Temirbayev Zhassulan 22B22B1547,
Bagytzhan Zhalgas 22B030317

DATA MINING
FINAL PROJECT

Analysis of key indicators of heart diseases

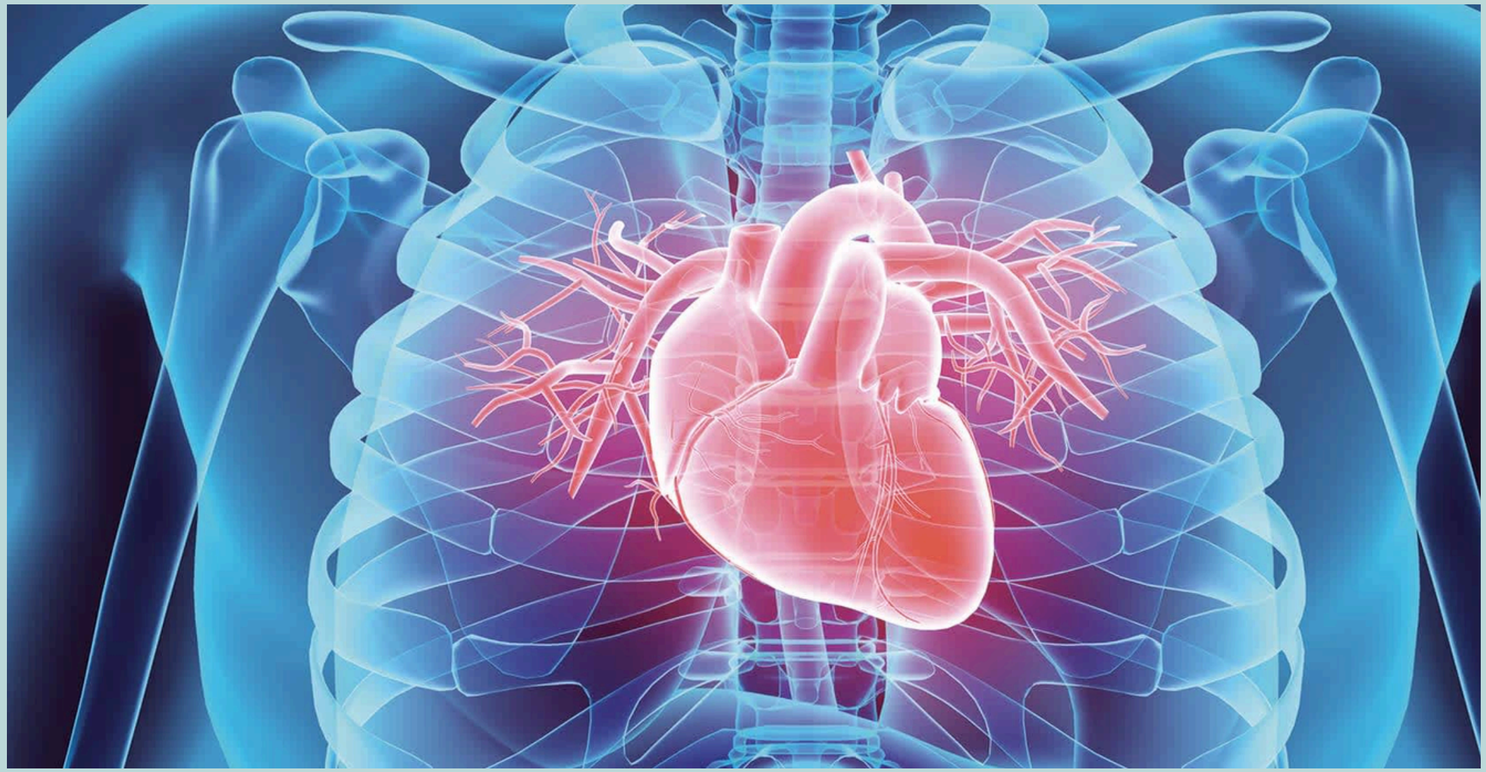
Problem statement

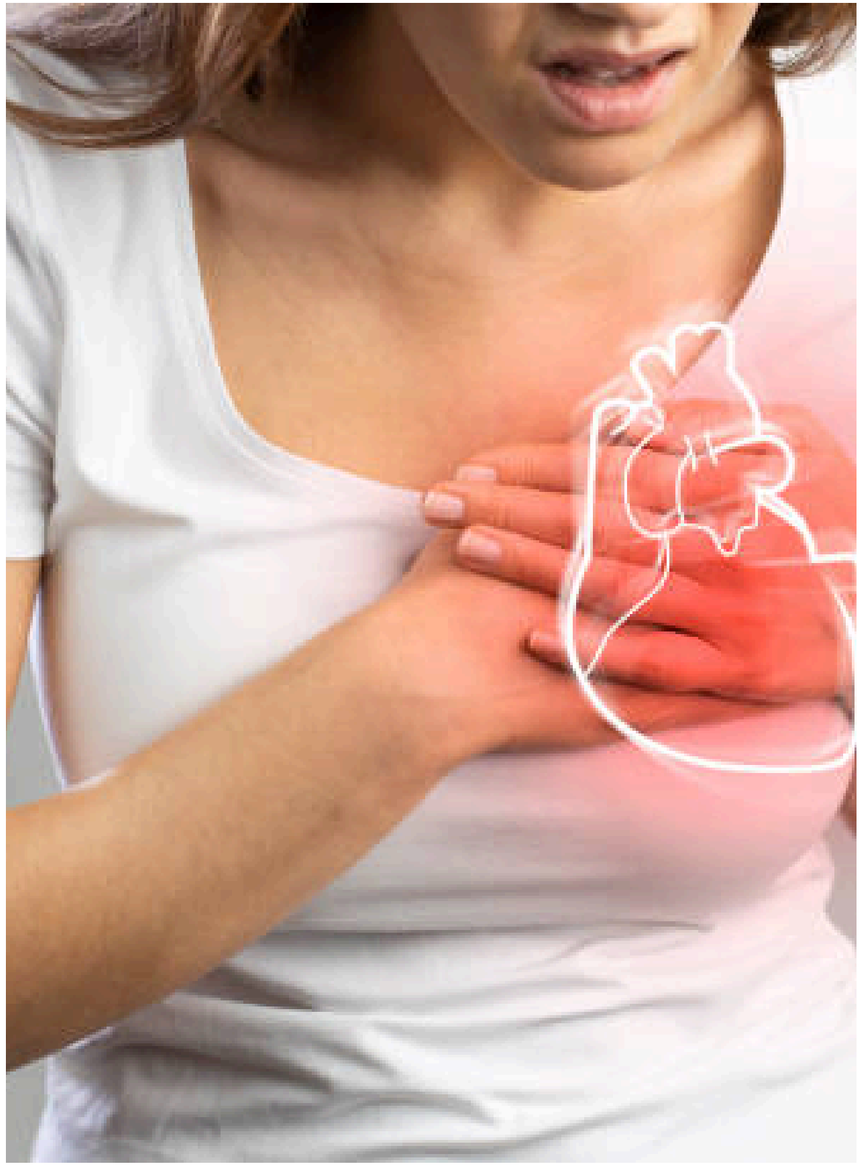
This project analyzes a health dataset to identify patterns and correlations in heart diseases. The dataset contains 445132 health records across different demographic groups

The primary challenge involves preprocessing and analyzing highly imbalanced data with mixed formats (boolean, string, numeric) to extract meaningful insights and predictive factors for various medical conditions.

Key objectives include:

- Data preprocessing categorical variables with Yes/No, boolean, and mixed text responses
- Analysis of relationships between demographics, lifestyle factors, and health outcomes
- Identification of patterns in healthcare access across different social groups
- Perform predictive modeling of health risks





Actuality and relevance

Increasing healthcare costs, aging populations, and growing health disparities, make the analysis of population health data ever more critical.

The relevance is amplified by several factors:

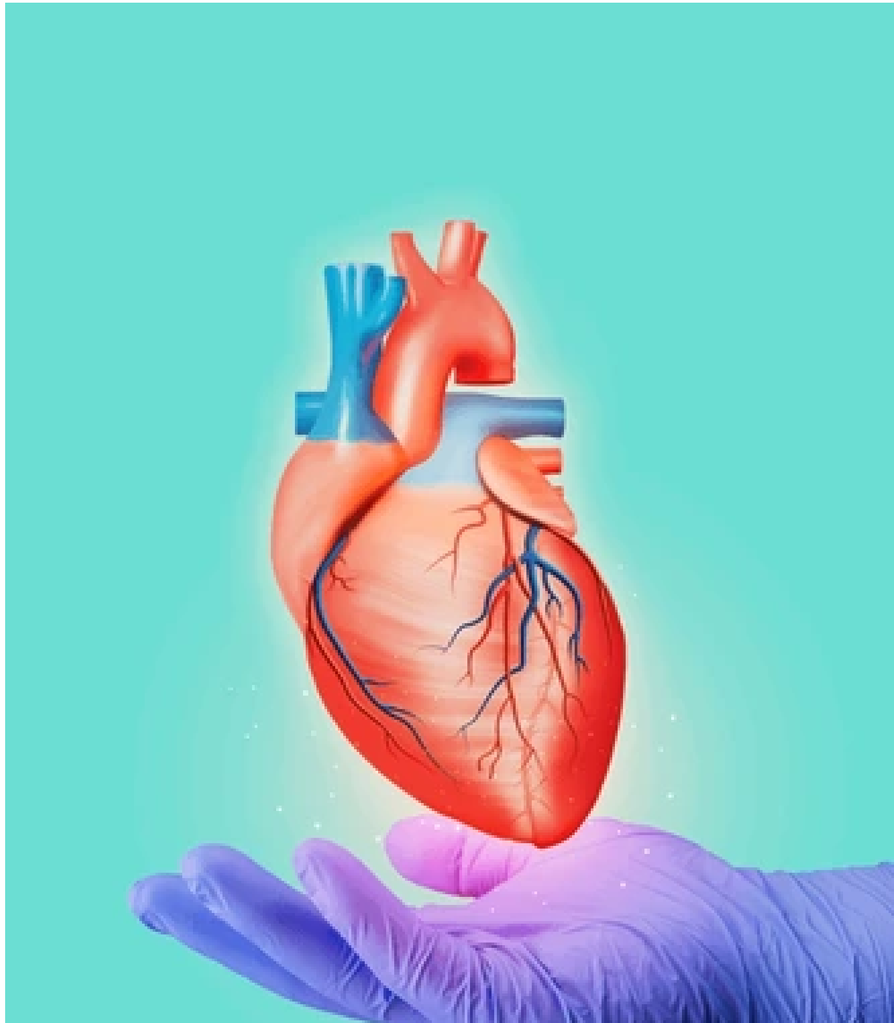
1. Healthcare worldwide shift toward wholistic care and preventive medicine, understanding health patterns is essential for effective policy development
2. Equitable healthcare access is a priority for governments and organizations in the modern political landscape.
3. Digitization of health records creates unprecedented opportunities for extraction of actionable insights from data.
4. Rising rates of chronic diseases necessitate better understanding of risk factors and preventive strategies
5. The COVID-19 pandemic exposed vulnerabilities in healthcare systems and new datasets can help understand long-term impacts.

Novelty and originality

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.



Literature overview

1. Previous studies, such as those by Xu et al. (2023), 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. Previous studies often use SMOTE (Synthetic Minority Over-sampling Technique) to generate fake data points. While effective, this can introduce noise

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

1. We integrated external GDP per Capita data to detect if regional economic prosperity correlates heart disease rates
2. 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

Data and methods

Data Source:

- heart_2022_with_nans.csv (Sourced from 2022 annual CDC survey data of 400k+ adults related to their health status) with 445132 entries,

1. Data preprocessing

- imputation of missing values
- Encoding of categorical features
- Duplicate Removal

2. Data imbalance handling

- XGBoost as our primary classification algorithm against randomforest

3. Evaluation in form of

- Recall
- F1-Score
- ROC-AUC

Results

	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

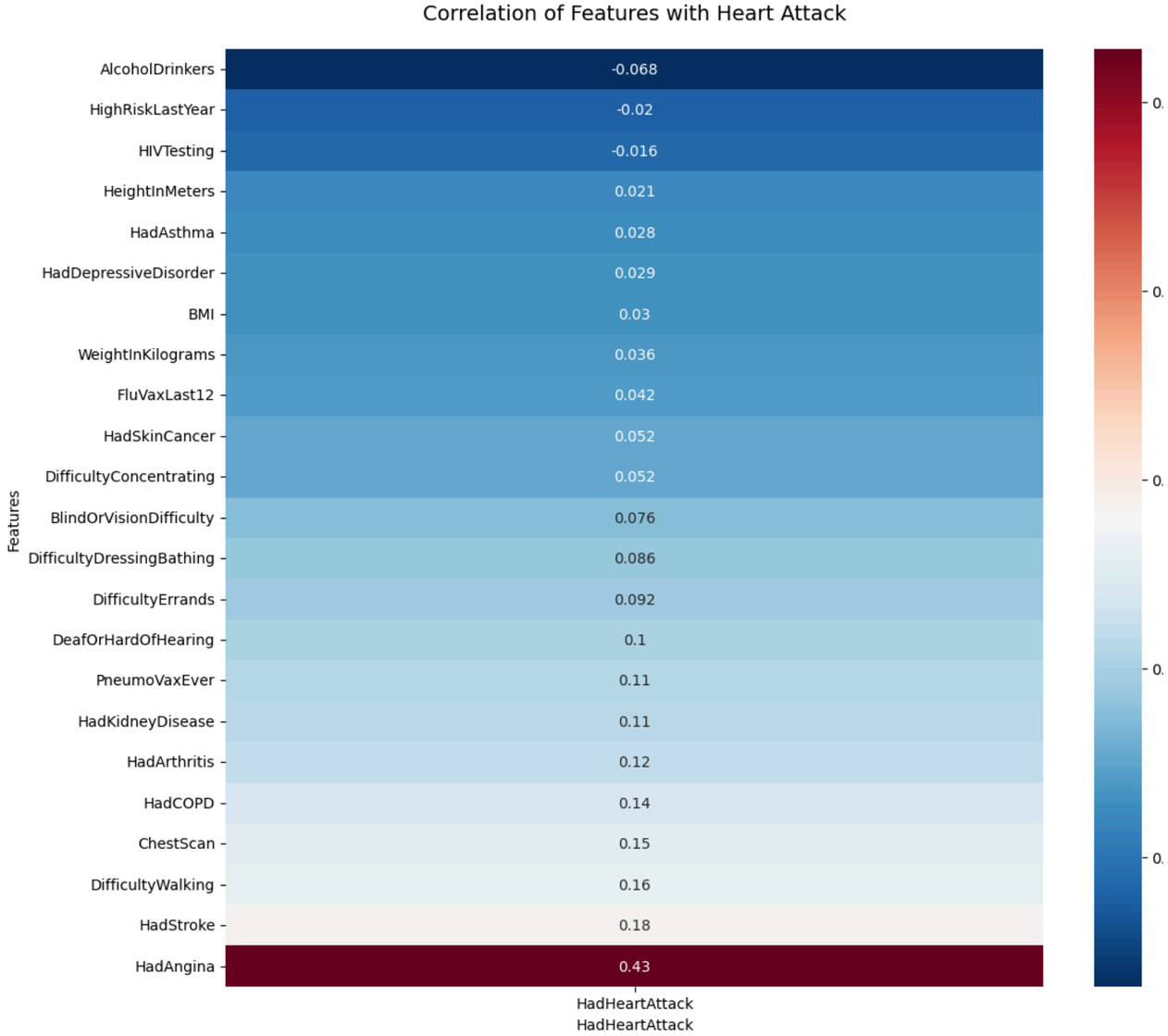
ROC-AUC: 0.8112817578254973

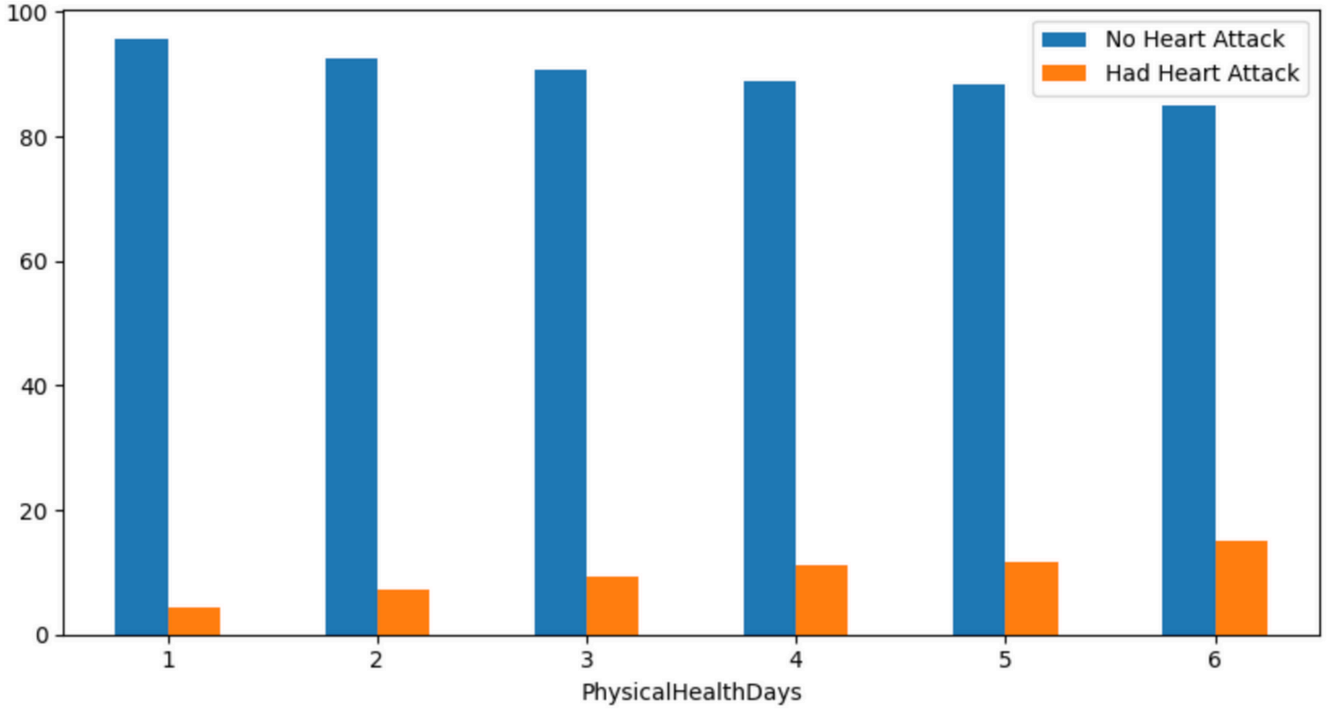
The model successfully identified 47% of all heart attack victims in the test set.

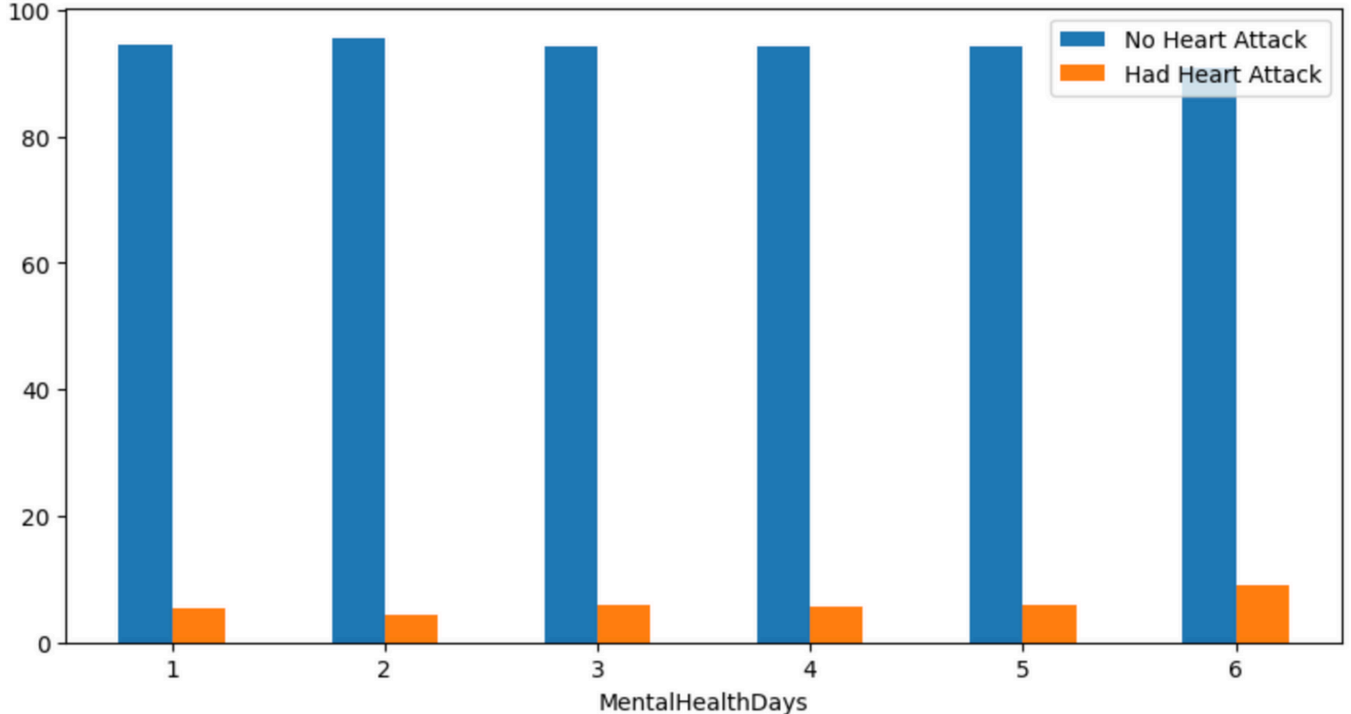
The analysis of feature importance revealed distinct drivers of cardiac risk:
Age Category
General Health
State GDP Per Capita.
BMI & Sleep

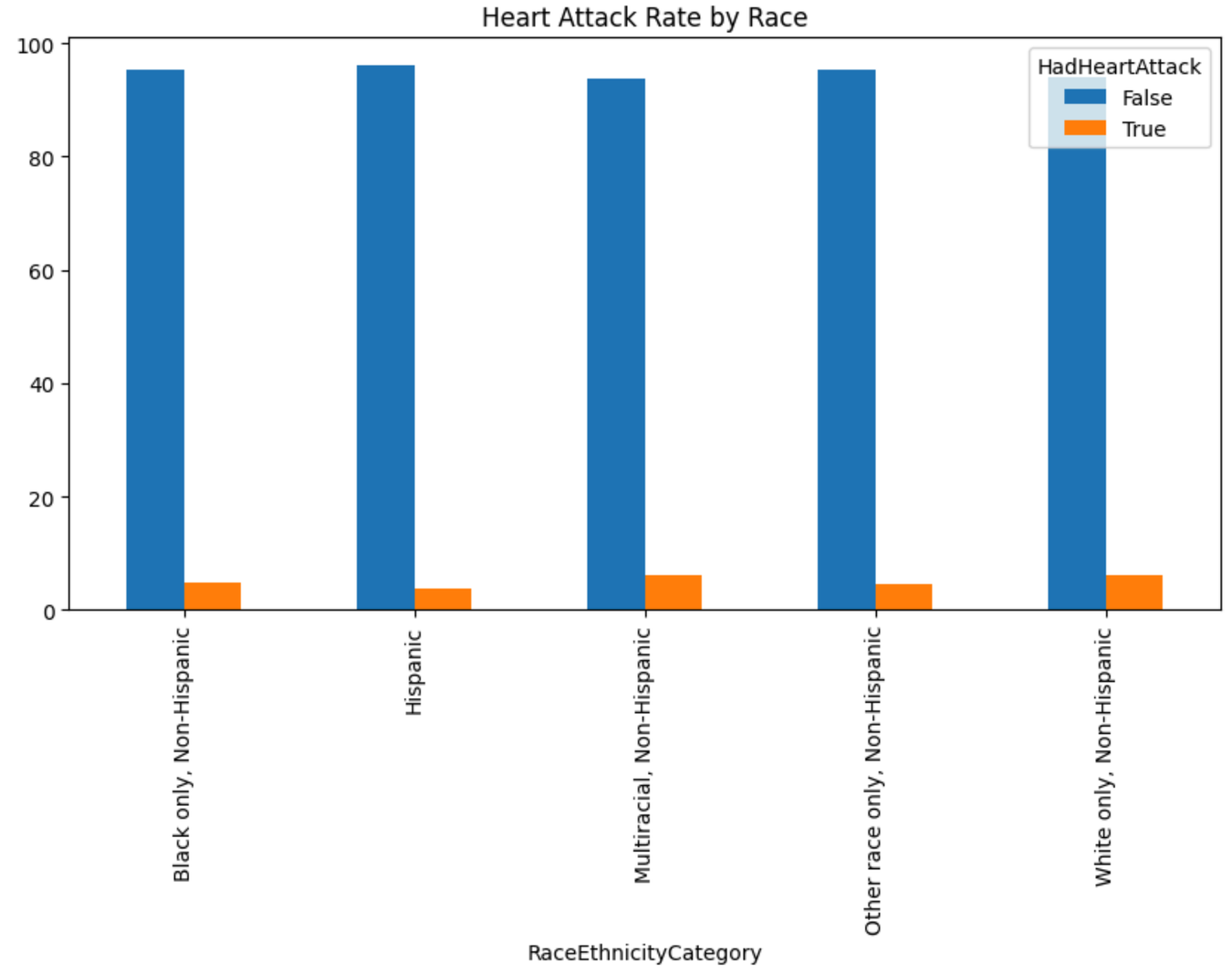
AWe trained a regression model to predict continuous BMI values based on lifestyle factors.
RMSE (Root Mean Square Error): 5.65
High error suggests that genetics or detailed diet data (missing from BRFSS) play a larger role than simple activity metrics.

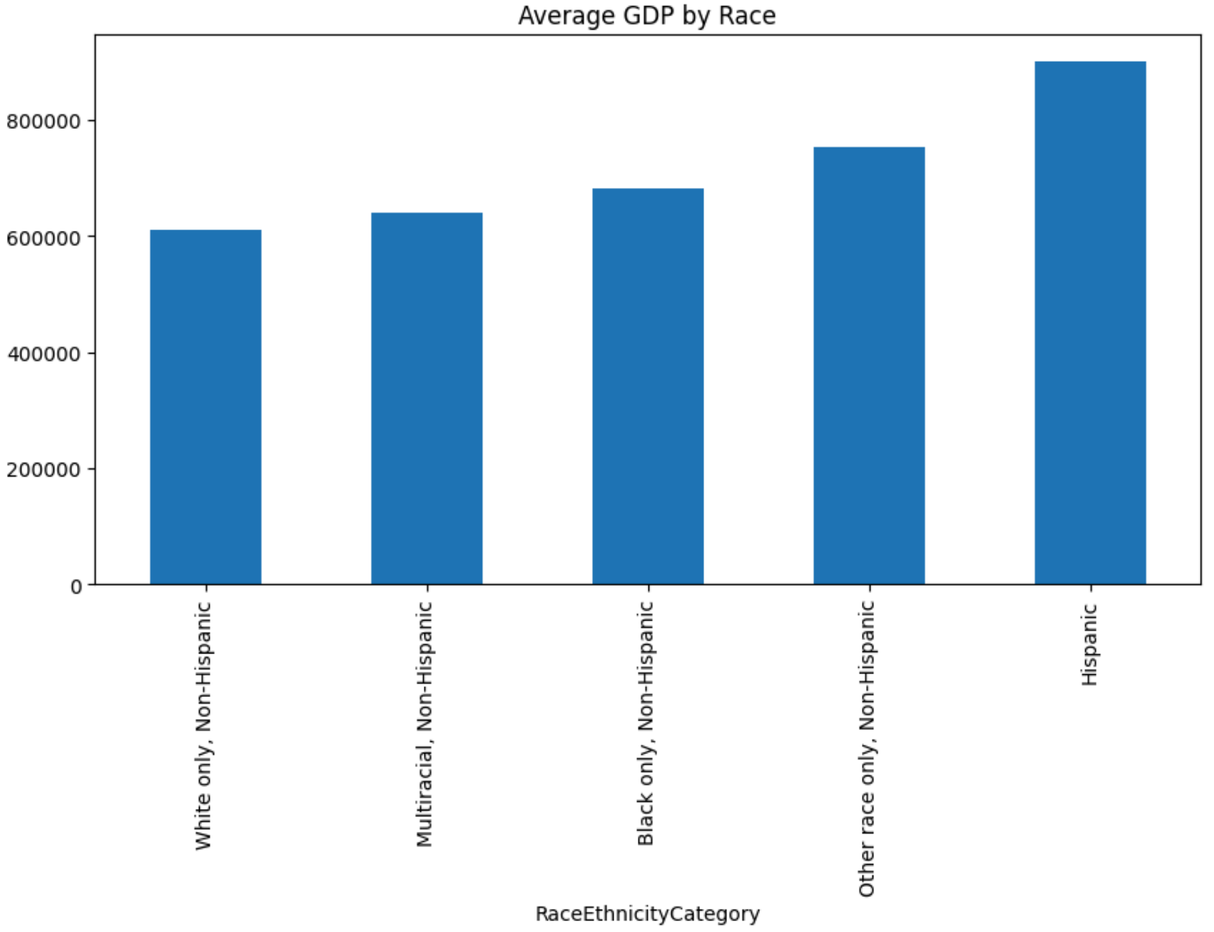
Visualizations




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

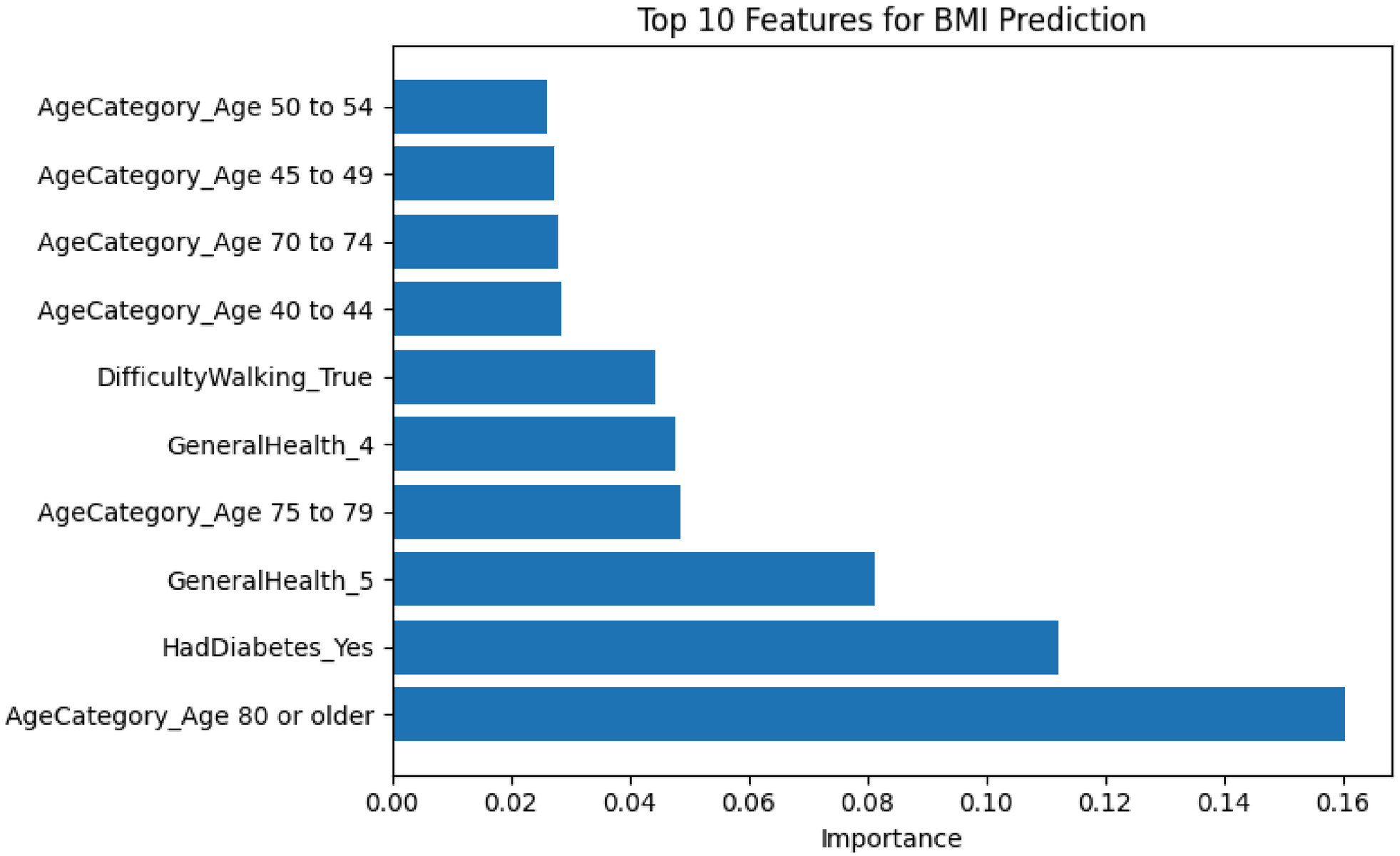

No Heart Attack
Had Heart Attack
100
80
60
40
20
0
1
2
3
4
5
6
MentalHealthDays


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


Average GDP by Race
800000
600000
400000
200000
0
White only, Non-Hispanic
Multiracial, Non-Hispanic
Black only, Non-Hispanic
Other race only, Non-Hispanic
Hispanic
RaceEthnicityCategory

	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


Top 10 Features for BMI Prediction
AgeCategory_Age 50 to 54
AgeCategory_Age 45 to 49
AgeCategory_Age 70 to 74
AgeCategory_Age 40 to 44
DifficultyWalking_True
GeneralHealth_4
AgeCategory_Age 75 to 79
GeneralHealth_5
HadDiabetes_Yes
AgeCategory_Age 80 or older
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
Importance

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:

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