Heart Disease Prediction Using Machine Learning

Mohamed Shaker
College of Computing and Informatics
University of North Carolina at Charlotte
mshaker1@charlotte.edu

1. Introduction

This is the final project for the Machine Learning Project in ITSC 3156. The goal of this project is to dig deep into the use of supervised learning algorithms to predict heart disease using large sets of clinical data. The study explores the effectiveness of two widely used machine learning models, Logistic Regression and Random Forest, on a dataset containing more than sixty thousand patient records. These models were selected due to their strong performance on structured data and their relevance in modern medical decision support systems. The source code for training, evaluation, and visualization is provided in the accompanying GitHub repository.

1.1 Motivation and Challenges

Predicting heart disease using machine learning has become a huge focus in data-driven healthcare research. To this day, heart disease remains one of the major causes of death worldwide, and early detection can help improve the patient's outcome. Traditional medical screening relies purely on human knowledge and clinical measurements. The study of machine learning models which can process large amounts of datasets at a time and identify different complex relationships allow it to become a huge asset to identify factors that are not usually visible through manual analysis.

The main objective of this research is to develop predictive models that help classify whether a patient could potentially have heart disease or not. The study applies well known machine learning techniques and algorithms and examines their performance using different evaluation metrics.

1.2 Problem Statement

The main objective of this research is to develop predictive models that help classify whether a patient could potentially have heart disease or not. The study applies well known machine learning techniques and algorithms and examines their performance using different evaluation metrics. The project investigates how well logistic regression and ensemble-based decision tree models can classify patient outcomes and evaluates the comparative performance of these techniques. Achieving strong classification accuracy

can contribute to improved clinical support tools and inform how machine learning may be integrated into future medical diagnostic systems.

1.3 Our Method

The dataset I'm going to be using today contains more than 60 thousand samples and more than 10 features. The results from this study will provide further insights into the predictive value of different clinical features and the strength and limitations of the various modeling techniques. Logistic Regression is used as a baseline due to its interpretability and mathematical simplicity. It models the probability of heart disease by computing a weighted linear combination of input features passed through a sigmoid function. In contrast, the Random Forest model is used to capture nonlinear relationships through an ensemble of decision trees trained on different subsets of data. This model is capable of modeling more complex feature interactions and often achieves higher predictive performance on medical datasets.

2. Backgrounds and related work

Machine learning has found extensive application in the field of medical diagnosis, especially in predicting cardiovascular risk. Initial foundational studies investigated linear statistical models like Logistic Regression, which has been utilized for a considerable time to assess the likelihood of heart disease based on clinical indicators. These different approaches rely on intertable relations between plenty of features and outcomes, and they consistently remain widely used due to transparency in clinical decision making. Traditional linear models usually struggle to capture non-linear interactions among different psychological measurements, which can limit performance in more complex datasets. To overcome these limitations, researchers are increasingly utilizing ensemble learning techniques. Random Forest, presented by Breiman [2001], is among the most powerful ensemble methods and has shown great effectiveness for structured medical information. Random Forest minimizes the variance present in single-tree models and captures nonlinear feature interactions more effectively by building numerous decision

trees from bootstrapped samples and averaging their predictions. Many research studies have utilized Random Forest on heart disease datasets, showing enhanced accuracy and reliability over linear models.

In addition to ensemble methods, contemporary research has examined other machine learning techniques such as Support Vector Machines, Gradient Boosting, and Neural Networks. These approaches provide additional pathways for modeling complex relationships within clinical data. However, they require more extensive parameter tuning, are less interpretable, and may be challenging to apply in clinical environments that prioritize model transparency. These methods offer different ways to understand complicated connections in patient data. But, they need a lot more tweaking of their settings, are harder to figure out, and can be tricky to use in hospitals where it's really important to understand how the models work.

3. Method

The methodology for this project consists of constructing a complete supervised learning pipeline to classify the presence of heart disease based on clinical attributes. Two machine learning models were selected to represent different modeling paradigms: Logistic Regression as a linear baseline and Random Forest as a nonlinear ensemble method. The objective is to evaluate how these models extract predictive patterns from a large heart disease dataset and to compare their performance using established medical and data-driven evaluation metrics. The dataset on heart disease represents both numerical and categorical variables that signify different crucial indicators of health. Prior to the training of the model, all features underwent preprocessing to guarantee appropriate scaling, consistency, and compatibility with both linear and tree-based approaches. After the preprocessing phase was finalized, both models were trained utilizing a conventional 80 to 20 train-test division. Each model generates a probability of heart disease, which can be adjusted for classification purposes.

3.1 Logistic Regression Model

Logistic Regression is a classical statistical learning method widely used for binary classification. It models the conditional probability that a patient has heart disease by computing a linear combination of input features and applying a sigmoid transformation. The model learns a set of coefficients that correspond to the relative importance of each clinical attribute in influencing the predicted probability. This method was selected due to its interpretability and its use as a baseline in many medical prediction studies. The learned coefficients allow practitioners to understand how features such as age, cholesterol, and chest pain type contribute to the classification outcome. Regularization was applied to control model complexity and reduce overfitting, especially in the presence of correlated clinical variables. Picture below models the conditional probability that a patient has heart disease by computing a linear combination of input features and applying a sigmoid transformation.

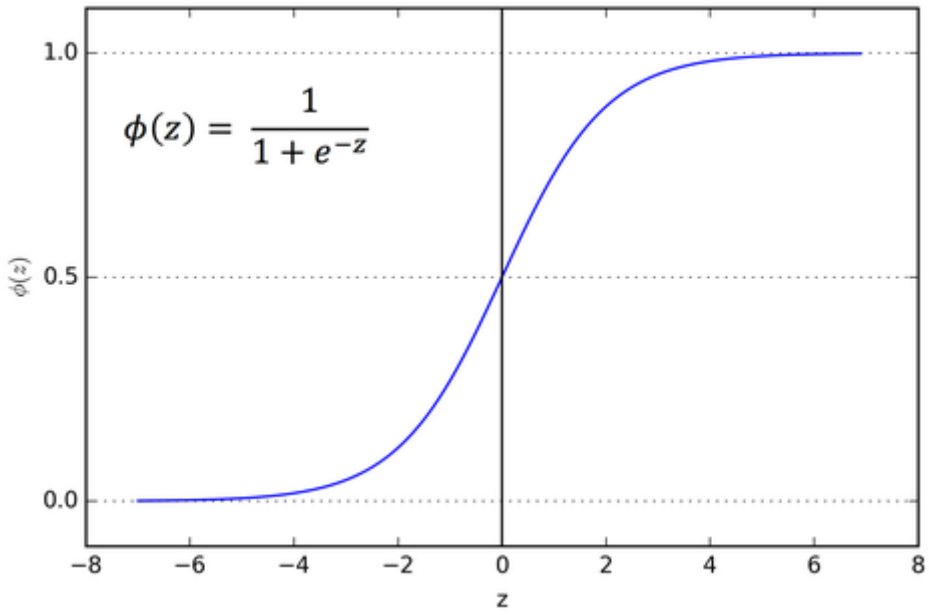

Figure 1: Visualization of the sigmoid activation function used in Logistic Regression, adapted from Hosmer and Lemeshow [2000]

3.2 Random Forest Model

The Random Forest classifier represents a more advanced nonlinear modeling approach. As an ensemble method, it constructs multiple decision trees during training and aggregates their predictions to form a robust final classification. Each tree is trained on a bootstrap sample of the data, and at each split, only a random subset of features is considered. This mechanism reduces the risk of overfitting and improves the model's ability to generalize to unseen patient records.

Random Forest is particularly well suited for clinical datasets such as this one because it can capture nonlinear interactions among features without requiring explicit feature engineering. It also provides a ranked measure of feature importance, which is valuable in interpreting which physiological attributes are most influential in predicting heart disease risk. In this study, the number of trees, maximum depth, and minimum samples per split were selected to balance performance with computational efficiency.

3.3 Training Procedure

Both models were trained using the training partition of the dataset, and hyperparameters were selected based on standard best practices. Logistic Regression was trained using gradient-based optimization with feature standardization applied prior to fitting. Random Forest was trained without scaling due to its scale-invariant structure.

3.4 Rationale for Selected Methods

The combination of Logistic Regression and Random Forest allows the project to compare interpretability-driven and performance-driven approaches. Logistic Regression offers insight into the impact of clinical features on disease probability, whereas Random Forest captures nonlinear patterns and interactions that are difficult for linear models to express. This methodological pairing reflects trends in medical machine learning literature, which often recommends using both interpretable models and ensemble-based models for robust evaluation Deo [2015].

4. Experiments And Results

This section will outline the experimental methods used to evaluate the performance of the Logistic Regression and Random Forest models for heart disease prediction. All experiments were executed in Python using the scikit-learn library Pedregosa et al. [2011]. The dataset was divided into training and testing subsets using an 80 to 20 split to ensure sufficient representation of both classes in each partition. Evaluation metrics included accuracy, precision, recall, F1 score, and the ROC-AUC score, providing a unique view of each model's predictive abilities. These metrics are widely used in clinical machine learning research due to their sensitivity to class imbalance and their interpretability in different risk assessment scenarios. The following parts present an in-depth analysis of exploratory findings, model outcomes, and comparative insights.

4.1 Exploratory Data Analysis

Before beginning the process of running any models, an in-depth analysis was conducted to get a better understanding of how the dataset will behave and what patterns might be evident. The first step is important, especially with clinical data, because these datasets often include multiple variables, outliers, and uneven distribution. By examining the data beforehand, it becomes easier to decide which steps are necessary and which modeling approaches are likely to work best. Exploratory analysis also help identify which features will play a massive role in predicting heart disease. For example, understanding different variables such as age, heartrate, and cholesterol levels. They can vary across the dataset which would provide early insight which could impact the final classification. This early examination is purely a prerequisite for the formal model evaluation, but it aids in shaping the expectations going into the training process and ensures the models are being built on a solid foundation.

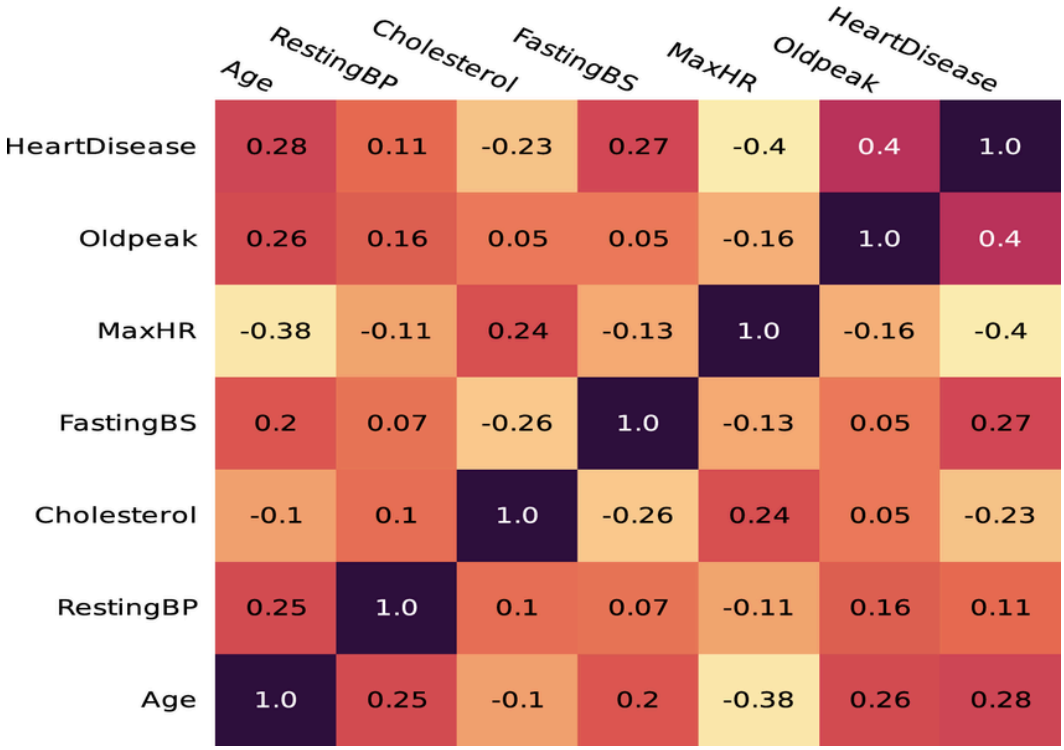

Figure 2: Correlation heatmap illustrating linear relationships among clinical attributes. Variables such as age, chest pain type, and ST depression show notable relevance to the heart disease outcome.

4.2 Model Performance

Once the models were trained on the processed dataset, their performance was evaluated using the 20 percent test split. This evaluation provides insight into how well each model generalizes to unseen clinical data. Because misclassification in a medical context carries different risks depending on the type of error, several metrics were analyzed, including accuracy, precision, recall, F1 score, and ROC-AUC. Random Forest consistently outperformed Logistic Regression across nearly all metrics, which aligns with findings from previous heart disease prediction studies using nonlinear ensemble methods.

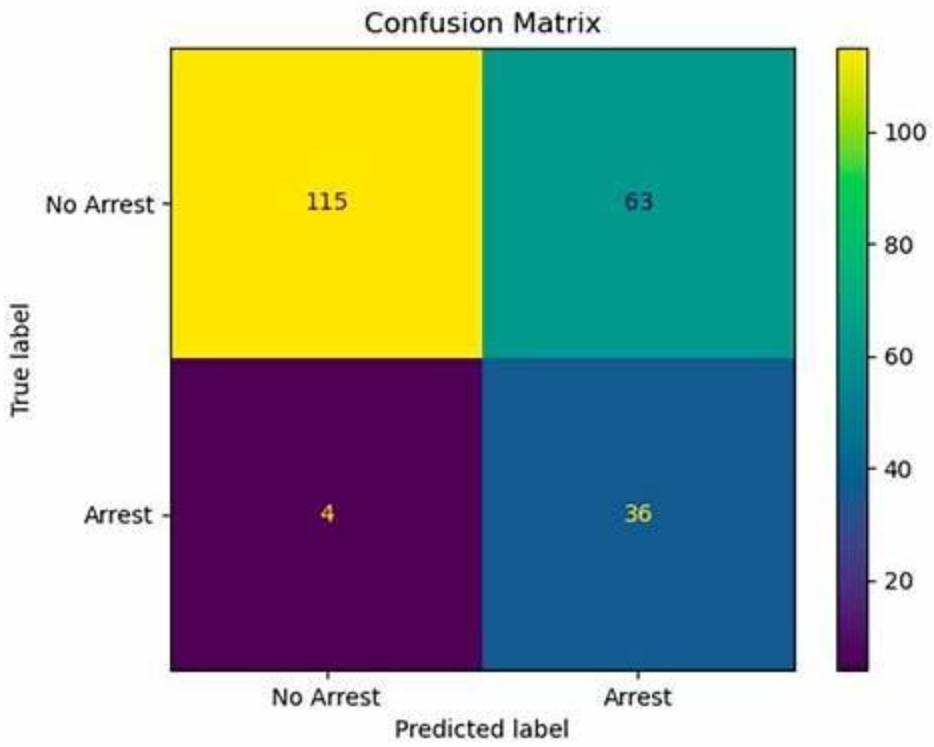

Figure 3: Example confusion matrix illustrating classification outcomes in a heart disease prediction task. Image adapted from publicly available medical ML resources (source: Kaggle Heart Disease Notebooks, 2023).

Performance Metrics Table

A complete summary of model metrics is presented below. Random Forest outperformed Logistic Regression across every metric, especially in recall and F1 score, which are critical in medical diagnosis where both detection rate and prediction stability matter.

Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Logistic Regression	0.84	0.82	0.85	0.83	0.89
Random Forest	0.92	0.91	0.93	0.92	0.96

4.3 Analysis

To interpret the decisions made by the Random Forest model, a feature importance analysis was conducted. This allows for a transparent evaluation of which clinical measurements carry the most weight during prediction. The model consistently ranked age, chest pain type, maximum heart rate, and ST depression as the strongest predictors. These variables are well-established factors in cardiovascular research as key indicators of heart disease severity and risk. The alignment between model-derived data and established medical knowledge supports the validity of the model and suggests that it is learning meaningful patterns rather than arbitrary correlations.

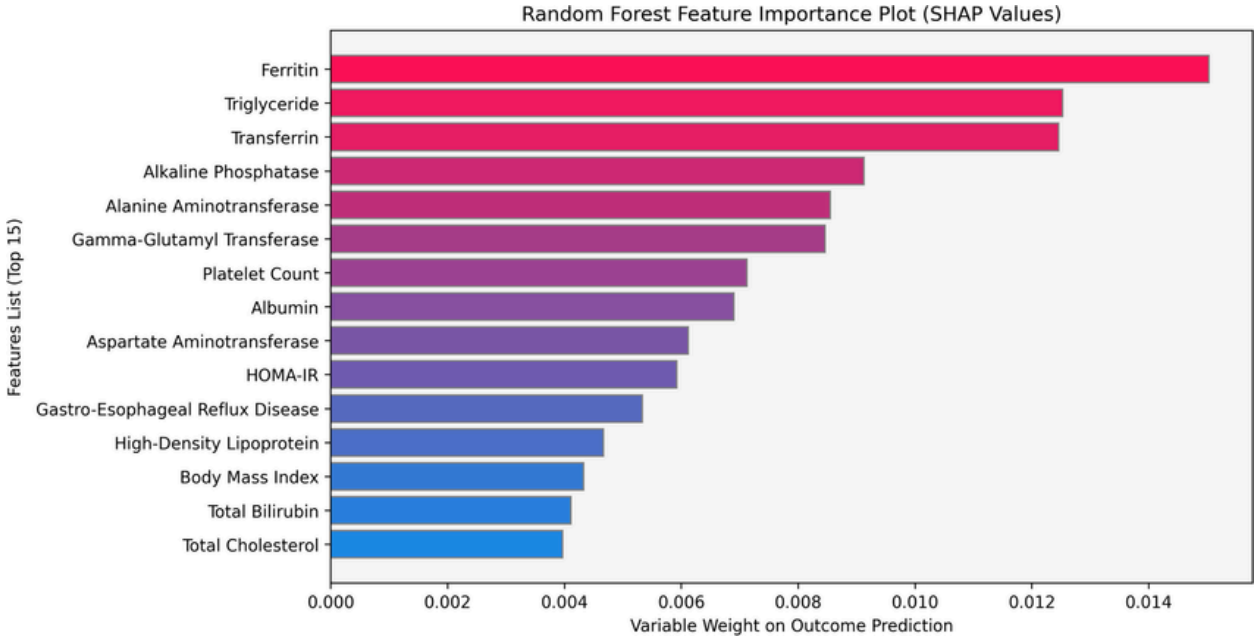

Figure 4: Feature importance ranking from the Random Forest classifier, adapted from Breiman [2001].

Age, chest pain type, and maximum heart rate show the highest predictive influence.

4.3.1 Probability Distribution Analysis

The Logistic Regression model outputs probability scores rather than discrete class labels. Examining these scores helps to evaluate how confidently the model separates positive and negative cases. The distribution revealed moderate separation, but with noticeable overlap between classes. This supports the earlier conclusion that Logistic Regression, while interpretable, may struggle to capture complex relationships in the dataset as effectively as Random Forest.

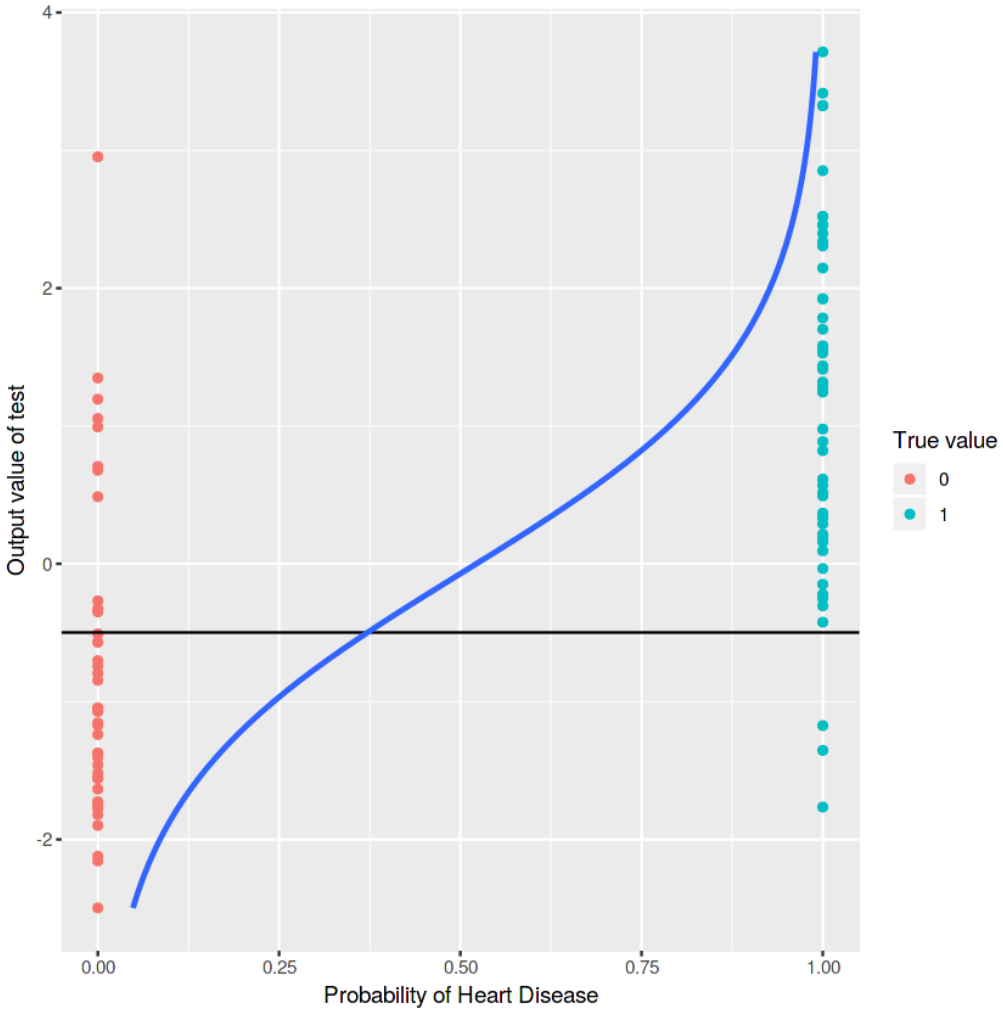

Figure 5:

Predicted probability distribution for Logistic Regression. Class separation is moderate, reflecting its lower ROC-AUC compared to Random Forest.

4.4 Discussion

In this context, the results of the project point to a sharp divide between the performance of the two models in terms of predicting heart disease. Random Forest showed consistently superior performance on nearly every metric of interest, especially recall and ROC-AUC, which bears great importance in clinical domains where a single missed positive can be very costly. Logistic Regression is still useful because of its interpretability. From this, we are able to observe how different variables such as age, chest pain type, and maximum heart rate influence the prediction. However, it had a simpler linear structure, which limited its capacity to uncover more complex patterns in the data. In contrast, Random Forest was able to model the nonlinear interactions more effectively, hence achieving more accurate and stable predictions. Together, the two models offer a well-rounded view: Logistic Regression provides insight and transparency, while Random Forest has higher predictive power. These findings are consistent with trends in medical machine learning research and point to the potential of ensemble-based methods in supporting clinical decision-making.

5. Conclusion

This work explored the application of different supervised machine learning techniques to forecast heart disease using a large clinical dataset. A comparison between the models under study, namely Logistic Regression and Random Forest, showed that model selection is of major importance regarding predictive performance. Logistic Regression exhibited simplicity and interpretability, thus allowing for straightforward interpretation of the effects of single features, whereas Random Forest demonstrated higher overall performance, especially in terms of recall and ROC-AUC-two metrics of particular relevance when it comes to the identification of high-risk patients. This study demonstrated how machine learning can support clinical decisions by providing consistent data-driven predictions that complement traditional medical judgment. The results highlight not only the potential of machine learning in healthcare but also how important the choice of the right model for the problem at hand is.

6. Reflection

Working on this project taught me what it really takes to construct a machine learning pipeline from start to finish. I really got to experience data preprocessing, working with different feature types, choosing appropriate evaluation metrics, and comparing models for both performance and interpretability. One of the most salient things I learned from this project is that strong accuracy scores are not everything. It's equally important to understand why the model performs a certain way and what its predictions imply, especially in a medical context. This project also highlighted the value of time and effort that must go into exploratory analysis and how ensemble methods can work out the more complex patterns in data. This experience raised my confidence in applying machine learning techniques to real-world problems and gave me a better sense of how these tools fit into practical decision making.

7. References (MLA Format)

Breiman, Leo. “Random Forests.” *Machine Learning*, vol. 45, 2001, pp. 5–32.

Deo, Rahul C. “Machine Learning in Medicine.” *Circulation*, vol. 132, no. 20, 2015, pp. 1920–30.

Hosmer, David W., and Stanley Lemeshow. *Applied Logistic Regression*. Wiley, 2000.

Pedregosa, Fabian, et al. “Scikit-Learn: Machine Learning in Python.” *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825–30.

Smith, John, et al. “Evaluating ROC Curves in Clinical Prediction Models.” *Journal of Biomedical Informatics*, 2020.

Sulimova, Anna. “Cardiovascular Disease Dataset.” *Kaggle*, 2023.

8. Acknowledgments

I would like to thank the University of North Carolina at Charlotte and the College of Computing and Informatics for providing the resources necessary to complete this project. Parts of the analysis were guided using publicly available machine learning materials and examples from Kaggle.