Applying Statistical Learning Methods

to Heart Attack Risk Prediction

By Isaac Mower
April 25th, 2025

Introduction

Predicting heart attack risk using machine learning presents a complex but important challenge in medical data analysis. In this paper, a range of classification models were trained to evaluate their ability to identify individuals at risk and to determine important predictors in those at risk of having a heart attack, each selected based on their unique strengths in handling the noisy, nonlinear, and often imbalanced nature of clinical datasets. These included traditional statistical models like Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA), which offer interpretability and computationally efficient results, as well as non-parametric methods like k-Nearest Neighbors (k-NN) that detect local patterns in the data. More advanced techniques such as Random Forests and LASSO logistic regression were chosen for their abilities to model complex interactions and perform embedded feature selection, respectively. Finally, Support Vector Machines (SVMs) were introduced to test their flexibility in constructing optimal decision boundaries in high-dimensional space. Throughmulti-faceted evaluation of these models, the goal was not only to assess predictive performance but also to explore the feasibility of machine learning as a tool to assist in early identification of heart attack

risk, where clinical stakes are high and false negatives can be life-threatening[1]. An important secondary goal of this paper is to identify important features in predicting heart attack risk.

The goal of this paper is to implement a variety of statistical learning methods to a heart attack risk data set in order to create a model that individuals can quickly use to determine if they are at risk of having a heart attack. After using the model, they can then go see a doctor for more in-depth things like an actual diagnosis and lifestyle change recommendations. To achieve this, a variety of classification models were employed in this study to predict heart attack risk, each selected for its unique strengths in addressing the challenges inherent in medical data. Linear Discriminant Analysis and Quadratic Discriminant Analysis were included as interpretable baseline methods that assume normally distributed predictors, with QDA offering added flexibility through class-specific covariance structures. The k-Nearest Neighbors algorithm, a non-parametric method, was used to capture local patterns in the data based on proximity in feature space, making it suitable for nonlinear relationships. Random Forests were chosen for their robustness to noise, ability to capture nonlinear interactions, and built-in feature importance measures, which are valuable in identifying key clinical variables. LASSO logistic regression was applied for its dual role in prediction and feature selection, particularly useful in high-dimensional settings where model simplicity and interpretability are essential. Finally, Support Vector Machines were selected for their capacity to construct flexible decision boundaries and generalize well in high-dimensional spaces, especially when enhanced with kernel functions and class imbalance handling. To train and tune these models I used the

---

[1] Richard Cummins and Mary Hazinzki. *Guidelines Based on Fear of Type II Errors.* August 22nd, 200. American Heart Association Journals. https://www.ahajournals.org/doi/10.1161/circ.102.suppl_1.I-377

following packages in Python: pandas, sklearn, numpy, and matplotlib. Collectively, these models offer a comprehensive framework for evaluating predictive performance and identifying important risk factors in a clinically meaningful context.

Data

The data set used to train and evaluate the models consisted of 9,652 observations of 25 different predictor variables ranging from lifestyle descriptors to blood/heart measurements. The response variable was binary, with a 1 being at risk for a heart attack and 0 being not at risk of a heart attack. To begin the modeling process, the dataset was subjected to a series of careful data cleaning and modification steps to ensure its suitability for statistical learning techniques. First, a duplicate response variable was identified and removed to prevent redundancy and data leakage during model training. In addition, the 'Gender' variable, originally a categorical feature with values 'Male' and 'Female,' was encoded numerically, with 'Male' assigned a value of 1 and 'Female' assigned a value of -1. This transformation preserved the categorical meaning while rendering the data compatible with the mathematical operations required by machine learning algorithms.

After that, the dataset was examined for missing values. Any missing observations were imputed using the median value of the corresponding feature, a robust choice that is less sensitive to outliers compared to mean imputation. This approach helped maintain the integrity of the data distribution while avoiding loss of samples through row-wise deletion. The data was then partitioned into training (60%), validation (20%), and test (20%) sets, using random sampling to preserve the imbalanced class distribution present in the response variable. All

cleaning decisions were made to maximize data quality and to ensure the fair, unbiased evaluation of all subsequent models.

LDA

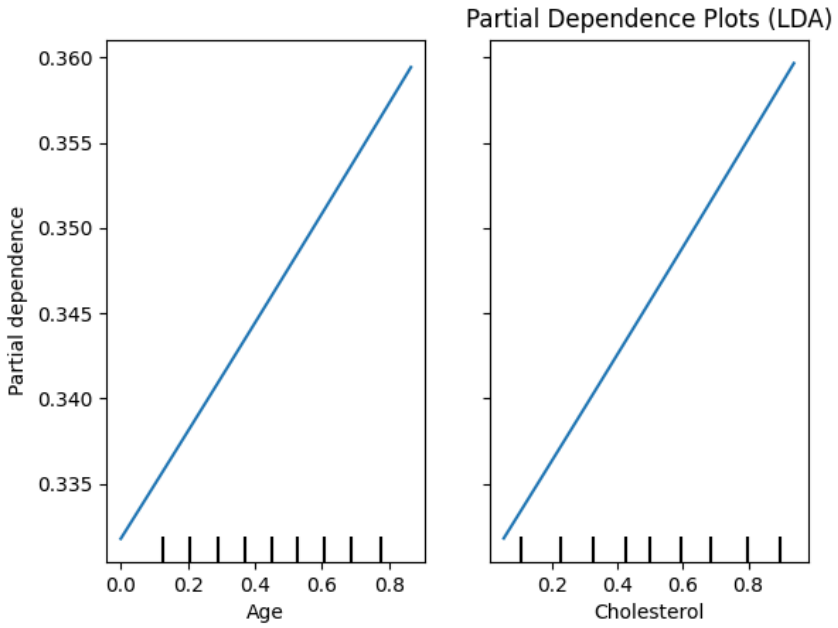


Figure 1: Partial dependence plot for LDA

Model	Accuracy	Recall	Specificity	Precision	F1 Score	MSE	AUC
LDA	0.655	0	1	0	0	0.223	0.531

Table 1: Accuracy metrics for LDA

Linear Discriminant Analysis was implemented as a baseline classification method for predicting heart attack risk. LDA assumes that the predictor variables are normally distributed within each class and that the covariance matrices are identical across classes, leading to linear decision boundaries. By modeling the class conditional distributions and using Bayes' theorem, LDA classifies observations based on maximizing the posterior probability of class membership.

After training on the scaled training data, LDA was evaluated on the test set to establish a reference point for model complexity and classification performance against more flexible algorithms. As seen in *Table 1,* LDA performed terribly. The model simply said that no one was at risk of a heart attack and left it at that. As a result there is no reason to use this model because randomly guessing would likely perform just as well. The partial dependence plots in *Figure 1*, show the rise in heart attack risk with the two most significant variables (Age and Cholesterol) in LDA. While our model performs poorly, it is a good sign that it is finding a statistically significant relationship between predictors like Age and Cholesterol that medical experts use when determining who is at risk of having a heart attack.

QDA

Model	Accuracy	Recall	Specificity	Precision	F1 Score	MSE	AUC
QDA	0.612	0.278	0.788	0.408	0.331	0.240	0.547

Table 2: Accuracy metrics for QDA

Quadratic Discriminant Analysis was also implemented to relax the assumption of equal covariance matrices made by LDA. By allowing each class to have its own covariance matrix, QDA is capable of producing quadratic decision boundaries, thereby capturing more complex relationships between the predictors and the heart attack risk outcome. While QDA is inherently more flexible and can model class heterogeneity better than LDA, it typically requires more data to accurately estimate the separate covariances without overfitting. Our dataset had plenty of data with 9,652 samples. As with LDA, QDA was trained on the training data and evaluated on the test set to provide a direct performance comparison and to understand the trade-off between model flexibility and variance. *Table 2* shows the results of this evaluation. The results are not

good. However, QDA performed best out of the models that were tested. While the model had the lowest overall accuracy at 61.2%, it had the highest recall at 27.8%. While neither of these metrics are in any way competent or useful, it does show that QDA attempted to properly assess who is at risk for a heart attack instead of just saying no one is at risk, which is much more useful than models that predicted everyone to not be at risk of a heart attack.

k-Nearest Neighbours

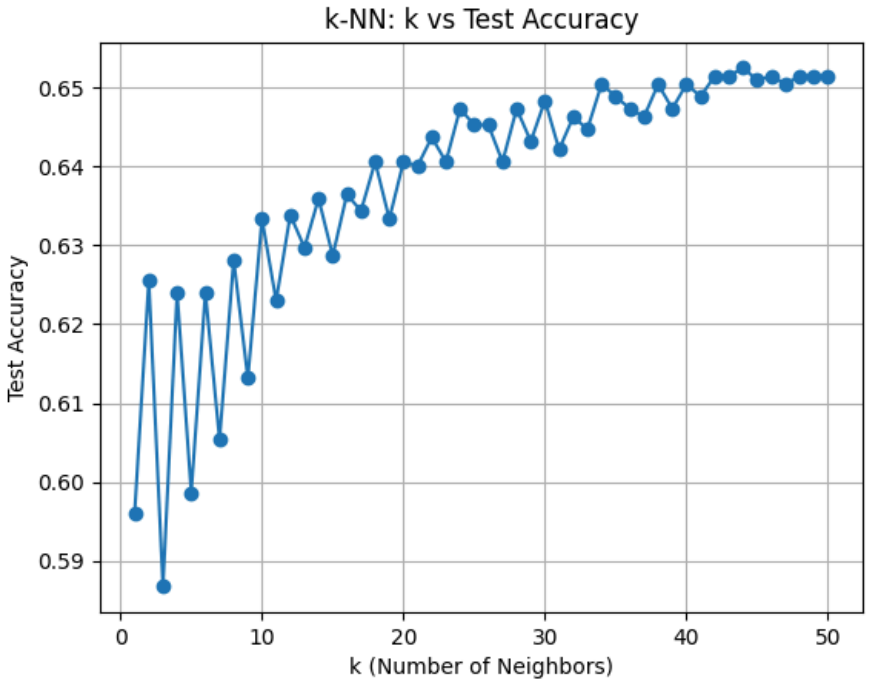


Figure 2: Number of Neighbours vs. Test Accuracy

Model	Accuracy	Recall	Specificity	Precision	F1 Score	MSE	AUC
k-NN, k=25	0.645	0.071	0.949	0.420	0.121	0.227	0.555

Table 3: Accuracy metrics for k-NN

The k-Nearest Neighbors classifier was explored by evaluating a wide range of k-values from 1 to 50, aiming to identify the optimal neighborhood size that balances bias and variance. After analyzing the test set performance across different values of k, a k-value of 25 was selected based on its relatively strong test accuracy and the diminishing returns in increased k's, as shown in *Figure 2*. A final k-NN model with k=25 was then trained on the training data and evaluated on the test set. Scaling of the features was crucial for k-NN due to its reliance on distance calculations. Overall, k-NN provided a simple, non-parametric benchmark method for comparison with more sophisticated classifiers. As shown in *Table 3,* K-NN did not perform well and had a low overall accuracy of 64.53%. K-NN also had an abysmal Recall of 7.05%. K-NN did not perform well in any of the other metrics that were evaluated except for Specificity. K-NN had a high specificity of 94.9%. High specificity is explained when we look at the confusion matrix and see that k-NN predicted the overwhelming majority of people to not be at risk of a heart attack.

Random Forests

A Random Forest classifier was constructed to leverage the ensemble power of multiple decision trees trained on bootstrapped samples of the data. Each split within a tree was made by considering a random subset of features, which helps to decorrelate the individual trees and improves generalization. The initial Random Forest model was trained using all available features, which capitalizes on the method's ability to handle high-dimensional data, capture complex nonlinear interactions, and offer robustness to outliers and noise.

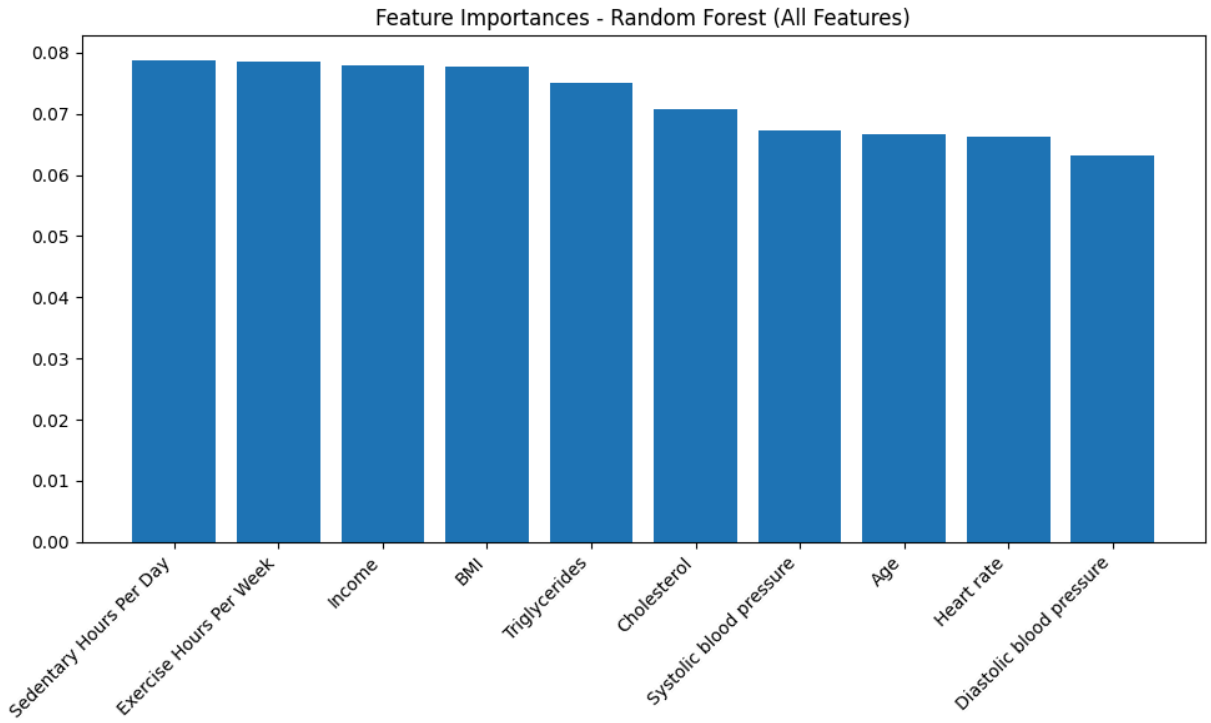


Figure 3: Feature importance according to Random Forests

Model	Accuracy	Recall	Specificity	Precision	F1 Score	MSE	AUC
Random Forests	0.664	0.074	0.976	0.613	0.131	0.214	0.603
Random Forests Top 10	0.665	0.120	0.953	0.571	0.198	0.210	0.615

Table 4: Accuracy metrics for Random Forests

To further enhance Random Forest's interpretability and potentially improve performance, feature selection was performed based on the computed feature importances from the full Random Forest model. The top ten most important features, ranked by their contribution to reducing impurity, were selected, and a new Random Forest was trained using only this reduced feature set. According to *Figure 3*, the ten most important features in descending order were: Sedentary Hours Per Day, Exercise Hours Per Week, Income, BMI, Triglycerides, Cholesterol, Systolic Blood Pressure, Age, Heart Rate, and finally, Diastolic Blood Pressure. This second model performed just as poorly as the first. *Table 4*, shows that both of these models

had an accuracy in the mid 66%s. Random Forests had disgusting Recall values, with the Top 10 Features Forest having the higher Recall at 12%. Neither of these metrics are impressive in any way, and the rest of Random Forests' metrics follow this unimpressive pattern.

LASSO Logistic Regression

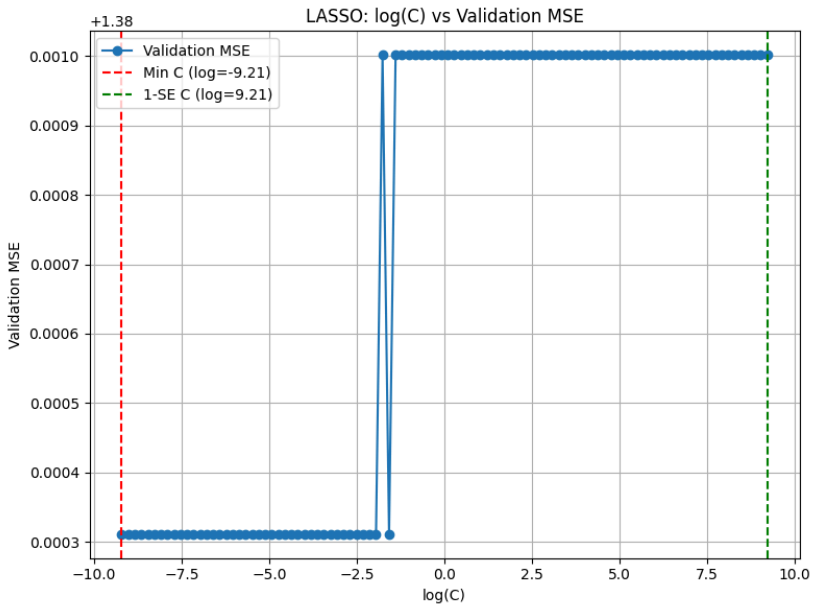


Figure 4: Validation MSE vs natural log of regularization parameter

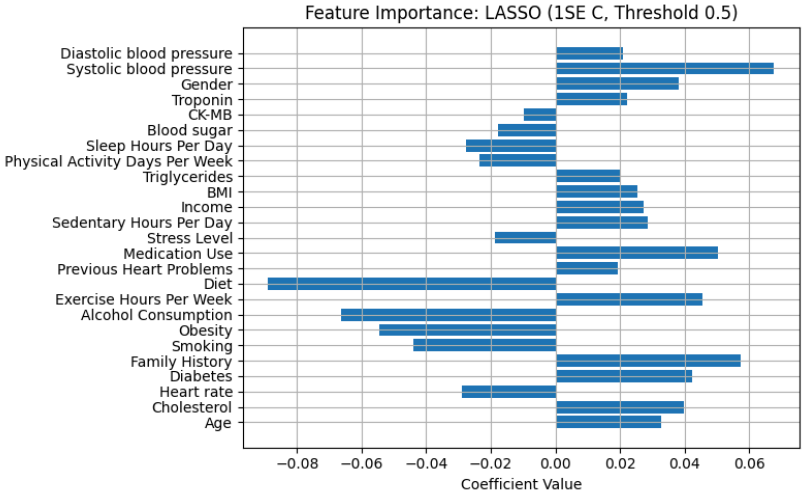
LASSO logistic regression was implemented to perform simultaneous classification and embedded feature selection by penalizing the absolute magnitude of the model coefficients. A wide range of regularization strengths was evaluated using five-fold cross-validation to find the value of the penalty parameter (lambda) that minimized the validation mean squared error. In addition, a second lambda value was selected using the 1-standard-error (1-SE) rule, which favors simpler models that achieve performance within one standard error of the best validation score. This dual approach balances predictive accuracy with less in-depth models, providing interpretable feature sets while minimizing overfitting. *Figure 4* shows us the Validation MSE

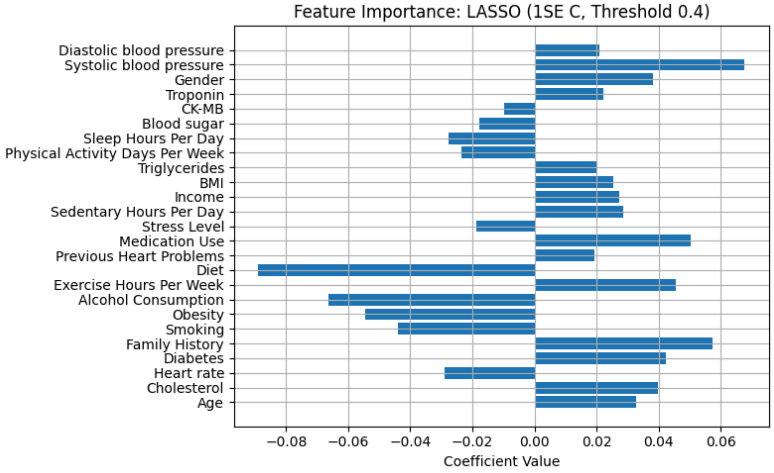
against the natural log of a variety of regularization hyperparameters along with the selected minimum regularization parameter and the regularization parameter one standard error away from the minimum.

Model	Accuracy	Recall	Specificity	Precision	F1 Score	MSE	AUC
Lambda_min (Thresh 0.5)	0.655	0	1	0	0	0.226	0.500
Lambda_1se (Thresh 0.5)	0.654	0	0.999	0	0	0.227	0.520
Lambda_min (Thresh 0.4)	0.655	0	1	0	0	0.226	0.500
Lambda_1se (Thresh 0.4)	0.621	0.118	0.886	0.354	0.178	0.227	0.520

Table 5: Accuracy metrics for the four LASSO models

Four LASSO models were ultimately trained to explore how different combinations of regularization strength and decision thresholds affect performance. Two models used the lambda that minimized validation error (one with a 0.5 threshold, one with a 0.4 threshold), and two models used the 1-SE lambda with the same threshold options. Adjusting the threshold was crucial in the context of heart attack risk prediction: a threshold of 0.5 maintains a balanced perspective between sensitivity and specificity, while a threshold of 0.4 emphasizes sensitivity. Lowering the threshold increases the likelihood of predicting a positive class (at-risk patients), reflecting the clinical reality that false negatives (failing to identify high-risk patients) are considerably more dangerous than false positives.





Figures 7 (top) and 8 (bottom): Coefficient values for predictors of heart attack risk for different LASSO models

Feature selection via LASSO was valuable, as many predictors in the dataset could contribute noise or redundancy. By shrinking less important coefficients to zero, LASSO isolated a subset of predictors most strongly associated with heart attack risk. Analyzing the nonzero coefficients from each model provided further insights into which clinical and lifestyle variables were consistently important across different model specifications, supporting both model

interpretability and potential clinical relevance. Both *Figures 7 and 8* show that the most significant predictors of heart attack risk were Diet, Alcohol Consumption, Systolic Blood Pressure, and Obesity. These significant predictors are pretty consistent with medical experts findings in the area[2].

Support Vector Machines

Support Vector Machines were implemented to build a flexible classifier capable of separating classes through both linear and nonlinear decision boundaries. Initially, a baseline SVM with a radial basis function (RBF) kernel was trained using standardized input features. The RBF kernel projects the data into a higher-dimensional space to capture complex patterns, which is particularly useful for nonlinear relationships among predictors. However, in the initial untuned configuration, the model struggled with recall, which reflects the difficulty in capturing the at risk class.

To improve classification performance and account for the imbalance in the binary response variable, a systematic hyperparameter tuning procedure was conducted using grid search with ten-fold cross-validation. The grid included a range of penalty parameters (C values) and kernel types ('linear' and 'rbf'). To ensure visibility into model progress and reduce training time, the search was optimized by restricting the grid to a reasonable number of candidate configurations. Additionally, the use of class_weight='balanced' in the SVM classifier helped mitigate the dominance of the majority class by reweighting the penalty associated with

[2] Multiple Authors. *Heart Disease Risk Factors*. CDC. https://www.cdc.gov/heart-disease/risk-factors/index.html

misclassifying the minority class. This is particularly relevant in medical applications such as heart attack risk prediction, where failing to identify an at-risk patient (a false negative) can be far more consequential than issuing a false positive.

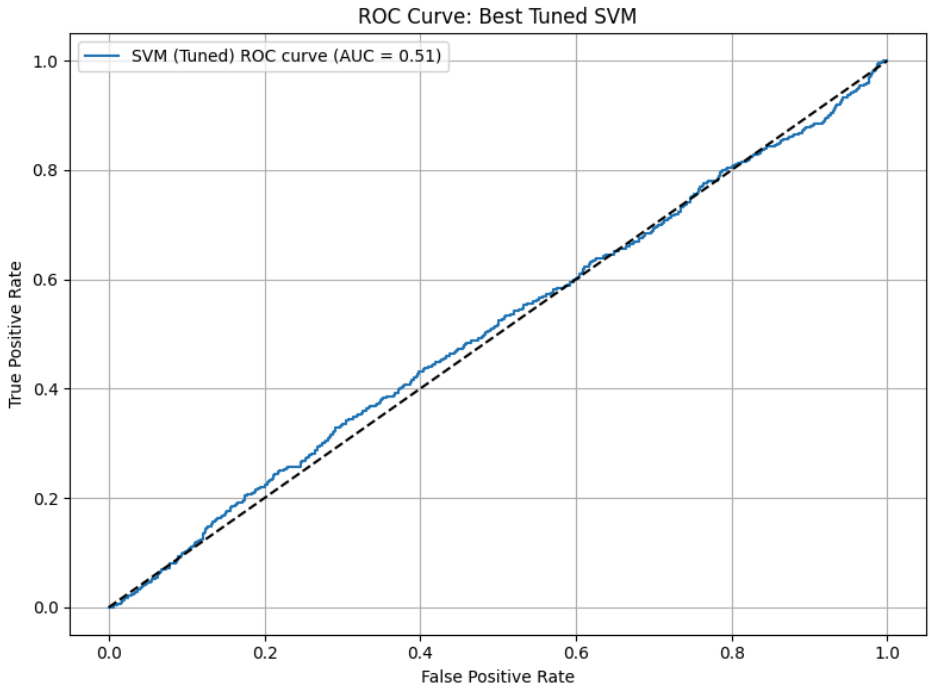


Figure 9: ROC curve for tuned SVM

Model	Accuracy	Recall	Specificity	Precision	F1 Score	MSE	AUC
Basic SVM	0.655	0	1	0	0	0.224	0.569
Tuned SVM	0.655	0	1	0	0	0.226	0.508

Table 6: Accuracy metrics for SVMs

The best-performing SVM configuration was selected based on cross-validated accuracy and then evaluated on the held-out test set. While SVMs do not naturally provide measures of feature importance, their strength lies in the ability to construct clear and maximally separating decision boundaries between classes. However, this strength was not displayed here. *Figure 9* shows that the best tuned SVM performs only slightly better than random chance most of the

time. The rest of the time it performs as well as or worse than random chance. *Table 6* shows that in both SVMs, the models classified everyone as not at risk for a heart attack. This resulted in the exact same poor metrics for both SVMs. Accuracy was deceptively high due to the model's default to the majority class and the recall was zero, making the model clinically useless. Despite the implementation of class balancing and hyperparameter tuning, the SVM ultimately failed to capture meaningful distinctions in the at-risk class.

Results

Model	Accuracy	Recall	Specificity	Precision	F1 Score	MSE	AUC
LDA	0.655	0	1	0	0	0.223	0.531
QDA	0.612	0.278	0.788	0.408	0.331	0.240	0.547
k-NN, k=25	0.645	0.071	0.949	0.420	0.121	0.227	0.555
Random Forests	0.664	0.074	0.976	0.613	0.131	0.214	0.603
Random Forests Top 10	0.665	0.120	0.953	0.571	0.198	0.210	0.615
Lambda_min (Thresh 0.5)	0.655	0	1	0	0	0.226	0.500
Lambda_1se (Thresh 0.5)	0.654	0	0.999	0	0	0.227	0.520
Lambda_min (Thresh 0.4)	0.655	0	1	0	0	0.226	0.500
Lambda_1se (Thresh 0.4)	0.621	0.118	0.886	0.354	0.178	0.227	0.520
Basic SVM	0.655	0	1	0	0	0.224	0.569
Tuned SVM	0.655	0	1	0	0	0.226	0.508

Table 7: Accuracy metrics for all models

According to *Table 7,* the models performed poorly in every category but specificity. This is because in most of the models that were tested, (6 out of the 11) one or zero people were predicted to be at risk for a heart attack. This is obviously problematic. A possible explanation is the imbalance in our response variable. Only 35.5% of the people in our dataset were at risk for a heart attack. While the prior probabilities of the response variable were accounted for whenever possible, it appears to have had little to no positive effect on the predictive ability of the models. The other main metric I will be evaluating the models on is Recall. This is because in this scenario, telling someone that they are not at risk for a heart attack when they are at risk is much

worse than telling someone they are at risk for a heart attack when they are not at risk. The models had an atrocious average Recall of 6.01% with the highest Recall belonging to QDA with a Recall of 27.8%. These horrendous numbers indicate the poor predictive ability of the above models in correctly identifying people who are at risk of having a heart attack.

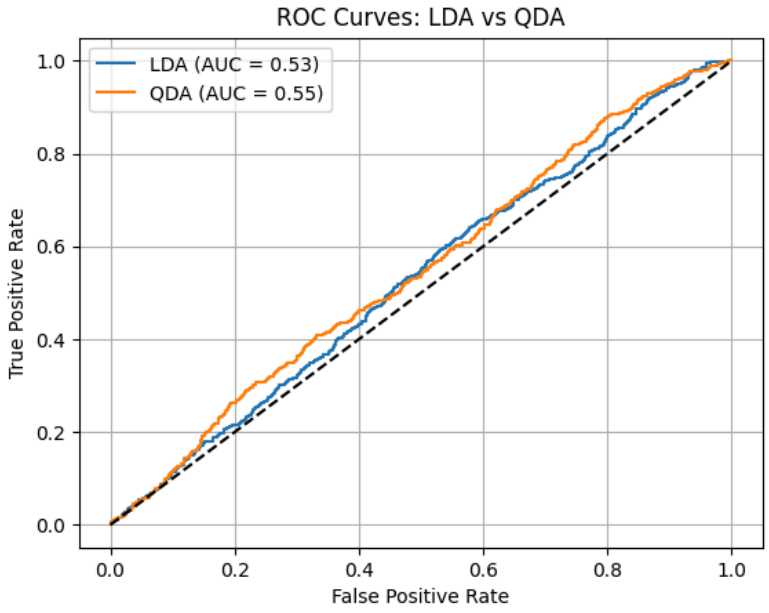


Figure 10: ROC curve of both LDA and QDA

Figure 10 further illustrates this poor model performance by showing the ROC curves of both LDA and QDA. The curves show that both LDA and QDA were only slightly better than randomly guessing. We can see a similar poor ROC curve in *Figure 9.* Interestingly enough, QDA performed best out of the models that were tested. While the model had the lowest overall accuracy at 61.2%, it had the highest recall at 27.8%. While neither of these metrics are in any way competent or useful, it does show that QDA attempted to properly assess who is at risk for a heart attack instead of just saying no one is at risk. The other "best" performing model was Random Forests with the 10 most significant features. This model had the highest accuracy of all

11 models with an accuracy of 66.5%. It also made an attempt to predict who is at risk for a heart attack. However, this was a poor attempt as the model only achieved a recall of 12.0%, which was the second highest recall among all of the models.

This overall trend highlights a critical flaw in how most models approached the classification task: favoring the majority class (not at risk) at the expense of identifying those truly at risk. In medical applications like this one, such behavior is not just a statistical issue but a real-world hazard. Despite trying different techniques, including using only the most significant features or adjusting for class imbalance, the models consistently failed to generalize well to the minority class. This could indicate that either the available features lack the necessary predictive power to distinguish those at risk, or the models themselves are not well-suited for handling imbalanced data in this context. Moving forward, alternative approaches, such as oversampling the minority class, using anomaly detection methods, or exploring more advanced ensemble techniques, may be necessary to improve the recall and overall reliability of these models.

Conclusion

One challenge of creating the models was maximizing recall because a false negative is very problematic when determining who is at risk for having heart attacks. I tried a variety of different probability thresholds and eventually settled on 0.4 since the recall plateaued when the threshold was lower, and the accuracy/sensitivity plummeted lower after that point. In the future, I would train and tune a more time-consuming and accurate model like neural networks. Neural networks typically outperforms the models I used, though they are more computationally expensive. I have no reason to believe they would not increase accuracy here because the

accuracy could not get much worse. I would also try training models with balanced classification data to avoid the poor predictive capability of the models in this paper. One major limitation of my analysis is that tI am not an expert in heart attack risk and I did not have access to experts in this field. This access would have allowed me to make more informed decisions throughout the model training process, as well as have a better interpretation of the models' results.

Despite the diversity of methods tested, the models in this analysis performed poorly overall, particularly in identifying patients who are truly at risk of a heart attack. While most models achieved high specificity, this came at the cost of alarmingly low recall, with the average recall across all models just 6.01%. This is especially problematic in a healthcare setting, where failing to flag high-risk individuals (false negatives) could have serious consequences. The best-performing model, QDA, only reached a recall of 27.8%, while most other models hovered close to zero. Class imbalance likely contributed to these results, with just 35.5% of observations labeled as "at risk," and many models defaulting to predicting the majority class. While adjusting probability thresholds helped marginally, the results suggest a need for more powerful approaches. Ultimately, this paper showcases both the potential and the limitations of traditional machine learning methods in critical health prediction tasks, an area where further research could yield massive benefits.

References

Multiple Authors. *Heart Disease Risk Factors.* CDC. https://www.cdc.gov/heart-disease/risk-factors/index.html

Richard Cummins and Mary Hazinzki. *Guidelines Based on Fear of Type II Errors.* August 22nd, 200. American Heart Association Journals. https://www.ahajournals.org/doi/10.1161/circ.102.suppl_1.I-377