MACHINE LEARNING MODELS FOR PREDICTION OF COVID-19 INFECTION IN NORTH MACEDONIA

Natasha Stojkovikj[1], Maja Kukusheva Paneva[2], Cveta Martinovska Bande[1], Dushan Bikov[1]

Abstract. COVID-19 is a respiratory illness caused by the SARS-CoV-2 virus that affected the population worldwide and has undoubtedly been one of the most devastating global crises this century. This pandemic had a significant impact on healthcare system, economics, and daily social life and has been a top priority for governments worldwide. North Macedonia, particularly, has experienced a significant burden, with a notably high percentage of deaths and periods where the healthcare system teetered on the brink of collapse. Moreover, asymptomatic cases delayed or missed diagnoses increasing the risk of COVID-19 infection. In this paper machine learning (ML) algorithms to predict the spread of COVID-19 are used. These approaches offer promising ways for diagnosing and prognosing patients affected by COVID-19 pandemic. ML algorithms can analyze large volumes of data to forecast disease outbreaks and identify individuals at high risk of negative outcomes. A diverse set of learning algorithms, including Decision Tree, Support Vector Machine (SVM), Naïve Bayes, Multilayer Perceptron, Logistic Regression and Random Forest (RF) are used. By using these advanced techniques and epidemiological data, this paper aims to develop accurate and reliable models for forecasting the spread of COVID-19. These algorithms are trained and tested on an epidemiology data set that contains only positive COVID-19 cases in Nort Macedonia obtained from Public Health Institute of North Macedonia. The data set encompasses 14 features, including 2 demographic variables (age and gender) and 12 clinical indicators: pregnancy, pneumonia, cardiovascular diseases (CVDs), diabetes, pneumonia, hepatitis, neuromuscular, hypothyroid/ Hashimoto's, immunodeficiency/ HIV, cancer, chronic kidney disease (CKDs) and the outcome (deceased or recovered).

The performance of the classification models was evaluated with commonly used metrics precision, recall and F1 score. Precision is the ratio of correctly predicted positive cases to the total predicted positive cases. It measures the accuracy of positive predictions. Recall is the ratio of correctly predicted positive cases to all cases in actual class. It measures the ability of the model to find all the relevant cases within a dataset. The F1 score is the harmonic mean of precision and recall. It combines both precision and recall into a single metric.

Our analyses show that Support Vector Machines and Multilayer Perceptron with precision 76%, recall 75% and F1 score 75% have better evaluation values than the other classifiers. Similar results are obtained with Naïve Bayes (precision 74%, recall 74% and F1 score 74%), Logistic Regression (precision 71%, recall 71% and F1 score 71%), and Decision Trees (precision 71%, recall 70% and F1 score 70%). The research demonstrated that Machine Learning (ML) can achieve a notable degree of accuracy in predicting COVID-19 outcome.

Keywords. Machine Learning (ML), COVID-19, pandemic, dataset.

[1]N. Stojkovikj and C. Martinovska Bande and D. Bikov

Faculty of Computer Science,

Goce Delcev University, Stip,

N. Macedonia

e-mail: natasa.stojkovik@ugd.edu.mk, cveta.martinovska@ugd.edu.mk, dusan.bikov@ugd.edu.mk

[2]M. Kukusheva Paneva

Faculty of Electrical Engineering

Goce Delcev University, Stip,

N. Macedonia

e-mail: maja.kukuseva@ugd.edu.mk