Heart Disease Severity Prediction Using Machine Learning

# Table of Contents

1.**Introduction**

2.**Methods**  
 2.1 Importing Libraries and Dataset  
 2.2 Exploratory Data Analysis (EDA)  
 2.3 Data Preprocessing  
 2.4 Handling Class Imbalance  
 2.5 Feature Engineering and Selection  
 2.6 Model Building and Training  
 2.7 Evaluation Metrics  
 2.8 Explainability with SHAP  
 2.9 ROC Curve Analysis  
 2.10 Deployment

3.**Results**

4 .**Discussion**

5.**Model Performance**

6.**Feature Impact**

7.**Class Imbalance Handling**

8.**Limitations**

9.**Conclusion**

10.**Contributions**

11.**References**

12.**Appendices**

## 1. Introduction

Cardiovascular disease remains one of the primary causes of mortality globally, presenting considerable difficulties for public health systems and clinical choices. Prompt and precise identification of heart conditions can significantly enhance patient outcomes and maximize healthcare resources. Historically, many predictive models in healthcare emphasize binary classification determining if a patient has heart disease or not. In practical clinical situations, grasping the seriousness of heart disease is significantly more important, as it informs the immediacy of intervention, choices regarding hospitalization, and planning for long-term care. This project seeks to advance beyond binary diagnosis by creating a multiclass machine learning model that can predict the severity of heart disease, classified into five levels (0 to 4) using various clinical and demographic factors. Utilizing the UCI Cleveland Heart Disease dataset, we employ sophisticated preprocessing methods, including SMOTE for balancing classes and standardizing features, and then conduct thorough model experimentation with algorithms such as Random Forest, XGBoost, LightGBM, and CatBoost. SHAP explainability techniques are utilized to promote transparency in predictions and underscore the key features that impact severity classification. This project showcases how the integration of strong algorithms with clinical data analysis can result in significant decision-support systems within healthcare analytics.

## 2. Methods

The approach taken in this project comprises multiple clearly outlined steps designed to create a trustworthy and understandable machine learning model for predicting heart disease severity. The comprehensive method incorporates data preprocessing, exploratory analysis, correction of class imbalance, model training, evaluation, and interpretability.

**2.1 Importing Libraries and Dataset**

We started by bringing in key Python libraries like pandas, numpy, matplotlib, seaborn, and scikit-learn for manipulating, visualizing, and modeling data. Libraries such as xgboost, lightgbm, catboost, and shap were employed for advanced model training and interpretation. The UCI Machine Learning Repository provided the Cleveland Heart Disease dataset.

**2.2 Exploratory Data Analysis (EDA)**

A comprehensive EDA was conducted to analyze the distribution and correlations of features:

* A pie chart was created to illustrate the class imbalance in the target variable (num).
* A correlation heatmap was utilized to identify linear relationships among input variables.
* Boxplots and stacked bar graphs were generated to evaluate how factors such as age and type of chest pain affected the severity of heart disease.

A colorful squares with black text

AI-generated content may be incorrect.

Heatmap of Heart Disease Correlations: Mapping the connections between input features to recognize significant correlations and potential multicollinearity issues.

**2.3 Data Preprocessing**

Data preprocessing involved several crucial steps:

* Missing values were replaced by the mean.
* The Interquartile Range (IQR) method was used to cap outliers in features like chol and trestbps.
* Categorical variables (e.g., cp, thal, slope) were converted via one-hot encoding.
* StandardScaler was utilized to normalize numerical features, guaranteeing uniform input ranges for model training.

**2.4 Handling Class Imbalance**

The dataset exhibited a strong class imbalance, especially in severity classes 3 and 4. To address this:

* We employed SMOTE (Synthetic Minority Oversampling Technique) to artificially create additional samples for minority categories.
* This phase was essential to prevent favoritism towards the majority class (class 0) and to guarantee that the model learned uniformly across all severity levels.

**2.5 Feature Engineering and Selection**

We carried out fundamental feature engineering by modifying and encoding features to prepare them for the model. Feature importance rankings derived from Random Forest and CatBoost models were utilized together with correlation analysis to keep the most significant predictors.

**2.6 Model Building and Training**

Several classification models were developed to assess performance:

* Random Forest – foundational ensemble model.
* XGBoost and LightGBM – gradient boosting libraries known for their outstanding generalization.
* CatBoost – selected for its excellent performance with categorical features and clarity in interpretation.
* Stacking Ensemble – merged the advantages of the leading models by utilizing logistic regression as a meta-learner.

Every model was trained on the SMOTE-adjusted, normalized dataset with an 80/20 train-test division.

**2.7 Evaluation Metrics**

Every model was assessed using:

* Precision – Total number of correct forecasts.
* Macro F1 Score – Average F1 score for all classes, giving each class the same weight.
* Weighted F1 Score – F1 score adjusted according to support (distribution of classes).

Outcomes were analyzed using a grouped bar chart that summarizes all the models.

**2.8 Explainability with SHAP**

To understand the decisions made by the model:

* SHAP (SHapley Additive exPlanations) was utilized with the CatBoost model.
* SHAP summary plots illustrated global feature significance.
* Waterfall plots demonstrated how single features influenced particular predictions.

**2.9 ROC Curve Analysis**

We additionally calculated ROC curves for every severity class utilizing a One-vs-Rest classifier configuration. This assisted in assessing the sensitivity and specificity of the models, especially CatBoost and XGBoost.

**2.10 Deployment**

The completed CatBoost model was implemented with Streamlit, enabling users to enter features (such as age, gender, cholesterol) and obtain the forecasted severity category. This local implementation showcased the practical use of the model in a healthcare environment.

## 3. Results

The performance of each model is summarized below.

A table with numbers and text

AI-generated content may be incorrect.

A graph of different colored bars

AI-generated content may be incorrect.

This table outlines the accuracy, F1 score, and ROC AUC (micro-average) for every model evaluated. CatBoost demonstrated the greatest accuracy, whereas the Stacking Ensemble obtained the highest F1 Score and a strong ROC AUC, reflecting a well-rounded and efficient prediction ability across various heart disease severity levels.

A graph of a diagram

AI-generated content may be incorrect.

Waterfall Plot illustrating single predictions through SHAP values. These graphs depict the impact of important factors such as age, chest pain type (cp), cholesterol levels, and resting blood pressure (trestbps) on the predicted severity of heart disease by the model. Red bars elevate the prediction, whereas blue bars decrease it. This improves the model's clarity and fosters clinical confidence in automated forecasts.

A diagram of a classifier

AI-generated content may be incorrect.

The CatBoost model exhibited excellent classification proficiency for Class 0 (AUC = 0.94) and a moderate performance for the other classes, achieving a micro-average ROC AUC of 0.84, which suggests generally dependable results.

## 4. Discussion

The findings of this project highlight the intricacies associated with forecasting heart disease severity through machine learning methods. Several models were examined, such as Random Forest, XGBoost, LightGBM, CatBoost, and a Stacking Ensemble. Among the options, CatBoost stood out as the most efficient, attaining the top micro-average ROC AUC of 0.84 and surpassing the others in terms of accuracy and interpretability. This model effectively managed categorical features and showed resilience against imbalanced classes, a persistent issue during the project. The distribution of classes showed a significant imbalance, as Class 0 (no or minimal severity) accounted for more than 50% of the dataset, while Class 4 (the most severe cases) made up less than 5%. This disparity greatly affected model training and performance indicators, particularly for minority classes where precision and recall stayed diminished even with resampling attempts. The analysis of feature correlation pinpointed trestbps, chol, ca, and thalach as some of the most significant predictors. Additional SHAP analysis aided in understanding how specific features influenced predictions locally, strengthening the CatBoost model's credibility in representing non-linear relationships. Even with various enhancements such as hyperparameter tuning, PCA, and feature engineering, the performance on minority classes is still restricted, indicating a possible requirement for synthetic oversampling or more detailed features. Moreover, the online deployment of the CatBoost model via an intuitive interface showcases the practical relevance of this work in a real-world healthcare context, enabling users to enter patient information and obtain heart disease severity predictions right away. In general, the project illustrates how machine learning can significantly aid in early diagnosis and risk assessment in cardiovascular health, while also highlighting the need to tackle class imbalance and improve model generalization for its clinical application.

## 5.Model Performance

To assess the predictive power of different algorithms, we analyzed the effectiveness of five unique machine learning models: Random Forest, XGBoost, LightGBM, CatBoost, and a Stacking Ensemble. Every model was trained with a stratified and balanced edition of the dataset, and their results were evaluated using Accuracy, F1 Score, and Micro-Average ROC AUC, which is particularly useful in cases of imbalanced classification.

The CatBoost Classifier achieved the highest accuracy (55.7%) and showed consistent performance across all metrics, indicating its effectiveness in capturing both categorical and numerical patterns in the dataset. Stacking Ensemble, while somewhat less accurate, obtained the highest F1 Score (0.251), demonstrating its strength in managing class overlap and complexity via meta-model integration. Random Forest and XGBoost were similar in accuracy but recorded slightly lower F1 Scores, likely due to difficulties in distinguishing minority classes. LightGBM, despite its fast training speed, yielded the lowest accuracy (49.2%) and F1 score (0.185), indicating that it might not be the best choice for this particular multi-class classification task without further parameter adjustment or feature engineering.

A comparative visualization of all the model outcomes, featuring a performance bar chart and a ROC Curve for CatBoost, showed that while all models performed adequately on the dominant class (Class 0), their predictive capabilities markedly decreased for the higher severity classes. This emphasizes the importance of selecting evaluation metrics beyond accuracy when handling imbalanced class distributions. In summary, the CatBoost model was chosen as the final model not only for its performance metrics but also for its compatibility with explainability tools like SHAP, which enhanced its transparency and interpretability in healthcare settings.

## 6.Feature Impact

To obtain a better understanding of how various clinical factors affected the prediction of heart disease severity, we utilized CatBoost's inherent feature importance in conjunction with SHAP (SHapley Additive exPlanations) values. This enabled us to analyze the role of each feature both globally (overall significance) and locally (influence on specific predictions).

Of all the features, resting blood pressure (trestbps) stood out as the most significant variable. It consistently emerged with significant importance in both model-driven rankings and SHAP visual representations. Patients with higher trestbps values were typically linked to more severe severity classes. The count of major vessels (ca) and age were also important predictors. Elderly individuals and those with greater blood vessel obstruction were more prone to being categorized into higher severity levels.

Cholesterol (chol) and the highest heart rate reached (thalach) played a moderate role in the prediction, frequently interacting with age and chest pain type (cp) to affect the model's outcomes. Although categorized, the type of chest pain was shown to significantly impact the class shifts between low and moderate severity, especially evident in the SHAP waterfall plots.

Attributes such as sex, fasting blood sugar (fbs), and resting ECG findings (restecg) ranked as some of the least significant in the final model, exhibiting minimal differences in SHAP values throughout the population.

The SHAP waterfall plots for single predictions illustrated how feature values influenced the prediction to lean towards a specific class. For instance, an elevated age and trestbps may elevate the anticipated severity score, whereas a standard cp type could lessen it. This degree of openness not only enhances model reliability in clinical settings but also points out possible opportunities for medical intervention.

## 7.Class Imbalance Handling

In our dataset on heart disease severity, we observed a notable class imbalance, with Class 0 (indicating no or minimal heart disease) prevailing in the target variable, representing more than 50% of the total records. The other classes, particularly Class 4 (the most severe heart disease), were lacking in representation. This uneven distribution created a risk of skewing the model in favor of the majority class, which could harm performance for minority classes and result in deceptive accuracy measurements.

To address this problem and guarantee equitable learning in every class, we implemented the subsequent strategies:

Segmented Division

We carried out stratified train-test divisions to preserve the same class distribution in both the training and testing datasets. This guaranteed that every severity category was appropriately represented, preventing biased evaluation outcomes.

SMOTE-based Synthetic Oversampling

We utilized SMOTE (Synthetic Minority Oversampling Technique) to artificially boost the sample size in minority classes. SMOTE creates additional samples by interpolating between existing instances, thus enhancing the diversity of underrepresented classes without replicating data. This greatly enhanced the model's capacity to generalize for uncommon instances (Classes 3 and 4).

Class Weighting in Algorithms

For models such as CatBoost and Logistic Regression, we applied class weights to impose a greater penalty on the misclassification of minority classes. This method encouraged the algorithm to focus more on these less represented categories during training.

Macro-Averaged Metrics for Assessment

Rather than depending only on accuracy, we presented macro-averaged F1 scores and ROC AUC scores, which consider all classes equally irrespective of their occurrence. This offered a more accurate representation of model performance, particularly in imbalanced situations.

Even with these efforts, the precision and recall metrics for Classes 3 and 4 continued to be comparatively lower, emphasizing the ongoing issue of scarce data in those areas. Nonetheless, our method effectively enhanced the recall for minority classes while maintaining the overall robustness of the model.

## 8.Limitations

Although the project effectively showcased the capabilities of machine learning in forecasting heart disease severity, multiple limitations were noted that could impact the model's generalizability and clinical relevance.

Dataset Class Imbalance

The initial dataset showed a major class imbalance, where most instances were in Class 0 and only a small number were in Classes 3 and 4. Even though we utilized SMOTE to oversample minority classes, the synthetic data might not completely represent the intricacies of actual patient diversity, possibly impacting model generalization.

Restricted Dataset Size

The UCI Cleveland dataset includes only 303 samples, which is quite limited for training effective multiclass classification models. This constraint might have limited the models' capacity to completely grasp subtle patterns in uncommon severity categories.

Streamlined Feature Collection

The dataset comprises merely 13 features, several of which are fundamental demographic or diagnostic variables. It does not include more comprehensive clinical details like imaging data, medication history, or lifestyle elements, which could greatly enhance predictive accuracy and context.

Model Effectiveness on Underrepresented Classes

Even with resampling and evaluation methods, the model's performance for severe classes (3 and 4) continued to be poorer regarding precision and recall. This highlights a larger issue in creating equitable models in healthcare, where detecting crucial yet infrequent cases is of utmost importance.

Only Local Deployment

The model was implemented locally with Streamlit. Although useful for showcasing, this restricts accessibility and scalability. A deployment of a cloud-based or integrated hospital system would be necessary for practical use.

Absence of Immediate Verification

The model has not been validated on actual clinical data or future patient inputs, indicating that its accuracy and practicality in a real-world hospital environment are still unproven.

# 9. Conclusion

This project effectively showcases the capabilities of machine learning in forecasting the severity of heart disease, advancing beyond conventional binary classification methods. Utilizing the UCI Cleveland dataset, we established a thorough pipeline that comprised data cleaning, exploratory analysis, feature engineering, class imbalance management through SMOTE, and meticulous model evaluation across multiple advanced algorithms. Of the models evaluated, CatBoost proved to be the most precise and interpretable, particularly

when coupled with SHAP explainability, which enabled clear insights into the contributions of features.

In spite of obstacles like dataset imbalance and small sample size, employing macro-averaged metrics and SHAP plots allowed for a fair assessment of model reliability and fairness. The final implemented model not only attained a commendable micro-average AUC of 0.84 but also provided a useful, locally-deployable resource for forecasting severity levels in clinical environments.

In summary, this study demonstrates that well-optimized machine learning models, when combined with suitable data preprocessing and interpretability instruments, can function as efficient decision-support systems in healthcare analytics—especially for the early detection and risk classification of cardiovascular disorders.

# 10. Contributions

|  |  |
| --- | --- |
| Shreyas Joshi | Data Preparation & Cleaning: Addressed missing values, transformed categorical variables, standardized the dataset, and maintained consistency throughout all feature columns.Exploratory Data Analysis (EDA): Created several informative visuals such as the heart disease severity pie chart, correlation heatmap, age-specific boxplot, and chest pain type analysis. These images were utilized to comprehend data distribution and patterns.Model Creation & Training: Developed and refined machine learning models including Random Forest, XGBoost, LightGBM, and CatBoost. Executed hyperparameter optimization and oversampling to address class imbalance issues.SHAP Explainability & Feature Significance: Utilized SHAP waterfall plots to clarify model predictions for each patient and illustrated the feature importance of CatBoost.Model Evaluation & Reporting: Assessed all models utilizing metrics like Accuracy, F1 Score, and ROC AUC. Developed performance comparison graphs and analyzed ROC curves.Documentation & Final Report: Wrote significant portions of the final report, covering methodology, results, discussion, and conclusion, and made certain that all graphs and tables were accurately positioned and formatted.Presentation Creation: Developed and organized the final PowerPoint slides, produced visuals for the summary, and maintained the project's logical progression. |
| Ayush Raj Saxena | Literature Review: Conducted background research on heart disease prediction studies using machine learning and helped shape the project's problem statement and justification.Data Analysis Support: Assisted in exploring trends and patterns in the dataset. Helped in interpreting the results of EDA visuals such as chest pain type distribution and age-wise analysis.Model Support & Validation: Contributed to the selection and evaluation of suitable classification models. Validated model performance metrics and contributed to ensemble method selection.Meeting Documentation: Took charge of weekly meeting minutes (1 to 6), summarizing project progress, key tasks completed, upcoming objectives, and member responsibilities.Proofreading and Editing: Reviewed report drafts, checked for consistency, and ensured that technical content was clearly explained and free from errors.Presentation Coordination: Supported the creation of slides for introduction, literature review, problem definition, and contributions. Provided feedback on visuals and ensured proper formatting.Deployment: Created and evaluated a local Streamlit web app for immediate prediction of heart disease severity. |

# 11. References

- UCI Heart Disease Dataset: https://archive.ics.uci.edu/ml/datasets/heart+Disease

- Scikit-learn Documentation: https://scikit-learn.org/

- CatBoost Documentation: https://catboost.ai/

- SHAP Documentation: https://shap.readthedocs.io/

- XGBoost Documentation: https://xgboost.readthedocs.io/

- LightGBM Documentation: <https://lightgbm.readthedocs.io/>

# 12.Appendices

* Finalpaper(G7).docx: Contains the final report
* Healthcare(G7).ipynb: Contains all the code for our project including importing, preprocessing, exploratory data analysis, machine learning algorithms and evaluation metrics.