–––––؛Graphical user interface, application

Description automatically generatedLogo, company name

Description automatically generated

**Big Data Project**

Supervised by: **Dr. Lydia Wahid**

**Eng. Omar Samir Galal**

Presented by:

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

**Spring 2023**

**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.**

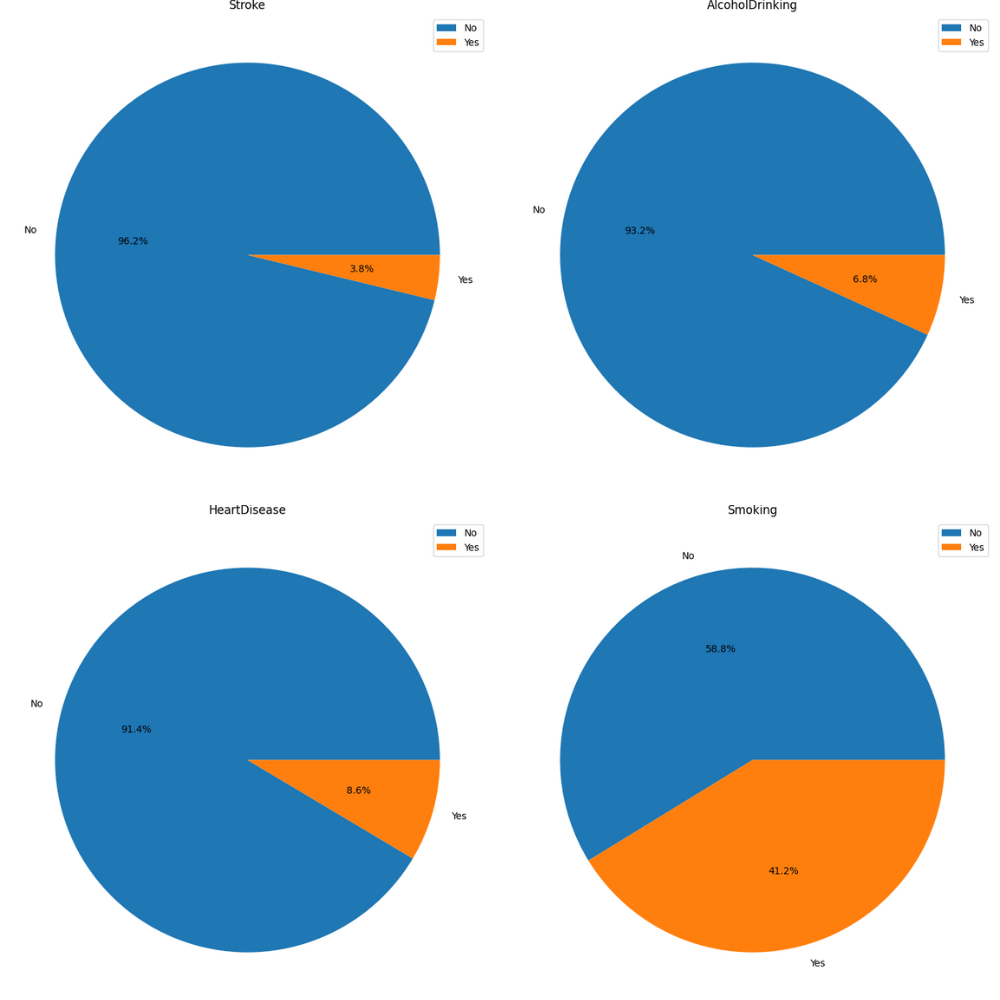
**Project pipeline**

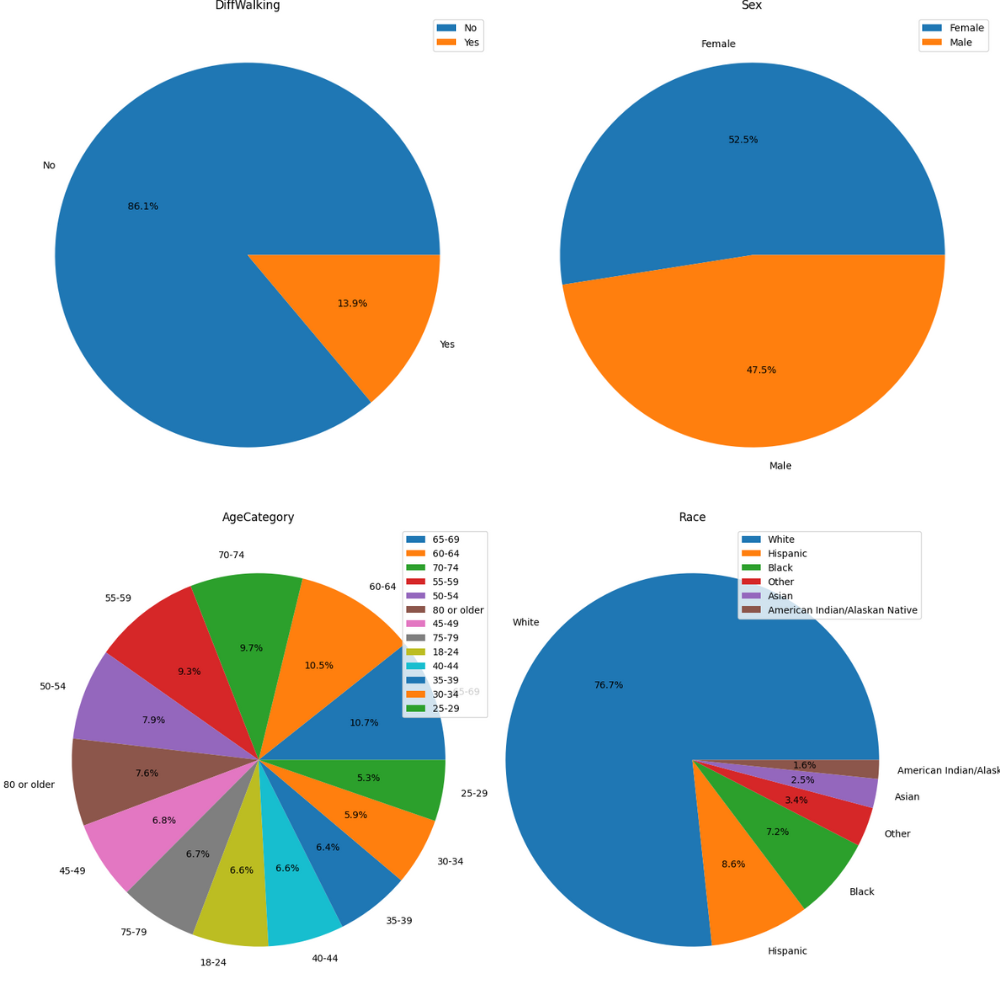
Data preprocessing

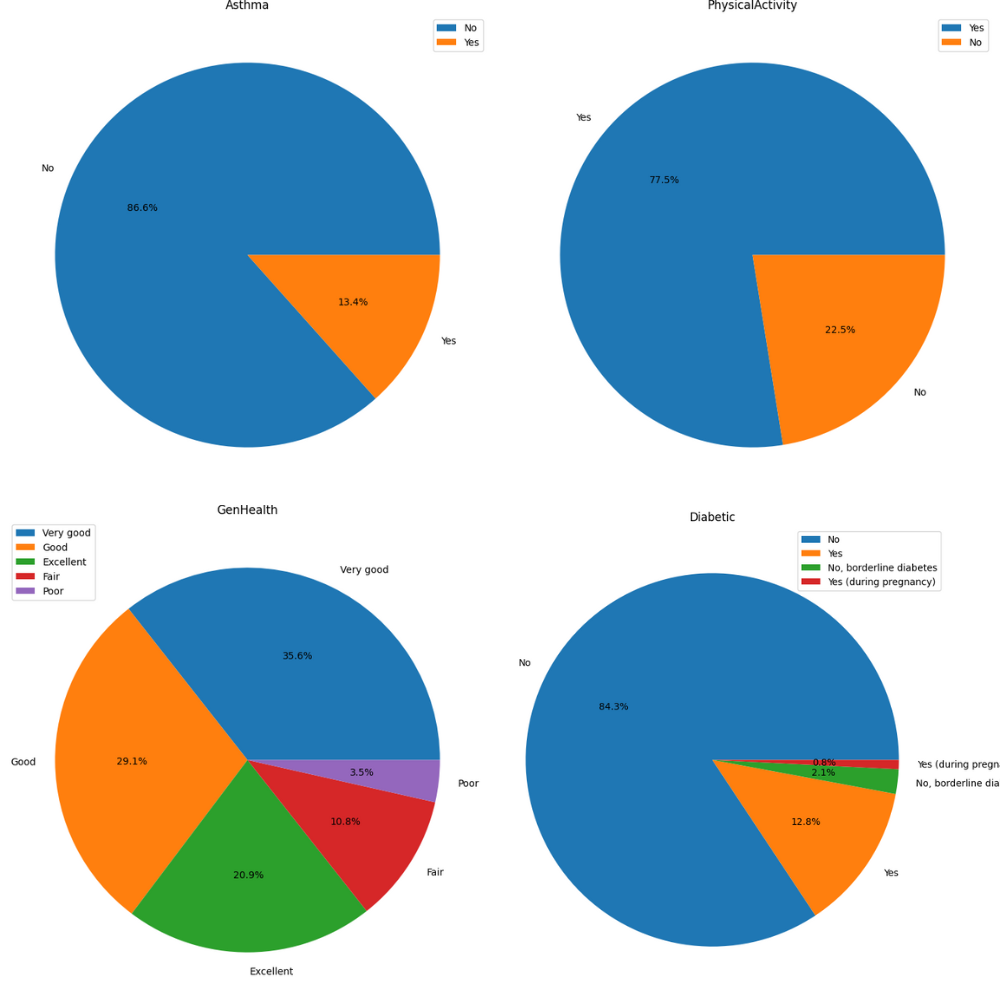
* Every categorical class changed to numerical
  + change columns that contain values of Yes / No with 0 / 1
  + change sex column that contain Male / Female with 1 / 0
  + convert AgeCategory, Race, Diabetic and GenHealth columns with increasing values
* Check If non-values exist and deal with it

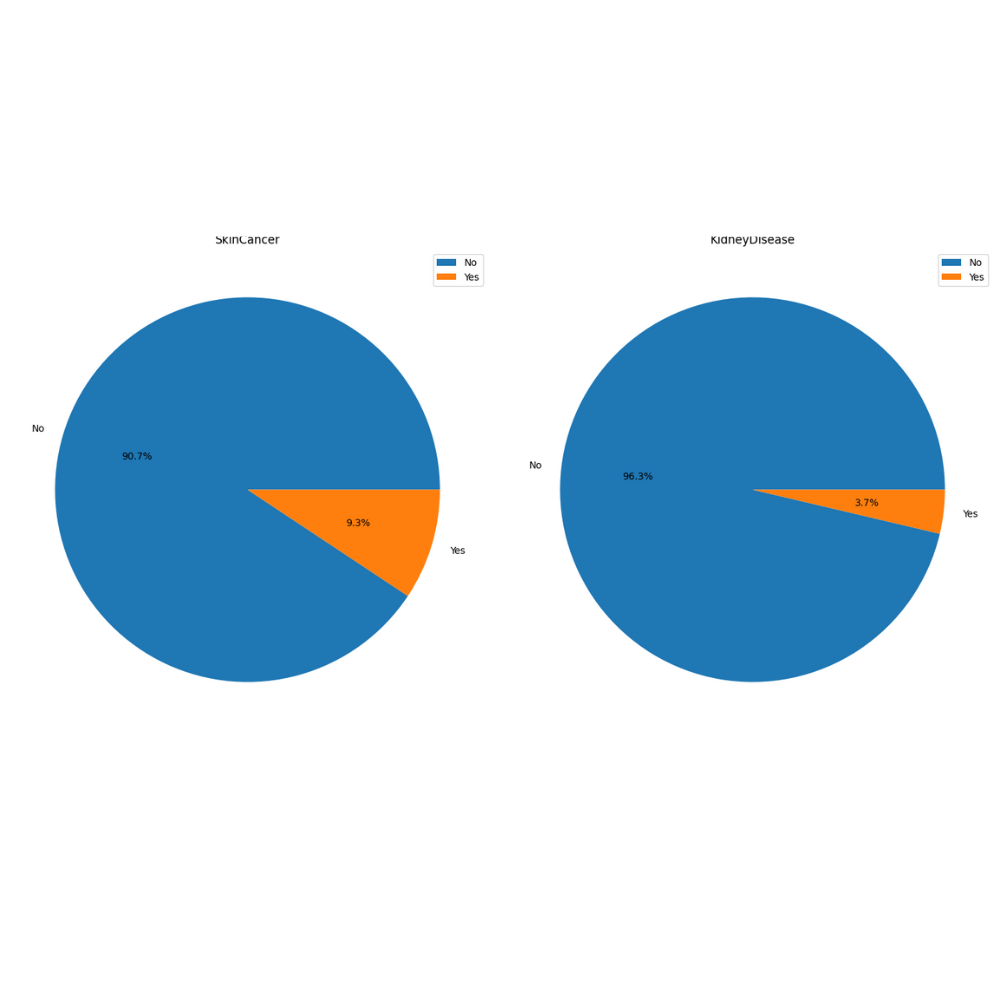
Data visualization

***Show distribution for each categorical class***

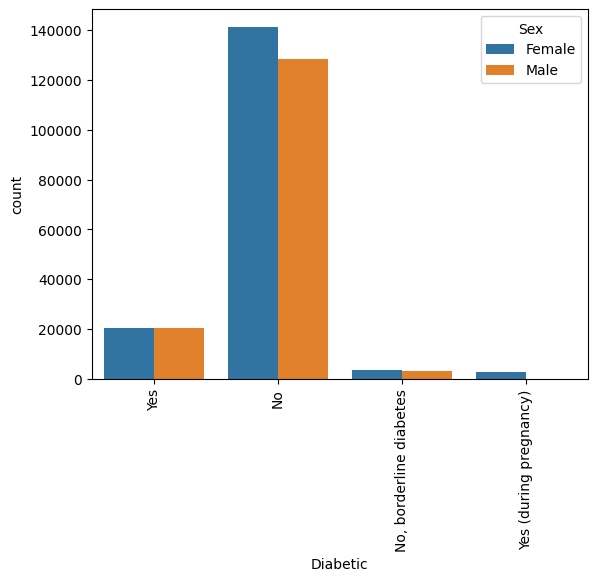
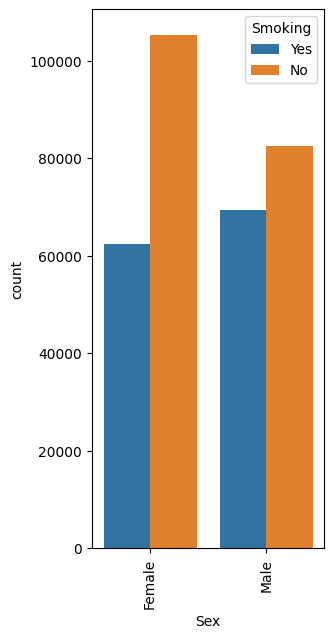




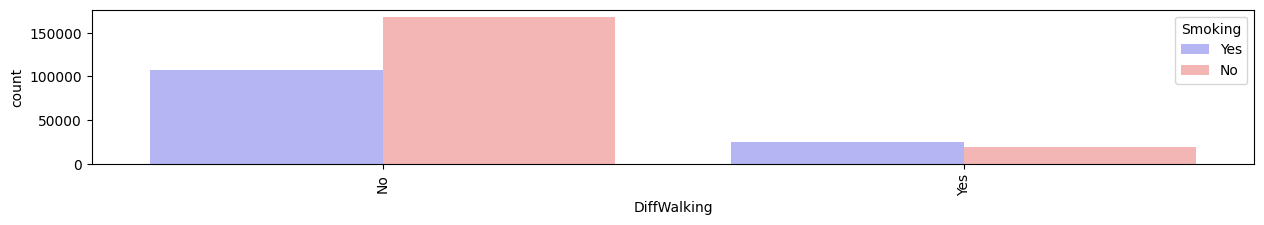
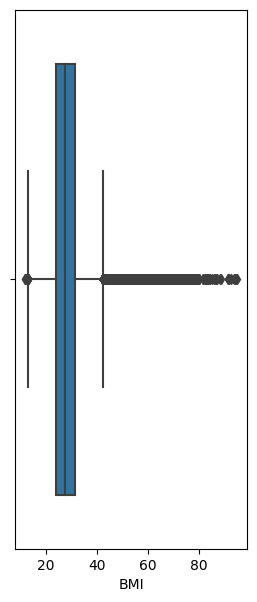
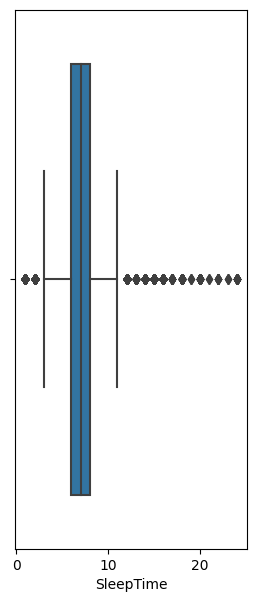
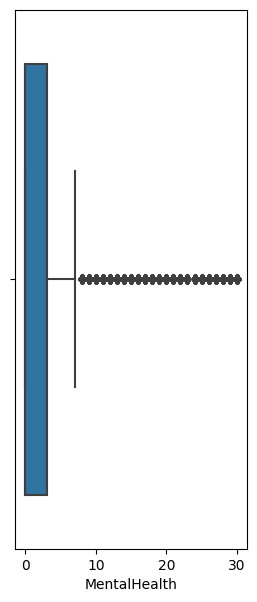
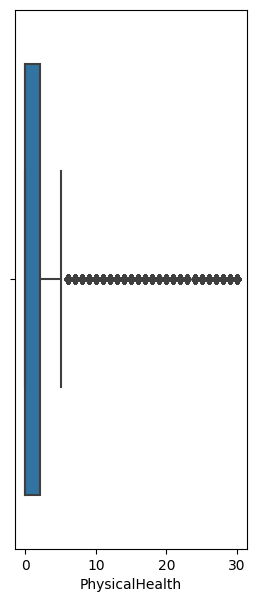




Extracting insights from data.

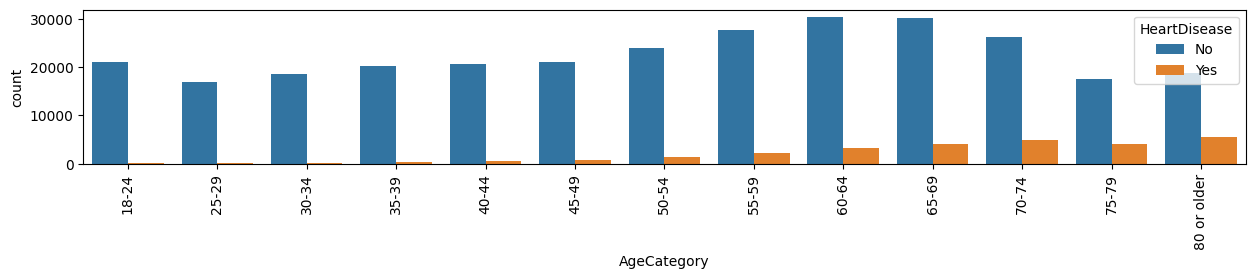
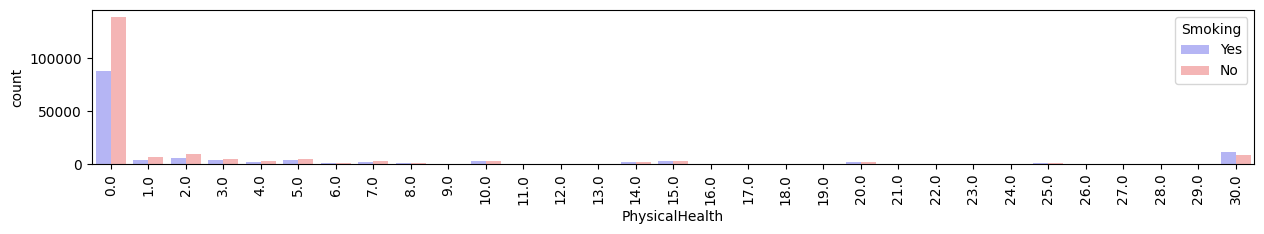


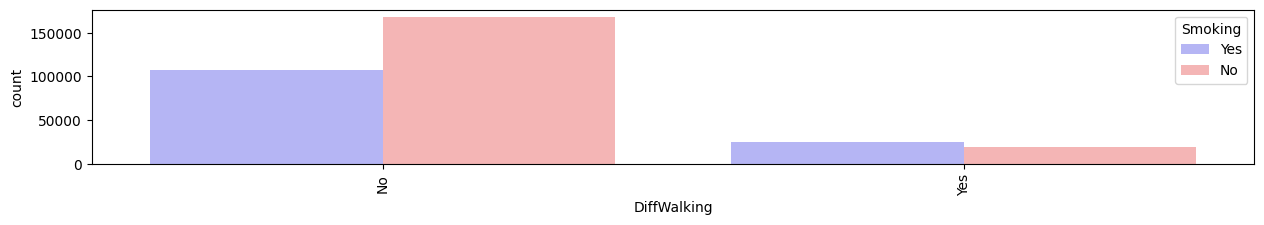
* first graph shows male have slightly more diabetes than female
* second graph shows female are smoking less than man



* first graph shows bmi average is between 20 and 40 and there are many outliers
* fourth graph shows is between 5 and 8 hours and there are outliers

Association Rules





1-no smoking → good physical health

2-more age → more heart disease

3- smoking → diff walking

Model/Classifier training

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

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

1. 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.

1. 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.

Results and Evaluation

We got 75% F1 score from the Naive Bayes we implemented using map-reduce.

And we got 76%, 62% and 76% from the ready made models logistic regression, Naive Bayes and SVMrespectively from MLlib in PySpark

Unsuccessful trials that were not included in the

final solution.

Any Enhancements and future work

We want to implement KNN using map-reduce.