# Chapter 17
# Comparison of Machine Learning Algorithms in Predicting the COVID-19 Outbreak

**Asiye Bilgili**
https://orcid.org/0000-0002-9776-5119
*Halic University, Turkey*

## ABSTRACT

*Health informatics is an interdisciplinary field in the computer and health sciences. Health informatics, which enables the effective use of medical information, has the potential to reduce both the cost and the burden of healthcare workers during the pandemic process. Using the machine learning algorithms support vector machines, naive bayes, k-nearest neighbor, and C4.5 algorithms, a model performance evaluation was performed to identify the algorithm that will show the highest performance for the prediction of the disease. Three separate training and test datasets were created 70% - 30%, 75% - 25%, and 80% - 20%, respectively. The implementation phase of the study was carried out by following the CRISP-DM steps, and the analyses were made using the R language. By examining the model performance evaluation criteria, the findings show that the C4.5 algorithm showed the best performance with 70% training dataset.*

## INTRODUCTION

In Wuhan, Hubei state in China, cases of pneumonia of unknown causes began to appear as of December 2019. According to World Health Organization (WHO), a pandemic occurs when above-normal number of cases are observed in a region or community (Definitions: Emergencies, 2021). WHO declared COVID-19 as a pandemic in March 2020 as several cases were observed in many countries. From the beginning of the epidemic until August 9, 2021, WHO has reported 4.293.591 deaths and 202.608.306 confirmed cases of COVID-19 worldwide (WHO Coronavirus (COVID-19) Dashboard, 2021). In these

DOI: 10.4018/978-1-7998-8674-7.ch017

confirmed cases, no symptoms were detected that could clearly distinguish a specific COVID-19 from other respiratory tract infections, and symptoms of fever, cough, fatigue, joint pains, loss of smell and taste, and shortness of breath were generally observed in COVID-19 cases (Wu, et al., 2020). The pandemic remains serious despite the measures taken and the vaccinations developed. Significant delays in treatments due to the initial diagnostic kit insufficiency, increased the necessity of receiving support from the interdisciplinary field of health informatics. Techniques such as non-clinical data mining, artificial intelligence, machine learning algorithms, expert systems in health informatics have a critical role in COVID-19 case diagnosis increasing the number of studies carried out in this area (Muhammad, et al., 2020). Although machine learning techniques were widely used in the field of computers in the beginning, it has been extensively used in a wide variety of fields such as education, health, justice, arts, manufacturing, marketing and automotive in recent years (Fry, 2019). In the health sector, machine learning techniques are used in different areas such as identification of effective genes in diseases, assessment of diabetes risk and early diagnosis of diseases.

## BACKGROUND

### Machine Learning

Machine learning, a sub-branch of artificial intelligence (AI), is a system that investigates the study and construction of algorithms that can make predictions on data for multidimensional biomedical and mathematical data analysis (Alwabel & Zeng, 2021). Machine learning algorithms work on the principle of creating a model to make data-based predictions and decisions based on existing examples, instead of following program instructions one-to-one like rule-based algorithms (Naqa & Murphy, 2015). The machine learning process, which starts with the collection of data from different sources, generates models based on the training data received by the algorithms. In other words, when the machine learning algorithm is trained with data, the machine learning model emerges. For example, a predictive algorithm creates a predictive model (Data Science and Machine Learning, 2021). Machine learning algorithms are basically classified in four ways as supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning.

### Supervised Machine Learning

Supervised machine learning is provided by training the algorithm through labeled data. While most of the available data is used to train the algorithm, the remaining is used to test the trained algorithm (Kotsiantis, 2007). Support Vector Machine, Artificial Neural Networks, Naive Bayes, K-Nearest Neighbor, Linear Regression, Logistic Regression, Principal Component Regression, Decision Tree, Random Forest Algorithms are among the commonly used controlled machine learning algorithms (Bilgin, 2017). Supervised machine learning algorithms are used in both classification and regression problems. Determining the class of an observation whose class is unknown is called classification. On the other hand, modeling mathematically the relationship between one or more independent variables and the dependent variable is called regression (Özkan & Selçukcan Erol, 2017; Metlek & Kayaalp, 2020). In this context, if supervised machine learning algorithms are divided into classification and regression, Support Vector Machines, Decision Tree, Naive Bayes, K-Nearest Neighbor Algorithms fall in the category of classifi-