Predicting Cardiovascular Heart Disease

AUTHOR: Caitlin Jansson
DATE: 8 APRIL, 2021
DATASET : https://www.kaggle.com/christofel04/cardiovascular-study-dataset-predict-heart-disea
GIT REPOSITORY: https://github.com/CJEJansson/Springboard_Projects/Predicting CHD
SLIDES: View ⦿ Download

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 ratio of male to female patients is unbalanced, with the dataset being approximately 43% male. The ratio of smokers to non-smokers is also unbalanced, with the dataset being 50.5% nonsmokers. When reviewing the feature for the number of cigarettes smoked per day, it was found that a null value was reported if a smoker did not divulge the number of cigarettes each day. All non-smokers were assigned zero values. Of those patients that did smoke, the majority reported smoking one pack a day (assuming 20 cigarettes per pack).

During exploratory analysis it was determined that the ages of patients do in fact range from 32-70, which is consistent with 20 years of observation for patients starting at approximately age 30. Patients ranged in education level from some high school to a higher education degree, while some did not report their education level. 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.

To simplify later analysis, 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. This will allow for analysis as though the data is continuous, rather than discrete. Converting all continuous data to discrete values (eg. blood pressure to Low, Normal, High, etc) was considered, but it was decided to take the more expedient route. The use of actual continuous values for health measurement data will also allow for potentially more accurate models during the model development phase.

A summary of observations for minimum and maximum values can be seen in Appendix A, in the table "Summary of Dataset by Feature".

INITIAL FINDINGS:

After cleaning, the dataset was explored via Exploratory Data Analysis (EDA) and hypothesis testing was used by leveraging inferential statistics. This analysis was started by using a series of

As expected, demographic data was not of much interest, with the exception of age. Age had the highest correlation with risk of developing Cardiovascular Heart Disease in the next ten years. Other significant factors were blood pressure, body weight, and other preexisting health conditions. This was expected as it is consistent with the results of the Framingham Heart Study's researcher identified milestones [2]. Those milestones were reviewed after completing data exploration and statistical analysis, and were consistent with the findings of this project.

Three methods were utilized to study the correlation between each feature and the risk of developing CHD in the next decade. First a correlation barplot was created to quickly visualize the magnitude of each relationship, seen below in Figure 1. Then a heatmap was created and modified to return only the bottom half of the results - to eliminate duplication. It is not necessary to report the entire heatmap as these variables are symmetric. If the dataset were asymmetric (the value of x is known but y cant be determined, and if the value of y is known the x is guaranteed), it would be necessary to return the entire heat map. This map can be seen in Appendix A. The full heat map can be seen in the attached code.

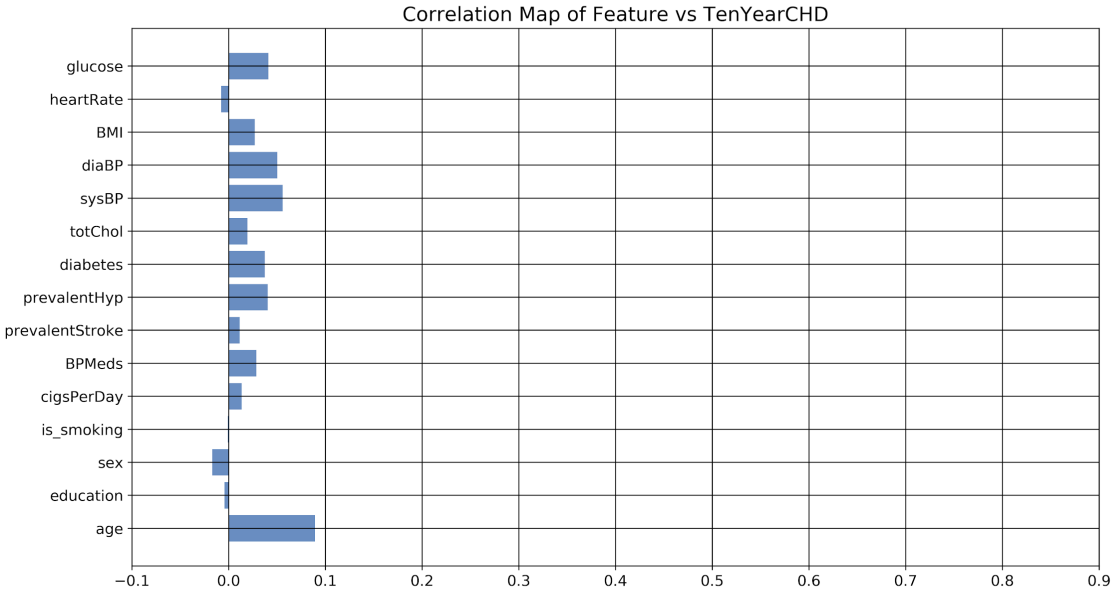


FIGURE 1: Barplot of correlation between each feature and the risk of developing CHD in the next ten years.

Finally, the correlation was visualized by plotting distribution plots to show the influence of each feature on the predicted value (risk of developing cardiovascular heart disease in the next 10 years). To do this, each feature will be plotted using a Kernel Density Estimation (KDE). The KDE shows the non-parametric probability density function of each data feature. This plot is created for each scenario - one for cases where CHD is observed, and a second for cases in which CHD

is not observed. The amount of overlap of each density function is then compared to determine correlation. If there is a high overlap of both density functions for the feature, it is implied that there is little to no correlation between the dependent(CHD) and independent(each feature) variables. The less overlap between plots, the more significant the correlations.

For those variables that are discrete, each category will have 2 plots. For example, when looking at the KDE for gender, there are 4 curves. To determine correlation one must compare the CHD curves for male, and then consider the CHD curves for female patients. The comparison for gender vs heart disease must be visualized using a different method. The completed KDE can be seen in Appendix A for all features in the dataset. This showed consistent results with the other methods.

Those variables found to have a significant correlation were unsurprisingly, age, previous medical conditions (stroke, hypertension- and treatment of hypertension with blood pressure medication, diabetes), number of cigarettes smoked per day, and body mass (BMI). Of note, it was interesting that there was stronger correlation between the number of cigarettes smoked per day and the risk of developing CHD than when only studying whether or not a patient smoked. As expected, there is an increased risk of CHD with smoking, and this risk increases with the number of cigarettes smoked per day, as seen in Figure 2.

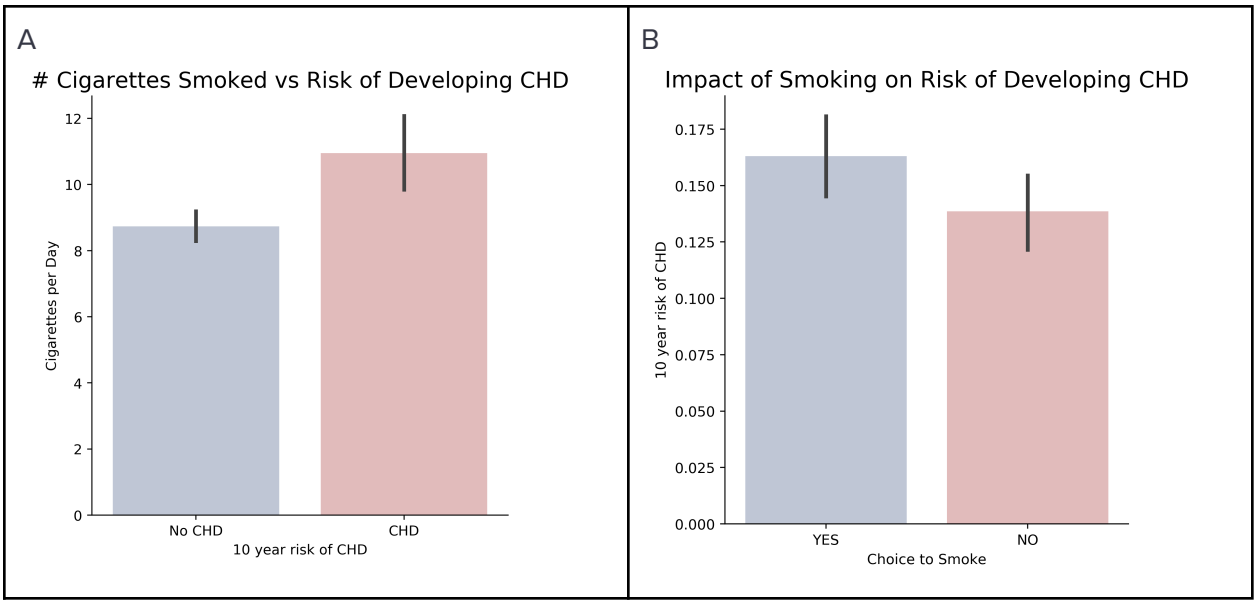
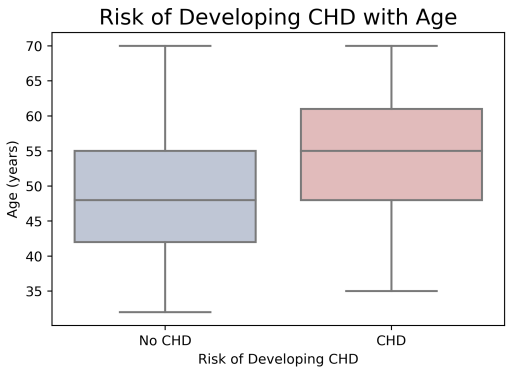


FIGURE 2 A) Risk of developing CHD as compared to # of cigarettes smoked per day; B) Risk of developing CHD with smoking

The risk of developing CHD was then investigated as patients age, since this was the most significant feature shown in the correlation plot. There was a distinct relationship between age and risk of developing CHD, showing that, on average, as patients age the risk increases. This relationship can be seen in the figure, to the right. This relationship was further explored to study the effect of age on prevalence of stroke, hypertension, and diabetes, all of which increased with age. Also of note, risk of developing CHD increased with prevalence of



diabetes, with prevalence of stroke, and with prevalence of hypertension. These plots can be seen in Appendix A.

As blood pressure is known to have an effect on heart health and risk of stroke as well as other health problems, the effect of blood pressure was also investigated with regards to risk of CHD. It was found, as expected, that on average the risk of developing CHD increases with increased blood pressure. This can be seen in Figure 3 below:

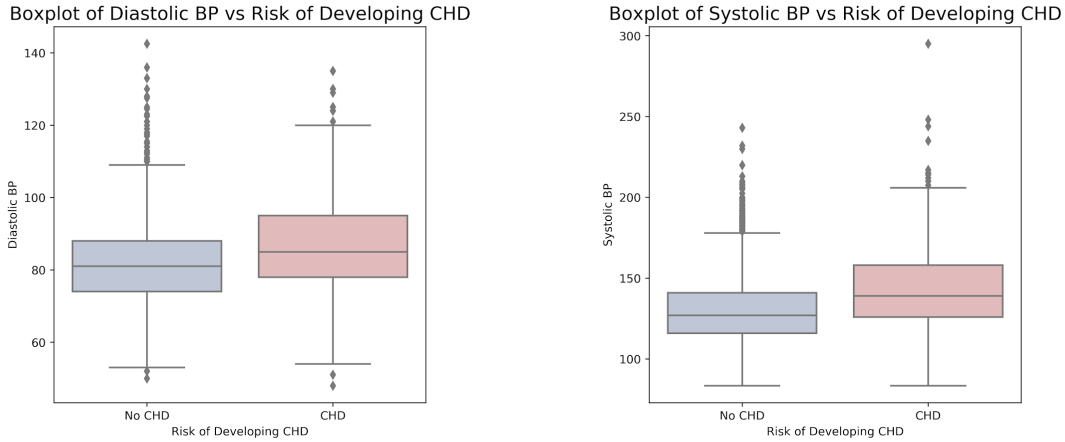


FIGURE 3 Effect of blood pressure, both systolic and diastolic vs risk of developing cardiovascular heart disease.

Gender was investigated, and it was found that, on average, males have a higher risk of developing heart disease, and the difference in this risk increases with age. Body weight also has an effect on risk of developing CHD, with risk increasing as BMi increases. These results can both be seen below in Figures 4, below, and Figure 5 on the next page.

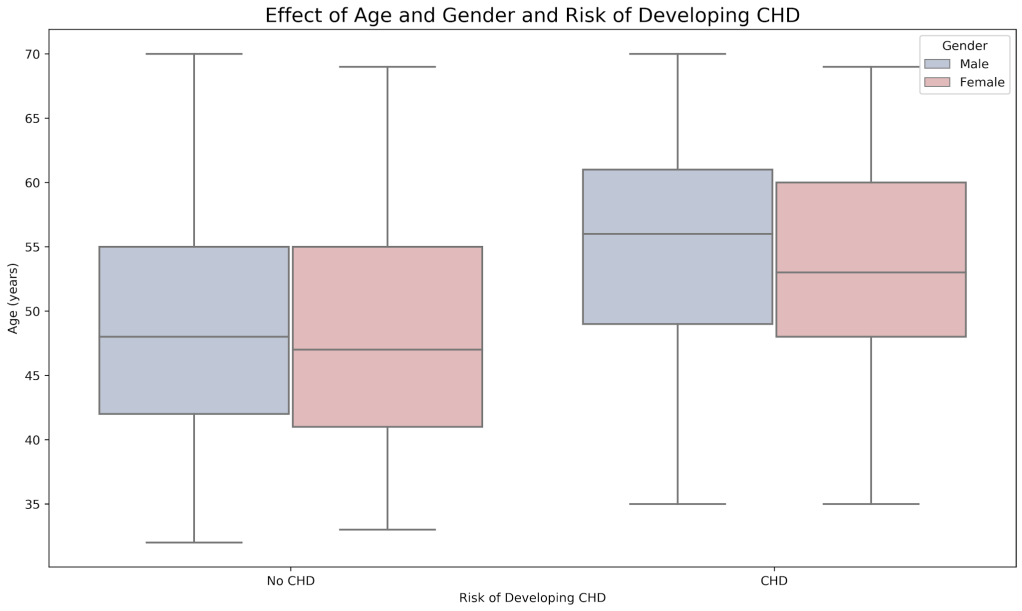


FIGURE 4 Effect of blood age and gender vs risk of developing cardiovascular heart disease.

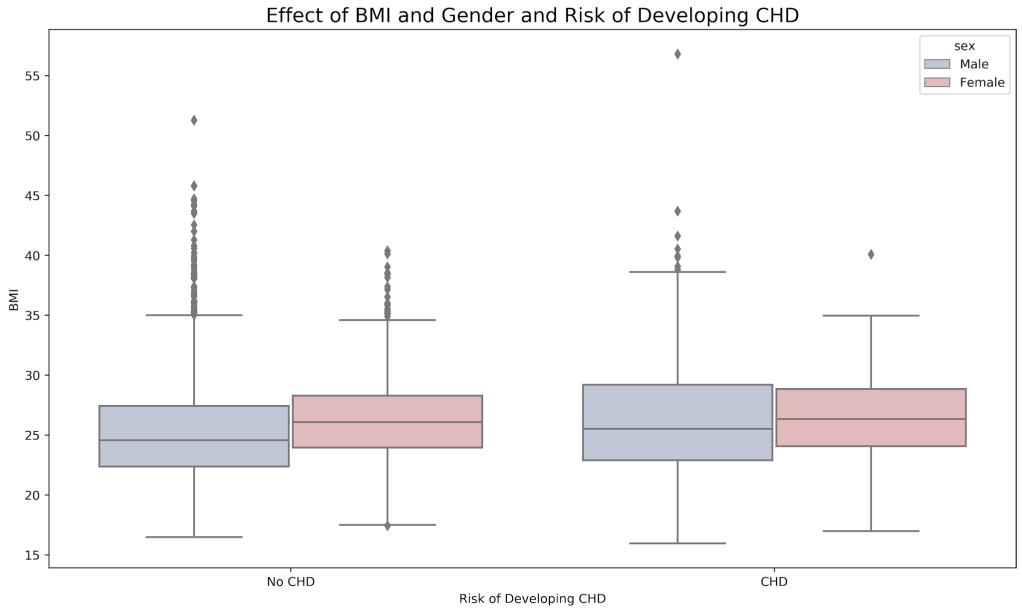


FIGURE 5 Effect of BMI and gender vs risk of developing cardiovascular heart disease.

The results of EDA were consistent with those found during the correlation analysis, and consistent with those observations made by other researchers who have explored this dataset. Now it's time to verify these findings via hypothesis testing, the details of which are discussed below.

STATISTICAL ANALYSIS :

After completing exploratory data analysis, Statistical analysis was conducted to perform high level hypothesis testing on each of the variables in the dataset. This was done to verify the correlation relationships revealed graphically, and will also serve to validate/verify findings from the Framingham Heart Study reported by researchers. Two types of statistical testing will be conducted, Chi-Squared testing and T-testing, based upon variable type.

Chi-Squared testing is used in situations when testing statistical independence or association between categorical variables. The following categorical variables will be tested using this method: education level, gender, choice to smoke, use of blood pressure medication, prevalence of stroke, prevalence of hypertension. Hypotheses tested were set up to determine whether there was or was not an association between the feature and risk of developing CHD.

T-testing was used to compare the means between groups of continuous data. This method was chosen rather than ANOVA because only two groups of data are being compared to determine if they are different. The response variables that will be tested using this method are continuous.

These variables included age, number of cigarettes smoked per day, total cholesterol, blood pressure, BMI, heart rate, and blood glucose. Hypotheses were set up to determine if patients in each group (with and without CHD) have the same risk of developing heart disease given the value of each feature.

During hypothesis testing, the majority of results were consistent with those found during the correlation analysis. Relationships that were interesting and of note were found. For example, hypothesis testing showed the association between smoking and CHD was not significant. However, when testing the number of cigarettes smoked per day and CHD three was a significant relationship found. While this is consistent with correlation values, the difference in significance was interesting. Another interesting observation was that a patient's resting heart rate had no significant association with risk of developing CHD. This is interesting because an elevated resting heart rate can sometimes be indicative of heart disease or overall wellness. It is entirely possible there is still some small relationship, but that it is not significant given the value of alpha utilized, which was 5%.

Overall statistical results can be seen in Appendix A and are shown as compared to the correlation values from the heat map.

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. This was done differently than in the EDA section, to allow for the application of certain models by converting all values to ordinal categorical and by "removing" null values. 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.

DATASET LIMITATIONS:

While the dataset does a good job of providing a basic feel for the Framingham data set, it's clearly been reduced from the original for the application. It would be nice to be able to know which features changed over time as patients checked in with researchers to have a better idea on what features change over time of those that affect risk.

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 could be leveraged by the client to target specific patient groups and ultimately lower the risk of developing CHD. Other details that limited the analysis include:

- Knowing whether patients reduced or increased smoking over time.
- Knowing how change in weight over time affects risk of developing CHD.
- Understanding how age at onset of smoking affects risk of CHD and being able to correlate this information to the data analyzed.
- Knowing that age at onset of obesity and high blood sugar/diabetes affect risk of developing CHD and being able to correlate this information to the data analyzed.

It's also possible that the reason the binary categorical for smoking (is_smoking) was found to reduce accuracy of the model is because the patient data wasn't captured over time. This eliminated the ability to compare people who have never smoked to those who have reduced smoking, which has a large effect on heart disease, and associated complications due to development of pulmonary issues. By only showing a single point in time with this datasets, the zero values in the number of cigarettes per day are likely assumed to be the same as zero values in whether or not the patient has ever smoked, which may not be a valid assumption.

OPPORTUNITIES FOR FUTURE WORK:

This project is as comprehensive as possible, given the constraints. However, there are opportunities to expand on this work, including, but not limited to:

- Building an interactive Application to allow for entry of new patient data by physician for risk prediction
- Building a model to predict amount of risk and projected decline if recommendations for healthy lifestyle are followed
- Building an interactive application for patients to track improvements given recommendation plan and risk improvement/elimination
- Sourcing alternative data from the Framingham Heart Data and utilizing it to improve models and expand on exploration as mentioned in dataset limitations.
- Looking at each patient over time to build a more comprehensive model.
- Expanding the dataset to include other patient data and effects of other comorbidities on the risk of developing CHD.

SUMMARY:

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.

The data was studied using the same machine learning models on two different selections 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 the 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.

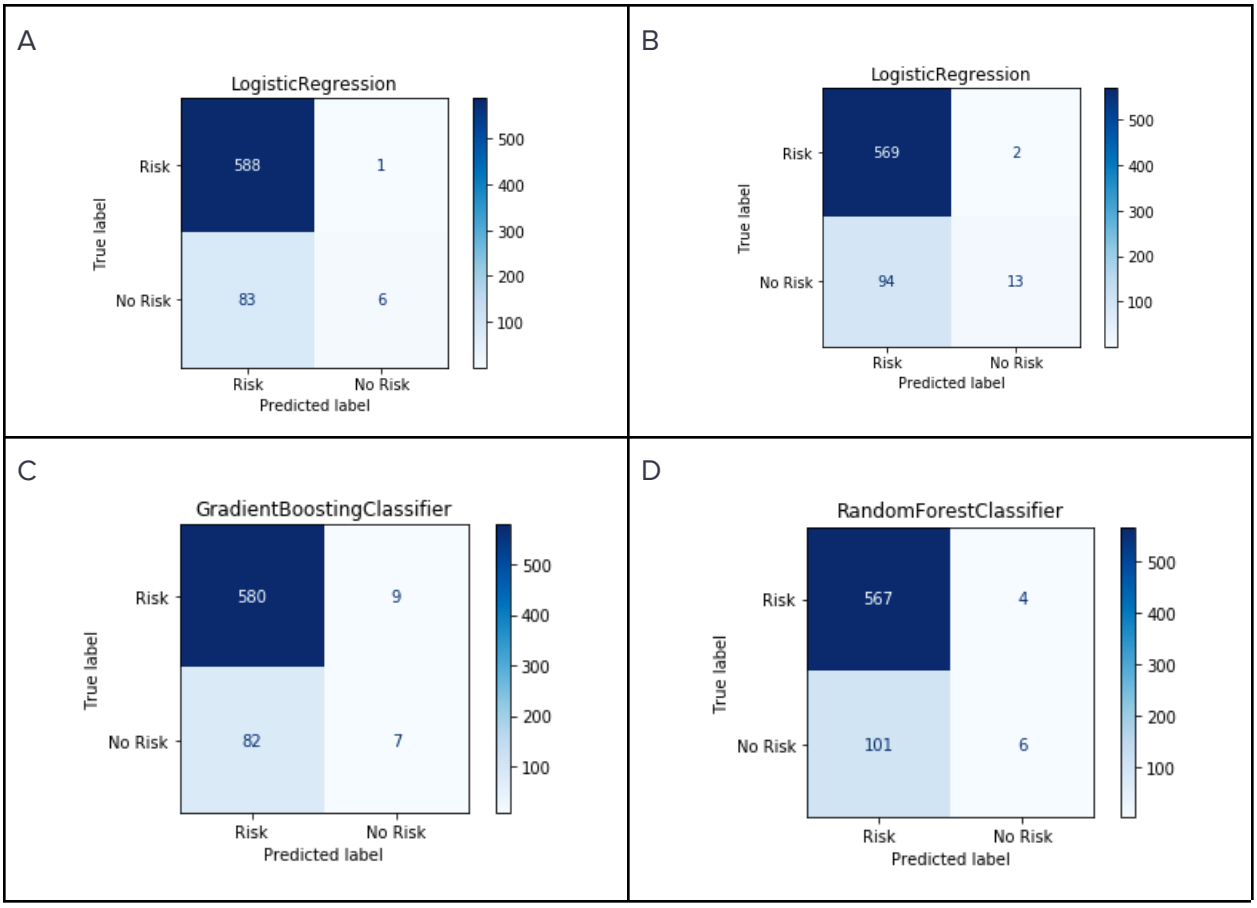


FIGURE 1: Confusion matrices for A) Logistic Regression- All Features, B) Logistic Regression- Reduced Features, C) Gradient Boosting All Features, D) Random Forest Reduced Features

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 B.

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

A

SUMMARY OF DATASET BY FEATURE:

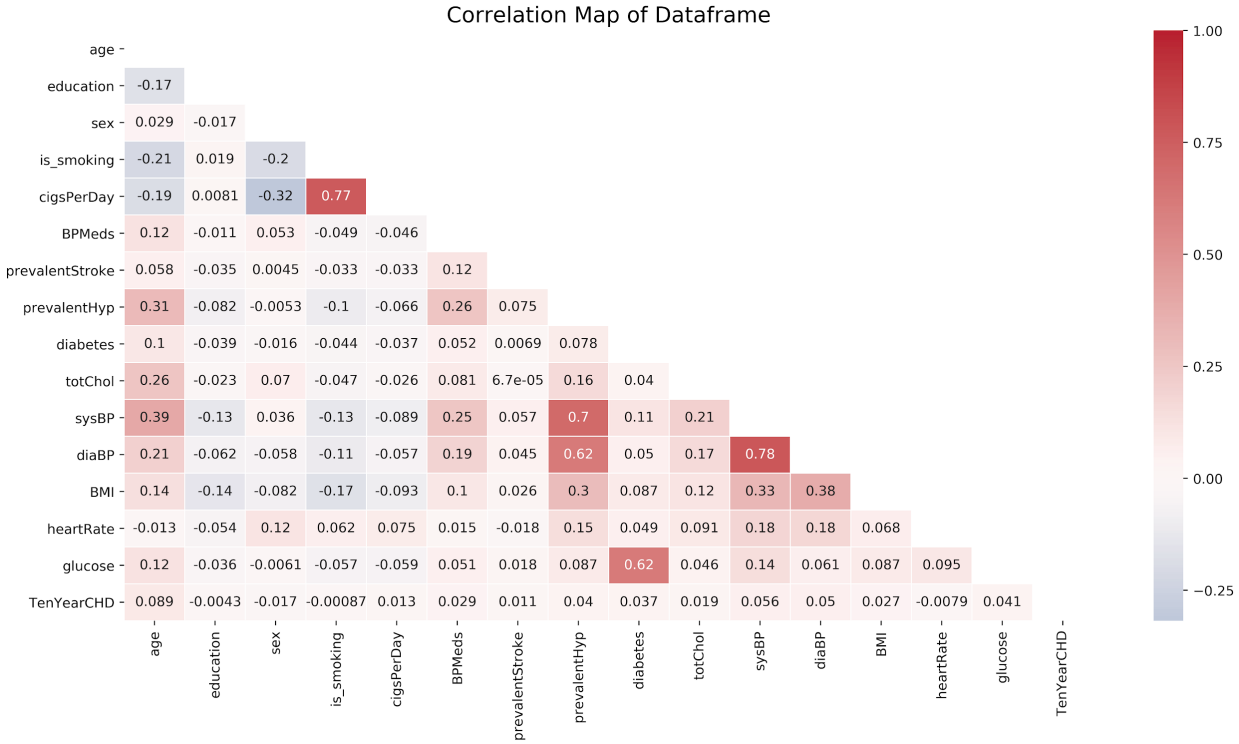
Feature	Values	Data Type	# Null	Description
Age	32-70	Continuous	None	Patient age in years, only whole numbers
Education	1, 2, 3, 4	Discrete	None	Education Level: 1-Some High School, 2-High School Diploma/GED, 3-College, 4-Degree
Sex	M, F	Discrete	None	Patient gender (M/F or 0=M, 1=F)
is_smoking	Yes, 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:

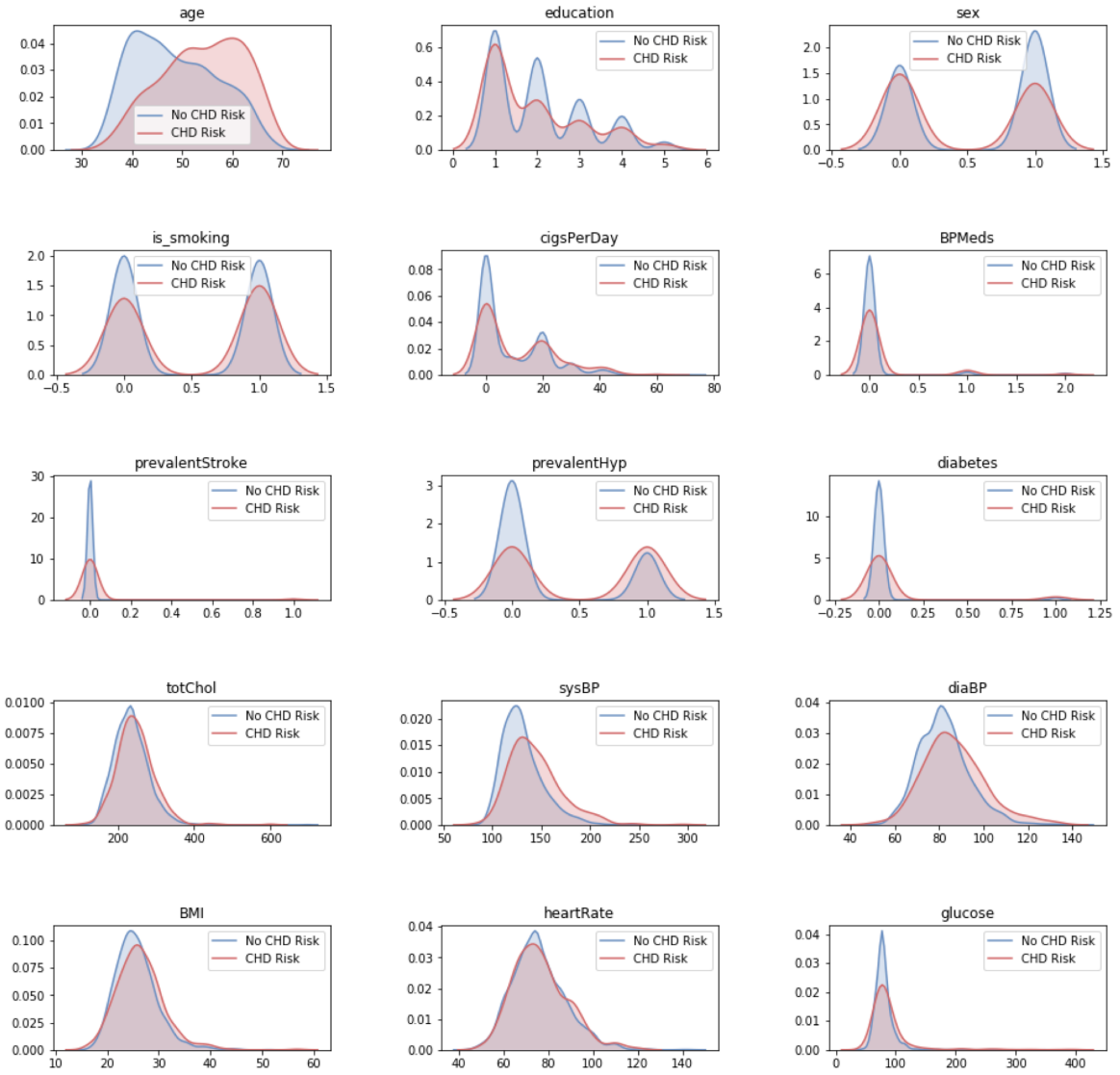
*Correlation value reported vs risk of developing CHD in the next 10 years

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

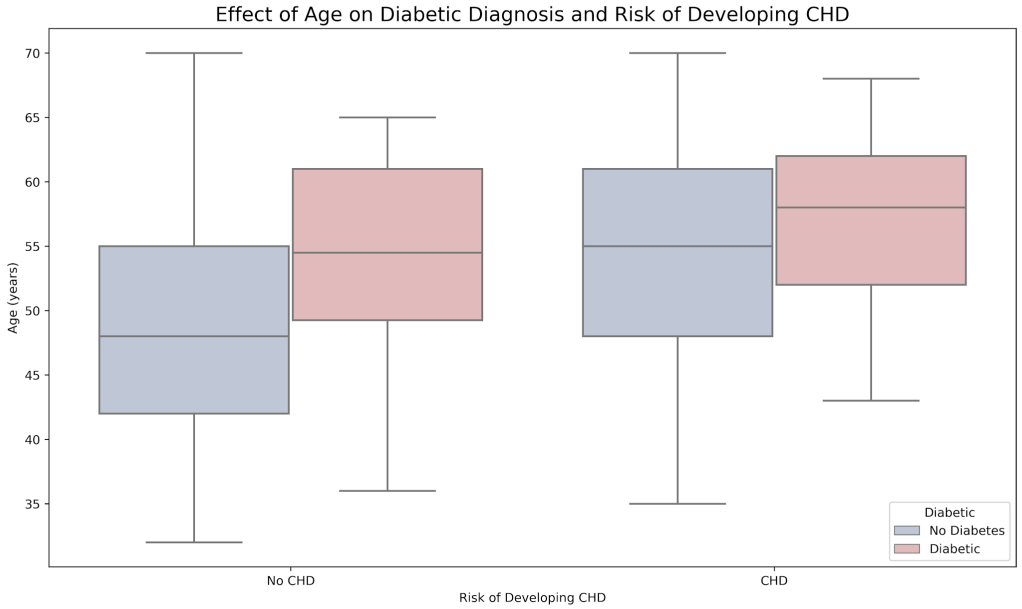
Correlation Heatmap showing correlation between features:

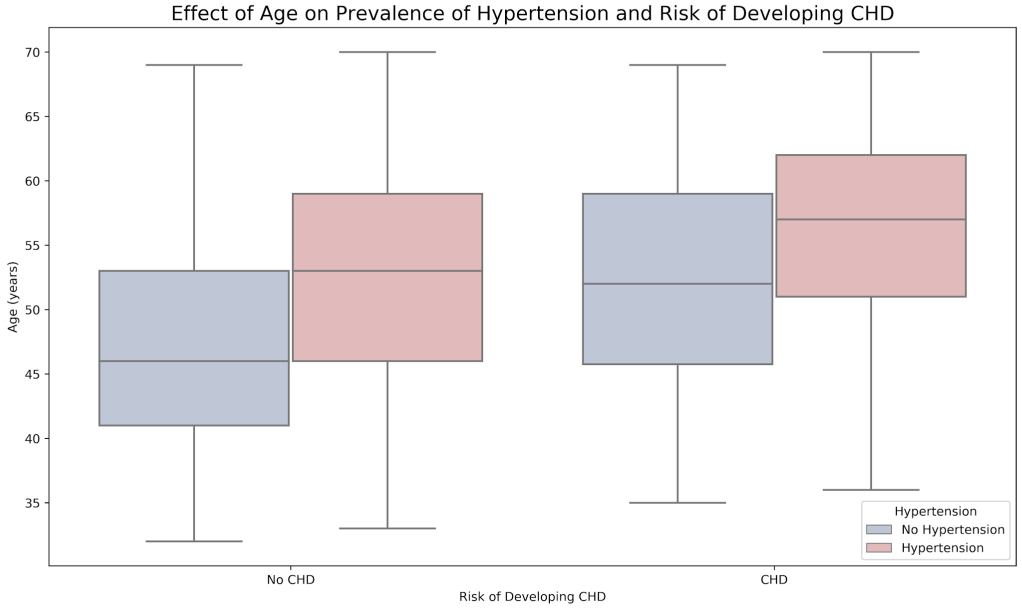


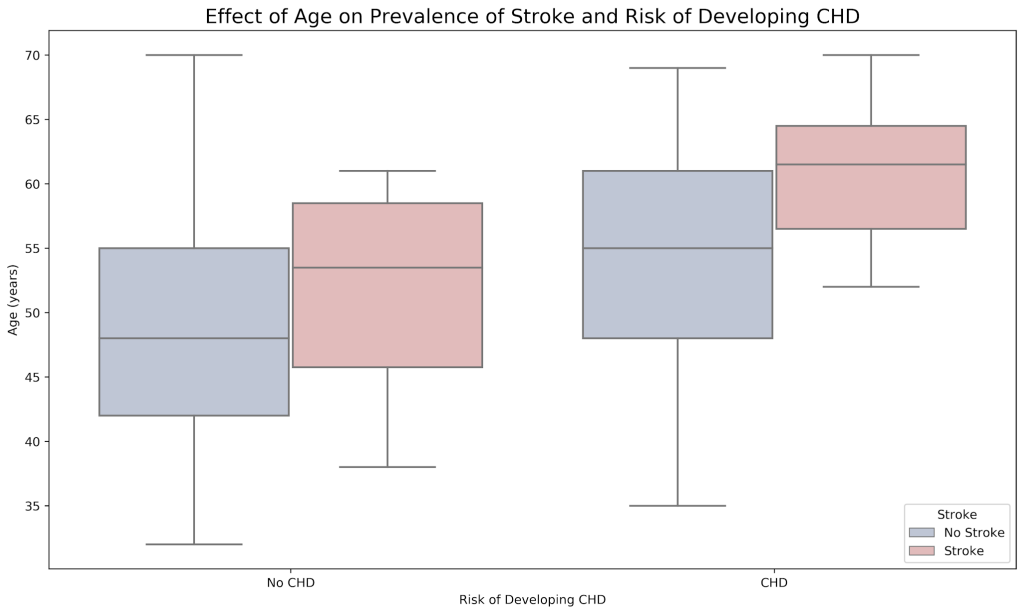
Kernel Density Estimation Distribution For Each Feature vs TenYearCHD showing correlation between features:



Boxplots of Features vs Risk of Developing CHD (by Feature)

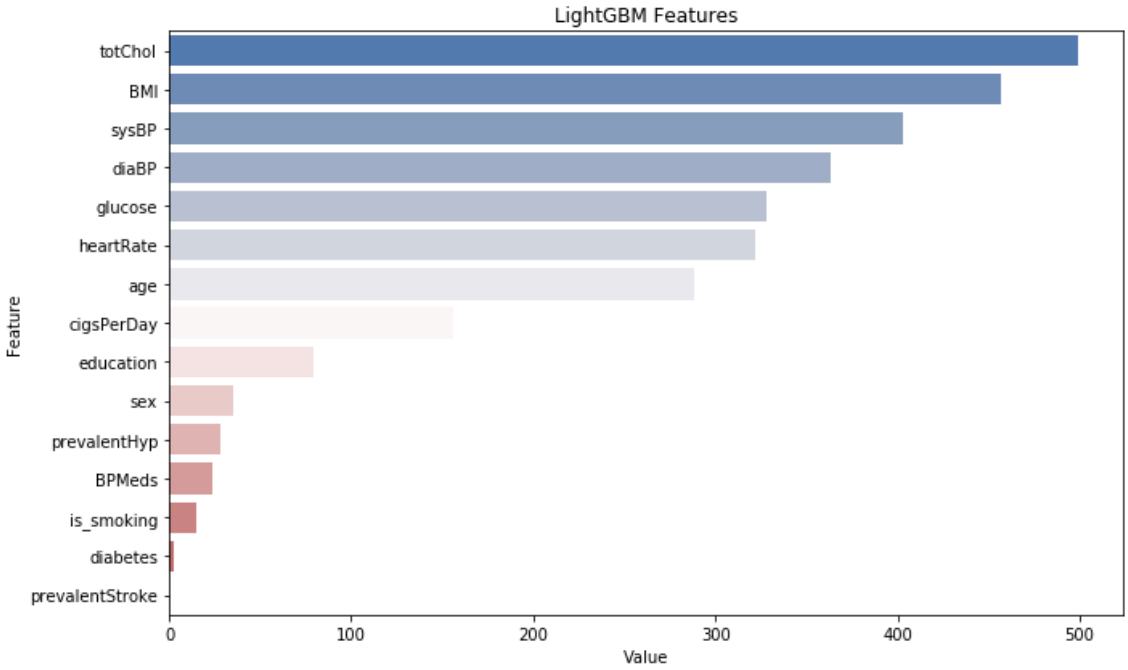




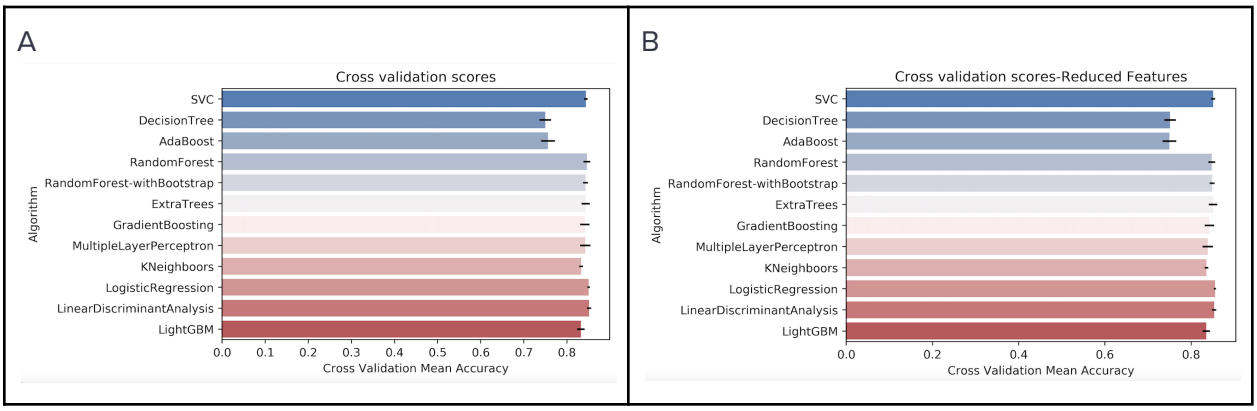

Effect of Age on Prevalence of Stroke and Risk of Developing CHD
Age (years)
70
65
60
55
50
45
40
35
No CHD
CHD
Risk of Developing CHD
Stroke
No Stroke
Stroke

Appendix B

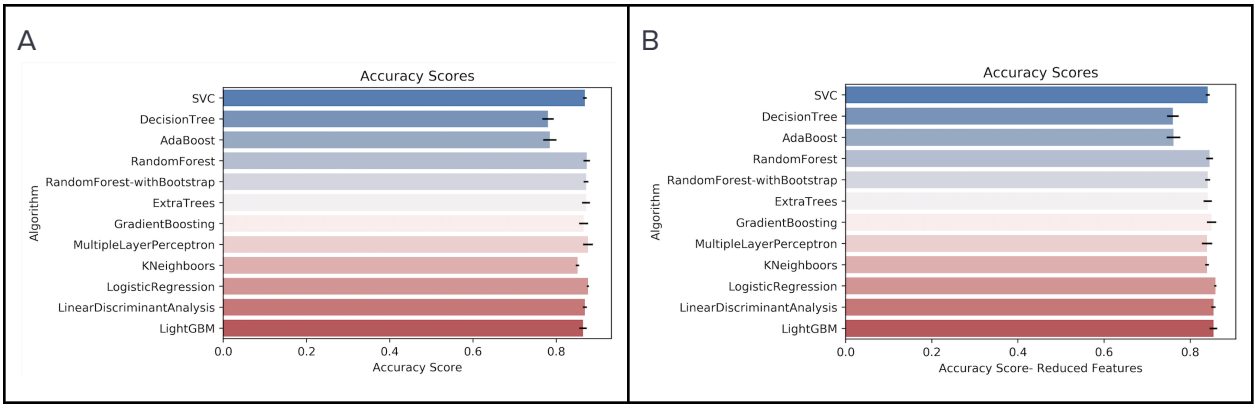
LIGHTGBM FEATURE SELECTION:



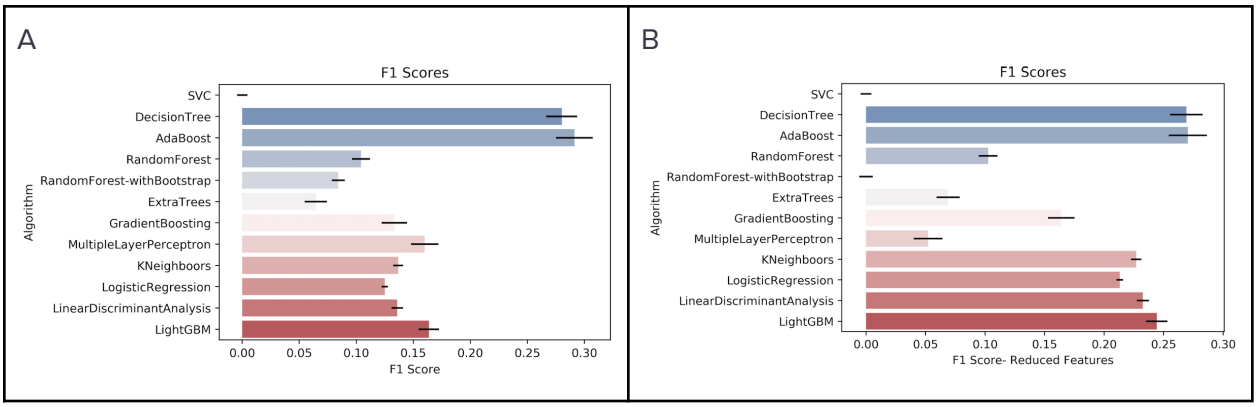
CROSS VALIDATION ACCURACY SCORES: a) all features, b) reduced features



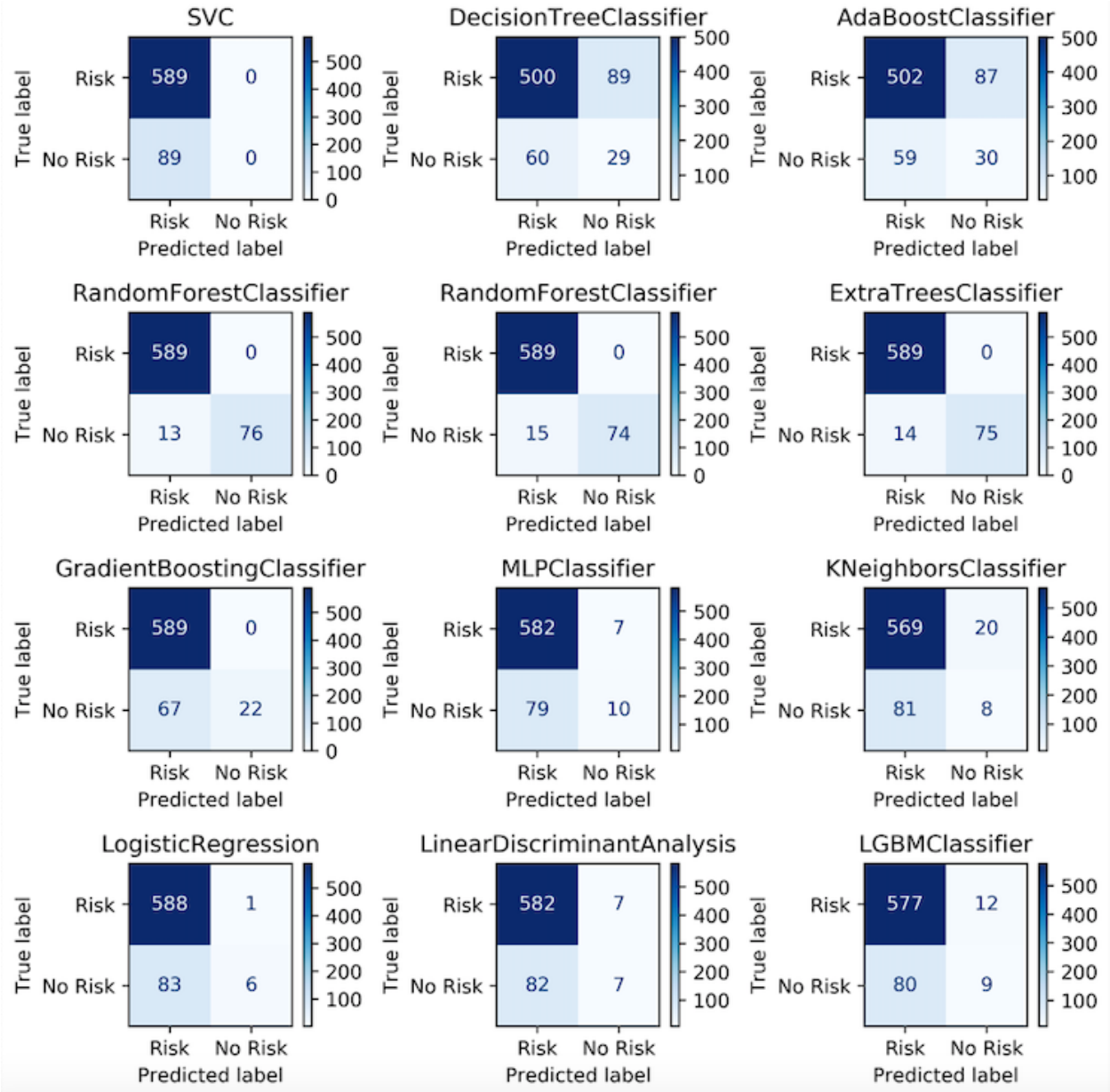
ACCURACY SCORES: a) all features, b) reduced features



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



CONFUSION MATRICES - ALL MODELS, ALL FEATURES:



CONFUSION MATRICES - ALL MODELS, REDUCED FEATURES: