Introduction

In this report, the team focuses on developing a predictive model to determine the likelihood of mortality among heart disease patients. Given the categorical nature of most variables in the dataset, I opted for boosting models instead of traditional regression models. Boosting models are particularly effective in handling categorical data and improving predictive performance.

By prioritizing precision and recall, the team aims to minimize misclassifications and provide a reliable tool for healthcare providers to make informed decisions. This approach ensures that the model is both accurate and practical for predicting patient outcomes.

Data Exploration

In the data exploration phase, we focused on understanding the structure and quality of our dataset. One key step was encoding categorical variables. We determined that variables with fewer than 10 unique values were better treated as categorical. Encoding these columns ensures that the model correctly interprets their categorical nature, which enhances the overall predictive performance.

We also addressed the issue of missing data. Utilizing a custom function designed to explore metadata across datasets, we assessed the completeness of our data. The function revealed that there were no missing values in our dataset, allowing us to proceed with confidence in the integrity of the data. This thorough exploration laid a solid foundation for the subsequent modeling process, ensuring that our models were built on clean and well-understood data.

Modeling Plan

Our independent variables include Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, and ST_Slope. These variables were selected based on their potential influence on heart health. For instance, Age and Sex are fundamental demographic factors that can affect heart disease risk. ChestPainType and ExerciseAngina are direct indicators of heart stress and potential blockages, while RestingBP and Cholesterol levels are critical markers of cardiovascular health. MaxHR (maximum heart rate) and Oldpeak (ST depression induced by exercise) are valuable for assessing heart function under stress. RestingECG and ST_Slope provide electrocardiogram insights, indicating abnormalities that could suggest heart disease.

The target variable, HeartDisease, was chosen because it directly represents the health outcome we aim to predict. Accurately predicting the presence of heart disease can significantly impact patient care, enabling early interventions and personalized treatment plans.

The modeling plan involves using these independent variables to build a robust predictive model. We will test various machine learning algorithms, particularly focusing on boosting models due to their efficacy with categorical data and their ability to handle the complexities of the dataset. We will evaluate the models based on precision and recall to ensure that our predictions minimize false positives and negatives. After thorough testing and validation, the model with the best performance metrics will be selected to predict heart disease, thereby aiding healthcare professionals in making informed decisions.

Performance Metrics Summary

Model	Precision (Test Data)	Recall (Test Data)
XGBoost	0.9	0.89
Decision Tree	0.9	0.89
CatBoost	0.89	0.87
HistGradientBoostingClassifier	0.92	0.87

XGBoost

One of the first models our team used was XGBoost. This model is widely used for boosting and has good performance in various classification and regression tasks. However, after considering it at length we decided not to proceed with it due to the potential of overfitting and its precision is not high enough. However, the overall precision score was 0.90, indicating room for improvement. Additionally, there were concerns about the model's ability to generalize well to unseen data, as boosting models can sometimes overfit to the training data.

Why Decision Tree is Considered

Decision trees provide clear, interpretable decision paths, making it easy for healthcare professionals to understand and trust the model. They highlight key factors, uncover new medical insights, and focus on important clinical variables. They handle complex, non-linear relationships, require minimal preprocessing, and manage incomplete data effectively. With high precision and recall, they ensure accurate identification of conditions, minimizing false positives and negatives. Their consistent accuracy across datasets makes them a reliable tool for healthcare decision-making.

Rationale for Considering and Ultimately Not Choosing CatBoost

CatBoost was initially considered for its robust handling of categorical data, which aligns well with the nature of our dataset. Known for its ability to efficiently handle categorical features without the need for extensive preprocessing, CatBoost also offers high predictive performance with minimal tuning. During the evaluation, CatBoost demonstrated solid precision and recall scores, indicating its capability to effectively identify high-risk patients and reduce false negatives.

However, despite its strong performance, CatBoost was ultimately not chosen because it fell slightly short in precision compared to the HistGradientBoostingClassifier. While CatBoost achieved a precision of 0.89, the HistGradientBoostingClassifier outperformed it with a precision of 0.92. Precision is a critical metric in our context as it helps minimize false positives, which is crucial for accurately predicting patient mortality. Therefore, although CatBoost is a powerful and efficient model, the superior precision of the HistGradientBoostingClassifier made it the preferred choice for our final predictive model.

Model Selection

The **HistGradientBoostingClassifier** was selected due to its superior precision, which is crucial for minimizing false positives in predicting mortality among heart disease patients. Specifically, it achieved:

- **Precision:** 0.92
- **Recall:** 0.87

These metrics demonstrate that the HistGradientBoostingClassifier provides the highest precision among the evaluated models while maintaining a high recall. Its robust handling of categorical data and computational efficiency further reinforce the decision to choose this model as the optimal choice for real-world healthcare applications. This ensures the model is reliable and practical for predicting patient outcomes accurately.

Conclusion and Recommendations

In this report, we developed a predictive model to determine the likelihood of mortality among heart disease patients. Through thorough data exploration and model evaluation, the HistGradientBoostingClassifier emerged as the optimal choice due to its superior precision and balanced performance across key metrics.

To ensure the ongoing improvement and relevance of this project, we recommend several specific steps. Firstly, expanding the dataset to include more diverse populations will enhance the model's generalizability and robustness. Secondly, conducting feature engineering to investigate additional features such as lifestyle factors and genetic data can provide deeper insights and improve predictive accuracy.

Continuously refining the model with new data and exploring advanced techniques, such as ensemble learning, will help maintain its performance. Clinical validation in real-world settings is essential to assess the model's practical utility and identify areas for further enhancement. Finally, developing user-friendly tools, such as dashboards or mobile applications, will facilitate the integration of the model into clinical practice, making it accessible for healthcare providers.

By implementing these recommendations, the project can evolve into a comprehensive tool for predicting heart disease outcomes, ultimately improving patient care and healthcare efficiency.