CARDIOVASCULAR HEART DISEASE INDICATOR

Riya Bharamaraddi, Kayla Martinez, Aiko Sherman, Manuella Shomba
CSSE/MA415 Machine Learning

13 May 2024

ABSTRACT

Cardiovascular heart disease remains the leading cause of death globally, claiming millions of lives each year. By identifying individuals at high risk before symptoms manifest, healthcare professionals can intervene with preventative measures and lifestyle modifications. In this study, we aim to predict if one will develop cardiovascular heart disease based on certain features. We also aim to understand what factors have the greatest impact on determining if one will develop the disease. A variety of models were built and their accuracies were compared. These models include Logistic Regression, KNN Classification, Decision Tree Classification, Gradient Boost Classification, PCA, and Random Forest Classification.

1.INTRODUCTION

Cardiovascular heart disease is the leading cause of death worldwide, with an individual dying every 33 seconds in the United States due to the condition. Heart disease also costs the United States about $239.9 billion each year, due to medical bills, services, and lost productivity. For this reason, a large portion of our medical resources have been aimed to reduce the presence and effect of cardiovascular heart disease.

If an understanding of what and how many different factors affect the likelihood of developing heart disease, it is possible that it could save many lives and billions of dollars. The use of machine learning models in health care has been on the rise. Accurate machine learning models in healthcare can lead to better diagnosis based on medical history, produce treatment options for patients, predict the chances that an individual will develop a specific disease, produce automatic and personalized medicines, or even help in the automation of surgeries and treatments [1].

Predicting the presence of cardiovascular health disease is challenging due to inadequate training data or lack of explainability. Machine learning models can be poor at recognizing a shift in context or data, meaning that it is possible for there to be a mismatch between training and operational data. The challenge of having a large, but balanced and represented dataset is also essential to having a successful model. Also, many machine learning models lack interpretability, meaning that it is essential for a clinician to overlook the predictions of the model [2].

To limit the possible challenges in our model, we use a large dataset that contains an equivalent amount of individuals who have heart disease and those who do not. The dataset is also one that we believe matches the distribution in the human population. This can be shown in Figure 1, which shows the BMI distribution, and Figure 2, which shows the age distribution of our dataset. The interpretability of each model was considered too, to fully grasp an understanding of how much oversight would be necessary to deploy the model into the medical field.

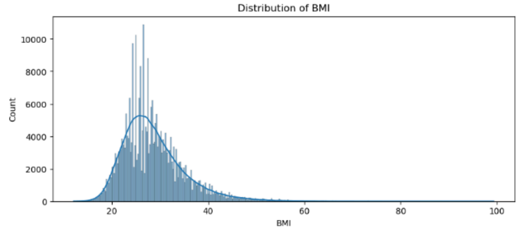

Figure 1: Exploratory plot displaying the distribution of the numerical feature ‘BMI,’ included in the dataset used.

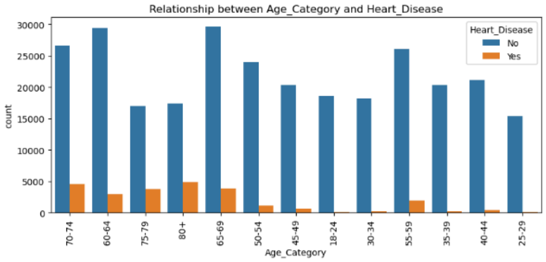

Figure 2: Exploratory plot displaying the relationship between the categorical features ‘Age_Category’ and ‘Heart_Disease,’ included in the dataset used.

In this study, we create multiple classification models to predict if an individual has heart disease or not. The inputs for our model include multiple categorical and numerical inputs that include information about the individual’s health history and lifestyle choices.

2. LITERATURE REVIEW

In order to get an understanding of what features might lead an individual to a Heart Disease, we looked at several articles and journals to give us a baseline on where to take this project. Two articles taken from the CDC [3] and [4] refer to these features as the leading risk factors contributing to heart disease; high blood pressure, diabetes, smoking, unhealthy diet, and increased age. Each of these features were represented in our dataset so it gave us a good starting point.

In a study by Padmanabhan, Yuan, Chada, and Nguyen [5], they used models such as Logistic Regression, Gradient Boost, and Decision Tree on their datasets. They found that Logistic Regression worked well and took less time to run than others. This provided us not only with what models could be best to use, but something to compare our work to if the models would also work best for us.

In another study by Yunxing Jiang, additional models were explored too, such as K-Nearest Neighbors (KNN), Random Forest, and Extreme Gradient Boost [6]. These models reported an accuracy between 0.770 and 0.872, which gives us a rough estimate of what accuracies we can expect from our models. However, they used a different dataset that is also smaller, so our results will likely vary somewhat.

Madhumita Pal conducted a study to analyze the results of a KNN and its applications to the healthcare industry [7]. The outliers and null values were removed to help increase the accuracy and performance of the model, proving that this is essential. The results were presented as a confusion matrix, which proves a reliable way to present a model’s results in a classification problem.

3.PROCESS

3.1 Data Source

The data source used was taken from Kaggle and includes features related to patients’ lifestyl, age, sex, general health, checkup frequency, exercise habits, smoking history, and the presence of various diseases. This dataset has been focused on features directly related to lifestyle factors and was originally taken from the 2021 Center for Disease Control’s Behavioral Risk Factor Surveillance System (BRFSS) which contains 304 features.

The original dataset started with 18 features and 308, 854 records. We used this data set to create multiple classification models, where we aim to predict if an individual has heart disease or not. The dataset included one target (Heart Disease) with two classes (Yes or No). However, the dataset was extremely unbalanced, with only 24,971 individuals having heart disease.

3.2 Preprocessing

The dataset came preprocessed and cleaned from the original BRFSS dataset. The original dataset from the BRFSS contained 304 unique variables but was cut down to 18 based on what was believed to contribute to one’s risk for developing heart disease by the author.

The dataset contains no missing values, which were counted using the isnull().sum() function. Numerical features were standardized using the StandardScaler function from the sklearn package and categorical features were one-hot encoded using the get_dummies function of the Pandas package.

3.3 Feature extraction or engineering

Feature extraction was an important part of creating our models, as it could help reduce run time and possibly increase the accuracy of our models. Feature extraction was already performed on the Kaggle dataset we used to only include features that are believed to have an impact on the development of heart disease, which originated from the CDC’s BRFSS. Feature extraction was not performed in our experiment.

3.4 Classifiers/regressors and tuning

First, we created an L2 (Ridge) Logistic Regression model due to its simplicity and explainability.

Logistic regression applies linear regression to classification problems. Each class gets its own linear regressor. Additionally, a non-linear Sigmoid function is added to the end, since the model does not fit the data perfectly. The Sigmoid function allows the function to map the values from 0 to 1.

In addition to the L2 logistic regression model, an L1 (Lasso) logistic regression model was created, improving the interpretability of the logistic regressor's results. Features such as Age_Category_50-54, Depression_Yes, and Smoking_History_Yes had feature weights pushed to zero, allowing the model to be interpreted much easier, in terms of feature weights. The accuracy of the model remained unchanged when compared to the L2 Ridge logistic regressor. The bootstrap method was also used to determine the top features by estimating at 95% confidence and comparing it to L1 and L2 logistic regression.

For better interpretability of the features and reduced dimensionality of the data, Principal Component Analysis (PCA) was performed on our best working model which was Logistic Regression. The test accuracies of both models were then compared to each other to determine if PCA helped improve our model's performance in terms of predictive accuracy.

A KNN Classifier model was run with the standard K=5, which looks at the closest five points and categorizes if the patient has a heart disease or not. Table 3 shows the results of the train and test accuracy which both beat the baseline of 50%.

A Decision Tree Classifier was another model that was explored. This model is a greedy algorithm, meaning that the best feature is placed at the root and then it recurses on each child by picking the next best feature. The depth of the decision tree must be tuned since overfitting is likely to occur for deep trees.

A Gradient Boost Classifier was used in addition to the other models. This model is known to reduce bias by consecutively ditting the deviations of the estimator. The number of boosting stages performed, the learning rate, and the maximum depth are all hyperparameters that were tuned with this model.

A Random Forest Classifier was created. This model limits the number of features that can be used to generate each tree to help address independence. Many parameters include the number of trees, the maximum number of features, the maximum depth, the minimum sample split, which is the minimum number of samples required to split an external node, and the minimum sample leaf, which is the minimum number of samples required to be a leaf node.

3.5 Post-processing

After running all 5 models, we observed that age and overall health were both top features. All five models agree that age and general health are top most important features that indicate the presence of heart disease. Comparing our results and common features identified to the CDC predictors from the literature review, we see that it matches with their results that increasing age and poor health are risks that lead to heart disease.

4. EXPERIMENTAL SETUP AND RESULTS

To train each other on the models, we used 80% of the data for training and the remaining 20% for testing. The data set was split using sklinear's train_test_split function and by setting the random state to zero. Shown in Table 1 is the distribution of individuals containing heart disease in the training and testing set.

Table 1: Amount of data entries in the testing and training dataset. An even amount of both classes are represented in each set.

	Heart Disease	No Heart Disease
Training	19,983	19,983
Testing	4,988	4,988

Both the training and test datasets were then balanced to obtain an even amount of "Yes" and "No" targets. This was completed by totaling the number of entries with heart disease in each data set and then selecting the equivalent number of entries without heart disease. The first entries without heart disease in the set were selected to be used.

The baseline accuracy was calculated and resulted in a baseline accuracy of 50%. Then, a total of five different models were created and compared to each other and the baseline in the experiment.

We decided to start with a Logistic Regression model due to its simplicity and explainability. We used both L1 and L2 Regulation. The accuracy of both types of regulation are equal. The top ten important features are also the same for L1 and L2 penalties. The most and least important features as a result of the L1 Logistic Regression model are shown in Figure 3.

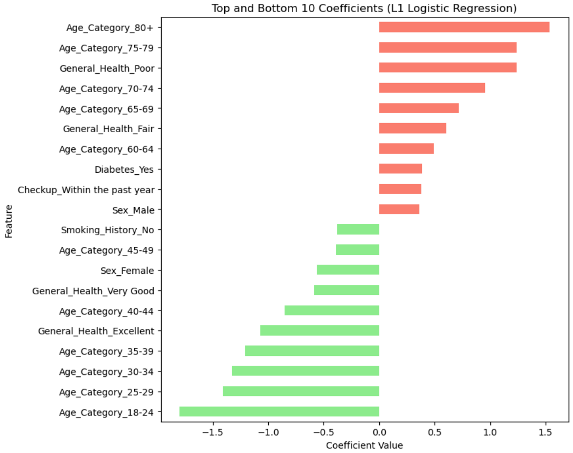

Figure 3: Plot of top ten and bottom ten feature weights for the L1 (Lasso). The resulting coefficient value is shown on the horizontal axis, while the features are represented on the vertical axis.

The highest coefficient values overall are related to the age category features. The features relating to general health are also impactful in the Logistic Regression models.

PCA was performed on L1 regularization and the obtained results are shown in Figure 4 below.

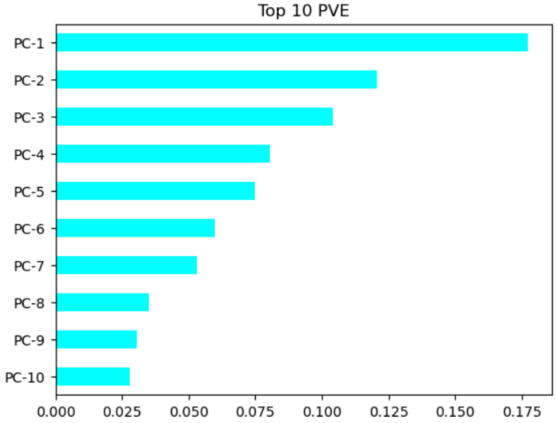

Figure 4: Top ten Proportion Variance Explained (PVE) by performing PCA on Logistic Regression. About 0.175 of the variance can be captured by PC-1.

The top ten PVE values show that 75% of the variance was captured in the top ten dimensions. The testing accuracy is 76.59% which is about the same as that for the original Logistic Regression model which was 76.59%.

We used the bootstrap method to generate a new sampling set and estimated it with a 95% confidence interval to then get the top ten features. The features matched the top 10 from the L1 Logistic Regression model as well, and can be seen in Figure 3 and Table 2.

Table 2: Results from using bootstrap at 95% confidence interval. The top ten most important features are displayed.

	0.025	**0.975**
Age Category 80+	1.693	1.693
Age Category 75-79	1.396	1.396
General Health Poor	1.181	1.181
Age Category 70-74	1.110	1.109
Age Category 65-69	0.871	0.871
Age Category 60-64	0.646	0.646
General Health Poor	0.553	0.553
Sex Male	0.418	0.418
Diabetes Yes	0.333	0.333
Checkup Within Past Year	0.324	0.324

.

The results from using bootstrapping match the results from the Logistic Regression models without bootstrapping.

A Decision Tree model was then created to see if a higher accuracy could be obtained. The maximum depth of the decision tree was tuned as a hyperparameter using grid search. We conducted a wide search, with values ranging from 1 to 100 to help ensure that we did not over or underfit the tree. It was found that a maximum depth of 9 resulted in the best results. A training accuracy of 0.741 and a testing accuracy of 0.726 were obtained from the Decision Tree model.

After, a gradient Boosting model was created, with the hope of reducing bias within the model. We started with a wide search including a learning rate that varied from 0.01 to 1, the number of boosting stages from 100 to 2,000, and the maximum depth from 1 to 40. It was found that the best parameters for Gradient Boosting are a learning rate of 0.01, 3,400 boosting stages, and a maximum depth of 2. The training accuracy was found to be 0.764 and the testing accuracy was found to be 0.767.

Next, we trained a Random Forest Classifier, where we varied multiple parameters. The number of estimators varied first from 10 to 800 trees to understand a reasonable value to begin our search. The results are shown in Figure 5.

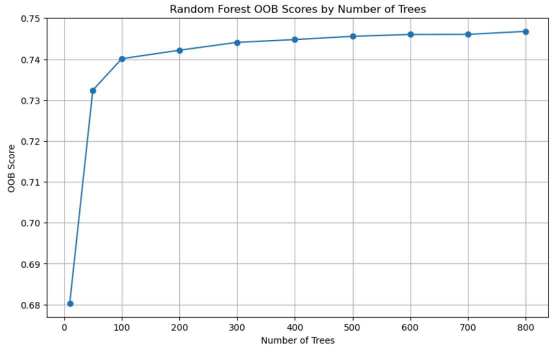

Figure 5: Compares the Out-Of-Bag (OOB) score, shown on the vertical axis, for Random Forest Estimator to the number of trees, shown on the horizontal axis. The OOB score levels off at about 300 trees.

The maximum number of features between “log2” and “sqrt”. The maximum depth ranged from 10 to 100 nodes deep. The minimum sample split varied from 2 to 10 and the minimum sample leaf varied from 1 to 4. We started with the same search technique used in the Gradient Boost model, where we first performed a wide search and incremented the parameters by a relatively large amount. This was followed by multiple interactions of refining the search to a smaller range of values. As a result, it was found that the most optimal parameters are four for the maximum number of features, 550 trees, 23 for the maximum depth, and 5 for the minimum sample split and minimum sample lead.

The results for all the models are shown in Table 3. The Gradient Boost model has the highest testing accuracy.

Table 3: Training and testing accuracies for each model explored in the experiment. The highest testing accuracy is with the Gradient Boost Model.

Model	**Training Accuracy**	**Testing Accuracy**
Logistic Regression	0.759	0.755
KNN Classifier	0.801	0.713
Decision Tree Classifier	0.776	0.729
Gradient Boost	0.768	0.762
Random Forest Classifier	0.812	0.758

5. DISCUSSION

It is important to note that some of the models created are more interpretable than others. For example, our Logistic Regression, KNN Classifier, and Decision Tree Classifier models are most interpretable, due to their simplicity or ability to be easily visualized. Other models, such as the Random Forest Classifier, are less interpretable due to the inability to trace back results. However, the Gradient Boost model provided the highest test accuracy.

Since our Logistic Regression model is one of the most interpretable and has a relatively high testing accuracy we analyzed its results more in depth. L1 regression was used as its more interpretable model since it pushes small feature weights to zero and provides the same accuracy as L2. Table 4 shows the resulting confusion matrix for the training set.

Table 4: Confusion Matrix for Logistic Regression model on training data set. The incorrect predictions are in blue.

	No	**Yes**
No	14,708	5,275
Yes	4,194	15,789

A confusion matrix was also created to show the errors with the testing dataset. The results are shown in Table 5.

Table 5: Conduction Matrix for Logistic Regression model on testing data set. The incorrect predictions are bolded.

	No	**Yes**
No	3,646	1,342
Yes	1,059	3,929

Both Tables 4 and 5 showed multiple false positives and false negatives. With the test set, we obtained a precision of 74% for the "Yes" class and 78% for the "No" class.

To understand what features led to the mispredictions shown in Table 4, we plotted each feature with its correct and incorrect classifications to see what features were leading to the most errors.

We found that the top two features that contributed to errors were age and general health. The misclassifications sorted by age are shown in Figure 6 using the testing dataset.

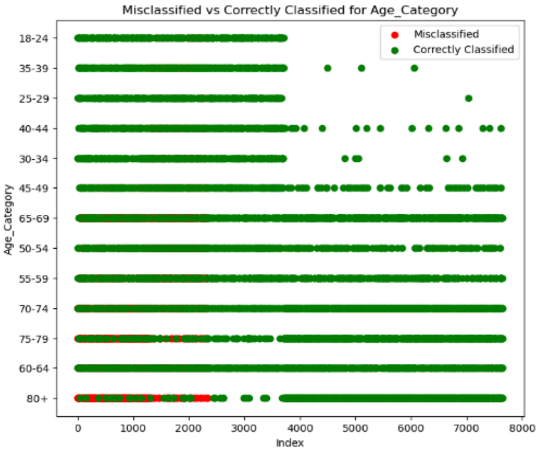

Figure 6: Results of Logistic Regressor sorted by age category. Most of the misclassifications were in the 80+ category.

The same plot was created for general health, as shown in Figure 7, which uses the testing dataset.

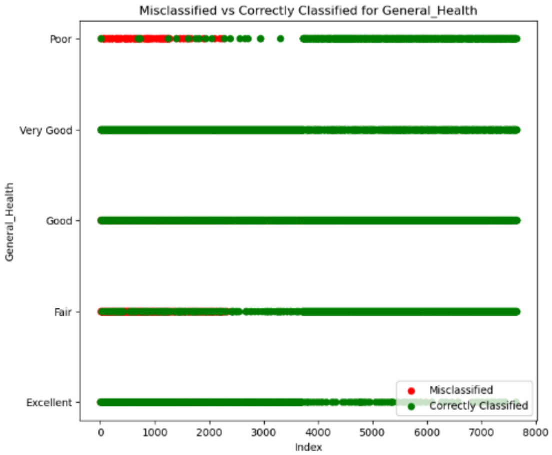

Figure 7: Results of Logistic Regression sorted by general health categories. Most of the misclassifications were in the poor health category.

The findings in Figures 6 and 7 can be backed up by looking at specific data entries, shown in Table 6.

Table 6: Two example data entries. Entry 1 leads to a correct prediction using the Logistic Regression model, while entry 2 leads to an incorrect prediction.

Feature	**Entry 1**	**Entry 2**
General Health	Very Good	Good
Check-Up	Within Past Year	Within Past Year
Exercise	Yes	Yes
Skin Cancer	No	No
Other Cancer	No	No
Depression	No	No
Diabetes	No	No
Arthritis	No	Yes
Sex	Female	Female
Age Category	18-24	80+
Height (cm)	160.00	178.00
Weight (kg)	72.57	90.72
BMI	38.34	28.70

Smoking History	No	No
Alcohol Consumption	5.00	0.00
Fruit Consumption	16.00	16.00
Green Vegetable Consumption	4.00	0.00
Fried Potato Consumption	16.00	16.00

Entry 1 is a data entry that would predict the correct class of no hearth heart disease. This individual is not in the 80+ age category or in the poor general health category. On the other hand, Entry 2 is an individual where the model incorrectly predicts that the individual does have heart disease. This individual is in the 80+ age category, which helps explain why the model predicted incorrectly.

However, not all incorrect predictions can be explained by the age or general health category. It is possible that our dataset does not include all relevant features for predicting heart disease. For example, we do not have a way to capture an individual's genetic predisposition to heart disease. Also, our data originates from a survey, meaning that the individuals could have misreported information.

When we tried to improve the accuracy of our Logistic Regression model using PCA, the test accuracy was lower than the accuracy of the model without PCA. This suggests that PCA does not help improve the model's predictive accuracy performance. A potential reason could be that the original features might have already captured most of the variance in the data, or that the relationship between the features and the target variable is not as well-suited for dimensionality reduction.

6. CONCLUSIONS AND FUTURE WORK

In this study, we compare the performance of various machine learning models, including Logistic Regression, KNN Classification, Decision Tree Classification, Gradient Boost Classification, PCA, and Random Forest Classification. Extreme Gradient Boost Classification provided the highest test dataset accuracy, while L1 logistic regression provided the most interpretable model, while still having a relatively high accuracy.

One major weakness of this study is the lack of knowledge of numerous features in this dataset. Units for numerical features such as alcohol consumption and vegetable consumption were not provided, limiting the real-world applicability of our findings. Identifying individuals at high risk and implementing preventative measures using our results may not be achievable if features used in the model cannot be interpreted.

Given the opportunity to work on this project for a longer period of time, an initial step would be to reach out to the CDC regarding their Behavioral Risk Factor Surveillance System survey and the features provided in the dataset. Not only that but to rerun our models without the features that lead to most of the misclassification and see if we can generate better results.

We also ran an analysis to identify which features were primarily responsible for misclassification, as depicted in Figures 6 and 7, coincidentally aligning with our two most critical features. Given more time, we would have refined our models by re-running them with modified feature sets after removing those causing misclassification, aiming to improve accuracy. This iterative process could have allowed us to pinpoint the optimal feature combinations for better predictive performance.

Having established that our dataset is survey-based and could potentially introduce bias into our results, future endeavors could involve selecting a dataset with more concrete and objective measures to explore potential enhancements in accuracy and robustness. By using datasets with more definitive outcomes, we aim to mitigate any inherent biases associated with survey-based data and strive for more reliable and generalizable conclusions.

7. KEY CHALLENGES AND LESSONS LEARNED

We had over 300,000 records when we got the dataset but since it was very unbalanced, we had to reduce the amount so that each yes and no had the same amount of cases. In the beginning, we began comparing our model accuracies to our initial baseline accuracy, causing some confusion with model interpretation. We learned that beginning with a somewhat balanced dataset is ideal for the interpretability of accuracy values.

Our dataset only contains features that the creator believes to impact one's risk for heart disease. However, it is possible that other features that were dropped from the original BRFSS dataset have an impact too. Also, our dataset is based on survey results, meaning that the individual's response might not actually reflect their habits. From this, we have learned that completing initial research on what factors are commonly identified as risk factors may have aided in a more comprehensive understanding of heart disease prediction. Using only lifestyle factors may have limited our models' ability to fully characterize the presence of heart disease.

Units for many features such as vegetable consumption and alcohol consumption were not identified, even after reading documentation available through the CDC. This somewhat limited our ability to fully interpret the data and limited our ability to use our models to make predictions. If we were not limited by time constraints, the next step would be to contact the CDC for more information on variables and features provided in the dataset. Since we did not look into this until late in the project, we did not have enough time to reach out to the CDC for more information. We have learned that defining features and understanding the metrics/units of our dataset should be a part of our preliminary work.

High run times required for grid search, which limited our ability to narrow in on specific hyperparameter values. Narrowing the values more would likely give us slightly higher accuracy, but these changes are likely negligible, meaning that it was not necessary for our study. We were able to perform multiple interactions of each model that used grid search, narrowing down the values with each successive iteration. By starting with a logarithmic search, we were able to complete a wide search for the first iteration of the model.

Finally, we tried implementing an Extreme Gradient Boost model, as it is known to provide a higher accuracy and be more efficient than Gradient Boosting. However, this model leads to extremely high run times, even with a narrow grid search. We even let the model run overnight but were still unable to obtain results. If we had additional time, we would have liked to have removed some of the less impactful features and attempted to run the Extreme Gradient Boosting model again.

8. REFERENCES

[1] A. Nayyar, L. Gadhavi, and N. Zaman, "Chapter 2 - Machine learning in healthcare: review, opportunities and challenges," *ScienceDirect*, Jan. 01, 2021. https://www.sciencedirect.com/science/article/pii/B9780128212295000112

[2]R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, and K. Tsaneva-Atanasova, "Artificial intelligence, bias and clinical safety," *BMJ Quality & Safety*, vol. 28, no. 3, pp. 231–237, Jan. 2019, doi: https://doi.org/10.1136/bmjqs-2018-008370.

[3] "Heart Disease and Stroke | CDC." https://www.cdc.gov/chronicdisease/resources/publications/factsheets/heart-disease-stroke.htm

[4] "Know your risk for heart disease | cdc.gov," *Centers for Disease Control and Prevention*, Mar. 21, 2023. https://www.cdc.gov/heartdisease/risk_factors.htm

[5] M. Padmanabhan, P. Yuan, G. Chada, and H. Van Nguyen, "Physician-Friendly Machine Learning: A Case Study with Cardiovascular Disease Risk Prediction," *Journal of Clinical Medicine*, vol. 8, no. 7, p. 1050, Jul. 2019, doi: 10.3390/jcm8071050.

[6]M. Pal, S. Parija, G. Panda, K. Dhama, and R. K. Mohapatra, "Risk prediction of cardiovascular disease using machine learning classifiers," *Open Medicine*, vol. 17, no. 1, pp. 1100–1113, Jan. 2022, doi: https://doi.org/10.1515/med-2022-0508.

[7] Y. Jiang *et al.*, "Cardiovascular Disease Prediction by Machine Learning Algorithms Based on Cytokines in Kazakhs of China," *Clinical Epidemiology*, vol. Volume 13, pp. 417–428, Jun. 2021, doi: https://doi.org/10.2147/clep.s313343.