Heart Disease Prediction

Fall 2022 SI 670 Final Report

University of Michigan
Jiaqi Li, Wenjie Wu, Ziqin Tian

Abstract

“About 697,000 people in the United States died from heart disease in 2020—that’s 1 in every 5 deaths”[1][2], which would be a hidden danger for all human-beings. Would people, like us, who live under relatively large pressure have a higher possibility of heart disease? The aim of this project is to predict if the person gets heart disease based on the individual physical and mental health condition, by using machine learning models. The dataset utilized for this project is the 2021 Behavioral Risk Factor Surveillance System(BRFSS) Survey Data from CDC. Also, some future directions and limitations are discussed at the end of this report.

Keywords: Heart Diseases; Machine Learning; Classification

1. Introduction

Heart disease is a group of diseases involving the heart and blood vessels. Heart diseases include blood vessel disease, such as coronary artery disease; irregular heartbeats (arrhythmias); heart problems people are born with (congenital heart defects); disease of the heart muscle; and heart valve disease. There are some common symptoms such as chest pain, chest tightness, chest pressure and chest discomfort (angina), shortness of breath, pain in the neck, jaw, throat, upper belly area or back [3]. People might not be diagnosed with coronary artery disease until they have a heart attack, angina, stroke or heart failure. A heart attack can cause major health issues and possibly death if it is not treated right away. Around the world, heart attacks are a common cause of death. Cardiovascular diseases are estimated to be associated with 17.5 million deaths worldwide. In middle- and low-income nations, cardiovascular illnesses cause more than 75% of deaths [4][5]. Risk factors for heart disease include age, sex, family history, smoking, unhealthy diet, high blood pressure, high cholesterol, diabetes, obesity, lack of exercise, stress, poor dental health, etc. [3] Hence, our SI 670 project goal is to explore the most important indicators of heart disease and use these indicators to predict whether people have heart disease or not. The rest of the paper is structured as follows: Section 2 describes the code and data used in this project, as well as the data cleaning, exploratory data analysis and the algorithms and models we used. Section 3 describes the evaluation and results analysis. Section 4 summarizes the related work. Finally, Section 5 concludes the paper along with future potential work.

2. Methods

2.1 What's in our code

Our code for this project is made up of three key parts: data processing, exploratory data analysis (EDA), and models. In the data processing stage, we used the SAS Transport Format (.XPT file) and read this format in Python. For further research and model runs, we scoped down the dataset to only what we needed and treated the missing data. In the EDA part, data summary, unique values in each column, and visualization of the relationship between target variable and features are what we did to understand the data we have. In the modeling stage, we split data into training and testing datasets, encode categorical variables, run the benchmark and other models and tune hyperparameters.

2.2 The dataset we used

Our dataset is collected by the Behavioral Risk Factor Surveillance System (BRFSS), which is "the nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services"[6]. It comprises the results from 49 states, the District of Columbia, Guam, Puerto Rico, and the US Virgin Islands in 2021. There are 303 distinct factors in all, including answers to the questionnaire and calculated variables[7]. These variables are related to health-related risk behaviors and personal health conditions, which would provide us useful information for indicators on heart disease from the internal health conditions. Since the dataset is broad enough for our project, it is the only dataset we used.

2.3 Data Cleaning

Since this questionnaire is large and sophisticated which contains around 300 questions, we firstly studied the files which introduce each column and understood how they made calculations to get calculated variables. One thing we'd like to mention is that the range for calculated BMI from the original dataset is 1 - 9999 which is really abnormal as a BMI index, so we looked back to data before calculation and found some unit converting problems. As a result, we converted the units ourselves and calculated the BMI. The more detailed variable matching processing is shown in Appendix Table 1.

At the data cleaning stage, we prepared the data for each characteristic to make them more organized for the EDA and modeling. To deal with the missing values, since we have a relatively large dataset and aim to prevent side effects and poor performance caused by missing values imputation [8], we opt to delete these missing data. For answers in the questionnaire such as "Refused", and "Don't know / Not sure", we classified these data as null values as well since they will not give us more information. The last stage would employ a dataset containing 334,205 valid records with 18 columns.

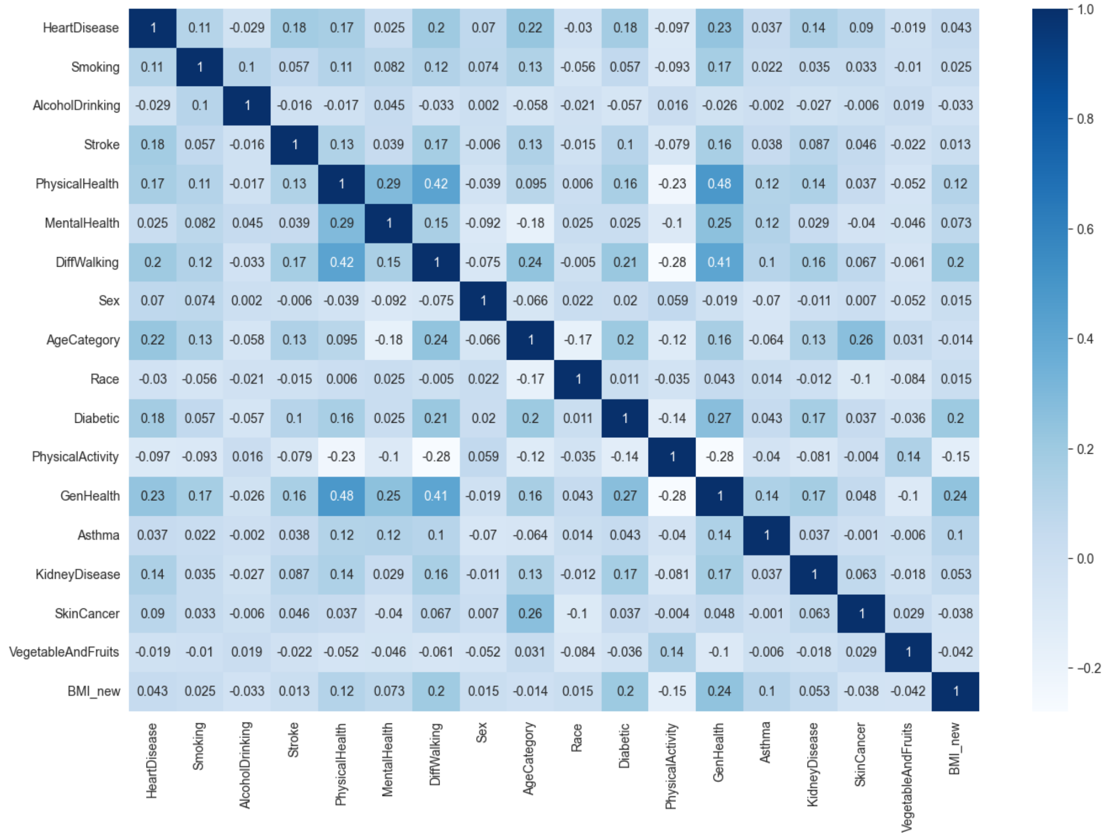
2.4 Exploratory Data Analysis (EDA)

2.4.1 Data Overview

There are 18 columns and 334205 rows in our dataset. “HeartDisease” is our target label. We can observe 27,546 cases of heart disease and 306,659 cases without heart disease. 14 features are categorical features (including “Smoking”, “AlcoholDrinking”, “Stroke”, “DiffWalking”, “Sex”, “AgeCategory”, “Race”, “Diabetic”, “PhysicalActivity”, “GenHealth”, “Asthma”, “KidneyDisease”, “SkinCancer”, “VegetableAndFruits”). Three features are numerical features (including “BMI”, “PhysicalHealth”, “MentalHealth”).

2.4.2 Features Correlation

For all the features, we plotted the correlation Heatmap (Fig. 1) between each feature. We also plotted the importance histogram (Fig. 2) of the correlation between heart disease and other features. Most relevant features of heart disease are people's general health, age, difficulty walking or climbing, etc.



	HeartDisease	Smoking	AlcoholDrinking	Stroke	PhysicalHealth	MentalHealth	DiffWalking	Sex	AgeCategory	Race	Diabetic	PhysicalActivity	GenHealth	Asthma	KidneyDisease	SkinCancer	VegetableAndFruits	BMI_new
HeartDisease	1	0.11	-0.029	0.18	0.17	0.025	0.2	0.07	0.22	-0.03	0.18	-0.097	0.23	0.037	0.14	0.09	-0.019	0.043
Smoking	0.11	1	0.1	0.057	0.11	0.082	0.12	0.074	0.13	-0.056	0.057	-0.093	0.17	0.022	0.035	0.033	-0.01	0.025
AlcoholDrinking	-0.029	0.1	1	-0.016	-0.017	0.045	-0.033	0.002	-0.058	-0.021	-0.057	0.016	-0.026	-0.002	-0.027	-0.006	0.019	-0.033
Stroke	0.18	0.057	-0.016	1	0.13	0.039	0.17	-0.006	0.13	-0.015	0.1	-0.079	0.16	0.038	0.087	0.046	-0.022	0.013
PhysicalHealth	0.17	0.11	-0.017	0.13	1	0.29	0.42	-0.039	0.095	0.006	0.16	-0.23	0.48	0.12	0.14	0.037	-0.052	0.12
MentalHealth	0.025	0.082	0.045	0.039	0.29	1	0.15	-0.092	-0.18	0.025	0.025	-0.1	0.25	0.12	0.029	-0.04	-0.046	0.073
DiffWalking	0.2	0.12	-0.033	0.17	0.42	0.15	1	-0.075	0.24	-0.005	0.21	-0.28	0.41	0.1	0.16	0.067	-0.061	0.2
Sex	0.07	0.074	0.002	-0.006	-0.039	-0.092	-0.075	1	-0.066	0.022	0.02	0.059	-0.019	-0.07	-0.011	0.007	-0.052	0.015
AgeCategory	0.22	0.13	-0.058	0.13	0.095	-0.18	0.24	-0.066	1	-0.17	0.2	-0.12	0.16	-0.064	0.13	0.26	0.031	-0.014
Race	-0.03	-0.056	-0.021	-0.015	0.006	0.025	-0.005	0.022	-0.17	1	0.011	-0.035	0.043	0.014	-0.012	-0.1	-0.084	0.015
Diabetic	0.18	0.057	-0.057	0.1	0.16	0.025	0.21	0.02	0.2	0.011	1	-0.14	0.27	0.043	0.17	0.037	-0.036	0.2
PhysicalActivity	-0.097	-0.093	0.016	-0.079	-0.23	-0.1	-0.28	0.059	-0.12	-0.035	-0.14	1	-0.28	-0.04	-0.081	-0.004	0.14	-0.15
GenHealth	0.23	0.17	-0.026	0.16	0.48	0.25	0.41	-0.019	0.16	0.043	0.27	-0.28	1	0.14	0.17	0.048	-0.1	0.24
Asthma	0.037	0.022	-0.002	0.038	0.12	0.12	0.1	-0.07	-0.064	0.014	0.043	-0.04	0.14	1	0.037	-0.001	-0.006	0.1
KidneyDisease	0.14	0.035	-0.027	0.087	0.14	0.029	0.16	-0.011	0.13	-0.012	0.17	-0.081	0.17	0.037	1	0.063	-0.018	0.053
SkinCancer	0.09	0.033	-0.006	0.046	0.037	-0.04	0.067	0.007	0.26	-0.1	0.037	-0.004	0.048	-0.001	0.063	1	0.029	-0.038
VegetableAndFruits	-0.019	-0.01	0.019	-0.022	-0.052	-0.046	-0.061	-0.052	0.031	-0.084	-0.036	0.14	-0.1	-0.006	-0.018	0.029	1	-0.042
BMI_new	0.043	0.025	-0.033	0.013	0.12	0.073	0.2	0.015	-0.014	0.015	0.2	-0.15	0.24	0.1	0.053	-0.038	-0.042	1

Fig. 1: Feature Correlation Heatmap

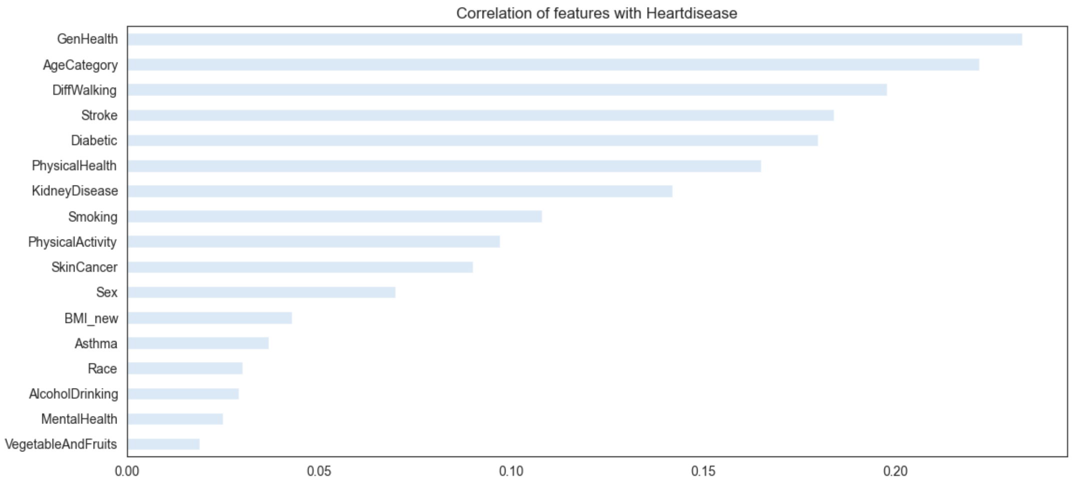


Fig. 2: Correlation between Heart Disease and Other Features

2.4.3 Categorical Features

For categorical features, we plotted histogram diagrams of the relationship between each feature and heart disease. In the present example shown in Figure 3, the numbers 1 to 13 represent the different age categories (1: 18-24; 2: 25-29; 3: 30-34; 4: 35-39; 5: 40-44; 6: 45-49; 7: 50-54; 8: 55-59; 9: 60-64; 10: 65-69; 11: 70-74; 12: 75-79; 13: 80+). From all the figures we inferred initial analysis:

- Males have a higher rate of heart disease than females.
- People who have a smoking history are more likely to have heart disease than people who don't smoke.
- People over 65 years old are more likely to have heart disease.
- People with Kidney Disease also have a higher rate of heart disease.
- People with Stroke also have a higher rate of heart disease.
- People with Diabetic also have a higher rate of heart disease.
- People who have serious difficulty walking or climbing have a higher rate of heart disease.
- People who do exercise have lower rates of heart disease.
- People who say they are in fair/poor health have higher rates of heart disease.

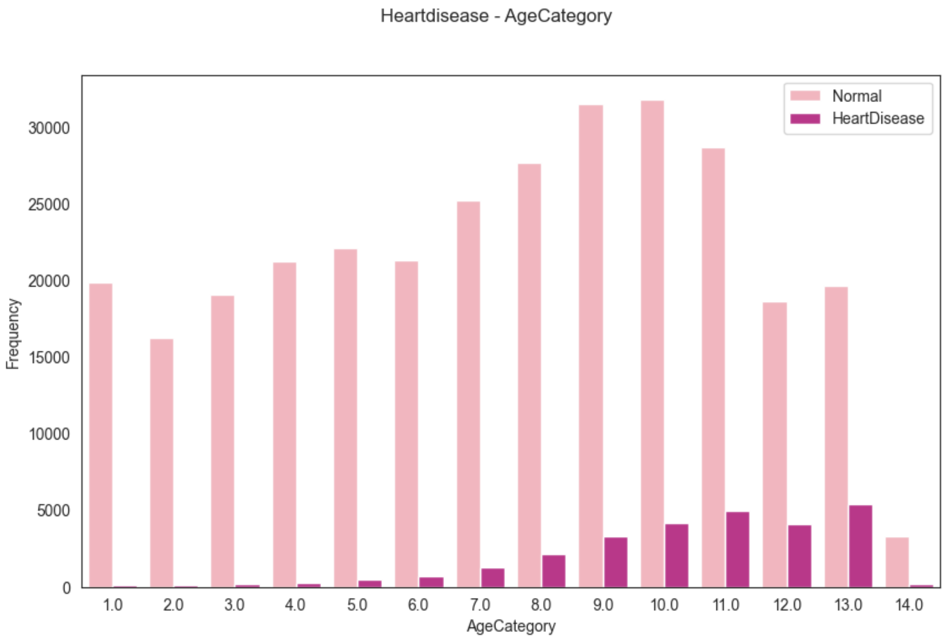


Fig. 3: Heart Disease - Age Category

2.4.4 Numerical Features

For numerical features, we plotted the kernel density estimate (KDE) for each feature with heart disease. As shown in figure 4, people with heart disease have a higher body mass index (BMI) than people without heart disease. In order to keep sufficient data, we assigned null values to 0 for PhysicalHealth and MentalHealth, so the "0" showing in the figures of PhysicalHealth and MentalHealth means null. People with heart disease are more likely to have had more days of poor physical and mental health in the past 30 days (Fig. 5, 6). Based on all the results of our EDA, we performed further modeling and prediction.

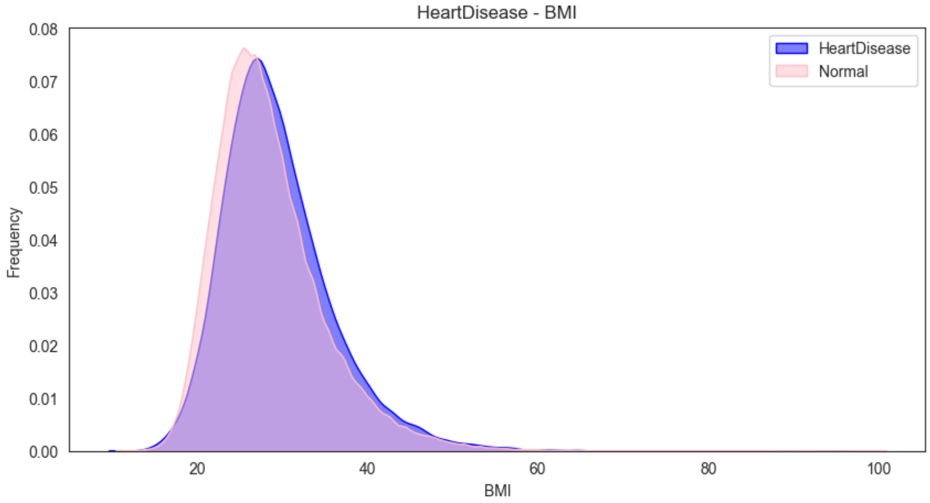


Fig. 4: HeartDisease - BMI

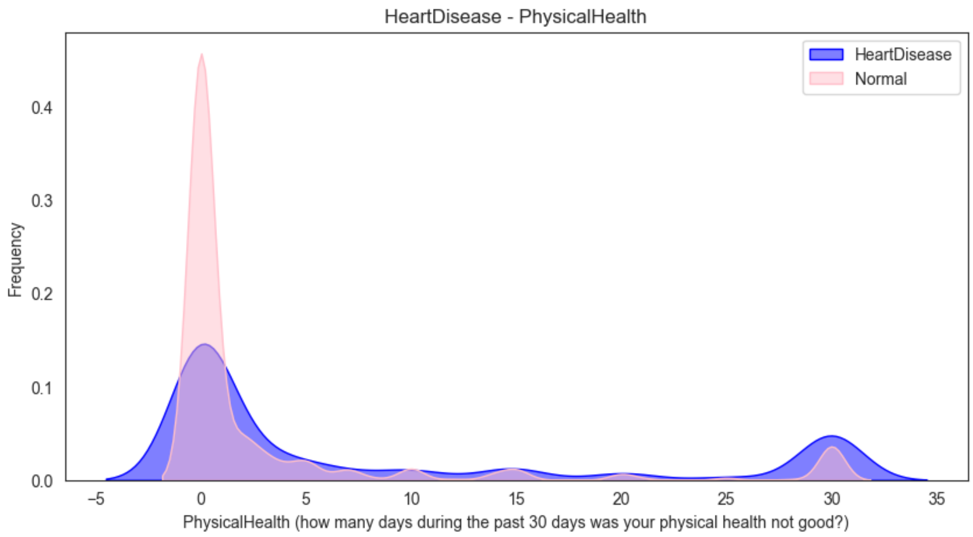


Fig. 5: HeartDisease - PhysicalHealth

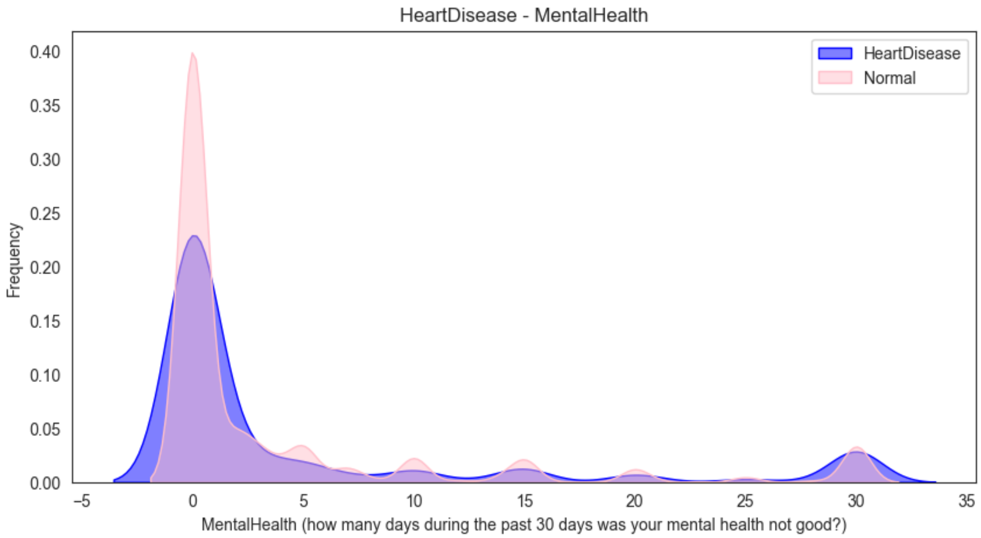


Fig. 6: HeartDisease - MentalHealth

2.5 Data Preparation

2.5.1 Training / Testing Data Split

Since our mission is to classify and predict the probability of getting heart disease, we used the column "HeartDisease " as our target variable and used the remaining variable as the training features to predict the target. Since we have 306,659 people with no heart disease while only 27,546 people reported heart disease, we clearly have imbalanced data that will affect the performance of the model. To solve the problem, we used a sampling method called Synthetic Minority Oversampling Technique, or SMOTE for short. SMOTE works by oversampling the examples in minority classes for a model to effectively learn the decision boundary. After we get the appropriate dataset, we split the training and testing data and shuffle when splitting.

2.5.2 Encoding and Scaling

Our dataset has two mixed data types: categorical data and numerical data. To better predict the target, we need to encode all the categorical data. We first recognized all the categorical data and listed those variables. Then we use OneHotEncoder to encode the categorical variables we listed. We merged new encoded variables with the original table and dropped the categorical variables. Moreover, to avoid a larger range of data playing a more decisive role while training the model, we performed the standard scaler on both training features and testing features.

2.6 Models

2.6.1 Benchmark

We use Logistic Regression without setting any specific parameters and scaling as our benchmark. The following is the result of the benchmark model:

Accuracy	Precision	Recall	F1 Score
0.7668	0.7563	0.7843	0.7700

The accuracy stands for the ratio of correct predictions. 0.9094 of accuracy means 90.94% of predictions are correct. The benchmark performs really well, especially for the dataset in real-world applications. Precision in this context is referred to as positive predictive value. Recall is referred to as true positive rate or sensitivity. In the medical subject, we prefer a higher recall to detect as many people as possible who are likely to have heart disease, even with a lower precision.

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

Fig. 7: Precision and Recall

The F1 score is the harmonized average of the precision and recall score. The higher the F1 score, the higher both precision, and recall scores. We want to maximize the F1 score as much as possible.

2.6.2 Initial Models

Before we run models, we applied principal component analysis to reduce the dimension of the data. In order to find the best number for the parameter of PCA, we use a loop to iterate through 2 to 8 to find the best number. Then we ran 14 machine learning models in total to explore the best model. Since it is a binary classification problem that needs to predict whether or not the target has a chance of diagnosing heart disease, we did not perform regression models. The models we used and the results are discussed below.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.7195	0.7242	0.7051	0.7145
K Nearest Neighbor Classifer	0.8676	0.8594	0.8776	0.8684
Support Vector Classifier	0.8580	0.8815	0.8257	0.8527
Random Forest	0.9018	0.88	0.9295	0.9041
Decision Tree	0.8510	0.8419	0.8626	0.8521
Gradient Boosting	0.8295	0.8084	0.8617	0.8342
Ada Boost	0.7966	0.7881	0.8089	0.7984
Bagging	0.8847	0.8792	0.8908	0.8850
Extra Trees	0.9027	0.8814	0.9295	0.9048
XGBoost	0.8921	0.8726	0.9172	0.8944
LightGBM	0.8838	0.8712	0.8996	0.8852
Naive Bayes	0.6875	0.709	0.6311	0.6679
Neural Network	0.8624	0.8611	0.8626	0.8619
Bernoulli Naive Bayes	0.6936	0.6928	0.6910	0.6919

From the results above, the top 5 performance models based on accuracy are random forest, extra trees, XGBoost, LightGBM, and Bagging.

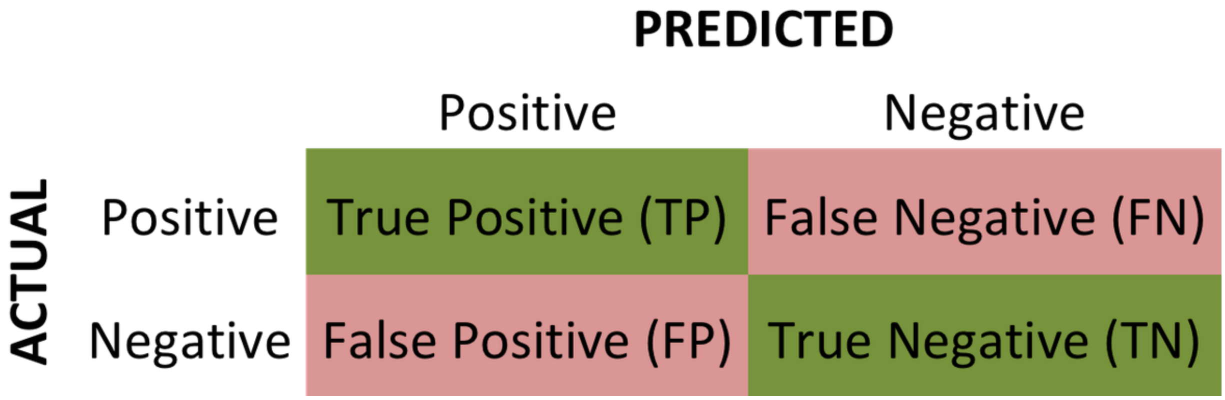
2.6.3 Tuning Model

We tried to do GridSearch for every model to match the best parameters, but the runtime is too long. We also tried to use other computational methods like using the Great Lake, and accelerating using Google Colab GPU, but those chucks still need a significant amount of time to run. Due to this limitation, we randomly sampled 5000 rows for training. We did a grid search for each model to find the best parameters, given the parameter of PCA is 8, which has the top 5 performance from the initial models. Instead of doing a simple cross validation, we did a repeated Stratified K Fold, which allows improving the estimated performance of a machine learning model, by simply repeating the cross-validation procedure multiple times, and reporting the mean result across all folds from all runs. This mean result is expected to be a more accurate estimate of the model's performance.

3. Evaluation and Analysis

3.1 Evaluation Methods

Our main evaluation method is to use the confusion matrix, accuracy, recall, precision, and F1 score to evaluate the results. In general, accuracy can be a good indicator of evaluation, since the accuracy score can reflect the performance of our models directly and is easy to interpret. However, given the medical background, we also need to consider the recall rate of our model. The goal of this project was to develop a machine learning model that could predict the probability of heart disease using as many variables as possible for 17 possible risk factors. A challenge in developing a prediction model with this type of data is the balance between accuracy and recall. Under this background, the precision score is interpreted as the proportion of the models' predictions of heart disease where heart disease is actually present. And the recall rate represents the proportion of all cases of heart disease that the model accurately predicted. Since failure to "capture" even one case of heart disease can lead to death, the model should emphasize recall scores. It is best not to "miss" any potential heart disease patients, even if this means "flagging" some patients who do not actually have heart disease.



		PREDICTED	
		Positive	Negative
ACTUAL	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Fig. 8: Confusion Matrix

3.2 Best Model & Result

After we got the best parameters for the top 5 models, we used Voting Classifier to ensemble the final model. It simply aggregates the results of each classifier into a voting classifier and predicts the coucome category based on the highest majority vote. The idea is that instead of creating separate dedicated models and finding out the accuracy of each model, we create a single model, train through these models, and predict the output based on the combined majority vote of each output category. We used soft voting in the parameter, which the result is based on the average of probability given to the classifier, instead of the highest majority votes in the hard voting. The result of the best model is listed below:

Accuracy	Precision	Recall	F1 Score
0.9075	0.8936	0.9242	0.9086

Compared to the benchmark model, our final model significantly improved the performance, no matter in which evaluation indicators.

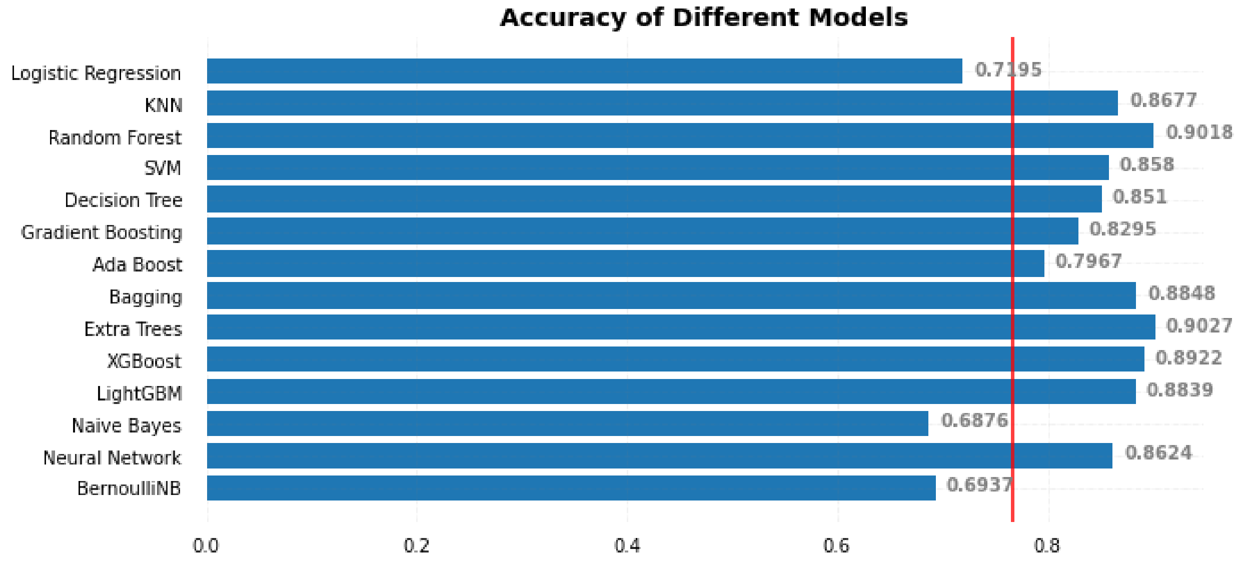


Fig. 9: Accuracy of different Models

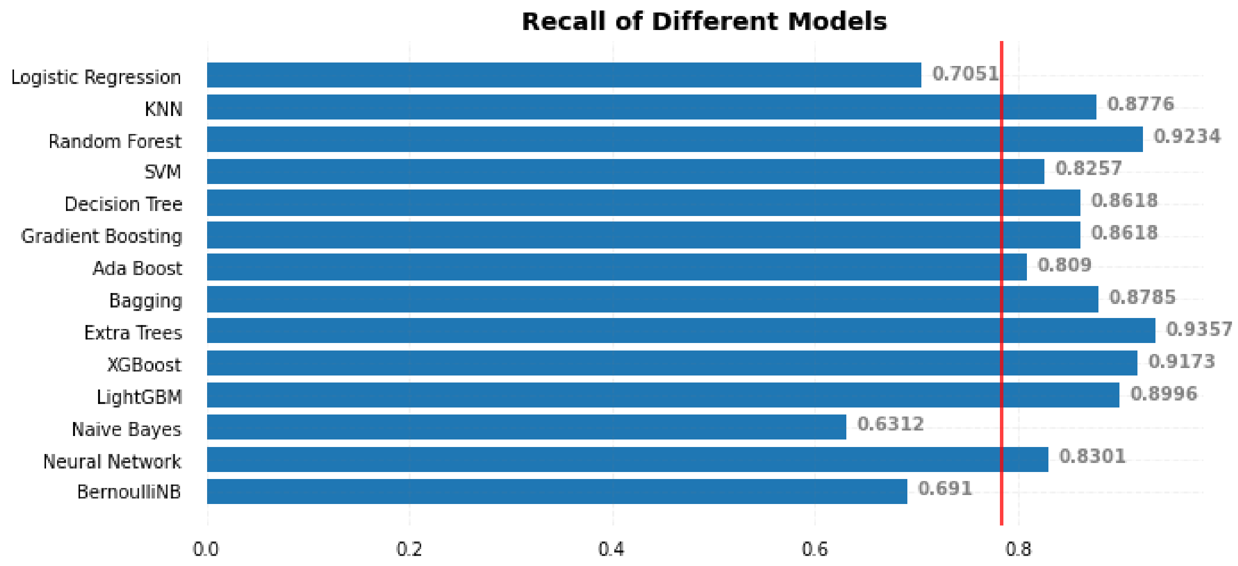


Fig. 10: Recall of different Models

4. Related Work

Trying to understand more about machine learning applications for heart disease, we reviewed some publications and related works to support our own work. The most inspirational one would be a research published on Biomedical Signal Processing and Control [9]. Their research explored numerous features and ML algorithms [9]. The models employed in their work gave us a strong background of the model choice, such as Gradient Boost, AdaBoost, Logistic Regression, Decision Tree, Random Forest, Neural Network, and Naive Bayes. Further, the features picked in their work provided us with one of the strategies of feature selection, such as Fasting Blood Sugar, Serum cholesterol. As a result, we added Diabetes and Asthma as our features. Another interesting point caught our attention is that the categorical medical characteristics are more effective compared to the numerical medical features in heart disease prediction with limited data [9], thus more categorical indicators are included in our project. Furthermore, we were inspired by another work [4] that the drawback in the existing heart disease prediction works is that they focus more on the application of classification techniques, rather than on various data cleaning and pruning techniques to prepare a dataset suitable for mining. An unclean dataset with missing values performs substantially worse than one that has been properly cleaned and pruned. Hence, we devoted great efforts to pre-processing and preparing the data to enhance accuracy.

5. Discussion and Conclusion

5.1 Overall results

Overall, our final model improved significantly compared to the benchmark model. The model can correctly predict 90.75% of the situation, and 92.42% of total relevant results correctly classified by our model. This is a pretty high accuracy and recall rate in the real world. In general, if a patient is classified as 1 - high probability of getting heart disease, he/she should pay attention to heart health. Moreover, we examed the top 5 features that can affect the heart health. The top 5 features are: Smoking, Alcohol Drinking, Stroke, Physical sex, and Age. This is very much in line with common sense. Men who smoke and drink and have a history of strokes are more likely to have heart disease as they get older. The full list of factors that have significantly effect to the heart disease can refer to Appendix C.

5.2 Summary of What We Learned

Through this project, we improved our experience in obtaining, processing, and preparing real data in the real world, as well as improving the accuracy by referring to existing works. In addition, we gained more experience in selecting appropriate methods for EDA and model prediction. Finally, we had a great time working together as a team and learning from each other to successfully complete the project.

5.3 Future Work

If we have another 6-12 months to continue working on the project, we may use better approaches to handle missing values. More importantly, we are likely to explore in depth the complex features in our dataset and analyze how they affect the predictions. Real-world data is far more complex than scikit-learn data, and we need to have more careful consideration when making predictions. For example, we dropped the gestational diabetes data within the diabetes feature due to time limitations. This may lead to a decrease in the accuracy of our prediction model for diabetes during pregnancy. Another example, White people represent the majority in the dataset, the unbalanced data distribution and gaps of the data may lead to biases in the prediction model for different races. Last but not least, since our dataset is too large, we have to sample from the dataset to do model training. However, if we have more computational resources, we would be able to use the whole dataset as training data to get more accurate predictions. These are limitations that exist in this project. If given more time, we hope to make a more comprehensive, practical and equitable prediction model for everyone in the future.

Reference

[1] “Multiple cause of death, 1999-2020 request,” Centers for Disease Control and Prevention. [Online]. Available: https://wonder.cdc.gov/mcd-icd10.html. [Accessed: 30-Oct-2022].

[2] C. W. Tsao, A. W. Aday, Z. I. Almarzooq, A. Alonso, A. Z. Beaton, M. S. Bittencourt, A. K. Boehme, A. E. Buxton, A. P. Carson, Y. Commodore-Mensah, M. S. V. Elkind, K. R. Evenson, C. Eze-Nliam, J. F. Ferguson, G. Generoso, J. E. Ho, R. Kalani, S. S. Khan, B. M. Kissela, K. L. Knutson, D. A. Levine, T. T. Lewis, J. Liu, M. S. Loop, J. Ma, M. E. Mussolino, S. D. Navaneethan, A. M. Perak, R. Poudel, M. Rezk-Hanna, G. A. Roth, E. B. Schroeder, S. H. Shah, E. L. Thacker, L. B. VanWagner, S. S. Virani, J. H. Voecks, N.-Y. Wang, K. Yaffe, and S. S. Martin, “Heart disease and stroke statistics—2022 update: A report from the American Heart Association,” Circulation, vol. 145, no. 8, 2022.

[3] https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

[4] Learning, M. (2017). Heart disease diagnosis and prediction using machine learning and data mining techniques: a review. Advances in Computational Sciences and Technology, 10(7), 2137-2159.

[5] https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1

[6] “CDC - about BRFSS,” Centers for Disease Control and Prevention, 16-May-2014. [Online]. Available: https://www.cdc.gov/brfss/about/index.htm. [Accessed: 08-Dec-2022].

[7] “2021 BRFSS Questionnaire .” The Behavioral Risk Factor Surveillance System, 08-Jun-2022.

[8] M. N. N. Ramli, A. S. Yahaya, N. A. Ramli, N. F. F. M. Yusof, and M. M. A. Abdullah, “Roles of imputation methods for filling the missing values: a review,” Advances in Environmental Biology, vol. 7, no. 12, p. 3861, Oct. 2013.

[9] C. Pan, A. Poddar, R. Mukherjee, and A. K. Ray, “Impact of categorical and numerical features in ensemble machine learning frameworks for heart disease prediction,” Biomedical Signal Processing and Control, vol. 76, p. 103666, 2022.

Appendix A

Variable in original dataset	Notes	Variable in cleaned dataset	Notes
_MICHD *Calculated variable for respondents who have ever reported having coronary heart disease (CHD) or myocardial infarction (MI).*	1: detected 2: not detected	**HeartDisease** *- categorical*	0: not detected 1: detected
_SMOKER3 *Calculated variable for four-level smoker status: everyday smoker, someday smoker, former smoker, non-smoker*	1: everyday 2: some days 3: former 4: never smoked 9: missing/no/dk	**Smoking** *- categorical* *Did you smoke at least 100 cigarettes in your entire life?*	0: no 1: yes
_RFDRHV7 *Calculated variable for heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)*	1: no 2: yes	**AlcoholDrinking** *- categorical* *Are you a heavy drinker?* *(adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)*	0: no 1: yes
CVDSTRK3 *(Ever told) (you had) a stroke?*	1: yes 2: no 7: don't know / not sure 9: refused	**Stroke** *- categorical* *(Ever told) (you had) a stroke?*	0: no 1: yes
PHYSHLTH *Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?*	Number of days (01-30) 88: None 77: Don't know/not sure 99: Refused	**PhysicalHealth** *- numerical* *Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good?*	0: no 1-30: days physical health were not good during the past 30 days

MENTHLTH *Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?*	Number of days (01- 30) 88: None 77: Don’t know/not sure 99: Refused	**MentalHealth** *- numerical* *Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?*	0: no 1-30: days mental health were not good during the past 30 days
DIFFWALK *Do you have serious difficulty walking or climbing stairs?*	1: yes 2: no 7: Don’t know / Not sure 9: refused	**DiffWalking** *- categorical* *Do you have serious difficulty walking or climbing stairs?*	0: no 1: yes
_SEX *Calculated variable for calculated sex variable.*	1: male 2: female	**Sex** *- categorical*	0: female 1: male
_AGEG5YR *Calculated variable for fourteen-level age category.*	1: 18-24 2: 25-29 3: 30-34 4: 35-39 5: 40-44 6: 45-49 7: 50-54 8: 55-59 9: 60-64 10: 65-69 11: 70-74 12: 75-79 13: 80+ 14 not found	**AgeCategory** *- categorical*	1: 18-24 2: 25-29 3: 30-34 4: 35-39 5: 40-44 6: 45-49 7: 50-54 8: 55-59 9: 60-64 10: 65-69 11: 70-74 12: 75-79 13: 80+ 14 not found
_RACE	1: White only, non-Hispanic 2: Black only, non-Hispanic 3: American Indian or Alaskan Native only, Non-Hispanic 4: Asian only,	**Race** *- categorical*	1: White only, non-Hispanic 2: Black only, non-Hispanic 3: American Indian or Alaskan Native only, Non-Hispanic

	non-Hispanic 5: Native Hawaiian or other Pacific Islander only, Non-Hispanic 6: Other race only, non-Hispanic 7: Multiracial, non-Hispanic 8: Hispanic 9: Don’t know/Not sure/ Refused		4: Asian only, non-Hispanic 5: Native Hawaiian or other Pacific Islander only, Non-Hispanic 6: Other race only, non-Hispanic 7: Multiracial, non-Hispanic 8: Hispanic 9: Don’t know/Not sure/ Refused
DIABETE4 *(Ever told) (you had) diabetes*	1: yes 2: yes, but female told only during pregnancy 3: no 4: no, prediabetes or borderline diabetes 7: don’t know / Not sure 9: refused	**Diabetic** *- categorical* *(Ever told) (you had) diabetes*	0: no 1: yes
_TOTINDA *Calculated variable for adults who reported doing physical activity or exercise during the past 30 days other than their regular job.*	1: Had physical activity or exercise 2: No physical activity or exercise in last 30 days 9: Don’t know/Refused/ Missing	**PhysicalActivity** *- categorical* *Did you have physical activity or exercise during the past 30 days other than their regular job.*	0: no 1: yes
GENHLTH *Would you say that in general your health is—*	1: Excellent 2: Very Good 3: Good 4: Fair 5: Poor Do not read:	**GenHealth** *- categorical* *Would you say that in general your health is—*	1: Excellent 2: Very Good 3: Good 4: Fair 5: Poor

	7: Don’t know/Not sure 9: Refused		
ASTHMA3 *(Ever told) (you had) asthma?*	1: Yes 2: No 7: Don’t know / Not sure 9: Refused	**Asthma** *- categorical* *(Ever told) (you had) asthma?*	0: no 1: yes
CHCKDNY2 *Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?*	1: Yes 2: No 7: Don’t know / Not sure 9: Refused	**KidneyDisease** *- categorical*	0: no 1: yes
CHCSCNCR *(Ever told) (you had) skin cancer?*	1: Yes 2: No 7: Don’t know / Not sure 9: Refused	**SkinCancer** *- categorical*	0: no 1: yes
_VEGLT1A *Calculated variable for consume vegetables 1 or more times per day.*	1: Consumed vegetables one or more times per day 2: Consumed vegetables less than one time per day 9: missing	**VegetableAndFruits** *- categorical* *Did you consume vegetables one or more times per day*	0: no 1: yes
/	/	**BMI_new** *- numerical* Calculated from (WTKG3 / 100) / (HTM4/ 100)^2 Based on the bmi formula	

Table 1: variables in original and cleaned dataset

Appendix B

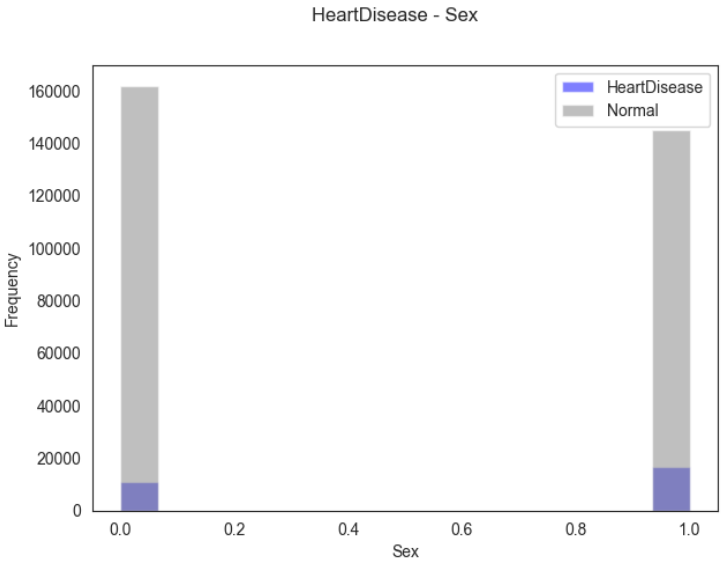


Fig. 1: Heart Disease - Sex
(1: male, 0: female)

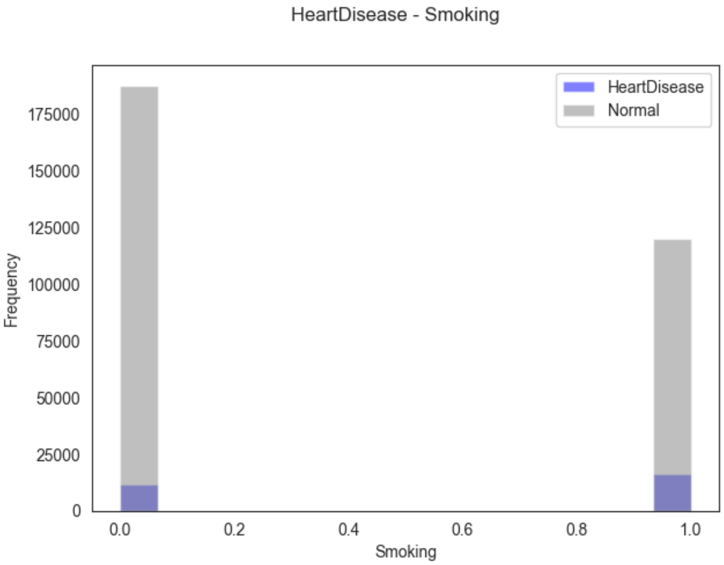


Fig. 2: Heart Disease - Smoking
(1: People who have smoking history 0: People do not have smoke history)

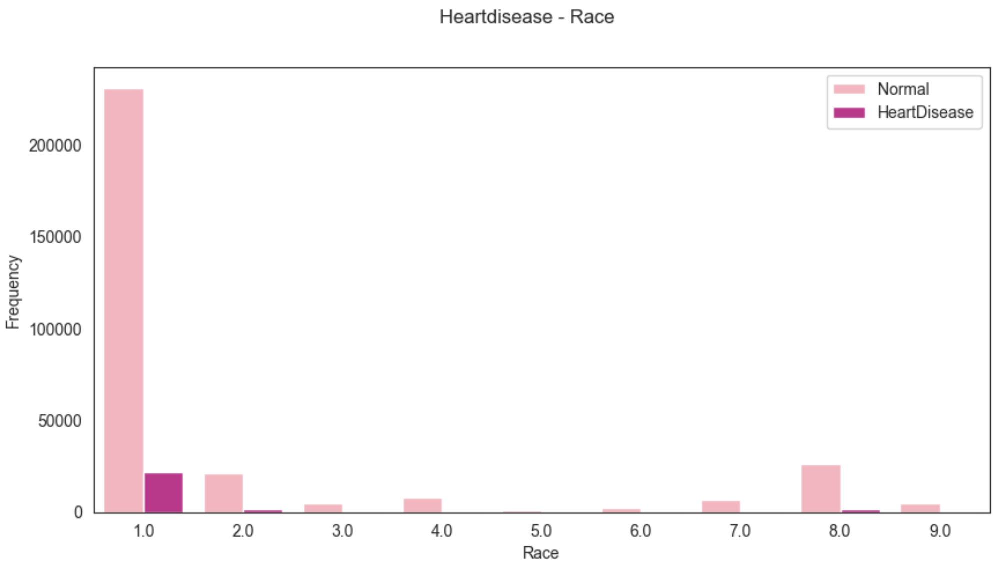


Fig. 3: Heart disease - Race
(1: White only, non-Hispanic; 2: Black only, non-Hispanic; 3: American Indian or Alaskan Native only, Non-Hispanic; 4: Asian only, non-Hispanic; 5: Native Hawaiian or other Pacific Islander only, Non-Hispanic; 6: Other race only, non-Hispanic; 7: Multiracial, non-Hispanic; 8: Hispanic; 9: Don't know/Not sure/ Refused)

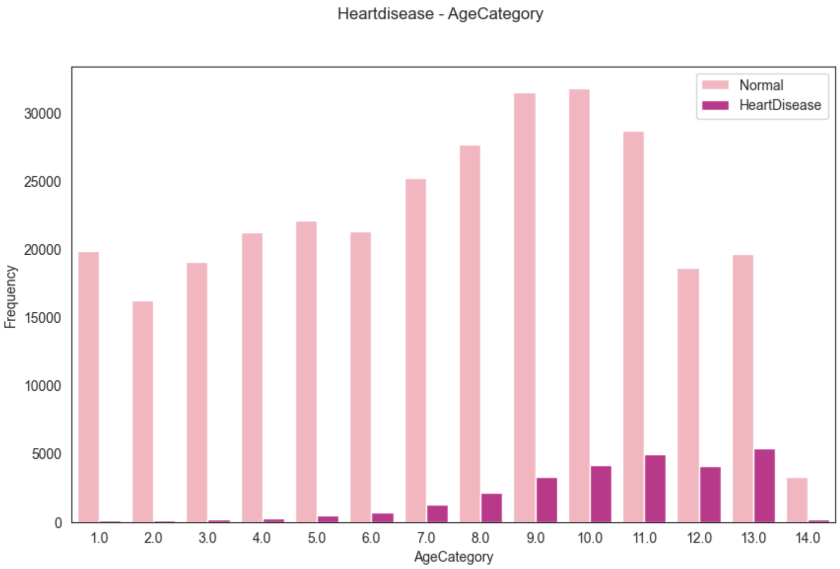


Fig. 4: Heart Disease - Age Category
(Age category: 1: 18-24; 2: 25-29; 3: 30-34; 4: 35-39; 5: 40-44; 6: 45-49; 7: 50-54; 8: 55-59; 9: 60-64; 10: 65-69; 11: 70-74; 12: 75-79; 13: 80+; 14 not found)

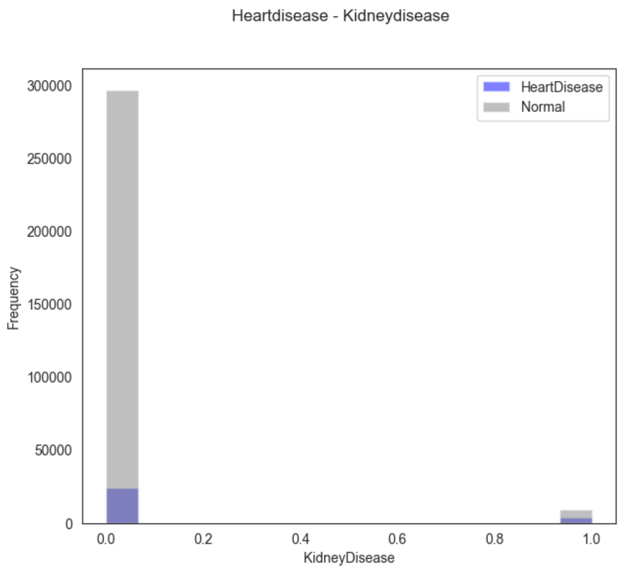


Fig. 5: Heart Disease - Kidney Disease
(1: people who have Kidney disease [not including kidney stones, bladder infection or incontinence], 0: people who do not have Kidney disease)

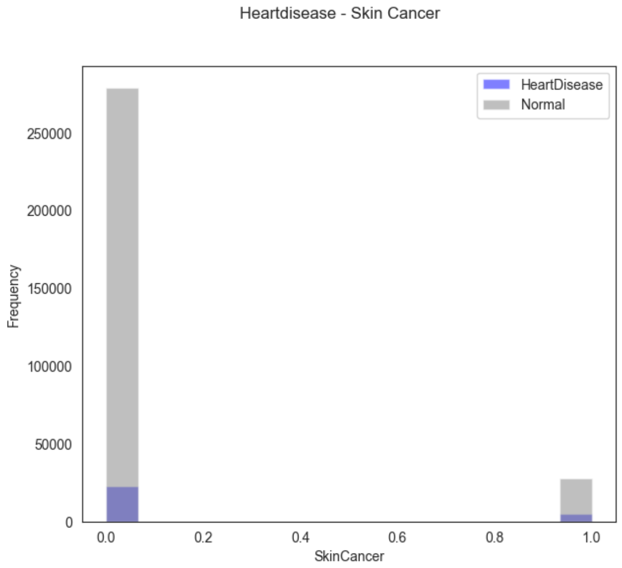


Fig. 6: Heart disease - Skin Cancer
(1: people who have skin cancer, 0: people do not have skin cancer)

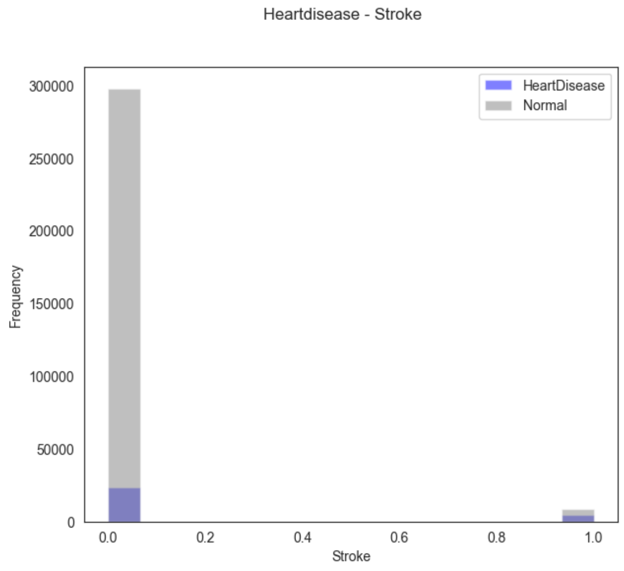


Fig. 7: Heart Disease - Stroke
(1: people who have Stroke, 0: people do not have Stroke)

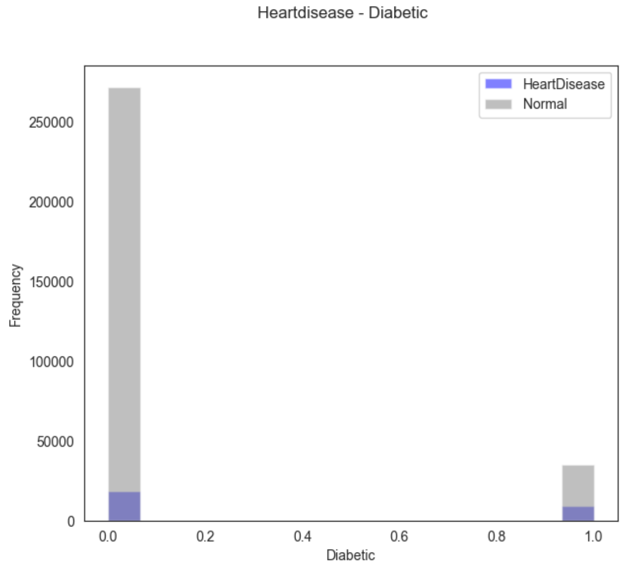


Fig. 8: Heart Disease - Diabetic
(1: people who have Diabetic, 0: people do not have Diabetic)

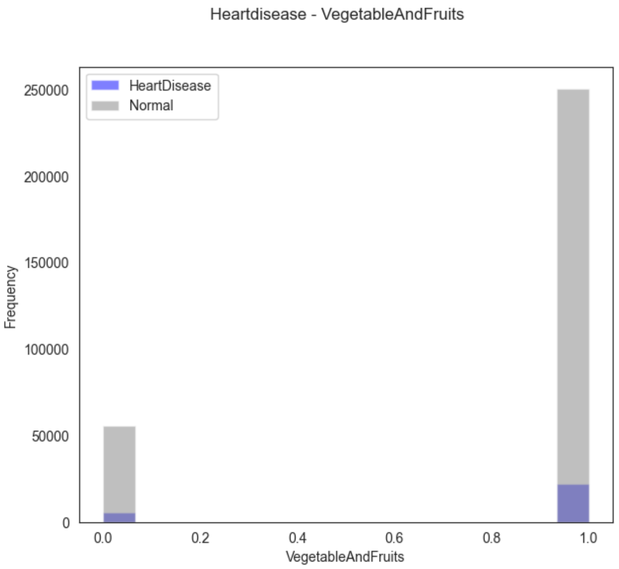


Fig. 9: Heart Disease - Vegetable And Fruits
(1: consume vegetables one or more times per day, 0: consumed vegetables less than one time per day)

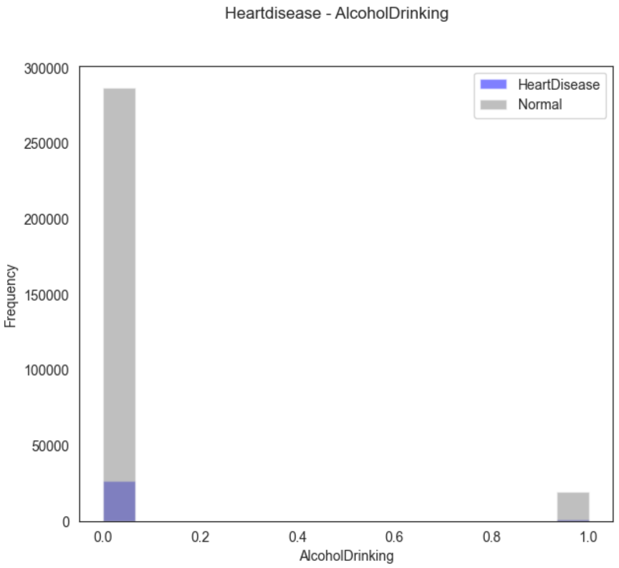


Fig. 10: Heart Disease - Alcohol Drinking
(1: Heavy drinkers [adult men having more than 14 drinks per week and adult women having more than 7 drinks per week], 0: Not heavy drinkers)

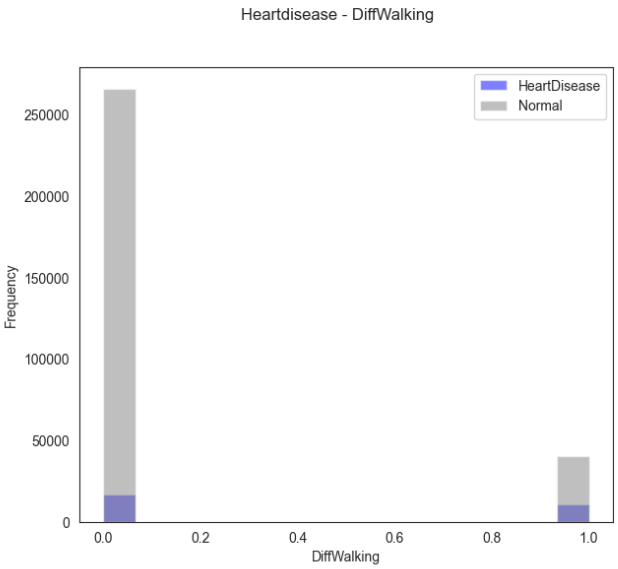


Fig. 11: Heart Disease - DiffWalking
(1: have serious difficulty walking or climbing stairs, 0: do not have serious difficulty walking or climbing stairs)

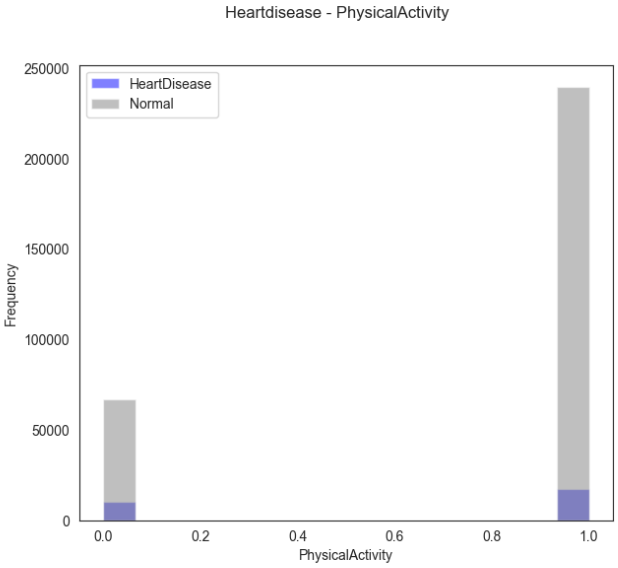


Fig. 12: Heart Disease - Physical Activity
(1: People who reported doing physical activity or exercise during the past 30 days other than their regular job, 0: People who did not report doing physical activity or exercise during the past 30 days other than their regular job)

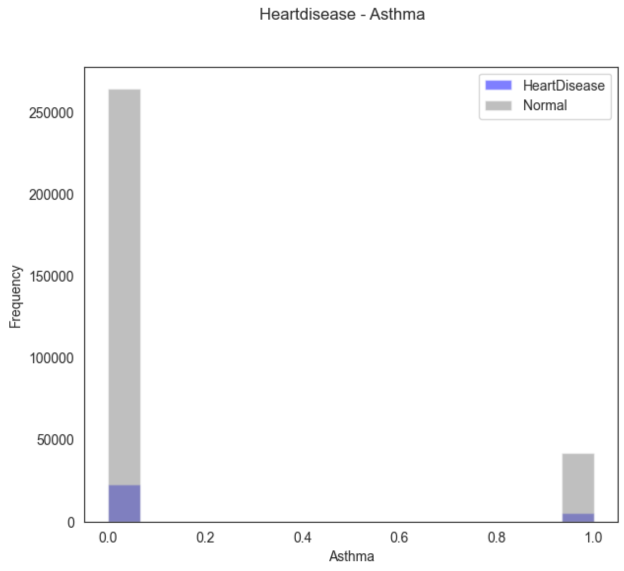


Fig. 13: Heart Disease - Asthma
(1: people who have Asthma, 0: people who do not have Asthma)

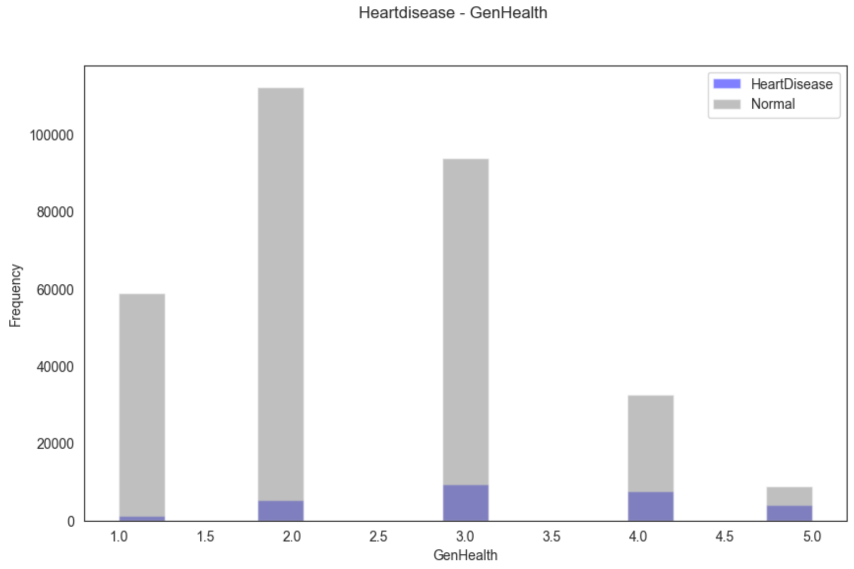


Fig. 14: Heart Disease - GenHealth
(People say in general their health is: 1: Excellent, 2 Very Good, 3 Good, 4 Fair, 5 Poor)

Appendix C

Importance of factor	Factors that can lead to heart disease
1	Smoking
2	AlcoholDrinking
3	Stroke
4	Sex
5	AgeCategory
6	Diabetic
7	PhysicalActivity
8	GenHealth
9	Asthma
10	KidneyDisease
11	SkinCancer
12	BMI