remaining 30% saved for testing. A seed was set to ensure the same data was used for all modeling. Each model was trained using 5-fold cross-validation via the CrossValidator method from pyspark.ml.tuning.

Models

Logistic Regression

We started our modelling with a logistic regression model and this model did not have a regularization penalty. The model was built using the training dataset and all the variables after performing the feature engineering and upsampling described. Our engineered label column was used as the dependent binary variable for heart disease. After running the model to make predictions on the test dataset, the initial results were promising with accuracy of 0.965, AUROC of 0.995, and a F1 precision of 0.965 for predicting the risk of heart disease.

LASSO

Given the promising results of the Logistic regression model, we decided to apply a regularization penalty as our data had over 140 variables. With such a high number of variables, we felt using the Lasso penalization instead of Ridge penalization would be more appropriate since it assigns a zero weighting to variables that have low relevance to the model. 0.01, 0.1, 0.3 were used as regularization parameters, with 0.1 being selected as the best based on the cross validation results. As shown below, the results for this model improved upon the logistic regression model with accuracy of 0.989, AUROC of almost 1, and a F1 precision of 0.989 for predicting the risk of heart disease.

Random Forest

We then looked at models outside of logistic regression, and Random Forest was the next model we built. In this model, we performed our cross validation with 50, 100, and 300 trees with the max depth ranging from 5, 7, and 10. Based on the cross validation, the model with the best results had 300 trees and 10 nodes. Below are the results from this model.

SVM

Finally, we built a Linear Support Vector Machine (SVM) model with regularization parameters ranging from 0.05, 0.1, and 0.3. As shown below, this model produced the least accurate results. The best-performing model had a regularization parameter of 0.1.

Results

Confusion Matrices

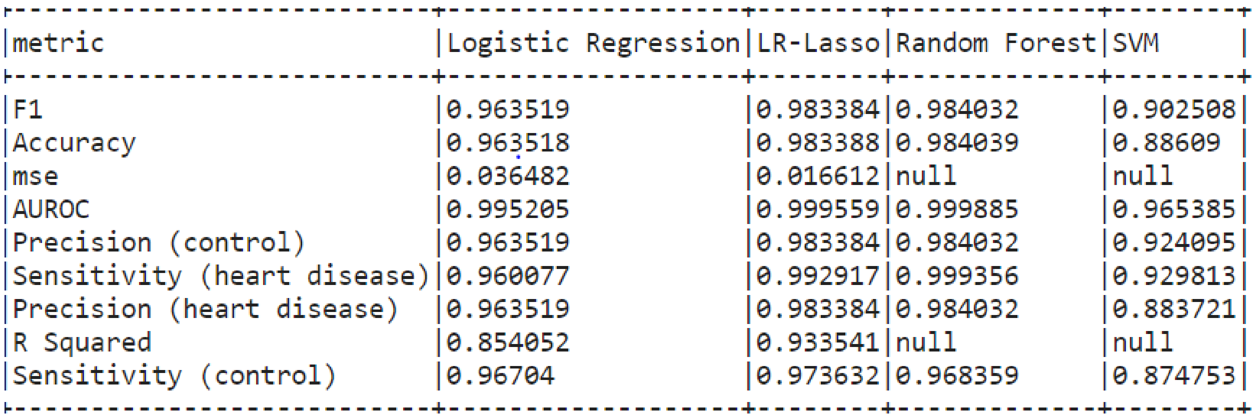
The raw values of the predicted and actual labels of the testing sample for each model are detailed in the charts below:

	Logistic Regression		**LASSO**	
	Predicted Positive	Predicted Negative	Predicted Positive	Predicted Negative
Actual Positive	1467	50	1477	40
Actual Negative	62	1491	11	1542

	Random Forest		**Linear Support Vector Machine (SVM)**	
	Predicted Positive	Predicted Negative	Predicted Positive	Predicted Negative
Actual Positive	1469	48	1327	190
Actual Negative	1	1552	109	1444

Evaluation Metrics

The table below contains a variety of ML evaluation metrics for each of the four models.



```
+---------------------------+-------------------+--------+-------------+--------+
|metric                     |Logistic Regression|LR-Lasso|Random Forest|SVM     |
+---------------------------+-------------------+--------+-------------+--------+
|F1                         |0.963519           |0.983384|0.984032     |0.902508|
|Accuracy                   |0.963518           |0.983388|0.984039     |0.88609 |
|mse                        |0.036482           |0.016612|null         |null    |
|AUROC                      |0.995205           |0.999559|0.999885     |0.965385|
|Precision (control)        |0.963519           |0.983384|0.984032     |0.924095|
|Sensitivity (heart disease)|0.960077           |0.992917|0.999356     |0.929813|
|Precision (heart disease)  |0.963519           |0.983384|0.984032     |0.883721|
|R Squared                  |0.854052           |0.933541|null         |null    |
|Sensitivity (control)      |0.96704            |0.973632|0.968359     |0.874753|
+---------------------------+-------------------+--------+-------------+--------+
```

Discussion

Based on the models' performance, any one of the studied methods could be used to accurately predict the risk of heart disease in patients over the age of 12. The LASSO regression and Random Forest were the most successful models tested, each with under 2% of observations misclassified. Least successful was the SVM, with around 88.6% accuracy. The most concerning aspect of this model's performance was the 190 false negatives (around 3.9% of the total testing group), each of which represents a patient with heart disease that was missed. Overall, our group feels that LASSO would be the best choice for predicting heart disease, since it requires fewer computational resources during model training. This makes it ideal for implementation in healthcare settings that may rely on older computers. The results also indicate that the LASSO method was slightly better than random forest at minimizing false negatives, ensuring that as many patients as possible are flagged for further cardiovascular testing and early interventions that can improve quality of life.

Future Opportunities

With the models we have built and the information available from the survey, there are a few areas we would be interested in exploring further. One such area would be to take a closer look at the individual variables to further limit the models to necessary variables or performing Principal Component Analysis (PCA) on the data. It would also be interesting to examine how well or differently the various models perform for different demographic groups, such as is random forest just as accurate for men as it is for women or if the performance varies across different racial groups. In order to perform such analysis, we would bring in more years of data from the NHANES database as it would give us more data points to train and test our models and deduce more accurate conclusions.

Bibliography

Data Source: NHANES study 2017-2018, CDC (each csv sourced from the "Data, Documentation, and Codebooks" section)
https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2017

Background information used to narrow down variables: Madjid and Fatemi 2013, *Components of the Complete Blood Count as Risk Predictors for Coronary Heart Disease*. Texas Heart Institute Journal. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3568280/

Spark Documentation on Classification/Regression:
https://spark.apache.org/docs/latest/ml-classification-regression.html