HEART DISEASE PREDICTION BASED ON PERSONAL KEY INDICATORS

by

D.M.K.M. Dissanayake

dmkasun22@gmail.com | kasund@uom.lk

A project report submitted in partial fulfilment of the requirements for the

Machine Learning Foundations program conducted by

Data Science Academy

ABSTRACT

Heart diseases are the leading cause of death globally according to the World Health Organization (WHO) [1]. According to the same source, an estimated 17.9 million people died from heart diseases in 2019, representing 32% of all global deaths that year. Preventing heart diseases through early recognition and healthy lifestyles can bring the above-mentioned numerical figures to a more reduced amount. Especially in recognizing the risk at early stages and altering the lifestyles to become more healthier and by doing so mitigating the risk. A machine learning model is developed to predict the likelihood of a person get confronted with a heart disease based on his or hers provided specific personal information. The report mainly discusses about above-mentioned machine learning model and the techniques which have been used to build it, milestones surpassed before escorting the data to build the model, strategies and approaches used to evaluate the model and measures taken to develop the total effort to a comprehensive high-level application.

INTRODUCTION

Heart disease may occur due to numerous reasons. However, some lifestyle factors and medical conditions can substantially increase the risk. High blood pressure, high cholesterol, smoking, a high intake of alcohol, overweight and obesity, dietary choices and low activity levels are some of the bunch. The recognition of the likelihood of encountering a heart disease can actually help to alter the aforementioned habitual lifestyles and avoid those medical conditions.

Therefore, as a convenient method to get a likelihood prediction of having to confront a heart disease based on some key personal indicators can be extremely useful in recognition of the need to change above mentioned habitual lifestyles and avoid medical conditions.

Here the model is presented as a binary classification model since the actual prediction is to whether the user confront a heart disease or not. Although the output contains a score of likelihood of getting confronted with a heart disease as a probability.

DATA USED IN BUILDING THE MODEL

In order to build the subjected machine learning model, a relevant dataset was used. It contained the needed details regarding personal key indicators for heart diseases such as high blood pressure, high cholesterol, smoking, diabetic status, obesity (high BMI), not getting enough physical activity and drinking too much alcohol.

The dataset consists a total of 18 columns and 17 of them represent key indicators and risk factors for heart diseases. The dataset originally comes form the Centres for Disease Control and Prevention (CDC) and is a major part of the Behavioural Risk Factor Surveillance System (BRFSS), which conducts annual telephone surveys to gather data on the health status of U.S. residents. The dataset is available free of charge on Kaggle website [2].

Following table 1 presents the column names and a description of each column of the dataset which has been used to develop the subjected machine learning model.

Table 1 Column Descriptions

Column Name	Description
HeartDisease	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
BMI	Body Mass Index (BMI)
Smoking	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]
AlcoholDrinking	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)
Stroke	(Ever told) (you had) a stroke?
PhysicalHealth	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? (0-30 days)
MentalHealth	Thinking about your mental health, for how many days during the past 30 days was your mental health not good? (0-30 days)

DiffWalking	Do you have serious difficulty walking or climbing stairs?
Sex	Are you male or female?
AgeCategory	Fourteen-level age category
Race	Imputed race/ethnicity value
Diabetic	(Ever told) (you had) diabetes?
PhysicalActivity	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job
GenHealth	Would you say that in general your health is...
SleepTime	On average, how many hours of sleep do you get in a 24-hour period?
Asthma	(Ever told) (you had) asthma?
KidneyDisease	Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?
SkinCancer	(Ever told) (you had) skin cancer?

METHODOLOGY

Classical machine learning model building approach has been taken in providing the solution. Development, evaluation and deployment of the model, all have been done using Jupyter kernel running Python 3.10.

Exploratory Data Analysis

Using the ‘describe’ function, 4 numerical columns have been identified. They are ‘BMI’, ‘PhysicalHealth’, ‘MentalHealth’ and ‘SleepTime’. Distributions of those numerical columns have been observed using histogram plots and kernel density estimate (KDE) plots. KDE plots for the above-mentioned numerical columns have been drafted against the ‘HeartDisease’ column to get an understanding about the distribution relation to one-another.

Following figures (Figure 1 to 4) shows some of the depicted graphs used for the exploratory data analysis.

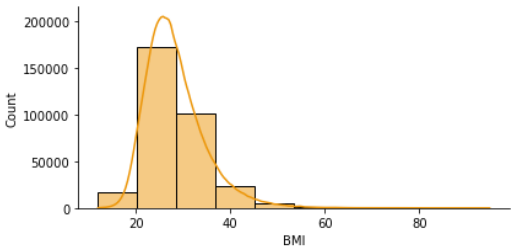

Figure 1. Histogram of BMI

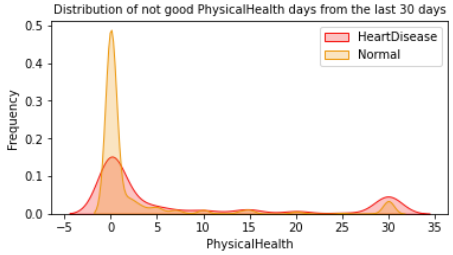

Figure 2. Distribution of 'PhysicalHealth'

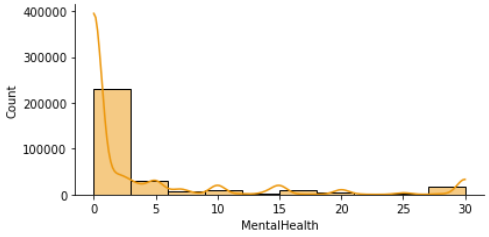

Figure 3. Histogram of MentalHealth

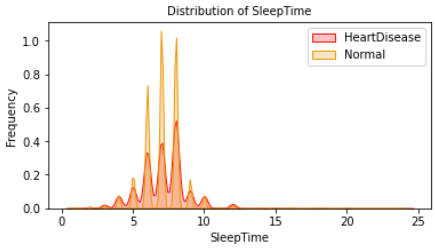

Figure 4. Distribution of 'PhysicalHealth'

For the analysis of the categorical columns, histograms have been observed for each column against the 'HeartDisease' column. Following figure 5 displays those histograms.

Figure 5. Histograms of Categorical Columns

In order to understand about the correlation of the numerical columns, a correlation matrix and a heat map have been observed. Figure 6 depicts the observed correlation matrix.

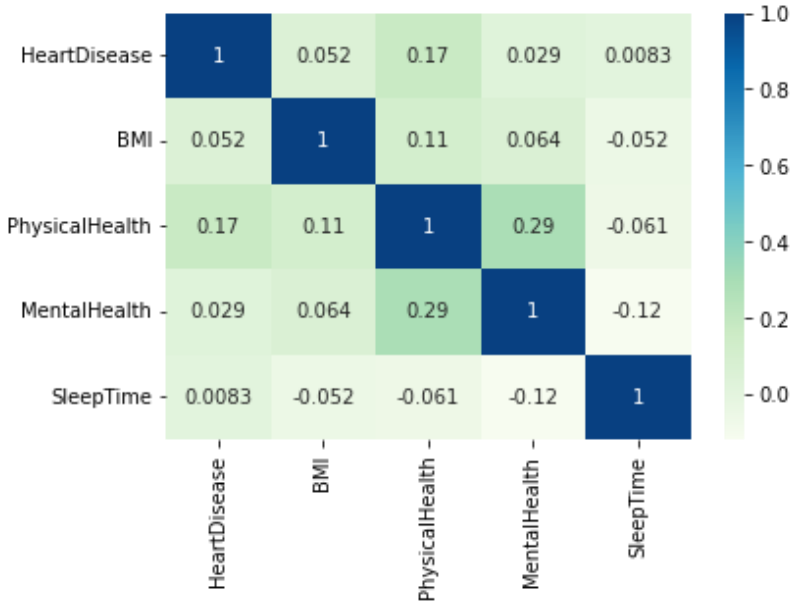

Figure 6. Heat Map

As you can see from the above figure 6, 'PhysicalHealth' column which describes the number of days the patient felt 'not good' physically has the highest correlation with the 'HeartDisease' column.

Feature Engineering

One hot encoding was done to the binary categorical columns initially then followed by other multi-categorical columns. In order to create a column for each category within multi-categorical columns, 'pd.get_dummies' function has been used. Then the original multi-categorical columns which contained the categories as rows have been deemed redundant therefore been removed.

Numerical column data then went through a process of standardization in order to remove the bias which can spawn due to the scale difference. 'StandardScaler' function is used to get this task done.

Then the data went through the process of checking for any columns which contained null values. Since there was no evidence of columns contained any null values there wasn't any reason to drop any rows. For binary columns, a data type re-assignment was done to preserve the memory usage.

Model Training

After selecting the X and the y variables, train, test and split method was used to separate the data into two sets. It is a necessary separation regarding evaluating the model. Best to use different data set to test the model than the one used to train it.

A model training function has been defined in order to train the model. It would take the selected model and the data as input and then return the model evaluating metrics such as accuracy, precision, f1score and roc_auc.

To find the best suitable model as the solution, couple of models and different set of hyperparameters for some models were used in training.

Random Forrest Classifier, which was one of the models used was applied a grid search in order to find the best hyperparameters. Although, the best model and the best hyperparameter configuration was selected manually observing evaluation metrics. The selected model is the Random Forrest Classifier model with 500 number of trees and 10 maximum depth hyperparameter configurations which was named ‘rf3’.

Then the selected model was saved using Pickle. Then a function called Score Function was defined to get the probability values for the predictions. Then a post-processing function is defined to get the predicted probability of having a heart disease out of 1.

Then finally App Prediction Function is defined as the final high-level solution which will get the input data containing personal key indicators and predict the probability of having a heart disease as a percentage.

RESULTS

When selecting the best model to use, the selected model as well as all the other models were evaluated using evaluation metrics. Following figure 7 depicts the resulted values.

	model_name	model	accuracy	precision	f1_score	roc_auc
0	lgr1	LogisticRegression(n_jobs=3, verbose=1)	0.914477	0.526646	0.887088	0.837163
1	dt1	DecisionTreeClassifier(min_samples_leaf=2)	0.886542	0.280865	0.877842	0.614661
2	rf1	(DecisionTreeClassifier(max_features='auto', r...	0.903626	0.338404	0.882206	0.784058
3	rf2	(DecisionTreeClassifier(max_features='auto', r...	0.904074	0.346863	0.882892	0.790890
4	rf3	(DecisionTreeClassifier(max_depth=10, max_feat...	0.914633	0.627551	0.877644	0.828468
5	rf4	(DecisionTreeClassifier(max_depth=20, max_feat...	0.914352	0.529699	0.884244	0.831120

Figure 7. Model Evaluation Results

After creating the final App Prediction Function, a local sever was created including the Pre-Processing, Score, Post Processing and App Prediction Functions using a python script to demonstrate the solution. Then using an API request user can get the predicted probability.

CONCLUSION

In conclusion, a successful machine learning model building procedure can be observed throughout this project containing all the essential ingredients including data pre-processing, model training and evaluating, post processing and up to the very last step of developing an inference pipeline. Several machine learning models were built and evaluated in order to find the optimum solution. Even though there was a close result between last two Random Forrest Classifier Methods rf3 and rf2, the precision decided the most suitable model between the two. The solution can be deployed to the users since it has finalized as a high-level application.

DISCUSSION

Even though the other evaluation metrics showed a better result, precision can be improved in our solution. Un-balanced training data may have a say in this, although the 3 out of 4 evaluation metrics giving more than 80 percent results should be taken as a positive.

Un-balanced data issue can be solved using re-sampling (under-sampling or over-sampling) or using deep learning as to build the model

REFERENCES

[1] World Health Organization (WHO), "Cardiovascular diseases (CVDs)," 11 June 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)#:~:text=Key%20facts,to%20heart%20attack%20and%20stroke.. [Accessed 28 April 2022].

[2] Kaggle, "Personal Key Indicators of Heart Disease," Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease?resource=download. [Accessed 10 April 2022].