**Predicting myocardial infarction risk using supervised and unsupervised machine learning methods**

Katherine Shi (kshi3), Katelyn Bechler (kbechler), Vy Ho (vivianho)

## 1 Abstract

Myocardial infarction is the leading cause of death in the United States, while disparities in cardiovascular care signal a need for improved methods of risk stratification. Using publicly available data from 303 cardiovascular patients, we explored patient phenotypes with k-means and DBSCAN clustering methods and trained classification models including logistic regression, gradient boosted trees, SVM, elastic net, random forest, and neural networks to predict MI risk. The best performing model was logistic regression, with auROC of 0.87, and sensitivity of 0.94.

## 2 Introduction

Myocardial infarction (MI) is the leading cause of death in the United States, incurring over 2,300 deaths per day (1). Despite well-established guidelines on risk reduction, sex disparities exist in both preventative and interventional cardiovascular care. Women with high risk of MI are less likely to receive preventative medications such as statin treatment for high cholesterol, and those who develop an MI wait longer for life-saving surgery than their male counterparts (2). These findings highlight a need for new, more inclusive risk stratification methods for MI prediction in order to better inform MI prevention. We first apply unsupervised clustering methods to a publicly available dataset to characterize MI risk phenotypes, followed by supervised learning methods to build a classifier to assess individual heart attack risk. We hope to improve early detection and treatment of MI in patients with atypical presentations.

## 3 Related Work

Previous efforts to predict MI risk include MI-3, which used a gradient boosting machine model to predict likelihood of MI in patients admitted to the hospital with serial troponin monitoring (3). As our model is designed to be deployed in the outpatient setting based on test results obtained on an ambulatory basis, our work differs in regards to clinical context. Similarly, Commandeur et al utilized an extreme gradient boosting model to predict long-term risk of MI and cardiac death based on clinical data and imaging measurements including coronary arterial calcium and epicardial adipose tissue (4). As coronary arterial calcium and epicardial adipose tissue measurements require manual curation by a radiologist which is not widely available, we seek to train a model on more accessible testing results.

## 4 Dataset and Features

Our dataset is the publicly available Heart Attack and Analysis & Prediction Dataset obtained from Kaggle, which contains 13 independent variables and a binary outcome variable for high and low MI risk for 303 patients (5). The provided features and data types are described in Table 1.

**Table 1. Variables in the Heart Attack and Analysis & Prediction Dataset**

| Variable | Variable Type | Description |
|---|---|---|
| risk | Categorical | 0 = low MI risk; 1 = high MI risk |
| age | Continuous | Age in years |

| | | |
|---|---|---|
| sex | Categorical | 0 = female; 1 = male |
| cp | Categorical | 0 = asymptomatic; 1 = typical anginal chest pain; 2 = atypical anginal chest pain; 3 = non-anginal chest pain |
| trtbps | Continuous | Systolic blood pressure in mmHg |
| chol | Continuous | Serum cholesterol in mg/dL |
| fbs | Continuous | Fasting blood sugar in mmol/L |
| restecg | Categorical | Rest ECG results: 0 = normal; 1=ST-T wave abnormality; 2=probable or definite left ventricular hypertrophy |
| slp | Categorical | Slope of Peak Exercise ST Segment: 0 = downsloping; 1 = flat; 2 = upsloping |
| thalachh | Continuous | Maximum heart rate in thallium stress test, in beats per minute |
| exng | Categorical | Exercise Induced Angina: 0 = not present; 1 = present |
| caa | Categorical | Patent coronary arteries on angiogram (minimum 0, maximum 3) |
| oldpeak | Categorical | ST Depression: 0 = not present; 1 = present |
| thall | Categorical | Thallium Stress Test Result: 1 = fixed defect; 2 = normal; 3 = reversible defect |

## 5 Methods

*5.1 Data exploration*

A correlation matrix of features and outcome was used to obtain an overview of the data. Boxplots, histograms, and logistic regression were used for visual inspection and statistical correlation to examine specific relationships between variables. Principal component analysis (PCA) was performed for dimension reduction and assessment of clustering patterns in a lower dimension.

*5.2 Clustering analyses*

Three unsupervised learning clustering methods (PCA, k-means, DBSCAN) were implemented to explore underlying patterns within the dataset. The Adjusted Rand Index (ARI), a measurement of binary classification accuracy, was utilized to evaluate model performance in hyperparameter tuning for k-means and DBSCAN. In k-means, two centroids were used given the binary outcome of interest. Data was dimensionally reduced with PCA and hyperparameter tuning was performed over 1-10 components in PCA, optimizing for ARI. For DBSCAN, the epsilon hyperparameter was iterated from 1-100. Python packages NumPy and decomposition, cluster, metrics from Scikit-learn were used (6-8).

*5.3 Model building*

Data was randomly split into a training (60%), validation (20%) and test (20%) set. We trained 7 classes of models on our training set and validated on our validation set. Our models were logistic regression; elastic net; gradient boosted machine (GBM); bagging; random forest (RF); support vector machine (SVM) with linear, radial, and polynomial kernels; and a fully-connected neural network (NN). An elastic net model was tuned using 3-repeats of 10-fold cross validation. Alpha search grid ranged from [1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 0.0, 1.0, 10.0, 100.0] and L1 ratio parameters were between 0 and 1. GBM learning rate was tuned over [0.05, 0.075, 0.1, 0.25, 0.5, 0.75, 1]. Our bagging model was tuned with tree numbers ranging from [10, 50, 100, 300,

500, 700, 1000, 5000]. The search grid was expanded around the number of trees with the highest F1 score on the validation set. A second round of tuning was performed with tree numbers ranging from [32,34,36,38,40,42,44,46,48,50,52,54,56,58,60,62,64,66,68,70]. Hyperparameters for SVM with linear kernel consisted of cost [1e-3, 1e-2, 1e-1, 1, 10, 100, 1000, 10000] and gamma [1e-6, 1e-5, 1e-4, 1e-3, 1e-2, 1e-1, 1, 10, 100, 1000, 10000]. For SVM with polynomial kernel, the cost range was the same and degree range was [2, 3, 4, 5, 6]. The Random Forest model was generated using cross-validation in a randomized search grid and then a full grid search for hyperparameter tuning. The final hyperparameters tuned included depth, features, leaf samples, split samples, among others. Last, we trained a NN with 4 hidden layers each with ReLU activation. The number of neurons were 18, 15, 10, and 8 in our hidden layers. We trained with 150 epochs, 10 samples per batch, and cross entropy loss. Data were scaled using min-max scaling prior to training our elastic net and SVM models. Where applicable, models were trained using 3 repeats of 10-fold cross validation. The best model from each class was selected based on best or best average F1 score.

*5.4 Model evaluation*

The models with the highest F1 scores on the validation set were selected for assessment on our test set. Test performance was evaluated with AUROC, precision, recall, F1 scores, and AUROC and PR curves.

## 6 Results and Discussion

*6.1 Data exploration*

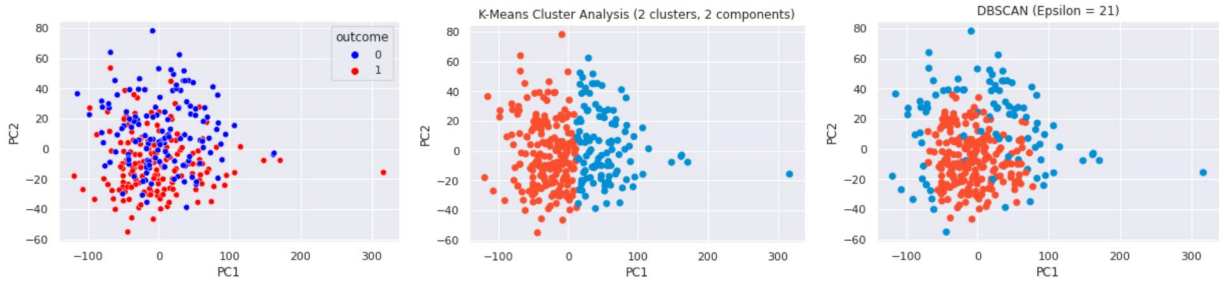
The upper register of cholesterol levels (> 350 mg/dL) was dominated by men though there was no difference in cholesterol levels overall between patients that did and did not have exertional angina. As expected, we saw a positive correlation between chest pain and ST segment slope on EKG with high risk of heart disease which are diagnostic symptoms and exam findings for MI. We saw a negative correlation between chest pain and sex which confirms literature that find women often have atypical presentations and do not present with chest pain. There was a stronger association between male sex and MI risk though it should be noted that women are underrepresented in this dataset (96 of 303 patients).

*6.2 Cluster Analysis*

PCA analysis using the first two principal components showed no separation of the two outcome classes (Fig. 1A). The optimal model in k-means yielded an ARI of 0.02 with 2 components and 2 clusters. The maximum ARI was obtained after 2-component PCA. Figure 1B demonstrates the distribution of the k-means clusters. In comparison, the optimal DBSCAN model produced a higher ARI of 0.06, using an epsilon of 21. Figure 1C illustrates the DBSCAN clustering pattern. Although DBSCAN had the best performance, both models had an ARI of less than 0.1, suggesting poor performance overall. This is explained by the poor separability in PCA as the resultant components were used as axes for clustering. Since methods performed close to the threshold of randomness (0), supervised techniques were explored.

**Figure 1: Unsupervised clustering analysis.**

A) Principal Component Analysis B) K-Means Clustering (ARI = 0.02) C) DB SCAN (ARI = 0.06).




*6.3 Model Tuning*

The logistic regression model performed well across all metrics, with an AUC of 0.91, accuracy of 0.82, precision of 0.83, and recall of 0.89. The best GBMmodel was with a learning rate of 0.75. Our SVM model using a linear kernel had a relatively lower accuracy and precision of 0.80 and 0.79, respectively. Our optimal elastic net model had an alpha of 0.01 and an L1 ratio of 0.28. The optimal bagging model had 68 trees. Our optimal random forest had a max depth of 88, max features of square root, the min number of data points placed in node before split of 9, the min number of data points allowed in leaf node of 1, number oftrees was 1065, and bootstrap was true. Our optimal SVM (radial) model had a cost of 10 and gamma of 0.1.

**Table 2: Model performance metrics on validation and test set.**

| | Validation Set | | | | Test Set | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | **AUROC** | **Precision** | **Recall** | **F1** | **Accuracy** | **AUROC** | **Precision** | **Recall** | **F1** |
| Logistic regression* | 0.91 | 0.83 | 0.89 | 0.86 | 0.87 | 0.86 | 0.84 | 0.94 | 0.89 |
| Gradient boosting* | 0.84 | 0.85 | 0.85 | 0.85 | 0.84 | 0.83 | 0.81 | 0.91 | 0.86 |
| SVM (linear) | 0.79 | 0.79 | 0.85 | 0.82 | -- | -- | -- | -- | -- |
| SVM (radial)* | 0.83 | 0.83 | 0.89 | 0.86 | 0.46 | 0.5 | 0.00 | 0.00 | 0.00 |
| SVM (polynomial) | 0.76 | 0.75 | 0.89 | 0.81 | -- | -- | -- | -- | -- |
| Elastic Net | 0.81 | 0.80 | 0.89 | 0.84 | -- | -- | -- | -- | -- |
| Bagging* | 0.84 | 0.85 | 0.85 | 0.85 | 0.80 | 0.80 | 0.78 | 0.88 | 0.83 |
| Random Forest* | 0.84 | 0.85 | 0.85 | 0.85 | 0.79 | 0.78 | 0.78 | 0.85 | 0.81 |
| Neural Net | 0.82 | 0.85 | 0.81 | 0.83 | -- | -- | -- | -- | -- |

*Models with top two F1 scores on validation.

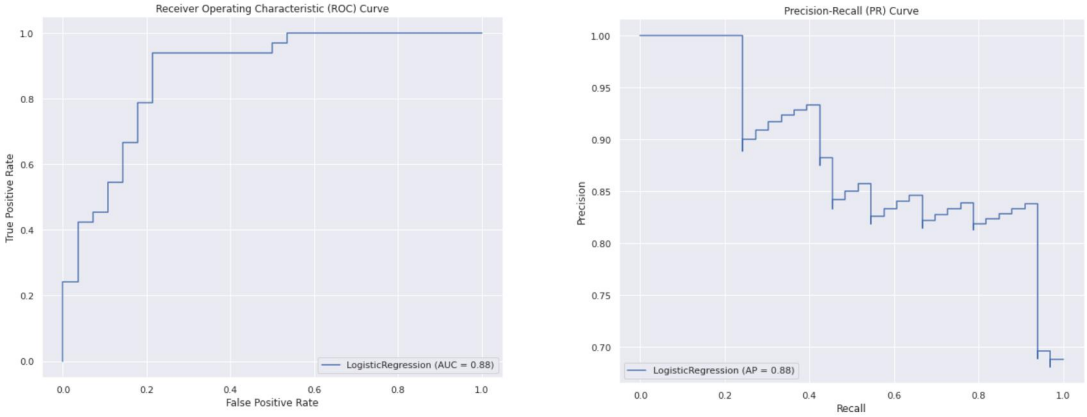
*6.4 Model Performance*

After assessing the above models on the validation datasets, we selected the top performing models based on the top two F1 scores (0.85, 0.86) to evaluate on the test sets. This included the logistic regression, gradient boosting, bagging, random forest, and SVM with radial kernel models. See table 2 for performance metrics of each model on the validation set and the best models on the test set. Of the five model evaluated on the test set, logistic regression had the best performance across all metrics. Both bagging and random forest models had a slight drop in performance in all metrics. SVM with radial kernel performed very poorly and does worse than a majority class classifier (majority class, 1 = high risk MI). A majority class classifier would have an

accuracy of 54% whereas the SVM with radial kernel had an accuracy of 46% on the test set. The drop off in performance in these three models on the test set, especially the SVM, is likely due to overfitting on the training data. The GBM model interestingly maintained performance on the test set and had improved recall (0.85 on validation, 0.91 on test). This can be explained by the mechanics of GBM -- it is an amalgamation of iterative tree models. GBM starts with a shallow tree and each iteration focuses on the areas of the training data with the largest residual errors (9). The final model is a summation of the trees built in each iteration and with a slow learning rate, GBM is robust to overfitting.

**Figure 2: Logistic Regression model evaluated on test set.**

A) AUROC Curve B) PR Curve




## 7 Conclusion/Future Work (1-2 paragraphs)

In conclusion, the logistic regression model had the best performance with auROC 0.87 and sensitivity 0.94. In the context of clinical risk stratification, high sensitivity is preferred to increase detection of patients who could benefit from additional cardiovascular care resources, particularly in the setting of aforementioned care inequity and ongoing preventable death from MI. In regards to the superior performance of logistic regression in comparison to other candidate models, we acknowledge that our dataset is artificially 'cleaned' for the purposes of a data competition and thus contains little latent noise that could lead to overfitting. If additional variables based on real-world electronic health record data were to be added, more complex models may perform better.

Future work includes the expansion of the dataset to include additional clinically relevant variables, such as serum homocysteine level or menopause status, which have been shown to impact MI risk in select populations (10,11). In addition to incorporating additional variables, further external validation of our model on external datasets should be undertaken to improve interoperability and implementation across healthcare institutions and contexts. This could involve curating new datasets in different health systems and evaluating model performance.

## 8 Contributions

**Vy:** Literature review, clinical expertise, model building (K-means clustering, DBSCAN)
**Katie:** Preliminary coding and experiments (EDA), model building (bagging, neural net)
**Katelyn:** Preliminary coding and experiments (EDA), model building (SVM, random forest)
**All:** Discussion for project definition, methods and analysis, proposal for next steps, model building, model evaluation, results and discussion analysis

## 9 References


(1) Virani, Salim S., et al. "Heart disease and stroke statistics—2020 update: a report from the American Heart Association." *Circulation* 141.9 (2020): e139-e596.

(2) McClellan, Mark, et al. "Call to action: urgent challenges in cardiovascular disease: a presidential advisory from the American Heart Association." *Circulation* 139.9 (2019): e44-e54.

(3) Than, Martin P., et al. "Machine learning to predict the likelihood of acute myocardial infarction." *Circulation* 140.11 (2019): 899-909.

(4) Commandeur F, Slomka PJ, Goeller M, Chen X, Cadet S, Razipour A, McElhinney P, Gransar H, Cantu S, Miller RJH, Rozanski A, Achenbach S, Tamarappoo BK, Berman DS, Dey D. Machine learning to predict the long-term risk of myocardial infarction and cardiac death based on clinical risk, coronary calcium, and epicardial adipose tissue: a prospective study. Cardiovasc Res. 2020 Dec 1;116(14):2216-2225. doi: 10.1093/cvr/cvz321. PMID: 31853543; PMCID: PMC7750990.

(5) Rahman R. (2012, March). Heart Attack Analysis & Prediction Dataset. Retrieved April 4 2021 from https://www.kaggle.com/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

(6) Pedregosa et al. 'Scikit-learn: Machine Learning in Python', JMLR 12, pp. 2825-2830, 2011.

(7) Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke & Travis E. Oliphant. Array programming with NumPy, Nature, 585, 357–362 (2020), DOI:10.1038/s41586-020-2649-2 (publisher link)

(8) Real Python. K-Means Clustering in Python: A Practical Guide. *Real Python*, Real Python, 8 Jan. 2021, realpython.com/k-means-clustering-python/.

(9) Natekin A. & Knoll A. (2013). Gradient boosting machines, a tutorial. *Front. Neurorobot,* 4;7:21.

(10) Ganguly, P., & Alam, S. F. (2015). Role of homocysteine in the development of cardiovascular disease. *Nutrition journal*, *14*(1), 1-10.

(11) Gonzales, T. K., Yonker, J. A., Chang, V., Roan, C. L., Herd, P., & Atwood, C. S. (2017). Myocardial infarction in the Wisconsin Longitudinal Study: the interaction among environmental, health, social, behavioural and genetic factors. *bmj Open*, *7*(1).