Prediction for Heart Disease

1st Chen Yao
University of Rochester
Rochester, United States
cyao10@u.rochester.edu

2nd Naman Bharara
University of Rochester
Rochester, United States
nbharara@u.rochester.edu

3nd Yiwei Han
University of Rochester
Rochester, United States
yhan32@u.rochester.edu

4nd Hanyang Zhang
University of Rochester
Rochester, United States
hzh112@u.rochester.edu

5nd Yunran Yin
University of Rochester
Rochester, United States
yyin19@u.rochester.edu

***Abstract*—Many factors contributes to heart disease. We first used preprocessing techniques on variables, and used different machine learning methods to predict whether a person under certain variables have heart disease.**

***Index Terms*—Key Words: Heart Disease Prediction, Logistic Regression, Random Forest, Neural Network, One-Hot Encoding, Oversampling**

I. Introduction

Our heart disease prediction project established a model utilizing two algorithms to increase the accuracy of our data's prediction. To predict an individual's cardiovascular health result, we used features from the Centers for Disease Control and Prevention's (CDC) Personal Key Indicators of Heart Disease dataset, accessible on Kaggle [1]. The first reason we chose to make a prediction project for heart disease, especially from this dataset, is due to the variety of possible causes and factors in it. Second, we discovered numerous earlier research on the causes of heart disease [2]. However, relatively few models have been developed to predict and calculate the weight of an individual's health outcomes. Thus, our ultimate goal is to develop the most accurate model and identify the most weighted predictive variables for heart disease outcomes. For people at a high risk of heart disease, we believe that our prediction model will be helpful in alerting them of their predisposition and motivate them to make lifestyle changes that will reduce the risk. When we used oversampling in all of our models while building our models, we saw that the accuracy increased from 70 percent to 90 percent. We used three models to seek the highest accuracy initially. We achieved an accuracy of 0.9164 for the Logistic Regression model, accuracy of 0.9156 for Deep learning, and an accuracy of 0.9159 for Random Forest.

II. Data

The dataset we used is called Personal Key Indicators of Heart Disease from Kaggle [1]. It contains roughly 400k results from CDC 2020 annual survey data regarding heart disease. This dataset originally contains 300 variables but is now being reduced to 17 variables which are considered the most important to heart disease. The dataset is stored in heart_2020_cleand.csv, and it has 319795 rows and 18 columns of data. There are 17 columns of variables, and 1 target column, which is whether that person has heart disease. We extracted several important variables that we found more significantly related to having heart disease.

For categorical data, we did distribution analysis to see whether they have certain relationship with having heart disease. For age, we divided into four parts, female and male with heart disease and without heart disease. From the graph below, we see that there's noticeable larger proportion of male having heart disease than female.

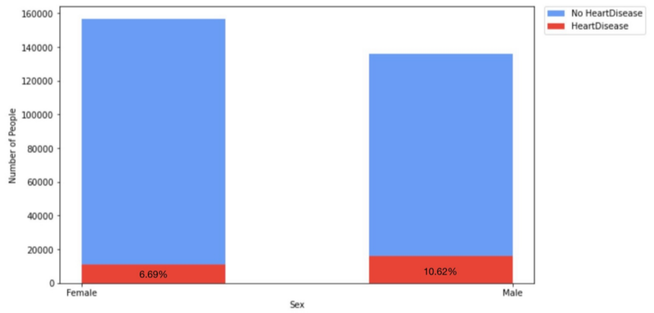


Fig. 1. Percentage of Heart Disease in Different Sex Group

For smoking, we also divided into 4 parts, smoking and non-smoking with heart disease and without heart disease. We can also see the proportion of people smoking with heart disease is double to the proportion of people not smoking. So there is a positive relation with smoking and having heart disease.

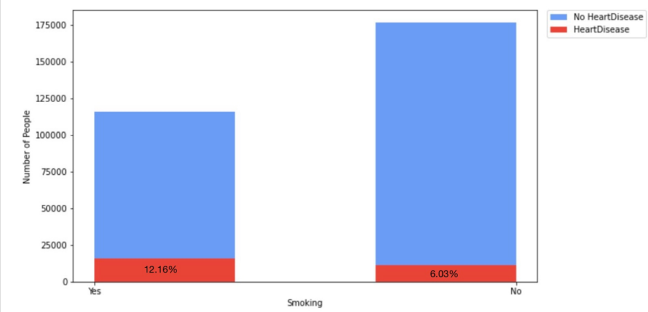
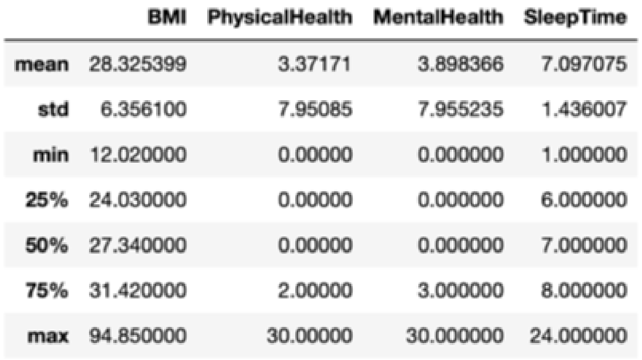


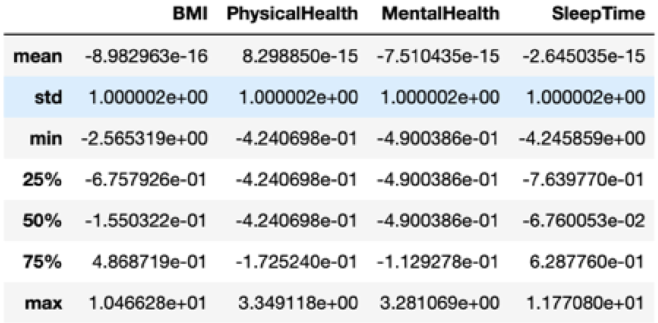
Fig. 2. Percentage of Heart Disease in Different Smoking Group

There are 4 numerical variables “BMI”, “PhysicalHealth”, “MentalHealth” and “SleepTime”. We first find the distribution of those variables. As the table indicated, the distributions of “PhysicalHealth” and “MentalHealth” are similar, the min, mean, max and std does not vary a lot. However, comparing to other two variables “BMI” and “SleepTime”, the distribution is different, the range and std both vary a lot. So we did standardization to the four columns to make the distribution in the same range.



	BMI	PhysicalHealth	MentalHealth	SleepTime
mean	28.325399	3.37171	3.898366	7.097075
std	6.356100	7.95085	7.955235	1.436007
min	12.020000	0.00000	0.000000	1.000000
25%	24.030000	0.00000	0.000000	6.000000
50%	27.340000	0.00000	0.000000	7.000000
75%	31.420000	2.00000	3.000000	8.000000
max	94.850000	30.00000	30.000000	24.000000

Fig. 3. Summary of Numerical Data before Standardizing



	BMI	PhysicalHealth	MentalHealth	SleepTime
mean	-8.982963e-16	8.298850e-15	-7.510435e-15	-2.645035e-15
std	1.000002e+00	1.000002e+00	1.000002e+00	1.000002e+00
min	-2.565319e+00	-4.240698e-01	-4.900386e-01	-4.245859e+00
25%	-6.757926e-01	-4.240698e-01	-4.900386e-01	-7.639770e-01
50%	-1.550322e-01	-4.240698e-01	-4.900386e-01	-6.760053e-02
75%	4.868719e-01	-1.725240e-01	-1.129278e-01	6.287760e-01
max	1.046628e+01	3.349118e+00	3.281069e+00	1.177080e+01

Fig. 4. Summary of Numerical Data after Standardizing

We did a box-plot to see the range of BMI of people with or without heart disease. We can see that the mean, 1st quadrant, 3rd quadrant, lower bound and upper bound of BMI for people with heart disease are all higher than the ones without. As a result, there is a positive relation between BMI and heart disease.

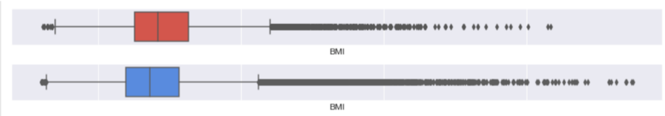
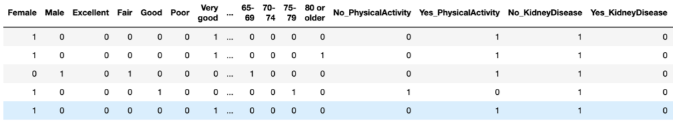


Fig. 5. Box Plot Distribution of BMI with and without Heart Disease

We also used one-hot encoding to processed all categorical data. This can partition each category within each features to different columns. For the category each row lies within, it would use 1, and the ones that not, use 0 to indicate. Afterwards, it is more convenient for us to do correlation matrix analysis, since it can only take numerical data. The darker the color in the correlation matrix in the row or column with “HeartDisease”, the more relation it has with heart disease. For example, we can see that “Yes_stroke”, “Yes_DiffWalking” and “80 or older” are darker, so they would contribute to heart disease more than other variables.



Female	Male	Excellent	Fair	Good	Poor	Very good	...	65-69	70-74	75-79	80 or older	No_PhysicalActivity	Yes_PhysicalActivity	No_KidneyDisease	Yes_KidneyDisease
1	0	0	0	0	0	1	...	0	0	0	0	0	1	1	0
1	0	0	0	0	0	1	...	0	0	0	1	0	1	1	0
0	1	0	1	0	0	0	...	1	0	0	0	0	1	1	0
1	0	0	0	1	0	0	...	0	0	1	0	1	0	1	0
1	0	0	0	0	0	1	...	0	0	0	0	0	1	1	0

Fig. 6. Categorical Data after One-Hot Encoding

The target column “HeartDisease” is binary, which contains ”yes” and ”no”, which indicates whether a person has heart disease. Moreover, from the pie chart distribution of ”HeartDisease” column, we can see the dataset is very unbalanced. Only 8.56% of people report that they have heart disease, while the majority don’t. Since the data is unbalanced, it is very likely that regular prediction model would predict more people without heart disease, we need to do oversampling method to this data set to perform more accurate prediction base on variables. We use oversampling to fit the training dataset, which makes the number of having or not having heart disease balance.

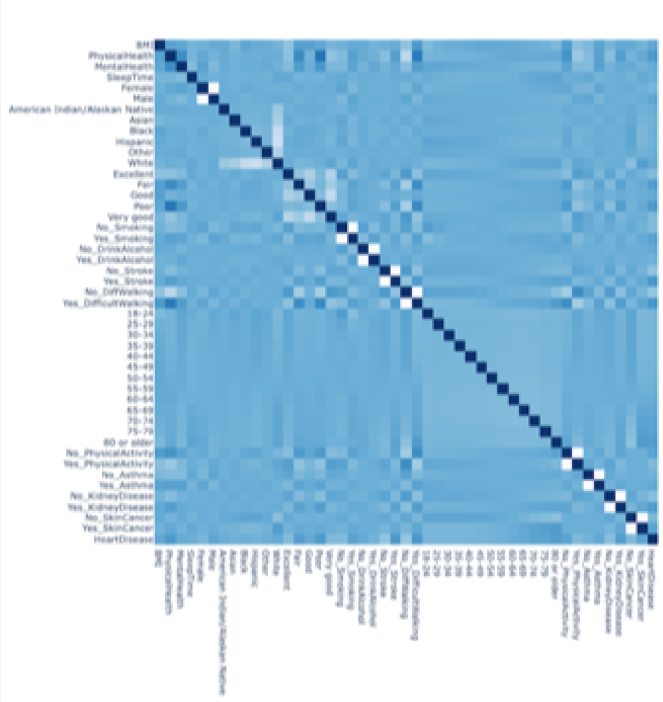


Fig. 7. Correlation Plot

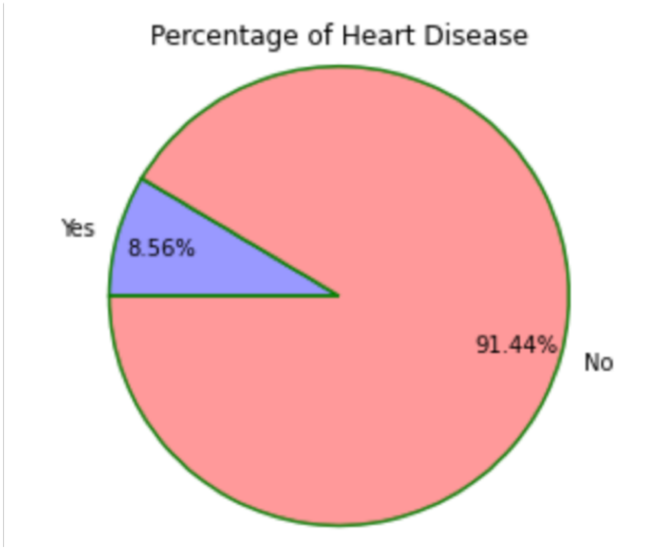


Fig. 8. Percentage of Heart Disease

III. Literature Review

Existing research on the topic was utilized in order to guide the study. According to a paper by Rachel Hajar, the major risk factors associated with coronary artery disease were having high blood pressure, high cholesterol levels, smoking, being overweight, smoking, diabetes, lack of physical activity, stress, and an unhealthy diet [4]. These factors are considered as controllable factors, however the article also mentions the presence of uncontrollable factors, such as age, sex, race, and family history [4]. A study on the risk factors associated with Cardiovascular disease in Nevadans found similar results, in this study BMI, was found to be the most prevalent factor followed by hypertension. Another factor that the study found was that males were 1.64 times likely to have Cardiovascular disease than females, further suggesting the sex can play a role in cardiovascular disease. The study also noted that individuals with high cholesterol were 2.67 times more likely to have cardiovascular disease. Notably the study found that hypertension increased to magnitude of risk to be 6.18 times more likely than a normal individual [5]. The research conducted by Jonathan C. Brown, Thomas E. Gerhardt, and Edward Kwon, reaffirms the previously mentioned factors, however the findings also mentioned additional risk factors, such as kidney disease, liver disease, and arthritis [8]. As a result, these factors were noted to be factors of interest due to their prevalence in existing research. A study utilizing Machine Learning and Deep learning to predict heart disease was studied to consider which methods, and factors should be utilized for the prediction model. This paper utilized a different dataset, however, the research provided a basis for our inclusion of Logistic Regression, Random Forest, and Neural Networks for our prediction models. Additionally, the paper provided a basis of feature selection, such as the inclusion of age, sex, and overall health. The results of this paper also provided a basis for the usage of confusion matrices as a form of evaluation for the approaches used within this paper [6]. Another paper that shared a similar goal of heart disease prediction utilized the Random Forest Classifier and was able to achieve an accuracy of 83%, which suggested that further refinement of the algorithm, and improvements to the data preprocessing could make the algorithm a viable option for the prediction of heart disease [7].

IV. Goals

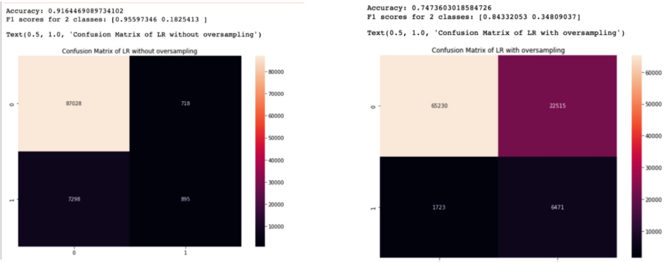
Our preliminary goal is to use multiple personal health data to predict the possibility of having heart disease. If our models can successfully sift out the patient with heart disease, the models might save people's lives. Besides the predicative task, we also planned to perform Logistic Regression, which has explanatory coefficients. Thus, we could possibly better understand about the key indicators of Heart Disease, by evaluating the values of coefficients of Logistic Regressor. Additionally, in order to achieve better performance of our predication, we will compare multiple models with different sampling techniques.

V. Results

A. Logistic Regression

The first method we used is Logistic Regression. Logistic regression is a tool used for discovering the relationship between the dependent variable and the independent variables. The package we are using here is from sklearn linear model. It is pretty straightforward as we set the parameter random_state to 265 and implement trained and tested variables. We ran the model for both the original data set and the dataset after over sampling. The accuracy for original data set is

0.9164469089734102, and the one for over sampled data set is 0.7473603018584726. The F1 score for having heart disease without over sampling is 0.1825413, without over sampling is 0.34809037. Below is our confusion matrix for Logistic Regressiossionn using for calculating.

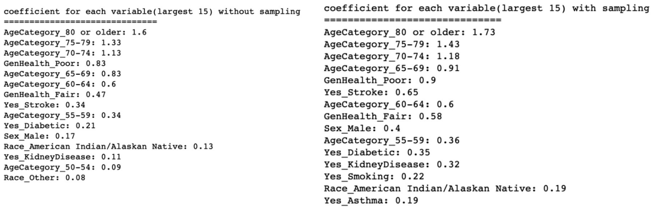


```
Accuracy: 0.9164469089734102
F1 scores for 2 classes: [0.95597346 0.1825413 ]

Text(0.5, 1.0, 'Confusion Matrix of LR without oversampling')
```

```
Accuracy: 0.7473603018584726
F1 scores for 2 classes: [0.84332053 0.34809037]

Text(0.5, 1.0, 'Confusion Matrix of LR with oversampling')
```

Fig. 9. Confusion Matrix of Logistic Regression Method with/without Oversampling

We also modeled the top 15 coefficients that has the to relations and best predict the heart disease shown in the table below. We can see both with or without sampling, the top variables that contribute to heart disease is basically the same. And as our description in the data, we showed that male has for heart disease than female, it is proved here that "Sex_Male" is one of the top 15 most important factors. Same as the darker correlation matrix spot " Yes_Stroke" and "AgeCategory_80 or older", which both appear in the top 15 important categories.



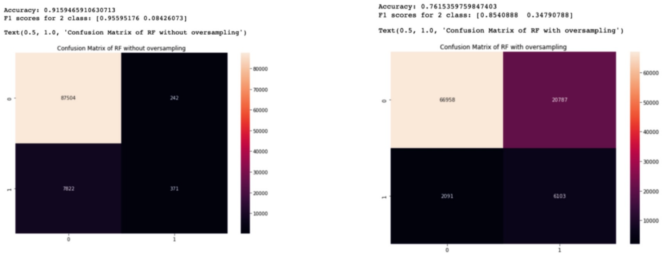
```
coefficient for each variable(largest 15) without sampling
==============================
AgeCategory_80 or older: 1.6
AgeCategory_75-79: 1.33
AgeCategory_70-74: 1.13
GenHealth_Poor: 0.83
AgeCategory_65-69: 0.83
AgeCategory_60-64: 0.6
GenHealth_Fair: 0.47
Yes_Stroke: 0.34
AgeCategory_55-59: 0.34
Yes_Diabetic: 0.21
Sex_Male: 0.17
Race_American Indian/Alaskan Native: 0.13
Yes_KidneyDisease: 0.11
AgeCategory_50-54: 0.09
Race_Other: 0.08
```

```
coefficient for each variable(largest 15) with sampling
==============================
AgeCategory_80 or older: 1.73
AgeCategory_75-79: 1.43
AgeCategory_70-74: 1.18
AgeCategory_65-69: 0.91
GenHealth_Poor: 0.9
Yes_Stroke: 0.65
AgeCategory_60-64: 0.6
GenHealth_Fair: 0.58
Sex_Male: 0.4
AgeCategory_55-59: 0.36
Yes_Diabetic: 0.35
Yes_KidneyDisease: 0.32
Yes_Smoking: 0.22
Race_American Indian/Alaskan Native: 0.19
Yes_Asthma: 0.19
```

Fig. 10. Top 15 Coefficient in Logistic Regression with/without Oversampling

B. Random Forest

It is a classification algorithm made up of many decision trees. When building each individual tree, it employs bagging and feature randomness in an attempt to create an uncorrelated forest of trees whose prediction by committee is more accurate than any individual tree. Each individual tree in the random forest predicts a class, and the class with the most votes becomes our model's prediction.We understand that some trees may be incorrect during the prediction process. Many other trees, however, will be suitable so that the trees can move in the right direction as a group. The accuracy for original data set is 0.9159465910630713, and the one for over sampled data set is 0.7615359759847403. The F1 score for having heart disease without over sampling is 0.08426073, without over sampling is 0.34790788. Below is our confusion matrix for Random Forest using for calculating.



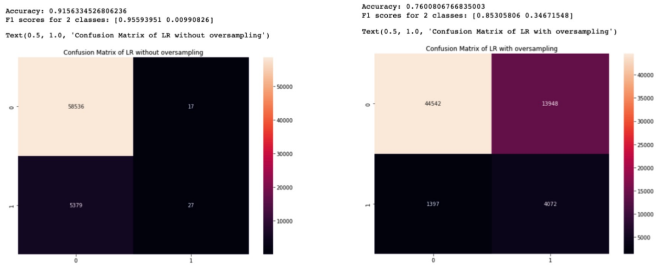
```
Accuracy: 0.9159465910630713
F1 scores for 2 class: [0.95595176 0.08426073]

Text(0.5, 1.0, 'Confusion Matrix of RF without oversampling')
```

```
Accuracy: 0.7615359759847403
F1 scores for 2 class: [0.8540888  0.34790788]

Text(0.5, 1.0, 'Confusion Matrix of RF with oversampling')
```

Fig. 11. Confusion Matrix of Random Forest Method with/without Oversampling

C. Neural Network

Our model has three layers: the input layer, the hidden layer, and the output layer. The input layer's vector is set to 48, the hidden layer's vector is set to 30, the output layer's vector is set to 12. After increasing the number of hidden layers, the accuracy of the result did not improve. We set learning rate to 0.01, and our function was programmed to run 10000 times in total. We used cross-entropy loss as our criterion. In addition, we use SGD to optimize the new weight. We used this model to train and test our data. The accuracy for original data set is 0.9156334526806236, and the one for over sampled data set is 0.7600806766835003. The F1 score for having heart disease without over sampling is 0.00990826, without over sampling is 0.34671548. Below is our confusion matrix for Neural Network using for calculating.



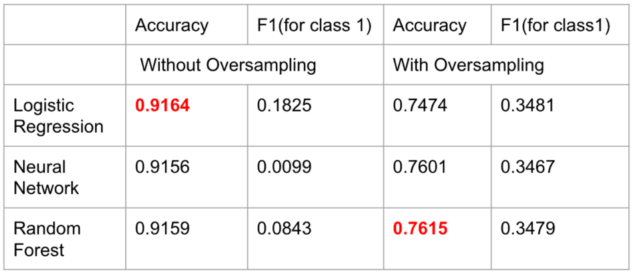
```
Accuracy: 0.9156334526806236
F1 scores for 2 classes: [0.95593951 0.00990826]

Text(0.5, 1.0, 'Confusion Matrix of LR without oversampling')
```

```
Accuracy: 0.7600806766835003
F1 scores for 2 classes: [0.85305806 0.34671548]

Text(0.5, 1.0, 'Confusion Matrix of LR with oversampling')
```

Fig. 12. Confusion Matrix of Neural Network Method with/without Oversampling

D. Comparison

Below is the table that summarize all the accuracy from different models as well as with or without over sampling. We can see Logistic Regression's accuracy is the highest in the original data set, while Random Forest's accuracy is the highest in the oversampled dataset. Moreover, we can see the accuracy for oversampling is generally lower than the accuracy for without oversampling. This is because our original dataset is very unbalance, most people do not have heart disease, so the model without overdamping would more likely to predict a person without heart disease. With oversampling method, we took more people with heart disease, thus make the dataset balanced. As a result, the model will not likely to predict people without heart disease, to be specific, the likelihood is the same. So the accuracy is reduced. For the F1 score for class 1(with heart disease), it is also higher with oversampling. For our question, we hope that the F1 score could be higher, since the higher the F1 score for class 1, the more likely we would find out a person has heart disease with our model. With oversampling significantly increase the F1 score for class 1, which is a better model and prediction. We could successfully judge whether a person have heart disease with the model, thus save more lives of people.



	Accuracy	F1(for class 1)	Accuracy	F1(for class1)
	Without Oversampling		With Oversampling	
Logistic Regression	**0.9164**	0.1825	0.7474	0.3481
Neural Network	0.9156	0.0099	0.7601	0.3467
Random Forest	0.9159	0.0843	**0.7615**	0.3479

Fig. 13. Summary of Accuracy and F1 Score

VI. Conclusion

Based on our results, all of our models achieved very good accuracy, without doing any sampling, because of unbalance of dataset. However, due to the desire of having better performance of identifying people who have heart disease, we also need to consider some metrics other than accuracy. In our project, we also used F1 score of class1 (people have heart disease) to measure the performance. To increase the F1 score for class1, we applied oversampling to our dataset. This technique greatly increased the F1 score, which means our model could find out patients having heart diseases better. Although improved model has some decrease in accuracy, the increment in F1 score enables the models to save more lives.

Besides the metrics of models, we also identified some important features that contributes to heart disease. Based on our result from Logistic Regressor, age is the feature having strongest positive correlation with the possibility of having heart disease. Other important features include general health status of patient, experience of having stroke, having diabetics or not, and so on.

References

[1] K. Pytlak, “Personal key indicators of heart disease” ,Kaggle, 16-Feb-2022. [Online]. Available: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease. [Accessed: 16-Apr-2022].

[2] Benjamin, Emelia J., et al. ”Heart disease and stroke statistics—2018 update: a report from the American Heart Association.” Circulation 137.12 (2018): e67-e492.

[3] “Know your risk for heart disease,” Centers for Disease Control and Prevention, 09-Dec-2019. [Online]. Available: https://www.cdc.gov/heartdisease/risk_factors.htm: :text=
These%20are%20called%20risk%20
factors,%2C%20high%20cholesterol%2C%20and%20smoking
.amp;text=Some%20risk%20factors%20for%20heart
,your%20age%20or%20family%20history. [Accessed: 02-May-2022].

[4] R. Hajar, “Risk factors for coronary artery disease: Historical perspectives,” Heart views : the official journal of the Gulf Heart Association, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5686931/. [Accessed: 02-May-2022].

[5] D.-M. T. Tran, N. Lekhak, K. Gutierrez, and S. Moonie, “Risk factors associated with cardiovascular disease among adult nevadans,” PLOS ONE. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0247105. [Accessed: 02-May-2022].

[6] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, “Prediction of heart disease using a combination of machine learning and Deep Learning,” Computational Intelligence and Neuroscience, 01-Jul-2021. [Online]. Available: https://www.hindawi.com/journals/cin/2021/8387680/. [Accessed: 02-May-2022].

[7] V. Chang, V. R. Bhavani, A. Q. Xu, and M. A. Hossain, “An artificial intelligence model for heart disease detection using machine learning algorithms,” Healthcare Analytics, 31-Jan-2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772442522000016sec2. [Accessed: 02-May-2022].

[8] “Risk factors for coronary artery disease - statpearls - NCBI bookshelf.” [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK554410/. [Accessed: 03-May-2022].

[9] “9 key risk factors for heart disease everyone should know,” Northeast Georgia Health System, 11-Apr-2022. [Online]. Available: https://www.nghs.com/2020/10/13/9-key-risk-factors-for-heart-disease-everyone-should-know/. [Accessed: 02-May-2022].

[10] By: IBM Cloud Education, “What is Random Forest?,” IBM. [Online]. Available: https://www.ibm.com/cloud/learn/random-forest. [Accessed: 02-May-2022].

[11] S. Swaminathan, “Logistic regression - detailed overview,” Medium, 18-Jan-2019. [Online]. Available: https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc. [Accessed: 02-May-2022].