Final Phase:

Title: Heart disease prediction

Student: Anibal Ruiz

Abstraction:

With the rampant increase in the heart disease rates at all ages, we need to do a study based on some data to be able to detect the symptoms of a heart stroke and thus prevent it. Thus we propose to classify the data which can give ud some predictions fot he vulnerability of a heart disease given basic symptoms like age, sex, sleep hours, BMI, etc. The machine learning algorithm Random Forest has proven to be one of the most accurate and reliable algorithm and hence used in the proposed system.

**Introduction:**

For this project I have chosen data about heart disease prediction because in the last months I’ve been trying to change my lifestyle and be healthier, and having good habits are probably the most important things we need to improve to accomplishes that change. The data is proposed to help in understanding the relationship between some attributes and having a heart disease. According to the CDC, 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. Detecting and preventing the factors that have the greatest impact on heart disease is very important in healthcare. Computational developments, in turn, allow the application of machine learning methods to detect "patterns" from the data that can predict a patient's condition.

First at all, my plan is to visualize the data. I would like to see graphs to discover things like noise. My prediction of the data is that if you are a smoker and have bad habits like drink alcohol or not sleeping enough time, you’ll probably suffer a heart disease.

**Data Description:**

The original dataset of nearly 300 variables was reduced to just about 20 variables. In addition to classical EDA, this dataset can be used to apply a range of machine learning methods, most notably classifier models (logistic regression, SVM, random forest, etc.). You should treat the variable "HeartDisease" as a binary ("Yes" - respondent had heart disease; "No" - respondent had no heart disease). But note that classes are not balanced, so the classic model application approach is not advisable. Fixing the weights/under sampling should yield significantly betters results.

The data has 17 attributes and the predicted class: heart disease: yes or no. the explanation of the attributes is:

1. Heart Disease: 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? ( The answer 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
5. Stroke: (Ever told) (you had) a stroke?
6. Physical Health: 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).
7. Mental Health: Thinking about your mental health, for how many days during the past 30 days was your mental health not good? (0-30 days).
8. DiffWalking: Do you have serious difficulty walking or climbing stairs?
9. Sex: Are you male or female?
10. Age Category: Fourteen-level age category.
11. Race: Imputed race/ethnicity value.
12. Diabetic: (Ever told) (you had) diabetes?
13. Physical Activity: 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. Kidney Disease: Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?
18. Skin Cancer: (Ever told) (you had) skin cancer?

Size of data: 18 columns and 319795 entries.

**Data Cleaning and preprocessing:**

**Load data into Python:**

RangeIndex: 319795 entries, 0 to 319794

Data columns (total 18 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 HeartDisease 319795 non-null object

1 BMI 319795 non-null float64

2 Smoking 319795 non-null object

3 AlcoholDrinking 319795 non-null object

4 Stroke 319795 non-null object

5 PhysicalHealth 319795 non-null float64

6 MentalHealth 319795 non-null float64

7 DiffWalking 319795 non-null object

8 Sex 319795 non-null object

9 AgeCategory 319795 non-null object

10 Race 319795 non-null object

11 Diabetic 319795 non-null object

12 PhysicalActivity 319795 non-null object

13 GenHealth 319795 non-null object

14 SleepTime 319795 non-null float64

15 Asthma 319795 non-null object

16 KidneyDisease 319795 non-null object

17 SkinCancer 319795 non-null object

dtypes: float64(4), object(14)

memory usage: 43.9+ MB

After loading the information data into python, this is the information I got. NO Null values in the data to clean for now. Four attributes (BMI, Physical Health, Mental health and Sleep time) are integers data type. They are also very important attributes for my prediction to heart disease.

The total of columns are 18 (attributes) and 319,795 entries. The target is Heart disease.

When you will look at the data sample and the summary, first of all, you will find that you need to convert a lot of features into numeric ones, so that the machine learning algorithms can process them. Furthermore, you will see that the features have widely different ranges, that you will need to convert into roughly the same scale. You can also spot some more features, that contain missing values (NaN = not a number), that you need to deal with.

**Any missing values:**

HeartDisease False

BMI False

Smoking False

AlcoholDrinking False

Stroke False

PhysicalHealth False

MentalHealth False

DiffWalking False

Sex False

AgeCategory False

Race False

Diabetic False

PhysicalActivity False

GenHealth False

SleepTime False

Asthma False

KidneyDisease False

SkinCancer False

dtype: bool

No Missing values.

**Converting data:**

Basically all (or almost all) attributes have binary data (yes or no). Some of them classified for categories. In order to make my work simpler, I’ll convert the data to integers. (At least the attributes I think would have an import ROLE in the prediction of heart Disease.

**HeartDisease**

First, I converted the heartDisease attribute as follows:

0 = No

1 = Yes

I would be easy for prediction and to using standard deviationa dn the mean. Also, for doing the classification.

**Age**

The age was classified for categories. To make it simple and easy for use, I converted the values to an specific age. For example, the people for “80 or older” I gave them an age of ‘80’. For the people “75-79” I gave them an age of 75, and so on. This will make my data simplier to read and understand.

**Sex**

Changed the attribute as follows:

Female = 0, and Male = 1

**Smoking**

The smoking attribute is also binary. So, I converted it as follows:

No = 0, Yes = 1.

**AlcoholDrinking**

The AlcoholDrinking is converted as follows:

No = 0, Yes = 1.

**SkinCancer**

The SkinCancer attribute converted as follows:

No = 0, yes = 1

**Asthma**

The Asthma attribute converted as follows:

No = 0, yes = 1

**KidneyCancer**

The KidneyCancer attribute converted as follows:

No = 0, Yes = 1

**DiffWalking**

The difficulty Walking converted as follows:

No = 0, Yes = 1.

**DATA EXPLORATORY AND VISUALIZATION**

**General Description:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Count | Mean | STD | Min | 25% | 50% | 75% | Max |
| BMI | 319795.0 | 28.33 | 6.36 | 12.02 | 24.03 | 27.34 | 31.42 | 94.85 |
| PhysicalHealth | 319795.0 | 3.37 | 7.95 | 0.00 | 0.00 | 0.00 | 2.00 | 30.00 |
| MentalHealth | 319795.0 | 3.898 | 7.96 | 0.00 | 0.00 | 0.00 | 3.00 | 30.00 |
| SleepTime | 319795.0 | 0.097 | 1.44 | 1.00 | 6.00 | 7.00 | 8.00 | 24.00 |
| HeartDisease | 319795.0 | 0.085 | 0.2797 | 0.00 | 0.00 | 0.00 | 0.00 | 1.00 |

After doing a description of the data (int datatypes), we can get the following:

* The mean of BMI (Body Mass Index) is: 28.33

A BMI of less than 18.5 means a person is underweight. A BMI of between 18.5 and 24.9 is ideal. A BMI of between 25 and 29.9 is overweight. A BMI over 30 indicates obesity.

* The Standard deviation of BMI is: 6.36

Low standard deviation means data are clustered around the mean, and high standard deviation indicates data are more spread out.

MOST IMPORTANT:

What percentage of people have a heart disease?

Approximately 8%

**CORRELATION TABLE:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | BMI | Physical Health | Mental Health | Sleep time |
| BMI | 1.000000 | 0.109788 | 0.064131 | -0.051822 |
| Physical Health | 0.109788 | 1.000000 | 0.287987 | -0.061387 |
| Mental Health | 0.064131 | 0.287987 | 1.000000 | -0.119717 |
| Sleep time | -0.051822 | -0.031387 | -0.119717 | 1.000000 |

First at all, almost every integer attribute, except sleep time, compared to each other we get a positive correlation, where it indicates that both attributes increase together.

* Sleep time compared to any attribute will get a negative correlation. It means whatever value tends to increase, the other decreases.

Also, the closer to 0 is a correlation, the weaker the linear relation is. Sleep time compared to any other attribute has no relation at all as their correlation is closer to 0.

Mental health and physical health are the attributes with more linear relation in the data.

AGE OF PEOPLE IN DATA:

65-69 34151

60-64 33686

70-74 31065

55-59 29757

50-54 25382

80 or older 24153

45-49 21791

75-79 21482

18-24 21064

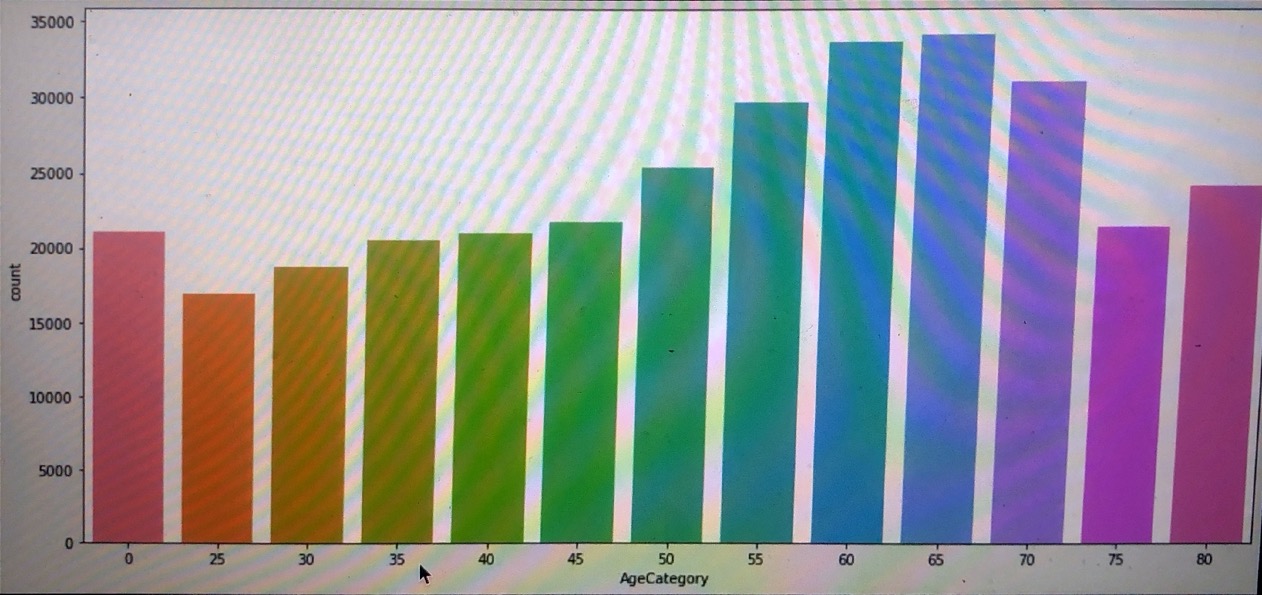
40-44 21006

35-39 20550

30-34 18753

25-29 16955

Name: AgeCategory, dtype: int64



Important notes:

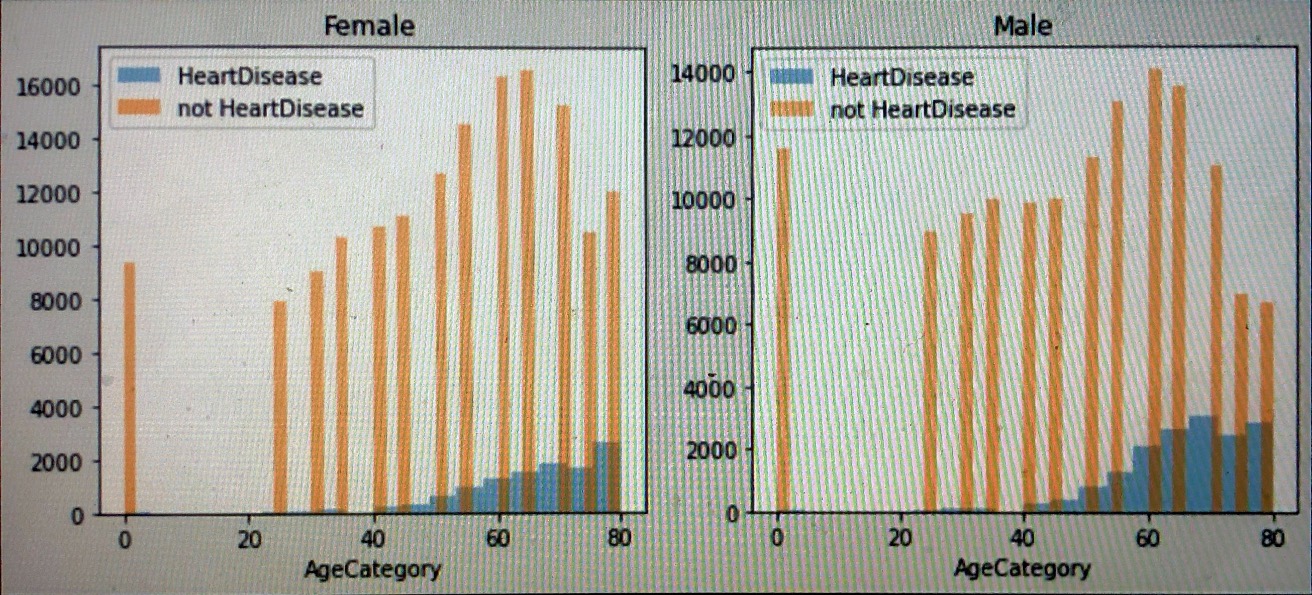
-All the participants are adults.

-Good number of participants for each range of ages. (Balance)

**-Most participants are in the range of “50 or older”. Which make infer that if you are older, you got more chances to have a heart disease.**

**AGE AND SEX**

Create histograms for male (heartDisease vs not heartDisease) and for female (heartDisease vs not heartDisease) where the x axis represents their age and y axis will represent the count of male/female in that age bucket who has heartDisease / not have heartDisease.

****

**This histogram is clear. The sex is not important for a heart disease. Either females or males have the same tendency to have heart problems when they are older. The most cases is between 60 and 80 years old.**

**Smoking, Gen Health and Sex (correlation):**

I did a correlation graph between Gen health and sex based on Smoking.

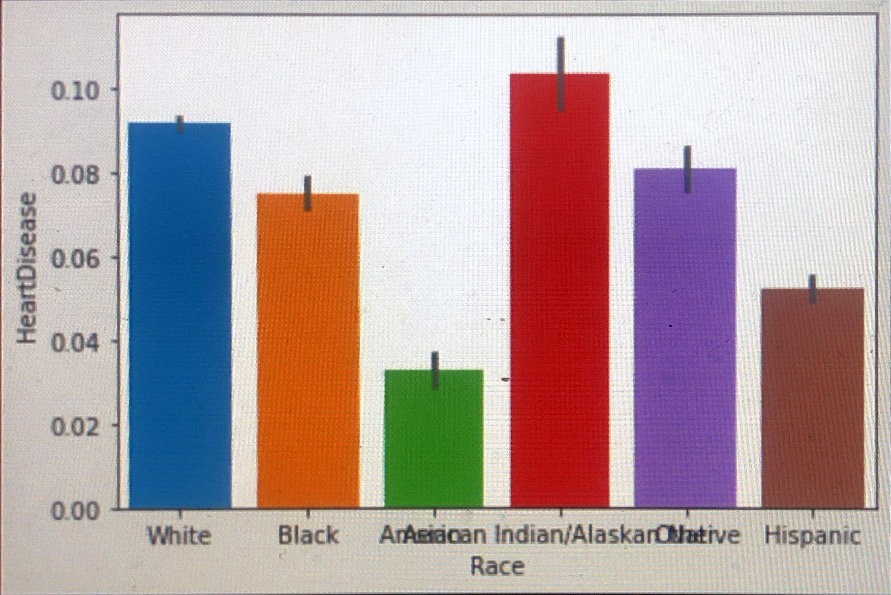
Some obvious results:

If you have a strong Gen Health (good or Very good), you are more likely not to have a heart disease even if you are a smoker.

If you have a weak Gen Health (fair or poor), you are more likely to have a heart disease even if you are NOT a smoker.

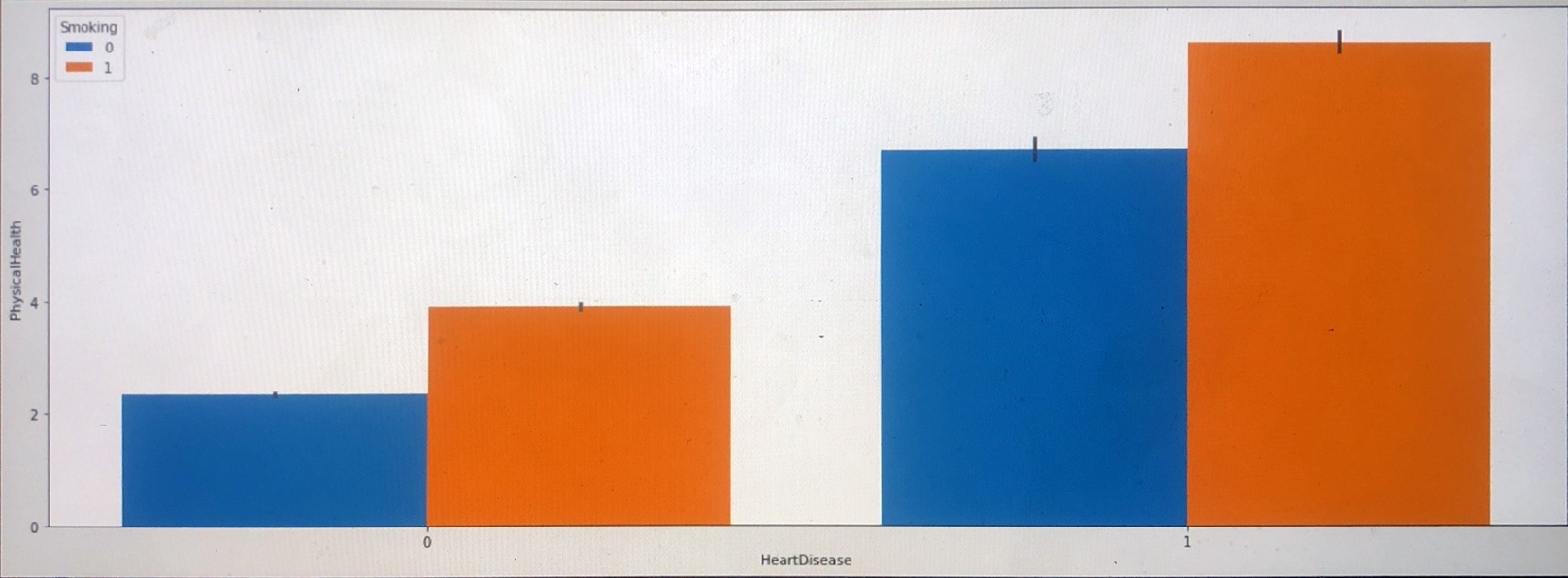
\*Men are most likely to suffer a heart disease than women. Just a little bit more likely.

**Race and Heart Disease:**

****

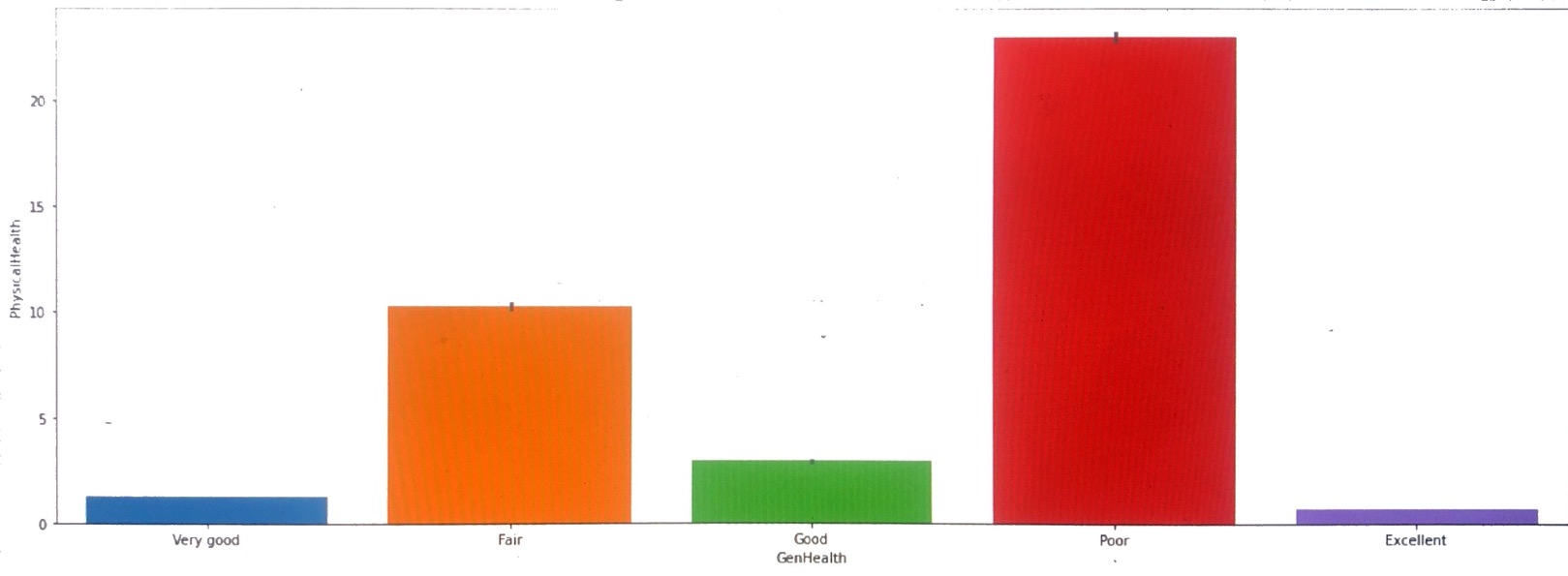
**American Indian/Alaska Native is the race most likely to have a heart disease. In second position white people. Asian and Hispanic are the less likely to have a heart disease.**

**PHYSICAL HEALTH AND SMOKING:**

****

Most people with heart disease have a bad physical health.

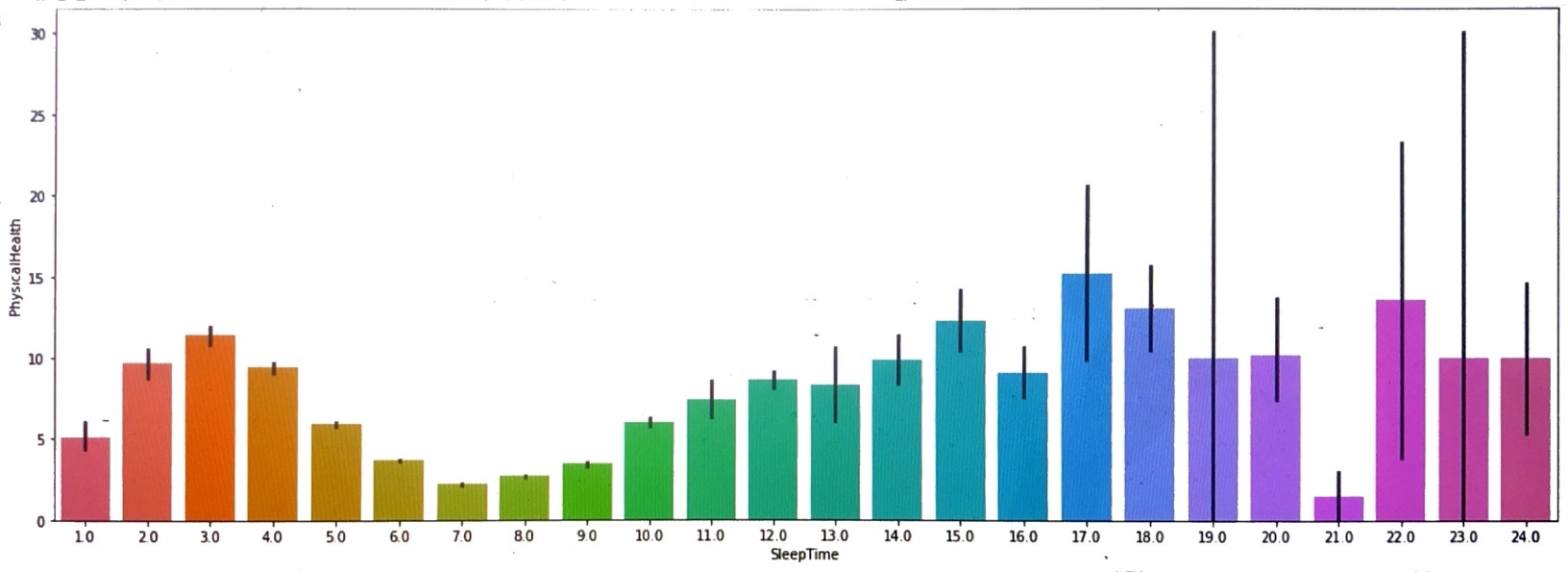
**GenHealth and PhysicalHealth:**



Having a good GenHealth lead to a better physical health.

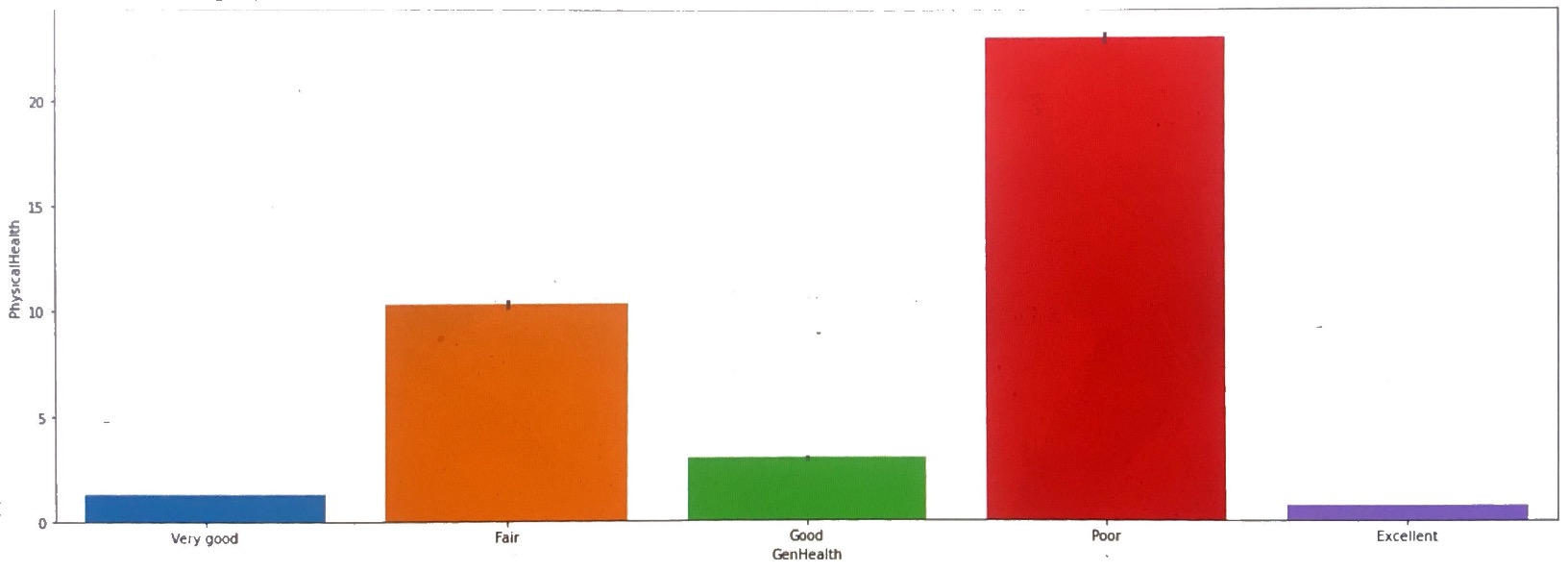
At this point, data is ready to be split in train and test data sets and make the classification.

**Sleep Time and Physical Health**



“If a person sleep between 6 and 9 hours average per day, that person has a good physical health.”

**Physical Health and Gen Health**



“Having a good GenHealth lead to better PhysicalHealth”

Final Phase: More changes to data before classification:

AgeCategory

‘7’ age categories were created to minimize the work.

0 to 11 -> 0

11 to 20 -> 1

21 to 30 -> 2

31 to 40 -> 3

41 to 50 -> 4

51 to 60 -> 5

61 to 70 -> 6

71 or greater -> 7

PhysicalHealth:

Converted as follows:

0 – No, 1 -yes

Stroke:

Converted as follows:

0 – no, 1 -yes

**● Data Modeling**

**Before doing any accuracy modeling, I did the MIN Scaler transformation for my training data.**

**70% for training, 30% for testing.**

**For my project, I’ll model my data using Random Forest with k-fold cross validation.**

Which algorithms are finally used, a brief introduction to them?

Random Forest: (main algorithm for my project)

Random Forest is a supervised learning algorithm. Like you can already see from it’s name, it creates a forest and makes it somehow random. The forest, It builds, is an ensemble of Decision Trees, most of the time trained with the “bagging” method. The general idea of the bagging method is that a combination of learning models increases the overall result. To say it in simple words: Random Forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Logistic Regression:

Logistic regression is a process of modeling the probability of a discrete outcome given an input variable. The most common logistic regression models is a binary outcome; something that can be two values such as true/false, yes/no, and so on. It is used in statical software to understand the relationship between the dependent variable and one or more independent variables by estimating the probabilities using a logistic regression equation.

K Nearest Neighbor:

K-Nearest Neighbor is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

Gaussian Naive Bayes:

Naive Bayes are a group of supervised machine learning classification algorithms based on the Bayes theorem. It is a simple classification technique but has high functionality. They find use when the dimensionality of the inputs is high. Complex classification problems can also be implemented by using Naive Bayes Classifier. One assumption taken is the strong independence assumptions between the features. These classifiers assume that the value of a particular feature is independent of the value of any other feature. In a supervised learning situation, Naive Bayes Classifiers are trained very efficiently. Naive Bayed classifiers need a small training data to estimate the parameters needed for classification. Naive Bayes Classifiers have simple design and implementation, and they can applied to many real life situations.

Perceptron:

The [Perceptron algorithm](https://en.wikipedia.org/wiki/Perceptron) is a two-class (binary) classification machine learning algorithm.

It is a type of neural network model, perhaps the simplest type of neural network model. It consists of a single node or neuron that takes a row of data as input and predicts a class label. This is achieved by calculating the weighted sum of the inputs and a bias (set to 1). The weighted sum of the input of the model is called the activation.

Linear Support Vector Machine:

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine.

Decision tree:

Decision Tree algorithm belongs to the family of supervised learning algorithms. Unlike other supervised learning algorithms, the decision tree algorithm can be used for solving **regression and classification problems** too. The goal of using a Decision Tree is to create a training model that can use to predict the class or value of the target variable by **learning simple decision rules** inferred from prior data(training data). In Decision Trees, for predicting a class label for a record we start from the **root** of the tree. We compare the values of the root attribute with the record’s attribute. Based on comparison, we follow the branch corresponding to that value and jump to the next node.

K-fold cross validation:

K-Fold Cross Validation randomly splits the training data into K subsets called folds. Let’s imagine you would split the given data into 4 folds (K = 4). Your random forest model would be trained and evaluated 4 times, using a different fold for evaluation every time, while it would be trained on the remaining 3 folds. The image below shows the process, using 4 folds (K = 4). Every row represents one training + evaluation process. In the first row, the model get’s trained on the first, second and third subset and evaluated on the fourth. In the second row, the model get’s trained on the second, third and fourth subset and evaluated on the first. K-Fold Cross Validation repeats this process till every fold acted once as an evaluation fold.

**● Evaluation**

Performance comparison of the different models.

Accuracy:

Random Forest: 96.2

Logistic Regression: 91.62

K Nearest Neighbor: 92.92

Gaussian Naive Bayes: 85.05

Perceptron: 84.51

Linear Support Vector Machine: 91.59

Decision tree: 96.21

Random Forest and Decision tree are the best models.

**Finding how Random Forest works with k-fold cross validation:**

Scores: [0.9014, 0.9032, 0.9015, 0.9022, 0.9020, 0.8998, 0.9019 0.8981, 0.9025, 0.9036]

Mean: 0.9016510590346509

Standard Deviation: 0.0015397808125594981

Our model has a average accuracy of 90% with a standard deviation of 0.15 %. The standard deviation shows us, how precise the estimates are.

Models comparison:

|  |  |  |
| --- | --- | --- |
| **Model** | **Mean** | **Standard Deviation** |
| KNN | 0.897564 | 0.001970 |
| Logistic Regression | 0.916196 | 0.000619 |
| Naïve Bayes | 0.850507 | 0.002169 |
| Perceptron | 0.910161 | 0.010409 |
| Linear SVM | 0.915906 | 0.000347 |
| Decision Tree | 0.889326 | 0.001480 |

Which model would you prefer to use based on your results? Why?

I prefer to use Random Forest because it has a good accuracy rate for my data and it is easier to understand. Using K-fold, Linear SVM and Logistic Regression, have the best mean but worse standard deviation than Random Forest. Also, Random Forest is based on Decision trees which I think it is easier to understand.

● **Discussion and Conclusion**

Discuss the results

The best machine models are Random Forest, Linear SVM, and Perceptron. Accuracy:

Random Forest: 96.2

Linear SVM: 91.59

Perceptron: 84.51

After using cross validation:

Random Forest: 90.0% accuracy and 0.15% standard deviation.

Linear SVM: 91% accuracy and 0.0347% standard deviation.

Perceptron: 91% accuracy and 0.0148% standard deviation.

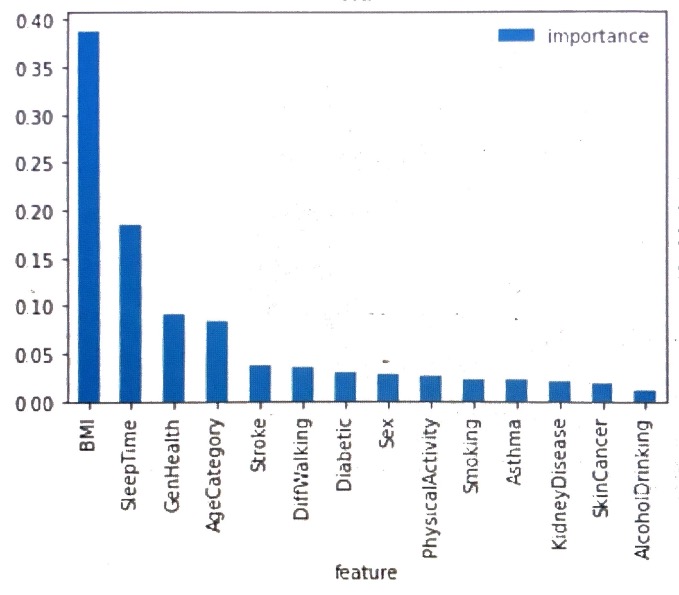
For simplicity and good results, I prefer to use Random Forest as my final model to classify the data.

  What have you learned from this project?

I learned that BMI (Body Mass Index) is the most important feature when prediction a heart disease. An overwheighted person is more feasible to get a heart disease.

The importance table looks like this:

|  | **importance** |
| --- | --- |
| **feature** |  |
| **BMI** | 0.388 |
| **SleepTime** | 0.185 |
| **GenHealth** | 0.091 |
| **AgeCategory** | 0.084 |
| **Stroke** | 0.037 |



At the beginning of my research, I thought AgeCategory, sleep hours, Smoking and Alcohol drinkers were at the top of the importance features. I was wrong on the last two.

Conclusion:

Heart disease kills roughly the same number of people on the United States each year as cancer, lower respiratory disease, and accidents combined. Prior studies have demonstrated lower all-cause mortality in individuals who are overweight compared with those with normal body mass index (BMI). Most importantly, heredity and gender also cause people to have heart diseases. Men have very higher risks to have heart disease than women, but after women reach the age of 65 they have about same risks as men do.

References:

https://www.kaggle.com/code/ahmedklabi/heart-disease-pred/notebook

2020 annual CDC survey data of 400k adults related to their health status.

<https://www.cdc.gov/heartdisease/index.htm> Center for Disease and Control prevention.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7521325/>