





# Prediction of Disease Stage by Machine Learning Classification Methods for Covid-19 Patients

**Merve Doğançay**   
Dokuz Eylül University  
Graduate School of Science  
Statistics, Buca, Izmir Türkiye

**Özlem Ege Oruç**   
Dokuz Eylül University  
Faculty of Science  
Department of Statistics  
Buca, Izmir Türkiye

**Melike Şırlancı**   
University of Colorado  
Anschutz Medical Campus  
Department of Pediatrics  
Anschutz Health, Colorado USA

**Zeynep Altın**   
Izmir University of Health Sciences  
Tepecik Training & Research Hospital  
Internal medicine, Konak Izmir, Türkiye

---

## Abstract

Supervised machine learning classification algorithms have been widely used in many fields in recent years. Especially, health is one of the most important areas where machine learning studies are carried out successfully. The aim of this study is to develop models that predict the disease stage of people who apply to hospital with the diagnosis of Covid-19.

Inadequacies such as intensive care occupancy, insufficiency of beds, and shortage of respiratory equipment are among these problems, and this has left healthcare workers faced with the overwhelming burden of patients. Therefore, estimating the disease stages of Covid-19 patients at an early stage is of great importance. The data set used in the study includes the clinical and laboratory data of the patients during in their admission to the hospital. It has been tried to develop models that predict disease stage by using Logistic Regression, Random Forest and Support Vector Machine algorithms in the data set. The random forest model with 9 variables was the best performing model.

With the models obtained, it will be ensured that the hospital management receives information in order to see the necessary treatment for low-risk or high-risk patients and to avoid medical system inadequacies.

*Keywords:* Covid-19, supervised machine learning, logistic regression, random forest, support vector machine, R.

---

## 1. Introduction

Artificial Intelligence (AI) is a comprehensive term referring to systems or machines that mimic human intelligence. Machine learning (ML), one of the subfields of AI, focuses on creating systems that can learn from or improve their performance based on data. At its core, ML

aims to learn intricate patterns in data (typically through a mathematical model) to perform predictions. In recent years, ML research has been widely used to solve decision-making and prediction-based problems by leveraging principles from mathematics and statistics. Classification, a fundamental and popular application of ML, involves assigning an unknown data point to a known group [Harrington \(2012\)](#). ML classification methods have become prevalent in various fields, with healthcare being one of the significant domains where ML has been successfully applied. Classifications conducted using ML, when executed accurately and effectively, offer significant benefits to both healthcare providers and patients.

The literature contains numerous studies in the healthcare sector that rely on classification-based ML methods. Some examples include using the principal component analysis and logistic regression to predict the outcomes of infertility treatments ([Milewska, Jankowska, Citko, Więsak, Acacio, and Milewski 2014](#)) comparing the accuracy of random forest, support vector machines, gradient boosting decision trees, and penalized regression models using miRNA expression data to classify dementia patients ([Shigemizu, Akiyama, Asanomi, Boroevich, Sharma, Tsunoda, Sakurai, Ozaki, Ochiya, and Niida 2019](#)), and using support vector machines and k-nearest neighbor algorithms to diagnose liver disorders, with the k-nearest neighbour algorithm achieving the highest accuracy at 80.68 [Ribeiro, Marinho, Velosa, Ramalho, and SanchesRibeiro \*et al.\* \(2011\)](#).

Coronaviruses comprise a broad family of enveloped, single-stranded, positive-sense zoonotic RNA viruses ([Gürlevik 2020](#)). They are named "coronaviruses" due to the crown-like appearance of spike protein protrusions on their envelope surfaces, with "corona" being Latin for "crown." Coronaviruses cause severe illnesses in various animal species and humans. In humans, coronaviruses lead to respiratory diseases such as Middle East Respiratory Syndrome (MERS) and Severe Acute Respiratory Syndrome (SARS). The new coronavirus disease is caused by the SARS-CoV-2 virus, and it has been named "Covid-19" by the World Health Organization (WHO). Covid-19 was first identified in December 2019 in the People's Republic of China and subsequently declared a pandemic by the WHO. The Covid-19 pandemic has led to a significant loss of lives worldwide and inflicted substantial damage on the healthcare sector. Even today, many individuals continue to test positive for Covid-19, and fatalities persist. Since the onset of the pandemic, healthcare systems in various countries have suffered major setbacks, with healthcare workers bearing the overwhelming burden of Covid-19 patients. Many countries experienced deficiencies in medical resources, including shortages of hospital beds, intensive care units, and respiratory equipment. The ability to rapidly identify high-risk patients and accurately allocate healthcare resources has become crucial in improving hospital capacity planning and ensuring timely treatment for patients.

### 1.1. Literature review

The global nature of the pandemic and the continuous updates in data available to researchers from different countries made artificial intelligence methods a promising tool in scientific studies. In this context, ML techniques have gained widespread use and promise to assist in the diagnosis process. These efforts are considered in three main subgroups: (a) computational epidemiology, aiming to predict disease trends and potential outbreaks; (b) early detection and diagnosis, aiming to differentiate individuals with and without Covid-19; and (c) disease progression, aiming to predict the progression of the disease, e.g., severity and the likely outcomes ([Syeda, Syed, Sexton, Syed, Begum, Syed, Prior, and Yu 2021](#)). Since our study aims to predict disease severity using routine clinical data, we provide a brief literature review of the studies predicting disease progression using ML techniques.

Many of these studies used classical ML methods for classification. The authors used random forests to predict the risk for patient deterioration, requiring different clinical data ([Cheng, Joshi, Tandon, Freeman, Reich, Mazumdar, Kohli-Seth, Levin, Timsina, and Kia 2020](#)), ([Burian, Jungmann, Kaissis, Lohöfer, Spinner, Lahmer, Treiber, Dommasch, Schneider, Geisler, Huber, Protzer, Schmid, Schwaiger, Makowski, and Braren 2020](#)), while ([Dou-](#)

ville, Douville, Mentz, Mathis, Pancaro, Tremper, and Engoren 2021) used random forests to determine whether Covid-19 patients need respiratory support. In (Ji, Yuan, Shen, Lv, Li, Chen, Zhu, Liu, Liang, Lin, Xie, Li, Chen, Lu, Ding, Zhu, Gao, Ni, Hu, Shi, Shi, and Dong 2020), the authors used logistic regression to predict disease progression based on clinical laboratory data with radiological features. Using logistic regression, the authors developed a model to predict outcomes for patients with severe Covid-19 infection, identifying age, high-sensitivity C-reactive protein levels, lymphocyte count, and d-dimer levels as informative factors during patient admission (Hu, Liu, Jiang, Shi, Zhang, Xu, Suo, Wang, Song, Yu, Mao, Wu, Wu, Shi, Jiang, Mu, Tully, Xu, Jin, Li, Tao, Zhang, and Chen 2021). To compare the predictive accuracy of a range of ML methods, some studies used methods including support vector machines, k-nearest neighbors, naive Bayes, and extreme gradient boosting in addition to the methods listed earlier Ayaz (2021), (Yadaw, Li, Iyengar, Bunyavanich, and Pandey 2020), (Wollenstein-Betech, Cassandras, and Paschalidis 2020). Others used more complex deep-learning techniques with similar purposes. In addition to random forests and classification and regression decision tree models, the authors used neural networks to predict patient deterioration (Assaf, Gutman, Neuman, Segal, Amit, Gefen-Halevi, Shilo, Epstein, Mor-Cohen, Biber, Rahav, Levy, and Tirosch 2020). Also, the authors used artificial neural networks to predict the recovery-death (Al-Najjar and Al-Rousan 2020), and the future need for a mechanical ventilation (Shashikumar, Wardi, Paul, Carlile, Brenner, Hibbert, North, Mukerji, Robbins, Shao, Westover, Nemati, and Malhotra 2021). The authors used hierarchical neural networks for diagnostic and prognostic analysis using computed tomography images. These studies developed the listed ML methods based on patient demographics and clinical data routinely collected for the medical treatment of patients.

The aim of this study is to develop models for predicting the disease risk status in individuals with Covid-19 using ML classification methods. In this study, supervised ML techniques, including logistic regression, random forest, and support vector machines, were employed to predict whether an individual's disease risk status belongs to one of two classes (low or high). The methods used in this study showed high accuracy in classifying disease risk status. We believe that these methods, which take into account specific clinical and laboratory results for each patient, have the potential to be a powerful tool to direct patient treatment strategy for improved health condition of hospitalized Covid-19 patients, In addition, these tools can be used to inform medical resource allocation and hospital capacity planning.

## 2. Material and methods

The main focus of this study is to predict the disease stage of Covid-19 patients using ML algorithms. Following the steps outlined in the Cross-Industry Standard Process for Data Mining (CRISP-DM) (Niakšu 2015), (Martinez-Plumed, Contreras-Ochando, Ferri, Hernández-Orallo, Kull, Lachiche, Ramírez-Quintana, and Flach 2021) approach, the prediction of disease stages for Covid-19 patients was carried out. The following steps were followed in the implementation:

**I. Problem Definition:** During the Covid-19 pandemic, deficiencies in the healthcare system placed an overwhelming burden on healthcare workers and resulted in numerous fatalities. Consequently, early identification of disease stages in Covid-19 patients during their hospitalization has become crucial for hospital management. In this context, the study aimed to predict the disease stage of individuals with a Covid-19 diagnosis upon hospital admission.

**II. Dataset:** The dataset used in the study was obtained from the Izmir Tepecik Training and Research Hospital database. The study population comprised 462 individuals admitted to the hospital with a Covid-19 diagnosis, with no personal identifying information available in the dataset. The dataset includes 52 variables related to the patients. Of these 52 variables, 24 pertain to the patients' laboratory data, while 28 relate to clinical data. To align with the study's objective, a dependent target variable representing the disease stages

of patients was created using the clinical data variables "clinical stage" and "alive/dead." The target variable was divided into two groups based on these two variables: low-risk (clinical stage 1-2 or alive/dead 1) and high-risk (clinical stage 3-4 or alive/dead 0). Subsequently, the "clinical stage" and "alive/dead" variables were removed from the dataset, and the target variable was added for further analysis. Upon examining the dataset, it is evident that there are numerous missing values, particularly in variables such as fibrinogen, ferritin, and Lactate Dehydrogenase (LDH). Figure 1 illustrates the proportions of missing values for each variable, revealing that the fibrinogen, ferritin, and LDH variables have more than %30 missing data. Based on this information, a decision has been made to either impute the missing values or remove variables with missing data from the dataset. When considering the descriptive statistics of the numerical variables in the dataset, it becomes apparent that some variables exhibit significant differences between their mean values and minimum or maximum values. Box plots were generated for these variables indicating the presence of outliers in the data. The next steps in the application involve addressing the missing values and handling outliers in the dataset, as well as conducting further data preprocessing and feature selection, before applying ML algorithms to predict Covid-19 disease stages.

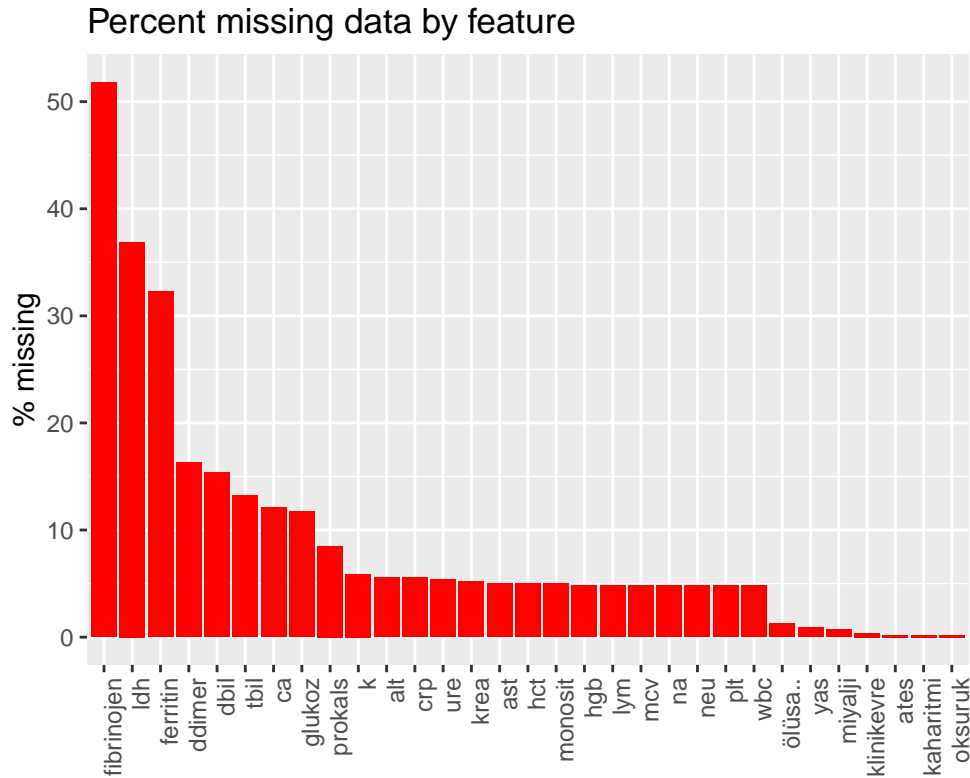


Figure 1: Percentages of missing features in the dataset

**III. Data Preparation:** In the data preparation step, the data preprocessing process was conducted to make the data analysis-ready. Firstly, the variables fibrinogen, ferritin, and LDH were removed from the dataset. Similarly, 22 observations with a substantial number of missing values were excluded from the dataset. The remaining missing values in the dataset were then imputed. For continuous (numerical) variables, the missing values were replaced with median values, while for categorical variables, they were replaced with mode values. Additionally, to prevent potential information loss in the dataset due to outliers, the missing value imputation process was applied to extreme values identified in the dataset. Thus, median values were used in place of the detected outliers. Consequently, after the data preprocessing step, the dataset contains 439 patients and 49 variables. Of these 439

observations, %51 fall into the high-risk group, while %49 belong to the low-risk group (Table 1). In the next step, which is modeling, analyses were conducted using this dataset.

Table 1: Percentage distribution of observations by disease stage

Target	Number of Patients	Percentage
Low Risk Group	213	0.4851936
High Risk Group	226	0.5148064

**IV. Modeling:** The aim of the study was to develop models using ML algorithms to predict the disease stage of Covid-19 patients. To enable a comparison between the models and considering model performance evaluation methods, the dataset was split into two subsets: %70 for training and %30 for testing. Subsequently, four separate models were constructed using 10-fold cross-validation on the training data [Refaeilzadeh, Tang, and Liu \(2009\)](#).

In our study, three of the most commonly used supervised ML methods for classification, namely logistic regression (LOGREG), random forest (RF), and support vector machines (SVM), were established without parameter selection. We preferred these methods because their purposes align with our study. Brief explanatory information about these methods is provided below:

Logistic regression is a statistical method used to solve classification problems. It is typically used for binary classification but can be adapted for multi-class classification problems as well. The primary objective is to understand how independent variables affect one or more dependent variables and make predictions on new data based on this knowledge. Logistic Regression is widely used, especially in fields such as medical diagnosis and identifying risk factors.

Random forest is a powerful ensemble learning method used for tasks like classification and regression. This method combines multiple decision trees to create a stronger and more stable model by aggregating the predictions of these trees. Random forest is commonly used for classification problems and has found successful applications in fields such as medicine and finance.

Support vector machines are algorithms designed to maximize the separation between classes and find the optimal class boundary. These algorithms aim to maximize the margin between classes. SVMs are known for their effectiveness, especially in dealing with high-dimensional data and handling non-linear classification through the use of kernel functions. They have versatile applications, including data mining, medical diagnosis, and financial forecasting.

### 3. Results

We used logistic regression, support vector machine, and random forest methods, enabling us to generate valuable insights for predicting the disease stage of Covid-19 patients. We used the GINI index [Dodge \(2008\)](#), which measures the homogeneity of classes created using the random forest algorithm to rank the importance of each variable. Therefore, the random forest model, utilizing the GINI index, determined the importance ranking of the 49 variables. As a result, an alternative random forest model (RF2) was created using only 9 variables (neu, wbc, crp, ure, ca, hct, ddimer, age, and hgb).

To determine which of the four models best aligns with the study's objective, model performance evaluation methods were applied. Initially, based on the average Area Under the Curve (AUC) value in the 10-fold cross-validation, it was determined that the model with the highest performance was the random forest model constructed with 9 variables (RF2) (Table 2).

There are some criteria used to evaluate the performance of classification models. These metrics are based on the error matrix. The error matrix is a matrix that provides information

Table 2: AUC based on 10-fold cross-validation

	<i>Min</i>	<i>Q1</i>	<i>Median=Q2</i>	<i>Mean</i>	<i>Q3</i>	<i>Max</i>
<i>RF</i>	0.7377778	0.8046875	0.8375000	0.8377361	0.8739583	0.9208333
<i>LOGREG</i>	0.6000000	0.7989583	0.8226944	0.8026944	0.8390278	0.8625000
<i>SVM</i>	0.7333333	0.8135417	0.8341667	0.8341667	0.8614583	0.8875000
<i>RF2</i>	0.7541667	0.8048958	0.8483611	0.8483611	0.9062500	0.9291667

about the accuracy of the predictions by allowing the comparison of the prediction values with the real values of the classification model.

Table 3: Error matrix

		ACTUAL VALUES	
		<i>Positive</i>	<i>Negative</i>
PREDICTIVE VALUES	<i>Positive</i>	TRUE POSITIVE (TP)	FALSE POSITIVE (FP)
	<i>Negative</i>	FALSE NEGATIVE (FN)	TRUE NEGATIVE (TN)

There are classification model performance evaluation measures based on the error matrix in Table 2. The accuracy of the classification model is calculated as follows.

- $$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

The success of the classification model in predicting positive situations is called sensitivity.

- $$\text{Sensitivity (True positive rate)} = \frac{TP}{TP + FN}$$

The success of the classification model in predicting negative situations is called specificity.

- $$\text{Specificity} = \frac{TN}{TN + FP}$$
- $$\text{False positive rate} = \frac{FP}{FP + TN}$$

Performance evaluation metrics obtained from the application of the models to the test data using the error matrix are presented in Table 4. According to this table, the random forest model constructed with 9 variables (RF2) achieved the best performance, with an accuracy of 0.85 in predicting the disease stage. This model can predict low-risk patients with a sensitivity of 0.92 and high-risk patients with a specificity of 0.79.

Table 4: Performance evaluation metrics based on the error matrix

		<i>Models</i>			
		<i>LOGREG</i>	<i>RF</i>	<i>RF2</i>	<i>SWM</i>
Performance evaluation criteria	<i>Accuracy</i>	0.7154	0.8462	0.8538	0.7846
	<i>Sensitivity</i>	0.7460	0.9048	0.9206	0.8254
	<i>Specificity</i>	0.6866	0.7910	0.7910	0.7463

Receiver Operating Characteristic (ROC) is a comparison method created using the false positive rate and true positive rate. To evaluate this curve, AUC is checked. AUC is the area



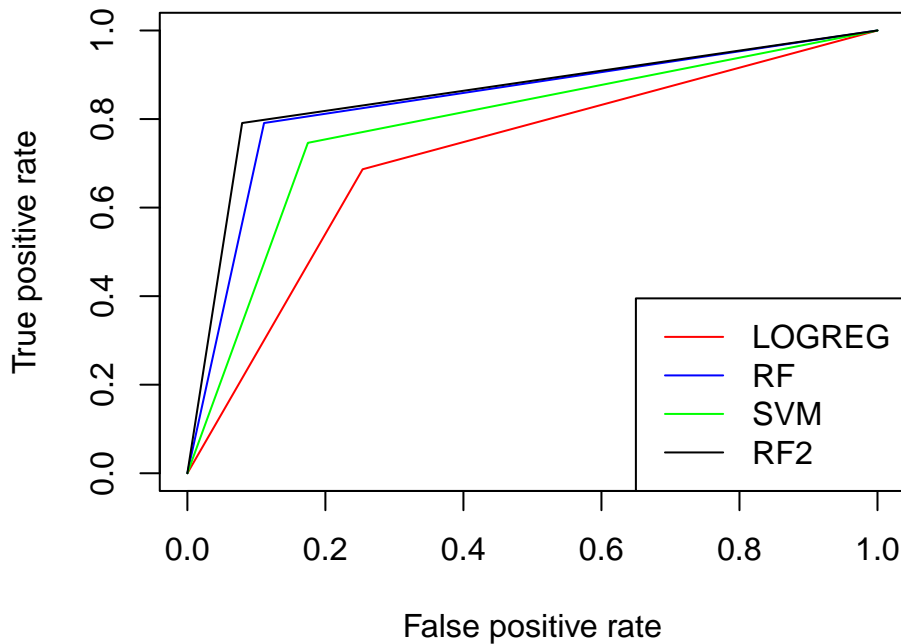


Figure 2: ROC curves of the models constructed on the data

under the ROC curve, and the larger this value is, the more effective the classification model is. ROC curves obtained on the test data set are given in Figure 2. As can be seen in the figure, the random forest (RF2) model drawn in black gave the best results.

When examining the AUC values obtained on the test dataset, it can be observed that the random forest model constructed with 9 variables (RF2) is a more effective model compared to the others (Table 5).

Table 5: AUC values for the test data

Methods	<i>LOGREG</i>	<i>RF</i>	<i>RF2</i>	<i>SVM</i>
<b>AUC Values</b>	0.72	0.84	0.86	0.79

Taking into account all model performance evaluation methods, it has been determined that the random forest model constructed with 9 variables (RF2) is the most suitable model for predicting the disease stage.

## 4. Discussion

The Covid-19 pandemic, since its inception and global spread, has significantly impacted daily life, posing adverse effects on both public health and the world economy. Despite efforts to reduce transmission risks through the development of vaccines and treatment methods, there remains no definitive cure for the disease. Furthermore, the prediction of how Covid-19 patients will be affected, specifically their disease stages and risk profiles, remains uncertain. This uncertainty places a substantial burden on healthcare professionals, complicating patient care decisions and increasing the cost of medical tests.

In this study, we aimed to predict the disease stages of Covid-19 patients at the time of their hospital admission by employing supervised ML classification algorithms. Four distinct models were developed, utilizing logistic regression, random forest, and support vector machines, while a 9-variable random forest model displayed superior performance based on a 10-fold cross-validation with an AUC of 0.848, accuracy of 0.8538 according to the confusion matrix, and an AUC of 0.86 on the test dataset.

All analyses were conducted using the R programming language, leveraging the open-source capabilities of RStudio. The results suggest that the models generated in this study have the potential to enable more rapid early-stage disease predictions for Covid-19 patients, allowing for expedited and cost-effective medical interventions by healthcare professionals.

In conclusion, this research contributes to the field of healthcare by introducing a novel approach for predicting disease stages in Covid-19 patients. It is anticipated that this approach will offer valuable insights to healthcare professionals, directing medical tests more effectively and initiating treatment promptly. This study may serve as a beacon guiding future research endeavors and the application of ML algorithms in healthcare-related studies.

## Acknowledgement

Authors are thankful to editor Prof. Gajendra K. Vishwakarma and anonymous reviewers for their valuable comments to improve the quality and content of manuscript.

## References

- Al-Najjar H, Al-Rousan N (2020). “A Classifier Prediction Model to Predict the Status of Coronavirus Covid-19 Patients in South Korea.” *European Review for Medical and Pharmacological Sciences*, **24**, 3400–3403.
- Assaf D, Gutman Y, Neuman Y, Segal G, Amit S, Gefen-Halevi S, Shilo N, Epstein A, Mor-Cohen R, Biber A, Rahav G, Levy I, Tirosh A (2020). “Utilization of Machine-learning Models to Accurately Predict the Risk for Critical COVID-19.” *Internal and Emergency Medicine*, **15**, 1435–1443. doi:10.1007/s11739-020-02475-0.
- Ayaz M (2021). *Detection of Covid-19 Patients Using Machine Learning Algorithms*. Pamukkale University Institute of Social Sciences, M.Sc thesis.
- Burian E, Jungmann F, Kaissis GA, Lohöfer FK, Spinner C, Lahmer T, Treiber M, Dommasch M, Schneider G, Geisler F, Huber W, Protzer U, Schmid RM, Schwaiger M, Makowski MR, Braren RF (2020). “Intensive Care Risk Estimation in COVID-19 Pneumonia Based on Clinical and Imaging Parameters: Experiences from the Munich Cohort.” *Journal of Clinical Medicine*, **18**(5), 1514. doi:10.3390/jcm9051514.
- Cheng FY, Joshi H, Tandon P, Freeman R, Reich DL, Mazumdar M, Kohli-Seth R, Levin M, Timsina P, Kia A (2020). “Using Machine Learning to Predict ICU Transfer in Hospitalized COVID-19 Patients.” *Journal of Clinical Medicine*, **9**(6), 1668. doi:10.3390/jcm9061668.
- Dodge Y (2008). *The Concise Encyclopedia of Statistics*. 1st edition. Springer Science.
- Douville NJ, Douville CB, Mentz G, Mathis MR, Pancaro C, Tremper KK, Engoren M (2021). “Clinically Applicable Approach for Predicting Mechanical Ventilation in Patients with COVID-19.” *British Journal of Anaesthesia*, **126**(3), 578–589. doi:10.1016/j.bja.2020.11.034.
- Gürlevik SL (2020). “Coronaviruses and New Coronavirus SARS-CoV-2.” *Journal of Pediatric Infection*, **14**(1), 46–48.



- Harrington P (2012). *Machine Learning in Action*. 1st edition. Manning.
- Hu C, Liu Z, Jiang Y, Shi O, Zhang X, Xu K, Suo C, Wang Q, Song Y, Yu K, Mao X, Wu X, Wu M, Shi T, Jiang W, Mu L, Tully DC, Xu L, Jin L, Li S, Tao X, Zhang T, Chen X (2021). “Early Prediction of Mortality Risk among Patients with Severe COVID-19, Using Machine Learning.” *International Journal of Epidemiology*, **49**(6), 1918–1929. doi:10.1093/ije/dyaa171.
- Ji M, Yuan L, Shen W, Lv J, Li Y, Chen J, Zhu C, Liu B, Liang Z, Lin Q, Xie W, Li M, Chen Z, Lu X, Ding Y, An P, Zhu S, Gao M, Ni H, Hu L, Shi G, Shi L, Dong W (2020). “A Predictive Model for Disease Progression in Non-severely Ill Patients with Coronavirus Disease 2019.” *European Respiratory Journal*, **56**(1), 2001234. doi:10.1183/13993003.01234-2020.
- Martinez-Plumed F, Contreras-Ochando L, Ferri C, Hernández-Orallo J, Kull M, Lachiche N, Ramírez-Quintana MJ, Flach P (2021). “CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories.” *IEEE Transactions on Knowledge and Data Engineering*, **33**(8), 3048–3061. doi:10.1109/TKDE.2019.2962680.
- Milewska AJ, Jankowska D, Citko D, Więsak T, Acacio B, Milewski R (2014). “The Use of Principal Component Analysis and Logistic Regression in Prediction of Infertility Treatment Outcome.” *Studies in Logic, Grammar and Rhetoric*, **1**(39), 7–23. doi:10.2478/slgr-2014-0043.
- Niakšu O (2015). “CRISP Data Mining Methodology Extension for Medical Domain.” *Baltic Journal of Modern Computing*, **3**(2), 92–109.
- Refaeilzadeh P, Tang L, Liu H (2009). *Cross-Validation*, pp. 532–538. Springer US, Boston, MA. ISBN 978-0-387-39940-9. doi:10.1007/978-0-387-39940-9\_565.
- Ribeiro R, Marinho R, Velosa J, Ramalho F, Sanches JM (2011). “Diffuse Liver Disease Classification from Ultrasound Surface Characterization, Clinical and Laboratorial Data.” In J Vitrià, JM Sanches, M Hernández (eds.), *Pattern Recognition and Image Analysis*, pp. 167–175. Springer Berlin Heidelberg, Berlin, Heidelberg. doi:10.1007/978-3-642-21257-4\_21.
- Shashikumar SP, Wardi G, Paul P, Carlile M, Brenner LN, Hibbert KA, North CM, Mukerji SS, Robbins GK, Shao YP, Westover MB, Nemati S, Malhotra A (2021). “Development and Prospective Validation of a Deep Learning Algorithm for Predicting Need for Mechanical Ventilation.” *Chest*, **159**(6), 2264–2273. doi:10.1016/j.chest.2020.12.009.
- Shigemizu D, Akiyama S, Asanomi Y, Boroevich KA, Sharma A, Tsunoda T, Sakurai T, Ozaki K, Ochiya T, Niida S (2019). “A Comparison of Machine Learning Classifiers for Dementia with Lewy Bodies Using miRNA Expression Data.” *BMC Medical Genomics*, **12**(1), 1–10. doi:10.1186/s12920-019-0607-3.
- Syeda HB, Syed M, Sexton KW, Syed S, Begum S, Syed F, Prior F, Yu F (2021). “Role of Machine Learning Techniques to Tackle the Covid-19 Crisis: Systematic Review.” **9**(1), 23811. doi:10.2196/23811.
- Wollenstein-Betech S, Cassandras CG, Paschalidis IC (2020). “Personalized Predictive Models for Symptomatic Covid-19 Patients Using Basic Preconditions: Hospitalizations, Mortality, and the Need for an ICU or Ventilator.” *International Journal of Medical Informatics*, **142**(104258), 1–8. doi:10.1016/j.ijmedinf.2020.104258.
- Yadaw AS, Li Y, Bose S, Iyengar R, Bunyavanich S, Pandey G (2020). “Clinical Features of Covid-19 Mortality: Development and Validation of a Clinical Prediction Model.” *The Lancet Digital Health*, **2**(10), 516–525. doi:10.1016/S2589-7500(20)30217-X.

**Affiliation:**

Özlem Ege Oruç  
Department of Statistics  
Dokuz Eylül University Faculty of Science  
Tinaztepe Campus 35360, Buca İzmir, Türkiye  
Email: [ozlem.ege@deu.edu.tr](mailto:ozlem.ege@deu.edu.tr)  
Url: <https://avesis.deu.edu.tr/ozlem.ege>