Heart Risk Analysis using Big Data and Machine Learning Techniques

Group –Cloud Clan

NAMES AND ROLES:

Sai Prudhvi Charan Pothumsetty - 16343752

- Identify and obtain the heart disease dataset.
- Verify data for completeness and consistency.
- Preprocess the dataset according to the random forest classifier.
- Splitting the dataset into training and testing datasets

Manchireddy Kavya Reddy - 16342232

- Perform feature transformation using Vector Assembler.
- Compare and select the best-performing feature extraction method.

Sai Deeksha Enukonda - 16343736

- Train the machine learning model using the training dataset.
- Evaluate the models using various evaluation metrics such as accuracy, precision, recall, and F-1 score.
- Select the best-performing model based on its evaluation metrics.
- Use the selected model for random forest analysis on new data.
- Show the visualizations using matplotlib and seaborn or by using power BI.

Collaboration Methods:

The team collaborated by assigning roles and responsibilities based on individual strengths and completing assigned tasks on a daily basis. This ensured that everyone was working efficiently and effectively towards the project objectives.

1. Abstract

The prediction of heart attack is a critical issue in the field of healthcare. The use of machine learning techniques has been proposed as a means to accurately predict the occurrence of heart attacks. In this study, the process of developing machine learning models for heart attack prediction using a dataset containing various risk factors. The dataset was preprocessed to clean the data, further the data is processed according to Random Forest Classifier and Naïve Bayes Classifier. Then the processed data undergoes feature selection, by selecting the most suitable features and therefore converting them into feature vector using one of the feature transformation techniques and then given as an input to different machine learning algorithms including random forest classifier and naïve bayes classifier to predict heart attack. The performance of each algorithm was evaluated using metrics such as accuracy, f1-score, precision and recall. This study demonstrates the potential of machine learning techniques on classification problems in predicting heart attacks and could help healthcare providers to identify individuals at risk of heart attack and take preventive measures.

The implementation was done in Google Colab using PySpark and Python and the in-built libraries such as matplotlib, etc., allowing us to perform heart stroke prediction on the entire heart disease dataset in a reasonable amount of time. We have also use Power BI to show the visualization on heart risk data.

This project provides a useful tool for heart stroke prediction on heart disease data using PySpark and machine learning algorithms. It demonstrates the effectiveness of PySpark as well as Python for processing large datasets and extracting insights from the obtained heart disease data.

2. Introduction

Every year, 13 million individuals suffer a stroke according to the World Stroke Organization, resulting in about 5.5 million deaths, making it the primary cause of fatality and disability across the world. This condition has severe implications on all aspects of life since it not only influence the patients but also their social circle, family, and work environment. Furthermore, it is a common misconception that stroke only affects specific demographics, but in reality, it can occur at any age or physical condition, regardless of gender.

In this project, the heart attack prediction on the heart disease dataset which contains multiple attributes aims to develop a model that can accurately predict the heart stroke field (Stroke) based on the other 17 fields (‘HeartDisease’, ’BMI’, ’Smoking’, ’AlcoholDrinking’, ‘PhysicalHealth’, ‘MentalHealth’, ‘DiffWalking’, ‘Sex’, ‘AgeCategory’, ‘Race’, ‘Diabetic’, ‘PhysicalActivity’, ‘GenHealth’, ‘SleepTime’, ‘Asthma’, ‘KidneyDisease’, ‘SkinCancer’).

To accomplish this, PySpark, a powerful distributed computing framework that enables us to process large datasets efficiently is being used along with some popular machine learning algorithms namely Random Forest Classifier and Naïve Bayes Classifier.

The project is implemented in Google Colab, which provides a free and easy-to-use environment for working with PySpark, Python and machine learning libraries. The results show that the models perform well on the heart disease dataset with better accuracy.

This project uses PySpark, Python and machine learning algorithms to provide useful tools for heart disease data. This demonstrates the effectiveness of PySpark and Python for processing the large datasets and extracting insights from healthcare data. The results of this project can be applied by healthcare industries to improve their services by better understanding the serious causes of heart strokes in their patients.

3. Project Scope/Purpose

The main scope of the project is to predict heart attack/stroke on the heart disease dataset using PySpark, Python and machine learning algorithms. The dataset contains different attributes such as BMI, Smoking, Alcohol Drinking, Stroke, Physical, Mental Health, Diffwalking, Sex, AgeCategory, Diabetic, Race, Physical Activity, Gen Health, Sleep time, Asthma, KidneyDisease, SkinCancer making it an ideal dataset for training and evaluating machine learning models. The project focuses on using machine learning algorithm such as Random Forest classifier and Naïve Bayes classifier for prediction of heart stroke.

The project involves cleaning and preprocessing the data, followed by data analysis to identify patterns and relationships between variables. Feature transformation is done to create a feature vector, and important features are selected based on the results of the analysis. The data is then splitted into training and testing sets, and the model is trained and evaluated using different metrics. Finally, the model is tested by passing a sample feature vector to verify its quality in predicting heart stroke/attack.

The implementation of the project is done using PySpark, a powerful distributed computing framework that allows for the efficient processing of large datasets. The use of PySpark enables us to perform heart stroke prediction on the entire heart disease dataset in a reasonable amount of time, which would not have been feasible using traditional machine-learning tools.

4. System Design

4.a. Architectural Diagram:

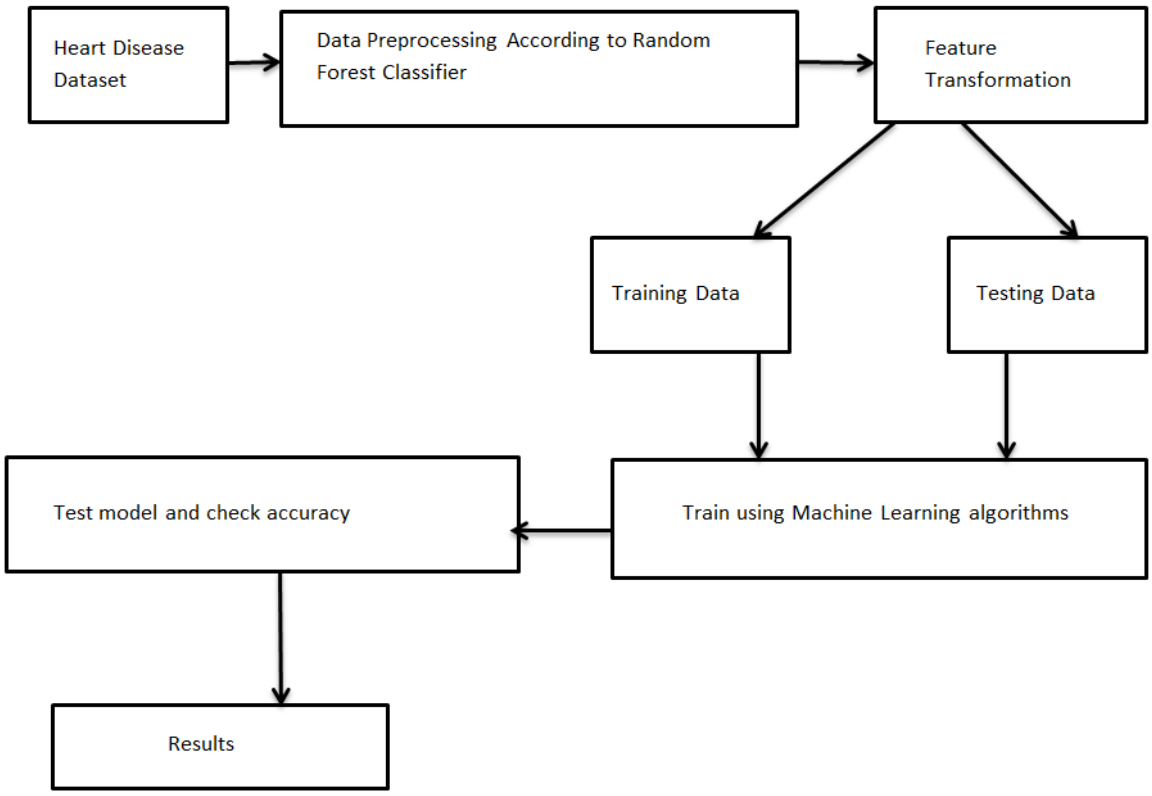


Fig 1: Architectural Diagram

5. Data Visualizations

Data Visualization using Pyspak :

Matplotlib: widely-used Python library offers comprehensive capabilities for generating static, interactive, and animated visual representations of data. Its broad range of plot types encompasses line plots, scatter plots, bar plots, histograms, heatmaps, and additional chart types.

Seaborn: Seaborn is a Python library used for data visualization that is constructed on Matplotlib. It offers a more advanced and user-friendly interface for generating statistical graphics that are both visually attractive and informative, unlike the standard plots produced by Matplotlib.

Visualizations:

1. Plot 1 - Bar plot/chart of number of persons having never diagnoised with major diseases like Asthma, Heart Disease, Skin Cancer, and Kidney Disease and got a Heart Stroke and it`s count is being visualized by grouping according to the General Health field.

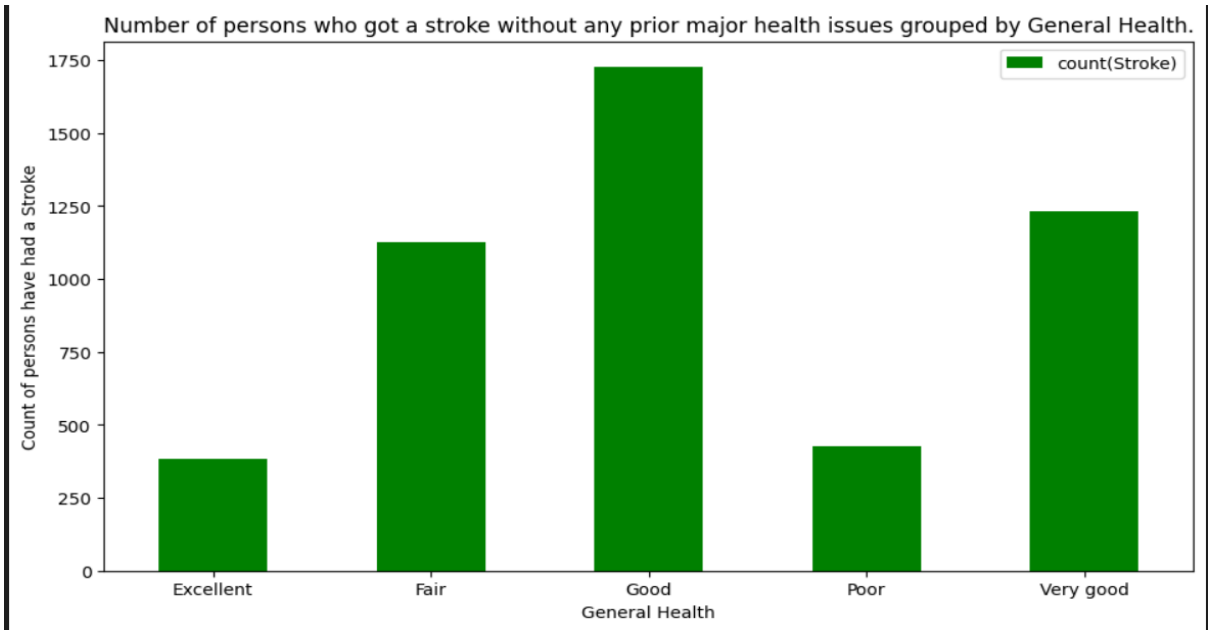


Fig 5: Bar plot for Number of persons who got a stroke without any prior major health issues grouped by General Health.

2. Plot 2 - Stacked Bar Plot/Chart of the count of the persons having Asthma and habits like Smoking, and got a Heart Stroke visualized by grouping according to the Age Category and the percentage of the Race for each "AgeCategory" Category.

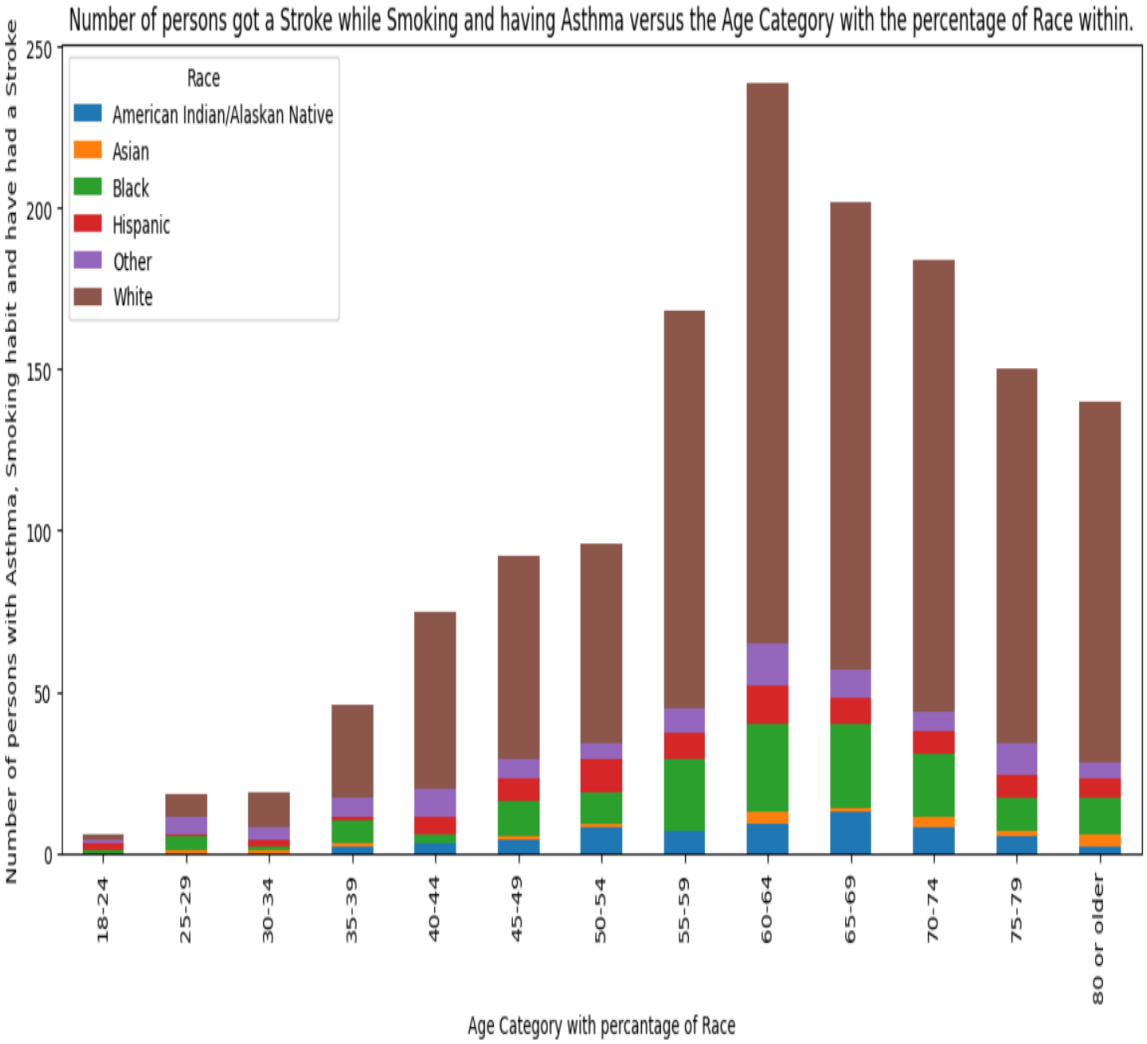


Fig 6: Stacked Bar Plot fir Nuber of persons got a stroke while smoking and having Asthama versus the Age Category with percentage of Race within.

3. Plot 3 - Histograms of the persons who got a Heart Stroke visualized against the BMI index and splitting the data according to Sex field (Male and Female) into two groups which produces two histograms.

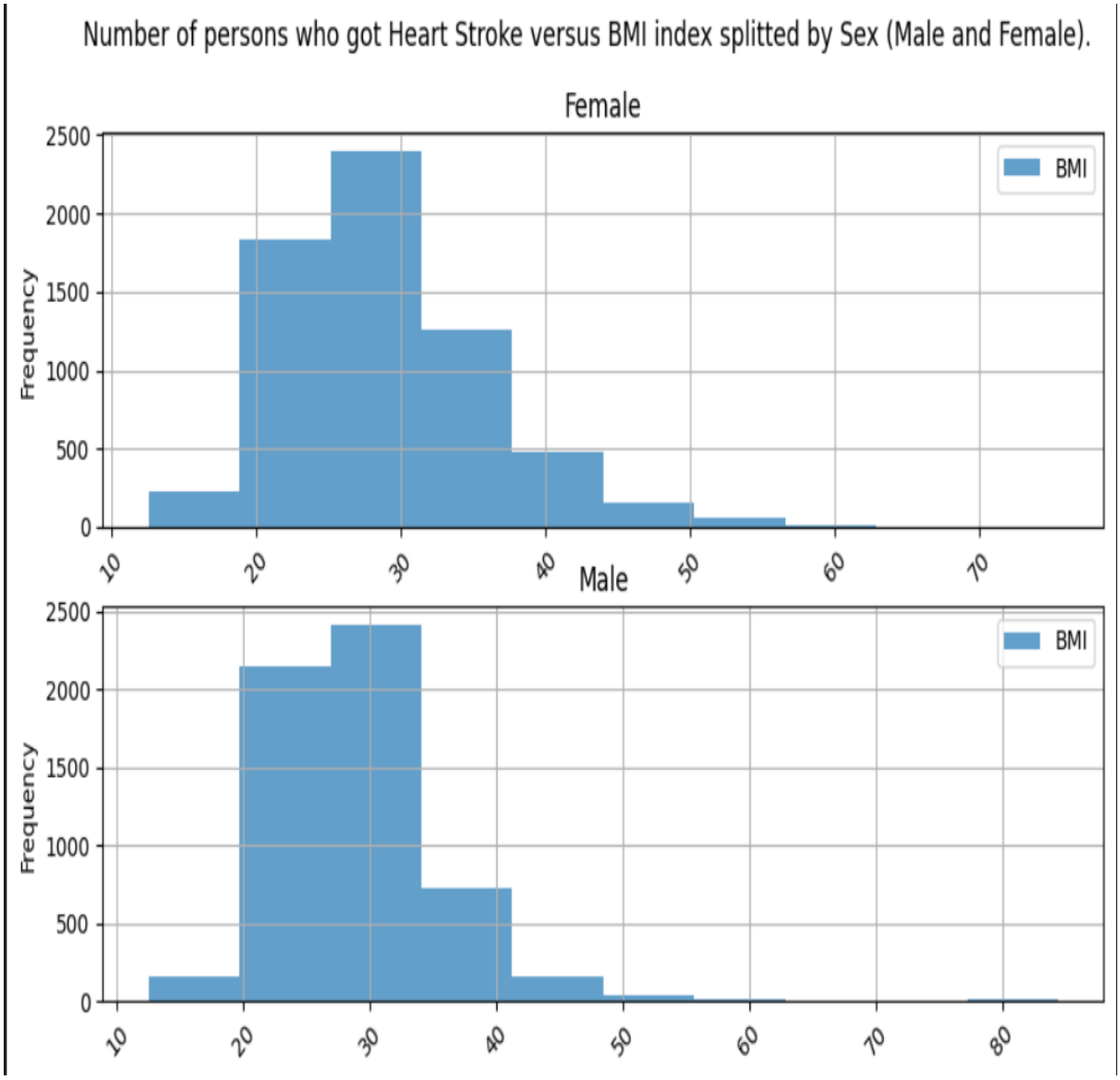


Fig 7: Number of persons who got heart stroke versus BMI Index splitted by Sex(Male and Female).

4. Plot 4 - Grouped Bar Plot/Chart of the count of the persons having previous Heart Stroke or a Heart Disease with Drinking or Smoking habit, and no major diseases visualized by grouping according to the Sex (Male and Female) and the percentage of the Stroke and Heart Disease within.

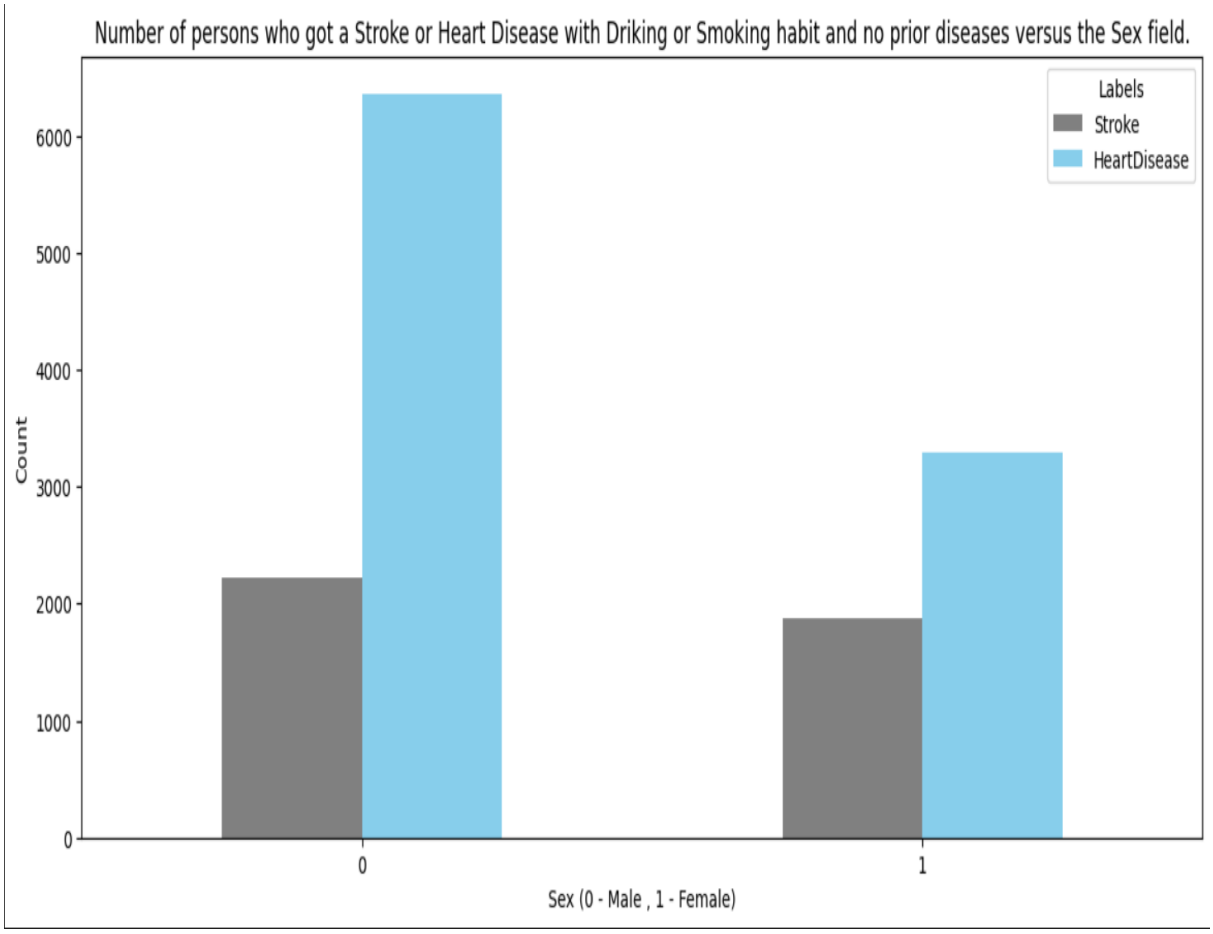


Fig 8: Grouped bar plot for Number of persons who got a stroke or Heart Disease with Drinking or smoking habit and no prior diseases versus the sex field.

5. Plot 5 - Pie Charts/ Plots of the persons have had a Heart Stroke visualized by grouping according to the General Health fields which invloves pregnancy and diabetes in gener.

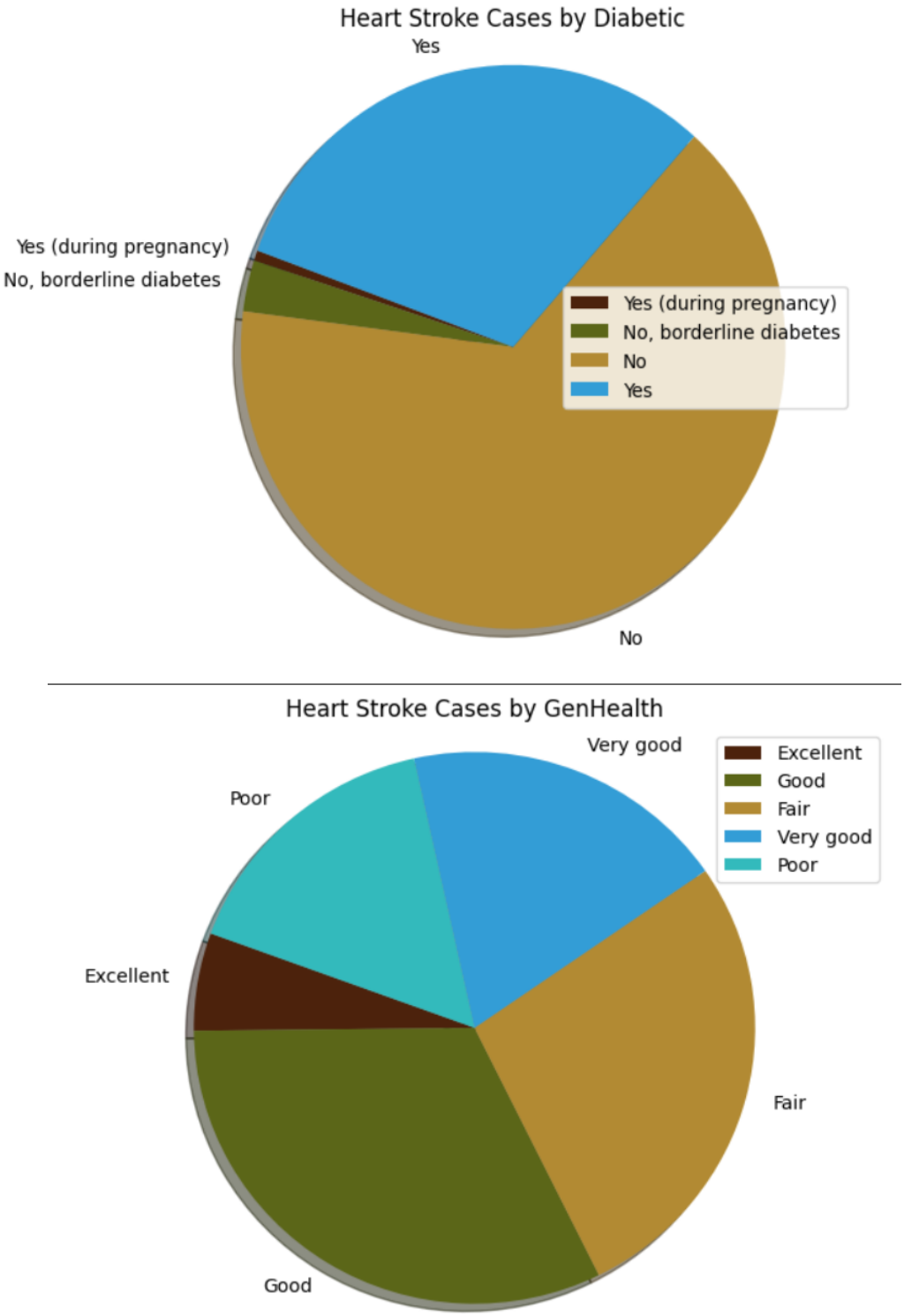
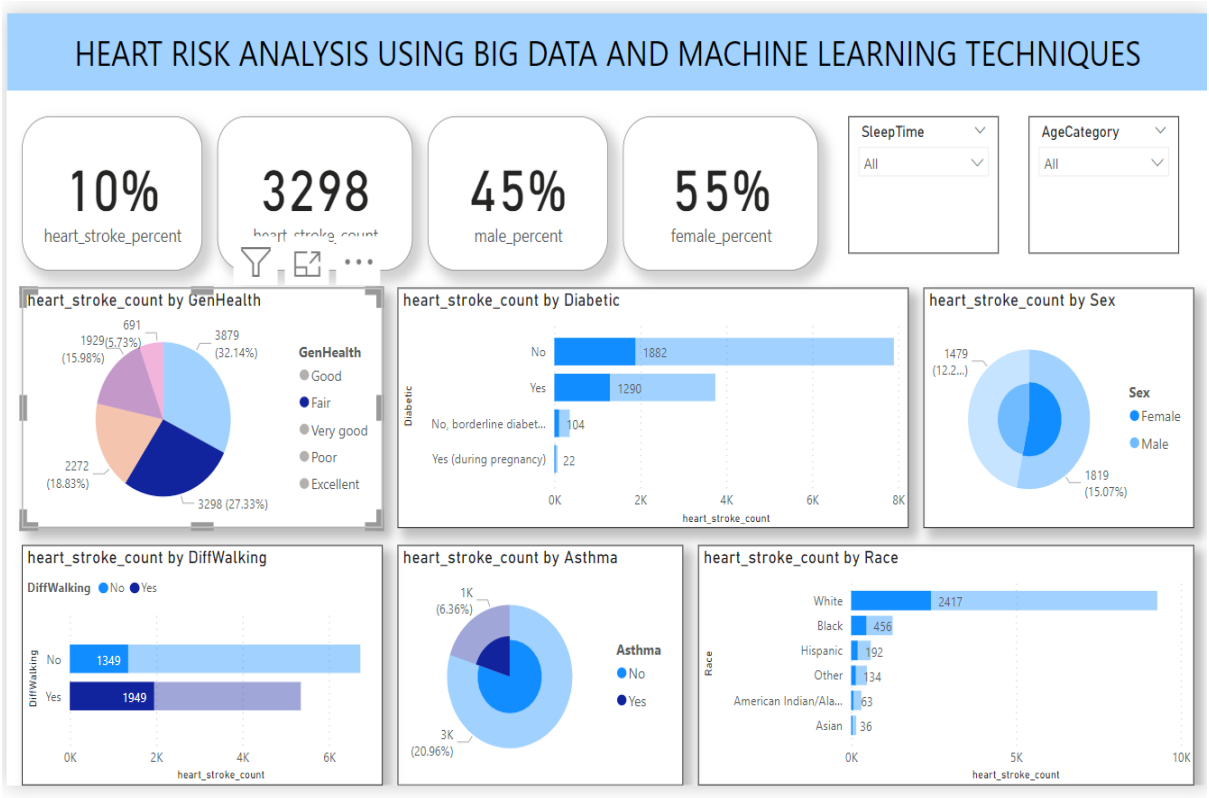
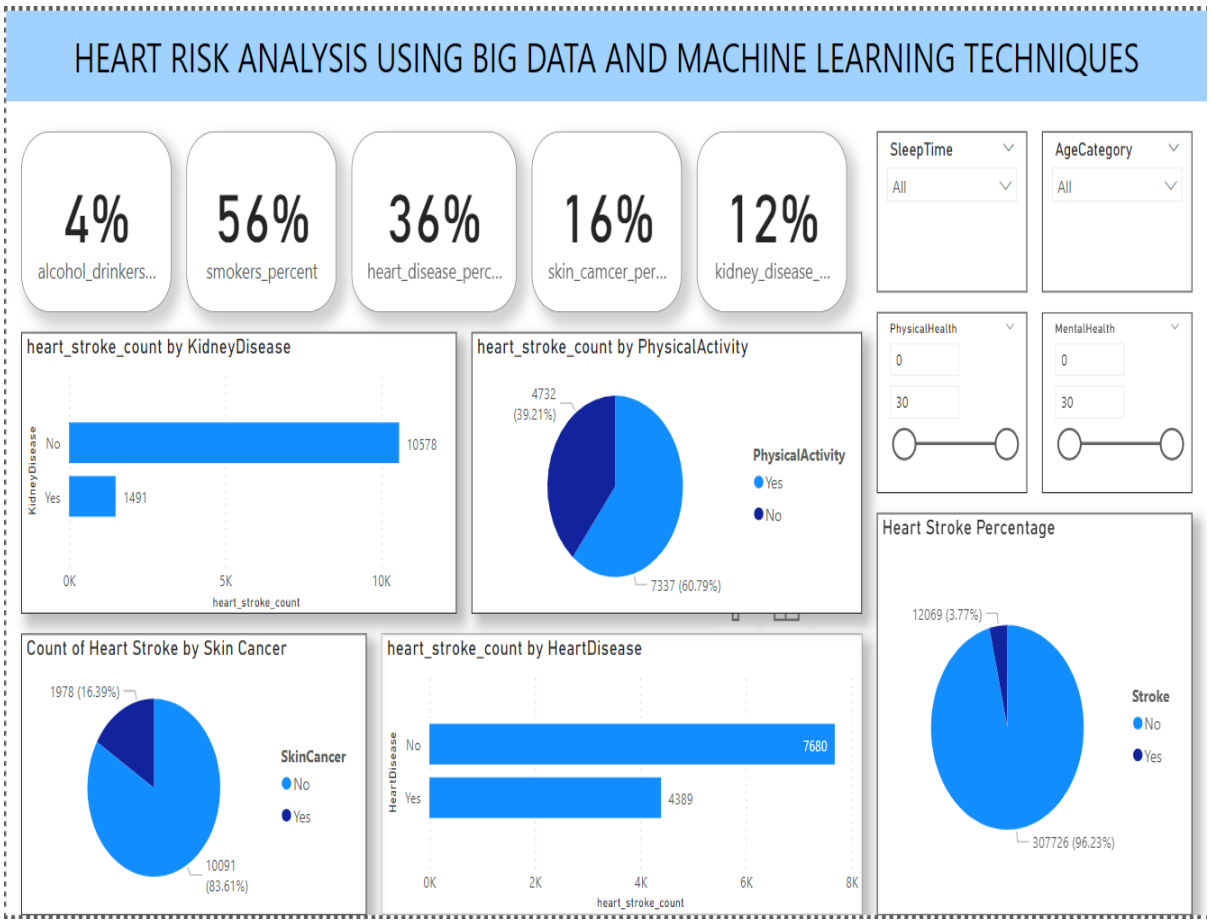


Fig 9: Pie Chart for Heart stroke cases by Diabetic and by GenHealth.

Data Visualization using Power BI:





Alcohols drinkers percentage

```
alcohol_drinkers_count = CALCULATE(COUNT('dataset'[AlcoholDrinking]), FILTER('dataset', 'dataset'[AlcoholDrinking] = "Yes" && 'dataset'[Stroke] = "Yes"))
```

Heart stroke count by Asthma

```
asthma_count = CALCULATE(COUNT('dataset'[Asthma]), FILTER('dataset', 'dataset'[Stroke] = "Yes" && 'dataset'[Asthma] = "Yes"))
```

Female Percentage

```
female_percent =
var num = CALCULATE(COUNT('dataset'[Sex]), FILTER('dataset', 'dataset'[Sex] = "Female" && 'dataset'[Stroke] = "Yes"))
var den = [heart_stroke_count]
var res = (num/den)
return res
```

Heart Disease percentage

```
heart_disease_percent =
var num = CALCULATE(COUNT('dataset'[Smoking]), filter('dataset', 'dataset'[HeartDisease] = "Yes" && 'dataset'[Stroke] = "Yes"))
var den = [heart_stroke_count]
var res = num/den
return res
```

heart stroke percentage by stroke

```
heart_stroke_count = CALCULATE(COUNT('dataset'[Stroke]), FILTER('dataset', 'dataset'[Stroke] = "Yes"))
```

Hear stroke percentage

```
heart_stroke_percent =
var num = [heart_stroke_count]
var den = COUNT('dataset'[Stroke])
var res = num/den
return res
```

Kidney disease percentage

```
kidney_disease_percent =
var num = CALCULATE(COUNT('dataset'[Smoking]), filter('dataset', 'dataset'[KidneyDisease] = "Yes" && 'dataset'[Stroke] = "Yes"))
var den = [heart_stroke_count]
var res = num/den
return res
```

Mae percentage

```
male_percent =
var num = CALCULATE(COUNT('dataset'[Sex]), FILTER('dataset', 'dataset'[Sex] = "Male" && 'dataset'[Stroke] = "Yes"))
var den = [heart_stroke_count]
var res = (num/den)
return res
```

skin cancer percentage

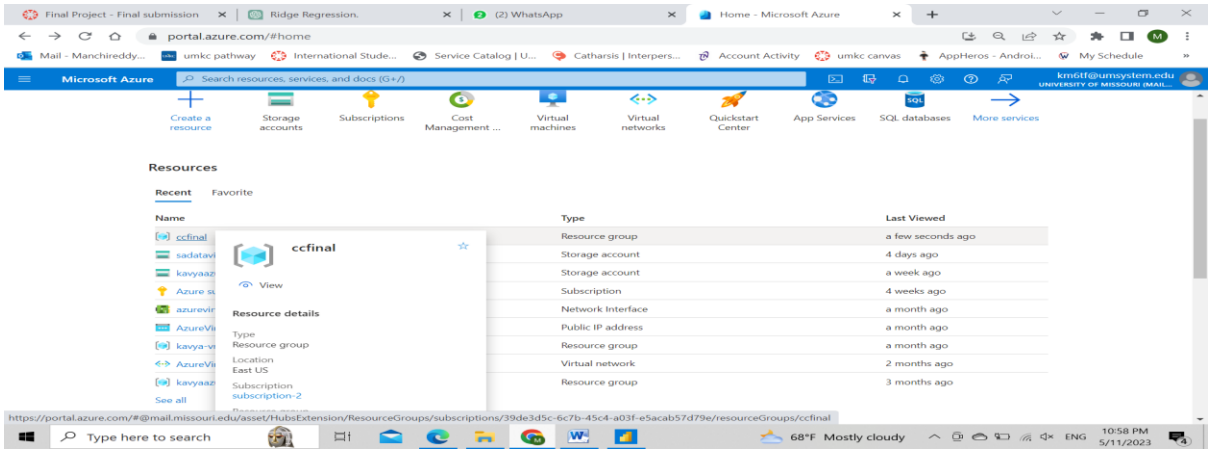
```
skin_camcer_percent =
var num = CALCULATE(COUNT('dataset'[Smoking]), filter('dataset', 'dataset'[SkinCancer] =
"Yes" && 'dataset'[Stroke] = "Yes"))
var den = [heart_stroke_count]
var res = num/den
return res
```

Smokers percentage

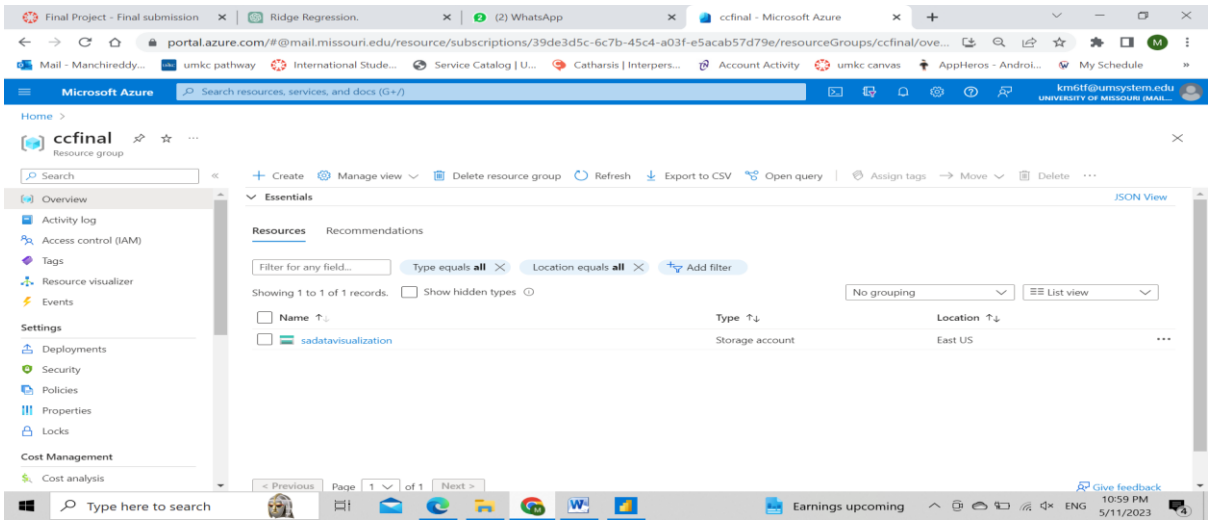
```
smokers_percent =
var num = CALCULATE(COUNT('dataset'[Smoking]), filter('dataset', 'dataset'[Smoking] = "Yes"
&& 'dataset'[Stroke] = "Yes"))
var den = [heart_stroke_count]
var res = num/den
return res
```

6. Steps

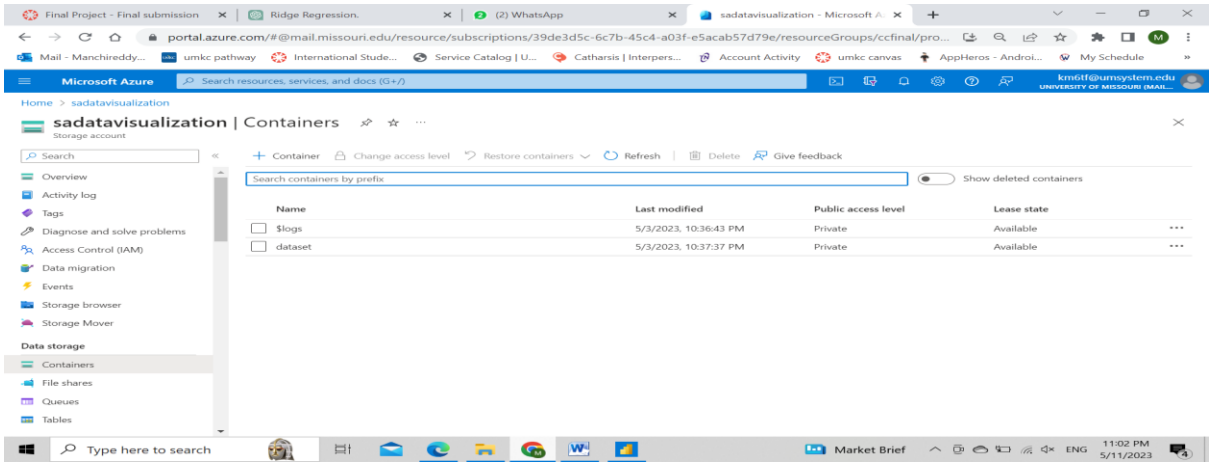
Step 1: created a resource group

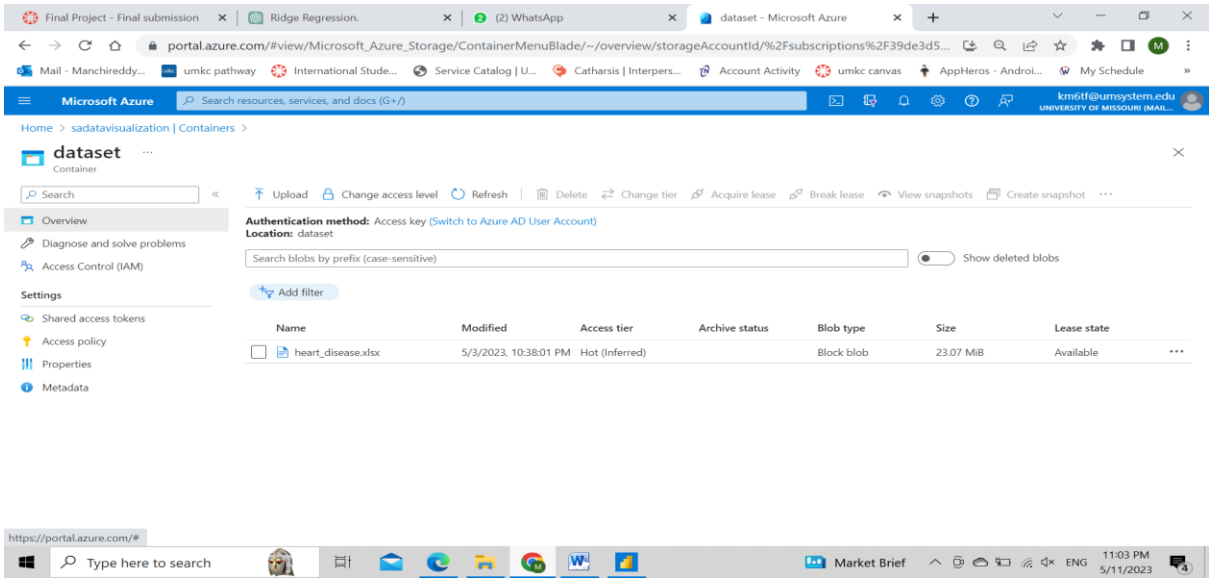


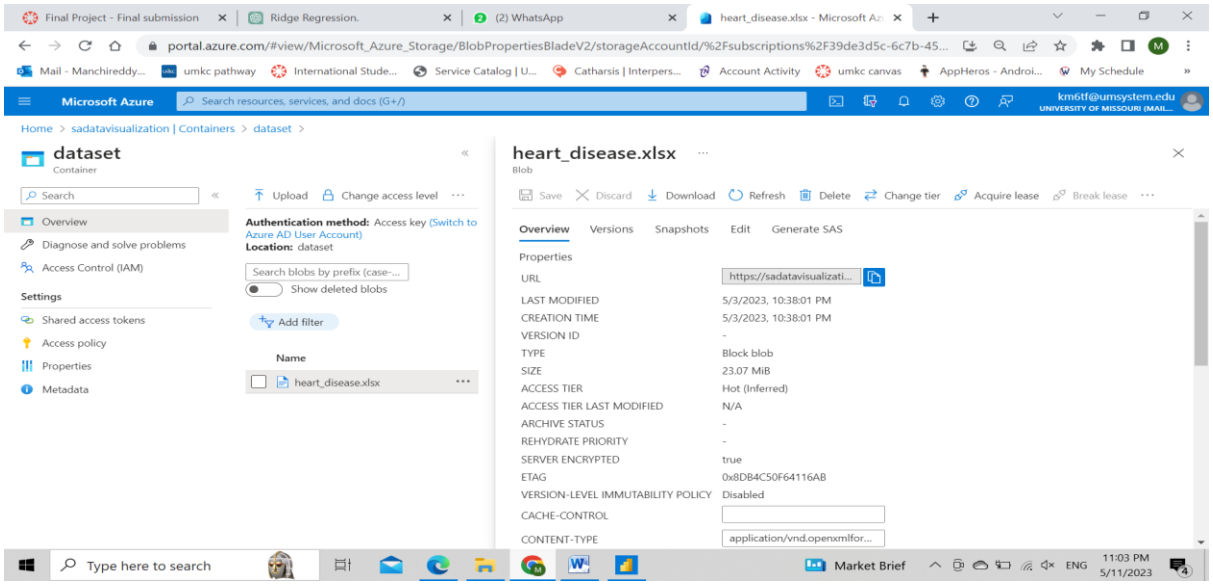
Step 2:created a strorage account in the resource group



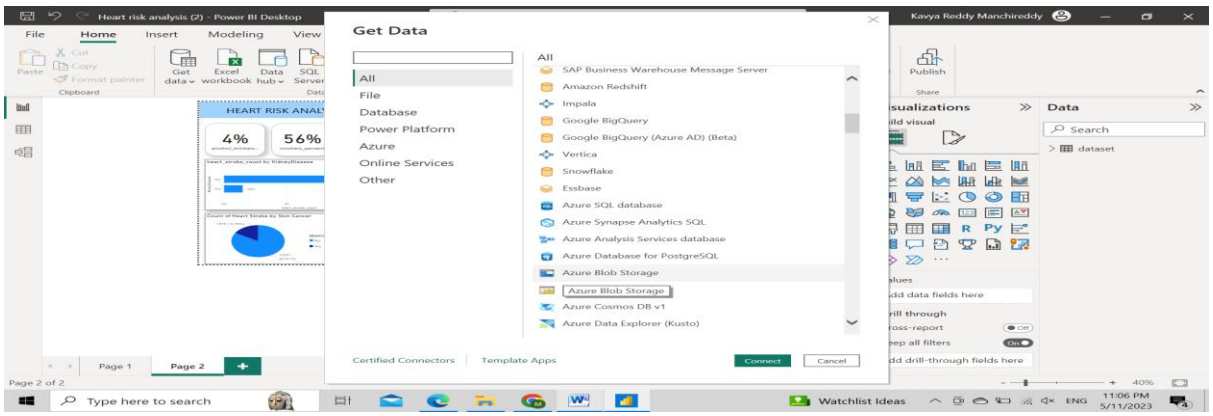
Step 3: uploaded the dataset in the blob storage



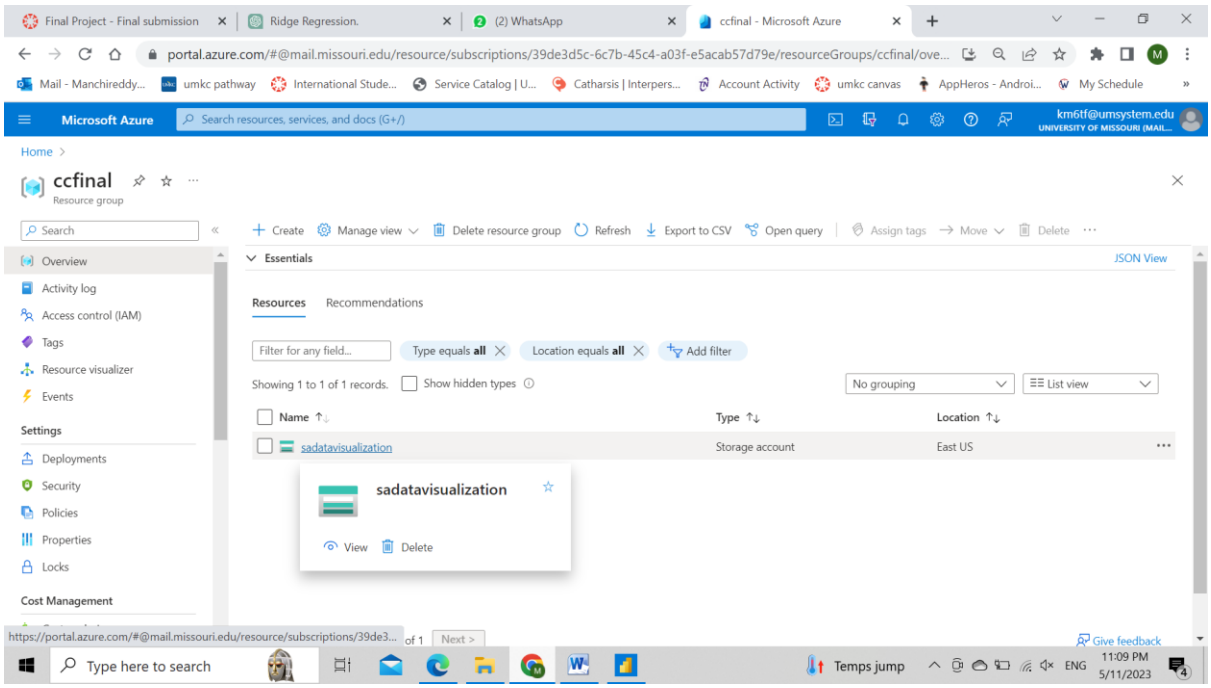




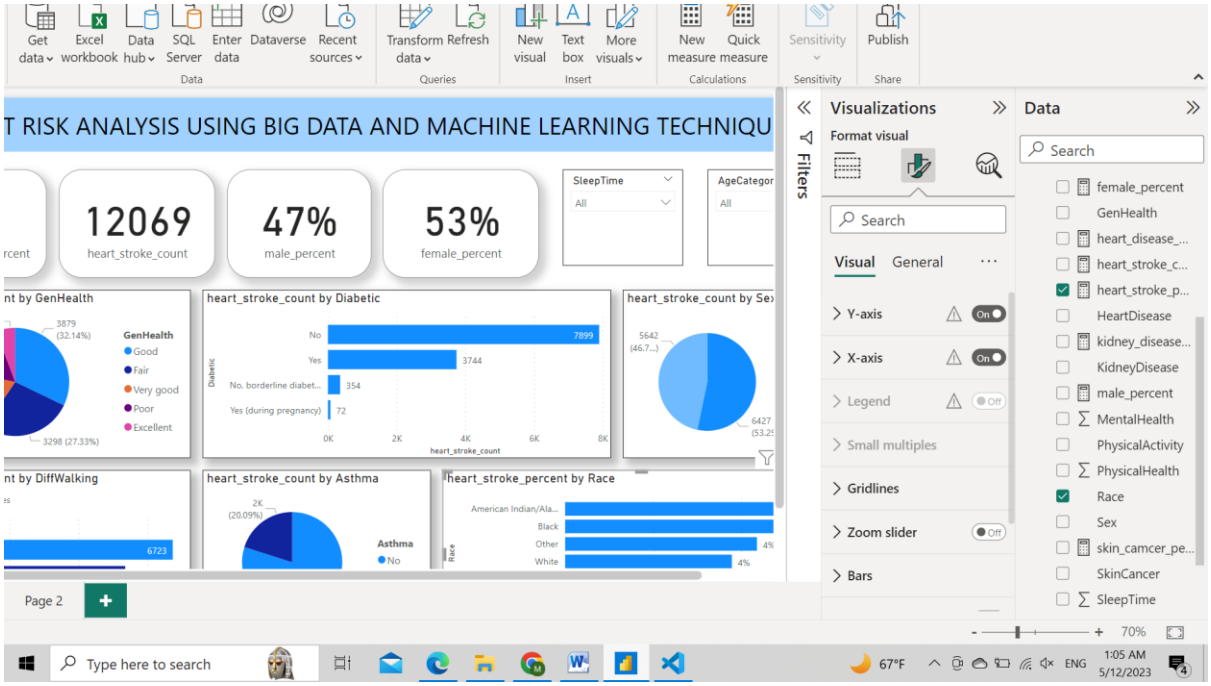
Step 4: install power BI desktop and in the get data select Azure Blob storage and click on connect.



Step 5: Give account name or URL of the blob storage as out blob storage name is sadatavisualization.



Step 6:Build the visualizations by selecting the attributes from the datset



7. Link to the codes

Github Link: https://github.com/manchireddy23/CCFinalProject.git

9. Conclusion

As per the dataset, the prediction is done on the feature Stroke (Heart Stroke), whether or with how much probability does a person can expect a heart stroke in the near future depending on various other effecting features like previous Heart Disease, Kidney Disease, Skin Cancer, Difficulty in walking, etc. On pre-processing the data and with the help of visualizations and pair-wise correlation between all the features and on the Stroke feature, a feature vector is created using feature transformation: ['HeartDisease','DiffWalking','AgeCategory'] as it is passed into machine learning models. Random Forest Classifier and Naive Bayes Classifier are the two machine learning models which have been selected to classify the Stroke feature (0 or 1), as they are the model which perform their best when comes to the classification problems.It can be observed that the accuracy provided by the Random Forest Classifier, i.e. around 85%, whereas the accuracy provided by the Naive Bayes Classifier i.e. around 84%. Both the models performed close to one another, but Random Forest Classifier clearly well suited for the existing problem of Stroke Classification and also shown the visualizations in power BI using various attributes in the dataset.

10. References

1. Heart Disease Dataset (https://www.kaggle.com/datasets/vaisakhnair/heart-disease-data)

2. Stroke Risk Prediction with Machine Learning Techniques (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9268898/)

3. Learn about Stroke. (https://www.world-stroke.org/world-stroke-day-campaign/why-stroke-matters/learn-about-stroke)

4. Elloker T., Rhoda A.J. The relationship between social support and participation in stroke: A systematic review. Afr. J. Disabil. 2018;7:1–9. (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6191741/)

5. Katan M., Luft A. Seminars in Neurology. Volume 38. Thieme Medical Publishers; New York, NY, USA: 2018. Global burden of stroke; pp. 208–211. (https://pubmed.ncbi.nlm.nih.gov/29791947/)