HEART ATTACK PREDICTION: What symptoms/health measurements increase risk of heart attack?

Problem Statement

We're embarking on a project to predict the risk of a heart attack using the 1988 heart disease dataset. Our primary goal is to develop a robust predictive model while conducting in-depth feature analysis to understand the significance of each attribute in determining the risk of a heart attack.

Central Questions

1. What factors truly impact an individual's susceptibility to a heart attack?
2. Can our predictive model accurately assess the risk of a heart attack using features from diverse databases?
3. What hidden patterns and correlations in the dataset provide valuable insights into the factors contributing to a heart attack?

Context

Heart attacks are a significant health concern globally, emphasizing the need for early intervention. This project aims to leverage a diverse dataset to unravel the complex relationships between various attributes and the risk of a heart attack.

Criteria for Success

Success in this project is defined by our ability to build a reliable predictive model, identify key features influencing heart attack risk, and effectively communicate these findings. The project aims to contribute valuable insights for proactive prevention measures.

Scope of Solution Space

The project involves exploring correlations, building predictive models, and conducting feature analysis to deepen our understanding of heart attack risk factors. The focus is on generating actionable insights for proactive prevention measures.

Constraints

While the dataset provides valuable information, there may be missing details and a lack of descriptions for some attributes. Additionally, certain demographic factors might not be included, limiting our analysis.

Stakeholders

Healthcare professionals and individuals seeking proactive measures to prevent heart attacks stand to benefit from our accurate predictive model and insightful feature analysis.

Data Source
Link: *https://www.kaggle.com/datasets/juledz/heart-attack-prediction*
The dataset is sourced from Kaggle, containing diverse attributes, from demographics to medical history and diagnostic results, providing a rich source for analysis.

Data Cleaning and Data Wrangling

Initially the data frame contained 14 columns and 1025 entries

Columns:

-Age (Numeric): Represents the age of the individual.

-Sex (Categorical): Indicates the gender of the individual (0: Female, 1: Male).

-Chest Pain Type (Categorical): Describes the type of chest pain experienced by the -individual, categorized into four types.

-Resting Blood Pressure (Numeric): Represents the blood pressure of the individual at rest. -Serum Cholesterol (Numeric): Indicates the cholesterol levels in milligrams per deciliter (mg/dL).

-Fasting Blood Sugar (Categorical): Indicates whether the fasting blood sugar is above 120 mg/dL (1: Yes, 0: No).

-Resting Electrocardiographic Results (Categorical): Describes the resting electrocardiographic results as normal, abnormal, or hypertrophy.

-Maximum Heart Rate Achieved (Numeric): Represents the highest heart rate achieved during testing.

-Exercise-Induced Angina (Categorical): Indicates whether angina was induced by exercise (1: Yes, 0: No).

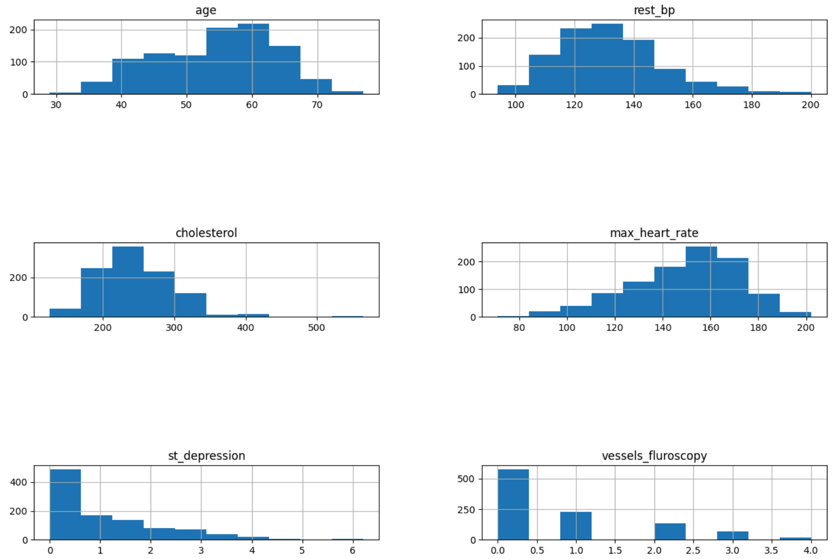
-Oldpeak (ST Depression) (Numeric): Represents the ST depression induced by exercise relative to rest.

-Slope of Peak Exercise ST Segment (Categorical): Describes the slope of the peak exercise ST segment as upsloping, flat, or downsloping.

-Number of Major Vessels Colored by Fluoroscopy (Numeric): Represents the count of major vessels colored by fluoroscopy (angiography).

-Thalassemia (Categorical): Describes thalassemia as normal, fixed defect, or reversible defect.

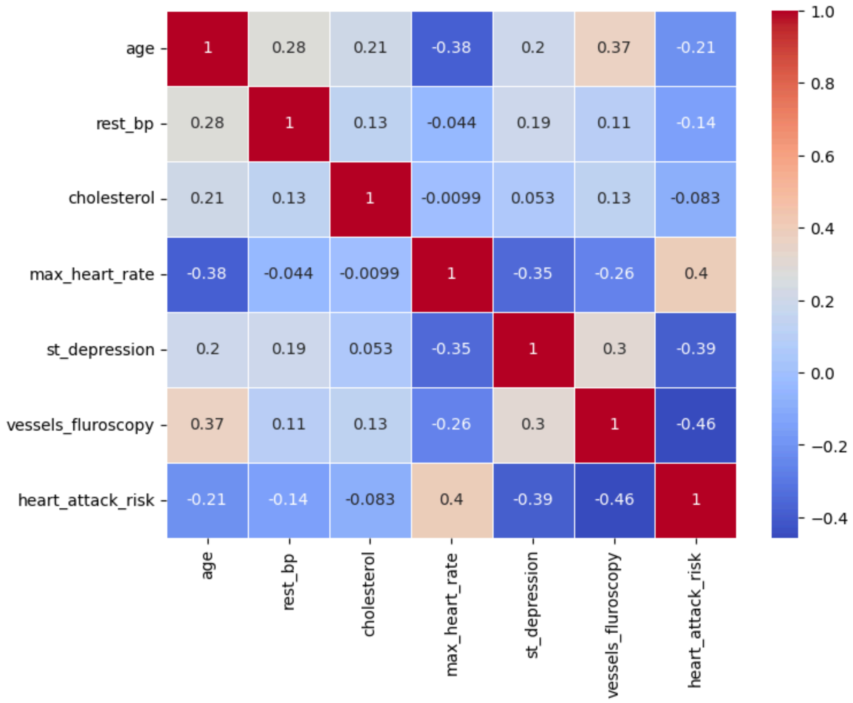
Since there were no null entries within the columns, no removals or drops needed to take place in that regard.



The histograms for the numerical variables above shows that most of the numerical columns are distributed normally without any outliers. However the vessels_fluroscopy represents the number of major vessels seen under fluoroscopy, and the maximum value is only 3 vessels. Since we have some values above 3, I decided to drop those columns for more usable data to remain in the dataset. The final cleaned dataset had 14 columns and 1007 entries.

Exploratory Data Analysis

Overall Correlation

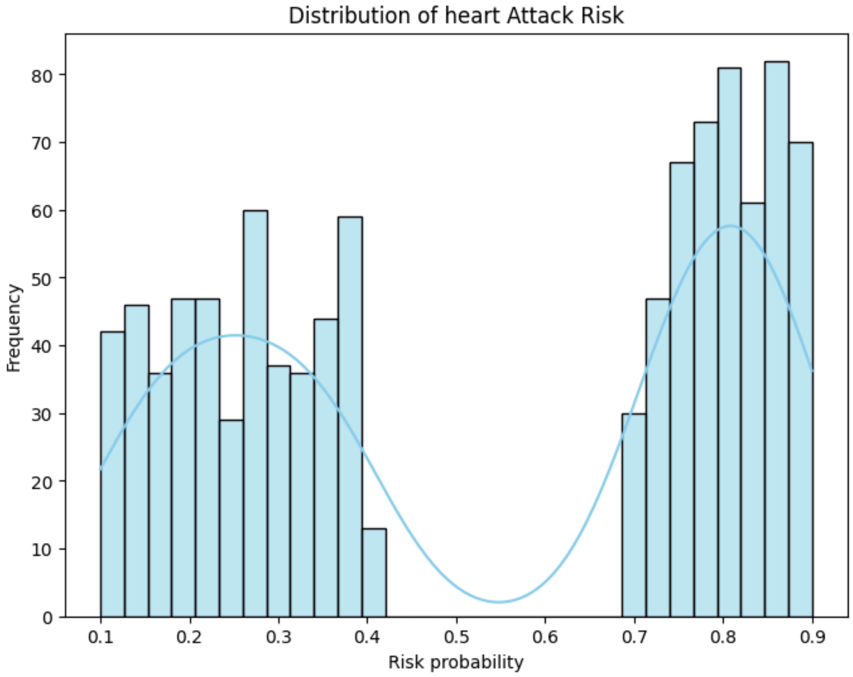


	age	rest_bp	cholesterol	max_heart_rate	st_depression	vessels_fluroscopy	heart_attack_risk
age	1	0.28	0.21	-0.38	0.2	0.37	-0.21
rest_bp	0.28	1	0.13	-0.044	0.19	0.11	-0.14
cholesterol	0.21	0.13	1	-0.0099	0.053	0.13	-0.083
max_heart_rate	-0.38	-0.044	-0.0099	1	-0.35	-0.26	0.4
st_depression	0.2	0.19	0.053	-0.35	1	0.3	-0.39
vessels_fluroscopy	0.37	0.11	0.13	-0.26	0.3	1	-0.46
heart_attack_risk	-0.21	-0.14	-0.083	0.4	-0.39	-0.46	1

The table above shows the correlation between the different numerical variables in the table. The main focus of this table should be the variables most correlated to the heart attack risk probability.

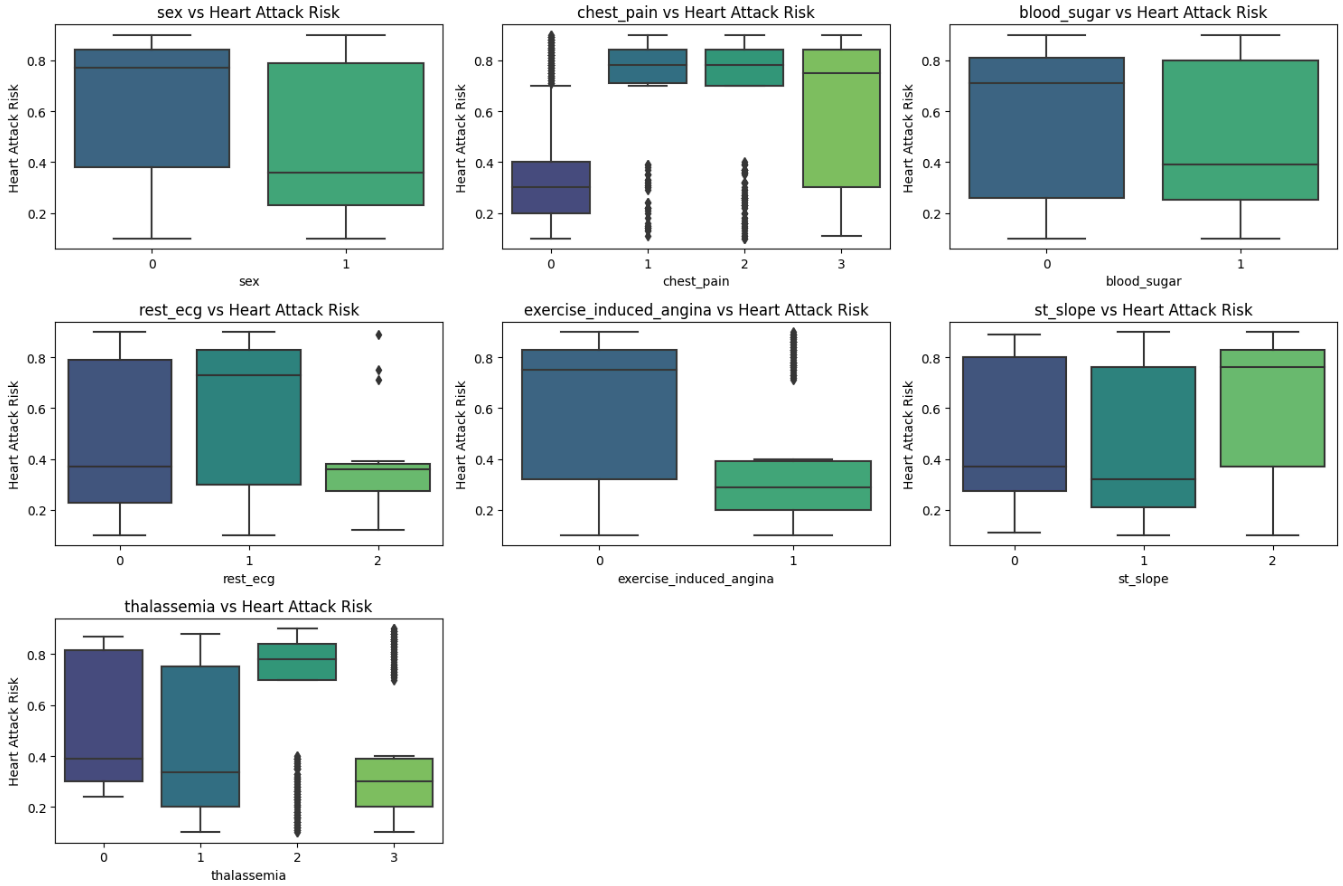
1. Number of Major Vessels (Fluoroscopy) and Heart Attack Risk: This correlation aligns with clinical expectations. A higher number of major vessels with issues (visible under fluoroscopy) is associated with a greater risk of cardiovascular events, including heart attacks.
2. ST Depression and Heart Attack Risk: Elevated ST depression during exercise is a recognized indicator of myocardial ischemia (insufficient blood supply to the heart). This aligns with the correlation, suggesting that individuals with higher ST depression might be at an increased risk of a heart attack.
3. Max Heart Rate and Heart Attack Risk: The correlation between maximum heart rate and heart attack risk could indicate that individuals with specific heart rate responses during exercise might have an elevated risk. Abnormal heart rate responses can be indicative of underlying cardiovascular issues.

Distribution of Heart Attack Risk



This graph points out that heart attack risk is split with values below 0.5, and with values above 0.5. There are no risk probabilities at 0.5 or 0.6. This will prove useful during the model stage, when I implement a classification model.

Box Plots: ***Categorical variables vs. Heart Attack Risk***



From the boxplots above we can see that Anginal(1) and non-anginal(2) chest pain show a high risk of heart attack from the box plot. When the Angina is exercise induced there is a lower risk of heart attack. Reversible defect thalassemia has the highest risk rates among the thalassemia types. While the individual variables can explain some of the correlations to heart attack risks, there may be multiple correlation, and or underlying factors influencing the rate. For example, a person with blood sugar, and thalassemia may have a higher risk rate of heart attacks.

Interpretations

1. Anginal(1) and Non-Anginal(2) Chest Pain: Consistent with medical knowledge, Anginal and Non-Anginal chest pain types are associated with a higher risk of heart attack. These types of chest pain can be indicative of coronary artery disease, which may lead to compromised blood flow to the heart.
2. Exercise-Induced Angina: Exercise-induced angina typically suggests stable angina, where chest pain occurs during physical exertion but is relieved by rest. Individuals with stable angina may have better-managed cardiovascular conditions, contributing to a lower risk of heart attack during exercise.
3. Reversible Defect Thalassemia: Thalassemia is a genetic blood disorder that can affect the production of hemoglobin. The association between reversible defect thalassemia and a higher risk of heart attack could be related to the impact of thalassemia on cardiac function.

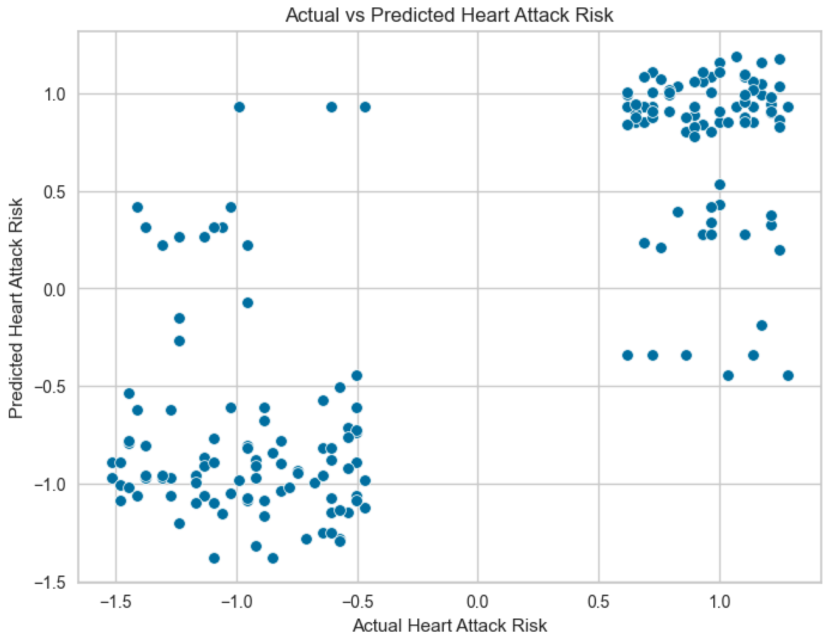
Individuals with thalassemia may experience complications affecting the heart, contributing to the observed risk.

Model Implementation

I began by encoding and scaling my variables. I encoded all the categorical variables, and ran a standard scaler for all the numerical variables. Furthermore, I split the data into training and testing sets, with the testing sets being 20% of the original data. The target feature was heart attack risk.

KNN Regressor

I wanted to get a baseline regression value prior to selecting the best model, and tuning hyperparameters. Using the k-NN regressor for this dataset provides a straightforward and intuitive baseline measure due to its simplicity and ease of implementation. As a non-parametric algorithm, k-NN doesn't assume specific data distributions, making it adaptable to complex relationships within the features. Its lack of linearity assumptions is advantageous for datasets where the correlation between features and the target variable is nonlinear.

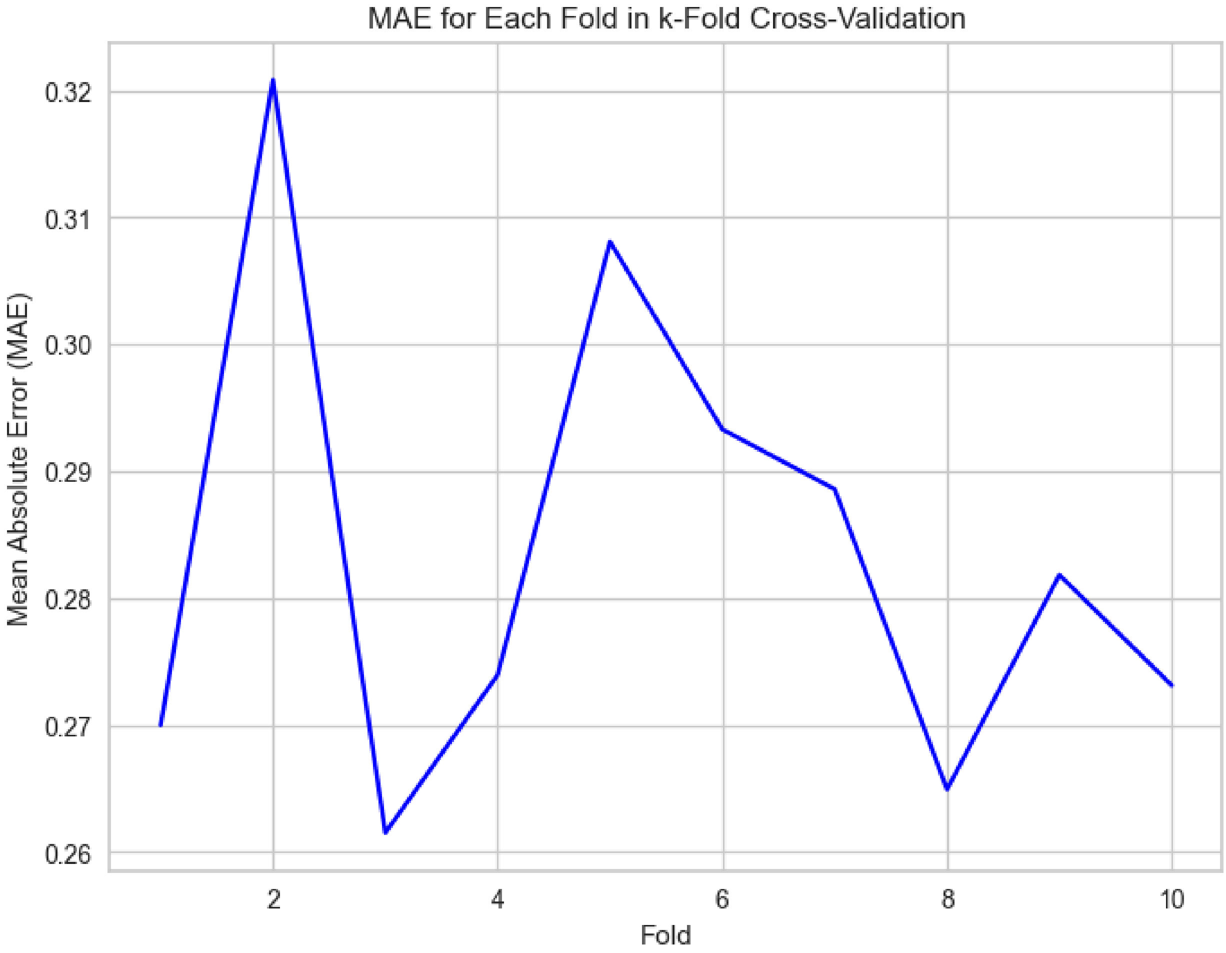


The Mean Squared Error (MSE) for the KNN Regressor model is calculated to be 0.3472. The MSE is a measure of the average squared difference between the predicted and actual values of the target variable, in this case, the heart attack risk.The MSE of 0.3472 suggests that, on average, the squared difference between the predicted and actual heart attack risk values is relatively moderate.

XGBoost Regressor

I felt that the knn regressor did not do a great job at predicting heart attack risk, and thought an ensemble model with hyperparameter tuning would fare better. Using the XGBoost regressor with

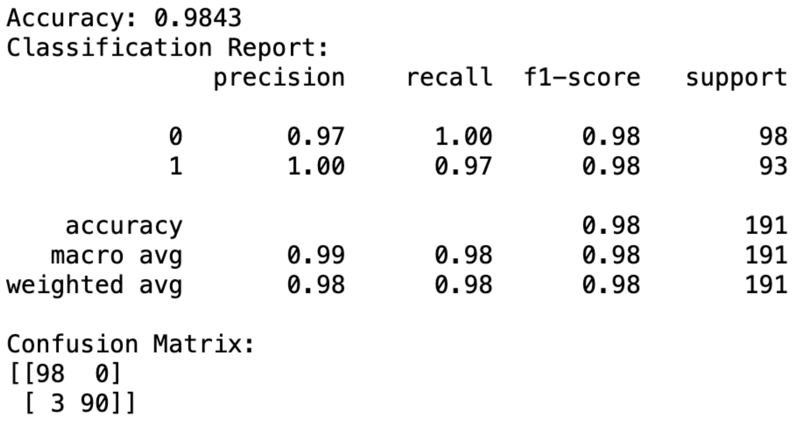
10-fold cross-validation provides a strong baseline measure for our dataset. XGBoost excels in capturing complex patterns and non-linear relationships within the data. Its ensemble approach helps prevent overfitting and enhances predictive accuracy. The 10-fold cross-validation ensures a thorough evaluation, training and validating the model on different parts of the data. Overall, XGBoost's flexibility and the comprehensive assessment through cross-validation make it a powerful choice for predicting heart attack risk probability in our dataset.



Achieving a Mean Absolute Error (MAE) of 0.26 on fold 8 during the 10-fold cross-validation is a promising result. This low MAE indicates that, on average, the XGBoost regressor's predictions were very close to the actual values in that specific fold. The average Mean Absolute Error (MAE) across the 10 folds is approximately 0.28, indicating overall good accuracy in predicting heart attack risk probability. Notably, the model performed exceptionally well in the eighth fold with the lowest MAE of 0.26. While some variability exists across folds, the consistent low MAE values suggest the XGBoost regressor is robust in capturing patterns within the dataset.

Random Forest Classifier

Since the heart attack probabilities were split between being above 0.5 and below 0.5, I split the risk into a binary variable with 0 being below 0.5 (lower risk) and 1 being above 0.5 (high risk). I thought that a classification model might better capture and predict risk, as the risk variable seems to represent collinearity and non linear progression. The algorithm's ability to handle non-linear relationships and provide feature importance scores is crucial in understanding factors influencing heart attack risk. Random Forest's resilience to outliers and simplicity in implementation further make it a practical baseline for classifying different risk levels in a medical context.



```
Accuracy: 0.9843
Classification Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.98        98
           1       1.00      0.97      0.98        93

    accuracy                           0.98       191
   macro avg       0.99      0.98      0.98       191
weighted avg       0.98      0.98      0.98       191

Confusion Matrix:
[[98  0]
 [ 3 90]]
```

The model achieved an impressive accuracy of 98.43%, correctly predicting heart attack risk in the majority of instances. In classifying low-risk cases (0), it demonstrated high precision (97%) and recall (100%), resulting in a balanced F1-score of 0.98. For high-risk cases (1), both precision and recall were perfect at 1.00, highlighting the model's precision and effectiveness in identifying instances with elevated risk. The confusion matrix further corroborates these findings, showcasing minimal misclassifications with only 3 false positives and no false negatives. Overall, these metrics indicate a robust and accurate model for heart attack risk prediction.

Further Exploration

Future Questions:

1. Temporal Analysis: How does the impact of certain features on heart attack risk change over time? Are there evolving trends or seasonal variations that affect susceptibility to heart attacks?

2. External Factors: To what extent do external factors, such as lifestyle changes or environmental influences, contribute to heart attack risk? How can these factors be incorporated into the predictive model to enhance accuracy?

3. Long-Term Predictions: Can the model be extended to predict long-term heart attack risk and how well does it perform in predicting risks over an extended period?

4. Comparative Analysis: How does our predictive model compare to existing risk assessment tools or traditional methods used in healthcare? What unique insights does our model provide that could enhance current practices?

5. Patient-Specific Recommendations: Can the model be refined to provide personalized recommendations for individuals based on their unique health profiles, ultimately improving the effectiveness of preventative interventions?

Future Considerations:

1. Ethical Implications: As predictive models are increasingly used in healthcare, what ethical considerations should be addressed, such as bias in the model, data privacy, and potential societal impact?

2. Model Interpretability: How can we enhance the interpretability of our model, making it accessible to healthcare practitioners, policymakers, and individuals seeking to understand the factors influencing heart attack risk?