Michal Chowaniak

DSC 540 Advanced Machine Learning Final Project

Various Machine Learning Classifiers used in identifying heart diseases.

1 ABSTRACT

Heart diseases are a considerable unsolved problem around the world. This paper proposes several machine learning classifiers to improve early diagnosis of heart diseases. The best performing classifiers for that task was Random Forest, which could correctly predict the existence of heart disease or lack of it in approximately 87 out of 100 cases. The most important characteristics selected were a level of fat in blood, smoking, angina, and chest pain.

2 INTRODUCTION

Heart disease is a broad name for various types of heart conditions, which can lead to more serious problems like heart attack, chest pain (angina) or stroke. All of them involve blocked or narrowed blood vessels (Figure 1) which supply oxygen to heart muscle. Per the Center for Disease Control and Prevention (1) heart diseases kill 610,000.00 (2) people every year in the United States. It is the common cause of death in many states in the USA (Figure 2).There are many factors responsible for heart diseases among them are age, sex, family history, lifestyle, other diseases. To analyze the heart diseases, I used a dataset from UCI Machine Learning Repository Heart Disease Data Set called Cleveland (3). The set was donated in 1988, it includes 14 variables which can be grouped into demographics, symptoms, various medical test results. To analyze this data, I used Sci-Kit Learn (4) libraries and a code temple written by Dr. Casey Bennett (5) with some modifications. The rest of this paper is organized in the following way: literature review, methodology, results, discussion, conclusion, future work, sources, appendix.

3 LITERATURE REVIEW

This subject and the dataset have been analyzed by others for many years.

Koodali et al (2015) (6) implemented J48 decision tree, Random Forest, Bagging, Rep Tree, Decision Stump, Cart, Naïve Bayes using Weka tool. They achieved the highest accuracy of 95.77 and recall of 0.974 on bagging methods however, they did not mention using cross-validation.

Mythili T. et al (2013) (7) chosen Support Vector Machine, Decision Trees, Logistic Regression they concluded that a decision tree is the best model for this classification problem and proposed a comparative study of multiple results.

Resul Das et al (2009) (8) used Neural Networks and various ensemble methods using SAS enterprise miner software, they achieved 89.1% accuracy, 80.95 sensitivity, and 95.91 specificity.

Akin Ozcift et al (2011) (9) investigated computer-aided diagnosis algorithms on 3 data sets for, diabetes, heart diseases and Parkinson's diseases using Rotation Forest classifier and they achieved an

accuracy of 74.47, 80.49, 87.13 compared to base classifier 72.15. 77.52, 84.43. Their most important variables were chest pain, resting electrocardiograph results and maximum heart rate achieved.

Abdullah, A. Sheik et al (2012) (10) researched random forest classifier in comparison to the decision tree. Random Forest accuracy was 63.33 versus 50.67 for decision tree.

4 METHODOLOGY

4.1 Preprocessing

Data set of heart diseases was obtained from UCI Machine learning repository. Cleveland dataset includes 14 variables and 303 observations, it was a mix of categorical and continuous variables. There were 5 missing values found by handling of value error during import of dataset, they were re-coded to value 9999, then replaced with the median value for the column using SimpleImputer. Because original data set target variables for no disease was 0, and for disease 1,2,3,4, the target variable was binned to binary 0-no disease, 1-disease present by replacing values of 2,3,4 with value 1. Target variables was balanced 0 – 52%, 1- 48%. Data was left not normalized.

4.2 Feature Selection

The features selection process included running various models (Decision Tree, Random Forest, Gradient Boosting, Ada Boost, Neural Networks, and SVC) using different types of feature selection methods. The best feature selection method was chosen based on a low number of features selected and high score for accuracy and AUC. The following feature selection methods were used.

Stepwise Backwards recursive selection using Random Forest – in this approach selection process starts with all features and at each step, a variable is eliminated recursively until an optimal model is selected.

Wrapper-based using Random Forest, Gradient Boosting, Support Vector Classification (SVC) – in this approach multiple models are built using different classifiers, on different sets of features and the best one is chosen. Classifiers are explained in the Methods section.

Filter feature selection method using Chi-squared - in this approach features are selected based on a statistical test of independence.

Figure (3) and (4) presents features selected by the above approaches alongside accuracy and AUC scores for the best performing classifier. Wrapper-based Support Vector Classifier (SVC) feature selection method was chosen because the number of features selected was only 7, and by looking at accuracy and AUC scores, all models performed better than in other features selection methods.

Figure (5) shows the most important features selected. Out of 7 selected following 3 had the highest Chi-squared score: 1. Level of triglycerides, 2. Smoker/Nonsmoker, 3. Exercise-induced angina.

4.3 Models

4.3.1 Random Forest

Random Forest is a boosting ensemble method which uses many decision trees, where at each split a random subset of features is chosen. There were 24 decision tree classifiers built using cross-validation

on preprocessed data by SVC wrapper method, Figure(6), each with different parameters. At first criterion entropy and a number of trees was tested. Peak performance of 0.83 for average accuracy and 0.91 for average AUC was observed when a number of trees were 1000. A bigger number of estimators did not improve the model and decreased average scores. Second criterion gini and number of trees was tested. Peak performance was reached when a number of trees reached 1500, but average scores were lower than for entropy. In next step minimum sample split was tested with criterion entropy and a number of trees of 1000. The peak scores were achieved with min split of 30 and were 0.87 for accuracy, 0.92 for AUC, 0.83 for recall and 0.88 for precision. The last step was an attempt to improve the best model by tuning parameters using GridSearchCV however model produced had lower scores of 0.85 for accuracy and 0.91 for AUC.

4.3.2 Gradient Boosting

Gradient Boosting is an ensemble of a weak learner, it creates weighted models and using gradient descent decides if and when to add them to the ensemble. There was 24 Gradient Boosting classifiers build using cross-validation on preprocessed data by SVC wrapper method, Figure(7). At the beginning parameter loss equal deviance was tested and various numbers of estimators. The best classifier was observed with a number of trees equal to 25 and it had 0.83 accuracy and 0.91 AUC. Next, the loss was changed to exponential and again a various number of trees was tested. The peak performance achieved was 0.84 for accuracy and 0.91 for AUC for a number of estimators equal to 25. Following steps were to run various models with a loss equal to deviance or exponential, a number of estimators equal to 25 and various values of learning rate and maximum depth. The best model with the above parameters and learning rate of 0.05 achieved an accuracy score of 0.85 and AUC of 0.90. It is model number 15 in figure(7). The last attempt to improve scores was to use GridSearchCV to run many models with various parameters. That attempt did not improve scores and produced similar accuracy of 0.84 and 0.90 AUC.

4.3.3 ADA Boost

Ada Boost is an ensemble of a weak learner method, which creates weighted models and it assigns greater weights to observations that are difficult to classify forcing classifier to focus on them. They were 12 Ada Boost models run using cross-validation and SVC wrapper feature selection. First, a number of estimators were changed, then learning rate and then best estimator. Out of several models run the best-performed model with a number of estimators/trees equal 100 which achieved an accuracy of 0.85 and AUC of 0.92. Next, the learning rate was changed for the above best model, but it did not improve scores. Also, base estimator parameter was changed from None to Random Forest which increased runtime only. The last phase was to deploy GridSearchCV to find parameters which produce a slightly better model, however, this attempt produced worse than average Ada Boost model (model # 12, Figure (8).

4.3.4 Neural Network

Neural Networks are methods inspired by brain's neurophysiology, they work by detecting complex nonlinear relationships in data, preferably not structured data, by creating layers of nodes (neurons) which can learn on their own errors. There were 19 Neural Networks build using cross-validation on preprocessed data by SVC wrapper method, Figure(9). At the beginning, various hidden layers were tested using solver adam. The peak performance was achieved by model number 3, Figure(9) with 2 hidden layers and 100 units. Accuracy was 0.84 and AUC was 0.90. Next, the model was tested with solver lbfgs, but it did not improve results. The following step was to check performance with various activation values, which also did not improve results. Finally, grid search produced a model with almost the same scores but was not able to improve the scores of my best model.

4.3.5 Support Vector Classifier (SVC)

Support Vector Classifier is a discriminative model which separates data in many dimensions into classes by utilizing different kernels among them linear, polynomial, sigmoid, radial basis function. There were 15 SVC models created using cross-validation on preprocessed data by SVC wrapper method, Figure (10), each with different parameters. At first, a different kernel like linear, poly, sigmoid, rbf was tested with various gamma values, maximum iteration was not limited and penalty C parameter was equal to 1. The peak scores were achieved by kernel linear and gamma equal to scale and were 0.83 for accuracy and 0.90 for AUC. Next GridSearchCV was applied for hyperparameters tuning and its results confirmed earlier scores, however, parameters were different.

4.4 Performance

4.4.1 Scores

4.4.1.1 Confusion Matrix

In order to understand accuracy and other metrics, it is helpful to understand the confusion matrix first.

Predicted/Actual	Fact: You have heart disease	Fact: You do not have heart disease
Your doctor thinks you have heart disease	TP- True Positive Your doctor told you that you have heart disease and you really have it	FP- False Positive Your doctor told you that you have heart disease, but in fact, you do not have it.
Your doctor thinks you do not have heart disease	FN – False Negative Your doctor told you that you do not have heart disease, but you have it	TN – True Negative Your doctor told you that you do not have heart disease, and in fact, you do not have it.

4.4.1.2 Accuracy

Accuracy is a percentage, or a fraction of observation classified correctly.

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}}$$

4.4.1.3 Recall

The recall is a measure of how many truly relevant results are returned. The recall is a percentage or fraction of relevant observation over total relevant observation.

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

4.4.1.4 Precision

Precision is a measure of relevancy. Precision is a percentage or fraction of relevant observation over retrieved observation.

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

4.4.1.5 The area under the curve (AUC)

The area under the curve (AUC) is a measure of classification performance.

$$\mathrm{AUC} = \frac{1}{2} * \left(\frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} + \frac{\mathrm{TN}}{(\mathrm{TN} + \mathrm{FP})} \right)$$

4.4.2 Cross-Validation

It is a sampling process to evaluate models. Data set is split according to k value, models are run on training data set of size k-1/k and tested on a sample of size 1/k. Then a different testing/training subset is chosen from the original split and process repeats k times. Then an average score is used for evaluation.

5 RESULTS

The best performing models for each classifier are presented in figure (11). The models were examined with 4 metrics which were accuracy, AUC, recall, and precision. The chart shows performance metrics for various machine learning models, higher score, better the model is. All 5 models, which were Random Forest, Gradient Boosting, Ada Boost, and Neural Networks, Support Vector Machine performed well. Random Forest is a random ensemble of simple decision trees, Gradient Boost and Ada Boost are ensembles of weak learners, Neural Network is a model inspired by natural intelligence, SVM is a model which separates data in many dimensions into classes. The best model chosen would be the one with the best scores, decent runtime, simplest, easy to explain. Although Gradient Boosting, Ada Boost, and Neural Networks performs similarly to Random Forest and produce a simpler model, in my opinion, the best model is Random Forest because it is easier to explain. The best Random Forest can correctly predict the existence of heart disease or lack of it in approximately 87 out of 100 cases. AUC of 0.92 tells that the probability that the Random Forest model classifies a patient as having heart disease is 92%. Recall tells that when a patient for a fact has heart disease, the Random Forest can correctly identify the illness in 83 out of 100 cases. Precision tells that when Random Forest classifier makes a prediction of a patient having heart disease, it predicts the existence of heart disease correctly in 88 out of 100 cases.

6 DISCUSSION

In most models Random Forest, Gradient Boosting, Ada Boost, and Neural Networks, Support Vector Machine scores were similar between 0.77 and 0.87 for average accuracy, except for some SVC models which scores were as low as 0.23 and between 0.82 and 0.92 for average AUC, except for some SVC models which scored as low as 0.14 for AUC. I suspect Kernel sigmoid in SVC, which is equivalent to 2 layers Neural Network, did not perform adequately, because it works well on wide datasets. Overall the best-performing feature selection method was SVC with cross-validation. The worst feature selection methods were stepwise backward removal, and chi-squared. Classification models which did not perform well were SVC with kernel sigmoid and Neural Network with solver lbfg, Gradient Boosting with loss set up to deviance or exponential. Models which performed well were Random Forest with criterion entropy, a large number of trees and large split also ADA Boost with base estimator equal None and Neural Network scores were very high for solver adam.

7 CONCLUSION

Heart diseases are common causes of deaths around the world, which in many cases could be prevented. Per WHO “cardiovascular diseases are the number 1 cause of death globally, an estimate of 17.9 million people died from cardiovascular diseases in 2016, representing 31% of all global deaths, of these deaths 85% are due to heart attack and stroke” (11). The above analysis shows that the Random Forest model can provide helpful information in Clinical Support Decision Systems. Additionally, the important lesson from this analysis for patients and health care practitioners are the most important risk factors of a heart disease, which are the following, high level of triglycerides, which is a measure of a fat in a blood, followed by smoking, angina, which is reduced blood flow to a heart muscle due to fat accumulated in arteries, and a chest pain. All those risk factors are associated with a sedentary lifestyle. To avoid heart diseases American Heart Association recommends (12) the following “stop smoking, choose good nutrition, lower cholesterol, lower blood pressure, be physically active every day”.

8 FUTURE WORK

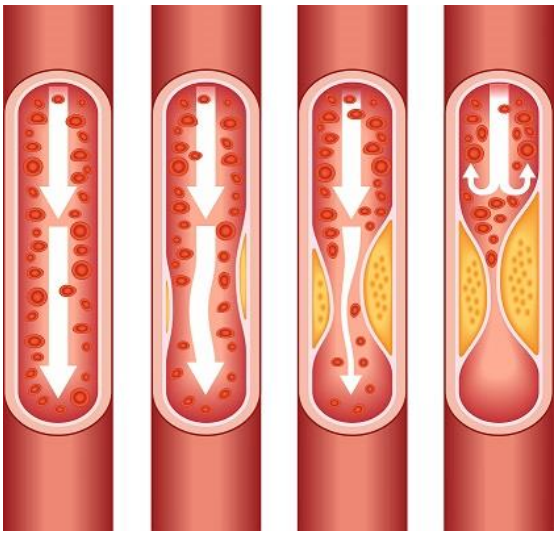
In the future, someone could improve the above analysis by creating more models per each classifier and investigate reasons why grid search was not able to improve any models. One of a reason might be a need to limit grid search running time due to the limited processing power capabilities of my laptop. Because of the grid search parameters were in the vicinity of manual best model parameters, it is possible that better model paraments were not so similar to parameters of the best model found manually.

9 SOURCES

1. Center for Disease Control and Prevention. (March 14 2019). Retrieved from https://www.cdc.gov/heartdisease/facts.htm
2. CDC, NCHS. Underlying Cause of Death 1999-2013 on CDC WONDER Online Database, released 2015. Data are from the Multiple Cause of Death Files, 1999-2013, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed Feb. 3, 2015.
3. UCI Heart Disease Data Set. (March 14 2019). Retrieved from https://archive.ics.uci.edu /ml/datasets/Heart+Disease.
4. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
5. Scikit-learn code template created by Casey Bennett 2018, www.CaseyBennett.com
6. Kodali Lohita, Adusumili Amitha Sree, Doreti Poojitha, T. Renuga Devi, A Umamakeswari . "Performance Analysis of Various Data Mining Techniques in the Prediction of Heart Disease" Indian Journal of Science and Technology Vol 8(35) December 2015.
7. Mythili T., Dev Mukherji, Nikita Padalia and Abhiram Naidu, "A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)", International Journal of Computer Applications (0975-8887) Volume 68-No.16, April 2013.
8. Resul Das, Ibrahim Turkoglu, Abdulkadir Sengur, "Effective diagnosis of heart disease through neural networks ensembles", Expert Systems with Applications 36 (2009) 7675-7680
9. Akin Ozcift, Arif Gulten, "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms", Computer Methods and Programs in Biomedicine 104 (2011) 443-451
10. Abdullah, A. Sheik, and R. Rajalaxmi. "A data mining model for predicting the coronary heart disease using random forest classifier." International Conference in Recent Trends in Computational Methods, Communication and Controls. 2012.
11. World Health Organization, "Cardiovascular diseases (CVDs) - key facts". (March 16, 2019). Retrieved from https://www.who.int/news-room/fact-sheets/detail/ cardiovascular-diseases-(cvds)
12. American Heart Association, "Lifestyle Changes for Heart Attack Prevention". (March 16, 2019). Retrieved from https://www.heart.org/en/health-topics/heart-attack/life-after-a-heart-attack/lifestyle-changes-for-heart-attack-prevention

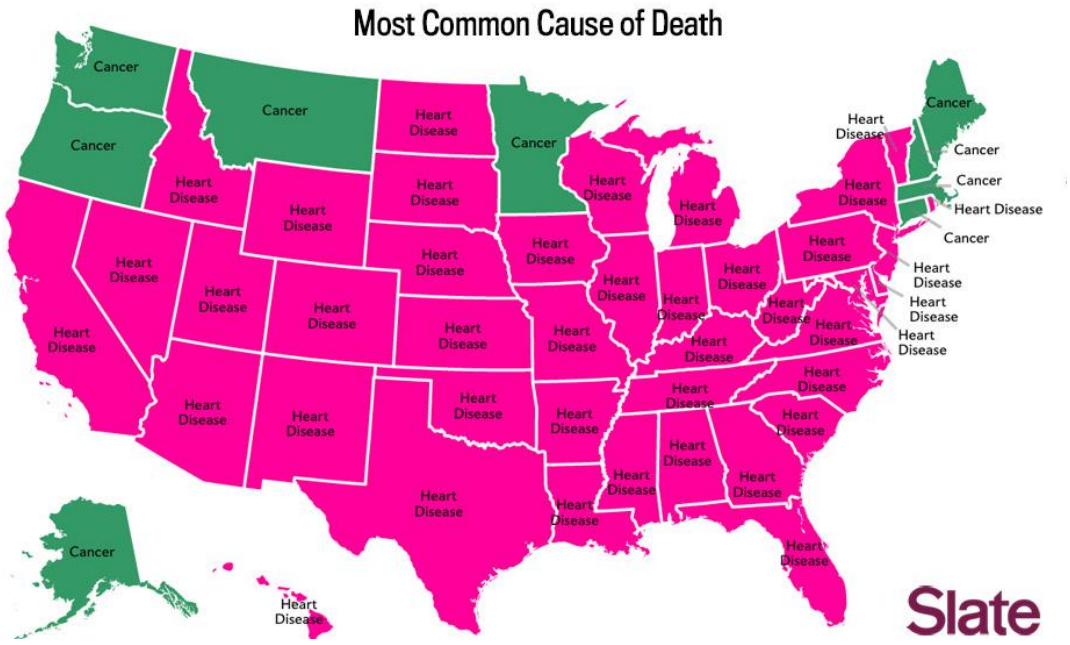
10 APPENDIX

Figure (1) Narrowing of blood vessel



Source: www.sdc.gov

Figure (2) Most Common Cause of Death



Source: https://www.ukprogressive.co.uk/live-in-alabama-heres-how-you-will-die/article27210.html

Figure (3). Feature selected by various methods

	Model	Feature Selected	Acc	AUC	runtime
Stepwise Backwards Removal, no cross validation	Decision Tree	['age', 'cp', 'thalach', 'oldpeak', 'thal']	0.63	0.62	0.0099
	Random Forest	['age', 'cp', 'thalach', 'oldpeak', 'thal']	0.76	0.84	0.19
	Gradient Boosting,	['age', 'cp', 'thalach', 'oldpeak', 'thal']	0.68	0.75	0.072
	ADA Boost	['age', 'cp', 'thalach', 'oldpeak', 'thal']	0.73	0.84	0.11
	Neural Network	['age', 'cp', 'thalach', 'oldpeak', 'thal']	0.82	0.88	0.39
	SVC	['age', 'cp', 'thalach', 'oldpeak', 'thal']	0.83	0.88	0.19
Wrapper Select via model Random Forest, no cross validation	Decision Tree	['age', 'cp', 'trestbps', 'chol', 'thalach', 'oldpeak', 'ca', 'thal']	0.79	0.79	0
	Random Forest	['age', 'cp', 'trestbps', 'chol', 'thalach', 'oldpeak', 'ca', 'thal']	0.85	0.89	0.22
	Gradient Boosting,	['age', 'cp', 'trestbps', 'chol', 'thalach', 'oldpeak', 'ca', 'thal']	0.83	0.90	0.07
	ADA Boost	['age', 'cp', 'trestbps', 'chol', 'thalach', 'oldpeak', 'ca', 'thal']	0.86	0.90	0.14
	Neural Network	['age', 'cp', 'trestbps', 'chol', 'thalach', 'oldpeak', 'ca', 'thal']	0.80	0.89	0.472
	SVC	['age', 'cp', 'trestbps', 'chol', 'thalach', 'oldpeak', 'ca', 'thal']	0.83	0.89	0.81
Cross Validation, feature selection Stepwise Backwards Removal	Decision Tree	['age', 'cp', 'thalach', 'oldpeak', 'thal']	0.71	0.70	0.02
	Random Forest	['age', 'cp', 'thalach', 'oldpeak', 'thal']	0.75	0.83	1.12
	Gradient Boosting,	['age', 'cp', 'thalach', 'oldpeak', 'thal']	0.75	0.83	0.3
	ADA Boost	['age', 'cp', 'thalach', 'oldpeak', 'thal']	0.78	0.85	0.7
	Neural Network	['age', 'cp', 'thalach', 'oldpeak', 'thal']	0.80	0.88	2.415
	SVC	['age', 'cp', 'thalach', 'oldpeak', 'thal']	0.80	0.87	1.404

Figure (4). Features selected by various methods

	Model	Feature Selected	Acc	AUC	runtime
Cross Validation, Wrapper Select via model Random Forest	Decision Tree	['age', 'cp', 'chol', 'thalach', 'oldpeak', 'ca', 'thal']	0.73	0.73	0.0299
	Random Forest	['age', 'cp', 'chol', 'thalach', 'oldpeak', 'ca', 'thal']	0.79	0.88	1.155
	Gradient Boosting,	['age', 'cp', 'chol', 'thalach', 'oldpeak', 'ca', 'thal']	0.77	0.88	0.31
	ADA Boost	['age', 'cp', 'chol', 'thalach', 'oldpeak', 'ca', 'thal']	0.80	0.88	0.72
	Neural Network	['age', 'cp', 'chol', 'thalach', 'oldpeak', 'ca', 'thal']	0.54	0.27	0.13
	SVC	['age', 'cp', 'chol', 'thalach', 'oldpeak', 'ca', 'thal']	0.82	0.89	8.752
Cross Validation, Wrapper Select via model Gradient Boosting	Decision Tree	['age', 'cp', 'oldpeak', 'ca', 'thal']	0.76	0.75	0.0199
	Random Forest	['age', 'cp', 'oldpeak', 'ca', 'thal']	0.79	0.87	1.1
	Gradient Boosting,	['age', 'cp', 'oldpeak', 'ca', 'thal']	0.77	0.87	0.29
	ADA Boost	['age', 'cp', 'oldpeak', 'ca', 'thal']	0.80	0.89	0.75
	Neural Network	['age', 'cp', 'oldpeak', 'ca', 'thal']	0.82	0.87	2.4
	SVC	['age', 'cp', 'oldpeak', 'ca', 'thal']	0.82	0.89	0.18
Cross Validation, Wrapper Select - SVC	Decision Tree	['sex', 'cp', 'fbs', 'exang', 'slope', 'ca', 'thal']	0.80	0.82	0.03
	Random Forest	['sex', 'cp', 'fbs', 'exang', 'slope', 'ca', 'thal']	0.82	0.90	1.198
	Gradient Boosting,	['sex', 'cp', 'fbs', 'exang', 'slope', 'ca', 'thal']	0.82	0.89	0.564
	ADA Boost	['sex', 'cp', 'fbs', 'exang', 'slope', 'ca', 'thal']	0.85	0.92	1.018
	Neural Network	['sex', 'cp', 'fbs', 'exang', 'slope', 'ca', 'thal']	0.79	0.90	3.378
	SVC	['sex', 'cp', 'fbs', 'exang', 'slope', 'ca', 'thal']	0.83	0.90	0.0499
Cross Validation, Univariate Feature Selection - Chi-squared k=5	Decision Tree	['thalach', 'exang', 'oldpeak', 'ca', 'thal']	0.75	0.74	0.03
	Random Forest	['thalach', 'exang', 'oldpeak', 'ca', 'thal']	0.80	0.86	1.11
	Gradient Boosting,	['thalach', 'exang', 'oldpeak', 'ca', 'thal']	0.77	0.85	0.29
	ADA Boost	['thalach', 'exang', 'oldpeak', 'ca', 'thal']	0.82	0.88	0.75
	Neural Network	['thalach', 'exang', 'oldpeak', 'ca', 'thal']	0.82	0.90	3.196
	SVC	['thalach', 'exang', 'oldpeak', 'ca', 'thal']	0.83	0.90	0.884

Figure (5). Most significant variables

Variable name	Description	Chi2
ca	serum cholestoral in mg/dl, level of triglycerides	84.12
thal	1 = smoker ; 0 = not smoker	65.48
exang	exercise induced angina (1 = yes; 0 = no)	38.05
cp	chest pain type -- Value 1: typical angina; Value 2: atypical angina; Value 3: non-anginal pain; Value 4: asymptomatic	15.14
slope	the slope of the peak exercise ST segment: 1: upsloping, 2: flat, 3: downsloping	8.24
sex	1= male, 0=famale	7.43
fbs	fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)	0.16

Figure (6) Random Forest scores

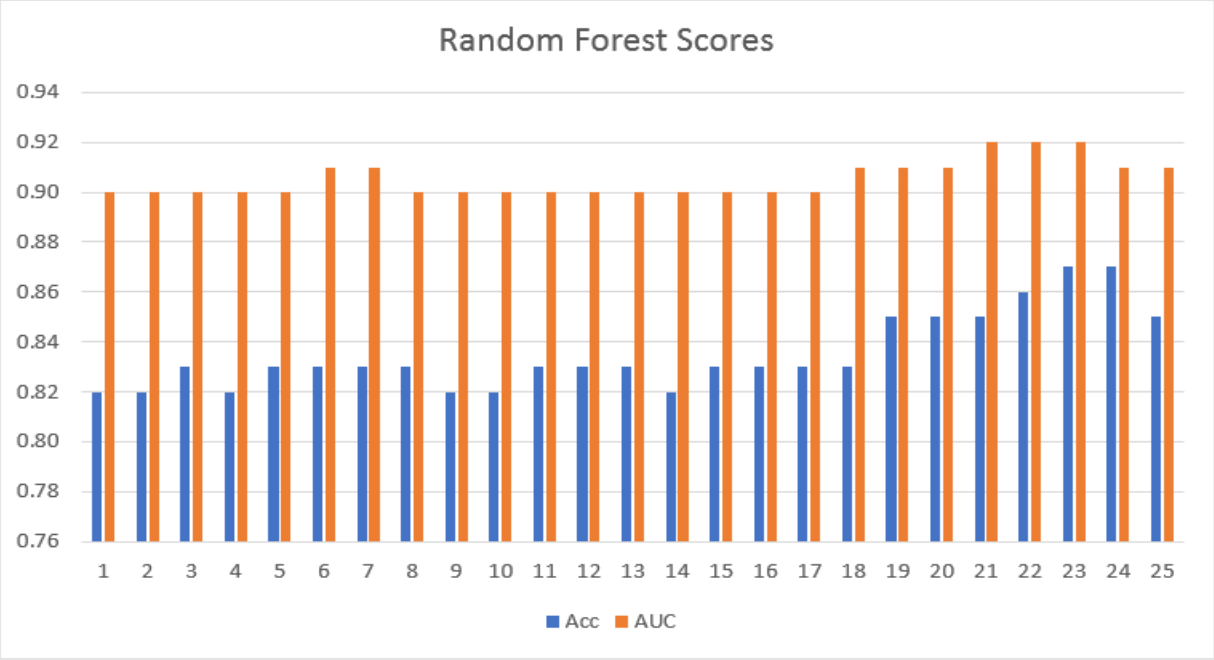


Figure (7) Gradient Boosting scores

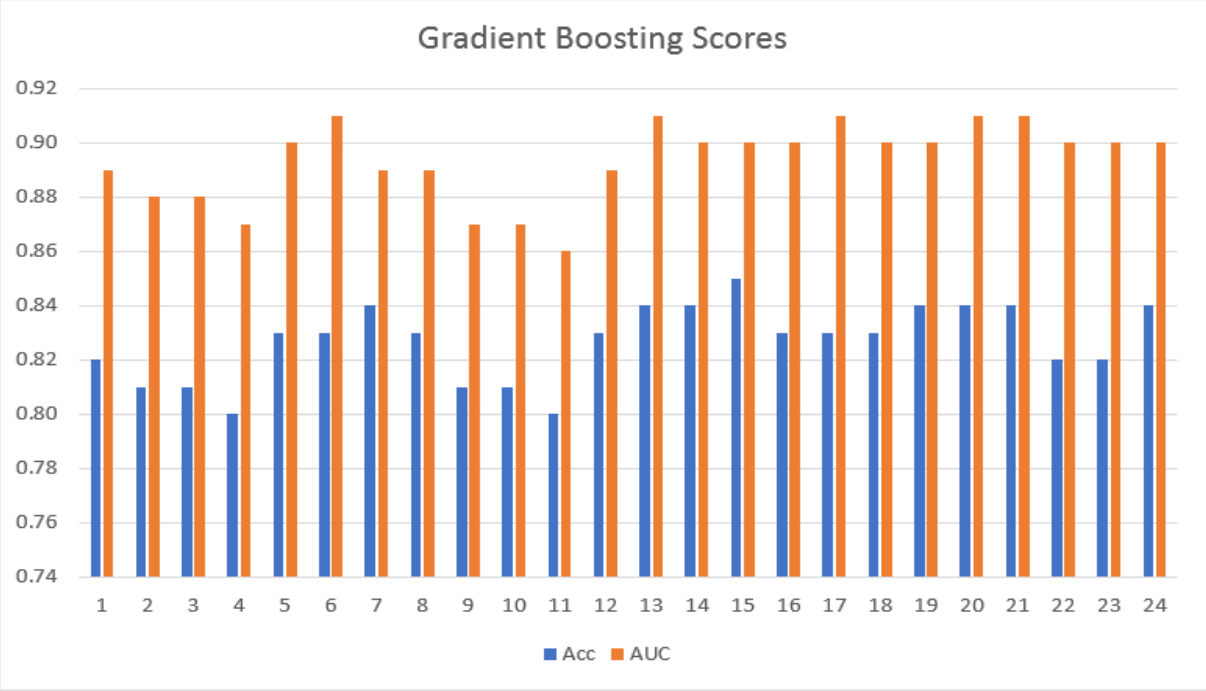


Figure (8) Ada Boost scores

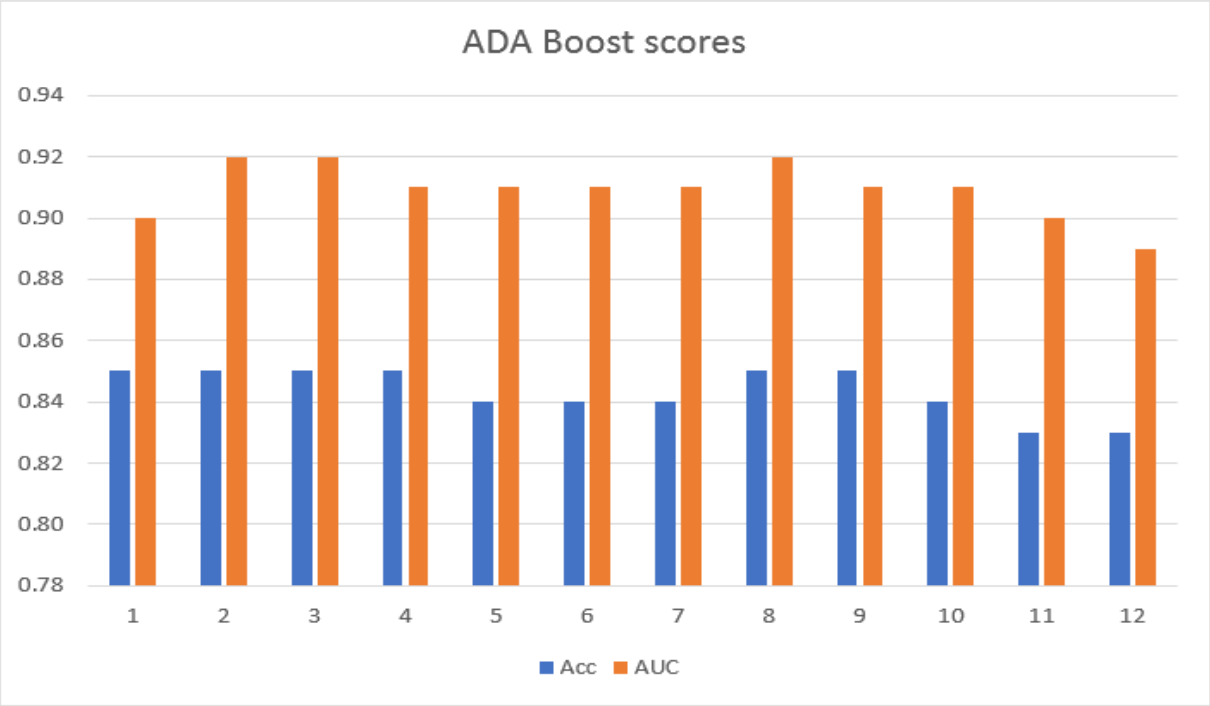


Figure (9) Neural Network scores

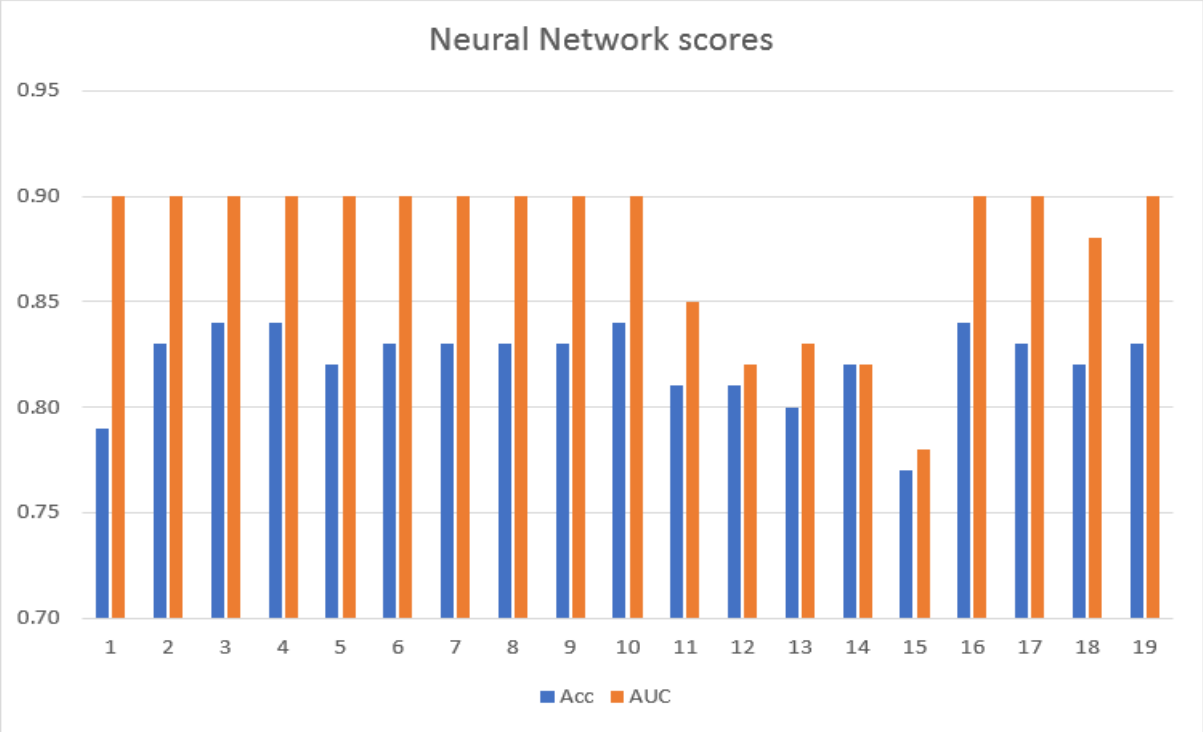


Figure (10) Support Vector Classifier scores

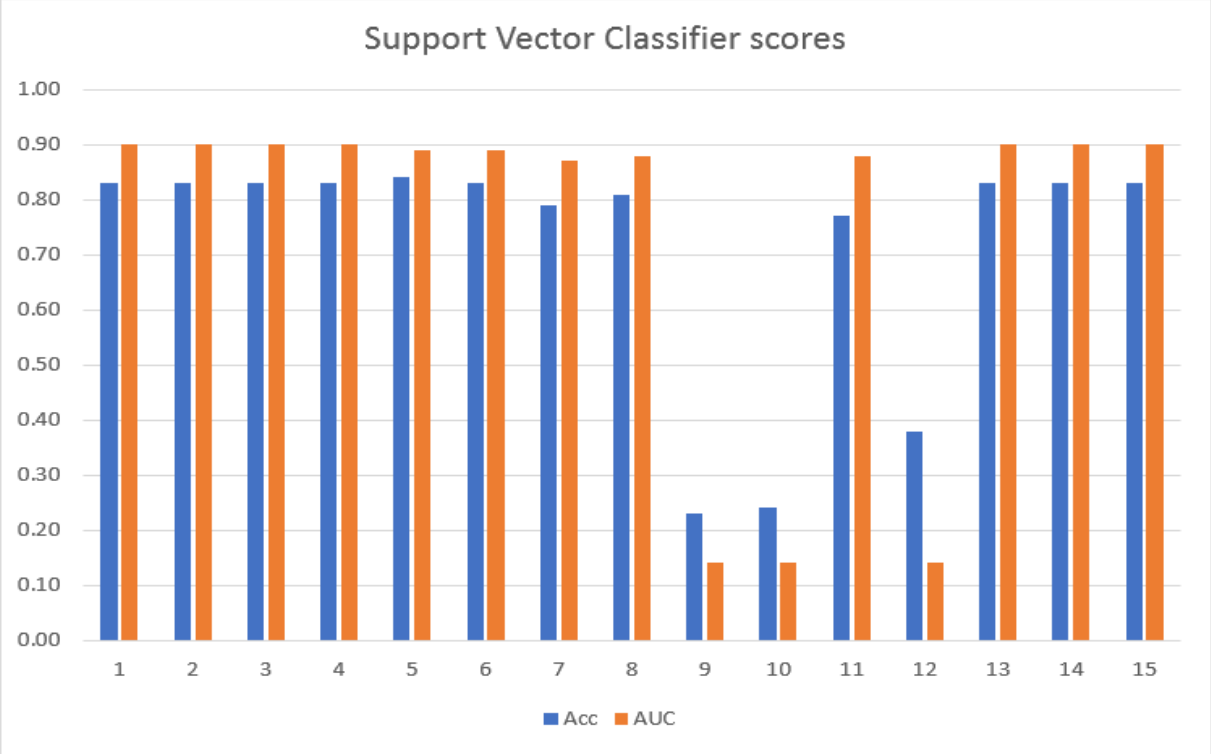
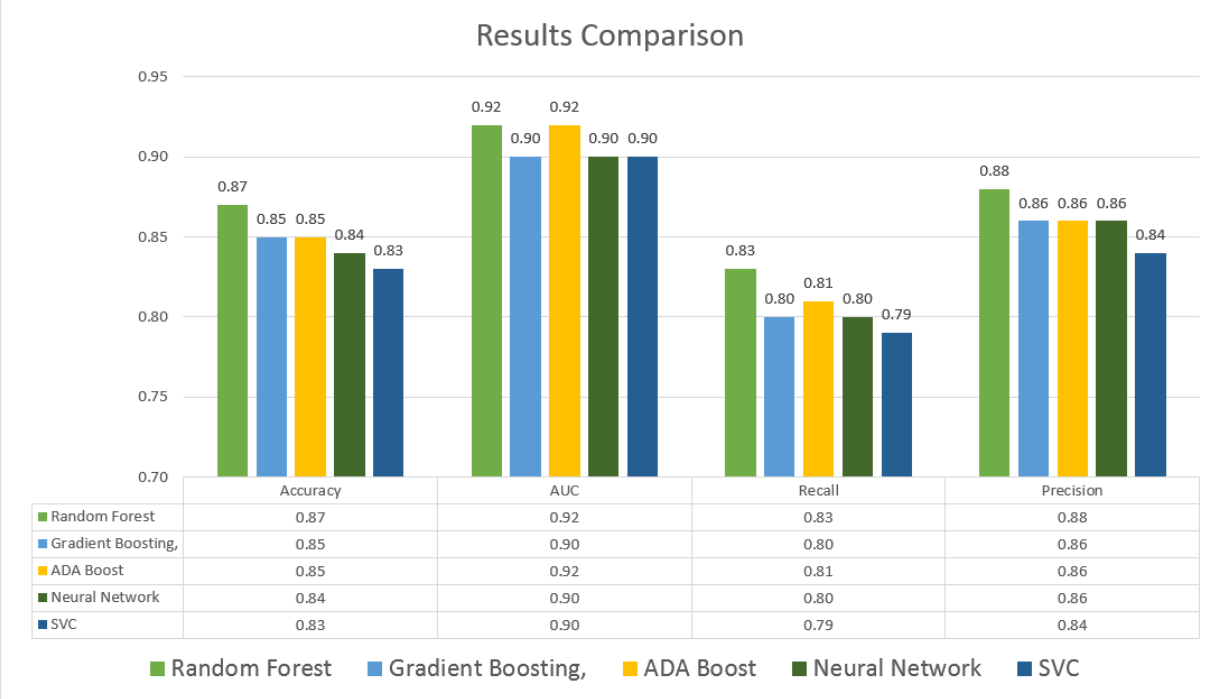


Figure (11) Result Comparison



	Accuracy	AUC	Recall	Precision
Random Forest	0.87	0.92	0.83	0.88
Gradient Boosting,	0.85	0.90	0.80	0.86
ADA Boost	0.85	0.92	0.81	0.86
Neural Network	0.84	0.90	0.80	0.86
SVC	0.83	0.90	0.79	0.84