Machine Learning-Powered Risk Stratification: Optimizing Resource Allocation Using Heart Attack Prediction in Hospitals

By - Chetan Sarda, Zixuan Zhu, Yinong Yao, Randeep Singh

Background, Motivation and Business Questions

Client's Ask:

Star City General Hospital is facing challenges in effectively diagnosing patients when it comes to their risk of having a heart attack. To address this issue, they have enlisted the expertise of our team of business analysts and data scientists to build a state-of-the-art model that takes in historical patient data and predicts whether the patient is at risk of heart disease.

Using this model, they have identified three major scopes for deployment:

1. Efficiently identify patients at high risk for heart disease to offer timely interventions.
2. Optimally allocate resources for the treatment of patients given the model's prediction and the patient's general health.
3. Help businesses design and implement effective wellness programs that address the most significant heart disease risk factors.

Business Value:

Annually, approximately 12 million people in the U.S. experience diagnostic errors in outpatient settings

Misdiagnoses frequently result in patient harm, with 64% of physicians reporting harm from misdiagnoses in up to 10% of cases. Also, 28% of reported diagnostic errors were life-threatening or led to death or permanent disability. According to the Improving Diagnosis in Healthcare reprot from the Institute of Medicine (IOM), diagnostic errors contribute to 10% of patient deaths and 6-17% of hospital adverse events.

Additionally, these errors account for 30% of the of annual healthcare spending in the U.S., estimated at approximately $750 billion. Patients and their families often face the most substantial financial burdens due to diagnostic errors. These costs manifest in ongoing expenses for continuous care resulting from long-term disabilities, as well as the loss of earnings attributed to either the premature death or the disability of the patient.

Effective prediction models can significantly reduce these errors, improving patient outcomes and hospital credibility. This project, focusing on heart disease diagnosis,

stands to significantly enhance patient safety and reduce the financial burdens stemming from misdiagnoses.

Some applications of this project to drive real business value for the hospital:

- Health Benefits for Corporate Wellness: The predictive model can be used to tailor health benefits and wellness initiatives to the specific needs of employees, based on their heart disease risk profiles, providing a strategic approach to corporate health management.
- Enhanced Diagnostic Accuracy and Legal Risk Mitigation: By analyzing a wide array of data points, the model aims to improve diagnostic effectiveness, thereby reducing legal risks associated with misdiagnosis.
- Resource Optimization in Patient Care: Utilizing the model to allocate medical resources more effectively, prioritizing high-risk heart disease patients and ensuring efficient treatment, while minimizing resource wastage on low-risk individuals.
- New Revenue Streams through Wellness Program Consultation: The hospital can leverage the insights from the model to offer consultancy services to employers, designing and implementing personalized wellness programs based on exploratory data analysis findings.

Problem Statements:

Risk Assessment Tools: How can healthcare providers efficiently identify patients at high risk for heart disease to offer timely interventions?

Why did we choose this project:

Heart disease (heart disease) is a group of diseases related to cardiovascular diseases, manifested by a violation of the normal functioning of the heart and may result from damage to its different components. These conditions can persist for a long time in a latent form without clinically manifesting themselves. Along with various tumours, these diseases are a significant cause of premature death in developed countries. The uninterrupted operation of the circulatory system, which consists of the heart as a muscle pump and a network of blood vessels, is a necessary condition for the normal functioning of the body.

According to the National Heart, Lung and Blood Institute in Framingham (USA), the most important factors in the development of cardiovascular disease in humans are obesity, a sedentary lifestyle, and smoking.

According to the CDC, heart disease ranks among the top causes of death for various racial groups in the US. Roughly 47% of Americans have at least one of three primary risk factors: high blood pressure, high cholesterol, or smoking. Other significant indicators include diabetic status, obesity (high BMI), insufficient physical

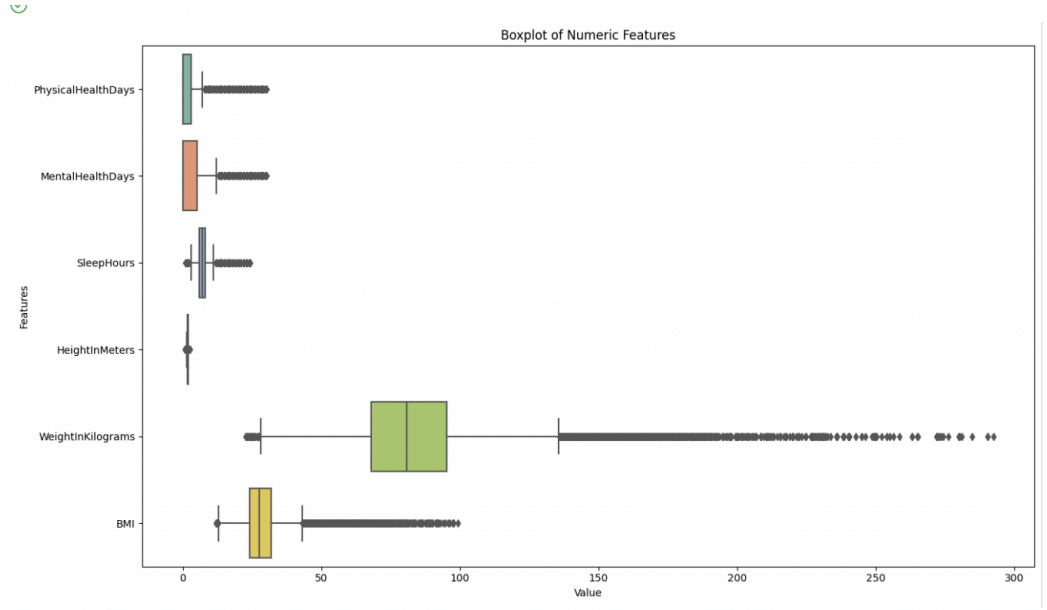
activity, or excessive alcohol consumption. Recognizing and preventing these influential factors is crucial in healthcare.

Given these alarming statistics, there is a pressing need to leverage computational advancements and machine learning methods to analyze vast datasets and discern patterns that can aid in predicting an individual's susceptibility to heart attacks. By identifying and understanding these influential factors, the goal is to develop a predictive model that can assist healthcare professionals in early intervention, personalized risk assessment, and targeted preventive measures. This approach aligns with the broader objective of improving public health outcomes and promoting proactive healthcare strategies.

Exploratory Data Analysis on Train Dataset

Train data set overview: There are 356,105 observations and 40 features in the training data set, with 34 categorical variables and 6 numeric variables.

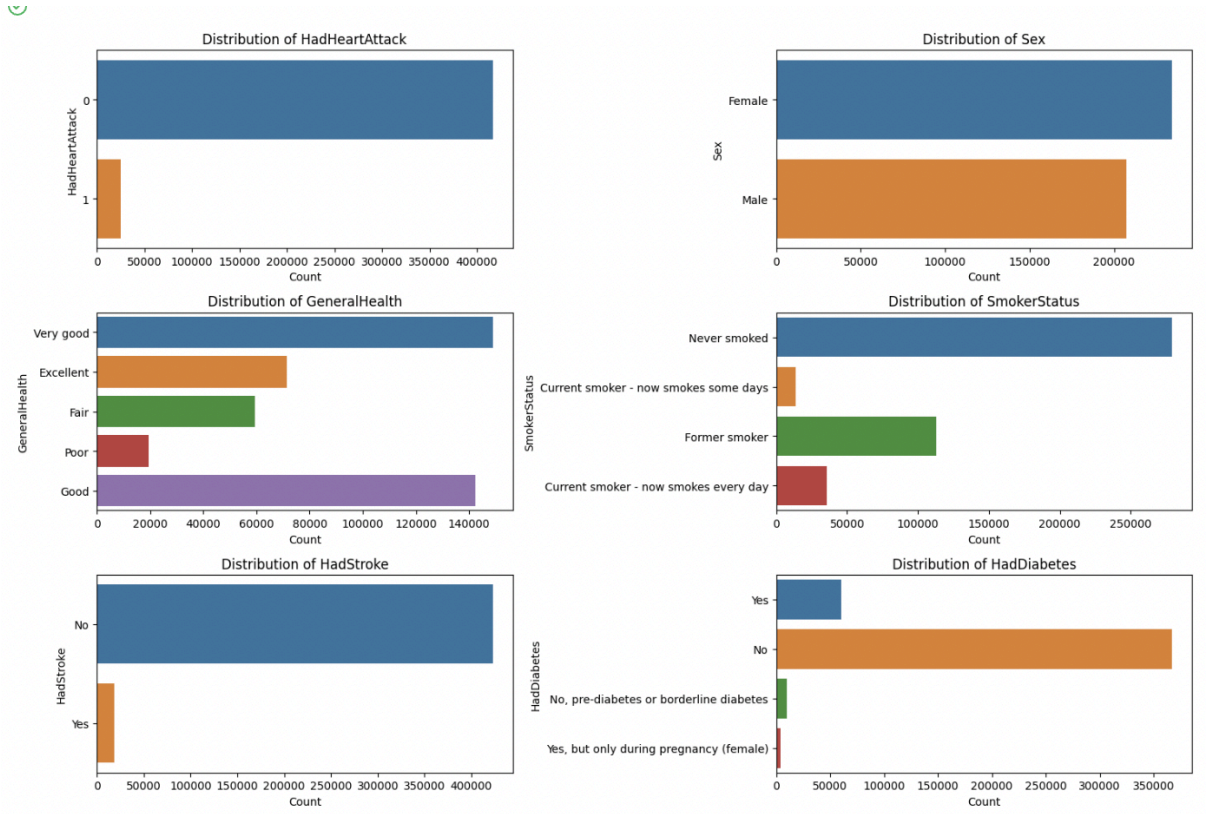
```
Int64Index: 356105 entries, 20878 to 121958
Data columns (total 40 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   State                     356105 non-null  object
 1   Sex                       356105 non-null  object
 2   GeneralHealth             355126 non-null  object
 3   PhysicalHealthDays        347377 non-null  float64
 4   MentalHealthDays          348802 non-null  float64
 5   LastCheckupTime           349403 non-null  object
 6   PhysicalActivities        355225 non-null  object
 7   SleepHours                351716 non-null  float64
 8   RemovedTeeth              347020 non-null  object
 9   HadHeartAttack            353654 non-null  object
 10  HadAngina                 352580 non-null  object
 11  HadStroke                 354852 non-null  object
 12  HadAsthma                 354687 non-null  object
 13  HadSkinCancer             353622 non-null  object
 14  HadCOPD                   354319 non-null  object
 15  HadDepressiveDisorder     353838 non-null  object
 16  HadKidneyDisease          354550 non-null  object
 17  HadArthritis              354010 non-null  object
 18  HadDiabetes               355221 non-null  object
 19  DeafOrHardOfHearing       339643 non-null  object
 20  BlindOrVisionDifficulty   338864 non-null  object
 21  DifficultyConcentrating   336757 non-null  object
 22  DifficultyWalking         336930 non-null  object
 23  DifficultyDressingBathing 337013 non-null  object
 24  DifficultyErrands         335608 non-null  object
 25  SmokerStatus              327768 non-null  object
 26  ECigaretteUsage           327589 non-null  object
 27  ChestScan                 311332 non-null  object
 28  RaceEthnicityCategory     344807 non-null  object
 29  AgeCategory               348861 non-null  object
 30  HeightInMeters            333257 non-null  float64
 31  WeightInKilograms         322474 non-null  float64
 32  BMI                       317042 non-null  float64
 33  AlcoholDrinkers           318820 non-null  object
 34  HIVTesting                303102 non-null  object
 35  FluVaxLast12              318417 non-null  object
 36  PneumoVaxEver             294362 non-null  object
 37  TetanusLast10Tdap         290112 non-null  object
 38  HighRiskLastYear          315546 non-null  object
 39  CovidPos                  315445 non-null  object
dtypes: float64(6), object(34)
```



The boxplot reveals general distribution about the numeric features in the dataset:

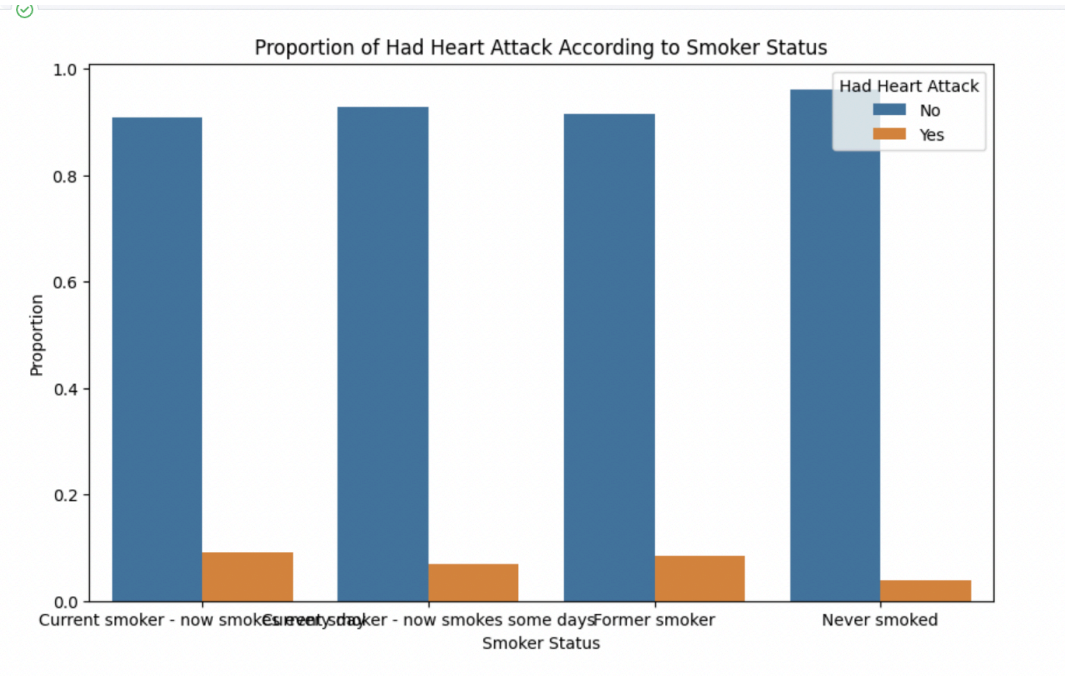
- **PhysicalHealthDays** and **MentalHealthDays** show significant outliers, indicating some individuals have many unhealthy days.
- **SleepHours** is evenly distributed, with a normal range for most but some extreme cases.
- **HeightInMeters** has a normal distribution without notable outliers.
- **WeightInKilograms** and **BMI** are right-skewed with outliers, suggesting a subset of the population is significantly heavier.

These insights can inform data preprocessing steps, such as outlier handling and feature scaling, to improve model performance.



- The bar charts show a clear data imbalance across some of the categorical features. The target variable, 'HadHeartAttack,' exhibits a severe imbalance with a significantly larger count of 'No' instances compared to 'Yes.'
- 'GeneralHealth' has a skewed distribution with 'Very good' and 'Good' being the most common.
- 'Sex' is more balanced.
- 'SmokerStatus' is imbalanced, dominated by 'Never smoked.'
- 'HadStroke' is heavily skewed towards 'No.'
- 'HadDiabetes' shows a major imbalance favouring 'No.' This imbalance may lead to biased model predictions, which should be addressed at the preprocessing part of the classification model building using methods like oversampling.

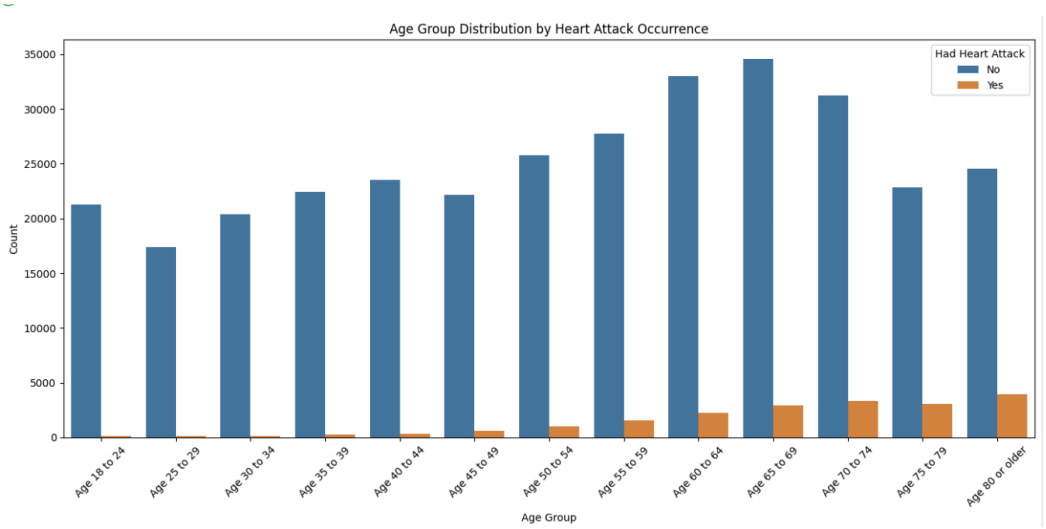
Categorical features visualization



The plot included different categories of smoking status such as current smokers, former smokers, individuals who have never smoked, and unknown status. There appeared to be a noticeable number of heart attack cases among current smokers. This suggests a potential link between active smoking and an increased risk of heart attacks.

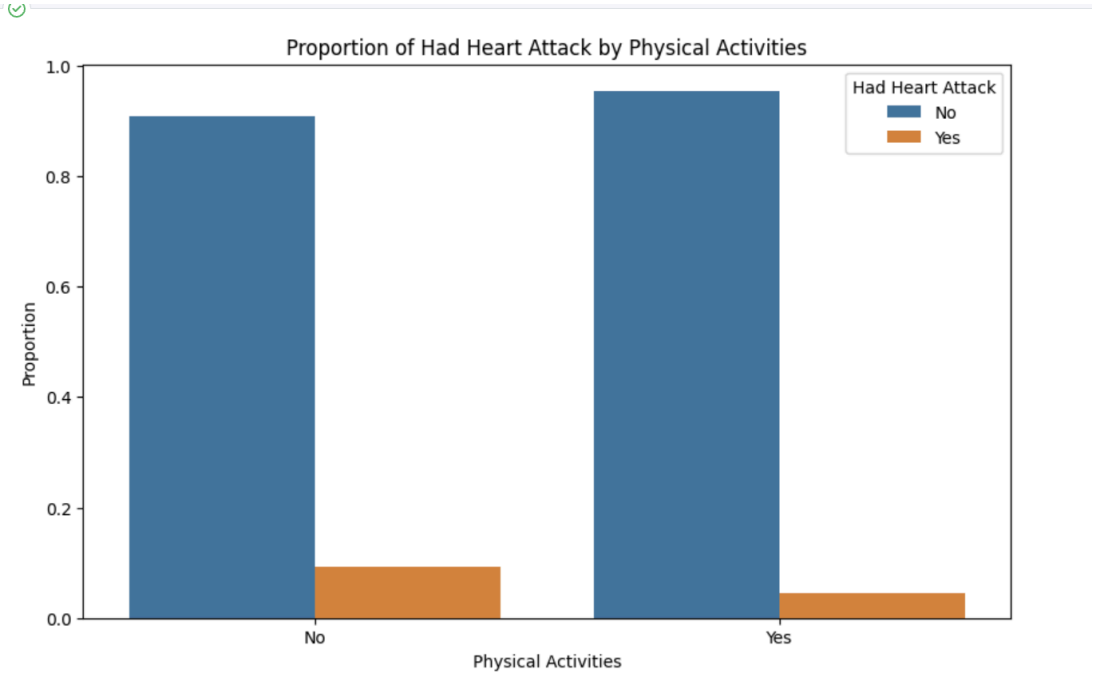
Former smokers group also showed a significant number of heart attack cases. It's important to note that former smokers might have quit recently or many years ago, and the duration since quitting can influence heart disease risk. The incidence of heart attacks in never smokers group was relatively lower compared to current and former smokers, implying a possible lower risk of heart disease among those who have never smoked

The plot indicates a correlation between smoking status and the occurrence of heart attacks. Current and former smokers showed a higher number of heart attack cases, aligning with medical research that identifies smoking as a significant risk factor for heart disease. The data supports the idea that smoking cessation may be beneficial in reducing heart disease risk, as seen in the difference between current and never smokers.



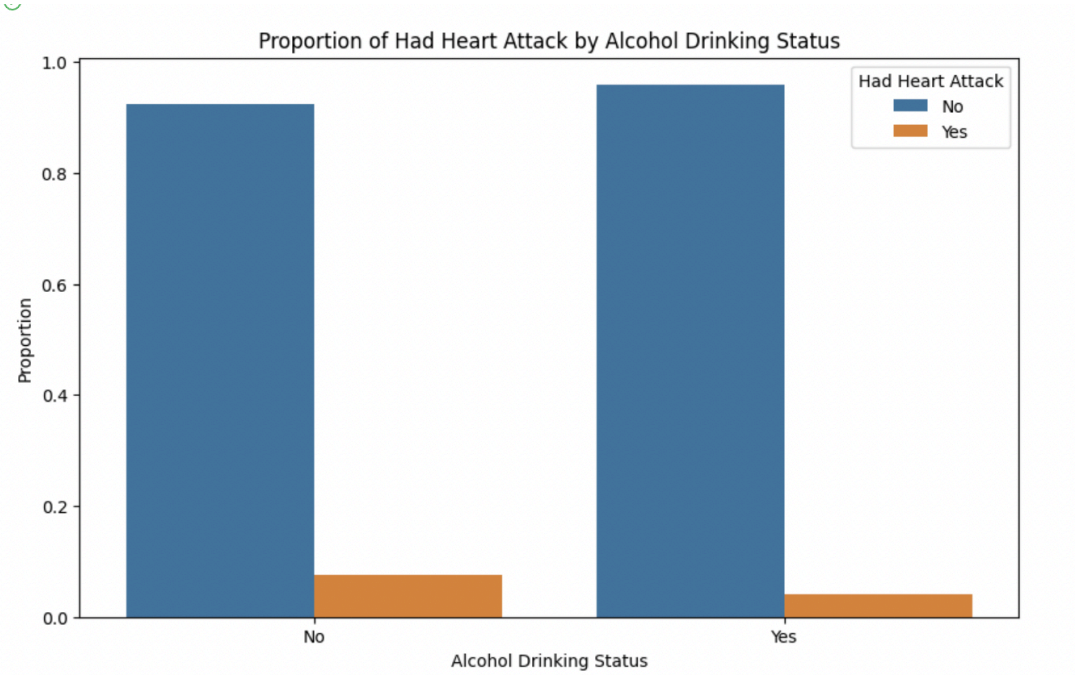
The data includes various age groups, sorted for clarity. There is a noticeable trend where the number of heart attack cases increases with age. Older age categories show a higher count of heart attack cases compared to younger ones. In the younger age groups, the number of heart attack cases is relatively low. There's a significant increase in heart attack cases in the older age groups, which aligns with the general medical understanding that heart attack risk increases with age.

The analysis revealed a mix of string and float types in the 'AgeCategory' column, with a small number of float values likely representing missing or improperly recorded data. Age is a crucial factor in predicting heart disease, as evidenced by the increase in heart attack cases with advancing age categories.

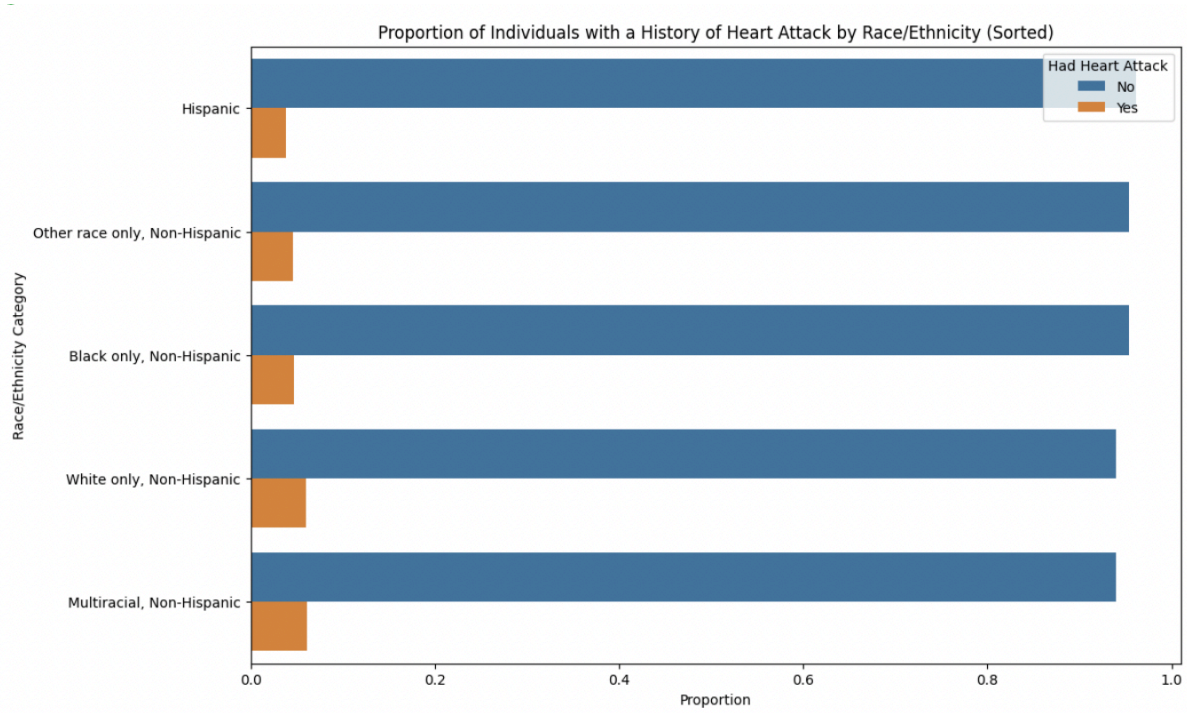


The plot categorizes individuals based on whether they engage in physical activities. Among those who engage in physical activities, the number of heart attack cases is relatively lower compared to those who do not.

There's a noticeable increase in the number of heart attack cases among individuals who do not engage in physical activities. The plot suggests that being physically active might be associated with a lower occurrence of heart attacks. This aligns with general health advice that regular physical activity can reduce the risk of heart disease. Physical activity status could be a significant predictor in a model for heart disease.
This visualization supports its inclusion and consideration in predictive analytics. Physical activity is a key lifestyle factor that can influence overall health, including heart health. While the plot shows a correlation, it's important to consider other lifestyle and health factors that might also play a role in heart disease risk.



Alcohol Drinking Status by their alcohol drinking status, which is binary (0 or 1). The color distinction within the bars indicates the number of individuals who have had a heart attack ('1') versus those who have not ('0'). The majority of individuals, as represented by the taller bar, do not consume alcohol (status '0'), and among them, a smaller proportion has had a heart attack. Conversely, the shorter bar indicates that among those who do consume alcohol (status '1'), there is also a proportion that has had a heart attack, albeit smaller in absolute numbers compared to non-drinkers. The chart suggests that within this dataset, a smaller number of people report consuming alcohol. Among both drinkers and non-drinkers, there are individuals who have experienced a heart attack, with the prevalence apparently lower in the drinking group.



This chart illustrates the proportion of individuals who have had a heart attack within different racial and ethnic categories, sorted by increasing proportion of those who had a heart attack. This plot shows that Multiracial has the highest proportion of people who have a heart attack while White only follows closely in second place.

Data and Statistical Question

We propose to do a Chi-Square Test of Independence to determine if the difference between the distribution of heart attack occurrences across different racial groups is statistically significant.

- Null Hypothesis (H_0): There is no association between race/ethnicity and the occurrence of heart attacks. This implies that the proportion of heart attack occurrences is the same across different racial/ethnic groups.
- Alternative Hypothesis (H_1): There is an association between race/ethnicity and the occurrence of heart attacks. This suggests that the likelihood of having a heart attack varies across different racial/ethnic groups.

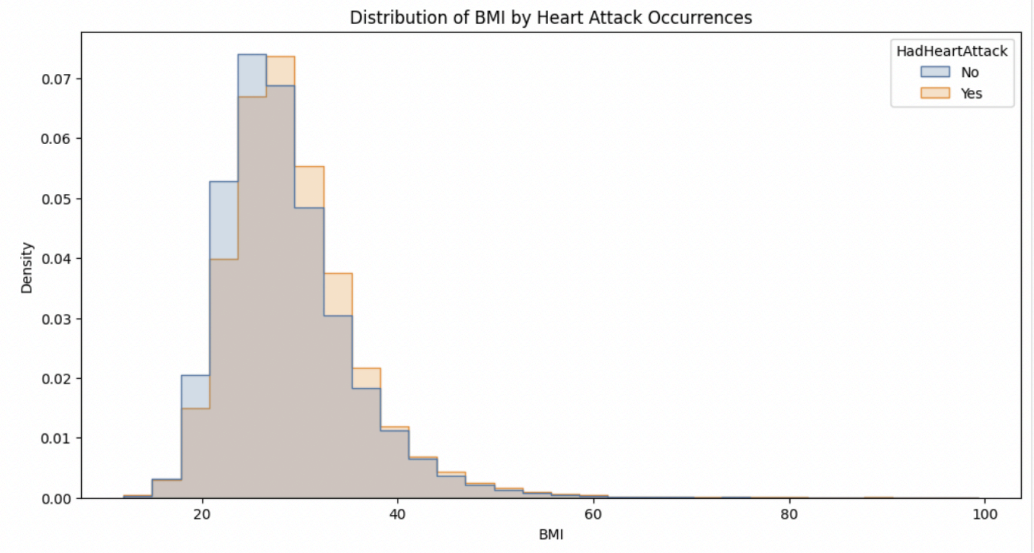
```python
from scipy.stats import chi2_contingency

# Performing the Chi-Square Test of Independence
chi2, p_value, _, _ = chi2_contingency(contingency_table_race)

# Calculating Cramér's V statistic
n = contingency_table_race.sum().sum()  # Total number of observations
min_dim = min(contingency_table_race.shape) - 1  # Minimum dimensionality
cramers_v = np.sqrt(chi2 / (n * min_dim))

chi2, p_value, cramers_v


```

```
(403.21850908798683, 5.608833261682354e-86, 0.03430455837894008)
```

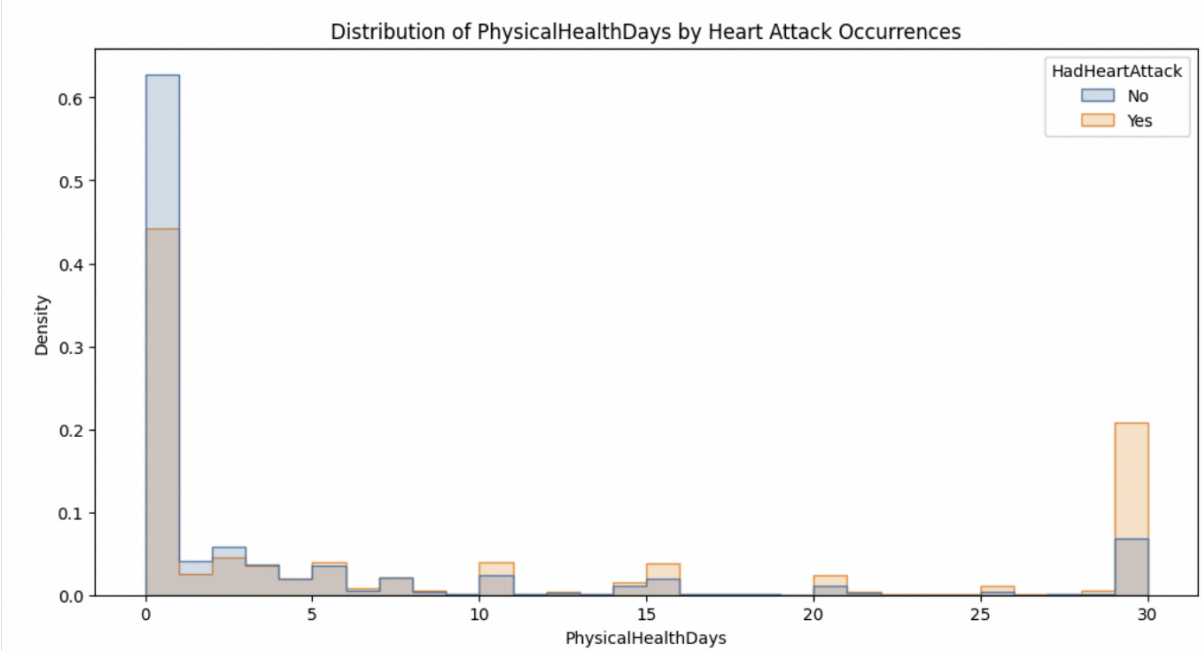
- The very low p-value practically 0 when considering standard significance levels like 0.05 or 0.01) suggests that there is a statistically significant association between race/ethnicity and the occurrence of heart attacks. This means that the likelihood of having a heart attack is not distributed the same across different racial and ethnic groups.

- The value of approximately 0.033 is quite small, indicating that while the association is statistically significant, the strength of this association is weak. This suggests that race/ethnicity, though related to heart attack occurrences, is not a strong predictor.

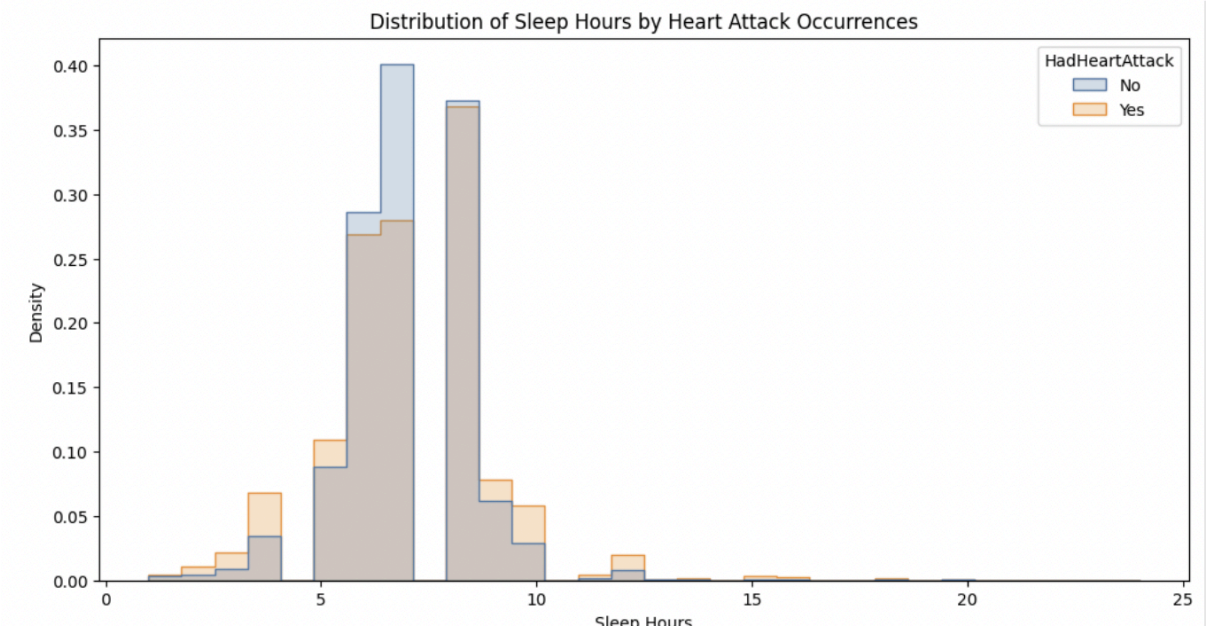
EDA on Numeric Features:



This histogram indicates differences in BMI distribution between those who have had a heart attack and those who have not. The distribution for those who had a heart attack tends to have a higher BMI (>30). This suggests that BMI could be a relevant feature for predicting heart attack occurrences because it appears that individuals with a higher BMI have a higher density of heart attack occurrences.



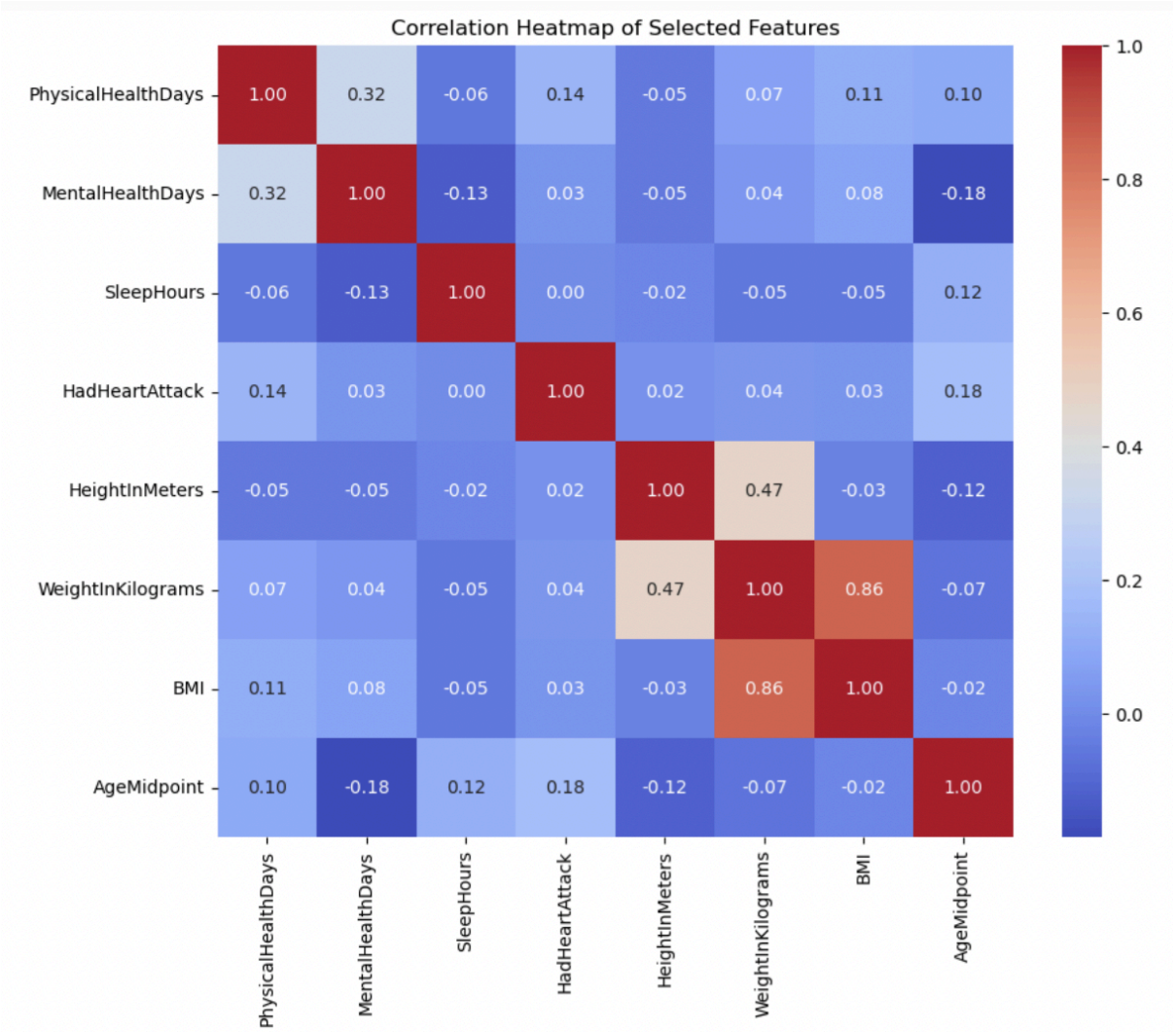
This histogram indicates that individuals who had more days in poor physical health ("PhysicalHealthDays") also had a higher incidence of heart attacks. This feature shows a clear difference in distribution between those who had a heart attack and those who did not, especially at 0 and 30 physical health days, making it a relevant factor for predicting heart attack in the model.



The histogram compares the distribution of sleep hours for individuals who have had a heart attack versus those who have not. Both groups show a similar pattern, with the highest density around 7-8 hours of sleep. There's no distinct difference between

the two groups across different amounts of sleep, therefore, sleep hours do not appear to be a good predictor of heart attack occurrence.

In general, the above three numeric predictors (**'BMI', 'PhysicaHealthDays, 'Sleep Hours''**) are suggested to be selected for the final classification model.



The correlation heatmap shows which features may affect the likelihood of having a heart attack.

HeightInMeters and **WeightInKilograms** have very low positive correlations (0.02 and 0.04, respectively) with HadHeartAttack. However, these features are highly correlated with **BMI** (0.47 and 0.86, respectively), which suggests it should not be used in conjunction with weight and height to prevent data redundancy.

Method & Results

Building the model:

DummyClassifer
The Dummy classifier gave the following accuracies:

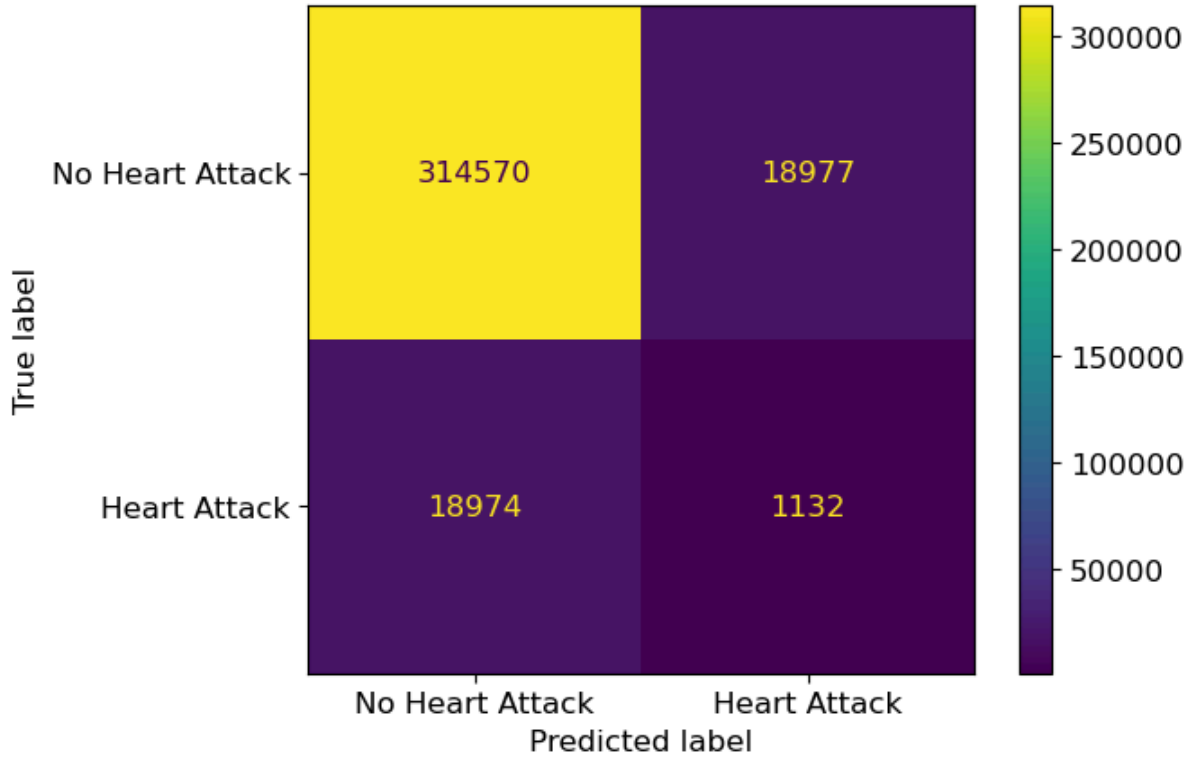
Training score: 0.8926885958835356
Test score: 0.8939647567127378

The model shows high accuracy, yet accuracy alone is insufficient for our heart attack prediction task. Accuracy does not fully capture our need to precisely identify individuals at high risk of heart attacks.

Scoring Metric Selection:

Given the critical nature of heart attack predictions, we prioritize minimizing false negatives to ensure that individuals at risk are not overlooked. Therefore, we use the ROC AUC metric, which helps us achieve a balance by effectively measuring the model's ability to distinguish between those who are at risk and those who are not, aiming for a high true positive rate while keeping false positives low.

Confusion Matrix for Dummy Classifier:



roc_auc: 0.49970353545305235

Feature pre-processing:

Preprocessing steps-
Dropped rows with missing HadHeartAttack values and converted this column from 'Yes'/'No' to 1/0.
Converted categorical features to 'category' data type for better processing.

Feature Engineering:
Identified ordinal features (like GeneralHealth, AgeCategory) and applied specific mappings to features such as HadDiabetes, CovidPos, TetanusLast10Tdap.

Used a ColumnTransformer to apply different preprocessing steps:

Numerical features: Imputed missing values with the median and scaled.
Categorical features: Imputed missing values with the most frequent category and applied one-hot encoding.
Ordinal features: Imputed missing values with the most frequent category, encoded them based on predefined order, and scaled.

We chose undersampling mainly due to concerns about overfitting and computational efficiency. The dataset is large, making oversampling impractical because it would significantly increase the data size and processing time. Also, wanted to avoid introducing synthetic data that could make my model overfit. By undersampling, this method efficiently balanced the classes without adding complexity or risking the model's generalizability on unseen data.

Feature Selection:

```
feature_selector = RFECV(ExtraTreesClassifier(n_estimators=100, random_state=42, n_jobs=-1, class_weight = 'balanced',
                                              min_impurity_decrease = 0.01), min_features_to_select=int(X_train_preprocessed.shape[1]*0.40),
                         step=7, cv=5, scoring=scoring_metric_chosen)

```

We used feature selection to pinpoint the most relevant features for predicting heart attacks, aiming to improve model accuracy and reduce training time. RFECV with an ExtraTreesClassifier was chosen because it's robust and handles both feature importance and interactions well.

ExtraTreesClassifier can quickly evaluate the importance of each feature due to its ensemble approach, offering a good balance between performance and computational efficiency. This method helped eliminate irrelevant features early on, focusing on the most impactful ones for the final model.

Model selection:

Candidate models:

LogisticRegression_balanced: Linear model with class weights balanced for imbalanced data. Max 5000 iterations, uses all CPU cores.

LogisticRegression: Similar to balanced version without class weight adjustment. Max 5000 iterations, all CPU cores used.

KNeighborsClassifier: Non-linear, votes from k nearest neighbors. Utilizes all CPU cores.

DecisionTreeClassifier: Builds decision trees.

SVC: Support Vector Classifier with balanced class weights, high iteration limit for convergence, addresses imbalanced data.

GaussianNB: Applies Bayes' theorem, suitable for continuous data and class imbalance.

RandomForestClassifier_balanced: Ensemble of decision trees with balanced class weights, parallel computation.

RandomForestClassifier: Ensemble approach without explicit class weight balancing, utilizes parallel processing.

Final Model:

	Train	Crossval	Hyperparameter	Gap
RandomForestClassifier	0.888267	0.878736	0.883258	0.005009
RandomForestClassifier_balanced	0.888267	0.878807	0.882910	0.005358
DecisionTreeClassifier	0.815204	0.717271	0.851093	0.035888
GaussianNB	0.777952	0.842636	0.852107	0.074155
LogisticRegression	0.799189	0.883595	0.883596	0.084406
LogisticRegressionBalanced	0.799189	0.883685	0.883689	0.084500
SVC	0.798244	0.874038	0.882751	0.084506
KNeighborsClassifier	0.999876	0.815769	0.793079	0.206797

RandomForestClassifier - final pick because it showed strong and consistent results between the training and validation phases, indicating good generalization without overfitting.

The balanced version didn't add much benefit, so we stuck with the standard RandomForest which was slightly better.

DecisionTreeClassifier was overfitting, evident from its large gap between training and validation scores.

GaussianNB and Logistic Regression (balanced or not) were good at generalizing, as seen by their validation scores, but they were not as performant as the RandomForest.

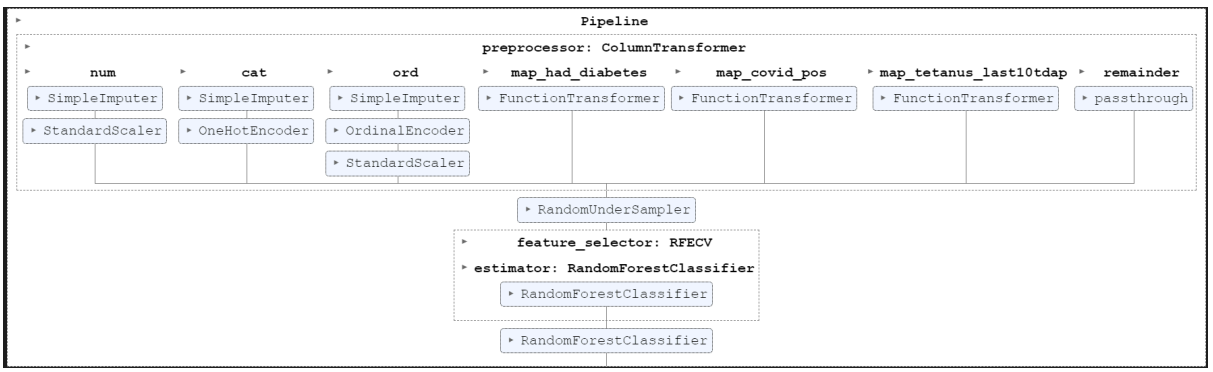
SVC was decent, but not as good as RandomForest, plus it had convergence issues that could complicate deployment.

KNeighborsClassifier performed almost too well on the training set, which was a clear sign of overfitting, making it unreliable for unseen data.

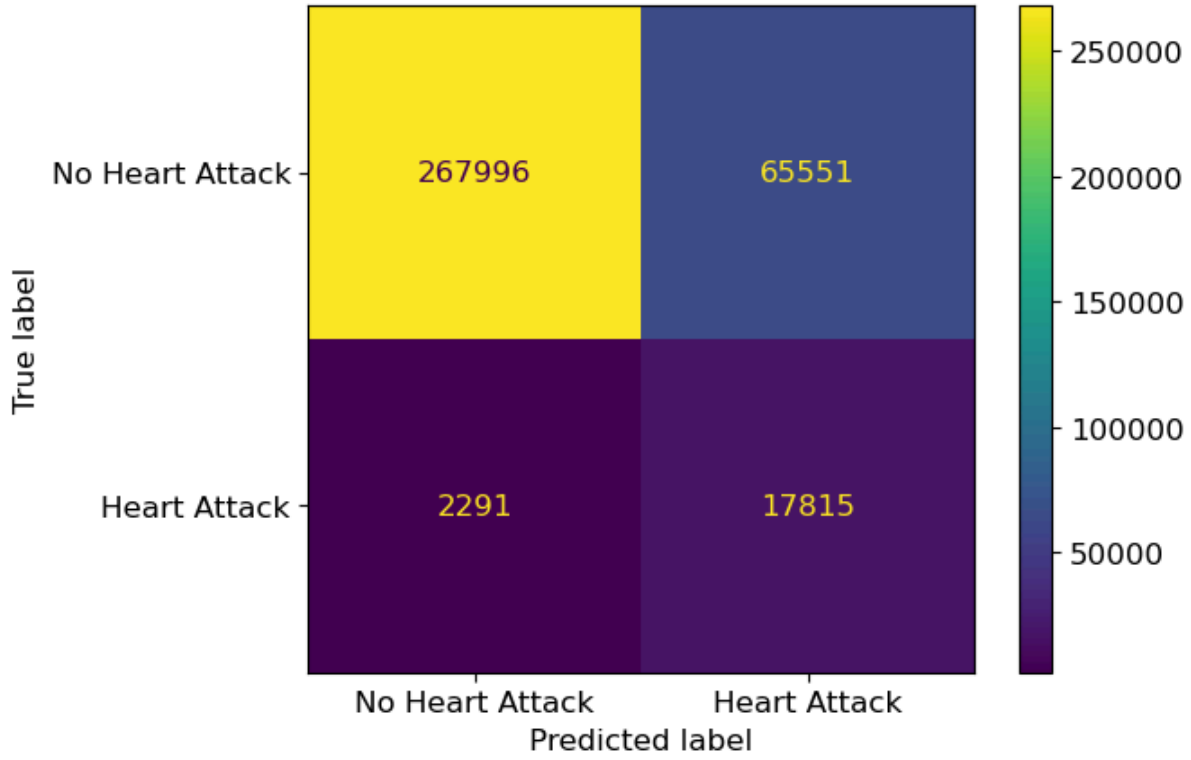
Overall, RandomForest hit the sweet spot between high performance and reliability, which is why it was my model of choice.

Final Model:

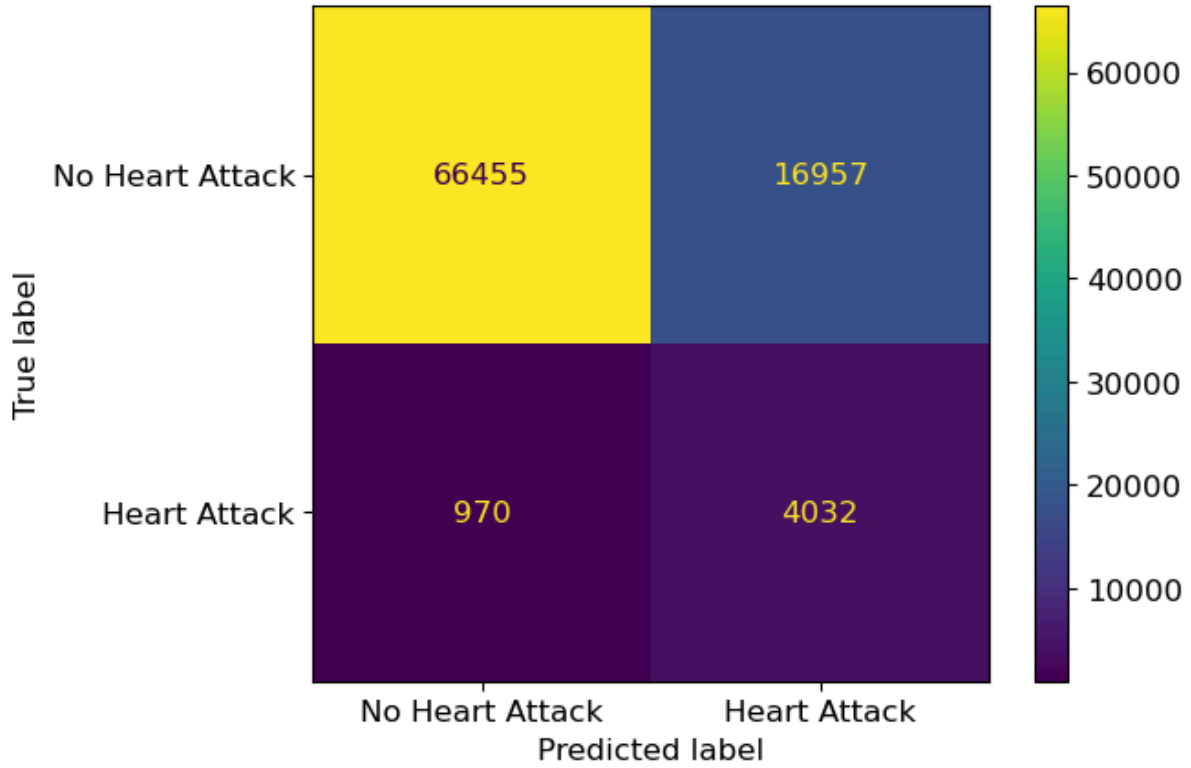
```
final_feature_selector = RFECV(RandomForestClassifier(n_jobs=-1, min_samples_split=grid_search_rfc.best_params_['classifier__min_samples_split'],
                                                      min_samples_leaf=grid_search_rfc.best_params_['classifier__min_samples_leaf'],
                                                      random_state=42), step=1, cv=3,
                               scoring=scoring_metric_chosen, min_features_to_select=int(0.4*X_train_preprocessed.shape[1]))
final_feature_selector.fit(X_train_selected, y_train_resampled)
final_pipeline = ImbPipeline([
    ('preprocessor', preprocessor),
    ('sampler', undersampler),
    ('feature_selector', final_feature_selector),
    ('classifier', RandomForestClassifier(n_jobs=-1, min_samples_split=grid_search_rfc.best_params_['classifier__min_samples_split'],
                                          min_samples_leaf=grid_search_rfc.best_params_['classifier__min_samples_leaf'],
                                          random_state=42))])
```



Confusion Matrix and Evaluation:



Training roc_auc score: 0.8447634440391152



Test roc_auc score: 0.8013927383537546

Classification Report:

```
print(classification_report(y_test, y_pred, target_names=['No Heart Attack', 'Heart Attack']))
```

✓ 0.1s

```
                precision    recall  f1-score   support

No Heart Attack       0.99      0.80      0.88     83412
   Heart Attack       0.19      0.81      0.31      5002

       accuracy                           0.80     88414
      macro avg       0.59      0.80      0.60     88414
   weighted avg       0.94      0.80      0.85     88414
```

Reflection on Performance:

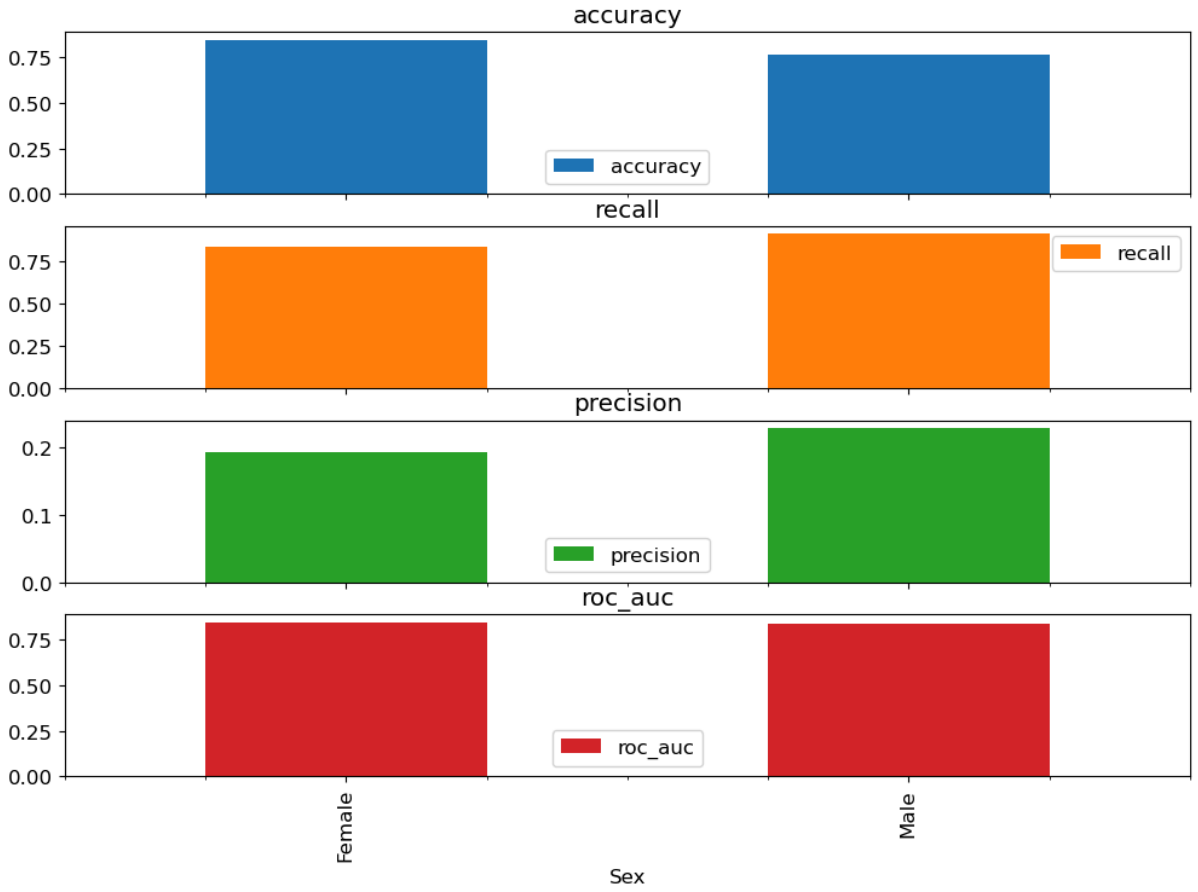
The model's ability to correctly identify whether someone will have a heart attack or not is quite good, with an ROC AUC score of about 0.80 for unseen data (test set). This means it's generally good at ranking individuals by their risk of having a heart attack, which is crucial for prioritizing who might need further medical evaluation.

The high recall of 0.81 for heart attack predictions tells us the model is very effective at identifying the majority of actual heart attack cases. In simpler terms, it doesn't miss many individuals who are at risk.

However, the precision for heart attack predictions is 0.19, indicating that when the model predicts a heart attack, it's correct about 19% of the time. This suggests we might be over-alerting, but in the context of heart attacks, we'd rather be safe and check more people than miss someone who is truly at risk.

Overall accuracy is 0.80, meaning that 80% of the model's predictions are correct. For non-heart attack predictions, the model is very reliable (precision of 0.99), meaning there's little chance of false alarms in this group.

Fairness

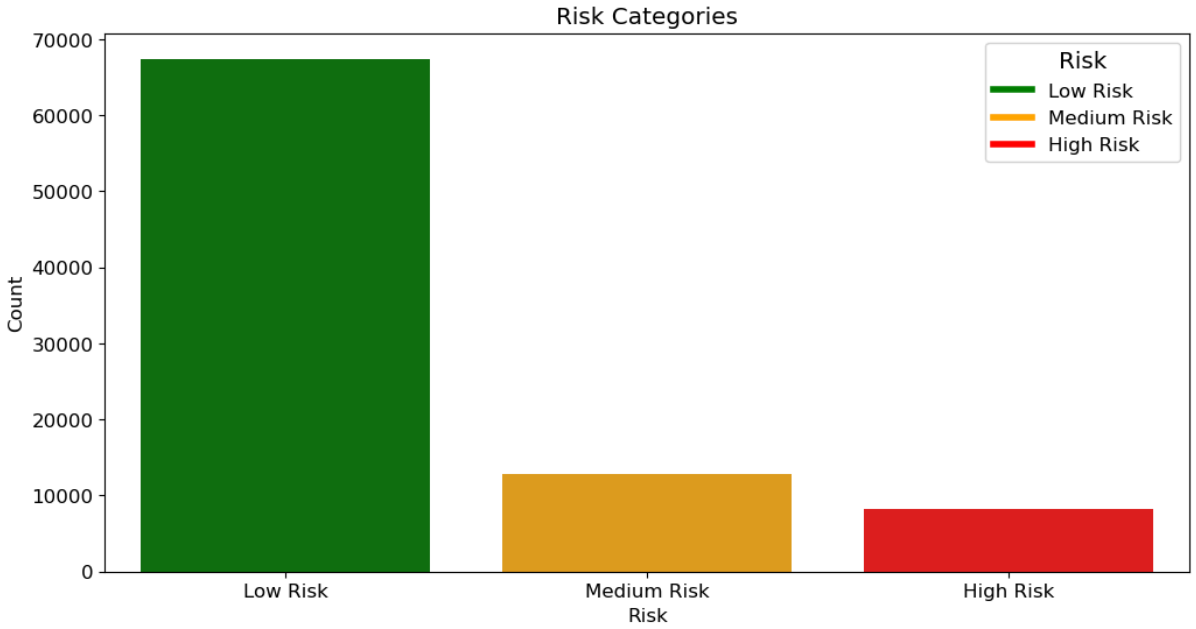


Fairness Assessment: The performance metrics—accuracy, recall, precision, and ROC AUC—are relatively balanced between females and males. This balance suggests that the model's predictive performance is consistent across genders, indicating no immediate bias based on sex. Therefore, there may not be a need for optimization specifically for fairness concerning gender.

The recall, precision, and roc_auc appear to be equal for both females and males. This is particularly important in medical predictions as it implies the model is equally good at identifying true heart attack cases across genders, minimizing the risk of missed diagnoses.

Overall, the graph suggests that the model's performance is fair and robust across different genders, which is a positive indication for the deployment of the model in diverse populations. It shows that the model's ability to predict heart attacks does not favour one gender over the other, which is crucial for ethical medical applications.

Risk Categorization:



This graph categorizes predictions into three risk levels: Low, Medium, and High Risk of heart attack.

It shows the count of individuals in each risk category.

A large number of individuals are categorized as Low Risk, a smaller number as Medium Risk, and even fewer as High Risk.

This categorization helps in prioritizing medical attention or further diagnostic testing based on risk levels.

Communication of Results and Advice

Suggestions on how to better allocate resources and prepare wellness programs based on EDA

1. Focus on High BMI and Physical Health Days: The data shows a higher incidence of heart attacks among individuals with a higher BMI and more days in poor physical health. This suggests that resources should be particularly allocated to interventions aimed at weight management and improving overall physical health. Wellness programs could include nutrition counselling, weight loss support groups, and physical activity challenges to address these issues.

2. Physical Activity Promotion: The analysis revealed a lower occurrence of heart attacks among individuals who engage in physical activities. This underscores the importance of incorporating physical activity into daily routines. Wellness programs can include gym memberships, fitness

challenges, and incentives for meeting physical activity goals to encourage more active lifestyles.

3. Address Age-Related Risks: The number of heart attack cases increases with age. Focusing on modifiable risk factors in older populations is crucial. Wellness programs for older employees can focus on heart health education, regular health screenings, and specific interventions targeting blood pressure, cholesterol levels, and other age-related concerns.

4. Alcohol Consumption Awareness: Although the data showed a smaller proportion of heart attack occurrences among alcohol consumers, it's essential to promote moderation. Wellness programs could offer education on the effects of excessive alcohol consumption on heart health and encourage responsible drinking habits.

5. Mental Health Support: Given the significance of PhysicalHealthDays, which may also relate to mental health, offering resources for mental health support can be beneficial. This could include stress management workshops, access to mental health professionals, and initiatives to reduce workplace stress.

Model Improvement Strategies:

Ensemble Learning:
We could explore combining different models to improve our predictions. This isn't just about using multiple models; it's about strategically combining them to cover each other's weaknesses.

Advanced Feature Engineering:
This involves both technical experimentation and consulting with medical experts to understand which data points might be most telling for heart attack risks.

Deep Learning:
Since we have significant data to train effectively, we can uncover complex patterns using deep learning.

References

Canvas - https://canvas.ubc.ca/courses/129201/assignments/1748562

Dataset- https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Steamlit App - https://kamilpytlak-heart-condition-checker-app-2r42q4.streamlit.app/

CDC Report - https://www.cdc.gov/heartdisease/risk_factors.htm

Statistics on Diagnostic errors - https://pinnaclecare.medium.com/the-human-cost-and-financial-impact-of-misdiagnosis-b50ead6f53f4

Code References in Notebook