Title

Early Warning Systems: Enhancing Heart Attack Prediction Accuracy

Abstract

The objective of this study is to develop predictive models for assessing heart attack risks using various machine learning algorithms. We utilized logistic regression, decision tree, Apriori algorithm, and two-step clustering to analyze a dataset containing health-related variables. The logistic regression model identified key predictors with moderate accuracy, while the decision tree highlighted cholesterol as the most significant factor. The Apriori algorithm demonstrated relationships between features, revealing patterns associated with increased heart attack risks. K-Means clustering provided insights into underlying data structures. Our results indicate that by implementing the models in healthcare systems can aid in early detection and personalized health interventions for at-risk individuals.

Introduction

Good health has consistently remained a paramount concern on a global scale, garnering widespread attention across all demographics. While historically more prevalent among adults, the advent of the Covid-19 pandemic has notably heightened good health awareness across all age groups. In light of this, our research problem focus will be specifically directed towards addressing the significant issue of heart attacks, with the key focus to reduce the risk of heart attack by producing prediction of heart attack likelihood enabling individual to take necessary precaution. A comprehensive, data-driven approach is required to accurately assess these risks and develop effective preventive strategies tailored to different demographic attributes.

Heart attack is one of the life-threatening diseases that frequently afflict adults, particularly those aged 40 and older. Conversely, younger adults typically experience a lower risk of heart attack. However, Henderson (2022) reported that “heart attack death rates took a sharp turn and increased for all age groups during the pandemic and the increase was most significant among individuals ages 25-44”. From this information, it does support our assumption that the covid-19 pandemic has contributed to heightened vulnerability of heart attack among various demographics. Henderson (2022) has furthered elaborated the reason heart attack has increased is due to “psychological and social challenges associated with the pandemic, including job loss and other financial pressures that can cause acute or chronic stress.” This may seem relevant since many individuals in that mentioned age group need to earn a living to support the family and the absence of reliable income sources inevitably leads to heightened stress and significant challenges.

We have evidence supporting the assumption that the risk of heart attacks affects younger people more significantly. Besides the impact of the COVID-19 pandemic, daily habits also play a crucial role. One prevalent habit among this age group is smoking. American Heart Association News (2021) reported that “young men who smoked had the highest long-term risk for heart attacks – 24%”. From this, it is noted that it is long-term risk, thus the occurrence of heart attack may be at a later age, and also depends on other underlying factors and habits each individual has. But at least we know that smoking will contribute to the risk of having heart attack in the long run. According to Fogoros (2022), some other lifestyle habits that may cause risk of heart attack is “poor diet, inactivity, and being overweight or obese”. While these habits alone may not directly cause a heart attack, when combined with other underlying health issues that individuals may have, they can contribute to an increased risk of heart attack.

Thus, the research focus is narrowed down to the following objectives – firstly, assess the likelihood of an individual developing heart disease from their lifestyle factor, in hope to reduce the risk of heart attack. Secondly, find out the main contributor(s) to heart attack disease.

The data mining goals to carry out the research objectives are firstly, to generate a predictive model to assess the likelihood of individuals towards developing heart disease by analysing comprehensive data encompassing their medical history, lifestyle factors, and demographic attributes. Secondly, the strategic focus is on leveraging advanced analytics to provide actionable insights that empower informed decision-making and promote proactive measures by exploring relationship between factors and heart attack in order to identify patterns and promote proactive measures for heart disease prevention. Lastly, apply feature selection to determine which is the main contributor to heart attack disease.

Literature Review

There are many underlying factors that may cause heart attacks, making it challenging to completely prevent them. However, risk can be reduced by adopting healthier daily lifestyle habits. Stewart et al., (2017) mentioned that quit smoking is the most cost-effective method to significantly reduce the risk of heart attack. Despite this, quitting smoking can be difficult, particularly for those with an addiction, and relapse is common. Therefore, medication can be a useful aid in quitting smoking and preventing relapse. According to Stewart et al., (2017), "the use of nicotine replacement therapy (NRT) and buproprion (a norepinephrine dopamine reuptake inhibitor) are universally recommended and have improve abstinence rates by 50–70%". The figure indicates significant improvement in smoking cessation. Quitting smoking is beneficial in all aspects and significantly reduces the risk of heart attacks. Some individuals do not have habit of smoking. Therefore, another potential solution is to emphasis on ways to maintain a healthy lifestyle. Rippe, (2018) talks about physical inactivity is a major risk factor for coronary heart disease (CHD), which can lead to heart attack. Physical inactivity in layman's term simply means no exercise. According to statistics found by Rippe, (2018), "compared with those who are very physically active, the risk of CHD in sedentary individuals is 150% to 240% higher." The statistics shown that the risk of heart attack is significantly higher in those who do not have any exercise. Therefore, based on Rippe, (2018), it is highly recommended that individuals should at least incorporate small amount of exercise because this will help to decrease the risk significantly. Furthermore, according to Campbell et al., (2012), there is strong evidence that dietary salt intake is a major contributor to elevated blood pressure and can lead to cardiovascular disease, which includes heart attack. It seems like it is the chemical in our diet that we need to be mindful of. Campbell et al., (2012) suggest that "salt intake should be between 9 and 12 g/day". However, a more ideal consumption of salt in the diet would be around "6 g/day because this can reduce coronary heart disease by 18%" (Campbell et al., 2012).

Research methodology

To effectively address the research objectives and ensure the reliability and validity of our findings, we adopted a comprehensive research methodology. This approach includes several steps and techniques starting with business understanding, data preparation and transformation, the application of various data mining methods, algorithm selection and implementation, and model building and validation. Each step was meticulously designed to thoroughly analyze the factors contributing to heart attack risks and to develop accurate predictive models.

Business understanding is the phase that we investigate the current situation and identify the business objectives / goals. In this research, the goal is to investigate the situation in relation to heart attack. The task carried out in this phase includes identify business objectives, carry out situation assessment which includes the resources, requirements, assumptions, constraints and risk, identify data mining goals and create a project plan.

Data understanding is the phase that allows us to get a better understanding of the data. In this phase, we have collected the initial data from Kaggle, which is an open- source data platform. This step is mainly about getting basic insights about the dataset such as summary statistics, and identify

any outliers or missing values. Also, this includes creating basic plots to have a better understanding of the data or certain variables visually.

Data preparation is the phase where we properly clean the data and take other necessary steps to prepare the dataset for further analysis in subsequent steps. This includes choosing relevant variables based on initial exploration, construction of new variables from existing variable(s), integration of two different data has taken place and lastly ensure format of the data is correct and suitable for subsequent analysis.

Data transformation in the context of the research is to reduce the number of variables by only keeping the important ones. This is done by using feature selection, which is an algorithm that helps to make decision on which variables are important and which are not, and it is our decision to make to decide which variables to keep.

Data Mining method selection is the process where we select and discuss the appropriate data mining methods which are most appropriate for our data mining objectives.

Algorithm selection is the process where we conduct exploratory analysis of the data mining algorithm in line with our data mining objectives. This is the research part to get to know the basic theory behind the algorithm. Once learn about the theory and selected the algorithm, then models are built with the chosen algorithm and model parameters are selected wisely.

Data Mining is the process where we ensure the model we have built run successfully without error. This phase is crucial because we get the result such as accuracy score and based on the output, we make important decision to address our objective and problem. This phase also includes searching for insightful patterns for analysis.

Interpretation is the phase where we communicate the results from each model. Not just that, visualization is also produced as a visual aid. More importantly, we also assess the results and come up with a final decision. However, in this phase, multiple iterations also take place to ensure the model is effective and in hope to produce better result.

Data Analysis

For the research, three different software stacks have been used. The software that was used are IBM SPSS Modeler 18.4, Spyder and PySpark in Jupyter notebook. These software tool enables to conduct statistical analysis such as summary statistics and regression analysis, data visualization and much more. Given that Kaggle is an open-source platform, and the data has been obtained from here, it is accessible to anyone, potentially leading to data alteration that could impact subsequent analysis results. Thus, to deal with risk of data alteration, we have performed checks such as anomalies and outliers and missing values in subsequent analysis. Initial understanding of the data, which includes describe the data, explore the data, visualization of the distribution of data and so forth, produce identical results for each of the three different software stacks. Thus, I will illustrate using the result produced in Spyder. This dataset encompasses a wide array of attributes, such as age, cholesterol levels, blood pressure, smoking habits, exercise patterns, dietary preferences, and additional factors.

```
# Load the csv file into Spyder

file = 'heart_attack_dataset.csv'

import pandas as pd
heart_data = pd.read_csv(file)
```

Figure 1

Firstly before I begin, I upload the dataset into Spyder with the lines of code in figure 1 above. I have assigned the name heart_data to the heart attack dataset that will be used for this project.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2499 entries, 0 to 2498
Data columns (total 27 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
```

Figure 2

```
#   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
0   Patient ID                        2499 non-null   object
1   Age                               2499 non-null   int64
2   Sex                               2499 non-null   object
3   Cholesterol                       2499 non-null   int64
4   Systolic                          2499 non-null   int64
5   Diastolic                         2499 non-null   int64
6   Heart Rate                        2499 non-null   int64
7   Diabetes                          2499 non-null   int64
8   Family History (1: Yes)           2499 non-null   int64
9   Smoking                           2499 non-null   int64
10  Obesity                           2499 non-null   int64
11  Alcohol Consumption               2499 non-null   int64
12  Exercise Hours Per Week           2496 non-null   float64
13  Diet                              2499 non-null   object
14  Previous Heart Problems (1 : Yes) 2499 non-null   int64
15  Medication Use                    2499 non-null   int64
16  Stress Level                      2498 non-null   float64
17  Sedentary Hours Per Day           2497 non-null   float64
18  Income                            2499 non-null   int64
19  BMI                               2496 non-null   float64
20  Triglycerides                     2499 non-null   int64
21  Physical Activity Days Per Week   2499 non-null   int64
22  Sleep Hours Per Day               2497 non-null   float64
23  Country                           2499 non-null   object
24  Continent                         2499 non-null   object
25  Hemisphere                        2499 non-null   object
26  Heart Attack Risk (1: Yes)        2499 non-null   int64
dtypes: float64(5), int64(16), object(6)
```

Figure 2.1

In figure 2.

In Figure 2, it is evident there are 2499 rows and 27 columns (circled in red). In Figure 2.1, it depicts the format of the attributes. The Dtype column provides information about the types of variables present. Attributes such as Sex, Country, Diet etc. are identified as object because it takes string value. Attributes such as Age, Systolic etc. are identified as integer (int64) since it only takes whole number. Attributes such as exercise per week, sleep hours per day etc. are identified as float since it can take decimal values. In my dataset, there are 8 attributes that are binary variables. Attributes such as heart rate, diabetes, previous heart problem etc. are represented with values [0,1], indicating binary variables. In Figure 3.2, these variables are being identified as int64 data type. In the context of our dataset, 0 signifies "No" and 1 signifies "Yes.

Index	count	mean	std	min	25%	50%	75%	max
Age	2499	53.55	21.29	18	35	54	71	103
Cholesterol	2499	261.6	80.15	120	194	261	330	400
Systolic	2499	135.67	26.59	90	113	136	159	180
Diastolic	2499	85.4	14.53	60	73	86	98	110
Heart Rate	2499	74.77	20.57	40	57	75	92	200
Exercise Hours Per Week	2496	9.98	5.78	0.01	4.91	10.13	14.98	40.55
Stress Level	2498	5.5	2.86	1	3	6	8	20
Sedentary Hours Per Day	2497	5.94	3.49	0.01	2.87	5.79	9.04	11.99
Income	2499	157514	80349.6	5000	87317	157722	225304	299769
BMI	2496	28.96	6.31	18	23.41	28.89	34.33	39.99
Triglycerides	2499	421.49	222.13	30	231	420	611.5	800
Physical Activity Days Per Week	2499	3.49	2.29	0	2	3	6	7
Sleep Hours Per Day	2497	7	2.03	2	5	7	9	20

Figure 3

Figure 3 above shows the summary statistics for the numerical attributes. It includes the upper and lower quartile too.

```
In [118]: heart_data['Sex'].value_counts()
Out[118]:
Sex
Male      1723
Female     776
Name: count, dtype: int64
```

Figure 4

In Figure 4 depicted above, it reveals that there are 1723 individuals classified as male and 776 individuals classified as female. Another way to show the count is with the aid of visualization. Figure 5 below is an example for the gender category.

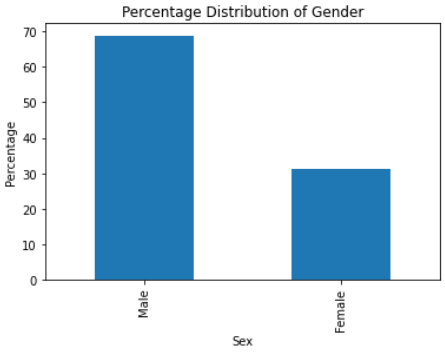


Figure 5

Once the basic understanding of the data has been conducted, it is crucial to verify the data quality. This is when checking for missing values and / or outliers take place. Upon further investigation, there are several variables that contain missing values (shown in figure 6). Not only that, outliers are also found in several variables (shown in figure 7). However, it's important to note that the variable "Smoking" is binary (taking values 0 or 1) and should not have any outliers due to its nature. The noticeable disparity in the proportion of smokers and non-smokers likely led Spyder to flag non-smokers (0) as outliers. The same variables identified to be extreme values are outliers, we can conclude that, in this context, the terms "outliers" and "extreme values" are synonymous.

```
In [95]: print(missing_values[missing_values > 0])
Exercise Hours Per Week    3
Stress Level               1
Sedentary Hours Per Day    2
BMI                        3
Sleep Hours Per Day        2
dtype: int64
```

Figure 6

```
In [139]: print(outlier_counts_filtered)
Heart Rate                   1
Smoking                    276
Exercise Hours Per Week      1
Stress Level                 1
Sleep Hours Per Day          1
dtype: int64
```

Figure 7

After reviewing our dataset, I have decided to retain all the variables and only remove the Patient ID variable column. The rationale behind this is because the patient ID seems irrelevant and unrelated to our analysis, as it is merely a unique identifier for each patient. Additionally, no sensitive information is included in the dataset that would identify the patients.

Missing values are not removed, instead imputation is applied. I have decided to impute missing values for the stress level variable using the mode because it is sensible to use mode imputation for categorical variable. Thus, the missing value will be replaced with the most frequent stress level. I have decided to impute missing value for continuous variables using the mean. For extreme values, it is best if we discard them. By removing the extreme values, it may improve our model performance in the subsequent analysis. In our heart attack dataset, there is the Age variable. However, for a more insightful analysis, I think it is beneficial to categorize the patients into age groups rather than analyzing individual ages. As a result, I created a new variable Age group, from existing Age variable. The new variable consists of three different age groups, which are Young Adults (age 18-30), Middle Age (31 – 59) and Old (60-90). It is evident that majority of the patients belong to the “Old” age group of age 60-90 (shown in Figure 8).

```
In [183]: cleaned_data['Age
group'].value_counts()
Out[183]:
Age group
Old             1044
Middle Age      1001
Young Adults     449
Name: count, dtype: int64
```

Figure 8

The target variable is *Heart Attack Risk*. The distribution is relatively skewed (63% for no risk (0) and 37% for presence of risk (1)). This might have a negative influence on the overall predictive power of our model in subsequent analysis. Thus, I have taken a further step to balance out the count for the target variable. This is dome by randomly selects a subset of instances from majority class to match the count of the minority class. This also result in reduced entries (rows). In the dataset, there are originally 27 features. However, some features might not be the main contributor to heart attack. Many features may lead to model to overfit. Therefore, in order to find out the variables that are potentially contributors, feature selection technique is used. Features are chosen based on the *f_regression* scoring function in Spyder, *RandomForestClassifier()* function in PySpark, and the feature selection node in SPSS Modeller. According to the three output, the selected features are previous heart problem, income, diabetes, cholesterol , systolic, age and stress level. These features are used for subsequent analysis.

We aim to discuss some data mining methods which align with our data mining objectives. But firstly, we will have a quick look into the data types for our subsequent data mining process. The data types are a mixture of continues variable and binary variable. There are various modelling techniques that can be used, such as classification etc. We will delve deeper to have a better understanding of some data mining methods. Ali (2023) talks about one of the data mining techniques is association. "Association technique involves looking for certain occurrences with connected attributes" (Ali, 2023). Like the term itself, it refers to the relationship between different variables. This technique could be useful especially when we wish to draw certain assumption to determine if one variable has an effect on another variable. The next common data mining method is classification. "Classification in data mining is a technique used to assign labels or classify each instance, record, or data object in a dataset based on their features or attributes" (Utkarsh, 2023). Again, from the term itself, this suggest grouping in common term. (Utkarsh, 2023) also mention that classification techniques can be divided into categories - binary classification and multi-class classification. Binary means it only has 2 distinct values, while multi-class means it has more than 2 values, as the word multi suggest. Another common data mining method is clustering. (Sehgal, 2022) describes clustering as putting items with similarity together but could look for values that are out of the norm range, such as outliers. Classification and clustering on simple term may sound alike, but in fact, both techniques are different. Firstly, classification is supervised learning, and clustering is unsupervised learning. "Supervised Learning works with the help of a well-labelled dataset, in which the target output is well known" (Baheti, 2021). On the other hand, unsupervised learning is the opposite. "Algorithm is trained using data that is unlabeled. The machine tries to identify the hidden patterns and give the response" (Baheti, 2021). Additionally, feature selection is also an option. Feature selection by its own term means select features that are consider important in relation to the target variable as often dataset could consist of many features but some may be redundant or irrelevant.

The first objective is to measure the likelihood of individuals towards developing heart disease. The target variable is heart attack risk, which is a binary variable that takes the value 0 or 1. In simpler term, we are interested to predict the chances of individuals developing heart disease, thus we will use classification method for this first objective. In the above subsection, it is mentioned that classification can handle binary variables, thus it is the appropriate method to tackle this particular objective.

The second objective is to provide actionable insights that empower informed decision-making by exploring relationship between factors and heart attack in order to identify patterns and promote

proactive measures for heart disease prevention based on findings. We will use association method to handle this objective. This is because we intend to find the significance of association between variables, which contributes to more effective strategies for heart disease prevention. We may also implement clustering method to identify distinct risk profile with the aim to make proactive measures more personalized.

The third objective is to determine the main contributor to heart attack disease. In order to achieve this objective, we will utilize the feature selection method. As mentioned in the subsection above, feature selection will only keep some features and filter out the rest that are being determined to be irrelevant or play an unimportant role to heart attack disease. Feature selection will only select a few variables that are potentially main contributors to heart attack disease.

Association technique is used to find relationship between features and produce result that can be used for the purpose of prediction and / or decision. Among association technique, there are a few algorithms. Apriori algorithm is one of them. “Apriori algorithm uses a bottom-up approach, starting with individual items and gradually building up to more complex itemsets” (All, 2023). (All, 2023) has mentioned that Apriori algorithm is straightforward and comprehensible, search for common patterns and association rules within extensive datasets. (All, 2023) has also discussed that apriori algorithm may produce a lot of association rules, potentially complicating result interpretation and thus difficult to understand. Classification techniques consists of many different algorithms. For instance, decision tree is one of them. “Decision tree works best for simple cases with few variables, thus often used as first line classification method” (Keserer, 2023). Decision tree algorithm is able to handle categorical variable without the need to transform it to a factor. This makes the process easier and straightforward since dataset may contain more than one type of data types. However, decision tree has some disadvantages. “It is a high variance algorithm, may easily overfit because it has no inherent mechanism to stop, thereby creating complex decision rule” (Kapil, 2022). From this piece of statement, this problem might arise if there are too many variables in the dataset. So, it does support the reference above quoted by (Keserer, 2023) that decision tree is a powerful algorithm for dataset with less variables.

On the other hand, according to (AIML.com, 2024), logistic regression which is another classification algorithm, is able to handle large number of features. (AIML.com, 2024) has also mentioned that different data types are acceptable for logistic regression, for instance continuous and categorical variables. For clustering technique, one of the common algorithms is the K-means clustering algorithm. “K-means minimize the variance of data points within a cluster” (McGregor, 2020). (McGregor, 2020) has also noted that K-means cluster is suitable for small dataset because the algorithm is designed to go through all data point. From this, we may determine that the disadvantage of this algorithm is it can be time consuming for large dataset since it needs to iterate over all the data.

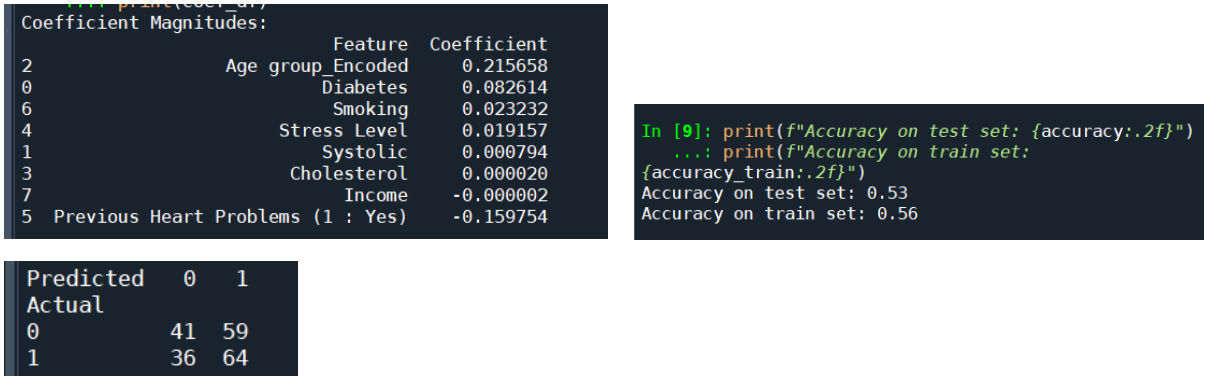
Moving on to select algorithm, the first objective is to predict the chances of individuals developing heart disease. The target variable is heart attack risk, which is a binary variable that takes the value 0 or 1. It is mentioned in section 5.2 that we will using the classification method. In particular, we will use **logistic regression** since the target is a binary variable. Also, we will implement the **decision tree** algorithm because this algorithm, as mentioned above, is a powerful algorithm for dataset with less variables. There are only 8 features and 1 target in our dataset, so it is sensible to use this algorithm for our analysis and hopefully derive interesting patterns from. The second objective is to provide actionable insights that empower informed decision-making by exploring relationship between factors and heart attack in order to identify patterns and promote proactive measures for heart disease prevention based on findings. As mentioned previously, this objective will be solved using the association method. In particular, we will try using the **Apriori** algorithm. As for clustering method, we will use the **K-Means clustering** algorithm. The third objective is to find the main contributor. We will use **feature selection** method, based on the f_regression scoring function in Spyder and

RandomForestClassifier function in PySpark, as this is a straightforward and easy method to find out the variables quickly.

Before building the model, the dataset has been split into the ratio 80:20 for train and test set. The reason behind the split is due to no other data to test our model. Once the splitting of dataset is done, we proceed to fit the first algorithm – Logistic Regression. We fit the logistic regression using X_train and Y_train. Subsequently, we also make prediction on the testing and training set and produce accuracy score for both sets. Also, we create a cross-tabulation of predicted vs. actual class for the target variable. The second algorithm is decision tree algorithm. The decision tree algorithm allows us to see the feature importance values based on how much each feature contributes to the model's decision-making process. However, decision trees do not provide coefficient magnitudes like logistic regression. But it can produce tree rules in a readable manner. Also, we create a cross-tabulation of predicted vs. actual class for the target variable. Moving on to the next algorithm – Apriori, itemsets that appear in at least 20% of the dataset will be considered frequent. It measures the likelihood of the consequent item(s) appearing in dataset containing the antecedent item(s). Lastly is the K-Means clustering method. The dataset is split into 3 distinct clusters and we create boxplot for each cluster to see the occurrence of features in each cluster.

Finding / Result Analysis

Logistic Regression findings:



```
Coefficient Magnitudes:
                          Feature  Coefficient
2                 Age group_Encoded     0.215658
0                         Diabetes     0.082614
6                          Smoking     0.023232
4                     Stress Level     0.019157
1                         Systolic     0.000794
3                      Cholesterol     0.000020
7                           Income    -0.000002
5  Previous Heart Problems (1 : Yes)    -0.159754
```

```
In [9]: print(f"Accuracy on test set: {accuracy:.2f}")
   ...: print(f"Accuracy on train set:
{accuracy_train:.2f}")
Accuracy on test set: 0.53
Accuracy on train set: 0.56
```

Predicted	0	1
Actual		
0	41	59
1	36	64

Figure 9 : Logistic regression coefficient summary (top left) & model accuracy (top right) & cross-tab output (bottom)

From the coefficient summary output, we can see there are multiple features that are considered significant to predict the target. From the cross-tab output, we can get an insight about the true negative and false positive values. Logistic regression belongs to predictive pattern, because it trains the dataset to predict on the test set. We can infer from the model accuracy score that logistic regression correctly predicts the heart attack risk 53% with test set and 56% with train set. Furthermore, we can infer from the coefficient values that not all features are significantly important in predicting the heart attack risk. Age group_Encoded and Diabetes are not significantly important variables in this logistic regression. From the cross-tab output, we can interpret the result that was tested on the test set as the cell (Actual = 0, Predicted = 0) indicates the model correctly predicted 41 instances of class 0 (no heart attack risk). On the other hand, the cell (Actual = 1, Predicted = 0) indicates the model incorrectly predicted 36 instances of class 0 when the actual class was 1 (heart attack risk presence) using the logistic regression algorithm. The model is not overfitting neither underfitting. The accuracy score for both test and train set are close. An accuracy of 53% on the test set suggests that the model's performance in making correct predictions on unseen data is relatively modest. The model is correct in its predictions for slightly more than half of the instances in the test set. As mentioned above, we can determine that some variables are not significantly important variables in this logistic regression based on their coefficient. If the coefficient is greater than p-value (usually 0.05), then we can conclude the variable is not significantly important.

Decision tree findings:

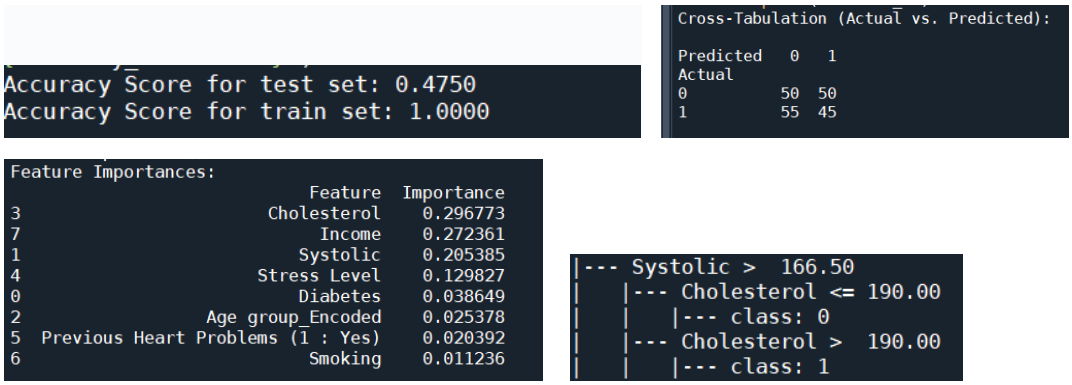
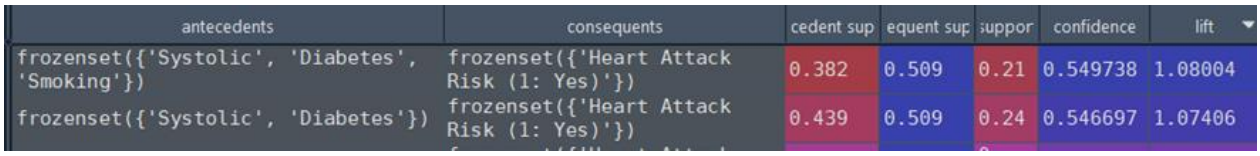


Figure 10: Decision tree accuracy score (top left), Cross-tab output (top right), feature importance (descending) (bottom left), rule set example (bottom right)

From the rule set, we can see the pattern where the algorithm list out possible set of rules, such as rules for heart attack labelled as class 1 or 0 and the feature importance in descending order. From the cross-tab output, we can get an insight about the true negative and false positive values. According to feature importance score, cholesterol is ranked as the most important features. This piece of information may imply that if an individual has cholesterol, then the chances of heart attack risk is high. As depicted, a cholesterol reading of less than or equal to 190, indicates class 0 (no risk), but a reading greater than 190 indicates class 1 (presence of risk). But this also depends on other factors such as systolic value. From the cross-tab output, the cell (Actual = 0, Predicted = 0) indicates the model correctly predicted 50 instances of class 0 (no heart attack risk). On the other hand, the cell (Actual = 1, Predicted = 0) indicates the model incorrectly predicted 55 instances of class 0 when the actual class was 1 (heart attack risk presence) using the decision tree algorithm. the accuracy for train and test set varies significantly. Train set has an accuracy score of 100% while test set has an accuracy score of 47.5%. This indicates the model is overfitting because train set has a much higher accuracy. The accuracy score is not good compare to logistic regression. There are several reasons why it has big difference compare to logistic regression and this might be because the algorithm works differently.

Apriori algorithm findings:



antecedents	consequents	cedent sup	equent sup	suppor	confidence	lift
frozenset({'Systolic', 'Diabetes', 'Smoking'})	frozenset({'Heart Attack Risk (1: Yes)'})	0.382	0.509	0.21	0.549738	1.08004
frozenset({'Systolic', 'Diabetes'})	frozenset({'Heart Attack Risk (1: Yes)'})	0.439	0.509	0.24	0.546697	1.07406

Figure 11: Apriori output

According to figure 11, we found the pattern that consequent is the target variable, and it is compared to other variables, to find the relationship between them. Apriori algorithm falls under association method. Thus, it depicts a descriptive pattern. We are able to derive information by looking at these patterns as there are many information we can get from the different columns. The target variable is the consequents and the feature(s) are the antecedents. It depicts a “If-then” relationship pattern. We can see that Smoking, Systolic and diabetes are the three prominent factors in modelling the heart attack risk. The confidence % is the two highest among the rest. We can interpret the result in a way by looking at the first row of figure 11 such as, if an individual has smoking habit, high systolic value and presence of diabetes, then heart attack risk is likely to be presence. By comparing the two rows, the first row has a higher confidence % and lift value. Therefore, we may infer that smoking habit may be a more prominent factor since this is the only

different variable between the two different antecedents. The values such as confidence % is not high but yet above 50%, and is able to capture some association / relationship between variables.

K-Means clustering findings:

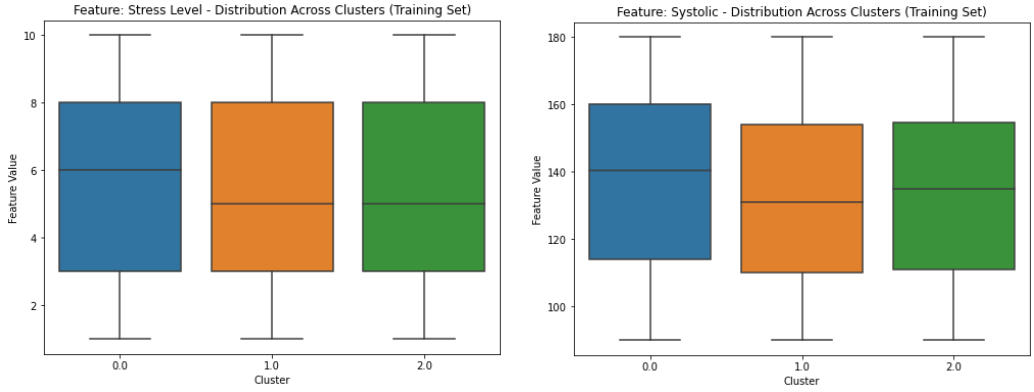


Figure 12: Cluster comparison

```
Silhouette Score (Training): 0.5591407265448741
Silhouette Score (Test): 0.560795673558588
```

Figure 12.1: Silhouette score output

The algorithm separates the data into 3 distinct cluster, and each cluster characteristics are depicted above. Clustering also belongs to descriptive pattern. We are able to find interesting pattern just by looking at the above image. Clustering also helps to reveal underlying relationship that may be interesting and surprising. For instance, in the boxplot, we can tell which cluster has a lower median value and derive further insights from there. The cluster comparison output depicts cluster 0 tends to have a higher median value for stress level and systolic, compare to the other 2 cluster. This could indicate that majority of the data points are larger for cluster 0, skewing the distribution towards higher values. Whereas, for example, cluster 1 has the lowest median value for systolic, which indicates that larger portion of systolic values are lower for cluster 1, compare to the other cluster. Silhouette score for both train and test data is 0.56 respectively. The cluster quality is considered fairly good since is it greater than 0.5. This score indicates that the features have captured some relevant and important patterns of the data.

Discussion

In this research, the primary objective was to predict the likelihood of an individual developing heart attack and to identify the main contributors to heart attacks using various machine learning algorithm. The valuable insights gained can be applied in several practical ways. In this section, we will discuss some proposed actions. Firstly, is to integrate the predictive model into the healthcare systems in order to identify high risk patients in early stage and implement proactive measures. This can alert patients even if they themselves have totally no clue that they are categorized as high risk or medium risk. Another way is to develop a health app so that individual can download in their device. This health app should be personalized as it can help them to monitor the risks and perhaps receive personalized health advice based on their habits or data. Besides that, pilot tests should be conducted in a few healthcare facilities first in order to refine the model and address any issues or error and make improvement before utilizing it formally. Once those are applied, it is important to be able to monitor the implementation. One key way is to regularly evaluate performance metrics such as accuracy to assess the model effectiveness. This is crucial because we do not want to misclassified patients into wrong categories. With the app, we can implement a feedback evaluation form for healthcare providers to report any issues encountered. Individuals can also provide feedback in terms of how they find the app, whether it is useful or no, and if the advice provided in the app helpful or no. After that, the next step would be how to maintain the implementation. This step

involves ongoing task to ensure the model remains effective and even make improvement. One action is to ensure the system is upgraded regularly to support the model functionality. Another way is to develop algorithm for handling errors in model prediction. This may occur so it is wise to develop such algorithm that has this function. It is also a good idea to provide training to healthcare staff to keep them updated about new features or any changes made in the model. This ensure they have good knowledge about it and are less likely to make mistakes. Lastly, is to enhance the model in the future. This may include collaboration with research institution to continue to improve the model based on their research and findings. In order to increase the model's predictive power, it is better if additional data are included, such as environmental data etc. However, these may include sensitive information, therefore, it is crucial that individuals build trust in the personalized health app before they input more sensitive information. Also, in the future, advanced algorithm could be implemented.

Conclusion

We have effectively implemented a variety of machine learning algorithms to predict the likelihood of heart attacks and to pinpoint the primary factors contributing to heart attack risk. By utilizing logistic regression, decision trees, the Apriori algorithm, and K-Means clustering, we gained valuable insights into the key influences on heart attack risks. Our models identified crucial predictors such as cholesterol levels, smoking habits, systolic blood pressure, and the presence of diabetes. Incorporating these predictive models into healthcare systems can facilitate the early identification of high-risk individuals and support proactive health management.

Reference

AIML. (2024). What are the advantages and disadvantages of logistic regression?. AIML.com. https://aiml.com/what-are-advantages-and-disadvantages-of-logistic-regression/

Ali, A. (2023). Top 10 Data Mining Techniques. Astera.

https://www.astera.com/type/blog/top-10-data-mining-techniques/

All, M. (2023). Association Rule Mining in Python Tutorial. Datacamp. https://www.datacamp.com/tutorial/association-rule-mining-python

American Heart Association News (2021). For smokers, fatal heart attack or stroke may be first sign of cardiovascular disease. American Heart Association News. https://www.heart.org/en/news/2021/11/17/for-smokers-fatal-heart-attack-or-stroke-may-be-first-sign-of-cardiovascular-disease#:~:text=Young%20men%20who%20smoked%20had,such%20strokes%20or%20heart%20failure.

Baheti, P. (2021). Supervised and Unsupervised Learning [Differences & Examples]. V7. https://www.v7labs.com/blog/supervised-vs-unsupervised-learning

Campbell, N., & He, F., & MacGregor, G. (2012). *Reducing salt intake to prevent hypertension*

and cardiovascular disease.

https://www.scielosp.org/pdf/rpsp/v32n4/08.pdf

Fogoros, R.N. (2022). Heart Attack Risks in Young People. Verywellhealth. https://www.verywellhealth.com/how-common-are-heart-attacks-in-young-people-3866059

Henderson, E. (2022). Heart attack death rates took a sharp turn and increased during the pandemic, study shows. News Medical Life Sciences. https://www.news-medical.net/news/20221024/Heart-attack-death-rates-took-a-sharp-turn-and-increased-during-the-pandemic-study-shows.aspx

Kapil, A.R. (2022). Decision Tree Algorithm in Machine Learning: Advantages, Disadvantages, and Limitations. AnalytixLabs.

https://www.analytixlabs.co.in/blog/decision-tree-algorithm/

Keserer, E. (2023). 8 Types of Machine Learning Classification Algorithms. Akkio https://www.akkio.com/post/5-types-of-machine-learning-classification-algorithms#:~:text=With%20an%20input%20training%20dataset,similar%20patterns%20in%20future%20data.

McGregor, M. (2020). 8 Clustering Algorithms in Machine Learning that All Data Scientists Should Know. FreeCodeCamp. https://www.freecodecamp.org/news/8-clustering-algorithms-in-machine-learning-that-all-data-scientists-should-know/

Rippe, J. (2018). Lifestyle Strategies for Risk Factor Reduction, Prevention, and Treatment of Cardiovascular Disease. *American Journal of Lifestyle Medicine*.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6378495/

Sehgal, A. (2022). Clustering Data Mining Techniques: 5 Critical Algorithms 2024. Hevo. https://hevodata.com/learn/clustering-data-mining-techniques/

Stewart, J., Manmathan, G., Wilkinson, P. (2017). Primary prevention of cardiovascular disease: A re view of contemporary guidance and literature. *Sage Journals.*

https://journals.sagepub.com/doi/10.1177/2048004016687211#bibr11-2048004016687211

Utkarsh. (2023). Classification in Data Mining. Scaler. https://www.scaler.com/topics/data-mining-tutorial/classification-in-data-mining/