Paper Title\* (use style: paper title)

\*Note: Sub-titles are not captured in Xplore and should not be used

Sammak Musabbir Hasan   
*School of Computer Sciences*  
*Universiti Sains Malaysia*Penang, Malaysia  
sammak@student.usm.my

Lee Jing Wen  
*School of Computer Sciences*  
*Universiti Sains Malaysia*Penang, Malaysia  
leejingw@student.usm.my Wang Huaixu  
*School of Computer Sciences*  
*Universiti Sains Malaysia*Penang, Malaysia  
huaixu.wang@student.usm.my

Lee Kar Choon  
*School of Computer Sciences*  
*Universiti Sains Malaysia*Penang, Malaysia  
karchoon\_lee@student.usm.my

*Abstract*—With the increase and successful implementation of machine learning in various predictive tasks in the real world and cardiovascular disease being one of the leading causes of death worldwide, attracting the interest in the medical field in applying the same technique on available data. By using openly available software and public domain data, data exploring, data manipulation, machine learning techniques implementation and evaluation will be done to create a predictive model on the risk of developing cardiovascular disease in patients. The process done using open-source tools R will be provided in this paper. Support Vector Machine is the best performing machine learning algorithm among the four algorithms selected with an accuracy of 73.33%.

Keywords—cardiovascular disease (CVD), machine learning, R,

# Introduction

Cardiovascular disease or in-short CVD is a type of diseases with the involvement of the heart or blood vessels. CVDs includes a wide variety of types such as myocardial infarction (heart attack), stroke, abnormal heart rhythms, stroke etc. The cause of CVD varies according to the disease, in general the main causes are diabetes, high blood cholesterol, high blood pressure, excessive alcohol consumption, smoking and physical inactivity. It was being said that, 80% of CVD deaths for males and 75% of females are accounted by coronary artery disease and stroke [1]. Cardiovascular diseases are also one of the leading causes of death globally [1].

According to the Deputy Health Minister Dr Lee Boon Chye, for 13 years from 2005 to 2017, cardiovascular disease (CVD) remains to be the leading cause of death among Malaysians [2]. The issue is that CVD is expected to increase in Malaysia in the near future due to the increase of Malaysian aged 65 in the population to 14.5% of the total population [2]. With a 54% increase of mortality rate due to heart disease over 10 years amounting to 13,503 deaths compared to 8,776 in 2007, the future prospect is indeed worrying [2].

Currently, screening is the most popular way in CVD identification. But there is a catch, screenings such as ECGs, myocardial perfusion imaging, cardiac stress testing and echocardiography are not recommended to be done among those with no CVD symptoms or at low risk [3][4]. With assumptions stated, it will be too late to detect CVD. Additional to that, biomarkers can be used to predict the risk of future CVD, but the biomarkers result are controversial [5]. The present cardiovascular disease detection in the medical field is yet mature enough and can be costly and time consuming in undergoing the test procedure.

Hence, there is a need to come up with a system to early detect or identify CVD among people using general medical data so that early treatment and preventive measures can be done in an efficient manner. With the advancement of technology, computational power, storage and memory improved drastically. Additional to that, statistical algorithms for machine learning is developing and substantial amount of medical record data is available. By using computers to undertake machine learning on the data we have, reliable predictions on CVD can be made.

# Background and literature review

Before the advent of ubiquitous application of machine learning and other modern data science methods, it had been long practiced according to the belief that the risk of cardiovascular diseases is based on linear relationship with countable factors[6], which was based on the limitations of data collection and prediction tools and has been already proved to be biased[8].

Some attempts [7] use biomarkers of large cohort that are difficult to collect and the interpretation of which is restricted to certain professionals. This kind of methods has been expected to be at least partly replaced by more simple and easier prediction models like those that are based on more understandable and available attributes such like age, blood pressure and alcoholism [9].

Muthuvel etc [9] summarized the researches of heart disease prediction using Machine Learning and other data analytics approaches. It is seen that in recent 5 years the main paradigm of research of this problem has been shifted to common attributes-based as mentioned beforehand.

The techniques in use are common ones like Multiple Linear Regression and Logistic Regression, Decision Tree, Naïve Bayes, Support Vector Machine, Artificial Neural Network (ANN). The combination of at least techniques enhances the accuracy from some 60% (the case of J48, a decision tree method) to more than 80%, for example [10]. However, the sensitivity (recall) is generally much higher than the precision (positive predictive value). This is typical in [8] where sensitivity is near 70% after improvement of techniques but precision is still lower than 20%. It is worth being noticed that the attributes used in recent years’ researches usually include medical diagnostic attributes such like electrocardiogram, serum and lipid contents, heart beats which are available after some instrumental diagnosis and key factors in early stages of diagnosing heart diseases. These more professional attributes indicate the common practice of leveraging the prediction of risks based on “better safe than sorry” principle and this results in high false negative rate in machine learning.

# data analysis and interpretation

## Source of the dataset

The dataset is obtained from Kaggle [12] which consists a total of 70000 record of patient’s data. In this dataset, there is a total of 11 features which can be categorized into 3 types of input features, Objective, Examination and Subjective. Objective type is based on factual information, Examination type is from the medical examination results and Subjective type is information obtained from patient.

## Identification of attribute

There are two more attributes created in this project, they are BMI which was calculated using attributes height and weight and BloodPressure which uses systolic and diastolic blood pressure to categorize into lower than normal (-1), normal (0) and higher than normal (1).

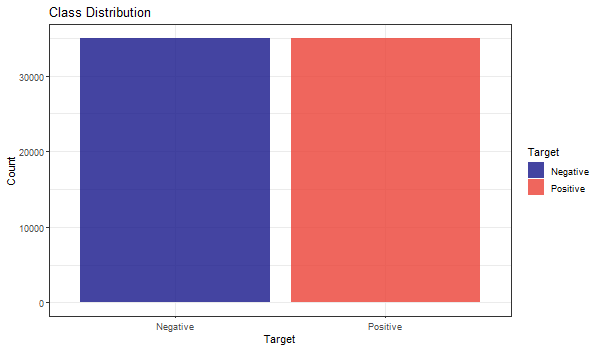
|  |  |  |
| --- | --- | --- |
| **Systolic blood pressure** | **Diastolic blood pressure** | **BloodPressure** |
| More or equal to 140 | More or equal to 90 | 1 |
| Less or equal to 90 | Less or equal to 60 | -1 |
| More than 90 and less than 140 | More than 60 and less than 90 | 0 |

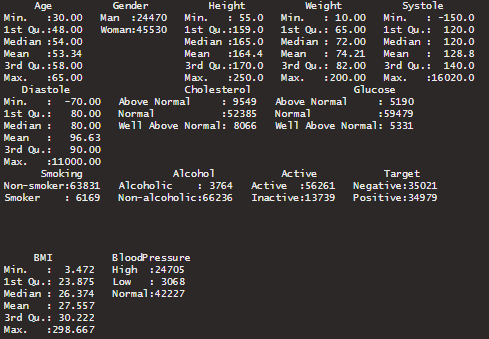
The summarized data description is stated at the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Attribute Name** | **Types of features** | **Data Type** |
| age | Age | Input, Objective feature | Integer (days) |
| gender | Gender | Input, Objective feature | Categorical code |
| height | Height | Input, Objective feature | Integer (cm) |
| weight | Weight | Input, Objective feature | Float (kg) |
| ap\_hi | Systolic blood pressure | Input, Examination feature | Integer |
| ap\_lo | Diastolic blood pressure | Input, Examination feature | Integer |
| cholesterol | Cholesterol | Input, Examination feature | 1: normal  2: above normal  3: well above normal |
| gluc | Glucose | Input, Examination feature | 1: normal  2: above normal  3: well above normal |
| smoke | Smoking | Input, Subjective feature | Binary |
| alco | Alcohol intake | Input, Subjective feature | Binary |
| active | Physical activity | Input, Subjective feature | Binary |
| cardio | Presence or absence of cardiovascular disease | Target Variable | Binary |
| BMI | BMI | Input, Objective feature | Float |
| BloodPressure | Blood Pressure | Input, Examination feature | Integer |

# Data exploring

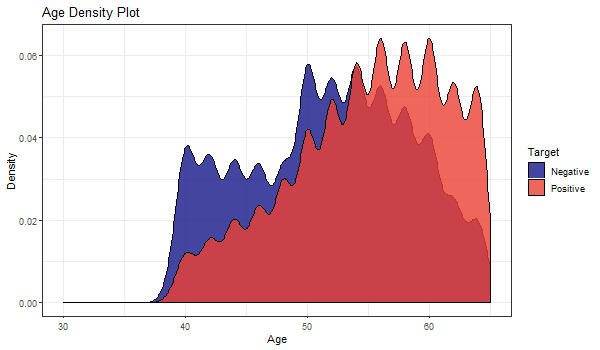
The class distribution of the target attribute, cardio is balanced as seen in the figure below:

****



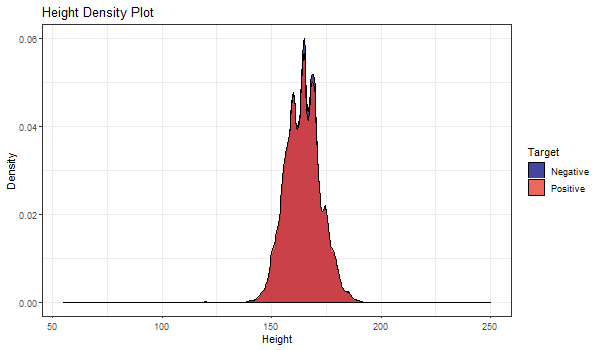
## Density plot

### Age Density Plot

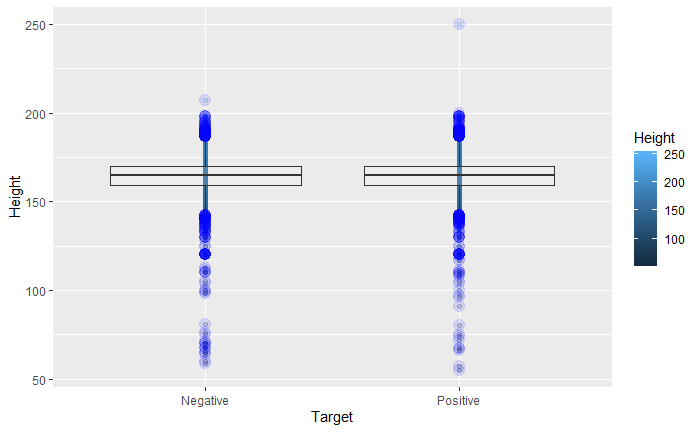


The peak density of the distribution is above 0.06 at the age range of 57 to 60. The distribution displayed multimodality characteristics with multiple peaks. It can be determined that the patients with cardiovascular diseases are more present in high age group.

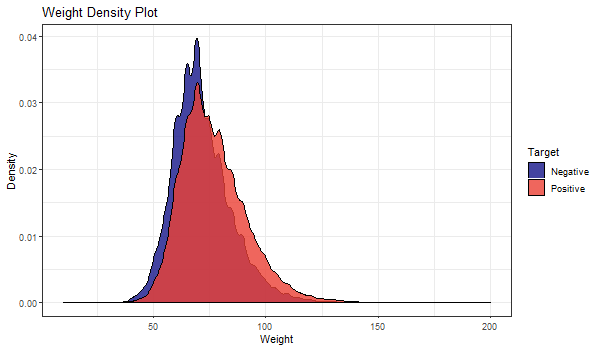
### Height Density Plot



The peak density of the distribution is at the height of 164cm. The distribution displayed multimodality characteristics with multiple peaks. There seems to be no trend in identifying whether patients have cardiovascular diseases in the height attribute, since the distribution are the same for both classes. Outliers are present in the Height data.

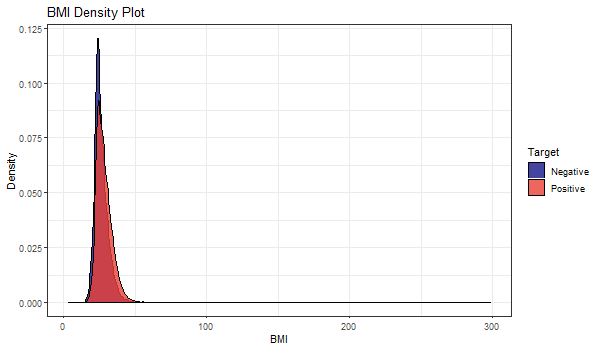


### Weight Density Plot



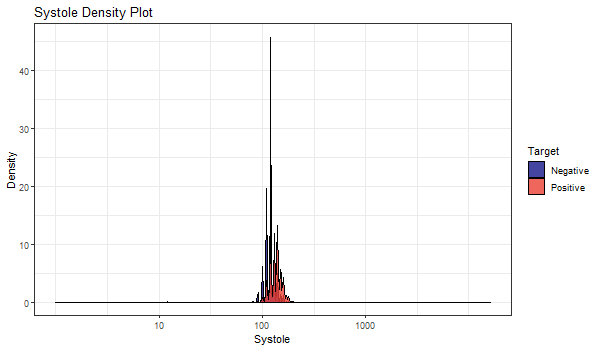
Peak weight density of the distribution can be seen to be higher in the data where patients do not have cardiovascular disease. The distribution displayed multimodality characteristics with multiple peaks. As the weight increases starting from around 72kg, there is a higher chance the patient has cardiovascular disease. The mean weight of patients with CVD was slightly higher than those without. There are a few outliers in the weight attribute.

### BMI Density Plot



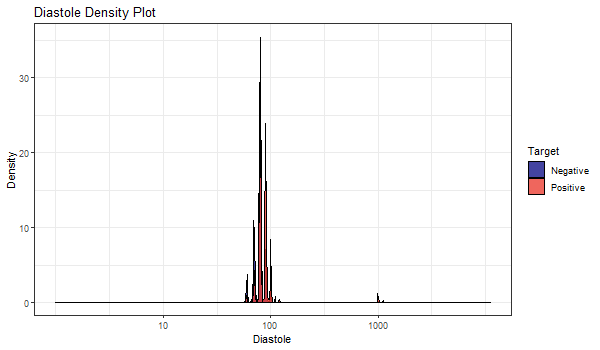
Since BMI is calculated using attribute weight and height, we can see that the BMI distribution shows unimodality characteristic. The peak density of the distribution is above 0.11 where patients do not have cardiovascular disease. It can be determined that patients with cardiovascular diseases are more present when BMI is at 28 onwards. The mean of distribution where patients has CVD was slightly higher compared to patients without CVD. Since BMI is calculated from Height and Weight where outliers are present in both attributes, leading the BMI data to have outliers too.

### Systole Density Plot



With a glance, Systole featured a lot of outliers and seems to have invalid values since the x-axis covers a wide range which is out of the normal medical range. The mean of Systolic blood pressure can be seen to be slightly higher in CVD affected patients.

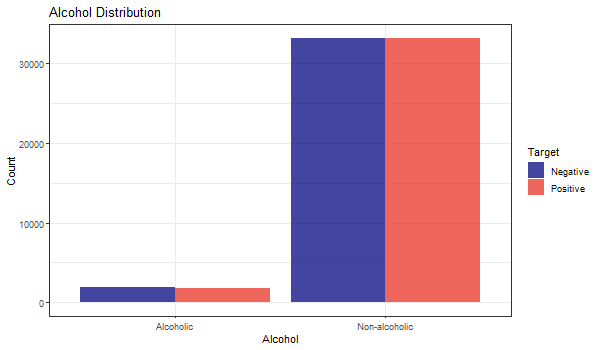
### Diastole Density Plot



Attribute Diastole has the same trend and characteristic found in Systole and should be taken care of.

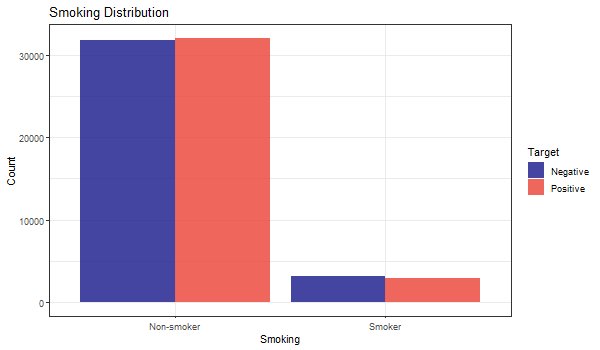
## Bar Plot

### Alcohol Distribution



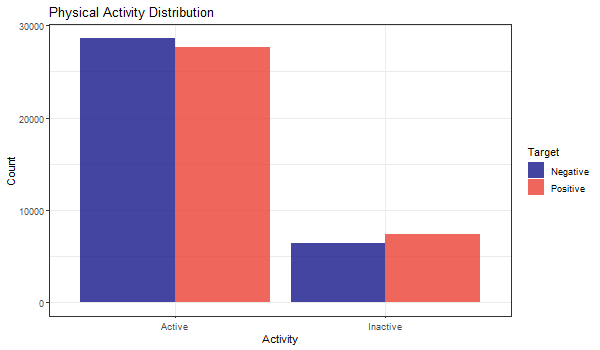
The proportion of having cardiovascular diseases is not seen to be positively correlated to alcoholism as both groups have the proportion of negatives higher, but the contract is not pronounced.

### Smoking Distribution



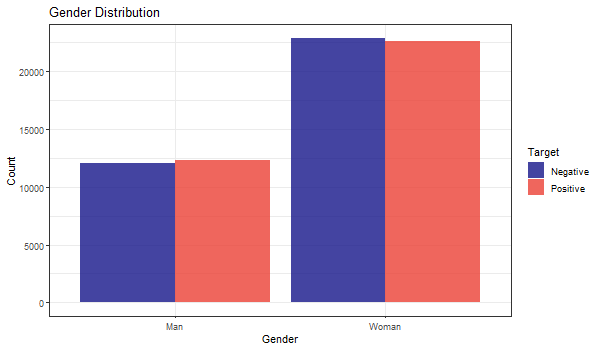
Like Alcohol Distribution, it seems hard to put smoking as a powerful indicator due to same non-obvious distribution.

### Physical Activeness Distribution



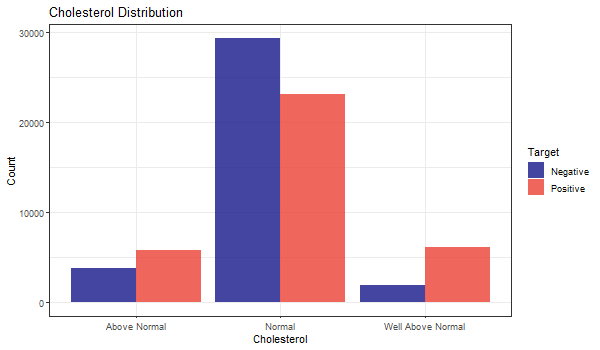
There is a trend that people who are not active in physical activity are more prone to have cardiovascular diseases.

### Gender Distribution



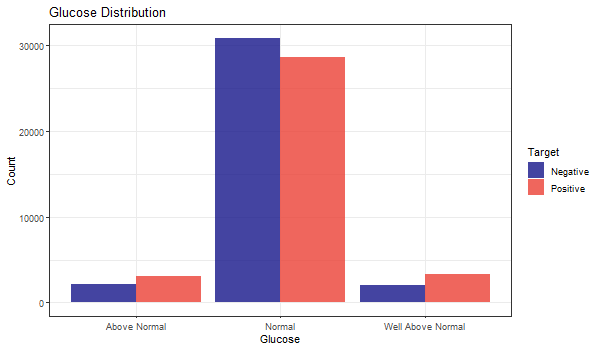
CVD were slightly more prevalent in male patients than their counterpart.

### Cholesterol Distribution



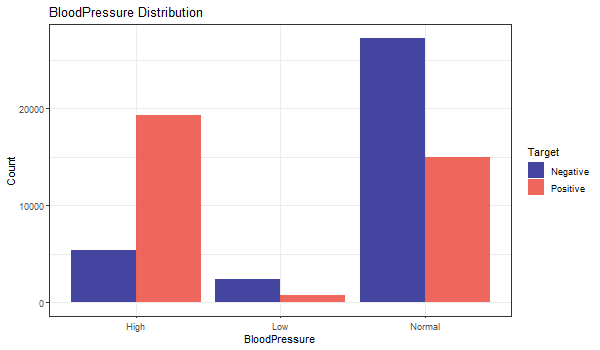
Cholesterol is seen as an obvious attribute and the well above normal category level is positively correlated to cardiovascular diseases.

### Glucose Distribution



Glucose is seen as an obvious attribute as well where levels above normal has larger among of patients having cardiovascular diseases.

### Blood Pressure Distribution



The target is strongly correlated with Blood Pressure as those have lower-than-normal value have very low percentage being positive and higher-than-normal instances showing patients having a very high rate for having CVD.

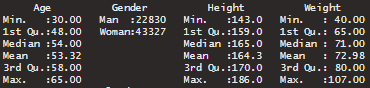
# Data preparation

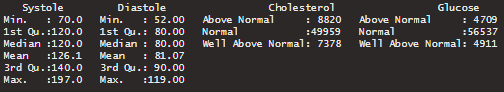
## Handling missing or null data points

There are no missing values in the data, hence further data processing for this is not required.

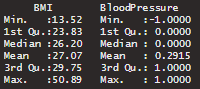
## Outliers

As seen in the previous R summary statement and density plots, there are outliers in some attribute such as Systole (ap\_hi) and Diastole (ap\_lo) that are negative in values which is impossible, weight attribute which has a minimum weight of 10kg and maximum value of 200kg and height attribute with a maximum of 250cm and minimum of 55cm which does not fit in the normal range. The outliers are handled by only retaining in 5 times of the Inter Quartile Range (IQR), since 1.5 times IQR removes too many observations that we deem to be useful. With the updated R summary shown below as comparison, the attributes which has outliers like Height, Weight, Systole and Diastole looks way better after the outliers are handled:





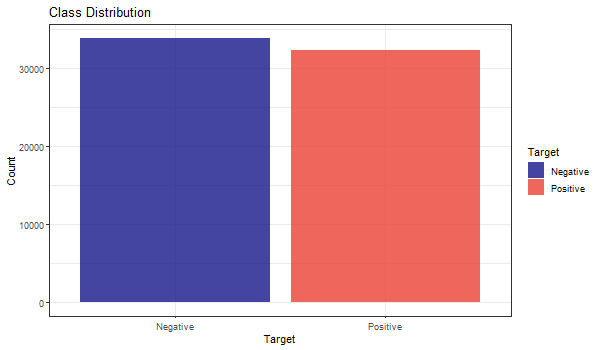




## Z-score normalization

Z-score also known as zero-mean is the conversion of values to a common scale where the average is zero with a standard deviation of one. This process is very useful if the data has outliers which is what was found in our data. By computing, the value of A in this case, , is normalized to z. Formula of z-score is shown below, where and are the mean and standard deviation of the attribute respectively:

The difference between Negative count and Positive count in the Target class increased a little after the outlier removal process, but it is insignificant in impacting the modelling process.

****

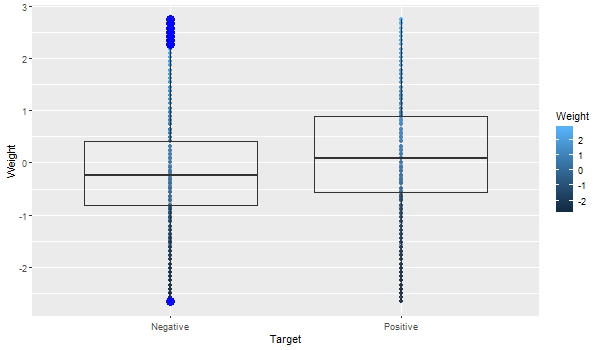
# feature selection

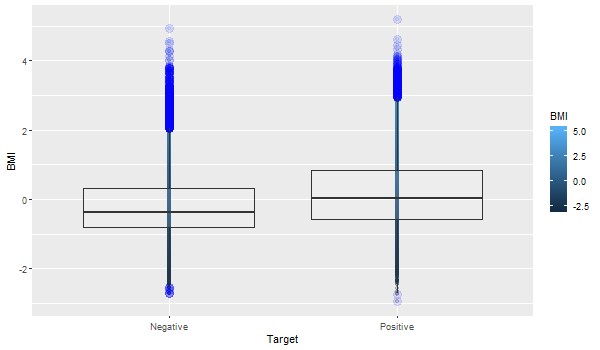
Most of the feature selection methods can be categorized in two categories, which are, wrapper methods and filter methods[11]. Wrapper methods evaluate a model by plugging different sets of features in order to find out the optimal subset for which the performance is maximum. Wrapper methods are indeed search algorithms that take features as inputs and output the optimal subset of features. There are various wrapper methods available, for example, recursive feature elimination, genetic algorithms, simulated annealing etc. On the other hand, filter methods find out the relevance of the features before modelling the data and models the data subsequently only with important features. In other words, only features with important relationship are retained for training.

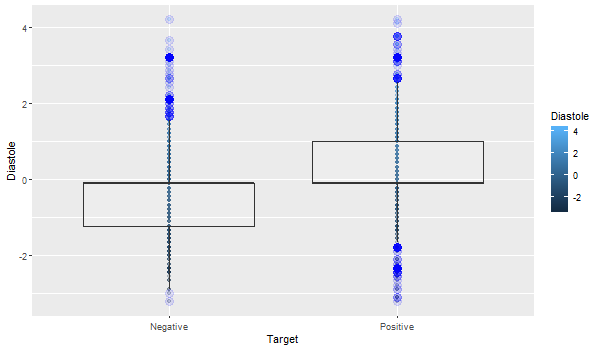
However, there are both advantages and disadvantages of both methods. Filter methods are less computationally demanding task than its counterpart, but it does not directly justify the performance of the model. As this method evaluates each feature separately, important interactions between features is not quantified. In contrast, wrapper methods are computationally intensive, but there is no risk of overfitting.

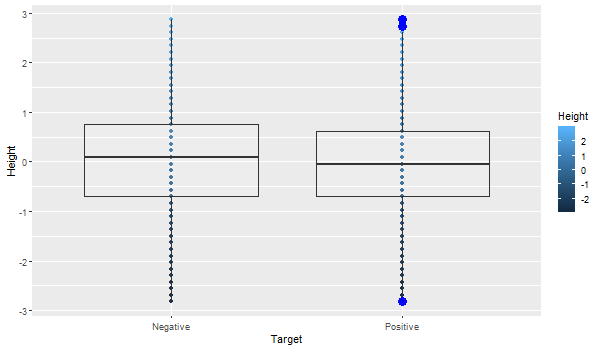
In this study, a wrapper method Recursive Feature Elimination (RFE) is applied because during the exploratory analysis no feature was found to have significant predictive power over the target. So we trained each model with different subsets of feature to identify the most effective ones.

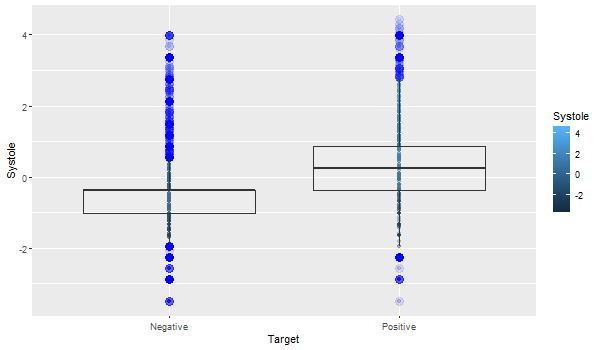
## Box plot

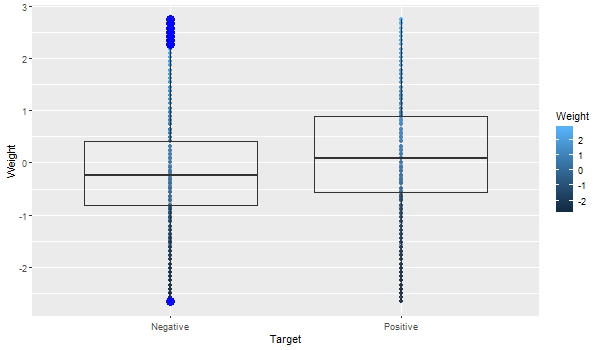










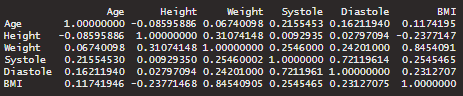


## Checking Near Zero Variance Attributes

Sometimes features may only have a single unique value. For many models, this may cause the model to crash or the fit to be unstable. Similarly, features may have only a few unique values that occur with very low frequencies. The concern here that these predictors may become zero-variance predictors when the data are split into cross-validation/bootstrap sub-samples or that a few samples may have an undue influence on the model. These “near-zero-variance” predictors may need to be identified and eliminated prior to modeling. However, no features had near zero variance.

## Correlation matrix

Some models might show improved performance if the level of correlation between the predictors is reduced. Only BMI was found to have a strong correlation with Weight, which is obvious.



## Checking Linearly Dependent Features

No features were found to be linearly dependent among each other.

## Recursive feature selection

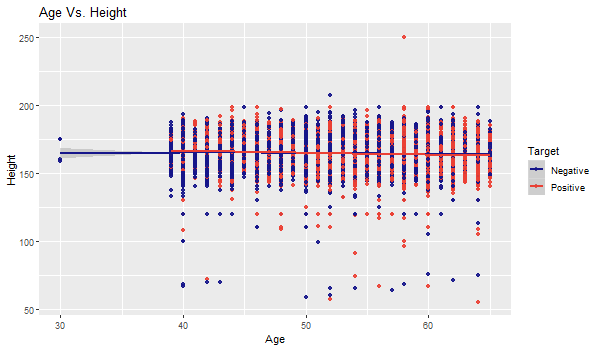
Recursive feature selection using Naïve Bayes where 15% of the dataset was used and resampling over the data is performed by using 5 folds cross-validation method. The result we obtained are as below:

|  |  |
| --- | --- |
| **Rank by importance** | **Attribute** |
| 1 | Systole |
| 2 | Diastole |
| 3 | BloodPressure |
| 4 | Age |
| 5 | BMI |
| 6 | Weight |
| 7 | Cholesterol |
| 8 | Active |
| 9 | Height |
| 10 | Smoking |
| 11 | Gender |
| 12 | Glucose |
| 13 | Alcohol |

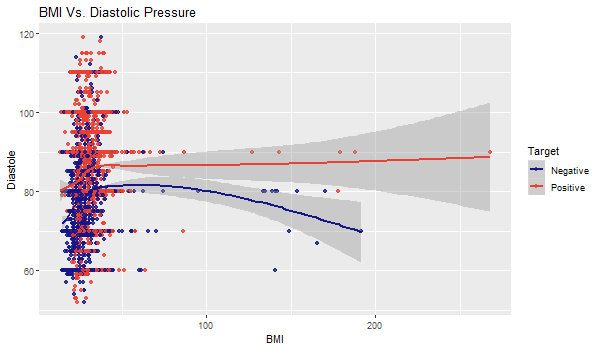
## Regression Analysis

The study on relationship between selected dependent attributes with the independent attribute are done using regression analysis.

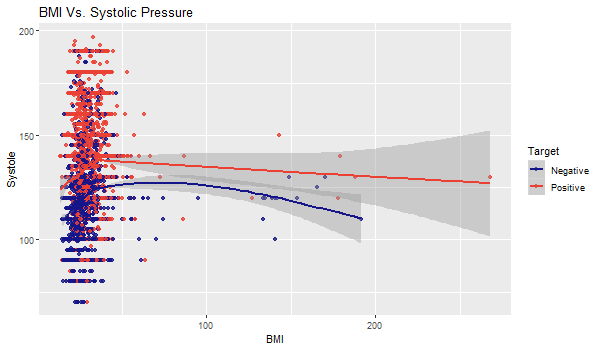
### Age versus Height



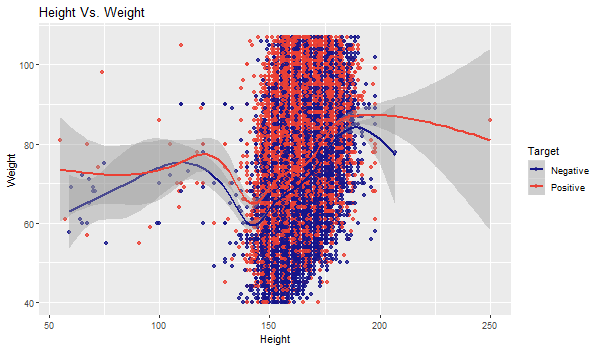
### BMI versus Diastolic Pressure



### BMI versus Systolic Pressure



### Height versus Weight



Base on the chart, sampling data spread concentrating on the range from 120 to 200, which proves that the data is correct according to the standard height of human beings.

The positive results also spread the same range of height and this proves that height is not a key factor to CVD. However, the positive tested results spread as much as possible at higher weight sampling data. We believed that Weight is the one of the key factors to Cardiovascular Disease.

It was decided that the top 8 attributes from recursive feature selection would be used in this project; Systole, Diastole, BloodPressure, Age, BMI, Weight, Cholesterol and Active.

# Machine learning

## Cross-validation

This statistical technique is implemented on our data where data is partitioned into subsets, with a subset used for model training and other subset used to evaluate the performance of the model. Cross-validation will aid in reducing variance and avoid model overfitting. Useful for small dataset like what we are using in this project.

## Model Selection

Our aim is to predict whether a person is at risk of developing cardiovascular disease, which is a classification problem. Because we are classifying patients into two groups, that is positive and negative, this problem is a binary classification problem. There are many algorithms available for binary classification problems. For example, Naïve Bayes, Decision Tree, Logistic Regression, Support Vector Machine, etc. For this project, we chose to use all the aforementioned algorithms to select the best performing algorithm.

* Logistic Regression

Logistic regression is a parametric model due to it having a finite set of parameters. Particularly, parameters are the regression coefficients. It is appropriate since relationship between dependent binary variable and one or more independent variable no matter it is nominal, interval, ordinal or ratio-level can be done.

* Decision Tree

Decision Tree or in short DT is a widely used non-parametric model. DT uses a tree-like model with branches, leaf nodes and root node to implement conditional control statements.

* Support Vector Machine (SVM)

SVM is a non-parametric model and makes less assumptions about the data. For this reason, even if the real-world data do not follow the training data distributions in future, it will still give a fair result.

* Naïve Bayes

In contrast to SVM and DT, Naïve Bayes is a parametric model and has several assumptions about the data, for example, it assumes that the features are independent of each other.

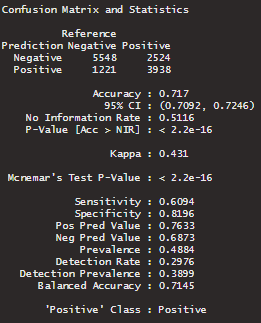
## Model Evaluation

For model evaluation, we, first, established the null model, which is the lower bound of the model. As it is a classification problem, we selected null model to be the most common of all target classes. Then we calculated the Bayes rate which is the upper bound of the model. We also constructed the best single variable model possible and compared it against our final models. For performance measurement, we constructed confusion matrices and calculated accuracy, precision, recall, f1 score, specificity, and sensitivity for all the models. However, in this case, misclassification of someone who is not at risk of developing disease into at risk or positive would not be much of a problem because taking preventive measures are not discourageable. In contrast, if we classify somone who is indeed at risk into negative, it would be a problem. So, we wanted the precision or sensitivity to be as high as possible.

# results and discussion

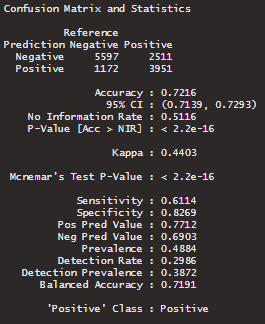
## Single Variable Model

Single variable prediction using "Systole" attribute alone yields around 71.7% accuracy. This is the null model which our models must beat.

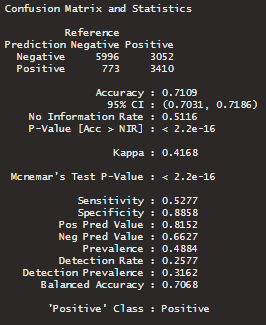


## Algorithm

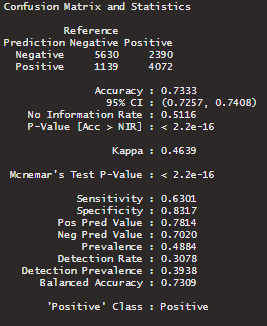
### Naïve Bayes



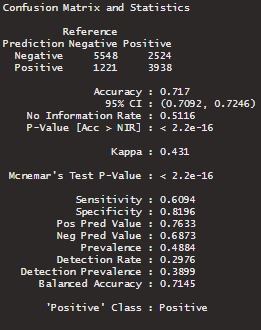
### Logistic Regression



### Support Vector Machine (SVM)



### Decision Tree



## Summary table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **P-Value** | **Sensitivity** | **Specificity** |
| Single Variable (Null Model) | 0.717 | 2.2e-16 | 0.6094 | 0.8196 |
| Naïve Bayes | 0.7216 | 2.2e-6 | 0.6114 | 0.8269 |
| Logistic Regression | 0.7109 | 2.2e-16 | 0.5277 | 0.8858 |
| SVM | 0.7333 | 2.2e-16 | 0.6301 | 0.8317 |
| Decision Tree | 0.717 | 2.2e-16 | 0.6094 | 0.8196 |

Sensitivity is also referred as the True Positive rate.

Specificity is False Positive rate.

With single variable model as the null model, the 71.7% accuracy will be used as our baseline. The p-value is 2.2e-16 which is less than the threshold 0.05, with this condition we reject the null hypothesis and agree that the result is statistically significant. The null model also has a sensitivity value of 0.6094 and specificity value of 0.8196

Naive Bayes model achieved accuracy of 72.16%  
whereas the p-value is 2.2e-16 which is less than the threshold of 0.05. Hence, we reject the hypothesis, and the result is deemed to be statistically significant. Sensitivity and specificity value are 0.6114 and 0.8269 respectively, higher than the null model. It shows slightly higher accuracy (around 0.46%) compared to the baseline accuracy too.

Next, Logistic Regression model achieved an accuracy of 71.09% lower by around 0.61% if compared to the null model. No difference for p-value between the null model and Navies Bayes model at 2.2e-16. Sensitivity is 0.5277 lower than the null model, but specificity is higher than the null model at 0.8858 in comparison to 0.8196.

Accuracy of 73.33% is achieved by Support Vector Machine (SVM) model. p-value is still the same among the models at 2.2e-16. The sensitivity value is at 0.6301 and specificity at 0.8317.

Lastly, Decision Tree model’s accuracy and p-value is at 71.7% and 2.2e-16 respectively, similar value to the null model. Both sensitivity and specificity value are the same for Decision Tree model and the null model.

The highest accuracy at 73.33% is achievable using SVM.

SVM also has the highest sensitivity value of 0.6301.

Quite surprising, Logistic Regression has the highest specificity value at 0.8858.

# conclusion

The project initially consists of 11 predictors and after further research, another 2 predictors that we deem useful are computed, BMI is calculated using Height and Weight  
and Blood Pressure Level from Systole and Diastole. These calculated predictors was proven to be the among 8 predictors to train the model via recursive feature selection. The other predictors are Systole, Diastole, Age, Weight, Cholesterol and Active. By exploring the data, we can see that there are outliers in some of the attribute which request for further processing.

A total of 4 machine algorithm are chosen to build the model, they are Naïve Bayes, Logistic Regression, SVM and Decision Tree with the implementation of cross-validation. The model accuracy ranges from 71% to 73% with some model performing worse compared to the null model, we can say that there is no overfitting in our models.

Hence, we can conclude that the best performing model is Support Vector Machine model with the highest accuracy of 73.33%.

# Future work

Exploring the data using more plots to get better understanding. So that data processing can be done in a much confident manner.

Creating better machine learning models using dataset which is larger in size and having more dependent attributes such as family history, ethnicity, etc. and exploring other feature selection technique and implementing them.

Utilizing different machine learning algorithm and tuning the parameters available on algorithms used to improve the model performance.

##### References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use “Ref. [3]” or “reference [3]” except at the beginning of a sentence: “Reference [3] was the first ...”

For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].

1. Global atlas on cardiovascular disease prevention and control. (2011). [ebook] WHO; World Heart Federation; World Stroke Organization. Available at: https://www.who.int/cardiovascular\_diseases/publications/atlas\_cvd/en/ [Accessed 1 Dec. 2019].
2. The Star Online. (2019). *Heart disease ‘leading cause of death’*. [online] Available at: https://www.thestar.com.my/news/nation/2019/01/25/heart-disease-leading-cause-of-death [Accessed 1 Dec. 2019].
3. BMJ 2016;353:i2416
4. PMID: 25775317  DOI:[10.7326/M14-1225](https://doi.org/10.7326/M14-1225)
5. Wang, T.J., Gona, P.N., Larson, M.G., Tofler, G.H., Levy, D., Newton‐Cheh, C., Jacques, P.F., Rifai, N., Selhub, J., Robins, S.J., Benjamin, E.J., D'Agostino, R.B., & Vasan, R.S. (2006). Multiple biomarkers for the prediction of first major cardiovascular events and death.
6. Lloyd-Jones, D.M., Leip, E.P., Larson, M.G., d’Agostino, R.B., Beiser, A., Wilson, P.W., Wolf, P.A. and Levy, D., 2006. Prediction of lifetime risk for cardiovascular disease by risk factor burden at 50 years of age. Circulation, 113(6), pp.791-798.
7. Wang, T.J., Gona, P., Larson, M.G., Tofler, G.H., Levy, D., Newton-Cheh, C., Jacques, P.F., Rifai, N., Selhub, J., Robins, S.J. and Benjamin, E.J., 2006. Multiple biomarkers for the prediction of first major cardiovascular events and death. New England Journal of Medicine, 355(25), pp.2631-2639.
8. Weng, S.F., Reps, J., Kai, J., Garibaldi, J.M. and Qureshi, N., 2017. Can machine-learning improve cardiovascular risk prediction using routine clinical data?. PloS one, 12(4), p.e0174944.
9. Muthuvel, Marimuthu & Abinaya, M & Hariesh, K & Madhankumar, K & Pavithra, V. (2018). A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach. International Journal of Computer Applications.
10. Jaymin Patel, Prof. Tejal Upadhyay, Dr.Samir Patel,“Heart Disease Prediction using Machine Learning and Data Mining Technique”, International Journal of Computer Science and Communication, September 2015-March 2016, pp.129-137.
11. G. John, R. Kohavi, and K. Pfleger, “IrreleJohn, G., Kohavi, R., & Pfleger, K. (1994). Irrelevant Features and the Subset Selection Problem. Icml, 121–129. Retrieved from http://machine-learning.martinsewell.com/feature-selection/JohnKohaviPfleger1994.pdfvant Features and the Subset Selectio,” Icml, pp. 121–129, 1994.
12. Kaggle.com. (2019). *Cardiovascular Disease dataset*. [online] Available at: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset [Accessed 3 Dec. 2019].

**IEEE conference templates contain guidance text for composing and formatting conference papers. Please ensure that all template text is removed from your conference paper prior to submission to the conference. Failure to remove template text from your paper may result in your paper not being published.**