**A picture containing jewelled headdress, chandelier

Description automatically generated**

STUDENT ID: GAU20508418

DATA SCIENCE

COURSEWORK- 2

DAtaset: COVID-19 SEVERITY CHECKER

name: MEHRAJ HARISH GAUD

**TABLE OF CONTENTS**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SR NO. | | CONTENT | PAGE NO. | |
| 1. | **ABOUT DATASET** | | | **2** |
| 2. | | **HYPOTHESIS** | **2** | |
| 3. | **DATA PRE-PROCESSING MANUALLY** | | | **3** |
| 4. | | **DATA PER-PROCESSING**  **IN PYTHON** | **5** | |
| 5. | **TECHNIQUE USED** | | | **5** |
| 6. | | **REFLECTION** | **13** | |

**ABOUT DATASET:**

In this report, we will discuss the dataset about COVID-19 which covers all the possible signs and symptoms of the disease, the severity of the disease, and many other related factors. According to WHO, there are five main symptoms of COVID-19 fever, dry cough, difficulty in breathing, sore throat, and tiredness. This data set is from the following URL:

<https://www.kaggle.com/datasets/iamhungundji/covid19-symptoms-checker>

and was published by **Bilal Hungund**.

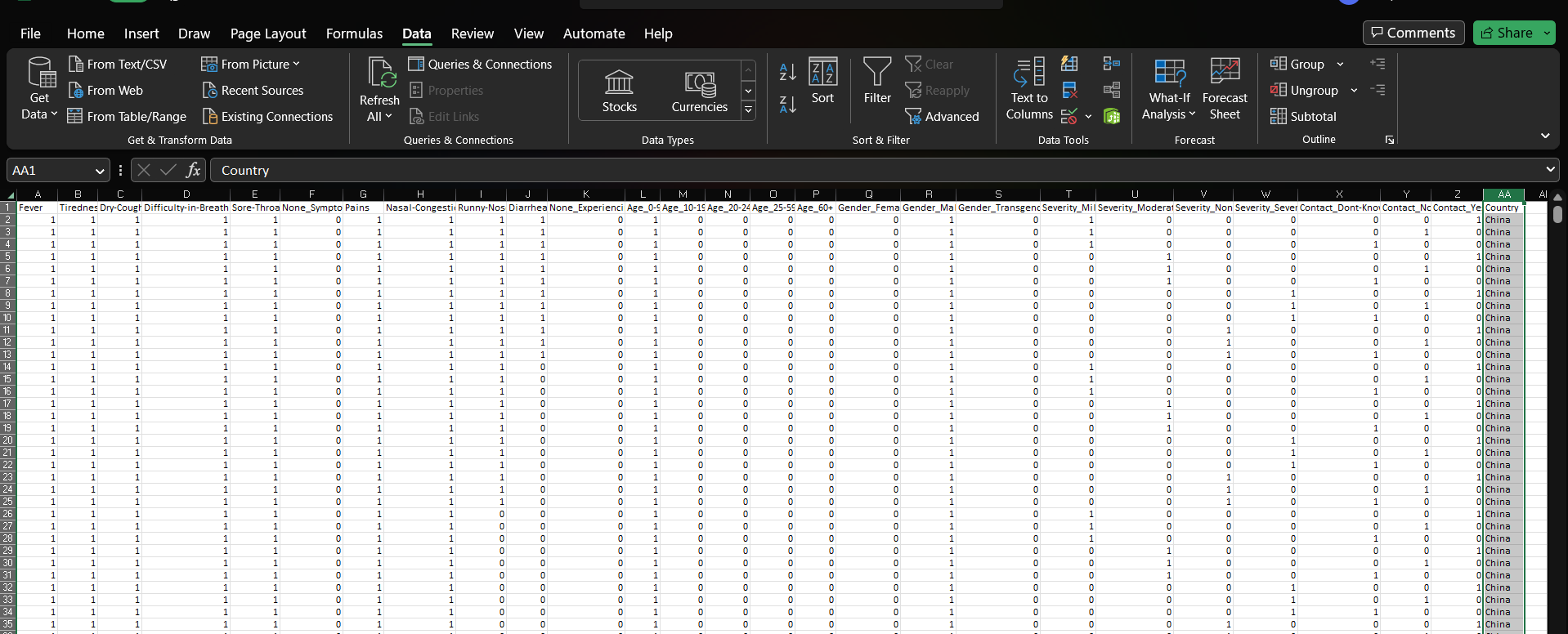
**HYPOTHESIS:**

|  |  |
| --- | --- |
| GOAL | Predict the severity of COVID-19. |
| HYPOTHESIS 1 | Medical experts are trying to find which age and gender are most prone to the severity of this disease and can alert those aged and gender to prevent the spread of the disease and can provide them with regular self-check-up kits. |
| HYPOTHESIS 2 | Medical experts are trying to find the major symptoms which can eventually lead to severe COVID, and therefore can’t be ignored. |
| HYPOTHESIS 3 | Medical experts are also trying to determine whether the virus is spreading through contact so that contact with things and people can be reduced or controlled through different measures like wearing masks, gloves, glasses, sanitizing, etc. |

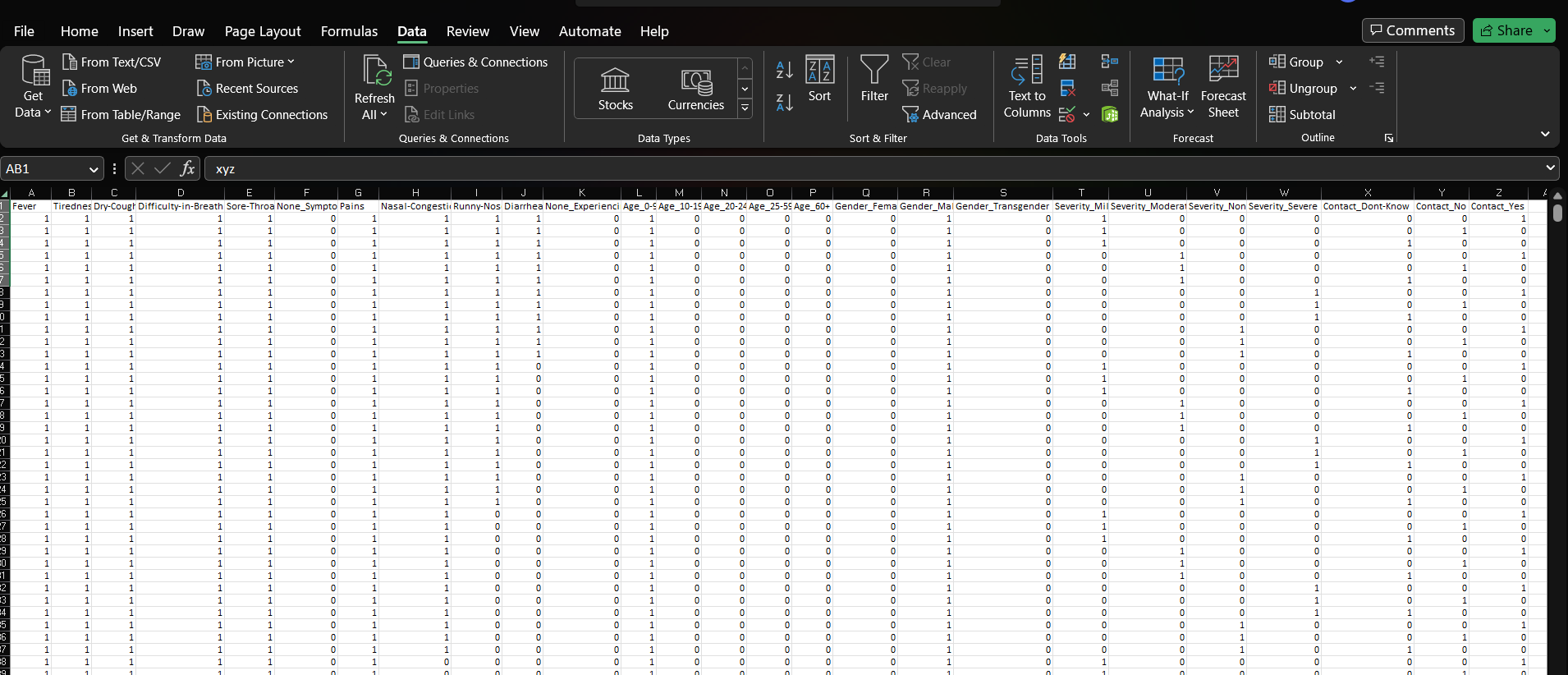
**DATA PRE-PROCESSING MANUALLY:**

In this dataset, we use several pre-processing steps to remove unnecessary data which might not be useful for our analysis or to conclude what we want. So, the following are the steps taken:

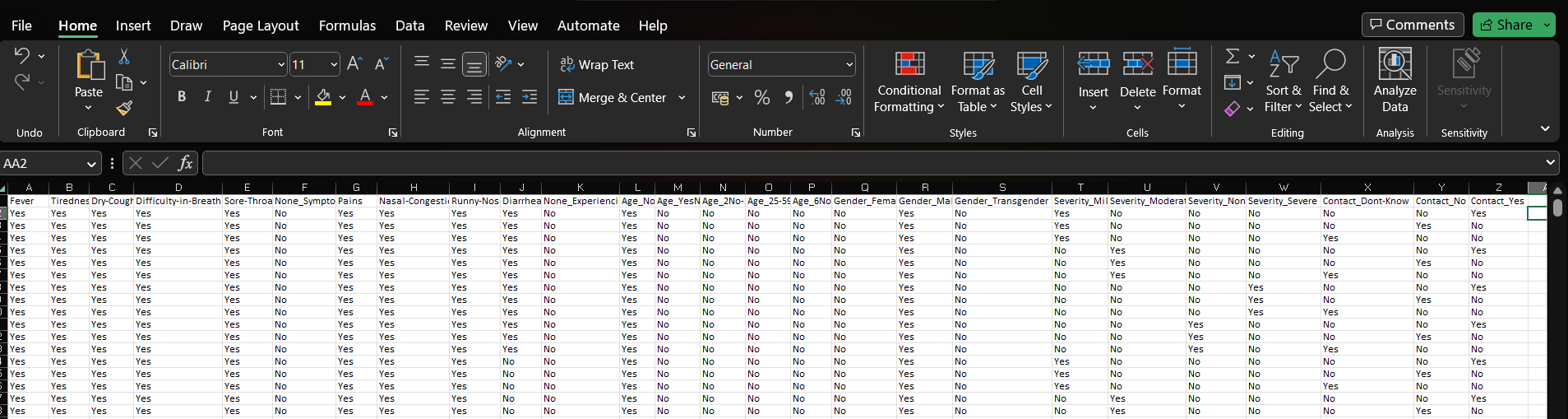
1. I have removed the country column from the dataset as it didn’t help me predict the necessary result.



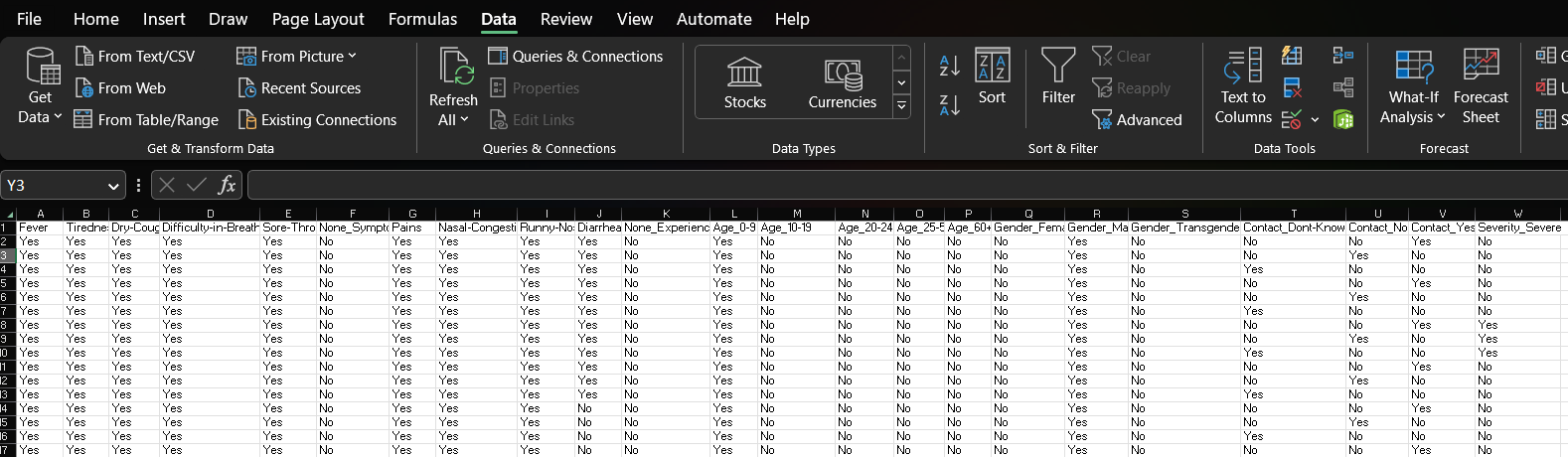
1. I normalized the data by filling up some of the empty remaining columns to make it normal.



1. I have changed all the 1’s to Yes and all the 0’s to No for a better understanding of the user.



1. I have removed the mild, moderate, and no severity columns leaving the severe column only used for prediction as there can only be one labelled column that can be used for prediction.



**DATA PRE-PROCESSING IN PYTHON:**

1. Look for any null values in the dataset:

Table

Description automatically generated

1. Number of rows in each column:

Text

Description automatically generated

**TECHNIQUE -USED AND SOLUTION:**

Diagram

Description automatically generated

To begin the prediction following are the steps:

* Load database into Python using a pandas’ data frame.
* Extract data from the database into a variable.
* Using a label encoder encodes values into a variable.
* Gather them into an array for further prediction.
* Calculate the accuracy of your prediction.
* By giving raw input check if your model is working or not.

**Hypothesis 1:**

In this hypothesis, we are going to compare age group and gender which is prone to severe COVID.

First, we can check which age group is more prone to the severity of the disease:

Shape

Description automatically generated with low confidence

As we can see in the above image all age groups are equally prone to the severity of the disease, therefore all age groups should be aware and alert about it also we can see from the result that not all age groups are at high risk of getting severe COVID-19 which is a good result but not taking care of it might result in worst.

Second, we are going to check the gender which is prone to the severity of the disease:

A picture containing letter

Description automatically generated

In this case, as well all the gender groups are equally vulnerable to severe COVID, but similarly, the percentage of each gender prone to severe COVID is not high but being irresponsible can increase the chances of being severe.

Third, we trained the dataset to check its accuracy and error rate:

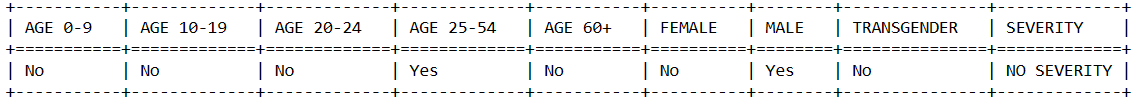
ACCURACY:



ERROR RATE:



Fourth, predicting using external input:



**Hypothesis 2:**

In the second hypothesis, we check what symptoms affect or contribute the highest towards severe COVID. We will divide this part into two categories symptoms and signs, where we will analyse signs and symptoms separately.

First, we will see which symptoms are important to look after and have the highest contribution to the severity of COVID:

Table

Description automatically generated

By, analysing the above table we say that a few symptoms like:

* DRY COUGH (14.0625%)
* TIREDNESS (12.5%)
* DIFFICULTY IN BREATHING (12.5%)

Above are the symptoms which are the major symptoms that cannot be ignored and are the major causes of severe COVID. Other symptoms like FEVERE and SORE THROAT are less in percentage but are also effective and can help build the severity therefore are needed to be cured soon as possible and found any of the major symptoms should be immediately referred to the hospital. Moreover, we can see NONE-SYMPTOMS also contribute to COVID in a minimal percentage but still they exist which can be fatal as a person might have COVID but seem normal.

Second, we will now analyse major signs which lead to severe COVID:

A picture containing table

Description automatically generated

Analyzing the above table, we can say that the following are the major signs:

* NASAL CONGESTION (13.6364%)
* RUNNY NOSE (13.6364%)

Above are the most common signs which are the major symptoms that contribute to severe COVID-19. Furthermore, signs like PAIN and DIARRHEA are the least affecting for severe COVID but can be impactful in the long run. Moreover, NONE-EXPERING has almost no or negligible effect on severe COVID.

Third, we train the dataset to check its accuracy for signs and symptoms separately:

**SYMPTOMS:**

ACCURACY:



ERROR RATE:



**SIGNS:**

ACCURACY:

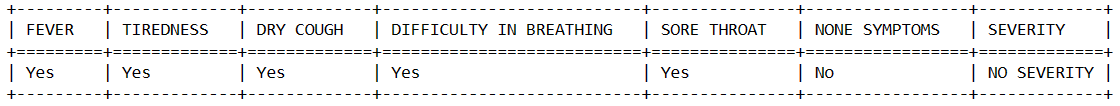


ERROR RATE:



Fourth, predicting using external input:

**SYMPTOMS:**



**SIGNS:**

A picture containing text, antenna

Description automatically generated

**Hypothesis 3:**

Finally, we want to predict whether or not coming in contact with different things and people can lead to severe COVID or can contact be one of the major factors to spread COVID, and if the probability for the spread is higher due to contact, then the government can give out rules and regulations to wear equipment that can avoid contact and stop the spread of the virus and use of appropriate sanitizing liquids to kill the virus.

In the final hypothesis, we will analyse the role of contact in spreading the virus and its severity. To do so we have the following steps:

First, we will analyse the table generated in Python to get a rough idea:

A picture containing table

Description automatically generated

By looking at the above table we can say that even though they all share the same percentage avoiding contact can prevent the spread of the disease massively.

Second, now let’s what is the accuracy of the trained dataset for the same:

ACCURACY:



ERROR RATE:



Third, predicting using external input:

A picture containing text, antenna

Description automatically generated

Lastly, excluding all the hypotheses the overall accuracy of the prediction using all the columns for prediction is:

ACCURACY:



ERROR RATE:



**REFLECTION:**

In conclusion to the report and the coursework, I would present a reflection on my work. In this report, I have a dataset for the COVID symptom checker on which I have performed the Naïve Bayes classifier on it, where using Naïve Bayes probabilistic measurers we can predict the severity of COVID. In this coursework, the most difficult part was to choose the appropriate dataset for prediction and choose the correct technique for it as it will determine all your results for the work.

The pre-processing of the data was the easiest part as the dataset I got was cleaned and was pre-processed already. The implementation part was interesting as I got to learn about many different libraries in Python that are used for such predictions and how to use these libraries to train your dataset for prediction. Working on this coursework also taught me how predictions and these models can be useful for further prediction to stop or control these kinds of viruses which be fatal and cause millions of deaths.

Moreover, the hypothesis I build helped me a lot in the implementation part as it gave me some basis and clues to work with. Also, it helped me to put forward my results with some clarity and proof.

Overall, this coursework was a fun experience where I learned a lot about data science and that data can be used for good causes. Also, python can be a useful tool for analyzing data and careful prediction on all types of data using different techniques.