Covid-19 severity prediction

ML phase II report

Moneer Zaki
Computer Engineering
The American University in Cairo
moneerzaki@aucegypt.edu

Mario Ghaly
Computer Engineering
The American University in Cario
mariomamdouh@aucegypt.edu

ABSTRACT

In this phase of our machine learning project, we conducted a comprehensive analysis of our chosen dataset to prepare it for model training. Our dataset comprises approximately 1,000,000 rows and 20 features, primarily consisting of boolean values. We examined each feature individually, considering its correlation with COVID severity and its relevance to our predictive model. We encountered several preprocessing challenges, including handling null values, inconsistencies in data representation, and addressing incomparable values. Through a systematic approach, we addressed these issues by removing irrelevant columns, converting certain attributes to boolean values, and predicting missing values using similarity-based techniques. Our final dataset, consisting of 18 relevant features, underwent thorough cleaning and preprocessing, resulting in a dataset free of null values and accurately representing real-world scenarios. Metrics post-cleaning revealed data completeness and accuracy, with all columns exhibiting boolean values except for two. This meticulous process ensures the readiness of our dataset for model training, paving the way for accurate predictions of COVID severity.

Chosen Dataset, its Description, and Rationale for choosing it.

The dataset we've selected is comprehensive, with 21 columns and 1,048,575 rows, making it suitable for training our model. Most of the data consists of Boolean values (1, 2), simplifying its use in the model. However, before proceeding, we need to address some preprocessing issues.

First, we will generally examine all columns for a better understanding of why it is of good correlation with covid severity:

"**USMR**" denotes the level of care (first, second, or third) the patient received, while "**medical unit**" specifies the type of institution within the National Health System providing care, offering complementary insights into the healthcare services received for example, hospital, clinic, or medical center. There are 13 possible values for this feature,

"**INTUBED**" indicates the type of treatment administered to the patient in the medical unit. In other words, whether the patient has been connected to a ventilator or not. And "**ICU**" indicates whether the patient was admitted to an Intensive Care Unit.

"**patient type**" values are 1 for patient returned home and 2 for hospitalization. Also, there is "**pregnancy**" feature to indicate if the patient is pregnant.

"**Age**," "**Sex**," and "**Date Died**" serve as parameters for assessing the likelihood of COVID severity. Older age is associated with higher severity rates, while the correlation between sex and severity requires further investigation through a correlational model.

Then, there are **11 diseases—pneumonia**, diabetes, COPD, asthma, immunosuppression, hypertension, cardiovascular issues, obesity, chronic kidney disease, and tobacco use—that are expected to correlate with COVID severity, and a final Boolean column to reflect if the patient has another disease not in those. Those 11 features indicate whether the patient has/had or doesn't have this disease For all the binary features in general: 1 indicates "yes" and 2 indicates "no".

Other reasons why this dataset was chosen are because there are not so many NULL values in different features, which will be shown in following tables. Furthermore, features have meaningful correlations with covid severity as these diseases are known for being related to how severe the coronavirus is on the patient, and this is also shown in the following tables.

The correlation table provided will aid in understanding the relationships between these variables. Further parameters will be discussed subsequently.

Full Analysis of Dataset Features

As we discussed part of the analysis we have done before, the data is very related to what we need the model to predict at the end.

Moreover, we have done more descriptive analysis on the initial dataset without any modifications, as shown in the following screenshots. All data features are of type int64.

Statistical Analysis

	USMER	MEDICAL_UNIT	SEX	PATIENT_TYPE	INTUBED
count	1.048575e+06	1.048575e+06	1.048575e+06	1.048575e+06	1.048575e+06
mean	1.632194e+00	8.980565e+00	1.499259e+00	1.190765e+00	7.952288e+01
std	4.822084e-01	3.723278e+00	4.999997e-01	3.929041e-01	3.686889e+01
min	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
25%	1.000000e+00	4.000000e+00	1.000000e+00	1.000000e+00	9.700000e+01
50%	2.000000e+00	1.200000e+01	1.000000e+00	1.000000e+00	9.700000e+01
75%	2.000000e+00	1.200000e+01	2.000000e+00	1.000000e+00	9.700000e+01
max	2.000000e+00	1.300000e+01	2.000000e+00	2.000000e+00	9.900000e+01

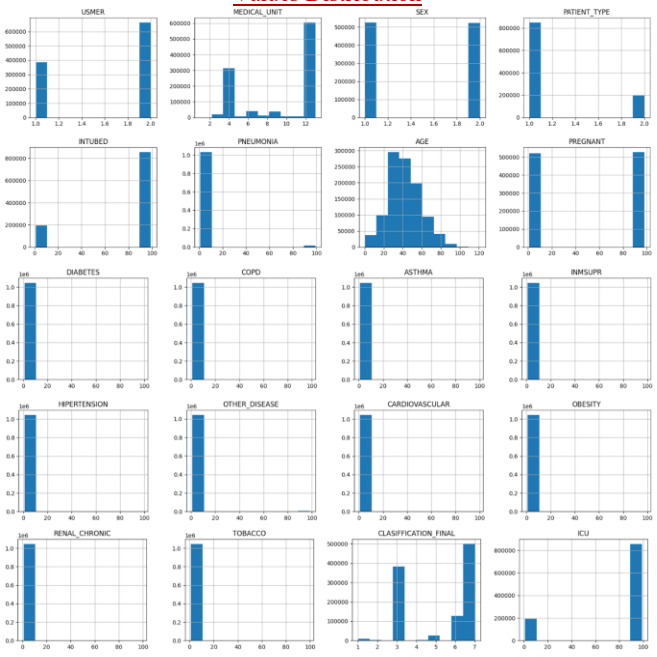
	PNEUMONIA	AGE	PREGNANT	DIABETES	COPD
count	1.048575e+06	1.048575e+06	1.048575e+06	1.048575e+06	1.048575e+06
mean	3.346831e+00	4.179410e+01	4.976558e+01	2.186404e+00	2.260569e+00
std	1.191288e+01	1.690739e+01	4.751073e+01	5.424242e+00	5.132258e+00
min	1.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
25%	2.000000e+00	3.000000e+01	2.000000e+00	2.000000e+00	2.000000e+00
50%	2.000000e+00	4.000000e+01	9.700000e+01	2.000000e+00	2.000000e+00
75%	2.000000e+00	5.300000e+01	9.700000e+01	2.000000e+00	2.000000e+00
max	9.900000e+01	1.210000e+02	9.800000e+01	9.800000e+01	9.800000e+01

	ASTHMA	INMSUPR	HIPERTENSION	OTHER_DISEASE
count	1.048575e+06	1.048575e+06	1.048575e+06	1.048575e+06
mean	2.242626e+00	2.298132e+00	2.128989e+00	2.435143e+00
std	5.114089e+00	5.462843e+00	5.236397e+00	6.646676e+00
min	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
25%	2.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00
50%	2.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00
75%	2.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00
max	9.800000e+01	9.800000e+01	9.800000e+01	9.800000e+01

	CARDIOVASCULAR	OBESITY	RENAL_CHRONIC	TOBACCO
count	1.048575e+06	1.048575e+06	1.048575e+06	1.048575e+06
mean	2.261810e+00	2.125176e+00	2.257180e+00	2.214333e+00
std	5.194850e+00	5.175445e+00	5.135354e+00	5.323097e+00
min	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
25%	2.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00
50%	2.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00
75%	2.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00
max	9.800000e+01	9.800000e+01	9.800000e+01	9.800000e+01

	CLASIFFICATION_FINAL	ICU
count	1.048575e+06	1.048575e+06
mean	5.305653e+00	7.955397e+01
std	1.881165e+00	3.682307e+01
min	1.000000e+00	1.000000e+00
25%	3.000000e+00	9.700000e+01
50%	6.000000e+00	9.700000e+01
75%	7.000000e+00	9.700000e+01
max	7.000000e+00	9.900000e+01

Values Distribution



In addition, Date of birth couldn't not be represented graphically, so this its values counts

9999-99-99 971633

A specific date 76942

Name: DATE_DIED, Length: 401, dtype: int64

Despite the correlation between our dataset and the output of our model—COVID severity—we encountered several issues:

Firstly, two columns—ICU and INTUBED—contain a significant number of null values, approximately 80%, posing challenges for analysis.

Secondly, the presence of a "pregnant" value of 97 for males is incomparable and must be addressed to ensure consistency in the dataset.

Thirdly, while some columns have fewer null values, they still require careful handling during the training phase to avoid any adverse effects on the model's final output.

Finally, discrepancies in correlation values need to be addressed to ensure the accuracy and reliability of our analysis.

This represents only a portion of the analysis conducted. For a more detailed description and the complete analysis, please refer to the provided Colab link and GitHub repository. Moreover, the correlations specified in the table are after changing null values for males in PREGNANT feature to be not pregnant, and DATE_DIED feature to be DEAD(boolean whether the patient died or not) in order to be able to get a meaningful correlation.

Full description of the cleaning and preprocessing steps

		Unique values	Number of unknown rows	Correlation percentage	What should be done (preprocessing)	Order of edits
1	USMER	1,2		2.8839849290264 43%	N/A	0
2	MEDICAL_UNIT	1 to 13		2.8839849290264 43%	N/A	0
3	SEX	1, 2		5.7782009301862 89%	N/A	0
4	PATIENT_TYPE	1, 2		18.336965819874 866%	N/A	0
5	DATE_DIED	A lot of values		19.608503305344 406%	To Boolean	1
6	INTUBED	1, 2, 97, 99	97 -> 848544 99 -> 7325	19.307514681825 51%	Remove (80% unknowns)	2

7	PNEUMONIA	1, 2, 99	16003	7.535121763674622%	Needs prediction	4
8	AGE	0 to 121		15.263746029745228%	N/A	0
9	PREGNANT	1, 2, 97, 98	97 (converted to 0) -> 52351198 -> 3754	5.780946979693501%	pregnant males, 97 to 2. unknown female, 98 to either 1 or 2	1 4
10	DIABETES	1, 2, 98	3338	0.47387655178426175%	Needs prediction	3
11	COPD	1, 2, 98	3003	1.0336483259646734%	Needs prediction	3
12	ASTHMA	1, 2, 98	2979	1.1177889087449102%	Needs prediction	3
13	INMSUPR	1, 2, 98	3404	0.9411781788913769%	Needs prediction	3
14	HIPERTENSION	1, 2, 98	3104	0.6020237325700838%	Needs prediction	3
15	OTHER_DISEASE	1, 2, 98	5045	1.1142710652587846%	Needs prediction	3
16	CARDIOVASCULAR	1, 2, 98	3076	1.2142899794549882%	Needs prediction	3
17	OBESITY	1, 2, 98	3032	0.6924411016774965%	Needs prediction	3
18	RENAL_CHRONIC	1, 2, 98	3006	1.1342198411331843%	Needs prediction	3
19	TOBACCO	1, 2, 98	3220	1.256707629152292%	Needs prediction	3
20	CLASIFFICATION_FINAL	1 to 7			Values from 4 to 7 have the same meaning	5
21	ICU	1, 2, 97, 99	97 -> 848544 99 -> 7488	19.3163200132346%	remove (too many unknowns) done	2 done

Analysis for the data:

As seen from table above, we started by getting out all unique values of each feature and number of instances for each value in the second column. We also extracted the percentage of correlation between each feature and the output feature to check if there is a feature correlated with the label by a very high number, the feature should be removed. However, for our dataset, this was not the case for any feature as the highest correlation is around 19%.

After extracting all the previous values and analysis we have detected the kind of feature preprocessing that needs to be done in such feature. For example:

ICU and **INTUBED**, have around 80% of its values are none. Thus, the procedure to be considered in such case is column or feature removal

Death_Date, logically won't help in predicting Covid severity. Thus, the procedure to be considered is to convert those values to boolean value if the patient died or not.

PREGNANCY, for males the value is 97 because they are incomparable, thus, the procedure to be considered is to convert all values to be 2 which is not pregnant as this is the logical thing that males are not pregnant.

CLASSIFICATION_FINAL, values starting from 4 up to 7 do have the same meaning according to the data description in the website – which is that the patient is not a carrier of COVID -- Thus, all of them need to be compressed to have a value of 4.

Finally, all other features with None values, 97, 98, 99, need to be modified for model training. Thus, the procedure to be considered in this part is to predict the values of such missing info according to the similarity between those patients with none values and other patients with actual values.

For example, DIABETES patients with None values (97) will be of values 1 or 2 depending on their similarity, euclidean distance, with the average of patients with DIABETES value of 1 and the average of patients with DIABETES value of 2, but first, all features were normalized before applying Euclidean distance, and we predict the Boolean value normalized. Finally, we have a dataset of normalized values, so we do inverse transform to return the dataset to its original values.

Applying the same technique to the rest of diseases.

Excluding instances of patients with None values, 97, 98, 99, in all the predicted columns, like those which have None values in all columns.

It is worth mentioning that we noticed at first that there are some diseases that are highly correlated to each other from the correlation matrix, so we tried at first to fill the NULL values using Euclidean distance but for these highly correlated features only. However, after we did this, the correlations between each other reduced significantly, which meant that these correlations were for the NULL values, meaning that our initial assumptions were not correct. Nevertheless, we kept this dataset, and it is written in the code as method1.

Last but not least, we have all our binary features as type integers with values 1 and 2 not 0 and 1 as we believe it won't be of a great significance.

Final list of chosen features

USMER, MEDICAL_UNIT, SEX, PATIENT_TYPE, DATE_DIED, PNEUMONIA, AGE, PREGNANT, DIABETES, COPD, ASTHMA, INMSUPR, HIPERTENSION, OTHER_DISEASE, CARDIOVASCULAR, OBESITY, RENAL_CHRONIC, TOBACCO.

Rationale for adding or removing any feature

DATE_DIED: became a Boolean attribute to become more reasonable for Covid Severity prediction as there is no relation between the Death Date and Covid Severity predictoin

ICU and **INTUBED** were removed because of a 80% of unknown instances

Description of the dataset size and metrics post cleaning

Data size:

21 columns - 2 removed - 1 final output = 18 columns
1,048,575 rows or instances of patients

Metrics post cleaning:
Data Completeness:

around 10,000 was predicted throughout the whole dataset in all columns. With an average of 3,000 in some columns and 2,000 of them are common among all columns. That means if a patient has a none value in DIABETES, probably the same patient would have a none value in other diseases as well.
No Null values in the dataset we have now.

Data Accuracy:

Data is very representative of the real world as discussed in the above sections.

Data types:

All of the columns, features are of boolean values except for 2 of them which are a range of values. The 2 columns are AGE and MEDICAL_UNIT.

Full Analysis of the Final Cleaned Dataset

Statistical Analysis

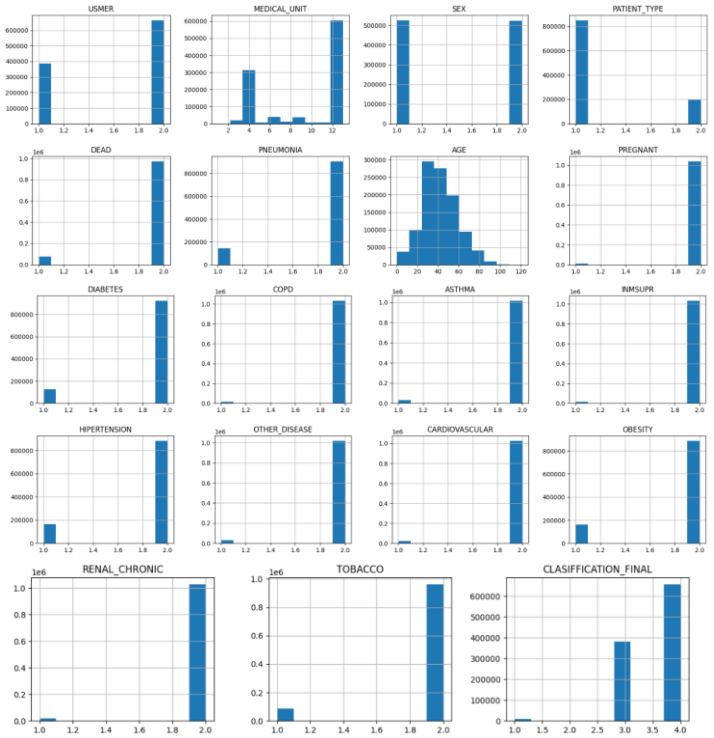
	USMER	MEDICAL_UNIT	SEX	PATIENT_TYPE	DEAD
count	1.048575e+06	1.048575e+06	1.048575e+06	1.048575e+06	1.048575e+06
mean	1.632194e+00	8.980565e+00	1.499259e+00	1.190765e+00	1.926622e+00
std	4.822084e-01	3.723278e+00	4.999997e-01	3.929041e-01	2.607556e-01
min	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
25%	1.000000e+00	4.000000e+00	1.000000e+00	1.000000e+00	2.000000e+00
50%	2.000000e+00	1.200000e+01	1.000000e+00	1.000000e+00	2.000000e+00
75%	2.000000e+00	1.200000e+01	2.000000e+00	1.000000e+00	2.000000e+00
max	2.000000e+00	1.300000e+01	2.000000e+00	2.000000e+00	2.000000e+00

	PNEUMONIA	AGE	PREGNANT	DIABETES	COPD
count	1.048575e+06	1.048575e+06	1.048575e+06	1.048575e+06	1.048575e+06
mean	1.862833e+00	4.179410e+01	1.989799e+00	1.879414e+00	1.984490e+00
std	3.440238e-01	1.690739e+01	1.004858e-01	3.256452e-01	1.235681e-01
min	1.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
25%	2.000000e+00	3.000000e+01	2.000000e+00	2.000000e+00	2.000000e+00
50%	2.000000e+00	4.000000e+01	2.000000e+00	2.000000e+00	2.000000e+00
75%	2.000000e+00	5.300000e+01	2.000000e+00	2.000000e+00	2.000000e+00
max	2.000000e+00	1.210000e+02	2.000000e+00	2.000000e+00	2.000000e+00

	ASTHMA	INMSUPR	HIPERTENSION	OTHER_DISEASE
count	1.048575e+06	1.048575e+06	1.048575e+06	1.048575e+06
mean	1.969019e+00	1.985110e+00	1.843466e+00	1.971275e+00
std	1.732665e-01	1.211117e-01	3.633613e-01	1.670318e-01
min	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
25%	2.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00
50%	2.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00
75%	2.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00
max	2.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00

	CARDIOVASCULAR	OBESITY	RENAL_CHRONIC	TOBACCO
count	1.048575e+06	1.048575e+06	1.048575e+06	1.048575e+06
mean	1.978874e+00	1.846083e+00	1.980853e+00	1.917601e+00
std	1.438038e-01	3.608698e-01	1.370414e-01	2.749725e-01
min	1.000000e+00	1.000000e+00	1.000000e+00	1.000000e+00
25%	2.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00
50%	2.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00
75%	2.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00
max	2.000000e+00	2.000000e+00	2.000000e+00	2.000000e+00

Values Distribution



Correlations

Feature	Correlation
USMER	0.5810295642010892%
MEDICAL_UNIT	4.553922310851523%
SEX	5.1765038825045355%
PATIENT_TYPE	17.532570007042782%
DEAD	18.870111807043873%
PNEUMONIA	17.582168918887405%
AGE	14.64825037631760%
PREGNANT	0.6018364253746976%
DIABETES	8.875245295503493%
COPD	1.1029871972147935%
ASTHMA	1.6151815786611659%
INMSUPR	0.4132704366255451%
INMSUPR	0.4132704366255451%
HYPERTENSION	8.134929922326833%
OTHER_DISEASE	0.2727896017981867%
CARDIOVASCULAR	1.3664018563477038%
OBESITY	6.511289790085584%
RENAL_CHRONIC	1.5504070954501303%
TOBACCO	1.5703446410291817%