Big Data Project

Supervised by :
Eng. Omar Samir

Team members

Ahmed Emad
Ahmed Hafez
Nour Moustafa
Youssef Shabrawy

Agenda

- Project Idea
- Data preprocessing
- Data Visualization
- Association Rules
- Models
- Models Evaluation
- Future Work

Project Idea

(Business Part)

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

In this Project, we develop a predictive model for the early detection of heart disease.

Data preprocessing

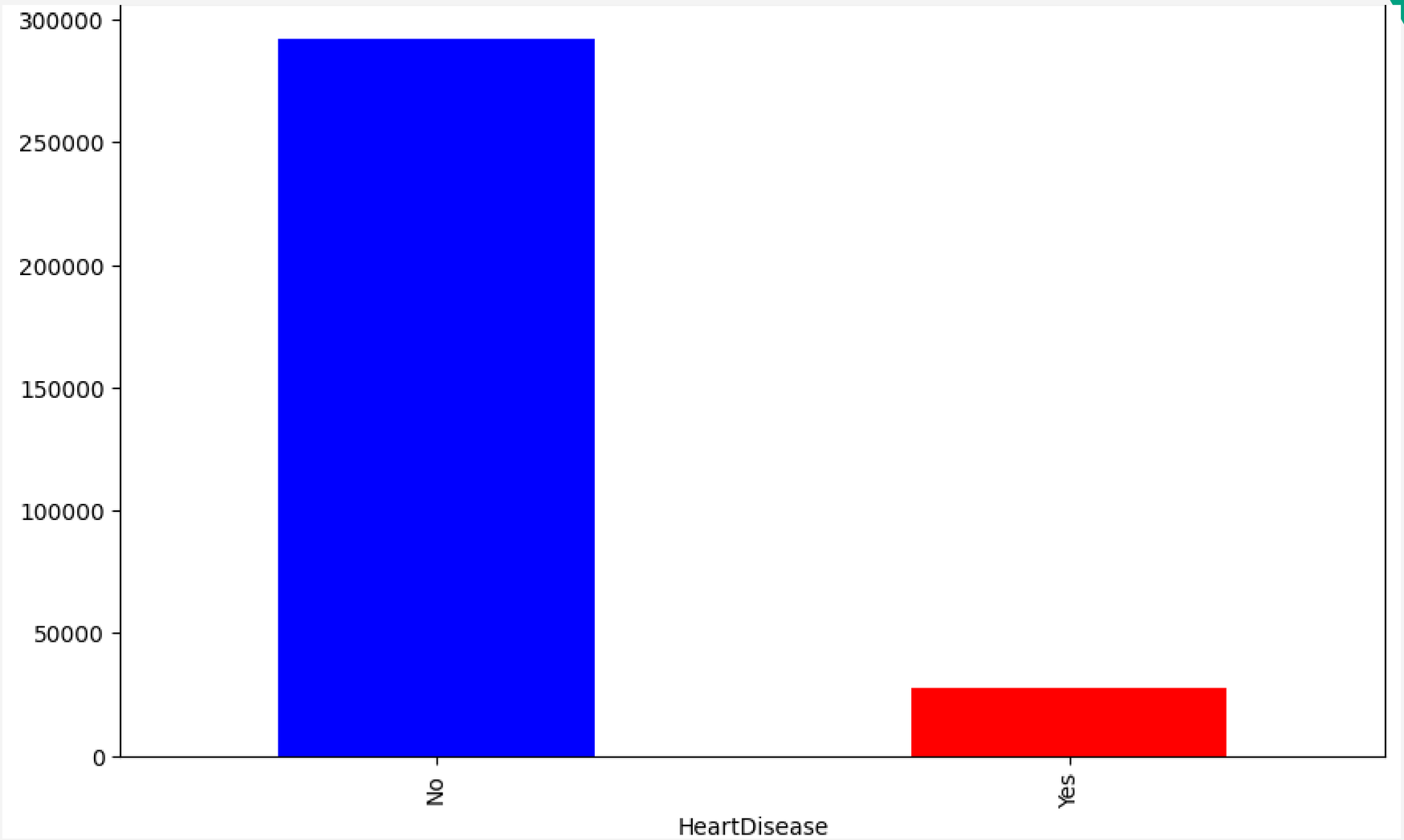
(technical Part)

check for rows that do not contain values

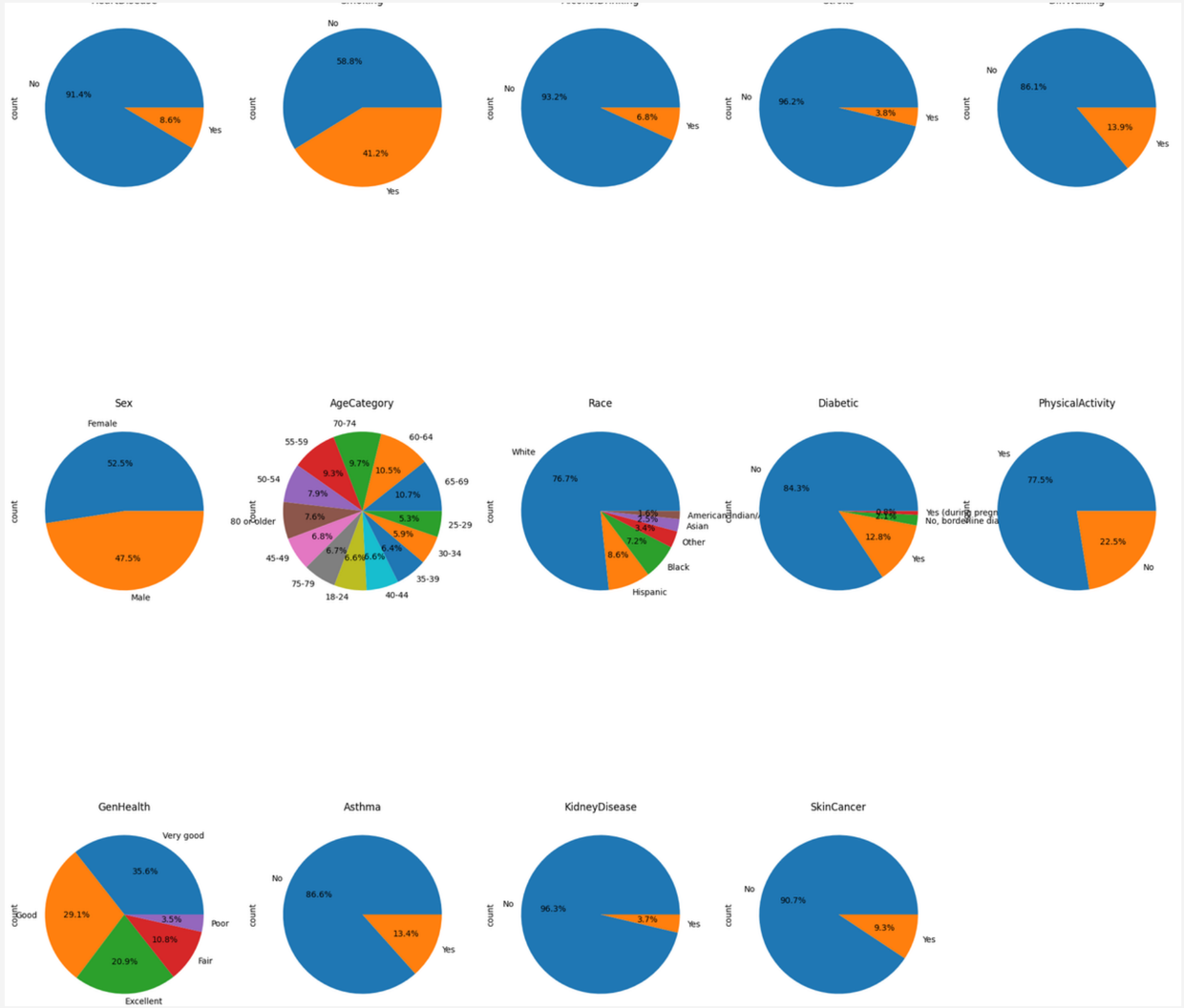
convert all non-numerical (14 features) to numerical ones

Data Visualization

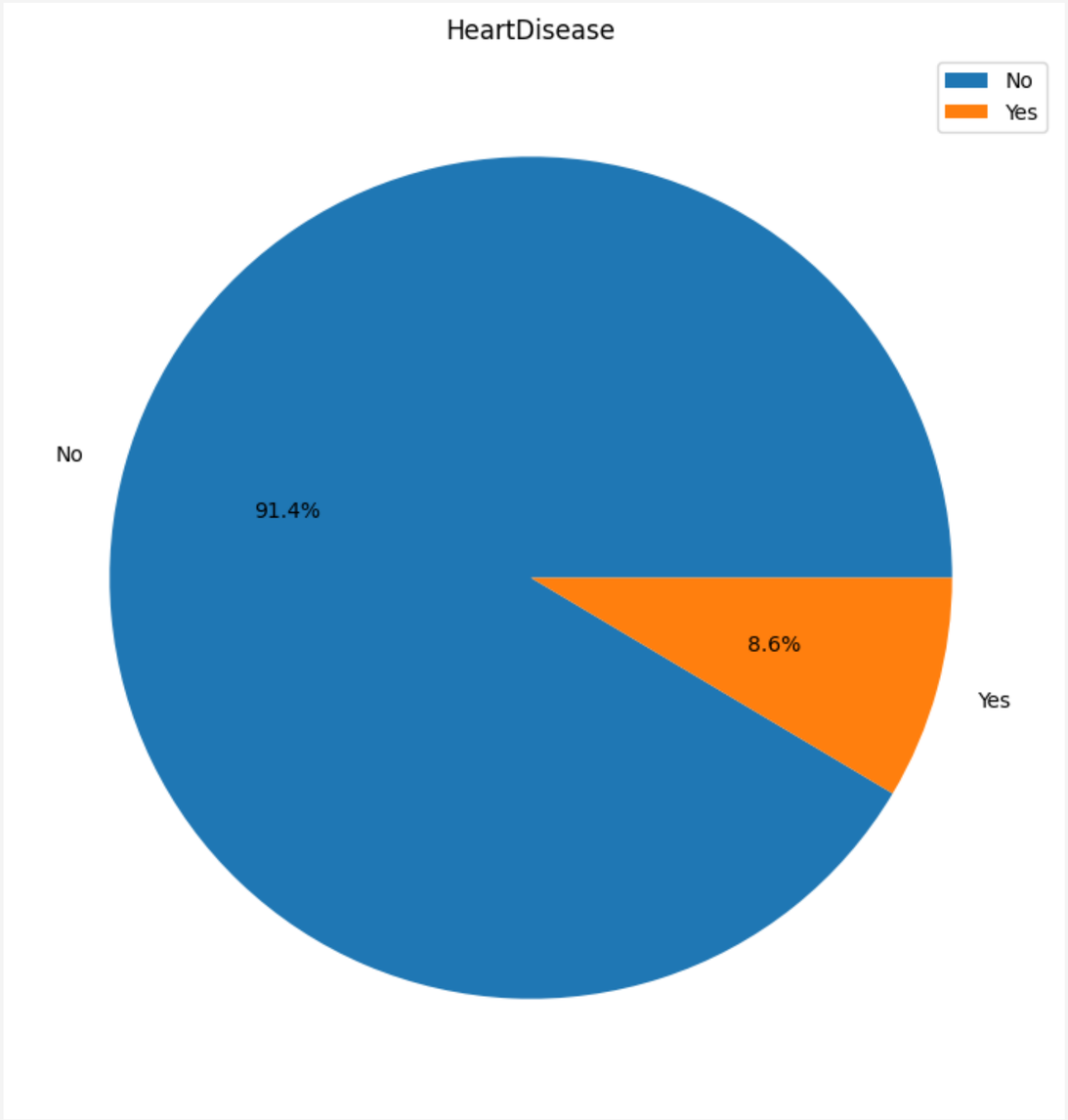
Target Class

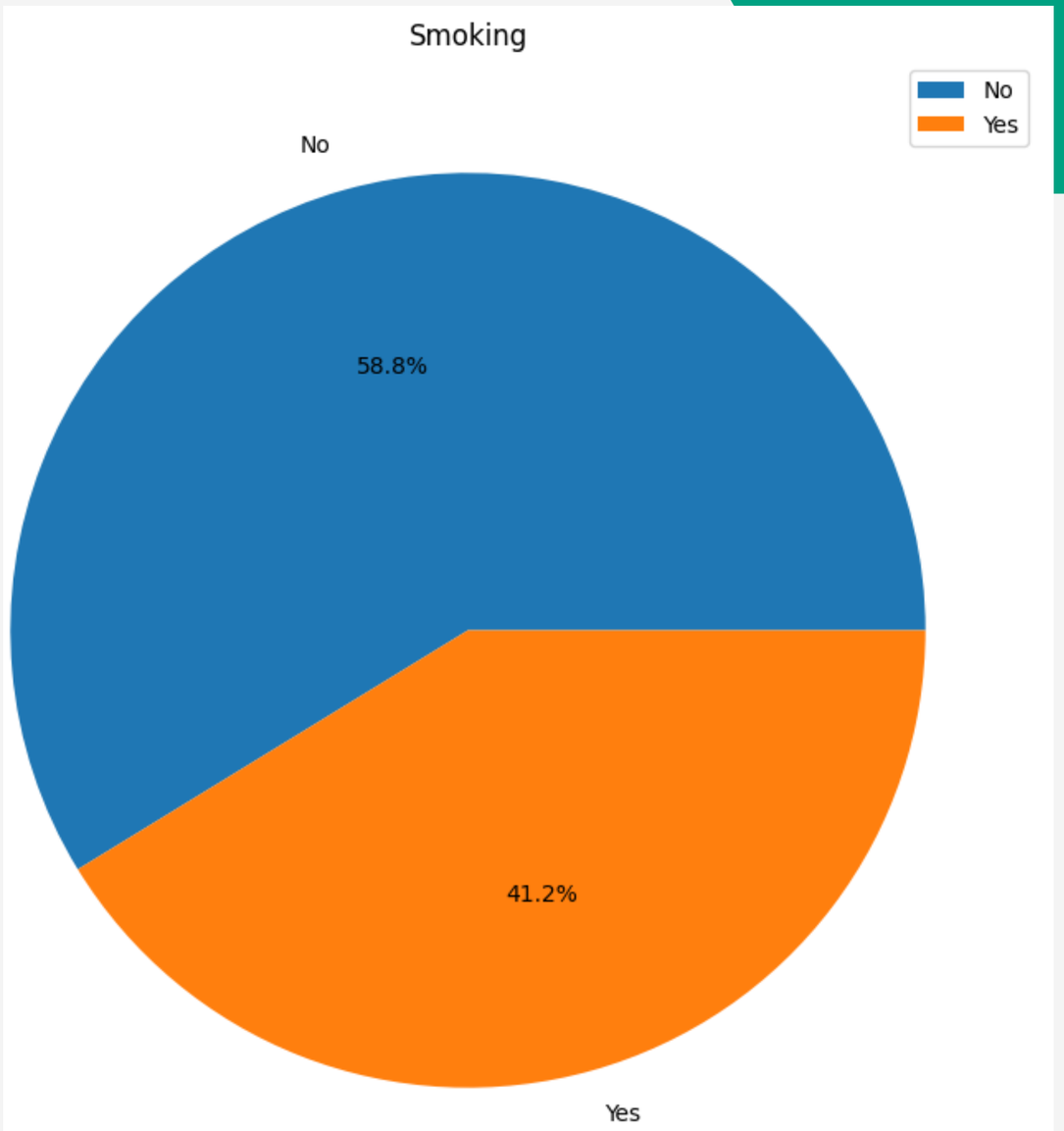


Categoral Classes Distribution

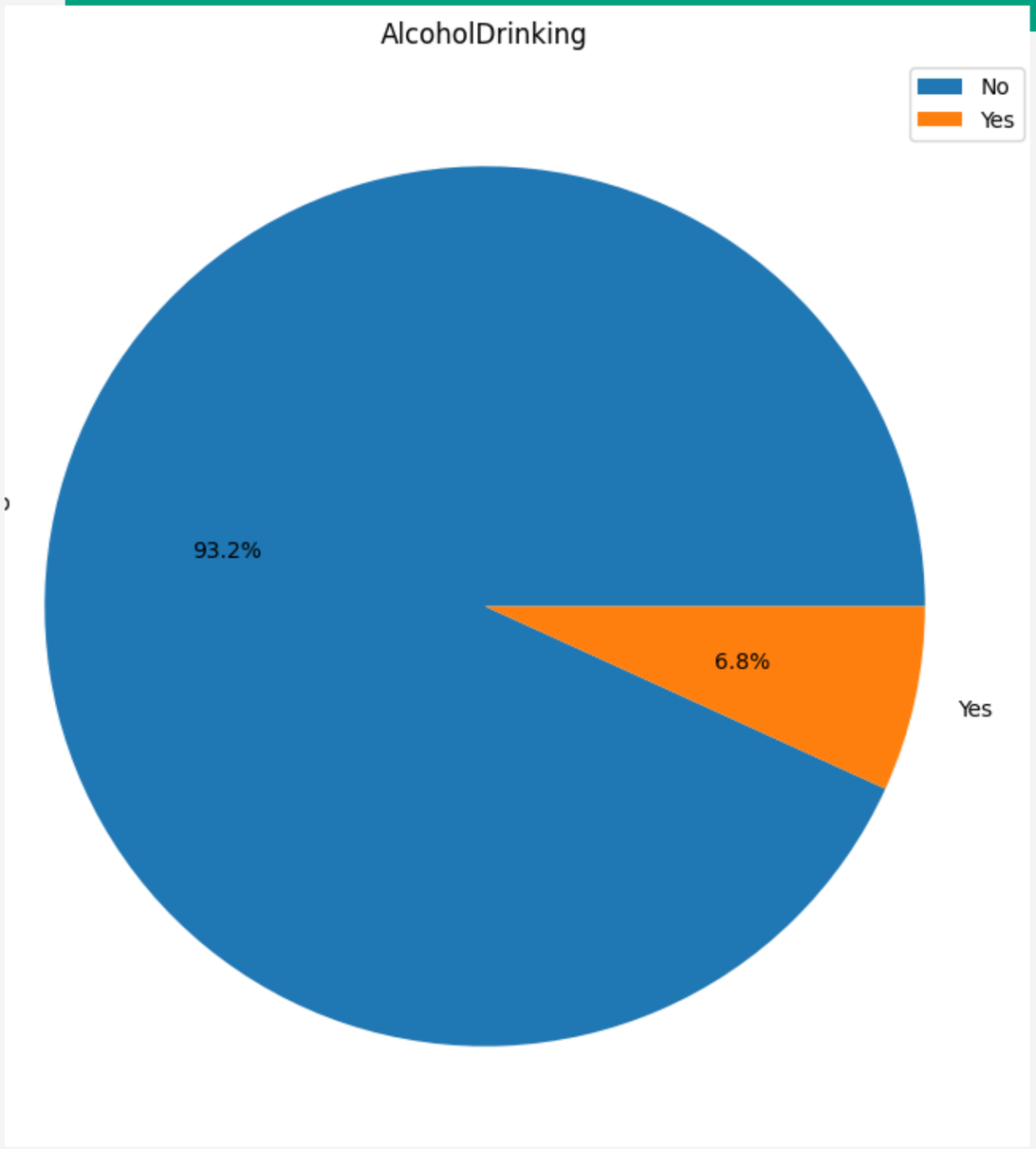


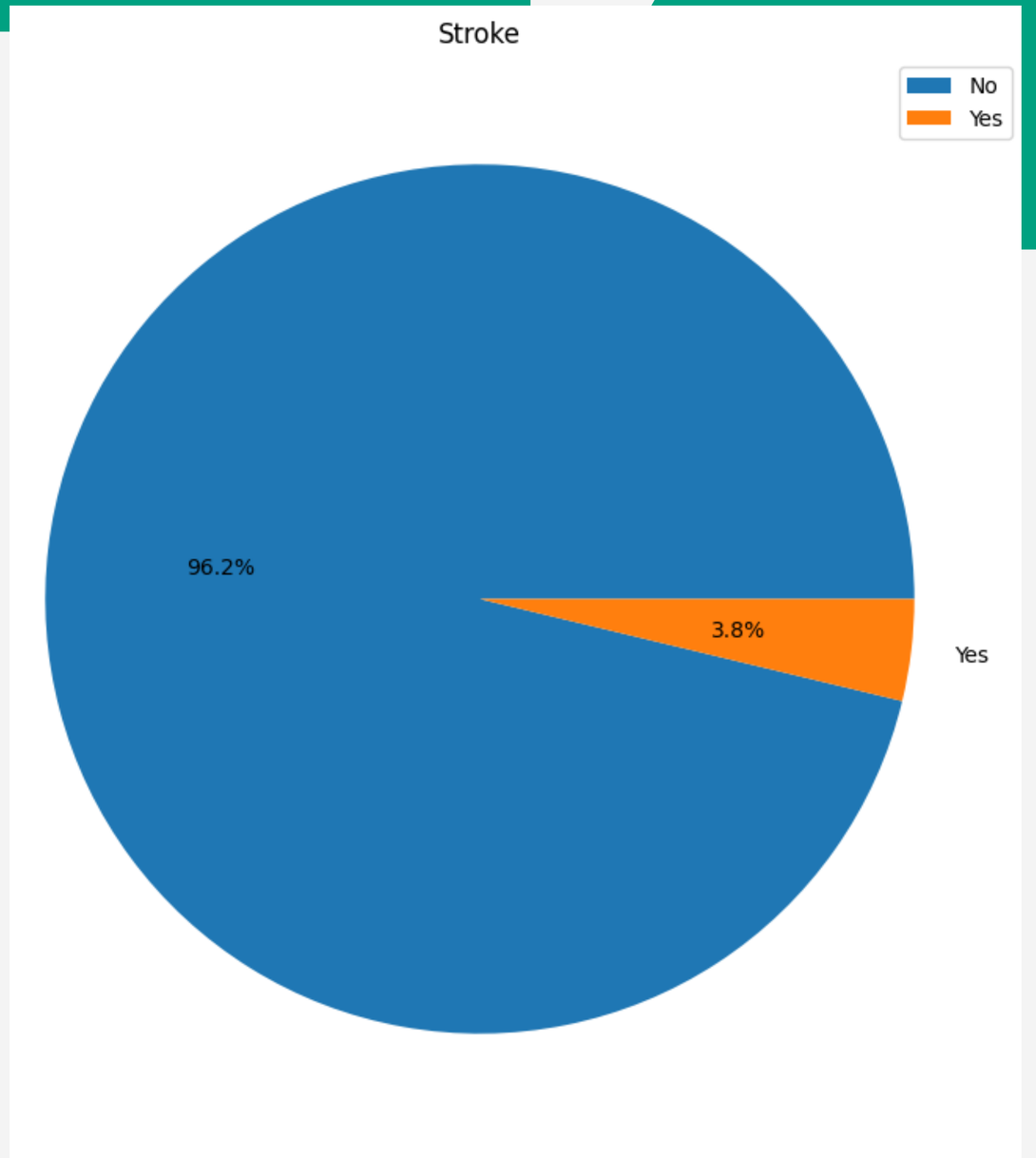
Categoral Classes Distribution



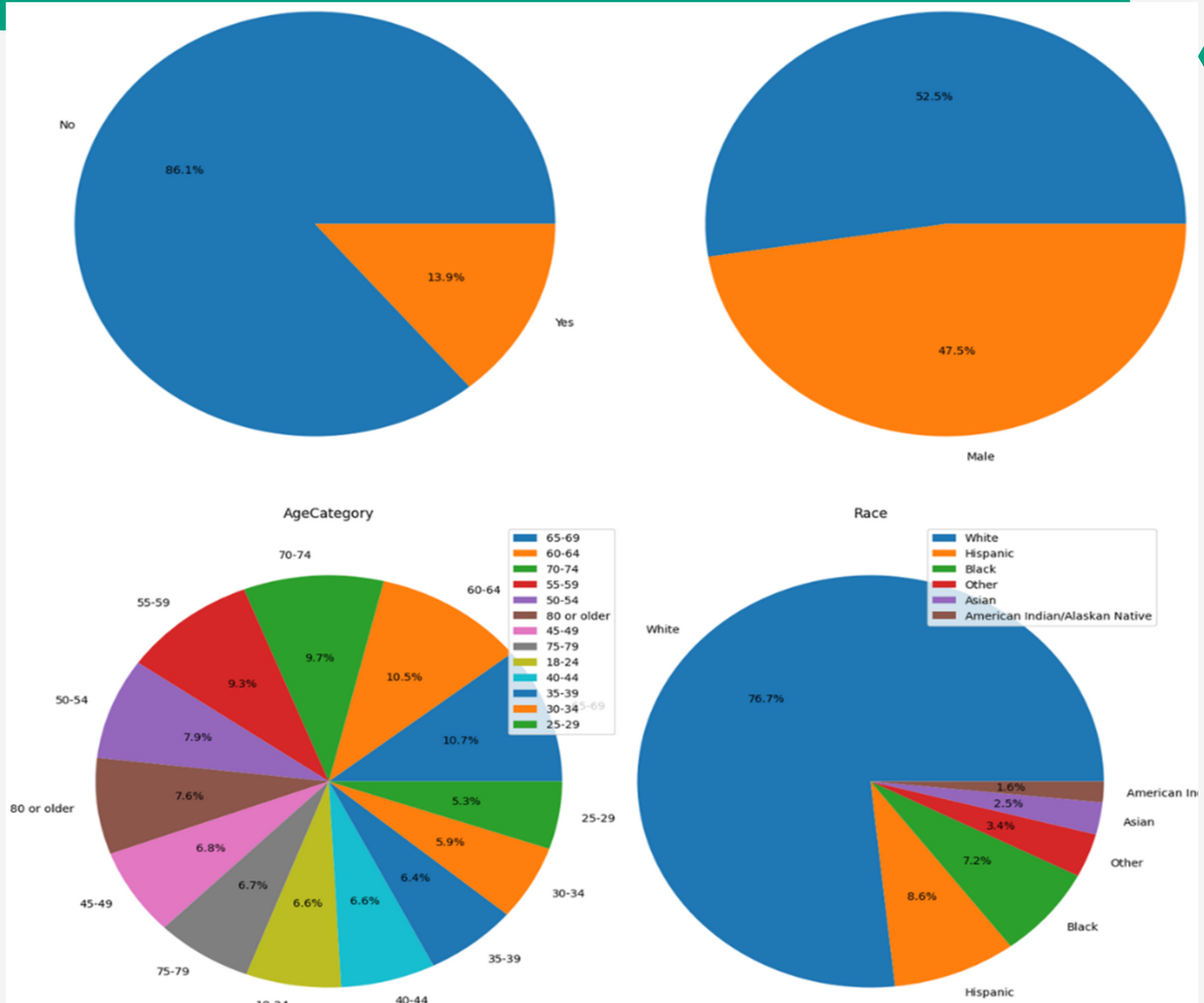


Categoral Classes Distribution

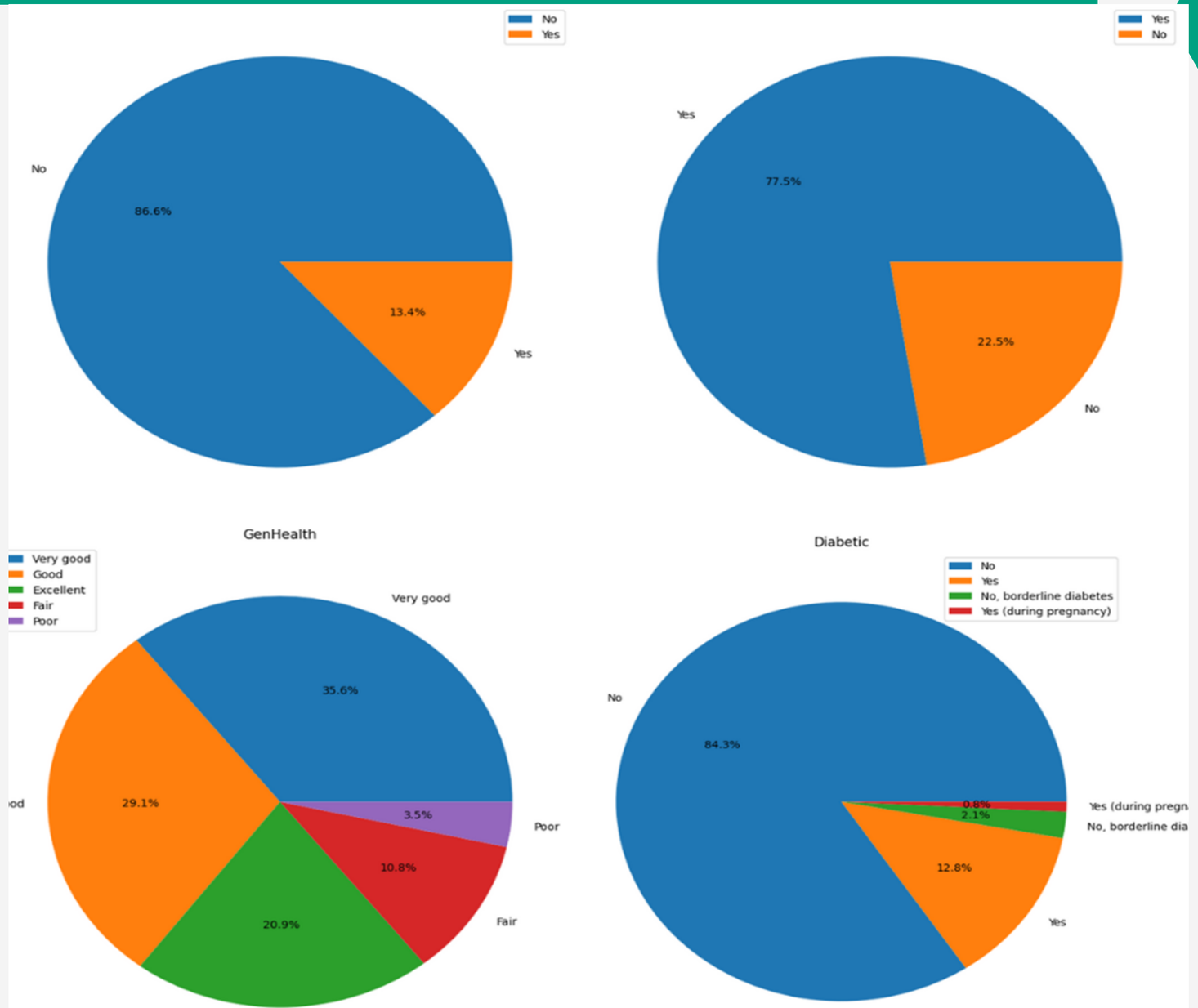




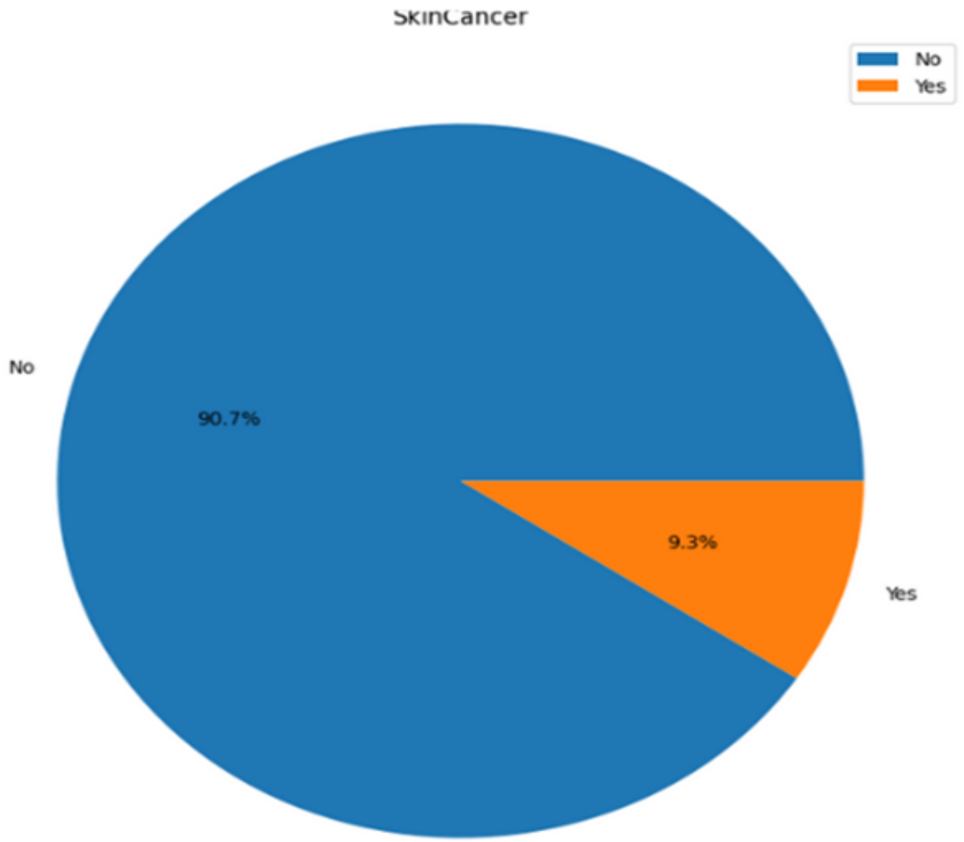
Categoral Classes Distribution

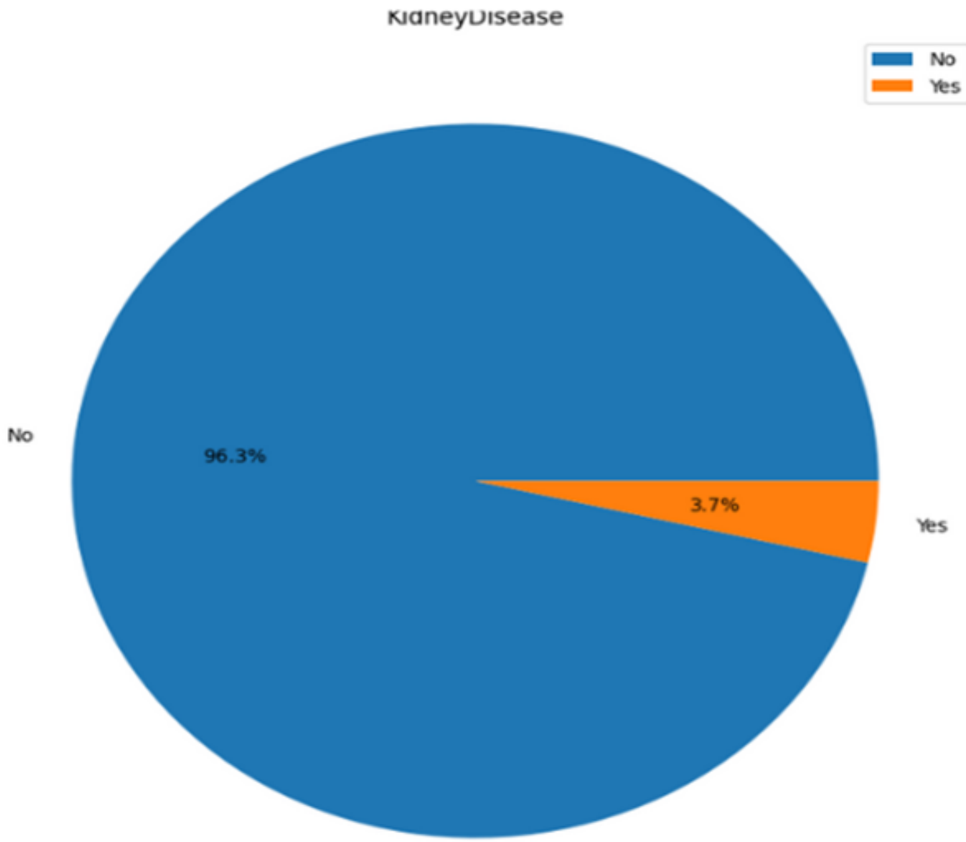


Categoral Classes Distribution

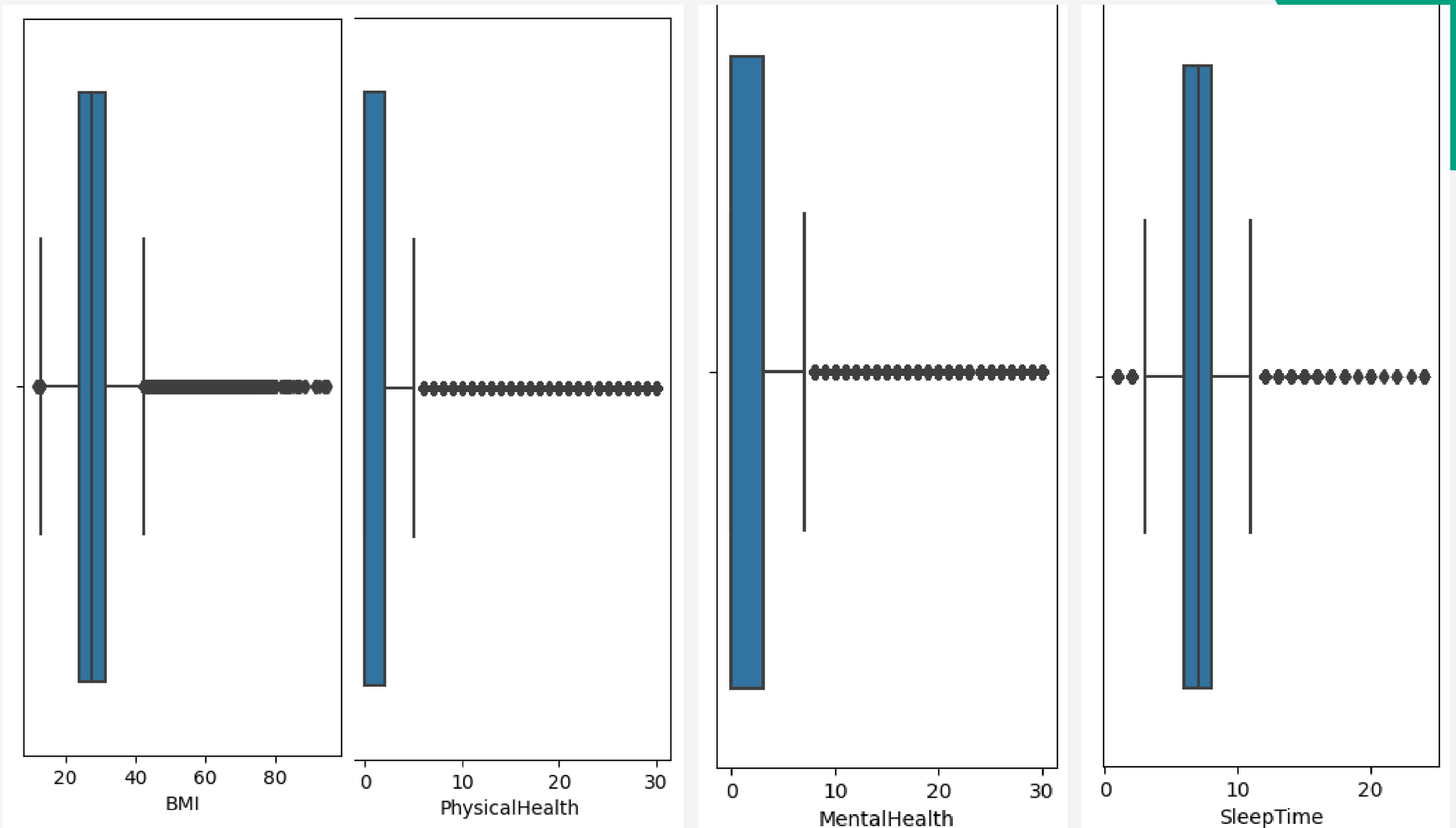


Categoral Classes Distribution

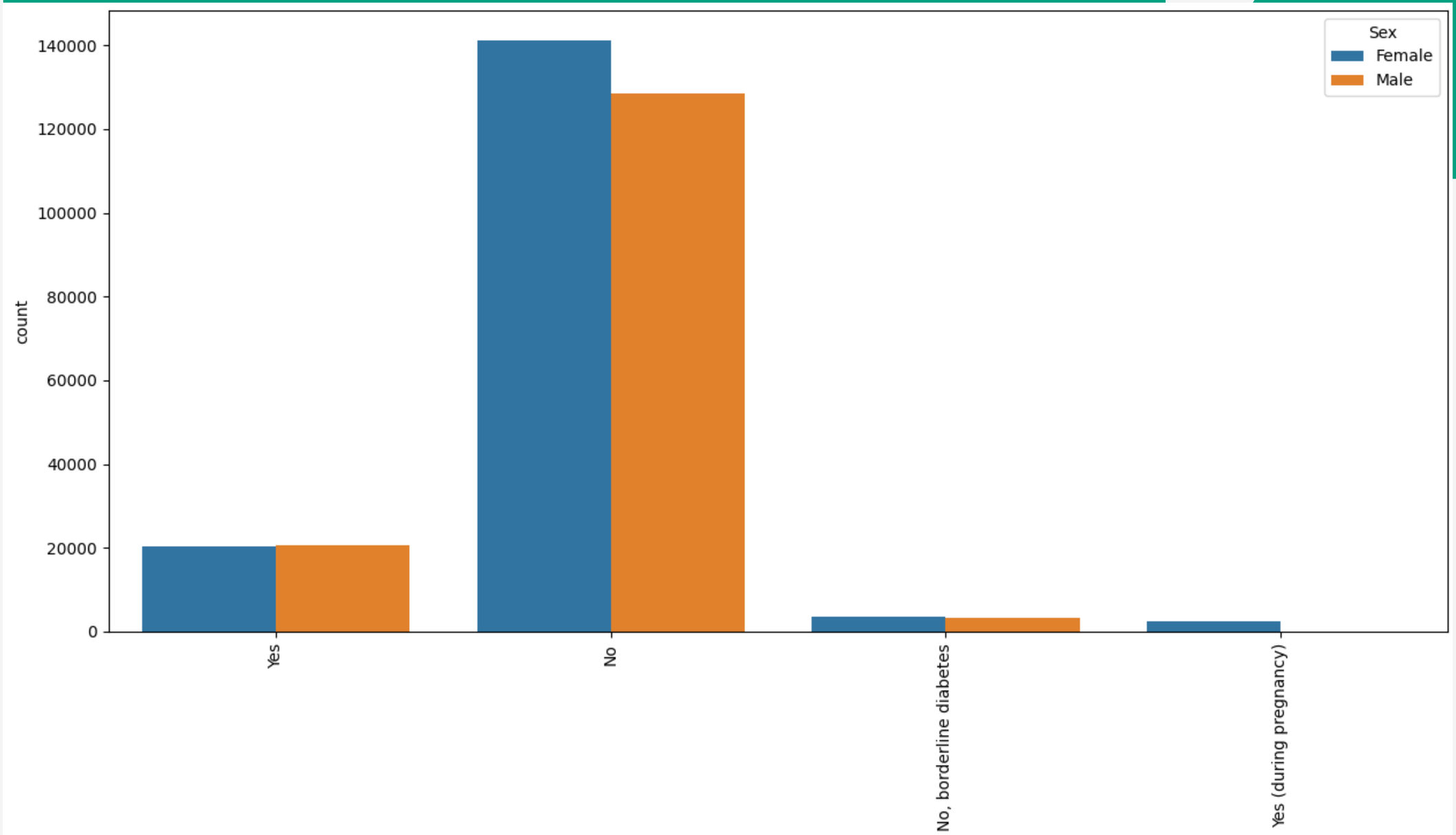




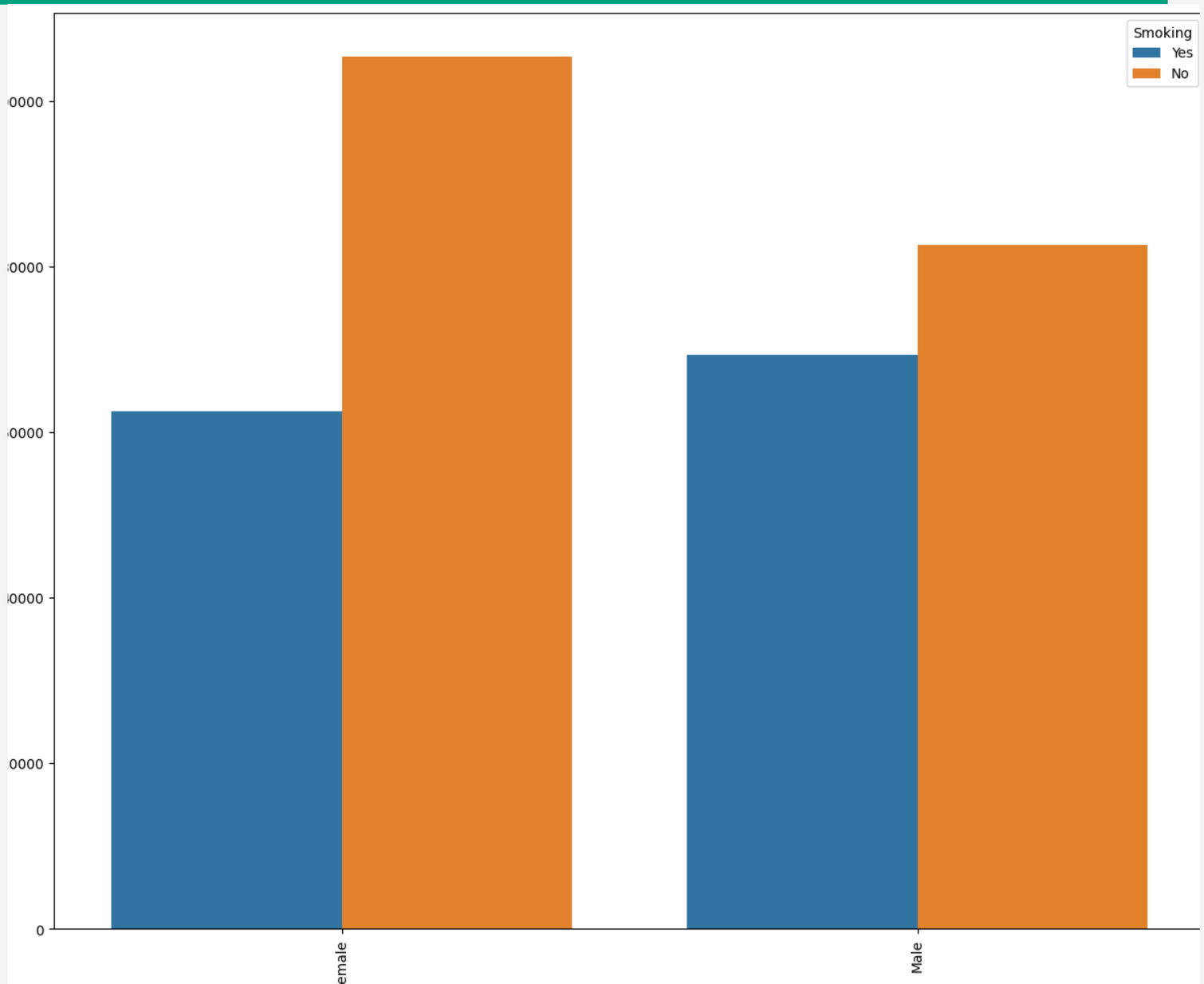
Numeric Classes Distributions



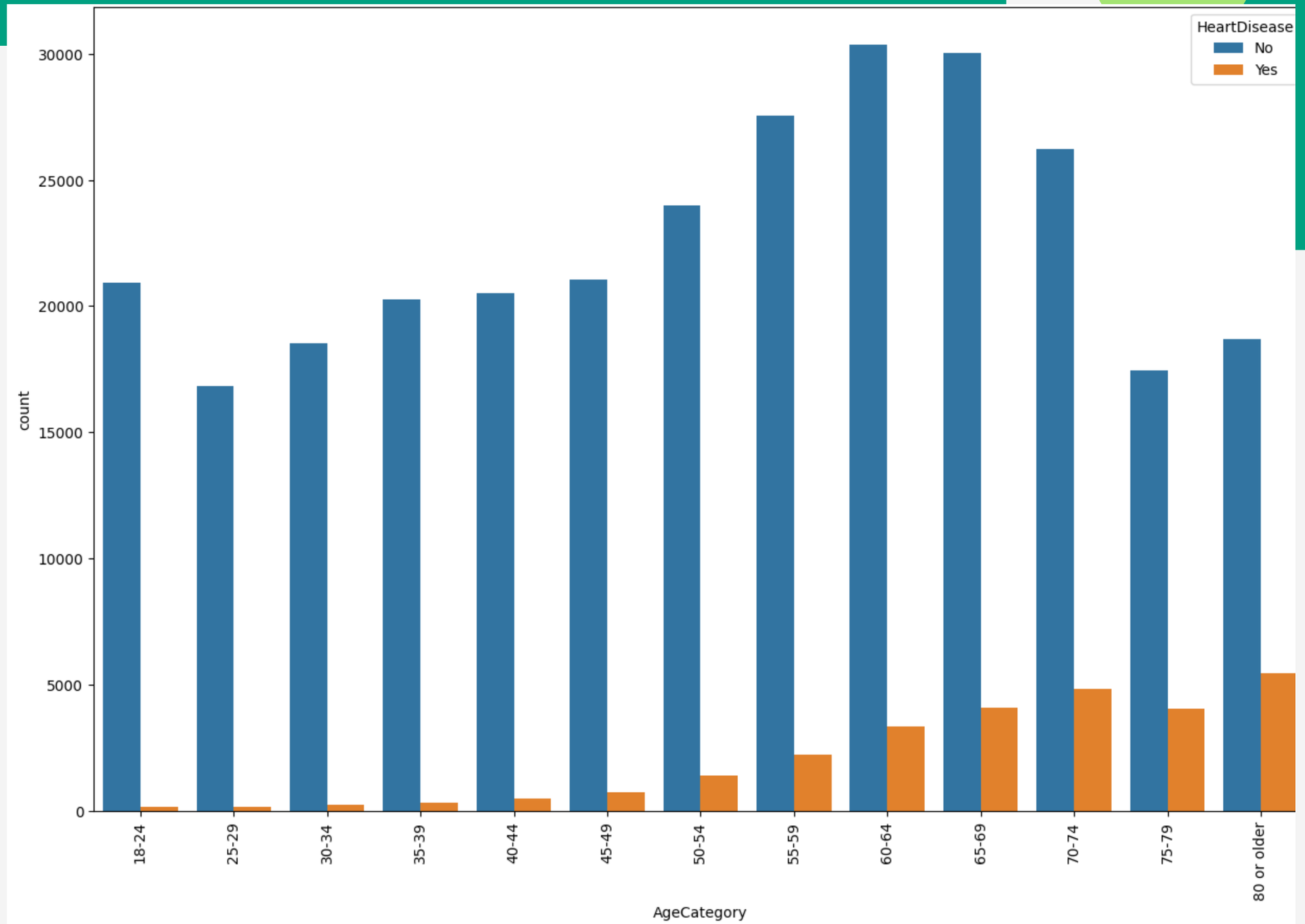
Variables Relations



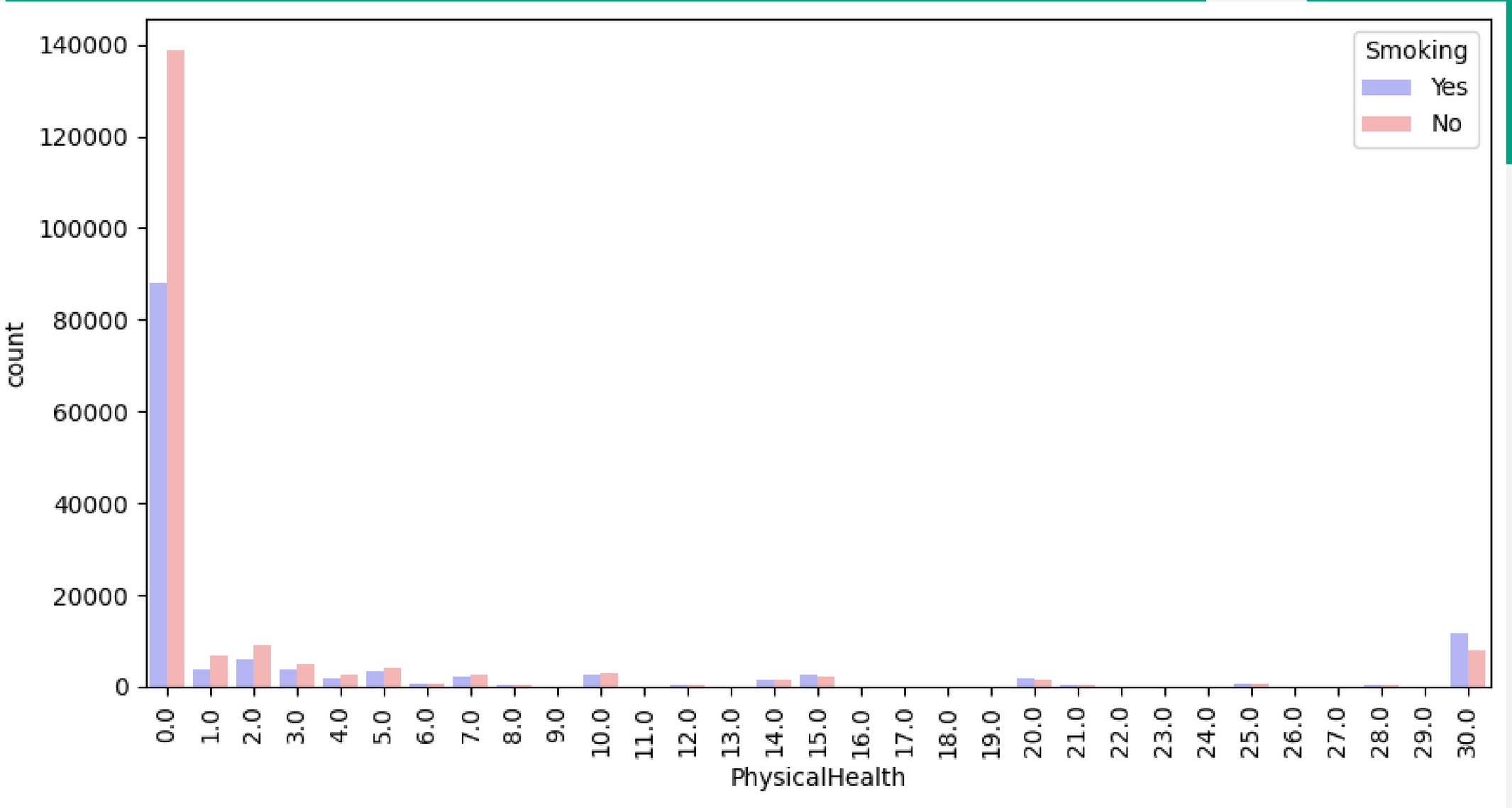
Variables Relations



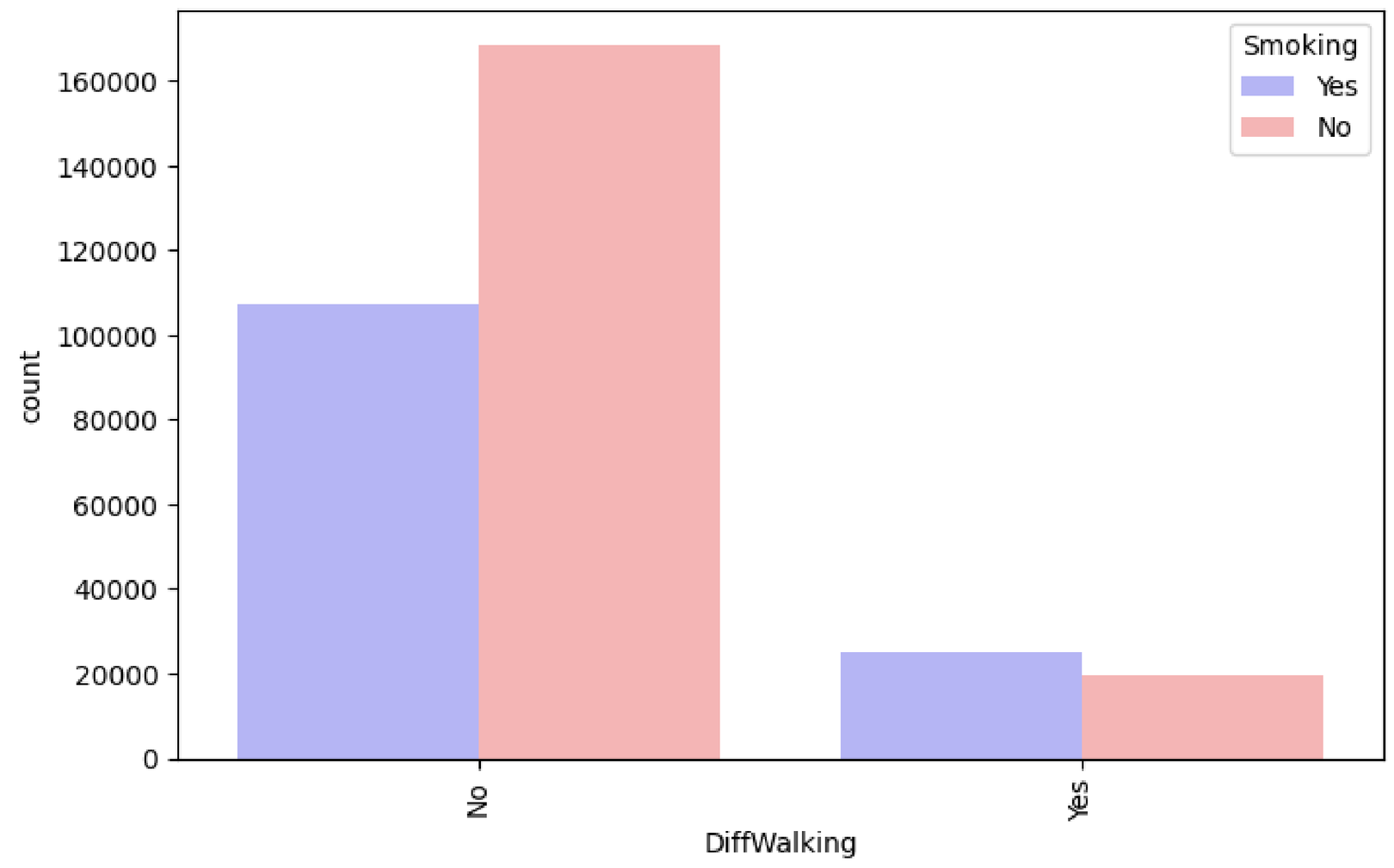
Variables Relations



Variables Relations



Variables Relations



Association Rules

HeartDisease and PhysicalHealth

Excellent -> No
No -> Excellent

HeartDisease and DiffWalking

No Rules

HeartDisease and SleepTime

medium -> No
No -> medium

HeartDisease and Smoking

Yes -> No

HeartDisease and Sex

No -> Female
Female -> No
Male -> No

HeartDisease and Age

No Rules

Models

Logistic Regression

- We used the model from ML-lib in pySpark.
- We did oversampling using RandomOversampler to balance the classes.
- Converted the data from pandas dataframe to spark dataframe.
- So we had to combine each row in the data to be a python list to feed it to the model to train on it.

Naive Bayes

- We used the model from ML-lib in pySpark.
- We did oversampling using RandomOversampler to balance the classes.
- Converted the data from pandas dataframe to spark dataframe.
- So we had to combine each row in the data to be a python list to feed it to the model to train on it.

SVM

- We used the model from ML-lib in pySpark.
- We did oversampling using RandomOversampler to balance the classes.
- Converted the data from pandas dataframe to spark dataframe.
- So we had to combine each row in the data to be a python list to feed it to the model to train on it.

Naive Bayes using map-reduce

- We did oversampling using RandomOversampler to balance the classes.
- Converted the data from pandas dataframe to rdd dataframe which convert tubular form to list of the rows.
- First, we calculated the prior probabilities of each class.
  - Map phase we generated a key-value pair <key=class, value=1>
  - Reduce phase we aggregated the values with the same key so we have a key-value pair <key=class, value=totalCount>

Naive Bayes using map-reduce

- Second, we calculated the conditional probabilities of each variable.
  - Map phase we generated a key-value pair <key= (featureValue,class), value=1>
  - Reduce phase we aggregated the values with the same key so we have a key-value pair <key= (featureValue,class), value=totalCount>
  - Another map phase to generate the conditional probabilities<key=(featureValue,class), value=totalCount/classCount>

Models Evaluation

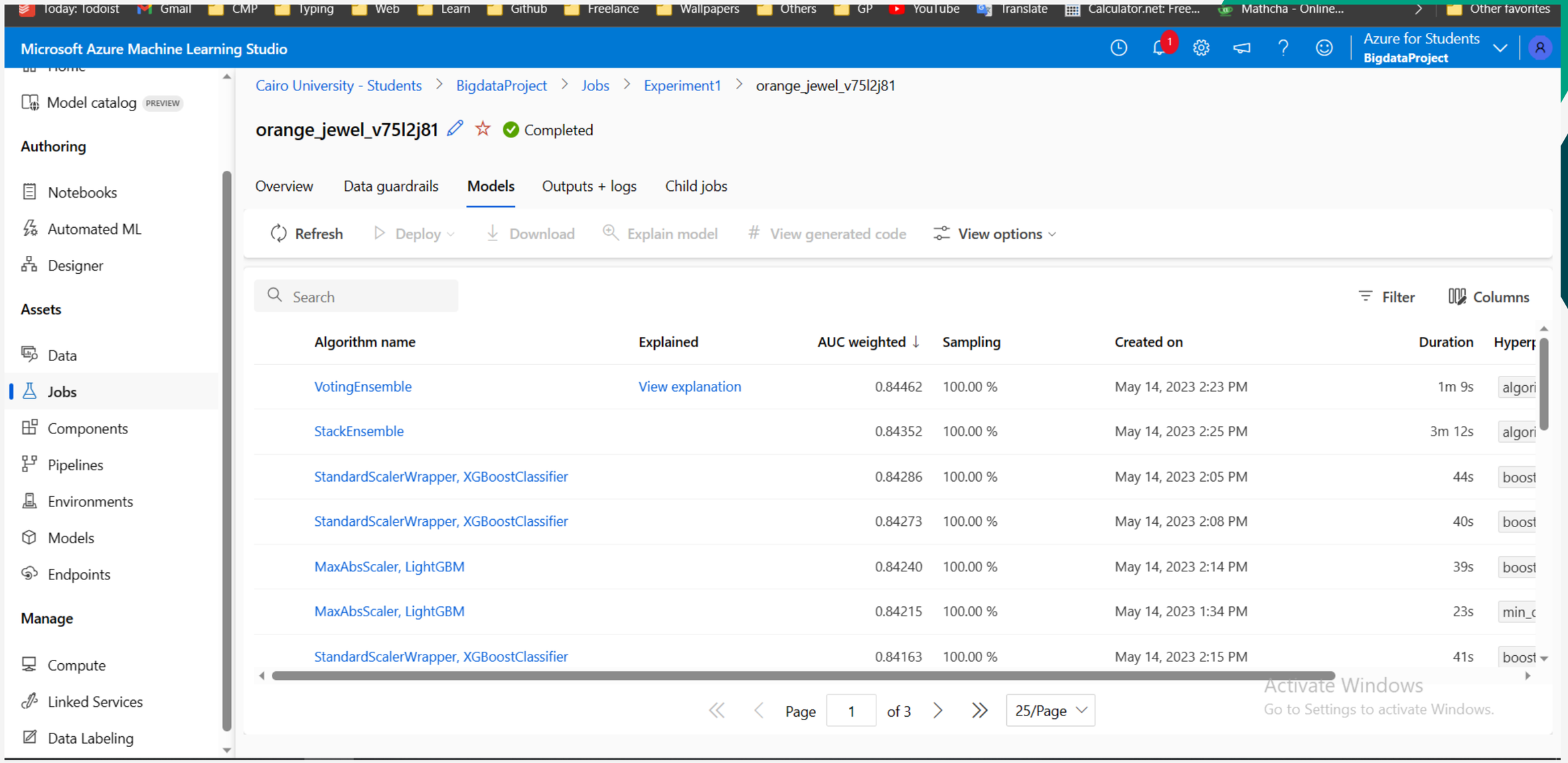
	Accuracy (F1-score)	Macro Avg
Logistic Regression	76%	76%
Naive Bayes	65%	62%
SVM	76%	76%
Naive Bayes Map-reduce	75%	75%

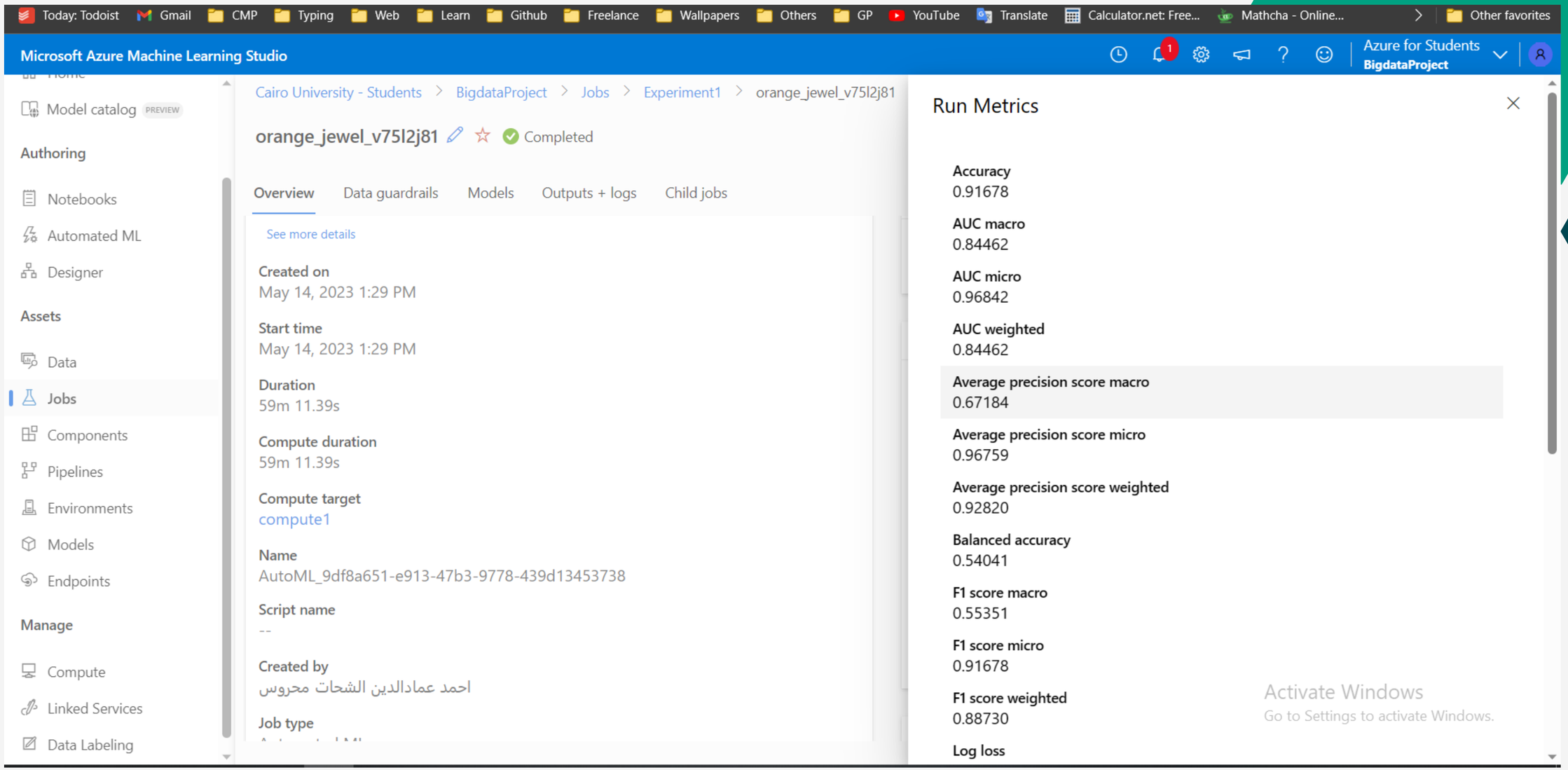
Future Work

Future Work

- We want to implement KNN using map-reduce

Bonus


Today: Todoist
Gmail
CMP
Typing
Web
Learn
Github
Freelance
Wallpapers
Others
GP
YouTube
Translate
Calculator.net: Free...
Mathcha - Online...
Other favorites
Microsoft Azure Machine Learning Studio
Azure for Students
BigdataProject
Model catalog PREVIEW
Authoring
Notebooks
Automated ML
Designer
Assets
Data
Jobs
Components
Pipelines
Environments
Models
Endpoints
Manage
Compute
Linked Services
Data Labeling
Cairo University - Students > BigdataProject > Jobs > Experiment1 > orange_jewel_v75l2j81
orange_jewel_v75l2j81
Completed
Overview
Data guardrails
Models
Outputs + logs
Child jobs
Refresh
Deploy
Download
Explain model
View generated code
View options
Search
Filter
Columns
Algorithm name
Explained
AUC weighted
Sampling
Created on
Duration
Hyperp
VotingEnsemble
View explanation
0.84462
100.00 %
May 14, 2023 2:23 PM
1m 9s
algori
StackEnsemble
0.84352
100.00 %
May 14, 2023 2:25 PM
3m 12s
algori
StandardScalerWrapper, XGBoostClassifier
0.84286
100.00 %
May 14, 2023 2:05 PM
44s
boost
StandardScalerWrapper, XGBoostClassifier
0.84273
100.00 %
May 14, 2023 2:08 PM
40s
boost
MaxAbsScaler, LightGBM
0.84240
100.00 %
May 14, 2023 2:14 PM
39s
boost
MaxAbsScaler, LightGBM
0.84215
100.00 %
May 14, 2023 1:34 PM
23s
min_c
StandardScalerWrapper, XGBoostClassifier
0.84163
100.00 %
May 14, 2023 2:15 PM
41s
boost
Page 1 of 3
25/Page
Activate Windows
Go to Settings to activate Windows.


Today: Todoist
Gmail
CMP
Typing
Web
Learn
Github
Freelance
Wallpapers
Others
GP
YouTube
Translate
Calculator.net: Free...
Mathcha - Online...
Other favorites
Microsoft Azure Machine Learning Studio
Azure for Students
BigdataProject
Model catalog
PREVIEW
Authoring
Notebooks
Automated ML
Designer
Assets
Data
Jobs
Components
Pipelines
Environments
Models
Endpoints
Manage
Compute
Linked Services
Data Labeling
Cairo University - Students
BigdataProject
Jobs
Experiment1
orange_jewel_v75l2j81
orange_jewel_v75l2j81
Completed
Overview
Data guardrails
Models
Outputs + logs
Child jobs
See more details
Created on
May 14, 2023 1:29 PM
Start time
May 14, 2023 1:29 PM
Duration
59m 11.39s
Compute duration
59m 11.39s
Compute target
compute1
Name
AutoML_9df8a651-e913-47b3-9778-439d13453738
Script name
--
Created by
احمد عمادالدين الشحات محروس
Job type
Run Metrics
Accuracy
0.91678
AUC macro
0.84462
AUC micro
0.96842
AUC weighted
0.84462
Average precision score macro
0.67184
Average precision score micro
0.96759
Average precision score weighted
0.92820
Balanced accuracy
0.54041
F1 score macro
0.55351
F1 score micro
0.91678
F1 score weighted
0.88730
Log loss
Activate Windows
Go to Settings to activate Windows.

Thank You