Cairo University

Spring 2023

Faculty of Engineering
Cairo University

Big Data Project

Supervised by:

Dr. Lydia Wahid

Eng. Omar Samir Galal

Presented by:

Team 12

Name	Sec	B.N.
Ahmed Emad	1	7
Ahmed Mahmoud Hafez	1	11
Nour El-din Moustafa	2	33
Youssef Atef Abdo	2	41

Idea:

Heart disease is the leading cause of death globally, and its prevalence is increasing rapidly. Early detection and prevention of heart disease are crucial for reducing mortality rates and improving the quality of life for patients.

We suggest a big data project that aims to develop a predictive model for the early detection of heart disease.

Project pipeline

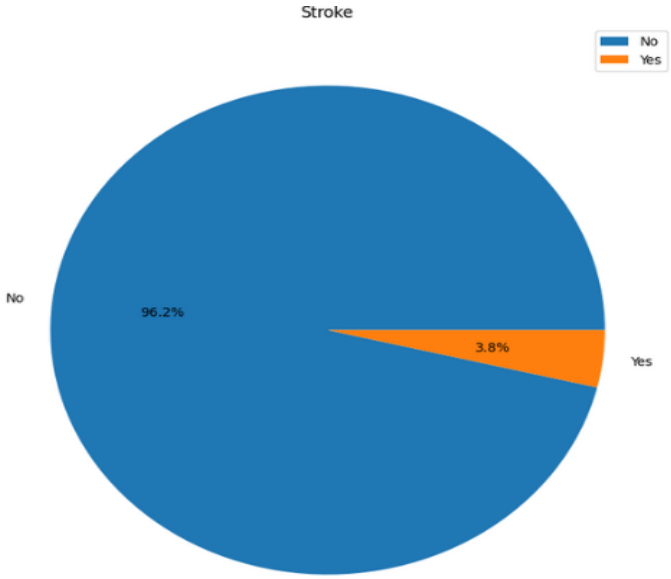
1- Data preprocessing

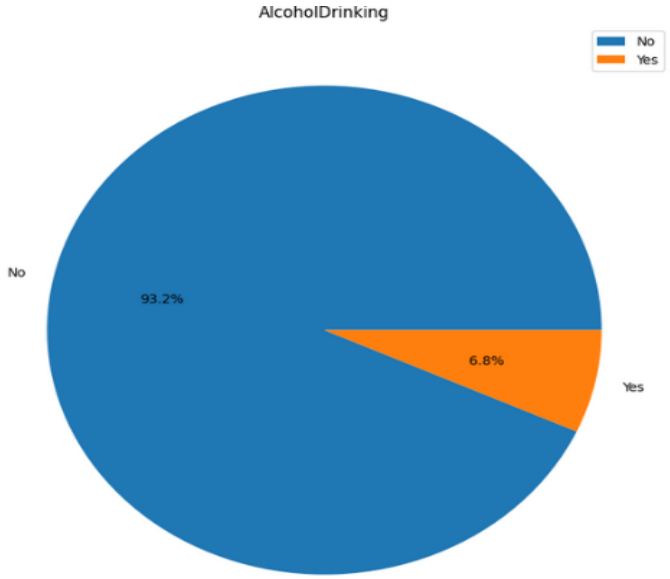
- Every categorical class changed to numerical
  - change columns that contain values of Yes / No with 0 / 1
  - change sex column that contain Male / Female with 1 / 0
  - convert Age Category, Race, Diabetic and GenHealth columns with increasing values

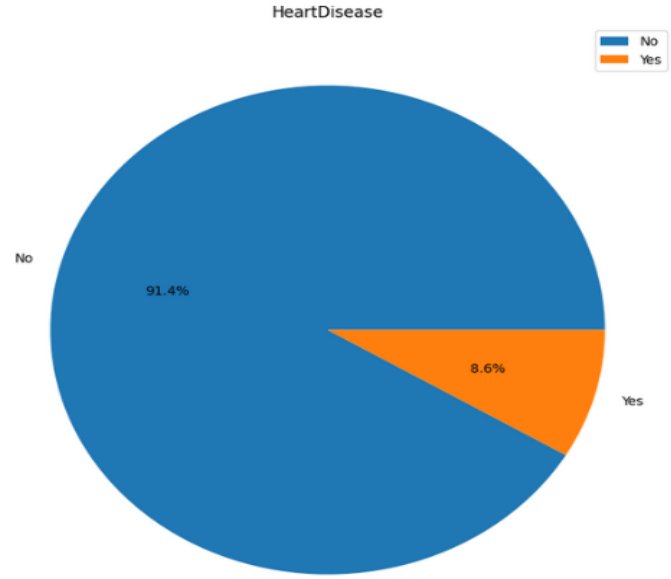
- Check If non-values exist and deal with it

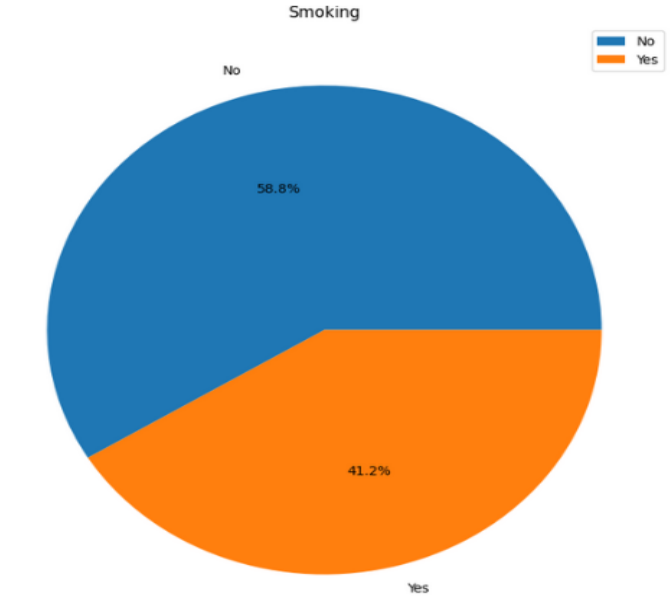
2- Data visualization and insights

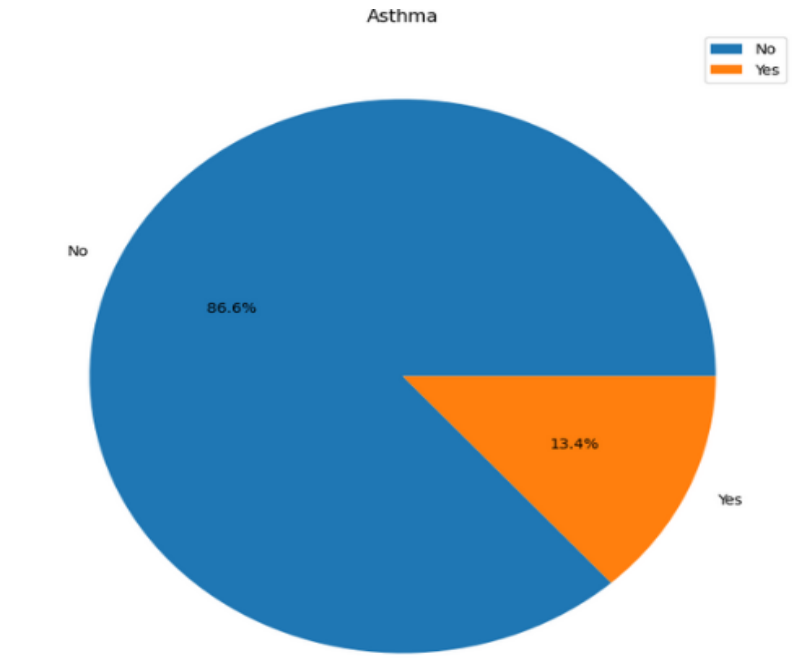
Show distribution for each categorical class

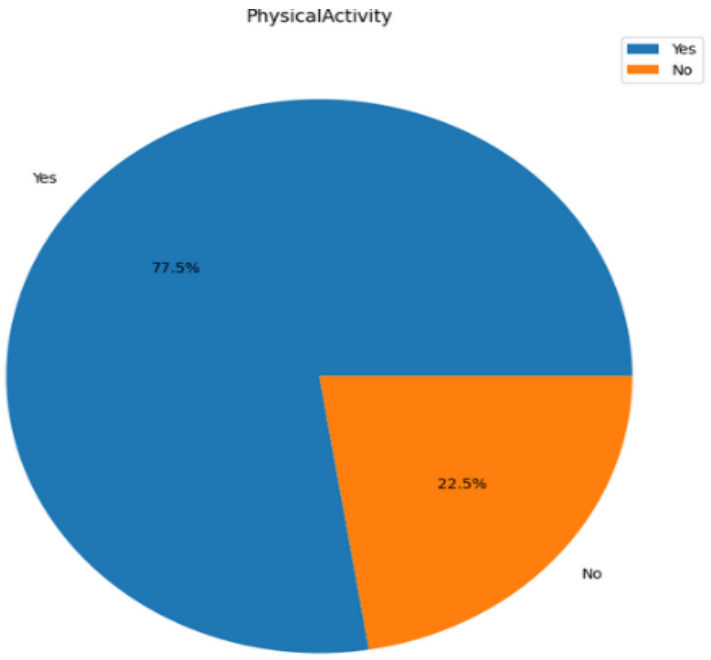

Stroke
No
Yes
No
96.2%
3.8%
Yes

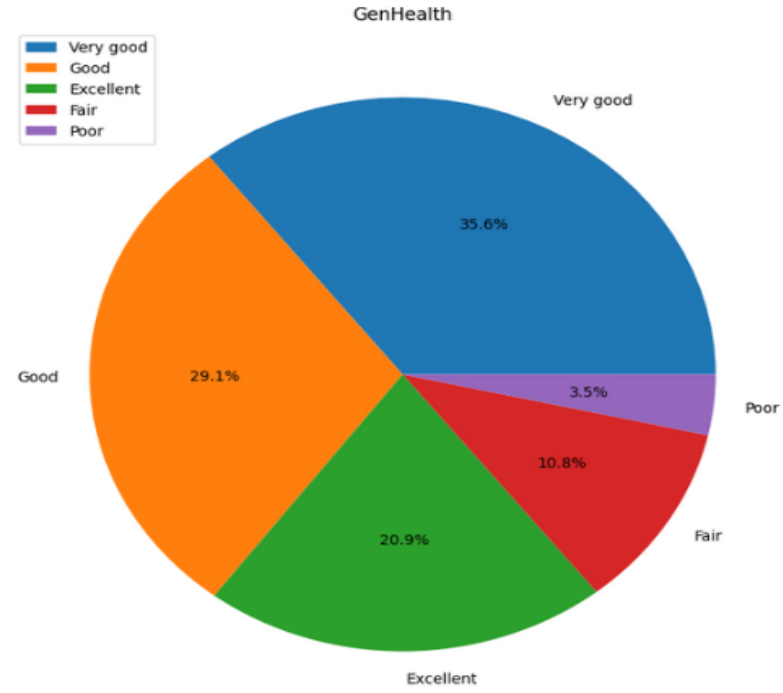

AlcoholDrinking
No
Yes
No
93.2%
6.8%
Yes

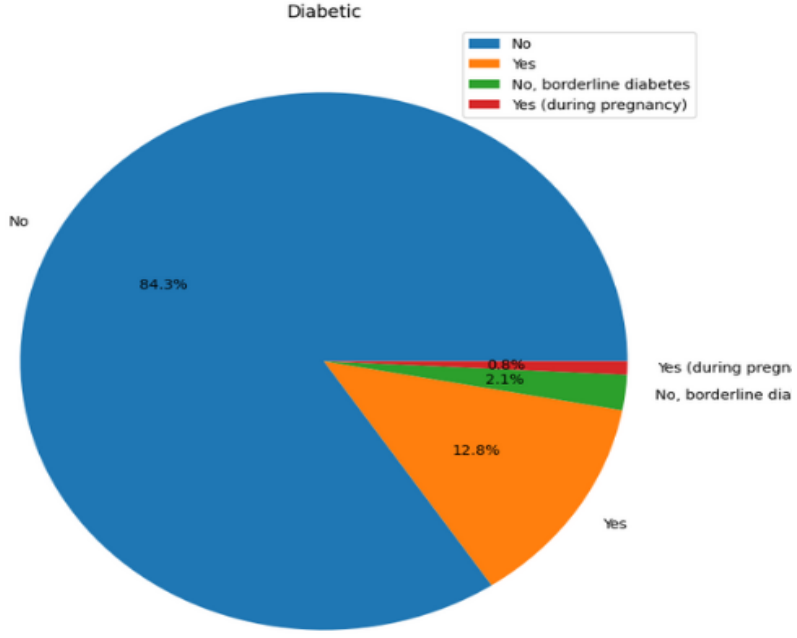

HeartDisease
No
Yes
No
91.4%
8.6%
Yes

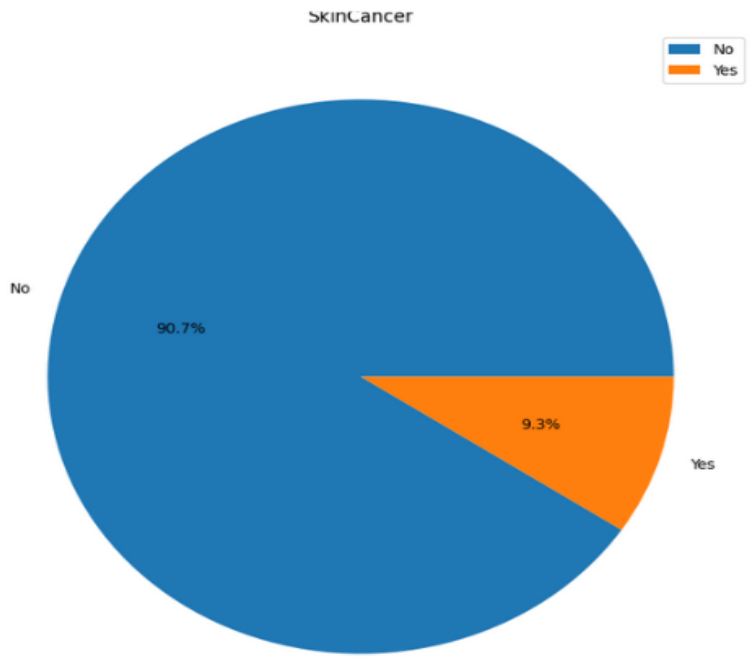

Smoking
No
Yes
No
58.8%
41.2%
Yes

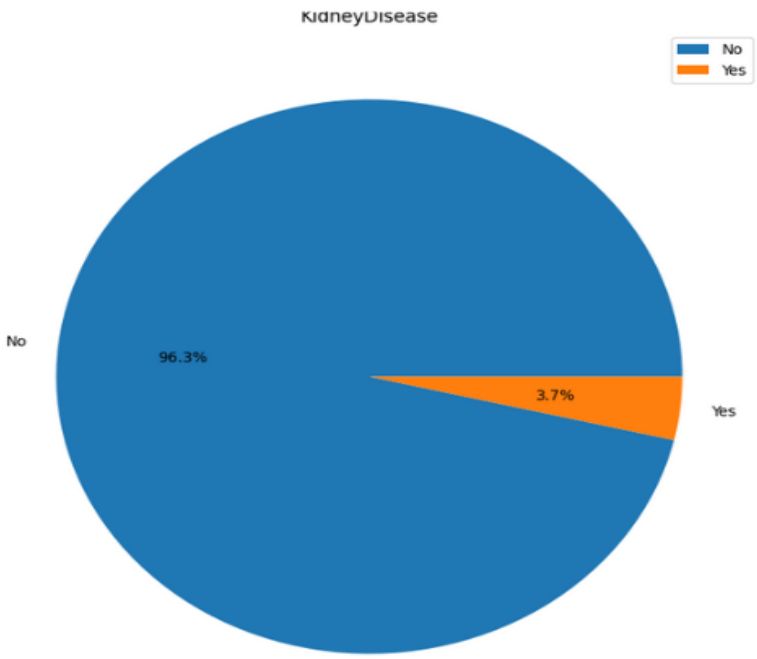

Asthma
No
Yes
No
86.6%
13.4%
Yes

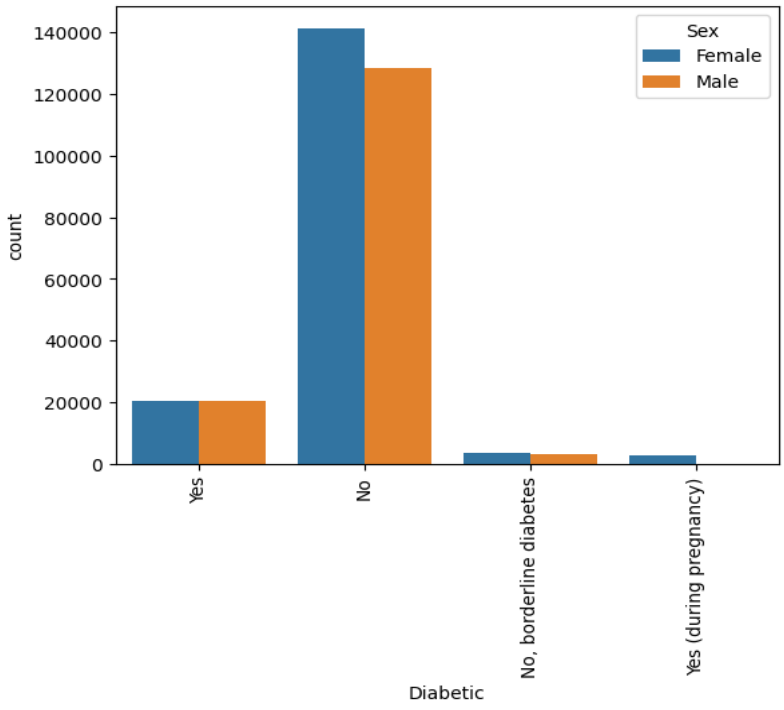

PhysicalActivity
Yes
No
Yes
77.5%
22.5%
No


GenHealth
Very good
Good
Excellent
Fair
Poor
Very good
35.6%
Good
29.1%
3.5%
Poor
10.8%
Fair
20.9%
Excellent

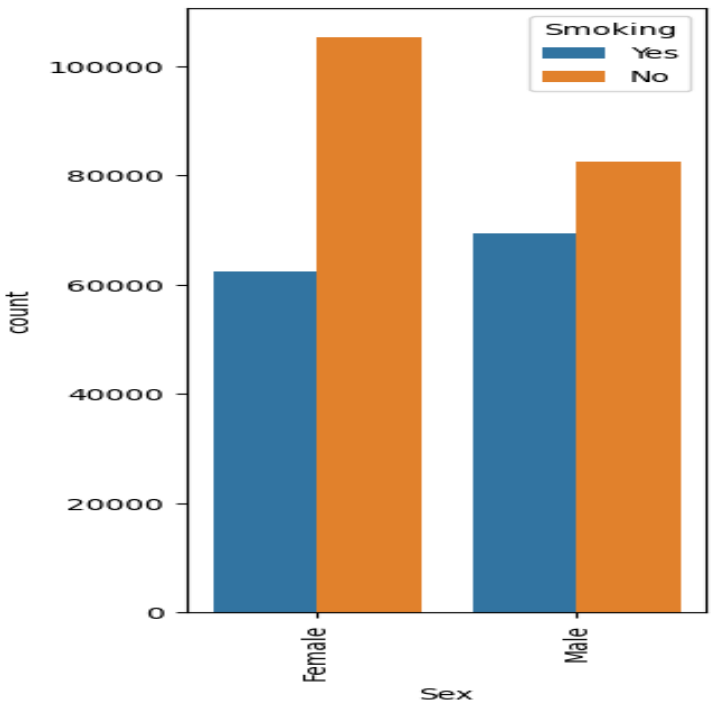

Diabetic
No
Yes
No, borderline diabetes
Yes (during pregnancy)
No
84.3%
0.8%
Yes (during pregn
2.1%
No, borderline dia
12.8%
Yes


SkinCancer
No
Yes
No
90.7%
9.3%
Yes

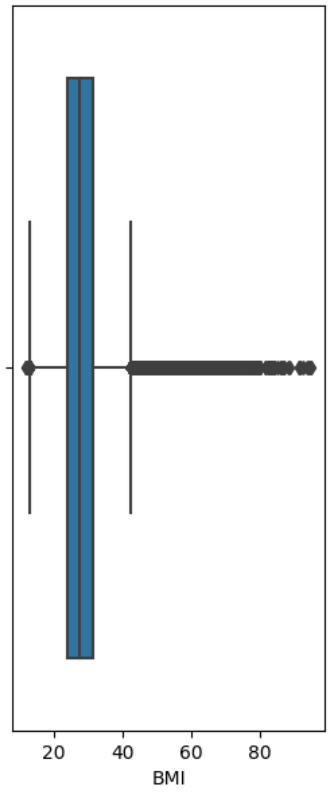
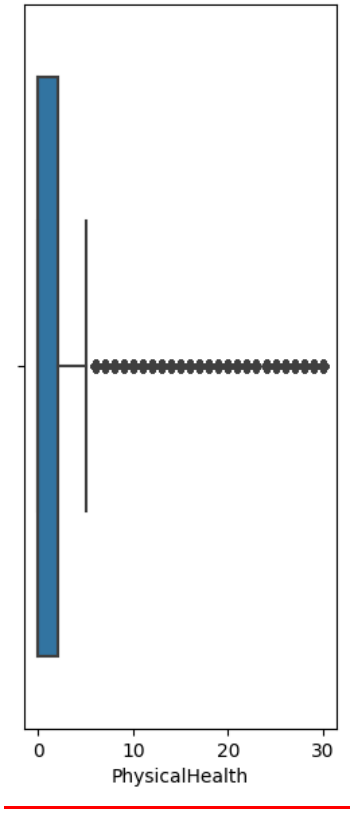
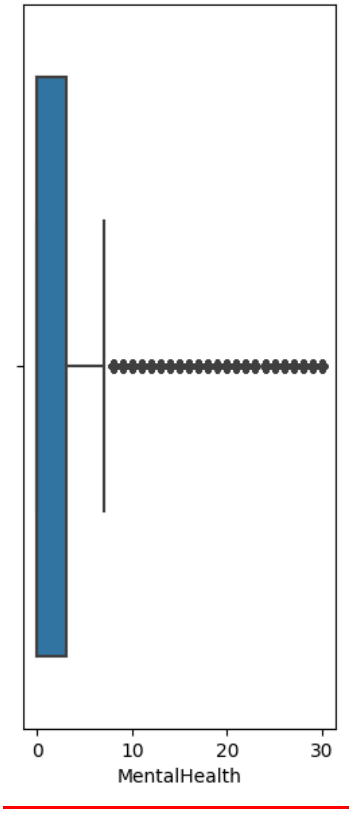
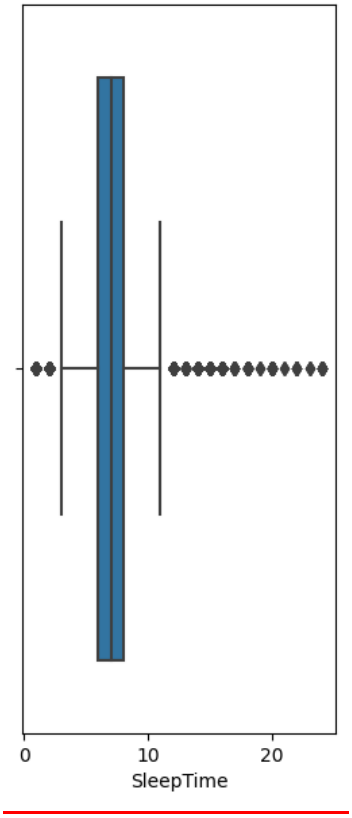

KidneyDisease
No
Yes
No
96.3%
3.7%
Yes

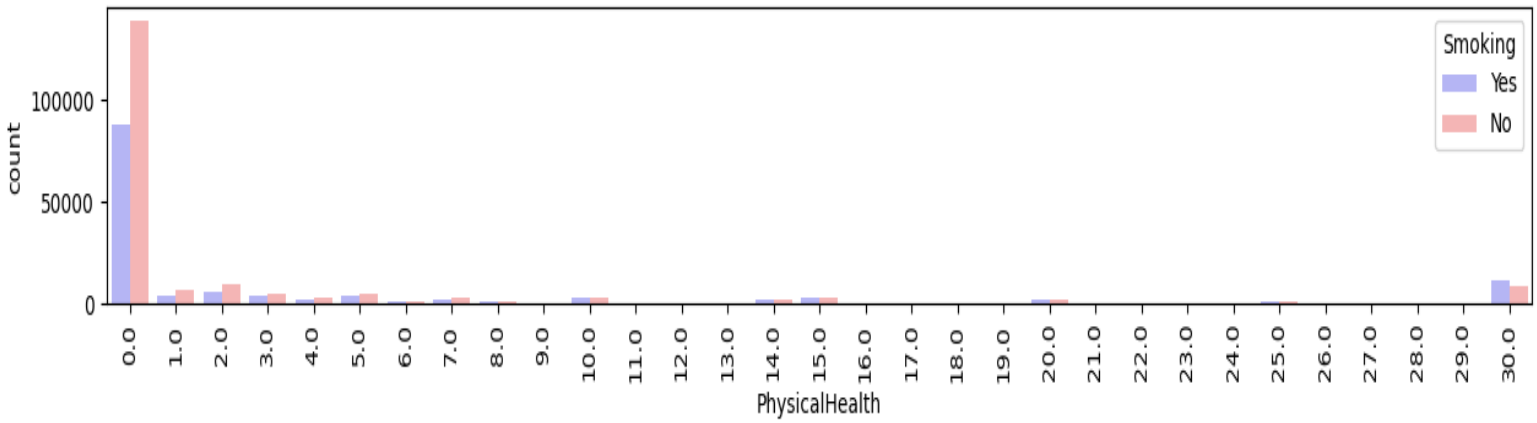


this graph shows male have slightly more diabetes than female

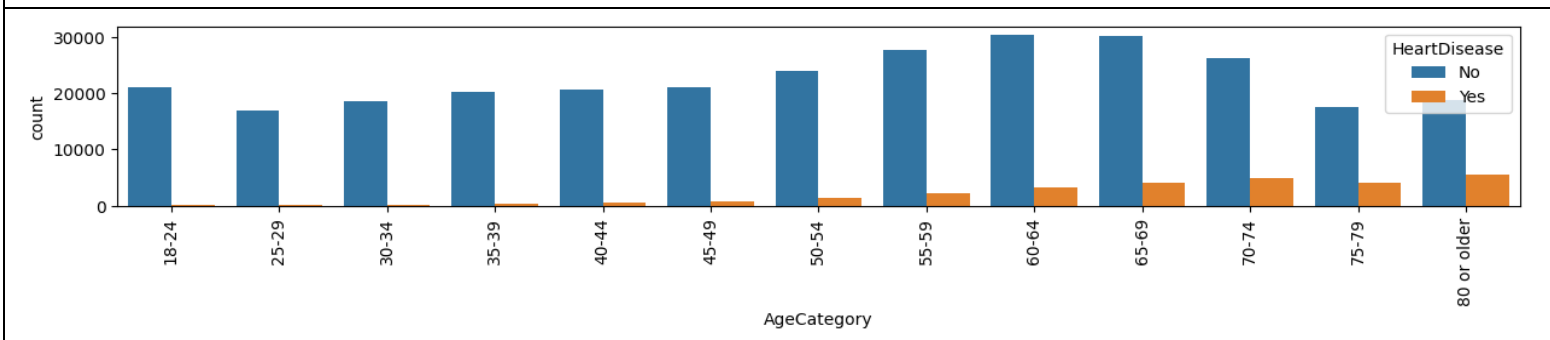


this graph shows female are smoking less than man

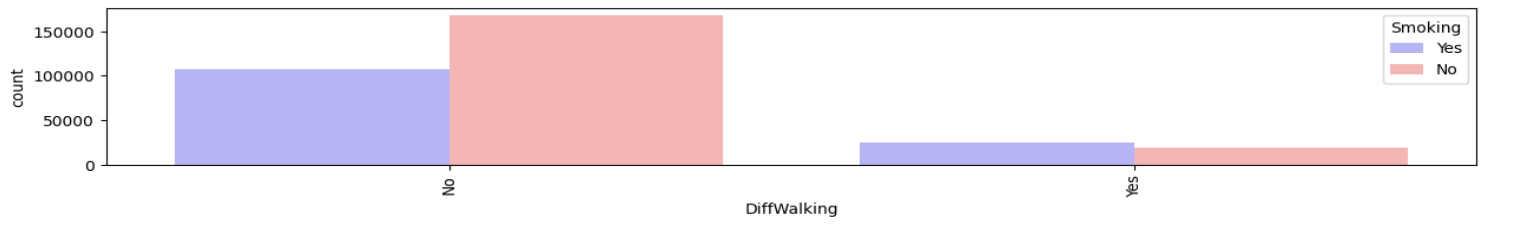
			
BMI average is between 20 and 40 and there are many outliers	Distribution of physical health	Distribution of Mental health	between 5 and 8 hours and there are outliers



1 - no smoking → good physical health

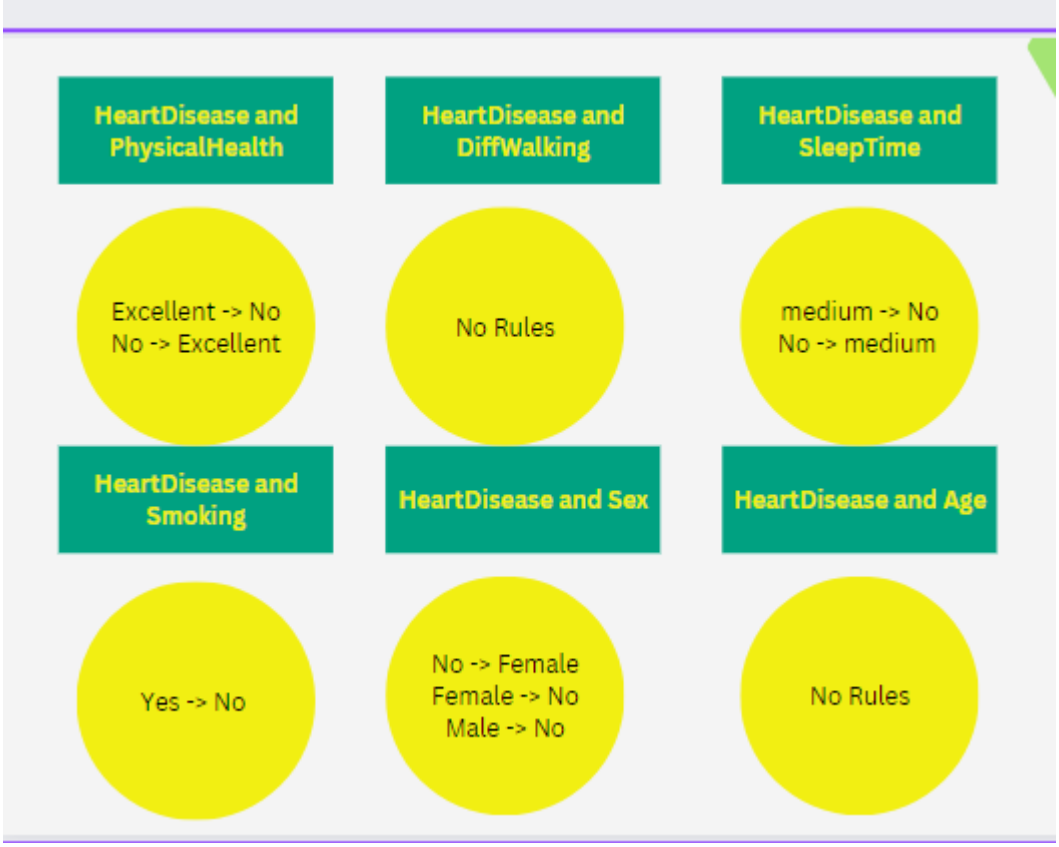


2- the bigger the age → more heart disease to have



3- smoking → diff walking

Association Rules



- ***Excellent health leads to low heart disease***
- ***Sleeping fair hours (5-8) of sleeping saves you from heart diseases***
- ***Gender has no effect of increasing probability of Heart diseases***

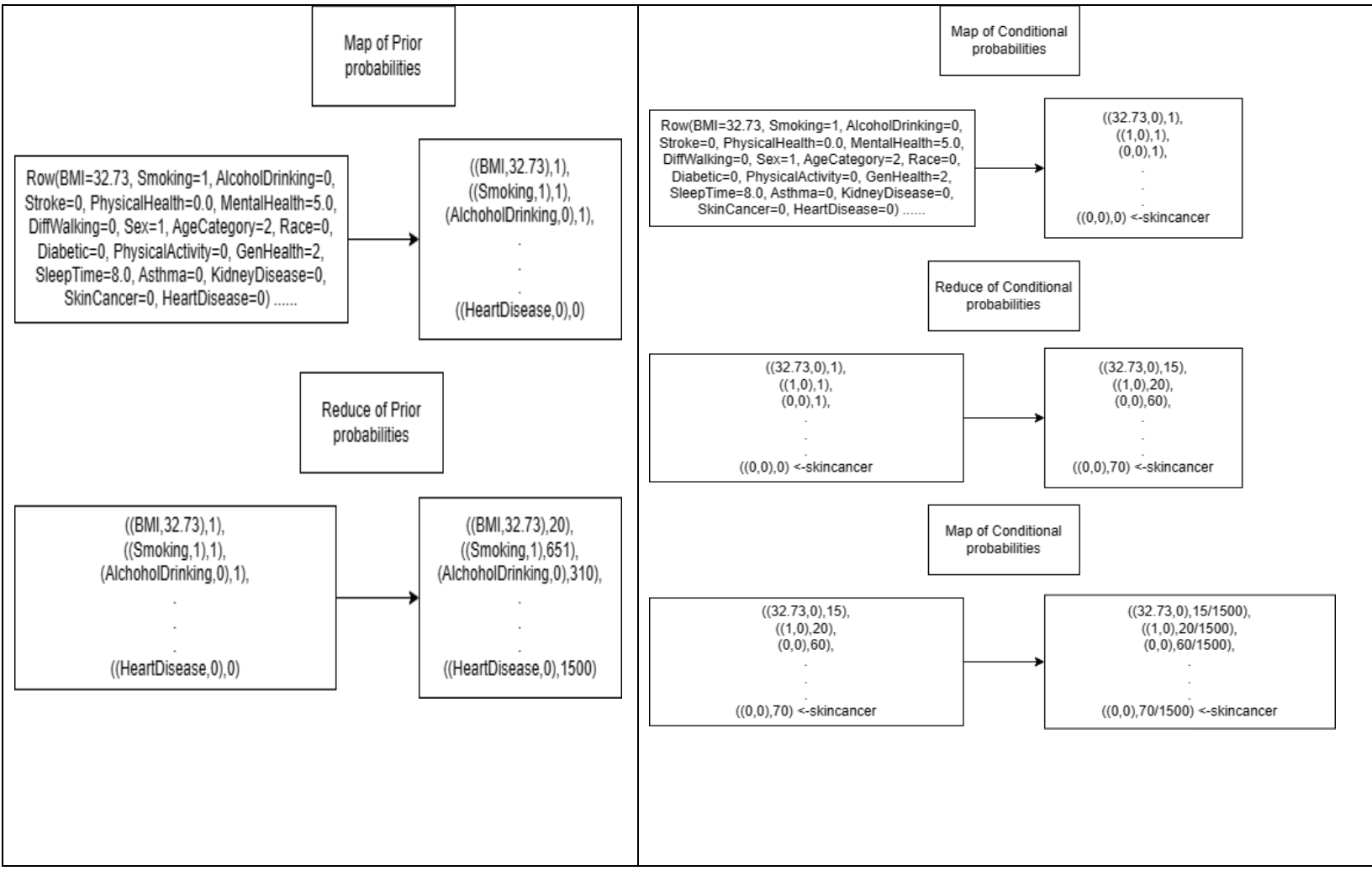
Model/Classifier training

- For all models we used: we did RandomOversampler to balance the classes.

1- We used Naive Bayes , SVM and logistic regression from MLlib in PySpark. Those are ready made models.

2- Then we used map-reduce functions to implement Naive Bayes:

- We calculated the prior probabilities of the features and the classes we have.
    - The map phase was used to generate key-value pair <feature, 1>
    - The reduce phase was used to aggregate the number of each attribute value.
- We calculated the conditional probabilities of each feature given each class.
    - The map phase was used to generate key-value pair <(feature, class), 1>
    - The reduce phase was used to aggregate.
    - Then another map phase was done to calculate the conditional probabilities by dividing the totalCount of the (featureValue, class) by the totalCount of the class.



3-Results and Evaluation

Model	Accuracy (F1-score)	Macro Avg
Logistic Regression	76%	76%
Naive Bayes	65%	62%
SVM	76%	76%
Naive Bayes	75%	75%

Unsuccessful trials

Just a note that SVM takes a very long time

Any Enhancements and future work

We want to implement KNN using map-reduce.