Heart Attack Risk Prediction

Yadnesh Mankame L00179080

Abstract

Nowadays heart attack is increasing because of the daily lifestyle. The report represents the approach to predicting heart attack risk based on the medical as well as daily routine of a specific person. The prediction is based on features like the age of a person, what will be cholesterol level, blood pressure, the count of heart rate, if there is any history of a person's family who has heart-related issues then there are many chances of a heart attack risk, does the particular person smoke, do they consume alcohol, what are weekly exercise hours, regular diet can also play an import role, do the person facing some heart-related issues, is there any heavy medication use, what is the stress level, higher body mass index can cause the risk more, if the exercise hours per week is low then the another feature can check is there any physical activity days per week so that it can get balanced and the risk will get minimise, sleeping hours is also an important factor this is because if there is no enough sleep then it may increase stress level and can be at a higher risk.

1 Introduction

The risk of heart attack plays a serious role in a person's life. It can be caused by family history, daily routine, or various factors like medication. The main aim of this report is to determine the relationships between various factors that may cause the risk of heart attack. In this project, the prediction of heart attack risk is generated based on age, sex, cholesterol level, blood pressure, heart rate, family history, smoking(Yes/No), alcohol consumption(Yes/No), exercise hours per week, regular diet, previous heart problems, medication use, stress level, body mass index, physical activity days per week, sleep hours per day based on those various factors model will predict is there any risk of heart attack or no.

Age and sex are the two main factors that can be referred to the health of the heart and blood vessels, which include coronary heart disease, stroke, heart failure, heart arrhythmia, and heart valve problems. If diabetes increases there will be more risk of heart disease, which highlights the significance of controlling this metabolic disorder.

Family History plays an important risk factor for cardiovascular diseases(Heart-Related Problems). Daily lifestyle routine can also impact cardiovascular diseases which can include smoking, alcohol consumption, exercising regularly, and following certain dietary unhealthy guidelines. Additional factors that improve risk assessment include Body Mass Index, medication use, stress level, and past heart-related issues. This report takes a

comprehensive approach, utilizing machine learning methods to create predictive models that incorporate these various characteristics. The objective is to develop the best model that can predict the heart attack risk of a particular person with the help of various features present in the data set.

It is getting more and more important to identify and manage heart attack risk factors with the rise of cardiovascular diseases. With the help of this report, we can improve our knowledge of the complex relationships between different characteristics and heart attack risk. With the help of this research, we can prevent heart attacks by taking the actions that are required.

2 Data Set Description with EDA

2.1 Age with Heart Attack Risk:

Need to know at which age the risk of attack is higher the figure represents a bar plot on the X-axis it shows the age and on the Y-axis it shows the count of Heart Attack Risk(HAR) at a particular age.

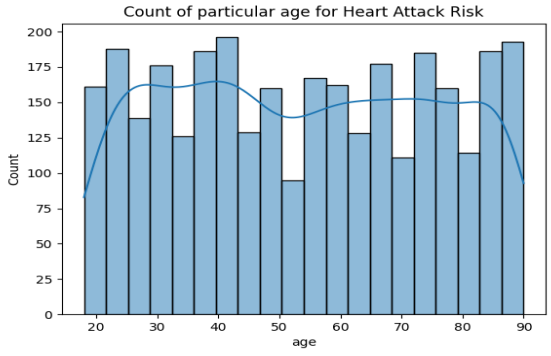

Fig 1: Age with Heart Attack Risk

In the **Fig 1** we can see that the HRS is higher between 40 and 50 age which is around 200

2.2 Diabetes against diet plan:

In this section, we can get to know how daily diet can affect diabetes which means if the diet is not proper does that affect diabetes, or else will get to know further which attribute affects more.

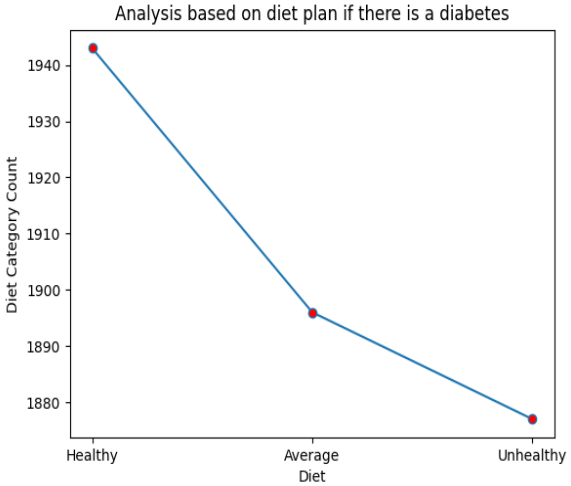

Fig 2: Diabetes based on diet plan

In the above diagram which is **Fig 2**, we can clearly see that the diet is healthy but still the diabetes count is higher and for an unhealthy diet the count is lower.

2.3 How alcohol consumption can affect diabetes:

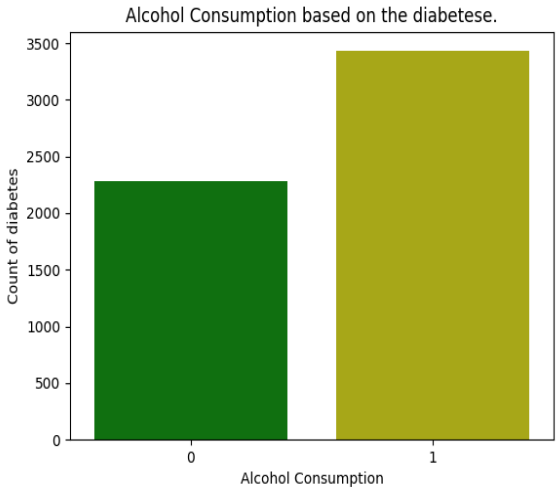

Fig 3: Diabetes based on alcohol consumption

In **Fig 3** '0' means no to alcohol consumption and '1' means yes for alcohol consumption. The count of alcohol consumption is higher if there is diabetes which means alcohol affects more diabetes.

2.4 Attack risk against diet plan:

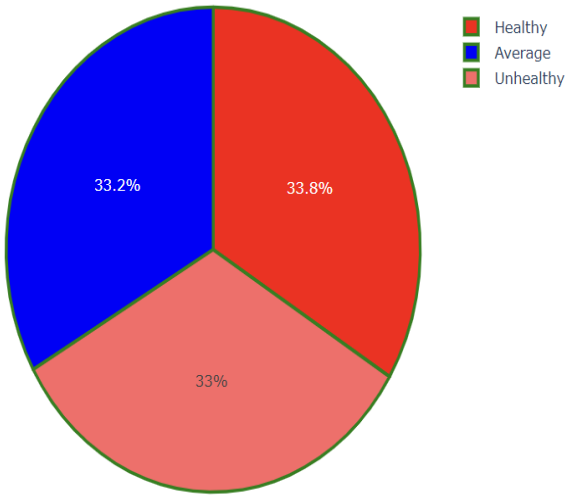

Fig 4: Attack risk with diet plan.

Fig 4 shows the percentage of each category of diet who have a risk of heart attack. There is not a lot of difference in the comparison of those diet plans, each diet plan affects the same therefore based on these analyses we can clearly say that the diet doesn't affect the risk of heart attack.

2.5 Attack risk for specific gender

As stated earlier age and sex are the two main factors that can affect HRS based on the analysis the **Fig 5** shows the count of a specific gender as per the HRS.

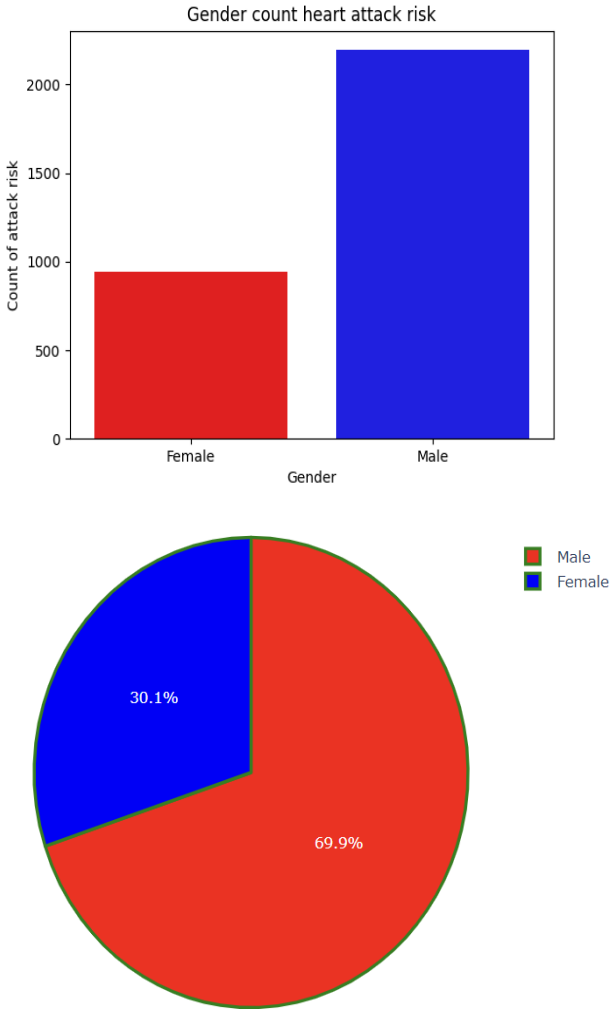

Fig 5: Attack risk with gender category.

The count of males is higher than the count of females category which means that the male category is required to take more precautions as needed.

3 Methods

3.1 Data Extraction

My platform is Google Colab(GC), in that my very first step is to import the required libraries like pandas with the alias name pd to import the data sets for further use to fulfill the requirements, alias is the name that we can use further in the code an important thing is to give the meaningful alias name. The other libraries for visualization are seaborn with alias sns, matplotlib.pyplot with alias plt and plotly.express with the alias px. In the GC we can extract the data from the drive by importing the drive from google.colab.

The next step is to extract the data set in the variable which will be a Data Frame(DF) by using pandas from Google Drive(GD) by using path.

3.2 Data Cleaning:

There are no null values in the dataset so no need to worry about the null values. Dropped not required columns by using the drop function, transformed the columns to lowercase, and cleared the columns that contained blank spaces and special characters for a better look if we want to upload the data into the database then there will be no issue. Split the Blood Pressure(BP) column with the help of slash, by using the split method and lambda function as the data before slash is the systolic BP, and the data after slash will be diastolic BP. Select the columns in the serial order as required. After describing DF it shows some values as NaN that is because the data type of a particular column is an object.

3.3 Feature Engineering

Feature engineering plays a vital role in building and implementing machine learning models. Features will be an input to a model to train and get the best prediction here the features are age, sex, cholesterol, blood pressure,

heart rate, diabetes, family history, smoking, alcohol consumption, exercise hours per week, diet, previous heart problems, medication use, stress level, body mass index, physical activity days per week, sleep hours per day and heart attack risk. The Ordinal Encoder converts a string to numeric values based on the values present within the column.

3.3.1 Feature Transformation

Diet, sex, and blood pressure need to be converted into an integer or float data type for this transformation I used an Ordinal Encoder.

3.3.2 Feature Reduction

Looking at the volume of the data given with too many features it is recommended to use the dimension reduction method. The number of features associated with the dimensionality of the dataset, and the challenges we face with the high dimensionality is the "curse of dimensionality". Two main issues with high dimensional data are data sparsity and distance concentration. Data Sparsity: With increasing features, different no of combinations required to train our data also increases, lesser combinations can lead to overfitting and lack of generalizability.

Distance concentration: As the number of features increases, the distance between all pair of points decreases, which makes it difficult to judge if two observations are more or less similar to each other.

Principal component analysis (PCA) (Dimensionality reduction technique)is an unsupervised learning algorithm that transforms high-dimensional data into a smaller number of features using principal components (PCs). The goal is to preserve as much variance and information from the original data in a lower-dimensional space.

We implement PCA on standardized data. In PCA we get principal components which are eigenvectors, and the variance ratio is given by eigenvalues which are described in descending order, a component with higher variance comes first in the list.

We are training our correlated data set with PCA and taking a number of components = 4 after that, we are splitting our dataset into 80:20 data splits for training and testing respectively.

```
pca = PCA(n_components=4)

X_pca = pca.fit_transform(X)

X_train,X_test, y_train,y_test = train_test_split(X_pca, y, test_size=0.20,random_state=0)
```

Fig 6: Train Test Split Dataset

In PCA we get principal components which are our eigenvectors, and the variance ratio is given by eigenvalues which are described in descending order, so that the component with higher variance comes first.

3.3.3 Import required libraries

```
from sklearn.preprocessing import OrdinalEncoder
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LogisticRegression
import sklearn.metrics
from sklearn.naive_bayes import GaussianNB
from sklearn.decomposition import PCA
from sklearn.metrics import confusion_matrix
from sklearn.preprocessing import StandardScaler
```

Fig 7: Import Required Libraries and Algorithms

For the model training, we have to import the required libraries as mentioned in **Fig 7**.

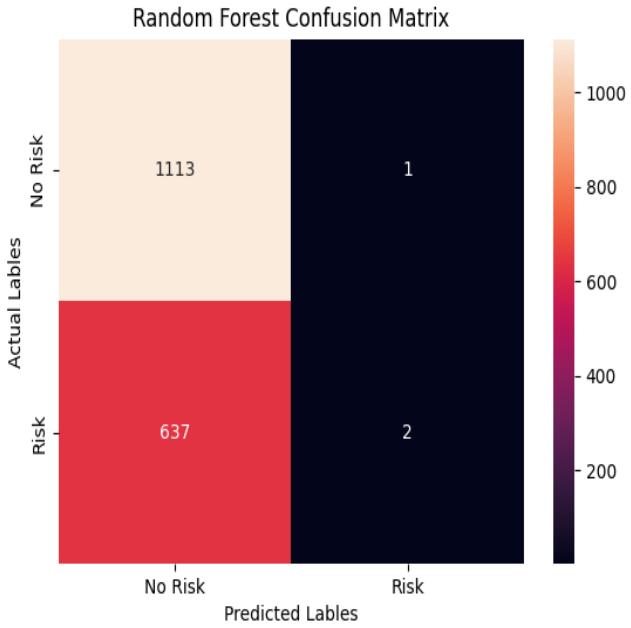

Fig 8: Random Forest Confusion Matrix

3.3.4 Prediction Model Implementation

Random Forest Classifier(RFC): Random forest and Ensemble learning method: I am applying random forest to avoid issues like overfitting and high variance. Bootstrap Agrregating Bagging()is an ensemble learning method, that trains multiple base models using a subset of the dataset with a subset of features independently of each model parallelly, in which a subset of the dataset is taken randomly. In a random forest, different decision trees(base models) are called to classify new points, and each of these trees reports its classification, and random forest returns the most popular classification and average in case of regression, this concept of voting is called aggregation. Based on this model implementation **Fig 8** shows a confusion matrix Random Forest Classifier.

Logistic Regression(LR): To predict the probability of a HAR which is a categorical dependent variable based on given independent features variables that are based on health, and dietary results. Logistic regression uses the logistic function to transform the input variables into a probability value between 0 and 1, which represents the likelihood of the dependent variable being 1 or 0 which means true or false. **Fig 9** shows a confusion matrix Logistic Regression.

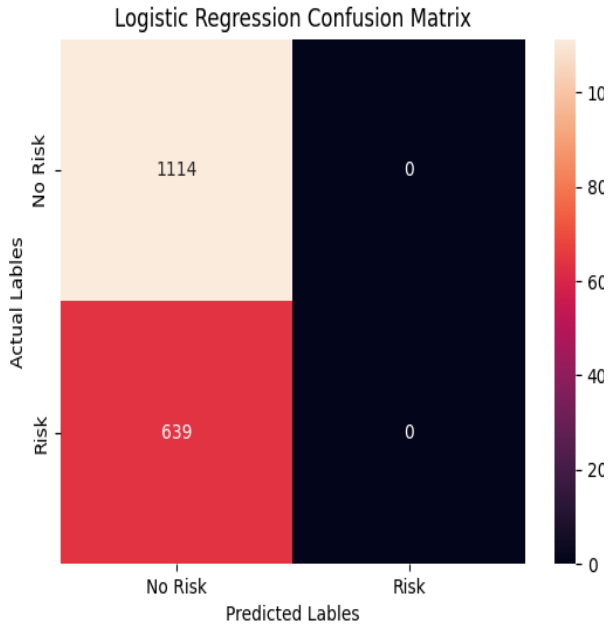

Fig 9: Logistic Regression Confusion Matrix

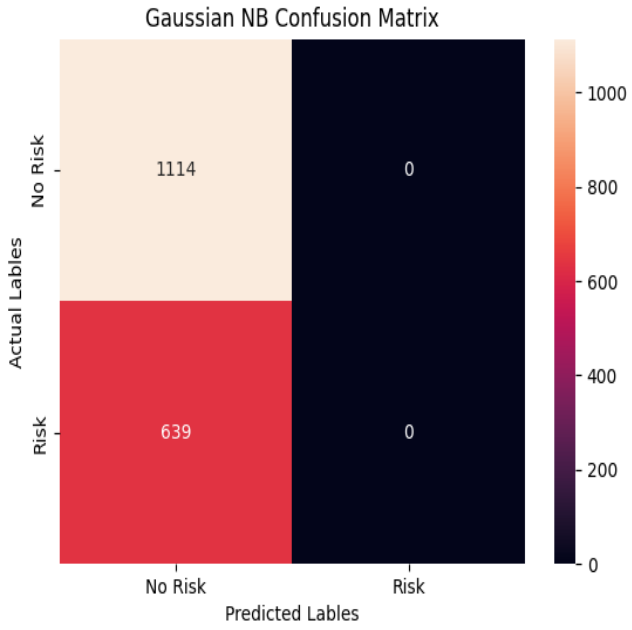

Fig 10: Gaussian NB Confusion Matrix

Gaussian NB(GNB): If the data is discrete means we have to predict discrete values then we don't have to apply the Gaussian Naive Bayes Classifier(GNBC). If the features are a type of continuous variables and model the probability of those variables then we have to use GNBC. After feature engineering, we have features that are converted into continuous values so from those we have to identify which features belong to which class means belongs to HRS which is denoted by '1' or there is no risk which belongs to '0' **Fig 10** shows a confusion matrix for Gaussian NB.

3.3.5 Hyperparamater Tuning

Grid Search CV: We are implementing Grid Search CV which is a popular Hyperparameter tuning technique to find the best set of parameters for a machine learning model, we configure hyperparameter settings before training out data. In Grid Search we create a grid with each value of maximum of depth that we want to test down the left and each value of learning rate across the top the intersection square of each of these is a model that we need to run, running a model for every cell in the grid with the hyperparameter is known as Grid Search. In Grid Search CV and after applying grid search we also evaluate the model performance for each combination using cross-validation.

3.3.6 Voting Classifier

A Machine Learning(ML) model known as a voting classifier(VC) is trained on a large ensemble of models and predicts the output class which means the highest appearance of the desired class. VC compares the results of every classifier here we are feeding RFC, LR, and GNB from those three VC will predict the output class according to voting largest majority.

4 Conclusion

This report investigates the risk of heart attack based on various features and algorithms which predict the value. There is an issue in the data as I tried to apply PCA but need to figure out how many components which I have to take it is showing me a straight line within a graph. The count of class imbalance means that the count of no risk is higher than the risk of attack. By using EDA(Exploratory Data Analysis) get the result of various questions like what is the exact heart attack risk(HAR) age? Another analysis is the count of diabetic patients based on the dietary plan and at which age the risk of diabetes will play a role. Do the exercise hours matter to reduce the HRS? To answer those questions I used libraries like matplotlib, seaborn, etc. But while training the models on the data set I got to know that the feature that contains cholesterol details affects more to the prediction of the HRS. Even after trying some hyperparameter tuning by using Grid Search CV the prediction percentage is not increasing which is 64 percent. Further from this report, I would like to say that daily dietary as well as sex and the age affect the HRS so need to get the precautions of our health.

5 References

[**1**] Cardiovascural health(25th November 2023)

[**2**] Codecademy(29th November 2023)

[**3**] GeeksForGeeks(1st December 2023)