



Predicting Cardiovascular Heart Disease

Caitlin Jansson
April 6, 2021
Milestone 2 Report

TABLE OF CONTENTS

- 1. Problem Statement**
- 2. Dataset**
- 3. Results of Data Analysis**
- 4. Machine Learning Results**
- 5. Conclusions**
- 6. References**
- 7. Thanks**
- 8. Appendix - Additional Figures**
 - a. Figures and Graphs from Data Analysis**
 - b. Figures and graphs from Machine Learning**



PROBLEM STATEMENT

Cardiovascular heart disease (CHD) is the leading cause of death annually worldwide.

Cardiovascular diseases can be managed if caught early and simple lifestyle changes are made.

The intent of this project is to explore 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.
- Develop insights regarding CHD by exploring patient data
- Make recommendations for a patient reeducation program to reduce risk of developing CHD



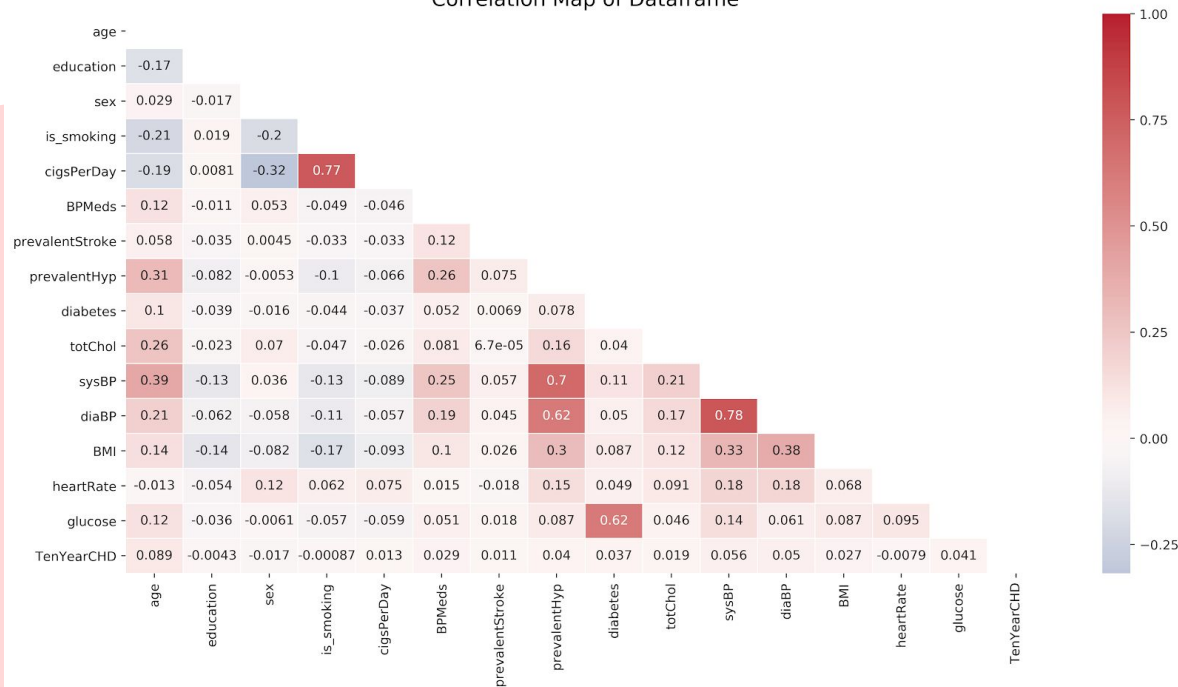
THE DATASET

- 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
- The subset of data utilized contains information on over 4,200 patients.
 - 43% Male
 - 50.5% Nonsmokers
 - Patients ranging 32-70 years old
- Data is 15% Patients At Risk and 85% patients with no risk of developing CHD in the next 10 years
 - Due to imbalance will have to use different tactics during model building

Find the Data Here: <https://www.kaggle.com/christofel04/cardiovascular-study-dataset-predict-heart-disease>

FEATURES INVESTIGATED

Correlation Map of Dataframe



- Age
- Education Level
- Gender
- Smoker vs Nonsmoker
 - # of cigarettes smoked daily
- Prevalence of
 - Stroke
 - Hypertension
 - Diabetes
- Cholesterol
- Blood Pressure
- BMI
- Heart Rate
- Blood Sugar

Exploratory Data Analysis Results

The risk of developing Cardiovascular Heart Disease increases with:

1. Smoking, the more a patient smokes the higher the risk.
2. Cholesterol, and higher cholesterol means higher risk.
3. Blood Pressure
4. Prevalence of Stroke
5. Prevalence of Hypertension
6. Diabetes

CONCLUSIONS

Patient groups at highest risk, to target with reeducation material include:

1. Patients over 50
2. Males
3. Patients who smoke
4. Patients with diabetes and hypertension.

CONCLUSIONS

Recommendations for reeducation material:

1. Managing diet and exercise to reduce blood pressure, manage blood sugar, manage weight can reduce risk!
2. Cutting back on smoking can reduce risk: even if the patient only reduces consumption.

In-Depth Analysis Results

Multiple Models were built and Tested For Accuracy





- Model accuracy was determined by comparing correct predictions and incorrect predictions.
- Models were built using 2 methods: (1) using all patient data and (2) using patient data without prevalence of stroke and if a patient smokes or not
- Model efficacy was determined by
 - (1) Least number of patients at risk incorrectly categorized, and
 - (2) Least number patients not at risk incorrectly categorized
 - (3) Correctly categorized patients

Model Results and Use

Actual Risk		Predicted Risk	
		At Risk	No Risk
At Risk	# Pts <u>at risk</u> correctly categorized 511	# Pts <u>at risk</u> categorized as <u>not at risk</u> 0	Most important metric!
No Risk	# Pts <u>not at risk</u> categorized as <u>at risk</u> 0	# Pts <u>not at risk</u> correctly categorized 2878	

2nd most important metric!

Model Results and Use

Actual Risk	At Risk	# Pts <u>at risk</u> correctly categorized 	# Pts <u>at risk</u> categorized as <u>not at risk</u> 
	No Risk	# Pts <u>not at risk</u> categorized as <u>at risk</u> 	# Pts <u>not at risk</u> correctly categorized 
		At Risk	No Risk
		Predicted Risk	

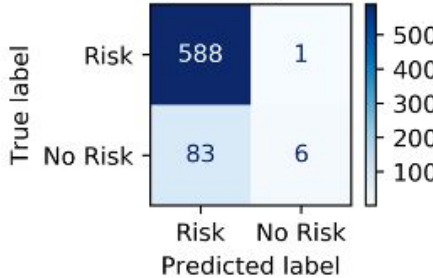
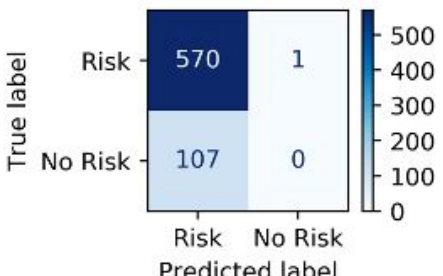
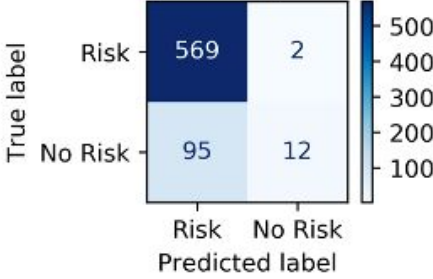


- 1) Recommend these patients for immediate lifestyle change
- 2) Consider additional measures as patient is AT RISK



- 1) Recommend these patients for immediate lifestyle change-

Final Model Results

All Data	<p>SVC</p>  <p>Accuracy = 84.4%</p>	<p>RandomForestClassifier</p>  <p>Accuracy = 84.6%</p>	<p>LogisticRegression</p>  <p>Accuracy = 85.5%</p>
	<p>SVC</p>  <p>Accuracy = 85.1%</p>	<p>RandomForestClassifier</p>  <p>Accuracy = 84.8%</p>	<p>LogisticRegression</p>  <p>Accuracy = 85.5%</p>
	<p>Best Model</p>		

Final Conclusions from Analysis

1. Best model accuracy was achieved by removing prevalence of stroke and whether or not patient smokes
2. Accuracy of categorizations is $\sim 85.5\%$
3. Recommendations for patient treatment based upon final categorization in resulting matrix

REFERENCES

- [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>

THANKS

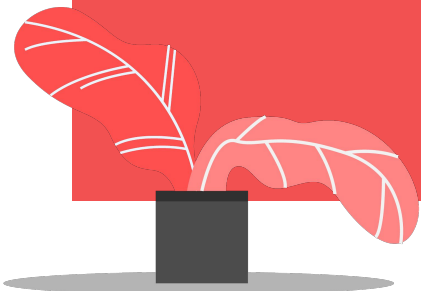
Does anyone have any questions?

caitlinjeansson@gmail.com

https://github.com/CJEJansson/Springboard_Projects



CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, infographics & images by **Freepik**



APPENDIX A:

Exploratory Data Analysis and Statistics

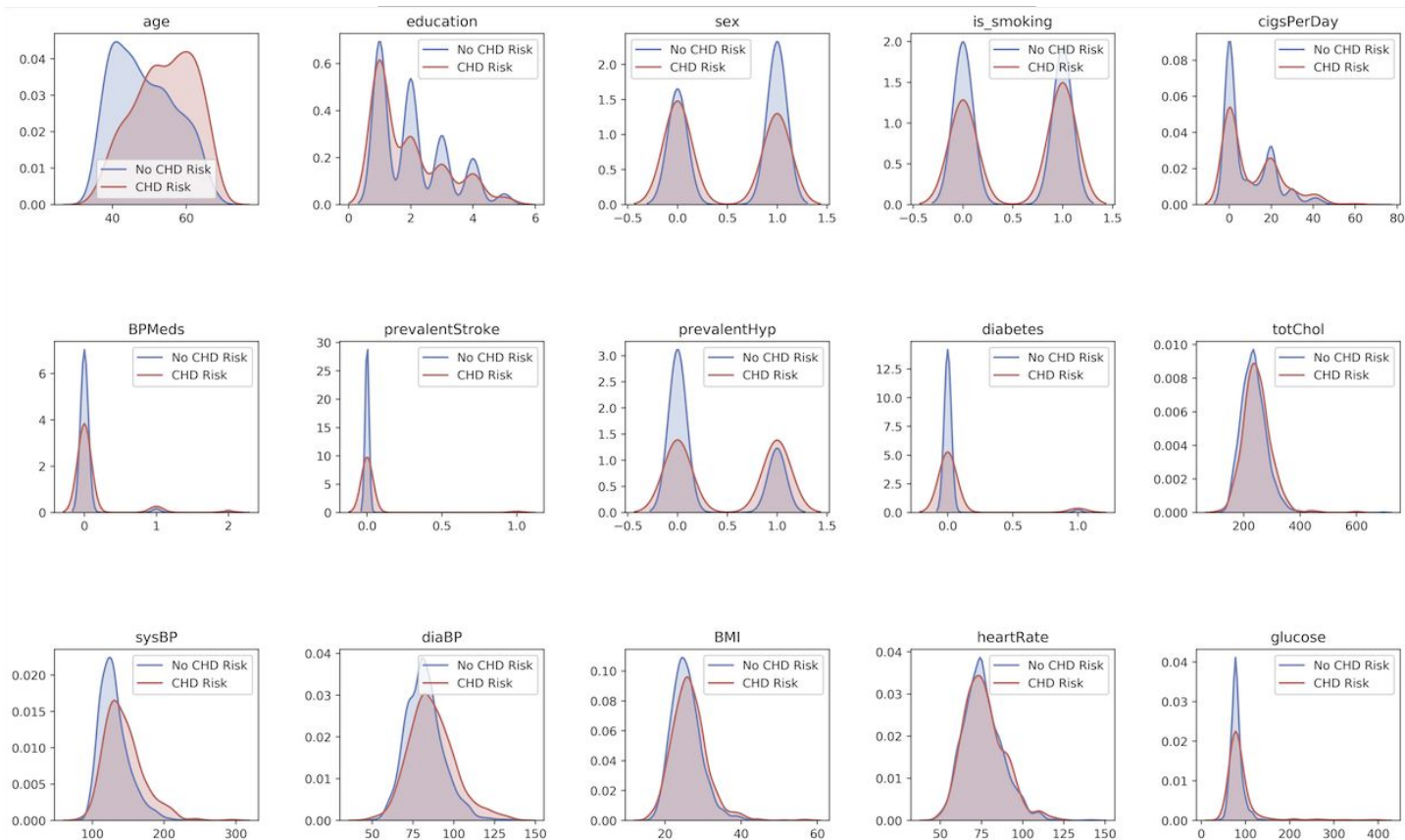
SUMMARY OF DATASET BY FEATURE

Feature	Values	Data Type	# of 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 occurrences 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)

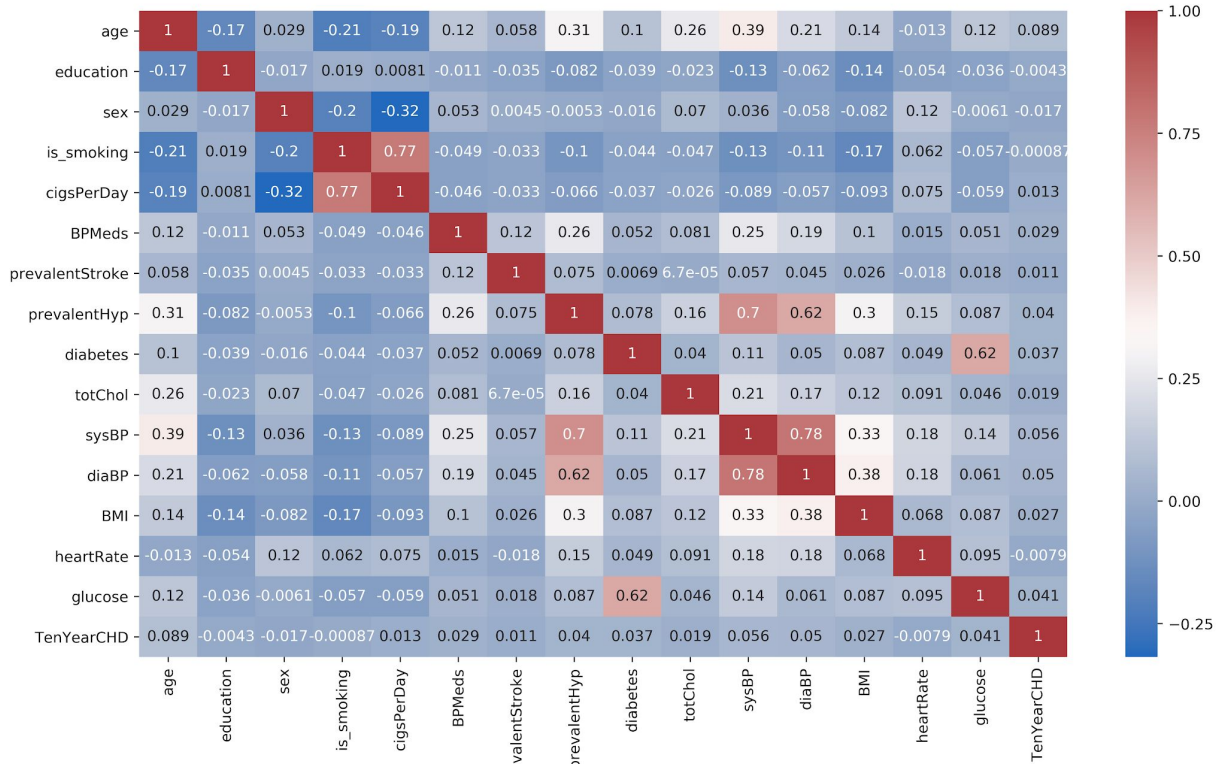
SUMMARY OF STATISTICAL ANALYSIS BY FEATURE

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

KDE Distributions by Feature vs Risk of Developing CHD

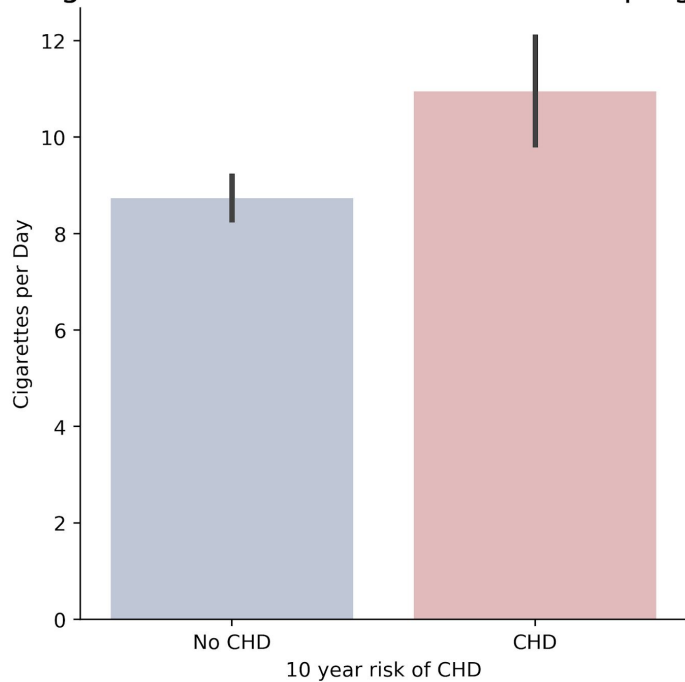


Full Correlation Heatmap

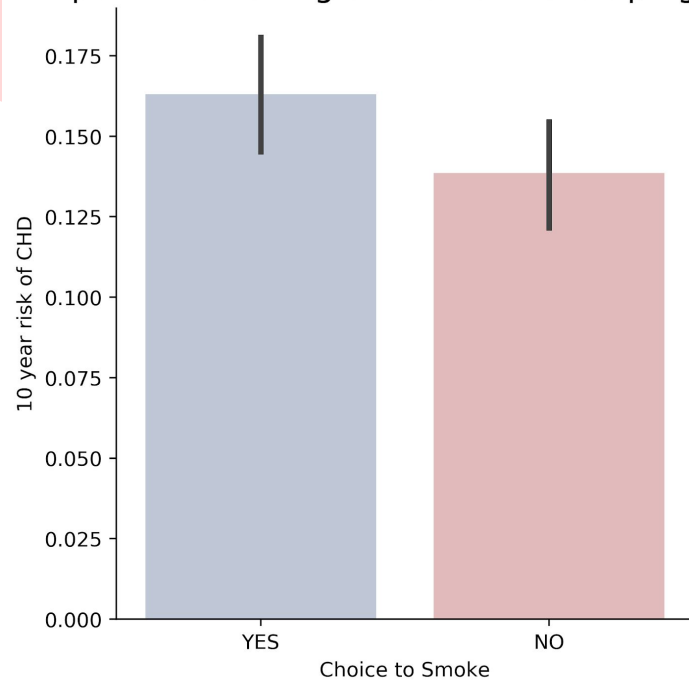


INITIAL FINDINGS: Smoking and CHD

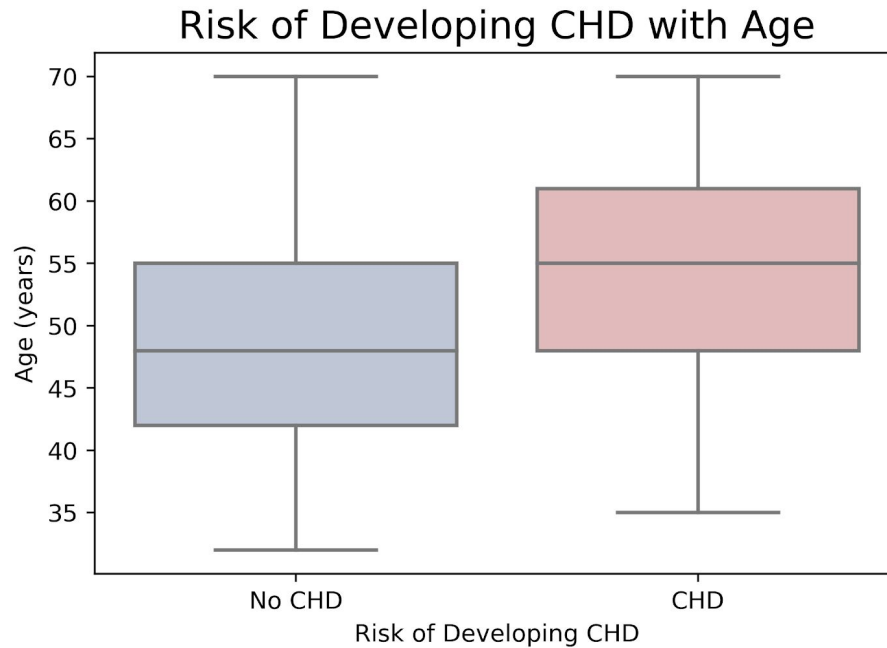
Cigarettes Smoked vs Risk of Developing CHD



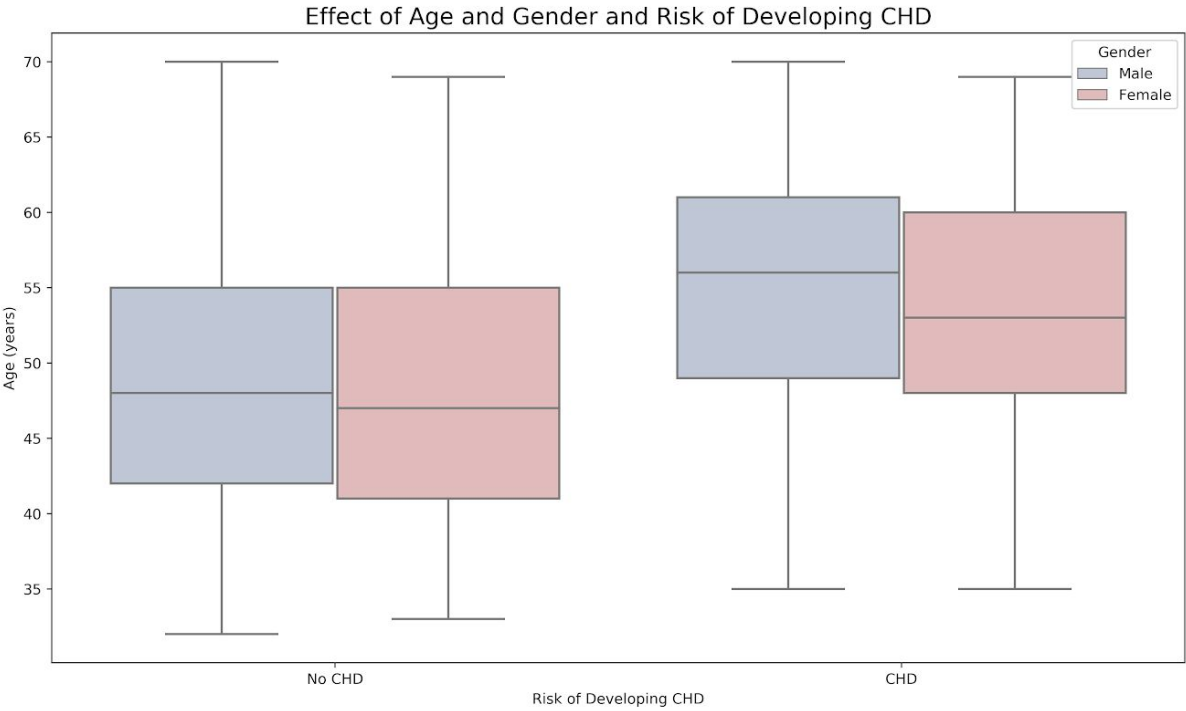
Impact of Smoking on Risk of Developing CHD



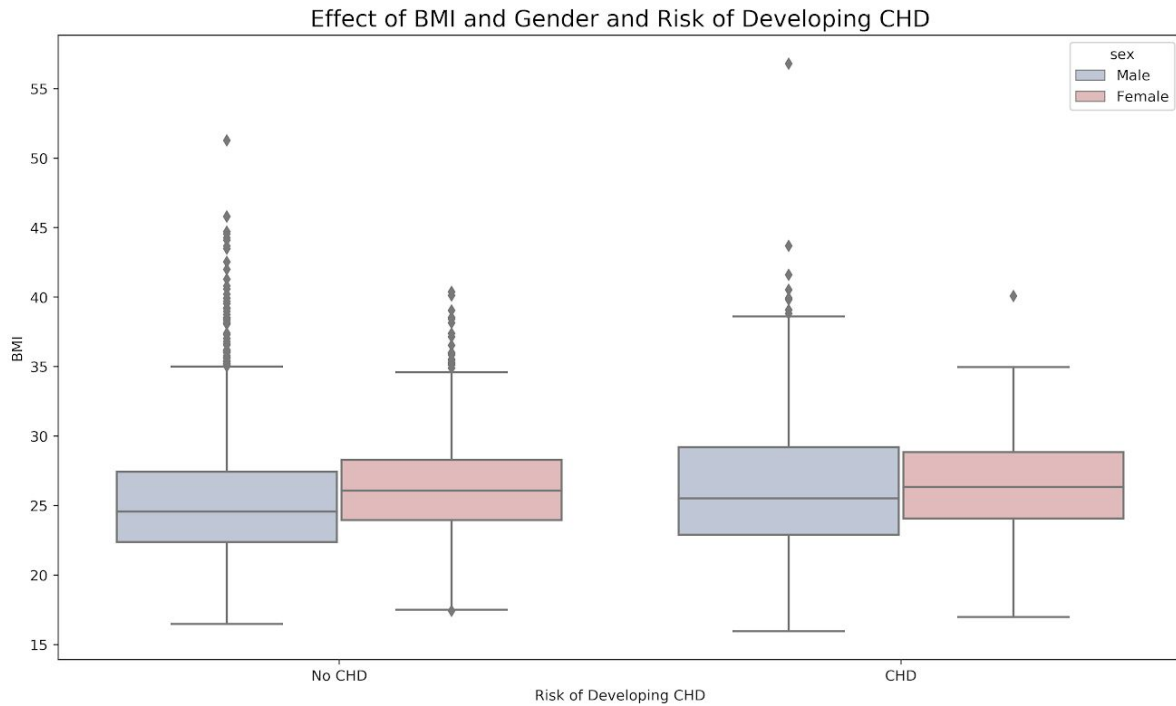
INITIAL FINDINGS: AGE AND CHD



INITIAL FINDINGS: AGE, GENDER AND CHD

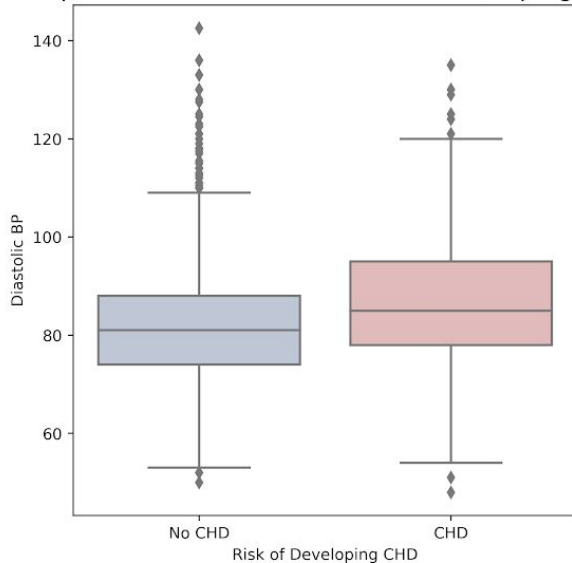


INITIAL FINDINGS: BMI, GENDER AND CHD

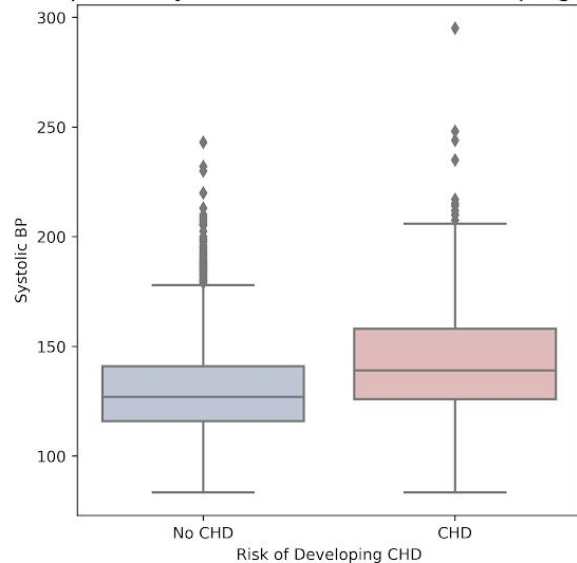


INITIAL FINDINGS: BLOOD PRESSURE AND CHD

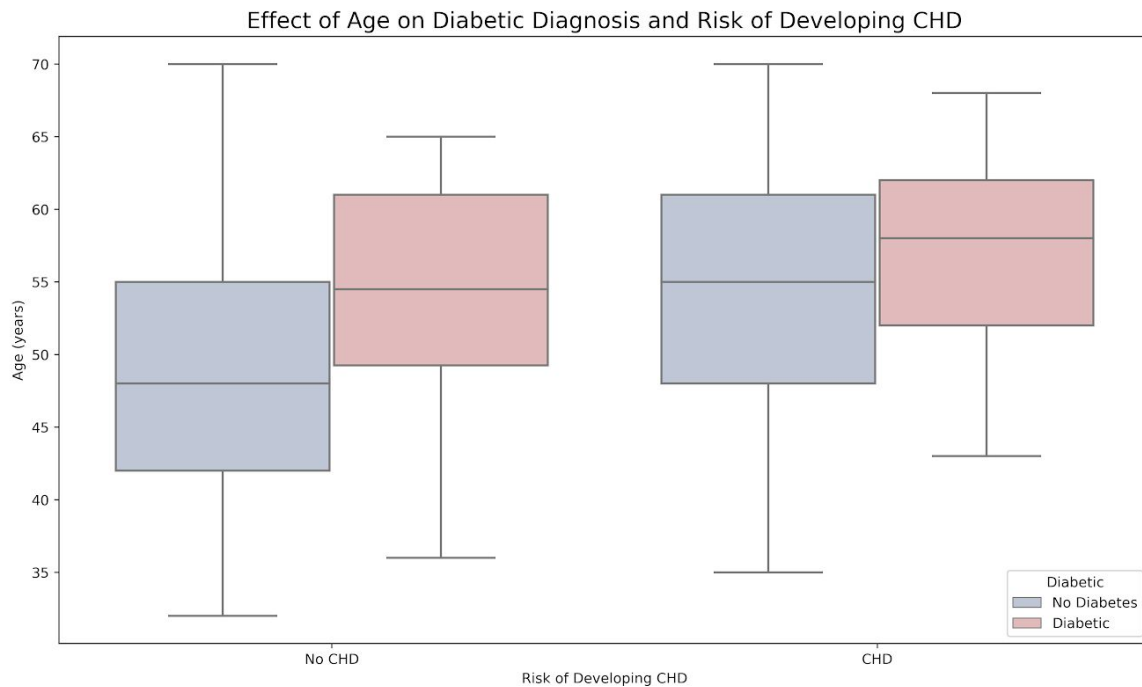
Boxplot of Diastolic BP vs Risk of Developing CHD



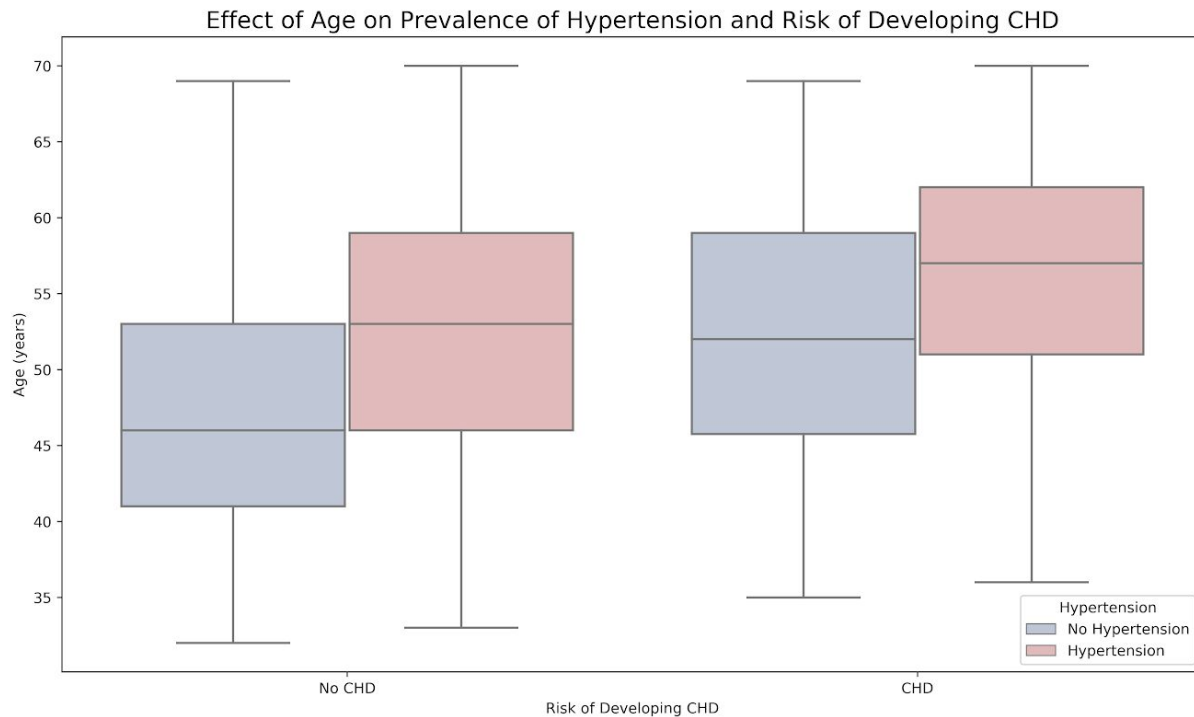
Boxplot of Systolic BP vs Risk of Developing CHD



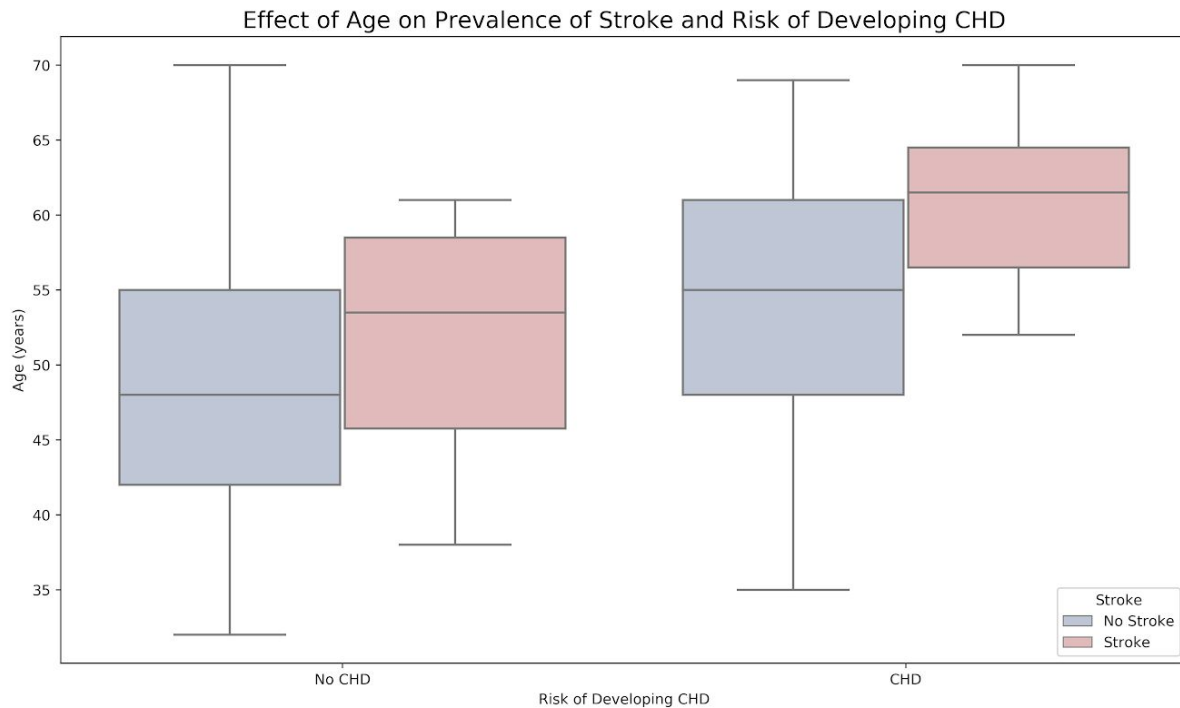
INITIAL FINDINGS: AGE, DIABETES AND CHD

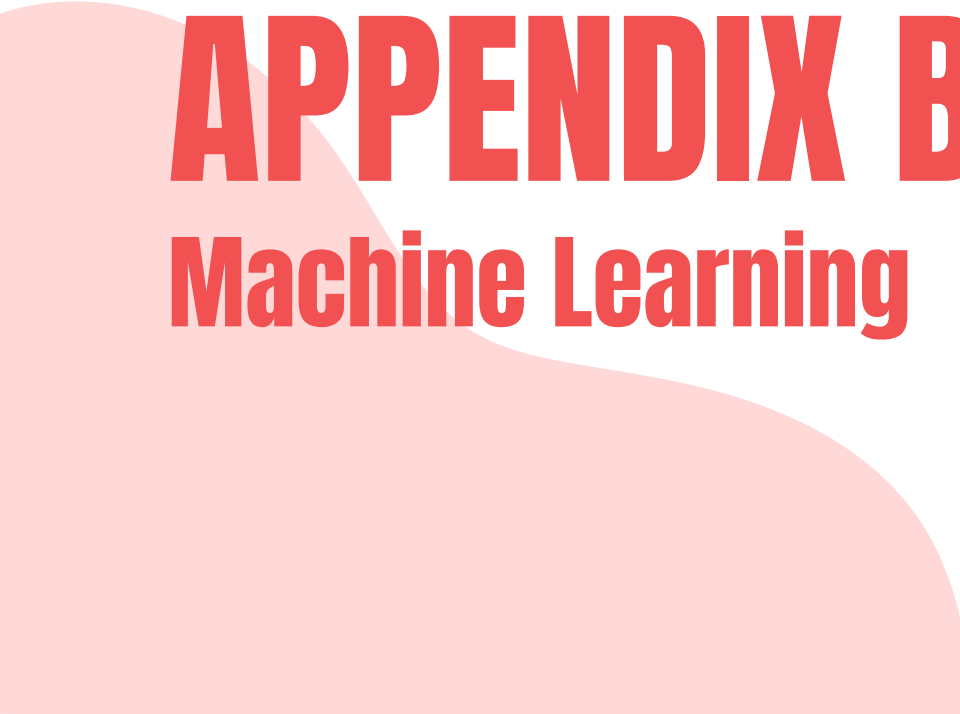


INITIAL FINDINGS: AGE, HYPERTENSION AND CHD



INITIAL FINDINGS: AGE, STROKE AND CHD

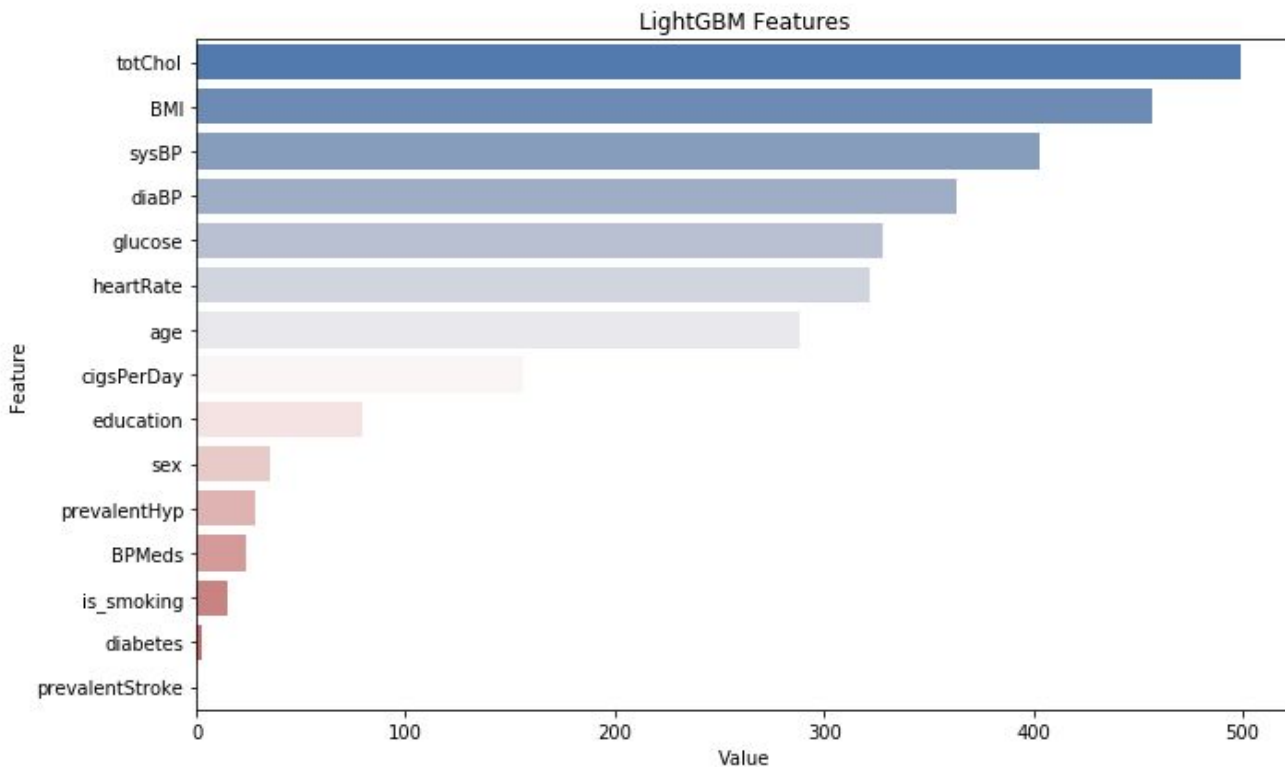


A large, soft pink abstract shape with organic, flowing edges occupies the bottom-left portion of the slide, partially overlapping the text.

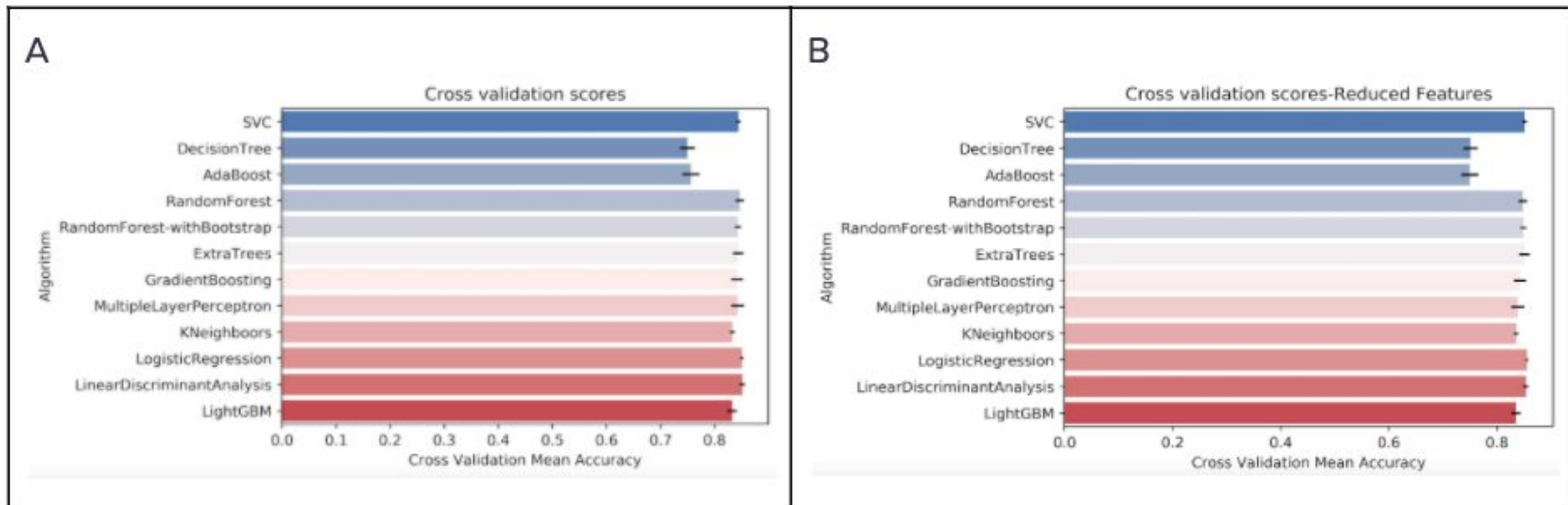
APPENDIX B:

Machine Learning

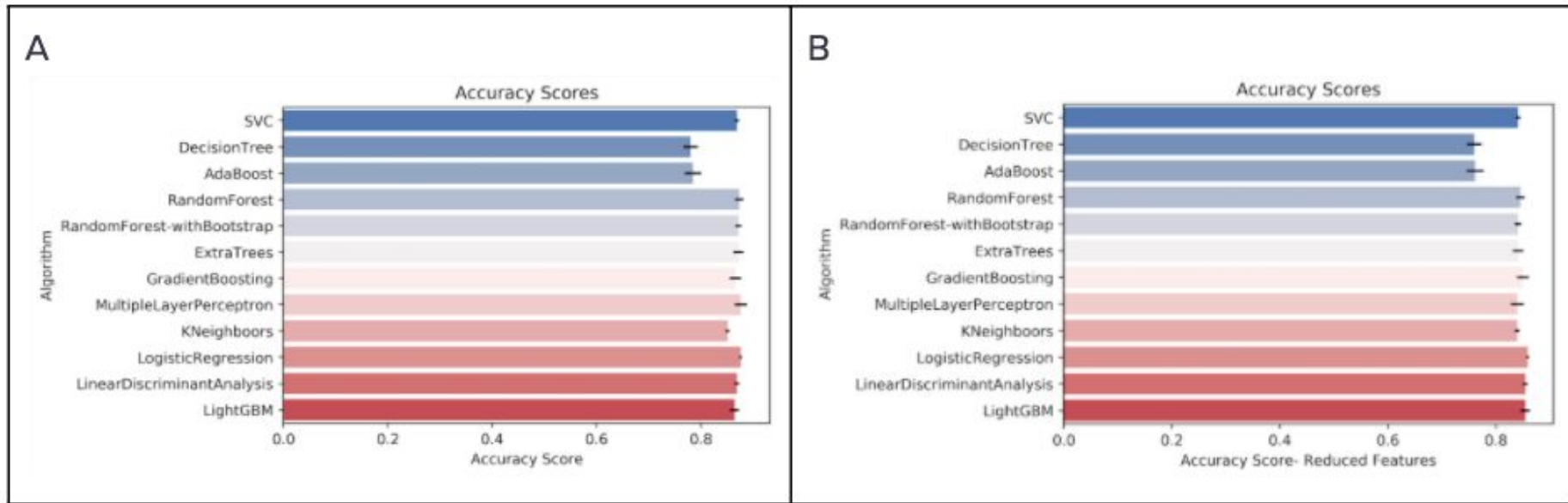
Light GBM Feature Selection



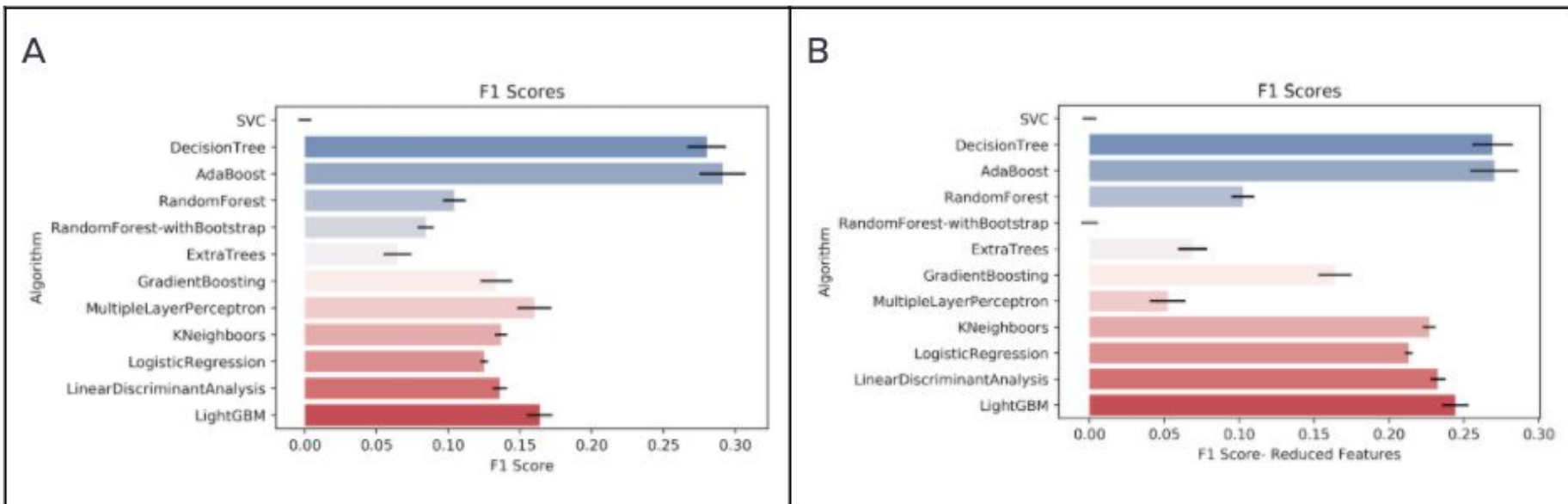
Cross Validation Accuracy Scores



Accuracy Scores

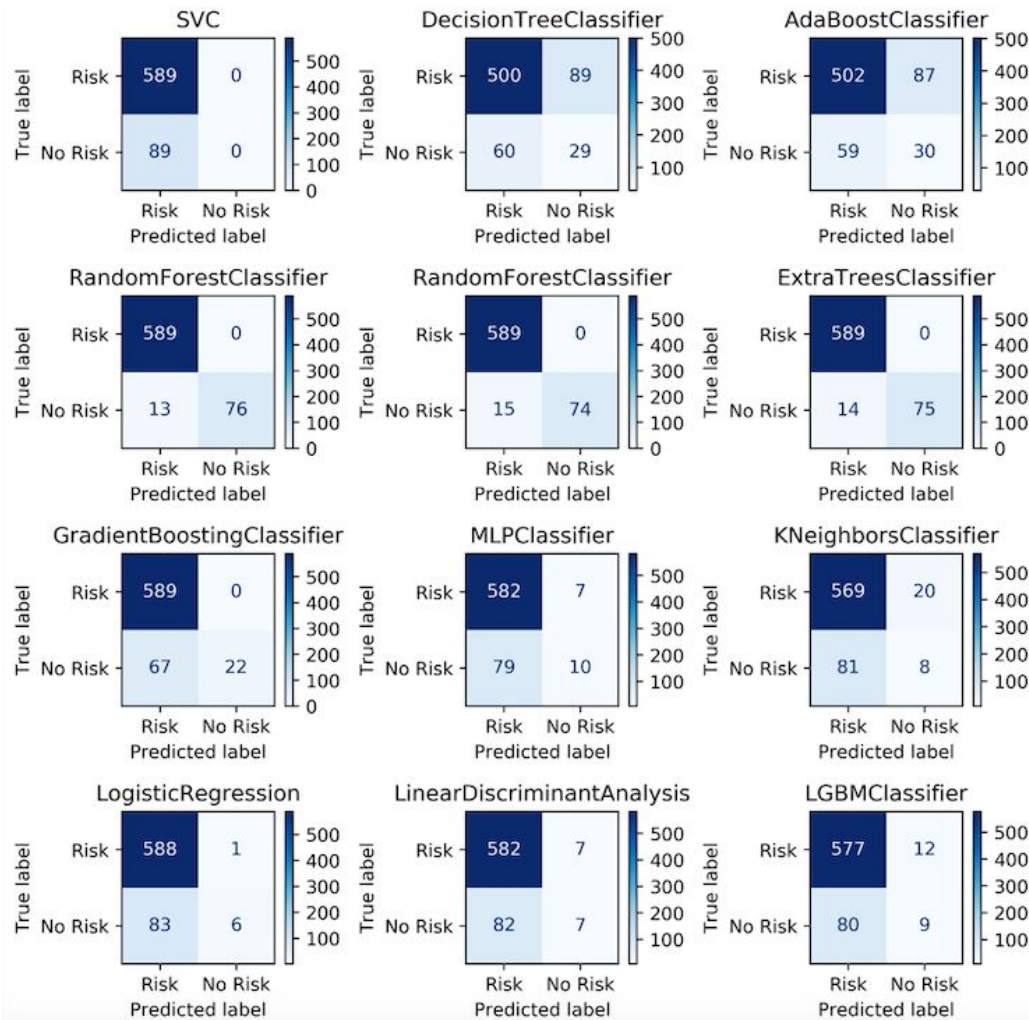


F1 Scores

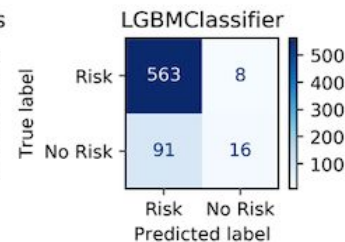
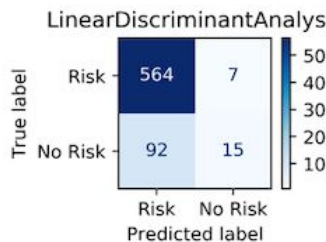
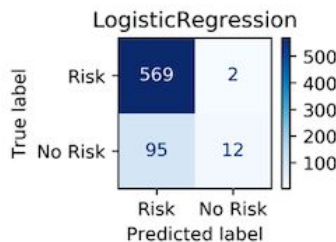
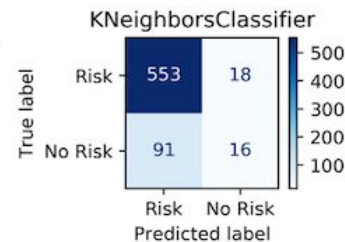
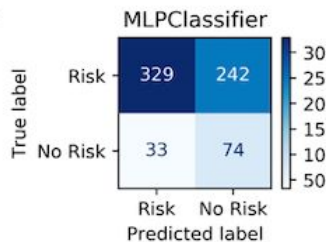
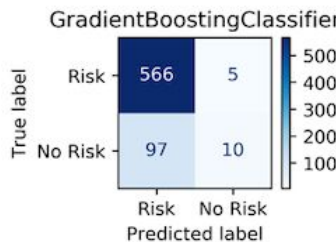
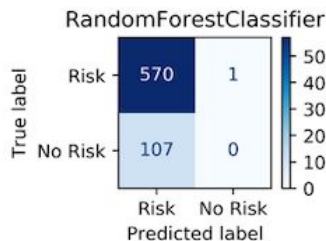
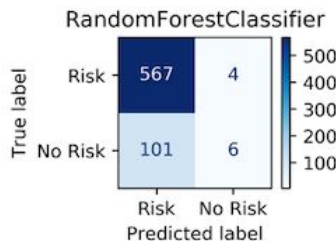
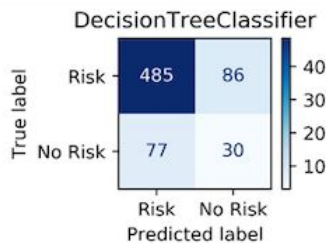
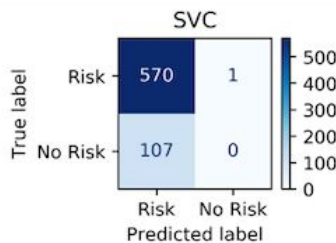


Confusion Matrices:

All Models, All Features



Confusion Matrices: All Models, Reduced Features



Summary of Accuracy Scores: 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
	KNeighbors	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
	KNeighbors	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)					