Predicting Risks of COVID-19 Patients With Data

Zheqi Liu, Tsinghua University
Supervisor: Professor Hien Tran

ABSTRACT

Predicting the risk of Covid-19 is very important for those infected because we can warn the vulnerable to take action in advance. With a dataset provided by the Mexican Government, which has information on 1,048,575 COVID-19 patients' symptoms, status, and medical history, and various machine learning models, we can now predict whether a patient died of COVID-19.

DATASET OVERVIEW

This dataset has no not-a-number values but there are several missing values.

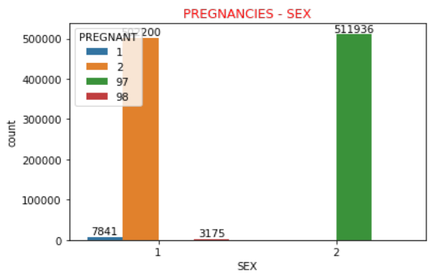
1) We have some features that we expect them to have just 1 and 2 but these features have more values. For example, the feature "PNEUMONIA" has 1,2, and 99. 99 represents nan values. So we will just take 1 and 2.

2) In the "DATE_DIED" column, we have "9999-99-99" values which represent alive patients so take this feature as "alive" and other numbers as "dead".

DATA PROCCESSING

1. Get rid of the missing values of features (because in these features missing values are rare) except "INTUBED", "PREGNANT", and "ICU".

2. Prepare the "DEATH" column using the date-died column.



3. We see that all "97" values are for males and males can not be pregnant so we will convert 97 to 2 and get rid of 98.

4. Drop the "INTUBED" column and the "ICU" column since there are so many missing values. Also, we have the "patient_type" column which tells whether the patient was hospitalized. Those intubated or in ICU had to be hospitalized and this compensates for the missing "ICU" and "INTUBED" data.

USMER	medical_unit	Sex	asthma	inmsupr	hipertension
0.12	-0.15	-0.08	-0.02	0.05	0.21
patient_type	pneumonia	age	cardiovascular	obesity	renal_chronic
-0.52	0.47	0.32	0.08	0.06	0.12
pregnant	diabetes	COPD	other_disease	tobacco	classification_final
-0.02	0.22	0.09	0.06	0.01	-0.20

5.Calculate correlation coefficients with "DEATH" and drop l ow related features (absolute value<0.1).

6. Prepare the categorical features ("MEDICAL_UNIT","CLASIFFICATION_FINAL") which are not binary by turning them into one-hot form. And scale the AGE column values into (-1,1)

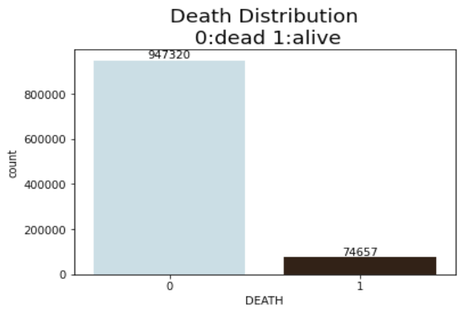
METHOD AND RESULT

To perform classification, I used various machine learning models including logistic regression, multiple layer perceptron (MLP), support vector machine (SVM), linear discriminant analysis (LDA), random forest, and decision tree to see which one has the best result in terms of f1-score. The higher the f1-score goes, the better the result is.

With imbalanced data, by using logistic regression, we have a 0.96 f1-score for those who survived but we only have a 0.51 for those who died.

This is because this imbalanced dataset has too many patients that survived COVID.

After using undersampling and SMOTE, we recieve much better average f1-scores.



Death distribution

	Logistic regression	MLP (2000 epoches)	Random forest
Undersampling	0.905	0.910	0.908
SMOTE	0.909	0.913	0.908
	SVM	LDA	Decision tree
Undersampling	0.904	0.896	0.911
SMOTE	-	0.900	0.918

Augmented data

DISCUSSION

1. In the correlation coefficients part, hospitalization is most related to death, followed by pneumonia, diabetes, and hypertension. People with diabetes and hypertension need to take care and people with pneumonia should be hospitalized.

2. In the result section, the decision tree model has the best f1-score in both undersampled data and the SMOTE data, which means that decision tree is most suitable for this task.

3. The SMOTE data have a slightly better result than the undersampled data which can be explained by the fact that the SMOTE data have far more samples.

4. The SMOTE data are not suitable for SVM because SVM model needs a lot of calculation and the speed of calculation is very slow when the dataset is big.

5. The original dataset is not balanced and we have a much better result after we balance it. So it is obvious that we examine the dataset and balance the imbalanced ones.

DATASET AND CODE

This dataset can be found on kaggle.
Link: https://www.kaggle.com/datasets/meirnizri/covid19-dataset
If you want to run the code, go to https://github.com/azishabibi/covid
Contact: liuzheqi20@mails.tsinghua.edu.cn