Data Analysis of Factors Affecting Heart Disease

Savankumar Goswami
Nikhil Valse
Purv Dishor
Wang Qifang
Department of Business, Long Island University
Gurpreet Singh
11th November 2022

Abstract

According to Amiri at al. Ischemic heart disease (IHD) and stroke are the leading causes of death and disability worldwide, and CVDs more generally constitute a substantial contribution to both. Estimates from the World Health Organization's Global Burden of Disease Study 2015 are used to examine the extent of the total CVD burden, which includes 13 primary causes of cardiovascular death and 9 secondary risk factors**.** According to Baggio, B. (2000) The 2019 Global Burden of Disease Study (GBD). Using all available population-level data sources on incidence, prevalence, case fatality, mortality, and health risks, GBD, an ongoing multinational collaboration to provide comparable and consistent estimates of population health over time, produced estimates for 204 countries and territories from 1990 to 2019. Total prevalence of CVD nearly quadrupled between 1990 and 2019; from 271 million (95% UI: 257 to 285 million) to 523 million (95% UI: 497 to 550 million); and from 12.1 million (95% UI:11.4 to 12.6 million) to 18.6 million (95% UI:17.1 to 19.7 million) fatalities due to CVD by Barbiellini Amidei, C.(2022) Years spent disabled increased from 17.7 million (95% UI: 12.9 to 22.5 million) to 34.4 million (95% UI:24.9 to 43.6 million) worldwide. Disability-adjusted life years (DALYs) and years of life lost also increased significantly worldwide. Since 1990, the amount of DALYs attributable to IHD has gradually increased, with 2019 expected to see 182 million (95% UI: 170 to 194 million) DALYs, 9.14 million (8.40 to 9.74 million) deaths, and 197 million (99% UI: 178 to 220 million) prevalent cases. Since 1990, the number of DALYs attributed to stroke has gradually increased, with 2019 expected to see 143 million (95% UI: 133 to 153 million) DALYs, 6.55 million (95% UI: 6.00 to 7.02 million) deaths, and 101 million (95% UI: 93.2 to 111 million) prevalent instances of stroke. Diseases of the heart continue to account for the lion's share of mortality worldwide. For practically all countries outside of high-income countries, the burden of cardiovascular disease (CVD) continues to

climb, and unfortunately, the age-standardized rate of CVD has begun to rise in several regions where it had previously been dropping in high-income countries.

Background:

According to Kang, D. (2022) Cardiovascular diseases (CVDs) are one of the leading causes of death worldwide, accounting for 31% of deaths from NCDs. The prevalence of cardiovascular diseases (CVDs) is rising dramatically in both developed and developing countries as a direct result of rapid changes in human lives. There are an estimated 422.7 million people who suffer from CVDs, and 17.9 million fatalities per year are due to cardiovascular diseases, as reported in the most recent assessment of GBD (Global Burden of Disease).

In Eastern Mediterranean communities, including Iran, the burden of CVDs and their linked risk factors, as well as the rising trend, is particularly concerning. Recent reports indicate that ischemic heart disease is the leading cause of mortality in Iran, with a prevalence of 5.9% for coronary heart disease (CHD) and a prevalence of 3% for stroke. Given the alarmingly high rates of cardiovascular disease (CVD) occurrence and prevalence, there has been a growing movement to identify the most important factors contributing to this epidemic. Obesity, lack of physical activity, glucose intolerance, hypertension, mental stress, and smoking are among the most significant risk factors of cardiovascular diseases.

The development of other cardiovascular risk factors, such as glucose intolerance, dyslipidemia, and thrombus formation, is facilitated by smoking, the second largest modifiable risk factor of CVDs. Smoking directly hurts and affects the cardiac vasculature. Iran has been the most effective country in the Eastern Mediterranean region in implementing World Health Organization (WHO) policies for tobacco control; yet, the latest estimates indicate that smoking prevalence is still high, with around a fifth of Iranian men continuing to smoke.

For this project, we have selected the dataset ”Personal Key Indicators of Heart Disease” made available at the link[1] on Kaggle. The dataset contains Key Indicators of Heart Disease and was created from 2020 annual CDC survey data of 400k adults related to their health status. CDC is the Centers for Disease Control and Prevention. According to them, Heart Diseases are among leading causes for death so detecting and preventing the factors that have the greatest impact on heart disease is very important in healthcare.

Dataset Description:

The dataset was originally collected through telephonic surveys by CDC and a part of the Behavioural Risk Factor Surveillance System (BRFSS). Original dataset consisted of 401,958 rows and 279 columns with a major portion being questions asked to respondents regarding their health. In this particular dataset, only features that directly or indirectly influenced the heart and contributed to heart diseases were retained.

So, only 18 variables exist in the data that we selected. 9 of those variables are of boolean nature. 4 of them are decimals and 5 of them are strings.

The details of those variables are given below:

Heart Disease: It contains Yes or No depending upon the answer of respondents that they ever have/had any coronary heart disease (CHD) or myocardial infarction (MI)

BMI: Body Mass Index of respondent.

Smoking: Has the respondent smoked 5 or more packs of cigarettes in their lifetime.

AlcoholDrinking: Is the respondent a heavy drinker?

Stroke: If the respondent has ever suffered a stroke.

PhysicalHealth: If the respondent has any physical injury or health issue in the past 30 days and if so, for how many days?

MentalHealth: If the person has had issues related to mental health in the past 30 days and if so, how many days.

DiffWalking: Does the person have difficulty while walking?

Sex: Gender of the person

AgeCategory: Fourteen level age categories have been defined for this variable.

Race: This variable details the race of the respondent.

Diabetic: This variable tells whether the a person is diabetic or not

PhysicalActivity: This variable details whether a person has performed any physical activity or exercise apart from their job in the last 30 days.

Gen Health: This variable tells the opinion of the person about their own general health.

Sleep Time: This variable details the average sleeping hours of the person.

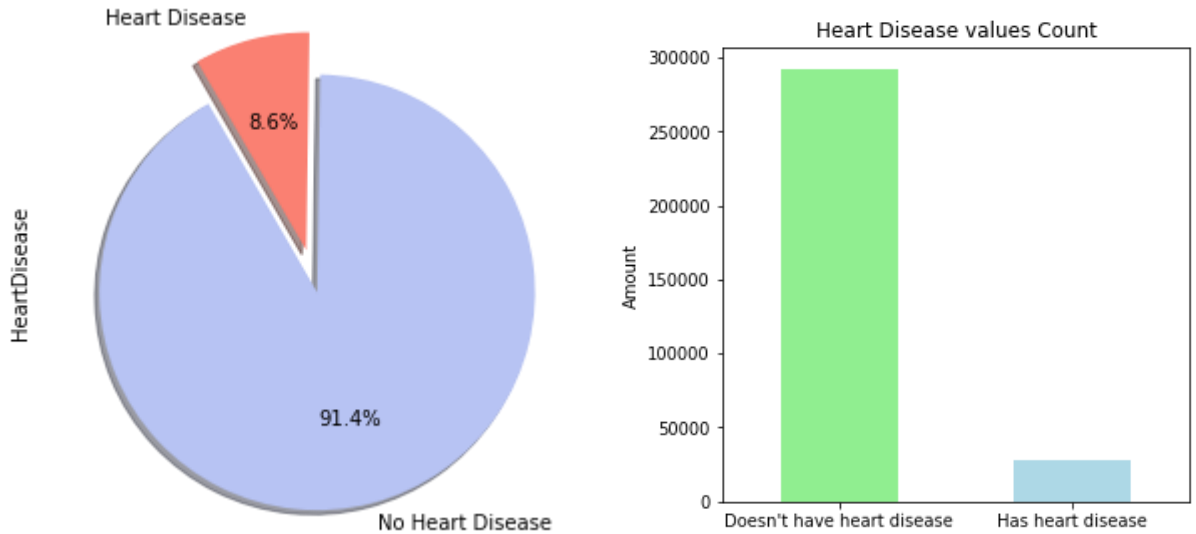
Asthma: This variable contains information whether a person has asthma or not.

Kidney Disease: It contains information if the person has kidney disease or not.

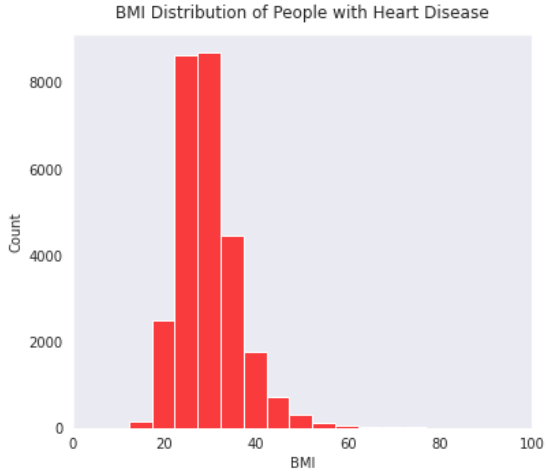
Skin Cancer: This explains whether a person has or ever had skin cancer.

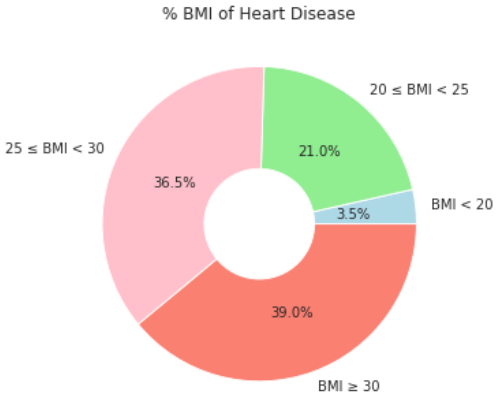
Dataset analysis, visualisation, feature correlation, insights extracted from data visualisation

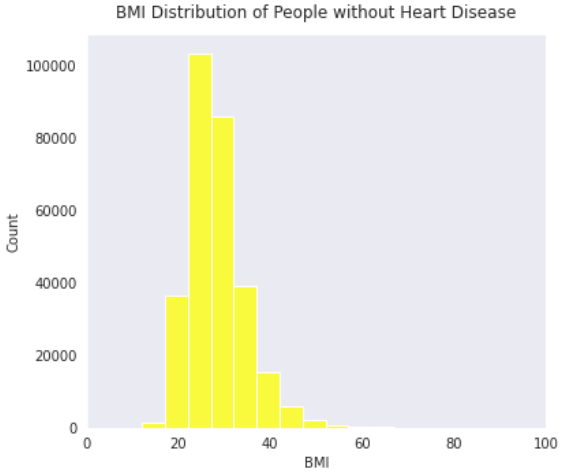
1. Count Comparison of People Having Heart Disease

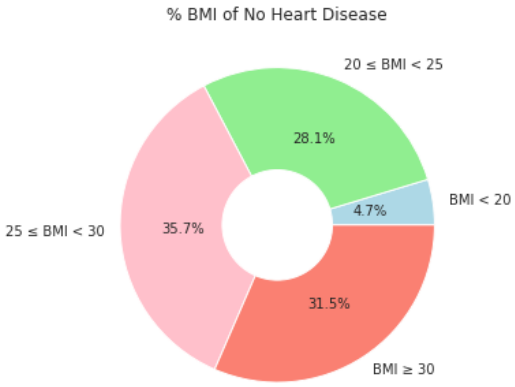


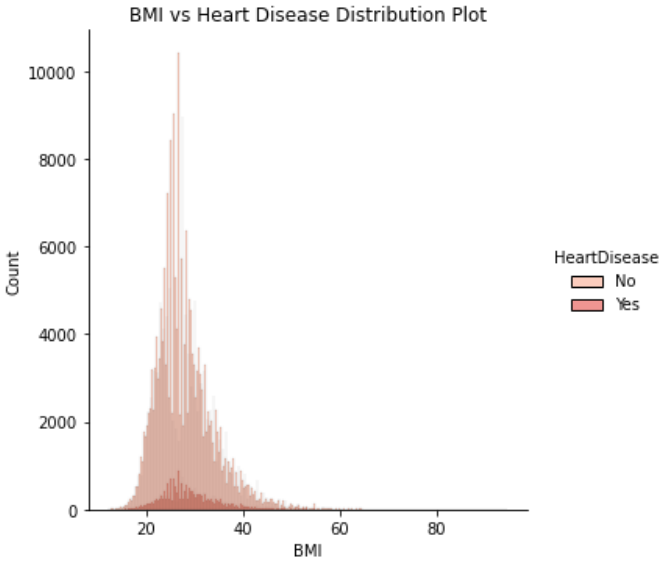
2. BMI and Heart Disease Comparison



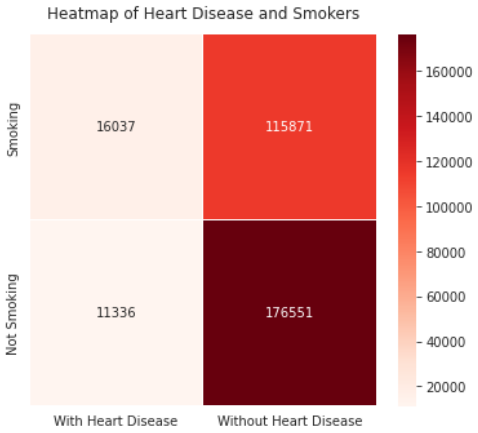


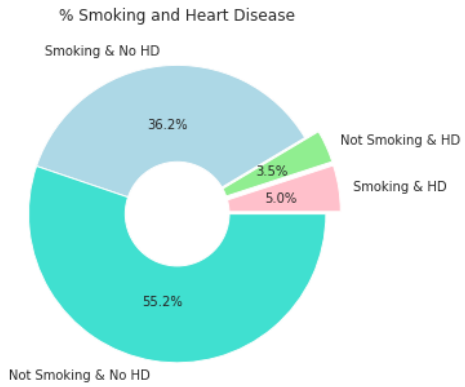


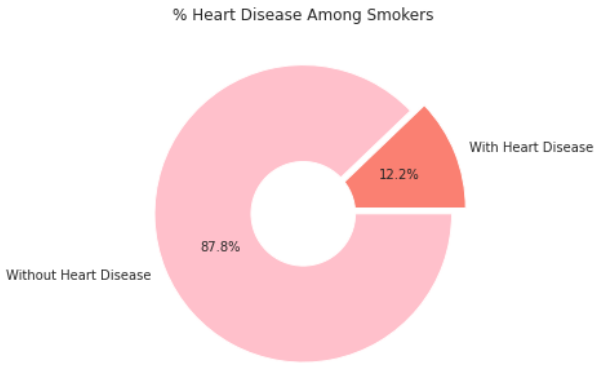


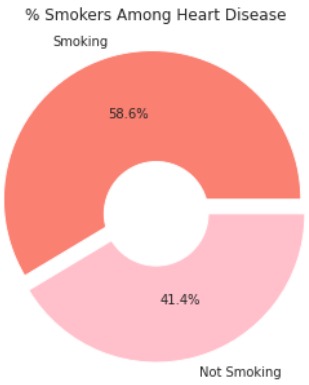


3. Impact of Smoking on Heart Disease:

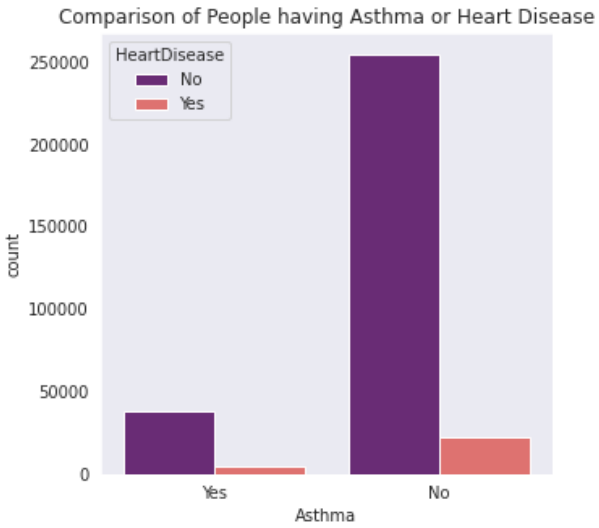




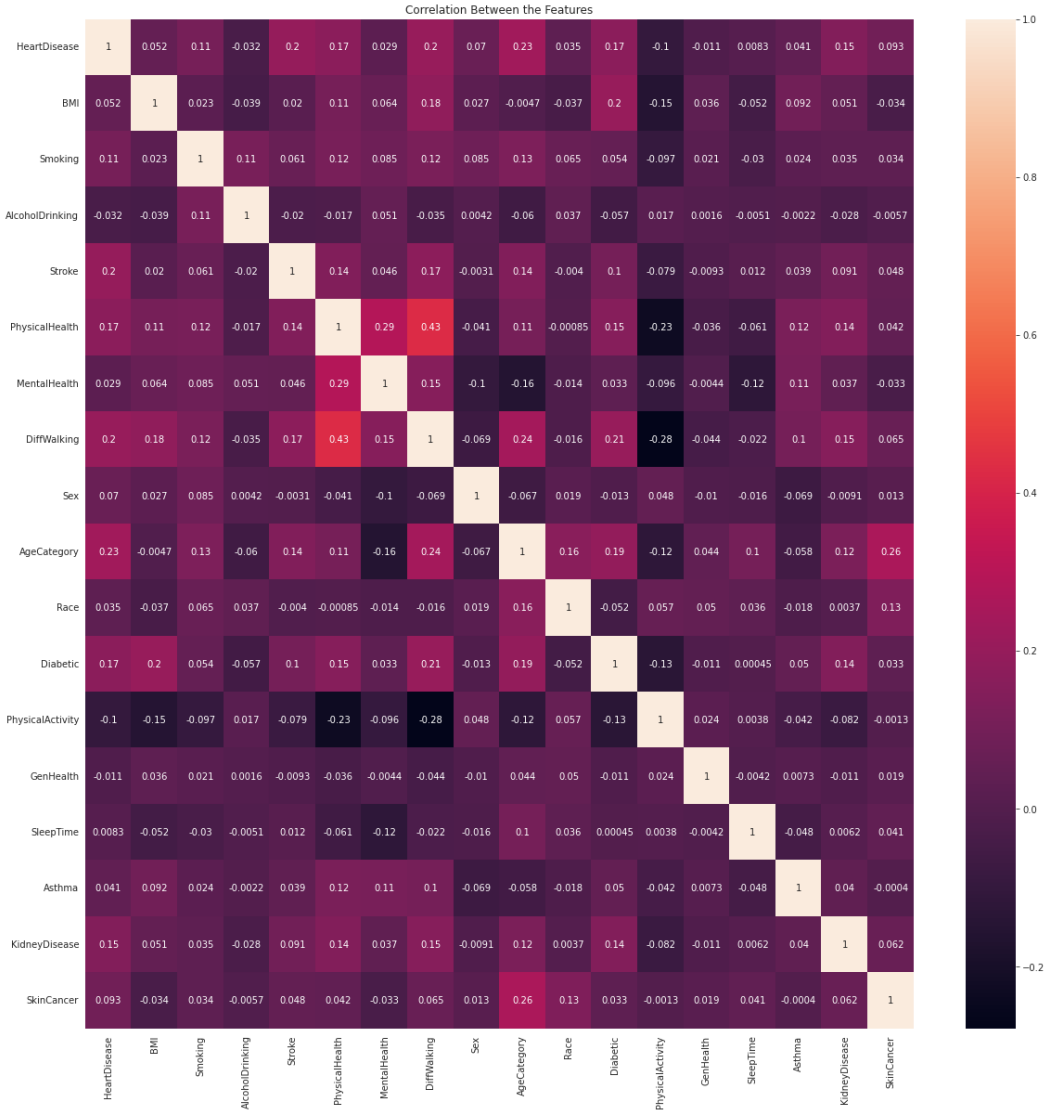




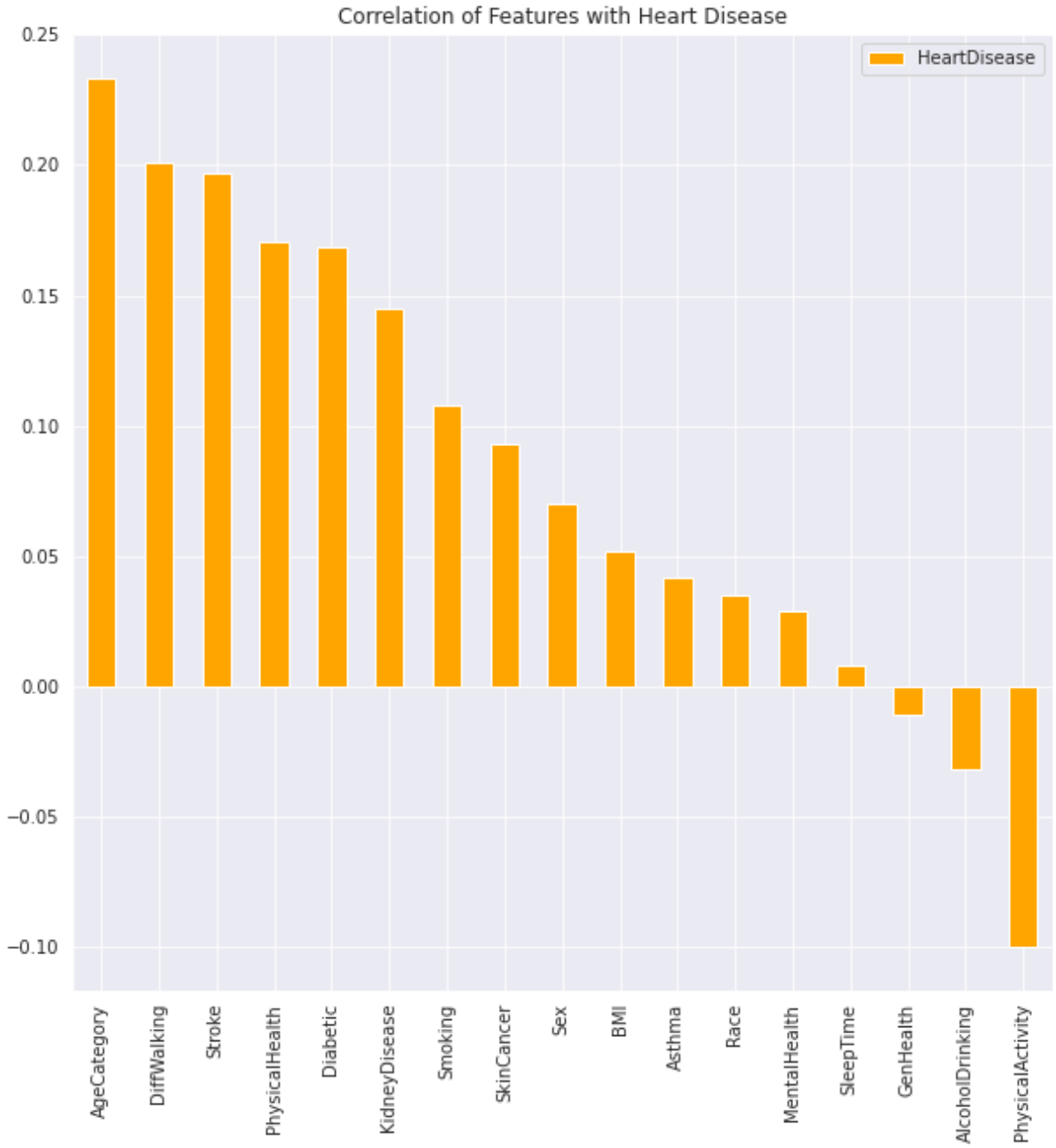
4. Asthma and Heart Disease:



5. Correlation Between Features:



6. Correlation Of Features with Heart Disease:



Data Analysis:

We analysed the data we selected. Some of the features were numerical in nature like BMI and PhysicalHealth and MentalHealth while most of the features were categorical. Categorical features were also of three natures, binary (Sex, Stroke etc), nominal(Race, Diabetic) and Ordinal(AgeCategory, GenHealth). Data was found to be highly imbalanced with the majority

class being people without Heart Disease. Some features visualised have been shown in above pictures. We also checked our data for any null or nan values and found out there was no missing data.

Insights From the Data Analysis:

We analysed various features and their impact and contribution to our target variable of heart disease. We found out that features like smoking behaviour, diabetes, stroke and physical health contribute a lot to heart disease.

We performed data analysis and visualised some of the features and extracted correlation between the features as well as the correlation of each input feature with our target variable and based on the insights we found out the features we can drop as well as the preprocessing steps we need to perform on the data.

Methodology:

Data preparation including transforms, scaling, re-shaping and any feature selection to reduce dimensionality. Summary of cleaned/pre-processed data

Since our data had mixed variables, we selected the categorical variables to encode them. We used sklearn's LabelEncoder() to encode the categorical variables to convert them into numerical variables. The correlation was computed after performing the label encoding.

From correlation results, we saw that features 'AlcoholDrinking', 'GenHealth, ''PhysicalHealth' and 'SleepTime' are not much related to Heart Disease and thus we dropped these features.

We saw that our classes were highly imbalanced. Highly imbalanced data is not suitable for model training as the model tends to favour the majority class so we decided to balance our data. We used two approaches so we could compare each approach while implementing the model and see which worked better. The approaches included:

Oversampling: Upsampled the minority class using Smote technique with created new data points.

Undersampling: Removing points from the majority class to balance both classes.

Since some of the numerical variables have larger values compared to others and such features if not scaled tend to affect the model performance so we used StandardScaler() to scale both the data.

Summary of the cleaned data: We created two copies of the data. One contained reduced sample points with equal classes. The other contained increased sample points with minority class upsampled. All the categorical features were label encoded and the data was standardised using sklean's StandardScaler().

Modelling:

For model training, we used the following algorithms:

Random Forest: A decision tree combines some decisions, whereas a random forest combines several decision trees. Random forest is suitable when we have large datasets. In our dataset, the number of data points are 300K+ and they further increase for the oversampling case. So Random Forest will be a good choice.

KNN: KNN works well with large amounts of data. So in comparison with RF we also chose to work with KNN.

Results using RF:

Three models using RF algorithms were trained. Models were trained on datasets in following ways:

- Without sampling
- With undersampling of majority class
- With over sampling of minority class

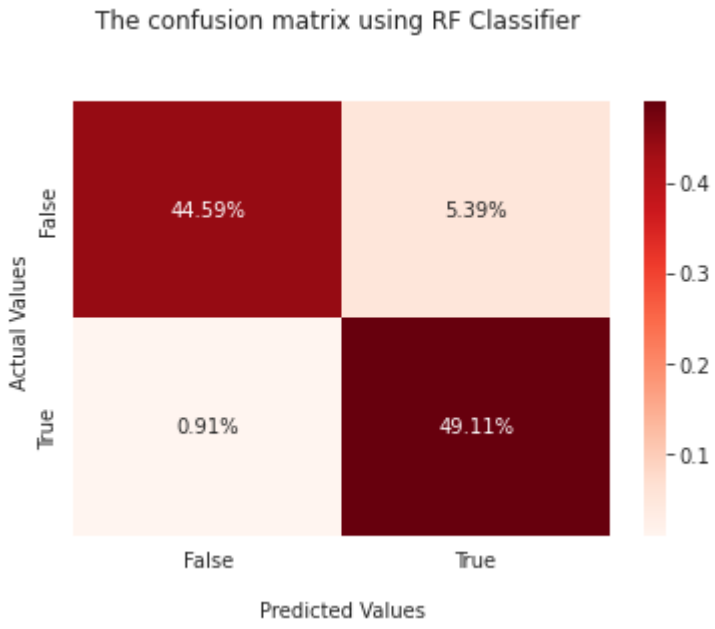
The results for the models using RF are given below:

```
Accuracy of Train before sampling: 98.0596944917838
Accuracy of  Test before sampling: 89.81222345565128

_____________________________________________________
Accuracy of Train after RandomUndersampling: 96.94035985021463
Accuracy of Test after RandomUndersampling: 71.04109589041096

_____________________________________________________
Accuracy of Train after Random Oversampling: 96.91648410366017
Accuracy of Test after Random Oversampling: 93.70004018158657

_____________________________________________________
```

The model with the data oversampled worked best in our case. The detailed results for the best performing model is given as:

[[52154 6307]

[1062 57446]]




```
The Classification Report for RF Classifier:
              precision    recall  f1-score   support

           0       0.98      0.89      0.93     58461
           1       0.90      0.98      0.94     58508

    accuracy                           0.94    116969
   macro avg       0.94      0.94      0.94    116969
weighted avg       0.94      0.94      0.94    116969
```

KNN Classifier:

KNN classifier was able to predict the heart disease with precision of 82% for the best model. Same three models were trained for KNN as well. In this case also, the top performing model was the model with the oversampling of minority class.

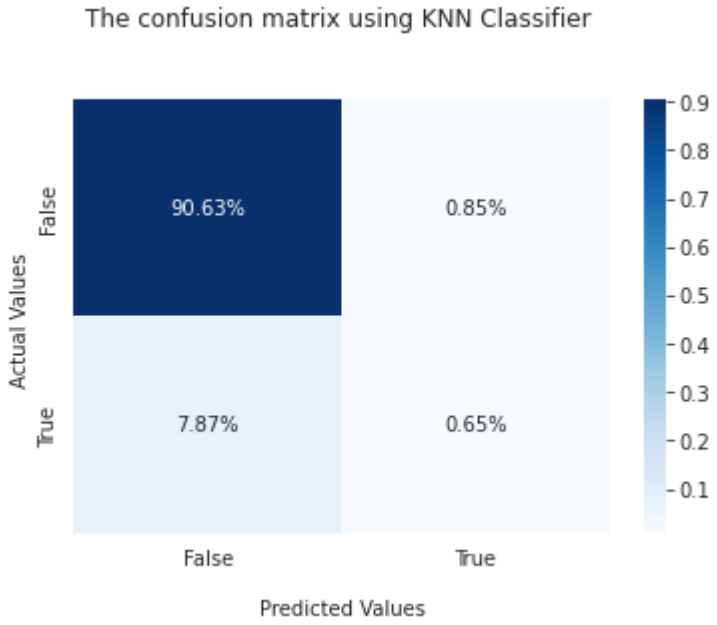
The results of three models showing accuracy score are:

```
Accuracy of Train before sampling: 92.11135258524993
Accuracy of  Test before sampling: 91.28191497678199

_____________________________________________________________
Accuracy of Train after Random Undersampling: 78.4934697232624
Accuracy of Test after Random Undersampling: 70.7579908675799

_____________________________________________________________
Accuracy of Train after Random Oversampling: 90.47117285599786
Accuracy of Test after Random Oversampling: 87.69674016192324

_____________________________________________________________
```

The detailed result for the best model is:

[[57968 544]

 [5032 415]]



```
The Classification Report for KNN Classifier:
              precision    recall  f1-score   support

           0       0.95      0.80      0.87     58461
           1       0.82      0.96      0.89     58508

    accuracy                           0.88    116969
   macro avg       0.89      0.88      0.88    116969
weighted avg       0.89      0.88      0.88    116969
```

Data Visualization and Manipulation:

We are categorizing the BMI column to get better insights into the data, it will also help in better prediction in the ML model after mutating it this way. By discretizing the continuous variable, we are also reducing the noisiness of the variable. With binning, the speed of the algorithm also gets improved.

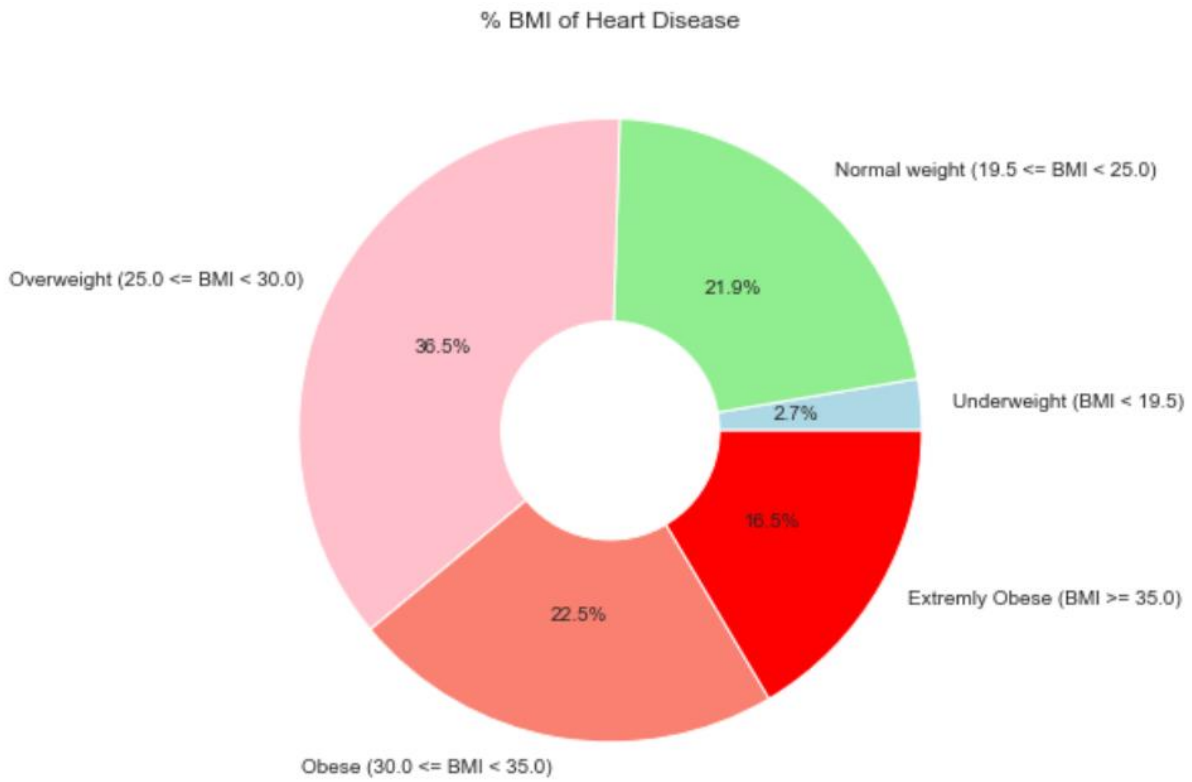
We are visualizing the BMI variable with new categories created so that we get the statistics about the types of people depending on their BMIs among the two categories that are With heart disease and without heart disease.

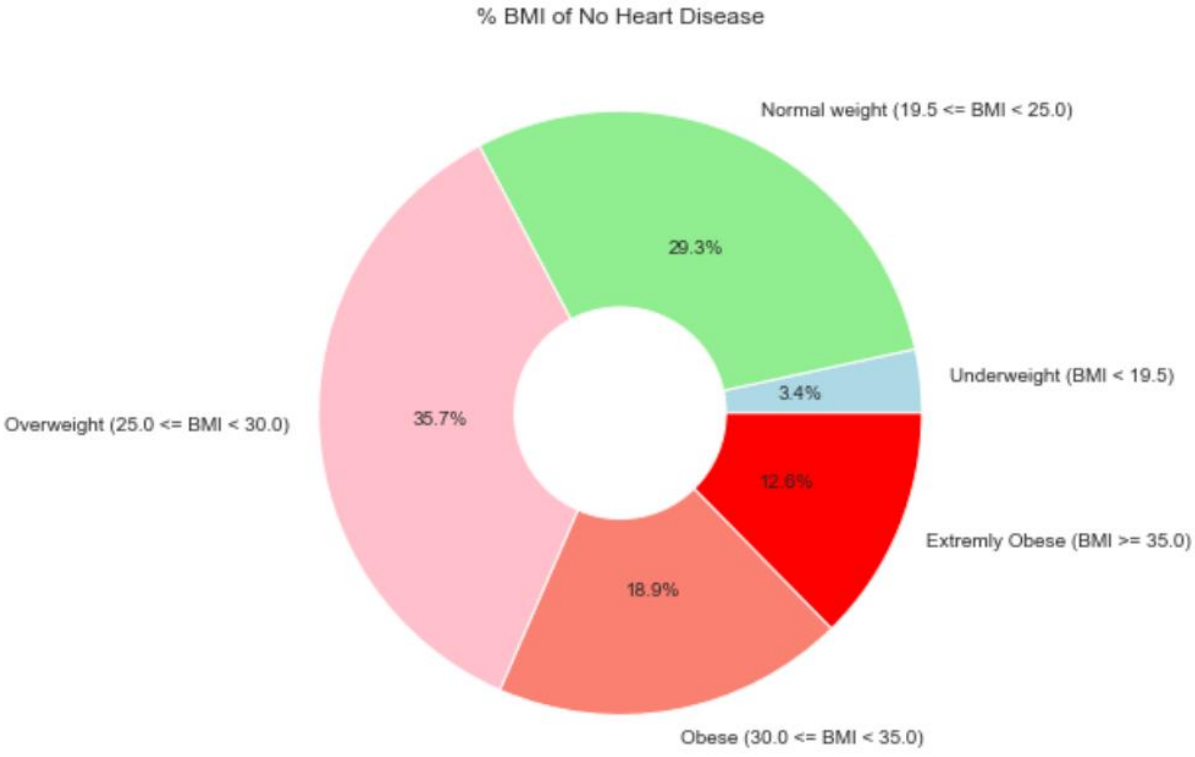
BMI categories are as follows,

Underweight --- (BMI < 19.5), Normal weight --- (19.5 <= BMI < 25.0),

Overweight --- (25.0 <= BMI < 30.0), Obese --- (30.0 <= BMI < 35.0),

Extremely Obese --- (BMI >= 35.0)


% BMI of Heart Disease
Normal weight (19.5 <= BMI < 25.0)
21.9%
Underweight (BMI < 19.5)
2.7%
Extremly Obese (BMI >= 35.0)
16.5%
Obese (30.0 <= BMI < 35.0)
22.5%
Overweight (25.0 <= BMI < 30.0)
36.5%


% BMI of No Heart Disease
Normal weight (19.5 <= BMI < 25.0)
29.3%
Underweight (BMI < 19.5)
3.4%
Extremly Obese (BMI >= 35.0)
12.6%
Obese (30.0 <= BMI < 35.0)
18.9%
Overweight (25.0 <= BMI < 30.0)
35.7%

Now we do mutate the actual data-frame for the model,

```
bins = [0, 19.5, 25, 30, 35, np.inf]
names = ['Underweight (BMI < 19.5)', 'Normal weight (19.5 <= BMI < 25.0)',
'Overweight (25.0 <= BMI < 30.0)', 'Obese (30.0 <= BMI < 35.0)', 'Extremely
Obese (BMI >= 35.0)']
df['BMICategory'] = pd.cut(df['BMI'], bins, labels=names)
df.drop('BMI',axis=1,inplace=True)
```

Now, our new BMICategory column has distribution like this.

```
df['BMICategory'].value_counts()
```

```
Overweight (25.0 <= BMI < 30.0)     114355
Normal weight (19.5 <= BMI < 25.0)   92243
Obese (30.0 <= BMI < 35.0)           61169
Extremly Obese (BMI >= 35.0)         41379
Underweight (BMI < 19.5)             10649
Name: BMICategory, dtype: int64
```

Feature Encoding:

```
from sklearn.preprocessing import LabelEncoder
order_cols = ["BMICategory", "AgeCategory", "HeartDisease"]
no_order_cols = [ "Smoking", "AlcoholDrinking", "Stroke", "DiffWalking",
            "Sex", "Race", "Diabetic", "PhysicalActivity",
            "GenHealth", "Asthma", "KidneyDisease", "SkinCancer"]

# Label encoding
for col in order_cols:
    df[col] = LabelEncoder().fit_transform(df[col])

# One-hot encoding
for col in no_order_cols:
    dummy_col = pd.get_dummies(df[col], prefix=col)
    df = pd.concat([df, dummy_col], axis=1)
    del df[col]
```

We have now converted categorical columns into numeric form so as to make them machine-readable. Logistic regression can then decide in a better way how those labels must be operated. It is an important pre-processing step for the dataset in supervised learning.

We use normal label encoding for the categories which have scaling orders, and one-hot encoding for the categorical variables having less or no scaling orders.

Race_America...	Race_Asian	Race_Black	Race_Hispanic	Race_Other	Race_White
0	0	0	0	0	1
0	0	0	0	0	1
0	0	0	0	0	1
0	0	0	0	0	1
0	0	0	0	0	1
0	0	1	0	0	0
0	0	0	0	0	1
0	0	0	0	0	1
0	0	0	0	0	1

For example, in the Race column above, the normal encoding would assign the numbers from 0 to 5. So, the ML algorithm reads 0 as less significant and 5 as more significant or better parameter. Instead of that we create six columns by six different races and add 1 to every column for the corresponding race, otherwise zero. We should try to mitigate this problem by using One-hot encoding. It also provides more nuanced predictions than single labels.

```
df.head()
```

✓ 0.5s

	HeartDisease	PhysicalHealth	MentalHealth	AgeCategory	SleepTime	BMICategory	Smoking_No	Smoking_Yes
0	0	3.0	30.0	7	5.0	4	0	1
1	0	0.0	0.0	12	7.0	1	1	0
2	0	20.0	30.0	9	8.0	3	0	1
3	0	0.0	0.0	11	6.0	1	1	0
4	0	28.0	0.0	4	8.0	1	1	0

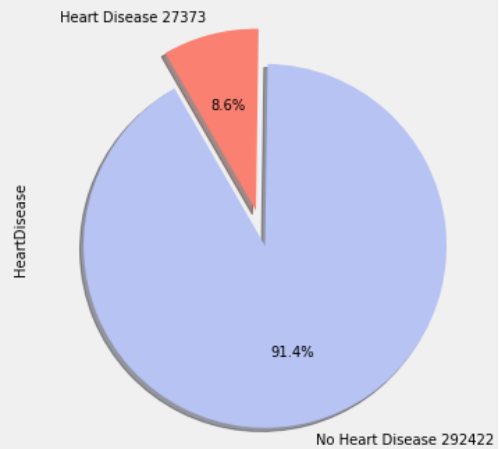
5 rows × 39 columns

Mitigating the Imbalanced predictor variable problem:

We first need to separate independent variables from the dependent variable which we are going to predict.

```
X = df.drop('HeartDisease',axis=1)
Y = df['HeartDisease']
```

As we know from the previous visualizations, we have roughly 90 percent of the rows in the data represent the people with no heart disease and only 10 percent represent the heart disease and their parameters.



This can affect our prediction model in negative way because algorithm can not find the pattern in the insufficient data points of the people with heart disease.

We are using SMOTE (Synthetic Minority Oversampling Technique) which creates the synthetic observations of the minority class, which are similar but tweaked randomly with the nearest values.

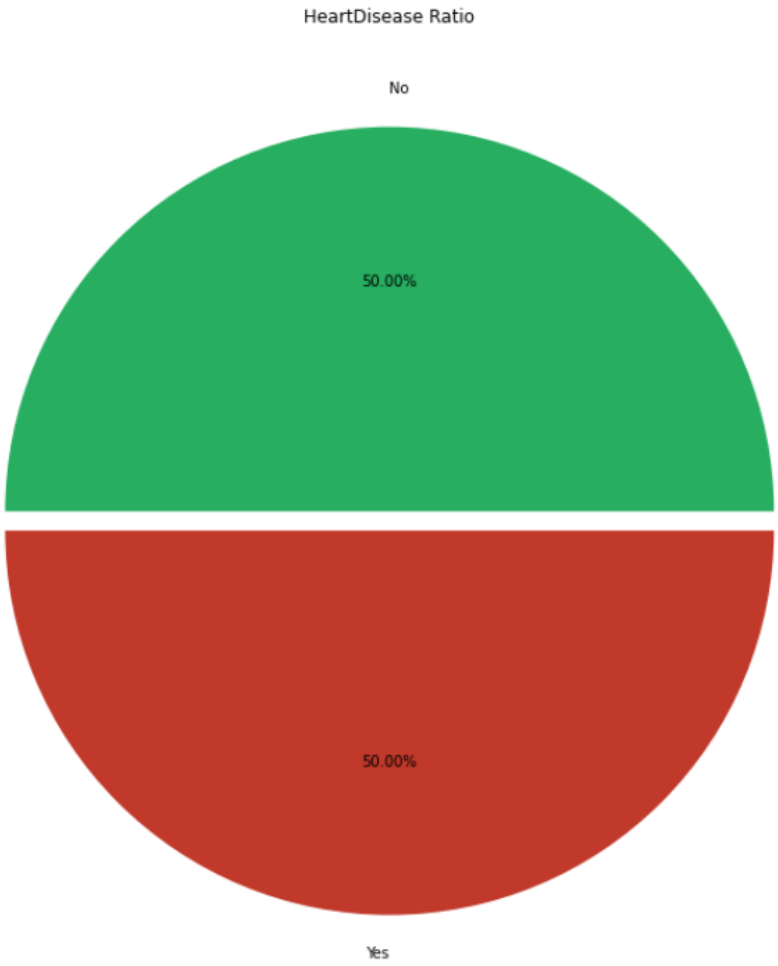
Then we split the data in the ratio of 75:25 between training and testing data as follows,

```
from sklearn.model_selection import train_test_split
from imblearn.over_sampling import SMOTE
over = SMOTE(k_neighbors=2000)
X, Y = over.fit_resample(X, Y)
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.25,
random_state=2022, stratify=Y)
```

With random_state=None, we get different results across different executions and the shuffling process is out of control. So, we use specific random state to avoid this, in our case we are using 2022. So every time the train and split row selection will be similar.

We can now see in the Figure below that the data is balanced between records of people with heart disease and without heart disease.

```
fig, ax = plt.subplots(1, 1,figsize =(15, 12))
ax.pie(Y.value_counts(),autopct='%1.2f%%',labels=['No',"Yes"],
explode=(0,0.05),colors=['#27ae60','#c0392b'])
ax.set_title('HeartDisease Ratio')
fig.show()
```



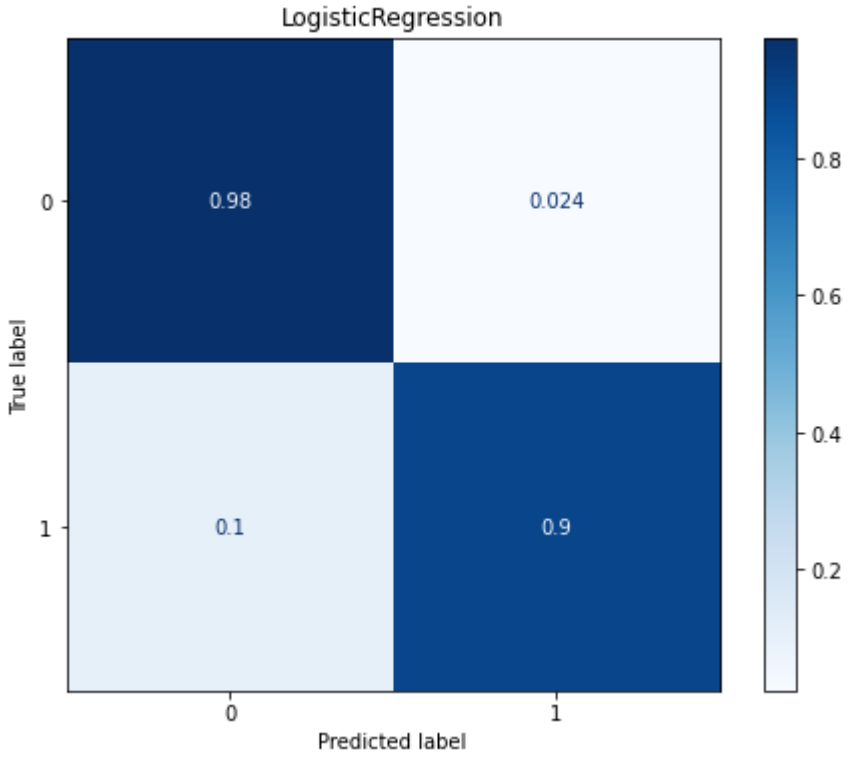
Implementing the actual model:

For implementing the actual model on our balanced data, we need to import several libraries in order to split the data between train and test sets and finding out the predictions and its metrics.

```
from sklearn import metrics
from sklearn.linear_model import LogisticRegression
```

Now we are going to train the algorithm on 75 percent train data, and predict the output for the remaining 25 percent test data.

```
fit = LogisticRegression(random_state=2022).fit(X_train, y_train)
fig, ax = plt.subplots(figsize=(8, 6))
ax.set_title("LogisticRegression")
metrics.plot_confusion_matrix(fit, X_test, y_test, cmap='Blues',
normalize='true', ax=ax)
plt.show()
```



Upper Left: True Positives, Upper Right: False Negatives

Lower Left: False Positives, Lower Right: True Negatives

Interpretation:

As we can see in the confusion matrix above, we got 98 percent predicted results as true positive which is great achievement and 90 percent true negative.

Speaking of the errors, we have Type I error which are false positives, we have nearly 10 percent predictions positive for heart disease when the person is not having heart disease.

For the Type II error we have only 2.4 percent predictions which got result as negative when it should be positive. This type of error can prove to be dangerous because person can become unaware that he has the heart disease even though he might be in trouble internally. Fortunately, we have least amount of Type II error.

Other Metrics:

```
recall = round(metrics.recall_score(y_test, fit.predict(X_test))*100, 2)
print('recall', recall)
precision = round(metrics.precision_score(y_test, fit.predict(X_test))*
100, 2)
print('precision', precision)
accuracy = round(metrics.accuracy_score(y_test, fit.predict(X_test))*
100, 2)
print('accuracy', accuracy)
```

```
recall 89.53
precision 97.43
accuracy 93.59
```

Accuracy: Accuracy is a metric for evaluating classification models. Basically, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Precision is defined as the ratio of the total number of correctly classified positive classes divided by the total number of predicted positive classes. Or, out of all the predictive positive classes, how much we predicted correctly. Precision should be high.

$$Precision = \frac{TP}{TP + FP} \text{ or } \frac{\text{True Positive}}{\text{Predictive Results}}$$

Recall (Sensitivity): Recall is defined as the ratio of the total number of correctly classified positive classes divide by the total number of positive classes. Or, out of all the positive classes, how much we have predicted correctly. Recall should be high.

$$\textit{Recall} = \frac{TP}{TP + FN} \text{ or } \frac{\text{True Positive}}{\text{Actual Results}}$$

References

1. Amiri, P., Mohammadzadeh-Naziri, K., Abbasi, B., Cheraghi, L., Jalali-Farahani, S., Momenan, A. A., . . . Azizi, F. (2019). Smoking habits and incidence of cardiovascular diseases in men and women: Findings of a 12 year follow up among an urban eastern-Mediterranean population. *BMC Public Health, 19*(1). doi:10.1186/s12889-019-7390-0
2. Baggio, B. (2000). Ischemic renal disease: Impact of cardiovascular risk factors and smoking. *Contributions to Nephrology,* 68-74. doi:10.1159/000060043
3. Barbiellini Amidei, C., Trevisan, C., Dotto, M., Ferroni, E., Noale, M., Maggi, S., . . . Sergi, G. (2022). Association of physical activity trajectories with major cardiovascular diseases in elderly people. *Heart, 108*(5), 360-366. doi:10.1136/heartjnl-2021-320013
4. Borgenicht, N. (2022). Uber explores partnerships with taxi industry. doi:10.4135/9781529609219
5. Kang, D., Sung, J., Kim, D., Jin, M., Jang, E., Yu, H. T., . . . Joung, B. (2022). Association between exercise habit changes and mortality following a cardiovascular event. *Heart*. doi:10.1136/heartjnl-2022-320882
6. Mohan, B., Sharma, S., Sharma, S., Kaushal, D., Singh, B., Takkar, S., . . . Wander, G. S. (2017). Assessment of knowledge about healthy heart habits in urban and rural population of Punjab after SMS campaign—a cross-sectional study. *Indian Heart Journal, 69*(4), 480-484. doi:10.1016/j.ihj.2017.05.007
7. Murray, R. P., Connett, J. E., Tyas, S. L., Bond, R., Ekuma, O., Silversides, C. K., & Barnes, G. E. (2002). Alcohol Volume, drinking pattern, and cardiovascular disease morbidity and mortality: Is there a U-shaped function? *American Journal of Epidemiology, 155*(3), 242-248. doi:10.1093/aje/155.3.242

8. Navarro-Prado, S., Schmidt-RioValle, J., Montero-Alonso, M. A., Fernández-Aparicio, Á, & González-Jiménez, E. (2018). Unhealthy lifestyle and nutritional habits are risk factors for cardiovascular diseases regardless of professed religion in university students. *International Journal of Environmental Research and Public Health*, *15*(12), 2872. doi:10.3390/ijerph15122872
9. Preventing heart disease. (2022, August 29). Retrieved November 17, 2022, from https://www.hsph.harvard.edu/nutritionsource/disease-prevention/cardiovascular-disease/preventing-cvd/
10. Review for "global burden and attributable risk factors of acute lymphoblastic leukemia in 204 countries and territories in 1990–2019: Estimation based on Global Burden of Disease Study 2019". (2021). doi:10.1002/hon.2936/v2/review2
11. Rippe, J. M., & Angelopoulos, T. J. (2019). Lifestyle Strategies for Risk Factor Reduction, prevention and treatment of cardiovascular disease. *Lifestyle Medicine,* 19-36. doi:10.1201/9781315201108-2
12. Robinson, S. (2021). Cardiovascular disease. *Priorities for Health Promotion and Public Health,* 355-393. doi:10.4324/9780367823689-16
13. Tourlouki, E. (2009). Dietary habits and cardiovascular disease risk in middle-aged and elderly populations: A review of evidence. *Clinical Interventions in Aging,* 319. doi:10.2147/cia.s5697.