Georgia State University

CSC4780/6780&DSCI4780 – Fundamentals of Data Science

Fall-2024

Final Project Report

Heart Attack Prediction

The Prediction Wizards

Benedict Antonio Mervyn

Venkata Dheeraj Bhogi

Nikitha Rajendran

Table of Contents

1. Business Understanding

1.1 Business Problem

Cardiovascular diseases (CVDs), including heart attacks and strokes, are the leading causes of mortality worldwide, accounting for approximately 17.9 million deaths each year, with a significant portion of these fatalities occurring in individuals under the age of 70. Given the widespread nature of this problem, early detection and timely intervention are crucial to improving patient outcomes. Heart attacks, in particular, can be life-threatening if not detected early, making it imperative for medical practitioners to identify individuals at high risk before an incident occurs. Current diagnostic methods, although effective, can often be time-consuming and may not always provide immediate insights for timely interventions.

The business problem we aim to tackle is to predict the likelihood of a heart attack in patients using clinical and demographic data. A predictive model will allow healthcare providers to identify individuals at high risk, enabling early intervention, personalized treatment plans, and ultimately saving lives. By implementing a data-driven solution, we aim to support healthcare professionals in their decision-making processes, potentially transforming the way heart attacks are predicted and managed in clinical settings.

1.2 Dataset

The dataset consists of 1,319 patient samples, each containing a combination of clinical and demographic features related to cardiovascular health. The below mentioned eight features provide detailed insights into the physiological state of each patient, including vital signs, biochemical markers, and demographic factors. Together, they allow for the analysis and prediction of heart attack risk, making this dataset a valuable resource for understanding and modeling cardiovascular conditions.

Features:

- age: The age of the patient (integer).
- gender: Gender of the patient (boolean, 0 for female & 1 for male).
- impulse: The heart rate of the patient (integer).
- pressurehigh: It is the pressure in blood vessels when the heart beats known as systolic blood pressure (integer).
- pressurelow: It is the pressure in blood vessels when the heart is at rest between beats known as diastolic blood pressure (integer).
- glucose: A measure of glucose in the blood (float).
- kcm: A marker enzyme found in heart muscle cells that can indicate heart injury known as Creatine Kinase-MB (float).
- troponin: It is a protein complex in heart muscle cells that regulates contraction and is released into the bloodstream when the heart muscle is damaged.

Label:

- ➢ class: The label feature indicating whether the patient has experienced a heart attack or not (Boolean: positive and negative).

1.3 Proposed Analytics Solution

The goal of this project is to develop a machine learning model capable of accurately classifying patients as “at risk” or “not at risk” for a heart attack. A diverse range of machine learning algorithms will be implemented, tested, and evaluated to ensure the robustness and accuracy of the solution.

These algorithms include:

- ➢ **Logistic Regression:** A baseline linear model widely used for binary classification.
- ➢ **Support Vector Machine (SVM):** Implemented to find the optimal hyperplane that best separates the two different classes in the feature space.
- ➢ **Random Forest:** An ensemble method combining multiple decision trees for robust classification and reduced overfitting.
- ➢ **Gradient Boosting:** A sequential learning model designed to optimize classification performance by reducing errors iteratively.
- ➢ **XGBoost:** An advanced gradient boosting technique with features like regularization and tree optimization for enhanced accuracy and speed.
- ➢ **LightGBM:** A fast and efficient gradient boosting algorithm capable of handling large-scale data with high performance.

The models will be evaluated using metrics like accuracy, precision, recall, F1-score, and ROC-AUC to identify the best-performing one. The final model will help healthcare providers assess heart attack risk efficiently, enabling early intervention and personalized care, with potential integration into clinical decision-making tools. Through these efforts, the project aims to contribute to the broader goal of improving patient outcomes and reducing the prevalence of heart attacks through timely and data-driven decision-making.

2 Data Exploration and Preprocessing

2.1 Data Quality Report

The dataset contains in total 8 descriptive features which are continuous except for gender feature which is encoded to 0 and 1 for female and male respectively. Please find below data quality reports for continuous and categorical features.

Continuous features:

Feature	Desc.	Count	% of Missing	Card.	Min.	Q1	Median	Q3	Max.	Mean	Std. Dev.
age	Continuous	1319	0.0	75	14	47.0	58.0	65.0	103	56.19	13.65
impluse	Continuous	1319	0.0	79	20	64.0	74.0	85.0	1111	78.34	51.63
pressurehight	Continuous	1319	0.0	116	42	110.0	124.0	143.0	223	127.17	26.12
pressurelow	Continuous	1319	0.0	73	38	62.0	72.0	81.0	154	72.27	14.03
glucose	Continuous	1319	0.0	244	35.0	98.0	116.0	169.5	541.0	146.63	74.92
kcm	Continuous	1319	0.0	700	0.321	1.655	2.85	5.805	300.0	15.27	46.33
troponin	Continuous	1319	0.0	352	0.001	0.006	0.014	0.0855	10.3	0.36	1.15

Categorical features:

Feature	Desc.	Count	% of Missing	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode Perc
gender	Categorical	1319	0.0	2	1	870	65.96	0	449	34.04
class	Categorical	1319	0.0	2	positive	810	61.41	negative	509	38.59

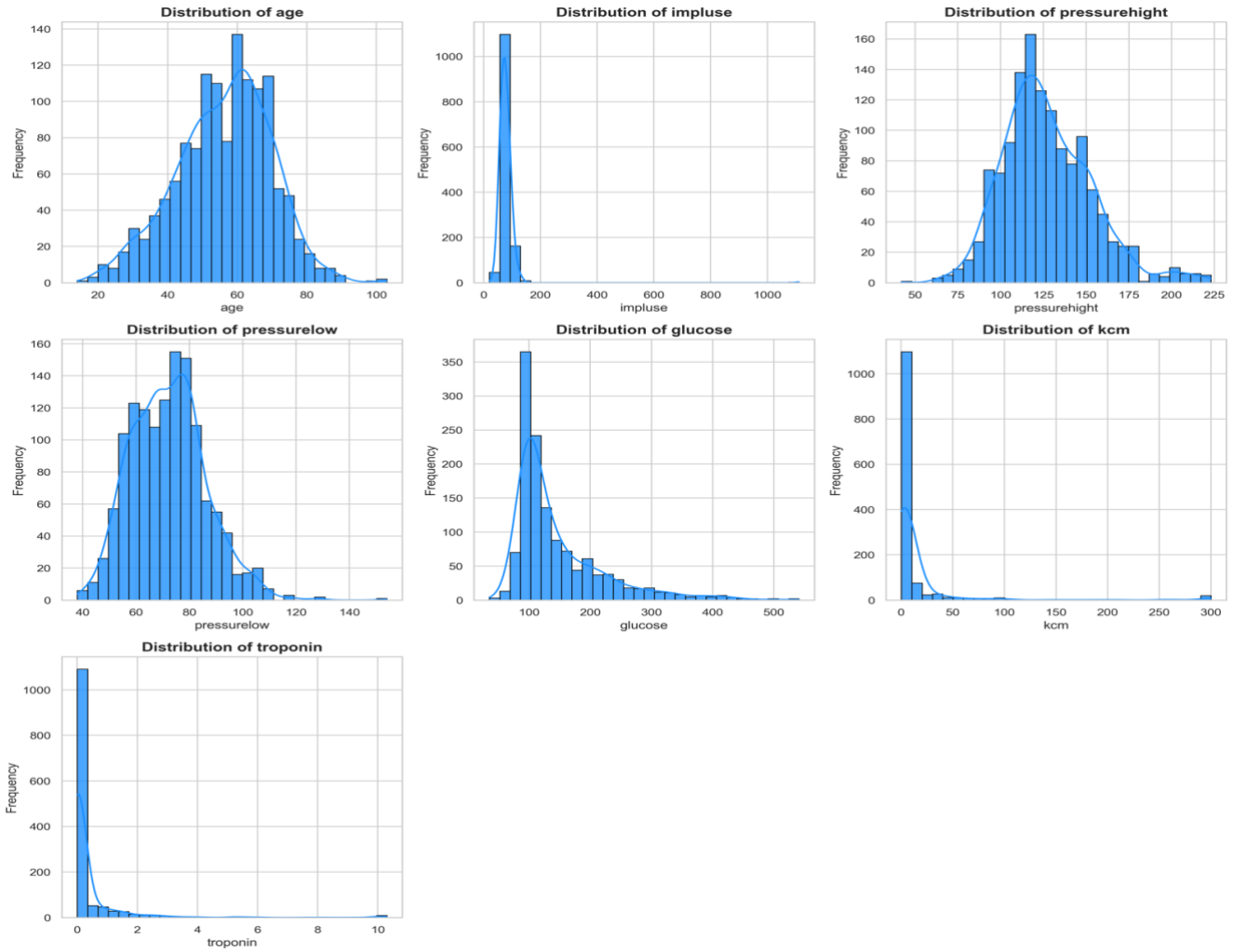

Figure: Continuous features distribution.

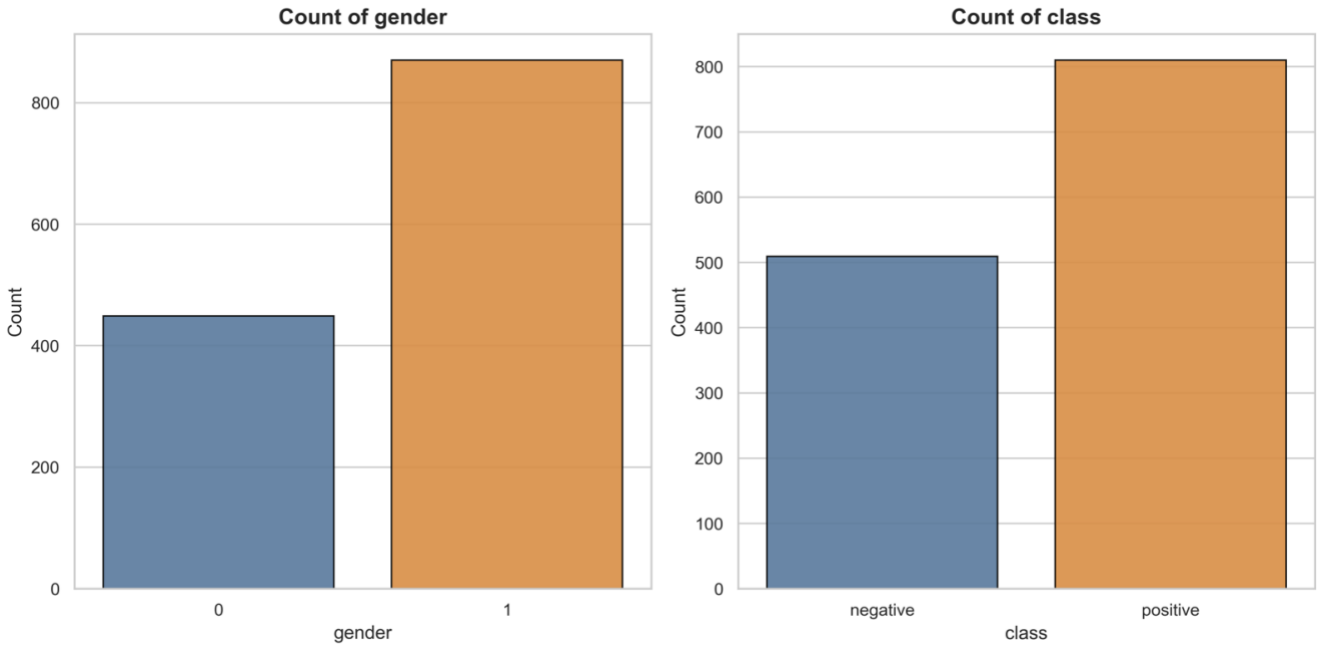

Figure: Categorical features distribution

2.2 Missing Values and Outliers

The dataset was confirmed to have no missing values based on the data quality report. However, several features contained extreme outliers that required handling:

- **Impulse**: Values exceeding 200 bpm, considered rare and likely measurement errors, were capped to maintain data consistency while retaining valid observations.
- **Pressurehigh and Pressurelow**: Outliers were capped to address potential errors while preserving significant values indicative of high-risk cardiovascular conditions.
- **Glucose**: High values, likely from diabetic patients, were capped to manage extreme outliers while retaining critical data points related to elevated risk.
- **KCM and Troponin**: Significant outliers were retained, as they provide essential information for heart attack prediction. Log transformations were applied to normalize their distributions.

To address these outliers, the Interquartile Range (IQR) method was employed, and values outside the acceptable range were capped. This approach effectively managed outliers while preserving valuable information crucial for diagnosing heart conditions.

2.3 Normalization

Following normalization was applied to prepare the dataset for modeling:

- Scaling was performed on age, impulse, pressurehigh, and pressurelow using StandardScaler to standardize them to a mean of 0 and a standard deviation of 1.

- ➢ Log normalization was applied to glucose, kcm, and troponin to reduce skewness and normalize their distributions.

These normalization techniques ensured that the features were on comparable scales, improving consistency and model performance.

2.4 Feature Selection and Transformations

To enhance the predictive capabilities of the models, several feature engineering techniques were applied, introducing derived and interpretable features that enriched the dataset:

- ➢ **Blood Pressure Ratio (bp_ratio)**: A new feature was created by calculating the ratio of systolic blood pressure to diastolic blood pressure (pressurehigh / pressurelow). This ratio highlights the balance between the two pressures, providing additional cardiovascular insights.
- ➢ **Impulse Feature Categorization**: The impulse (heart rate) feature was categorized into three ranges based on clinically relevant thresholds:

1. Low: Heart rate below 60 beats per minute (indicative of bradycardia).
2. Normal: Heart rate between 60 and 100 beats per minute (healthy range).
3. High: Heart rate above 100 beats per minute (indicative of tachycardia).
   These thresholds ensure the heart rate data is interpretable and aligned with medical standards.

- ➢ **Age Grouping**: The age feature was grouped into three categories using predefined thresholds:

1. Young: Ages 0–35 years.
2. Middle-Aged: Ages 36–60 years.
3. Elderly: Ages above 60 years.

These groupings simplify the age variable while preserving its influence on heart attack risk factors.

By applying these transformations, the dataset was enriched with derived features that capture meaningful relationships, improving the models' predictive accuracy and interpretability.

3. Model Selection and Evaluation

3.1 Evaluation Metrics

The performance of the machine learning models was evaluated using several key metrics to ensure robust and reliable results. These metrics include:

- ➢ **Accuracy**: The percentage of correct predictions over the total predictions made by the model.
- ➢ **Precision**: The ratio of true positive predictions to the total predicted positives, focusing on minimizing false positives.
- ➢ **Recall**: The ratio of true positive predictions to the total actual positives, emphasizing the identification of true cases.
- ➢ **F1-Score**: The harmonic mean of precision and recall, providing a balanced measure.
- ➢ **ROC-AUC**: The area under the Receiver Operating Characteristic curve, assessing the model's ability to distinguish between classes.

3.2 Models

The following machine learning models were implemented and evaluated for predicting heart attack risks:

- **Logistic Regression**: A linear baseline model to provide a benchmark for performance.
- **Support Vector Machine (SVM)**: Implemented to find the optimal hyperplane that best separates the two different classes in the feature space.
- **Random Forest**: An ensemble learning technique to enhance robustness through aggregated decision trees.
- **Gradient Boosting**: A sequential ensemble method focused on minimizing errors iteratively.
- **XGBoost**: An advanced gradient boosting algorithm with built-in regularization to prevent overfitting.
- **LightGBM**: A highly efficient gradient boosting technique known for its speed and scalability.

3.3 Evaluation

3.3.1 Evaluation Settings and Sampling

In this project, the dataset was carefully split into training (70%) and testing (30%) sets to ensure that models were properly trained and evaluated on separate data. The training set was used to fit the models, while the testing set allowed for unbiased evaluation of their performance, ensuring that the models' generalization capabilities were accurately assessed.
To maintain consistency and reproducibility across multiple runs, a random state of 42 was used for data splitting, ensuring that the same random partitioning occurred each time the models were trained. This approach minimizes variability in model performance that may arise from different data splits. The dataset was balanced through this method, ensuring that rare classes were adequately represented in both the training and testing sets, enabling the models to learn the underlying patterns for all classes more effectively. This setup contributed to the robustness of the models' performance metrics, especially when evaluating precision, recall, and F1-scores for the different risk categories.

3.3.2 Hyper-parameter Optimization

Hyperparameter optimization was conducted to enhance model performance, employing GridSearchCV for exhaustive searches in smaller parameter spaces and RandomizedSearchCV for quicker exploration in larger spaces. Each model's critical parameters were fine-tuned using 5-fold cross-validation with metrics like ROC-AUC and F1-Score guiding the selection process.

Key Optimized Parameters:

- Random Forest: Achieved optimal performance with n_estimators=200, max_depth=20, min_samples_split=5, and min_samples_leaf=2.
- Gradient Boosting & XGBoost: Best results with learning_rate=0.05/0.1, n_estimators=200, max_depth=6, and additional parameters like subsample=0.8 for XGBoost.
- LightGBM: Optimal settings included num_leaves=63, max_depth=10, learning_rate=0.05, and

min_data_in_leaf=20.

- ➢ Logistic Regression: Achieved best performance with C=1, penalty='l2', and solver='liblinear'.

This structured tuning process ensured robust and optimized models, leading to high accuracy and effective predictions.

3.3.3 Evaluation

Each model was trained on the training data and evaluated on the unseen test dataset, with ROC curves plotted for each to visualize their discriminative ability. The ROC curves provided a comprehensive view of the models' classification performance at various thresholds by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR).The results revealed significant differences in model performance.

- ➢ **Random Forest** emerged as the top performer, achieving the highest AUC, indicating its superior ability to distinguish between positive and negative cases. Additionally, it demonstrated the highest F1-Score, reflecting a balanced performance between precision and recall, which is crucial in imbalanced classification problems.
- ➢ **The Support Vector Machine (SVM)** also performed well, particularly excelling in Precision and Recall, but did not match the overall robustness of Random Forest.
- ➢ **Logistic Regression** showed acceptable performance but struggled with lower AUC and F1-Score.
- ➢ **Gradient Boosting Machine (GBM)** showed higher variance in performance across different metrics.While it did perform well on some evaluation criteria, its results were less stable compared to the other models. GBM's performance fluctuated significantly depending on the evaluation metric, indicating it might be sensitive to the choice of hyperparameters.
- ➢ **XGBoost** performed strongly with an Accuracy of 0.98, ROC-AUC of 0.9841, Precision of 0.98, Recall of 0.99, and F1-Score of 0.99. Its performance was enhanced by hyperparameter tuning, making it a competitive choice.
- ➢ **LightGBM**, despite producing warnings about splits with positive gain, still showed reasonable results. LightGBM had some variability too in its performance, but it was still a strong contender in terms of general predictive power.

Confusion matrices were analyzed to examine the trade-offs between false positives and false negatives for each model. Cross-validation revealed that Random Forest, SVM, and XGBoost were stable and consistent, while Logistic Regression, GBM and LightGBM showed more variability. Hyperparameter tuning further optimized Random Forest, improving its overall performance.

Overall, Random Forest was selected as the best-performing model due to its superior performance across multiple metrics, including ROC-AUC, Precision, Recall, and F1-Score. To optimize its performance further, hyperparameter tuning techniques like GridSearchCV were applied, leading to improved results. This model's consistently high performance made it the best choice for the task. XGBoost performed second best, with impressive metrics serving as a strong alternative depending on specific requirements.

4 Results and Conclusion

Class 0 represents instances where a heart attack is predicted to be negative, and Class 1 represents instances where a heart attack is predicted to be positive.
The table below summarizes the performance metrics for all models:

Model	Accuracy	Precision (Class 1)	Precision (Class 0)	Recall (Class 1)	Recall (Class 0)	F1-Score	ROC-AUC
Logistic Regression	78.0%	80.0%	74.0%	85.0%	66.0%	0.82	0.86
SVM	77.0%	80.2%	71.0%	83.0%	68.0%	0.81	0.87
Random Forest	98.2%	98.0%	99.0%	99.0%	97.0%	0.99	0.99
Gradient Boosting	97.5%	97.9%	99.0%	98.9%	96.0%	0.97	0.98
XGBoost	97.8%	98.2%	97.0%	98.1%	97.0%	0.98	0.98
LightGBM	97.0%	98.0%	97.0%	98.0%	96.0%	0.97	0.98

- ➢ Random Forest proved to be the best-performing model, achieving the highest accuracy (98.2%) and ROC-AUC (0.99). This indicates its robustness in distinguishing between heart attack risks and non-risks. Its strong performance suggests it could serve as a reliable tool for risk prediction in a healthcare setting.
- ➢ XGBoost showed comparable results, with only slight differences from Random Forest. This model demonstrated excellent performance, making them a viable alternative in cases where tuning or specific requirements are needed.
- ➢ Logistic Regression and SVM served as reliable baseline models but showed limitations in capturing complex patterns which led to lower overall performance compared to the ensemble methods. while
- ➢ GBM and LightGBM performed well but lacked stable predictions indicating it might be sensitive to the choice of hyperparameters.

Based on the analysis, **Random Forest is the most effective model** for predicting heart attack risks, due to its superior performance across key metrics. For healthcare applications, this model can be recommended for deployment to predict and assess patient risks, providing clinicians with reliable, high-accuracy predictions.

However, XGBoost and Gradient Boosting may also serve as valuable alternatives, offering

competitive performance with slightly different configurations.

The model choice should be guided by the specific use case, resource availability, and the need for model explainability, as ensemble models like Random Forest and XGBoost provide strong generalization across different datasets and can be adapted with further tuning for improved performance.

Future Scope:

- ➢ Dataset Enrichment: Incorporating lifestyle factors such as smoking, exercise, and diet for a more comprehensive analysis.
- ➢ Deployment: Developing a web-based application for real-time predictions.
- ➢ Explainability: Integrating SHAP or LIME for model interpretability, enhancing clinical relevance and trust.