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[Lloyd-Jones 2006], which was based on the limitations of data collection and prediction tools and has been already proved to be biased[Weng2017].

Some attempts [Wang T.j 2006] 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. [Muthuvel 2018]

Muthuvel etc [ ] summarised 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 [Jaymin Patel 2016]. However the sensitivity (recall) is generally much higher than the precision (positive predictive value). This is typical in Weng[] 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

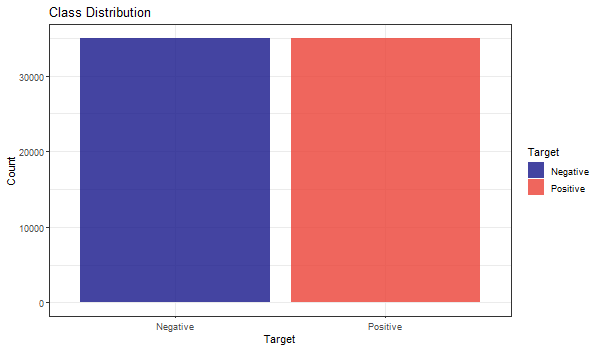
The dataset is obtained from Kaggle which consists a total of 70000 record of patient’s data. In this dataset, there is a total of 11 features which can be categorized into 3 types of input features, Objective, Examination and Subjective. Objective type is based on factual information, Examination type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

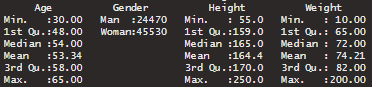
## Identification of attribute

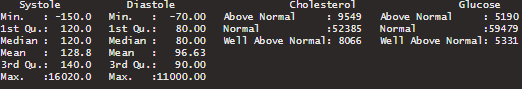
The summarized data description is stated at the table below:

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

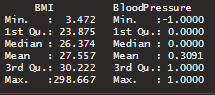
The class distribution of the target attribute, cardio is balanced as seen in the figure below:

****









# Data preparation

## Data processing??

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 |

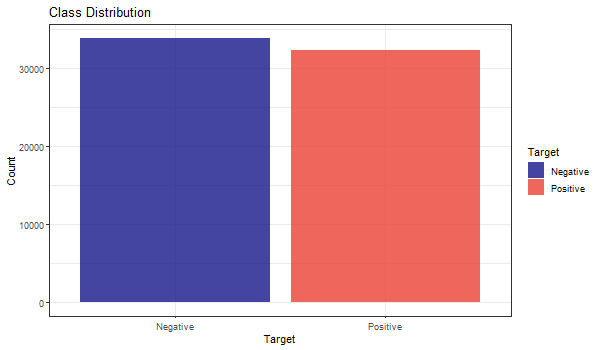
|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute** | **Attribute Name** | **Types of features** | **Data Type** |
| BMI | BMI | Input, Objective feature | Float |
| BloodPressure | Blood Pressure | Input, Examination feature | Integer |

## Handling missing or null data points

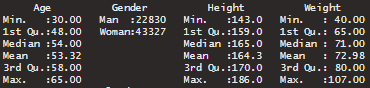
There are no missing values in the data, hence further data processing for this is not required.

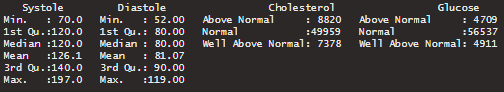
## Outliers

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 the range of 25% to 75% quantile with a fixed multiplier of respective attribute. 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.

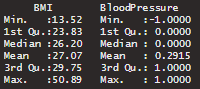
****

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:







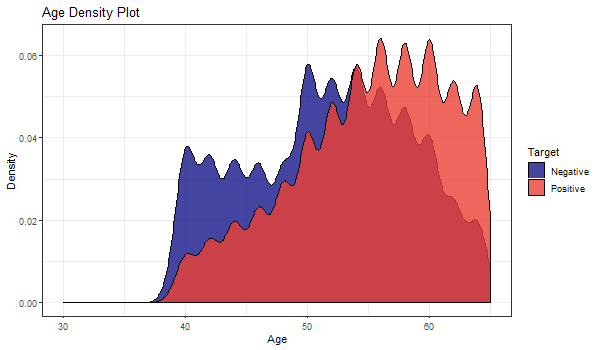


# Data visualization

## Density Plot

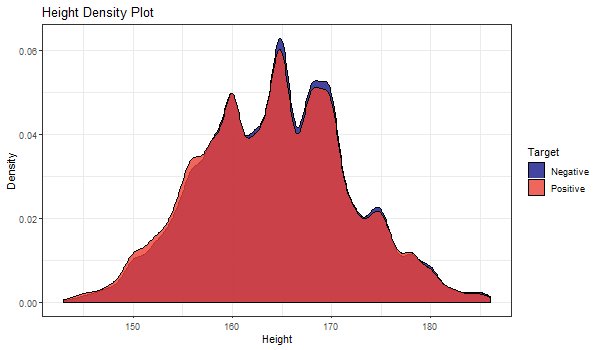
Why Density plot? Description needed

### Age Density Plot



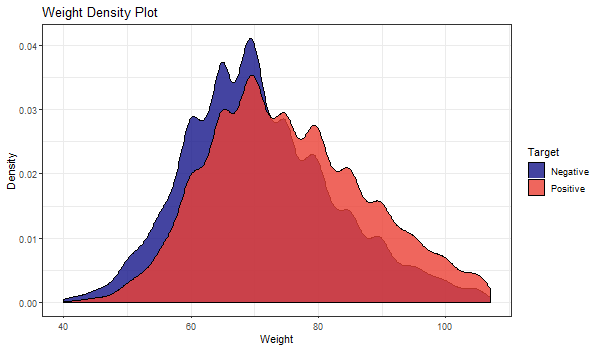
The peak density of the distribution is above 0.06 at the age range of 57 to 60. The distribution displayed multimodality characteristics with multiple peaks. It can be determined that the patients with cardiovascular diseases are more present in high age group.

### Height Density Plot



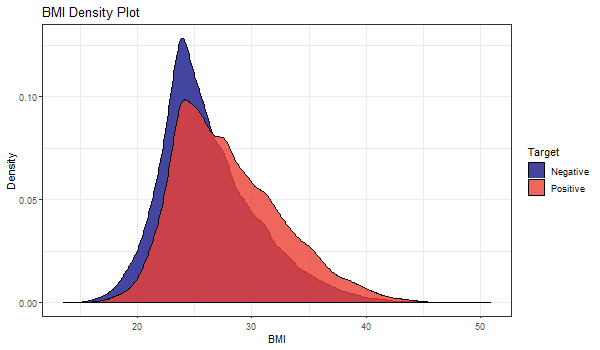
The peak density of the distribution is at the height of 164cm. The distribution displayed multimodality characteristics with multiple peaks. There seems to be no trend in identifying whether patients have cardiovascular diseases in the height attribute.

### 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.

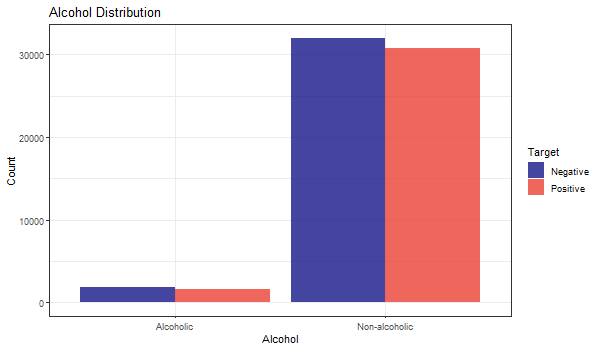
### BMI Density Plot



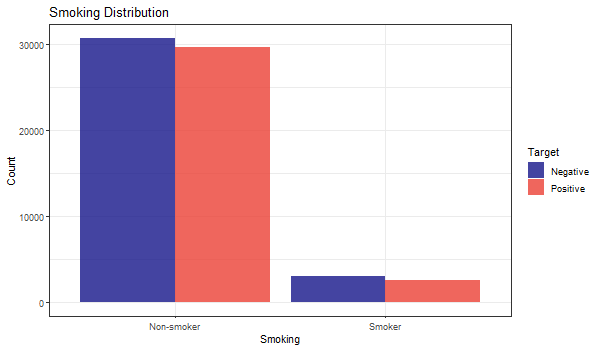
Since BMI is calculated using attribute weight and height, we can see that the BMI distribution shows unimodality characteristic which is much more helpful. The peak density of the distribution is above 0.11 where patients do not have cardiovascular disease. It can be determined that patients with cardiovascular diseases are more present when BMI is at 28 onwards.

## Distribution plot

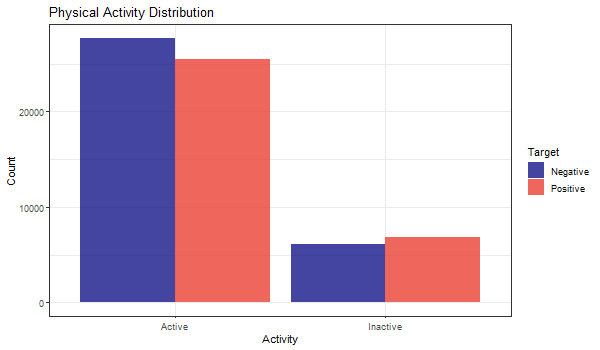
### Alcohol Distribution



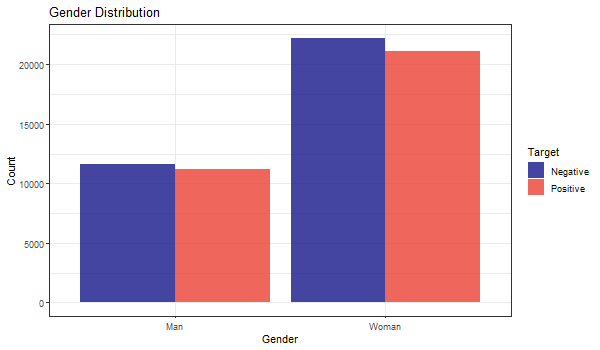
### Smoking Distribution



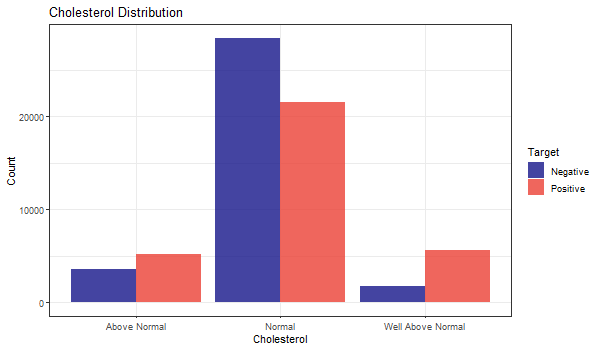
### Physical Activity Distribution



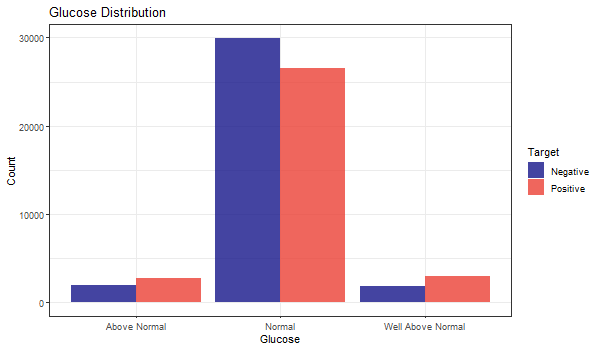
### Gender Distribution



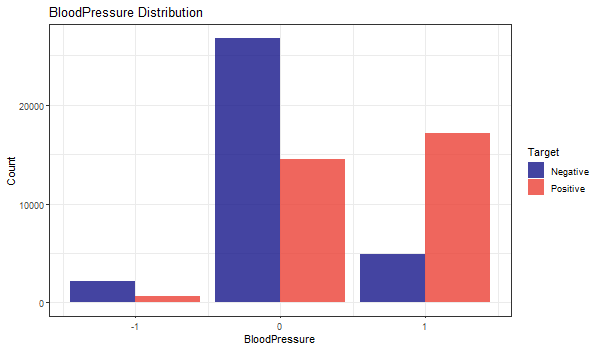
### Cholesterol Distribution



### Glucose Distribution



### BloodPressure Distribution



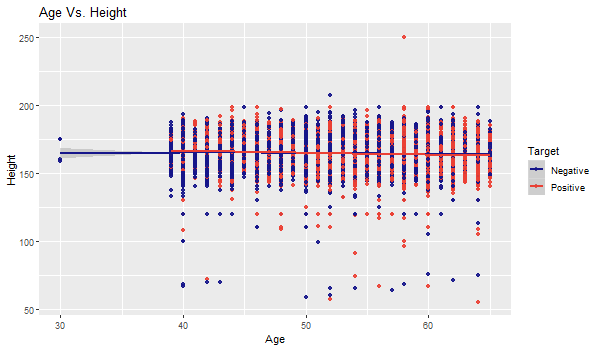
# feature selection

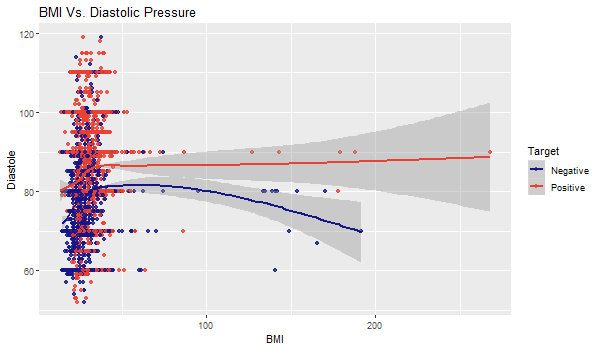
Most of the feature selection methods can be categorized in two categories, which are, wrapper methods and filter methods [1]. Wrapper methods evaluate a model by plugging different sets of features in order to find out the optimal subset for which the performance is maximum. Wrapper methods are, indeed, 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.

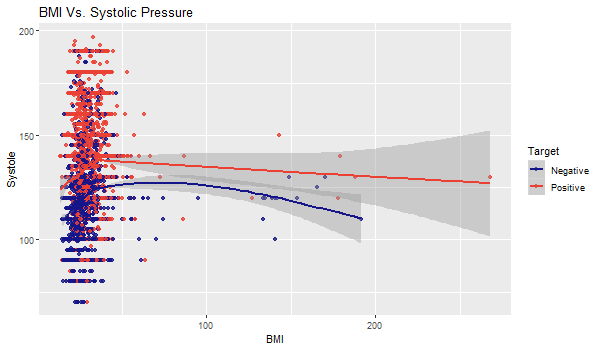
However, there are both advantages and disadvantages of both methods. Filter methods are less computationally demanding task than its counterpart, but it does not directly justify the performance of the model. As this method evaluates each feature separately, important interactions between features is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

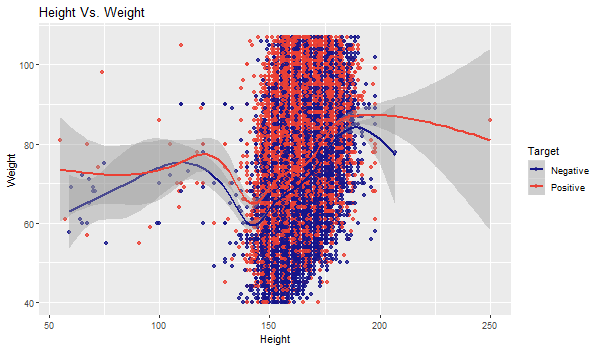
In this study, a wrapper method Recursive Feature Elimination (RFE) was applied because during the exploratory analysis no feature was found to have significant predictive power over the target. So, we trained each model with different subsets of feature to identify the most effective ones.

[1] 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.









# 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
* Decision Tree
* Support Vector Machine

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, 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.

## Model Validation

Standard procedures were maintained for model validation. The data was split into three groups for training, testing, and calibration. K-fold cross validation was applied during modeling. Significance tests were performed on the models and their p-values were compared.

# results and discussion

# conclusion

# 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.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### 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.

**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.**