Capstone 2 Milestone Report #2

TITLE:	Predicting Cardiovascular Heart Disease
AUTHOR:	Caitlin Jansson
DATASET :	Cardiovascular-study-dataset-predict-heart-disease
GIT REPOSITORY:	CJEJansson/Springboard_Projects/Capstone_2
SLIDES:	Milestone 2 Report Out - In Depth Analysis

PROBLEM STATEMENT

Cardiovascular heart disease (CHD) is the leading cause of death annually worldwide. Cardiovascular diseases can, however, be managed if caught early and simple lifestyle changes are made. This project explores a set of data for patients measuring known factors for heart disease to develop a machine learning model to predict risk of developing heart disease within the next ten years.

GENERAL OVERVIEW OF THE DATASET:

The data is provided from the Kaggle.com HME Workshop on Oct 3, 2020 targeted at increasing prediction of risk of developing CHD within the next ten years. This dataset is a subset of the Framingham, MA heart study data set. This data consists of a large group of initially "healthy" patients between the ages of 30-59 who were then tracked for 20 years to determine if they developed CHD [1]. The subset of data utilized in this project is divided into a test (80%) and train (20%) dataset and contains information on over 4,200 patients. The majority of the feature data has been converted to ordinal format (typically numeric), and there is no descriptive data included.

Overall the data was relatively clean, after a brief exploration. The dataset is imbalanced when looking at risk of developing CHD in the next ten years, with the ratio of Risk:No Risk being 511:2878, so only approximately 15% of patients in the dataset have a risk of developing Cardiovascular Heart disease.

DATA EXPLORATION AND STATISTICAL ANALYSIS:

Overall, the data showed results expected, and the results found here were consistent with those results found by the researchers doing the study. The risk of developing CHD is increased with smoking, and the more a patient smokes per day the higher the risk. High cholesterol and high blood pressure (looking at both systolic and diastolic) increase the risk of developing CHD. Increase of developing CHD increases with age. Patients with a prevalence of stroke, hypertension, and diabetes all also increase the risk of developing CHD with age.

Knowing from the full study that risk of CHD can be decreased, it would have been interesting to study this as part of the project, but this data was not included in the provided dataset for analysis. These insights can be leveraged by the client to target specific patient groups and ultimately lower the risk of developing CHD. It is hoped that these insights will lend themselves to the development of an accurate risk prediction engine as the project continues.

A complete summary of the figures generated and statistical findings from Exploratory Data Analysis can be found in Appendix B.

IN-DEPTH ANALYSIS:

DATA PREPARATION:

After completion of Exploratory Data Analysis (EDA) and Statistical Analysis, a more in-depth analysis was started by developing several Machine Learning Models. To prepare the data for application of Machine Learning models, the data had to be adjusted to fill null values in all features other than the target variable, in this case TenYearCHD. The combined test and train data was preprocessed by performing the following steps:

1. Both gender and smoking status were converted to ordinal values, consistent with the rest of the data set. Values were assigned to Male/Female as 0/1, and Non-smoker/Smoker as 0/1.
2. Null values for education were filled with a 5, to correspond to other values 1-4
3. Null values for BP meds were filled with a 2, to correspond to No-0, Yes-1
4. Null values in the number of cigarettes per day, total cholesterol, BMI, heart rate, and glucose (blood sugar) were filled with the median value calculated from non-nulls.
5. Split the data back into the original train/test datasets where "train" is populated using entries with non-null target values, and "test" contains null values for TenYearCHD.

The train data was used to conduct the entirety of the analysis for this project, due to the presence of values in the target variable. Analysis was conducted in two ways, first by splitting on the variable "is_smoking", since that was the most balanced variable in the dataset. Data was split into train/test, saving 20% of the data for testing. Several models were then run on this data. As a second analysis, the results of the LightGBM feature selection (Appendix B) were used to remove the 2 least significant features is_smoking and prevalent_stroke, and analyzing the data, again using an 80/20 split.

To determine accuracy of the models, 3 scoring systems were used.

1. **Accuracy Score**: used to measure all the correctly identified cases. This method was used because both classes are equally important, but also because the client specifically requested accuracy.
2. **F-1 score**: this metric gives a better measure of incorrectly classified cases than the accuracy score. Of the two, the F-1 score is very important in this case, as the false negative and false positive classifications are more important than the true positive/negative values. Unfortunately, because of the imbalance in the dataset, the number of false positives and negatives were very high, meaning all of the F1-scores were

close to zero. However, there is still some value in the results, particularly as the client is primarily concerned with incorrectly classified at-risk patients.

3. **Cross Validation:** utilized because of the unbalanced nature of the dataset. K-fold cross validation was utilized, with the number of folds set equal to 5. To satisfy the client's requirement, accuracy was used as the scoring technique. An average of the scores was taken across all folds, as was the standard deviation, to allow for complete representation of the results.

MODELS SELECTED FOR TESTING

An attempt was made to use imbalance's overfitting and underfitting to try and address the imbalance nature of the dataset. However, accuracy scores when using this method were approximately 68%, which was significantly lower than the other results of models tested. For this reason this method was not pursued further. Confusion matrices for these results can be seen below.

Regression Algorithms:

Logistic Regression was chosen because the dataset is a categorical classification problem. Despite the lack of strong linear relationship between features, it was decided that it was worth trying to apply this model. The solver method chosen was randomized and the maximum number of iterations was increased to ensure the model would converge. Overall, this model performed very well, with only one patient who was at risk classified as non-risk. The cross validation accuracy was 85.0% for all features, and 85.5% when using reduced features.

Instance Based Algorithms:

Given that the problem is looking for similarities between patients who show risk over time, it was decided to test some instance based algorithms.

A Support Vector Classifier was chosen, agan to pursue best fit, but also because the solver could be randomized rather than focusing on only a linear solver, as in LinearSVC. This method was chosen over SVM for the speed of the algorithm. This model performed best at classification of at-risk patients, but the accuracy was significantly lower overall for those patients not at risk. Cross validation scores for this model were 84.4% for all features and 85.1% for reduced features.

K-Nearest Neighbors was also chosen, again to focus on the patterns between instances and try to define whether there were clear boundaries easily. This model performed relatively well, but was prone to misclassifying at-risk patients. Cross-validation accuracy scores were 83.3% for all features and 83.6% for reduced features.

Decision Tree Algorithms:

Historically, decision tree algorithms tend to have good results on kaggle competitions. For this reason Random Forest, Decision Tree, and ExtraTrees were chosen.

Random forest was utilized two ways. As just a basic Random Forest implementation, and also by utilizing bagging, to try and account for the imbalance in the dataset. When compared, the version of random forest without using bagging performed better, but only marginally, with one less misclassified no-risk patient. Cross validation accuracy for random forest without bagging was 84.6% for all features, and 84.8% for reduced features. For random forest with bagging accuracy was 84.3% for all features and 84.9% for reduced features.

Decision tree performed significantly less well than random forest. Cross validation accuracy scores were 75.0% for all features and 75.1% for reduced features.

While Extra trees performed better than both Random Forest and Decision Tree, and also correctly classified the majority of at-risk patients, it did not perform as well as either of the others at correctly classifying no-risk patients. Cross-validation accuracy scores were 84.4% for all features and 85.1% for reduced features.

Dimensional Reduction Algorithms:

To try and exploit the inherent in the structure, rather than focusing on the imbalance, one dimensional reduction algorithm was implemented. Overall, this model performed very well. Models implemented included Linear Discriminant Analysis - which had one of the highest scores at 85.1%. Unfortunately, a large number of patients with risk were misclassified - 7 total, which moved the preference for this model down, as correctly classified risk patients were considered the most important indicator by the client.

Ensemble Algorithms:

To try and prevent overfitting, and given the imbalanced nature of the dataset, some ensemble algorithms were implemented to determine their efficacy. These models allow for a combination of weak learners to form a stronger model. Those models chosen included Gradient Boosting, AdaBoost and Light GBM.

AdaBoost did not perform well, with cross validation accuracy scores of 75.6% for all features and 75.0% for reduced features. It was one of the few models that performed worse when features were reduced.

Gradient boosting performed in the top 3 models for correct classification. It performed almost as well as the logistic regression and SVC models. Cross-validation accuracy scores were 84.2% for all features and 84.3% for reduced features.

LightGBM performed fairly well, and was useful for eliminating lowest significance features when building models. Accuracy scores were not in the bottom 3 scores, with cross validation accuracy of 83.2% for all features and 83.5% for reduced features.

Artificial Neural Network:

One algorithm from this class was applied, Multiple Layer Perceptron. This model was chosen because it is good with pattern matching in classification problems. This is another model that performed worse when reducing the features in the dataset. It was in the lower half of accuracy scores, with cross validation accuracy of 84.3% for all features and 83.9% for reduced features.

SUMMARY:

The data was studied using the same machine learning models but a different selection of features. Overall, the reduced feature data had higher deviations when looking at results of cross validation. However, when looking at accuracy, the reduced features models tended to have higher accuracies. This is possibly due to the model overfitting the data. The best model for the dataset without overfitting would require access to the solutions for the train data, which are not available at this time. Access has been requested with no response to date.

Logistic Regression was found to be the most accurate of models used when only considering within group results. That is to say when only looking at models analyzing all features or models analyzing reduced features. When considering both modeled datasets together, Gradient boosting or Random forest were the most accurate models. They had a cross validation accuracy of approximately 84% +/- 1%.

Given the clients expressed goals, and knowing the risk of failing to identify a risk of CHD, it was determined that that model accuracy should be prioritized by least number of patients classified as having no risk who actually have risk. Based on confusion matrix results for these four models (Fig. 1) the best model is to use Logistic Regression with the entire feature set. This results in the least number of patients with risk who are not alerted, which is most critical. Accuracy for this dataset is 85% when using the cross-validation accuracy score. Even though the reduced data yielded better results for accuracy when patients who do not have a risk are identified as having a risk, those patients at risk who were not identified were determined to be the most critical.

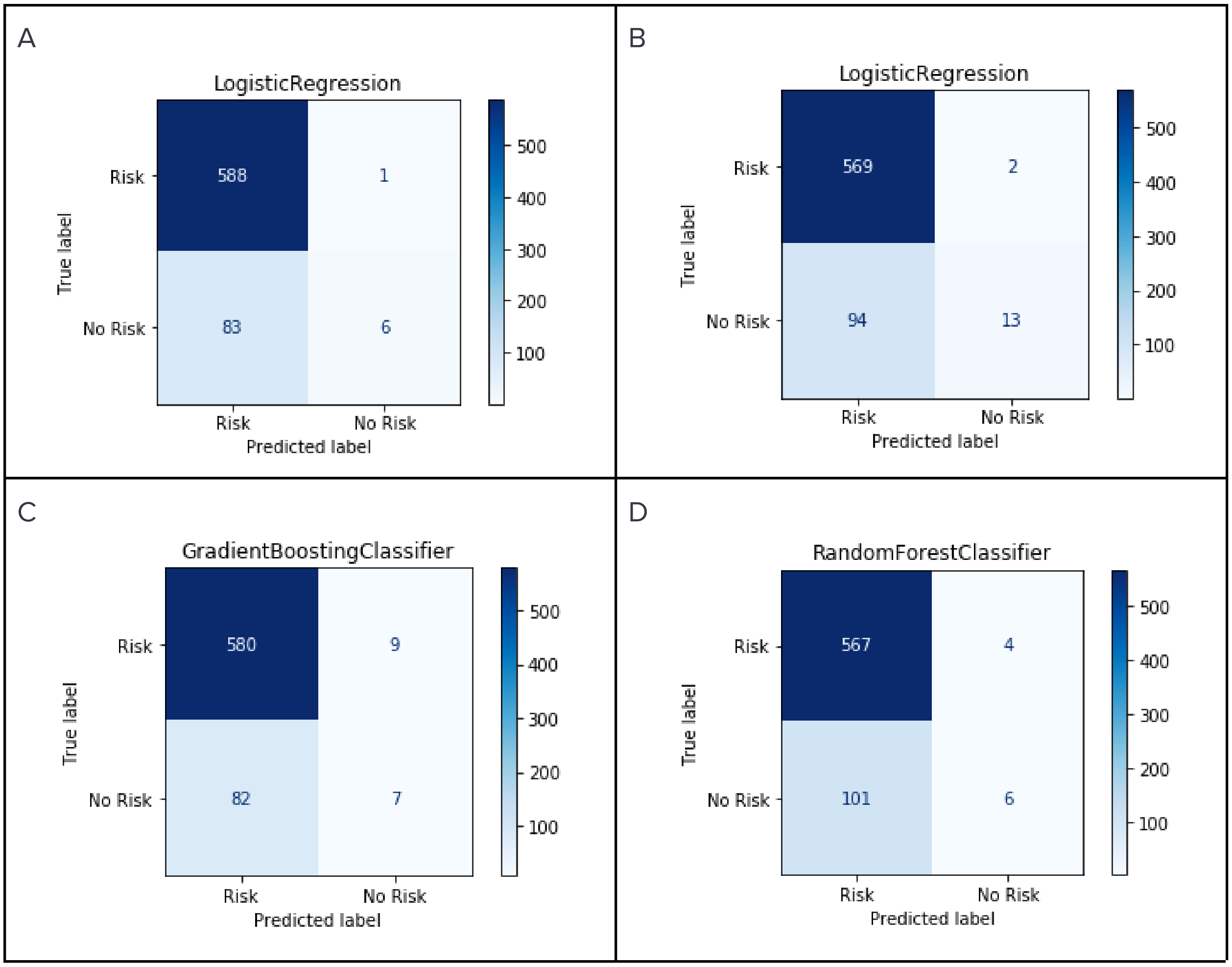
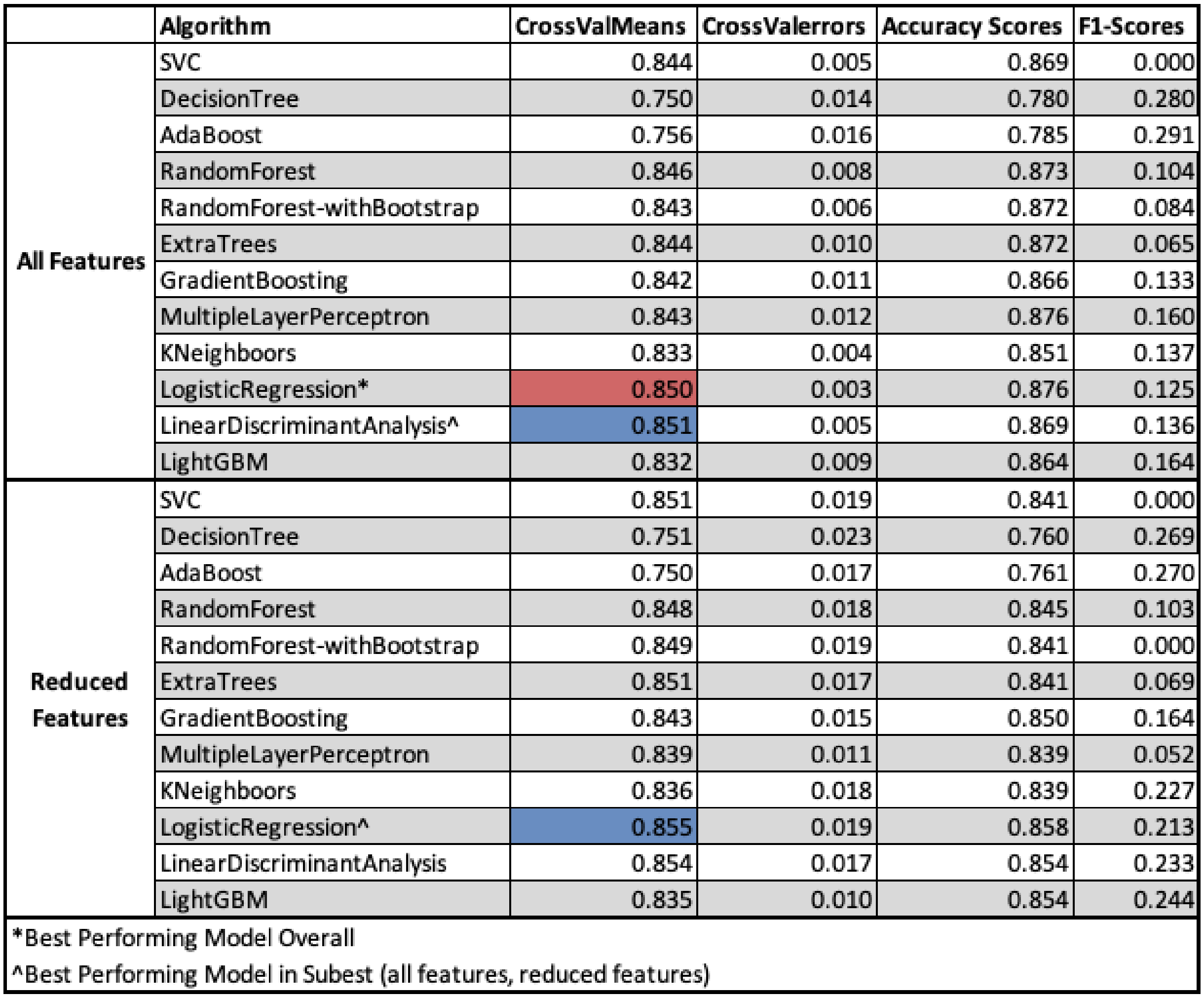


FIGURE 1: Confusion matrices for A) Logistic Regression- All Features, B) Logistic Regression- Reduced Features, C) Gradient Boosting All Features, D) Random Forest Reduced Features

SVC was not chosen, despite having zero at risk patients misidentified, because of the F1 score of zero, and the overall lower accuracy scores.

A complete summary of the model accuracy scores can be seen in Table 1 below. Additional details regarding accuracy scores can be seen in Appendix A.

TABLE 1: Summary of Accuracy Scores for All Models



	Algorithm	CrossValMeans	CrossValerrors	Accuracy Scores	F1-Scores
All Features	SVC	0.844	0.005	0.869	0.000
	DecisionTree	0.750	0.014	0.780	0.280
	AdaBoost	0.756	0.016	0.785	0.291
	RandomForest	0.846	0.008	0.873	0.104
	RandomForest-withBootstrap	0.843	0.006	0.872	0.084
	ExtraTrees	0.844	0.010	0.872	0.065
	GradientBoosting	0.842	0.011	0.866	0.133
	MultipleLayerPerceptron	0.843	0.012	0.876	0.160
	KNeighboors	0.833	0.004	0.851	0.137
	LogisticRegression*	0.850	0.003	0.876	0.125
	LinearDiscriminantAnalysis^	0.851	0.005	0.869	0.136
	LightGBM	0.832	0.009	0.864	0.164
Reduced Features	SVC	0.851	0.019	0.841	0.000
	DecisionTree	0.751	0.023	0.760	0.269
	AdaBoost	0.750	0.017	0.761	0.270
	RandomForest	0.848	0.018	0.845	0.103
	RandomForest-withBootstrap	0.849	0.019	0.841	0.000
	ExtraTrees	0.851	0.017	0.841	0.069
	GradientBoosting	0.843	0.015	0.850	0.164
	MultipleLayerPerceptron	0.839	0.011	0.839	0.052
	KNeighboors	0.836	0.018	0.839	0.227
	LogisticRegression^	0.855	0.019	0.858	0.213
	LinearDiscriminantAnalysis	0.854	0.017	0.854	0.233
	LightGBM	0.835	0.010	0.854	0.244

*Best Performing Model Overall

^Best Performing Model in Subest (all features, reduced features)

CITATIONS

[1] https://www.thelancet.com/journals/lancet/article/PIIS0140-6736%2813%2961752-3/fulltext

[2] https://web.archive.org/web/20170710152157/https://www.framinghamheartstudy.org/index.php

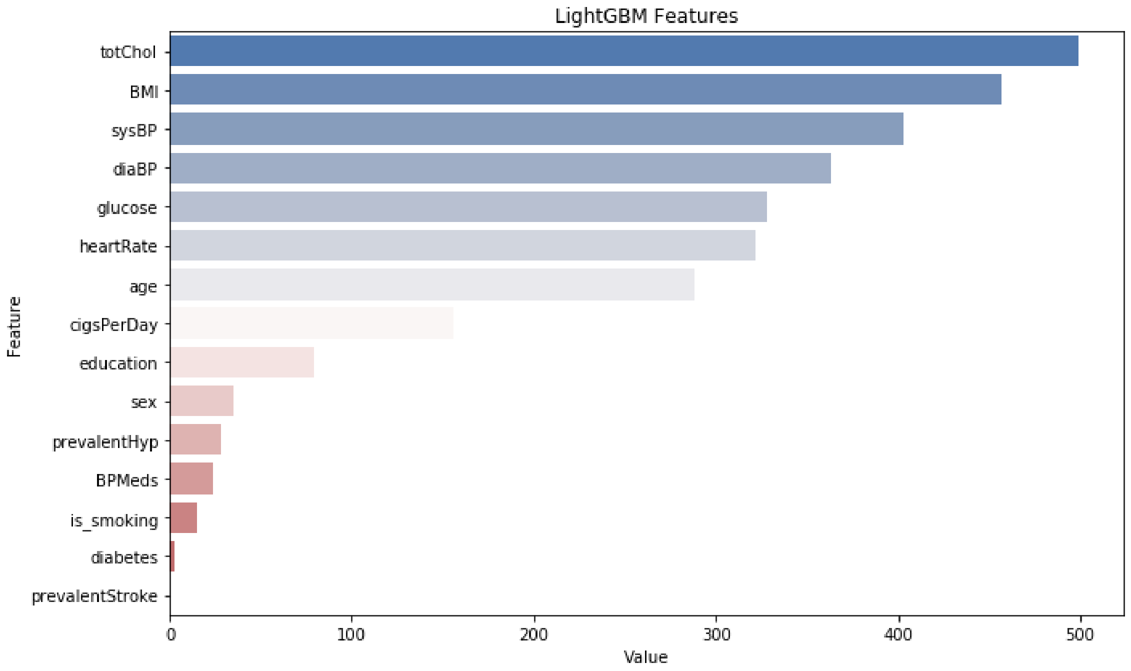
[3] https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

[4] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics

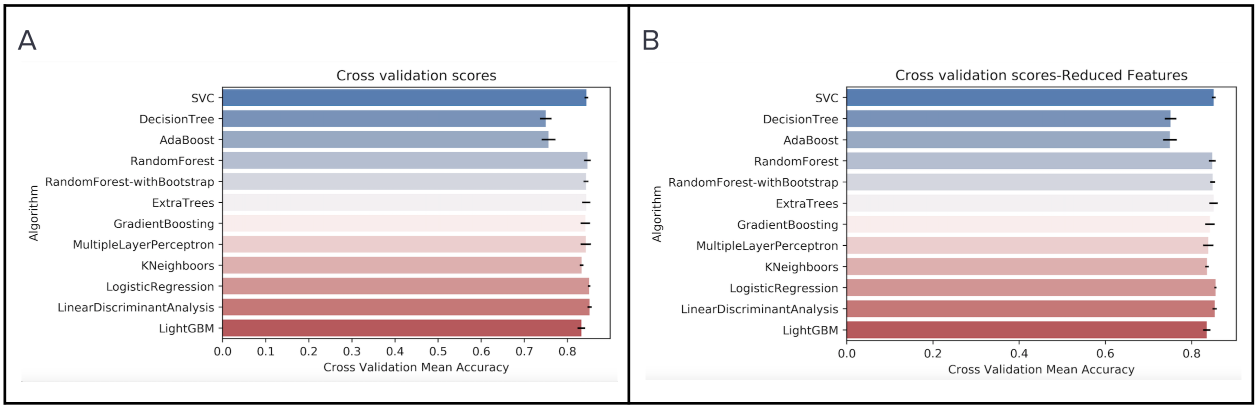
[5] https://towardsdatascience.com/cross-validation-430d9a5fee22

Appendix A

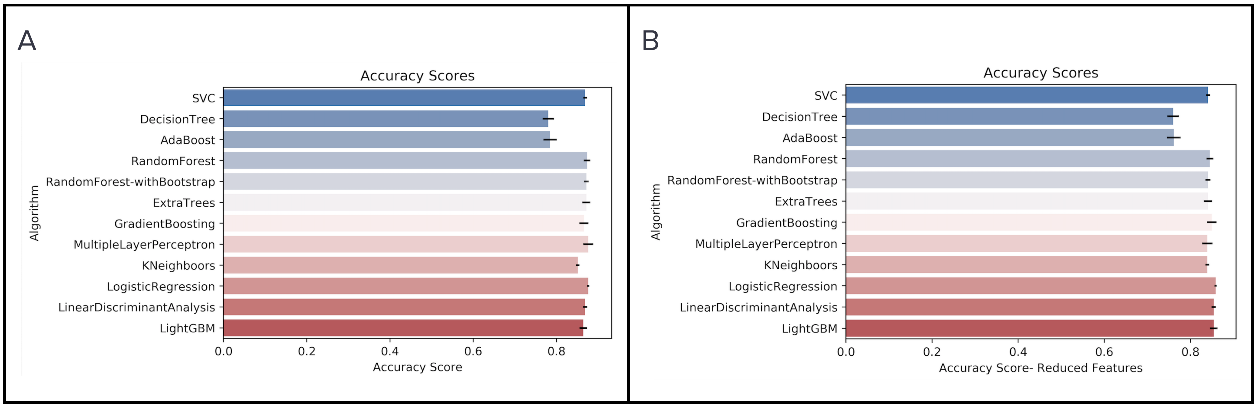
LIGHTGBM FEATURE SELECTION:



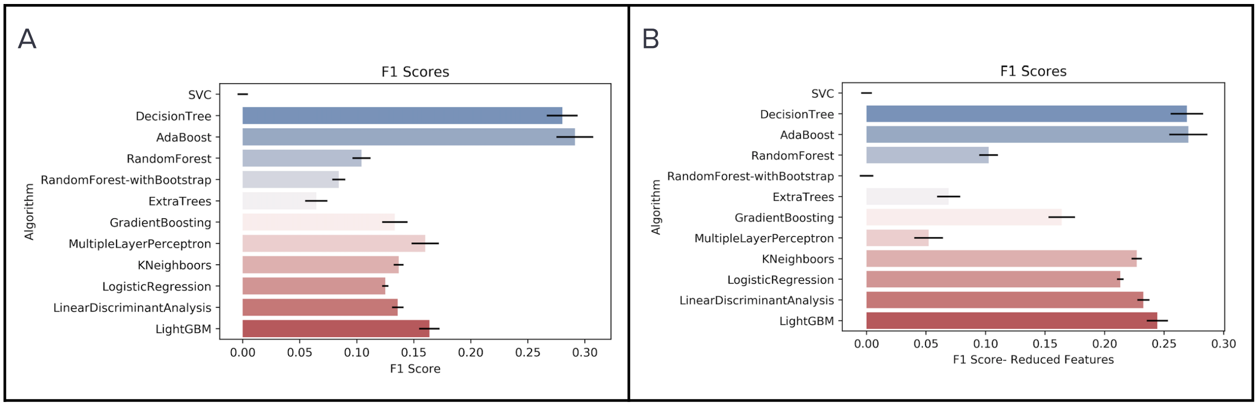
CROSS VALIDATION ACCURACY SCORES: a) all features, b) reduced features



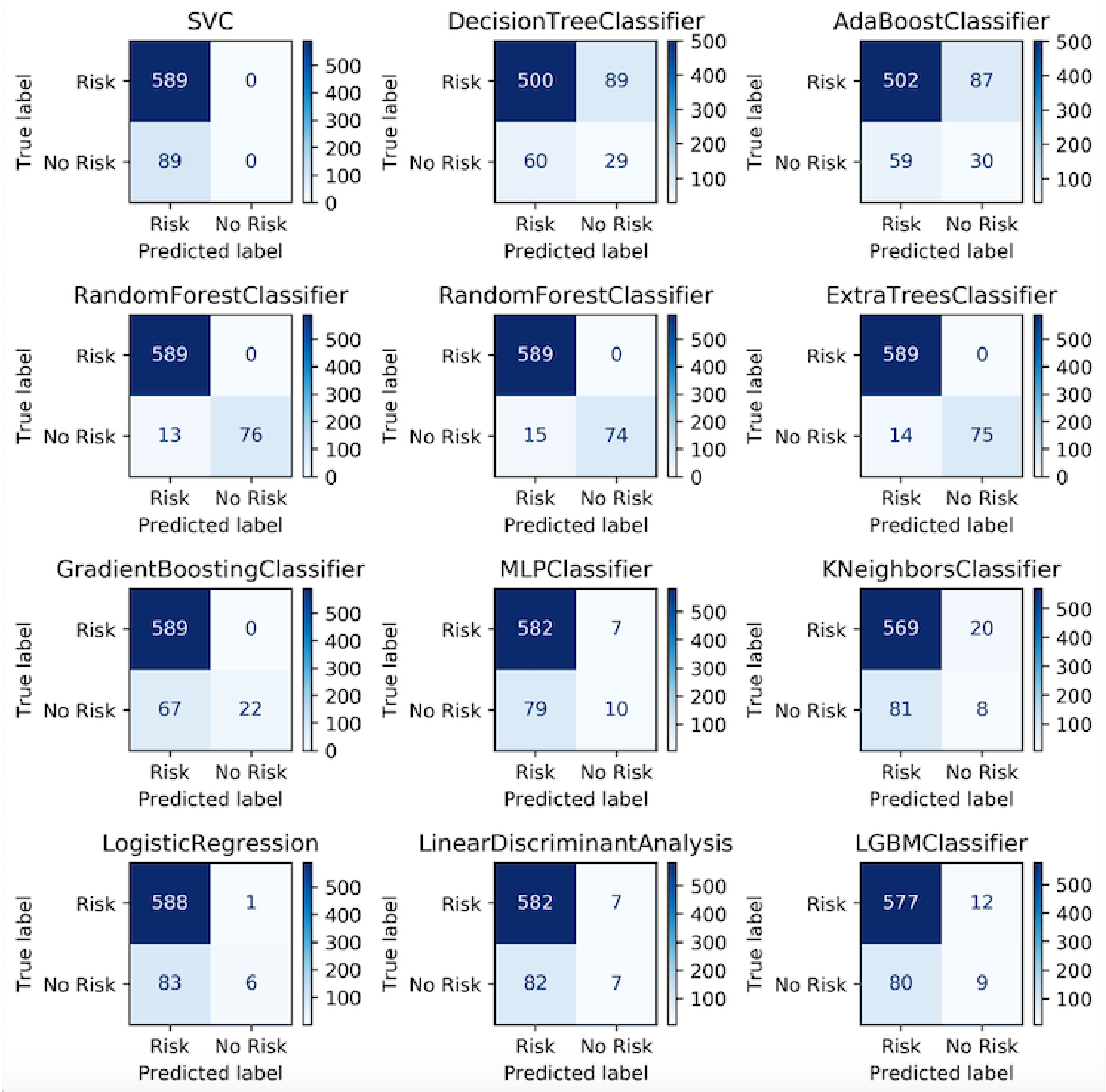
ACCURACY SCORES: a) all features, b) reduced features



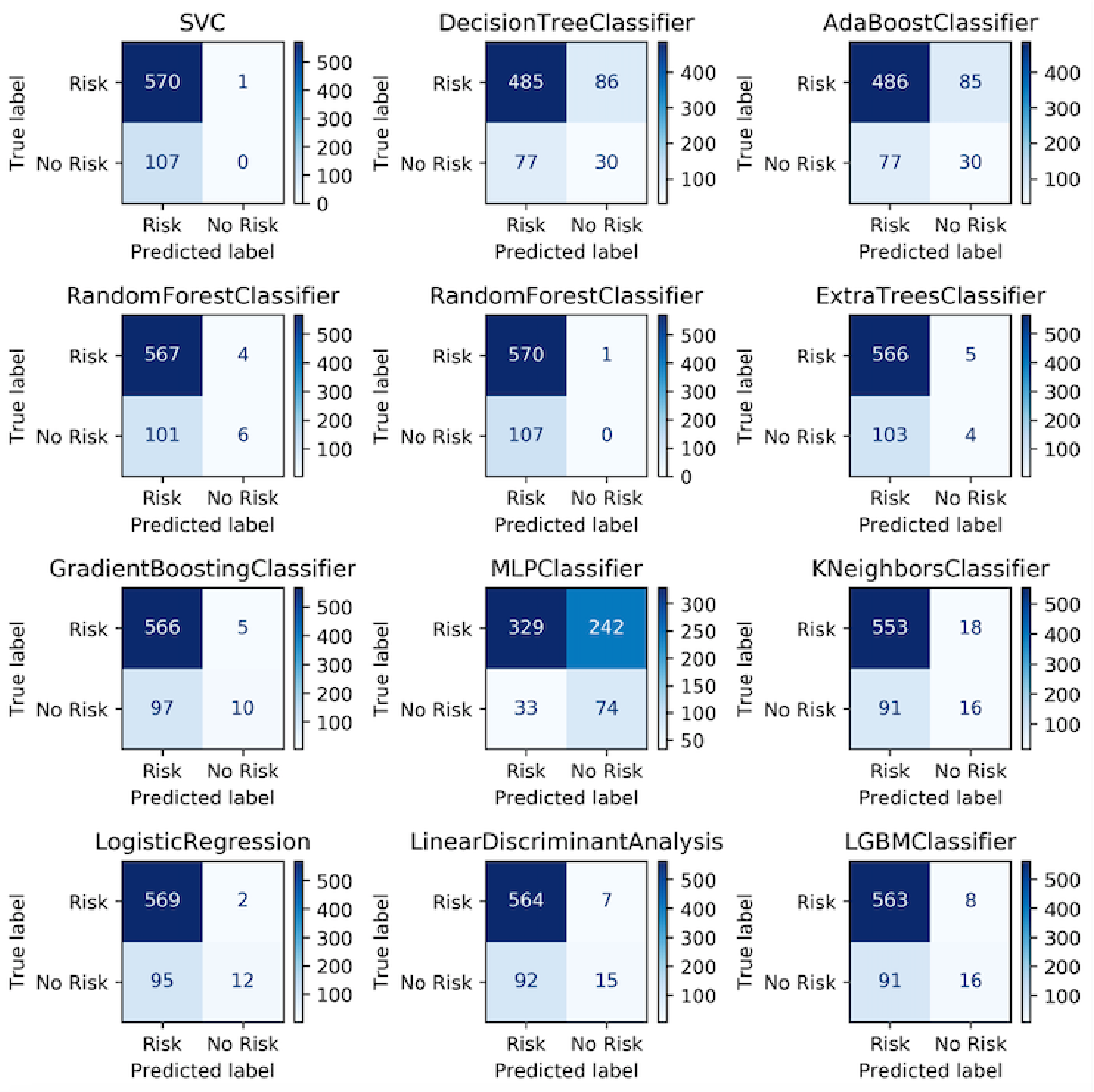
F1 SCORES: a) all features, b) reduced features



CONFUSION MATRICES - ALL MODELS, ALL FEATURES:



CONFUSION MATRICES - ALL MODELS, REDUCED FEATURES:



Appendix B

SUMMARY OF DATASET BY FEATURE:

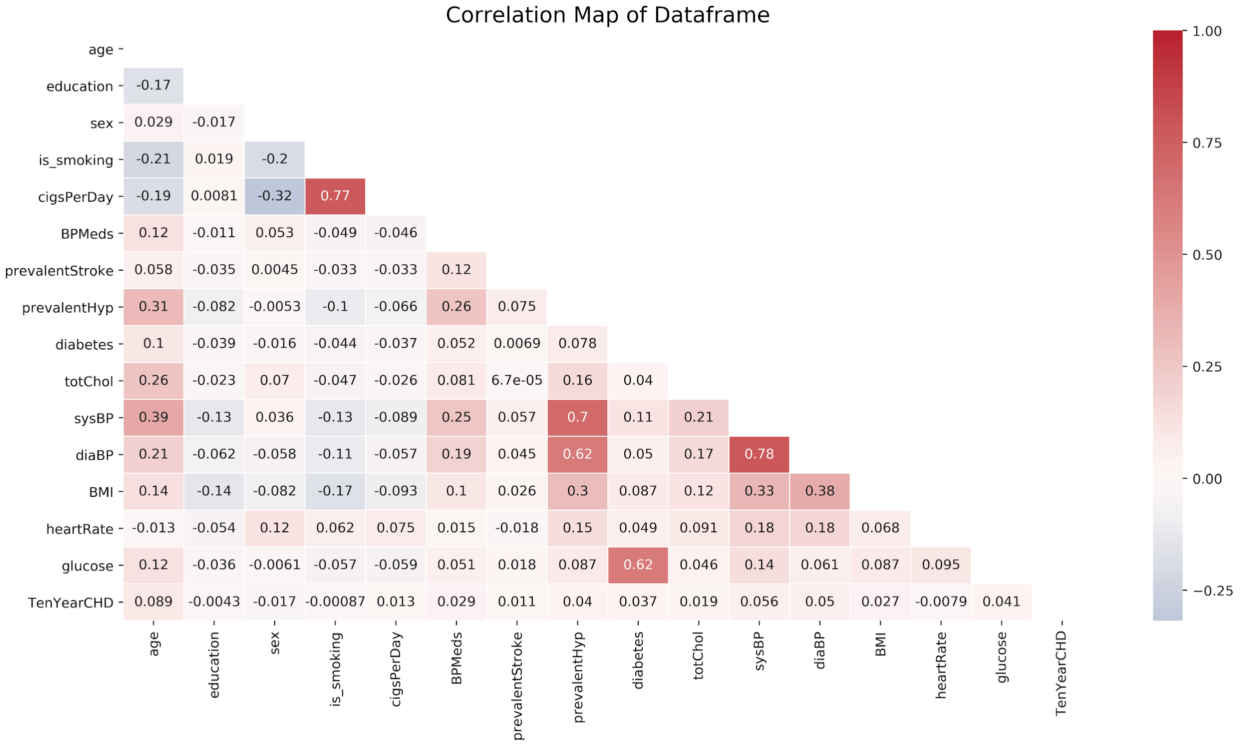
Feature	Values	Data Type	# Null	Description
Age	32-70	Continuous	None	Patient age in years, only whole numbers
Education	1, 2, 3, 4	Discrete	None	Education Level: 1-Some High School, 2-High School Diploma/GED, 3-College, 4-Degree
Sex	M, F	Discrete	None	Patient gender (M/F or 0=M, 1=F)
is_smoking	Yes, No	Discrete	None	If the patient is a current smoker (Yes/No or 1=yes, 0=no)
Cigs per Day	0-70	Continuous	29	Number of Cigarettes smoked per day (null = unknown)
BP Meds	0, 1	Discrete	53	Whether the patient is taking Blood Pressure Medications (0=no, 1=yes, null=unknown
prevalentStroke	0, 1	Discrete	None	Prevalence of stroke (0=none, 1=has had occurences of stroke)
prevalentHyp	0, 1	Discrete	None	Prevalence of hypertension (0=none, 1= has prevalence hypertension
diabetes	0, 1	Discrete	None	If the patient has diabetes (0=no, 1=yes)
totChol	107-696	Continuous	50	Total Cholesterol
sysBP	83.5-295	Continuous	None	Systolic Blood Pressure
diaBP	48-142.5	Continuous	None	Diastolic Blood Pressure
BMI	15-54-56.8	Continuous	19	Body Mass Index
heartRate	44-143	Continuous	1	Resting heart rate in beats per minute (bpm)
glucose	40-394	Continuous	388	Blood glucose level. (mg/dL)
TenYearCHD	0,1	Discrete (calculated)	848 (test data)	Risk of developing CHD in next decade (0=no risk, 1=risk)

STATISTICAL RESULTS OF FEATURE SPECIFIC INVESTIGATION:

*Correlation value reported vs risk of developing CHD in the next 10 years

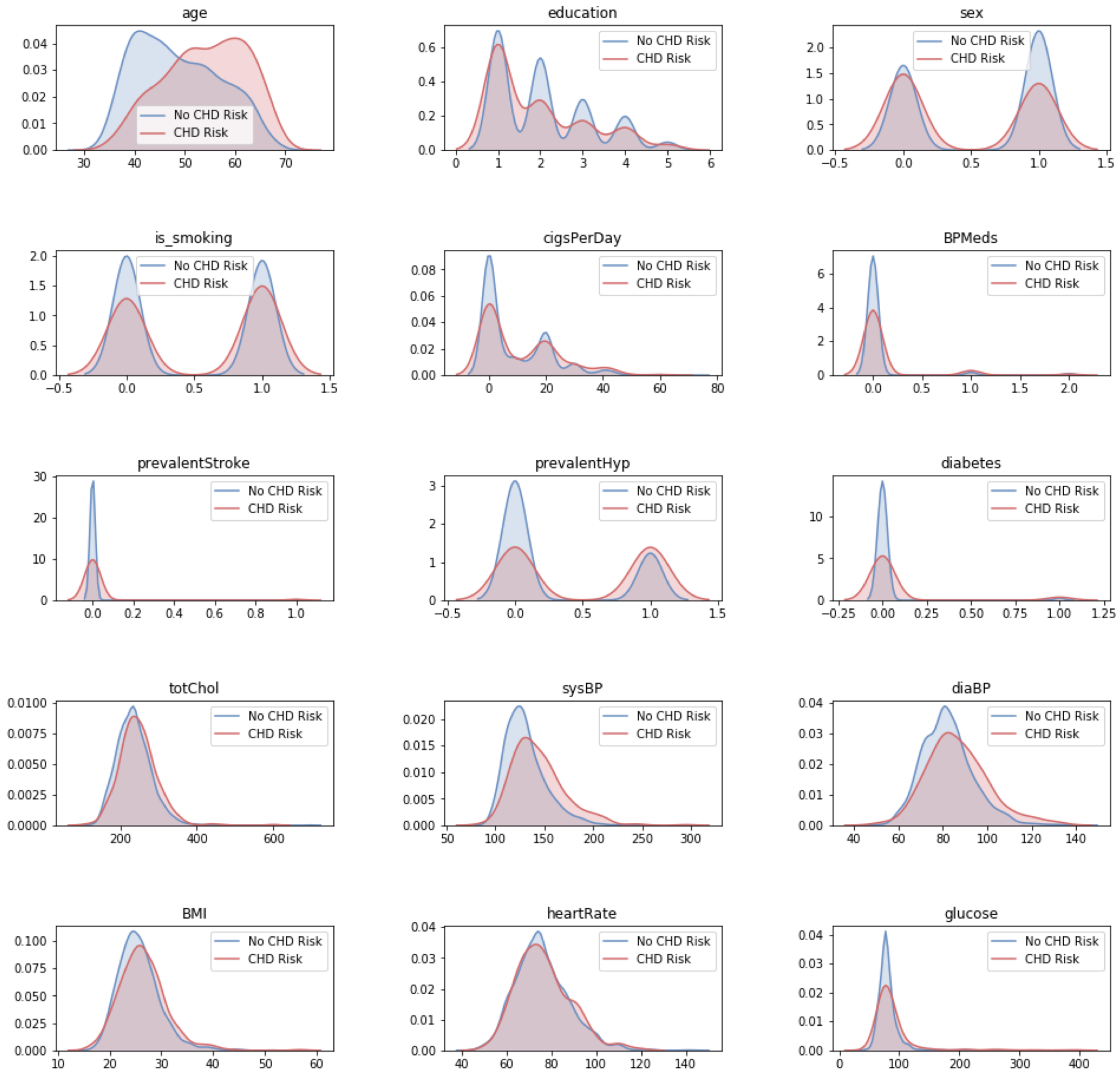
Feature	Correlation	Hypothesis Testing		
		Test	*Result*	*P-Value*
Age	0.0890	T-Test	Reject Ho - Correlated	1.85E-38
Education Level	-0.0043	Chi-Square	Reject Ho - Correlated	4.87E-04
Gender	-0.0170	Chi-Square	Reject Ho - Correlated	3.37E-06
Smoker	-0.0009	Chi-Square	Fail to Reject	9.07E-02
Number of Cigarettes per Day	0.0130	T-Test	Reject Ho - Correlated	5.37E-04
Use of BP Meds	0.0290	Chi-Square	Reject Ho - Correlated	2.07E-06
Prevalence of Stroke	0.0110	Chi-Square	Reject Ho - Correlated	9.32E-05
Prevalence of Hypertension	0.0400	Chi-Square	Reject Ho - Correlated	1.03E-21
Diabetic	0.0370	Chi-Square	Reject Ho - Correlated	1.27E-08
Total Cholesterol	0.0190	T-Test	Reject Ho - Correlated	5.31E-07
Systolic BP	0.0560	T-Test	Reject Ho - Correlated	5.52E-24
Diastolic BP	0.0500	T-Test	Reject Ho - Correlated	8.90E-12
BMI	0.0270	T-Test	Reject Ho - Correlated	4.45E-04
Heart Rate	-0.0079	T-Test	Fail to Reject	2.47E-01
Blood Sugar	0.0410	T-Test	Reject Ho - Correlated	3.41E-06

Correlation Heatmap showing correlation between features:



	age	education	sex	is_smoking	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	totChol	sysBP	diaBP	BMI	heartRate	glucose	TenYearCHD
age																
education	-0.17															
sex	0.029	-0.017														
is_smoking	-0.21	0.019	-0.2													
cigsPerDay	-0.19	0.0081	-0.32	0.77												
BPMeds	0.12	-0.011	0.053	-0.049	-0.046											
prevalentStroke	0.058	-0.035	0.0045	-0.033	-0.033	0.12										
prevalentHyp	0.31	-0.082	-0.0053	-0.1	-0.066	0.26	0.075									
diabetes	0.1	-0.039	-0.016	-0.044	-0.037	0.052	0.0069	0.078								
totChol	0.26	-0.023	0.07	-0.047	-0.026	0.081	6.7e-05	0.16	0.04							
sysBP	0.39	-0.13	0.036	-0.13	-0.089	0.25	0.057	0.7	0.11	0.21						
diaBP	0.21	-0.062	-0.058	-0.11	-0.057	0.19	0.045	0.62	0.05	0.17	0.78					
BMI	0.14	-0.14	-0.082	-0.17	-0.093	0.1	0.026	0.3	0.087	0.12	0.33	0.38				
heartRate	-0.013	-0.054	0.12	0.062	0.075	0.015	-0.018	0.15	0.049	0.091	0.18	0.18	0.068			
glucose	0.12	-0.036	-0.0061	-0.057	-0.059	0.051	0.018	0.087	0.62	0.046	0.14	0.061	0.087	0.095		
TenYearCHD	0.089	-0.0043	-0.017	-0.00087	0.013	0.029	0.011	0.04	0.037	0.019	0.056	0.05	0.027	-0.0079	0.041	

Kernel Density Estimation Distribution For Each Feature vs TenYearCHD showing correlation between features:


age
No CHD Risk
CHD Risk
education
sex
is_smoking
cigsPerDay
BPMeds
prevalentStroke
prevalentHyp
diabetes
totChol
sysBP
diaBP
BMI
heartRate
glucose