Scientific Research Publishing

Journal of Computer and Communications, 2022, 10, 18-25
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Framework Development Using Data Mining Techniques to Predict Mortality Risk during Pandemic

Debjany Chakraborty, Md Musfique Anwar

Computer Science and Engineering Department, Jahangirnagar University, Dhaka, Bangladesh
Email: debjany.cuetcse@gmail.com, manwar@juniv.edu

How to cite this paper: Chakraborty, D. and Anwar, M.M (2022) Framework Development Using Data Mining Techniques to Predict Mortality Risk during Pandemic. *Journal of Computer and Communications*, **10**, 18-25.
https://doi.org/10.4236/jcc.2022.108002

Received: January 23, 2022
Accepted: August 9, 2022
Published: August 12, 2022

Copyright © 2022 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Open Access

Abstract

The corona virus, which causes the respiratory infection Covid-19, was first detected in late 2019. It then spread quickly across the globe in the first months of 2020, reaching more than 15 million confirmed cases by the second half of July. This global impact of the novel coronavirus (COVID-19) requires accurate forecasting about the spread of confirmed cases as well as continuation of analysis of the number of deaths and recoveries. Forecasting requires a huge amount of data. At the same time, forecasts are highly influenced by the reliability of the data, vested interests, and what variables are being predicted. Again, human behavior plays an important role in efficiently controling the spread of novel coronavirus. This paper introduces a sustainable approach for predicting the mortality risk during the pandemic to help medical decision making and raise public health awareness. This paper describes the range of symptoms for corona virus suffered patients and the ways of predicting patient mortality rate based on their symptoms.

Keywords

Sequential forward Feature Selection, Symptom Categorization, Decision Tree, Attribute Selection Measure

1. Introduction

The global impact of the novel coronavirus (COVID-19) requires accurate forecasting about the spread of confirmed cases as well as continuation of analysis of the number of deaths and recoveries. Forecasting requires a huge amount of data. At the same time, forecasts are highly influenced by the reliability of the data, vested interests, and what variables are being predicted. This paper introduces a

sustainable approach for predicting the mortality risk during pandemic to help medical decision making and raise public health awareness. This paper describes the range of symptoms of the patients who suffered in corona virus and the ways of predicting patient mortality rate based on their symptoms.

In this study, we will propose a data-driven predictive algorithm based on data mining to determine the health risk and predict the mortality risk of patients with COVID-19. The algorithm predicts the mortality risks based on patients' physiological conditions, symptoms, chronic medical history, duration of illness and demographic information. This model will help hospitals and medical facilities in the following way:

- who needs to get attention first?
- who has higher priority to be hospitalized,
- triage patients when the system is overwhelmed by overcrowding,
- eliminate delays in providing the necessary care.
- take immediate decisions by observing the most alarming symptoms

2. Background Study

Sajana *et al.* [1] applied a non-invasive machine learning techniques to facilitate the doctors for ordering the hazard in dengue patients. They have conducted a comparison study among Simple Classification and Regression Tree (CART), Multi-layer perception (MLP) and C4.5 algorithms, based on which demonstrating that Simple CART algorithm shows 100% accuracy for classification of affected or unaffected patient. In this paper they have investigate various papers of different authors and made a list of the comparison between them in tabular form.

Krishna *et al.* [2] in their research paper mainly focused on a data mining technique that had an objective of creating a prediction model, using decision tree for predicting the chances of occurrences of diseases in an area, this model also identifies different significant parameters which can be used to help for the creation of model. They have taken both rural and urban area data, classified data set though decision tree construction method and they showed finally that out of 344 dengue cases 48.93% are from tribal areas. There are very limited cases from Hill, Rural and Urban areas of East Godavari District *i.e.*, dengue cases are reported mainly in Tribal and Hill areas.

Mahdavi *et al.* [3] aimed to develop and compare prognosis prediction machine learning models based on invasive laboratory and noninvasive clinical and demographic data from patients' day of admission. Wanyan *et al.* proposed a novel framework that utilizes relational learning based on a heterogeneous graph model (HGM) for predicting mortality at different time windows in COVID-19 patients within the intensive care unit (ICU) [4]. Friedman *et al.* [6] performed analysis by introducing a publicly available evaluation framework for assessing the predictive validity of COVID-19 mortality forecasts and track the model performance as well.

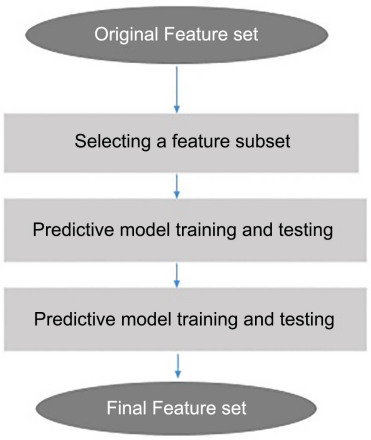

Figure 1. Workflow of sequential forward feature selection.

3. Proposed Methodology

We have applied step by step process for completing our whole framework development as shown in **Figure 1**.

3.1. Data Preprocessing

We first need to apply preprocessing steps to remove noisy information from the dataset [5] [7] [9] [10] [11].

- We used a dataset of more than 117,00 laboratory-confirmed COVID-19 patients from 76 countries around the world including both male and female patients with an average age of 56.6 [8].
- The original dataset contained 32 data elements from each patient, including demographic and physiological data.
- At the data cleaning stage, we removed useless and redundant data elements such as data source, admin id, and admin name.
- Data imputation techniques were used to handle missing values.

3.2. Sequential forward Feature Selection (SFS)

The primary purpose of feature selection is to find the most informative features and eliminate redundant data to reduce the dimensionality and complexity of the model.

In SFS variant features are sequentially added to an empty set of features until the addition of extra features does not reduce the criterion. Here, starting from the empty set, sequentially add the feature $x+$ that maximizes $J(Yk+x)$ when combined with the features Yk hat have already been selected.

3.3. Symptom Categorization of Covid-19 Patient

The Nobel corona virus can cause a range of symptoms, ranging from mild illness to pneumonia. Symptoms of the disease are fever, cough, sore throat and headaches. In severe cases difficulty in breathing and deaths can occur. Here we

have classified the symptoms into three categories: mild, severe, critical. The symptoms of these three categories are written below:

Mild Case: Fever, Dry Cough

Severe Case: Fever, Dry Cough, Diarrhea

Critical Case: Fever, Dry Cough, Diarrhea, Pneumonia, Shortness of Breath, Respiratory Failure (Figure 2).

3.4. Decision Tree Construction

We follow decision tree construction method to identify Decision Tree is a popular classifier which is simple and easy to implement. It requires no domain knowledge or parameter setting and can handle high dimensional data. The results obtained from Decision Trees are easier to read and interpret. The drill through feature to access detailed patients profiles is only available in Decision Trees.

- Select the best attribute using Attribute Selection Measures (ASM) to split the records.
- Make that attribute a decision node and breaks the dataset into smaller subsets.
- Start tree building by repeating this process recursively for each child until one of the condition will match:

- All the tuples belong to the same attribute value.
- There are no more remaining attributes.
- There are no more instances.

A decision tree is a tree where each node represents a feature (attribute), each link (branch) represents a decision (rule) and each leaf represents an outcome (categorical or continues value). The whole idea is to create a tree like this for the entire data and process a single outcome at every leaf (or minimize the error in every leaf) (Figure 3).

Conditions for stopping partitioning:

- All samples for a given node belongs to the same class
- There are no remaining attributes for further partitioning—majority voting is engaged for classifying the leaf. There are no samples remaining.

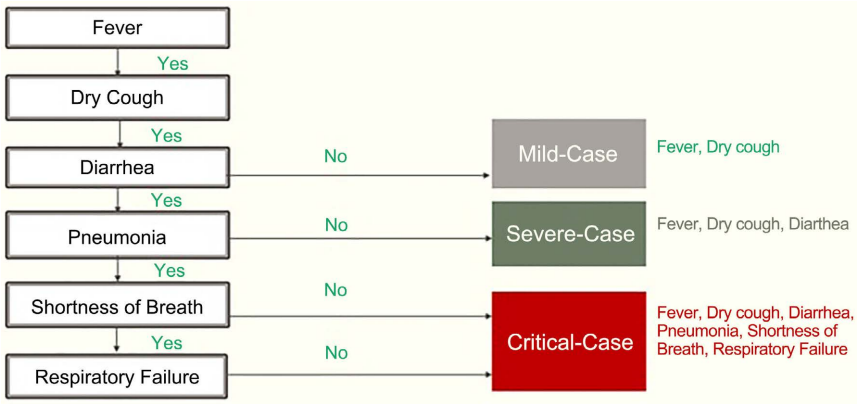

Figure 2. Schematic decision tree.

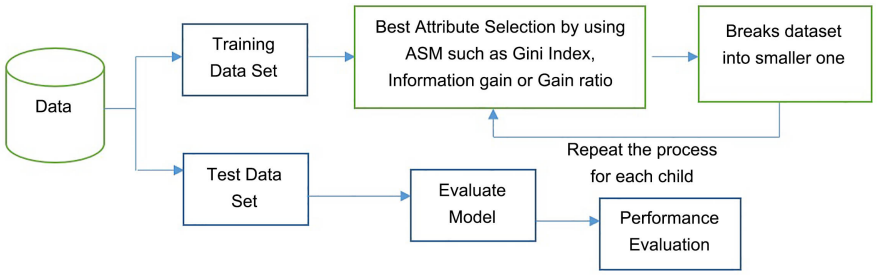

Figure 3. Process of decision tree construction.

3.5. Attribute Selection Measure (ASM)

By explaining the given dataset, attribute Selection Measure provides a rank to each feature (or attribute). We will identify the splitting attribute through best score attribute. But we need to define the split points in case of a continuous-valued attribute. Here,

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \quad (1)$$

$$Gain(A) = Info(D) - Info_A(D) \quad (2)$$

where, $Info(D)$ is the average amount of information needed to identify the class label of a tuple in D. The term $|D_j|/|D|$ is the weight of the j-th partition. $InfoA(D)$ is the expected information required to classify a tuple from D based on the partitioning by A and the attribute A with the highest information gain where, $Gain(A)$ is chosen as the splitting attribute at node N.

Next, we build schematic decision tree as shown in Figure 4.

Here, **Mchronic** means Multi-chronic diseases: Those patients who have more than one chronic disease history. Again, **Chronic** means those patients who have one chronic disease.

4. Experimental Results

We have divided the original dataset[1] into 60:40 split (Train:Test). The validation in the auto model is a multi-hold out set validation. The model has trained on 60% data and the 40% test data has been divided into subsets. Once the model is trained, it has been used to make predictions on each of the subsets independently and the performance of these subsets has been averaged (Figures 5-7).

Table 1 presents the performance evaluation of the proposed model with the real time data. This performance is based on the following criteria:

- The performance is calculated on a 40% hold out set which has not been used for any of the performed model optimizations.
- This hold-out set is then used as input for a multi-hold-out-set validation where we calculate the performance for 7 disjoint subsets.
- The largest and the highest performance are removed and the average of the remaining 5 performances is reported here.

[1]https://www.worldometers.info/coronavirus/coronavirus-death-rate/#hfr)

- Although this validation is not as thorough as a full cross-validation, this approach strikes a good balance between runtime and model validation quality.

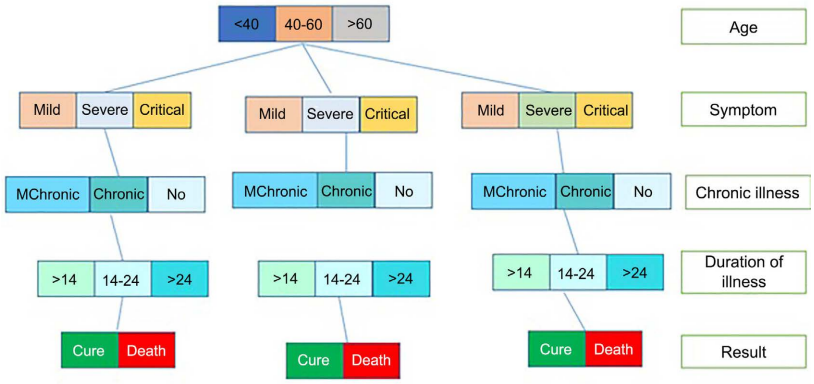

Figure 4. Schematic decision tree.

Row No.	Age	Sex	Chronic illne...	Duration of il...	Symptom	Result
9	30	Female	None	11	Mild	Cured
10	46	Female	None	16	Severe	Cured
11	31	Male	Gout	15	Severe	Cured
12	48	Male	High BP	26	Mild	Dead
13	80	Male	Cancer	23	Critical	Cured
14	30	Female	None	11	Mild	Cured
15	46	Female	None	16	Severe	Cured
16	31	Male	Gout	15	Severe	Cured
17	48	Male	High BP	26	Mild	Dead
18	80	Male	Cancer	23	Critical	Cured
19	30	Female	None	11	Mild	Cured
20	46	Female	None	16	Severe	Cured
21	31	Male	Gout	15	Severe	Cured
22	48	Male	High BP	26	Mild	Dead

Figure 5. Schematic decision tree.

Row No.	survived	prediction(survived)	confidence(no)	confidence(yes)	Sex	Chronic illne...	Symptom	Age	Duration of illness
1	no	yes	0.417	0.583	Male	High BP	Mild	48	26
2	yes	yes	0.083	0.917	Male	Cancer	Critical	80	23
3	yes	yes	0.000	1.000	Female	None	Mild	30	11
4	yes	yes	0.000	1.000	Female	None	Severe	46	16
5	yes	yes	0.000	1.000	Male	Gout	Severe	31	15
6	yes	yes	0.083	0.917	Male	Cancer	Critical	80	23
7	yes	yes	0.000	1.000	Female	None	Mild	30	11
8	yes	yes	0.000	1.000	Female	None	Severe	46	16
9	yes	yes	0.000	1.000	Male	Gout	Severe	31	15
10	yes	yes	0.000	1.000	Female	None	Severe	46	16
11	yes	yes	0.000	1.000	Male	Gout	Severe	31	15
12	yes	yes	0.083	0.917	Male	Cancer	Critical	80	23
13	yes	yes	0.000	1.000	Female	None	Mild	30	11
14	yes	yes	0.000	1.000	Male	Gout	Severe	31	15
15	yes	yes	0.000	1.000	Male	Gout	Severe	31	15
16	no	yes	0.417	0.583	Male	High BP	Mild	48	26
17	yes	yes	0.083	0.917	Male	Cancer	Critical	80	23

Figure 6. Schematic decision tree.

Age	Integer	0	Min 30	Max 80	Average 46.955
Sex	Polynominal	0	Least Female (67)	Most Male (90)	Values Male (90), Female (67)
Chronic illness	Polynominal	0	Least High BP (30)	Most None (67)	Values None (67), Cancer (30), ...[2 more]
Duration of illness	Integer	0	Min 11	Max 26	Average 18.102
Symptom	Polynominal	0	Least Critical (30)	Most Severe (67)	Values Severe (67), Mild (60), ...[1 more]
Result	Polynominal	0	Least Dead (30)	Most Cured (127)	Values Cured (127), Dead (30)

Figure 7. Schematic decision tree.

Table 1. Performance evaluation.

Death category	Proposed System	Original case
Overall	24.05%	15%
Male	9.05%	4.07%

5. Conclusions

This work proposes a framework to predict the mortality risk during pandemic in order to make proper medical decisions as well as generate public health awareness. Our observation is that data mining excels at categorizing data, especially once it has been exposed to large amounts of data on the subject. That creates great promise for data mining when it comes to diagnostics—medical imaging analysis and patient medical records, genetics, and more can all be combined to improve diagnostic outcomes and medical decision support.

The future scope of this study is to design and develop an industry level Hospital Management ERP solution by incorporating our framework in the proper way in our traditional hospital management system software to minimize cost.

Acknowledgements

We are thankful to the Department of Computer Science and Engineering, Jahangirnagar University.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Sajana, T., Navya, M., Gayathri, Y. and Reshma, N. (2018) Classification of Dengue Using Machine Learning Techniques. *International Journal of Engineering Technology*, **7**, 212-218. https://doi.org/10.14419/ijet.v7i2.32.15570

[2] Varma, M., Krishna, S. and Rao, N.K. (2015) Dengue Data Analysis Using Decision

Tree Model. *International Conference on Emerging Trends in Science Technology Engineering and Management*, 9th & 10th, October 2015, 404-414.

[3] Mahdavi, M., Choubdar, H., Zabeh, E., Rieder, M., Safavi-Naeini, S., *et al.* (2021) A Machine Learning Based Exploration of COVID-19 Mortality Risk. *PLOS ONE*, **16**, e0252384. https://doi.org/10.1371/journal.pone.0252384

[4] Wanyan, T. and Vaid, A. and De Freitas, J.K., *et al.* (2020) Relational Learning Improves Prediction of Mortality in COVID-19 in the Intensive Care Unit. *IEEE transactions on Big Data*, **7**, 38-44. https://doi.org/10.1109/TBDATA.2020.3048644

[5] Anwar, M.M., Liu, C. and Li, J. (2018) Uncovering Attribute-Driven Active Intimate Communities. In: Wang, J., Cong, G., Chen, J. and Qi, J., Eds., *Databases Theory and Applications. ADC* 2018, Springer, Cham, 109-122. https://doi.org/10.1007/978-3-319-92013-9_9

[6] Friedman, J., Liu, P., Troeger, C.E., *et al.* (2021) Predictive Performance of International COVID-19 Mortality Forecasting Models. *Nature Communications*, **12**, Article No. 2609. https://doi.org/10.1038/s41467-021-22457-w

[7] Anwar, M.M., Liu, C., Li, J. (2017) Discovering and Tracking Active Online Social Groups. In: Bouguettaya, A., et al., Eds., *Web Information Systems Engineering. WISE* 2017. Springer, Cham, 54-69. https://doi.org/10.1007/978-3-319-68783-4_5

[8] Stephany, F., Stoehr, N., *et al.* (2020) The CoRisk-Index: A Data-Mining Approach to Identify Industry-Specific Risk Assessments Related to COVID-19 in Real-Time. *SSRN Electronic Journal*, 18 p. https://dx.doi.org/10.2139/ssrn.3607228

[9] Ahmed, M.S., Aurpa, T.T. and Anwar, M.M. (2021) Detecting Sentiment Dynamics and Clusters of Twitter Users for Trending Topics in COVID-19 Pandemic. *PLOS ONE*, **16**, e0253300. https://doi.org/10.1371/journal.pone.0253300

[10] Das Badhan, C., and Anwar, M.B., *et al.* (2021) Attribute Driven Temporal Active Online Community Search. *IEEE Access*, **9**, 93976-93989. https://doi.org/10.1109/ACCESS.2021.3093368

[11] Anwar, M.M., Liu, C. and Li, J. (2018) Discovering and Tracking Query Oriented Active Online Social Groups in Dynamic Information Network. *World Wide Web*, **22**, 1819-1854. https://doi.org/10.1007/s11280-018-0627-5