Tarbiat Modares University

The International Journal of Humanities

Received: 2 July 2021
Accepted: 11 March 2022
Published: 2 October 2022

[1] Associate Professor, Department of Remote Sensing and Geographical Information System, Faculty of Geography, University of Tehran, Iran.
E-mail: nneysani@ut.ac.ir

[2] Department of Remote Sensing and Geographical Information System, Faculty of Geography, University of Tehran, Iran (Corresponding Author).
E-mail: Mehdifarrokhi7@gmail.com

How to cite this article:
Neysani Saman, Najmeh; Farokh Anari, Mehdi (2022). The Prediction of Low and High-Risk Zones of Tehran during COVID-19 by Using the Random Forest Algorithm, *The International Journal of Humanities* (2022) Vol. 29 (4): (23-35).

https://eijh.modares.ac.ir/article-27-31831-en.html

RESEARCH ARTICLE

The Prediction of Low and High-Risk Zones of Tehran during COVID-19 by Using the Random Forest Algorithm

Najmeh Neysani Samani[1], Mehdi Farokh Anari[2]

Abstract: **The Coronavirus disease (Covid-19) is one of the infectious and contagious ones called 2019-nCoV acute respiratory disease. Its outbreak was first reported on December 31, 2019, in the Chinese city of Wuhan that quickly spread throughout the country within a few weeks and spread to several other countries, including Italy, the United States, and Germany, within a month. This disease was officially reported in Iran on February 19, 2020. It is important to detect and analyze high risk zones and establish regulations according to the data and the analyses of Geographic Information System (GIS) in epidemiological situations. Meanwhile, the GIS, with its location nature, can be effective in preventing the breakdown of Covid-19 by displaying and analyzing the dangerous zones where people infected with the disease. In fact, recognizing regions based on the risk of getting the disease can influence social restriction policies and urban movement rules in order to prepare daily and weekly plans in different urban regions. In this applied and analytical research, high and low risk zones of Tehran have been identified by using the random forest algorithm which is used for both classification and regression. The algorithm builds decision trees on data samples and then predicts data from each of them, and finally chooses the best solution. In this research, 7 effective criteria have been used in the level of risk of regions toward Covid-19 virus, which is: subway paths and bus for rapid transits, hospitals, administrative and commercial complexes, passageways, population densities and urban traffic. After providing the map of high-risk zones of Covid-19, the Receiver Operating Characteristic curve (ROC) has been used for evaluation. The area under the curve (AUC) obtained from ROC shows an accuracy of 98.8%, which means the high accuracy of this algorithm in predicting high and low zones toward getting the Covid-19 disease.**

Keywords: Covid-19; Location Analysis; Random Forest Algorithm; Epidemiology.

Introduction

Nowadays, doctors pay attention to the treatment of most diseases, and almost with the development of medicine, the solution for those diseases has been determined. The epidemics of disease bring great problems to the medical community and people alike. From the end of 2019 until now, a new virus from the coronavirus family has been threatening the human society. Concerns about this new virus are because of the fact that the world has faced a dangerous epidemic for the third time (Wilson and Chen, 2020). Epidemiology has various applications in the field of geographical identification of vulnerable regions, one of its main applications is to facilitate the identification of geographical regions and vulnerable groups who are exposed to infectious diseases (Ghaedamini Asadabadi, et al. 2012). In fact, geographic epidemiology is a part of descriptive epidemiology that pays attention to the geographical aspects of an epidemic (Elliottet et al. 2006). According to this issue, the risk of Covid-19 still exists in societies, and one of the most important factors of its outbreak is gathering and scattering of people suffering from the disease, The geographic information system (GIS) can help the medical community by preventing the breakdown of the Covid-19 virus. Therefore, the most important part of controlling epidemic diseases is controlling the population and social distancing using the location data of the people suffering from the disease.

One of the important tools to control the population and social distancing is statistical data, and since the statistical tables are not capable of showing the location of distancing as well as high and low risk zones, the maps have a significant performance in this field (Bell & Broemeling, 2000; Cliff, 1995). The GIS is a system that processes the data of the reference location and gathers those related to the phenomena, which is somehow associated with the location (Rezaeiyan, 2004). The applications of the GIS in the field of health include crisis management, analyzing traffic accidents, studying and analyzing spatial data showing the relationship between the outbreak of diseases and vulnerable factors of the environment, designing and implementing health programs, epidemiologic studies of parasitology, determining the coverage of vaccination and immunization programs as well as determining the pattern of any disorders and disabilities in vulnerable populations (Erdogan et al, 2008; Moss, et al, 2006).

However, the geographic information system is not a perfect solution for the prevention and awareness of citizens about the disease and its outbreak process, however it can be said that by using computers and mobile phones, today the GIS can allow users analyze the breakdown of the disease in his/her geographical region (Field and Grigsby, 2002). In other words, the GIS is a tool to gather, store, integrate, manage, recycle, analyse and also show the location data, which can be used in epidemiologic research and health policies (Odwyer & Burton, 1998; Scholten, & De Lepper, 1991). Predicting methods for geographic issues and crises have developed greatly in recent years. The random forest algorithm is among the innovations used in this research, which can predict with a high percentage. This algorithm has been used to predict geographical issues such as landslides and floods. But this algorithm has not been used in the field of Covid-19 until now. The main purpose of this research is to analyze and predict hot and cold zones by using the random forest algorithm in order to prepare a daily plan. In this research, 7 effective criteria for the infection rate of the Covid-19 virus were used in order to predict hot and cold zones, which are: subway paths and bus rapid transits, hospitals, administrative and commercial complexes, passageways, population densities and urban traffic, and also, patients infected and suspected to covid-19 have been used as an independent criterion.

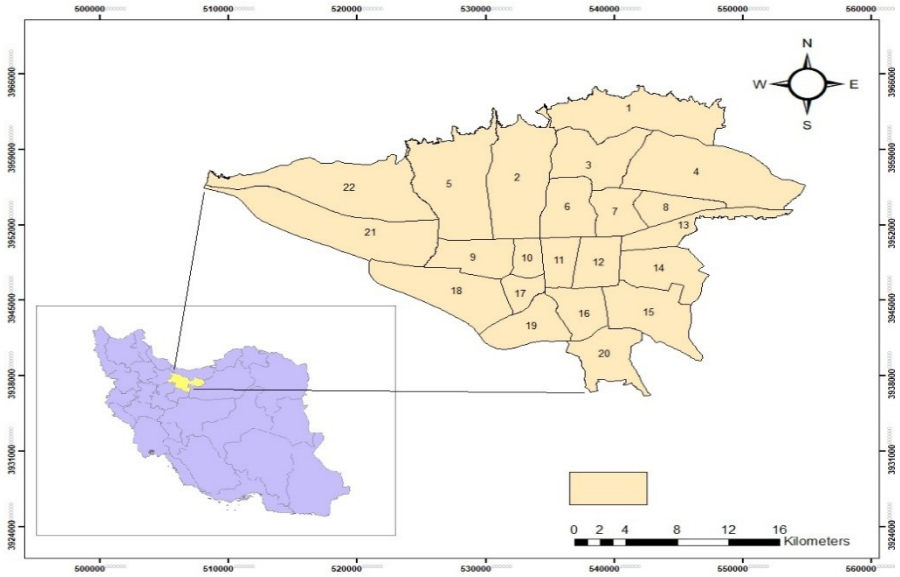

Fig. 1. Study Area (Tehran)

Studied Region

The metropolis of Tehran is geographically located from 51°17 E to 51°33 eastern longitudes and from 35°33 to 35°44 N northern latitude (Fi. 1) (Tehran Municipality website). The climate of this region is mainly influenced by the altitude, so it is more than 3000 meters in the northern part and 900 meters in the southern part. Tehran is the political center and the capital of Iran and is located on a relatively flat plain with an area of 68,995 hectares and a population of 8,737,510 persons (Iran Statistical Yearbook, 2016-2017). This city is the most important economic hub in the country.

Methodology

The type of current study is applied research in terms of purpose. The methodology of this research is descriptive-analytical. The process of this research is shown in the flow chart, Figure 2. Choosing influential criteria in the breakdown of corona disease in this research 7 very important criteria were determined by experts, and the latest articles related to corona were extracted, and their maps were drawn.

The identification of influential criteria is required for the procedure of research and the determination of hot and cold zones; therefore, the influential factors in this field were extracted through performing a library study by the writers and considering the social and natural conditions of Tehran. The independent criteria were considered: active subway paths and bus rapid transits, the hospitals for Covid-19 patients, active administrative and commercial complexes, the passageways of citizens, population densities in the regions of Tehran and urban traffic. The dependent criteria which are considered in this research: Covid-19 patients were identified by location and areas that had low risks of Covid-19, according to the opinion of an expert. To implement the research, in the first step of the information layers, the mentioned criteria were provided from the system of Tehran Municipality and the system of the Ministry of Health and were applied to the maps to perform location analysis by using the analytical distance function. In the second step, the aforementioned information layers were normalized through the following formula to avoid influencing each other and to convert to zero and one values:

$$x = \frac{\mathbf{xi - x\ min}}{\mathbf{x\ max - x\ min}}$$

In the mentioned formula, x is the normalized value of the considered layer, xi is the value of the input layer, X min is the minimum value of the layer, and X max is the

maximum value of the considered layer. In the next step, the criteria influencing the increase in the breakdown of Covid-19 are given as the input layers to the network.

In the third step, 7 mentioned parameters were introduced as input to the random forest algorithm. After preparing the input indicators, 988 training samples were selected to provide training points. These training points are divided into three parts: The first part is for training, the second part is for stopping the calculations when the error increases, and the third part is for verifying the random forest. From the 988 training points, 70% points (691 points) are used for training the algorithm, 15% (148 points) for the validation, which are used to calibrate the model, and finally, the remaining 15% of points (148 points) are used for evaluation and conclusions. Figure 4 shows the distribution of the training points of the random forest algorithm. After selecting the inputs and training points of the network, the number of hidden decision trees is determined. There are many methods for choosing the number of tree sets; one of the most efficient methods is the trial and error method which was used in this research. In this method, the best choice for the number of two sets of decision trees with 1000 trees has the best performance. Finally, the map of hot and cold zones has been identified, and also, in the end, the training data has been tested and evaluated through the ROC diagram, which is provided in an acceptable way.

Random Forest Algorithm (RF1)

The random forest algorithm has usability in the issues related to classification, regression and unsupervised training (Liaw and Wiener, 2002). Random forests are a modern type of basic tree that includes a lot of classification and regression trees (Breiman, 2001). This method has been used in the different fields of location analyses, including processing and classifying the satellite and aerial images [Ghasemi, et al. 2016; Pal, 2005], landslide risk zoning (Trigila et al., 2015), the potentiometric and Vulnerability Evaluation of groundwater (Rahmati et al., 2016), and always has an appropriate performance. The predicting model of the random forest is established based on the averaging of the results derived from all related decision trees and classifies accurately for many sets of data (Ebrahimkhani et al. 2010).

In other words, in a random forest with an input vector, each tree is classified and the output is labeled with classes that the majority consider acceptable.

Data Analysis

The covid-19 disease requires social distancing and informing people about hot and cold zones due to its high contagiousness. Therefore, in this research, 7 criteria are used to identify hot zones daily. Seven effective and key parameters include the distance from urban passageways (Fig. 3a), population density (Fig. 3b), distance from commercial and administrative complexes (Figure 3c), distance from BRT or rapid bus transits (Figure 3d), distance from the hospitals admitting Covid-19 patients (Fig. 3c), the distance from the subway (Figure 3c), traffic (Figure 3h). And also, the dependent criterion that patients suspected of covid-19 has been selected as the training dependent criterion of the algorithm (Fig. 4).

The data related to BRT, subway, urban passageways, and administrative centers were obtained from the website of Tehran Municipality and mapped in this study. The data relating to the number of infected and suspected cases of Covid-19 was extracted from the Mask application, which is the data approved by the Ministry of Health of Iran. This data is related to May 29, 2020. The official statistics of the patients of this day is 224 patients in Tehran, although, in this research, the data related to suspected individuals of the disease were also used. The data related to traffic and Covid-19's hospitals (admission of Covid-19 patients) were obtained from the traffic control company and the website of the Ministry of Health, respectively. The data relating to the patients was recorded on May 29, 2020, because newer data is not available by location and city separation. The reason for using the mentioned 7 criteria is that the crowd in these areas is more than in other areas, and considering that the World Health Organization has stated that social distancing is one of the important things in the control of Covid-19 (WHO, 2020), and therefore, we have used it.

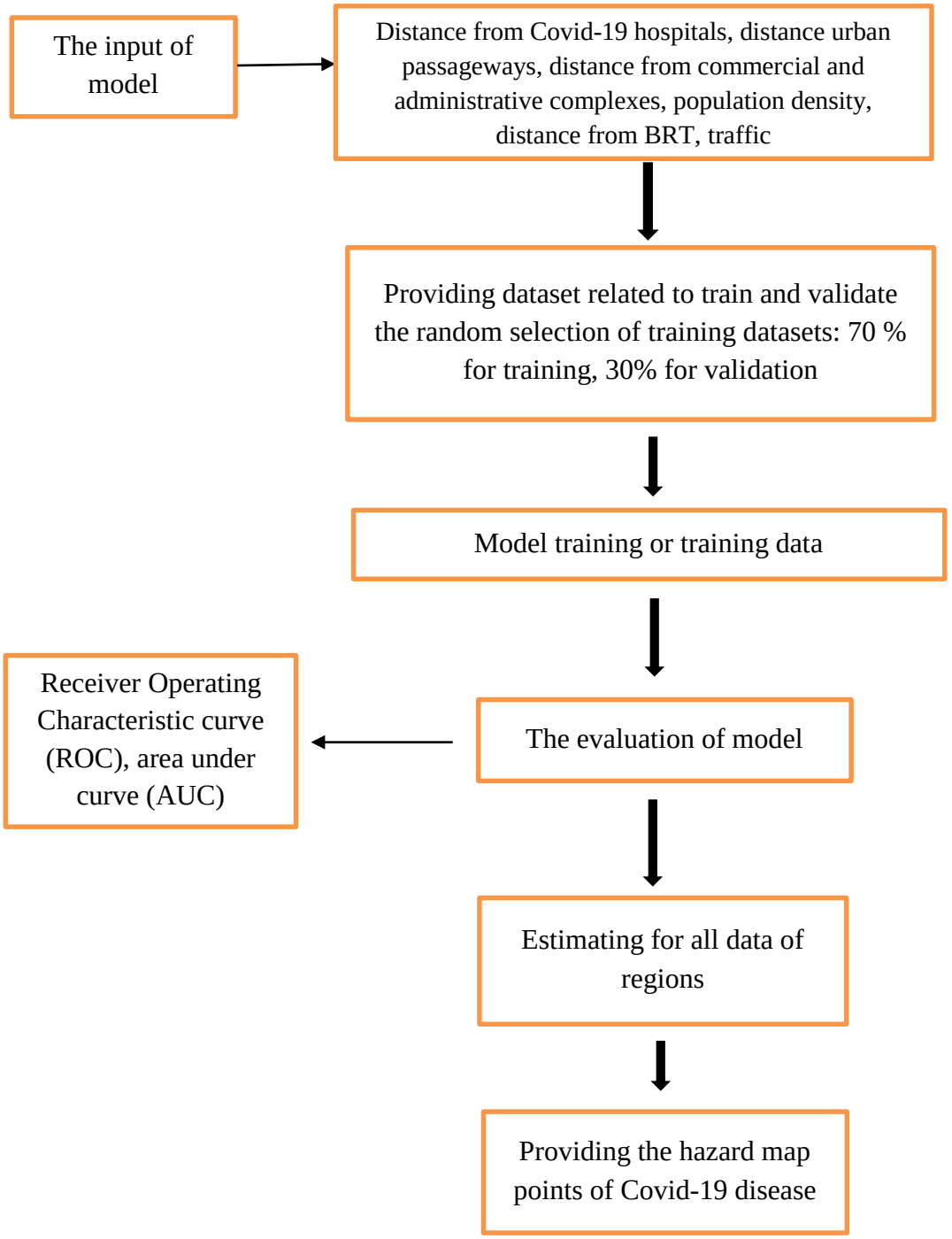

Fig. 2. The flow Chart of the Implementation Process of the Random Forest Algorithm

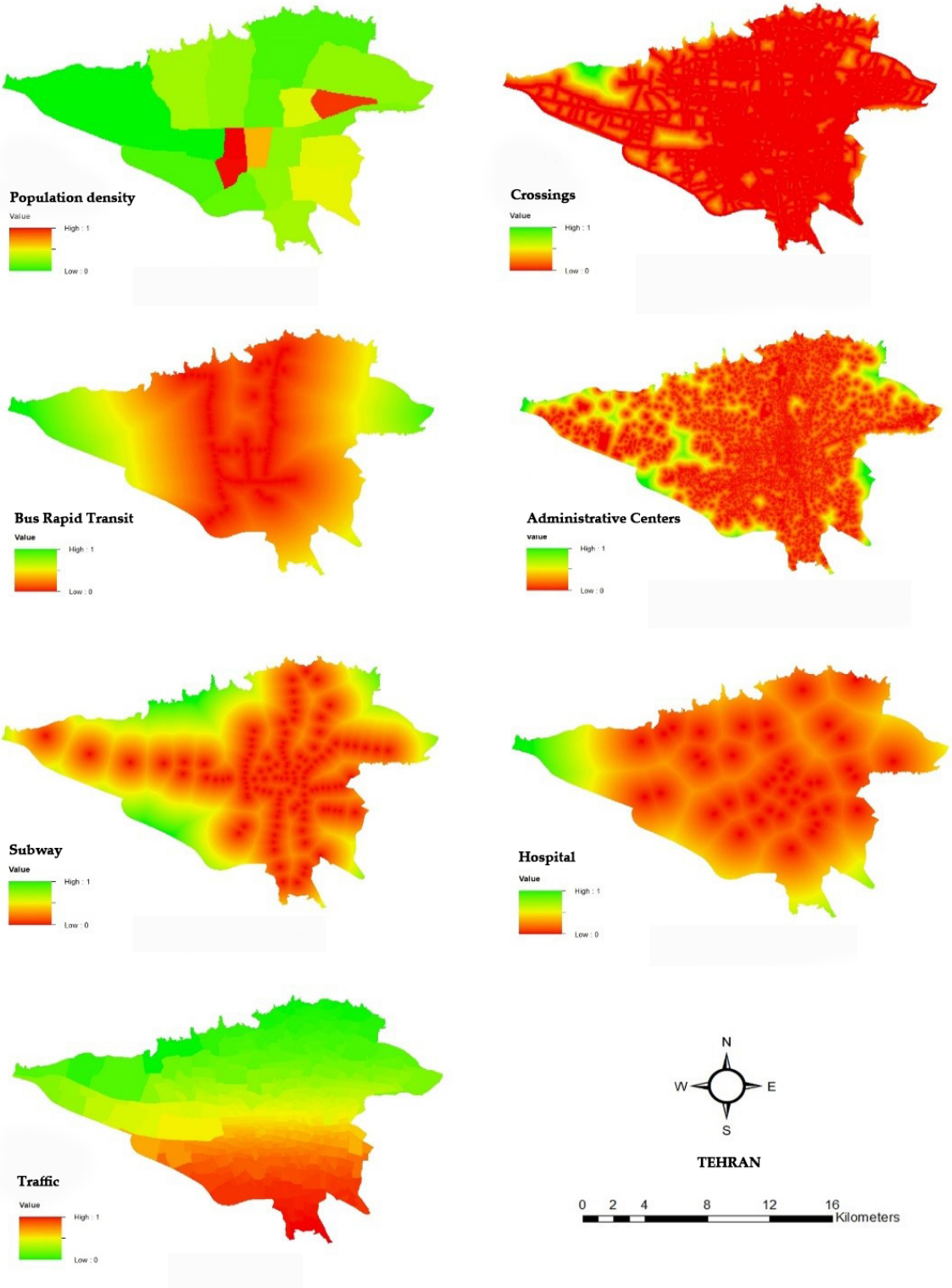

Fig. 3. Criteria Affecting the Risky Areas of the Covid 19 Disease

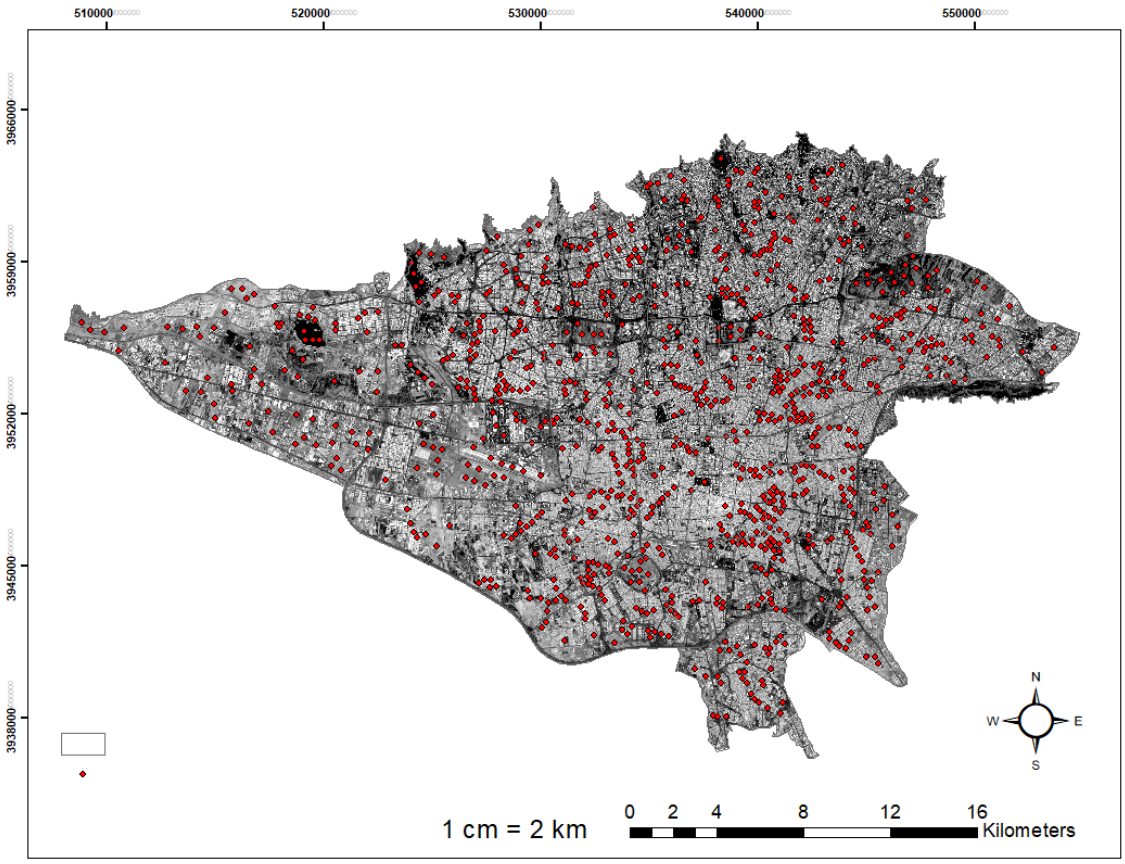

Fig. 4. Training Examples of Random Forest Algorithm

Results and Discussion

Results derived from the random forest bring to the hazard map of the Covid-19 virus (figure 5). Derived results into 5 classes which are: safe, less dangerous, medium, dangerous, and very dangerous. The results show that the most dangerous area is the center area of Tehran. The Receiver Operating Characteristic Curve (ROC) is used to evaluate the accuracy of the random forest. This curve is one of the important criteria for the evaluation of the performance of classified or multi-layer models. This appropriate criterion can measure the models at different thresholds. In fact, this curve is a curve based on possibility and is one of the most efficient methods in providing the characteristic of the determination, possible identification and prediction of systems which shows the accuracy in a qualitative form (WHO, 2020). The ROC curve is a curve in which the Y axis is formed by TPR, and the X axis is formed by FPR. TPR is the abbreviation of True Positive Rate, and it means the correct rate, which is also called sensitivity. The value of TPR is calculated through the following formula:

$$\text{TPR} = \frac{\mathbf{TP}}{\mathbf{TP + FN}}$$

The specificity of this curve is that its value depends on the following relation:

$$\text{Specificity} = \frac{\mathbf{TN}}{\mathbf{TN + FP}}$$

FPR is the component of the X-axis in the Receiver Operating Characteristic Curve, which is the abbreviation for False Positive Rate means the incorrect positive rate and selected test values, whose value can be calculated through the following relations:

$$\text{FPR} = 1 - \text{Specificity} = \frac{\mathbf{FP}}{\mathbf{TN - FP}}$$

The best model is one in which Receiver Operating Characteristic Curve is close to one. This means that the closer it is to one, the more accurate and appropriate the measurement is. Based on the Receiver Operating Characteristic Curve, the value of the area under the curve (AUC) of the studied area using the random forest algorithm was estimated to be 2.98, which shows the very good evaluation of the random forest algorithm in identifying hazard zones of Covid-19 (Fig. 6).

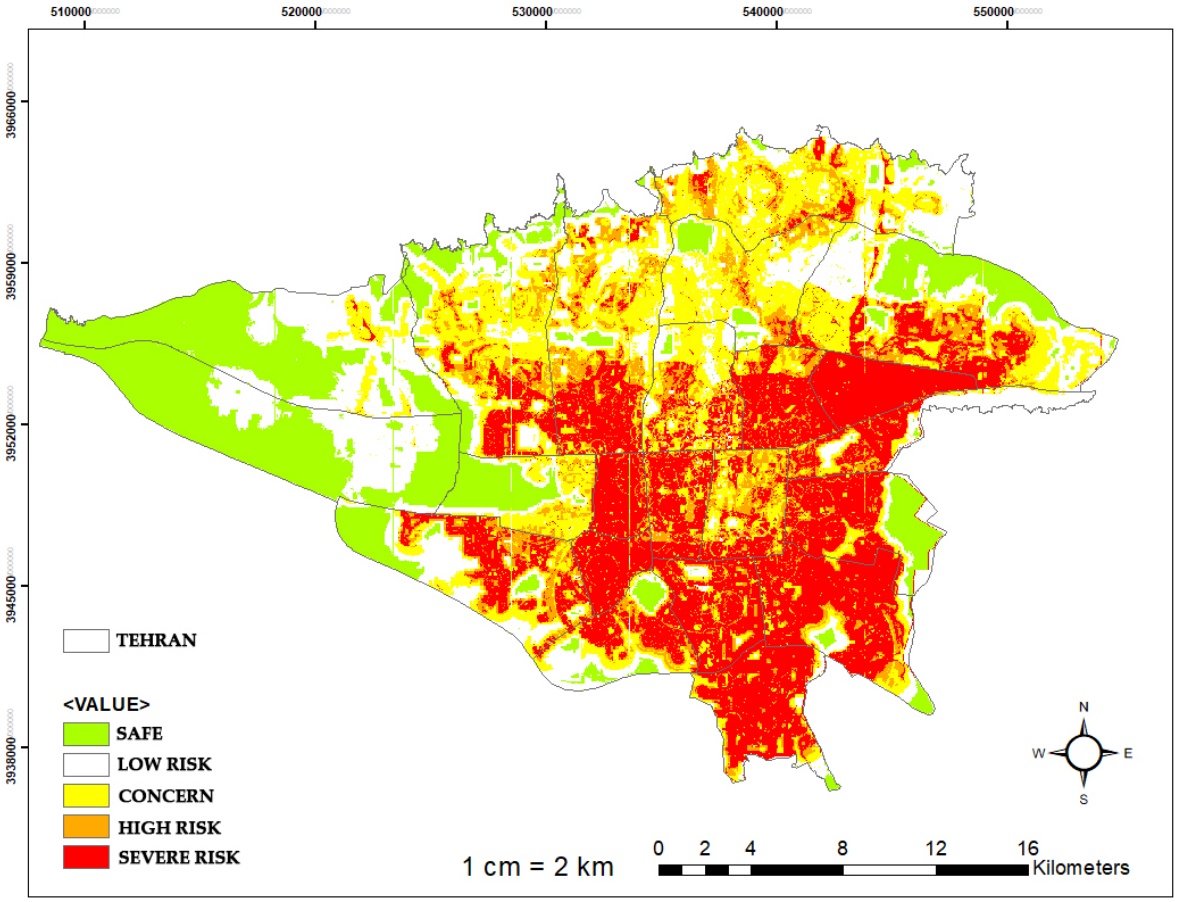

Fig. 5. Classification of Dangerous Areas of the Covid-19 virus

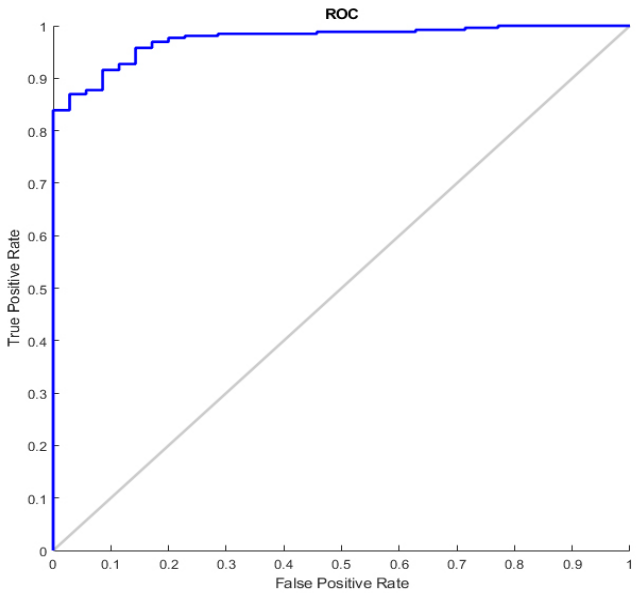

Fig. 6. Relative performance detection curve

Conclusion and Suggestions

The covid-19 virus has changed the lifestyle of people in society, the economy and most importantly, the type of transportation due to the severity of its contagion. Meanwhile, there are several groups of people influencing the prevention of the breakdown of Covid-19. The Geographical Information System science can be a solution to social distancing and rules for this issue. The influential individuals are people and government institutions. Predicting hot zones in order to prevent people from entering hot zones and establishing Covid-19 policies in order to block areas so that people do not enter those areas, and informing people about this issue can be perfect solutions and help to prevent the breakdown of the disease. It will be effective to create different software in order to provide information in line with the prediction. In this research, using different tools such as the random forest algorithm and the geographic information system (GIS), hot zones where the possibility of the outbreak is high were discovered and displayed on the map. By the geographic prediction of the hazard of Covid-19, policies can be adopted daily or weekly to control crowds, and on the other hand, people can avoid visiting hot zones. For example, by providing a daily hazard map using the random forest algorithm, it is possible to change people's travel plans to these areas and control the disease outbreak.

References

[1] Ebrahimkhani, Somayeh; Afzali, Mehdi; Shokoohi, Ali, (2010). Prediction and investigation of road accidents factors using data mining algorithms, *Danesh Entezami Zanjan*, Vol. 1, Issue 1, January 2012, 111-127 (in Persian).

[2] Ghasemi, Akbar; Fallah, Asghar; Shataee Joibari, Shaban, (2016). Evaluation of four algorithms for estimation of canopy cover of mangrove forests by using aerial imagery, *RS & GIS for Natural Resources*, Vol. 7, Issue 2, Summer 2016: 1-15 (in Persian).

[3] Statistical Center of Iran, (2016-2017). Iran Statistical Yearbook

[4] Bell, B.; & Broemeling, L., (2000). "A Bayesian analysis for spatial processes with application to disease mapping", *Stat Med*, 19: 974–989.

[5] Breiman, L., (2001). Random forests Machine Learning 45(l), 5-32.

[6] Chen W, Xie X, Wang J, Pradhan B, Hong H, Bui DT, Duan Z, Ma J., (2017). A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility. Catena, 151: 147-160.

[7] Cliff, A., (1995). "Analyzing geographically related disease data", *Stat Methods Med Res*, 4: 93–101.

[8] Elliott, P.; Cuzik, J.; English, D.; & Stern, R., (1996). *Geographical and environmental epidemiology*, 1st edition, England, Oxford University Press.

[9] Erdogan S, Yilmaz I, Baybura T, Gullu M. Geographical information systems aided traffic accident analysis syste case study: City of Afyonkarahisar. Accid Anal Prev. 2008; 40(1): 174-81. Doi: 10.1016/j.aap.2007.05.004.

[10] Field MJ, Grigsby J., (2002). Telemedicine and remote patient monitoring. JAMA. 2002; 288(4):423-5. doi:10.1001/jama.288. 4.423

[11] Ghaedamini Asadabadi, R.; Tofighi, S.; Ghaedamini, H.; Azizian, F.; Amerieon, A.; & Shokri, M., (2012). "A review of some infectious diseases distribution based on geographic information system (GIS) in the area of Chahar Mahal and Bakhtiari", *Journal of Police Medicine*, 1(2), PP: 113-123.

[12] Liaw A, Wiener M., (2002). Classification and regression by randomForest. R news, 2(3): 18-22.

[13] Moss MP, Schell MC, Goins RT. Using GIS in a first national mapping of functional disability among older American Indians and Alaska Natives from the 2000 census. Int J Health Geogr. 2006; 5:37. Doi: 10.1186/1476- 072X-5-37

[14] Odwyer, L.; & Burton, D., (1998). "Potential meets reality: GIS & public health research in Australia", *Aust J Public Health*, 22, PP: 819–823.

[15] Pal M., (2005). Random forest classifier for remote sensing classification. International Journal of Remote Sensing, 26(1): 217-222.

[16] Rahmati O, Pourghasemi HR, Melesse AM., (2016). Application of GIS-based data driven random forest and maximum entropy models for groundwater potential mapping: a case study at Mehran Region, Iran. Catena, 137: 360-372.

[17] Rezaeian M., (2004). An introduction to the practical methods for mapping the geographical morbidity and mortality rates. Tollo-e-Behdasht 2004; 2-41-51 (in Persian).

[18] Scholten, H.J.; & De Lepper, M.J., (1991). "The benefits of the application of geographical information systems in public & environmental health", *World Health Stat Q*, 44: 160– 170.

[19] Trigila A, Iadanza C, Esposito C, Scarascia- Mugnozza G., (2015). Comparison of Logistic Regression and Random Forests techniques for shallow landslide susceptibility assessment in Giampilieri (NE Sicily, Italy). Geomorphology, 249: 119-136.

[20] Wilson M. E, Chen L. H., (2020)Travellers give wings to novel coronavirus (2019-nCoV). JTM. 2020. doi:10.1093/jtm/taaa015.

[21]. World Health Organization (WHO) (2020). Coronavirus disease 2019 (COVID-19) Situation Report.

[22] Tehran Municipality website, https://www.tehran.ir

The International Journal of Humanities (2022) Vol. 29 (4): (23-35)

The International Journal of Humanities

پیش‌بینی منطقه‌های کم‌خطر و پرخطر تهران نسبت به بیماری کووید ۱۹ با بهره‌گیری از الگوریتم جنگل تصادفی

نجمه نیسانی سامانی[۱] و مهدی فرخ اناری[۲]

تاریخ دریافت: ۱۴۰۰/۸/۱۱
تاریخ پذیرش: ۱۴۰۰/۱۲/۲۰
تاریخ انتشار: ۱۴۰۱/۷/۱۰

[۱] دانشیار گروه سنجش‌ازدور و سیستم اطلاعات جغرافیایی، دانشکده جغرافیا، دانشگاه تهران، تهران، ایران.
E-mail: nneysani@ut.ac.ir

[۲] گروه سنجش‌ازدور و سیستم اطلاعات جغرافیایی، دانشکده جغرافیا، دانشگاه تهران، ایران.
E-mail: Mehdifarrokhi7@gmail.com

چکیده: کووید ۱۹ یکی از بیماری‌های عفونی و واگیردار است که بیماری تنفسی حاد ان‌کاو-۲۰۱۹ نامیده می‌شود. گسترش بیماری کوید ۱۹ اولین‌بار در ۳۱ دسامبر سال ۲۰۱۹ در ووهان چین گزارش شد که طی چند هفته، به‌سرعت در سرتاسر چین و طی ۱ ماه به چندین کشور دیگر همچون ایتالیا، ایالات متحده امریکا و آلمان گسترش یافت. این بیماری در ایران به صورت رسمی در ۳۰ بهمن ۱۳۹۸ تأیید شد. شناسایی و تحلیل مناطق پرخطر و ایجاد مقررات باتوجه به داده‌ها و تحلیل‌های سیستم اطلاعات جغرافیایی (GIS) در شرایط اپیدمیولوژیک اهمیت دارد. در این میان سیستم اطلاعات جغرافیایی با ماهیت مکانی خود می‌تواند در جلوگیری از گسترش ویروس کووید ۱۹ با نمایش و تحلیل مناطق خطرناک در ابتلا شدن افراد، مؤثر باشد. شناخت مناطق براساس میزان خطر ابتلا به بیماری می‌تواند برای ارائه سیاست‌های محدودیت‌گذاری اجتماعی و قوانین تردد شهری به‌منظور تهیه برنامه روزانه و هفتگی در مناطق مختلف شهری مؤثر باشد. در این پژوهش کاربردی و تحلیلی، با استفاده از الگوریتم جنگل تصادفی به شناسایی مناطق پرخطر و کم‌خطر در شهر تهران پرداخته شده است. در این پژوهش از ۷ معیار مؤثر در خطرپذیری مناطق نسبت به ویروس کووید ۱۹ استفاده شده است که عبارت‌اند از: مسیرهای مترو و اتوبوس‌های تندرو، بیمارستان‌ها، مراکز اداری و تجاری، معابر، تراکم جمعیت و ترافیک شهری. پس از تهیه نقشهٔ مناطق پرخطر ویروس کووید ۱۹، برای ارزیابی از منحنی تشخیص عملکرد نسبی (ROC) استفاده شده است. سطح زیر منحنی (AUC) به‌دست‌آمده از منحنی تشخیص عملکرد نسبی، نشان‌دهندهٔ دقت ۹۸/۸ در صد، است که نشان‌دهندهٔ دقت بالای این الگوریتم در جهت پیش‌بینی مناطق پرخطر و کم خطر نسبت به ابتلای بیماری کووید ۱۹ است.

واژه‌های کلیدی: کووید ۱۹، تحلیل مکانی، الگوریتم جنگل تصادفی، اپیدمیولوژی.