Table of Contents

[List of figures 1](#_Toc99954992)

[List of Tables 1](#_Toc99954993)

[1- Data description and Research question 2](#_Toc99954994)

[2- Data preparation and cleaning 3](#_Toc99954995)

[3- Exploratory data analysis 4](#_Toc99954996)

[3.1- Univariate data analysis 4](#_Toc99954997)

[3.2- Bivariate data analysis 6](#_Toc99954998)

[3.3- Checking the normality of distribution 8](#_Toc99954999)

[3.4- Using unsupervised learning algorithm for doing EDA 8](#_Toc99955000)

[4- Machine learning prediction 9](#_Toc99955001)

[4.1- K nearest neighbors 9](#_Toc99955002)

[4.2- Implementation of KNN 11](#_Toc99955003)

[5- High performance computational implementation 11](#_Toc99955004)

[5.1- Implementation 11](#_Toc99955005)

[6- Performance evaluation and comparison of the methods 12](#_Toc99955006)

[7- Discussion of the findings 13](#_Toc99955007)

[7.1- Discussion 13](#_Toc99955008)

[7.2- Limitations 13](#_Toc99955009)

[8- Data management plan and Author contribution statement 13](#_Toc99955010)

[References 13](#_Toc99955011)

[Appendix 14](#_Toc99955012)

# List of figures

[Figure 1- Boxplot of numerical column 4](#_Toc99952142)

[Figure 2- Histogram for the numerical columns 5](#_Toc99952143)

[Figure 3- Barplot of the categorical columns 6](#_Toc99952144)

[Figure 4- Correlation plot 8](#_Toc99952145)

[Figure 5- Plotting the optimal no of clusters 8](#_Toc99952146)

[Figure 6- Cluster plot 9](#_Toc99952147)

[Figure 7: How is a new data point categorized. (Javatpoint, 2017) 10](#_Toc99952148)

[Figure 8: Calculating the distance between two points. (Javatpoint, 2017) 10](#_Toc99952149)

# List of Tables

[Table 1- Description of the features 2](#_Toc99952187)

[Table 2-Comparison of machine learning algorithms 12](#_Toc99952188)

# 1- Data description and Research question

In this case study, the Framingham heart study cohort data set is used to train and test the models for heart disease classification. The data is collected by examining the town residents from Framingham, Massachusetts, from an ongoing cardiovascular study (Kaggle,2021). This dataset contains over 4,000 records and 16 columns. This dataset was collected initially to study the prevalence of several cardiovascular diseases and the risk factors associated with them. This dataset helps us explore the patterns of CVD and the risk factors over time that affect our lives. The Framingham study was started in 1948 under the supervision of the U.S. Public health Service, which was then transferred under the control of the new National Heart Institute. Participants, including men and women from the town of Framingham, were sampled, and studied to identify the concept of risk factors and the joint effects of CVD, which also facilitates specialized studies.

Table 1- Description of the features

|  |  |  |
| --- | --- | --- |
| Feature name | Data type | Description |
| male | Integer | Represents the gender of the participant |
| age | Integer | Represents the age of the participant |
| education | float | Represents the education level of participant |
| currentSmoker | Integer | If the participant is a current smoker or not |
| cigsPerDay | float | Number of cigarettes the participant smokes per day on an average |
| BPMeds | float | Checks if the patient is on blood pressure medication |
| prevalentStroke | Integer | Checks if participant had a stroke previously |
| prevalentHyp | Integer | Checks if patient is hypertensive |
| diabetes | Integer | Checks if participant has diabetes |
| totChol | float | The total cholesterol level of participant |
| sysBP | float | Checks the systolic blood pressure of the patient |
| diaBP | float | Checks the diastolic blood pressure of participant |
| BMI | float | Checks body mass index of the participant |
| heartRate | float | Check the heart rate of the patient |
| glucose | float | Checks the glucose level in the body of patient |
| TenYearCHD | Integer | Checks for the ten-year risk of coronary heart disease in the patient |

Based on the problem statement, I have framed a research question that I am going to answer with the help of the is project is given down below.

To identify the most relevant/risk factors of heart disease as well as predict the overall risk of whether the patient has a 10-year risk of future coronary heart disease (CHD).

# 2- Data preparation and cleaning

In this stage, the data preparation is done, which includes data preprocessing. Some of the steps that we will follow in data preprocessing are given below.

* Checking for the column with wrong data types assigned and fixing typos errors.
* Checking for duplicate values in the dataset. To identify duplicate values duplicated function is used, and then we remove it.
* Checking for the missing values in the dataset and imputations on missing values.
* Check for anomalies in the dataset and remove them. To identify the anomalies, a boxplot is used, and then the Interquartile range's value is used to remove them (Ngare and Kennedy, 2019).

With the help of the glimpse function, I can observe that categorical variables are given in integer type. To find out the duplicate values duplicated function is used; the dataset contains zero duplicated values. To find out the missing values, the data explorer library is used. Several columns contain missing values, such as glucose, education, bpMeds, totChol, cigsPerDay, BMI, and heart rate. The mice library was used for computing the missing values.

To find out the outliers box plot was made. From the below figure, you can see the columns like BMI, glucose, totChol, cigsPerDay, diaBP, heart rate, and sysBP. Most of the columns were having outliers.

Diagram

Description automatically generated

Figure 1- Boxplot of numerical column

After that, the outlier values were stored in a result variable, and then we removed them from the dataset. After the data preprocessing was done, then I moved to the data analysis section.

# 3- Exploratory data analysis

In this section, EDA was conducted to uncover the hidden pattern and extract important features from the dataset. The data analysis helps improve the analytical process's transparency (Ho Yu, C, 2010).

In this project, some of the steps taken in doing EDA are given below.

1. Performing univariate and bivariate data analysis on the numerical and categorical variables with the help of scatter plots, density plots, bar charts, and many other graphs were used.
2. A correlation plot was used to check the relationship between the variables.
3. The Shapiro test is used to check the normality of data distribution.
4. I am using the unsupervised learning technique like principal component analysis for doing the exploratory data analysis.

## 3.1- Univariate data analysis

**For Numerical features**

The histogram was made for univariate analysis to analyze the numerical columns. The graphs of the numerical column are shown down below.

Chart, histogram

Description automatically generated

Figure 2- Histogram for the numerical columns

Some of the insights that I have obtained from the above figure are given below.

* Most people have an average age lies in the range of 40 to 50. There are few people whose age is greater than 65.
* The average Body Mass Index (BMI) is around 25 to 26. Few people's BMI exceeds over 35. You can observe that the BMI column does not follow the normal distribution.
* On average, nine cigarettes are consumed in a single day. But few people consume more than 20 cigarettes every single day. From the graph, you can observe that the values are skewed to the right.
* The average Diastolic blood pressure (diaBP) is 82, and the highest diabP is 142. Similarly, the average systolic blood pressure (sysBP) is 128, and the highest is 295.
* The average glucose level in the patient's body is 78, the minimum glucose level is 40, and the maximum glucose level is more than 150.
* In the heart rate column, most of the value is under 80. Some of the patients have more than 140 beats per minute. And at last, the average cholesterol level is 234, and the maximum is more than 350.

**For Categorical features**

To analyze the categorical column bar plot is used. The figure is given down below.

Graphical user interface, diagram

Description automatically generated

Figure 3- Barplot of the categorical columns

There are few numbers of patients who are on BpMeds or have diabetes or prevalent stroke. From the bar plot, you can observe that most of the columns have imbalanced numbers of samples between the classes. So before applying any machine learning model, you have to balance the classes first.

## 3.2- Bivariate data analysis

To perform bivariate analysis scatter plot was used. It helps us to observe the relationship between variables. With the help of the ggplot library scatter plot was made.

Chart

Description automatically generated

Figure 4- Scatter plot of variables

From the graph, you can observe that when the person's cholesterol level increases, the probability of having heart disease also increases. Similarly true for all the other variables, and another insight that I have found is that males are more likely to suffer from CHD than females.

The Pearson coefficient is used to make the correlation plot to examine the relationship between the variables. The below figure depicts that the systolic blood pressure (sysBp) features are correlated with the prevalentHyp and diaBP. You can also observe that males are more likely to smoke than females. Age feature has also shown a medium correlation with sysBP. There is also a medium correlation of the target variable with respect to age, and there is no relationship between smoking and age. Also, there is no correlation between education with all the given features.A picture containing chart

Description automatically generated

Figure 5- Correlation plot

## 3.3- Checking the normality of distribution

There are many statistical tests that assume that data is normally distributed. To check this assumption, I have performed the Shapiro Wilk test. If the significance level or p-value is greater than 0.05, then the data follows the normal distribution. To implement this test dplyr package was used. From the results, we can conclude that all the numerical column p-value is less than 0.05, which means the numerical columns do not follow the normal distribution.

## 3.4- Using unsupervised learning algorithm for doing EDA

In this section, exploratory data analysis is done with the help of K means clustering. To determine the optimal no of clusters, a plot was made to visualize the average silhouette, and a within-cluster sum of square statistics was used. From the below plot, you can observe that the optimal no of the cluster is 2.

Chart, line chart

Description automatically generated

Figure 6- Plotting the optimal no of clusters

Chart

Description automatically generated

Figure 7- Cluster plot

In the above plot, you can see that most of the points of other clusters overlap with each other, and the silhouette width of the second cluster is greater than the first cluster.

# 4- Machine learning prediction

## 4.1- K nearest neighbors

The k nearest neighbors is a supervised machine learning algorithm that is easy to implement and simple to solve both the classification and regression problems. The simple idea behind KNN is that it assumes that similar data exists in a closed form and near to each other like neighbors. Therefore, it works on gathering the nearby data points and works on this assumption in those cases that the algorithm proves to be useful. To implement this logic, the KNN algorithm needs to calculate the distance between two data points. KNN is robust to a dataset containing noise and is considered more effective when the training data is huge. However, the value of K always needs to be determined.

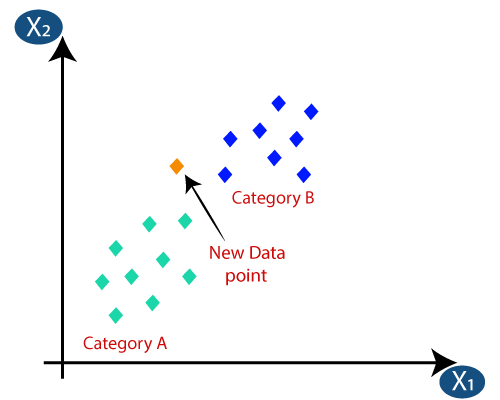


Figure 8: How is a new data point categorized. (Javatpoint, 2017)

The first step in KNN is to calculate the distance between the query example and the current example from the data and then add the distance and index of the example to the ordered collection. Once the ordered collection is achieved, it needs to be sorted among the indices in ascending order of distance (Harrison, 2018). Then from the sorted indices first K entries are to be selected. For the KNN classifier problem, the mode of the K labels is fetched as output.

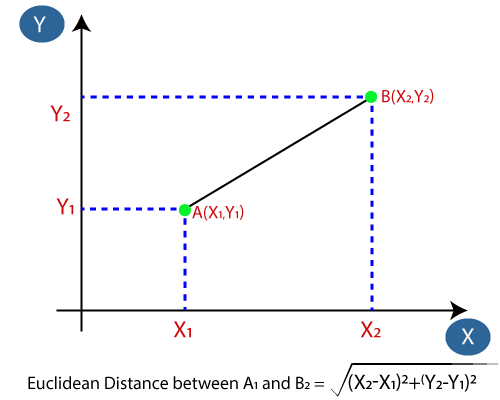


Figure 9: Calculating the distance between two points. (Javatpoint, 2017)

Choosing the right value of K is important while working with KNN, and therefore to choose an appropriate value of K, it is possible that we might need to run the algorithm several times and choose the value which reduces the number of errors encountered while maintaining the algorithm’s ability to make the predictions. As the value of k decreases to 1, the predictions become less stable, so if the value of K is increased, the predictions will become more stable.

One of the applications of KNN is a movie recommendation system in which KNN computes the distance between the target movie and all the available movies which are present in the database, and it outputs the top K recommended movies to the user.

## 4.2- Implementation of KNN

* After the exploratory data analysis is completed, the next step is to normalize the numerical column for that feature scaling is used.
* The second step is to split the dataset into training and testing in the ratio of 75:25.
* The third step is to implement the KNN algorithm. But before that K value is computed with the help of an elbow plot. I have observed that K \_\_\_ gives the minimum error.
* The fourth step is to call the KNN model with parameters like k, distance measure, and train on the training data.
* The model is evaluated on the testing data with the caret Package.
* The target column is imbalanced in nature, so to resolve this issue, downsampling was used. It reduces the no of training samples that falls under the majority class. At k=41, it gives the best accuracy.

# 5- High performance computational implementation

It stands for high-performance computations implementation. It is very beneﬁcial when we have to solve complex Computational problems like predicting accurate diagnosis for diabetes patients. Due to parallel processing, the speed is relatively high. One of the applications of the HPCI technique is fraud detection.

## 5.1- Implementation

* First, pyspark and findspark are installed in the Google Colab environment through the pip command.
* Second, I have created a Spark session with the help of SparkSession.builder() and set the app name as Colab.
* Third, spark.read.csv is used to read the CSV files and set the header to True, and printSchema() is used to get the schema of each column in tree format. From the tree, you can observe that most columns are in string format.
* Fourth, in this, you have to drop the irrelevant values from the dataset created at the time of loading data.
* Fifth, all the columns have string types, so to correct the data types, I have defined a list of the numerical column and categorical columns.
* Sixth, the dataframe is converted to pandas, and I have encoded the categorical column except for the target variable. Then feature scaling is used.
* Seventh, a vector assembler combines all the input columns and creates a single vector column named features.
* Eighth, StringIndexer is used to encode the target column, which is TenYearCHD, and renamed it as a target. Both vector assembler and stringIndexer are added to the pipeline, created at the seventh step.
* Ninth, the dataset is split into training and testing in the ratio of 75:25.
* Some of the following steps have been taken to implement the K nearest neighbors. First, to reduce the dimensions of the features, principal component analysis is used; after that, extracted features are converted to an array and broadcast to each node. After that KNN algorithm is implemented, which sorts the matrix by row and outputs the first k labels. The label appears the most as the predicted label for the test point.
* Evaluate the model using the metrics accuracy, precision, recall, and f1 score.

# 6- Performance evaluation and comparison of the methods

Now it's time to evaluate the models. For this work, classification metrics like confusion matrix, Accuracy, Precision, and recall were used.

The KNN algorithm performed with K value set to 10. It gives an accuracy of 88.41% on the test dataset, the recall value of model is 1 it means the model is correctly detecting all the data points whereas the true negative rate (specificity) is very low around 0.13 which means our model is too eager to find the positive result.

The target variables are imbalanced. So, to tune the parameters trainControl method was used in which 5-fold cross validation was used to balance the target class down sampling. The KNN model runs for 20 different K values and finds that the k=29 model performs the best. It gives an accuracy of 86.7% on the test dataset.

Table 2-Comparison of machine learning algorithms

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithm | Specificity | Recall | Accuracy | Time taken |
| K nearest neighbors (K=10) | 13.60% | 100% | 88.41% | 0. 09sec |
| KNN with parameter tunning | 2.40% | 100% | 86.91% | 17 sec |
| KNN(HPCI) | 0% | 49.6% | 85.11% | 0.00011 sec |

# 7- Discussion of the findings

## 7.1- Discussion

CHD is a common disease and often leads to death due to various complications. In this work, we identify the most crucial factors that lead to CHD and predict the patient's overall risk. Features like age, prevalentHyp, sysBP are the major factors in predicting the coronary heart disease. To find the most important factors with respect to the targeted variable a correlation heatmap was made. After that, a predictive model was made to predict CHD among patients. The target column is imbalanced, so the oversampling technique is performed. Then the KNN model is used which achieves an accuracy of 88.41% with the k value set to 10. The model is tuned and 5-fold cross-validation is used which achieves an accuracy of around 87% on the testing dataset.

For the HPCI same algorithm is also used and it achieves lower performance than the traditional knn algorithm with an accuracy of 85.11%. This model also takes less time as compared to the machine learning model.

7.2- Limitations

There are only a few features are given for the patient. Therefore, to improve, you can also add additional features like medications history and heart imaging. You can also use an autoencoder for feature extraction and then apply a classification algorithm to improve the performance. I have used only a machine learning model on the imbalance data in this work.

# 8- Data management plan and Author contribution statement

# References

captainozlem (2021) *Framingham\_CHD\_preprossing\_data*, *Kaggle.com*. Kaggle. Available at: https://www.kaggle.com/code/captainozlem/framingham-chd-preprossing-data/data (Accessed: April 4, 2022).

Ngare, Kennedy. (2019). Heart Disease Prediction System.

Ho Yu, C. (2010) “Exploratory data analysis in the context of data mining and resampling,” *International journal of psychological research*, 3(1), pp. 9–22. doi: 10.21500/20112084.819.

Harrison, O. (2018) ‘Machine Learning Basics with the K-Nearest Neighbors Algorithm | by Onel Harrison | Towards Data Science’, *towardsDatascienc*, pp. 1–16. Available at: https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761.

Javatpoint (2017) ‘K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint’. Available at: https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning.

# Appendix