BALKAN JOURNAL OF ELECTRICAL & COMPUTER ENGINEERING, Vol. 9, No. 2, April 2021

A Proposed Ensemble Model for The Prediction of Coronavirus Anxiety Scale of Migrant Workers

Emek Guldogan

***Abstract*—This study aimed to evaluate the potential negative effects of the scattered migrant worker population on the anxiety level by estimating the coronavirus anxiety scale (CAS) of the COVID-19 anxiety scale with Gradient Boosting Tree (GBT). In this study, a public data set achieved from a questionnaire [developed using the Coronavirus Anxiety Scale (CAS)] was used to conduct on 1350 people over phone calls. GBT model was constructed for predicting the CAS score of migrant workers based on input variables including demographical data. Hyperparameters of the GBT model were tuned using Optimize Parameters (Evolutionary) operator, which seeks the optimum values of the selected parameters by an evolutionary computation approach. Hyperparameters of the GBT model were 50 for the number of trees, 5 for minimal depth, 0.044 for learning rate, and 1.0E-5 for minimum split improvement. A total of 1500 people, 758 (56.1%) male, and 592 (43.9%) female, participated in this study. The experimental findings demonstrated that the GBT yielded a root mean square error of 3.547±0.235, the absolute error of 2.943±0.154, relative error lenient of 31.54%±0.82%, squared error of 12.623±1.691 and correlation of 0.577±0.130. Variable importance values for each input were calculated from the model-based results of the GBT model. The largest importance was achieved for income and the lowest was estimated for Covid-19 Infection. The calculated importances can be evaluated the potential impacts on the CAS score. In future works, different algorithms can be built for detailed predictions about COVID-19-related anxiety levels**

Index Terms— **Coronavirus anxiety scale, COVID-19, gradient boosting tree, regression task.**

I. Introduction

At the end of December 2019, a novel coronavirus was identified as the cause of pneumonia cases in Wuhan, China. The disease was later named COVID-19 (coronavirus disease 2019) and the causative agent was designated as SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2). The number of cases significantly increased within a short period and the disease has spread worldwide. COVID-19, which was declared as a global epidemic by WHO on March 11 due to its spread and the severity of the clinic, is the newest infection table caused by coronaviruses in humans [1, 2]. Fever, cough, and fatigue are the most common symptoms at the onset of COVID-19, but other symptoms include headache, hemoptysis, diarrhea, dyspnea, and lymphopenia [3]. The symptom list is expanding day by day. In more severe cases, causes pneumonia, Acute Respiratory Distress Syndrome (ARDS), multiple organ failure, and death [4].

With the COVID-19 pandemic first seen in China and then spread all over the world, humanity has perhaps entered a unique process in history. Education, commercial and social activities have been suspended almost all over the world. Restrictions on transportation, social isolation, quarantine measures and curfews have become a routine of life, and the concept of "new normal" has been introduced. Recent studies on the psychological and social effects of COVID-19 caused by the new coronavirus have also revealed that this disease causes radical changes in the vital conditions of many societies and is associated with negative psychological outcomes. As a result of all these, mental problems have emerged in society, especially anxiety disorders, depression, and post-traumatic stress disorder [5-7].

Methods that enable inference from data stacks and generation of information are included in the data mining discipline. Briefly, data mining is defined as the process of generating information by discovering patterns in data [8]. Data mining includes a combination of techniques from different disciplines such as database technology, statistics, machine learning, pattern recognition, neural networks, data visualization, and spatial data analysis [9].

Machine learning, one of these techniques, is a subfield of artificial intelligence that aims to make predictions about new data when they are exposed to new data by performing data-based learning [10]. Machine learning includes the design and development processes of algorithms that aim to perform data-based learning. The model created with the existing data set and the algorithm used is set up to give the highest performance. For this reason, many machine learning methods have been developed. Some of these approaches are capable of estimation and prediction (regression), some of them are clustering and some of them are capable of classification [11].

This study aimed to evaluate the potential negative effects of the scattered migrant worker population on the anxiety level by estimating the scores of the COVID-19 anxiety scale with the Gradient Boosting Tree (GBT).

EMEK GÜLDOĞAN, is with Department of Biostatistics and Medical Informatics, Faculty of Medicine, İnönü University, Malatya, Turkey,(e-mail: emek.guldogan@inonu.edu.tr).

https://orcid.org/0000-0002-5436-8164

Manuscript received March 9, 2021; accepted April 27, 2021.
DOI: 10.17694/bajece.893672

II. Material and Method

A. Public Dataset

The public dataset was received from the related website [12]. The data was constructed with the assistance of a questionnaire developed by employing the Corona Virus Anxiety Scale (CAS). Out of 1350 valid responses elicited through the telephone interview, we have established this knowledge. The data, which is considered to be the peak pandemic period, were collected from June to August 2020. We conclude that the time was critical because it was marked by the unstable environment with badly affected job markets and the government's fear of reintroducing travel restrictions on workers. The knowledge thus provides the interested researchers with insights and guidance to carry out our relevant studies on the mental health of disadvantaged groups such as migrant workers. The data will also encourage researchers and academics to examine the effect on the mental health of vulnerable groups of such pandemics and unpredictable events. This knowledge will allow the investigator to consider the potential adverse effects of the scattered population of migrant workers on the level of anxiety.

Age, gender, marital status, education, income, and Covid-19 infected or not status included in the relevant dataset were predictor variables and the output/target feature was CAS score.

B. The Coronavirus Anxiety Scale (CAS)

The risk of COVID-19 for migrant workers led to concerns with their mental health. The aim of public data collection, with a focus on migrant workers as the subject of the investigation, is to improve the users' awareness of the data on the effect of the pandemic on the mental health issues of internal migrant workers. Our research explores the psychological problems faced by migrant workers and the stress they encountered during the pandemic using the Corona Virus Anxiety Scale (CAS), which Lee originally developed in 2019. The scale uses four cognitive, mental, behavioral, and psychological aspects [13].

C. Gradient Boosted Tree (GBT)

GBT is a very effective and accurate machine learning algorithm. In many machine learning tasks, GBT achieves state-of-the-art efficiency, such as multi-class classification, click prediction, and learning to rank [14]. With the advent of big data (in terms of both the number of features and the number of instances) in recent years, GBT faces new challenges, especially in the trade-off between precision and performance. To estimate the information gain of all the possible split points, traditional GBT implementations need to search all data instances for each element. Therefore, both the number of features and the number of instances would be proportional to their computational complexity. When managing big data, this makes these implementations very time-consuming [15].

In the current study, the GBT model was constructed for predicting the CAS score of migrant workers based on input variables described earlier. Hyperparameters of the GBT model were tuned using Optimize Parameters (Evolutionary) operator, which seeks the optimum values of the selected parameters by an evolutionary computation approach [16]. Hyperparameters of the GBT model were 50 for the number of trees, 5 for minimal depth, 0.044 for learning rate, and 1.0E-5 for minimum split improvement.

D. Validation the Model

The GBT model was trained via the validation technique of bootstrapping. After bootstrapping a sample of training data set to estimate the statistical performance of a learning model, the bootstrapping validation operator performs validation and is primarily used to estimate how accurately a model can perform in reality. Performance metrics employed in this study were given below.

E. Performance Evaluation Criteria

Performance evaluation for the related models was assessed using root mean square error, absolute error, relative error lenient, squared error, and correlation. The elaborate results of the relevant formulas are described in the connected studies [17, 18].

III. Results

A total of 1500 people, 758 (56.1%) male, and 592 (43.9%) female, participated in this study. The number of those up to the age of 30 was 728 (53.9%), 345 (25.6%) of those aged 31-40, and 277 (20.5%) of those aged 41 and over. Singles are 823 (61.0), partners are 527 (39.0). 786 (58.2%) people did not receive a formal education, 446 (33.0) people dropped out and 118 (8.7%) people finished school. Those with income up to Rs 15000 are 415 (30.7%), those between Rs 15001-20000 are 877 (65.0%), and 58 (4.3%) are more than Rs 20000. The number of those with negative COVID-19 test is 1020 (75.6%), the number of positive ones is 330 (24.4%).

Table 1 summarizes the baseline characteristics of the subjects enrolled in the current study. The highest CAS score among the age categories was 31-40 years old [8.82+4.21 (8.37-9.26)], which was not significantly from the other age classes ($p=0.145$). CAS score for females enrolled in the study was significantly higher than that for males ($p<0.0001$). Similarly, median CAS scores were statistically significant between the categories of marital status and COVID-19 infection, and among the classes of income ($p<0.05$). However, the relevant scores were not significantly different among the education categories ($p=0.694$).

The importance levels of the variables concerning CAS score were presented in Table 2. The largest importance level (0.175) was calculated for income and the lowest level (0.004) was estimated for Covid-19 Infection.

Evaluation metrics for the proposed model were summarized in Table 3. GBT yielded a root mean square error of 3.547±0.235, the absolute error of 2.943±0.154, relative error lenient of 31.54%±0.82%, squared error of 12.623±1.691 and correlation of 0.577±0.130. The other metrics were presented in the related table.

TABLE I
BASELINE CHARACTERISTICS OF THE SUBJECTS ENROLLED IN THE CURRENT STUDY

Variable	Variable Categories	n	%	CAS Score [Mean ± S.D. (95% C.I. Median-(Min-Max)]	p
Age	Up to 30 Years old	728	53,9	8.33+4.4 (8.01-8.65)	0.145*
	31-40 Years old	345	25,6	8.82+4.21 (8.37-9.26)	
	41 and Above	277	20,5	8.72+4.04 (8.24-9.2)	
Gender	Male	758	56,1	8.11+4.54 (7.79-8.44)	<0.0001**
	Female	592	43,9	9.08+3.86 (8.76-9.39)	
Marital Status	Single	823	61,0	9.11+4.22 (8.82-9.4)	<0.0001**
	Partnered	527	39,0	7.64+4.23 (7.28-8)	
Education	No formal education	786	58,2	8.54+4.23 (8.24-8.83)	0.694*
	School drop outs	446	33,0	8.61+4.36 (8.2-9.02)	
	School completed	118	8,7	8.23+4.35 (7.44-9.02)	
Income	Up to Rs.15000	415	30,7	7.43+4.22 (7.02-7.84)	<0.0001*
	Rs. 15001-Rs. 20000	877	65,0	8.9+4.24 (8.62-9.18)	
	More than Rs. 20000	58	4,3	10.9+3.45 (9.99-11.8)	
COVID-19 Infection	Yes, tested positive	1020	75,6	12.39+2.6 (12.11-12.67)	<0.0001**
	No, not tested positive	330	24,4	7.29+3.97 (7.04-7.53)	

*: Mann Whitney U Test, **: Kruskal Wallis H test

TABLE II
THE IMPORTANCE LEVELS OF THE VARIABLES FOR CAS SCORE

Variable	Importance level
Income	0.175
Marital Status	0.086
Education	0.049
Gender	0.047
Age	0.016
Covid-19 Infection	0.004

TABLE III
EVALUATION METRICS FOR THE PROPOSED MODEL

Metrics	Value	Micro average
Root mean square error	3.547±0.235	3.553 ± 0.000
Absolute error	2.943±0.154	2.943 ± 1.990
Relative error lenient	31.54%±0.82%	31.55% ± 23.39%
Squared error	12.623±1.691	12.624 ± 14.861
Correlation	0.577±0.130	0.581

IV. Conclusions

The current study aimed to evaluate the potential negative effects of the scattered migrant worker population on the anxiety level by estimating the scores of the COVID-19 anxiety scale with the GBT model. In this context, this paper has presented a new GBT model for COVID-19 CAS score under the regression task. Moreover, the GBT model is employed for both regression and classification problems during the last years. An experimental analysis was performed to predict CAS score outcome based on the demographic and COVID-19 Infection data. The experimental findings demonstrated that the GBT yielded a root mean square error of 3.547±0.235, the absolute error of 2.943±0.154, relative error lenient of 31.54%±0.82%, squared error of 12.623±1.691 and correlation of 0.577±0.130. Therefore, the proposed model may be employed as an appropriate tool to predict classify CAS score concerning the COVID-19 pandemic. Additionally, variable importance values for each input were calculated from the model-based results of the GBT model. The largest importance was achieved for income and the lowest was estimated for COVID-19 Infection. The calculated importances can be evaluated the potential impacts on the CAS score. In future works, different algorithms can be built for detailed predictions about COVID-19-related anxiety levels.

References

[1] Ş. Alp and S. Ünal, "Yeni Koronavirüs (SARS-CoV-2) Kaynaklı Pandemi: Gelişmeler ve Güncel Durum," Flora Dergisi, vol. 25, 2020.
[2] M. Hasöksüz, S. Kiliç, and F. Saraç, "Coronaviruses and sars-cov-2," Turkish journal of medical sciences, vol. 50, pp. 549-556, 2020.
[3] W. J. Wiersinga, A. Rhodes, A. C. Cheng, S. J. Peacock, and H. C. Prescott, "Pathophysiology, transmission, diagnosis, and treatment of coronavirus disease 2019 (COVID-19): a review," Jama, vol. 324, pp. 782-793, 2020.
[4] H. A. Rothan and S. N. Byrareddy, "The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak," Journal of autoimmunity, p. 102433, 2020.
[5] P. Hyland, M. Shevlin, O. McBride, J. Murphy, T. Karatzias, R. P. Bentall, et al., "Anxiety and depression in the Republic of Ireland during

the COVID-19 pandemic," Acta Psychiatrica Scandinavica, vol. 142, pp. 249-256, 2020.
[6] N. Salari, A. Hosseinian-Far, R. Jalali, A. Vaisi-Raygani, S. Rasoulpoor, M. Mohammadi, et al., "Prevalence of stress, anxiety, depression among the general population during the COVID-19 pandemic: a systematic review and meta-analysis," Globalization and health, vol. 16, pp. 1-11, 2020.
[7] Y. Erdoğdu, F. Koçoğlu, and C. Sevim, "COVID-19 pandemisi sürecinde anksiyete ile umutsuzluk düzeylerinin psikososyal ve demografik değişkenlere göre incelenmesi," Klinik Psikiyatri Dergisi, vol. 23, 2020.
[8] I. H. Witten and E. Frank, "Data mining: practical machine learning tools and techniques with Java implementations," Acm Sigmod Record, vol. 31, pp. 76-77, 2002.
[9] D. J. Hand and N. M. Adams, "Data Mining," Wiley StatsRef: Statistics Reference Online, pp. 1-7, 2014.
[10] R. Polikar, "Ensemble learning," in Ensemble machine learning, ed: Springer, 2012, pp. 1-34.
[11] M. Atalay and E. Çelik, "Büyük Veri Analizinde Yapay Zekâ Ve Makine Öğrenmesi Uygulamalari-Artificial Intelligence and Machine Learning Applications in Big Data Analysis," Mehmet Akif Ersoy Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, vol. 9, pp. 155-172, 2017.
[12] S. K. G, Naveen; Jothi, M, "Data set on Impact of COVID-19 on Mental Health of Internal Migrant Workers in India: Corona Virus Anxiety Scale (CAS) Approach," V2 ed: Mendeley Data, 2021.
[13] R. Choudhari, "COVID 19 pandemic: mental health challenges of internal migrant workers of India," Asian journal of psychiatry, vol. 54, p. 102254, 2020.
[14] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, et al., "Lightgbm: A highly efficient gradient boosting decision tree," Advances in neural information processing systems, vol. 30, pp. 3146-3154, 2017.
[15] C. Krauss, X. A. Do, and N. Huck, "Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500," European Journal of Operational Research, vol. 259, pp. 689-702, 2017.
[16] R. Studio, "Visual workflow designer for the entire analytics team," ed, 2019.
[17] A. Niknafs, B. Sun, M. Richter, and G. Ruhe, "Predictions in Time Series with Repeated Patterns, Using Piecewise Linear Regression," Technical Report SEDS-TR-094/2011, University of Calgary2011.
[18] J. H. Chan, M. Joshi, R. Tang, and C. Yang, "Trinomial or binomial: Accelerating American put option price on trees," Journal of Futures Markets, vol. 29, pp. 826-839, 2009.

BIOGRAPHIES

EMEK GÜLDOĞAN obtained his BSc. degree in Computer Engineering from Middle East Technical University in 2001. He received MSc. degree in biostatistics and medical informatics from the Inonu University in 2005, and Ph.D. degrees in biostatistics and medical informatics from the Inonu University in 2017. He is currently working as an assistant professor of the Department of Biostatistics and Medical Informatics at Inonu University and as the information processing manager at Turgut Özal Medical Center. His research interests are cognitive systems, data mining, machine learning, deep learning.