Heart Attack Prediction Report

Introduction

The goal of this project is to develop a predictive model that assesses the likelihood of individuals experiencing heart attacks based on health-related parameters. Using machine learning techniques, we aim to identify key risk factors and build an accurate predictive system that can assist in early diagnosis and prevention. The success of this project will be measured by the accuracy, precision, recall, and F1-score of the predictive models. The key stakeholders for this study include healthcare professionals, researchers, and patients who may benefit from early detection.

Approach

Data Acquisition and Wrangling

The dataset used for this project was sourced from Kaggle, specifically the **Heart.csv** dataset. It contains the following features:

- **Age**: Age of the individual
- **Sex**: Gender (1 = male, 0 = female)
- **ChestPainType**: Type of chest pain experienced
- **RestingBP**: Resting blood pressure (mm Hg)
- **Cholesterol**: Serum cholesterol (mg/dl)
- **FastingBS**: Fasting blood sugar (> 120 mg/dl, 1 = true, 0 = false)
- **RestingECG**: Resting electrocardiogram results
- **MaxHR**: Maximum heart rate achieved
- **ExerciseAngina**: Exercise-induced angina (1 = yes, 0 = no)
- **Oldpeak**: ST depression induced by exercise relative to rest
- **ST_Slope**: The slope of the peak exercise ST segment
- **Target**: The dependent variable indicating the presence of heart disease (1 = heart disease, 0 = no heart disease)

After loading the dataset, several preprocessing steps were undertaken:

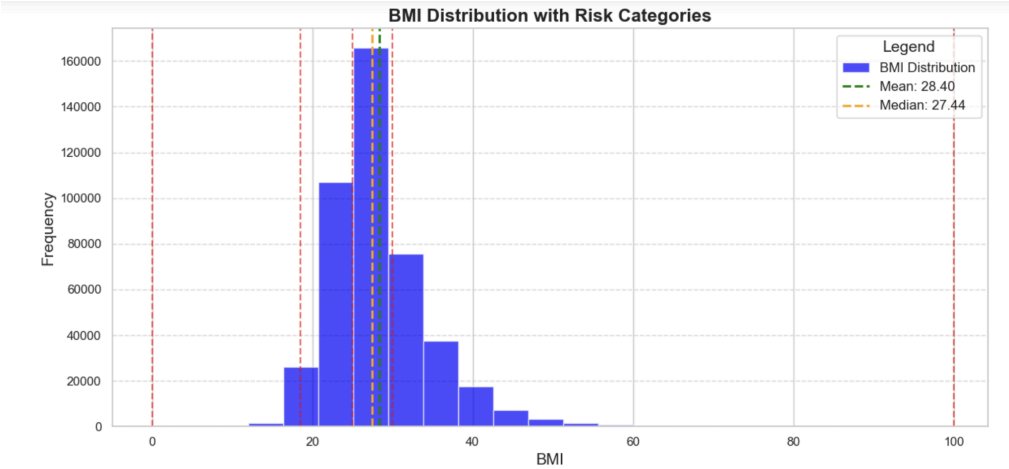
- Handling missing values (if any were found)
- Encoding categorical variables
- Normalizing numerical features
- Splitting the dataset into training and testing sets

The cleaned dataset was then stored in a Pandas DataFrame for further exploratory data analysis (EDA).

Storytelling and Inferential Statistics

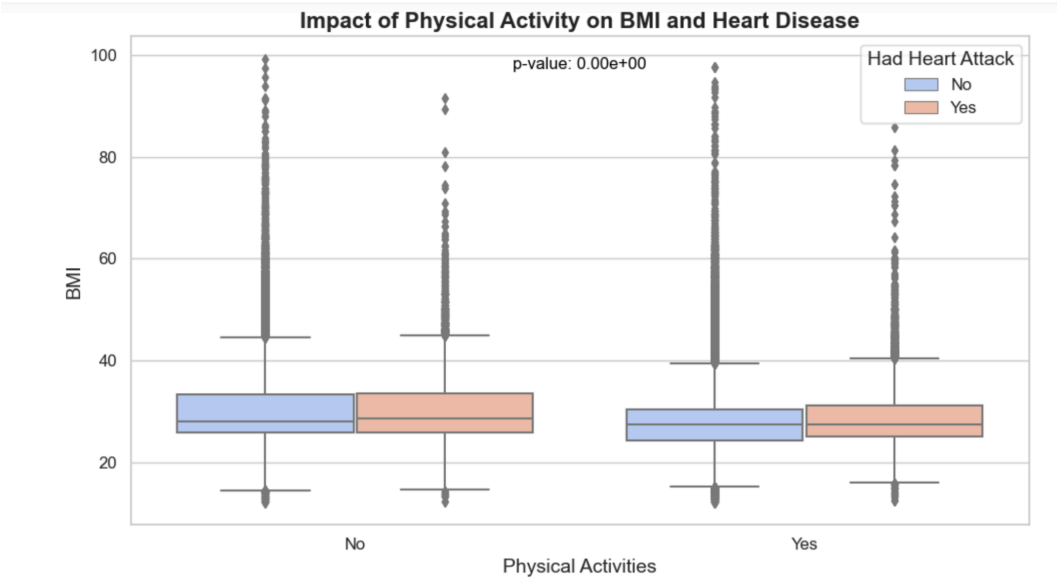
BMI Distribution with Risk Categories

This histogram provides insights into the distribution of BMI values across the dataset, divided into standard risk categories: Underweight, Normal, Overweight, and Obese. Most individuals fall in the Overweight or Obese categories, which are known risk factors for heart attacks. The chart helps identify the prevalence of high-risk BMI levels and validates its importance as a predictor in the model. By including BMI and categorizing it appropriately, the model can better account for the role of body composition in heart-related conditions.



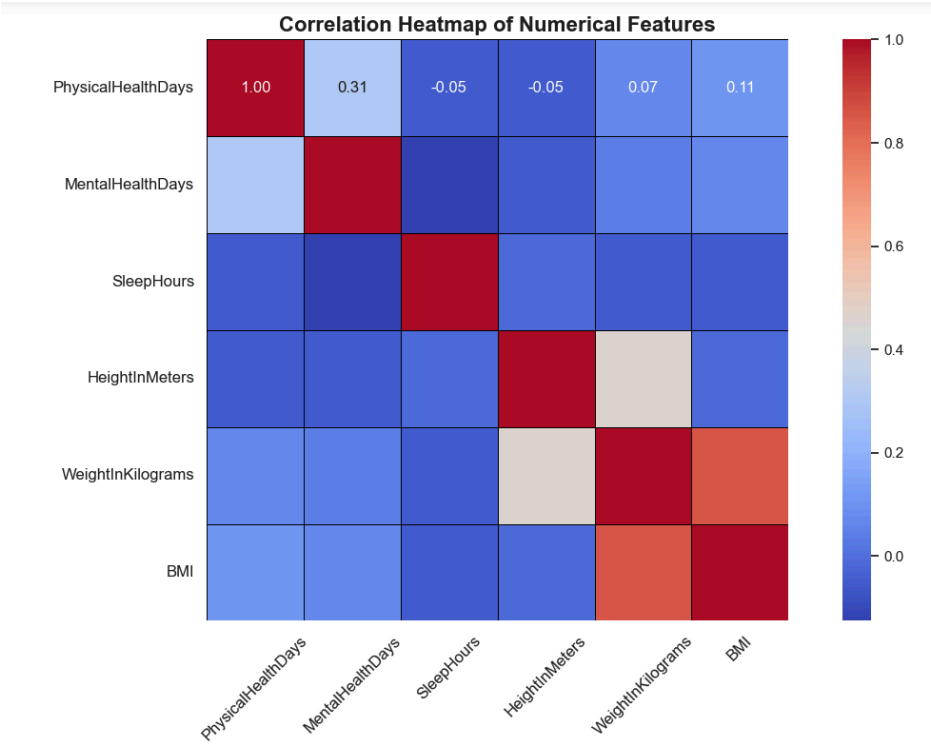
Impact of Physical Activity on BMI and Heart Disease

This boxplot shows how physical activity influences BMI and its association with heart attack occurrences. The p-value confirms that the difference in BMI between physically active and inactive individuals is statistically significant. Moreover, individuals with no physical activity tend to have higher BMI values and a greater likelihood of heart attacks, as shown by the distribution of cases. This visualization emphasizes that physical activity is an essential lifestyle factor to include in the model, as it directly impacts BMI and indirectly contributes to heart health.



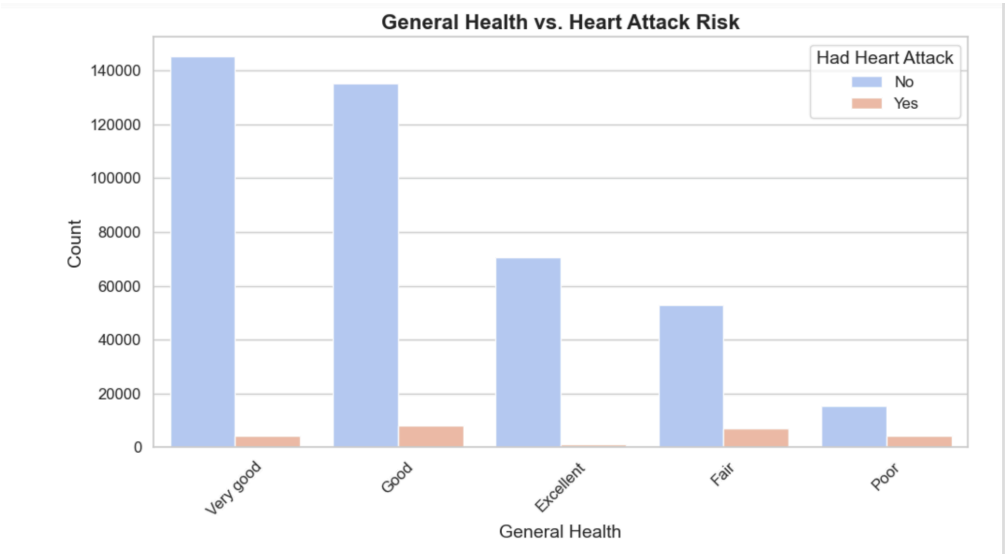
Correlation Analysis of Numerical Features

The correlation heatmap reveals the relationships between numerical features in the dataset, such as BMI, HeightInMeters, and WeightInKilograms. It highlights that BMI is strongly correlated with WeightInKilograms, which aligns with how BMI is calculated. However, it also shows weak correlations between other features (e.g., PhysicalHealthDays and MentalHealthDays). In terms of factors influencing heart attack risk, this heatmap can be a guide to focus on features with stronger relationships to the target variable (HadHeartAttack) or to remove redundant features to reduce multicollinearity.



General Health vs. Heart Attack Risk

This bar chart shows how individuals' self-reported general health categories (Very Good, Good, Excellent, Fair, Poor) relate to heart attack occurrences. The key insight here is that poorer health ratings (Fair or Poor) are associated with a higher proportion of heart attacks compared to better health ratings. This indicates that general health, as a subjective measure, is a critical factor to include in the model because it provides a direct signal about a person's overall health status and their likelihood of experiencing heart-related issues.



Baseline Modeling

Logistic Regression

The Logistic Regression classifier was trained using default hyperparameters without any optimization. It achieved an overall accuracy of **75.55%**, with the majority class (Low Risk) being predicted with high precision (98.84%) but lower recall (75.45%). The minority class (High Risk) suffered significantly, with a precision of only 11.32% and recall of 77.96%, resulting in a poor F1-score of 19.77%. This discrepancy reflects a strong bias toward the majority class, with a substantial number of True cases incorrectly classified as False.

Random Forest Classifier (Default Parameters)

The Random Forest model, trained with default parameters, achieved a high overall accuracy of **96.13%**, but completely failed to identify high-risk cases (True), with precision and recall for this class at **0%**. All 5,155 instances of the High Risk class were misclassified as Low Risk. This demonstrates the inability of the default Random Forest model to handle imbalanced data.

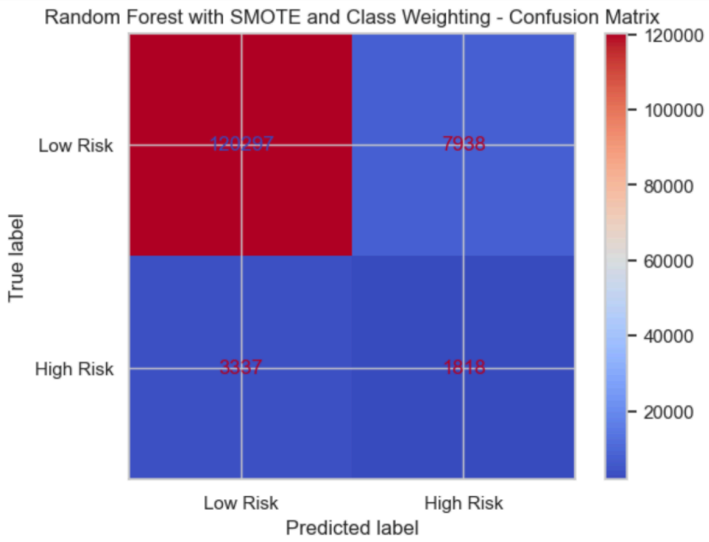
Extended Modeling

Random Forest with SMOTE and Class Weighting

To address the class imbalance, SMOTE was applied to oversample the minority class (High Risk). Additionally, class weights were adjusted to assign higher penalties to misclassifications of the minority class. GridSearchCV was used to optimize hyperparameters, prioritizing the F1-score. The best parameters were:

- max_depth: 20
- max_features: 'sqrt'
- min_samples_leaf: 1
- min_samples_split: 2
- n_estimators: 100

The optimized model achieved an overall accuracy of **92%**. For the Low Risk class, the precision was **97%**, recall was **94%**, and the F1-score was **96%**. However, for the High Risk class, the precision was **19%**, recall was **35%**, and the F1-score was **24%**. While the model performed well for Low Risk cases, it still struggled with High Risk cases, indicating room for further improvement.




```
Best Parameters for Random Forest with SMOTE: {'max_depth': 20, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min
_samples_split': 2, 'n_estimators': 100}

Classification Report:
              precision    recall  f1-score   support

    Low Risk       0.97      0.94      0.96    128235
   High Risk       0.19      0.35      0.24      5155

    accuracy                           0.92    133390
   macro avg       0.58      0.65      0.60    133390
weighted avg       0.94      0.92      0.93    133390
```

Findings

The optimized Random Forest model demonstrated improved performance compared to the baseline model but still faced challenges in identifying High Risk cases. While accuracy and F1-scores were reasonable for the Low Risk class, the High Risk class metrics revealed significant limitations in addressing the imbalanced dataset effectively.

Conclusions and Future Work

Conclusions

- The Random Forest model with SMOTE and class weighting improved classification performance for the minority class compared to the baseline models.
- However, the precision and recall for High Risk cases remained suboptimal, limiting the model's ability to generalize effectively.

Future Work

1. **Incorporating Additional Features**: Including more clinically relevant features could enhance model performance.

2. **Exploring Advanced Techniques**: Leveraging ensemble methods or deep learning models to better capture patterns in the data.
3. **Dynamic Sampling Techniques**: Investigating advanced resampling techniques to address class imbalance without introducing noise.
4. **Deploying Interpretability Tools**: Using SHAP or LIME to provide insights into model predictions, aiding healthcare professionals in understanding results.

Recommendations for Stakeholders

Based on the analysis and findings:

For Healthcare Professionals:

- Utilize the model as a supplementary tool to identify individuals at low risk of heart attacks. The high recall for the Low Risk class ensures a reliable identification of these cases.
- Exercise caution when relying on the model for identifying high-risk individuals due to the lower precision and recall in this category. Additional clinical tests or evaluations should complement the predictions.

For Researchers:

- Focus on improving the model's ability to handle imbalanced datasets by exploring advanced sampling methods or leveraging domain-specific features.
- Consider using interpretable machine learning tools (e.g., SHAP, LIME) to provide transparent insights into the model's predictions.

For Policymakers and Public Health Officials:

- Develop targeted awareness campaigns to encourage lifestyle changes, such as increased physical activity and weight management, given their strong correlation with heart attack risks.
- Fund initiatives aimed at collecting richer datasets with more balanced distributions to improve model generalizability and effectiveness.

For Patients:

- Leverage insights from the model as part of preventive care plans, especially for managing BMI and physical activity levels.
- Understand that the model is a supporting tool and not a definitive diagnostic instrument. Collaboration with healthcare providers is essential for accurate assessments.