Using Machine Learning to Predict Heart Attack Risk

Cheyanne Gardner & Paul Robles

Northwood University: DeVos Graduate School of Business

MGT665: Solving Problems with Machine Learning

Dr. Itauma

May 4, 2025

Using Machine Learning to Predict Heart Attack Risk

**1-Abstract**

Heart disease is still one of the top causes of death around the world, which makes early detection really important. This project uses machine learning to predict the risk of a heart attack based on clinical and lifestyle data. We tested three models: Logistic Regression, Decision Tree, and Random Forest. Since the original data had more low-risk cases, we used SMOTE to balance it out and help the models better detect high-risk individuals. We measured how well each model performed using accuracy, precision, recall, and F1 score. Logistic Regression had the highest recall (0.866), which means it was best at catching high-risk cases, even though it also flagged more false positives. The Decision Tree offered a good balance between performance and understanding how it made predictions. Random Forest had the best accuracy but wasn’t as good at identifying high-risk cases. Overall, Logistic Regression seems most useful for early screenings, especially if backed up by follow-up testing. This shows how machine learning could play a big role in improving early detection and care in healthcare.

**2-Introduction:**

Heart attacks continue to be a leading cause of death worldwide, making early risk identification a critical priority in healthcare. This project applies machine learning techniques to predict an individual’s risk of experiencing a heart attack, using a combination of clinical indicators and lifestyle factors. The goal is to leverage data-driven models to improve early detection and enable more proactive interventions.

Classification models were explored, including Logistic Regression, Decision Trees, and Random Forests. To address class imbalances within the dataset, the Synthetic Minority Over-sampling Technique (SMOTE) was applied, helping to improve the models' sensitivity to high-risk cases. By evaluating model performance across key metrics such as Accuracy, Precision, Recall, and F1 score, this project aims to identify the most effective approach for real-world clinical applications.

The full project code is available here:

<https://github.com/CGardner1998/Final/tree/main>

**3-Related work:**

The first study that was looked at was a study in *Scientific Reports that* used machine learning to predict heart disease with 95.5% accuracy. The researchers used health data from four different places: Cleveland, Switzerland, Long Beach, and Hungary. They cleaned and organized the data using special techniques to improve the results. Then, they tested how well their model worked using common evaluation tools like accuracy and precision. The study showed that machine learning can be a powerful tool for predicting heart problems (Paharia, 2024).

The second study, found on the National Institutes of Health website, used machine learning to look at heart health risks using data from yearly work checkups. The model created personal risk charts to help doctors catch problems early. The study shows how regular health data can be used to help prevent heart disease with the help of machine learning (Maatheson, 2022).

The last study that was looked at was predicting heart disease risk using logistic regression. One thing that stood out from this study was that the model did not have a high sensitivity; the threshold was also changed. From 0.5 to 0.2. There was no noted difference. This does make sense as all the data that we are dealing with does lead to negative outcomes and the models are not going to have enough information to have positive outcomes. (Evelyn, 2023)

**4-Methodology:**

The dataset was imported from the local directory. Data cleaning included handling missing values through median imputation, encoding categorical variables numerically, and splitting the dataset into training and testing sets using an 80/20 ratio.

Three classification models were used: Logistic Regression, a tuned Decision Tree, and a Random Forest. Model performance was evaluated using four key metrics: Accuracy, Precision, Recall, and F1 Score.

Initial results indicated poor performance, particularly in terms of precision and recall, suggesting that class imbalance between low-risk and high-risk patients was negatively impacting model sensitivity. To address this, SMOTE was applied to the training data to balance the classes.

Following the application of SMOTE, the models were retrained on balanced data. This adjustment led to improvements in recall and overall predictive performance, especially in identifying high-risk individuals.

**5-Results:**

A screenshot of a computer code

AI-generated content may be incorrect.

5.1 The Logistic Regression model, after adjusting the classification threshold, achieved the highest recall but at the cost of lower accuracy. This indicates that while the model successfully identified most high-risk individuals, it also produced a higher number of false positives.

5.2 The tuned Decision Tree model offered a better balance between recall and overall accuracy, leveraging its optimized parameters to improve interpretability and moderate predictive performance.

5.3 The Random Forest model, while achieving the highest overall accuracy, had the lowest recall, suggesting it was less effective at detecting high-risk cases in this healthcare context.

Based on these results, the threshold-adjusted Logistic Regression is most effective at maximizing recall, while the tuned Decision Tree offers a better balance between sensitivity and overall accuracy. The tuned Random Forest, despite its higher accuracy, is less suitable for clinical use due to its lower recall rate.

Also, while running the model, a warning appeared indicating that the system could not locate the number of physical CPU cores due to a missing file, so it defaulted to using the logical core count instead. This did not have an impact on the model’s performance.

**6-Discussion:**

To interpret the results of this study, it is important to first understand what the evaluation metrics are and how they impact the data. Accuracy is how well the model predicts the correct class. Precision is the proportion of true positives among the predicted positives. Recall is the proportion of actual positives correctly predicted, and F1-Score is the harmonic mean of precision and recall.

Logistic Regression: We can see that it has a high recall at 0.866 which is ideal for screening patients. However, it had a low precision at 0.352 and an accuracy of 0.381 which can have a lot of false positives.

Decision Tree: This model had a moderate recall of 0.436 and an accuracy of 0.501

Random Forest: This model had the highest accuracy at 0.544. However, the recall was extremely low at 0.232. This model would not be suitable for clinical use because of the many missed diagnostics.

**6.1-Key Metrics:**

Accuracy: Random Forest had the highest accuracy among the rest of the models at 0.544 which is relatively low and can reflect certain imbalance challenges.

Precision: Throughout all the models the precision was between 0.315 and 0.352. This indicates that there are high false positives.

Recall: Logistic Regression was by far the highest in this metric at 0.866 which can be critical for medical applications.

F1 Score: Logistic Regression was also in the lead with a score of 0.501 which balances recall and precision.

**6.2-Implications:**

Looking at these models for a bigger picture and using them clinically, Logistic Regression would be viable for screening patients. This should not be the single determining factor; follow-up tests are extremely encouraged. Decision Tree and Random Forest should not be used clinically because they both have low recall. These findings suggest significant implications for healthcare decision-making. We can say that high recall in Logistic Regression can potentially reduce mortality, but the false positives can strain resources. The other models did not perform well, and their usage is limited.

**6.3Limitations:**

Some limitations that could have impacted the dataset was that SMOTE was used on this dataset. Using SMOTE was critical to helping improve the model’s sensitivity to high-risk cases. However, it could have introduced noise which may have affected the recall on the model Random Forest. Some of the dropped features such as “Country” and “Continent” could have been relevant to this project.

**6.4-Areas for Future Improvement:**

Improving Data: Improving the current data that was used can be done by creating easy-to- use features. For instance, Systolic and Diastolic blood pressure can be combined and used as an average blood pressure as these two numbers go hand in hand. With that being said, it would be a good idea to see if those features are so alike that they can throw off the data. The features “Country” and “Continent” should be used to group the patients even further. The patients can be placed into a high- or low-income region that can show us if their access to health care matters when it comes to accessing their heart risk.

Refining models: Trying different thresholds can be a good idea to see if there are any places where there is a good balance, where there is a high recall and better precision. Another thing that can also be done is setting a higher “min\_samples\_leaf”, which potentially avoids overfitting and makes the tree less particular about the data that it is looking for.

**7-Conclusion:**

This heart attack risk prediction project evaluated Logistic Regression, Decision Tree and Random Forest on the dataset provided by Kaggel (Banerjee, 2024). The high recall of 0.866 that was generated by Logistic Regression made this model suitable for screening. However, the low precision of 0.352, and low accuracy of 0.381 would require patients have follow-up tests. Decision Tree had a recall of 0.436 which had some decent interpretable rules, but it missed a lot of at-risk patients. Lastly, Random Forest had a low recall of 0.232 which is unusable in this setting even though the accuracy was decent at 0.544. SMOTE did address the imbalance, but it could have potentially added noise. If data is improved along with tuning models, the project can become a valuable tool for early at-risk heart attack detection, which would eventually save lives and optimize the healthcare resources out there.

**References:**

Banerjee, S. (2024, May 11). *Heart attack risk prediction dataset*. Kaggle. <https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset/data>

Evelyn. (2023, July 9). *Predicting heart disease risk using logistic regression*. Medium. [https://medium.com/@evelyn.eve.9512/predicting-heart-disease-risk-using-logistic-regression-7ed58ae2abfd#:~:text=Logistic%20regression%20is%20a%20method,given%20dataset%20containing%20independent%20variables](mailto:https://medium.com/@evelyn.eve.9512/predicting-heart-disease-risk-using-logistic-regression-7ed58ae2abfd#:~:text=Logistic%20regression%20is%20a%20method,given%20dataset%20containing%20independent%20variables).

Matheson, M. B., Kato, Y., Baba, S., Cox, C., Lima, J. A. C., & Ambale-Venkatesh, B. (2022, November 11). *Cardiovascular risk prediction using machine learning in a large Japanese cohort*. Circulation reports. <https://pmc.ncbi.nlm.nih.gov/articles/PMC9726526/>

Paharia, P. T. (2024, April 5). *New Machine Learning Model achieves breakthrough in heart disease prediction with over 95% accuracy*. News. <https://www.news-medical.net/news/20240405/New-machine-learning-model-achieves-breakthrough-in-heart-disease-prediction-with-over-9525-accuracy.aspx>