Introduction

The goal of this study was to develop predictive models that can identify people who are at high risk for heart disease and lead the development of tailored interventions to lower their risk. The Analysis starts with Data Wrangling to preparing the dataset for analysis, then proceeded on to exploratory data analysis to find patterns and trends in the data. Based on the available information, three machine learning models were employed: logistic regression, decision tree, and random forest to predict the likelihood of heart disease. Overall, the research sheds light on the factors that contribute to heart disease and gives a data-driven approach to designing effective prevention interventions. Healthcare professionals can build focused interventions that aid individuals by employing machine learning algorithms to predict heart disease risk.

The Behavioral Risk Factor Surveillance System (BRFSS) is a surveillance system in the United States that collects data on numerous health-related behaviors, conditions, and practices. The Centers for Disease Control and Prevention (CDC) conducts the BRFSS survey yearly, using a standardized questionnaire to collect information from a representative sample of adults in each state.

The BRFSS questionnaire includes questions about chronic illnesses (such as heart disease and diabetes), risk behaviors (such as smoking and alcohol consumption), physical activity, diet, and healthcare access. The questionnaire has both closed-ended and open-ended questions and is distributed via phone, mail, or web-based survey.

Each variable/question in the BRFSS dataset is described briefly below:

1. HeartDisease: This binary variable reflects whether the responder has ever reported having coronary heart disease or a heart attack.
2. BMI: The body mass index of the respondent is computed as weight in kilos divided by height in meters squared.
3. Smoking: This binary variable asks if the responder has ever smoked at least 100 cigarettes in their life. If the answer is "Yes," the responder is classified as a current or former smoker.

4. AlcoholDrinking: This binary variable asks whether the respondent is a heavy drinker, defined as an adult man who consumes more than 14 alcoholic beverages per week or an adult woman who consumes more than 7 alcoholic beverages per week.
5. Stroke: This binary variable inquires whether the respondent has ever been informed by a healthcare provider that they have suffered a stroke.
6. PhysicalHealth: This variable asks respondents to report how many days in the last 30 days their physical health was poor.
7. MentalHealth: For this variable, respondents are asked to report the number of days in the previous 30 days when their mental health was poor.
8. DiffWalking: This binary variable inquires as to whether the respondent has significant difficulty walking or ascending stairs.
9. Sex: This variable specifies whether the responder is male or female.
10. AgeCategory: This variable divides the respondent's age into fourteen categories.
11. Race: The race/ethnicity of the respondent is indicated by this attribute.
12. Stroke: This binary variable inquires whether the respondent has ever been informed by a healthcare provider that they have suffered a stroke.
13. PhysicalHealth: This variable asks respondents to report how many days in the last 30 days their physical health was poor.
14. MentalHealth: For this variable, respondents are asked to report the number of days in the previous 30 days when their mental health was poor.
15. DiffWalking: This binary variable inquires as to whether the respondent has significant difficulty walking or ascending stairs.
16. Sex: This variable specifies whether the responder is male or female.
17. Diabetic: This binary variable inquires whether the respondent has ever been notified by a healthcare practitioner that they have diabetes.

18. PhysicalActivity: This binary variable asks whether the respondent engaged in physical activity or exercise other than their normal employment in the previous 30 days.
19. GenHealth: This variable asks respondents to rank their overall health as outstanding, very good, good, fair, or bad.
20. SleepTime: Respondents are asked to report the average number of hours of sleep they get in a 24-hour period for this variable.
21. Asthma: This binary variable inquires whether the responder has ever been told by a healthcare practitioner that they suffer from asthma.
22. Kidney Disease: This binary variable inquires whether the respondent has ever been advised by a healthcare provider that they have kidney disease.
23. AgeCategory: This variable divides the respondent's age into fourteen categories.
24. Race: The race/ethnicity of the respondent is indicated by this attribute.
25. SkinCancer: This binary variable asks if the respondent has ever been advised by a medical expert that they have skin cancer.

Data Wrangling and Cleaning

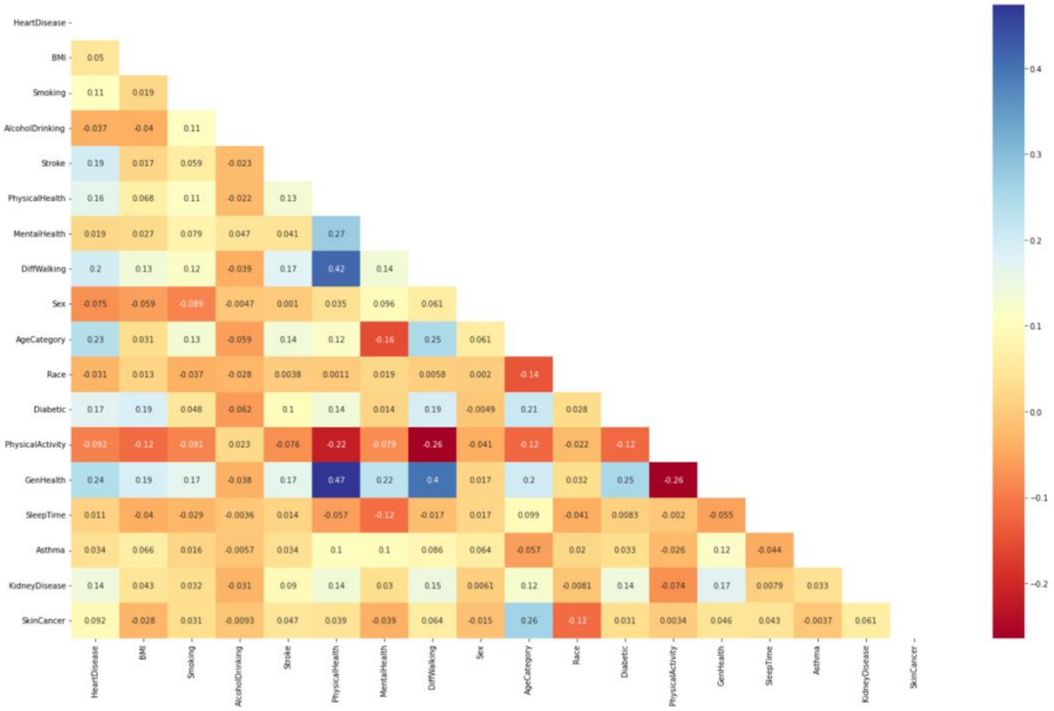
During the data cleaning process, the variable data types were appropriate for the study. The Data was cleansed by assessing the null values if any. We did, however, find and remove 18078 duplicate records, which could have biased our study. Moving on, we looked at the variable countplots and saw that our dataset was unbalanced, with much more patients without cardiac disease than with it. This imbalance, which could lead to skewed forecasts, must be rectified looking at the connections between factors, notably smoking, alcohol intake, and heart disease. According to the findings, smokers are more likely to suffer heart disease than those who drink alcohol. This discovery could be important in the prevention and treatment of heart disease. It was discovered that men were more likely than women to suffer heart disease, and the risk rose with age. These findings support prior study on the subject and may aid in the early detection and prevention of heart disease. It's observed that several BMI values were excessively high during outlier analysis, which could be attributable to data entry problems. We used the IQR technique to set up a fence outside of Q1 and Q3 to deal with these outliers.

Finally, looking at unbalanced data, in which the majority class (negative class) was much bigger than the minority class (positive class). As a result of this imbalance, approaches such as Logistic Regression, Decision trees and Random Forests that allow class weights to be set during model training are proposed. We can give the minority class more weight and make the algorithm more responsive to it this way. Overall, this data cleaning procedure was critical in ensuring that our data was error-free, bias-free, and outlier-free. It served as the foundation for our investigation, and the findings can be used to prevent and control heart disease, which is our major goal.

Exploratory Data Analysis

The summary statistics provide important insights into the distribution of variables relevant to our business problem of predicting factors influencing heart health. Body mass index (BMI) is a major risk factor for heart disease. The dataset's mean BMI is 27.83, which puts it in the overweight range according to BMI recommendations. The standard deviation of 5.39 suggests that BMI readings vary significantly among survey participants. The minimum BMI is 12.65, which is extremely low, and the maximum BMI is 43.08, which is extremely high. According to the quartile values, 50% of survey participants have a BMI between 23.83 and 31.19. Physical and emotional wellness are also significant aspects of heart health. On a scale of 0 to 30, the typical values for physical and mental health are 3.45 and 4.03, respectively. The standard deviation of these variables is relatively large, indicating that there is significant variety in the survey participants' responses. Sleep is also an important factor in heart health. The average sleep period of 7.09 hours is within the adult recommended sleep duration of 7-9 hours. The standard deviation of 1.46 suggests that sleep time varies among survey participants. The shortest sleep time is one hour, which is extremely short, and the highest sleep time is 24 hours, which is extremely long and most likely due to data entering errors. Overall, these summary statistics give useful information regarding the distribution and variability of the important variables associated with heart health. They will aid in our knowledge of the elements that lead to heart disease and in the development of predictive models. We proceed with Transforming categorical variables into classes for our business problem rather than employing one-hot encoding because it can help address the issue of imbalanced data. When dealing with imbalanced data, one-hot encoding can generate many binary variables, which can be problematic because it can leave the minority class with inadequate samples for model training, resulting in poor performance. In our scenario, we have a severely skewed dataset, with many more people without heart disease than people with heart illness. By grouping comparable categories together and lowering the number of variables included in the model, transforming the data into classes can assist address this issue. This can improve model efficiency and reduce overfitting while also guaranteeing that the minority class is not overlooked during model training.

Data Analysis and Modeling



A correlation matrix was created to understand the relationship between the variables. The variables Gen Health, Physical Health and Diff walking are significantly correlated. Though, There exists no multicollinearity among our variables. In our context, Logistic Regression, Decision tree and random forest models are well-suited for imbalanced multiclassification problems due to their ability to balance classes through various techniques such as splitting criteria, ensemble techniques, and class weighting. These models can produce more accurate and robust predictions by balancing the classes and giving more importance to the minority classes.

Logistic Regression Model

Based on several independent variables, the logistic regression model is used for predicting the likelihood of having heart disease. Our logistic regression model suggests that several independent variables play a significant role in predicting the chance of acquiring heart disease. With a coefficient of 1.01, having a heart issue increases a person's risk of getting a stroke. The overall health grade is particularly important because the risk of having heart disease rises by 1.61 times for every unit increase in grade. The probability of acquiring heart disease increases by 1.36 units for every unit higher in age category, making age another significant factor. Heart disease is 1.54 times more common in diabetics than in non-diabetics. It's interesting to note that those who declare drinking alcohol have 0.77 times lower odds of acquiring heart disease than those who don't. Finally, smoking is a statistically significant predictor of heart disease with an odds ratio of 0.3433 and a p-value of 0.000. If "Smoking" has a positive coefficient, it suggests that smoking raises our risk of heart disease. Overall, these independent variables offer important information about the risk factors for heart disease. With a pseudo R-squared of 0.261, the whole model can account for 21.61% of the variation in the outcome variable. The LLR p-value, which is less than 0.05, shows that the model fits the data reasonably well. According to the classification report and confusion matrix, the model can predict the outcome variable with an accuracy of 91%. The model has a significant false negative rate since the recall score for the positive class (having heart disease) is only 10%. This shows that the model might not be effective at detecting people with heart disease.

	P	N
T	105138	848
F	9426	996

Decision Tree Model

In comparison to the previous model, the decision tree model with 1:5 class weights for the minority class performs better in terms of precision and recall metrics. In the minority class, the precision and recall are 0.27 and 0.55, respectively, meaning that 27% of the projected positive instances are true positive cases and that the model accurately identified 55% of the true positive cases. However, with values of 0.95 and 0.86, respectively, the precision and recall for the majority class continue to be strong. The model's overall accuracy is 0.83, which indicates that 83% of the examples are correctly categorised by the model. However, accuracy by itself is not an appropriate metric to assess unbalanced datasets because the accuracy measure can be dominated by the dominant class. Therefore, to get a more thorough view of the model's performance, we should also take precision, recall, and F1-score into account for both classes. In conclusion, the decision tree model that incorporates class-specific weights has enhanced the model's performance for the minority class. The minority class's F1-score is 0.37, showing that there are still false positives and false negatives in the forecasts, therefore there is still space for improvement. To

	P	N
T	90827	15159
F	4710	5712

further enhance the performance of the model, it may be beneficial to investigate different methods like Random Forest. The top predictor variables are AgeCategory, GenHealth, Stroke, Diabetic, Smoking, and SleepTime, in descending order of feature relevance as determined by the random forest model. The most significant predictor variable, AgeCategory, indicates that older age groups are more likely to get a stroke. The second most significant predictor variable, GenHealth, measures general health status and shows that people with poor general health are more likely to experience a stroke. Stroke is the fourth most significant predictor variable, which makes logical given that a prior stroke would raise the risk of a subsequent one. The fifth and sixth most significant variables, respectively, are the presence of diabetes and smoking, both of which are recognized risk factors for stroke. Shorter sleep duration may increase the likelihood of having a stroke, according to SleepTime, the sixth most significant predictive variable.

Random Forest Model

In comparison to the prior Decision Tree Model, the Random Forest Model performed better. In comparison to the Decision Tree Model, the Accuracy Score of the Random Forest Model with the Adjusted Class Weight is 87%, which is higher. The Random Forest Model now has higher precision and recall scores for both classes. Class 1's minority class has a precision score of 0.33, which is marginally higher than the Decision Tree Model's precision score of 0.27. The recall score for the minority class is 0.42, which is higher than the Decision Tree Model's recall score of 0.55. The top predictor variables are AgeCategory, GenHealth, DiffWalking, Stroke, Diabetic, PhysicalHealth, and Sex, according to the feature importances of the Random Forest Model. AgeCategory has the greatest relevance score of 0.34, making it the most crucial predictor variable for identifying an individual's health status. The importance scores for GenHealth and DiffWalking are likewise very high (0.24 and 0.12, respectively). This suggests that a person's general health and their ability to walk comfortably are crucial indicators of their current state of health. Stroke and diabetes each have significance scores of 0.09 and 0.08, indicating that they both increase the likelihood of having a health condition in a person. PhysicalHealth and Sex both

	P	N
T	97083	8903
F	6062	4360

have significance values of 0.04 and 0.03, respectively, showing that they are important factors in figuring out a person's level of health.

Overall, the Random Forest Model with the adjusted class weight and feature importances offers a better comprehension of the crucial elements that affect a person's state of health. Based on their demographics and health-related data, it can be used as a prediction model to identify people who are more likely to develop a health issue. With the use of this knowledge, programs and treatments may be created that are specifically aimed at preventing the onset of chronic illnesses and enhancing general health and wellbeing.

Model	Accuracy	F1 Score(0)	F1 Score(1)
Logistic Regression	91%	95%	16%
Decision Tree	83%	90%	37%
Random Forest	87%	93%	37%

Conclusion

The decision tree and random forest models outperformed the logistic regression model in terms of their capacity to balance classes and predict the minority class, according to our analysis of the data from the three models (logistic regression, decision tree, and random forest). Additionally, we discovered that across all three models, AgeCategory, GenHealth, DiffWalking, Stroke, and Diabetic consistently ranked as the best predictors.

Recommendations

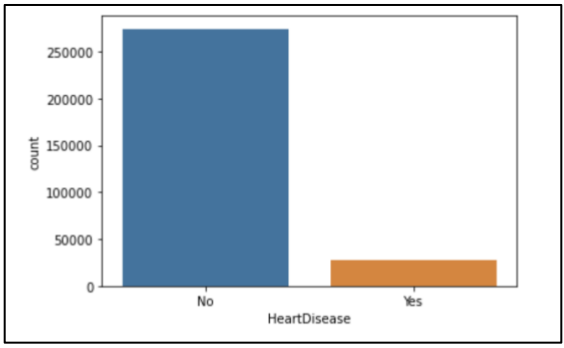
Based on these findings, we advise healthcare professionals to concentrate on these key factors when determining a patient's stroke risk. Healthcare professionals should pay special attention to the age, general health state, mobility, stroke history, and diabetes, and Smoking habits status of their patients. To help identify patients who are at higher risk for stroke, these characteristics should be included in their risk assessment methods. Additionally, we advise healthcare professionals to think about utilizing decision tree or random forest models to help with risk assessment. These models have proven to be quite effective at predicting the risk of stroke, particularly when the classes are unbalanced. Healthcare professionals can make more educated judgments on stroke prevention and treatment by incorporating these models into their clinical decision-making processes.

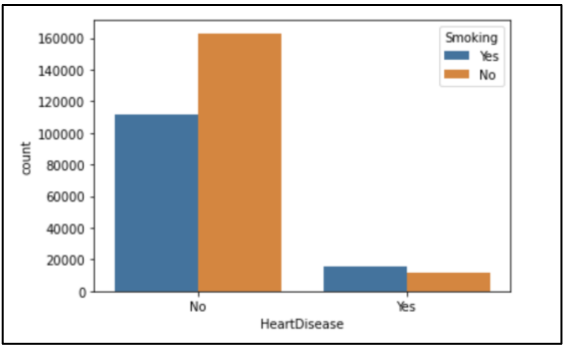
References

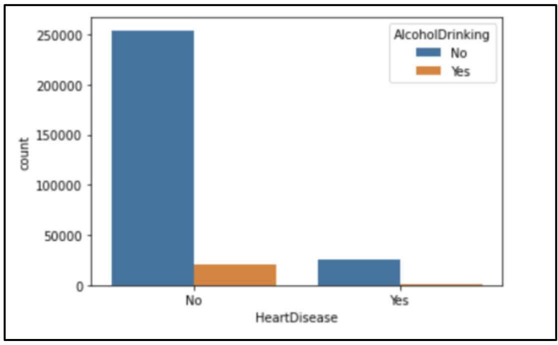
1. Chawla, N. V. (2009). Data mining for imbalanced datasets: An overview. In O. Maimon & L. Rokach (Eds.), Data mining and knowledge discovery handbook (pp. 853-867). Springer. https://doi.org/10.1007/978-0-387-09823-4_45
2. Dr.D.Ramyachitra and P.Manikandan. (2014). Imbalanced Dataset Classification and Solutions: A Review. International Journal of Computing and Business Research (IJCBR), 5(4), 1-10. ISSN (Online): 2229-6166.

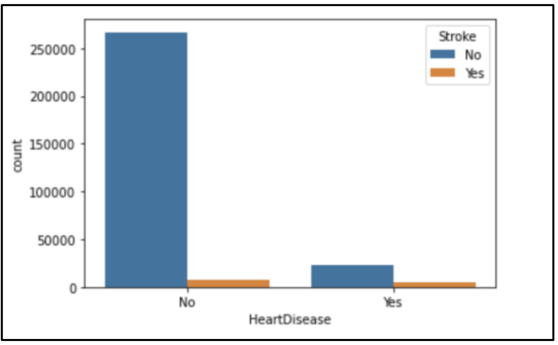
Appendix

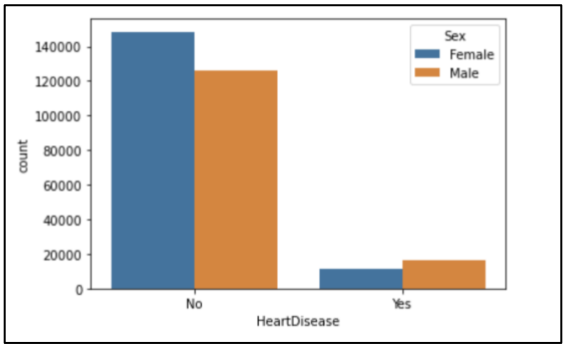
1. Countplots for the predictor variable

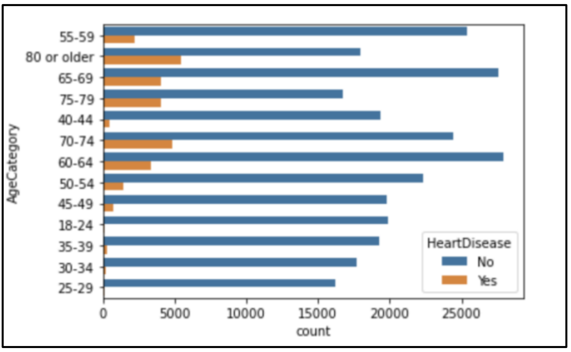


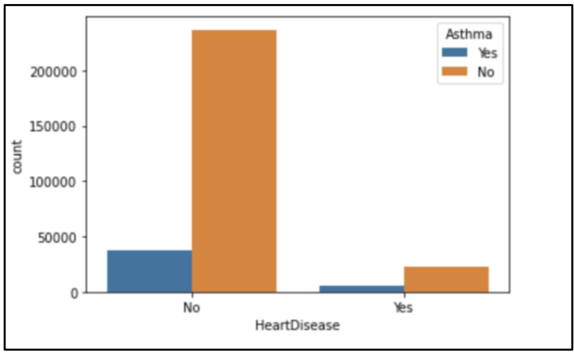


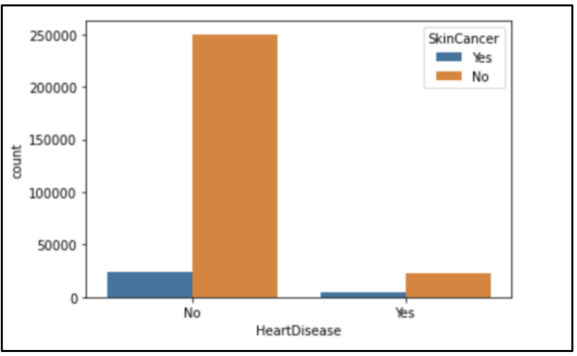


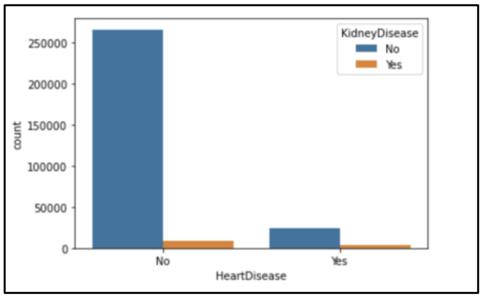




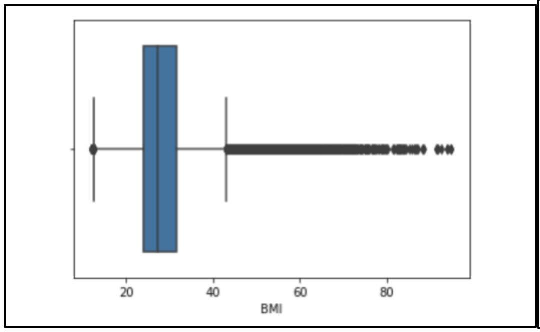


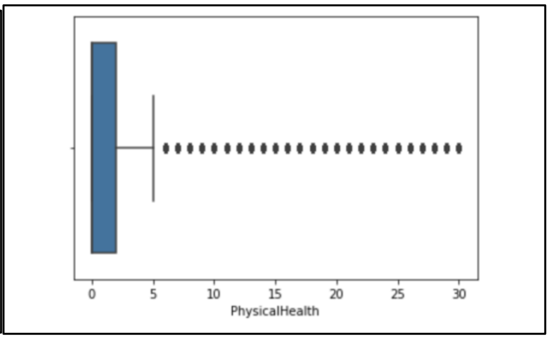


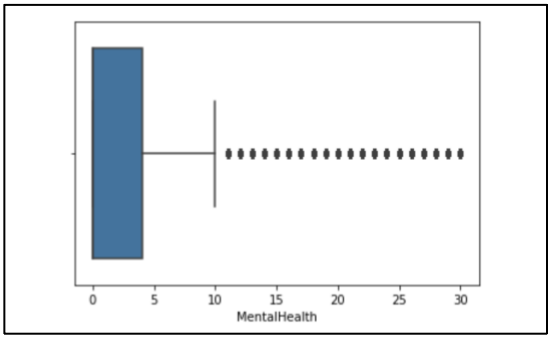


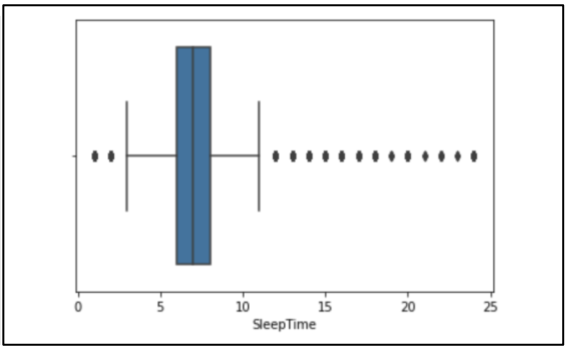


2. Outlier Analysis





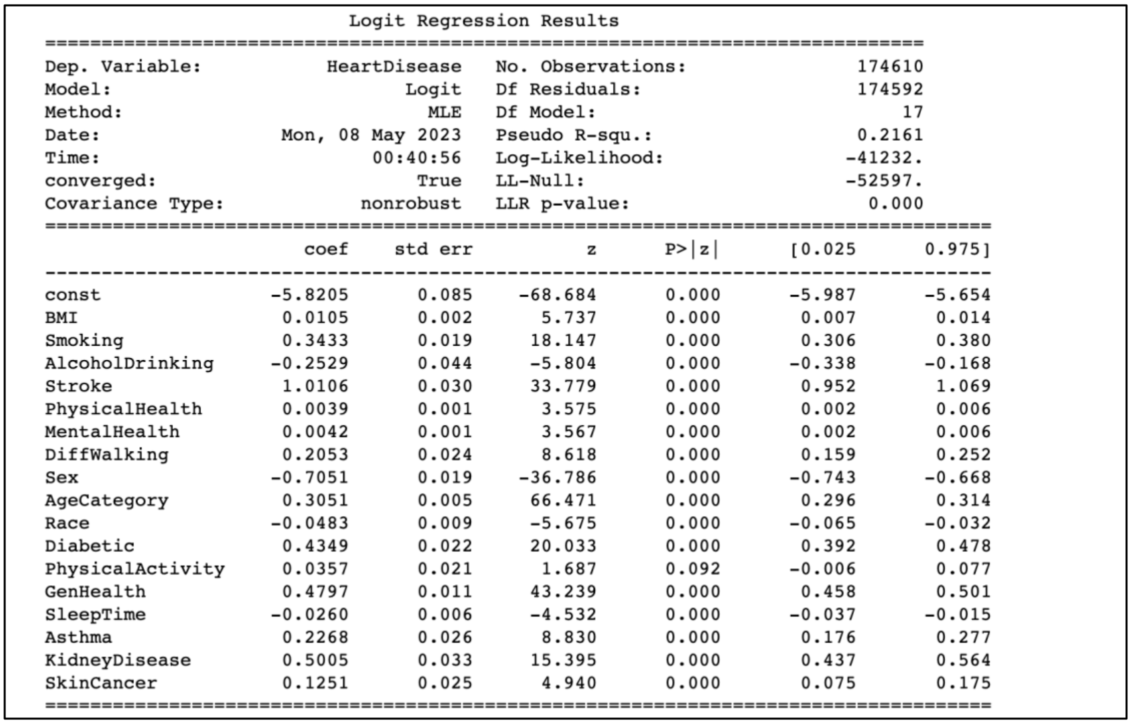




3. Summary Statistics

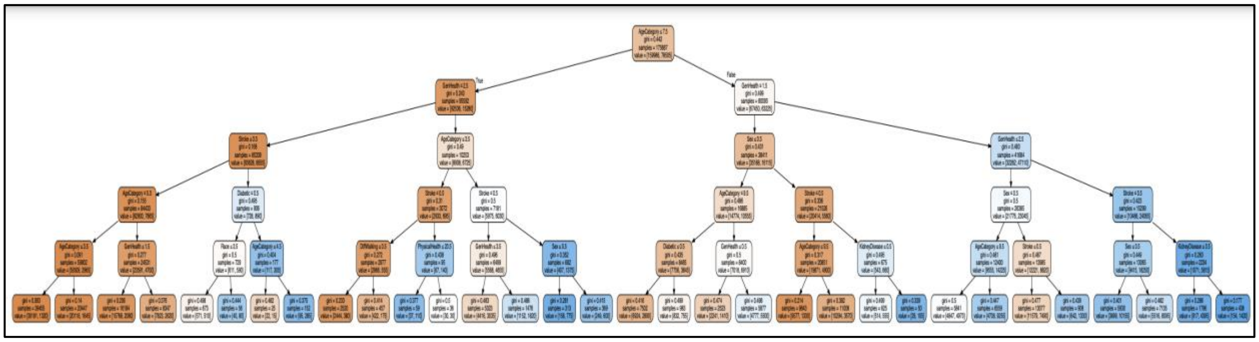
	BMI	PhysicalHealth	MentalHealth	SleepTime
Count	292812.00	292812.00	292812.00	292812.00
Mean	27.83	3.45	4.03	7.09
Std. Dev.	5.39	8.00	8.04	1.46
Min.	12.65	0.00	0.00	1.00
25%	23.83	0.00	0.00	6.00
50%	27.26	0.00	0.00	7.00
75%	31.19	2.00	4.00	8.00
Max.	43.08	30.00	30.00	24.00

4. Logistic Regression Model

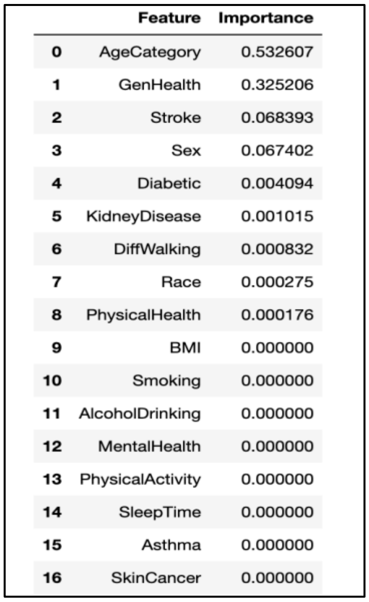


```
                           Logit Regression Results
==============================================================================
Dep. Variable:           HeartDisease   No. Observations:               174610
Model:                          Logit   Df Residuals:                   174592
Method:                           MLE   Df Model:                           17
Date:                Mon, 08 May 2023   Pseudo R-squ.:                  0.2161
Time:                        00:40:56   Log-Likelihood:                -41232.
converged:                       True   LL-Null:                       -52597.
Covariance Type:            nonrobust   LLR p-value:                     0.000
========================================================================================
                      coef    std err          z      P>|z|      [0.025      0.975]
----------------------------------------------------------------------------------------
const              -5.8205      0.085    -68.684      0.000      -5.987      -5.654
BMI                 0.0105      0.002      5.737      0.000       0.007       0.014
Smoking             0.3433      0.019     18.147      0.000       0.306       0.380
AlcoholDrinking    -0.2529      0.044     -5.804      0.000      -0.338      -0.168
Stroke              1.0106      0.030     33.779      0.000       0.952       1.069
PhysicalHealth      0.0039      0.001      3.575      0.000       0.002       0.006
MentalHealth        0.0042      0.001      3.567      0.000       0.002       0.006
DiffWalking         0.2053      0.024      8.618      0.000       0.159       0.252
Sex                -0.7051      0.019    -36.786      0.000      -0.743      -0.668
AgeCategory         0.3051      0.005     66.471      0.000       0.296       0.314
Race               -0.0483      0.009     -5.675      0.000      -0.065      -0.032
Diabetic            0.4349      0.022     20.033      0.000       0.392       0.478
PhysicalActivity    0.0357      0.021      1.687      0.092      -0.006       0.077
GenHealth           0.4797      0.011     43.239      0.000       0.458       0.501
SleepTime          -0.0260      0.006     -4.532      0.000      -0.037      -0.015
Asthma              0.2268      0.026      8.830      0.000       0.176       0.277
KidneyDisease       0.5005      0.033     15.395      0.000       0.437       0.564
SkinCancer          0.1251      0.025      4.940      0.000       0.075       0.175
========================================================================================
```

5. Decision Tree Model

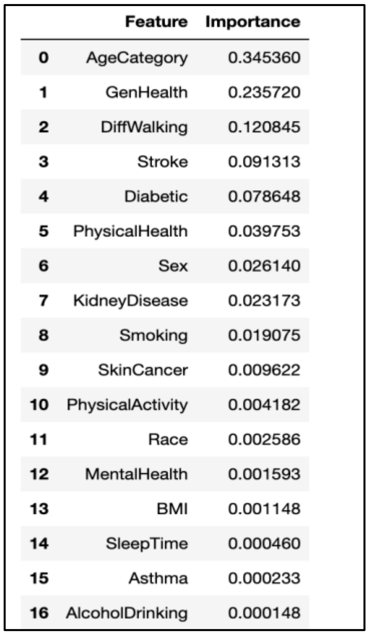


6. Feature Importances: Decision Tree model



	Feature	Importance
0	AgeCategory	0.532607
1	GenHealth	0.325206
2	Stroke	0.068393
3	Sex	0.067402
4	Diabetic	0.004094
5	KidneyDisease	0.001015
6	DiffWalking	0.000832
7	Race	0.000275
8	PhysicalHealth	0.000176
9	BMI	0.000000
10	Smoking	0.000000
11	AlcoholDrinking	0.000000
12	MentalHealth	0.000000
13	PhysicalActivity	0.000000
14	SleepTime	0.000000
15	Asthma	0.000000
16	SkinCancer	0.000000

7. Feature Importances: Random Forest Model



	Feature	Importance
0	AgeCategory	0.345360
1	GenHealth	0.235720
2	DiffWalking	0.120845
3	Stroke	0.091313
4	Diabetic	0.078648
5	PhysicalHealth	0.039753
6	Sex	0.026140
7	KidneyDisease	0.023173
8	Smoking	0.019075
9	SkinCancer	0.009622
10	PhysicalActivity	0.004182
11	Race	0.002586
12	MentalHealth	0.001593
13	BMI	0.001148
14	SleepTime	0.000460
15	Asthma	0.000233
16	AlcoholDrinking	0.000148