**Data Analytics and Prediction of Heart Disease**

**Fatemeh Allameh**

**1179093**

# **Introduction**

Heart disease is one of the leading causes of death worldwide that affects not only the heart but other major parts of the body. According to the CDC (Centers for Disease Control and Prevention), about 697,000 people in the United States died from heart disease in 2020, that 1 in every 5 deaths [0]. Therefore, detecting and preventing the factors that have the greatest impact on heart disease is so crucial in healthcare. In this study, I used the Personal Key Indicators of Heart Disease dataset [1 ] that contains 320K rows and 18 columns which is a smaller version and derived by the 2020 annual CDC survey data of 400k adults. For each patient (row), it contains the health status of that individual. This dataset is a binary classification, however it is notable that the classes are not balanced, so to address this, I applied different kinds of approaches to cope with the Imbalanced Class issue.

## **About the Project**

The main target of conducting this project is to classify a patient’s condition involving personal key indicators of heart disease to predict whether he/she is prone to this disease. In this project, I am going to div into the causes of heart disease and its relations with other health indicators, drawing insights and exploring the data aiming get a better picture of the leading causes (identifying the most critical factors associated with it). Then, after preprocessing and preparing the data to be normalized and trying to resolve the problem of Imbalanced Class I will use different models on our data to automate heart disease detection and then compare their results. In the end, experimental setup and the evaluation of various applied methods and models and comparing related metrics will be done to build a good prediction with higher Accuracy.

## **About the Dataset**

The Personal Key Indicators of Heart Disease dataset contains 320K rows and 18 columns. It is a cleaned, smaller version of the 2020 annual CDC (Centers for Disease Control and Prevention) survey data of 400k adults. For each patient (row), it contains the health status of that individual. The data was collected in the form of surveys conducted over the phone. Each year, the CDC calls around 400K U.S residents and asks them about their health status, with most questions being yes or no questions. Below is a description of the features collected for each patient:

|  |  |  |
| --- | --- | --- |
| No | Feature Name | Description |
| 1 | HeartDisease | Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI) |
| 2 | BMI | Body Mass Index (BMI) |
| 3 | Smoking | Have you smoked at least 100 cigarettes in your entire life?  [Note: 5 packs = 100 cigarettes] |
| 4 | AlcoholDrinking | Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week |
| 5 | Stroke | (Ever told) (you had) a stroke? |
| 6 | PhysicalHealth | For how many days during the past 30 days having any physical illness and injury? |
| 7 | MentalHealth | For how many days during the past 30 days, was your mental health not good? |
| 8 | DiffWalking | Do you have serious difficulty walking or climbing stairs? |
| 9 | Sex | Are you male or female? |
| 10 | AgeCategory | Fourteen-level age category |
| 11 | Race | Imputed race/ethnicity value |
| 12 | Diabetic | (Ever told) (you had) diabetes? |
| 13 | PhysicalActivity | Adults who reported doing physical activity or exercise during the past 30 days other than their regular job |
| 14 | GenHealth | Would you say that in general your health is... |
| 15 | SleepTime | On average, how many hours of sleep do you get in a 24-hour period? |
| 16 | Asthma | (Ever told) (you had) asthma? |
| 17 | KidneyDisease | Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease? |
| 18 | SkinCancer | (Ever told) (you had) skin cancer? |

# **Data Overview with EDA (Exploratory Data Analysis)**

This part of our project refers to exploring the data, the relation of heart disease with other features on our data, and the patterns and insights hidden inside the data. We will visualize the data in various ways in our exploration.

## **2.1 Data Reading**

In the first step, the dataset must be loaded and read to a DataFrame to be analyzed which below table has shown the head of it.

2.2 **Numerical Description**

There are only 4 features that are numerical including BMI, PhysicalHealth, MentalHealth and SleepTime. In the below table, some basic statistical details like Min, Max, Mean, Std and Percentile of all these 4 numerical features are shown.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | BMI | PhysicalHealth | MentalHealth | SleepTime |
| Count | 319795.000000 | 319795.00000 | 319795.000000 | 319795.000000 |
| Mean | 28.325399 | 3.37171 | 3.898366 | 7.097075 |
| Std | 6.356100 | 7.95085 | 7.955235 | 1.436007 |
| Min | 12.020000 | 0.00000 | 0.000000 | 1.000000 |
| 25% | 24.030000 | 0.00000 | 0.000000 | 6.000000 |
| 50% | 27.340000 | 0.00000 | 0.000000 | 7.000000 |
| 75% | 31.420000 | 2.00000 | 3.000000 | 8.000000 |
| Max | 94.850000 | 30.00000 | 30.000000 | 24.000000 |

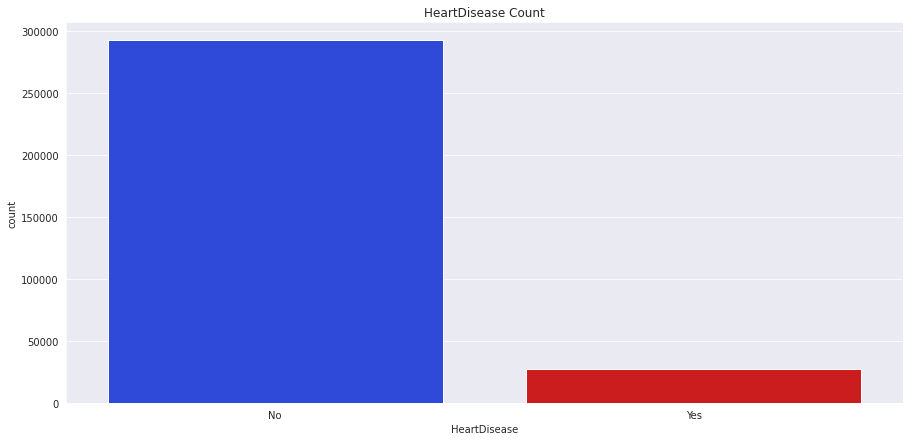
2.3 **Categorical Description**

Most features are categorical (the number of them is 14) and most of them consist of two categories (yes or no) while some of them have more than two categories like Agecategory, Race, Diabetic and Genhealth including 13, 6, 4 and 5 categories, respectively. For the prediction analysis, HeartDisease is considered as the target class label in which the number of instances for majority class (No) is 292422, and the number of records for minority class (yes) is 27373.

2.4 **Initial Data Assessment**

After an initial inspection and analysis, some following results was observed. The most important one was existence of a severe class imbalance (heart disease vs healthy) since only 8.56% of people have heart disease and 91.44% does not have. It means that approximately 9 in 100 people suffer from heart disease. Below figure shows the imbalanced class distribution of the dataset where the number of observations belonging to positive class (Yes) is significantly lower than the negative class (No).

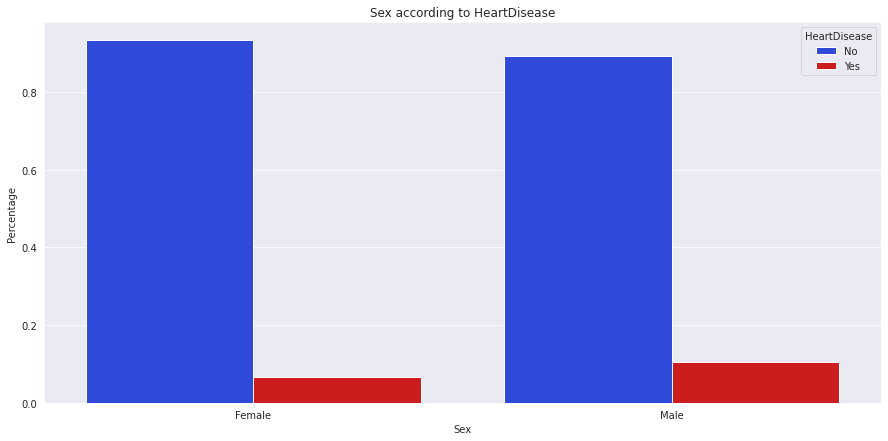
|  |  |  |  |
| --- | --- | --- | --- |
| Class | | Frequency | Percentage |
| Heart Disease | 27373 | | 8.56% |
| No Heart Disease | 292422 | | 91.44% |
| Total | 319795 | | 100% |



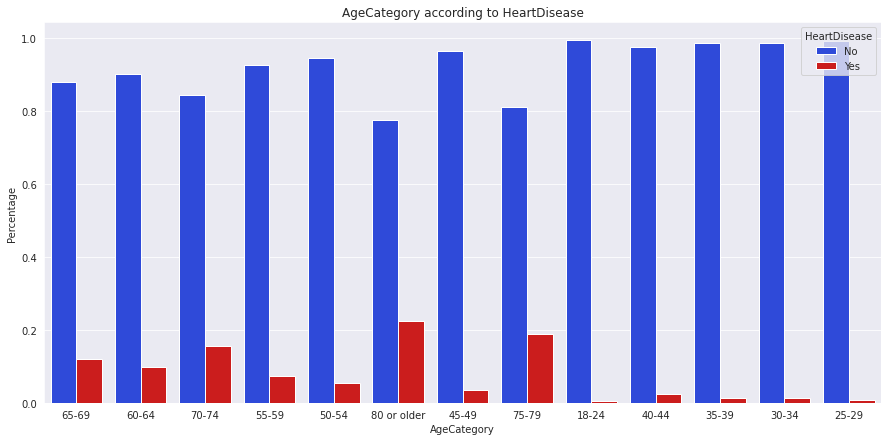
Moreover, the dataset has no missing values and do not need to handle it. Out of other notable points we can mention to the BMI feature which is skewed due to being out of normal range and PhysicalHealth and MentalHealth features which as well are severely skewed due to the number of zeros.

2.5 **Top Research Questions**

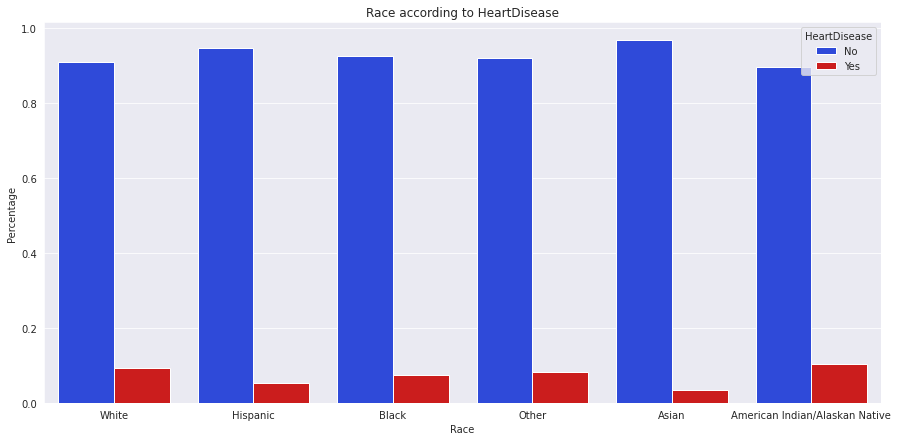
**2.5.1 Relationship between Sex and Heart Disease: Are males more likely to suffer from heart disease?** As we can see, in our data, males are more susceptible to the heart disease.



**2.5.2 Relationship between Age and Heart Disease: Are older individuals more susceptible to heart disease?** According to the below figure, we can see that age plays a major role in heart disease, as the amount of heart disease patients increases with age. The most susceptible people to the heart disease are those who are greater than 70 years old.

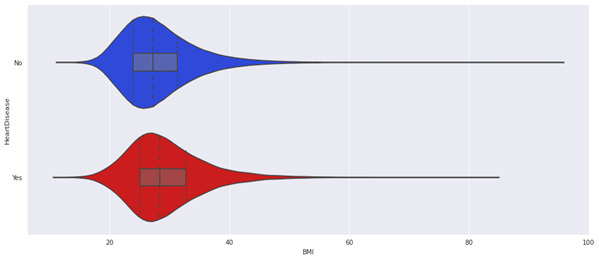


**2.5.3 Relationship between Race and Heart Disease: What is the percentage of heart disease among races?** In this plot, we can see the most people have the heart disease belongs to the Native Americans which percentage of heart disease is the highest (> 10%), followed by Whites (~9%). The least percentage of heart disease (~3%) is among Asians.

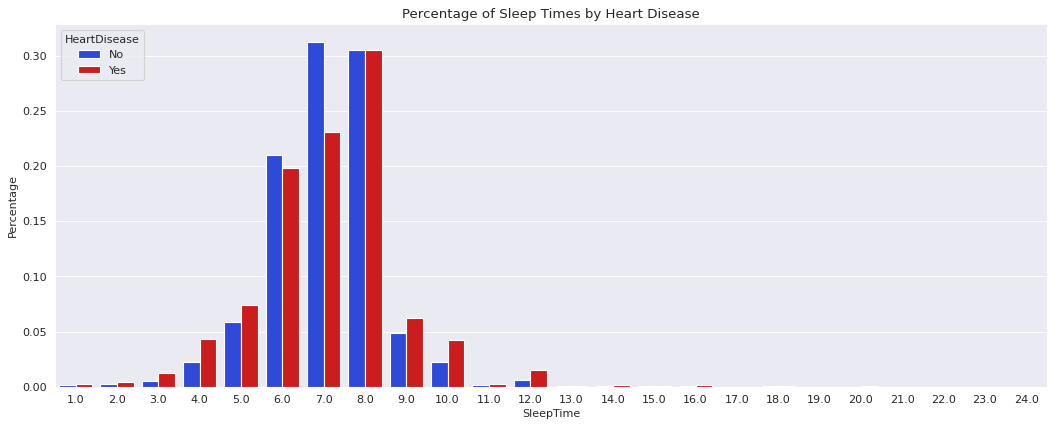
****

**2.5.4 Relationship between BMI and Heart Disease: Is the BMI of individuals affect on heart disease?**

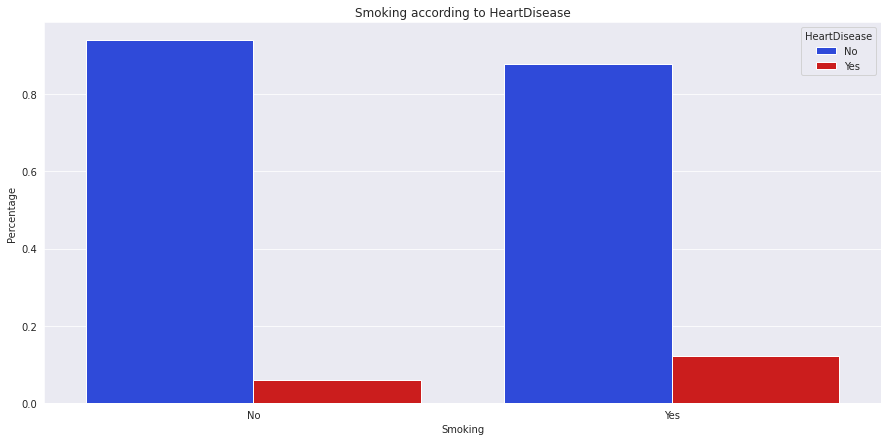
Based on below figure, both distributions are normal distributions and are in the same range which is from 12 to 94. However, the BMI distribution of individuals who suffer from heart disease is slightly shifted towards higher values in comparison to the distribution of those who don't. So, we can observe that BMI affect on heart disease.

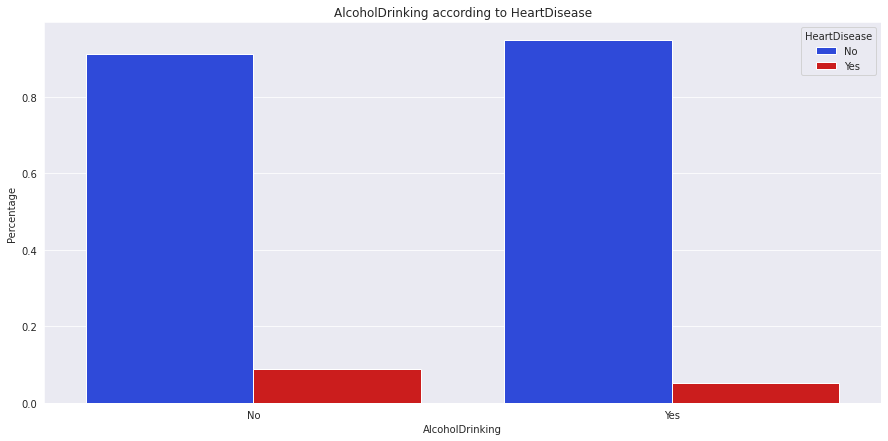
****

**2.5.5 Relationship between Sleep and Heart Disease: Is the distribution of sleep time among heart disease patients different?** Abnormal sleep duration is more prevalent in heart disease patients. Even though heart disease patients make 8.5% of the sample, they have higher percentages of sleep less than 6 hours or more than 9 hours, which is considered abnormal.

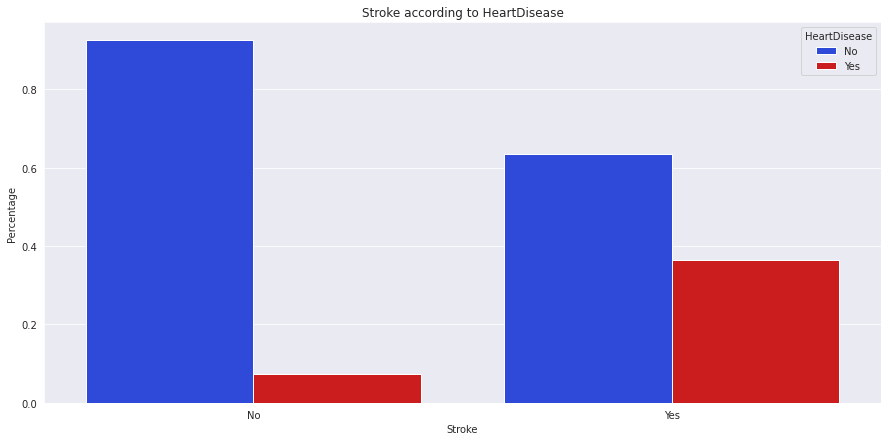
****

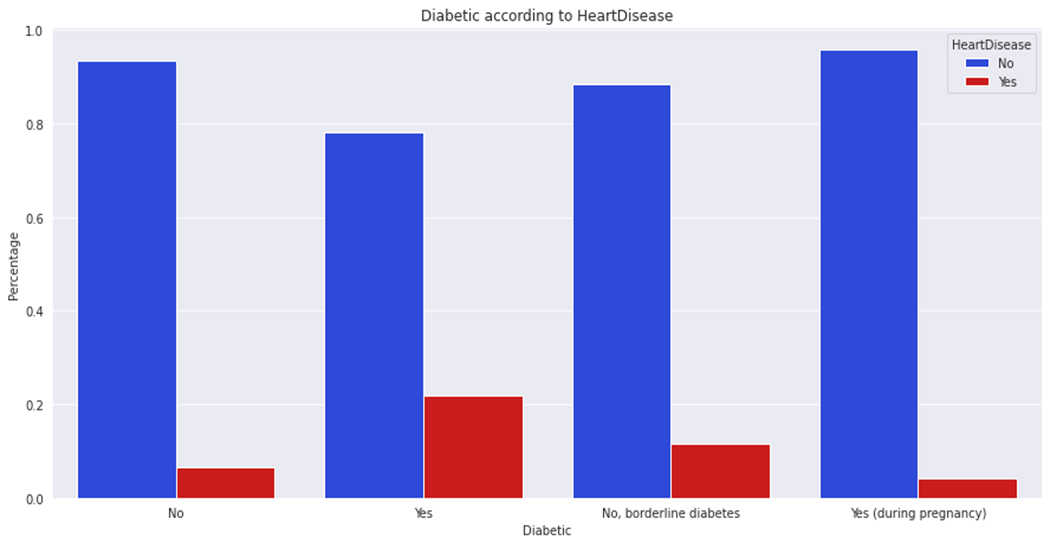
**2.5.6 Relationship between Alcohol Drinking and Smoking with Heart Disease: Do people with heart disease smoke more? Do people with heart disease consume more alcohol?** We can observe that the people who are smoking are more susceptible to the heart disease. However, it is interesting that some people who are not drinking alcohol have a heart disease too.

****

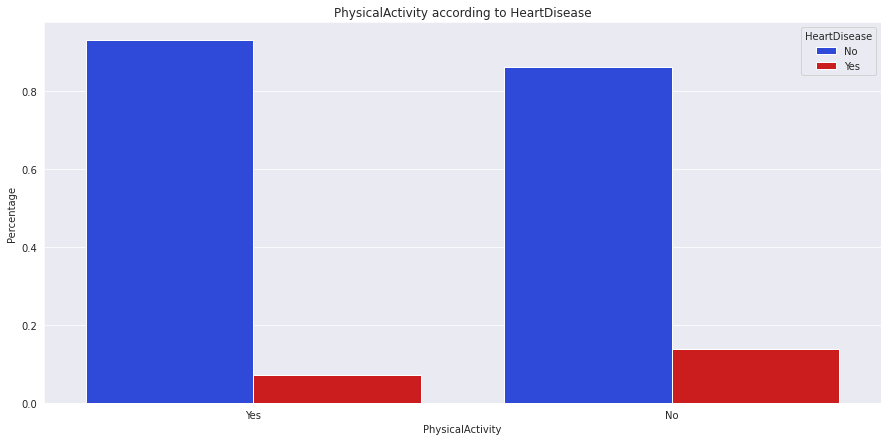
****

**2.5.7 Relationship between some diseases like Stroke or Diabetic with Heart Disease: Does having a Stroke affect the chances of heart disease?** **Does being diabetic increase the chances of heart disease?** It is obvious that Stroke is highly correlated with heart disease and people who have Diabetic or even they are in the borderline of it got more infected to the heart disease.

****

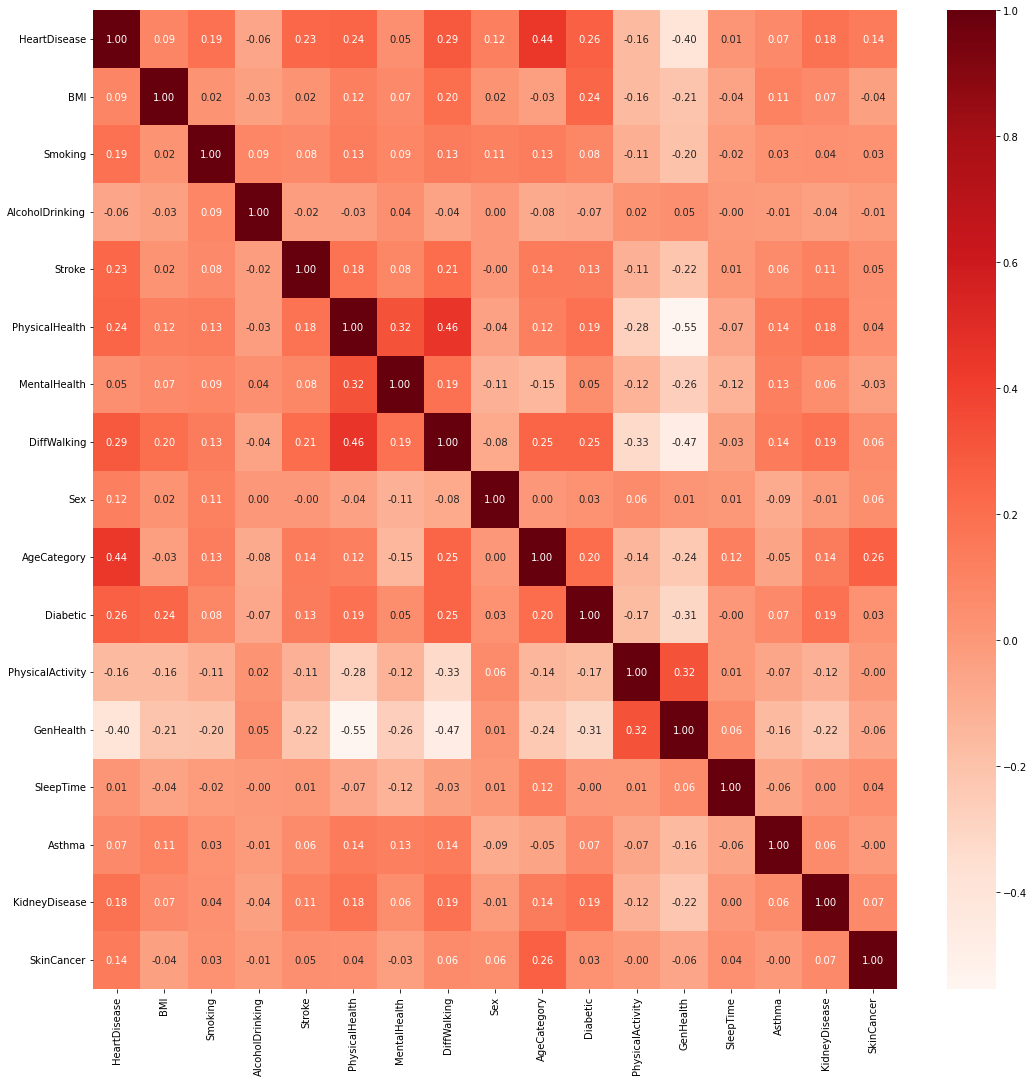
****

**2.5.8 Relationship between Physical Activity and Heart Disease? Do heart disease patients have less physical Activity or not?** People who reported doing physical activity or exercise during the past 30 days other than their regular job have a less heart disease compared to those who didn't make any physical activity. A larger percentage of heart disease patients are not physically active.

****

**2.6 Correlation Matrix**

A correlation matrix is a table showing correlation coefficients between features, each cell in the table shows the correlation between two variables [ ]. It is used to investigate the dependence between multiple variables at the same time. Below figure indicates the Correlation Matrix of all features of heart disease data set. As we can see, most of the correlation coefficients are smaller than 0.3 , and since in general if most of the correlation coefficients be smaller than 0.3, PCA (Principle Component Analysis) [ ] which is a popular technique for reducing the dimensionality of large datasets will not assist. Therefore, we did not see the necessity of using PCA to reduce the dimensionality of our dataset.



**3 Data Preprocessing**

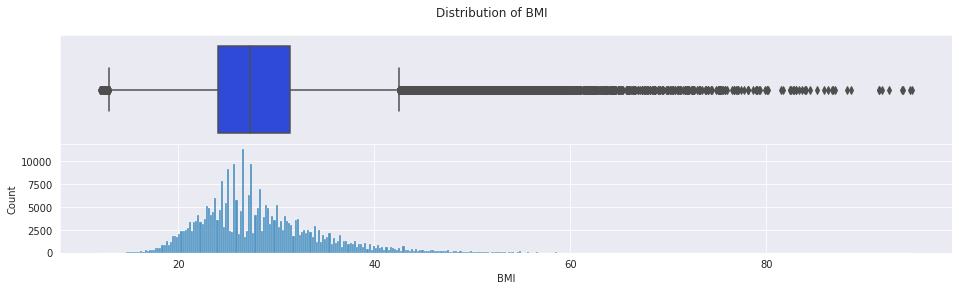
Data Preprocessing is the first and crucial step while creating a machine learning model which is a process of preparing the raw data and making it suitable for it. Therefore, it refers to manipulation or dropping of data before it is used in order to ensure or enhance performance model [tableau.com]. This section could be consisted of different stages such as Data Cleaning, Data Transformation and Data Normalization/Standardization. In the following, these stages will be explained regarding our heart disease dataset.

**3.1 Data Cleaning**

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset [tableau.com]. It could be composed of finding duplicate data, handling missing values, detection of Outliers and removing them. In our dataset, after checking the existence of missing values and also duplicate data, we concluded that there is not any such kind of data in this dataset in order to handle it, therefore this step is omitted from our analysis.

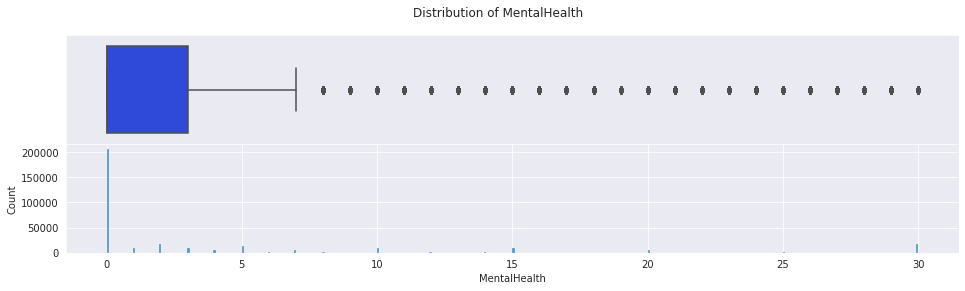
**3.2 Outliers Detection**

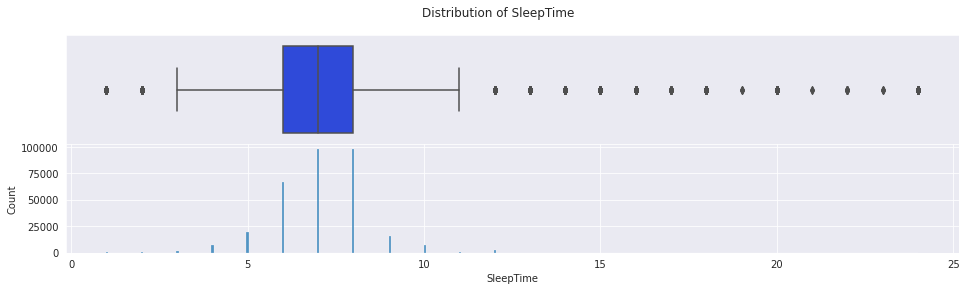
In statistics, an outlier is a data point that differs significantly from other observations [1]. In other words, in data analytics, outliers are values within a dataset that vary greatly from the others, they are either much larger, or significantly smaller. In our dataset, we should explore outliers only in the numerical features since there is no such concept of an outlier in the categorical data. In descriptive statistics, a box plot or boxplot is a method for graphically demonstrating the locality, spread and skewness groups of numerical data through their quartiles [2]. Below, the distribution of four numerical features of heart disease dataset is indicated that we can see that all numerical variables are skewed and contain outliers.



Table

Description automatically generated

****

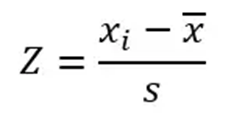
****

And to be more specific, we detected some observations that could be considered as outliers with using the IQR (Inter-Quartile Range) measure. IQR is a measure of statistical dispersion, which is the spread of the data [] and it is defined as the difference between the 75th and 25th percentiles of the data. The data points which fall below Q1 – 1.5 IQR or above Q3 + 1.5 IQR are outliers. So, finally we got 113661 outliers in total from all our numerical features.

|  |  |  |
| --- | --- | --- |
| Numerical Feature | Outlier Num | Outlier Percentage |
| BMI | 10396 | 3.25% |
| PhysicalHealth | 47146 | 14.74% |
| MentalHealth | 51576 | 16.13% |
| SleepTime | 4543 | 1.42% |
| Total | 113661 |  |

**3.3 Outliers Removal**

After detecting the outliers, we removed them by using Z-score which is a highly efficient method of detecting and removing outliers. Z-scores are the number of standard deviations above and below the mean decrease for each value. The standard cut-off value for finding outliers is any Z-score greater than +3 or less than -3 from zero which is considered as outlier [3]. To calculate the Z-score for an observation, it takes the original data, subtract the mean, and divide by the standard deviation. Hence, mathematically, the formula is:



As a result of detecting and removing the outliers of our dataset, we reached to a dataset with the size of 294402 \*17.

**3.4 Data Transformation**

Data transformation is the process of converting data from one format or structure into another [tableau.com]. As all machine learning algorithms are based on mathematics, we need to convert all the columns into numerical format. Heart Disease dataset is composed of a combination of categorical and numerical features. As mentioned before, the dataset is involving 4 numerical attributes and 14 categorical attributes. Out of these 14 categorical features, one is the target label which consists of “Yes” for positive class and “No” for negative class. And 9 of them are binary variables, means that they just have 2 unique values (Yes and No). However, 4 of them are multi categorical variables and have 4,5,6 and 13 unique values respectively. In our analysis, we use One-Hot Encoding technique for transformation of these multi categorical variables to numerical variables and by using Label Encoding method we converted binary categorical features to numerical ones. In the end, by doing this, we had a dataset with the dimension of 294402 \* 50.

**3.5 Feature Scaling**

Feature Scaling or Standardization is an important technique that is mostly performed as a preprocessing step before many machine learning models, to standardize the range of features of an input data set. In this step we used StandardScaler which removes the Mean and scales each feature to unit Variance to standardize our numerical features before using classifiers to train the data.

**3.6 Splitting the dataset**

After doing the above-mentioned steps, we split the dataset into train set and test set by the ratio of 75% and 25%, respectively which ended up having 220801 observations for the train data and 73600 observations was left for the test data.

1. **Solving Imbalance Class Problem**

A classification data set with skewed class proportions is called imbalanced. Classes that make up a large proportion of the data set are called majority classes. Those that make up a smaller proportion are minority classes [4]. As a matter of fact, imbalance class problem is naturally arise in many real-world applications like disease diagnosis, fraud detection, computer security, where the positive class occurs with reduced frequency. Most machine learning algorithms assume data equally distributed. So, when we have a class imbalance, the machine learning classifier tends to be more biased towards the majority class, causing bad classification of the minority class [5]. Therefore, effective classification with imbalanced data is crucial. As mentioned previously, the heart disease dataset has an extreme imbalance class problem. So, we tried to cope with this problem by conducting different methods such as Random Over-Sampling [ ], Random Under- Sampling [ ] and SMOTE (Synthetic Minority Over-Sampling Technique) [ ].

5 **Training Models**

Training is the most important step in machine learning, in which the prepared data is passed to a machine learning model to find patterns and make predictions. We applied different supervised learning models such as Naïve Bayes [ ], KNN [ ], Decision Tree [ ] and Random Forest [ ] to predict heart disease.

6 **Evaluation of Models**

After training various model, we must check to see how they are performing. This was done by testing the performance of the models on previously unseen data (our test set). In order to see the effect of the work we intended to do; we used these models before doing any special task on the data set and we noted the results. In below table, these results are shown.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Heart Disease Precision | No Heart Disease Precision | Heart Disease Recall | No Heart Disease Recall | Heart Disease F1 | No Heart Disease F1 | Accuracy |
| Naïve Bayes | 0.22 | 0.97 | 0.72 | 0.75 | 0.34 | 0.85 | 0.75 |
| KNN | 0.39 | 0.92 | 0.04 | 0.99 | 0.07 | 0.95 | 0.91 |
| Decision Tree | 0.23 | 0.93 | 0.25 | 0.92 | 0.24 | 0.92 | 0.86 |
| Random Forest | 0.33 | 0.92 | 0.13 | 0.97 | 0.18 | 0.95 | 0.90 |

**6.1 Performance Metrics**

|  |
| --- |
| **TP**: The number of predictions where the classifier correctly predicts the positive class as positive.  **TN**: The number of predictions where the classifier correctly predicts the negative class as negative.  **FP**: The number of predictions where the classifier incorrectly predicts the negative class as positive.  **FN**: The number of predictions where the classifier incorrectly predicts the positive class as negative. |

Confusion Matrix is generally used for classification problems which is a tabular way of visualizing the performance of prediction models []. As it is shown in below it consists of four components TP (True Positive), TN (True Negative), FP (False Positive) and FN (False Negative). As the heart disease prediction is a binary classification problem it has only two classes to classify, a positive and a negative class.

|  |  |  |  |
| --- | --- | --- | --- |
|  | | **Actual Class** | |
| **Positive** | **Negative** |
| **Predicted Class** | **Positive** | **TP** | **FN** |
| **Negative** | **FP** | **TN** |

When using classification models in machine learning, one metric that is often used to assess the quality of a model is Accuracy, which is the percentage of correct predictions to the total number of input samples. However, it works well only if there are equal number of samples belonging to each class. Hence, evaluating the performance based on Accuracy metric is not a suitable measure if the data is imbalanced. Therefore, in addition the Accuracy, Precision, Recall, F1-Score were considered for the performance evaluation of different models. The values of Accuracy, Precision, Recall, F1-Score were calculated as follows:

Precision: It refers to what fraction of predictions as a positive class were actually positive.

Recall: It refers to what fraction of all positive samples were correctly predicted as positive by the classifier. It is also known as True Positive Rate (TPR) or Sensitivity.

F1-Score: It combines Precision and Recall into a single measure. Mathematically it’s the harmonic mean of Precision and Recall.

Furthermore, in disease diagnosis like heart disease prediction, reducing false negatives is so crucial because it could be the difference between life and death. On the other hand, false positives are not as important, as further tests to the patient would reveal the misdiagnosis. Therefore, we were looking for high recall for the heart disease class. Below is the comparison of the models after data cleaning and trying to balance imbalanced class based on the above-mentioned metrics:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Model | Class Imbalance  Technique | Heart Disease Precision | No Heart Disease Precision | Heart Disease Recall | No Heart Disease Recall | Heart Disease F1 | No Heart Disease F1 | Accuracy |
| Naïve Bayes | Under-Sampling | 0.18 | 0.97 | 0.78 | 0.71 | 0.29 | 0.82 | 0.71 |
| Over-Sampling | 0.18 | 0.98 | 0.78 | 0.71 | 0.29 | 0.82 | 0.71 |
| SMOTE | 0.15 | 0.98 | 0.83 | 0.63 | 0.26 | 0.77 | 0.64 |
| KNN | Under-Sampling | 0.19 | 0.96 | 0.65 | 0.78 | 0.30 | 0.86 | 0.77 |
| Over-Sampling | 0.19 | 0.95 | 0.49 | 0.83 | 0.28 | 0.89 | 0.81 |
| SMOTE | 0.19 | 0.95 | 0.49 | 0.83 | 0.28 | 0.89 | 0.81 |
| Decision Tree | Under-Sampling | 0.14 | 0.96 | 0.66 | 0.67 | 0.23 | 0.79 | 0.67 |
| Over-Sampling | 0.19 | 0.93 | 0.20 | 0.93 | 0.20 | 0.93 | 0.88 |
| SMOTE | 0.19 | 0.94 | 0.27 | 0.90 | 0.22 | 0.92 | 0.86 |
| Random Forest | Under-Sampling | 0.18 | 0.97 | 0.71 | 0.73 | 0.29 | 0.84 | 0.73 |
| Over-Sampling | 0.24 | 0.93 | 0.18 | 0.95 | 0.21 | 0.94 | 0.89 |
| SMOTE | 0.19 | 0.94 | 0.27 | 0.91 | 0.22 | 0.92 | 0.86 |

**Hyperparameter Tuning**

Once we created and evaluated our models, we did hyperparameter tuning to see if its accuracy can be improved in any way.

**K-fold Cross-validation**

7 **Conclusion and Future Work**

Most cardiovascular diseases can be prevented by addressing behavioural risk factors such as tobacco use, unhealthy diet and obesity, physical inactivity, and harmful use of alcohol. Hence, it is important to detect cardiovascular diseases as early as possible. To identify the causes of heart disease, we analyzed the Personal Key Indicators of Heart Disease which had 17 indicators of heart disease of 319,795 surveyed individuals in the USA. During our investigation we identified that age is a major factor in heart disease. Furthermore, heart disease can be seen more in smokers (~12%), kidney disease (~30%), stroke victims (~48%), skin cancer patients (~18%), people who have difficulty in walking (~18%), and diabetics (~25%). Finally, after experimenting with different models we concluded that Naïve Bayes with SMOTE yields the best Recall for the Heart Disease Class (83%), Random Forest with Over-Sampling yields the best Recall for the No Heart Disease Class (95%). When it comes to the F1-Score metric, KNN with Under-Sampling yields the best for the Heart Disease Class (30%) and Random Forest with Over-Sampling yields the best for the No Heart Disease Class (94%).

**References**

0. Tsao CW, Aday AW, Almarzooq ZI, Beaton AZ, Bittencourt MS, Boehme AK, et al. Heart Disease and Stroke Statistics—2022 Update: A Report From the American Heart Association. Circulation. 2022;145(8):e153–e639.

<https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease>

1. Maddala, G. S. (1992). "Outliers". Introduction to Econometrics (2nd ed.). New York: MacMillan. pp. 89. ISBN 978-0-02-374545-4. An outlier is an observation that is far removed from the rest of the observations.

2. C., Dutoit, S. H. (2012). Graphical exploratory data analysis. Springer. ISBN 978-1-4612-9371-2. OCLC 1019645745

3. <https://towardsdatascience.com/outlier-detection-part1-821d714524c>

4.https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data

5. <https://towardsdatascience.com/class-imbalance-a-classification-headache-1939297ff4a4>

Jolliffe, Ian T.; Cadima, Jorge (2016-04-13). "Principal component analysis: a review and recent developments". Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 374 (2065): 20150202. doi:10.1098/rsta.2015.0202. PMC 4792409. PMID 26953178.