International Journal of Electrical and Computer Engineering (IJECE)
Vol. 12, No. 2, April 2022, pp. 1754~1758
ISSN: 2088-8708, DOI: 10.11591/ijece.v12i2.pp1754-1758

A data mining analysis of COVID-19 cases in states of United States of America

Özerk Yavuz
Department of Management Information Systems, Faculty of Business Administration, Halic University, Istanbul, Turkey

Article Info

Article history:

Received Jan 7, 2021
Revised Sep 9, 2021
Accepted Oct 2, 2021

Keywords:

Coronavirus
COVID-19
Data mining
Epidemic
Machine learning
Pandemic
Quantitative analysis

ABSTRACT

Epidemic diseases can be extremely dangerous with its hazarding influences. They may have negative effects on economies, businesses, environment, humans, and workforce. In this paper, some of the factors that are interrelated with COVID-19 pandemic have been examined using data mining methodologies and approaches. As a result of the analysis some rules and insights have been discovered and performances of the data mining algorithms have been evaluated. According to the analysis results, JRip algorithmic technique had the most correct classification rate and the lowest root mean squared error (RMSE). Considering classification rate and RMSE measure, JRip can be considered as an effective method in understanding factors that are related with corona virus caused deaths.

Corresponding Author:

Özerk Yavuz
Department of Management Information Systems, Faculty of Business Administration, Halic University
Beyoğlu, Istanbul, 34445, Turkey
Email: ozerkyavuz@halic.edu.tr; ozerky@gmail.com; ozerk@alumni.bilkent.edu.tr

1. INTRODUCTION

Epidemic diseases can be extremely dangerous with its hazarding short term and long-term effects. So, understanding factors associated with it and applying precautions just in time would have positive influence on prevention of the spread of disease, can save many lives and eliminate negative consequences. In 2019, outbreak of a new coronavirus, causing the respiratory illness had been identified in Wuhan, China. Virus later had seen in different countries and regions as well [1]. The name "coronavirus" is derived from Latin corona, meaning "crown" or "wreath" [2]. As stated by Unhale Coronaviruses are a group of enveloped viruses. They make up a large family of viruses that can infect birds and mammals, including humans [1]. As Syed indicated some of the common symptoms highlighted in literature are runny nose, headache, cough, sore throat, fever, a general feeling of being unwell [3]. Human coronaviruses most commonly spread from an infected person to others through air by coughing and sneezing, close personal contact [3]. Common signs of infection include fever, cough, and respiratory difficulties. Serious cases can lead to even death [3]. Studies and researches continue worldwide for the treatment of the virus worldwide. Washing hands with soap and water, paying attention to social distancing, avoidance of touching eyes, nose, or mouth with contaminated hands are suggested. Specialists are raising awareness on the precautions associated with the disease and applying several medical approaches for the treatment of the form it is seen in human beings [3].

2. RESEARCH METHOD

Purpose of data mining is to extract knowledge and insights from large amounts of data. In doing so a systematic approach is followed [4]. As it is illustrated in Figure 1, data mining process is composed of some set of steps [5]. These include business understanding, data understanding, data preparation, model building, testing and evaluation with deployment as indicated by Shearer [5]. Business understanding refers to the analysis and elicitation of business needs and characteristics, data understanding is the analysis of the data that is going to be examined, data preparation represents the pre-processing and organization of the data, model building refers to the applying model building approaches from data as in classification and clustering algorithmic methods, in the model testing and evaluation indicates the assessment of different models used and fitting to the data, finally deployment refers to the finalization and generation of the data mining analysis results to the stakeholders [5].

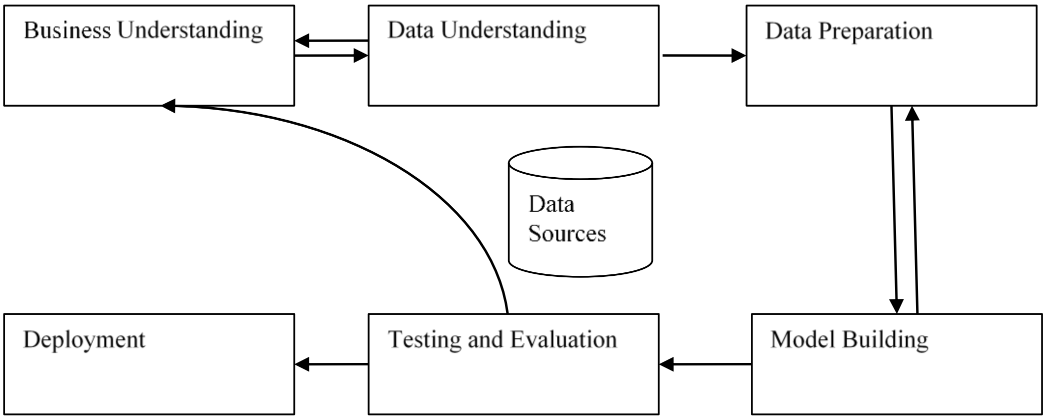


Figure 1. Data mining cycle

2.1. Data gathering and processing

Following the literature review of the study, a research model composed of variables age, gender, corona virus deaths, state have been used for the analysis. For data set, public records of United States (US) Department of Health and Human Services-Centers for Disease Control and Prevention has been used composed of 1378 instances and four variables as shown in Table 1 [6]–[9]. Age group refers to the age range, sex is the physiological and sexual characteristic of an individual, covid-19 deaths indicate the number of deaths and finally state represents the 50 federal territories which adhere and relate to the central United States government [6]–[9].

Table 1. List of attributes

Age Group	Nominal
Sex	Nominal
COVID-19 Deaths	Numeric
State	Nominal

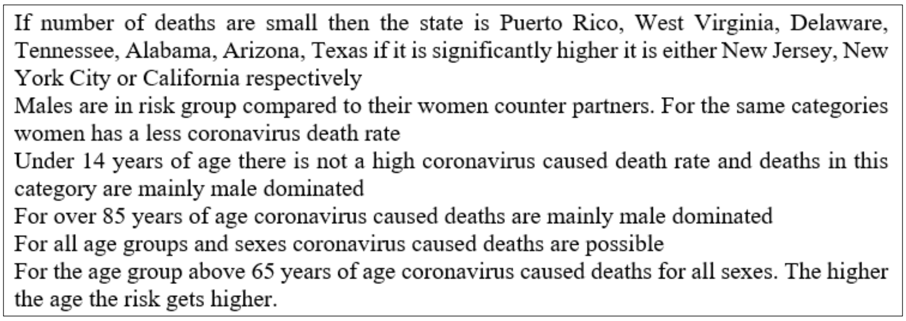
3. FINDINGS

In this research comparison of performances of several data mining algorithms have been made and rules discovered by data mining algorithms have been shared [10], [11]. In this study, comparison of the algorithms of J48, JRip, Part, OneR Method, Multilayer Perceptron, Bayesian networks have been made. In testing the research model with each of the data mining approaches 66 percent of the data has been used for the training whereas remaining part of the data set has been used for the testing of the model. Among different data mining approaches J48 had the values (RMSE=0.35; precision=0,53; correct classification rate=48.94%; incorrect classification rate=51.05). JRip had the values (RMSE=0.34; precision=0,491; correct classification rate=53.29%; incorrect classification rate=46.70). Part had the values (RMSE=0.35; precision=0.49; correct classification rate=48.94%; incorrect classification rate=51.05). OneR had the values (RMSE=0.54; precision=N/A; correct classification rate=41.16%; incorrect classification rate=58.83). OneR had the values (RMSE=0.54; precision=N/A; correct classification rate=41.16%; incorrect classification rate=58.83). Multilayer Perceptron had the values (RMSE=0.50; precision=0.42; correct classification rate=35.75%; incorrect classification rate=64.24). Bayesian networks had the values (RMSE=0.37; precision=N/A; correct classification rate=42.61%; incorrect classification rate=57.38) [12]–[15].

Among all the algorithms, JRip had the most correct classification rate with 53.29% and a precision 0.491. It also had the lowest RMSE with a value of 0.34 [16]–[18]. Comparison of data mining methods used can be seen in Table 2. Some of the rules discovered by applied algorithms are as follow and can be found in Figure 2 in detail. If number of deaths are small then the state is Puerto Rico, West Virginia, Delaware, Tennessee, Alabama, Arizona, Texas if it is significantly higher it is either New Jersey, New York City or California respectively. Males are in risk group compared to their women counter partners. For the same categories women has a less coronavirus death rate. Under 14 years of age there is not a high coronavirus caused death rate and deaths in this category are mainly male dominated. For over 85 years of age coronavirus caused deaths are mainly male dominated. For all age groups and sexes coronavirus caused deaths are possible. For the age group above 65 years of age coronavirus caused deaths for all sexes. The higher the age the risk gets higher.

Table 2**.** Comparison of the data mining methods

Method	RMSE	Precision	Correctly classified %	Incorrectly classified %
J48	0.35	0.535	48.94	51.05
JRip	0.34	0.491	53.29	46.70
Part	0.35	0.49	48.94	51.05
Oner Method	0.54	N/A	41.16	58.83
Multilayer percept.	0.50	0.42	35.75	64.24
Bayesian networks	0.37	N/A	42.61	57.38



If number of deaths are small then the state is Puerto Rico, West Virginia, Delaware, Tennessee, Alabama, Arizona, Texas if it is significantly higher it is either New Jersey, New York City or California respectively

Males are in risk group compared to their women counter partners. For the same categories women has a less coronavirus death rate

Under 14 years of age there is not a high coronavirus caused death rate and deaths in this category are mainly male dominated

For over 85 years of age coronavirus caused deaths are mainly male dominated

For all age groups and sexes coronavirus caused deaths are possible

For the age group above 65 years of age coronavirus caused deaths for all sexes. The higher the age the risk gets higher.

Figure 2. Rules discovered by data mining algorithms

4. DISCUSSION

Epidemic diseases can be extremely dangerous with its hazarding short term and long-term effects. So, understanding factors associated with it and applying precautions just in time would have positive influence on prevention of the spread of disease, can save many lives and eliminate negative consequences. In 2019, outbreak of a new coronavirus known as COVID-19 had been identified in Wuhan, China. Virus later had seen in different countries and regions as well. In the research process underlying reasons of covid-19 caused deaths have been examined using some of the data mining approaches following an intensive literature review. This is later followed with the model formation and applying the data mining techniques as suggested in literature. In the analysis part, relationship between different constructs have been examined. The model has been trained using 66 percent of the data whereas remaining part of the data has been used for testing of the model for each analysis approach [10], [19].

Data mining can be defined as the process of gaining insights and extracting knowledge from data. Knowledge discovery, prediction or forecasting can be in the focus of data mining activities. Jrip, part, OneR method, multilayer perceptron (neural networks) and Bayesian networks have been chosen as the data mining techniques applied [20]–[22]. Among them JRip is a rule learner alike in principle to the rule learner Ripper [23]. The part algorithm combines two common data mining strategies; the divide and conquer strategy for decision tree learning with the separate and conquer strategy for rule learning. OneR generates a one level decision tree that is expressed in the form of a set of rules that all test one particular attribute. Multilayer Perceptron is a version of the original perceptron model proposed by Rosenblatt in the 1950s and considered as a type of neural networks [23]–[28]. A perceptron (artificial neuron) is a function of several input perceptrons which is formed as a combination of input weights to the hidden layer perceptrons which lead them to the output layer. Finally graphical models such as Bayesian networks supply a general framework for dealing with uncertainly in a probabilistic setting and thus are well suited to tackle the problem of prediction [23]–[31].

5. CONCLUSION

In this study some of the factors that are related with corona virus caused deaths are analyzed using data mining techniques composed of supervised and unsupervised machine learning approaches. According to the analysis results, JRip had the most correct classification rate with 53.29% and a precision 0.491. It also had the lowest RMSE with a value of 0.34. Based on the classification rate and RMSE measure, JRip can be considered as an effective method in understanding factors that are related with corona virus caused deaths.

Some of the highlights discovered with classification and clustering data mining algorithms are as follow. If number of deaths are small then the state is Puerto Rico Puerto Rico, West Virginia, Delaware, Tennessee, Alabama, Arizona, Texas if it is significantly higher it is either New Jersey, New York City or California respectively. Males are in risk group compared to their women counter partners. For the same categories women has a less coronavirus death rate. Under 14 years of age there is not a high coronavirus caused death rate and deaths in this category are mainly male dominated. For over 85 years of age coronavirus caused deaths are mainly male dominated. For all age groups and sexes coronavirus caused deaths are possible. For the age group above 65 years of age coronavirus caused deaths for all sexes. The higher the age the risk gets higher. Of all the algorithms applied, Jrip had the most correct classification rate with 53.29%, a precision of 0.49 and lowest RMSE with a value of 0.34.

REFERENCES

[1] V. Kumar, K. Monika, R. Bharti, and N. Ali Khan, "A Review on Corona Virus and Covid-19," *International Journal of Pharmaceutical Sciences Review and Research*, vol. 65, no. 1, pp. 110–115, Nov. 2020, doi: 10.47583/ijpsrr.2020.v65i01.016.

[2] A. Reeha and S. Iqbal, "A review on covid-19 (coronavirus disease-2019): history, origin, spread, symptoms, treatment, prevention and impact," *International Journal of Recent Scientific Research*, vol. 11, no. 7, pp. 39141–39146, 2020, doi: 10.24327/ijrsr.2020.1107.5449.

[3] A. Syed, "Coronavirus: a mini-review," *International Journal of Current Research in Medical Sciences*, vol. 6, no. 1, pp. 8–10, 2020, doi: 10.22192/ijcrms.2020.06.01.002.

[4] E. Simoudis, "Reality check for data mining," *IEEE Expert*, vol. 11, no. 5, pp. 26–33, Oct. 1996, doi: 10.1109/64.539014.

[5] C. Shearer, "The CRISP-DM model: the new blueprint for data mining," *Journal of data warehousing*, vol. 5, no. 4, pp. 13–22, 2000.

[6] P. C. Y. Woo, Y. Huang, S. K. P. Lau, and K.-Y. Yuen, "Coronavirus Genomics and Bioinformatics Analysis," *Viruses*, vol. 2, no. 8, pp. 1804–1820, Aug. 2010, doi: 10.3390/v2081803.

[7] E. Erler, *Essays on amendment XIV*. The Heritage Foundation.

[8] S. Lemeshow and World Health Organization., "Adequacy of sample size in health studies," Chichester: John Wiley and Sons, pp. 1-239, 1990.

[9] D. Tyrrell and M. Fielder, *Cold wars: the fight against the common cold*. Oxford University Press, 2002.

[10] M. Rodríguez del Águila and A. González-Ramírez, "Sample size calculation," *Allergologia et Immunopathologia*, vol. 42, no. 5, pp. 485–492, Sep. 2014, doi: 10.1016/j.aller.2013.03.008.

[11] K. Blackmore and T. R. J. Bossomaier, "Comparison of See5 and J48.PART Algorithms for Missing Persons Profiling," in *Proceedings of the First International Conference on Information Technology and Applications (ICITA 2002)*, 2002, pp. 337–342.

[12] W. W. Cohen, "Fast Effective Rule Induction," in *Machine Learning Proceedings 1995*, Elsevier, 1995, pp. 115–123.

[13] T. Estola, "Coronaviruses, a new group of animal RNA viruses.," *Avian diseases*, vol. 14, no. 2, pp. 330–336, May 1970.

[14] E. Frank and I. H. Witten, "Generating accurate rule sets without global optimization," University of Waikato, Department of Computer Science, 1998.

[15] I. H. (Ian H. . Witten and E. Frank, *Data mining : practical machine learning tools and techniques with Java implementations*. San Francisco: Morgan Kaufmann, 2000.

[16] H. Ramchoun, M. A. J. Idrissi, Y. Ghanou, and M. Ettaouil, "New modeling of multilayer perceptron architecture optimization with regularization: An application to pattern classification," *IAENG International Journal of Computer Science*, vol. 44, no. 3, pp. 261–269, 2017.

[17] F. Rosenblatt, *The perceptron: A theory of statistical separability in cognitive systems (Project Para)*. [Washington]: [U.S. Dept. of Commerce Office of Technical Services], 1958.

[18] N. Saravana and V. Gayathri, "Performance and Classification Evaluation of J48 Algorithm and Kendall's Based J48 Algorithm (KNJ48)," *International Journal of Computer Trends and Technology*, vol. 59, no. 2, pp. 73–80, May 2018, doi: 10.14445/22312803/IJCTT-V59P112.

[19] M. Sasaki and K. Kita, "Rule-based text categorization using hierarchical categories," in *SMC'98 Conference Proceedings. 1998 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.98CH36218)*, 1998, vol. 3, pp. 2827–2830, doi: 10.1109/ICSMC.1998.725090.

[20] M. Taniguchi, M. Haft, J. Hollmen, and V. Tresp, "Fraud detection in communication networks using neural and probabilistic methods," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, 1998, vol. 2, pp. 1241–1244, doi: 10.1109/ICASSP.1998.675496.

[21] E. Venkatesan and T. Velmurugan, "Performance Analysis of Decision Tree Algorithms for Breast Cancer Classification," *Indian Journal of Science and Technology*, vol. 8, no. 29, Nov. 2015, doi: 10.17485/ijst/2015/v8i1/84646.

[22] C. Yau, *R tutorial with Bayesian statistics using OpenBUGS*, Kindle Edi. 2013.

[23] Ö. Yavuz, A. Karahoca, and D. Karahoca, "A data mining approach for desire and intention to participate in virtual communities," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 5, p. 3714, Oct. 2019, doi: 10.11591/ijece.v9i5.pp3714-3719.

[24] D. Karahoca, A. Karahoca, and Ö. Yavuz, "An early warning system approach for the identification of currency crises with data mining techniques," *Neural Computing and Applications*, vol. 23, no. 7–8, pp. 2471–2479, Dec. 2013, doi: 10.1007/s00521-012-1206-9.

[25] Ö. Yavuz, "Marketing implications of participative behavior in virtual communities," Bahcesehir University, Istanbul, 2018.

[26] Ö. Yavuz, "An early warning system approach for the identification of currency crises," Bahçeşehir University, Istanbul, 2009.

[27] Ö. Yavuz, “A public perceptions analysis with data mining algorithms,” in *International “Başkent” congress on physical, social and health sciences proceedings book*, 2021.
[28] Ö. Yavuz, “A Data Mining Analysis of Coronavirus Cases and Vaccinations in The City of London,” in *Engineering Science in the Changing and Transforming World: An Interdisciplinary Approach*, 2021, pp. 35–42.
[29] M.-C. Popescu, V. E. Balas, L. Perescu-Popescu, and N. Mastorakis, “Multilayer perceptron and neural networks,” *WSEAS Transactions on Circuits and Systems*, vol. 8, no. 7, pp. 579–588, 2009, doi: 10.5555/1639537.1639542.
[30] B. M. Ramageri, “Data mining techniques and applications,” *Indian Journal of Computer Science and Engineering*, vol. 1, no. 4, pp. 301–305, 2010.
[31] E. Grossi and M. Buscema, “Introduction to artificial neural networks,” *European Journal of Gastroenterology & Hepatology*, vol. 19, no. 12, pp. 1046–1054, Dec. 2007, doi: 10.1097/MEG.0b013e3282f198a0.

BIOGRAPHY OF AUTHOR

Özerk Yavuz received received his Ph.D. degree in Business Administration-Marketing from Bahceseshir University, Istanbul, M.Sc. degree in Computer Engineering from Bahcesehir University, Istanbul and his B.Sc. degree in Computer Technology and Information Systems from Bilkent University, Ankara. Several papers and articles of him have been published in respected and prestigious refereed, international scientific journals, books, book chapters, conference proceedings and presented in international conferences and congresses. Dr. Özerk Yavuz also has been referee, reviewer, moderator or editor of several notable, trusted international scientific journals and international, scientific, academic books. He is interested in management information systems, software engineering, computer engineering, data mining, virtual communities, virtual networks, marketing, management, and business administration. Dr. Özerk Yavuz has abroad and domestic working experiences in several institutions and countries, in various fields of business and higher education. He is interested in Salsa, Rumba, Cha-cha, East Coast Swing, Argentine Tango, American Tango, Vienna Waltz, Milonga and has been an active member of Bilkent University dance community. In his free time he loves travelling, swimming and enjoying different kitchens. Dr. Özerk Yavuz has worked with several respected and distinguished scholars, leaders and teenagers in his work life. He has been a member of several distinguished scientific communities, Bilkent University and Bahçeşehir University alumni organizations. He is currently working in Halic University, Faculty of Management, and Management Information Systems department as Asst. Prof. Dr. and continues his academic, administrative works. He can be contacted at email: ozerkyavuz@halic.edu.tr; ozerky@gmail.com; ozerk@alumni.bilkent.edu.tr.