MAKERERE  UNIVERSITY

SEMESTER ONE 2024/2025 ACADEMIC YEAR

SCHOOL OF COMPUTING AND INFORMATICS TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE

MASTER OF SCIENCE IN COMPUTER SCIENCE

MCS 7103

MACHINE LEARNING

ASSIGNMENT ONE

AGABA LUCKY

2024/HD05/21913U

2400721913

TASK;

RIGOROUS EXPLORATORY DATA ANALYSIS OF DATA CONCERNING HEALTHCARE, AND HEALTH POLICY ISSUES AFFECTING AMERICANS AGED 50 AND OLDER UTILIZING PYTHON LIBRARIES INCLUDING PANDAS, MATPLOTLIB, AND SEABORN

Table Contents

Introduction

Healthy aging involves maintaining physical, mental, and emotional well-being as we grow older. Key aspects include regular exercise to keep the body strong and agile, a balanced diet rich in nutrients, and staying mentally active through learning and social engagement. As the Government tries its level best to ensure that its citizens are healthy even as they age, it has tried to facilitate health centres though some people do not know the value of visiting them regularly. A common vice among the Ugandan population in particular is that they prioritize visiting health facilities when they are very ill.

Therefore, it is from this agenda that we can come up with a model that can help predict the number of times aged people visit the health facilities in a year.

This dataset is from the National Poll on Healthy Aging (NPHA) filtered down to develop and validate machine learning algorithms for predicting the number of doctors a survey respondent sees in a year. This dataset's records represent seniors who responded to the NPHA survey.

The dataset is a tabular dataset with the subject area of Health and Medicine associated with classification. It is a definite feature type with 714 instances and 14 features.

My Research Question before;

How can I explore and understand the National Poll on Healthy Aging data to effectively prepare it for machine learning analysis and which problem am I handling in particular?

Answer; I can achieve this by performing Exploratory Data Analysis (EDA) on this dataset and the problem being handled is predicting the number of doctors a survey respondent sees in a year to age or grow old when he or she is healthy.

My Research Questions after;

1. *What are the key features of my dataset, and what types of data are included?*

Answer; The dataset includes features such as age, gender, health conditions, and lifestyle factors.

2. *Are there missing values or data quality issues that need to be addressed?*

Answer; There are no missing values but some issues need to be addressed.

3. *What is my target Variable?*

Answer; The number of doctors visited is my target variable.

4. *How many instances and features do I have and are they really good enough for my dataset?*

Answer; I have 714 instances and 14 features and they are good enough to have the prediction achieved.

Below is a breakdown of the variables that were used;

Variable Name	Role	Type	Description
Number_of_Doctors_Visited	Target	Categorical	The total count of different doctors the patient has seen = { 1: 0-1 doctors 2: 2-3 doctors 3: 4 or more doctors }
Age	Feature	Categorical	The patient's age group = { 1: 50-64 2: 65-80 }
Physical_Health	Feature	Categorical	A self-assessment of the patient's physical well-being = { -1: Refused 1: Excellent 2: Very Good 3: Good 4: Fair 5: Poor }
Mental_Health	Feature	Categorical	A self-evaluation of the patient's mental or psychological health = { -1: Refused 1: Excellent 2: Very Good 3: Good 4: Fair 5: Poor }
Dental_Health	Feature	Categorical	A self-assessment of the patient's oral or dental health= { -1: Refused 1: Excellent 2: Very Good 3: Good 4: Fair 5: Poor }
Employment	Feature	Categorical	The patient's employment status or work-related information = { -1: Refused 1: Working full-time 2: Working part-time 3: Retired 4: Not working at this time }
Stress_Keeps_Patient_from_Sleeping	Feature	Categorical	Whether stress affects the patient's ability to sleep = { 0: No 1: Yes }
Medication_Keeps_Patient_from_Sleeping	Feature	Categorical	Whether medication impacts the patient's sleep = { 0: No 1: Yes }

Pain_Keeps_Patient_from_Sleeping	Feature	Categorical	Whether physical pain disturbs the patient's sleep = { 0: No 1: Yes }
Bathroom_Needs_Keeps_Patient_from_Sleeping	Feature	Categorical	Whether the need to use the bathroom affects the patient's sleep = { 0: No 1: Yes }
Uknown_Keeps_Patient_from_Sleeping	Feature	Categorical	Unidentified factors affecting the patient's sleep = { 0: No 1: Yes }
Trouble_Sleeping	Feature	Categorical	General issues or difficulties the patient faces with sleeping = { 0: No 1: Yes }
Prescription_Sleep_Medication	Feature	Categorical	Information about any sleep medication prescribed to the patient = { -1: Refused 1: Use regularly 2: Use occasionally 3: Do not use }
Race	Feature	Categorical	The patient's racial or ethnic background = { -2: Not asked -1: REFUSED 1: White, Non-Hispanic 2: Black, Non-Hispanic 3: Other, Non-Hispanic 4: Hispanic 5: 2+ Races, Non-Hispanic }
Gender	Feature	Categorical	The gender identity of the patient = { -2: Not asked -1: REFUSED 1: Male 2: Female }

Importing the necessary libraries;

The necessary Python libraries to be used while working with the dataset were imported as follows.

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Reading the dataset;

The dataset was loaded into Google Collab and this was achieved successfully.

```
df = pd.read_csv('/content/NPHA-doctor-visits.csv')
```

Data Wrangling

Performing a sanity check on the data;

Sanity checks on the dataset were performed by finding out the first five and last five rows of my dataset. It was discovered that there were 714 rows and 15 columns in the dataset. The columns and their data types were also identified where all of them had the same datatype integer, and for this, the *info()* method was used.

The number of unique elements in the dataset were checked using the *nunique()* method (*df.nunique()*) and it was discovered that all the elements in the dataset were unique. This was so helpful to decide which type of encoding to choose for converting categorical columns into numerical columns in case of any.

Handling missing values;

Checking for missing values, the *df.isnull().sum*() method was used, and it was realized that there were no missing values in the dataset. Duplicated values were also checked and it was realized that there were 42 duplicated values. These were later dropped using the *df.drop_duplicates()* method which brought the dataset to have 672 rows and 15 columns. This changed from the original dataset which had 714 rows.

The summary of the dataset was equally obtained using the pandas *describe()* method and this helped to apply basic statistical computations on the dataset like extreme values, count of data points, and standard deviation, among others. It was realized that any missing or NaN value is automatically skipped, giving a good picture of the distribution of data.

The garbage values (values whose data type is an object) were not identified in the dataset. Therefore, there were no garbage values in the dataset.

Exploratory Data Analysis (EDA)

Visualizing data relationships;

UNIVARIATE ANALYSIS

During the process of visualizing data, a histogram was used to understand the distribution of data.

Number of doctors visited column

Using a histogram, it was observed that most patients visit 2 – 3 doctors in a year, which is much higher than the frequency of those who visit 0-4 doctors and 4 and above doctors.

The Age column

The distribution of this feature shows that all the patients were of the same age category.

Physical health column

The distribution of this feature showed that patients' physical health status was good, very good, fair, excellent, and poor respectively. It was observed that the highest number of patients were good and very few were poor. The final observation was that the number of excellent patients was very minimal.

Mental health column

It was observed that the highest number of patients had very good mental health. It was also observed that a good number of patients following those that had very good mental health had excellent and good mental health status. It was also observed that there were very few with poor mental health and those who refused to respond to the survey.

Dental health column

The distribution of this feature showed that a high number of patients had very good and good dental health while the others were average. It was also observed that there were very few patients who refused to disclose their dental health status.

Employment column

The distribution of this feature showed that the highest number of patients were already retired. On the other hand, a few patients were full and part-time workers and a very minimal number were not working at all.

Stress Keeps Patient from Sleeping Column

The distribution of this feature showed that the highest number of patients are not affected by stress keeping them from sleeping. However, there is a minimal number that is affected by stress.

Medication keeps patients from sleeping column

From this feature, very few patients are affected by medication. On the other hand, the highest number is not affected by medication.

Pain keeps patients from sleeping column

The distribution of this feature shows that the highest number of patients is not affected by pain however there is also a number though not very many that are affected by pain preventing them from sleeping.

Bathroom needs keep patients from sleeping column

This feature shows that there is an equal number of patients who are affected and those that are not affected.

Unknown keep patients from sleeping column

The distribution of this features shows that a highest number of patients do not have other issues that keep them from sleeping. However, there are those that are affected.

Prescription sleep medication column

The distribution of this feature is skewed to the left and shows that the highest number of patients do not use prescription sleep medication.

Race column

This feature shows that the highest ethnic racial background are the whites.

Gender column

The distribution of this feature shows that the highest number of patients are female.

BIVARIATE ANALYSIS

This demonstrates how each feature relates to the target feature in the analysis. This attempts to ascertain the variables that have a greater influence on the target variable. It also clearly indicates whether there is a direct or inverse link between the two.

In addition to the histogram and boxplot, a scatter plot was also used to understand the relationship between my target variable "Number of doctors visited" and other variables. It was discovered that there is a positive relationship between my target variable and other variables in the dataset.

Number of doctors visited Vs Age

The scatter plot shows that as the age increases to a certain range, the number of doctors visited also increases.

Number of doctors visited Vs physical health

The scatter plot shows that an increase in age for different age categories leads to an increase in the number of doctors visited.

Number of doctors visited Vs mental health

The relationship between these two variables constantly increases. This illustrates that as the number of patients with mental health increases also the number of doctors visited increases.

Number of doctors visited Vs dental health

It stands out that as the number of patients with dental issues increases, the number of doctors visited also increases for the different age groups.

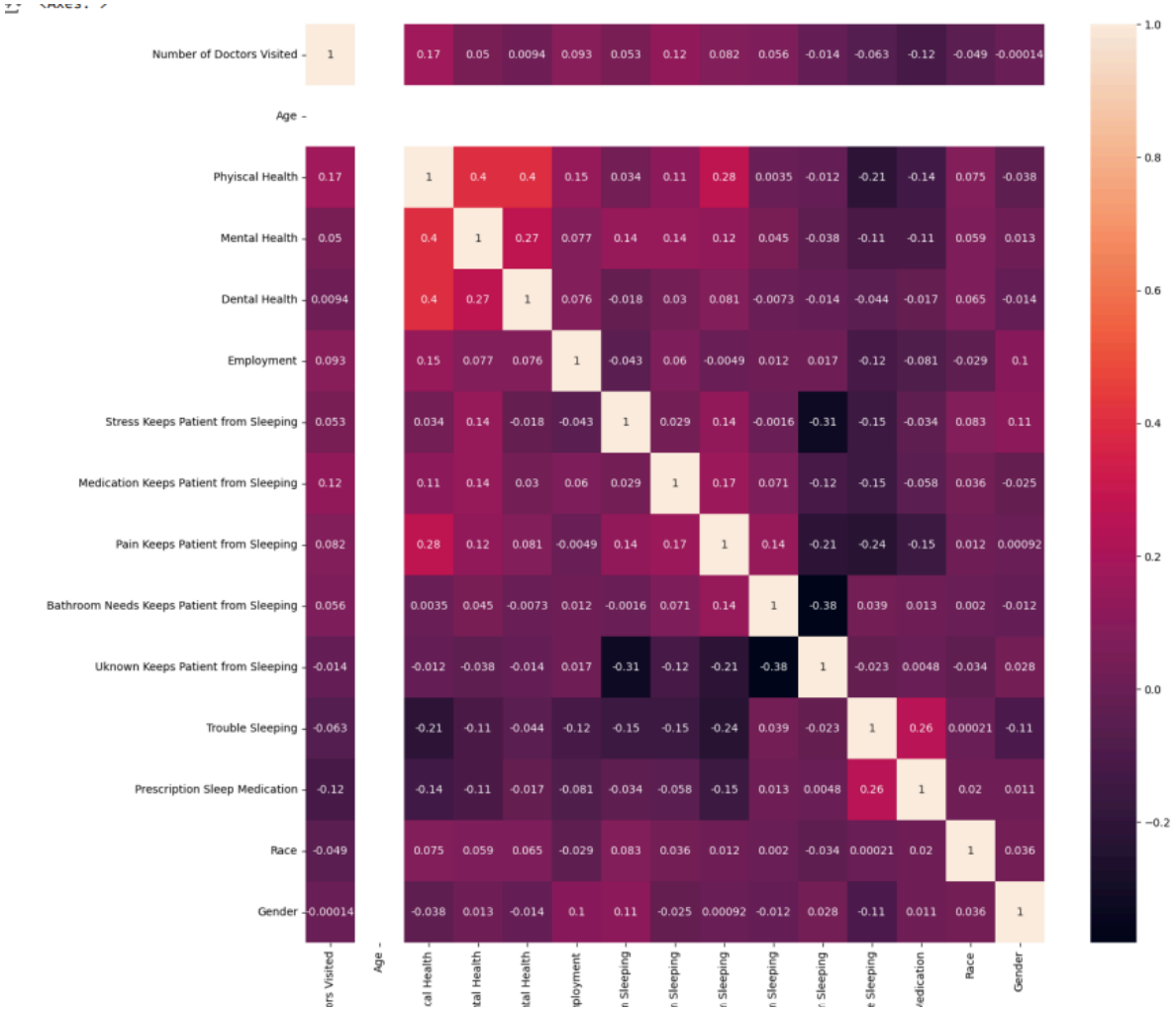
It was noted that the relationship between the number of doctors visited and all the other variables is directly proportional to the target variable. This means that as one increases, the other also increases.

Outliers' treatment

The boxplot was used to identify the outliers in the dataset. As a result of plotting the boxplot, it was realized that the dataset had some outliers. The outliers were basically in physical health, employment, stress keeping patients from sleeping, medication keeping patients from sleeping, pain keeping patients from sleeping, trouble sleeping, prescription sleep medication, and race. These outliers were considered not to have a bigger effect on the dataset after analyzing the relationship between the target variable and other variables which showed a positive correlation.

MULTIVARIATE ANALYSIS

A correlation matrix using a heatmap was constructed to help understand the correlation between the target variable and each of the other control variables. The variables with a correlation greater than 0.3 were considered to have a strong impact on the output variable.



Findings and insights about the dataset;

1. The dataset has no missing values and that makes it suitable to be used for the prediction.
2. The dataset has no duplicated values which makes it suitable for training the model.
3. From the visualization point of view, the data is well distributed and this makes it fit to be used to train the model for the prediction of the targeted variable.
4. The dataset has got few outliers which I have considered to be minor hence not affecting the final prediction after the model has been trained.
5. There is a positive correlation noticed between the target variable and other variables. This therefore makes the dataset fit to be used to train the model to arrive at the final prediction.

Recommendations

I recommend the Ministry of Health with support from the Government to encourage the aged people to at least try as much as possible to visit health facilities so that they can age while they are healthy.

Conclusion

According to the Exploratory Data Analysis, the main variables to consider while determining healthy aging are the number of doctors visited, physical health, mental health, dental health, and employment. It was discovered during visualization that the highest number of patients visit 2-3 doctors in a year. The more the number of doctors visited by a patient would enable one to age healthy due to the self-awareness that is discovered through visiting the doctors. With physical health, mental health and dental health, the more one grows older the more these features also come along. Therefore depending on the number of doctors one visits, it will determine the rate at which these features are controlled or eliminated. When it comes to unemployment, the more one grows older or ages, the more stressful one becomes due to unemployment. Sometimes because they could have been used to associating with friends at the workplace, and at a time when they retire, they no longer have such meetups with friends highly affects them in their aging.

With this dataset, therefore, we shall be able to predict the number of doctors one visits in a year for them to age while they are healthy. It is so interesting to discover how someone can age or grow old when he or she is still healthy.