Personal Key Indicators of Heart Disease

Babasaheb K. Thorve*, Marziya Z. Shaikh

Savitribai Phule Pune University Subcenter Ahilyanagar, Ahmednagar College, Ahilyanagar

b.thorve@gmail.com, mardhiya23@gmail.com.

ABSTRACT

Heart Disease is recognized as a serious health problem nowadays all over the world. A healthy lifestyle should be manageable based on useful or inspiring factors which prevent the risk of heart disease. Personal key indicators responsible for heart disease help us to live a stable and healthy life which decreases its risk. Some of the Risk factors which are responsible for coronary artery disease are Age, Smoking, High blood, Diabetes, Obesity, Lack of exercise, Poor sleep, Unhealthy diet, Family history. If we are able to identify personal key indicators which causes CAD then one can live healthy and stable lifestyle. Our aim is to find the personal key indicators which causes coronary artery disease. We have computed the risk ratio for the groups Alcohol drinkers, Physical activity, Diabetes, Smokers, difficulty in walking among the male and female who are exposed to heart disease and not exposed to heart disease. Oversampling technique is used to make the data balance. For the original data and balanced data, we fit the machine learning tools such as logistic regression, Decision Tree, K-nearest Neighbour and Random Forest. For this model precision, recall and f1-score is calculated. Form these matrices best suitable model is proposed.

KEYWORDS: *CAD, personal key indicators, balanced data, oversampling, Risk ratio.*

How to Cite: Babasaheb K. Thorve*, Marziya Z. Shaikh., (2025) Personal Key Indicators of Heart Disease, *Journal of Carcinogenesis*, *Vol.24*, *No.9s*, 213-221.

1. INTRODUCTION

Heart disease is a variety of issues that can affect your heart. When people think about heart disease, they often think of the most common type — coronary artery disease (CAD) and the heart attacks it can cause. But you can have trouble with different parts of your heart, like your heart muscle, valves or electrical system. When your heart isn't working well, it has trouble sending enough blood, oxygen and nutrients to your body. In a way, your heart delivers the fuel that keeps your body's systems running. If there's a problem with delivering that fuel, it affects everything your body's systems do.

Most of the types the heart disease is due to Narrowing of your heart's blood vessels because of fatty deposits (coronary artery disease), Abnormal heart rhythms (arrhythmias), Heart valve diseases, Abnormal heart muscle (cardiomyopathy), Heart squeezing and relaxation difficulties (heart failure), Heart issues you have at birth (congenital heart disease), Issues with the fluid-filled sac surrounding your heart (pericardium).

Heart disease is the top cause of death in the United States. In 2021, heart disease caused 1 in 5 deaths. That's nearly 700,000 people. Coronary artery disease, the most common type of heart disease, caused about 3, 75,000 of those deaths. Heart disease affects people from most ethnic backgrounds, regardless of gender.

According to the **World Health Organization (WHO)** and according to the NHLBI, NIH (National Institute of Health), unhealthy lifestyle habits can lead to coronary artery disease (CAD). Some of the risk factors for CAD are Age, Smoking, High blood, Diabetes, Obesity, Lack of exercise, Poor sleep, Unhealthy diet, Family history.

Depending on your heart issue, you may need to make changes to your daily life, take medication or have surgery. Possible preventive measures should be Changing of lifestyle, taking medicines, doing surgeries and participation in rehab program.

Motivation of study

According to the CDC (centres for disease control and prevention), heart disease is one of the leading causes of death for people of most races in the US (African Americans, American Indians and Alaska Natives, and White people). About half of all Americans (47%) have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other key indicator includes diabetic status, obesity (high BMI), not getting enough physical activity or drinking too much alcohol. According to many surveys, cardiovascular diseases is on hike and mortality rates are increasing rapidly day by day. In order to live a healthy life, one should be able to recognize indicators or factors which affect the health physically and mentally. Various personal features or indicators play a vital role in detecting causes of heart disease.

A healthy lifestyle should be manageable based on useful or inspiring factors which prevent the risk of heart disease. Personal key indicators responsible for heart disease help us to live a stable and healthy life which decreases its risk. Computational developments, allow the applications of Machine Learning methods to detect patterns from the data that can predict a patient's condition. AI is useful in predicting and fitting the proper model in order to recognize the personal key indicators of heart disease. For feature selection and for predicting the heart diseases, proper Machine Learning models have been established.

Detecting and preventing the factors that have greatest impact on heart disease is very important in healthcare as it also prevent mental deterioration as well financial loss which incurred due treatments of heart diseases. Identifying such factors will help people to detect the risk of heart disease and reduce it, that's why we chose this topic for research.

Data Description

The Dataset named 'HeartDisease_2020' were taken from Kaggle data source and used in study. Data consists of 319795 rows and 18 columns; has been collected from United States of America. The data is taken from https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease. There are 5756310 total data points.

There are 18 variables (11 binary, 3 strings and 4 numerical) with 18 columns and 319795 rows.
1) Heart Disease: Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI). Binary (Yes or No).
2) BMI: Body Mass Index. If your BMI is less than 18.5, it falls within the underweight range. If your BMI is between 18.5 to 24.9, it falls within the normal or Healthy Weight range. If your BMI is 25.0 to 29.9, it falls within the overweight range. If your BMI is 30.0 or higher, it falls within the obese range. Numerical
3) Smoking: Have you smoked at least 100 cigarettes (5 packs = 100 cigarettes) in your entire life? Binary (Yes or No).
4) Alcohol Drinking: Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week). Binary (Yes or No).
5) Stroke: (you had) a stroke? Binary (Yes or No).
6) Physical Health: Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days. Numerical
7) Mental Health: Thinking about your mental health, for how many days during the past 30 days was your mental health not good? Numerical
8) Diff Walking: Do you have serious difficulty walking or climbing stairs? Binary. (Yes or No).
9) Sex: Male or Female.
10) Age Category: Range of patient's age are given. Ordinal
11) Race: races in the US (African Americans, American Indians and Alaska Natives, and white people). Nominal
12) Diabetic: Person is having Diabetes yes, no, borderline, during pregnancy.
13) Physical Activity: Is there any physical activity or physical exercise going on or not, responses are in yes or no.
14) GenHealth: How is the General health or overall health of the patient. Nominal
15) Sleep Time: Duration of hour of sleep is recorded. Numerical
16) Asthma: Is patient is having asthma or not, responded in yes or no manner.
17) Kidney Disease: Is patient is having kidney disease, responded as Yes or no.
18) Skin Cancer: Is patient is having skin cancer, responded as Yes or no.

The response variable over this dataset is heart disease (Yes or No), which is dependent on other factors.
Since the data has 292422 data points / observations having **NO** Heart Disease and 27373 data points **having** heart disease. The classes over here are not balanced means equal number of observations are not parted into two classes (of Yes and No). There is a need to balance the classes in order to get an appropriate interpretation from the techniques applied to it. So, Oversampling (amount of data in minority class is made equal to majority class) & Undersampling (amount of data in majority class is made equal to minority class) Techniques are used to balance the dataset for further Machine Learning Techniques

2. LITERATURE REVIEW

R. Indrakumari et.al. (2020) applied exploratory data analysis techniques to predict and detect the patterns of heart disease data is, as it considered as analysing data that excludes inferences and modelling. EDA applied over to the data set includes Histogram, Bubble plot, Pir diagram, user defined filter. Various machine learning algorithms was applied which includes KNN, ANN, Naïve Bayes, Logistic Regression as they allow a researcher to find which factors influence heart disease.
In order to improve the performance and interpretability of classification algorithm, the continuous inputs is been transformed into distinct groups or bins. As the data is being made categorical for better interpretations, KNN does not work properly for it. However, to overcome this limitation, Chintan M. Bhatt et.al. (2023) applied K- modes algorithm. The accuracy after applying hyperparameters. Grid search CV, was for XG Booost(87.02%), multilayer perceptron (86.94%), random forest (86.92%) and decision tree (86.53%).

Edward Leonardo et.al. (2024) applied various ML models to detect which preventive measures can be taken. In order to achieve the balanced class distribution they used oversampling, under sampling and combined sampling. Gradient Boosting with undersampling gives better model having accuracy 74%.

Xuanye Wang (2022) made use of data preprocessing techniques such as converting the categorical values to numerical with the help of One Hot encoding. Sampling techniques such as oversampling, Under sampling is being used to make class distribution balanced. Before sampling the recall was 3% . This means that out of 100 people the model would recognise only 3 people having heart disease. In oversampling recall was increased to 82% and for under sampling it was 83% .

Nayab Akhtar (2021) applied various data visualisations tools on each feature variable. They applied KNN, ANN, Naïve Bayes, Decision Tree, Random Forest machine learning algorithms to the data. Among all those algorithms Naïve Bayes gave highest accuracy of 88%.

Harshit Jindal et.al. (2021) showed with the help of Data visualisations that risk of heart disease is more in 60-70 age of people. Male has high risk and asymptomatic popularity type of chest pain mainly causes heart diseases. They used data mining techniques such as logistic regression, KNN and Random Forest to the data. Among these models KNN fits good to the data with highest accuracy of 88.52%.

Niloy Biswas et.al. (2023) done data preprocessing and also applied feature selection. In classification and modelling they applied logistic regression, SVM, KNN, Decision Tree, Random Forest, Naïve Bayes. Random Forest provided the most optimistic performance with 94.51% accuracy.

Cardiovascular Sentiment Network is a deep learning took that uses personal key factors to forecast heart disease. Fatma M. Talaat (2024) used Probabilistic Neural Network (PNN), Back Propagation (BP) and General Regression Neural Network (GRNN). GRNN had the highest prediction accuracy.

Mohammadreza Hajiarbabi(2024) used various methods such as XGBoost, RD, Ensemble Learning and Neural Network for heart disease detection.

Ahmad Ayid and Huseyin Polat (2023) used Jellyfish algorithm to select the best features from the datasets. They used ANN, DT, Adaboost and SVM algorithm. Jellyfish algorithm gives best model with accuracy 98.4 %.

Methodology

- **One Hot Encoding:** One hot encoding is a technique that we use to represent categorical variables as numerical values in a machine learning model.
- **Risk Ratio:** A measure of the risk of a certain event happening in one group compared to the risk of the same event happening in another group is called as risk ratio. In cancer research, risk ratios are used in prospective (forward looking) studies, such as cohort studies and clinical trials. A risk ratio taking value 1 means there is no difference between two groups in terms of their risk of cancer, based on whether or not they were exposed to a certain substance or factor. A risk ratio having value greater than one or of less than one usually means that being exposed to a certain substance or factor either increases or decreases the risk of cancer. Or in other words the treatments being compared do not have the same effects.
- **Logistic regression Model fitting:** Binary variables are widely used in Statistics to model the probability of a certain class or event taking place. For example such as the probability of a team winning, probability of a patient being healthy, etc.In binary logistic regression there is a single binary dependent variable, coded by an indicator variable,

where the two values are labeled "0" and "1". On the other hand, the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). Logistic Regression is a statistical method that we use to fit a regression model when the response variable is binary. To assess how well a logistic regression model fits a dataset, we can look at the two matrices which are sensitivity and specificity. Sensitivity: The probability that the model predicts a positive outcome for an observation when indeed the outcome is positive. This is also called the “true positive rate.” Specificity: The probability that the model predicts a negative outcome for an observation when indeed the outcome is negative. This is also called the “true negative rate.” One way to visualize these two metrics is by creating a ROC curve, which stands for “receiver operating characteristic” curve. This is a plot that displays the sensitivity and specificity of a logistic regression model.

Decision Tree:
A decision tree is a flowchart-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules. It is important to know the measurements used to evaluate decision trees. The main metrics used are accuracy, sensitivity, specificity, precision, miss rate, false discovery rate, and false omission rate. All these measurements are derived from the number of true positives, false positives, true negatives, and false negatives obtained when running a set of samples through the decision tree classification model. Also, a confusion matrix can be made to display these results.

K-Nearest Neighbors
kNN, or the k-nearest neighbor algorithm, is a machine learning algorithm that uses proximity to compare one data point with a set of data it was trained on and has memorized to make predictions. kNN works off the assumption that similar points can be found near one another — birds of a feather flock together. As a classification algorithm, kNN assigns a new data point to the majority set within its neighbors. As a regression algorithm, kNN makes a prediction based on the average of the values closest to the query point. kNN is a supervised learning algorithm in which 'k' represents the number of nearest neighbors considered in the classification or regression problem, and 'NN' stands for the nearest neighbors to the number chosen for k.

In order to determine which data points are closest to a given query point, the distance between the query point and the other data points will need to be calculated. Distance can be calculated by different methods like **Euclidean distance**, **Manhattan distance**, **Minkowski distance, Hamming distance.**

- **Random Forest**

Random forests or random decision forests is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time. For classification tasks, the output of the random forest is the class selected by most trees. For regression tasks, the mean or average prediction of the individual trees is returned. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forests generally outperform decision trees, but their accuracy is lower than gradient boosted trees. However, data characteristics can affect their performance.

A big part of machine learning is classification — we want to know what class (a.k.a. group) an observation belongs to. The ability to precisely classify observations is extremely valuable for various business applications like predicting whether a particular user will buy a product or forecasting whether a given loan will default or not.

Data Analysis
We Construct multiple bar diagram to study the risk of the heart disease among the different races.

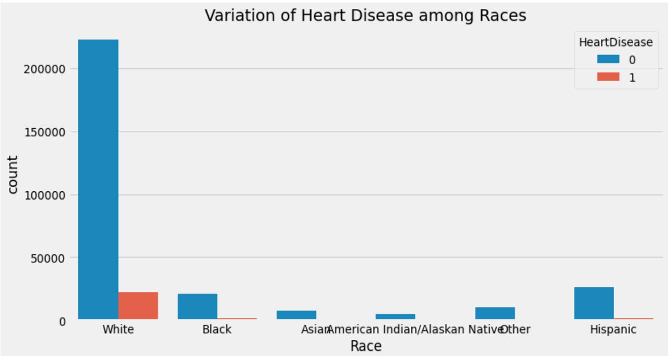


Fig 1: Variation of Heart Disease among races

From the Fig 1, we can infer that the risk of heart disease is more in White people races as compared to other races.

Risk Ratio: In the Table 1, we have computed the risk ratio for the groups Alcohol drinkers, Physical activity, Diabetes, Smokers, difficulty in walking among the male and female who are exposed to heart disease and not exposed to heart disease.

Table 1: Risk Ratio for Various Group

Sr. No.	Groups of exposed and unexposed to heart disease	Sex	Having heart disease Yes/No	Risk Ratio
1	Alcohol drinkers	Males	Exposed to heart disease or no	1.08630
		Females		0.551
2	Physical activity	Males	Exposed to heart disease or no	0.5590
		Females		0.4287
3	Diabetes	Males	Exposed to heart disease or no	3.4692
		Females		3.2577
4	Smokers	Males	Exposed to heart disease or no	2.0218
		Females		1.849
5	Difficulty in walking	Males	Exposed to heart disease or no	3.3874
		Females		4.3687

From the Table 1 we have following conclusions:

1) Females those who drink alcohol are 55% less likely to get the heart disease than the females who do not drink alcohol. Males those who drink alcohol are 108 times more likely to get the heart disease than the males who do not drink alcohol.
2) Females those who exercise (Daily Physical Activity) are 55% less likely to get the heart disease than the females those who don’t exercise. Males those who exercise (Daily Physical Activity) are 42% less likely to get the heart disease than the females those who don’t exercise.
3) Females those who have diabetes are **3.4692** times as likely to get the risk of heart disease than the females who don’t have diabetes. Males those who have diabetes would be **3.2577** times as likely to get the risk of heart disease than the males who don’t have diabetes.
4) Females those who smoke are **2.0218 times** more likely to get the heart disease than the females those who don’t smoke. Males those who smoke are **2.849 times** as likely to get the heart disease than the females those who don’t smoke.
5) Females those who have difficulty in walking are **3.3875 times** as likely to get the heart disease. Males those who have difficulty in walking are 4.3687 **times** as likely to get of heart disease.

Logistic Regression Analysis (in R)(without outliers):

There are some observations in the data that falls outside the usual pattern of the data. Such observations are called as outlier. We first removed the outliers from the data. Then by considering binary random variable HeaartDisease (Yes or No) as a response and Smoking, AlcoholDrinikng, Stroke, PhysicalHealth, MentalHealth, PhysicalActivity, GenHealth, SleepTime, Asthma, KidneyDisease, SkinCancer as dependent variables, we fit logistic regression model to the data.Following is the output of the logistic regression.

```
> summary(fit1)

Call:
glm(formula = HeartDisease ~ Smoking + AlcoholDrinking + Stroke +
    PhysicalHealth + MentalHealth + PhysicalActivity + GenHealth +
    SleepTime + Asthma + KidneyDisease + SkinCancer, family = binomial,
    data = dt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1775  -0.4104  -0.2969  -0.2274   3.2093

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -4.1785018  0.0479349 -87.170  < 2e-16 ***
Smoking             0.5132566  0.0137793  37.248  < 2e-16 ***
AlcoholDrinking    -0.4882044  0.0325298 -15.008  < 2e-16 ***
Stroke              1.3324271  0.0220389  60.458  < 2e-16 ***
PhysicalHealth      0.0085032  0.0008206  10.362  < 2e-16 ***
MentalHealth       -0.0173053  0.0008502 -20.355  < 2e-16 ***
PhysicalActivity   -0.1357359  0.0151423  -8.964  < 2e-16 ***
GenHealthFair       1.9987323  0.0314301  63.593  < 2e-16 ***
GenHealthGood       1.3924884  0.0287432  48.446  < 2e-16 ***
GenHealthPoor       2.4652423  0.0392369  62.830  < 2e-16 ***
GenHealthVery good  0.6779201  0.0298031  22.747  < 2e-16 ***
SleepTime           0.0393274  0.0051041   7.705 1.31e-14 ***
Asthma              0.0582381  0.0182636   3.189  0.00143 **
KidneyDisease       0.8514678  0.0237649  35.829  < 2e-16 ***
SkinCancer          0.6797698  0.0185551  36.635  < 2e-16 ***
---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

 (Dispersion parameter for binomial family taken to be 1)

     Null deviance: 186906  on 319794  degrees of freedom
 Residual deviance: 159894  on 319780  degrees of freedom
 AIC: 159924

 Number of Fisher Scoring iterations: 6
```

3. CONCLUSION:

From the above output, we can see that the p-value of all the features like Smoking, Alcohol Drinking, PhysicalHealth, MentalHealth, PhysicalActivity, GeenHealthFair, GenHealthGood, GenHeathpoor, GenHealthVerygood, SleepTime, Asthma, KidneyDisease, SkinCancer are less than 5% level of significance, indicating that to reject the null hypothesis. Rejecting the null hypothesis means that all the factors in the risk of heart disease are significantly important as per the data.

For the original data we fit the machine learning tools such as logistic regression, Decision Tree, K-nearest Neighbour and Random Forest. For this model precision, recall and f1-score is calculated. These model accuracies matrices are presented in the following table.

Table 2: Model Accuracies of Data (Original Data):

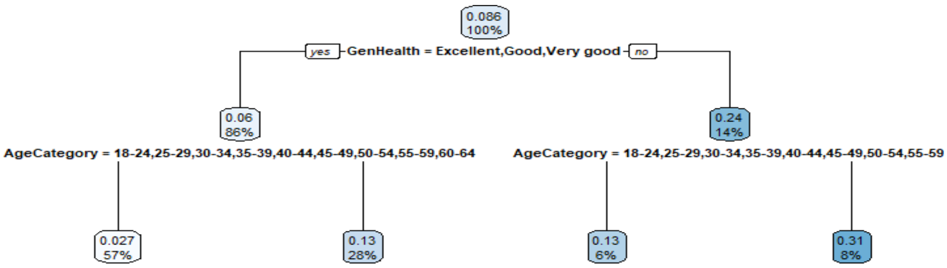
Logistic regression		Precision	Recall	f1-Score
	0	0.92	0.99	0.96
	1	0.53	0.09	0.15
	Accuracy			**0.92**
Decision Tree	0	0.91	1.00	0.96
	1	0.00	0.00	0.00
	Accuracy			**0.91**
K Nearest Neighbour	0	0.92	0.98	0.95
	1	0.32	0.08	**0.13**
	Accuracy			0.91
Random Forest	0	0.92	0.98	0.95
	1	0.34	0.12	0.17

	Accuracy			**0.91**	

Conclusion:

1) From the above table it is observed that the accuracy of the logistic regression is highest among others. So, logistic regression model fits good to the original data.
2) Accuracy score of logistic regression model is **92%**.
3) For logistic regression, True positive rate is 53% and True negative rate is 92% which means logistic regression model has predicted 92% of the data correctly.

Decision Tree by using R-studio:



__**Conclusion**__:

Person's whose general health is good are 86 % and rest whose general health is poor are 14 %. The individual whose general health is good falls in the age category of 18-40 with 57 %, whereas 28 % proportion of the people of age category 40-64 has very good health.

Since the data has 292422 data points / observations having **NO** Heart Disease and 27373 data points **having** heart disease. The classes over here are not balanced means equal number of observations are not parted into two classes (of Yes and No). There is a need to balance the classes in order to get an appropriate interpretation from the techniques applied to it. So, Oversampling (amount of data in minority class is made equal to majority class) & Under sampling (amount of data in majority class is made equal to minority class) Techniques are used to balance the dataset for further Machine Learning Techniques.

After oversampling the data, we fit the different classification machine learning model to classify the heart disease patient. For these models, we compute the different model accuracy matrices and are presented in Table no. 4.

Table 3: Model accuracies of the data (Over Sampling)

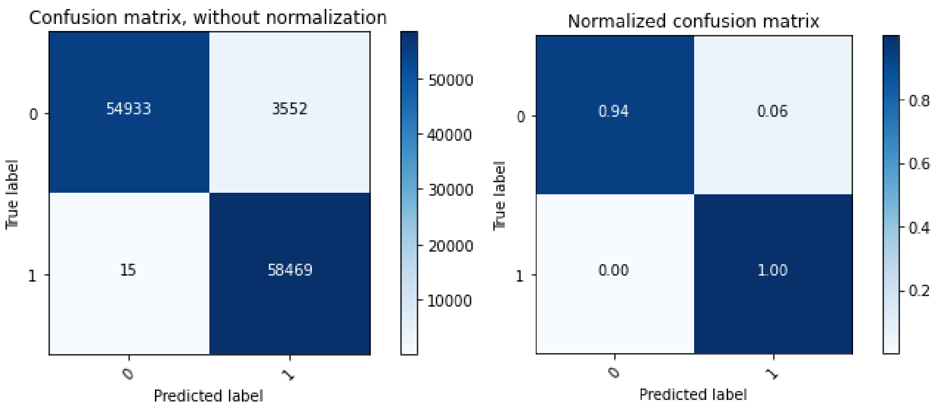
Logistic regression		Precision	Recall	f1-Score
	0	0.76	0.74	0.75
	1	0.75	0.76	0.75
	Accuracy			**0.75**
Decision Tree	0	0.72	0.72	0.72
	1	0.72	0.72	0.72
	Accuracy			**0.72**
K Nearest Neighbour	0	0.99	0.80	0.89

	1	0.83	0.99	0.91
	Accuracy			**0.90**
Random Forest	0	1.00	0.94	0.97
	1	0.94	1.00	0.97
	Accuracy			**0.97**

Conclusion: From the Table 3 following are the conclusions.

1) Accuracy score of logistic regression model when you used under sampling technique to balance the data is **75%** which is highest among other models.
2) So logistic regression model is best fit model to for the purpose of classification in case of under sampled data.
3) For the logistic regression model, True positive rate is 77% and True negative rate is 74% which means logistic regression model has predicted 75% of the data correctly.

Random Forest analysis in Python (over-sampling)



Conclusion:

1) Accuracy score of Random Forest model is 97%.
2) Here True positive rate is 100% and True negative rate is 94% which means Random Forest model has predicted 97% of the data correctly.

<u>Conclusions</u>

1) As per the graph, people among different races – white people have high heart disease.
2) Risk ratio of people of heart disease among exposed group (Smoking, Alcohol drinking, Diabetes and Physical Activity) has been calculated.
3) For the imbalance data logistic regression perform better with accuracy with 92 % accuracy but precision is very low.
4) Since the data was imbalanced, we have balanced the data by applying oversampling techniques. In Oversampled data Random Forest model has highest accuracy of 97 % and for under sampled data, logistic regression performs better with 75 % accuracy.

BIBLIOGRAPHY

[1] R. Indrakumari, T. Poongadi, Soumya Ranjan Jena (2020), Heart Disease Prediction Using Exploratory Data Analysis, Science Direct, 17(2020), 131-139.
[2] Chintan M. Bhatt , Parth Patel, Tarang Ghetia and Pier Luigi Mazzeo(2023), Effective Heart Disease Prediction Using Machine Learning Techniques, algorithms MDPI, 1-14.
[3] Edward Leonardo, Dr. Murugananthan Velayutham, and Justin Gilbert (2024), Data Analysis and Predictive Modelling on Heart Disease based on People's Lifestyle, EAI.EU, 10, 1-6.
[4] Xuanye Wang (2022), Heart Disease Classification Based on Personal Indicators Research Gate,1-8.

[5] Nayab Akhtar (2021) Heart Disease Prediction, Research Gate, 1-11.
[6] Harshit Jindal, Sarthak Agrawal, Rishabh Khera, Rachna Jain and Preeti Nagrath (2021), Heart disease prediction using machine learning algorithms, IOP Conf. Series: Materials Science and Engineering, 1022,1-11.
[7] Niloy Biswas, Md Mamun Ali, Md Abdur Rahaman, Minhajul Islam, Md. Rajib Mia, Sami Azam, Kawsar Ahmed, Francis M. Bui, Fahad Ahmed Al-Zahrani, Mohammad Ali Moni (2023), Machine Learning-Based Model to Predict Heart Disease in Early StageEmploying Different Feature Selection Techniques, Wiley, 1-30.
[8] Fatma M. Talaat (2024), Revolutionizing Cardiovascular Health: Integrating Deep Learning Techniques for Predictive Analysis of Personal Key Indicators in Heart Disease, Research Gate, 1-23.
[9] Mohammadreza Hajiarbabi(2024), Heart disease detection using machine learning methods: acomprehensive narrative review, Journal of Medical Artificial Intelligence,7,1-36.
[10] Ahmad Ayid Ahmad, Huseyin Polat (2023), Prediction of Heart Disease Based on Machine Learning Using Jellyfish Optimization Algorithm, MPDI, 13(14), 1-21.