Introduction

Heart disease is one of the leading causes of death worldwide. Identifying the factors that increase the risk of heart attacks can help in early prevention and treatment. In this project, we use machine learning to analyze a dataset related to heart attack risks. The goal is to predict the likelihood of a heart attack based on various health-related parameters such as age, cholesterol levels, blood pressure, and other risk indicators. By analyzing patterns in the data, we aim to develop models that can help medical professionals assess risk factors more effectively.

Dataset Exploration

The dataset consists of 50,000 records with 19 different features. These features include numerical variables such as age, cholesterol levels, resting blood pressure, and heart rate, as well as categorical variables like gender, smoking status, and family history

Preprocessing Steps:

- Before training the models, several preprocessing steps were performed to ensure the data was clean and suitable for machine learning:
- Handling missing values: Checked for and filled any missing data using appropriate imputation methods.
- Encoding categorical variables: Converted non-numeric values (such as "Male" and "Female") into numerical representations using label encoding.
- Feature scaling: Normalized numerical values to bring them to a common scale, ensuring that features with larger ranges did not dominate the model.
- Removing outliers: Detected and removed extreme values that could affect model performance.
- Feature selection: Identified and removed highly correlated features to avoid redundancy and improve model efficiency.

Model Training and Evaluation

The dataset was split into 70% training, 15% validation, and 15% test sets. The models were trained using Logistic Regression, Decision Tree, Random Forest, SVM, and XGBoost. Initially, simple models were trained without hyperparameter tuning. In the next stage, GridSearchCV was used to fine-tune model parameters. This resulted in a significant improvement in accuracy for the Decision Tree model, while other models showed only minor improvements.

Evaluation Metrics

To assess the performance of each model, we measured **accuracy** as the primary evaluation metric. The table below summarizes the results:

Model	Accuracy
Logistic Regression	50.00%
Decision Tree	49.89%
Random Forest	48.93%
SVM	50.04%
XGBoost	48.98%

For a more detailed evaluation, the Logistic Regression and Decision Tree models were further analyzed using additional metrics:

- **Confusion Matrix:** Provides insights into correct and incorrect predictions for each class.
- **Mean Squared Error (MSE):** Measures how far the predicted values deviate from actual values.
- **Classification Report:** Includes precision, recall, and F1-score to better understand model performance.

Conclusion

This project explored different machine learning techniques to predict heart attack risk. The models showed low predictive performance, indicating that either more relevant features are needed or that heart attack risk is influenced by many external factors not present in the dataset.One key takeaway from this study is the importance of hyperparameter tuning. The Decision Tree model, for instance, showed significant improvement after applying GridSearchCV to fine-tune parameters. This suggests that optimizing hyperparameters can make a major difference in model performance, highlighting the need for careful model tuning.

Despite these optimizations, the results indicate that the models used may be too simple for this type of complex medical prediction task. Given that heart disease is influenced by non-linear relationships and multiple interdependent factors, more advanced machine learning techniques, such as Neural Networks, may be better suited for capturing these intricate patterns.