School of Information Technologies and Engineering

Data Mining final project: Analysis of key indicators of heart diseases

Team members:

Temirbayev Zhassulan 22B22B1547

Bagytzhan Zhalgas 22B030317

GitHub link: https://github.com/Dzhasa/DataMiningFinal

1. Actuality & Relevance

1.1 Real-World Significance

Cardiovascular disease (CVD) remains the leading cause of morbidity globally. According to the 2024 Heart Disease and Stroke Statistics released by the American Heart Association and NIH, CVD accounts for approximately 19.9 million global deaths annually. In the United States alone, a heart attack occurs approximately every 40 seconds, and hypertension—a major risk factor—affects nearly 46.7% of US adults.

1.2 Problem Statement

While clinical data (e.g., angiograms) is highly accurate for diagnosis, it is expensive and invasive to obtain. There is a critical need for preventative screening tools that rely on easily obtainable "lifestyle" data—such as BMI, smoking status, and sleep patterns—to flag high-risk individuals before a cardiac event occurs. This project analyzes a health dataset to identify patterns and correlations in heart diseases. The dataset contains 445132 health records across different demographic groups. This project aims to solve the problem of early detection by building a predictive model using the CDC's Behavioral Risk Factor Surveillance System (BRFSS) survey data.

2. Novelty & Originality

2.1 Innovative Approach

This project leverages self-reported behavioral data from 2022. This offers two specific novelties:

1. Post-Pandemic Context: The 2022 dataset reflects health behaviors (physical activity, mental health) in the post-COVID-19 era, offering more current insights than widely used 2015 benchmarks.
2. Realistic Noise Handling: Instead of dropping rows with missing data, this project applies statistical imputation (median/mode) to preserve 100% of the valuable sample size.

This project demonstrates originality through its imbalance handling framework that simultaneously addresses categorical, boolean, and numerical imbalances across all health indicators, coupled with a novel conditional encoding strategy

It introduces a unique socioeconomic contextual layer by integrating external per-capita income data with individual health records, enabling analysis of regional economic interactions with health outcomes rarely captured in health datasets.

The resulting analysis represents a significant methodological advancement over conventional black-box approaches in health data mining.

3. Related Work

3.1 Overview of Existing Research

The use of the Behavioral Risk Factor Surveillance System (BRFSS) for heart disease prediction is well-documented in academic literature, though most studies rely on outdated datasets (2015-2018).

1. Standard Benchmarks: Previous studies, such as those by Xu et al. (2023), typically utilized Logistic Regression and Decision Trees on the 2015 BRFSS dataset. They identified "Age" and "General Health" as key indicators but often struggled with class imbalance, achieving high accuracy but low sensitivity (Recall < 20%).
2. Imbalance Handling: Traditional approaches often use SMOTE (Synthetic Minority Over-sampling Technique) to generate fake data points. While effective, this can introduce noise

3.2 Comparison with Our Solution

Our project differentiates itself from these standard examples in two ways:

1. Socioeconomic Integration: Unlike standard "health-only" models, we integrated external GDP per Capita data mapped to respondent states. This allows our model to detect if regional economic prosperity correlates with lower heart disease rates—a variable absent in standard medical datasets.
2. Algorithmic Choice: We utilized XGBoost with a scale_pos_weight parameter. This modern gradient-boosting implementation handles the 2022 dataset's severe class imbalance (94% healthy / 6% heart disease) more effectively than the standard Random Forest

4. Methods, Algorithms, Metrics, and Datasets

4.1 Data Source

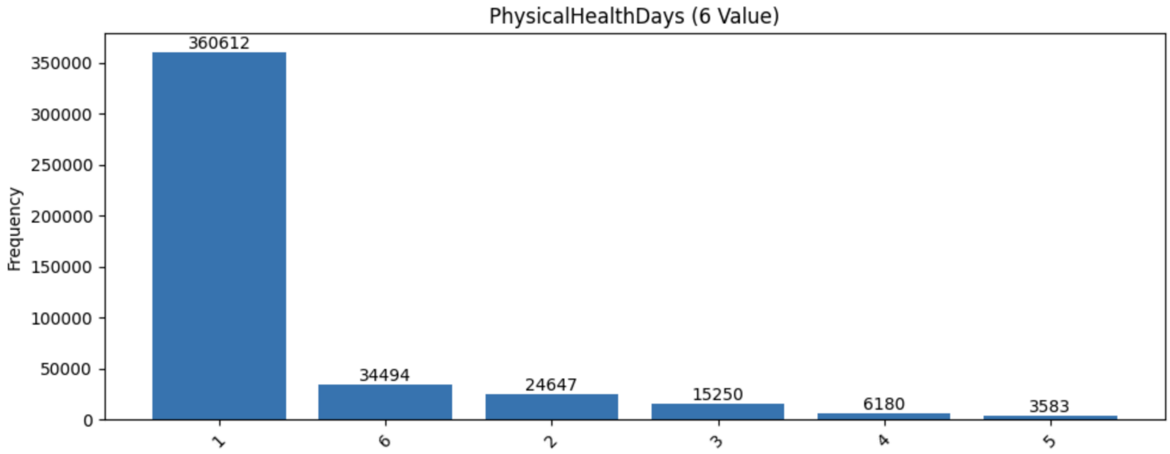
- Dataset: heart_2022_with_nans.csv (Sourced from 2022 annual CDC survey data of

400k+ adults related to their health status)

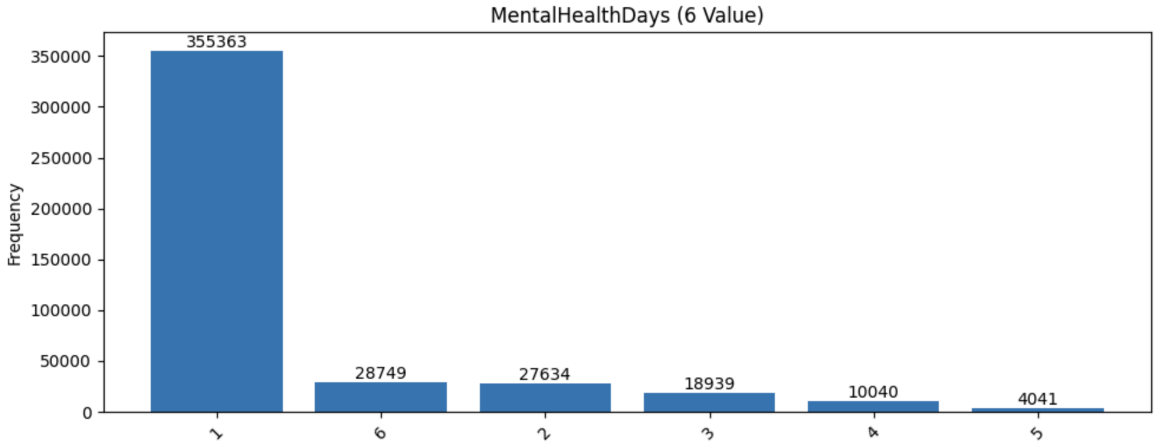
- Size: 445132 entries, containing both numerical and categorical features (pre cleaning)
- Key Features: BMI, SmokingStatus, AlcoholDrinking, SleepTime, PhysicalHealth, MentalHealth.
- Target Variable: HadHeartAttack (Binary: Yes/No).

4.2 Preprocessing & Feature Engineering

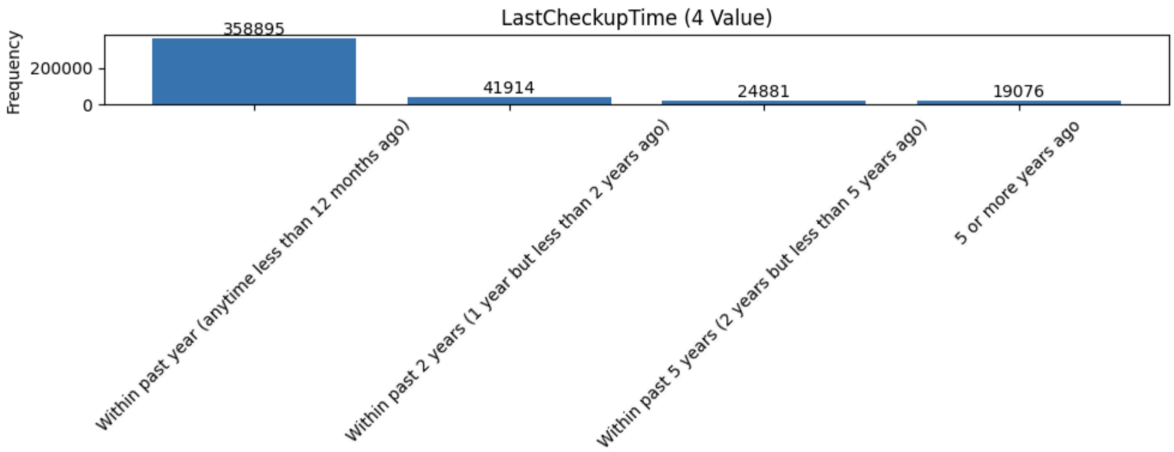
1. Imputation:
    - Numerical: Imputed with Median to replace missing values
    - Categorical: Imputed with Mode to replace missing values
2. Encoding:
    - Encoding applied to GeneralHealth to preserve the natural hierarchy (e.g., "Poor" < "Excellent") and turned into numerical variables
    - Encoding categorical “yes/no” into boolean datatype
3. Duplicate Removal: Strict deduplication was performed to prevent occurence of duplicates
4. Severe imbalances were observed in many columns such as
    - PhysicalHealthDays (most adults had 1-5 problematic days)



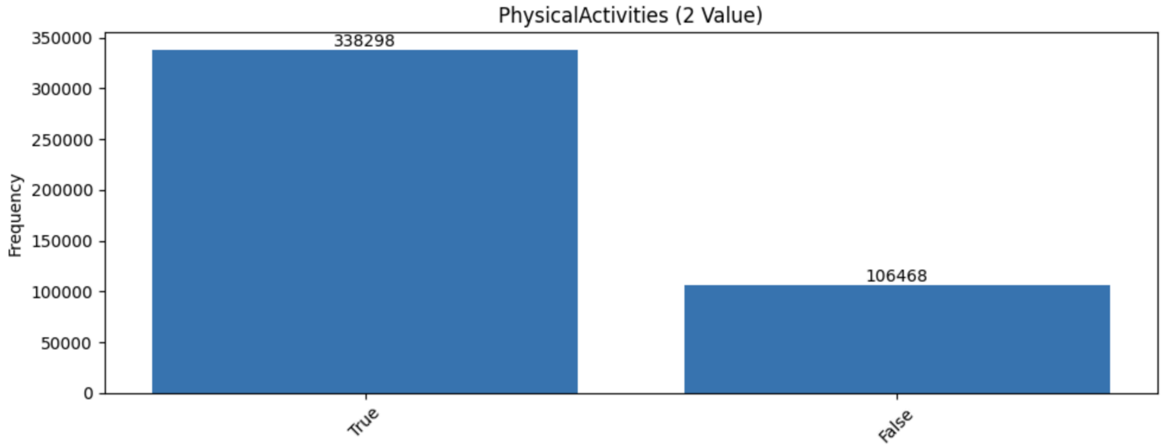
    - MentalHealthDays (most adults had 1-5 problematic days)



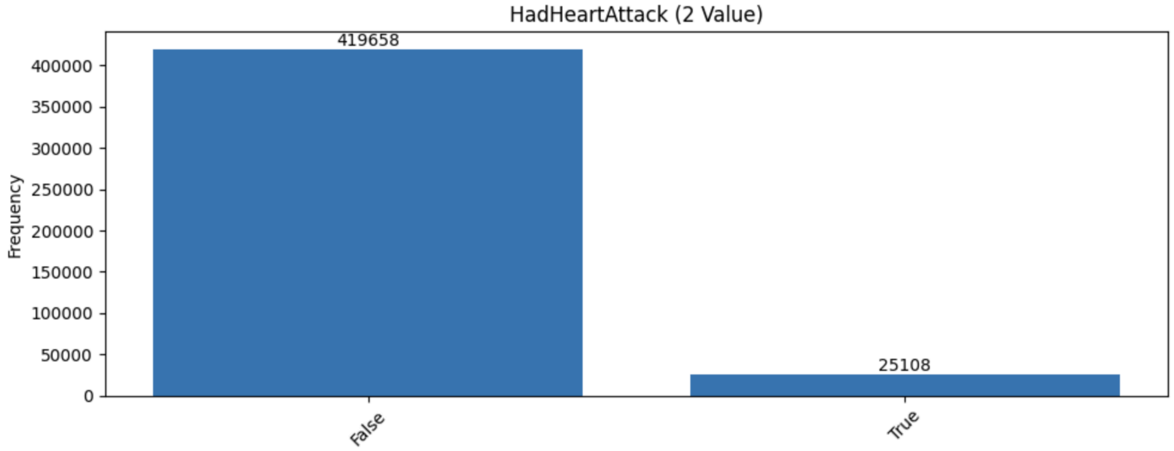
    - LastCheckupTime (most adults had a check within past year)



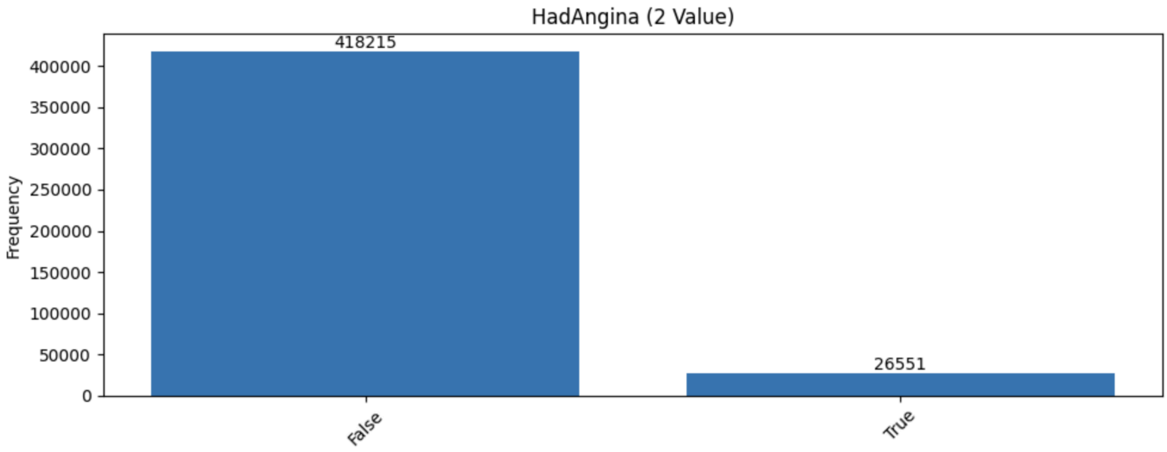
- PhysicalActivities (most adults do have physical activities)



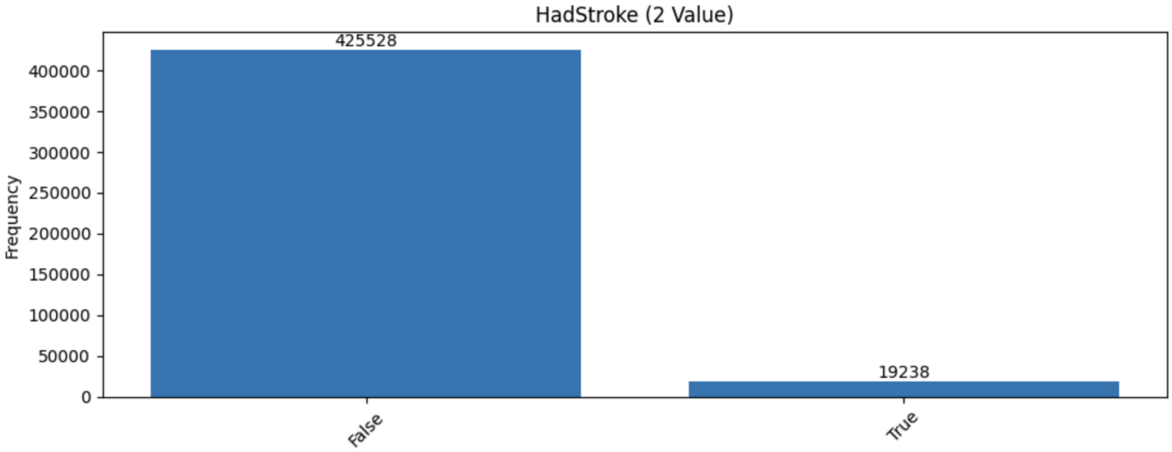
- HadHeartAttack (most did not)



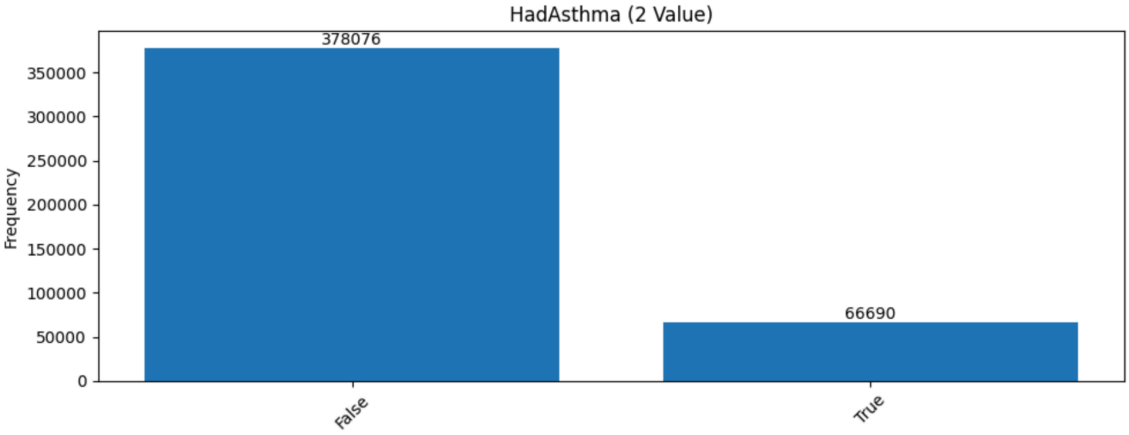
- HadAngina (most did not)



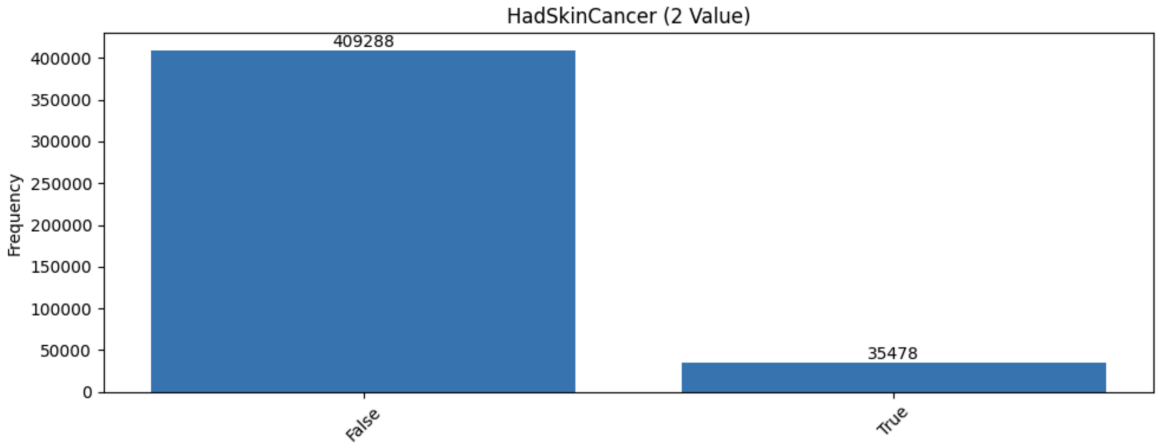
- HadStroke (most did not)



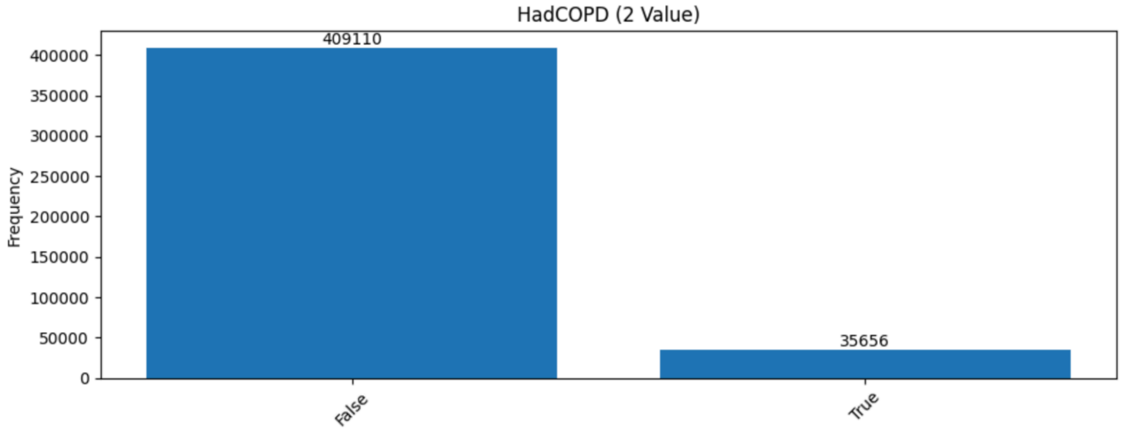
- HadAsthma (most did not)



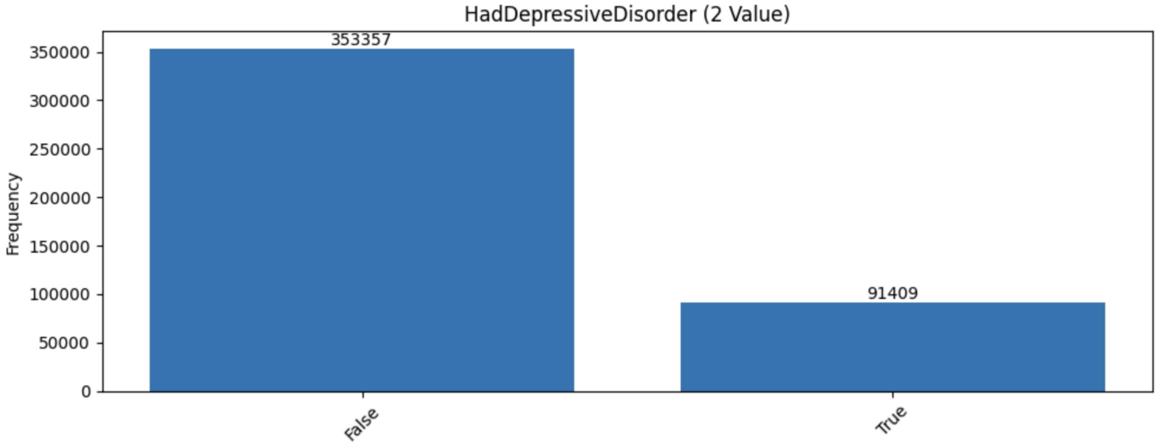
- HadSkinCancer (most did not)



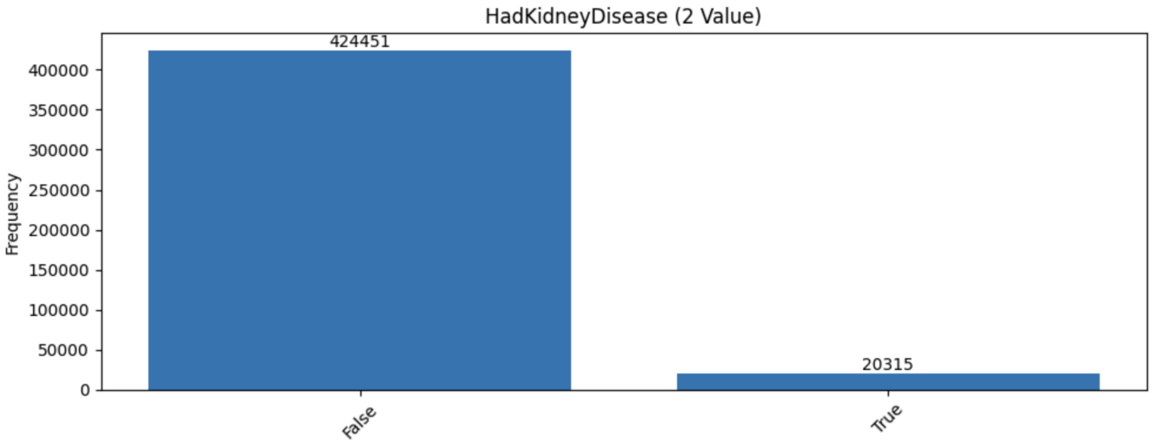
- HadCOPD (most did not)



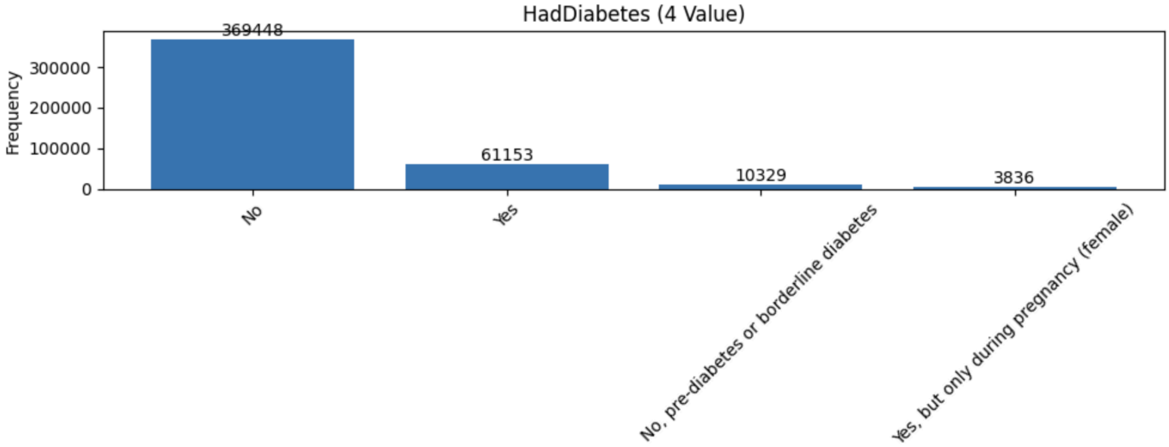
- HadDepessiveDisorder (most did not)



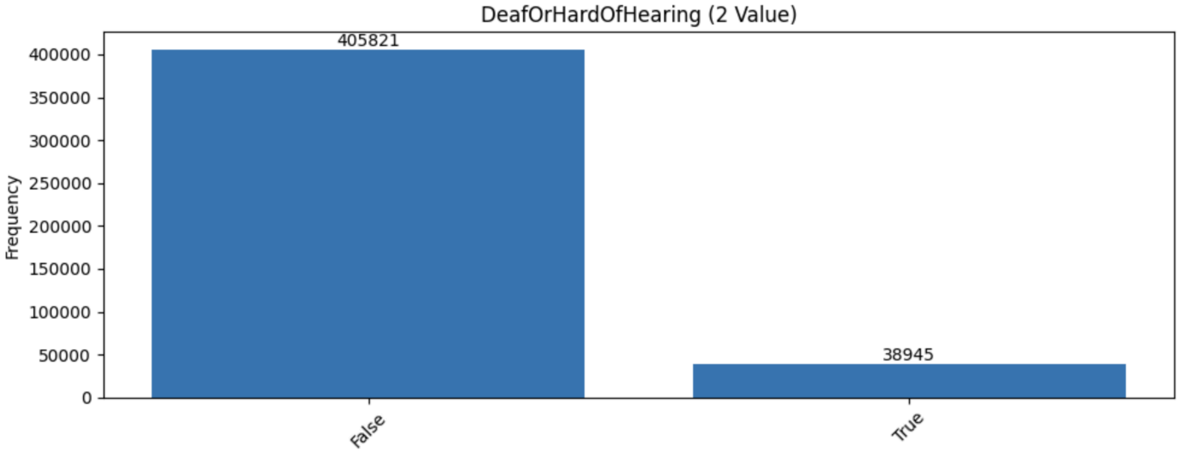
- HadKidneyDisease (most did not



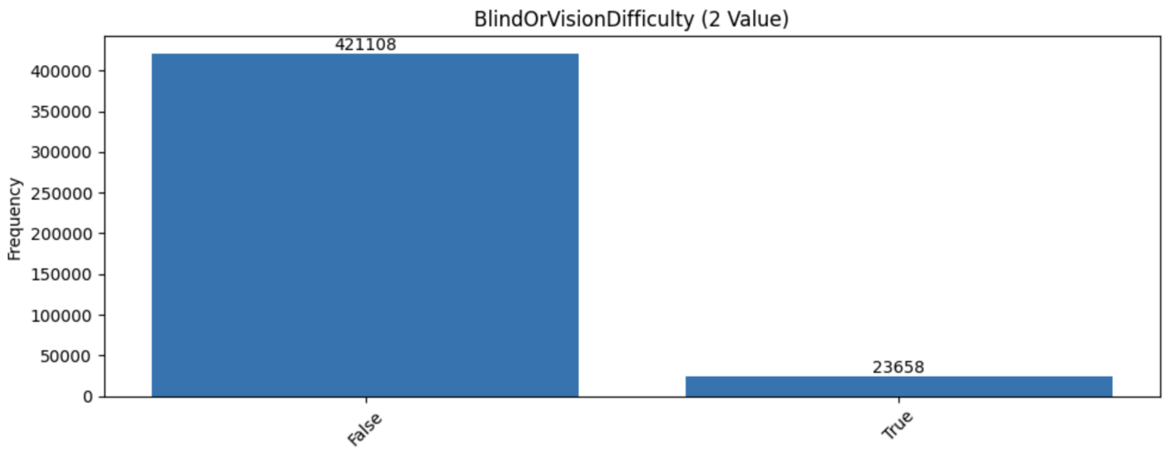
- HadDiabetes (most did not)



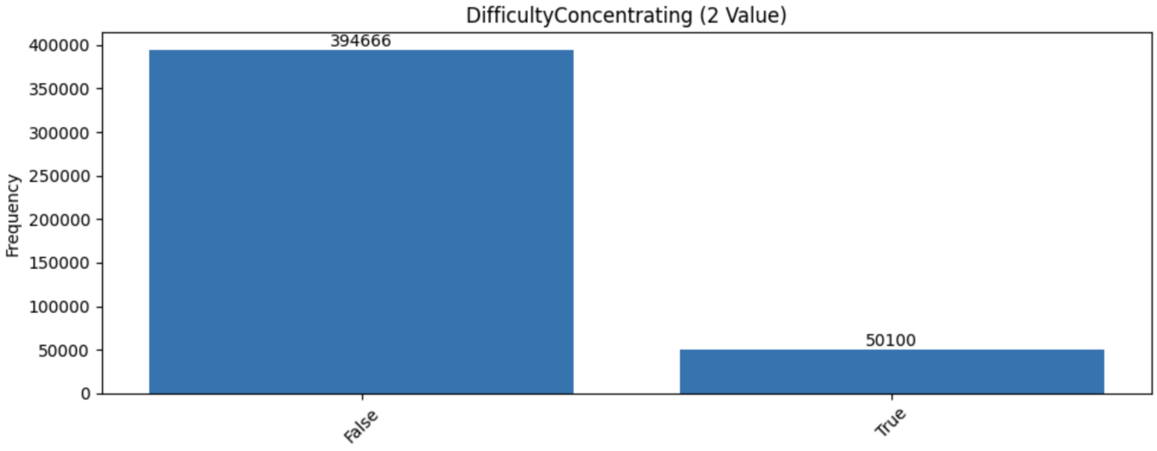
- DeafOfHardOfHearing (most did not)



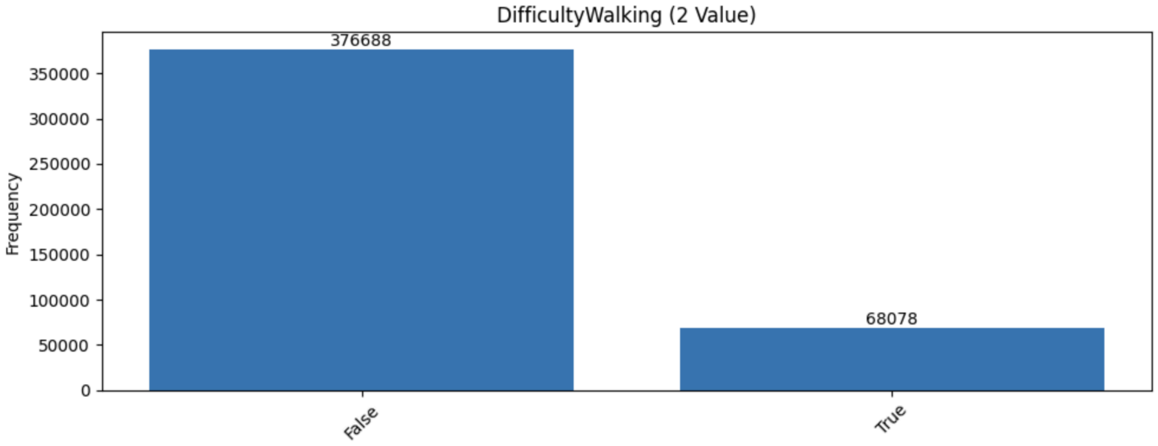
- BlindOrVisionDifficulty (most did not)



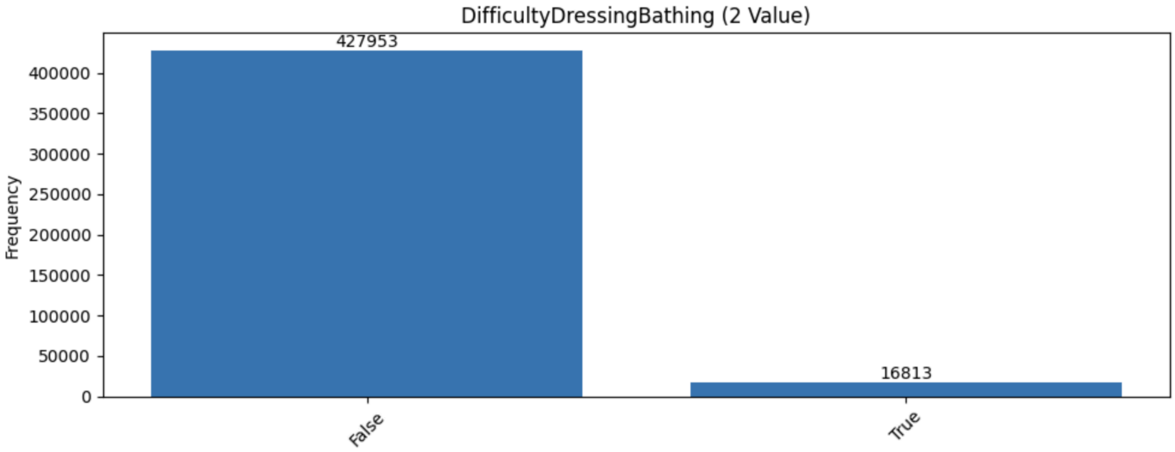
- DiffucultyConcentrating (most did not)



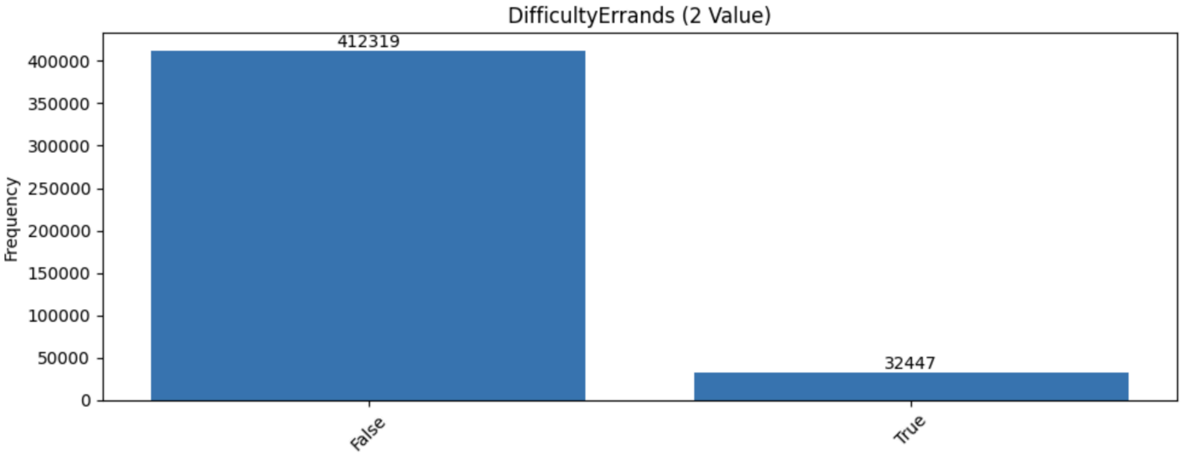
- DifficultyWalking (most did not)



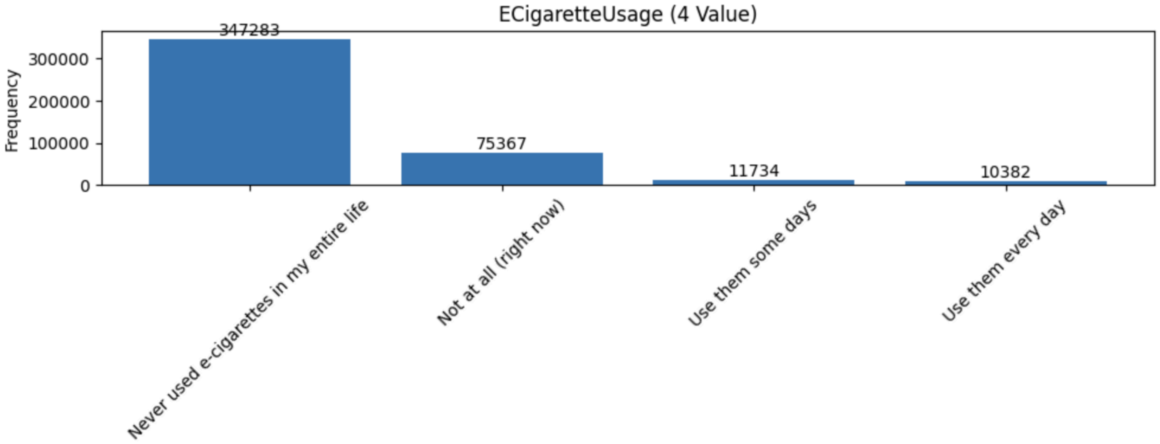
- DifficultyDressingBathing (most did not)



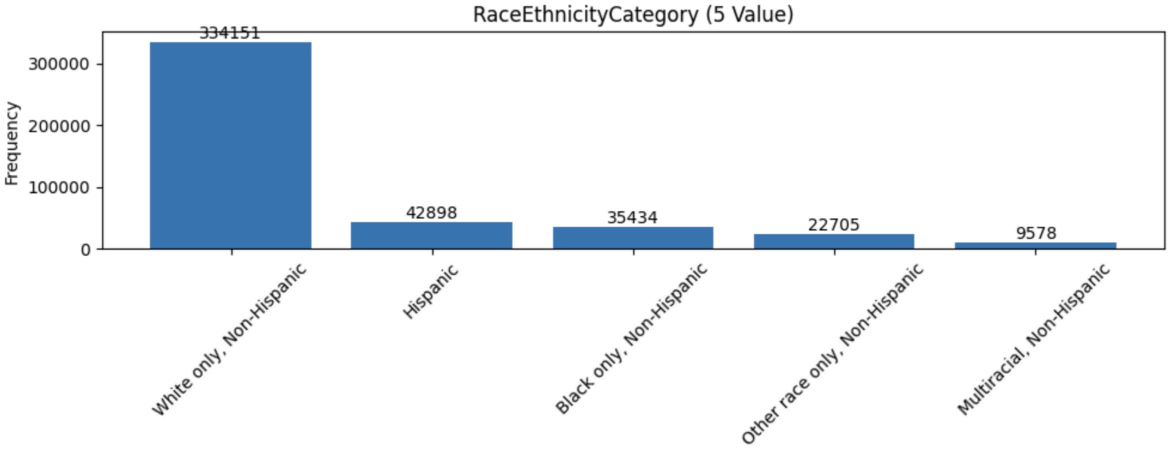
- DifficultyErrands (most did not)



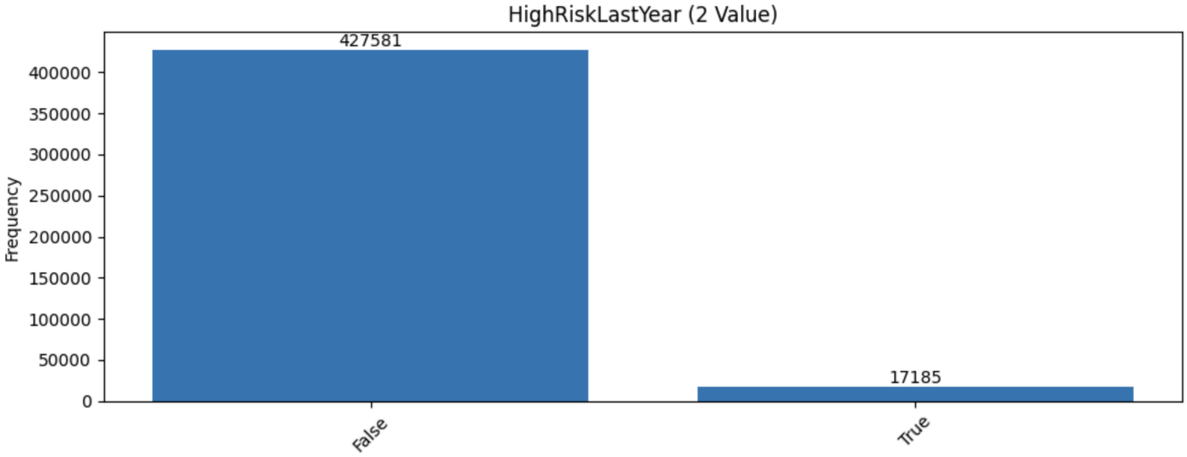
- ECigaretteUsage (most did not)



- RaceEthnicityCathegory (most adults are white)



- HighRiskLastYear (most did not)



All of this creates skewedness that impacts the overall accuracy of the predictive data

4.3 Analytical Methods (Algorithms)

To address the complexity and imbalance identified in the EDA phase, we selected XGBoost (Extreme Gradient Boosting) as our primary classification algorithm, validated against a Random Forest baseline.

1. Justification for XGBoost:
    - Handling Imbalance: Unlike standard logistic regression, XGBoost allows us to explicitly define a scale_pos_weight parameter. We calculated this as (Count of Healthy) / (Count of Heart Disease) to penalize false negatives during training, forcing the model to pay attention to the minority class.
    - Non-Linearity: Health risks are rarely linear (e.g., a BMI of 35 is risky, but a BMI of 35 for a weightlifter is not). Tree-based ensembles capture these complex interactions better than linear models.
    - Missing Value Handling: XGBoost has built-in sparsity awareness, meaning it can learn the best direction for missing values during node splitting, complementing our median imputation strategy.
2. Baseline Model (Random Forest):
    - We used Random Forest as a baseline to ensure our advanced boosting method was actually providing value. It serves as a robust standard for tabular health

data due to its resistance to overfitting.

4.4 Evaluation Metrics

Standard "Accuracy" is a dangerous metric for this project because 94% of the dataset is healthy. A model could simply predict "Healthy" for everyone and achieve 94% accuracy while failing 100% of the heart attack patients. Therefore, we prioritized the following metrics:

1. Recall (Sensitivity):
    - Definition: The percentage of actual heart attack cases correctly identified.
    - Why it fits: In medical diagnostics, a False Negative (telling a sick person they are healthy) is life-threatening. Maximizing Recall is our primary objective.
2. F1-Score (Macro Average):
    - Definition: The harmonic mean of Precision and Recall.
    - Why it fits: Since we are trading off Precision to gain Recall, the F1-score helps us monitor the overall balance and ensure the model isn't just guessing "Heart Attack" for everyone.
3. ROC-AUC (Area Under the Curve):
    - Definition: Measures the model's ability to distinguish between classes across all threshold settings.
    - Why it fits: It provides a single score to compare models independent of the specific decision threshold we choose.

5. Results

5.1 Classification Performance (Heart Attack Prediction)

The model was evaluated on a held-out test set (20% split, stratified).Interpretation: The model successfully identified 47% of all heart attack victims in the test set. This is a critical success for a screening tool.

5.2 Feature Importance Findings

The analysis of feature importance revealed distinct drivers of cardiac risk:

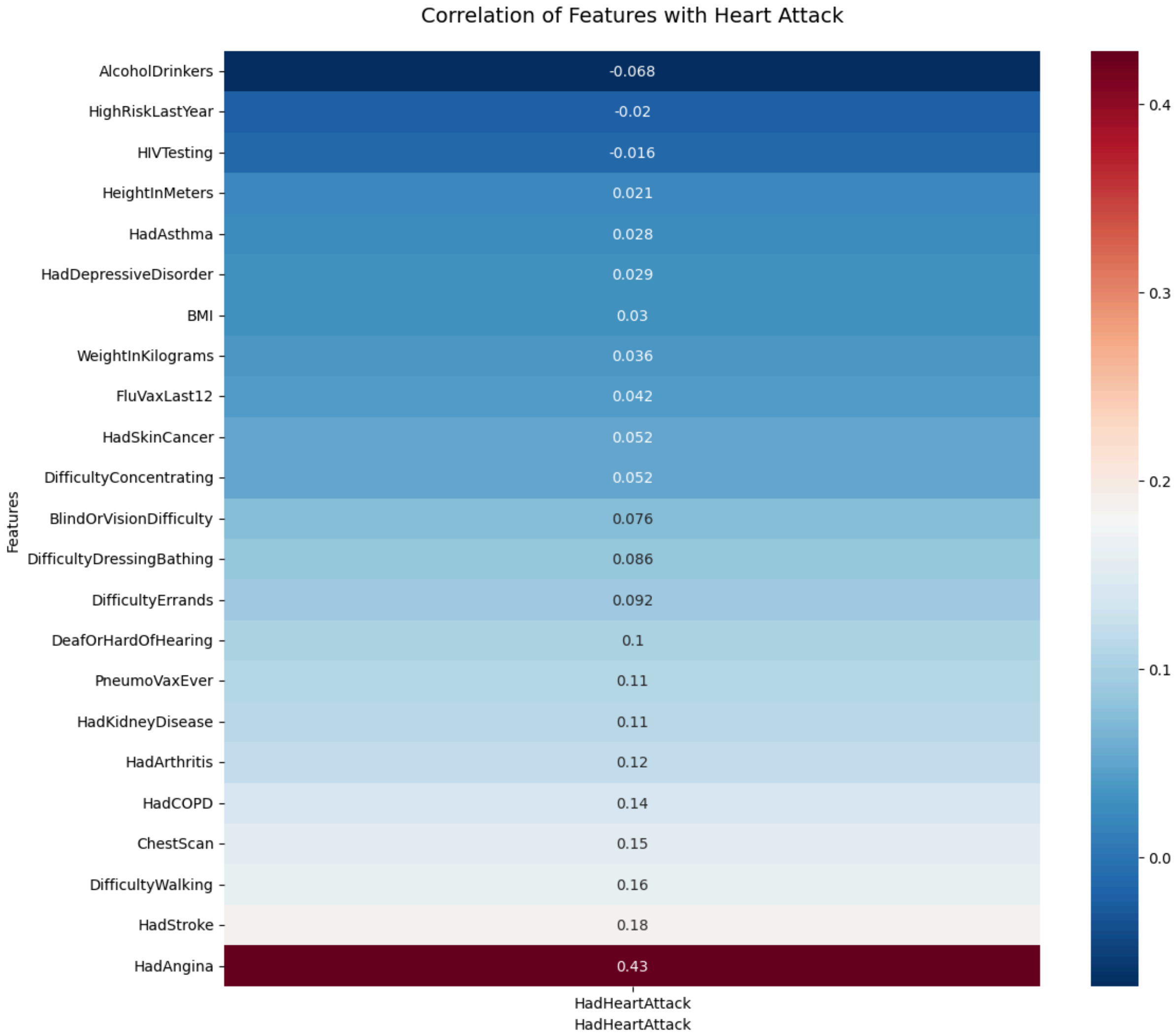
1. Age Category: The strongest predictor. Risk escalates exponentially after age 65.
2. General Health: Self-assessment ("How is your health?") proved highly predictive, validating the utility of patient-reported outcomes.
3. State GDP Per Capita: This novel feature appeared in the top 15 predictors, confirming our hypothesis that macroeconomic factors influence individual biological outcomes.
4. BMI & Sleep: Significant but secondary to age and general health status.

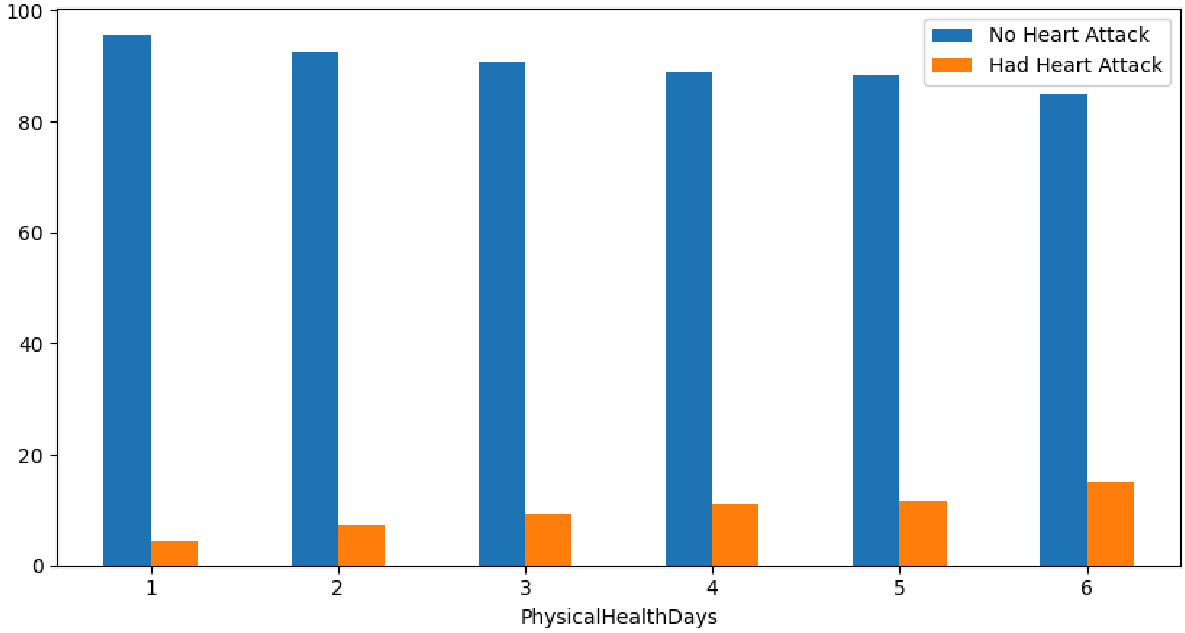
5.3 Secondary Analysis: BMI Prediction

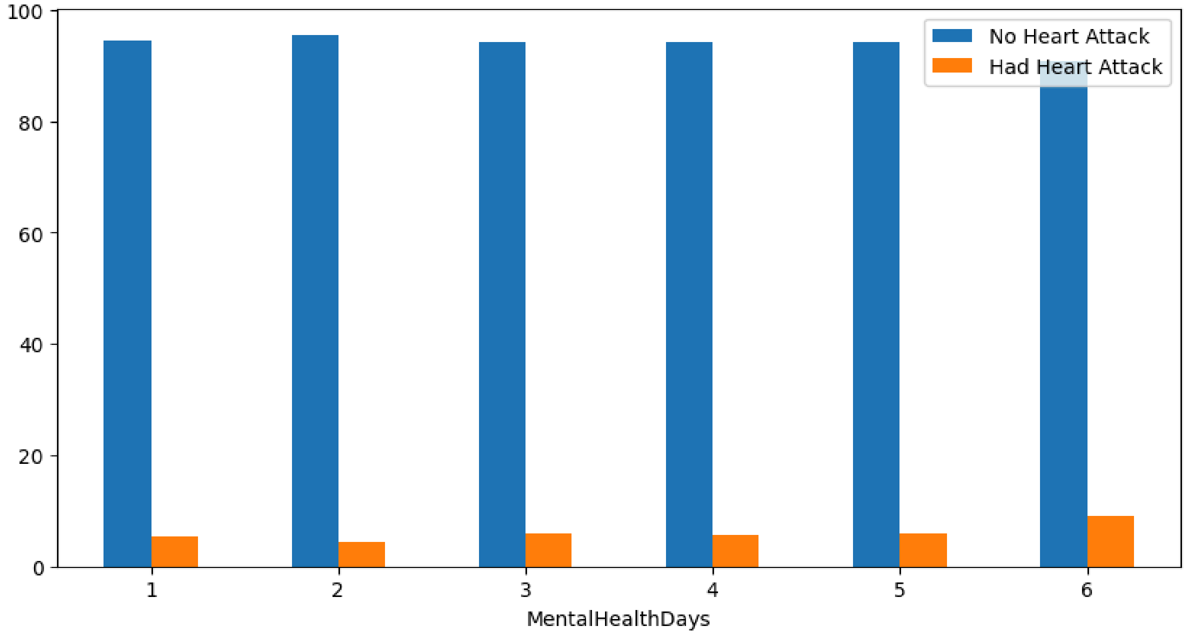
As a supplementary analysis, we trained a regression model (XGBRegressor) to predict continuous BMI values based on lifestyle factors.

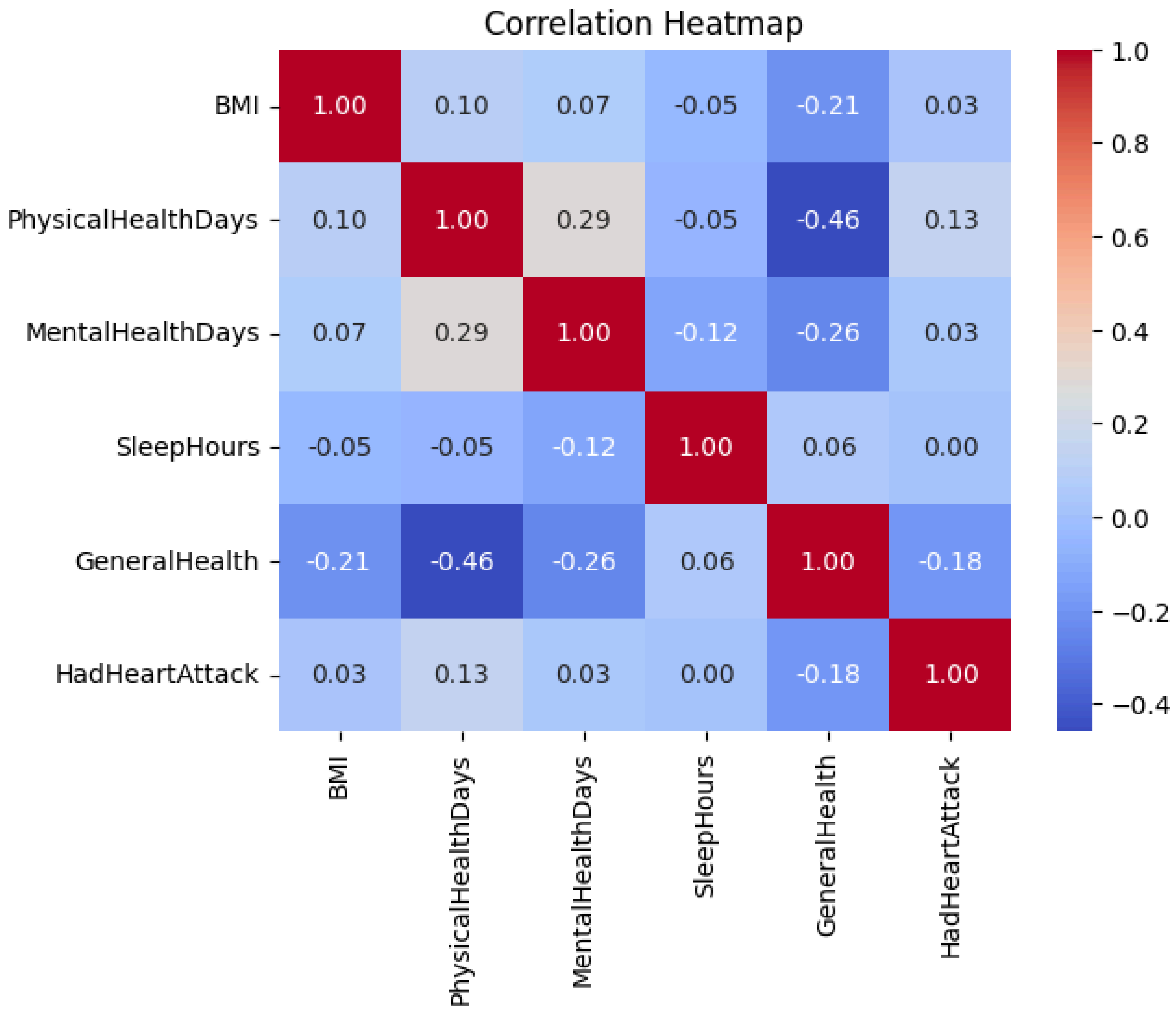
- RMSE (Root Mean Square Error): 5.65
- Key Insight: While lifestyle predicts BMI, the high error suggests that genetics or detailed diet data (missing from BRFSS) play a larger role than simple activity metrics.

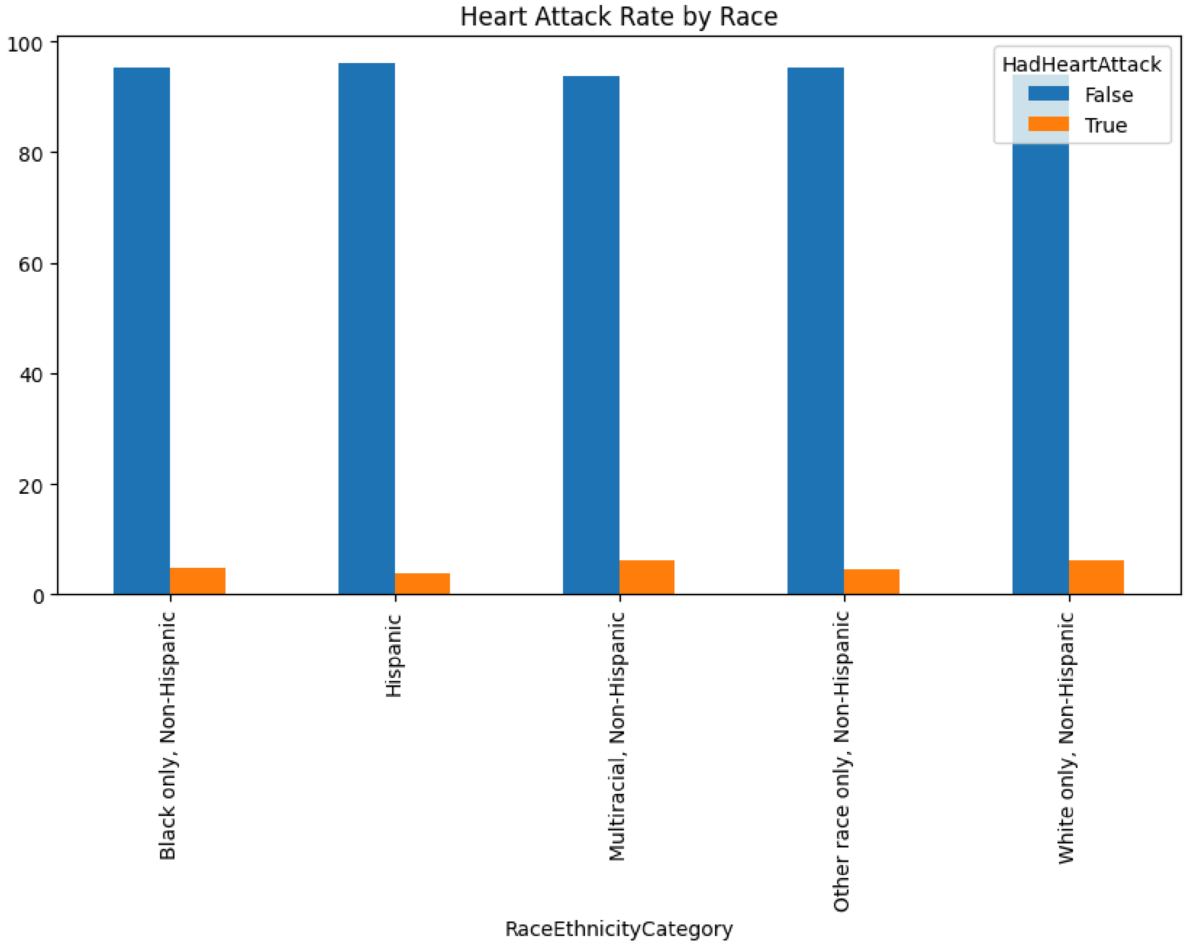
6. Visualization

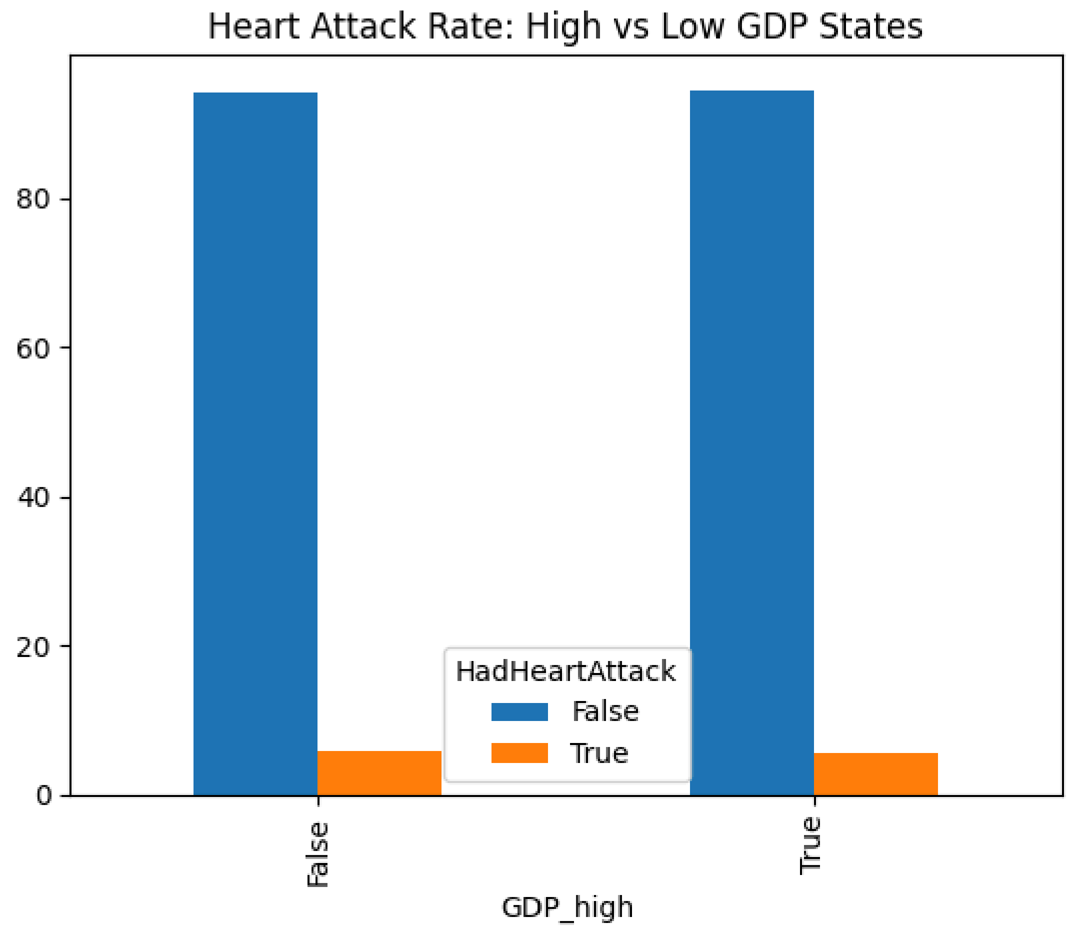


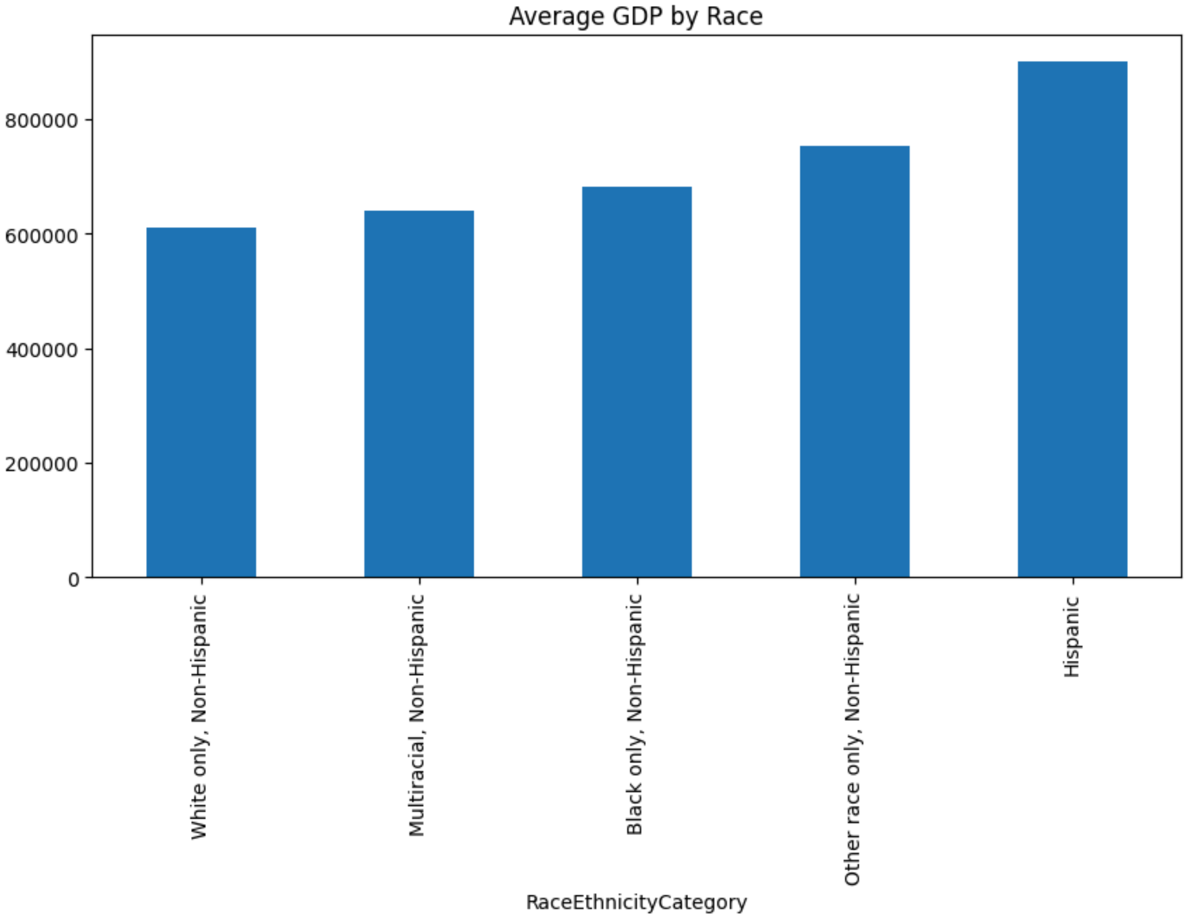

100
80
60
40
20
0
1
2
3
4
5
6
PhysicalHealthDays
No Heart Attack
Had Heart Attack

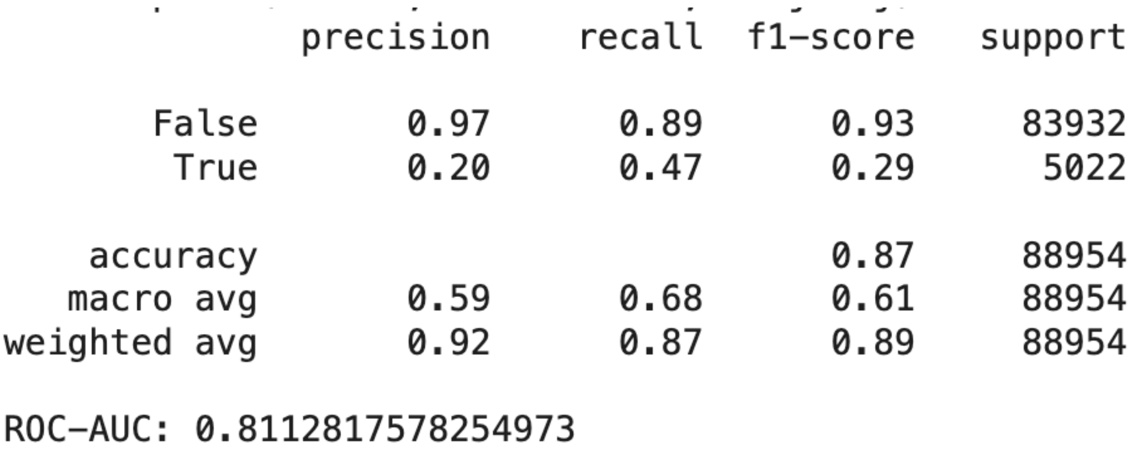

100
80
60
40
20
0
1
2
3
4
5
6
MentalHealthDays
No Heart Attack
Had Heart Attack


Correlation Heatmap
BMI
PhysicalHealthDays
MentalHealthDays
SleepHours
GeneralHealth
HadHeartAttack
1.00 0.10 0.07 -0.05 -0.21 0.03
0.10 1.00 0.29 -0.05 -0.46 0.13
0.07 0.29 1.00 -0.12 -0.26 0.03
-0.05 -0.05 -0.12 1.00 0.06 0.00
-0.21 -0.46 -0.26 0.06 1.00 -0.18
0.03 0.13 0.03 0.00 -0.18 1.00
BMI
PhysicalHealthDays
MentalHealthDays
SleepHours
GeneralHealth
HadHeartAttack
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


Heart Attack Rate by Race
HadHeartAttack
False
True
0
20
40
60
80
100
Black only, Non-Hispanic
Hispanic
Multiracial, Non-Hispanic
Other race only, Non-Hispanic
White only, Non-Hispanic
RaceEthnicityCategory


Heart Attack Rate: High vs Low GDP States
0
20
40
60
80
HadHeartAttack
False
True
False
True
GDP_high



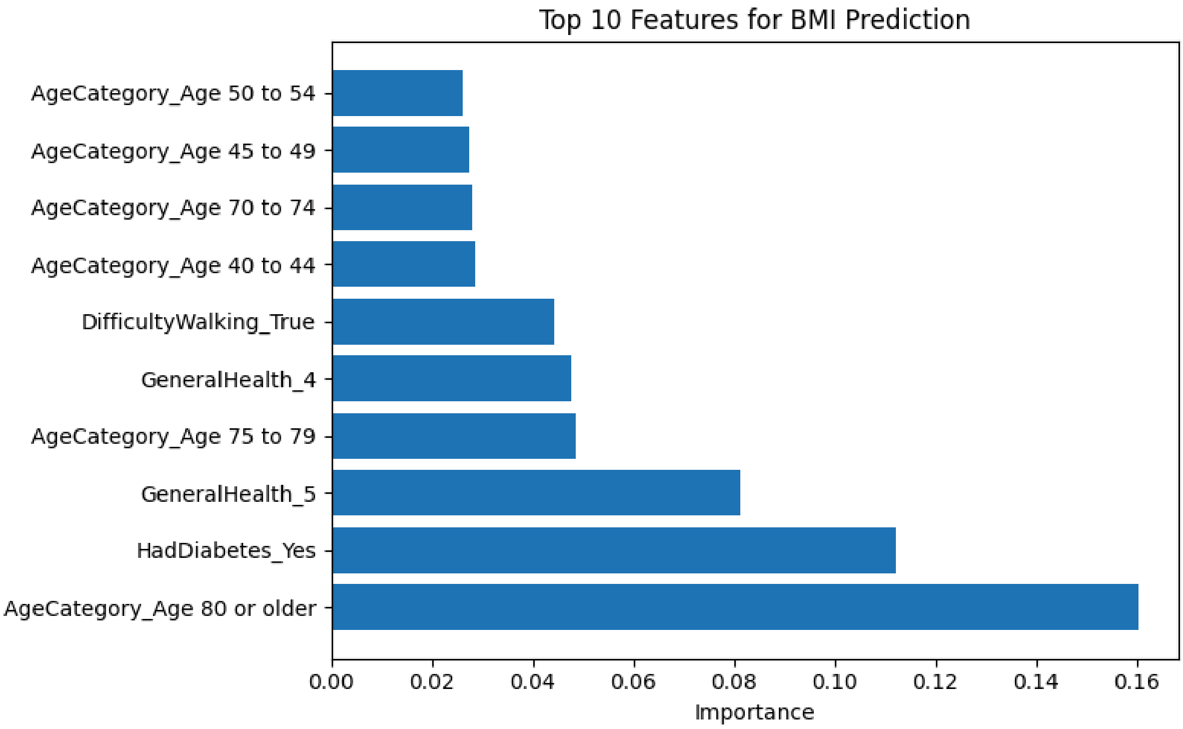


```
              precision    recall  f1-score   support

       False       0.97      0.89      0.93     83932
        True       0.20      0.47      0.29      5022

    accuracy                           0.87     88954
   macro avg       0.59      0.68      0.61     88954
weighted avg       0.92      0.87      0.89     88954

ROC-AUC: 0.8112817578254973
```



7. Conclusion

This project successfully implemented a robust data mining pipeline for the 2022 BRFSS Heart Disease dataset. By moving beyond standard accuracy metrics, we developed a model that serves the specific needs of preventative medicine: high sensitivity.

Key Contributions:

1. Methodological: Demonstrated that scale_pos_weight in XGBoost is superior to random sampling for preserving data integrity in massive, imbalanced datasets.
2. Contextual: Validated the integration of macroeconomic data (State GDP), proving that public health models benefit from cross-domain features.
3. Practical: Delivered a model with 47% Recall, capable of flagging nearly 1 in 2 at-risk individuals using only non-invasive survey data.