Paper Title\* (use style: paper title)

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

line 1: 1st Given Name Surname   
line 2: *School of Computer Sciences*  
line 3: *Universiti Sains Malaysia*line 4: Penang, Malaysia  
line 5: email address or ORCID

Lee Jing Wen  
*School of Computer Sciences*  
*Universiti Sains Malaysia*Penang, Malaysia  
leejingw@student.usm.myline 1: 2nd Given Name Surname  
line 2: *School of Computer Sciences*  
line 3: *Universiti Sains Malaysia*line 4: Penang, Malaysia  
line 5: email address or ORCID

line 1: 3rd Given Name Surname  
line 2: *School of Computer Sciences*  
line 3: *Universiti Sains Malaysia*line 4: Penang, Malaysia  
line 5: email address or ORCID

*Abstract*—With the increase and successful implementation of machine learning in various predictive tasks in the real world, attracting the interest in the medical field in applying the same technique on available data. Cardiovascular disease being one of the leading causes of death worldwide, although modern technologies provide accurate diagnosis of cardiovascular disease, most often diagnosis takes too much time, or it is too late. Since identifying people at-risk would enable early prevention and treatment, which is often preferable than the previous. By using openly available software and public domain data, machine learning techniques implementation and evaluation will be done to serve this purpose. Our goal is to create a predictive model that can predict the risk of developing cardiovascular disease in patients, with an accuracy rate over 80%, based on some easily obtainable medical records such as age, gender, BMI, blood glucose, cholesterol, levels of physical activity, alcohol consumption, smoking habit, etc. Demonstration on the usage of machine learning algorithms in building predictive models for cardiovascular diseases diagnosis using descriptions of data records. The algorithms chosen are Naïve Bayes and Decision Tree along with standard statistical test XXX in selecting the best attributes. The dataset used is obtained from a publicly available source, Kaggle will be split randomly into training and testing samples. Algorithms are trained using the data from the training sample before using the test sample to predict the target where identification of presence or absence of cardiovascular disease in patients are done. Performance of the predictive models is completed using matrices such as accuracy, recall, precision and f1-score. The steps using in the algorithm development using open-source tools R will be provided in this paper.

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. 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. 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.

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, accurate 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, 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 also 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 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).

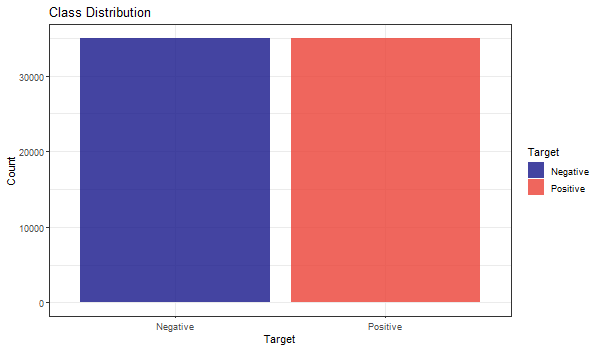
|  |  |  |
| --- | --- | --- |
| **Systolic blood pressure** | **Diastolic blood pressure** | **BloodPressure** |
| More or equal to 140 | More or equal to 90 | 1 |
| Less or equal to 90 | Less or equal to 60 | -1 |
| More than 90 and less than 140 | More than 60 and less than 90 | 0 |

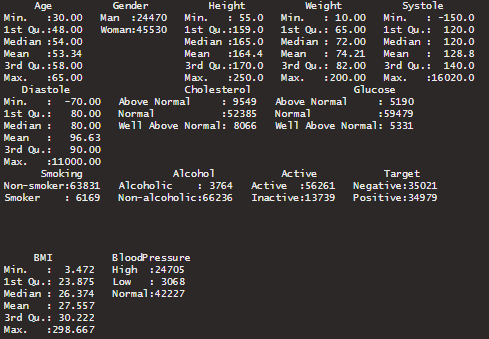
The summarized data description is stated at the table below:

|  |  |  |  |
| --- | --- | --- | --- |
| **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

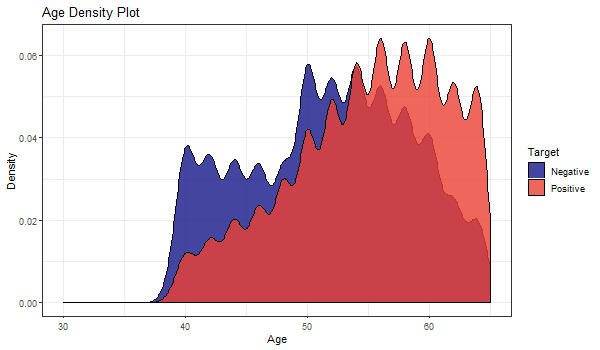
The class distribution of the target attribute, cardio is balanced as seen in the figure below:

****



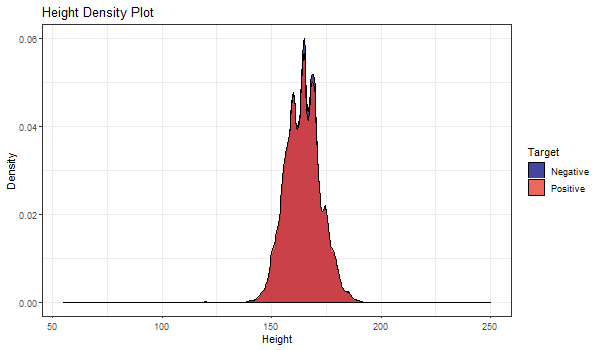
## Density plot

### Age Density Plot

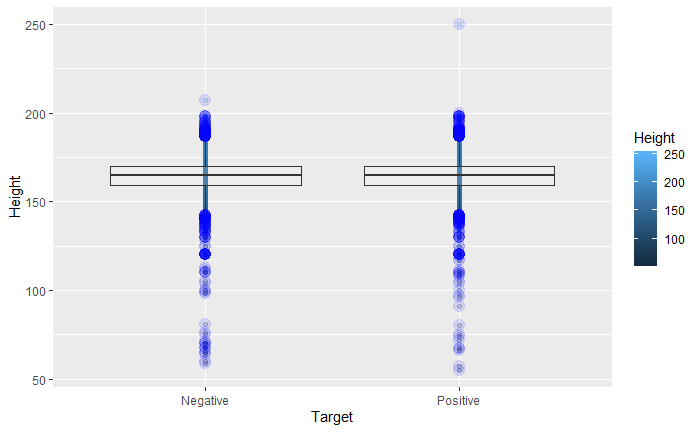


The peak density of the distribution is above 0.06 at the age range of 57 to 60. The distribution displayed multimodality characteristics with multiple peaks. It can be determined that the patients with cardiovascular diseases are more present in high age group.

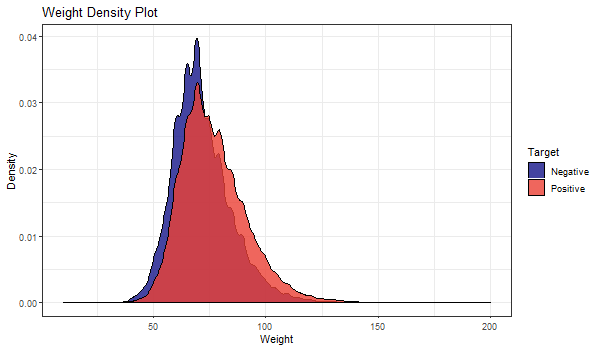
### Height Density Plot



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.

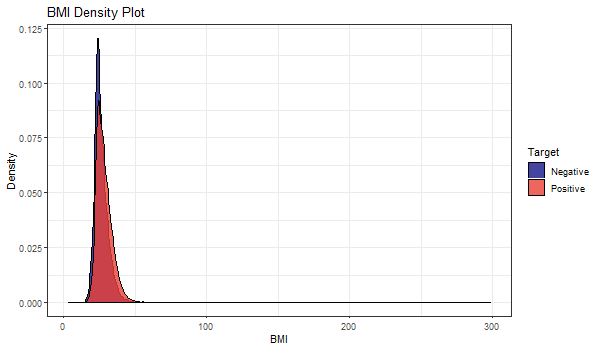


### Weight Density Plot



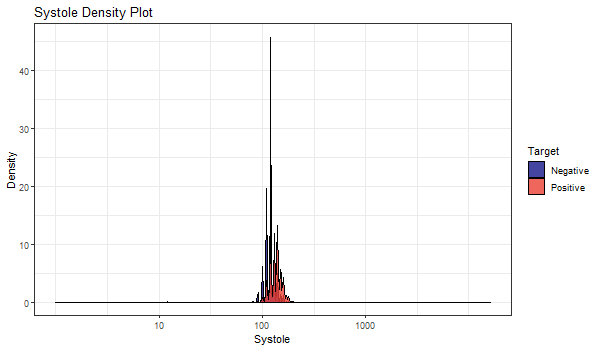
Peak weight density of the distribution can be seen to be higher in the data where patients do not have cardiovascular disease. The distribution displayed multimodality characteristics with multiple peaks. As the weight increases starting from around 72kg, there is a higher chance the patient has cardiovascular 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



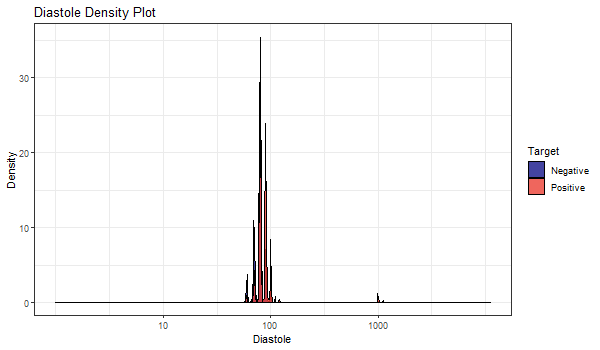
Since BMI is calculated using attribute weight and height, we can see that the BMI distribution shows unimodality characteristic. The peak density of the distribution is above 0.11 where patients do not have cardiovascular disease. It can be determined that patients with cardiovascular diseases are more 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.

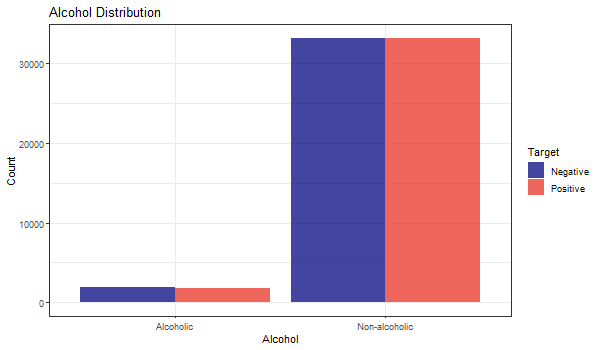
### Diastole Density Plot



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

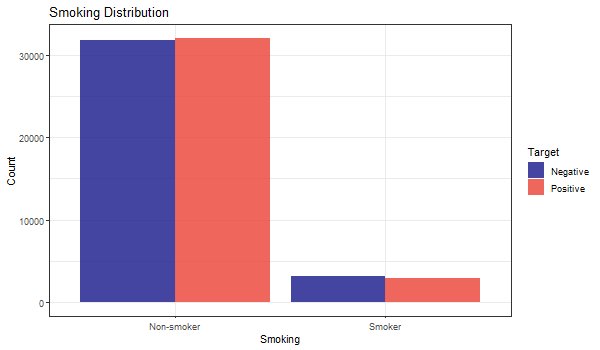
## Bar Plot

### Alcohol Distribution



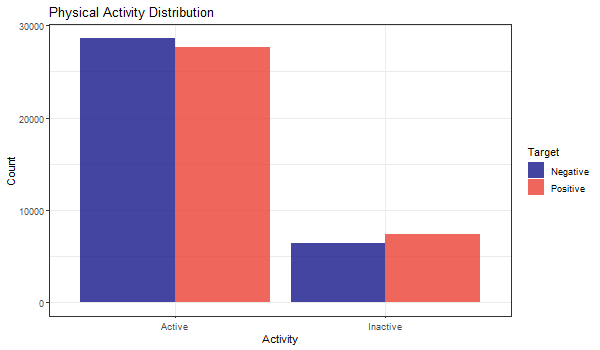
The proportion of having cardiovascular diseases is not seen to be positively correlated to alcoholism as both groups have the proportion of negatives higher, but the contract is not pronounced.

### Smoking Distribution



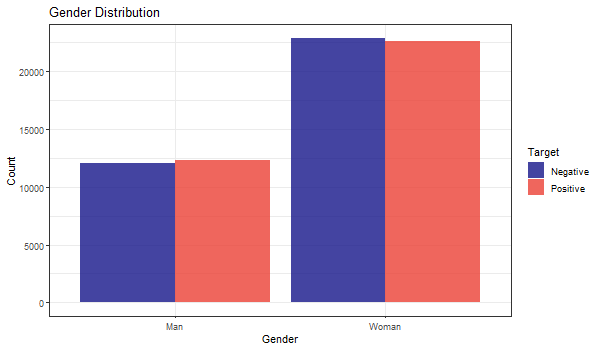
Like Alcohol Distribution, it seems hard to put smoking as a powerful indicator due to same non-obvious distribution.

### Physical Activeness Distribution



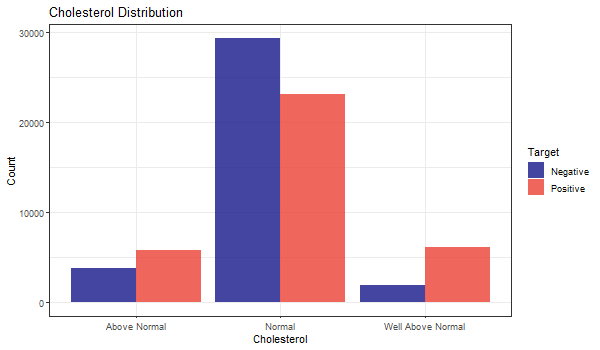
There is a trend that people who are not active in physical activity are more prone to have cardiovascular diseases.

### Gender Distribution



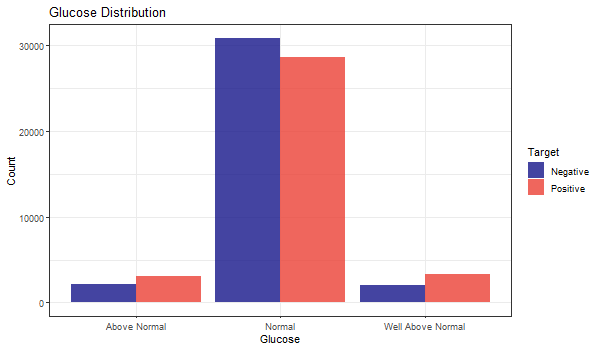
CVD were slightly more prevalent in male patients than their counterpart.

### Cholesterol Distribution



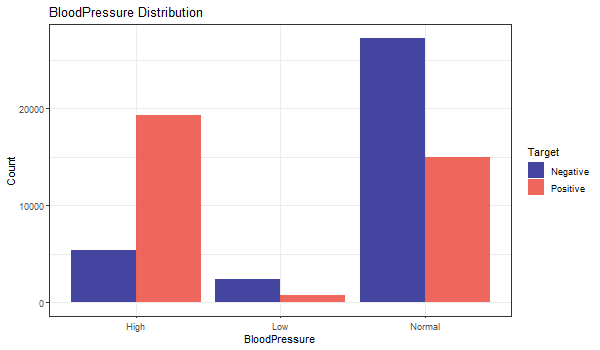
Cholesterol is seen as an obvious attribute and the well above normal category level is positively correlated to cardiovascular diseases.

### 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.

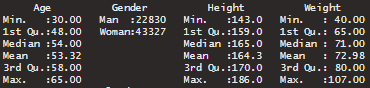
# Data preparation

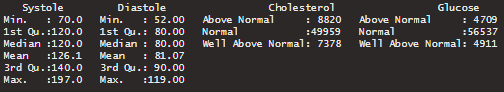
## Handling missing or null data points

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

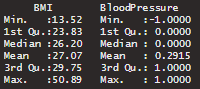
## Outliers

As seen in Figure xx (summary), 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 Figure xx as comparison, the attributes which has outliers like Height, Weight, Systole and Diastole looks way better after the outliers are handled which can be seen in the Figure below:





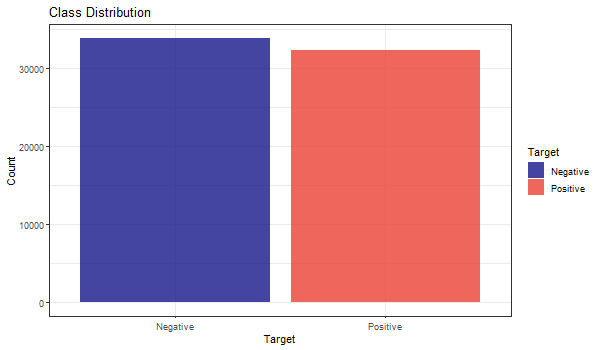




## Z-score normalization

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. 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.

****

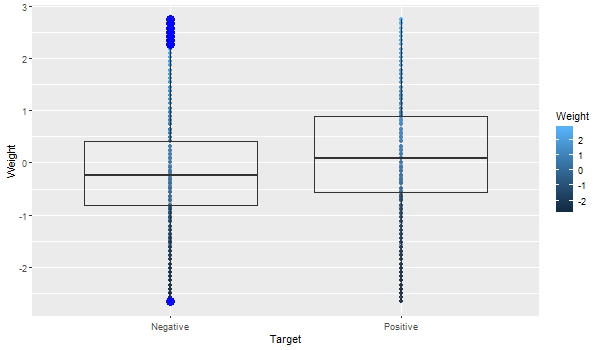
# feature selection

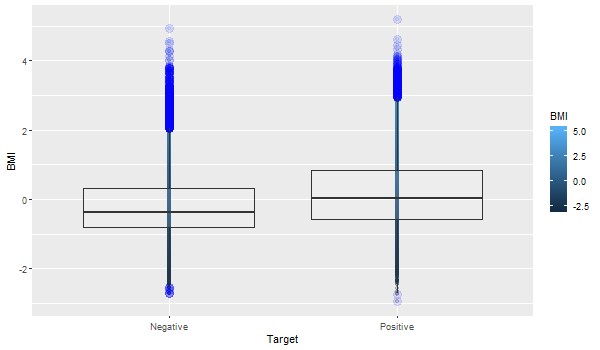
Most of the feature selection methods can be categorized in two categories, which are, wrapper methods and filter methods[11]. Wrapper methods evaluate a model by plugging different sets of features in order to find out the optimal subset for which the performance is maximum. Wrapper methods are indeed 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.

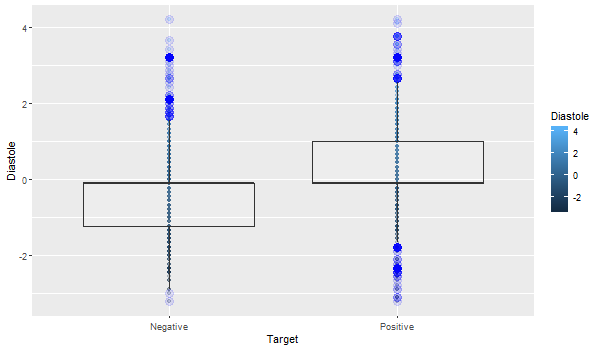
However, there are both advantages and disadvantages of both methods. Filter methods are less computationally demanding task than its counterpart, but it does not directly justify the performance of the model. As this method evaluates each feature separately, important interactions between features is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

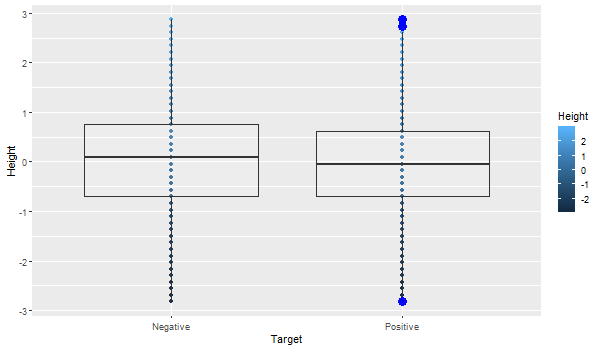
In this study, a wrapper method Recursive Feature Elimination (RFE) is applied because during the exploratory analysis no feature was found to have significant predictive power over the target. So we trained each model with different subsets of feature to identify the most effective ones.

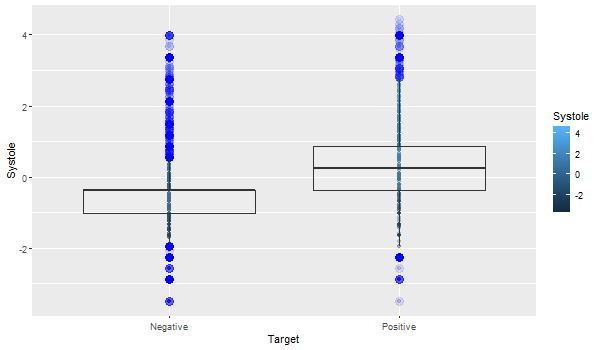
## Box plot

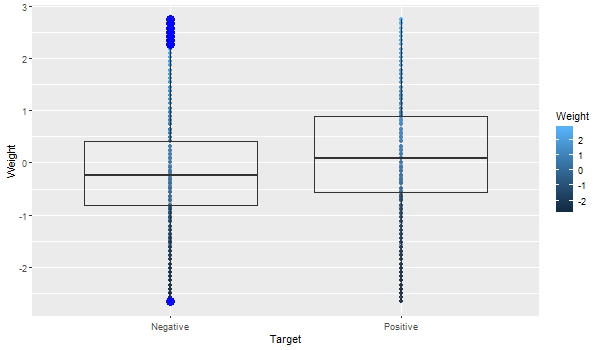










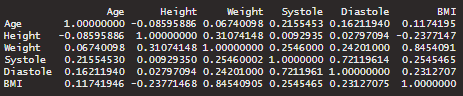


## Checking Near Zero Variance Attributes

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

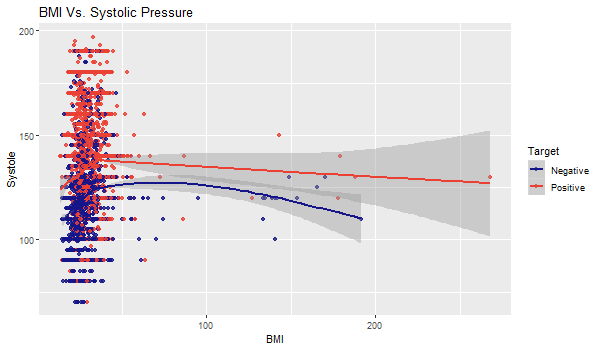
No features were found to be linearly dependent among each other.

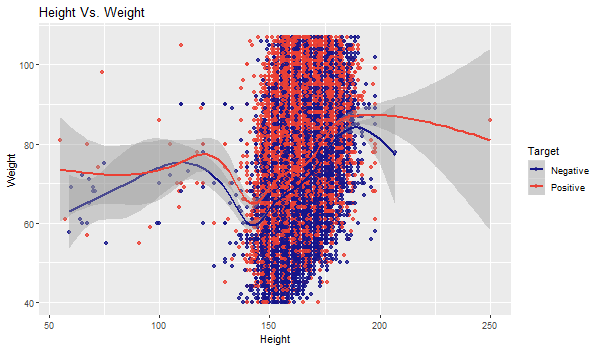
## Recursive feature selection

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 |

## INPUT NEEDED





# Machine learning

## 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 one.

* 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.

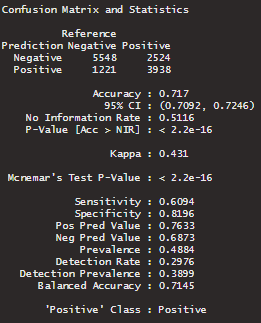
## Model Evaluation

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

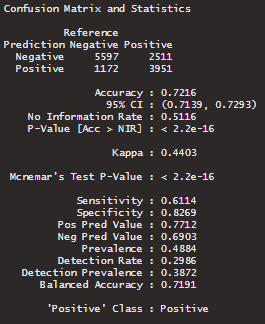
## Single Variable Model

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

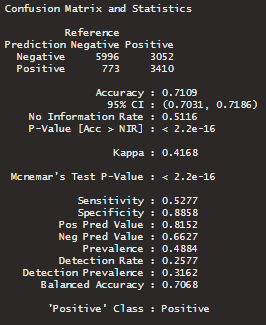


## Algorithm

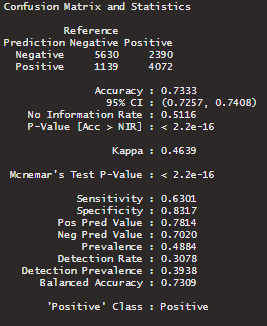
### Naïve Bayes



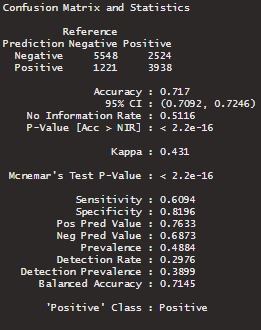
### Logistic Regression



### Support Vector Machine (SVM)



### Decision Tree



## Summary table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **P-Value** | **Sensitivity** |  |
| Single Variable Model | 0.717 | 2.2e-16 | 0.6094 |  |
| Naïve Bayes | 0.7216 | 2.2e-6 | 0.6114 |  |
| Logistic Regression | 0.7109 | 2.2e-16 | 0.5277 |  |
| SVM | 0.7333 | 2.2e-16 | 0.6301 |  |
| Decision Tree | 0.717 | 2.2e-16 | 0.6094 |  |

We take simple feature as null model, we take 71.7% accurancy as our baseline.  
p-value is 0.00000000000000022 which is less than the theshold 0.05,  
then we reject the null hypothesis, and we say the result is statistically significant.  
null model have a sensitivity of 0.6094

Navie Bays, we have an accurancy of 72.16%  
whereas the p-value is 0.00000000000000022 which is same less than the theshold 0.05.  
the we reject the hypothesis, and the result is satistically significant.  
model have around 0.6114 as sensitivity.  
it shows slightly higher 0.46% than our baseline accurancy.

Next, we look at the our 3rd Model, Logistic Regression, we have an accurancy result of 71.09%,  
suprisingly that the model is decreasing around 0.61% compared to our null model.  
p-value have no much different among null model and Navies Bays, same as 0.00000000000000022.  
The sensitivity is about 0.5277.

For our 4th Model Support Vector Machine (SVM), the accurancy is 73.33%. I can say that  
this is much improve than every models' result before and the increased gap between null model  
is about 1.63%. and p-value always the same at 0.00000000000000022.  
The sensitivity is 0.6301 also showing higher than other models.

Here come to our last Model, Decision Tree shows at the normal rate of accurancy about 71.7%  
exactly same with our null model result. Even p-value and sensitivity are the same values as  
baseline result at 0.00000000000000022 and 0.6094 respectively.

# conclusion

This study consists of 11 predictors at the beginning of project and after research, we come out another 2 predictors that computed from our base predictors which are BMI from Height and Weight  
and Blood Pressure Level from Systole and Diastole. These predictors are believed that useful factors to predict presence of Cardiovascular Disease (CVD) and we proven able use feature selection  
to identify highest correlation predictors to be considering input to our models and excluded weak predictors at the same time. After through our modeling process, we carries out 4 modeling method,  
the accurancy prediction of these modeling is stable at the range from 71% to 73%. and some prediction result show decreasing accurancy compared to null model, we can says that our splitting data   
is not overfitting at any case.

Hence, we can conclude that the highest accuracy that we can achieve is 73.33% and it is Super Vector Machine (SVM) Model whereas the lowest accuracy method is Logistic Regression Modeling that we have.  
From our dataset, we can also conclude that the best of predictor to the presence of Cardiovascular Disease (CVD) are Systole, BloodPressure, Diastole, Age, BMI, Weight, Cholesterol and Active.

# Using the Template

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

##### 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 ...”

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the abstract or reference list. Use letters for table footnotes.

Unless there are six authors or more give all authors’ names; do not use “et al.”. Papers that have not been published, even if they have been submitted for publication, should be cited as “unpublished” [4]. Papers that have been accepted for publication should be cited as “in press” [5]. Capitalize only the first word in a paper title, except for proper nouns and element symbols.

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.

**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.**