Introduction

Problem Identification

Heart diseases are the second leading cause of deaths in Canada. In 2012, the disease claimed 48,000 lives. A lot of the causes attributed to the spread and prevalence of the disease is as a result of poor lifestyle choices, stress, high blood pressure and cholesterol levels. In order to be able to cater to this growing need for accurate monitoring and evaluation of patients struggling with chest pains or any heart-related symptoms, there needs to be an efficient system in place that can accurately identify patients that are most likely to suffer from heart attacks.

Client: Health Care Professionals

The health system in Canada is a universal one. This means that every Canadian is able to access health care needs without incurring a personal cost. As a result, this system suffers from an efficiency problem and mismanagement of resources. This could mean that patients with a high probability of suffering from a heart attack are being turned away because hospitals do not have the resources or capacity to be able to cope with the growing demand.

Objective

In order to maximize efficiency and promote the optimal allocation of resources, I want to build a model that will predict which patients are most likely to suffer from a heart attack based on several health-related attributes as well as lifestyle indicators that I am provided with.

Data

I have access to data on a patient's **age, sex, chest pain type, resting blood pressure, cholesterol levels, fasting blood sugar, electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression, ST segment slope, fluoroscopy colored vessels, and thalassemia,** which I will be using as my independent variables (descriptive variables) in the analysis. My dependent variable (target variable) will be whether one of the patients' major coronary vessels has a blockage that is greater or smaller than 50%.

Dataset: UCI Machine Learning Repository/Heart Disease Dataset

Variable	Variable Description
Age	[integer]
Sex	[1,0] - Male/Female
Chest Pain Type	1- Typical Angina 2- Atypical Angina 3- Non-anginal Pain 4- Asymptomatic
Resting Blood Pressure	[integer mm Hg]
Serum Cholesterol Level	[integer mg/dl]
Fasting Blood Sugar	True | False - Criteria: >120mg/dl
Resting Electrocardiographic Results	0: Normal | 1: Having ST-T Wave Abnormality | 2: Showing Left Ventricular Hypertrophy
Maximum Heart Rate Achieved	[integer]
Exercise Induced Angina	Yes | No [0,1]
ST Depression	Depression of the ST induced by exercise relative to rest.
ST Segment Slope	Slope of the peak exercise ST segment - 1: Upsloping | 2: Flat | 3: Downsloping
Fluoroscopy Colored Vessels	Number of blood vessels colored by Flourosopy [1-3]
Thalassemia	3=Normal | 6=Fixed Defect | 7=Reversible Defect
Diagnosis of Heart Disease	>50% | <50% narrowing of major vessel [0,1]

Data Wrangling, Storytelling and Inferential Statistics

Part 1: Data Wrangling

Steps to Creating the Dataset

We started by downloading 4 processed datasets from the UCI Machine Learning Repository consisting of data from Cleveland, Long Beach, Switzerland and Hungary. This data is cross sectional and from the year 1988. There are 15 variables with 916 observations. We compiled and combined these datasets into a single working dataset in order to be processed further. We proceeded to inspect the dataset and clean it to make sure it was ready for analysis.

Dealing with Missing Values (Nans)

The accompanying documentation to the data informs us that values are considered missing if they have '-9' entries as well as '?' entries. We did not delete the rows of this dataset that had missing values. Instead, we chose to replace these values with a 'NaN' value (blank value) and keep the remaining row information. We do this in order to avoid taking away variation in our data which would reduce the power and accuracy of our analysis. However, it is important to note that we will eventually drop these rows when conducting supervised machine learning, due to the fact that supervised learning does not take these values into account. It is unwise to interpolate, back-fill or front-fill the data as we risk reducing its quality by altering it.

Correlations

We proceed to inspect whether there might be any high correlations between the variables in our data. We want to avoid the latter due to the fact that this might bias or skew our analysis. We observe no correlations above 50% within our data and therefore, do not have to drop any variables as a result. A full correlation table is exhibited later in this report.

Dropping Variables with Majority Missing Values

In the dataset, we dropped four variables due to the fact that the majority of its data was missing and therefore adds little to no value to our analysis. We now have 11 explanatory variables in our dataset.

Compiling the Cleaned Dataset

Now that we have cleaned the dataset and made it ready for analysis, we saved it as a new file for future uses. Here is a look at what our dataset looks like.

	age	sex	cp	rbp	scl	fbs	rer	mhra	eia	std	new_dhd
0	67.0	1.0	4.0	160.0	286.0	0.0	2.0	108.0	1.0	1.5	1
1	67.0	1.0	4.0	120.0	229.0	0.0	2.0	129.0	1.0	2.6	1
2	37.0	1.0	3.0	130.0	250.0	0.0	0.0	187.0	0.0	3.5	0
3	41.0	0.0	2.0	130.0	204.0	0.0	2.0	172.0	0.0	1.4	0
4	56.0	1.0	2.0	120.0	236.0	0.0	0.0	178.0	0.0	0.8	0

Part 2: Data Storytelling

In this section, we study our data further by analysing graphical representations. This will give us a deeper insight into what our data can tell us and will help us make important interpretations in our analysis.

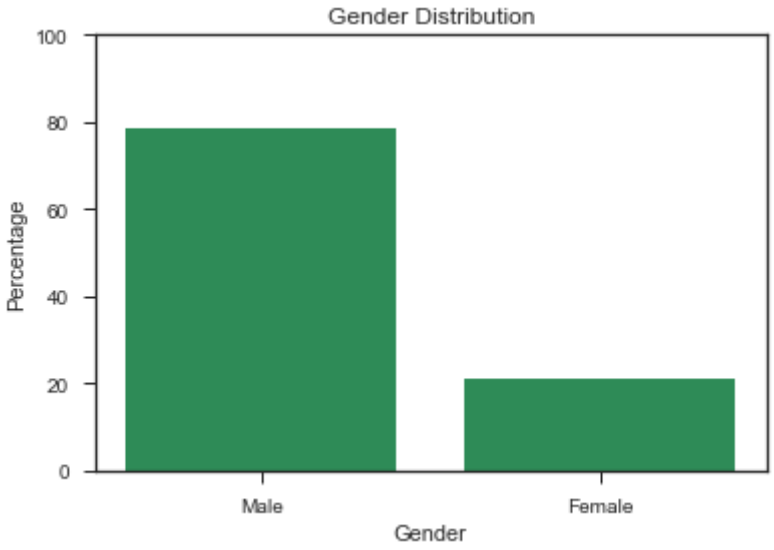
Patient Age Distribution

The average age of the patients in our dataset is between the range of 55-60 years old.

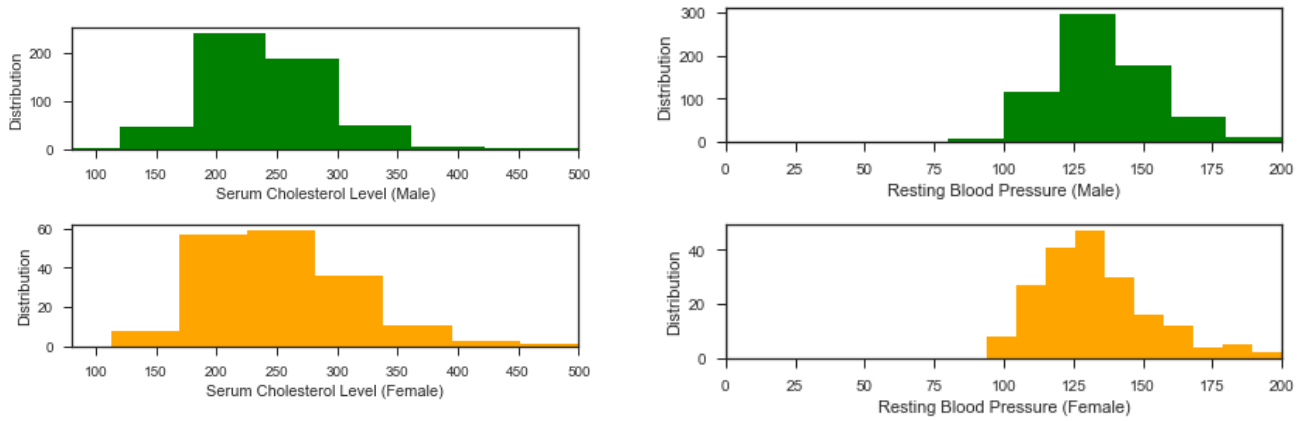
Analysis by Gender

Gender Distribution

The overwhelming percentage of male patients in this dataset is 78.8% as compared to the 21.2% of patients that are female.



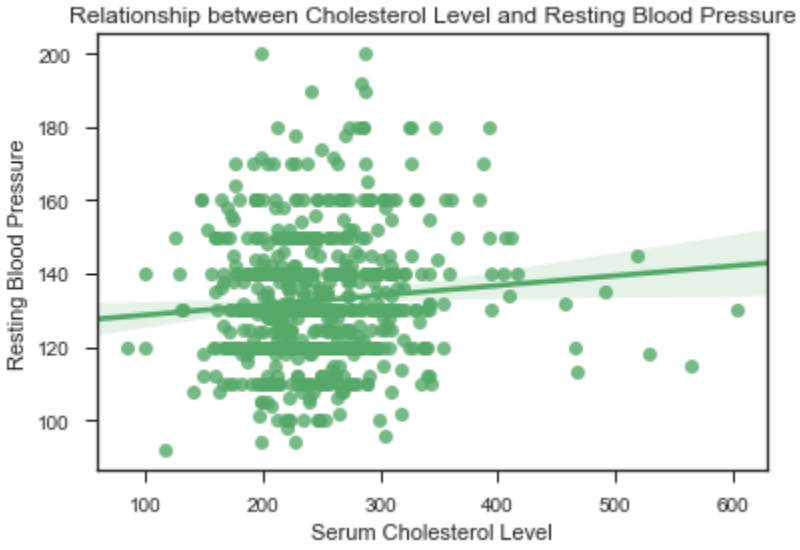
Resting Blood Pressure and Serum Cholesterol Levels by Gender



An analysis of the resting blood pressure of patients shows us that males and females exhibited similar resting blood pressures and serum cholesterol levels.

Analysis of Resting Blood Pressure and Serum Cholesterol Levels

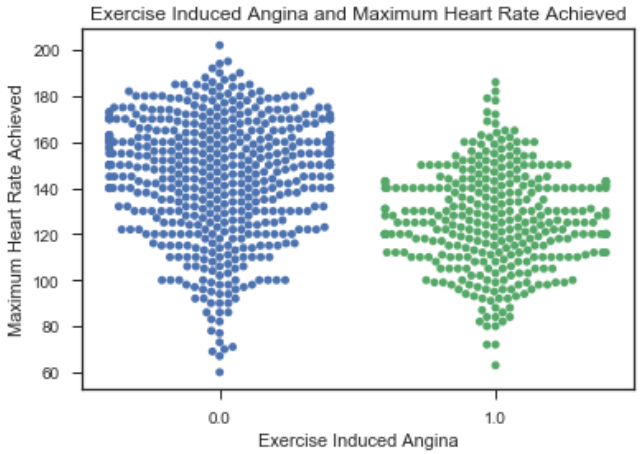
The mean resting blood pressure in our dataset is 132 mm Hg and the mean serum cholesterol level is 199.1 mg/dl.



We wanted to look further to see whether we could establish a relationship between a patient's resting blood pressure and serum cholesterol levels. We observe, based on the above plot, that there a slight positive relationship between both variables.

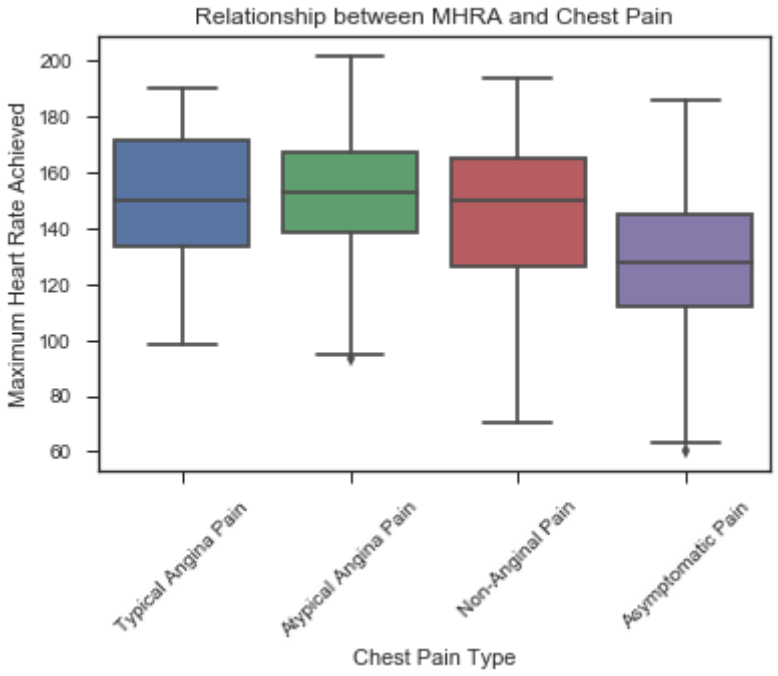
Analysis of Exercise Induced Angina and Maximum Heart Rate Achieved

Exercise Induced Angina (EIA) is pain caused in the chest region as a result of putting the patient under stress. We wanted to identify how the recorded Maximum Heart Rate Achieved (MHRA) during exercise changed with patients that suffered with EIA.



By studying the swarm plot above, we note that the ‘thickness’ of the swarm varies between whether a patient suffered with EIA or not. This means that, patients that suffered with EIA were less than those that did not. Also, observe the height of the swarm plot. The plot studying patients that did not suffer with EIA is ‘higher and thicker’ at the top, which tells us that the same patients were more likely to reach higher heart rates during exercise and were a lot more than the patients that did suffer with EIA. This makes sense, since the latter would not suffer chest pains during exercise and were able to put their bodies under more stress.

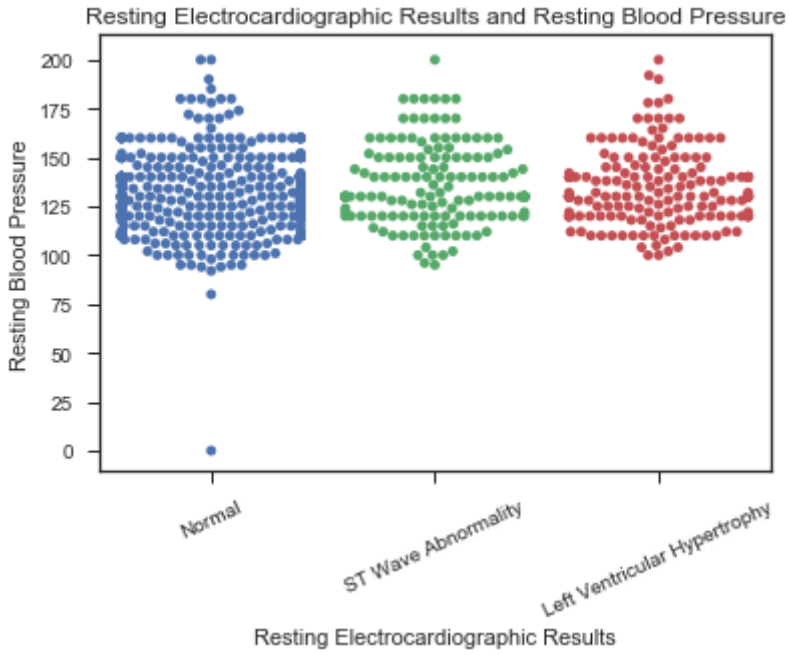
Analysis of Chest Pain and MHRA



We observe above that patients that suffered from Asymptomatic chest pain were less likely to reach higher heart rates when exercising. The remaining chest pain types seems to exhibit similar average heart rates.

Analysis of Electrocardiographic Results and Resting Blood Pressure

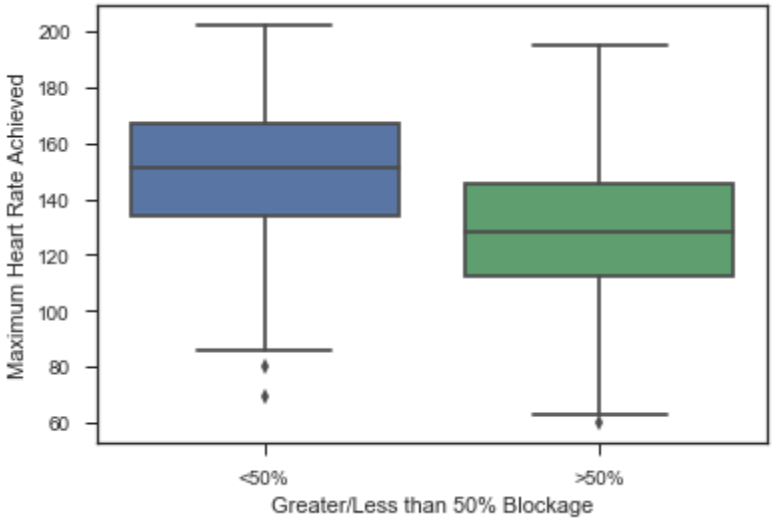
Electrocardiography is a method used to measure electrical impulses from the heart. We analyse these results and the respective patient's resting blood pressure.



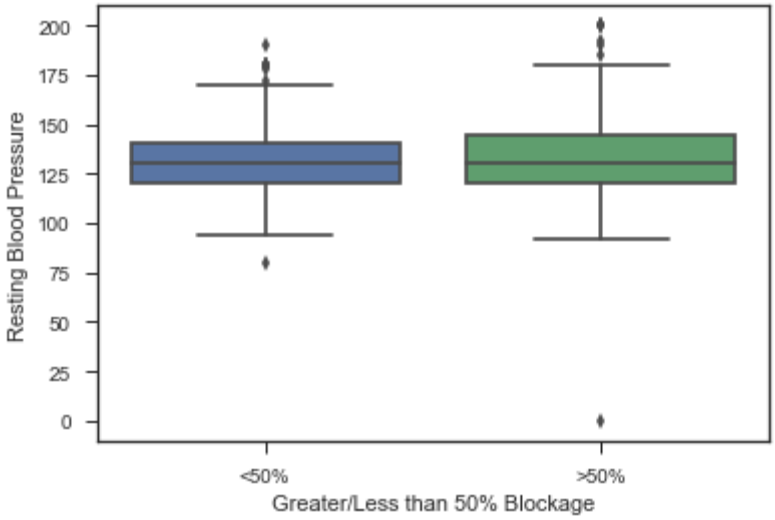
By studying the swarm plot above, we observe that most patients had normal electrocardiographic results. Also, we see that the height and distribution of the swarm plot for each criterion is fairly similar for both ST wave abnormality and left ventricular hypertrophy.

Analysis of the target variable (DHD) – the presence of >50% or <50% blockage in major coronary vessels

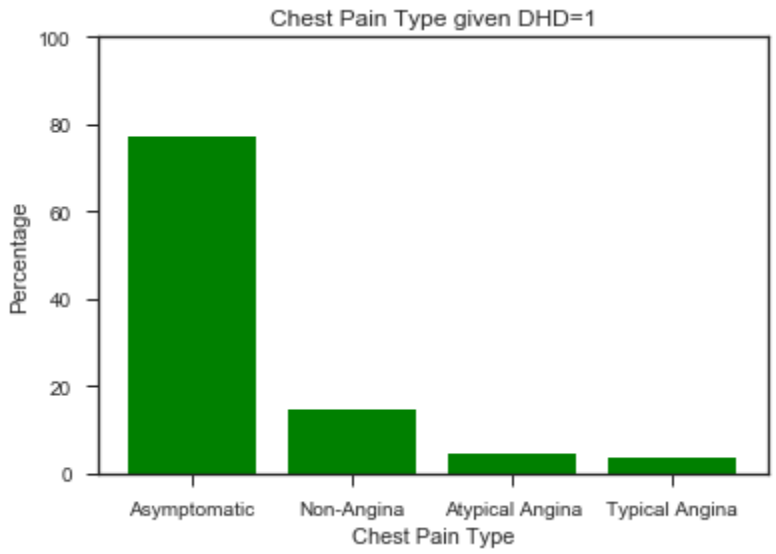
In our current dataset, 55% of patients have coronary vessels with more than 50% blockage.



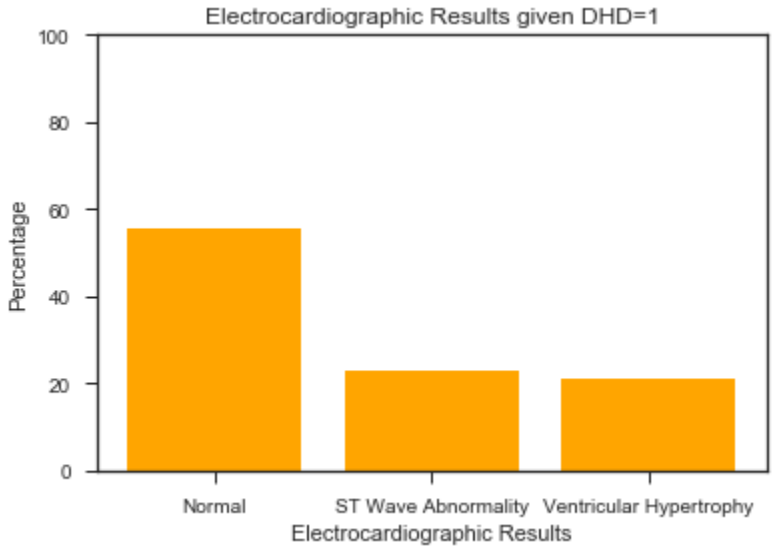
In our first plot, we see that patients with greater than 50% blockage were less likely to reach a MHRA when the patient's body was put under stress.



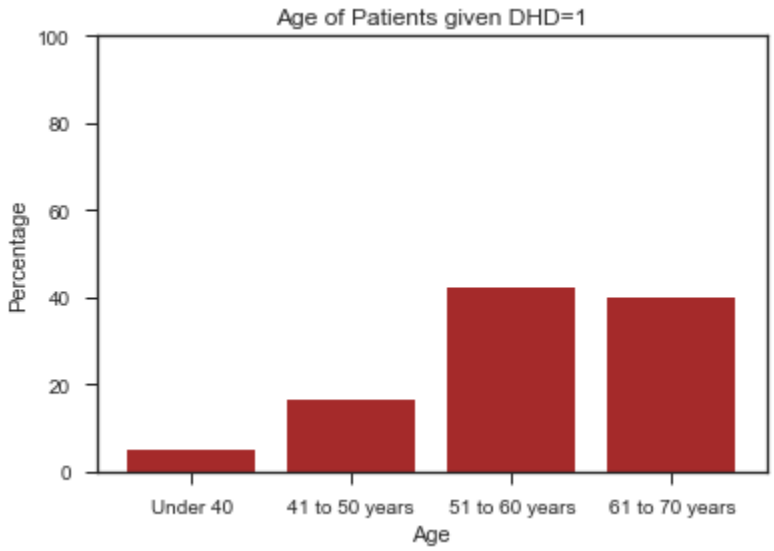
In this plot, we observe that patients with greater or less than 50% blockage were likely to have the same average blood pressure of approximately 130 to 135 mm Hg.



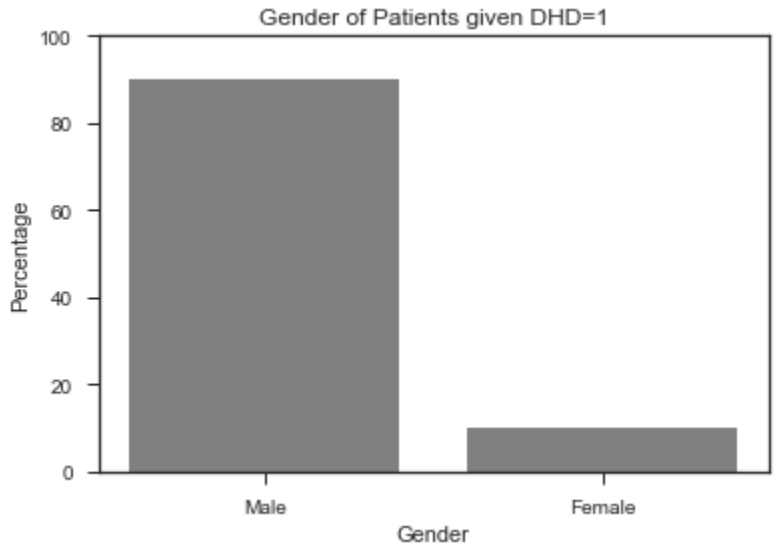
We separated the dataset such that we focused only on patients that suffer from a greater than 50% blocked coronary vessel (DHD=1). We then studied their respective chest pain types. According to the plot above, we observe that most patients suffered from an *asymptomatic* chest pains.



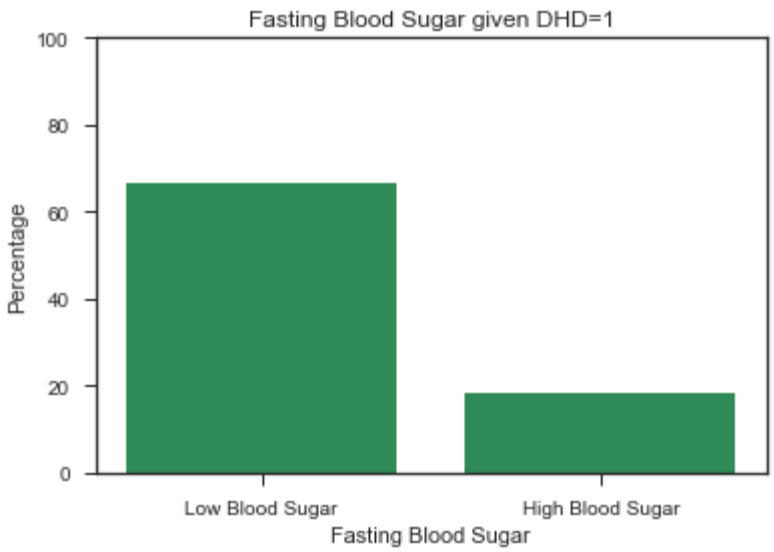
Here we observe that the overwhelming majority of patients have a normal electrocardiographic reading. This is quite alarming given that this is a an important indicator to the physician of any subtle abnormalities in the patient's heart function.



This plot provides insight into the age of the patients that suffer from a blocked coronary vessel. As shown in the plot, the majority of patients are between the ages of 51-70 years old.



The overwhelming majority of patients are male occupying more than 90% of the patient list. This is what we predicted, given that the original distribution of the data also showed a majority of male patients.



We observe that the patients are more likely to have a low blood sugar reading.

Now that we have had better graphical insight into the data and what it tells us about the relationship between variables as well as the characteristics of the patients that suffer from a a greater than 50% blocked vessel, we now start testing the data and challenge some of our hypothesis.

Part 3: Inferential Statistics

Grouping our data by our target variable and analysing means

We start by grouping the dataset by our target variable and study the means of our explanatory variables. It is worth noting that the majority of our explanatory variables are binary variables, so the interpretation of the mean is little less straightforward. However, for binary variables with '0' and '1' entries, the higher the mean, the more likely there were more '1' entries.

new_dhd	age	sex	cp	rbp	scl	fbs	rer	mhra	eia	std
0	50.542787	0.647922	2.767726	129.848329	227.735897	0.108861	0.545232	148.737789	0.141388	0.414433
1	55.936884	0.901381	3.648915	134.049145	176.667339	0.217593	0.651485	128.298729	0.595339	1.260086

We can observe that patients with greater than 50% blocked vessels were more likely to have higher resting blood pressures, however, the same patients averaged lower serum cholesterol levels. Also, the latter group of patients had a lower maximum heart rate reading, and were more likely to have exercise induced angina. Also, we note that these patients were older, with an average of 55 years in contrast to 50 years.

Two Sample t Tests

I separated the dataset into patients with above average and below average blood pressures. I wanted to study whether patients with different blood pressure levels were less likely to differ in their likelihood of whether they would have a greater than 50% or less than 50% blocked vessel. In other words, I hypothesized that both samples would have the same mean for the target variable.

test statistic	p-value
3.43	0.0006

This information provided tells us that we can confidently claim that the mean of the two samples are different and therefore, we can state that resting blood pressure is a significant factor to whether the patient has a major or minor vessel blockage.

Moreover, we proceeded to separate the dataset into two samples of patients based on their electrocardiographic results (normal and abnormal). Using the same test above we were provided with the following statistics.

test statistic	p-value
-2.98	0.002

In this case, we can also state, with confidence, that the means of the two samples with respect to whether the patient had major or minor vessel blockages were the same. Therefore, electrocardiographic were also significant in this case.

We proceed to test whether gender is related to a presence of a blockage in a coronary heart vessel.

test statistic	**p-value**
9.8	0

As in our previous exercises, we conclude that a patient's gender plays a role in whether they will have a major or minor blocked vessel.

The purpose of conducting this series of tests is for us to be able to state that the variables in our study add value to the analysis and will help us make accurate predictions and interpretations.

Analysis and Supervised Learning

Task

In this section, I will build machine learning models and fit the corresponding data in order to be able to predict whether a patient will suffer from a heart attack (given that more than 50% of their vessel is blocked) using my available explanatory variables (features).

Feature Selection

The features of our model are the explanatory variables that we will use in the model to predict our target variable. In this case, I plan on using all the explanatory variables as features. None of the variables are highly correlated with each other (as mentioned earlier) and each feature adds value to the analysis. Here is a correlation table for verification.

	age	sex	cp	rbp	scl	fbs	rer	mhra	eia	std	new_dhd
age	1.000000	0.058179	0.161971	0.240428	-0.091229	0.232110	0.208795	-0.366697	0.199167	0.254805	0.285509
sex	0.058179	1.000000	0.174879	0.001371	-0.197801	0.088660	-0.016069	-0.180025	0.182619	0.102933	0.308392
cp	0.161971	0.174879	1.000000	0.021998	-0.137328	0.046324	0.032497	-0.349508	0.416770	0.247330	0.473443
rbp	0.240428	0.001371	0.021998	1.000000	0.088187	0.159328	0.096457	-0.106844	0.151641	0.160970	0.109798
scl	-0.091229	-0.197801	-0.137328	0.088187	1.000000	0.025164	0.115322	0.235477	-0.036194	0.046726	-0.229004
fbs	0.232110	0.088660	0.046324	0.159328	0.025164	1.000000	0.128377	-0.054908	0.032062	0.052240	0.146087
rer	0.208795	-0.016069	0.032497	0.096457	0.115322	0.128377	1.000000	0.053739	0.032003	0.114411	0.065612
mhra	-0.366697	-0.180025	-0.349508	-0.106844	0.235477	-0.054908	0.053739	1.000000	-0.355075	-0.149286	-0.392371
eia	0.199167	0.182619	0.416770	0.151641	-0.036194	0.032062	0.032003	-0.355075	1.000000	0.392169	0.463133
std	0.254805	0.102933	0.247330	0.160970	0.046726	0.052240	0.114411	-0.149286	0.392169	1.000000	0.386574
new_dhd	0.285509	0.308392	0.473443	0.109798	-0.229004	0.146087	0.065612	-0.392371	0.463133	0.386574	1.000000

Model: Logistic Regression and Model Evaluation

In order to conduct this analysis, I will use a *logistic regression* approach. Logistic regressions are used when the target variable is a binary variable, as is the case in our analysis. Logistic regressions output results as probabilities where if the probability of an instant 'p' is greater than 0.5, the data will be labeled as 1.

In order for us to conduct supervised learning, we were required to drop NaN values in our dataset. We are left with a total of 737 observations.

I begin by splitting my data into a 70:30 (train:test) ratio. This means that I train my model on 70% of the data and use the remaining 30% to test my model's accuracy in prediction. I proceed to initiate a logistic regression, and then fit the training data. I test the model by predicting the test data. Below is a table with the analysis coefficients, standard errors and odds ratio.

Variable	Coefficient	Standard Error	Odds Ratio
Age	0.003	9.394	1.002
Sex	**1.108**	**0.425**	**3.023**
Chest Pain	**0.542**	**0.935**	**1.720**
Resting BP	0.003	18.597	1.003
Serum Cholesterol	-0.002	93.694	0.998
Fasting Blood Sugar	**0.404**	**0.356**	**1.499**
Resting Electrocardiographic Results	**0.277**	**0.839**	**1.32**
Maximum Heart Rate	-0.02	25.830	0.981
Exercise Induced Angina	**1.2**	**0.49**	**3.312**
ST Depression	0.701	1.08	2.018

It is important to note that the coefficients of a logistic regression are difficult to interpret in their original form. In order to interpret them, we derive the *odds ratio* from them and this can be interpreted. For example, we can say that a patient suffering from exercise induced angina is 3.3 times more likely to suffer from a heart attack.

In the table above, I have highlighted the important features in the analysis. This means that they have a significant impact on the target variable (high odds ratio) and a smaller standard error (more significant).

There are several ways to evaluate the accuracy or 'goodness' of a model. In this report, I will use a *confusion matrix, classification report, mean squared error, ROC Curve, AUC and R score,* for the purpose of evaluation.

Confusion Matrix and Classification Report

A standard confusion matrix gives us the information below:

True Positive	**False Negative**
False Positive	**True Negative**

The most important cells to observe here is the 'True Positive' and 'True Negative'. In the context of our analysis, a 'True Positive' is the number of times the model predicted that a patient will have a blocked coronary vessel. A 'True Negative' is the number of times the model predicts that the patient will not have a blocked coronary vessel. This is the confusion matrix from our analysis:

77	20
26	**99**

It is fair to say that our model performed well with the predictions, with the true positive rate (tpr) and true negative rate (tnr) equal to 79%.

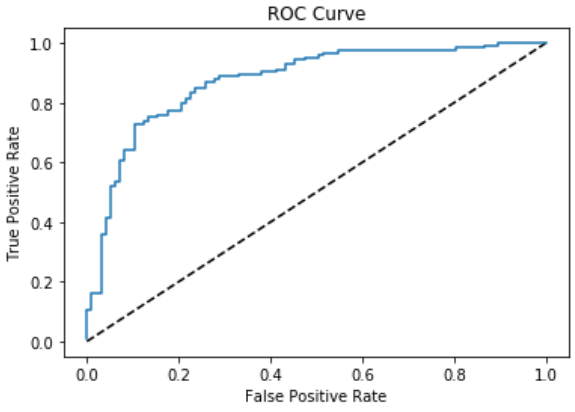
A classification report provides similar information to the confusion matrix; however, it is presented in a different format.

	Precision	**Recall**	**F1-Score**
0	0.75	0.79	97
1	0.83	0.79	125
Average	**0.8**	**0.79**	222

As we can see from the table above, a high precision and high recall rate is favourable. High precision rate can be considered as a true positive and recall rate as a true negative.

ROC Curve

We proceed to plot the ROC (Receiving Operator Characteristic) curve which informs us of the ratio of the true positive rate and the false positive rate. The higher the curve, and the more area beneath it, the more favourable our model is.



AUC and Mean Squared Error

We produce an AUC score by five-fold cross validation. This means that we portion our data into five units and choose a training set and test set and run our analysis, and repeat the same analysis with five different sets. The higher the AUC score, the more favourable is our model. The mean squared error (MSE) is a measure of noise or variation in the data. The higher the noise, the higher the MSE and the less favourable the results. The AUC results:

AUC Score	0.871	0.831	0.854	0.817	0.851

Our AUC score has produced favorable results over the five-fold cross validation.

Mean Squared Error of Training Set	Mean Squared Error of Test Set
0.188	0.207

In both cases we note an extremely small MSE, although the 'noise' does increase from the training set to the test set.

Model: Random Forest and Model Evaluation

A random forest model is a supervised learning algorithm that builds multiple decision trees and averages their results together in order to get a more accurate and stable prediction. Additional randomness is created within the model and this often leads to better results. A beneficial aspect of this model is its ability to highlight each feature's importance to the analysis. This will give us insight into which features contribute the most to our target variable.

We fit the model using the random forest regressor. An R score is used to determine how our training and test sets explain variability in the model.

R Score on Training Set	R Score on Test Set
0.88	0.3

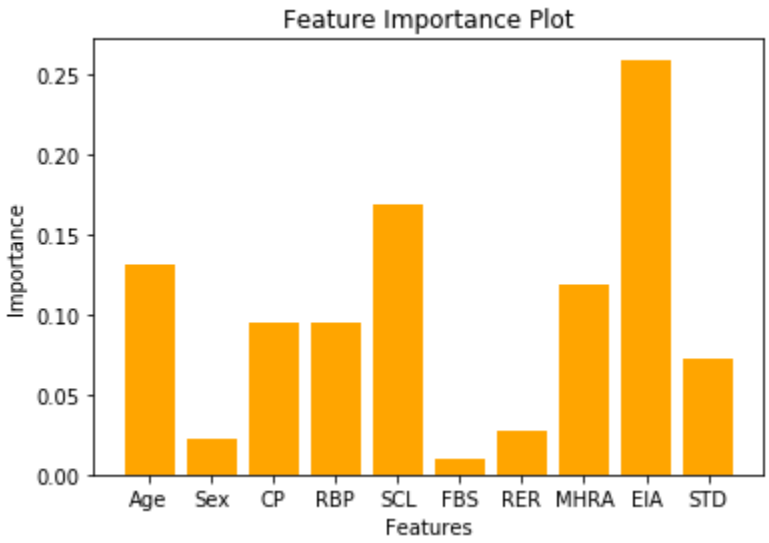
This tells us that our training set performed very well in explaining the variability in our model, however, the test set did not do well in comparison.

Now we will study the MSE and compare them to that of the logistic regression.

MSE on the Training Set	MSE on the Test Set
0.172	0.03

In comparison to the logistic model MSE, we observe a lower MSE for both our training and test sets when using a random forest regressor. The noise decreased from the training set to the test set. This is favourable, and shows us the model is a lot more stable in its functionality and prediction.

We conclude this section with some insight into feature importance, and this is displayed in the graph below.



We observe from the graph above that the three standout features in terms of importance are: Age, Serum Cholesterol Level and Exercise Induced Angina. Following the latter: Maximum Heart Rate Achieved, Resting Blood Pressure and Chest Pain were also very important features in determining whether a patient is likely to suffer from a heart attack.

Overall, the random forest regression performed poorly on the test data in terms of its ability to explain its variation. However, we observed smaller MSE for both the training and test sets in comparison to the logistic model we built earlier. This tells us the model is more stable in its predictions, and does not have a lot of noise.

Key Findings

When we initiated this study, we wanted to identify which features was an important predictor in whether a patient was likely to suffer from a heart attack. Using the analysis from both the random forest model we built earlier, we can state that the patient's *age, serum cholesterol levels, exercise induced angina* played major roles in determining whether a patient is likely to have a heart attack. The logistic model predicted that *sex, chest pain, fasting blood sugar, electrocardiographic results and exercise induced angina* were important features. The common feature that stood out in both analyses was *exercise induced angina*.

Limitations and Considerations

Our analysis provided valuable results, however, it is important to consider certain limitations that might have impacted our analysis. Firstly, we did not have many observations (737 observations). This greatly reduces the variation in our data and therefore does not encourage us to be confident in the accuracy of the predictions and results. Secondly, our dataset had many missing values. Some of the variables needed to be dropped due to the fact that 50% of their data was missing. This also impacts the accuracy and power of analysis. Moreover, we did not have enough data on women that suffered from a blocked coronary vessel. There was a significant ratio between both genders (80 male:20 female) and this will bias our results. This means that, we cannot attribute the conclusions of this analysis to the average female, and this is a major limitation in our analysis. Furthermore, the data is cross sectional and is approximately 30 years old. Since then, a lot has changed in terms of lifestyle choices, health awareness and general nutrition and this might not be taken into consideration in the analysis.

It is important that we interpret low MSE results of the logistic regressions carefully. Logarithmic formulae by nature reduces the noise in the analysis and brings extreme points closer towards its center. Although, initially, it may seem that the model is accurate and stable based on the latter (which could be true), it is important to consider the nature of the model when making interpretations.

Recommendations

In order to improve this analysis further, it would be beneficial to add more features to our model, and add more observations in order to add variation and power to our analysis. Additionally, it could serve the analysis well to transform the data into a longitudinal study as this will capture trends and changes in our data. Also, more recent data is needed in order to make this analysis more applicable to the contemporary health sector.

Conclusion

In this report, we began by identifying the problem, the client and the objective of the analysis. Moreover, we began by shedding light on the data cleansing and wrangling process in order to develop a fully functional and working dataset. Additionally, we presented different plots of variables and their respective relationships from the data and drew conclusions from them. Furthermore, we built two supervised learning machine models that helped construct, train, test and predict the target variable using the explanatory variables (features). We then proceeded to provide key insights and findings from our analysis, suggested improvements based on the limitations of our dataset and made recommendations for our client to consider if they were to proceed with this model as a standard for their patient intake. Medical professionals can use the analysis provided from this data and develop a system that caters to patients that need urgent care. This will amend the misallocation of resources that the health sector persistently struggles with.