Data Science Project

Project Team

Muhammad Abdul Nafay (19P-0117)
Muhammad Usman (19P-0116)

Table of Contents

About Dataset

Context

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. Most people infected with COVID-19 virus will experience mild to moderate respiratory illness and recover without requiring special treatment. Older people, and those with underlying medical problems like cardiovascular disease, diabetes, chronic respiratory disease, and cancer are more likely to develop serious illness.

During the entire course of the pandemic, one of the main problems that healthcare providers have faced is the shortage of medical resources and a proper plan to efficiently distribute them. In these tough times, being able to predict what kind of resource an individual might require at the time of being tested positive or even before that will be of immense help to the authorities as they would be able to procure and arrange for the resources necessary to save the life of that patient.

The main goal of this project is to build a machine learning model that, given a Covid-19 patient's current symptom, status, and medical history, will predict whether the patient is in high risk or not.

Content

The dataset was provided by the Mexican government (link). This dataset contains an enormous number of anonymized patient-related information including pre-conditions. The raw dataset consists of 21 unique features and 1,048,576 unique patients. In the Boolean features, 1 means "yes" and 2 means "no". values as 97 and 99 are missing data.

- Sex: female or male
- Age: of the patient.
- Classification: Values 1-3 mean that the patient was diagnosed with covid in different degrees. 4 or higher means that the patient is not a carrier of covid or that the test is inconclusive.
- Patient type: hospitalized or not hospitalized.
- Pneumonia: whether the patient already have air sacs inflammation or not.
- Pregnancy: whether the patient is pregnant or not.
- Diabetes: whether the patient has diabetes or not.
- Copd: Indicates whether the patient has Chronic obstructive pulmonary disease or not.
- Asthma: whether the patient has asthma or not.
- Inmsupr: whether the patient is immunosuppressed or not.
- Hypertension: whether the patient has hypertension or not.
- Cardiovascular: whether the patient has heart or blood vessels related disease.
- Renal chronic: whether the patient has chronic renal disease or not.
- Other disease: whether the patient has other disease or not.
- Obesity: whether the patient is obese or not.
- Tobacco: whether the patient is a tobacco user.
- Usmr: Indicates whether the patient treated medical units of the first, second or third level.
- Medical unit: type of institution of the National Health System that provided the care.
- Intubed: whether the patient was connected to the ventilator.
- Icu: Indicates whether the patient had been admitted to an Intensive Care Unit.
- Death: indicates whether the patient died or recovered.

Note

We have 2 files in csv. One is D1.csv in which most attributes are categorized (For example: Age). This file is used in WEKA (for us to select which machine algorithm to use in python). The other file, D2.csv is mostly used in data visualization as well as in machine learning in python, cause for some reason the data must be in numeric form for the ML algorithm in python to work.

D1.csv

SEX	AGE	PATIENT_TYPE	USMER	DIABETES	COPD	PNEUMONIA	ASTHMA	INMSUPR	HIPERTENSION	CARDIOVASCULAR	OBESITY	RENAL_CHRONIC	TOBACCO	OTHER_DISEASE	CLASIFFICATION
F	E	1	2	N	N	Y	N	N	Y	N	N	N	N	N	Y
M	E	1	2	N	N	Y	N	N	Y	N	Y	Y	N	N	N
M	A	2	2	Y	N	N	N	N	N	N	N	N	N	N	Y
F	A	1	2	N	N	N	N	N	N	N	N	N	N	N	N
M	E	1	2	Y	N	N	N	N	Y	N	N	N	N	N	Y
F	A	2	2	N	N	Y	N	N	N	N	N	N	N	N	Y
F	A	1	2	N	N	N	N	N	N	N	N	N	N	N	Y
F	A	1	2	Y	N	Y	N	Y	Y	N	N	Y	N	N	Y
F	A	2	2	Y	N	N	N	N	Y	N	Y	N	N	N	Y
F	A	2	2	N	N	N	N	N	N	N	N	N	N	N	Y
F	A	1	2	N	N	N	N	N	N	N	N	N	N	N	Y
M	A	2	2	N	N	N	N	N	N	N	N	N	N	N	Y
M	A	2	2	N	N	N	N	N	N	N	N	N	N	N	Y
M	A	1	2	N	N	N	N	N	N	N	N	N	N	N	Y
F	A	1	2	Y	N	N	N	N	N	N	N	N	N	N	Y
F	A	1	2	N	N	N	N	N	N	N	N	N	N	N	Y
F	E	2	2	N	N	Y	N	N	Y	N	N	N	N	N	Y
M	A	1	2	N	N	N	N	N	N	N	N	N	N	N	Y
M	A	1	2	N	N	N	N	N	N	N	N	N	N	N	Y

D2.csv

SEX	AGE	PATIENT_TYPE	USMER	DIABETES	COPD	PNEUMONIA	ASTHMA	INMSUPR	HIPERTENSION	CARDIOVASCULAR	OBESITY	RENAL_CHRONIC	TOBACCO	OTHER_DISEASE	CLASIFFICATION
1	3	1	2	2	2	1	2	2	1	2	2	2	2	2	Y
2	3	1	2	2	2	1	2	2	1	2	1	1	2	2	N
2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	Y
1	2	1	2	2	2	2	2	2	2	2	2	2	2	2	N
2	3	1	2	1	2	2	2	2	1	2	2	2	2	2	Y
1	2	2	2	2	2	1	2	2	2	2	2	2	2	2	Y
1	2	1	2	2	2	2	2	2	2	2	2	2	2	2	Y
1	2	1	2	1	2	1	2	1	1	2	2	1	2	2	Y
1	2	2	2	1	2	2	2	2	1	2	1	2	2	2	Y
1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	Y
1	2	1	2	2	2	2	2	2	2	2	2	2	2	2	Y
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	Y
2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	Y
2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	Y
1	2	1	2	1	2	2	2	2	2	2	2	2	2	2	Y
1	2	1	2	2	2	2	2	2	2	2	2	2	2	2	Y
1	3	2	2	2	2	1	2	2	1	2	2	2	2	2	Y
2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	Y
2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	Y

Part 1

This part primarily focuses on classifying whether a person has COVID or not, based on the features given of the patient. First, By using value-counts method and by data visualization, we checked how many unique values are they're of each attribute with its frequency. By this we had a clear idea of which attribute to remove and which to keep. The table below shows which attributes we removed along with its reasoning.

Attribute	Description on why it is removed
ICU	Missing values over 850,000, if this is filtered out, majority of the data is lost.
DATE_DIED	Irrelevant in our case, also we can't make this data in intervals (discrete).
PREGNANT	Missing values over 550,000. Also, if we remove this, majority of the females will be left in the whole data.
MEDICAL_UNIT	Irrelevant in our case.
INTUBED	Missing values over 850,000, if this is filtered out, majority of the data is lost.

Data Preprocessing

The following are the changes done:

1) It is common in demography to split the population into three broad age groups: children and young adolescents (under 15 years old) the working-age population (15-64 years) and. the elderly population (65 years and older)
   - Young Adolescents/Childrens = 1 (in D2.csv) OR C (in D1.csv)
   - Working-Age Populations = 2 (in D2.csv) OR A (in D1.csv)
   - Elderly Populations = 3 (in D2.csv) OR E (in D1.csv)
2) Reordering of columns
3) Encoding of data such that all 1's are Y (Yes) and 2's are N (No) in *D1.csv only*
4) Removing each record where ever the corresponding attribute had a value of 98 or 99

```
raw_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1048575 entries, 0 to 1048574
Data columns (total 21 columns):
 #   Column                Non-Null Count    Dtype
---  ------                --------------    -----
 0   USMER                 1048575 non-null  int64
 1   MEDICAL_UNIT          1048575 non-null  int64
 2   SEX                   1048575 non-null  int64
 3   PATIENT_TYPE          1048575 non-null  int64
 4   DATE_DIED             1048575 non-null  object
 5   INTUBED               1048575 non-null  int64
 6   PNEUMONIA             1048575 non-null  int64
 7   AGE                   1048575 non-null  int64
 8   PREGNANT              1048575 non-null  int64
 9   DIABETES              1048575 non-null  int64
 10  COPD                  1048575 non-null  int64
 11  ASTHMA                1048575 non-null  int64
 12  INMSUPR               1048575 non-null  int64
 13  HIPERTENSION          1048575 non-null  int64
 14  OTHER_DISEASE         1048575 non-null  int64
 15  CARDIOVASCULAR        1048575 non-null  int64
 16  OBESITY               1048575 non-null  int64
 17  RENAL_CHRONIC         1048575 non-null  int64
 18  TOBACCO               1048575 non-null  int64
 19  CLASIFFICATION_FINAL  1048575 non-null  int64
 20  ICU                   1048575 non-null  int64
dtypes: int64(20), object(1)
memory usage: 168.0+ MB
```

D1.csv

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1025152 entries, 0 to 1048574
Data columns (total 16 columns):
 #   Column          Non-Null Count    Dtype
---  ------          --------------    -----
 0   SEX             1025152 non-null  object
 1   AGE             1025152 non-null  category
 2   PATIENT_TYPE    1025152 non-null  int64
 3   USMER           1025152 non-null  int64
 4   DIABETES        1025152 non-null  object
 5   COPD            1025152 non-null  object
 6   PNEUMONIA       1025152 non-null  object
 7   ASTHMA          1025152 non-null  object
 8   INMSUPR         1025152 non-null  object
 9   HIPERTENSION    1025152 non-null  object
 10  CARDIOVASCULAR  1025152 non-null  object
 11  OBESITY         1025152 non-null  object
 12  RENAL_CHRONIC   1025152 non-null  object
 13  TOBACCO         1025152 non-null  object
 14  OTHER_DISEASE   1025152 non-null  object
 15  CLASIFFICATION  1025152 non-null  category
dtypes: category(2), int64(2), object(12)
memory usage: 119.3+ MB
```

D2.csv

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1025152 entries, 0 to 1048574
Data columns (total 16 columns):
 #   Column          Non-Null Count    Dtype
---  ------          --------------    -----
 0   SEX             1025152 non-null  int64
 1   AGE             1025152 non-null  category
 2   PATIENT_TYPE    1025152 non-null  int64
 3   USMER           1025152 non-null  int64
 4   DIABETES        1025152 non-null  int64
 5   COPD            1025152 non-null  int64
 6   PNEUMONIA       1025152 non-null  int64
 7   ASTHMA          1025152 non-null  int64
 8   INMSUPR         1025152 non-null  int64
 9   HIPERTENSION    1025152 non-null  int64
 10  CARDIOVASCULAR  1025152 non-null  int64
 11  OBESITY         1025152 non-null  int64
 12  RENAL_CHRONIC   1025152 non-null  int64
 13  TOBACCO         1025152 non-null  int64
 14  OTHER_DISEASE   1025152 non-null  int64
 15  CLASIFFICATION  1025152 non-null  category
dtypes: category(2), int64(14)
memory usage: 119.3 MB
```

Data Visualization

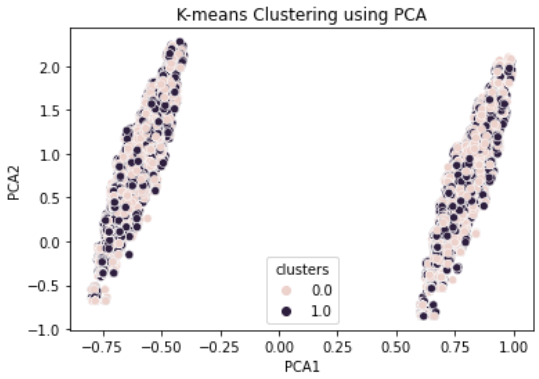
We have done data visualization of both raw data as well as clean data. This is entirely shown in the notebook.

Machine Learning

We have applied the following algorithms

ML Algo	Accuracy
Naïve-Bayes	64%
Logistic Regression	65%
K-Means Clustering using PCA	-
Decision Trees Classification	64%
Random Forest Trees Classification	66%

K-Means Clusters Visualization



The accuracy isn't great. But it is due to the fact of data impurity. There was very low correlation between the features and the class attribute which in our case was determining whether a patient has COVID or not.

```
USMER                0.028840
MEDICAL_UNIT         0.079981
SEX                 -0.057782
PATIENT_TYPE        -0.183370
INTUBED              0.193075
PNEUMONIA            0.075351
AGE                 -0.152637
PREGNANT            -0.057809
DIABETES            -0.004739
COPD                -0.010336
ASTHMA              -0.011178
INMSUPR             -0.009412
HIPERTENSION        -0.006020
OTHER_DISEASE       -0.011143
CARDIOVASCULAR      -0.012143
OBESITY             -0.006924
RENAL_CHRONIC       -0.011342
TOBACCO             -0.012567
CLASIFFICATION_FINAL 1.000000
ICU                  0.193163
Name: CLASIFFICATION_FINAL, dtype: float64
```

Part 2

Since there was less correlation between the features and the class attribute previously. This time we will focus on classifying the attribute “DEATH” which comes from DATE_DIED.

Data Preprocessing

If we have a date that is "9999-99-99", then that means this patient is alive, and vice versa.

1) We have some features that we expect them to have just 2 unique values but we see that these features have 3 or 4 unique values. For example, the feature "PNEUMONIA" has 3 unique values (1,2,99) 99 represents NaN values. Hence, we will just take the rows that includes 1 and 2 values.

2) In "DATE_DIED" column, we have 971633 "9999-99-99" values which represent alive patients so I will take this feature as a "DEATH" that includes whether the patient died or not.

In "INTUBED" and "ICU" features there are too many missing values so i will drop them. Also, we don't need "DATE_DIED" column anymore because we used this feature as a "DEATH" feature.

We have just one numeric feature which is called "AGE" the rest of them are categorical.

The following columns are dropped b/c they have very low correlation with “DEATH”.

- "SEX"
- "PREGNANT"
- "COPD"
- "ASTHMA"
- "INMSUPR"
- "OTHER_DISEASE"
- "CARDIOVASCULAR"
- "OBESITY"
- "TOBACCO"

We used “RobustScaler” from scikitlearn library to scale the attribute “AGE”. “RobustScaler” Scales the features using statistics that are robust to outliers.This Scaler removes the median and scales the data according to the quantile range.

(For more info, visit: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.RobustScaler.html)

Data Visualization

All the visualization is done in the notebook

Machine Learning Algorithm

The table below shows the ML algorithm that is applied and its corresponding accuracy

ML Algo	Accuracy
Logistic Regression	93%
Naïve-Bayes	91%

Since the correlation between the features and the class attribute were high. The accuracy comes out great as well!