# Three Different Machine Learners Used to Classify the Presence of Heart Disease

Kevin Graves

Department of Biomedical Informatics, University of Utah

## Introduction

A basic PudMed search for "predicting heart disease" came back with over 12,000 results. Many of the techniques used for prediction are based off the idea that using a small, pre-selected set of variables that are deemed clinically relevant and simple to calculate are the most useful things when creating a prediction model. This may be true in some cases, but this process precludes the possibility of discovering new predictive factors that have not been identified a priori. Furthermore, eliminating models that may be more complex computationally and harder to interpret clinically could be sacrificing models with a higher predictive ability (1).

It's been suggested that an optimal, general medical risk score would be one that applies to "any individual, group, or population and would not complicate the care-delivery process" (2). In order to get nearer this goal, a model that uses the strongest predictors is more valuable than one that uses the simplest ones. The $CHA_2DS_2VASc$ score is a common, and useful, risk score to determine risk in patients with atrial fibrillation, and it is something any clinician can easily calculate from a simple history review (3). However, resources are available in today's medical world to deliver risk-predictions in near real-time to electronic medical records that could not normally be calculated in a physician's head. More advanced machine learning techniques may be able to identify the presence of disease in patients.

The present study aimed to compare the accuracy of different machine learning models for classifying patients from the Statlog (Heart) dataset from the University of California Irvine (UCI) Machine Learning Repository as either presence, or no presence, of heart disease (4). Three different machine learning models were developed and tested on this dataset.

## Background

Classification plays an important role in clinical informatics. These methods allow clinicians to separate patients into states of disease. Accurate classification of the presence of a disease in a patient is crucial to ensure that the proper treatment and care is delivered. Epidemiological studies, such as the Framingham Heart Study, collect hundreds of variables for a population over long periods of time. Machine learning methods provide excellent tools to process, analyze, and evaluate all these data to elucidate risk factors, classify patients, and predict future events (5).

Previous studies have investigated the use of machine learning in cardiovascular event and disease classification (5-7). Several models have been presented that show the utility of machine learning techniques to classify the presence of disease. As the technology continues to be more and more available, there will be an increase in the demand for fast and accurate delivery of classification and prediction (8).

Two particular algorithms show up in several different studies; support vector machines (SVM) and random forest classifiers. Both of these methods have a long history of strong results in machine learning. Ambale-Venkatesh et al., utilized a random forest model (among many others) in comparison to a classic Cox Proportional Hazard regression model. This study dealt with a large sample of patients and they had over 700 variables to build into their model. While the random forest model that utilized all the variables ultimately performed the strongest, it is interesting to note that a similar random forest model that only utilized 20 variables was very close in performance. Variable selection is a delicate and difficult task when building a model and the balance of maintaining an accurate model while avoiding overfitting is a difficult task. Evaluation of the models using the C-index and Brier score showed that even with the complex

models, they were able to avoid overfitting. It is of note that the simpler model was nearly as powerful as the most complex one. When taking into account time to train and compile, a simpler, faster model may be just as useful to clinicians as one that is much more complicated.

Classification using regression trees is also a model that can make more intuitive sense to clinicians unfamiliar with machine learning techniques. Being able to move the regression trees into interpretable decision rules that can be applied to clinical practice is a valuable tool. Another reason many studies utilize decision trees is that it removes any assumption of linearity that many regression models tend to take on. This is an import feature the tree-based methods avoid making.

In this study, three different models were developed and trained on the same dataset. Random forest classifier, logistic regression, and XGBoost classifier were all used to identify the presence of heart disease based off of 13 variables. Due to the small nature of the dataset, bootstrapping with 1,000 iterations was used for every model and several parameters were ran for each model as well.

**Methods**

The data for this study was obtained from the UCI Machine Learning Repository. This dataset is a heart disease database similar to a database already present in the repository (Heart Disease databases) but in a slightly different form. The dataset is comprised of 13 variables as well as a binary variable that indicates the presence of heart disease. There are 270 instances in this dataset with no missing values. A breakdown of the demographics and variables is available in Table 1. Due to the small nature of this dataset and the fact that no values were missing, all values for all variables were included in the analysis. All analysis and training was done in

Jupyter Notebooks on an MacBook Pro with 16 GB of memory and a 2.6 GHz Intel core i7 processor.

Before the model was trained or ran on any machine learning algorithms, the chi-squared stats test was run on all the features and the class. Figure 1 shows a breakdown of the most discriminating features according to the chi-squared test. The first model developed utilized the Random Forest Classifier available in the scikit-learn library (9). As previously mentioned, there were only 270 instances available in this dataset, so a bootstrapping (sampling with replacement) method was employed. Seven different Random Forest Classifier models were built using a different number of trees (11, 51, 101, 201, 301, 401, 501). Each model was trained over 1,000 iteration. The original dataset was randomly sampled from 270 times to create the training set (In-the-bag) and tested on using the out-of-bag sample for each iteration. The overall training and test accuracy was calculated for each of the seven models. The training accuracy was plotted against the number of trees used in model and the results can be seen in Figure 2.

The second model built utilized Logistic Regression. Again, sci-kit learn was implemented to build the model and the same bootstrapping method used for the Random Forest Classifier was employed with 1,000 iterations. Eight different variations of the C parameter were tested (0.001, 0.01, 0.1, 1,1 0, 100, 200, 300) over 1,000 iterations each. Again, training accuracy versus the value of the C parameter was plotted and can be seen in Figure 3.

The final model was built using a scalable tree boosting system called XGBoost (10). For this model, a randomized search (using SK-learn's RandomizedSearchCV) for the best parameters was done first using 5-fold cross validation. The parameters of the best model were then selected for the final model. The bootstrapping method used by the previous two models

was employed and nine different learning rates were tested in a similar fashion to the other models. Accuracy versus learning rate was plotted and shown in Figure 4.

**Results**

The average age of patients included in the dataset was 54.4 years old with a standard deviation of 9.1 years, 67.8% were male, and 44.4% did have heart disease. 33.0% had a history of angina, 14.8% had a fasting blood sugar of greater than 120 mg/dL, which could be indicative of Type 2 diabetes. In general, the cohort of patients with heart disease were older, higher percentage of males, had a higher resting heart rate and serum cholesterol. The cohort with heart disease had a higher percent with angina and much higher percentage of diseased vessels on fluoroscopy. Interestingly, there was not much difference in the percentage with a high fasting glucose. Contingency table results for all three learners can be seen in Table 2 and the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, and area under the curve (AUC) for the best model from all 3 different kind of learners has been summarized in Table 3.

The random forest classifier performed best with a tree size of 201. Sensitivity and specificity were 0.77 and 0.87, respectively. PPV and NPV were both 0.82, while test accuracy was 82.11%. AUC for this model was 0.812. The test accuracy for the other models with different number of trees was 79.85% for 11 trees, 81.51% with 51 trees, 81.83% with 101 trees, 81.94% with 301 trees, 81.98% with 401 trees, and 82.03% with 501 trees.

The logistic regression model performed the best overall. The top performing model used a C parameter of 0.1. Sensitivity and specificity were 0.78 and 0.88, respectively. PPV was 0.84 while the NPV was 0.83. Test accuracy was 83.33% with an AUC of 0.828. The testing accuracy for the other C parameters was 71.98% when C=0.001, 79.38% when C=0.01, 83.11% when

C=1, 83.12% when C=10, 82.50% when C=100, 82.54% when C=200, and 82.52% when C=300.

After the random grid search, the final parameters used for model used an objective of binary:logistic and a booster using gradient boosted trees. The regularization alpha was set to 1, regularization lambda was set to 0.05, the number of estimators was 75, maximum depth was 1, and the gamma was 1 as well. The learning rate that performed the best with these parameters was 0.3. For the best model, the sensitivity and specificity were 0.77 and 0.87, which was the same as the random forest classifier model. PPV and NPV were 0.82 and 0.83, respectively. Accuracy was 82.54% while AUC was 0.820.

**Discussion**

Logistic regression performed the best of all learners. However, all the learners had accuracies in the low 80's with similar AUCs. This is not terribly surprising given the small dataset but should give hope for future work that even these simple models can be improved on with either more instances or perhaps more variables. All the learners did show better specificity than sensitivity, indicating that all the models were better at identifying those without heart disease as opposed the those with heart disease. Utilizing a classifier like this might be more productive in telling patient and clinicians that they don't have the disease as opposed to tell them that the disease is present. This could potentially save patients from undergoing further testing, which can be invasive, to determine the presence of heart disease. This would also free up more resources that can be devoted to working with patients who do have the disease.

These modest results are somewhat hopeful for the future of machine learning in clinical classification. This study showed that even with a small number of cases and limited variables, somewhat accurate classification can be achieved with different kinds of machine learning

algorithms. A simple model, such as the ones built for this study, should be easier to implement in a larger sense than more complex models that would require thousands of training cases. The use of a model with limited variables would also be simpler to introduce to clinicians so that they can see what features are powering the models. This may result in more trust being given to the classification being predicted by the model.

Future studies should focus on finding larger and more diverse datasets to train models on. Ultimately, these methods should be tested in a real-world, clinical setting to aid clinicians in decision-making or to help patients determine if they should be seeking care. Such a model could be put into production as an online risk calculator or be made available in an electronic healthcare record for clinicians to view and assess when determining patient history and treatment.

The biggest limitations of this study were in the scarcity of the data and that this dataset was derived from a larger dataset, any errors that were made in calculating some of the fields could have survived into this dataset without any knowledge. 270 instances and 13 variables is a very small dataset when working with machine learning. The results could have also been slightly improved by more finely tuning the parameters of the models being built. A grid search with cross validation for more than one variable would have potentially yielded stronger models for all the learners. However, this study was limited by computing resources, time, and the experience of the researcher.

**Conclusion**

This study showed that random forests, logistic regression, and XGBoost are all viable options for constructing a model for classifying patients as either having or not having heart disease. These models were built using a small dataset with a limited number of variables. Future

studies looking to improve upon this work should consider fine tuning of parameters and hyperparameters as well as utilizing a larger, more diverse dataset. A strong classification model could be used to identify patients with heart disease in clinical setting to prevent overuse of screening tests and more invasive methods.

# References


1. Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, et al. Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach. Academic emergency medicine : official journal of the Society for Academic Emergency Medicine. 2016;23(3):269-78.
2. Horne BD, May HT, Muhlestein JB, Ronnow BS, Lappe DL, Renlund DG, et al. Exceptional mortality prediction by risk scores from common laboratory tests. The American journal of medicine. 2009;122(6):550-8.
3. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest. 2010;137(2):263-72.
4. Dheera DaKT, Efi. {UCI} Machine Learning Repository. In: University of California I, School of Information and Computer Sciences, editor. 2017.
5. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, et al. Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. Circulation research. 2017;121(9):1092-101.
6. Austin PC, Tu JV, Ho JE, Levy D, Lee DS. Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes. Journal of clinical epidemiology. 2013;66(4):398-407.
7. Ross EG, Shah NH, Dalman RL, Nead KT, Cooke JP, Leeper NJ. The use of machine learning for the identification of peripheral artery disease and future mortality risk. Journal of vascular surgery. 2016;64(5):1515-22.e3.
8. Meyer A, Zverinski D, Pfahringer B, Kempfert J, Kuehne T, Sundermann SH, et al. Machine learning for real-time prediction of complications in critical care: a retrospective study. The Lancet Respiratory medicine. 2018.
9. Pedregosa FaV, G. and Gramfort, A. and Michel, V., and Thirion BaG, O. and Blondel, M. and Prettenhofer, P., and Weiss RaD, V. and Vanderplas, J. and Passos, A. and, Cournapeau DaB, M. and Perrot, M. and Duchesnay, E. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825-30.
10. Chen TaG, Carlos. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016:785-94.

Table 1. Demographics

| Variables | All (n=270) | Presence of HD (n=120) | No HD (n=150) |
|---|---|---|---|
| **Age (years)** | 54.4 (± 9.1) | 56.6 (±8.1) | 52.7(±9.5) |
| **Sex (male)** | 67.8% | 83.3% | 55.3% |
| **Resting SBP** | 131.3 (±17.9) | 134.4(±19.1) | 128.9(±16.5) |
| **Serum Cholesterol** | 249.7 (±51.7) | 256.5(±48.0) | 244.2(±54.0) |
| **Fasting Glucose >120** | 14.8% | 14.2% | 15.3% |
| **Max HR** | 149.7 (±23.2) | 138.9(±23.1) | 158.3(±19.3) |
| **History of Angina** | 33.0% | 55.0% | 19.2% |
| **No Diseased Vessels** | 59.3% | 33.3% | 80.0% |
| **1 Diseased Vessel** | 21.5% | 31.7% | 13.3% |
| **2 Diseased Vessels** | 12.2% | 21.7% | 4.7% |
| **3 Diseased Vessels** | 7.0% | 13.3% | 2.0% |

Table 2. Contingency Tables for All Three Learners

| Random Forest | True no HD | True HD | Totals |
|---|---|---|---|
| **Predicted no HD** | 47,629 | 10,359 | 57,988 |
| **Predicted HD** | 7,392 | 33,857 | 41,249 |
| **Totals** | 55,021 | 44,216 | 99,237 |

| Logistic Regression | True no HD | True HD | Totals |
|---|---|---|---|
| **Predicted no HD** | 48,163 | 9,769 | 57,932 |
| **Predicted HD** | 6,741 | 34,386 | 41,127 |
| **Totals** | 54,904 | 44,155 | 99,059 |

| XGBoost | True no HD | True HD | Totals |
|---|---|---|---|
| **Predicted no HD** | 47,711 | 9,938 | 57,988 |
| **Predicted HD** | 7,396 | 34,218 | 41,249 |
| **Totals** | 55,107 | 44,156 | 99,263 |

Table 3. Results of Best Model for All Three Learners

| | Random Forest | Logistic Regression | XGBoost |
|---|---|---|---|
| **Sensitivity** | 0.77 | 0.78 | 0.77 |
| **Specificity** | 0.87 | 0.88 | 0.87 |
| **PPV** | 0.82 | 0.84 | 0.82 |
| **NPV** | 0.82 | 0.83 | 0.83 |
| **Accuracy** | 82.11% | 83.33% | 82.54% |
| **AUC** | 0.812 | 0.828 | 0.820 |

Figure 1. Most Discriminant Features Using Chi-Squared Test.

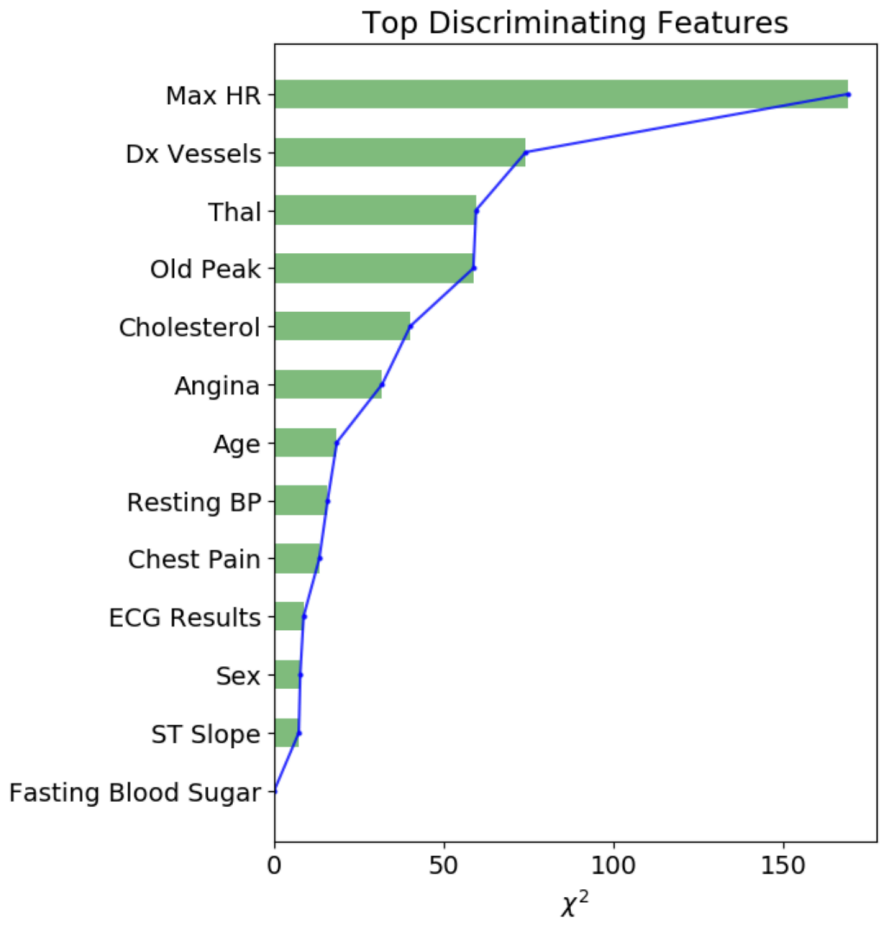



Figure 2. Training Versus Test Error for Random Forest Models

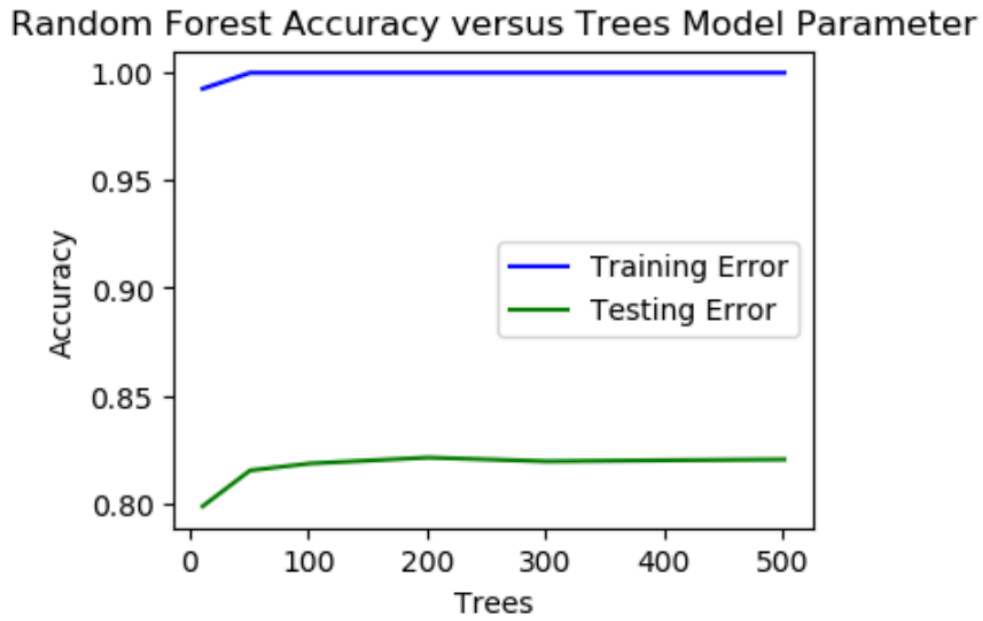



Figure 3. Training Versus Test Error for Logistic Regression Models

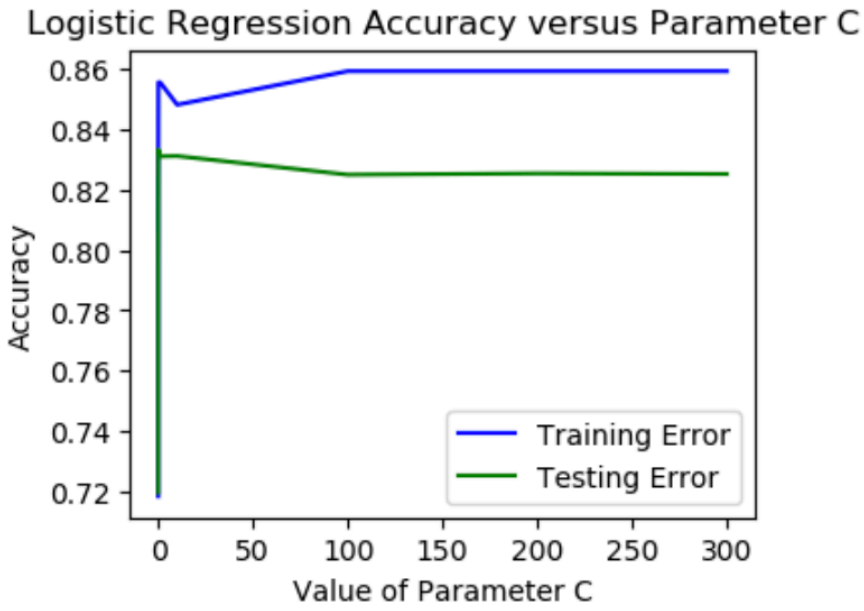
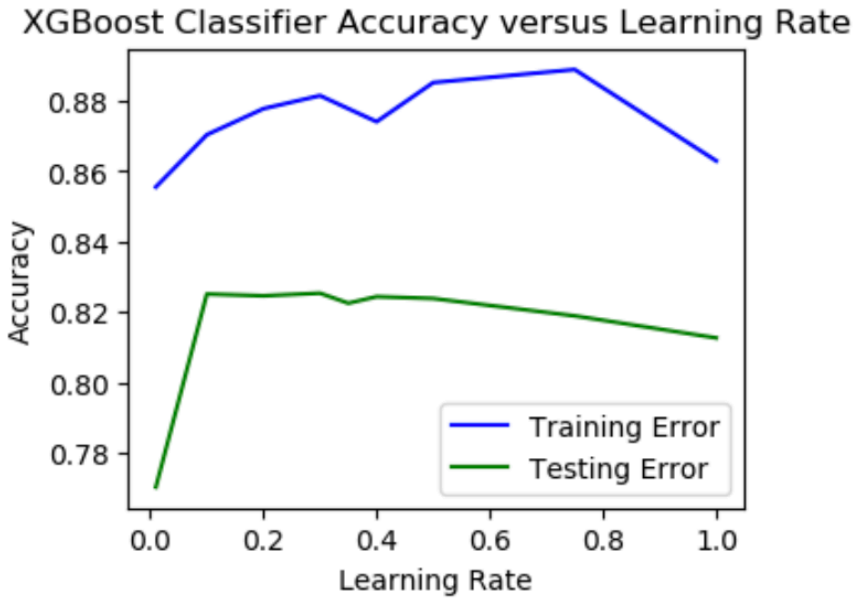



Figure 4. Training Versus Test Error for XGBoost Models


XGBoost Classifier Accuracy versus Learning Rate
Accuracy
0.88
0.86
0.84
0.82
0.80
0.78
0.0
0.2
0.4
0.6
0.8
1.0
Learning Rate
Training Error
Testing Error