DSC680 – Applied Data Science

Week Ten – Project Three White Paper

Author: Alysen Casaccio

Heart Disease Prediction Using 2022 CDC Survey Data

Business Problem

Heart disease remains a leading cause of death in the United States, making early prediction critical for improving patient outcomes. Timely interventions can significantly mitigate the impact of heart disease. This project aims to develop a machine learning model to predict the likelihood of heart disease based on key health indicators and lifestyle factors, leveraging data from the 2022 CDC survey.

Background

Heart disease has consistently been a significant public health issue, contributing to high mortality rates globally. Traditional diagnostic methods often rely on historical clinical data and patient-reported symptoms, which can be subjective and delayed. Using machine learning as a predictive tool offers the potential to analyze large datasets, identifying patterns and risk factors that may not be immediately evident, thus providing earlier and more accurate predictions. This project seeks to develop a model to improve heart disease prediction and early detection.

Data Explanation

The project utilizes the 2022 CDC survey data, which includes responses from 246,022 individuals. The dataset comprises variables such as state of residence, gender, general health, last checkup, physical activity, mental and physical health ratings, sleep hours, history of heart attacks, and other health-related factors.

Initial Survey Columns:

```
State                        object
Sex                          object
GeneralHealth                object
PhysicalHealthDays            int64
MentalHealthDays              int64
LastCheckupTime              object
PhysicalActivities           object
SleepHours                    int64
RemovedTeeth                 object
HadHeartAttack               object
HadAngina                    object
HadStroke                    object
HadAsthma                    object
HadSkinCancer                object
HadCOPD                      object
HadDepressiveDisorder        object
HadKidneyDisease             object
HadArthritis                 object
HadDiabetes                  object
DeafOrHardOfHearing          object
BlindOrVisionDifficulty      object
DifficultyConcentrating      object
DifficultyWalking            object
DifficultyDressingBathing    object
DifficultyErrands            object
SmokerStatus                 object
ECigaretteUsage              object
ChestScan                    object
RaceEthnicityCategory        object
AgeCategory                  object
HeightInMeters              float64
WeightInKilograms           float64
BMI                         float64
AlcoholDrinkers              object
HIVTesting                   object
FluVaxLast12                 object
PneumoVaxEver                object
TetanusLast10Tdap            object
HighRiskLastYear             object
CovidPos                     object
dtype: object
```

Data Prep

The dataset underwent several cleaning steps to ensure data accuracy and consistency. Missing values were checked and managed, duplicates were eliminated, and inconsistent data formats were standardized. For example, categorical variables like 'State,' 'Sex,' and 'GeneralHealth' were converted to numeric values to facilitate modeling. This conversion enabled machine learning algorithms to process the data effectively.

Data Dictionary

Some variables included:

HadHeartAttack: Target variable indicating if the respondent had a heart attack (Y/N).

AgeCategory: Age brackets of respondents.

GeneralHealth: Self-reported general health status.

PhysicalActivities: Engagement in physical activities.

SmokerStatus: Smoking habits and usage of e-cigarettes.

RemovedTeeth: Number of teeth removed due to health reasons.

	State	Sex	GeneralHealth	PhysicalHealthDays	MentalHealthDays	LastCheckupTime	PhysicalActivities	SleepHours	RemovedTeeth
0	Alabama	Female	Very good	4	0	Within past year (anytime less than 12 months ...	Yes	9	None of them
1	Alabama	Male	Very good	0	0	Within past year (anytime less than 12 months ...	Yes	6	None of them
2	Alabama	Male	Very good	0	0	Within past year (anytime less than 12 months ...	No	8	6 or more, but not all
3	Alabama	Female	Fair	5	0	Within past year (anytime less than 12 months ...	Yes	9	None of them
4	Alabama	Female	Good	3	15	Within past year (anytime less than 12 months ...	Yes	5	1 to 5

5 rows × 40 columns

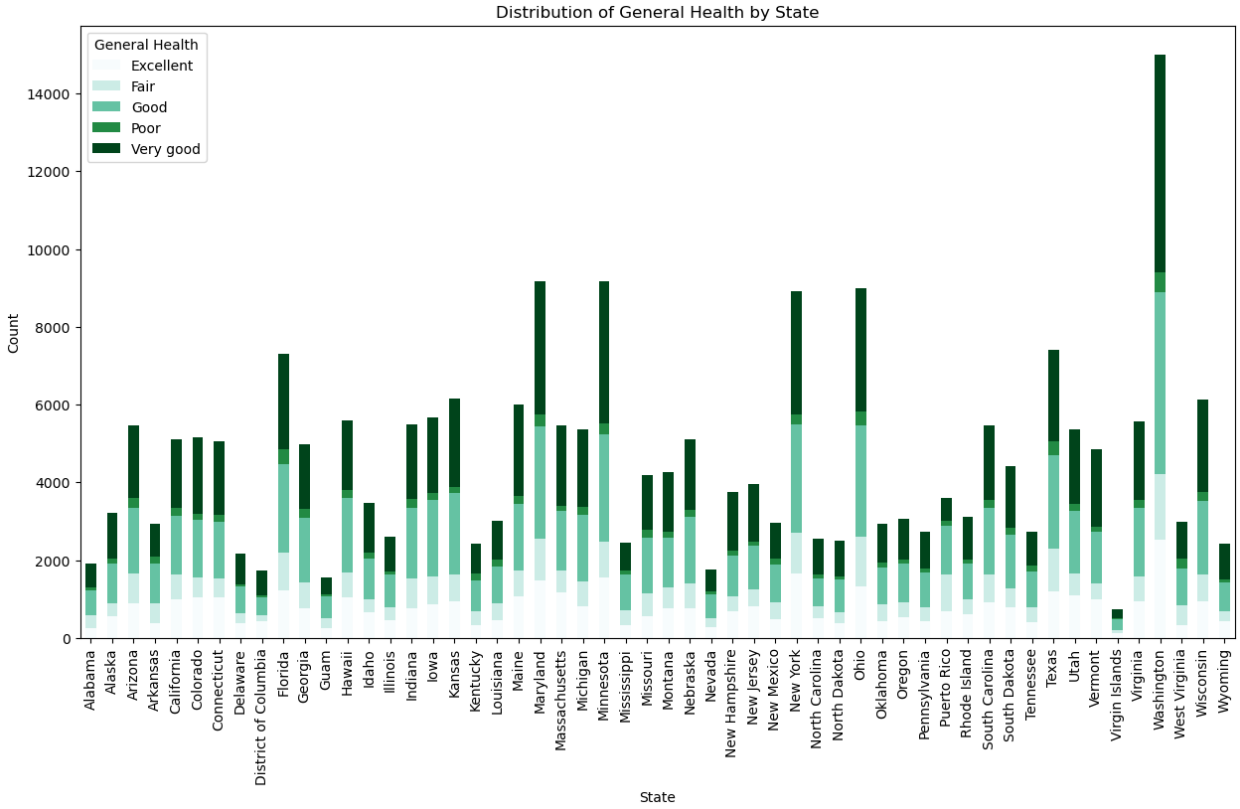
Methods

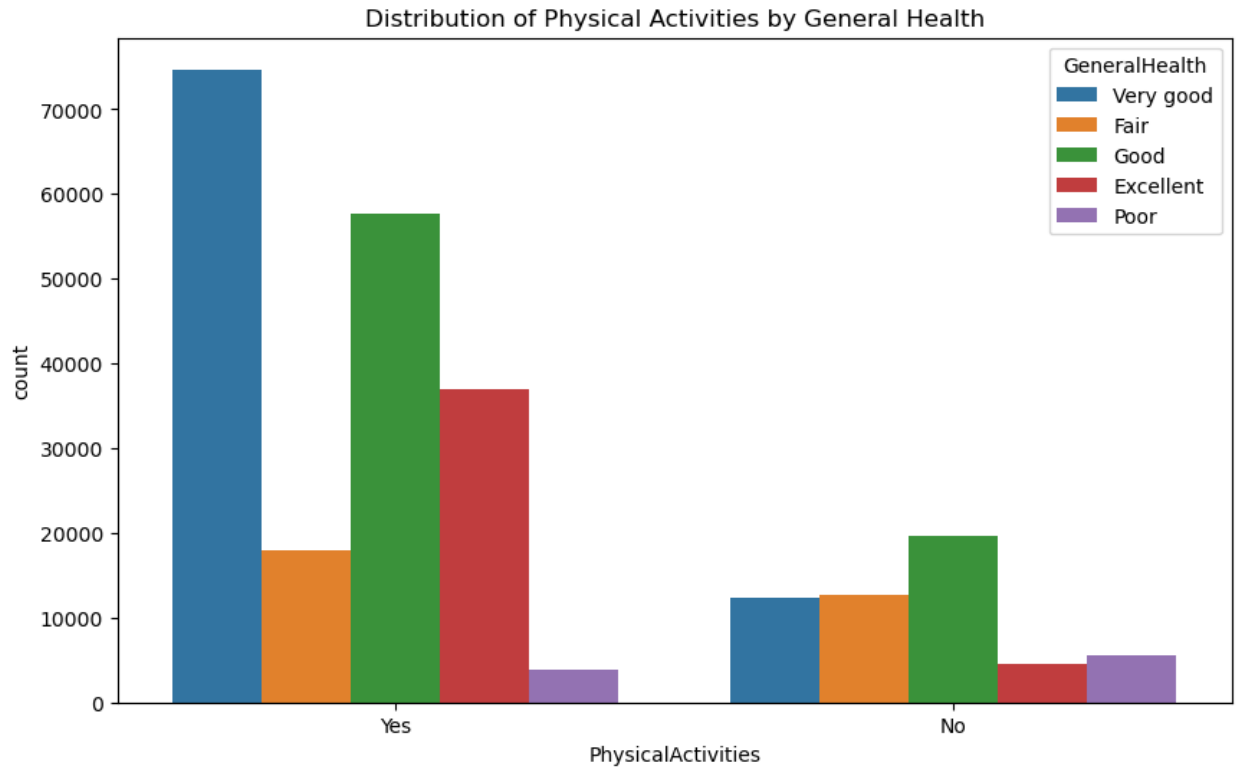
To develop a robust predictive model, the following methods were employed:

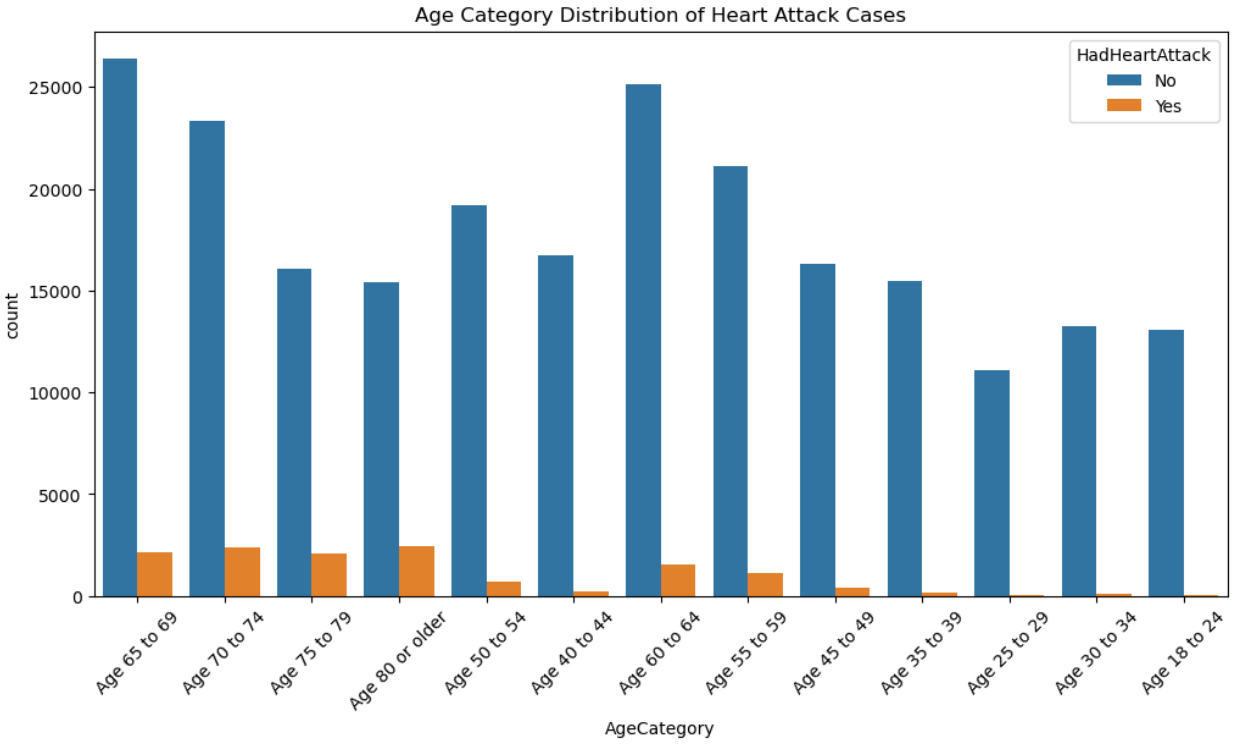
1. **Data Cleaning and Preparation:** The dataset was meticulously cleaned to ensure accuracy. Missing values were handled, duplicates were removed, and categorical variables were encoded. The data was then split into training and testing sets for model validation.
2. **Exploratory Data Analysis (EDA):** Various visualizations were created to understand the distributions and relationships of key features. Correlation analyses helped identify potential predictors of heart disease.
3. **Handling Class Imbalance:** The target variable 'HadHeartAttack' was highly imbalanced, with a minority of respondents having experienced a heart attack. To address this, SMOTE (Synthetic Minority Over-sampling Technique) was applied to balance the classes by oversampling the minority class.
4. **Modeling:** Multiple machine-learning models, including Logistic Regression, Random Forest, Gradient Boosting, XGBoost, and a Voting Classifier, were trained and evaluated. To optimize model performance, hyperparameter tuning was performed using GridSearchCV.
5. **Model Interpretation:** SHAP (SHapley Additive exPlanations) was used to interpret the model predictions and identify the most important features contributing to heart disease prediction.

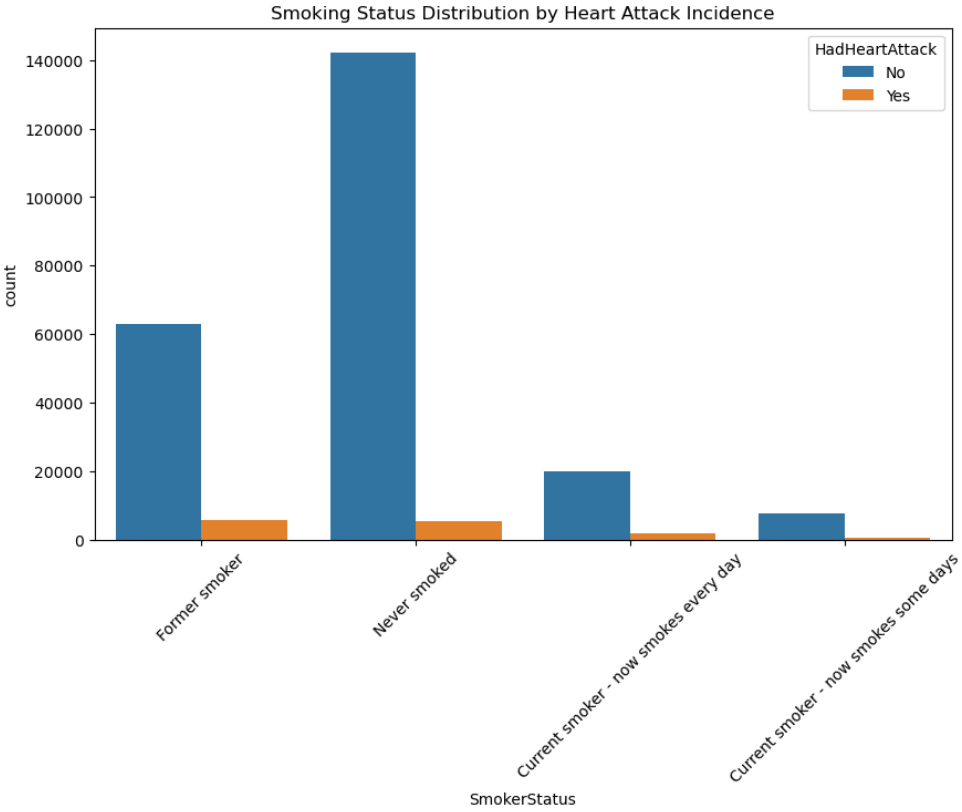
Exploratory Data Visualizations

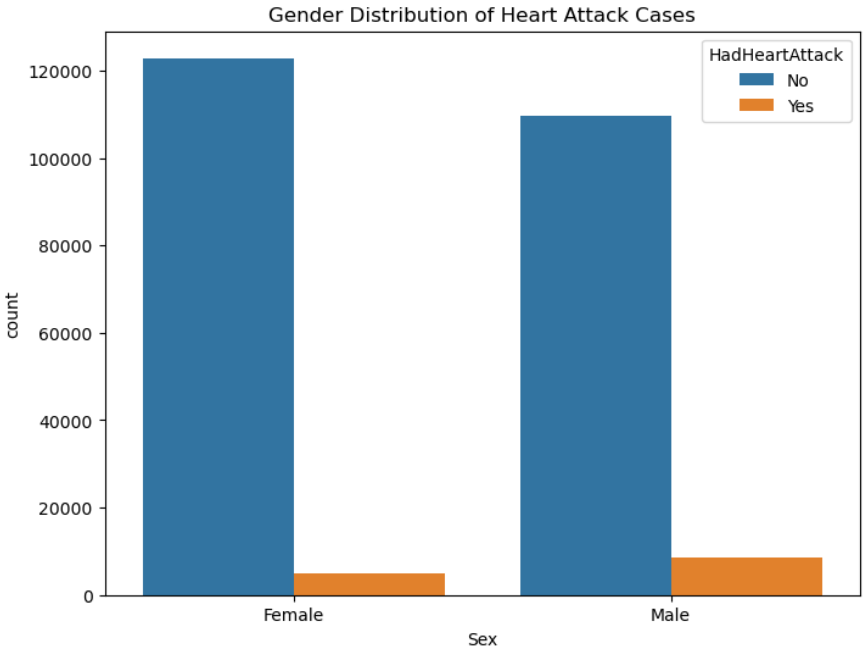
Visualizations summarized the distribution of heart attack cases by State of residence, Gender, Physical Health, and Smoking Status. This and the statistical analyses of numeric categories provided an initial understanding of the data and highlighted key trends and patterns.


Distribution of General Health by State
General Health
Excellent
Fair
Good
Poor
Very good
Count
0
2000
4000
6000
8000
10000
12000
14000
Alabama
Alaska
Arizona
Arkansas
California
Colorado
Connecticut
Delaware
District of Columbia
Florida
Georgia
Guam
Hawaii
Idaho
Illinois
Indiana
Iowa
Kansas
Kentucky
Louisiana
Maine
Maryland
Massachusetts
Michigan
Minnesota
Mississippi
Missouri
Montana
Nebraska
Nevada
New Hampshire
New Jersey
New Mexico
New York
North Carolina
North Dakota
Ohio
Oklahoma
Oregon
Pennsylvania
Puerto Rico
Rhode Island
South Carolina
South Dakota
Tennessee
Texas
Utah
Vermont
Virgin Islands
Virginia
Washington
West Virginia
Wisconsin
Wyoming
State


Distribution of Physical Activities by General Health
GeneralHealth
Very good
Fair
Good
Excellent
Poor
count
0
10000
20000
30000
40000
50000
60000
70000
Yes
No
PhysicalActivities


Age Category Distribution of Heart Attack Cases
HadHeartAttack
No
Yes
count
25000
20000
15000
10000
5000
0
Age 65 to 69
Age 70 to 74
Age 75 to 79
Age 80 or older
Age 50 to 54
Age 40 to 44
Age 60 to 64
Age 55 to 59
Age 45 to 49
Age 35 to 39
Age 25 to 29
Age 30 to 34
Age 18 to 24
AgeCategory


Smoking Status Distribution by Heart Attack Incidence
HadHeartAttack
No
Yes
count
140000
120000
100000
80000
60000
40000
20000
0
Former smoker
Never smoked
Current smoker - now smokes every day
Current smoker - now smokes some days
SmokerStatus



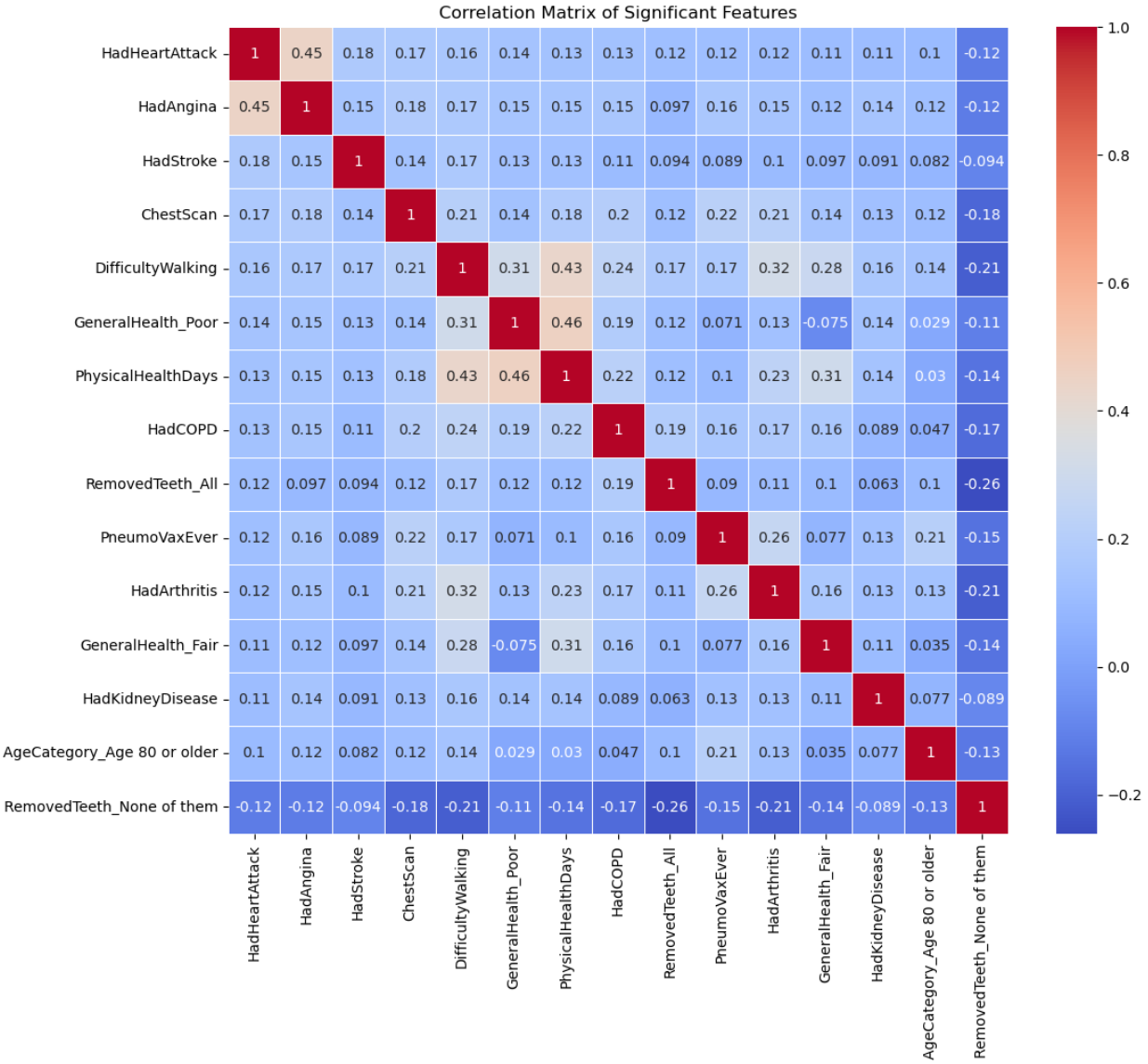
	PhysicalHealthDays	MentalHealthDays	SleepHours	HeightInMeters	\
count	246022.000000	246022.000000	246022.000000	246022.000000	
mean	4.119026	4.167140	7.021331	1.705150	
std	8.405844	8.102687	1.440681	0.106654	
min	0.000000	0.000000	1.000000	0.910000	
25%	0.000000	0.000000	6.000000	1.630000	
50%	0.000000	0.000000	7.000000	1.700000	
75%	3.000000	4.000000	8.000000	1.780000	
max	30.000000	30.000000	24.000000	2.410000	

	WeightInKilograms	BMI
count	246022.000000	246022.000000
mean	83.615179	28.668136
std	21.323156	6.513973
min	28.120000	12.020000
25%	68.040000	24.270000
50%	81.650000	27.460000
75%	95.250000	31.890000
max	292.570000	97.650000

Visualization

Various visualizations were created using Python to illustrate the data in the initial phases and as a part of the delivered prediction algorithm testing. The entire project included visualizations such as bar charts, heat maps, time series plots, scatterplots, histograms, and other visual analyses.

Early insights into the dataset included the correlation coefficients and a heat map was generated:



Class Balancing

Class imbalance is a common challenge that can significantly impact model performance, particularly when dealing with health-related data. In my heart disease prediction project, the target variable, HadHeartAttack, was highly imbalanced. Most respondents had not experienced a heart attack, while only a tiny fraction had, leading to a skewed dataset. This imbalance posed a risk of the models becoming biased towards the majority class, thereby failing to identify cases of heart disease accurately.

To address this issue, I employed the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is a powerful method for generating synthetic samples for the minority class by interpolating between existing ones. This approach helped to balance the dataset without simply duplicating existing samples, which could lead to overfitting.

Initially, I observed the distribution of the target variable. The dataset revealed a significant imbalance, with approximately 95% of respondents not having had a heart attack and only about 5% having experienced one.

I used the SMOTE function from the imblearn library to generate synthetic samples for the minority class. This involved creating new heart attack cases by interpolating between the nearest neighbors of existing cases.

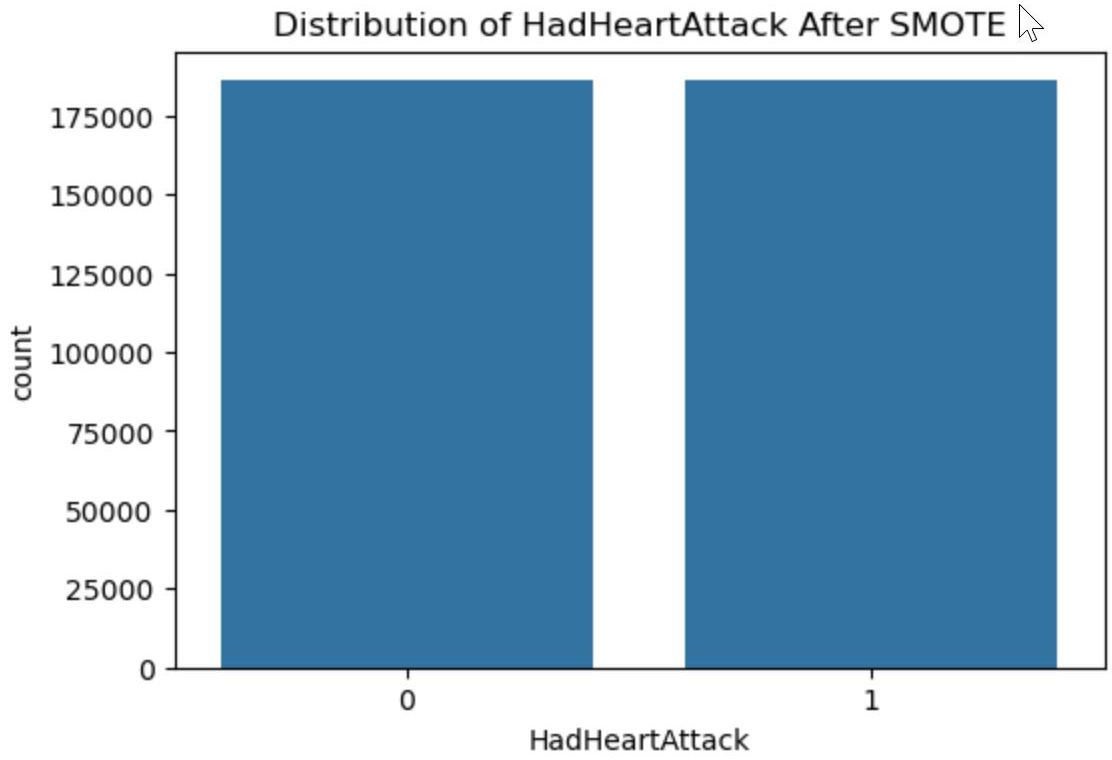
The SMOTE technique was applied to the training set, ensuring the class distribution was balanced before training models. This process involved:

- **Splitting the Data:** Dividing the dataset into training and testing sets, maintaining the original imbalance in the test set to evaluate model performance realistically.
- **Generating Synthetic Samples:** Applying SMOTE to the training set increased the number of heart attack cases to match the number of non-heart attack cases.

The application of SMOTE resulted in a balanced training dataset, which significantly improved the models' ability to learn and identify patterns associated with heart disease. By addressing the class imbalance, I enhanced the models' sensitivity to the minority class, leading to more reliable predictions.

Here is the visualization of the distribution of heart attacks after the application of SMOTE:

```
HadHeartAttack
0    0.5
1    0.5
Name: proportion, dtype: float64
```



Modeling Iterations

In this project, I employed multiple iterations of modeling to enhance the accuracy and reliability of my predictions. This narrative describes the path of iterative changes and the rationale behind each step and provides the confusion matrix and classification report for each attempt.

Logistic Regression

I began with a basic logistic regression model as the initial baseline. Logistic regression is an easy-to-interpret classification algorithm that estimates the probability of a binary outcome, in this case, whether a respondent has experienced a heart attack. However, the initial model struggled with the significant class imbalance in the dataset, resulting in high accuracy for the majority class but poor performance in identifying heart attack cases. This indicated that while the model could predict non-heart attack cases well, it did not effectively identify the minority class.

```
[[38717  7856]
 [ 1015  1617]]
              precision    recall  f1-score   support

           0       0.97      0.83      0.90     46573
           1       0.17      0.61      0.27      2632

    accuracy                           0.82     49205
   macro avg       0.57      0.72      0.58     49205
weighted avg       0.93      0.82      0.86     49205
```

Logistic Regression with Scaled Data

Recognizing the need for feature scaling in logistic regression, I applied standard scaling to the features to ensure they were on a comparable scale. This step improved the model's converging ability and slightly enhanced its performance.

```
[[38679  7894]
 [ 1071  1561]]
              precision    recall  f1-score   support

           0       0.97      0.83      0.90     46573
           1       0.17      0.59      0.26      2632

    accuracy                           0.82     49205
   macro avg       0.57      0.71      0.58     49205
weighted avg       0.93      0.82      0.86     49205
```

Random Forest with Class Weights

To better handle the class imbalance, I explored the Random Forest algorithm, a method known for its robustness and ability to handle class imbalance through class weights. By assigning higher weights to the minority class, I ensured that the model paid more attention to heart attack cases. This approach significantly improved the recall for the minority class, indicating a better ability to identify actual heart attack cases.

```
[[46278   295]
 [ 2265   367]]
              precision    recall  f1-score   support

           0       0.95      0.99      0.97     46573
           1       0.55      0.14      0.22      2632

    accuracy                           0.95     49205
   macro avg       0.75      0.57      0.60     49205
weighted avg       0.93      0.95      0.93     49205
```

Random Forest using Optimized Hyperparameters

To optimize the Random Forest model, I conducted hyperparameter tuning using GridSearchCV. This involved searching for the best combination of hyperparameters to further enhance model performance. Hyperparameter tuning allowed me to find the optimal settings for the algorithm, which could significantly improve its predictive power. The grid search identified the best parameters, and using these optimized settings, I retrained the Random Forest model, achieving better performance metrics.

```
Best parameters found: {'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}
[[46230   343]
 [ 2228   404]]
              precision    recall  f1-score   support

           0       0.95      0.99      0.97     46573
           1       0.54      0.15      0.24      2632

    accuracy                           0.95     49205
   macro avg       0.75      0.57      0.61     49205
weighted avg       0.93      0.95      0.93     49205
```

Gradient Boosting Model

I then moved on to Gradient Boosting, a method that builds models sequentially, with each model correcting errors from the previous ones. This approach is powerful for handling complex datasets and provided a significant boost in performance. Gradient Boosting effectively reduces bias and variance, leading to more accurate models.

```
[[45802   771]
 [ 1850   782]]
              precision    recall  f1-score   support

           0       0.96      0.98      0.97     46573
           1       0.50      0.30      0.37      2632

    accuracy                           0.95     49205
   macro avg       0.73      0.64      0.67     49205
weighted avg       0.94      0.95      0.94     49205
```

XGBoost Model

To leverage the strengths of Gradient Boosting further, I implemented XGBoost, a more advanced variant known for its efficiency and performance. XGBoost includes several optimizations and regularization techniques that help prevent overfitting. The XGBoost model performed well, particularly in handling the class imbalance through its built-in scale_pos_weight parameter.

```
[[37027  9546]
 [  592  2040]]
              precision    recall  f1-score   support

           0       0.98      0.80      0.88     46573
           1       0.18      0.78      0.29      2632

    accuracy                           0.79     49205
   macro avg       0.58      0.79      0.58     49205
weighted avg       0.94      0.79      0.85     49205
```

Gradient Boosting Model using Optimum Hyperparameters

I used GridSearchCV to optimize the Gradient Boosting model, like my approach with the Random Forest model. This step identified the best combination of parameters to maximize the model's performance. The hyperparameter tuning process for Gradient Boosting focused on finding the optimal number of trees, learning rate, and tree depth, which were critical factors influencing the model's accuracy and generalization.

```
[[45991   582]
 [ 1964   668]]
              precision    recall  f1-score   support

           0       0.96      0.99      0.97     46573
           1       0.53      0.25      0.34      2632

    accuracy                           0.95     49205
   macro avg       0.75      0.62      0.66     49205
weighted avg       0.94      0.95      0.94     49205
```

Best Parameters Gradient Boosting Model with Class Weights

I adjusted the class weights in this iteration to improve the model's sensitivity to the minority class. This iteration aimed to ensure that the optimized Gradient Boosting model performed well overall and specifically addressed the challenge of identifying heart attack cases. Adjusting the class weights helps in giving more importance to the minority class, so I was hoping to improve the recall and precision for that class.

```
[[45991   582]
 [ 1964   668]]
              precision    recall  f1-score   support

           0       0.96      0.99      0.97     46573
           1       0.53      0.25      0.34      2632

    accuracy                           0.95     49205
   macro avg       0.75      0.62      0.66     49205
weighted avg       0.94      0.95      0.94     49205
```

Combined Model: Gradient Boosting and XGBoost with Soft Voting Classifier

I combined these models using a soft voting classifier to harness the strengths of both Gradient Boosting and XGBoost. In soft voting, each model's predicted probabilities are averaged, and the final prediction is based on the highest probability. This approach leverages the complementary strengths of both models, potentially providing a more robust and accurate prediction. The combination of Gradient Boosting's sequential error correction and XGBoost's efficiency and regularization resulted in a highly effective model for predicting heart disease.

```
[[45991   582]
 [ 1964   668]]
              precision    recall  f1-score   support

           0       0.96      0.99      0.97     46573
           1       0.53      0.25      0.34      2632

    accuracy                           0.95     49205
   macro avg       0.75      0.62      0.66     49205
weighted avg       0.94      0.95      0.94     49205
```

Iteration Summary

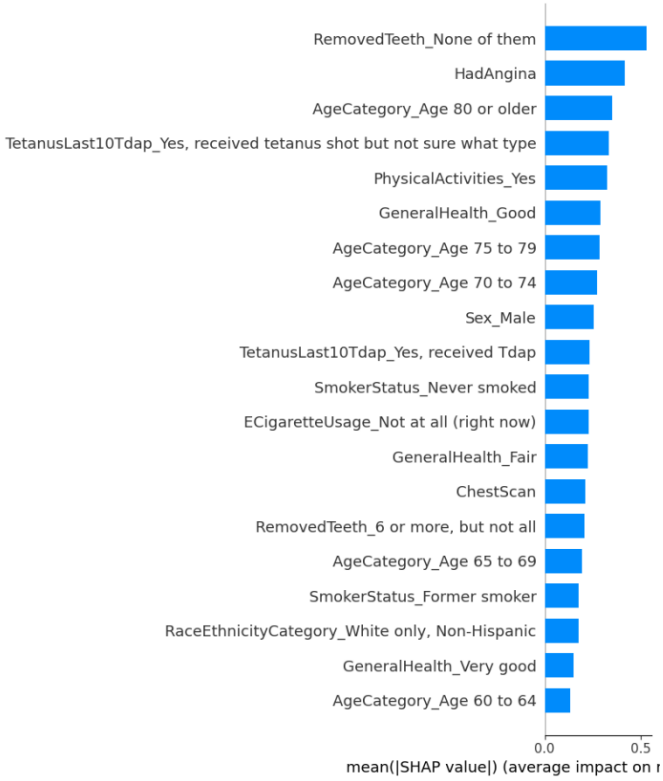
Through these iterative modeling steps, I progressively improved my ability to predict heart disease. Each iteration was built upon the previous one, addressing specific challenges such as class imbalance and model optimization. The final model I selected for this project was a combination of Gradient Boosting and XGBoost with a soft voting classifier. It provided a balanced and accurate prediction, demonstrating the power of ensemble methods and thorough model tuning in achieving reliable predictive performance.

Conclusion

The machine learning models developed in this project, particularly the Voting Classifier, successfully identified key risk factors for heart disease. The use of SHAP values provided detailed insights into feature importance, enhancing the transparency and interpretability of the models. The findings indicate that variables such as oral health, history of angina, age, e-cigarette usage, and self-reported general health significantly contribute to heart disease risk.

I personally find it interesting that oral health has been revealed as a top feature importance in the the final three model iterations. My nursing knowledge can relay that the bacteria that tends to cause tooth decay is the same bacteria that tends to attack heart structures like valves. I find this interesting because the data didn't mention this connection. It is something the model learned as it worked through its predictions. Although all healthcare providers are aware of this oral/cardiac connection, it is a causative factor that can be easily overlooked at the point of care without a project like this to point out the importance of it.

The visualization of top feature importances is presented here along with the list of top feature importances for the Gradient Boosting model, the XGBoost model, and the ensemble combination.



```
Top features by SHAP values for Gradient Boosting:
                                                feature  shap_value
92                           RemovedTeeth_None of them    0.528651
3                                           HadAngina    0.415738
114                        AgeCategory_Age 80 or older    0.352476
116  TetanusLast10Tdap_Yes, received tetanus shot b...    0.332960
89                              PhysicalActivities_Yes    0.325065
83                                  GeneralHealth_Good    0.288315
113                           AgeCategory_Age 75 to 79    0.286483
112                           AgeCategory_Age 70 to 74    0.272179
81                                            Sex_Male    0.256213
115               TetanusLast10Tdap_Yes, received Tdap    0.230731


Top features by SHAP values for XGBoost:
                                                feature  shap_value
92                           RemovedTeeth_None of them    0.731528
96               ECigaretteUsage_Not at all (right now)    0.510021
114                        AgeCategory_Age 80 or older    0.437403
83                                  GeneralHealth_Good    0.431772
113                           AgeCategory_Age 75 to 79    0.422174
3                                           HadAngina    0.408164
89                              PhysicalActivities_Yes    0.349890
115               TetanusLast10Tdap_Yes, received Tdap    0.343288
112                           AgeCategory_Age 70 to 74    0.338941
116  TetanusLast10Tdap_Yes, received tetanus shot b...    0.333658



Top combined features by SHAP values:
                                                feature  shap_value
92                           RemovedTeeth_None of them    0.630089
3                                           HadAngina    0.411951
114                        AgeCategory_Age 80 or older    0.394939
96               ECigaretteUsage_Not at all (right now)    0.368160
83                                  GeneralHealth_Good    0.360043
113                           AgeCategory_Age 75 to 79    0.354329
89                              PhysicalActivities_Yes    0.337477
116  TetanusLast10Tdap_Yes, received tetanus shot b...    0.333309
112                           AgeCategory_Age 70 to 74    0.305560
115               TetanusLast10Tdap_Yes, received Tdap    0.287009
```

Assumptions

Several assumptions were made during the project:

1. The survey responses accurately reflect the actual health statuses of the respondents.

2. The dataset is representative of the general population.
3. The selected features adequately capture the variables impacting heart disease risk.

Limitations

The study faced several limitations:

1. Potential biases in self-reported data could affect the accuracy of predictions.
2. The dataset was limited to variables collected in the CDC survey, potentially missing other relevant health factors.
3. The models may not capture all nuances of individual health conditions, leading to prediction variability.

Challenges

The project encountered various challenges:

1. Handling the significant class imbalance in the target variable required careful application of techniques like SMOTE.
2. Ensuring model interpretability while maintaining high accuracy was crucial.
3. Balancing precision and recall for the minority class to avoid high false positive or false negative rates.
4. Selecting the best, most well-balanced model was challenging due to the large number of iterations attempted.

Future Uses/Additional Applications

The methods and findings from this project can be applied to other health conditions with similar data structures. Integrating the model into healthcare systems can provide early warnings and enable timely interventions. The insights gained can inform public health policies and preventative measures, improving overall health.

Recommendations

Based on the findings, several recommendations are proposed:

- Further fine-tuning of the models to enhance performance.
- Implementing the model in real-world healthcare settings for validation and refinement.
- Continuously updating the model with new data to maintain accuracy and relevance.

Implementation Plan

1. **Model Deployment:**
- Integrate the model into a healthcare analytics platform.
- Provide training to healthcare professionals on using the model effectively.
2. **Monitoring and Maintenance:**
- Regularly update the model with new data.
- Monitor model performance and make necessary adjustments to ensure continued accuracy.
3. **User Training:**
- Educate healthcare providers on interpreting and using model predictions for patient care decisions.

Ethical Assessment

The project adheres to ethical standards in the following ways:

Patient Privacy: Ensuring all data is anonymized and compliant with HIPAA regulations.

Bias and Fairness: Addressing potential biases regarding race, socioeconomic status, and geographic location to ensure fair model predictions.

Accessibility: Making the model interpretable and actionable for healthcare providers.

10 Questions an Audience Would Ask Me:

1. How did you handle the class imbalance in the dataset?
2. What feature engineering techniques did you apply?
3. How did you ensure the model is not biased towards any demographic group?
4. What were the most significant predictors of heart disease according to your model?
5. How do you interpret the model's predictions in a real-world healthcare setting?
6. What measures did you take to ensure data privacy and security?
7. How did the model's performance vary with different algorithms?
8. Can the model be applied to other diseases or health conditions?
9. How do you plan to validate the model's predictions?
10. What challenges did you face during the data preprocessing stage, and how did you overcome them?

References

Abbott, D. (2014). Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst. Indianapolis: Wiley.

Madhumita Pal, S. P. (2022). Risk Prediction of Cardiovascular Disease Using Machine Learning Classifiers. Open Medicine, 1100-1113.

Morid MA, K. K. (2018 April). Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation. AMIA, 1312-1321.

Nadiah A. Baghdadi, S. M. (2023, September). Advanced Machine Learning Techniques for Cardiovascular Disease Early Detection and Diagnosis. Journal of Big Data, 10.

Sidra Abbas, S. O. (2024). Artificial Intelligence Framework for Heart Disease Classification. Scientific Reports, 3123.

Umarani Nagavelli, D. S. (2022). Machine Learning Technology-Based Heart Disease Detection Models. Journal of Healthcare Engineering.

Yar Muhammad, M. T. (2020). Early and Accurate Detection and Diagnosis of Heart Disease Using Intelligent Computational Model. Sci Rep. doi:https://doi.org/10.1038/s41598-020-76635-9