Predicting the Mortality of COVID-19 Patients*

Primary Topic: Data Mining
Course: 2020-201400174-1A Group: 25 – Submission Date: 2020-11-08

Deepika Jangamguravepalli Bramhanandareddy
University of Twente
d.jangamguravepallibramhanandareddy@student.utwente.nl

Sathvik Guru Rao
University of Twente
s.gururao@student.utwente.nl

ABSTRACT

This paper provides a study of the dataset provided by the Tongyi Hospital Wuhan, China. It contains COVID-19 patient biomarker information which helps to predict the severity of the COVID-19 patients. In the paper "predicting fatality of COVID-19 patients using logistic regression"[1] Feng et al, has used 3 biomarkers to identify the severity of COVID-19 patients with an accuracy of 90%. In our paper, we are using other biomarkers to see whether we can increase the accuracy of predicting the severity. We are using different Data mining models to compare the performance and give the best model which suits the prediction of mortality of the COVID-19 patients.

KEYWORDS

Data Mining, Feature Selection, Machine Learning, Supervised learning algorithm

1 INTRODUCTION

The epidemic COVID-19 has sparked a global alarm, it causes illness ranging in severity. The main problem the world is facing is identifying the COVID-19 cases, in this situation with the correct prediction, the treatment can be efficient.

The purpose of this paper is to determine up to what extent we can increase the accuracy by using other biomarkers and by constructing a data-driven decision support system using Data Mining models. We also want to know which features are relevant in this prediction of the mortality of covid -19 patients. With this as a goal, we framed the following two research questions,

- Which other biomarkers are important to decide the mortality of patients?
- Combine these features with the three test biomarkers used in the paper "predicting fatality of COVID-19 patients using logistic regression" and see how it affects the accuracy.

To achieve these research goals, we used various models such as Decision Tree, Random Forest, and SVM to predicting the accuracy of COVID-19 cases. We also used the feature selection algorithm such as XGBoost to strengthen the hypothesis.

The structure of the paper is categorized as follows. Firstly, we deliver an overview of the dataset and how the dataset is modified for better efficiency. Subsequently, we present the experiments we performed to evaluate the models and algorithms to determine the important features. Finally, the paper is concluded with a discussion about the results.

*Adapted from the ACM SigConf Template. More information about the template can be found at https://www.acm.org/publications/proceedings-template

2 BACKGROUND AND RELATED WORK

To develop an accurate model for predicting mortality in COVID-19 patients, statistical methods such as Decision tree, Random forest, and SVM should be known for understanding on which basis the classification was performed. The findings were helpful for guidance on the algorithms used and understand the role of different biomarkers in the disease pathogenesis of COVID-19. To get more insight into the biomarkers we consulted a few Doctors furthermore into the research.The selected biomarkers will directly associaled with the mortality for example if the CRP level increased the severity of the COVID-19 is high. The following figure represent the biomarkers and the severity of the COVID-19

Biomarker	Change in severe COVID-19 infection
CRP	Increase
SAA	Increase
IL-6	Increase
LDH	Increase
WCC	NLR increases LC decrease
D-dimer	Increase
Platelet count	Decrease
Cardiac troponin	Increase
Renal biomarkers	Urea & creatinine increase

CRP = C-reactive protein; SAA = serum amyloid A; IL-6 = interleukin 6; LDH = lactate dehydrogenase; WCC = White cell count.

Figure 1: Change in severe COVID-19 infection

Muhamed et al. [2]presented in their paper " The role of biomarkers in the diagnosis of COVID-19 - A systematic review" focuses on the many biomarkers with a strong association with mortality which helps clinically identifying the severity of disease earlier to improve the prognosis.
Yan Wang et al.[3] Presented in their paper "XG Boost risk model via feature selection and bayesian hyperparameter optimization" on exploring models based on the extreme gradient boosting(XGBoost) approach for feature selection(FS) during model training.

3 DATA PRE-PROCESSING

3.1 Dataset Description

The dataset provided by the Tongyi Hospital Wuhan, China consists of 375 patient information with 74 biomarkers. And the outcome is

classified as 0 as discharge and 1 as death. The data considering for training and test model consist of the last row information of each patient.

3.2 Missing values

The first step was checking for the missing values in the dataset. It was noted that there are a lot of missing values the features like Ferritin, HCV antibody quantification, HBsAg, HIV antibody, Interleukin, Interleukin 2, Tumor necrosis, and Interleukin 8 can be removed because 80% of the data is missing, filling this values will have an effect on the prediction of mortality.

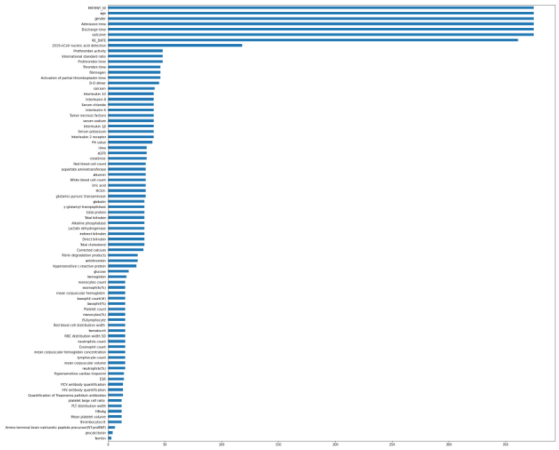

Figure 2: Missing Values

4 CLASS ANALYSIS

4.1 Class imbalance

While using Machine learning algorithms, the target variables should have the same number of instances of each class. When the number of instances of one class far exceeds the order, the problem "class imbalance" occurs. The count of class "outcome" of the feature is analyzed. From the figure 2, it is evident that the class is not having any imbalance.

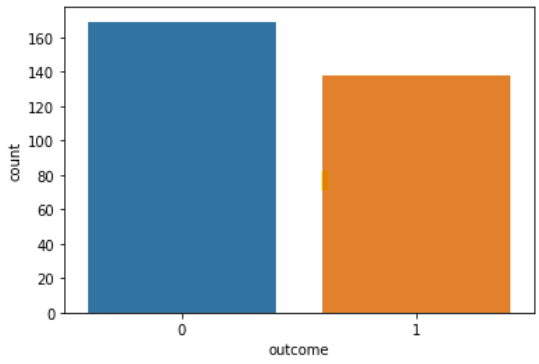

Figure 3: Class imbalance

4.2 Data split

The data split for the first research question, the training and test data were given, and for the research question two cross-validation approach was used to train and test the models

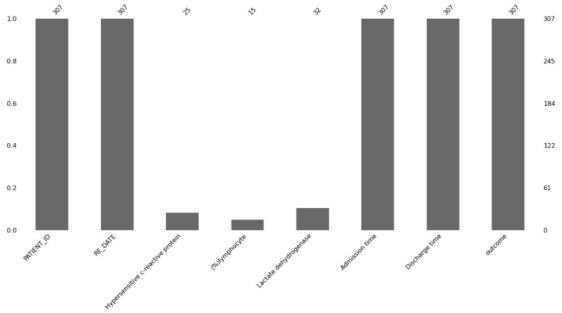

Figure 4: Training dataset

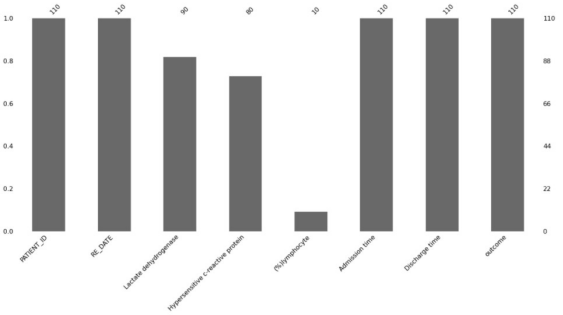

Figure 5: Testing dataset

4.3 Selected biomarkers form research

Ncov nucleic acid detection, Platelet count, D- Dimer, Cardin troponin, Lymphocyte count, Creatinine, White cell count, Neutrophil count, and Urea.

5 MACHINE LEARNING MODELS

Supervised machine learning algorithms are used for building prediction models. These models are trained by creating a dependency between the input features and their corresponding outcomes. When a new input feature is given to the model, it will predict the outcome based on the dependencies it was trained before. There are several supervised learning algorithms that can be used for prediction. In this paper Decision tree, Random forest, and support vector machine algorithm is implemented as these are most suited for categorical outcomes.

5.1 DECISION TREE

The decision tree is a predictive model used for categorical variables. The algorithm splits the data - the root node into subsets - leaf nodes until the leaf node has homogenous data. The splitting of the nodes is based on criteria. The leaf node gives the outcome. The criteria

used to split the nodes is "gini", which is the default criteria used by the decision tree classifier algorithm in sci-kit learn library. The gini impurity works based on the impurity in a given node. If all the elements in the node are of the same class, then the node is pure. Another parameter used in the algorithm is the splitter. The splitter used for this model is the "best" split. The best split chooses the best feature from all the features to split the nodes.

5.2 RANDOM FOREST

A random forest algorithm is a group of decision trees and the prediction of the outcome is based on votes from all the decision trees. Random forest reduces overfitting as the result depends on the number of votes from each decision tree. The number of trees generated can be tuned by using the parameter 'n_estimators' which is set to 100. The higher the number of trees, the stronger the prediction because there will be any number of decision trees to judge. As the random forest is a collection of decision trees, gini impurity is again used as the splitting criterion.

5.3 SUPPORT VECTOR MACHINE

The support vector machine algorithm(SVM) is a machine learning model that can be used for both classification and regression. It can be used for both continuous and categorical values. SVM creates a hyperplane which will separate the two classes. The algorithm finds points from both the classes that are close to the hyperplane. These points are called the support vectors. The greater the distance between the points and the hyperplane, the better the classification. SVM uses mathematical functions to convert the data features into the desired dimension to use it for the algorithm. This is called a kernel trick which helps in building models with more accuracy. The kernel used in our model is 'rbf' kernel which is the default kernel used by the sci-kit-learn library.

6 FEATURE SELECTION

One of the main process in a prediction model is the features that are used for training and testing. Features play a major role in accuracy and model performance. Using too many features may not only affects accuracy but also hinders the model performance. By using feature selection algorithms we can decrease the noise and computational cost during the training stage. Many studies have shown that by using the feature selection algorithm we can increase the classification performance

6.1 EXTREME GRADIENT BOOSTING

The extreme gradient Boost algorithm is a gradient boosted decision tree which is more efficient. In this paper, the XGBoost algorithm is used to get features importance of the features of the dataset. The algorithm takes all the features of the dataset and calculates the F-score which tells how the feature contributes to building the decision tree. The algorithm gives a ranked list of the important features.

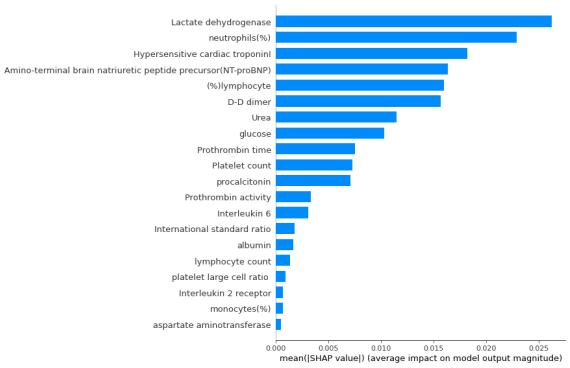

Figure 6: XGBoost Feature selection

7 METHODOLOGY

The first research question requires comparing the results of models using the three biomarkers used in the paper by Feng et al[1] and the biomarker found after some research by using research papers and doctors. The training data and test data are already given so the splitting of training data can be skipped. The training data with the three biomarkers are considered and trained on all the three models and evaluated on the test data. The performance of the evaluation is measured using accuracy.

The second research question is combining the three biomarkers given and the new nine biomarkers to check if the accuracy increases. The dataset used has '2019-nCoV nucleic acid detection','Platelet count','D-D dimer','Hypersensitive cardiac troponin-I','lymphocyte count','creatinine','White blood cell count','neutrophils count','Urea','time_in_hospital' biomarker. The training dataset is used with cross-validation so that all the algorithms have the same dataset when used for training. K- cross-validation is used to split the data K times where one subset is used in testing and the K-1 dataset is used in training the model. In this way the accuracy after each iteration is averaged. The test and training data will change in each iteration and will help effectively in modeling as the whole dataset will be used after all the iterations are complete. In this paper, the value of k is 10.

The XGboost algorithm was used to find important features of the whole dataset using the F-score of each attribute. The top ten of the important features were taken to create new data mining models. Some of the attributes in the top ten were common with the features which were selected earlier from research and doctor consultation. These biomarkers were again used for machine learning models.

8 RESULTS

The accuracy of the model with 3 given bio markers are as follows
accuracy DT : 78.07%
accuracy RF : 84.4%
accuracy SVM : 63.01%

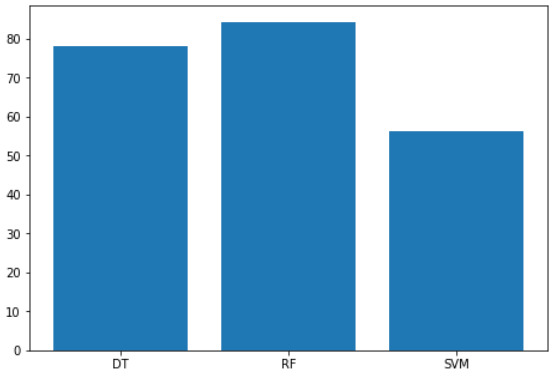

Figure 7: Accuracy using 3 biomarkers

The accuracy of the model with 9 selected bio markers are as follows
accuracy DT : 97.53%
accuracy RF : 97.63%
accuracy SVM : 74.1%

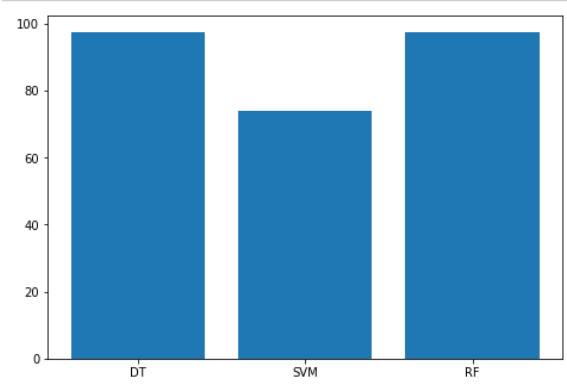

Figure 8: Accuracy using 7 selected biomarkers

The accuracy of the model combining all the selected and given bio markers are as follows
accuracy DT : 97.53%
accuracy RF : 97.63%
accuracy SVM : 74.1%

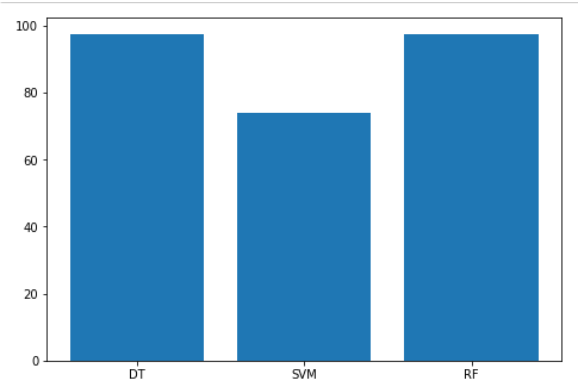

Figure 9: Accuracy using all selected and given biomarkers

The accuracy of the model with feature selection from XGBOOST
accuracy DT : 72.6%
accuracy RF : 74.2%
accuracy SVM : 57.4%

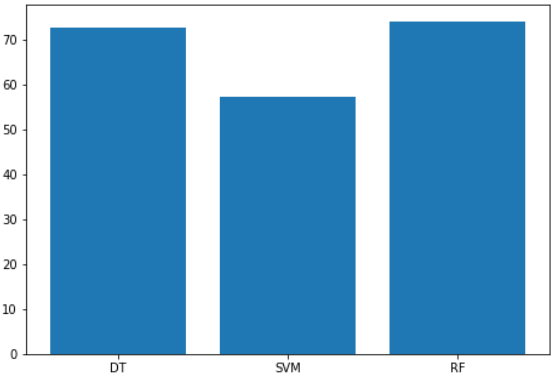

Figure 10: Accuracy using feature selection biomarkers

9 DISCUSSION

From the result obtained in section 8. All the models performed very well in predicting the mortality of COVID 19 cases. With respect to the first research question, the accuracies from the models are 78.07% for the decision tree, 84.4% for the random forest, and 63.01% for the support vector machine. Here random forest performed very well than other models. Then from the research, the nine selected biomarkers were used on the same models and the accuracies were 97.53% for the Decision tree, 97.63% for the random forest, and 74.1% for SVM which has increased the performance of prediction compared to using three biomarkers.

For the second research question, we combined these two sets of biomarkers, and the accuracy of predicting the mortality of COVID 19 patients remained the same as the accuracy of using the nine biomarkers found from research.

XGBoost feature selection algorithm was used to find important biomarkers using an F-score measure. The list of important features can be found in fig: 5. The features generated from the

algorithm and the features selected from the research were almost the same which strengthens the hypothesis of the first research question. The common feature among them are 'Lactate dehydrogenase','neutrophils(%)','Hypersensitive c-reactive protein', '(%)lymphocyte','D-D dimer','Urea','Platelet count'.

10 CONCLUSIONS

We can conclude that the research goals have been accomplished. By using different biomarkers the mortality can be predicted with higher accuracy. The accuracy was found using different data mining models. The random forest algorithm has worked well for the given dataset with an accuracy of 97%. The high accuracy may be because the missing values were replaced by the median value of respective features. The research can be improved if there were no or less missing data.

ACKNOWLEDGMENTS

We acknowledge the use of the data from Tongji Hospital in Wuhan,China which contains the patients information about the bio markers

REFERENCES

[1] Zhou, F., Chen, T., Lei, B. (2020).Do not forget interaction: Predicting fatality of COVID-19 patientsusing logistic regression. Retrieved fromhttps://arxiv.org/abs/2006.16942

[2] Muhammed Ker, The role of biomarkers in diagnosis of COVID-19 – A-systematic review, May 2020Life Sciences 254:117788, DOI: 10.1016/j.lfs.2020.117788

[3] Wang, Yan Ni, Xuelei. (2019). A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization.