Cairo University

Spring 2023

Faculty of Engineering
Cairo University

Big Data Project

Supervised by:

Dr. Lydia Wahid

Eng. Omar Samir Galal

Presented by:

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

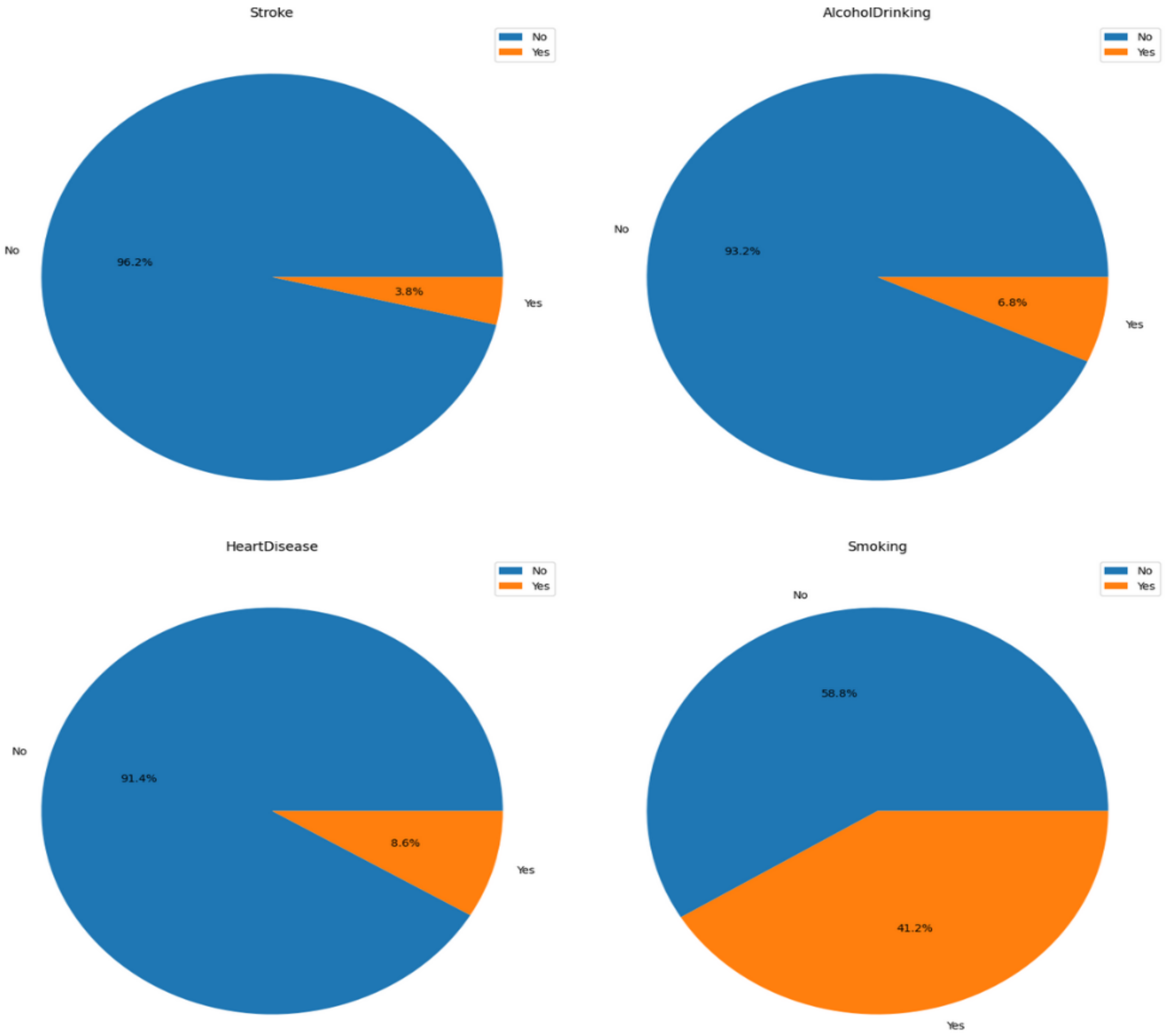
1- Data preprocessing

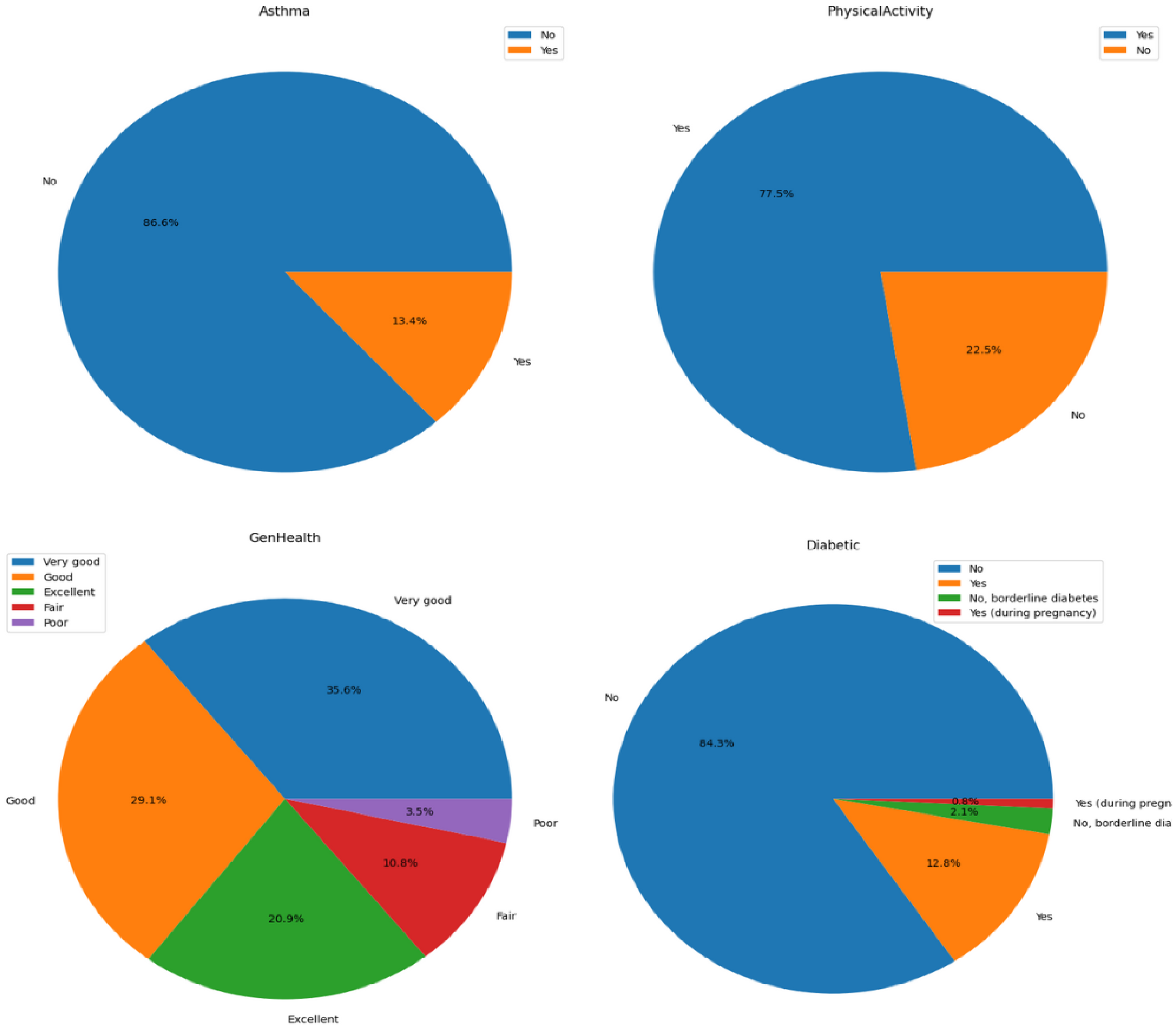
- Every categorical class changed to numerical
    - change columns that contain values of Yes / No with 0 / 1
    - change sex column that contain Male / Female with 1 / 0
    - convert Age Category, Race, Diabetic and GenHealth columns with increasing values

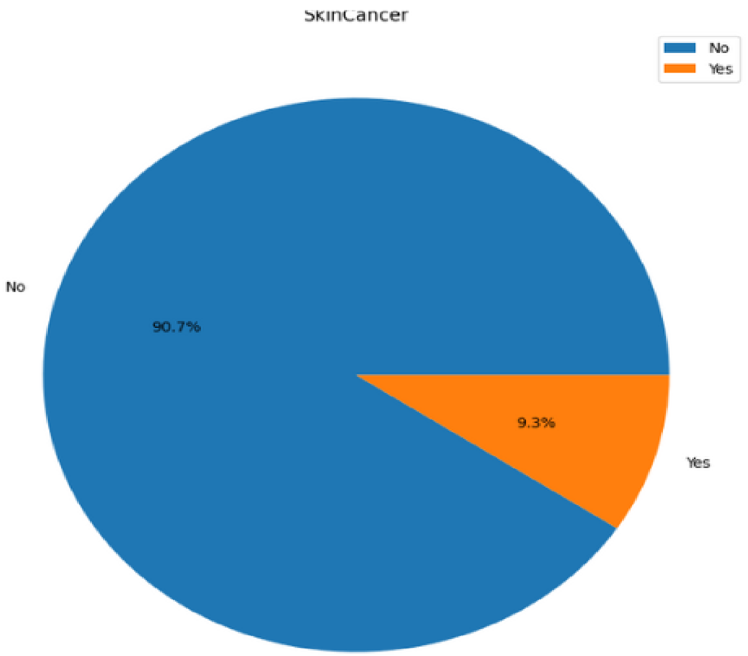
- Check If non-values exist and deal with it

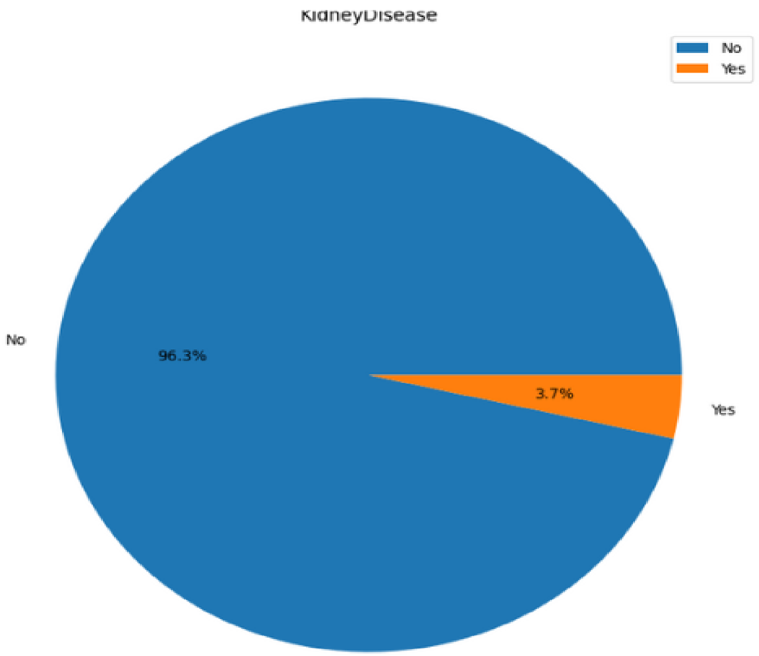
2- Data visualization

Show distribution for each categorical class

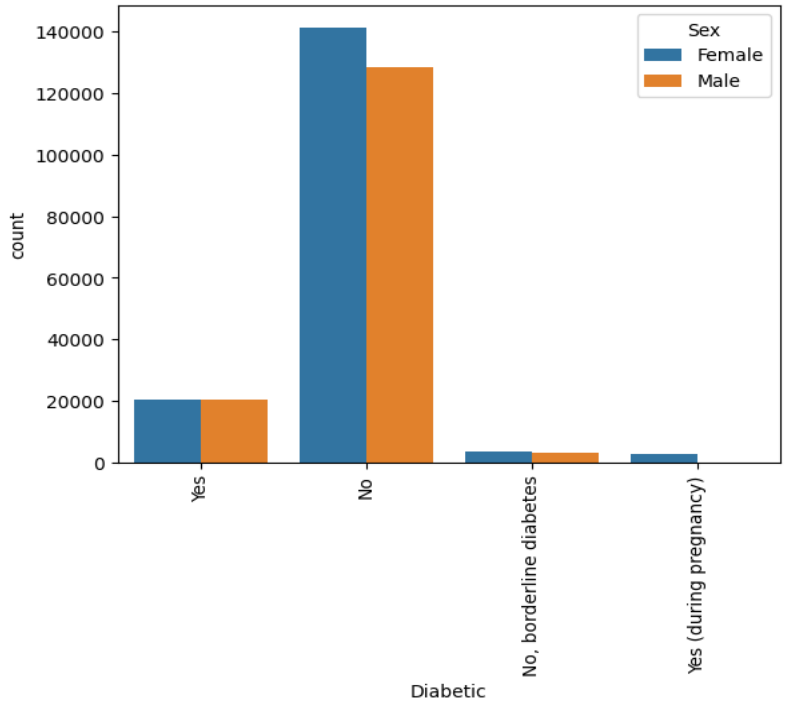
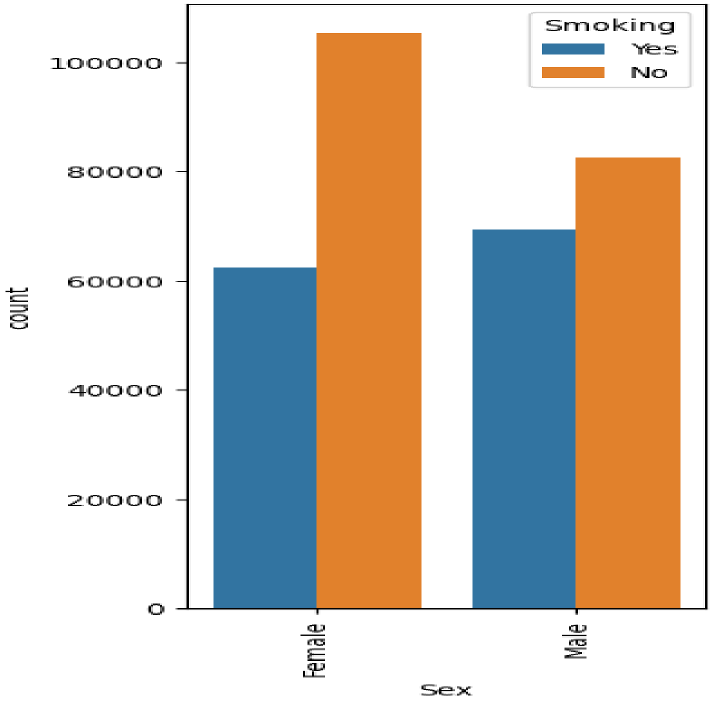

Stroke
No
Yes
No
96.2%
3.8%
Yes
AlcoholDrinking
No
Yes
No
93.2%
6.8%
Yes
HeartDisease
No
Yes
No
91.4%
8.6%
Yes
Smoking
No
Yes
No
58.8%
41.2%
Yes

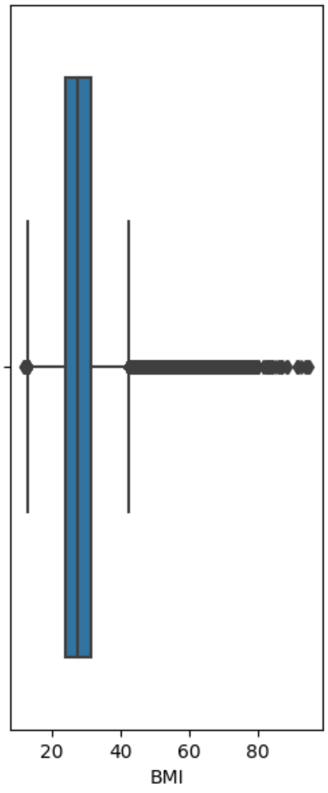
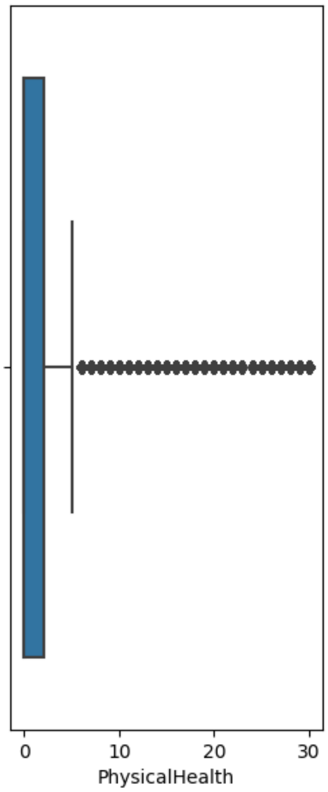
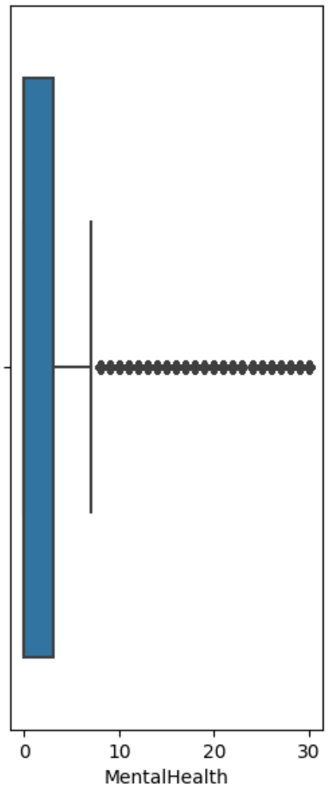
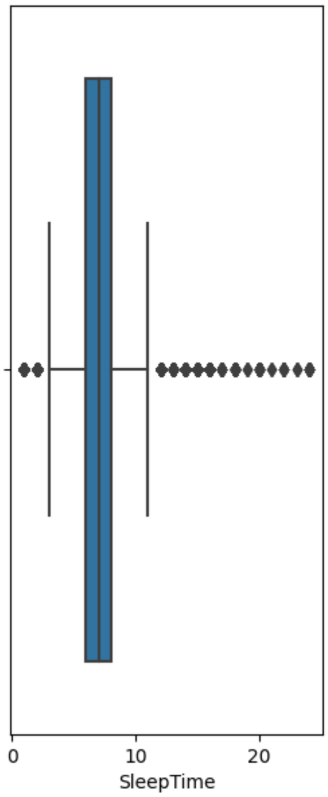

Asthma
No
Yes
No
86.6%
13.4%
Yes
PhysicalActivity
Yes
No
Yes
77.5%
22.5%
No
GenHealth
Very good
Good
Excellent
Fair
Poor
Very good
35.6%
Good
29.1%
3.5%
Poor
10.8%
Fair
20.9%
Excellent
Diabetic
No
Yes
No, borderline diabetes
Yes (during pregnancy)
No
84.3%
0.8%
Yes (during pregn
2.1%
No, borderline dia
12.8%
Yes



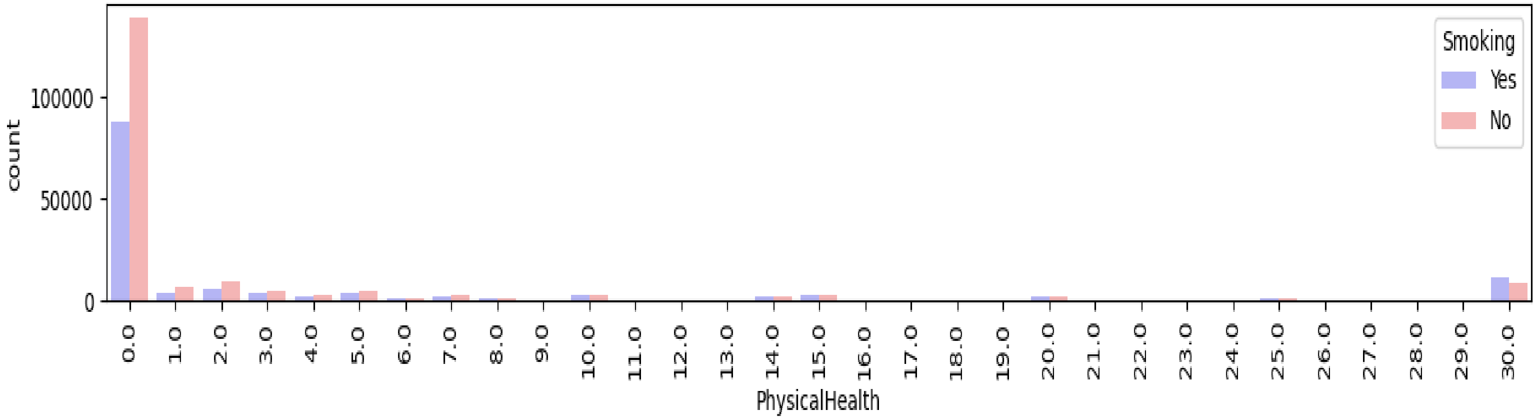


3- Extracting insights from data.

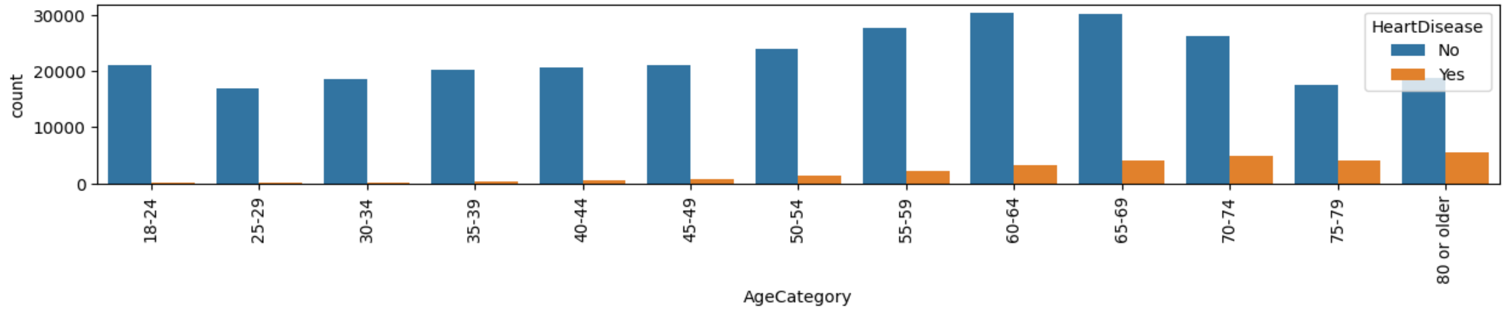
	
this graph shows male have slightly more diabetes than female	this graph shows female are smoking less than man

			
bmi average is between 20 and 40 and there are many outliers			between 5 and 8 hours and there are outliers

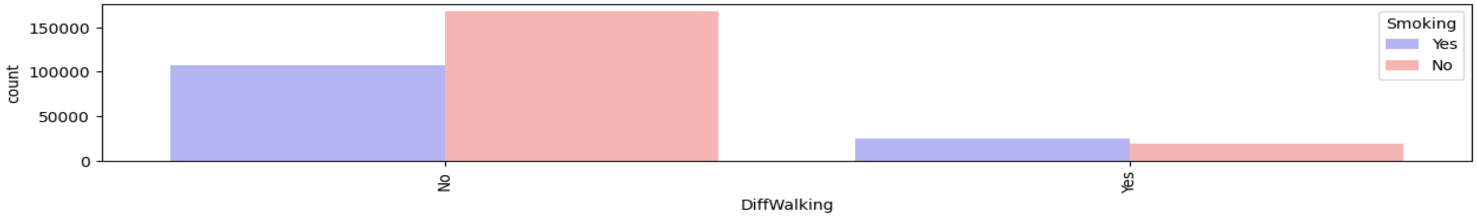
4-Association Rules



1 - no smoking → good physical health



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



3- smoking → diff walking

5- Model/Classifier training

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:

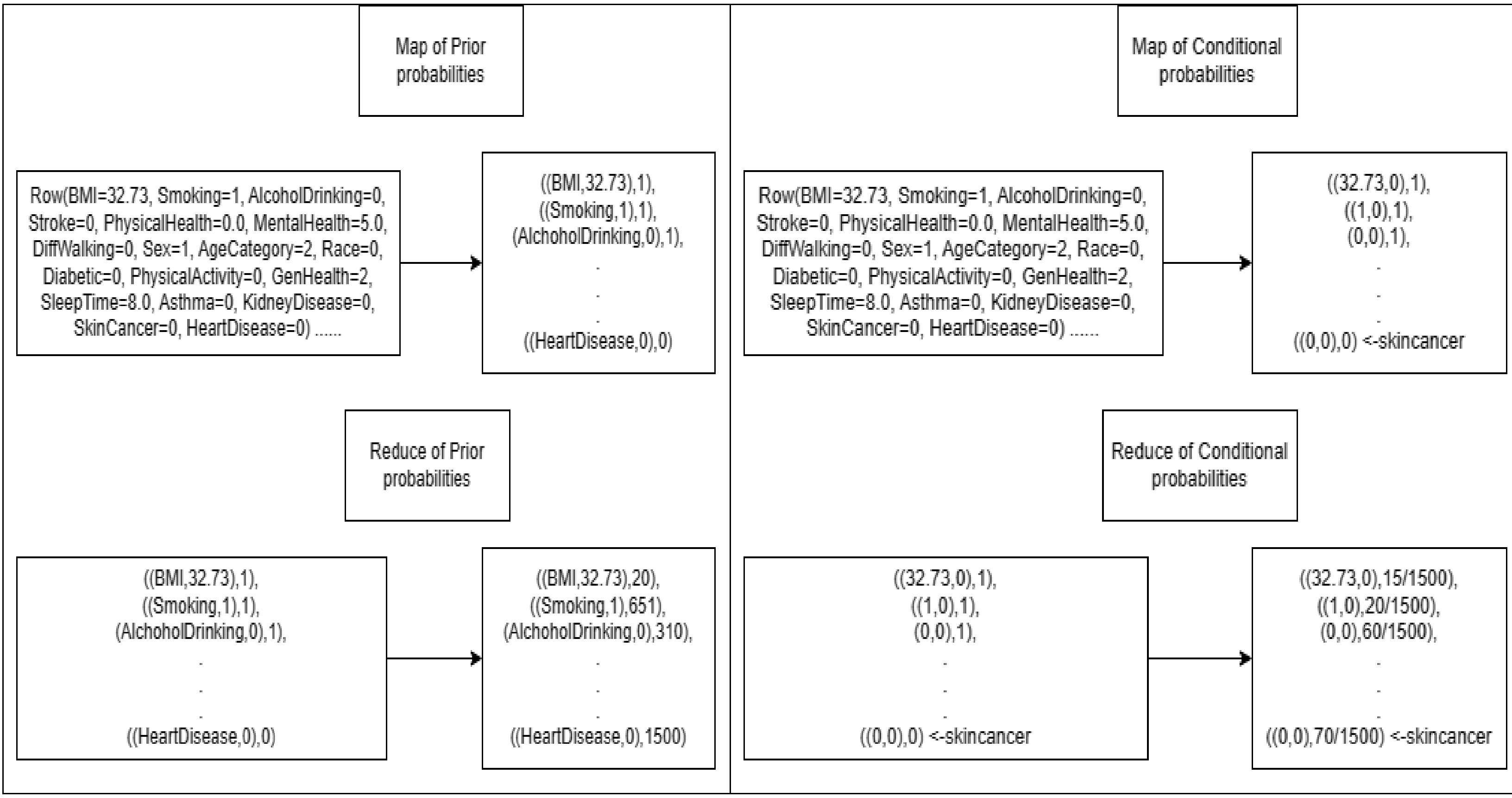
- We calculated the prior probabilities of the features and the classes we have.
- The map phase was used to generate key-value pair <feature, 1>

- The reduce phase was used to aggregate the number of each attribute value.

3- 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.


Map of Prior probabilities
Row(BMI=32.73, Smoking=1, AlcoholDrinking=0, Stroke=0, PhysicalHealth=0.0, MentalHealth=5.0, DiffWalking=0, Sex=1, AgeCategory=2, Race=0, Diabetic=0, PhysicalActivity=0, GenHealth=2, SleepTime=8.0, Asthma=0, KidneyDisease=0, SkinCancer=0, HeartDisease=0) ......
((BMI,32.73),1), ((Smoking,1),1), (AlchoholDrinking,0),1), . . . ((HeartDisease,0),0)
Reduce of Prior probabilities
((BMI,32.73),1), ((Smoking,1),1), (AlchoholDrinking,0),1), . . . ((HeartDisease,0),0)
((BMI,32.73),20), ((Smoking,1),651), (AlchoholDrinking,0),310), . . . ((HeartDisease,0),1500)
Map of Conditional probabilities
Row(BMI=32.73, Smoking=1, AlcoholDrinking=0, Stroke=0, PhysicalHealth=0.0, MentalHealth=5.0, DiffWalking=0, Sex=1, AgeCategory=2, Race=0, Diabetic=0, PhysicalActivity=0, GenHealth=2, SleepTime=8.0, Asthma=0, KidneyDisease=0, SkinCancer=0, HeartDisease=0) ......
((32.73,0),1), ((1,0),1), (0,0),1), . . . ((0,0),0) <-skincancer
Reduce of Conditional probabilities
((32.73,0),1), ((1,0),1), (0,0),1), . . . ((0,0),0) <-skincancer
((32.73,0),15/1500), ((1,0),20/1500), (0,0),60/1500), . . . ((0,0),70/1500) <-skincancer

6-Results and Evaluation

Model	Accuracy (F1-score)	
Naive Bayes	75%	Map-reduce
Naive Bayes	76%	from MLlib in PySpark
logistic regression	62%	
SVM	76%	

Unsuccessful trials that were not included in the final solution.

7-Any Enhancements and future work

- We want to implement KNN using map-reduce.