**Heart Disease Dataset**

**Mid-stage Project Report, MATH 8050, Fall 2022**

Sai Vipraj Telukoti, C17202671

Vamshi Krishna Mummadi, C27032539

Yaswanth Mulakala, C16676264

**Dataset Description: -**

Heart Disease is a major health threat, that is why it is important to know risk factors so that proactive measures can be taken before it’s too late. Several health conditions such as lifestyle, age and family history can increase risk for heart disease.

This report aims to highlight factors that contribute to heart disease by analyzing a dataset obtained from Kaggle.

This dataset dates from 1988 and consists of four databases:

1. Cleveland
2. Hungary
3. Switzerland
4. Long Beach V

It contains 14 attributes and 1025 observations. While working with this dataset it was found that, it did not record the data in the most efficient way. It was discovered that the ‘target’ variable in the data was swapped. This variable provides indication as to whether the heart disease is present or not. This inconsistency in the data could be attributed to human error.

To carry out further analysis, this variable was flipped back to reflect the correct proportions.

The dataset has 14 columns. So, it is too big to take a screenshot of and show it here. Here is the name of the columns and the explanation of each variable as described in Kaggle.

1. age: The age of a person
2. sex: The person’s gender (1 = male, 0 = female)
3. cp: The types of chest pain experienced (Value 1: typical angina, Value 2: atypical angina, Value 3: non-anginal pain, Value 4: asymptomatic)
4. trestbps: Resting blood pressure (mm Hg on admission to the hospital)
5. chol: Cholesterol measurement in mg/dl
6. fbs: Fasting blood sugar (if > 120 mg/dl, 1 = true; 0 = false)
7. restecg: Resting electrocardiographic measurement (0 = normal, 1 = having ST-T wave abnormality,

2 = showing probable or definite left ventricular hypertrophy by Estes’ criteria)

1. thalach: Maximum heart rate achieved
2. exang: Exercise induced angina (1 = yes; 0 = no)
3. oldpeak: ST depression induced by exercise relative to rest (‘ST’ relates to positions on the ECG plot)
4. slope: the slope of the peak exercise ST segment (Value 1: upsloping, Value 2: flat, Value 3: downsloping)
5. ca: The number of major vessels (0–3)
6. thal: A blood disorder called thalassemia (1 = normal; 2 = fixed defect; 3 = reversable defect)
7. target: heart disease (0 = no, 1 = yes)

**Modeling: -**

1. **Logistic Regression: -**

Logistic regression is a classification algorithm used to find the probability of event success and event failure. Logistic regression is used when the dependent variable is binary (0/1, True/False, Yes/No) in nature. Logit function is used as a link function in a binomial distribution.

Logistic regression is also known as Binomial logistics regression. It is based on sigmoid function where output is probability and input can be from -infinity to +infinity.

Logistic regression will be used to a generate a model that can predict the heart disease in patients. For the sake of this data, logistic regression is the ideal choice for two reasons. They are: -

Firstly, the target (dependent) variable in the model is binary; secondly, it is a predictive analysis and can be used to explain the relationship between one dependent binary variable with multiple predictors.

Since resting blood pressure and cholesterol were log transformed it may be difficult to interpret it. Hence, additional feature engineering step will be carried out. Standardization of these variables will enable us to interpret the effects in terms of standard deviation.

**a) Implementation of the logistic model: -**

Text

Description automatically generated with medium confidence

Since the model incorporated logistic regression, the coefficients above correspond to log odds value. These coefficients directly do not state anything except that there is a positive relationship. The p-value is very small, which indicates that this relationship is statistically significant. Therefore, to understand and interpret this relationship more accurately, odds ratio will be computed.

### b) Interpreting the Coefficients through Odds Ratio: -

### Graphical user interface, application Description automatically generated

Intercept will be ignored as it does not give any relevant information. However, rest of the coefficients will be interpreted as following:

(Intercept) age sex log1\_trestbpsstd ## 0.012 1.060 5.505 1.246 ## log1\_cholstd ## 1.302

* **Age**: odds ratio of 1.060 can be interpreted as - increasing age by 1 year is associated with 1.06x increase in odds of getting a heart disease. in other words, aging 1 year is associated with a 6% [100 x (1.060 - 1)] increase in the odds of getting a heart disease.
* **Sex**: It is important to note that the odds ratio of 5.505 represents the odds of going to from female to male (since 0 = female and 1 = male). Hence, it is interpreted as, being a male is associated with 5.505x increase in odds of getting a heart disease.
* **Chest Pain**: odds ratio of 0.347 can be interpreted as increase in chest pain level by 1 lead to 65.3% [100 x (1 - 0.347)] decrease in heart disease

Log1\_trestbps: the odds of getting a heart disease are 24.6% [100 x (1.246 - 1)] higher is you are 1 standard deviation above log blood pressure.

Log1\_chol: the odds of getting a heart disease are 30.2% [100 X (1.302 - 1)] higher if you are 1 standard deviation above the average log cholesterol.

### c) Evaluating Predictive Performance: -

To evaluate predictive performance, train/test framework will be used. It is and 80/20 split to the original ‘heartdata’. So, 80% training data and 20% test data.

Graphical user interface, text, application, email

Description automatically generated

Now the training and test sets are formed. The model will now be trained on the training set and then make predictions on the training and test sets.

**d) Predictions: -**

Now that the predictions are now stored for the train and test data sets, confusion Matrix function can be used to generate confusion matrix which will tabulate the correct and incorrect predictions versus the true values. Recall that since logistic regression was used in the model, classification metrics will be used.

### e) Confusion matrix for predictions on Training set and test set: -

Results between the training and the test datasets are very similar.

* **Accuracy**: is the portion of correct predictions. Accuracy of 0.6% means that, 60% of the heart rate predictions on the test set are correct.
* **Recall**: Based on the definition of recall which is the proportions of actual true that the model labelled true, it can be said that the model correctly identified 57% of the patients with heart disease
* **Precision**: is the proportion of positive predictors that are actually true. So, when the model predicted heart disease, it is true 62% of the time.
* In general, the results between the training and the test sets are very similar, point to the fact that our model can predict results on a dataset that it is very new to. In other words, overfitting was not a problem.

1. **K Means Clustering: -**

K Means Clustering is an Unsupervised Non-linear algorithm that cluster data based on similarity or similar groups. It seeks to partition the observations into a pre-specified number of clusters.

Segmentation of data takes place to assign each training example to a segment called a cluster. In the unsupervised algorithm, high reliance on raw data is given with large expenditure on manual review for review of relevance is given. It is used in a variety of fields like Banking, healthcare, retail, Media, etc.

**Theory**

K-Means clustering groups the data on similar groups. The algorithm is as follows:

* Choose the number K clusters.
* Select at random K points, the centroids (Not necessarily from the given data).
* Assign each data point to closest centroid that forms K clusters.
* Compute and place the new centroid of each centroid.
* Reassign each data point to new cluster.

**Implementation**: -

Implementation of Principle component analysis: - Here we are implementing Principle component analysis for the dataset in order to continue with the K-means clustering. Calculating the standard deviation from p=1 to p=14 in the below image. Rotation of (n \* k) of 14 \* 14 sets is implemented, which ranges from PC1 to PC14. This rotation is calculated for each and every variable that is present in the dataset.

Table

Description automatically generated

The below plot the relationship between variances.

Chart, line chart

Description automatically generated

**Implementation of K-means Clustering**: - With the help of principle component analysis, K-means clustering is implemented. The summary statistics of the implemented K-means clustering is given in the below picture.

Graphical user interface

Description automatically generated

**Visualization of K-means clustering**: - Using an ellipse structure construction, the visualization of k-means technique is implemented. The below figure shows the graphical representation between PC1 and PC2 with the three mentioned factors represented with various colors.

Chart, bubble chart

Description automatically generated

**Summary: -**

This project aimed at understanding the factors that contribute to heart disease. A model was built to predict whether a patient has heart disease or not using a dataset obtained from Kaggle. This dataset had to be adjusted for the ‘target’ variable as the values were found to be have switched. Below is a quick summary of the entire project by each variable:

**Age**

* This variable was recorded in years and displayed bimodal distribution.
* Density plot revealed that there are 3 groups. Less than 54 years (labelled as ‘adults’), greater than or equal to 54 and less than 70 years (labeled as ‘older adults’) and greater than or equal to 70 years (labelled as ‘elderly’)
* Older population in the dataset seem to have higher number of heart diseases than younger.
* Aging 1 year is associated with a 6% increase in the odds of getting a heart disease.

**Sex**

* Sex is related to heart disease, in that males have more heart disease.
* Being a male is associated with 5.5 increase in odds of getting a heart disease.

**Chest Pain**

* Side by side bar plots revealed even those that reported 0 chest pain had heart disease. This means that the level of chest pain experienced by the patient is not necessarily a result of heart disease.
* Odds ratio revealed that increase in chest pain level by 1 lead to 65.3% [100 x (1 - 0.347)] decrease in heart disease.

**Blood Pressure**

* The distribution of this variable was right skewed, so it was log transformed to make it normally distributed.
* Although those with heart disease seem to have lower blood pressure compared to those that do not, the odds of getting a heart disease are 24.6% higher is you are 1 standard deviation above log blood pressure.

**Cholesterol**

* This variable was right skewed, so it was log transformed to make it normally distributed.
* The odds of getting a heart disease is 30.2% [100 X (1.302 - 1)] higher if you are 1 standard deviation above the average log cholesterol.

Using logistic regression, a model was built that could determine whether a patient has heart disease or not. Based on analysis done on the data it was found that you are at higher risk of developing a heart disease if you are male, as you get older and if you have high blood pressure and blood cholesterol levels. Overfitting was not a problem with the model that was generated and so this model can confidently be used to predict whether a patient has heart disease or not.