COMPARING MACHINE LEARNING METHODS FOR COVID-19 FOOD SUPPLY QUANTITATIVE DATA

Özlem Kaymaz – Fatma Zehra Doğru

Abstract

Machine learning methods have been very popular recently, especially ensemble learning methods. Ensemble learning is a general meta-approach to machine learning in which multiple models are combined to produce better predictive performance. The model prediction errors are reduced using ensemble learning methods. In this study, we compare the performance of several important and widely used machine learning methods such as Boosting, and Bagging ensemble learning methods using the COVID-19 Healthy Diet Dataset from Kaggle. The dataset includes quantities of various food group supplies, nutrition values, obesity, undernourished percentages, and global COVID-19 cases. This dataset was collected to illustrate the relationship between healthy eating and COVID-19 cases ending with death. Firstly, we use the backward elimination method for selecting significant variables in the COVID-19 data. Then, we apply some of the machine-learning methods, namely Linear Model (LR), Support vector machine (SVM), K-nearest neighbor (KNN), Bagging ensemble methods (Bagged Trees, Random Forest (RF)) and Boosting ensemble methods (Gradient boosting machine (GBM), Extreme gradient boosting (XGboost), and Bayesian additive regression tree (BART)). Finally, we compare the performances of the machine learning methods with some measures such as R^2, root mean square error (RMSE), and mean absolute error (MAE).

Keywords: Machine learning, ensemble learning, boosting, bagging, COVID-19

JEL Code: C38, C52, C53

Introduction

The immune system is strengthened and diseases are prevented by a balanced, healthy diet. World Health Organization (WHO), Turkish Dietetic Association, and Food and Agriculture Organization (FAO) have all provided dietary advice to combat COVID-19 during the pandemic. However, some people's eating habits have also changed as a result of factors like

the isolation period, psychological issues like stress and anxiety, and inactivity. In this instance, they have contributed to a rise in disorders like diabetes and obesity. The possible risk of COVID-19 contracting these diseases has increased. As a result, nutrition becomes a crucial factor in COVID-19 treatment and illness prevention.

The link between COVID-19 and nutrition has been the subject of numerous investigations in the literature (see Zhao et al. (2020), Li et al. (2021), and Bousquet et al. (2020)). In this study, the COVID-19 Healthy Diet Dataset from Kaggle is used to evaluate the association between COVID-19 and nutrition using machine learning (ML) techniques. This dataset is available at "https://www.kaggle.com/datasets/mariaren/covid19-healthy-diet-dataset". Several researchers have used the same dataset, including García-Ordás et al. (2020), who used principal component analysis (PCA) and K-means for classification using machine learning algorithms, Shams et al. (2021a), who looked at correlations between variables using SVM and deep learning methods for classification, and Shams et al. (2021b), who used elastic net regression, PCA, and AdaBoost techniques.

In the present study, we compared the performance of ensemble and classical models using a few performance measures and studied relationships between various food types and the proportion of COVID-19 death cases.

1 Methodology

1.1 Machine Learning Algorithms

In this study, we apply some of the machine-learning methods, namely multiple linear regression (MLR), support vector machine (SVM), K-nearest neighbor (KNN), bagging ensemble methods (bagged trees, random forest (RF)), and boosting ensemble methods (Gradient boosting machine (GBM), extreme gradient boosting (XGboost), and Bayesian additive regression tree (BART)).

1.1.1 Multiple Linear Regression Model (MLR)

Let y be a dependent variable and $x_2, \dots, x_k$ be independent variables. The relationship between dependent and independent variables can be defined with the following multiple linear regression (MLR) model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \quad (1)$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are the regression coefficients and ϵ is the random error. It is known that the least squares estimation or maximum likelihood estimation methods can be used to estimate the regression coefficients in the model (1).

1.1.2 Support Vector Machine (SVM)

Support vector machine (SVM) is a supervised machine learning method based on the associated learning algorithms that ensure data analysis for classification and regression. This method was proposed by Cortes and Vapnik (1995). SVM aims to maximize the margin between support vectors and hyperplane and to find the optimal separating hyperplane between classes. In regression problems, support vector regression (SVR) is a developed algorithm for regression and functional approach. SVR uses a core function to transform classes into high-dimensional areas for linearly separable. That is, SVR is based on the classification of regression errors that are above or below a certain threshold. The basic kernel functions are linear, polynomial, and Gaussian functions.

1.1.3 K-Nearest Neighbor (KNN) Algorithm

The K-Nearest Neighbors (KNN) algorithm is a non-parametric method for classification and regression that was first defined by Fix and Hodges (1951). In recent years, it has been widely used in ML applications. The neighbors are selected from a training set, which is a collection of observations with similar features or values. Based on the value(s) provided, KNN chooses the values that are closest neighbors. It classifies the closest neighbors as a k number based on distances. The amount of closest neighbors to a given value is expressed as k. The value is then assigned to the category class with the highest frequency of repetition.

1.1.4 Bagged Tree Algorithm

Bootstrap aggregation or bagging is known as an ensemble method that is used to decrease the variance of a statistical learning method. This method, the earliest ensemble method, was offered by Breiman (Breiman, 1996). This approach relies on selecting a large number of training sets from the population, creating a unique prediction model for each training set, and averaging the results. These processes reduce the variance and increase the test set accuracy. Further, repeated samples from the (single) training data sets based on the bootstrap can be used to make this procedure more practicable. Then, B different bootstrapped training data sets are

generated and predictions of bagging can be computed by taking the average of the bth bootstrapped training data sets, $\hat{f}^{*1}(x), \hat{f}^{*2}(x), \ldots, \hat{f}^{*b}(x)$ which $\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{*b}(x)$.

1.1.5 Random Forest (RF) Algorithm

Using the random subspace technique, Ho (1995) created the first random decision forest algorithm, then Breiman (2001) improved it. Another ensemble technique used in classification and regression is random forest. By making a small modification that decorrelates the trees, it offers an enhancement over bagged trees. Several decision trees are enhanced on bootstrapped training samples, similar to bagging. But the bagged trees all resemble one another quite a bit, and they all make highly correlated predictions. This indicates that bagging does not significantly reduce variation over a single tree. This issue is solved by random forests by picking additional trees at random. As a result, when there is less correlation between random trees, performance is improved.

1.1.6 Gradient Boosting Machine (GBM)

Boosting methods work similarly to bagging methods; however, trees of gradient boosting machine (GBM) are not independent of the other trees. It means that each tree is grown using information from previously grown trees. GBM is also an additive modeling algorithm. Since it composes a model by iteratively adding M weak sub-models that are obtained to be based on the performance of the previous iteration. GBM repeats this process over and over, thus the model becomes stronger.

GBM is a result of Breiman's (1997) idea that boosting may be seen as an optimization technique for an appropriate cost function. Following this, explicit regression gradient boosting methods were established by Friedman (Friedman, 2001). While GBM trees function similarly to bagging approaches, they are not independent of one another. It implies that data from previously grown trees are used to guide the growth of each new tree. Since M weak sub-models that are generated based on the results of the previous iteration are added iteratively, the model is constructed. GBM keeps doing this, and as a result, the model gets stronger. The boosted model can be given as follows:

$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}^{b}(x) \tag{2}$$

where λ is the shrinkage parameter and B is the number of trees.

1.1.7 Extreme Gradient Boosting (XGboost)

Extreme Gradient Boosting (XGBoost) was first developed by Chen (2016) as a research project for the Deep Machine Learning Community (DMLC) group. It started out as a terminal application that could be configured with a libsvm, an open-source ML library, configuration file. Following the development of the Python and R packages, XGBoost has currently package implementations for Java, Scala, Julia, Perl, and more languages.

The additive modeling approach is known as XGBoost enhanced the gradient boosting process. Each faulty prediction is given a weight based on how poorly the learner performed. By adding up all of the basic learners' weights, a prediction is generated. Due to the inclusion of a regularization element, Xgboost is computationally more effective and less prone to overfitting. These two features improved the model's performance and predictive ability. Since then, it has become frequently utilized.

1.1.8 Bayesian additive regression tree (BART)

Another ensemble technique used in classification and regression issues is the Bayesian additive regression tree (BART) developed by Chipman et al. (2010). BART is connected to both bagging and boosting strategies. As in bagging and random forests, each tree is first constructed randomly. As in boosting, each tree then seeks to capture signals that the current model hasn't yet been able to account for. The method used for growing new trees is the key innovation in the BART. Here, this method uses the Bayesian approach to fitting an ensemble of trees. A new tree is produced from a posterior distribution when a tree is altered randomly each time.

1.2 Measures of Model Performance

To evaluate the model performance, R^2, RMSE, and MAE measures are used.

1.2.1 Coefficient of Determination (R^2)

It represents the change in the dependent variables explained by the model. It is calculated with the ratio of the total sum of squares and the sum of residual squares as in equation (3). R^2 is defined in 0 to 1 which has the following formula:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} y_i - \bar{y}}. \quad (3)$$

1.2.2 Root Mean Squared Error (RMSE)

Root mean squared error (RMSE) determines the root square of differences between the predicted and observed values. RMSE is calculated by using the following equation:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \quad (4)$$

1.2.3 Mean Absolute Error (MAE)

Mean absolute error (MAE) is computed by the average of the absolute values of the differences between the predicted and observed values. Its formula is given below:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|. \quad (5)$$

2 Application

This section consists of an application of a data set using the ML methods mentioned above.

2.1 Data Set

Nutrition is known to be effective in the occurrence of diseases. The effect of nutrition and/or diets on COVID-19 has been a major subject that has been researched and concern worldwide, particularly during the pandemic. Some food choices can positively affect people's mental and physical health, as well as some food harm with continuous consumption and even increase the possibility of death. In this study, we evaluate several ML methods on COVID-19 Food Supply Quantitative Data provided by Kaggle. The dataset includes quantities of various food group supplies, nutrition values, obesity, undernourished percentages, and global COVID-19 cases from 170 countries around the world.

2.2 Data Pre-Processing

Firstly, we explored the dataset for missing data, outliers, etc. We examined the descriptive statistics of the dataset to identify the observations. While there were no outliers in the data set, 7 missing observations were removed from the data set. After this stage, we performed variable selection to identify the important variables that affect the proportions of COVID-19 deaths. It consists of 32 variables related to quantities of various food group supplies, nutrition values, obesity, and undernourished percentages. The 7 variables related to the proportions of COVID-19 deaths selected by the backward elimination method were determined as follows; animal

products, miscellaneous foods, animal fats, stimulants, vegetable oils, vegetal products, and obesity.

2.3 Data Splitting

We divided the data set into 80% training set and 20% test set. Using training data, models were trained and optimized. The models were validated using a 10-fold cross-validation technique. The models were then tested and validated on the test set. For hyperparameter optimization, the grid search approach with cross-validation is also utilized. Grid search is used to fine-tune the hyperparameters of various conventional ML models by generating discrete grids within the hyperparameter domain and picking a set of parameters that provides the best performance.

3 Results

Table 1 shows the descriptive statistics of the significant variables in the model after the backward elimination method.

Tab. 1: Descriptive Statistics of Variables

Variables	Mean±Sd	Median	Min-Max
Animal Products	20.679±8.027	20.928	5.02-36.90
Animal Fats	4.185±3.322	3.317	0.034-14.937
Miscellaneous	0.05±0.069	0.030	0-0.456
Stimulants	0.645±0.705	0.383	0-3.38
Vegetable oils	18.663±6.751	18.351	4.955-36.419
Vegetable Products	29.320±8.026	29.07	13.10-44.98
Obesity	18.701±9.420	21.30	2.10-45.50
Deaths	0.039±0.049	0012	0-0.185

Source: own computations in the R software

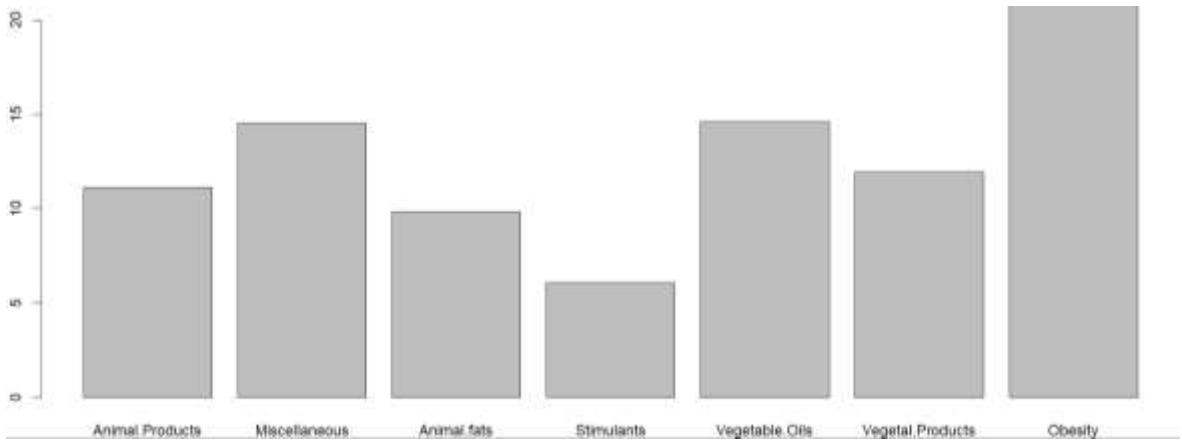
The performances of SVM, RF, and bagged tree models outperformed the performances of MLR, KNN, GBM, Xgboost, and BART models. The estimation results can be found in Table 2. This table consists of RMSE, MAE, and R^2 values. According to RMSE and MAE values, the best performance was obtained from the bagged tree model; on the other hand, the highest R^2 value was obtained from the SVM model.

Tab. 2: The Comparison of Model Performances

ML Models	R^2	RMSE	MAE
MLR	0.5406889	0.04042023	0.03014874
KNN	0.5098129	0.04055903	0.02832495
SVM	0.6470742	0.03791384	0.02707881
RF	0.6154767	0.03711299	0.02665569
Bagged Tree	0.61405	**0.03634035**	**0.02431243**
GBM	0.4984619	0.04147751	0.03177391
Xgboost	0.5608378	0.03791323	0.02621772
BART	0.512406	0.04102906	0.03121084

Source: own computations in the R software

Fig. 2: Variable importance plot for the COVID-19 Healthy Diet data set resulting from the bagged tree model



Source: own computations in the R software

Figure 2 displays the variable importance plot of the COVID-19 Healthy Diet data set yielded from the bagged tree model. The overall importance of each predictor is determined by calculating the total diminution in the residual sum of squares. As can be seen in Figure 2, obesity has a larger value so it is a more important predictor than others. The least significant predictor among the variables is the stimulant.

Conclusion

In this study, several ML methods, which have been widely used in recent years, were applied to the COVID-19 Healthy Diet data set. The data set includes quantities of various food group supplies, nutrition values, obesity, undernourished percentages, and global COVID-19 cases

from 170 countries around the world. The data set was collected to show the effect of healthy and sufficient nutrition on COVID-19 cases.

Before starting the application, we focused on the data pre-processing and variable selection. We determined significant and correlated variables by using the backward elimination method in COVID-19 cases ending with death. After that, MLR, KNN, SVM, RF, bagged tree, GBM, Xgboost, and BART models were used to estimate the COVID-19 cases ending with death due to nutrition. The performances of the models obtained as a result of these algorithms were evaluated with R^2, RMSE, and MAE metrics. According to the results of metrics, the lowest RMSE and MAE were found in the bagged tree model. The RF model follows the bagged tree model. Although the R^2 value of SVM was higher than the Bagged tree, in terms of model adequacy, the best model is the bagged tree model according to values of RMSE and MAE. As a result, the bagged tree model, which is one of the bagging ensemble methods, is estimated with lower error than classical methods and boosting ensemble methods.

References

Bousquet, J., Anto, J. M., Iaccarino, G., Czarlewski, W., Haahtela, T., Anto, A., Akdis, C. A., Blain, H., Canonica, G. W., Cardona, V., Cruz, A. A., Illario, M., Ivancevich, J. C., Jutel, M., Klimek, L., Kuna, P., Laune, D., Larenas-Linnemann, D., Mullol, J., … Zuberbier, T. (2020). Is diet partly responsible for differences in covid-19 death rates between and within countries? *Clinical and Translational Allergy, 10*(1).

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24, 123-140.

Breiman, L. (1997). Arcing the edge (pp. 1-14). Technical Report 486, Statistics Department, University of California at Berkeley.

Breiman, L. (2001). Random forests. *Machine learning*, 45, 5-32.

Chen, T. (2016). Story and lessons behind the evolution of XGBoost. 2016-03-10)[2020-06-24]. https://homes. cs. Washington. edu.

Chipman, H. A., George, E. I., & McCulloch, R. E. (2010). BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1), 266-298.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*, 273-297.

Fix, E., & Hodges, J. L. (1951). Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties (PDF) (Report). USAF School of Aviation Medicine, Randolph Field, Texas.

Friedman, J., H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189-1232.

García-Ordás, M. T., Arias, N., Benavides, C., García-Olalla, O., & Benítez-Andrades, J. A. (2020). Evaluation of country dietary habits using machine learning techniques in relation to deaths from covid-19. *Healthcare*, *8*(4), 371.

Ho, T. K. (1995). Random decision forests. In Proceedings of 3rd international conference on document analysis and recognition (Vol. 1, pp. 278-282). IEEE.

Li, G., Zhou, C., Ba, Y., Wang, Y., Song, B., Cheng, X., Dong, Q., Wang, L., & You, S. (2021). Nutritional risk and therapy for severe and critical COVID-19 patients: A multicenter retrospective observational study. *Clinical Nutrition*, *40*(4), 2154–2161.

Shams, M. Y., Elzeki, O. M., Abd Elfattah, M., Abouelmagd, L. M., Darwish, A., & Hassanien, A. E. (2021). Impact of covid-19 pandemic on diet prediction and patient health based on support vector machine. *Advances in Intelligent Systems and Computing*, 64–76.

Shams, Mahmoud Y., Elzeki, O. M., Abouelmagd, L. M., Hassanien, A. E., Elfattah, M. A., & Salem, H. (2021). Hana: A Healthy Artificial Nutrition Analysis Model during COVID-19 pandemic. *Computers in Biology and Medicine*, *135*, 104606.

Zhao, X., Li, Y., Ge, Y., Shi, Y., Lv, P., Zhang, J., Fu, G., Zhou, Y., Jiang, K., Lin, N., Bai, T., Jin, R., Wu, Y., Yang, X., & Li, X. (2020). Evaluation of Nutrition Risk and its association with mortality risk in severely and critically ill Covid-19 patients. *Journal of Parenteral and Enteral Nutrition*, 45(1), 32–42.

Contact

Özlem Kaymaz

Ankara University, Faculty of Science, Department of Statistics

Ankara Üniversitesi, Fen Fakültesi, İstatistik Bölümü, 06100 Beşevler /Ankara/Turkey

ozlgullu@ankara.edu.tr

Fatma Zehra Doğru

Giresun University, Faculty of Arts and Sciences, Department of Statistics

Giresun Üniversitesi, Gaziler Mah., Prof. Ahmet Taner Kışlalı Cd, 28200 Giresun/Turkey

fatma.dogru@giresun.edu.tr