John Attisha and Thavaselvi Ramalingam

COMPARISON OF ENSEMBLE MODELS FOR CLASSIFICATION OF CARDIOVASCULAR DISEASE

Abstract

According to the Centers for Disease Control, heart disease is the leading cause of death in the United States. Patients who are aware of their susceptibility to heart disease can benefit by working with their health provider to address their medical condition in a proactive manner. In this report, we aim to develop a method to classify cardiovascular disease in a patient, as well as to compare various ensemble models to identify which approach performs best with respect to our particular dataset. We attempt to solve two classification problems: multi-class classification, where each class corresponds to a different level of heart disease; and binary-class classification, where we aim to distinguish between presence and absence of cardiovascular disease. Three ensemble models are considered, including Random Forests, AdaBoost and Extremely Randomized Trees. A decision tree model was constructed in order to serve as a baseline comparison to the ensemble models. With regard to the multi-class problem, which was highly imbalanced, accuracies for the various models ranged between 57% and 60%. Sensitivity was high for the majority class (Class 0, which corresponds to no heart disease), but relatively low for all other classes. With regard to the binary-class problem, which was more balanced, accuracies for the various models ranged between 80% and 83%. Additionally, sensitivity improved substantially compared to the multi-class problem. Paired t-tests identified that the accuracy between the baseline decision tree and each of the ensemble models was statistically the same, indicating that a complex ensemble model was not necessarily required for this dataset.

Introduction

Over the past three decades, there has been a significant decrease in cardiovascular-related deaths within the United States, thanks in main part to clinical intervention and a reduction in risk factors.[1] According to the Centers for Disease Control, however, cardiovascular disease still remains the leading cause of death within the U.S.[2] Needless to say, patients with attributable risk factors would benefit in knowing if they are susceptible to getting heart disease. In addition to the domain expertise of medical professionals, proper data mining and machine learning techniques can be used to leverage patient information and predict whether an individual may or may not have cardiovascular disease.

In this project, we aim to solve two problems. First, we attempt to classify whether patients have cardiovascular disease based on related health attributes. Second, we aim to identify how various ensemble classifiers work with regard to our database and determine which algorithm performs best based on various performance metrics.

Related Work

Ensemble classifiers have become increasingly popular in the fields of machine learning and data mining, as they have shown to be more accurate than their individual base classifiers.[8] Construction of base learners for ensemble classifiers can be executed through a number of different methods. Techniques include, but are certainly not limited to, resampling of the training set (such as Bagging and Boosting), introducing some form of randomness (such as Random Forests), and manipulation of features (such as Rotation Forests). Random Forests combine the techniques of bagging and the random selection of features, often achieving high classification accuracy and reducing overall variance.[9] An example of an ensemble classifier which implements boosting is AdaBoost, which attempts to construct a succession of weak learners, with each subsequent base classifier focusing on incorrectly classified cases.[10]

The Cleveland Clinic dataset has been used to investigate sick and healthy factors which contribute to heart disease for males and females using association rule mining.[11] This research found that asymptomatic chest pain and presence of exercise-induced angina indicate the likely existence of heart disease in both men and women. This dataset has also been used to build a weighted fuzzy rule based clinical decision support for the diagnosis of heart disease.[12] Sumit et al showed that the use of simple Support Vector Machines with an optimal feature set improved the accuracy for the multi-class Cleveland dataset.[13] Optimal feature selection was done using an integer-coded genetic algorithm. Additionally, Akhil Jabbar et al proposed the use of k-nearest neighbors with a genetic algorithm to improve the accuracy for this dataset.[14] Our project is different from these works as we are implementing ensemble classifiers to improve the accuracy of the risk prediction.

Methodology

The Cleveland Clinic has released a dataset, available through the UCI Machine Learning Repository, which contains heart-related information on 303 patients.[3] There are a total of 13 cardiovascular-related attributes, of which six are numeric, three are binary categorical, and four are multi-class categorical. These attributes contain information such as a patient's resting blood pressure, electrocardiographic results, and the number of major vessels colored by fluoroscopy. A full description of the predictor attributes, as well as summary statistics, can be found in Tables 3 and 4 in the Appendix. With regard to the target variable, there are five levels: {0, 1, 2, 3, and 4}, where 0 indicates the absence of cardiovascular disease, and 1-4 indicate varying levels of cardiovascular disease (the higher the number, the worse the patient's condition). As seen in Table 1, the dataset is fairly imbalanced in that most people do not have heart disease.

Class	Count
0	164
1	55
2	36
3	35
4	13

Table 1 – Multi-Class Distribution of Cardiovascular Database

Most experiments with regard to this database have concentrated on distinguishing presence of heart disease from absence, thus transforming the task into a binary classification problem.[4] In this case, the class distribution becomes fairly balanced, as seen in Table 2. In this project, we perform two classification tasks, one on the imbalanced multi-class database, and one on the balanced binary-class database.

Class	Count
0	164
1	139

Table 2 – Binary-Class Distribution of Cardiovascular Database

With regard to the predictor variables, there were a handful of missing values in our database. The missing values were replaced by the mean of the variable, where the mean was calculated by taking into account only those observations which shared the same class label as the observation with the missing value. We felt that this would more appropriately approximate the missing value, rather than taking the overall mean of the variable. Next, in order to implement our various classifiers, all categorical variables had to be converted to dummy variables. This involved replacing integers (which represented specific labels) into the actual labels themselves, and then converting the database into a standard spreadsheet format. Data in the standard spreadsheet format was normalized via min-max normalization, such that all values were transformed to the range [0, 1]. Lastly, a second set of class labels was created for our target variable in order to perform the binary classification task. This was done by converting the multi-class target variable (consisting of {0, 1, 2, 3, 4}) into a binary-class target variable (consisting of {0, 1}, where 1 consisted of the previous labels {1, 2, 3, 4}).

In terms of the distribution of our data, ages ranged from 29 to 77 years, with a mean of 54 years. Out of the 303 patients in our database, 206 patients were male, and 97 were female. As seen in Figure 1, after transforming our class label to become binary, we see that men are much more likely to have heart disease as compared to women. With regard to resting blood pressure, we see in Figure 2 that these values exhibit a fairly normal distribution. The mean resting blood pressure is 132 mm Hg, with a standard deviation of 17.6 mm Hg. In terms of serum cholesterol levels and maximum heart rate achieved, we also find fairly normal distributions, with mean levels of 246 mg/dl and 149 beats per minute (bpm), respectively. Plots for these variables, as well as for other predictor variables, can be found in the Appendix at the end of this report. Examining the correlation matrix and scatterplot of our variables, we find somewhat interesting results. Focusing on the binary class label, we see that diagnosis of heart disease has the highest correlation with the dummy variable cp_asymptomatic, with a correlation value of 0.52. This dummy variable refers to patients who exhibit no signs of chest pain. It may be the case that individuals who do not have chest pain in general are still highly susceptible to having heart disease, perhaps unknowingly taking part in actions that may cause their medical condition to worsen. Surprisingly, we also see that exercise induced angina also has a relatively high correlation value with the diagnosis of heart disease (binary class label = 1), with a correlation value of 0.43. Patients who have chest pain while exercising should treat this symptom as a potential leading indicator of cardiovascular disease, and perhaps consider seeking medical attention.

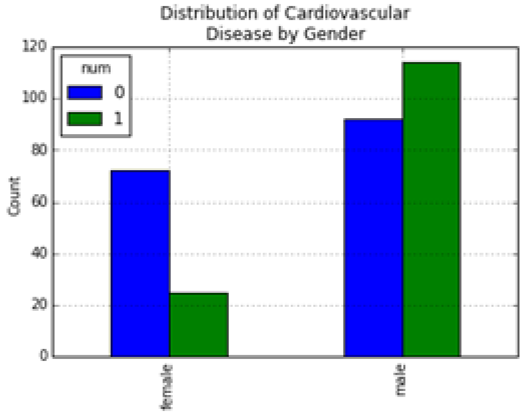


Figure 1

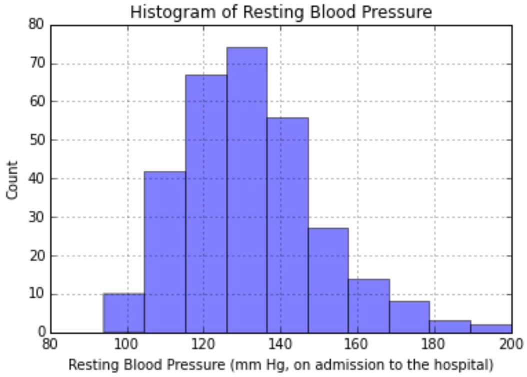


Figure 2

In terms of the methodology of our approaches, we implemented three ensemble models in order to identify which method would perform best with regard to our specific dataset. We also constructed a decision tree classifier which would serve as a baseline comparison to each of our ensemble models. As previously mentioned, our approach was performed on two datasets: a multi-class, imbalanced dataset; and a binary-class, balanced dataset. Each model was fine-tuned by running a grid search process, considering a wide spectrum of values for various parameters for each of our models. The three ensemble methods considered were AdaBoost, Random Forests and Extremely Randomized Trees.

AdaBoost fits a series of weak learners on data which is repeatedly modified such that greater focus is placed on incorrectly classified cases for subsequently constructed learners.[5] Predictions from each of the weak learners (decision trees in this particular case) are combined via weighted majority voting, resulting in a single overall classification. With regard to Random Forests, base classifiers (also decision trees in this case) are built in parallel, with each classifier being trained on a bootstrap sample of the training set. Additionally, a split in the node of a base classifier is chosen as the best split among a random subset of features. For this report, we have utilized Python's Scikit-Learn version of the Random Forest algorithm, which combines classifiers by averaging their probabilistic prediction (as opposed to the original Random Forest publication, which lets each classifier vote for a single class).[6] Lastly, Extremely Randomized Trees takes Random Forests and goes one step further with regard to how randomness is leveraged to determine the best split at a node. In addition to a random selection of candidate features, this method also identifies a random selection of candidate thresholds on which to consider a split. The most discriminative of these random thresholds is chosen as the threshold on which to split the node. [7, 15]

The data was split into 80% training and 20% testing. For each of the four classifiers (three ensemble classifiers and one decision tree classifier), a grid search was implemented in order to try every possible combination of the selected parameter values. 10-fold cross-validation was performed for each combination, and the best parameter values were chosen by determining the combination which resulted in the highest cross-validation accuracy. The best performing parameter values for each of the four classifiers were then used to construct the models, and each of the models were then tested on the remaining 20% of the data which was not used in the grid search process. The range of parameter values considered, and the final values selected for each classifier, can be seen below.

Model Construction for Binary-Class Problem

With regard to the decision tree, the following parameter values were considered for the grid search:
-Criterion: [Entropy, Gini]
-Max Depth: The range of numbers from [1, 20], by 1.
-Min Samples Leaf: The range of numbers from [2, 30], by 2.
-Min Samples Split: The range of numbers from [2, 20], by 2.
After running the grid search on every possible combination of parameter values listed above, and performing 10-fold cross-validation on each combination, it was determined that the best parameter values for the decision tree classifier were as follows. Criterion: Entropy; Max Depth: 4; Min Samples Leaf: 12; Min Samples Split: 2.

With regard to AdaBoost, the following parameter values were considered for the grid search:
-N Estimators: The range of numbers from [5, 75], by 5.
-Learning Rate: 30 evenly spaced numeric values in the range [0.0001, 1]
After running the grid search on every possible combination of parameter values listed above, and performing 10-fold cross-validation on each combination, it was determined that the best parameter values for the AdaBoost ensemble classifier were as follows. N Estimators: 70; Learning Rate: 0.55.

With regard to the Random Forest ensemble classifier, the following parameter values were considered for the grid search:
-Criterion: [Entropy, Gini]
-N Estimators: The range of numbers from [5, 50], by 5.
-Max Depth: The range of numbers from [1, 20], by 1.
-Min Samples Split: The range of numbers from [2, 20], by 2.
After running the grid search on every possible combination of parameter values listed above, and performing 10-fold cross-validation on each combination, it was determined that the best parameter values for the Random Forest ensemble classifier were as follows. Criterion: Entropy; N Estimators: 20; Max Depth: 17; Min Samples Split: 16.

With regard to the Extremely Randomized Trees ensemble classifier, the following parameter values were considered for the grid search:
-Criterion: [Entropy, Gini]
-N Estimators: The range of numbers from [5, 50], by 5.
-Max Depth: The range of numbers from [1, 20], by 2.
-Min Samples Split: The range of numbers from [2, 20], by 2.
After running the grid search on every possible combination of parameter values listed above, and performing 10-fold cross-validation on each combination, it was determined that the best parameter values for the Extremely Randomized Trees classifier were as follows. Criterion: Entropy; N Estimators: 5; Max Depth: 3; Min Samples Split: 6.

Model Construction for Multi-Class Problem

With regard to the decision tree, the following parameter values were considered for the grid search:
-Criterion: [Entropy, Gini]
-Max Depth: The range of numbers from [1, 20], by 1.
-Min Samples Leaf: The range of numbers from [2, 30], by 2.
-Min Samples Split: The range of numbers from [2, 20], by 2.
After running the grid search on every possible combination of parameter values listed above, and performing 10-fold cross-validation on each combination, it was determined that the best parameter values for the decision tree classifier were as follows. Criterion: Gini; Max Depth: 4; Min Samples Leaf: 12; Min Samples Split: 2.

With regard to AdaBoost, the following parameter values were considered for the grid search:
-N Estimators: The range of numbers from [5, 75], by 5.

-Learning Rate: 30 evenly spaced numeric values in the range [0.0001, 1]

After running the grid search on every possible combination of parameter values listed above, and performing 10-fold cross-validation on each combination, it was determined that the best parameter values for the AdaBoost ensemble classifier were as follows. N Estimators: 10; Learning Rate: 0.6.

With regard to the Random Forest ensemble classifier, the following parameter values were considered for the grid search:

-Criterion: [Entropy, Gini]
-N Estimators: The range of numbers from [5, 50], by 5.
-Max Depth: The range of numbers from [1, 20], by 2.
-Min Samples Split: The range of numbers from [2, 20], by 2.

After running the grid search on every possible combination of parameter values listed above, and performing 10-fold cross-validation on each combination, it was determined that the best parameter values for the Random Forest ensemble classifier were as follows. Criterion: Entropy; N Estimators: 40; Max Depth: 20; Min Samples Split: 21.

With regard to the Extremely Randomized Trees ensemble classifier, the following parameter values were considered for the grid search:

-Criterion: [Entropy, Gini]
-N Estimators: The range of numbers from [5, 50], by 5.
-Max Depth: The range of numbers from [1, 20], by 2.
-Min Samples Split: The range of numbers from [2, 50], by 3.

After running the grid search on every possible combination of parameter values listed above, and performing 10-fold cross-validation on each combination, it was determined that the best parameter values for the Extremely Randomized Trees classifier were as follows. Criterion: Gini; N Estimators: 20; Max Depth: 20; Min Samples Split: 26.

Results

Accuracy results obtained from the base and ensemble models for the multi-class dataset can be seen in Table 3. Both the baseline and ensemble classifiers provide similar results. The accuracy ranges between 57% and 60%.

	Decision Tree	Random Forest	AdaBoost	Extra Trees Classifier
Training	66.03 ± 0.64	71.52 ± 0.55	**62.58 ± 0.82**	71.57 ± 0.61
Testing	57.70 ± 2.72	59.42 ± 2.04	**59.83 ± 2.64**	58.60 ± 2.27

Table 3 – Multi-Class Accuracy results (Average over 20 trials)

As this dataset is unbalanced, the accuracy results could be misleading. Since our goal is to predict patients who have cardiovascular disease, performance measures like sensitivity and specificity were also evaluated. Tables 4 through 8 provide sensitivity and specificity values for each of the five classes.

	Decision Tree		Random Forest		AdaBoost		Extra Trees Classifier	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Training	92.26 ± 0.82	74.20 ± 2.05	99.11 ± 0.29	57.99 ± 2.14	94.32 ± 0.99	70.54 ± 2.72	97.88 ± 0.28	60.79 ± 2.05
Testing	86.49 ± 2.61	72.31 ± 3.17	97.47 ± 1.25	50.84 ± 3.93	92.79 ± 1.96	72.37 ± 4.17	96.54 ± 1.20	52.69 ± 3.50

Table 4 – Class 0: Sensitivity & Specificity Results (Averaged Over 20 Trials)

	Decision Tree		Random Forest		AdaBoost		Extra Trees Classifier	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Training	37.56 ± 4.73	87.45 ± 1.60	26.63 ± 4.58	96.25 ± 1.45	26.96 ± 4.41	85.21 ± 2.38	30.57 ± 4.46	95.03 ± 1.04
Testing	22.66 ± 5.63	82.88 ± 3.15	6.68 ± 3.55	92.19 ± 2.55	25.97 ± 6.71	82.62 ± 3.36	4.98 ± 2.03	92.04 ± 2.00

Table 5 – Class 1: Sensitivity & Specificity Results (Averaged Over 20 Trials)

	Decision Tree		Random Forest		AdaBoost		Extra Trees Classifier	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Training	42.62 ± 6.39	93.21 ± 1.72	52.89 ± 4.16	96.99 ± 0.66	27.46 ± 7.15	92.9 ± 2.37	49.09 ± 5.05	97.54 ± 0.5
Testing	26.80 ± 9.51	91.65 ± 3.03	21.78 ± 7.06	93.92 ± 1.77	23.98 ± 8.81	91.42 ± 2.40	17.04 ± 6.99	94.29 ± 1.37

Table 6 – Class 2: Sensitivity & Specificity Results (Averaged Over 20 Trials)

	Decision Tree		Random Forest		AdaBoost		Extra Trees Classifier	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Training	36.32 ± 8.2	93.62 ± 2.04	56.14 ± 4.05	96.15 ± 0.87	27.59 ± 8.20	93.77 ± 1.82	58.58 ± 3.13	95.38 ± 0.96
Testing	21.13 ± 8.7	90.67 ± 2.77	17.82 ± 7.69	92.34 ± 1.62	15.82 ± 6.71	93.64 ± 2.50	19.39 ± 5.87	90.47 ± 1.64

Table 7 – Class 3: Sensitivity & Specificity Results (Averaged Over 20 Trials)

	Decision Tree		Random Forest		AdaBoost		Extra Trees Classifier	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Training	2.92 ± 5.57	99.80 ± 0.37	6.65 ± 5.39	100 ± 0.00	4.33 ± 2.91	99.91 ± 0.08	10.41 ± 4.07	99.97 ± 0.04
Testing	0.00 ± 0.00	99.66 ± 0.64	0.00 ± 0.00	100 ± 0.00	0.00 ± 0.00	99.91 ± 0.17	0.00 ± 0.00	99.83 ± 0.23

Table 8 – Class 4: Sensitivity & Specificity Results (Averaged Over 20 Trials)

Results in the tables above highlight that the classifiers have high sensitivity for Class 0 (absence of cardiovascular disease), but sensitivity for Classes 1, 2 and 3 are very low, and sensitivity for Class 4 is zero. This shows that the classifiers do not predict Class 4. This is due to the fact that the number of instances in Class 4 is very low (13 instances in total).

In order to boost the sensitivity results for instances which possess cardiovascular disease, the data was converted into a binary class problem. Classes 1, 2, 3 and 4 were binned together, as they represent patients who have cardiovascular disease. Accuracy results obtained for the binary class problem can be seen in Table 9. Both the baseline and ensemble classifiers provide similar accuracy results. Random Forest has the highest mean test accuracy, though the confidence interval is relatively wide. AdaBoost provides similar accuracy results, with a relatively narrower range for the confidence interval.

	Decision Tree	Random Forest	AdaBoost	Extra Trees Classifier
Training	85.56 ± 0.36	90.33 ± 0.48	91.28 ± 0.48	83.55 ± 0.91
Testing	81.88 ± 1.66	**83.36 ± 2.18**	**82.62 ± 1.64**	80.08 ± 2.35

Table 9 – Binary Class Accuracy Results (Averaged Over 20 Trials)

In order to verify the each classifier's performance on the positive class, sensitivity and specificity measures were calculated, and can be seen in Table 10. Sensitivity and specificity results have improved dramatically compared to our multi-class problem, as the data is now much more balanced. The Random Forest classifier has relatively high values for both sensitivity and specificity.

	Decision Tree		Random Forest		AdaBoost		Extra Trees Classifier	
	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
Training	91.01 ± 0.92	79.08 ± 1.52	93.38 ± 0.65	86.75 ± 0.91	93.41 ± 0.51	88.75 ± 0.88	86.39 ± 1.45	80.16 ± 1.50
Testing	88.93 ± 2.67	73.87 ± 2.87	**85.94 ± 2.60**	**80.59 ± 3.47**	86.84 ± 2.77	77.72 ± 0.74	83.62 ± 3.75	76.73 ± 3.37

Table 10 – Sensitivity & Specificity Results (Averaged Over 20 Trials)

As many of the classifiers have similar accuracy, sensitivity and specificity results, it is important to determine whether these metrics are significantly different between each of these models. A paired t-test on the testing accuracies of these models was performed for both multi-class and binary class datasets. Results can be seen in Table 11. The p-values for all the multi-class model comparisons are greater than 0.05, and hence we fail to reject the null hypothesis. Therefore, the accuracies are not significantly different between the base and ensemble models. For binary-class problem, the p-value is less than 0.05 (5% significance level) for the Random Forest/Extra Trees comparison, as well as for the AdaBoost/Extra Trees comparison, indicating that the accuracies are significantly different between each of these two comparisons. On the other hand, accuracies are not significantly different between the decision tree classifier and each of the ensemble classifiers.

	Multi-Class		Binary-Class	
	t-statistic	p-value	t-statistic	p-value
Decision Tree & Random Forest	-1.34	0.19	-1.29	0.21
Decision Tree & AdaBoost	-1.80	0.08	-0.84	0.41
Decision Tree & Extra Trees	-0.82	0.42	1.74	0.09
Random Forest & AdaBoost	-0.54	0.59	0.68	0.50
Random Forest & Extra Trees	0.97	0.33	3.24	**0.004**
AdaBoost & Extra Trees	1.46	0.16	2.60	**0.017**

Table 11 – Accuracy Comparison: Paired t-test Results

Conclusion

Reviewing the performance of the classifiers, it is evident that with respect to the multi-class problem, accuracy is biased due to the presence of class imbalance in the dataset. The poor performance can also be attributed to the fact that the data points are not linearly separable with respect to class variable. All classifiers considered for this project assume linear separability in the data. Non-linear approaches, such as certain Support Vector Machines, may be applied to the multi-class dataset in order to attempt to improve accuracy results. Various balancing techniques (in addition to reverting to a binary classification problem as performed in our project), may also be helpful to improve overall performance for the multi-class dataset. One such approach to consider is the synthetic minority over-sampling technique (SMOTE).[16] With regard to the balanced, binary classification problem, all four classifiers performed markedly better in terms of the performance metrics considered for this report. With the class labels converted to a binary classification problem, the observations in the dataset become more linearly separable (as compared to the multi-class problem). Additionally, with a balanced dataset, it was identified that performance of baseline classifier was same as that of the three ensemble classifiers, indicating a complex ensemble model was not required for this particular dataset.

Team Member Responsibilities

With regard to the breakdown of responsibilities, both John and Thavaselvi collaborated on a number of the sections throughout the project. In general, work was divided equally between the two team members. John focused his efforts on data pre-processing, exploratory data analysis, development of plots and graphs, and various attempts at model construction. John substantially contributed to the Abstract, Introduction, Methodology, and Appendix sections, and edited the remaining sections of the report. Thavaselvi focused her efforts on model construction and model comparison, and substantially contributed to the Methodology, Results, and Conclusion sections, and edited the remaining sections of the report. The Related Work section was divided equally between the team members. Similarly, both team members contributed equally to the development of the presentation slides.

References

[1] Patel, S. A., Winkel, M., Ali, M. K., Narayan, K. M., & Mehta, N. K., “Cardiovascular mortality associated with 5 leading risk factors: national and state preventable fractions estimated from survey data”. *Annals of Internal Medicine, 163,* 4, 245-53. Jan. 2015.
http://depaul.worldcat.org.ezproxy.depaul.edu/oclc/19415870619473

[2] Heart Disease Facts. Centers for Disease Control and Prevention. 2015
http://www.cdc.gov/heartdisease/facts.htm

[3] Heart Disease Dataset. UCI Machine Learning Repository. 1988.
https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[4] Heart Disease Dataset Description. Paragraph 4. UCI Machine Learning Repository. 1988.
https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/heart-disease.names

[5] AdaBoost Scikit-Learn Documentation
http://scikit-learn.org/stable/modules/ensemble.html#adaboost

[6] Random Forest Scikit-Learn Documentation
http://scikit-learn.org/stable/modules/ensemble.html#random-forests

[7] Extremely Randomized Trees Scikit-Learn Documentation
http://scikit-learn.org/stable/modules/ensemble.html#extremely-randomized-trees

[8] David Optiz, Richard Maclin. “Popular Ensemble Methods: An Empirical Study,” Journal of Artificial Intelligence Research (199): 169-198.
http://www.duluth.umn.edu/~rmaclin/publications/opitz-jair99.pdf

[9] Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.
http://link.springer.com/article/10.1023%2FA%3A1010933404324

[10] Freund, Yoav, and Robert E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting." *Journal of computer and system sciences* 55.1 (1997): 119-139.
http://www.sciencedirect.com/science/article/pii/S002200009791504X

[11] Jesmin Nahar, Tasadduq Imam, Kevin S. Tickle, Yi-Ping Phoebe Chen. “Association rule mining to detect factors which contribute to heart disease in males and females”. Expert Systems with Applications: An International Journal archive Volume 40 Issue 4, March, 2013 Pages 1086-1093 .
https://www.researchgate.net/publication/257404638_Association_rule_mining_to_detect_factors_which_contribute_to_heart_disease_in_males_and_females

[12] P.K. Anooj. “Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules”. Journal of King Saud University - Computer and Information Sciences, Volume 24, Issue 1, January 2012, Pages 27–40
http://www.sciencedirect.com/science/article/pii/S1319157811000346

[13] Sumit Bhatia, Praveen Prakash, and G.N. Pillai. “SVM Based Decision Support System for Heart Disease Classification with Integer-Coded Genetic Algorithm to Select Critical Features”. Proceedings of the World Congress on Engineering and Computer Science 2008 WCECS 2008, October 22 - 24, 2008, San Francisco, USA
http://www.iaeng.org/publication/WCECS2008/WCECS2008_pp34-38.pdf

[14] M.Akhil jabbar, B.L Deekshatulua, Priti Chandra. “Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm”. International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA) 2013
http://arxiv.org/ftp/arxiv/papers/1508/1508.02061.pdf

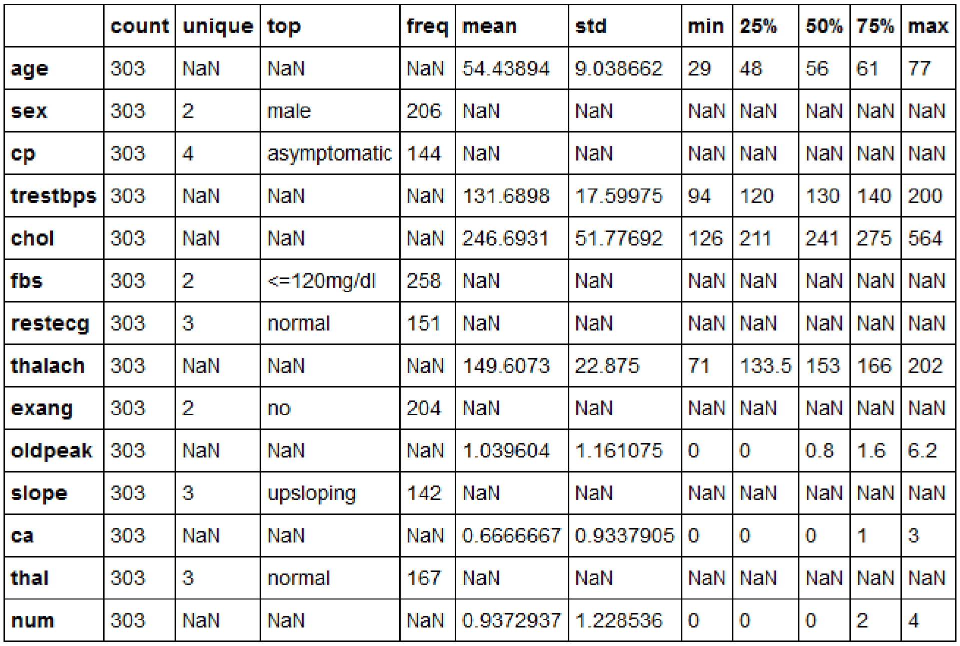
[15] Geurts, Pierre, Damien Ernst, and Louis Wehenkel. "Extremely randomized trees." *Machine learning* 63.1 (2006): 3-42.
http://link.springer.com/article/10.1007/s10994-006-6226-1

[16] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). “Smote: Synthetic minority over-sampling technique.” Journal of Artificial Intelligence Research, 16:321-357.
https://www.jair.org/media/953/live-953-2037-jair.pdf

Appendix

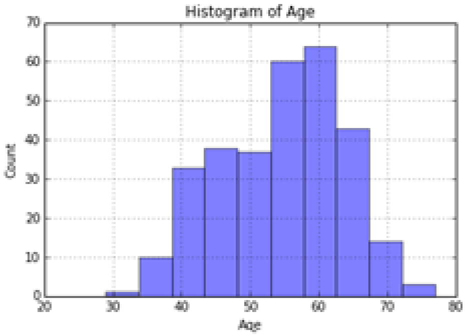
Number	Name	Description	Data Type
1	Age	Age (years)	Continuous integer values
2	Sex	Sex of patient	0 = Female 1 = Male
3	Cp	Chest pain type	1 = Typical Angina 2 = Atypical Angina 3 = Non-Anginal Pain 4 = Asymptomatic
4	trestbps	Resting blood pressure (mm Hg)	Continuous integer values
5	chol	Serum cholesterol (mg/dl)	Continuous integer values
6	fbs	Fasting blood sugar >120 mg/dl	0 = False 1 = True
7	restecg	Resting electrocardiographic results	0 = Normal 1 = Having ST-T Wave Abnormality 2 = Showing Probable or Definite Left Ventricular Hypertrophy
8	thalach	Maximum heart rate achieved	Continuous integer values
9	exang	Exercise induced angina	0 = No 1 = Yes
10	oldpeak	ST depression induced by exercise relative to rest	Continuous real values
11	slope	Slope of the peak exercise ST segment	1 = Upsloping 2 = Flat 3 = Downsloping
12	ca	Number of major vessels colored by fluoroscopy	Continuous integer values in range [0,3]
13	thal	Thalium stress test	3 = Normal 6 = Fixed defect 7 = Reversible defect

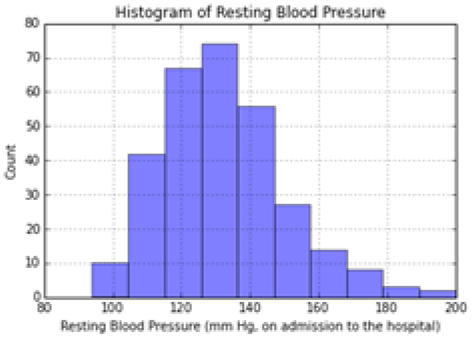
Table 12 – Description of Variables

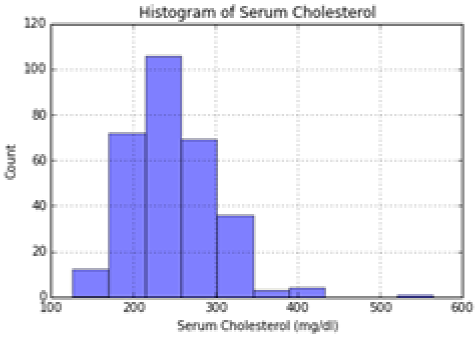


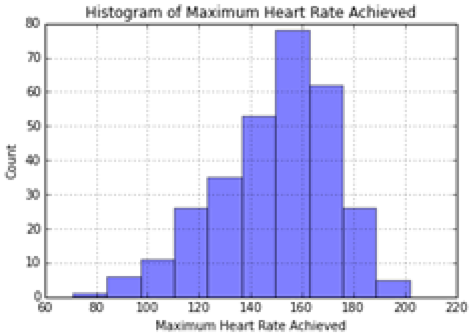
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
age	303	NaN	NaN	NaN	54.43894	9.038662	29	48	56	61	77
sex	303	2	male	206	NaN	NaN	NaN	NaN	NaN	NaN	NaN
cp	303	4	asymptomatic	144	NaN	NaN	NaN	NaN	NaN	NaN	NaN
trestbps	303	NaN	NaN	NaN	131.6898	17.59975	94	120	130	140	200
chol	303	NaN	NaN	NaN	246.6931	51.77692	126	211	241	275	564
fbs	303	2	<=120mg/dl	258	NaN	NaN	NaN	NaN	NaN	NaN	NaN
restecg	303	3	normal	151	NaN	NaN	NaN	NaN	NaN	NaN	NaN
thalach	303	NaN	NaN	NaN	149.6073	22.875	71	133.5	153	166	202
exang	303	2	no	204	NaN	NaN	NaN	NaN	NaN	NaN	NaN
oldpeak	303	NaN	NaN	NaN	1.039604	1.161075	0	0	0.8	1.6	6.2
slope	303	3	upsloping	142	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ca	303	NaN	NaN	NaN	0.6666667	0.9337905	0	0	0	1	3
thal	303	3	normal	167	NaN	NaN	NaN	NaN	NaN	NaN	NaN
num	303	NaN	NaN	NaN	0.9372937	1.228536	0	0	0	2	4

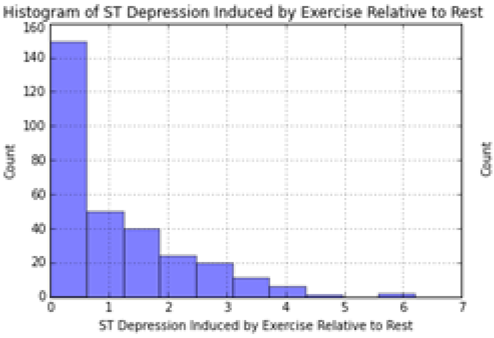
Table 13 – Summary Statistics of Variables

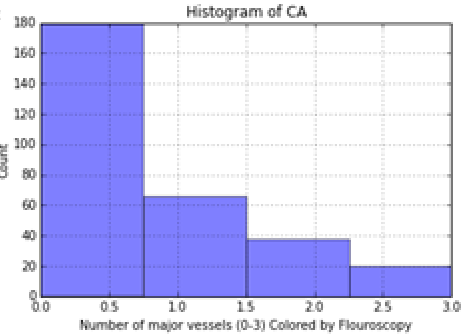

Histogram of Age
Count
0
10
20
30
40
50
60
70
20
30
40
50
60
70
80
Age

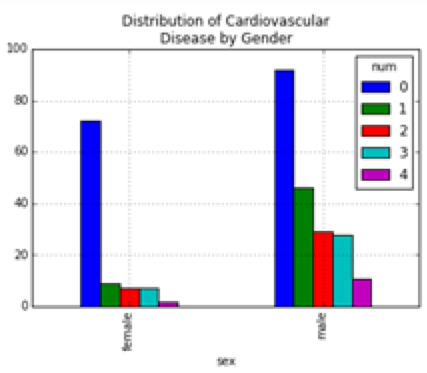

Histogram of Resting Blood Pressure
Count
0
10
20
30
40
50
60
70
80
80
100
120
140
160
180
200
Resting Blood Pressure (mm Hg. on admission to the hospital)

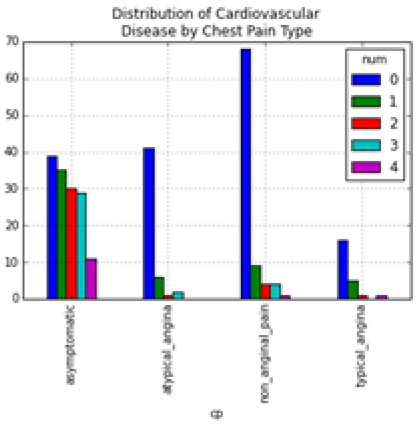

Histogram of Serum Cholesterol
Count
0
20
40
60
80
100
120
100
200
300
400
500
600
Serum Cholesterol (mg/dl)

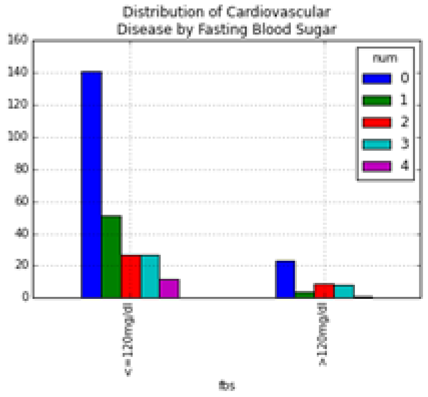

Histogram of Maximum Heart Rate Achieved
Count
0
10
20
30
40
50
60
70
80
60
80
100
120
140
160
180
200
220
Maximum Heart Rate Achieved

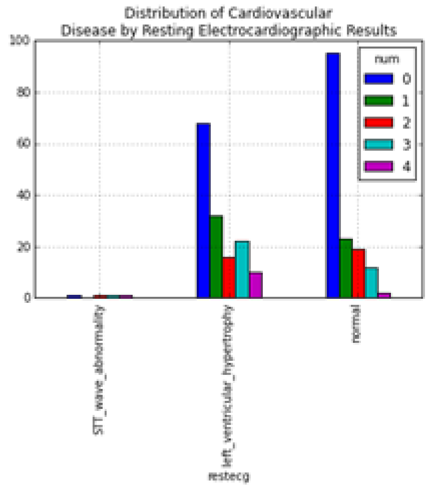

Histogram of ST Depression Induced by Exercise Relative to Rest
Count
0
20
40
60
80
100
120
140
160
0
1
2
3
4
5
6
7
ST Depression Induced by Exercise Relative to Rest

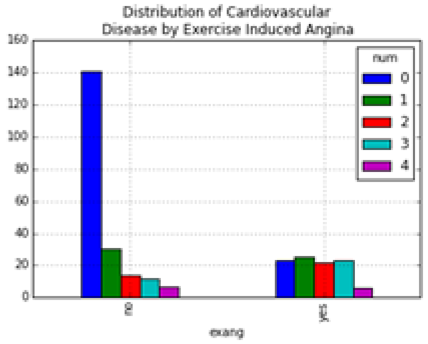

Histogram of CA
Count
0
20
40
60
80
100
120
140
160
180
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Number of major vessels (0-3) Colored by Flouroscopy

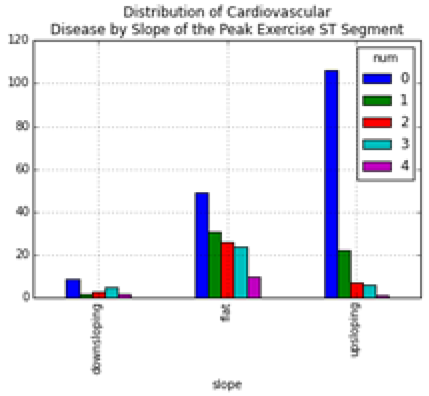

Distribution of Cardiovascular
Disease by Gender
100
80
60
40
20
0
num
0
1
2
3
4
female
male
sex

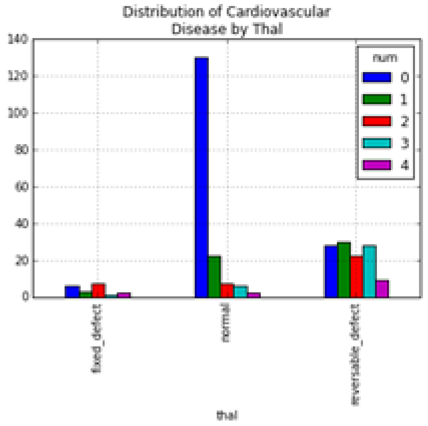

Distribution of Cardiovascular
Disease by Chest Pain Type
70
60
50
40
30
20
10
0
num
0
1
2
3
4
asymptomatic
atypical_angina
non_anginal_pain
typical_angina
cp

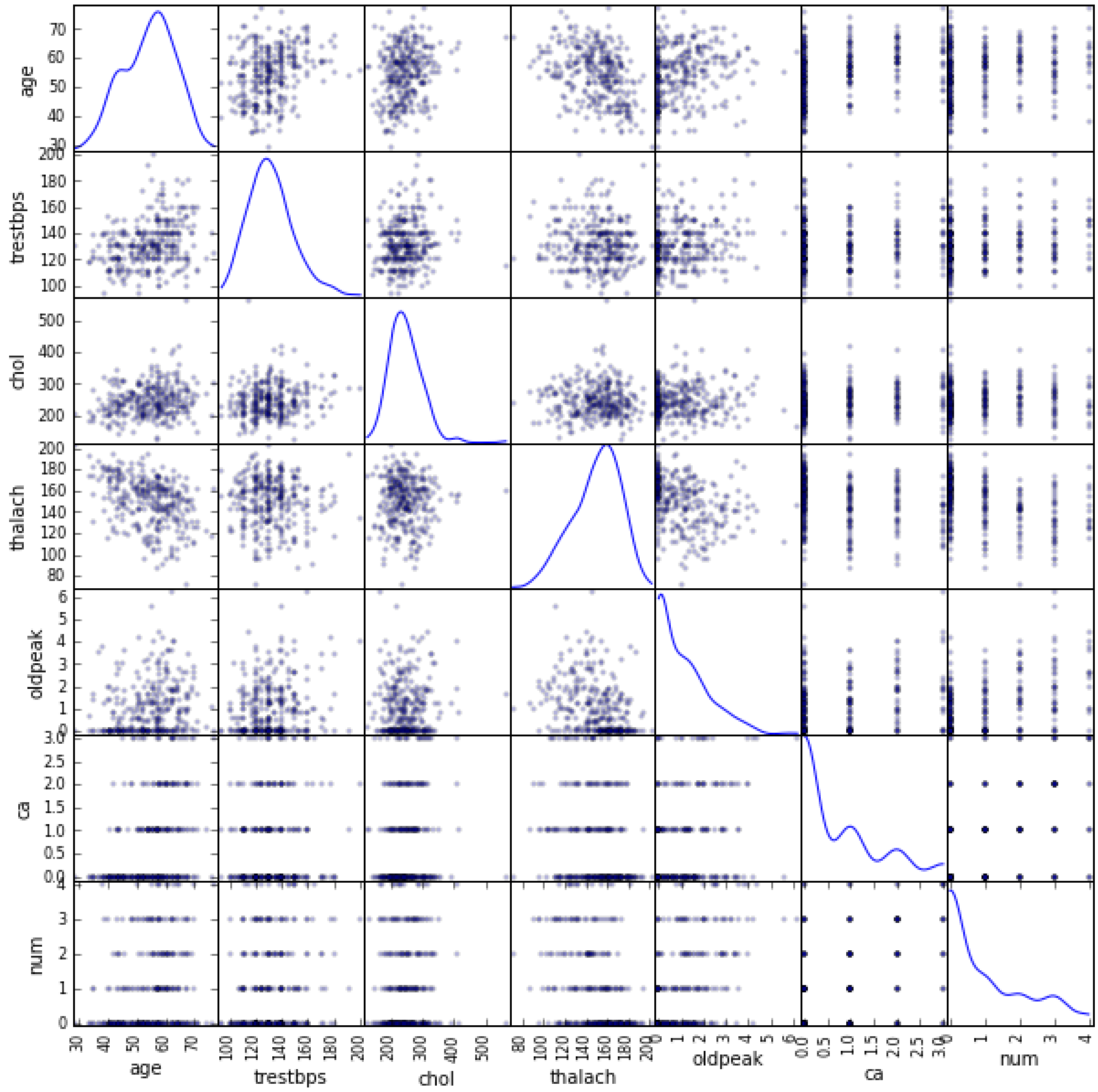

Distribution of Cardiovascular
Disease by Fasting Blood Sugar
160
140
120
100
80
60
40
20
0
num
0
1
2
3
4
<=120mg/dl
>120mg/dl
fbs

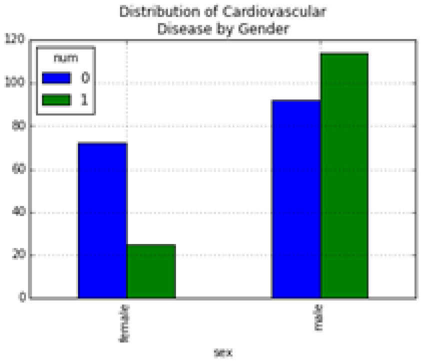

Distribution of Cardiovascular
Disease by Resting Electrocardiographic Results
100
80
60
40
20
0
num
0
1
2
3
4
STT_wave_abnormality
left_ventricular_hypertrophy
normal
restecg

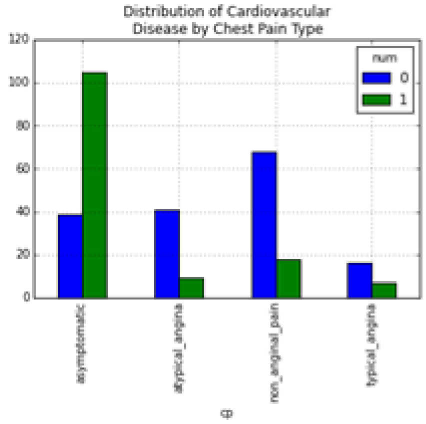

Distribution of Cardiovascular
Disease by Exercise Induced Angina
160
140
120
100
80
60
40
20
0
num
0
1
2
3
4
no
yes
exang

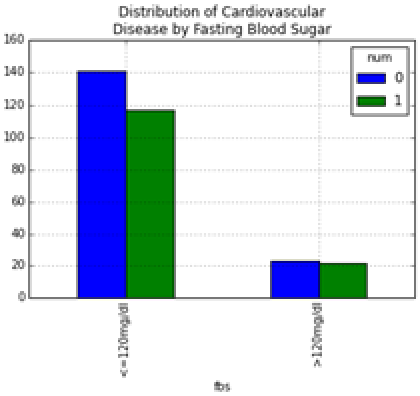

Distribution of Cardiovascular
Disease by Slope of the Peak Exercise ST Segment
120
100
80
60
40
20
0
num
0
1
2
3
4
downsloping
flat
upsloping
slope

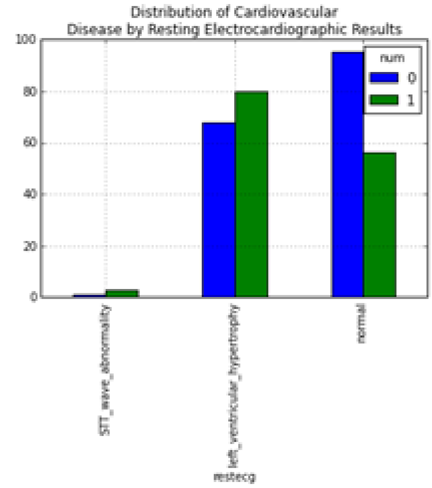

Distribution of Cardiovascular
Disease by Thal
140
120
100
80
60
40
20
0
num
0
1
2
3
4
fixed_defect
normal
reversable_defect
thal

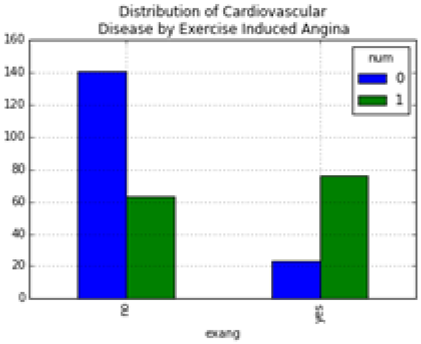

age
trestbps
chol
thalach
oldpeak
ca
num
30 40 50 60 70
100 120 140 160 180 200
200 300 400 500
80 100 120 140 160 180 200
0 1 2 3 4 5 6
0.0 0.5 1.0 1.5 2.0 2.5 3.0
0 1 2 3 4

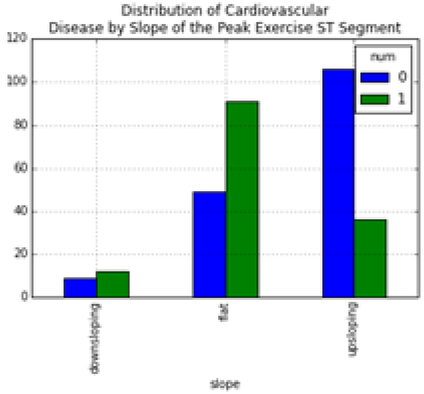

Distribution of Cardiovascular
Disease by Gender
num
0
1
120
100
80
60
40
20
0
female
male
sex

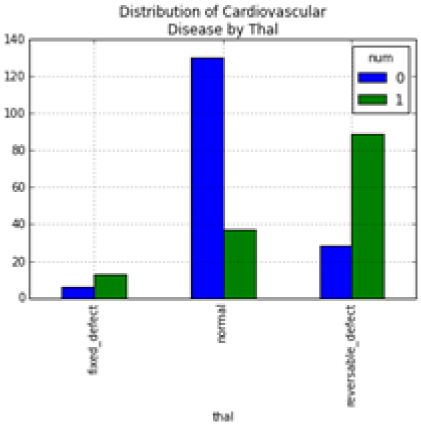

Distribution of Cardiovascular
Disease by Chest Pain Type
num
0
1
120
100
80
60
40
20
0
asymptomatic
atypical_angina
non_anginal_pain
typical_angina
cp

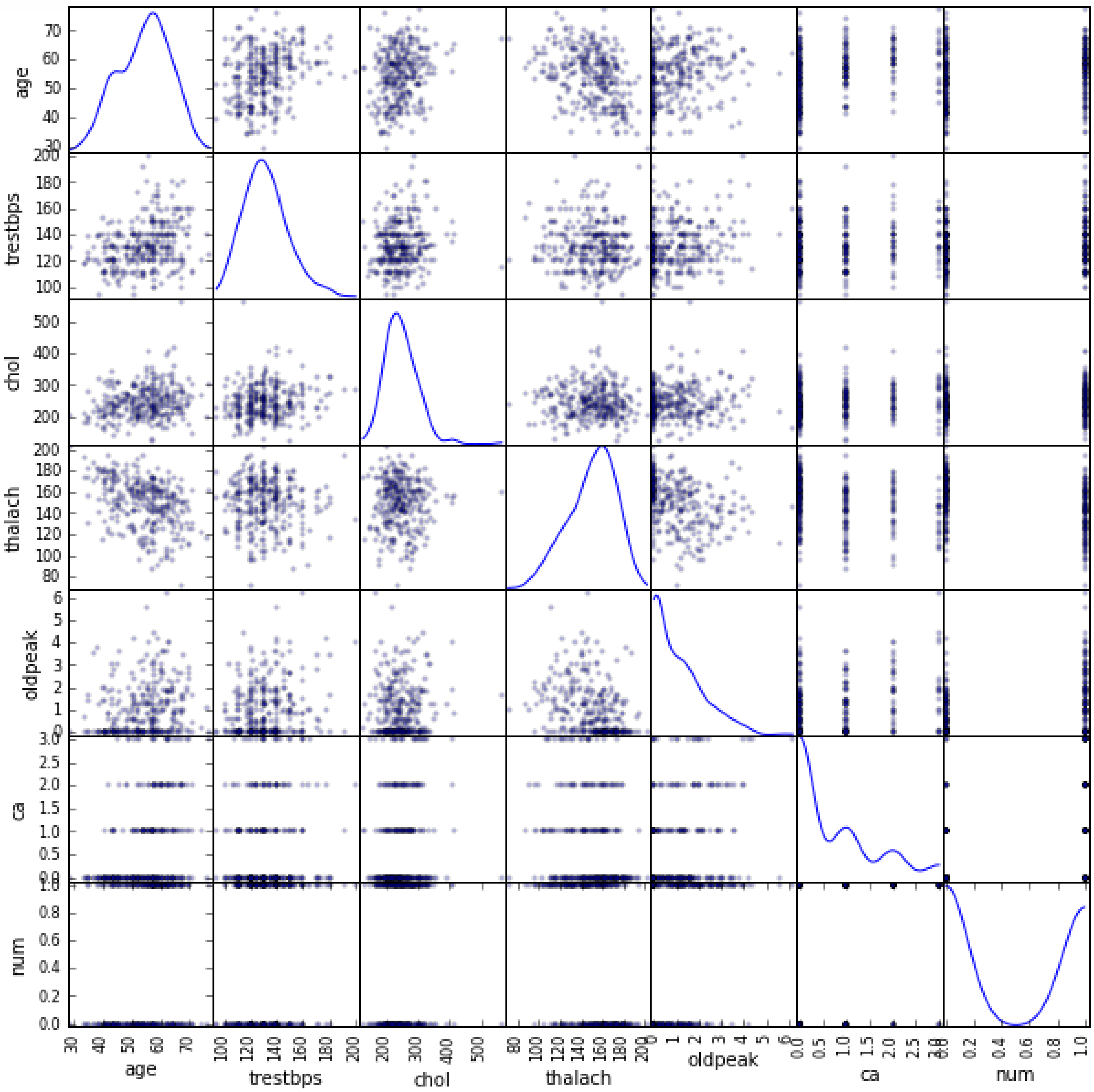

Distribution of Cardiovascular
Disease by Fasting Blood Sugar
num
0
1
160
140
120
100
80
60
40
20
0
<=120mg/dl
>120mg/dl
fbs

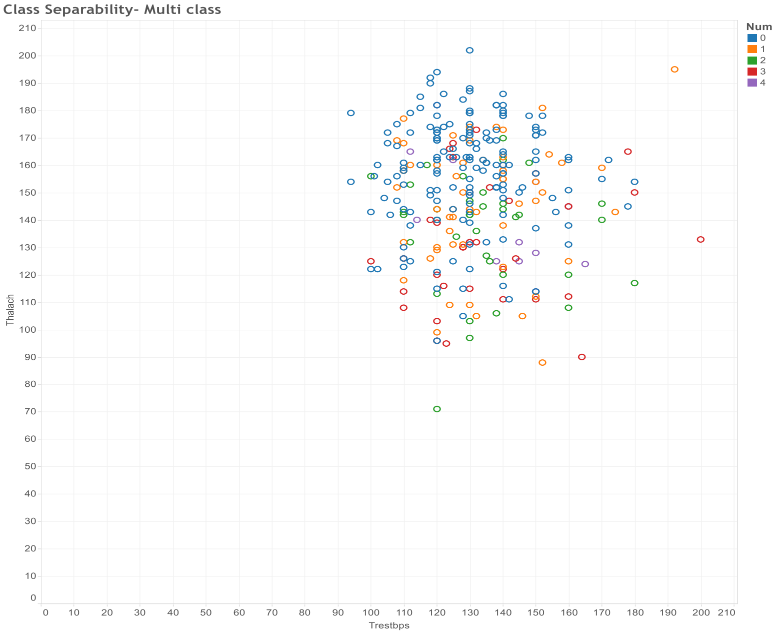

Distribution of Cardiovascular
Disease by Resting Electrocardiographic Results
num
0
1
100
80
60
40
20
0
STT_wave_abnormality
left_ventricular_hypertrophy
normal
restecg

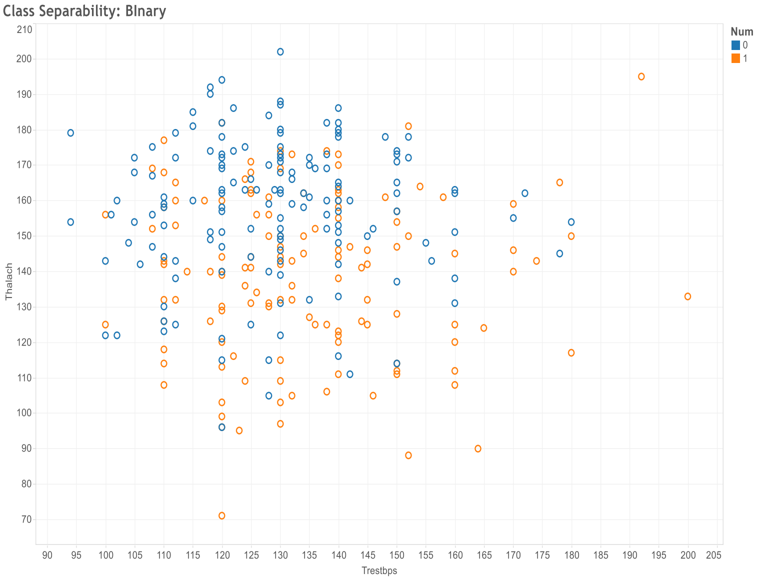

Distribution of Cardiovascular
Disease by Exercise Induced Angina
num
0
1
160
140
120
100
80
60
40
20
0
no
yes
exang


Distribution of Cardiovascular
Disease by Slope of the Peak Exercise ST Segment
num
0
1
120
100
80
60
40
20
0
downsloping
flat
upsloping
slope


Distribution of Cardiovascular
Disease by Thal
num
0
1
140
120
100
80
60
40
20
0
fixed_defect
normal
reversable_defect
thal


age
70
60
50
40
30
trestbps
200
180
160
140
120
100
chol
500
400
300
200
thalach
200
180
160
140
120
100
80
oldpeak
6
5
4
3
2
1
0
ca
3.0
2.5
2.0
1.5
1.0
0.5
0.0
num
1.0
0.8
0.6
0.4
0.2
0.0
30 40 50 60 70
age
100 120 140 160 180 200
trestbps
200 300 400 500
chol
80 100 120 140 160 180 200
thalach
0 1 2 3 4 5 6
oldpeak
0.0 0.5 1.0 1.5 2.0 2.5 3.0
ca
0.0 0.2 0.4 0.6 0.8 1.0
num


Class Separability- Multi class


Class Separability: BInary