# Encoding Symptoms to Predict Heart Disease

**Eva Balogun, Ross Curio, Maya Prasad, Anthony Termulo**

## ABSTRACT

Our research focuses on identifying significant risk factors for heart disease in adults aged 30-70 by analyzing heart disease datasets from 2020 and 2022. Together, we employed data preparation methods, including data cleaning, handling missing values, data type conversion, and standardization to ensure the integrity of the analysis. Three machine learning models were utilized: Logistic Regression, Random Forest, and Support Vector Machine (SVM). The logistic regression model identified smoking, difficulty walking, past strokes, diabetes, and sex as the most significant indicators of heart disease, with a high predictive accuracy of 91%. The Random Forest and SVM models also demonstrated high accuracies, with the SVM model showing exceptional performance with an accuracy of 94% in 2022. These findings offer valuable insights for public health authorities and the general public, enhancing awareness and informing targeted interventions to mitigate heart disease risk. The study concludes with potential future improvements to further refine the predictive models, aiming to optimize their accuracy and efficacy in predicting heart disease based on key indicators.

## INTRODUCTION & OBJECTIVE

Heart disease encompasses a range of conditions that affect the heart. Some common forms of heart disease include coronary artery disease, heart defects, and arrhythmias. It's a leading cause of mortality globally; Millions of lives are affected by heart disease each year, with the burden disproportionately impacting older individuals. The development of heart disease can be caused by a variety of elements, influenced by a combination of lifestyle factors and genetic predisposition. Research in this area aims to uncover the complex relationships between these elements.

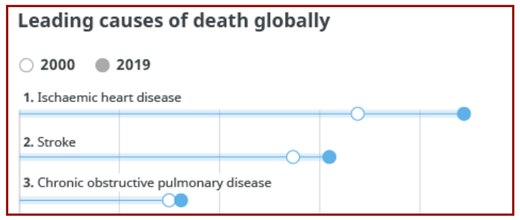



**Figure 1.** Graph from the World Health Organization displaying heart disease as the leading cause of deaths globally

For this project, we aim to analyze two datasets on heart disease from 2020 and 2022, specifically exploring the various factors that may influence its occurrence. Our primary goal of this project was to determine what the most significant risk factors are for heart disease in the adult population from ages 30-70, and how to minimize the risk of heart disease based on these factors, ultimately providing greater insight for healthcare professionals on the interplay between factors.

## DESIGN & METHODOLOGY

Our first step after acquiring data was to prepare the data in order to optimize the accuracy of our models. The data frame regarding heart disease indicators was cleaned with a wide variety of steps to ensure that the data was valid and ready for use. These common steps were taken during the EDA process:

- **Data Cleaning**: Encoding categorical data into integer values. This made it easier later on when applying predictive algorithms and machine learning techniques to the data
- **Missing Values**: Checking for missing data and decide whether to impute or remove rows with missing values. This will depend solely on the extent of missing data and its impact on the analysis. Based on our data, since the data frame was already cleaned, there were little to no missing values.
- **Data Conversion**: Ensured that all columns have appropriate data types. We utilized the python method astype() to convert the data within each column to its appropriate type, to ensure that it can be used for future analysis.
- **Standardization**: We then standardized or normalized numerical variables as necessary. This step ensures that all variables are on the same scale, which can improve the performance of our machine learning algorithms.
- **Invalid Data**: In order to further increase the accuracy of the data, we needed to check for any extraneous values. For example, we found that in the 2022 data set, 2% of respondents stated that on average received 24 hours of sleep everyday, but also managed to fit in time for daily physical activity. Values like these are invalid so we had to decide how we wanted to handle fixing the data in order for it to be the most precise for our models.

By preparing the data using these methods, we can ensure that when applying machine learning techniques later, the data is appropriately stored and structured so it can be efficiently processed.

For our research, the models that we found were suitable for our data were:

- **Logistic Regression**: Chosen for its suitability in binary outcomes and its ability to provide clear coefficients for each input feature, thereby clarifying their individual contributions to the prediction. This model will be trained with its feature importances assessed to understand the relative contribution of each feature to its predictive accuracy.
- **Random Forest**: Used for being explored for its ability to manage complex data relationships through the use of multiple decision trees, thereby reducing the risk of overfitting and accurately determining feature importance. Training this model will involve the RandomForestClassifier from the scikit-learn library, with evaluation based on accuracy, and hyper parameter tuning will focus primarily on the number trees. By reducing data variance and robust performance against overfitting, this model has the potential to achieve a high accuracy with our data.
- **Support Vector Machine**: Explored for its effectiveness in both classification and regression tasks; This SVM model will be trained using SVC class from the scikit-learn library, with attention to choosing the appropriate hyper parameters. Correct tuning of these parameters is expected to enhance the model's robustness against overfitting, thereby leading to high accuracies, especially when dealing with new data.

## REFERENCES

1) "The Top 10 Causes of Death", https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death, *World Health Organization*, 9 December 2022

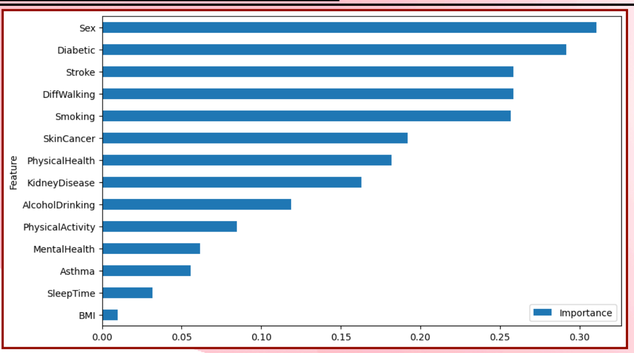
## RESULTS & EVALUATION

```
# partition the data
# can be replaced with full dataframe
X   = df_2020_subset[['BMI','Smoking','AlcoholDrinking','Stroke','PhysicalHealth','MentalHealth','DiffWalking','Sex',
                      'AgeCategory','Diabetic','PhysicalActivity','SleepTime','Asthma','KidneyDisease',
                      'SkinCancer']] #get the input features
y   = df_2020_subset['HeartDisease']                #get the target

X_train, X_test, y_train, y_test = train_test_split(X,                  #the input features
                                                    y,                  #the label
                                                    test_size=0.3,  #set aside 30% of the data as the test set
                                                    random_state=7, #reproduce the results
                                                    stratify=y      #preserve the distribution of the labels
                                                   )
```

```
# partition the data
# can be replaced with full dataframe
X   = df_2022_subset[['Sex', 'GeneralHealth','PhysicalHealthDays','MentalHealthDays','LastCheckupTime',
                      'PhysicalActivities','SleepHours','RemovedTeeth','Angina','Stroke','HadAsthma',
                      'SkinCancer','COPD','DepressiveDisorder','KidneyDisease','Arthritis','Diabetes','DeafOrHardOfHearing',
                      'BlindOrVisionDifficulty','DifficultyConcentrating','DifficultyWalking','DifficultyDressingBathing',
                      'DifficultyErrands','SmokerStatus','ECigaretteUsage','ChestScan','RaceEthnicityCategory',
                      'AgeCategory', 'AlcoholDrinkers','HIVTesting','FluVaxLast12','PneumoVaxEver','TetanusLast10Tdap','Hig
                      'CovidPos','HeightInMeters','WeightInKilograms','BMI']] #get the input features
y   = df_2022_subset['HeartAttack']                 #get the target

X_train, X_test, y_train, y_test = train_test_split(X,                  #the input features
                                                    y,                  #the label
                                                    test_size=0.3,  #set aside 30% of the data as the test set
                                                    random_state=7, #reproduce the results
                                                    stratify=y      #preserve the distribution of the labels
                                                   )
```




The code above represents the features used in our models as well as how we partitioned the data into training and test sets. For our models, we chose to separate 30% of the data for the test set. The bar graph above represents the relationship between each variable and its importance as an indicator of heart disease. This data was discovered through our logistic regression model and it can be seen that the top 5 most significant indicators are smoking, difficulty walking, past stroke, diabetic, and sex, with their coefficients ranging from 0.25-0.30. While low-risk indicators consisted of BMI, hours of sleep, asthma, mental health, and physical activity- which all had an importance of less than 0.10. Furthermore, the accuracy of this model was 91% which makes the predictions about these indicators trustworthy.

| | mean_train_score | std_train_score | mean_test_score | std_test_score |
|---|---|---|---|---|
| 2 | 0.849145 | 0.003064 | -0.018732 | 0.039526 |
| 1 | 0.846497 | 0.003399 | -0.021658 | 0.043926 |
| 0 | 0.841964 | 0.002790 | -0.029471 | 0.040732 |

| | mean_train_score | std_train_score | mean_test_score | std_test_score |
|---|---|---|---|---|
| 2 | 0.880871 | 0.002110 | 0.145429 | 0.032459 |
| 1 | 0.879816 | 0.001752 | 0.139987 | 0.031506 |
| 0 | 0.874854 | 0.002453 | 0.139356 | 0.029133 |

On the left, we have the mean_train_score values for the Random Forest model, with the top representing the scores from 2020, while the bottom is 2022. It can be seen in the tables that the mean_train_scores are more accurate for the 2022, compared to that of 2020. We also evaluated the data using a variety of n_estimators, and found that 200 yielded the highest accuracy for both years. Additionally, the mean squared error (MSE) was the same for both years, it being 0.092. Seeing that the average std_train_score is 0.03, by squaring it and comparing it to the MSE, we get 0.090 and 0.092. Due to the little difference between the variance and the MSE, this model is highly accurate.

The tables on the right represent the mean_train_score values for the Support Vector Machine. It can be seen again, that the mean_train_score values for 2022 yield higher accuracies to that of 2020. In this model for both years, the optimal C and gamma values were both 1. With these parameters, the 2020 model had an accuracy score of 90%, while the 2022 model had a score of 94%. Based on these statistics, the Support Vector Machine model appears to be the most accurate choice when it comes to predicting heart

| | mean_train_score | std_train_score | mean_test_score | std_test_score |
|---|---|---|---|---|
| 0 | 0.968643 | 0.000702 | 0.900571 | 0.000700 |
| 3 | 0.987214 | 0.000935 | 0.899000 | 0.002241 |
| 1 | 0.983929 | 0.000668 | 0.899000 | 0.001895 |
| 2 | 0.986214 | 0.000961 | 0.898571 | 0.002433 |
| 4 | 0.984964 | 0.000825 | 0.894857 | 0.001990 |

| | mean_train_score | std_train_score | mean_test_score | std_test_score |
|---|---|---|---|---|
| 0 | 1.0 | 0.0 | 0.938714 | 0.000286 |
| 1 | 1.0 | 0.0 | 0.938714 | 0.000286 |
| 20 | 1.0 | 0.0 | 0.938714 | 0.000286 |
| 21 | 1.0 | 0.0 | 0.938714 | 0.000286 |
| 22 | 1.0 | 0.0 | 0.938714 | 0.000286 |

## IMPACTS

- **High-Risk Individuals**: Adults ages 30-70, especially those with predisposing factors for heart disease can benefit from the findings from this research.
- **Public Health Authorities**: Our findings can provide a better understanding of the key risk factors and their prevalence. Health authorities can develop targeted campaigns and policies to reduce the incidence of heart disease among our population.
- **Public Awareness**: By publicizing our findings, the general public can become more aware of the risk factors associated with heart disease. This can lead to lifestyle changes at an individual level, potentially reducing the overall burden of heart disease.

## CONCLUSION

Through our research, we've completed our objective on determining the most significant risk factors for heart disease among adults aged 30-70. Based on the logistic regression model, it can be seen that there are variables that contribute more towards the risk of heart disease than others. These significant indicators consist of sex, history of diabetes, and strokes. We were also able to construct a highly accurate SVM model that can predict whether or not a patient suffers from heart disease, given a certain set of symptoms.

While our model was accurate with its predictions, there are some future improvements that we can consider. In order to optimize the accuracy of our models, we can attempt to find a better n_estimator value, as well as C and gamma values that further increase the accuracy of our current models.