Research Article

Türk Fen ve Sağlık Dergisi
Turkish Journal of Science and Health

| | |
|---|---|
| Volume | 6 |
| Number | 2 |
| Year | 2025 |
| Pages | 52-61 |

*e-ISSN: 2717-7173* https://dergipark.org.tr/tr/pub/tfsd

*Received:* 18.03.2025 *Accepted:* 12.05.2025 https://doi.org/10.51972/tfsd.1660325

# Comparison of Predictive Performances of Machine Learning Methods in the Diagnosis of Crimean-Congo Hemorrhagic Fever

**Esra Akaydın Gültürk [1]*, Hüdaverdi Bircan [2], Erdem Karabulut [3], Nazif Elaldı [4]**

[1]Sivas Cumhuriyet University Faculty of Medicine, Department of Biostatistics, Sivas, Türkiye
[2]Sivas Cumhuriyet University, Faculty of Economics and Administrative Sciences, Department of Business Administration, Sivas, Türkiye
[3]Hacettepe University Faculty of Medicine, Department of Biostatistics, Ankara, Türkiye
[4]Sivas Cumhuriyet University Faculty of Medicine, Department of Infectious Diseases and Clinical Microbiology, Sivas, Türkiye

**ABSTRACT:**
**Purpose:** This study aims to compare the performance results of the machine learning methods “Support Vector Regression, Random Forest, Regression Tree and Nearest Neighbor Regression models on the dataset of Crimean-Congo Hemorrhagic Fever Diagnosis.
**Materials and Methods:** The data of all patients who were hospitalized in Cumhuriyet University Faculty of Medicine, Infectious Diseases and Pediatrics service with the diagnosis of Crimean-Congo hemorrhagic fever between 2009 and 2011 were taken from the service records. During these three years, 6125 data entries were made for a total of 245 patients. A total of three groups of patient data were used in the study: adult, pediatric and all patients. Each scenario was repeated 1000 times with the bootstrap resampling method and the mentioned regression methods were applied in each repetition. To compare the performance of the regression models, the mean squared error and the percentage of explanatory variables were analyzed.
**Results:** Among the regression methods for the real data set, the regression model with the highest explanatory percentage and the lowest mean squared error was found to be the best performing regression model for all three groups.
**Conclusion:** As a result of the simulation study according to real data and scenario structures, the best prediction regression method was found to be support vector regression.
***Keywords:*** *K-nearest neighbor; support vector regression; random forest; regression tree; machine learning*

**Corresponding author: Esra Akaydın Gültürk, email: esra0709.eg@gmail.com*

## INTRODUCTION

Crimean-Congo Hemorrhagic Fever (CCHF) was first identified as a clinical entity in 1944 and 1945 in 200 Soviet soldiers assisting farmers harvesting in Crimea. A similar disease was seen in Congo in 1956, a virus was produced in 1967 and in 1969 it was determined that the causative agents of Crimean Hemorrhagic Fever and Congo Hemorrhagic Fever were the same virus, and the disease was named CCHF (Balinandi et al., 2024; Ergönül, 2006). Despite the presence of cases in neighboring countries since 1970, Tokat saw the first case of CCHF in Turkey in 2002. Later, cases were reported from the Central Anatolia and Eastern Black Sea Regions (Karti et al., 2004). CCHF is a viral infectious disease with a high mortality rate that has become common in Turkey in recent years and is transmitted by the bite of a tick infected with this virus (Seçmeer and Celik, 2010). In today's computer and technology-oriented world, databases contain vast amounts of information. The accessibility and abundance of this information makes data mining very important and necessary

(Han et al., 2012). Data mining makes major contributions to the world of science and technology with techniques based on machine learning, which is one of the tools used by decision support systems (Savaş et al., 2012; Ozgulbas and Koyuncugil, 2012). As in many fields, it provides support to the decisions of physicians, especially for the solution of problems in the field of health (Pala, 2013). Data mining is the process of discovering meaningful information from large data sets and making that information usable. This process is carried out using various techniques and tools and often requires the combination and utilization of various multidisciplinary sciences such as data analytics, machine learning, statistics, and artificial intelligence (Ersöz and Çınar, 2021). Machine learning is a sub-branch of artificial intelligence and is a field related to data training, model building, and analysis (Silahtaroğlu, 2008). Machine learning focuses on designing and developing algorithms and techniques that are easy for computers to learn. It does not aim to work on a specific data set. The algorithms it develops aim to solve every problem. In traditional statistical methods, a hypothesis is established and statistical tests are performed to accept or reject it, while machine learning searches for pattern structures whose existence is known but not certain (Ij, 2018). We analyze data mining methods into two parts: predictive and descriptive. While classification and regression models are used in predictive modeling, clustering, association, and sequential time patterns are widely used in descriptive modeling (Pupezescu and Ionescu, 2008). The main purpose of predictive modeling is to predict some fields in the database based on other fields. If the target variable to be predicted is a categorical variable, it is a classification problem. If the target variable to be predicted is a numeric (continuous) variable, it is a regression problem (Tüzüntürk, 2010).

Data mining works on "real" data for identification, prediction, or clustering using machine learning methods. In this direction, supervised and unsupervised learning algorithms are used (Palmer et al., 2011). Many regression algorithms in data mining can predict a continuous-valued target variable. The main techniques used for prediction are artificial neural networks, K-nearest neighborhood, decision trees, support vector machines, and random forest (Breiman, 2001). Support Vector Regression (SVR), developed by Vapnik, Chervonenkis, and others, basically aims to minimize the generalization error boundary in order to obtain a better generalized performance instead of directly minimizing the observed training error (Vapnik, 1999). With this feature, SVR is a method that aims to reduce the risk of overfitting by increasing the generalization ability. In addition to methods with strong theoretical foundations such as SVR, there are also algorithms that focus on data-based logical conditions. Among these algorithms, Classification and Regression Trees (CART) create tree-based structures that use logical rules to predict or classify cases. CART models are divided into two categories according to the type of dependent variable: classification trees are used when the dependent variable is categorical, and regression trees are used when the dependent variable is continuous. Regression trees are heuristic models that, starting from the root node, test for specific features at each internal node and present predictions at the leaf nodes. Thanks to these features, they are easy to interpret and visualize (Smola and Schölkopf, 2004).

Similar to CART models, K-Nearest Neighbors (KNN) regression, a heuristic and nonparametric method, predicts the target variable by averaging the target values of the k data points closest to the data point. The k parameter used here increases model flexibility by determining the accuracy and complexity level of the model. Due to its simplicity and nonparametric structure, KNN regression is a widely applicable method for different types of data (Frank et al., 2011). Another method that takes the advantages of methods such as KNN and CART even further is Random Forest Regression (RFR). RFR is an ensemble learning method that works by combining many decision trees together. In this method, each tree is constructed independently using randomly selected subsets of data and features, and the resulting predictions are obtained by averaging the predictions of all trees. Due to its high prediction accuracy and strong generalization ability, RFR is a powerful regression method preferred in many applications (Liaw and Wiener, 2002).

## MATERIAL and METHODS

### Purpose and Type of the Study

The aim of this study, which is planned as a methodological study, is to explain the features of "Support Vector Regression," "Regression Trees," "Random Forest," and "K-Nearest Neighbor" regression models commonly used in data mining, to compare the prediction performances of these four methods, and to evaluate the prediction performances of the models by using the bootstrap resampling method and simulation results of real data sets and different scenarios.

### Sampling and Participant

In our study, the data of all patients hospitalized with Crimean-Congo Hemorrhagic Fever (CCHF) in the Infectious Diseases and Pediatrics services of Cumhuriyet University Faculty of Medicine between 2009 and 2011 were obtained from the service records. In these three years, 6125 data entries were made for a total of 245 patients. The 245 patients included in the study consisted of an adult group of 113 and a pediatric group of 132.

### Implementation of the Study

Patient information on patients diagnosed with Crimean-Congo Hemorrhagic Fever (CCHF) was obtained from ward records. A total of three groups of patient data were used in our study: adult, pediatric, and all patients. The dependent variable was the duration of hospitalization, and the other independent variables were age, gender, place of residence, livestock status, conjunctivitis, jaundice, lung involvement, hepatomegaly, splenomegaly, change in consciousness, headache, myalgia, sore throat, nausea, vomiting, diarrhea, cough, fever, malaise, bleeding, rash, leukopenia, need for blood products, and number of symptom days. The total number of variables was 25, the number of qualitative variables was 22, and the number of continuous variables was 3. The results presented in this study were obtained using the bootstrap resampling method. In the simulation study, data sets were generated from a multivariate standard normal distribution using the mvrnorm function, where the mean vector was set to zero across all scenarios. Correlation matrices were constructed based on empirical values derived from the original data set, as well as from correlation values randomly generated from a uniform distribution. Categorical variables were formed using predetermined cut-off points, established according to frequency distributions observed in the original data set. For each simulation scenario, 1000 replications were executed, enabling robust evaluation and comparison of regression model performances. Additionally, a secondary simulation study was conducted to assess model performance under varying sample sizes (n=100, n=250, and n=1000) and correlation magnitudes categorized as low (r=0.00-0.20), medium (r=0.21-0.40), and high (r=0.41-0.60). The impact of these simulation conditions on model performance was systematically examined. Support Vector Regression: SVR is a powerful machine learning method that emerged by applying the principles of support vector machines to regression problems (Smola and Schölkopf, 2004). In particular, SVR can successfully model nonlinear relationships through kernel functions. Important advantages of SVR are its ability to model complex relationships and its robustness to overlearning. On the other hand, SVR is highly sensitive to parameter settings (e.g., parameter C and kernel function type) and can perform poorly when parameters are chosen incorrectly (Basak et al., 2007). For large data sets, the computational burden can increase, and model interpretation is more difficult compared to linear methods (Basak et al., 2007). K-Nearest Neighbor (KNN) algorithm: is a non-parametric machine learning method commonly used in classification and regression problems. The KNN method makes no assumptions about the distribution of the data and has a fundamentally simple, straightforward structure. This algorithm determines the outcome of the observation to be classified or predicted based on the values of the “k” nearest neighbors (Liaw and Wiener, 2002). The main advantages of the KNN algorithm are that it is easy to implement and its results can be intuitively interpreted. However, the algorithm has a high computational cost for large datasets, and its performance may suffer from the curse of dimensionality, especially for high-dimensional datasets. It also performs poorly when modeling complex data structures (Cover and Hart, 1967). Classification and Regression Trees (CART) is a heuristic method that builds models for the prediction of continuous variables by partitioning the dataset into smaller subsets according to certain rules (Hastie et al., 2009). Its major advantages are that it is easy to understand and interpret, can efficiently model complex non-linear data, and is robust to outliers. However, the major disadvantages of this method are that it is prone to overfitting and is highly sensitive to small data variations, producing different results (Quinlan, 1986; Hastie et al., 2009). Random Forest is a widely used machine learning method for both regression and classification problems. In regression problems, Random Forest generates a large number of independent decision trees and combines the predictions of each of these trees to arrive at a final prediction. This merging process is done by averaging the prediction values of all the generated trees. The key advantages of the random forest method are that it can successfully analyze data sets with complex or irregular data structures, provide consistent and stable results against small data changes, and reduce the problem of overlearning by using bootstrap sampling and random variable subsets. However, the disadvantages of this method include reduced prediction reliability in overly complex datasets and being adversely affected by unbalanced data (Sihombing et al., 2023).

**Matrix Evaluation**

Mean Squared Error (MSE) is defined as the average of squared differences between the predicted and actual values, indicating the magnitude of errors produced by the predictive model. Lower values of MSE reflect better predictive performance, demonstrating higher accuracy and reliability of the model (Sihombing et al., 2023). Coefficient of Determination ($R^2$) represents the proportion of the variance in the dependent variable explained by the regression model. $R^2$ values range from 0 to 1, where values closer to 1 suggest stronger explanatory power and greater capability of the model in capturing the observed variability in the data (Sihombing et al., 2023). In this study, the performance of regression models was evaluated by utilizing mean squared error (MSE) and coefficient of determination ($R^2$). Models with lower MSE values and higher $R^2$ values were considered to demonstrate superior predictive capability. Statistical AnalysisIn this study, R-Studio version 4.2.1, which has open-source code scripts, was used for analysis and evaluation of simulation outputs. The packages used in the related study, “e1071,” “mass,” “randomforest,” “rpart,” “corpcor,” “caret,” and “fnn,” were used, and the performance criteria of the regression models were analyzed. Ethical ApprovalThis study was conducted in accordance with the ethical principles stated in the Declaration of Helsinki. The study protocol was approved by the Non-Interventional Ethics Committee of Sivas Cumhuriyet University Faculty of Medicine (decision no: 2024-02/22, date: 22.02.2024).

## RESULTS

When the statistics of the demographic characteristics of the variables belonging to the CCHF data were analyzed over the groups, the mean duration of hospitalization was 9.83 ± 3.05 in pediatric patients, 8.62 ± 2.90 in the adult group, and 9.27 ± 3.04 in the whole patient group. When the number of days with symptoms was analyzed, it was 4.44±3.57 in pediatric patients, 5.22±2.32 in the adult group, and 4.80±3.07 in the whole patient group. When the mean age was analyzed, the mean age of the pediatric age group was 11.59± 3.95, the mean age of the adult age group was 46.65± 17.62, and the mean age of the whole patient group was 27.76± 21.39.Of the patients included in the study, 46.1% were adults and 53.9% were pediatric. When the frequency distribution according to the gender of the individuals included in the study is examined, 71.2% of pediatric patients were male and 28.8% were female, 43.4% of adult patients were male and 56.6% were female, and 55.4% were male and 44.6% were female in total. The rate of pediatric patients residing in Sivas province is 93.9%. In other provinces, this rate is 6.1%. Thirty-one percent of the adult patient group lives in Sivas. The rate of patients coming from other provinces is 69%. In generalmost patients are from Sivas province, with a rate of 64.9%. The rate of coming from other provinces is 35.1%. While the rate of animal husbandry in pediatric patients is 20.5%, this rate is 85.8% in the adult group. In the total group, this rate is 50.6%. The rate of conjunctivitis is 58.3% in the pediatric group, 47.in the adult group, and 46.in the total group. When we look at the jaundice variable, there is none in the child group; this rate is 7.1% in the adult group. In the total group, this rate was 3.3%. Pulmonary involvement was present in 72% of the pediatric group, but in the adult group it was very low, 5.3%. Hepatomegaly was present in 2.3% of the pediatric group, 22.1% of the adult group, and 11.4% of the total group. Splenomegaly was 1.5% in the pediatric group, 7.1% in the adult group, and 4.1% in the total group. Altered consciousness was 2.3% in the pediatric group, 3.5% in the adult group, and 2.9% in the total group. Complaint of sore throat was 4.5% in the pediatric group, 15% in the adult group, and 9.4% in the total group. The highest proportion of those who complained of headache was 81.4% in the adult group, 31.8% in the pediatric group, and 54.7% in the total group. Among those presenting with nausea, 81.4% were in the adult group, 61.2% in the pediatric group, and 61.2% in general. In the pediatric group, this rate is 43.9%. Of those who presented with vomiting, 71.7% were, 54.5% in the pediatric group and 62.4% in the total group. Most of the patients presenting with cough were in the pediatric group, 72.7%, while this rate was 27.3% in the adult group. Overall, this rate was 49.4%. Fever is a determinant variable that is the highest in the total groups. It was 96.2% in the pediatric group, 92% in the adult group, and 94.3% in the total group. Among patients presenting with complaints of fracture, 93.8% were in the adult group. In the pediatric group, this rate was 40.2%, and 64.9% in the total group. The complaint of muscle pain is higher in the adult group, 90.3%. This rate is 15.2% in the child group and 50.2% in the total group. The incidence of diarrhea, bleeding, rash, and leukopenia is generally low. Diarrhea was 17.4% in the child group, 34.5% in the adultp, and 25.3% overall. The incidence of bleeding was 17.4% in the pediatric group, 35.4% in the adult group, and 25.7% in the overall group. The incidence of rash and leukopenia was the same in the pediatric group; it was 12.1%. In the adult group, the incidence of rash was 21.2%, and the incidence of leukopenia was 28.3%. In general, the prevalence was 16.3% for rash and 19.6% for leukopenia. For the need for blood products, this rate is lower in the pediatric group (5.3%) and higher in the adult group (64.6%). In the total group, this rate is 32.7.Comparison of regression methods according to a real data set The MSE and $R^2$ of DVR, RF, RA, and KNN regression models according to patient groups are given in Table 1.

**Table 1.** Comparison of Regression Models According to Real Data Set

| | **Pediatri** | | **Adult** | | **Total** | |
|---|---|---|---|---|---|---|
| **Methods** | **MSE** | **$R^2$** | **MSE** | **$R^2$** | **MSE** | **$R^2$** |
| **DVR** | 2,00 | 0,84 | 0,30 | 0,98 | 1,35 | 0,89 |
| **RF** | 3,14 | 0,83 | 2,42 | 0,84 | 2,82 | 0,85 |
| **RA** | 6,88 | 0,26 | 5,90 | 0,29 | 6,23 | 0,32 |
| **KNN** | 2,60 | 0,80 | 2,20 | 0,79 | 2,54 | 0,79 |

**MSE:** Mean Square Eror

According to Table 1, $R^2$ for DVR in the pediatric group is 0.84, MSE = 2.00, while 0.98, MSE = 0.30, in the adult group. In the total patient group, $R^2 = 0.89$, 1.35. For RF, $R^2 = 0.83$ and MSE = 3.14 in the pediatric group. In the adult group, $R^2 = 0.84$ and MSE = 2.42; in the total patient group, $R^2 = 0.85$ and MSE = 2.82. For RA, $R^2 = 0.26$, MSE = 6.8 in the child group; $R^2 = 0.29$, MSE = 5.90 in the adult group; and $R^2 = 0.32$, MSE = 6.23 in the total patient group. KNN, $R^2$=0.80, MSE=2.60 for the child group and R2=0.79, MSE=2.20 for the adult group. In total, $R^2$=0.79, MSE=2.54. According to the simulation study, the comparison of the performance of the regression models for the total patient, child, and adult groups according to the simulation results obtained from the real data set with 1000 repetitions as the number of real data is shown in Table 2.

**Table 2.** Comparison of Regression Models According to the Results of 1000 Repeated Simulations up to the Number of Patient Groups

| | Pediatri | | Adult | | Total | |
|---|---|---|---|---|---|---|
| **Methods** | **MSE** | **$R^2$** | **MSE** | **$R^2$** | **MSE** | **$R^2$** |
| **DVR** | 0,069 | 0,95 | 0,035 | 0,98 | 0,054 | 0,96 |
| **RF** | 0,29 | 0,86 | 0,255 | 0,88 | 0,292 | 0,87 |
| **RA** | 0,649 | 0,34 | 0,621 | 0,37 | 0,67 | 0,32 |
| **KNN** | 0,81 | 0,20 | 0,82 | 0,19 | 0,80 | 0,20 |

**Table 3.** Comparison of Simulation Study and Regression Models for Different Scenarios

| **Sample size** | **methods** | **(r=0,00,20)** | | **(r=0,21-0,40)** | | **r=0,41-0,60)** | |
|---|---|---|---|---|---|---|---|
| | | **MSE** | **$R^2$** | **MSE** | **$R^2$** | **MSE** | **$R^2$** |
| **n=100** | **DVR** | 0,03 | 0,98 | 0,03 | 0,98 | 0,03 | 0,98 |
| | **RF** | 0,28 | 0,89 | 0,23 | 0,86 | 0,18 | 0,86 |
| | **RA** | 0,68 | 0,31 | 0,55 | 0,44 | 0,44 | 0,56 |
| | **KNN** | 0,52 | 0,42 | 0,93 | 0,08 | 0,92 | 0,14 |
| **n=250** | **DVR** | 0,05 | 0,97 | 0,04 | 0,97 | 0,04 | 0,96 |
| | **RF** | 0,28 | 0,89 | 0,23 | 0,85 | 0,18 | 0,85 |
| | **RA** | 0,70 | 0,30 | 0,57 | 0,43 | 0,46 | 0,54 |
| | **KNN** | 0,77 | 0,19 | 0,75 | 0,21 | 0,78 | 0,14 |
| **n=1000** | **DVR** | 0,10 | 0,92 | 0,10 | 0,92 | 0,10 | 0,90 |
| | **RF** | 0,30 | 0,89 | 0,23 | 0,84 | 0,20 | 0,83 |
| | **RA** | 0,94 | 0,13 | 0,74 | 0,26 | 0,56 | 0,43 |
| | **KNN** | 0,82 | 0,22 | 0,70 | 0,22 | 0,83 | 0,21 |

(r=0,00-0,20): low, (r=0,21-0,40): medium , (r=0,41-0,60): high

For DVR, $R^2$=0.95, MSE=0.069 in the child group, and $R^2$=0.98, MSE=0.035 in the adult group. In the total group, $R^2$ = 0.96 and MSE = 0.054. For RF, $R^2$=0.86, MSE=0.29 in the child group, and $R^2$=0.88, MSE=0.255 in the adult group. In the total group, $R^2$=0.87, MSE=0.292. For RA, $R^2$ = 0.34, MSE = 0.649 in the child group, and $R^2$ = 0.37, MSE = 0.621 in the adult group. In the total group, $R^2$ = 0.32 and MSE = 0.670. For KNN, $R^2$ = 0.20, MSE = 0.81 in the child group, and $R^2$ = 0.19, MSE = 0.82 in the adult group. For the total group, $R^2$ = 0.20, MSE = 0.80. For the Taking the correlation structure as (0.00-0.20) for n=100, MSE for DVR=0.03, $R^2$=0.98. MSE for RF=0.28, $R^2$= 0.89. MSE for RA=0.68, $R^2$= 0.31. For KNN, MSE=0.52, $R^2$=0.42. Taking the correlation structure as (0.21-0.40), MSE=0.03, $R^2$=0.98 for DVR. For RF, MSE=0.23, $R^2$=0.86. For RA, MSE=0.55, $R^2$=0.44. For correlation matrix of dependent and independent variables, 100, 250, and 1000 observations were obtained from a uniform distribution according to low (r=0.00-0.20), medium (r=0.21-0.40), and high (r=0.41-0.60) correlation relationships between them. According to these data, the comparison of the performances of SVR, RF, RA, and KNN models and the results of the effect of the correlation structure between the variables on the methods at 1000 repetitions for MSE and $R^2$ values are given in Table 3.

KNN, MSE=0.93 and $R^2$=0.08. When the correlation structure is taken as 0.41-0.60, MSE=0.03 and $R^2$=0.98 for DVR. For RF, MSE=0.18, $R^2$=0.86. For RA, MSE=0.44, $R^2$=0.56. For KNN, MSE=0.92, $R^2$=0.14. DVR for correlation structure (0.00-0.20) for n=250, MSE=0.05, $R^2$=0.97, RF: MSE=0.28, $R^2$=0.89, RA:

MSE=0.70, $R^2$=0.30, KNN: MSE=0.77, $R^2$=0.19. Taking the correlation structure as (0.21-0.40), DVR: MSE=0.04, $R^2$=0.97; RF: MSE=0.23, $R^2$=0.85; RA: MSE=0.57, $R^2$=0.43; KNN: MSE=0.75, $R^2$=0.21. Correlation structure (0.41-0.60) DVR: MSE=0.04, $R^2$=0.96; RF: MSE=0.18, $R^2$=0.85; RA: MSE=0.46, $R^2$=0.54; KNN: MSE=0.78, $R^2$=0.14. For n=1000, correlation structure (0.00-0.20): DVR: MSE=0.10, $R^2$=0.92; RF: MSE=0.30, $R^2$=0.89; RA: MSE=0.94, $R^2$=0.13; KNN: MSE=0.82, $R^2$=0.22. Correlation structure (0.21-0.40): DVR: MSE=0.10, $R^2$=0.92; RF: MSE=0.23, $R^2$=0.84; RA: MSE=0.74, $R^2$=0.26; KNN: MSE=0.70, $R^2$=0.22. Correlation structure (0.41-0.60): DVR: MSE=0.10, $R^2$=0.90; RF: MSE=0.20, $R^2$=0.83; RA: MSE=0.56, $R^2$=0.43. KNN: MSE=0.83, $R^2$=0.21. According to Table 4, the importance levels of the variables according to the random forest method are given. According to this, when we rank the variables in the total group in order of importance, removing the need for blood products from the data set created an increase of 4.571% in the MSE in the model, while removing the fever variable from the model increased the MSE by -0.82%.

**Table 4.**Importance ranking of variables according to total group data

| Variables | % Increased MSE |
|---|---|
| Blood product requirement | 4,5731 |
| Muscle pain | 4,5549 |
| Age | 3,9794 |
| Lung involvement | 3,5421 |
| Cough | 3,0831 |
| Livestock breeding | 2,9521 |
| Konjonktıvıt | 2,5541 |
| Sore throat | 2,3247 |
| Diarrhoea | 2,2918 |
| Number of symptom days | 2,0858 |
| Place of residence | 2,0231 |
| Leukopenia | 1,8049 |
| Hepatomegaly | 1,7989 |
| Gender | 1,5921 |
| Malaise | 1,3793 |
| Haemorrhage | 1,3763 |
| Nausea | 1,1370 |
| Splenomegaly | 0,9324 |
| Vomiting | 0,8605 |
| Change of consciousness | 0,6211 |
| Jaundice | 0,4295 |
| Headache | 0,4254 |
| Rash | 0,4101 |
| Fever | -0,8258 |

## DISCUSSION

In this study, the prediction performance of different data mining methods on child, adult, and total groups was compared based on the mean square error (MSE) and R-squared ($R^2$) values. The results show that especially the SVR method outperforms the other methods with high accuracy and a low error rate. This supports the high generalization capacity of SVR as stated in the studies on Support Vector Regression (SVR) by Vapnik (1998) and Smola and Schölkopf (2004). In the pediatric group, the SVR method achieved the lowest error and the highest accuracy with an MSE = 2.00 and $R^2$ = 0.84. This is in line with the literature showing that SVR stands out especially in small sample groups and in applications requiring high accuracy (Vapnik, 1999; Frank et al., 2011). Although the random forest (RF) method performed close to SVR on the child group, the MSE value of 3.14 reveals the superiority of SVR in the child group. This finding can be considered in parallel with Breiman's (2001) studies on RF, which suggest that RF is effective on large data sets (Breiman, 2001). The KNN method performed better than RA. Here, the regression model with the lowest explanation rate is RA. In the adult group, the SVR method showed the strongest performance in terms of prediction accuracy, reaching the lowest $R^2$ value (0.30) and the highest $R^2$ value (0.98) (Smola and Schölkopf, 2004). High accuracy capacity, achieved by focusing on reducing the generalization error of SVR, is consistent with the findings in the adult group (Smola and Schölkopf,2004). RF and KNN methods have higher MSE and lower $R^2$ values than SVR in this group, with MSE = 2.42 and $R^2$ = 0.84 for RF and MSE = 2.20 and $R^2$ = 0.79 for KNN. The performance of RF in this group is in line with studies that have shown that it is a reliable method for classification and regression in large data sets (Breiman, 2001; Liaw and Wiener, 2002). In the total group analysis, the SVR method provided the best results with MSE = 1.35 and $R^2$ = 0.89. The fact that SVR achieved the lowest error and highest accuracy rates in all groups shows that the method has the capacity to best explain the data sets as a generalizable model. (Vapnik, 1999; Frank et al., 2011) RF and KNN methods had similar MSE and $R^2$ values for the total group, with MSE = 2.82 and $R^2$ = 0.85 and MSE = 2.54 and $R^2$ = 0.79, respectively. This is in line with the work of Breiman (2001) and Quinlan (1986), who state that tree-based methods generally offer good generalization ability (Breiman, 2001; Quinlan, 1986). In general, the SVR method stands out as the most successful model by reaching the highest $R^2$ and lowest MSE values in all groups. These results show that SVR is the method that best models the relationships between the variables in the dataset and has high predictive power. The RA method, on the other hand, exhibited the poorest performance with low $R^2$ and high MSE values in all three groups. This is in line with the literature that finds the performance of regression trees inadequate in limited data sets (Quinlan, 1986; Cover and Hart, 1967). RF and KNN methods lag behind SVR. According to the 1000 replicate simulation results, the Mean Square Error (MSE) and Coefficient of Explanation ($R^2$) values evaluate the performance of four different methods, namely SVR, RF, RA, and KNN, on child, adult, and total groups. The SVR method has the lowest $R^2$ values and the highest $R^2$ values for all three groups. It provides high explanatory power with $R^2$=0.95 for the child group, $R^2$=0.98 for the adult group, and $R^2$=0.96 for the total group. The low MSE values indicate that the prediction error of the SVR is low. SVR stands out as the best-performing method on these three groups. This suggests that SVR is a generalizable modeling method in different age groups and in the total population. RF ranks second after SVR and has relatively low MSE values and high $R^2$ values. However, it shows a higher error rate compared to SVR. With $R^2$=0.88, it provides very good explanatory power in the adult group. The MSE value is higher than the SVR (MSE=0.255). For the total group, it lags behind SVR with $R^2$=0.87 and MSE=0.292. Although the RF method is slightly less sensitive than the DVR, it can still be considered a highly effective prediction tool. RA shows the highest error and the lowest $R^2$ values in terms of MSE values. This reveals that the RA method is insufficient in predictions for child, adult, and total groups. It provides low explanatory power with $R^2$=0.34 in the child group, $R^2$=0.37 in the adult group, and $R^2$=0.32 in the total group. This shows that the model has a weak relationship with the data. The RA method showed the lowest performance compared to the other methods on these three groups. The KNN method also shows a low performance similar to RA. The MSE values were calculated as 0.81 in the child group, 0.82 in the adult group, and 0.80 in the total group. $R^2$ values are 0.20 in the child group, 0.19 in the adult group, and 0.20 in the total group. This shows that the model offers low explanatory power and does not establish a strong enough relationship with the data. KNN ranks the lowest in terms of prediction performance with RA. The method did not fit the different age groups and the total group well. The findings of this study show that the SVR method has the highest $R^2$ values and the lowest MSE values for all three sample sizes and correlation levels. This indicates that SVR is the model that best explains the dataset. The RF method performs second best after SVR with the highest accuracy and lowest error values. The RA method, on the other hand, shows a lower performance compared to SVR and RF but gives better results than the KNN method. The KNN method has the lowest R² and highest MSE values at all sample sizes and correlation levels and shows the lowest performance. When the significance levels of the results are compared according to the random forest method, the error rate increases when they are excluded from the model as blood product requirement 4.57%, muscle pain 4.55%, age 3.98%, lung involvement 3.54%, and cough 3.08%. These findings have not only statistical but also clinical implications. In particular, the ability of high accuracy models such as SVR and RF to predict a clinically important variable such as length of stay provides a great advantage in terms of optimizing the treatment process in diseases such as CCHF, which have a rapid progression and risk of death. Thanks to early prediction, both patient management and efficient use of health resources will be possible. Therefore, the integration of the outputs of our study into clinical decision support systems could be an important step towards improving the quality of patient care in the future.

## CONCLUSION

While the SVR method provides the best performance in all cases, the RF method follows DVR with a close performance. These findings suggest that it may be appropriate to prefer SVR and RF methods, especially for datasets requiring high accuracy. While the RA method provides acceptable performance in certain situations, the KNN method offers a lower performance and lags behind the other methods. The methods discussed in this study offer advantages and disadvantages for different data types and problems. Especially in medical data analysis, Random Forest and Support Vector Regression have been found to give more successful results. However, if the accuracy of the selected model is to be improved, steps such as feature engineering, variable selection, and hyperparameter optimization should be applied. Variable importance ranking based on the random forest method revealed that the most important factors in predicting disease course are clinical indicators such as blood product requirement and muscle pain. On the other hand, variables such as jaundice, rash, and headache contributed little to the predictive power of the model.

**Acknowledgment**

In the study, only the records of the relevant period could be accessed, and the number of patients was all patients in that period, which is the limitation of the study. It is recommended that more recent and larger data sets should be used in future studies.

*- This study is derived from the doctoral dissertation titled “Performance Comparison of Estimation Methods Used In Data Mining” which was completed in 2016.*

**Conflict of Interest**

There is no conflict of interest in this study.

## REFERENCES

Balinandi, S., Mulei, S., Whitmer, S., et al. (2024). Crimean-Congo hemorrhagic fever cases diagnosed during an outbreak of Sudan virus disease in Uganda, 2022–23. PLOS Neglected Tropical Diseases, 18(10), e0012595. https://doi.org/10.1371/journal.pntd.0012595

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support vector regression. Neural Information Processing-Letters and Reviews, 11(10), 203-224.

Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32. https://doi.org/10.1023/A:1010933404324

Cover, T. M., & Hart, P. E. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27. https://doi.org/10.1109/TIT.1967.1053964

Ergönül, Ö. (2006). Crimean-Congo haemorrhagic fever. The Lancet Infectious Diseases, 6(4), 203-214. https://doi.org/10.1016/S1473-3099(06)70435-2

Ersöz, F., & Çınar, Y. (2021). Veri madenciliği ve makine öğrenimi yaklaşımlarının karşılaştırılması: Tekstil sektöründe bir uygulama. Avrupa Bilim ve Teknoloji Dergisi, 29, 397-414. https://doi.org/10.31590/ejosat.997235

Frank, E., & Hall, M. A. (2011). Data mining: Practical machine learning tools and techniques (3rd ed.). Burlington, MA: Morgan Kaufmann.

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Waltham, MA: Morgan Kaufmann Publishers.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). New York:

Springer. https://doi.org/10.1007/978-0-387-84858-7

Ij, H. (2018). Statistics versus machine learning. Nature Methods, 15(4), 233. https://doi.org/10.1038/nmeth.4642

Karti, S. S., Odabaşı, Z., Korten, V., et al. (2004). Türkiye'de Kırım-Kongo kanamalı ateşi. Emerging Infectious Diseases, 10(8), 1379-1384. https://doi.org/10.3201/eid1008.030928

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18-22.

Ozgulbas, N., & Koyuncugil, A. S. (2009). Financial profiling of public hospitals: An application by data mining. International Journal of Health Planning and Management, 24(1), 69-83. https://doi.org/10.1002/hpm.932

Pala, T. (2013). Tıbbi karar destek sisteminin veri madenciliği yöntemleriyle gerçekleştirilmesi. Fen Bilimleri Enstitüsü, Elektronik-Bilgisayar Eğitimi Ana Bilim Dalı. Yüksek Lisans Tezi: Marmara Üniversitesi, İstanbul-Türkiye

Palmer, A., Jiménez, R., & Gervilla, E. (2011). Veri madenciliği: Makine öğrenimi ve istatistiksel teknikler. Veri Madenciliğinde Bilgi Odaklı Uygulamalar, 373-396.

Pupezescu, V., & Ionescu, F. (2008). Advances in knowledge discovery in databases. Journal of Applied Economic Sciences, 4(6), 433-438.

Quinlan, J. R. (1986). Induction of decision trees. Machine Learning, 1(1), 81-106. https://doi.org/10.1007/BF00116251

Savaş, S., Topaloğlu, N., & Yılmaz, M. (2012). Veri madenciliği ve Türkiye'deki uygulama örnekleri. İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 11(21), 1-23.

Seçmeer, G., & Çelik, İ. H. (2010). Kırım Kongo Kanamalı Ateşi. Journal of Pediatric Infection/Çocuk Enfeksiyon Dergisi, 4(4), 156-160. https://doi.org/10.5152/ced.2010.15

Sihombing, P. R., Budiantono, S., Arsani, A. M., Aritonang, T. M., & Kurniawan, M. A. (2023). Comparison of regression analysis with machine learning supervised predictive model techniques. Jurnal Ekonomi Dan Statistik Indonesia, 3(2), 113-118. https://doi.org/10.11594/jesi.03.02.06

Silahtaroğlu, G. (2008). Data mining: Concepts and algorithms. İstanbul: Papatya Publishing.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. Statistics and Computing, 14(3), 199-222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Tüzüntürk, S. (2010). Veri madenciliği ve istatistik. Uludağ Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi, 29(1), 65-90.

Vapnik, V. N. (1999). An overview of statistical learning theory. IEEE Transactions on Neural Networks, 10(5), 988-999. https://doi.org/10.1109/72.788640