Project Documentation

CMP461
Spring 2022

Team #15

Name	Sec	BN
Saad Eldeen Mohamed Mohamed Ahmed Badr	1	26
Omar Ibrahim Elsayed Salama	2	7
Mohamed Adel Abdelmohsen	2	21
Ali Adel Esmail	2	5

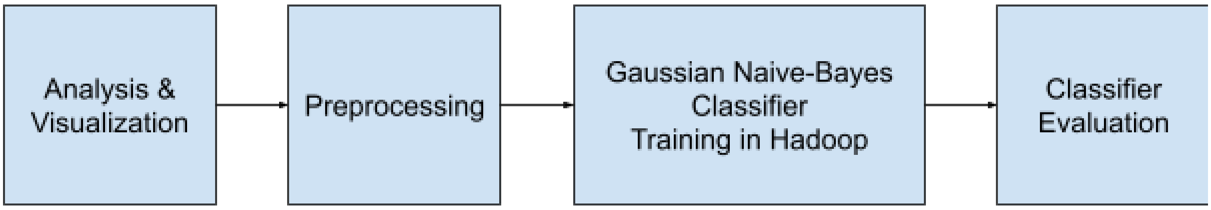
Brief problem description

Heart disease is one of the leading causes of death for people. 47% of Americans have at least 1 of 3 key risk factors for heart disease: high blood pressure, high cholesterol, and smoking. Other key indicators include 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.

We worked with Personal Key Indicators of Heart Disease | Kaggle, in the project we are providing:

- Analysis and visualization of the key indicators.
- Gaussian Naive Bayes classifier (training is done by mapreduce in hadoop).

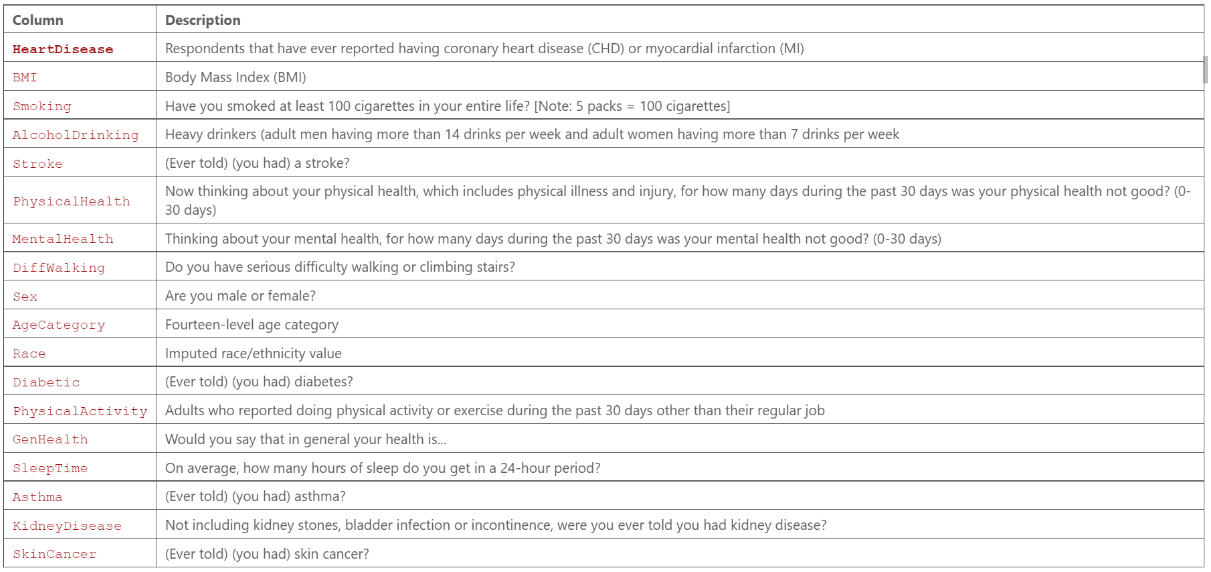
Project pipeline



Analysis and solution of the problem

Data preprocessing

These are the columns of the dataset



Column	Description
HeartDisease	Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
BMI	Body Mass Index (BMI)
Smoking	Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]
AlcoholDrinking	Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week
Stroke	(Ever told) (you had) a stroke?
PhysicalHealth	Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? (0-30 days)
MentalHealth	Thinking about your mental health, for how many days during the past 30 days was your mental health not good? (0-30 days)
DiffWalking	Do you have serious difficulty walking or climbing stairs?
Sex	Are you male or female?
AgeCategory	Fourteen-level age category
Race	Imputed race/ethnicity value
Diabetic	(Ever told) (you had) diabetes?
PhysicalActivity	Adults who reported doing physical activity or exercise during the past 30 days other than their regular job
GenHealth	Would you say that in general your health is...
SleepTime	On average, how many hours of sleep do you get in a 24-hour period?
Asthma	(Ever told) (you had) asthma?
KidneyDisease	Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?
SkinCancer	(Ever told) (you had) skin cancer?

Preprocessing responsibilities are:

- Encode **Diabetic** as numerical levels
  - No: 0 , pregnancy diabetic: 1, Borderline diabetes: 2, Yes: 3
    - pregnancy diabetic is temporary, so we encoded it close to “No”.
- Encode **GenHealth** as numerical levels
  - Poor: 0, Fair: 1, Good: 2, Very good: 3, Excellent: 4
- Convert **Age** to continuous feature instead of categorical
  - **AgeCategory** is in the shape of [min_age, max_age], so for each category we encoded it as the mean of its two ranges.
- Encode **Sex** as Male: 0, Female: 1
- Encode binary features as Yes: 1, No: 0
  - Binary Features: **Smoking**, **AlcoholDrinking**, **Stroke**, **DiffWalking**, **PhysicalActivity**, **Asthma**, **KidneyDisease**, **SkinCancer.**
- Encode target (**HeartDisease**) as Yes: 1, No: 0
- Drop **Race** feature as it is not significant.

Data visualization

Please refer to the notebook attached or view all visualizations at the end of this file.

Extracting insights from data.

Please refer to the notebook attached.

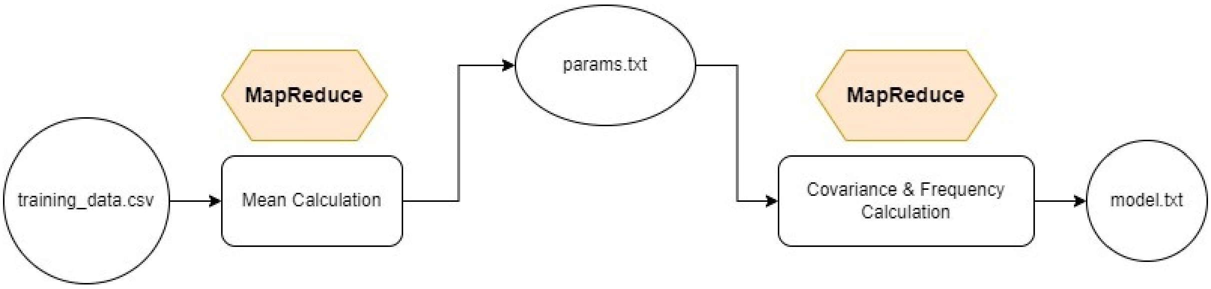
Some insights

- Males are more prone to heart disease.
- There is a negative correlation between general health and heart disease.
- There is a correlation between smoking and heart disease.
- Races appear as an insignificant feature regarding heart disease.
- Old people are more prone to Heart Disease.

Model/Classifier training

We used **Gaussian Naive-Bayes** classifier, and we did the training in hadoop using mapreduce.

- We splitted the data into **train** and **test** sets with a ratio *(0.7/0.3)*.
- Both train and test sets were unbalanced due to the unbalanced dataset *(**No**: 90% **Yes**: 10%)*, so our solution was to oversample from class **Yes** so that it has more values and undersample from class **No** to make it less dominant. This made us reach a ratio of *(**No**: 70% **Yes**: 30%)*.
- A MapReduce program running in hadoop did the training, which included finding means, covariances of features of each class and class probabilities to form **Gaussian Naive-Bayes** classifier.

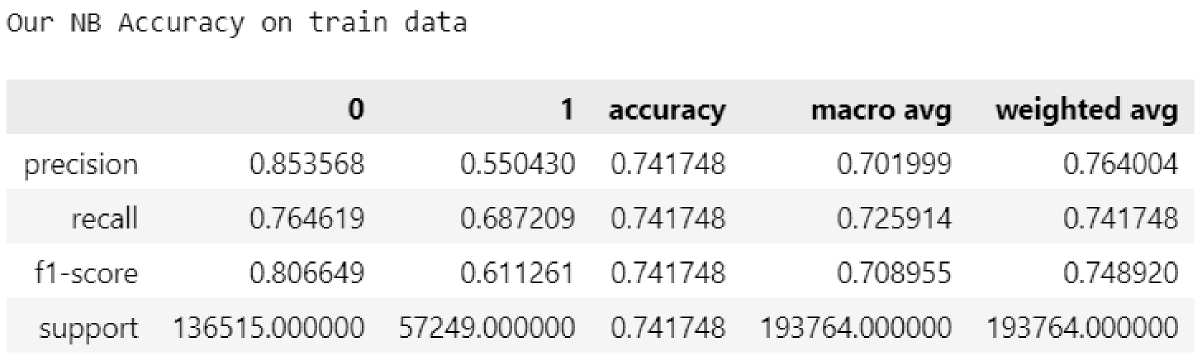


- For more details: 5 min video

Results and Evaluation

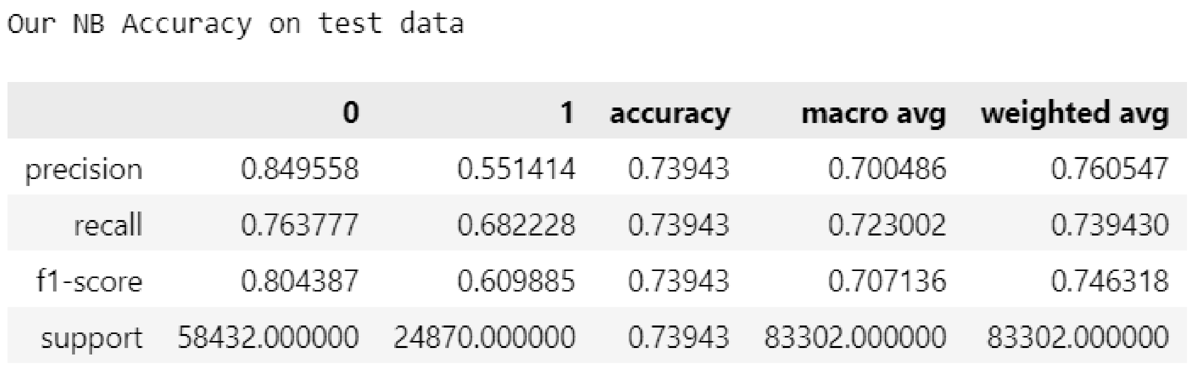
Gaussian Naive-Bayes classifier on train and test data (0.7/0.3).

(class 0: No Heart Disease - class 1: Yes Heart Disease)



```
Our NB Accuracy on train data
```

	0	1	accuracy	macro avg	weighted avg
precision	0.853568	0.550430	0.741748	0.701999	0.764004
recall	0.764619	0.687209	0.741748	0.725914	0.741748
f1-score	0.806649	0.611261	0.741748	0.708955	0.748920
support	136515.000000	57249.000000	0.741748	193764.000000	193764.000000



```
Our NB Accuracy on test data
```

	0	1	accuracy	macro avg	weighted avg
precision	0.849558	0.551414	0.73943	0.700486	0.760547
recall	0.763777	0.682228	0.73943	0.723002	0.739430
f1-score	0.804387	0.609885	0.73943	0.707136	0.746318
support	58432.000000	24870.000000	0.73943	83302.000000	83302.000000

Unsuccessful trials that were not included in the final solution

- We tried to encode **BMI** as a categorical feature instead of continuous but there was no difference in accuracy.
    - BMI is less than 18.5, it falls within the underweight range.
    - BMI is 18.5 to 24.9, it falls within the normal range.
    - BMI is 25.0 to 29.9, it falls within the overweight range.
    - BMI is 30.0 or higher, it falls within the obese range.
- At first, We thought of implementing a **KNN** classifier but it had a problem where we couldn’t train the model and learn its parameters, the map-reduce will need to run for each individual point to get classified, and either way its accuracy was very close to Naive Bayes.
- We also thought of implementing a **Random Forest** classifier but we decided to drop it as doing a mapreduce for it is super complicated.

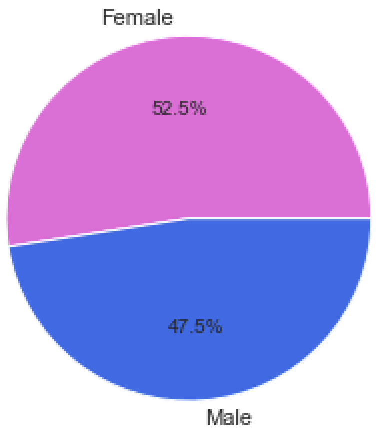
Future Work

- Implement different models, maybe more complex ones such as Random Forest and SVM.
- Research about state-of-art solutions of the unbalanced datasets.

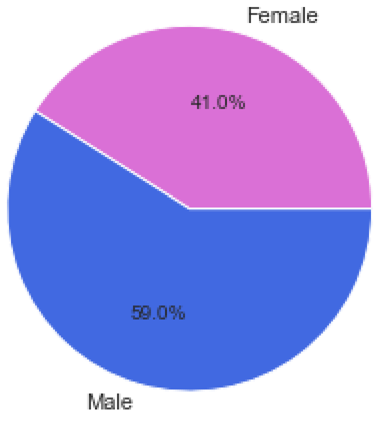
Visualizations

I - Categorical data

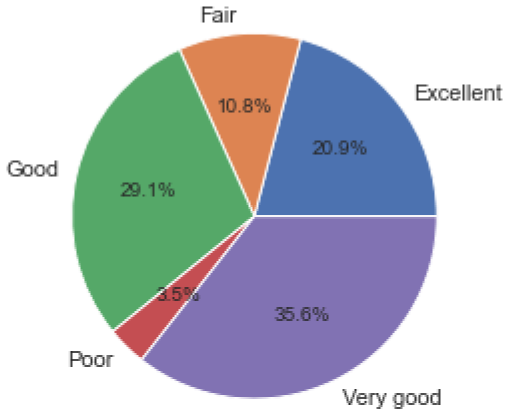
Total Males Vs Total Females



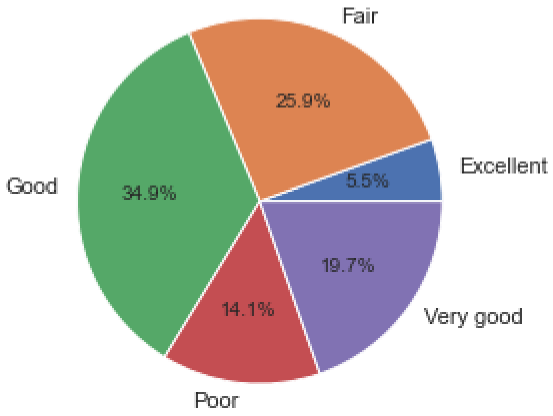
Male with Heart D. Vs Female with Heart D.

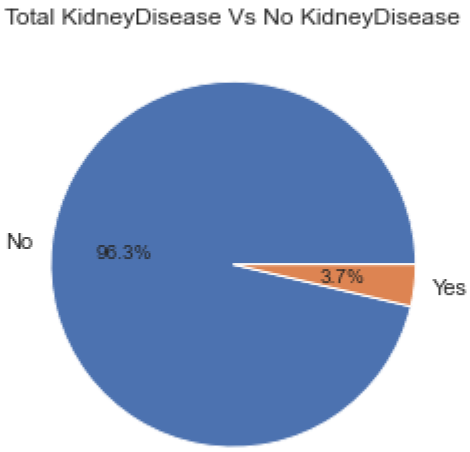


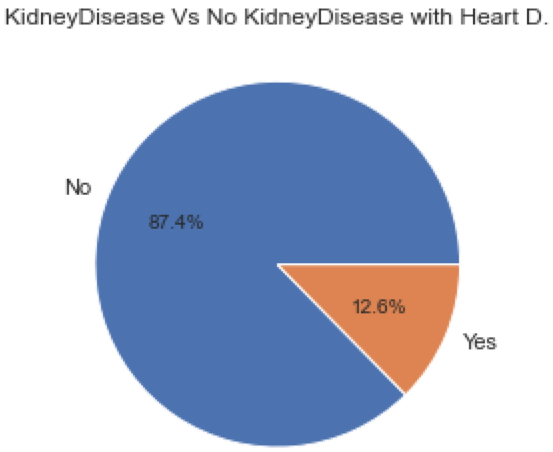
Total General Health count

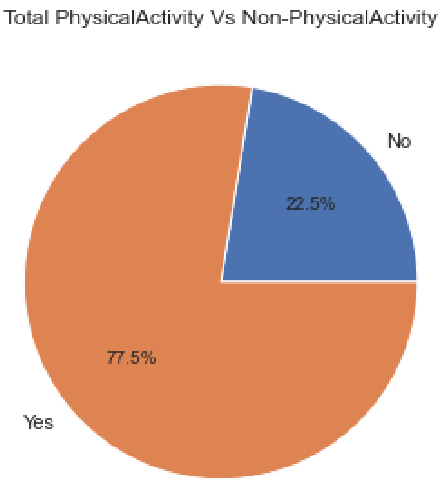


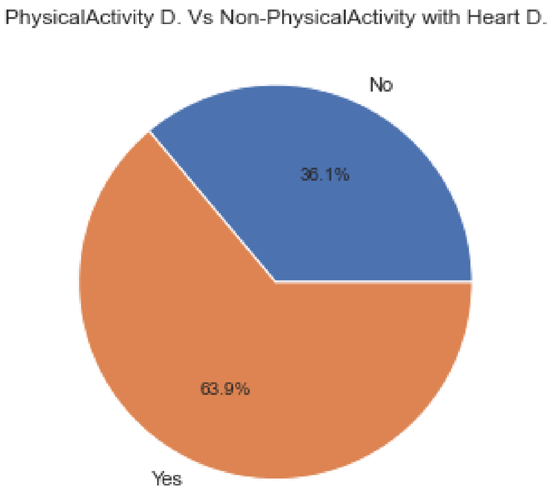
General Health with Heart D. count

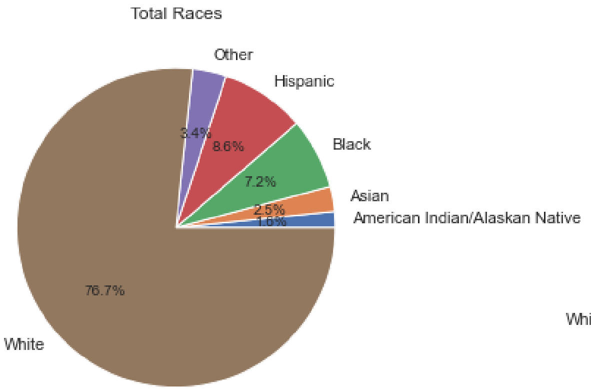


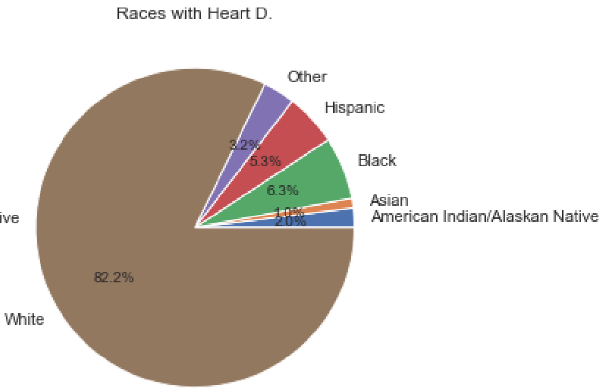

Total KidneyDisease Vs No KidneyDisease


KidneyDisease Vs No KidneyDisease with Heart D.

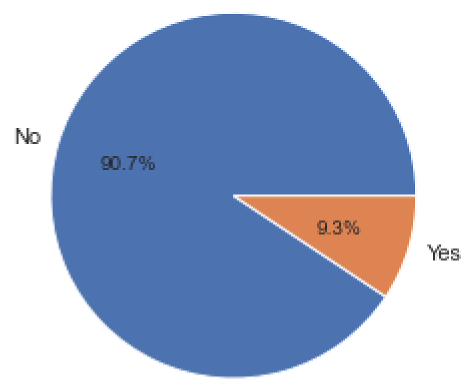

Total PhysicalActivity Vs Non-PhysicalActivity


PhysicalActivity D. Vs Non-PhysicalActivity with Heart D.

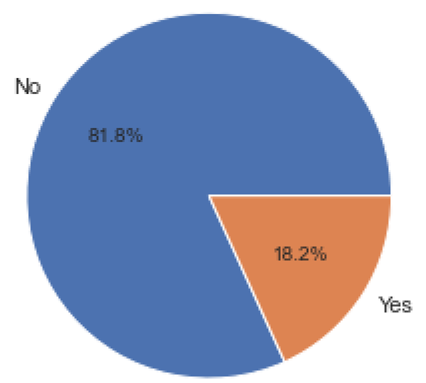

Total Races


Races with Heart D.

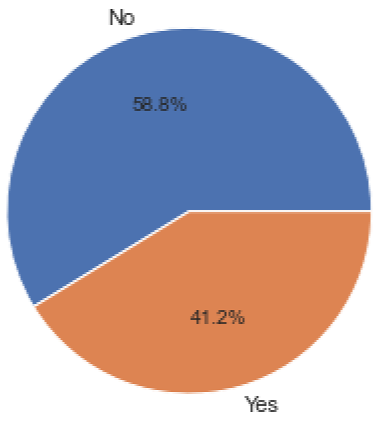
Total SkinCancer Vs No SkinCancer



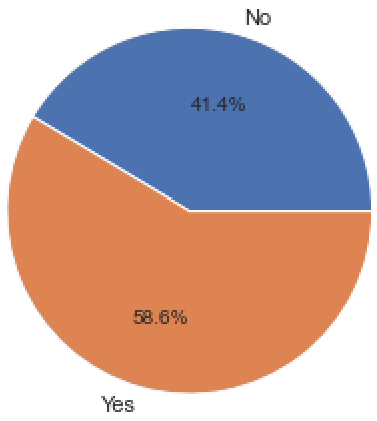
SkinCancer Vs No SkinCancer with Heart D.



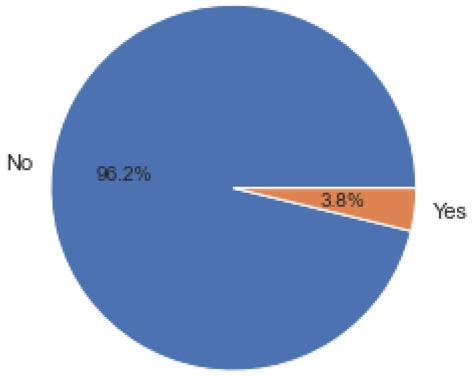
Total Smokers Vs Non-smokers



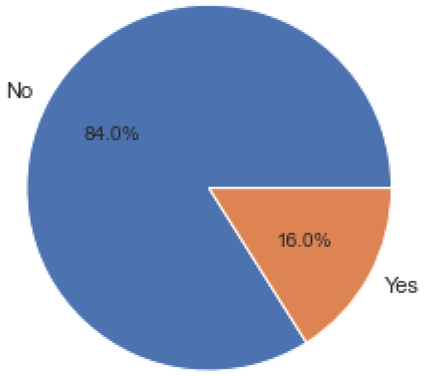
Smokers D. Vs Non-smokers with Heart D.

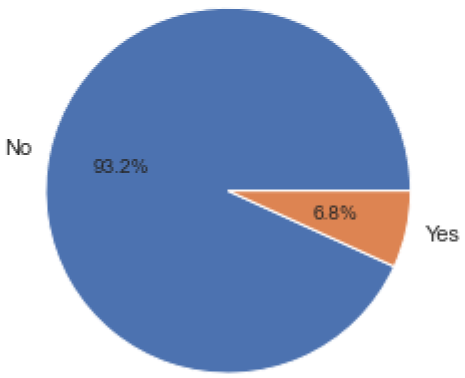


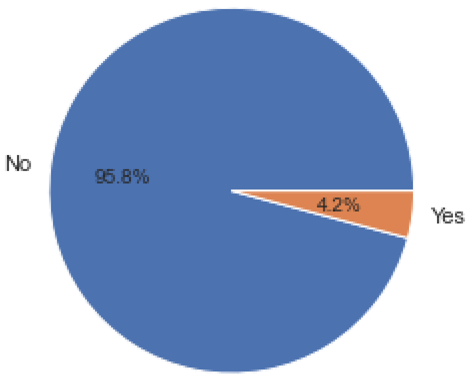
Total Stroke Vs No Stroke

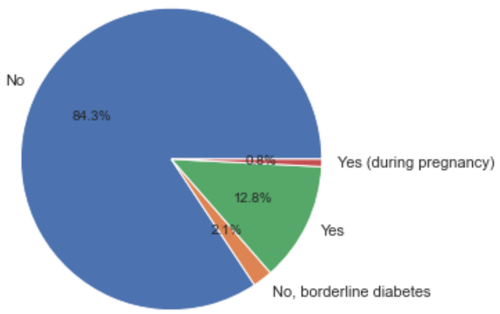


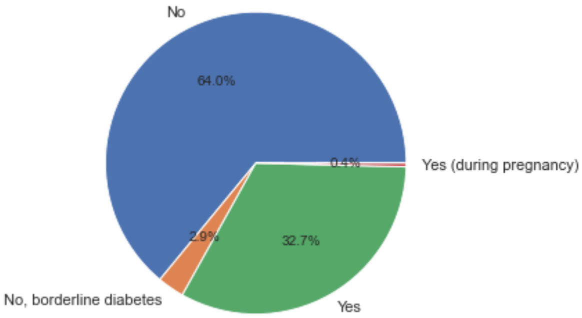
Stroke Vs No Stroke with Heart D. count

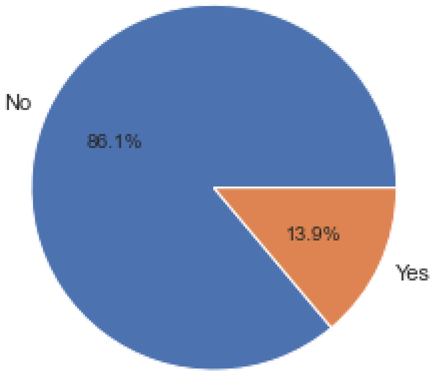


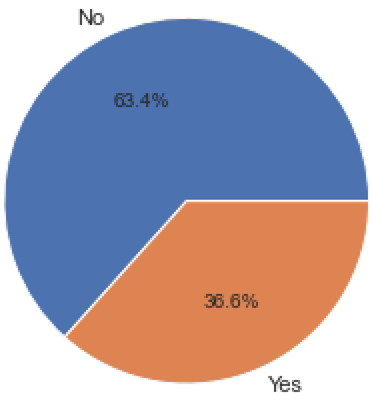

Total Alcoholic Vs Non-alcoholic

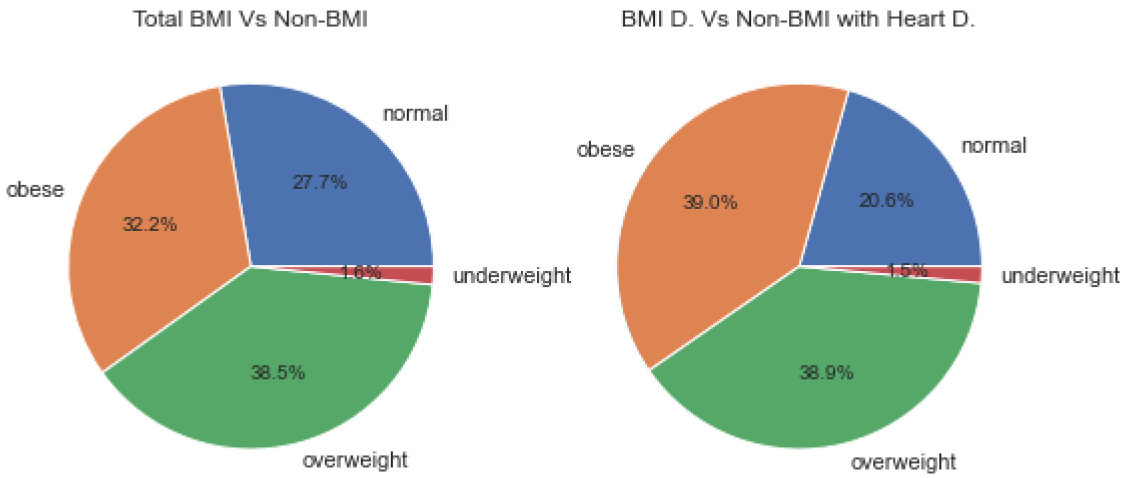

Alcoholic D. Vs Non-alcoholic with Heart D.

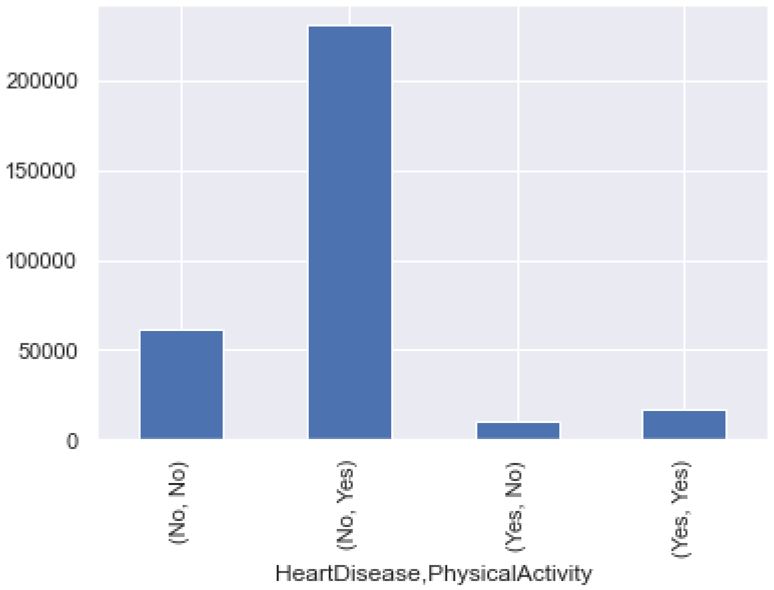

Total Diabetic Vs No Diabetic


Diabetic Vs No Diabetic with Heart D. count

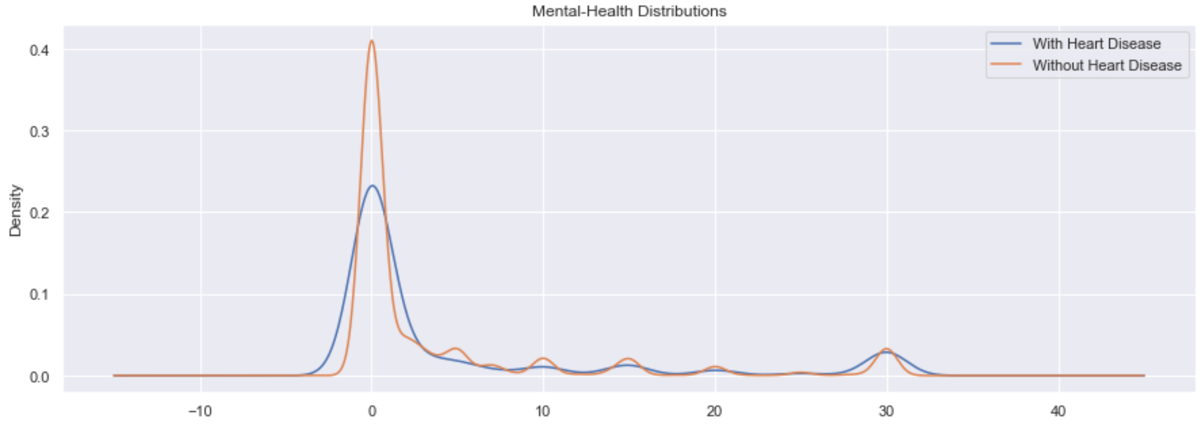

Total DiffWalking Vs No DiffWalking

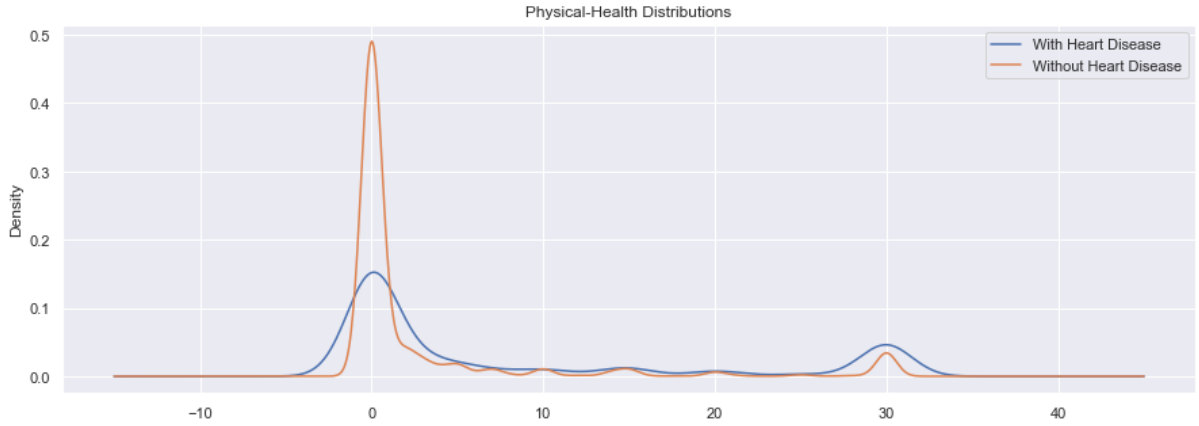

DiffWalking Vs No DiffWalking with Heart D.

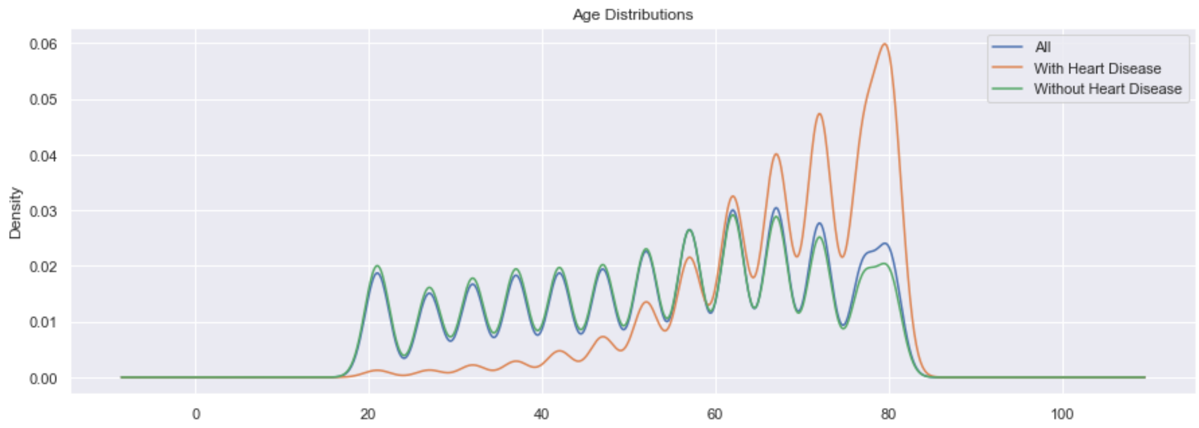

Total BMI Vs Non-BMI
normal
27.7%
obese
32.2%
1.6%
underweight
38.5%
overweight
BMI D. Vs Non-BMI with Heart D.
obese
39.0%
normal
20.6%
1.5%
underweight
38.9%
overweight

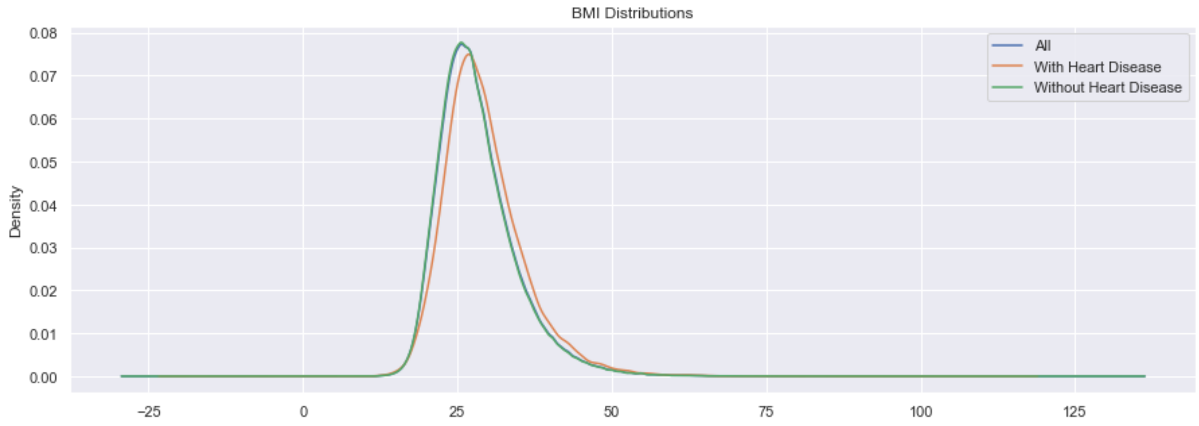

200000
150000
100000
50000
0
(No, No)
(No, Yes)
(Yes, No)
(Yes, Yes)
HeartDisease,PhysicalActivity

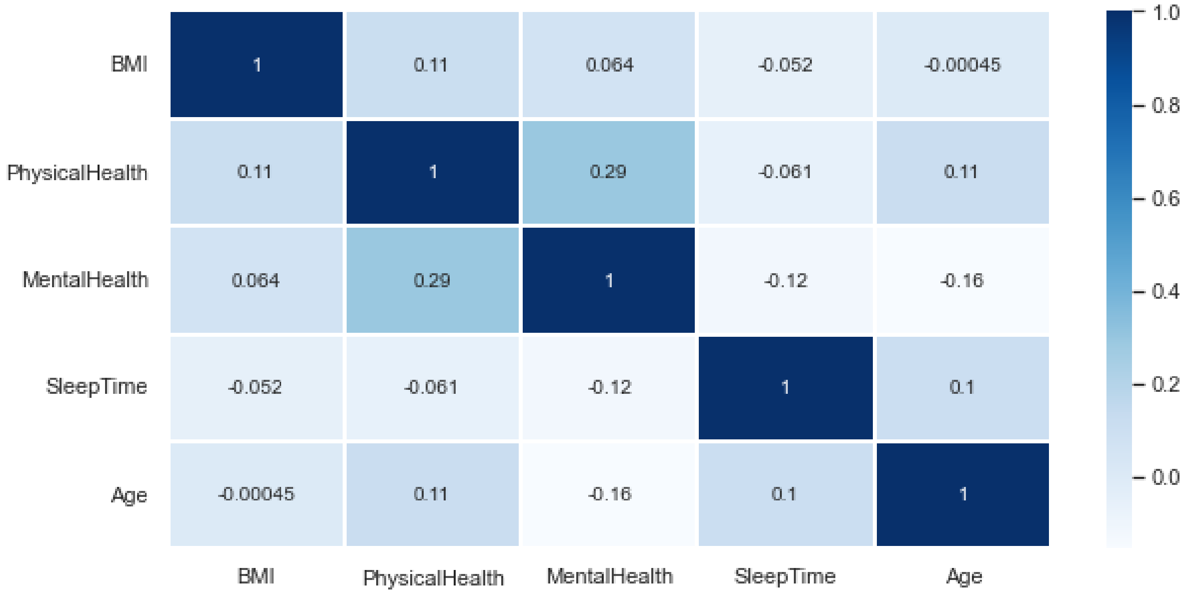
II - Continuous data











Correlation heatmap between the continuous key indicators