Türkiye Klinikleri Biyoistatistik Dergisi
Turkiye Klinikleri Journal of Biostatistics Turkiye Klinikleri J Biostat. 2023;15(3):171-81

ORIGINAL RESEARCH ORİJİNAL ARAŞTIRMA

DOI: 10.5336/biostatic.2023-99050

Explainable Machine Learning Methods for Person-Based Prediction in Simulated and Real Datasets: Methodological Research

Simüle ve Gerçek Veri Setlerinde Kişi Temelli Tahmin İçin Açıklanabilir Makine Öğrenmesi Yöntemleri: Metodolojik Araştırma

İrem KAR,[a] Batuhan BAKIRARAR[a]

[a]Department of Biostatistics, Ankara University Faculty of Medicine, Ankara, Türkiye

ABSTRACT Objective: The aim of this study is to build person-based prediction models for simulated and real datasets separately with the SHapley Additive exPlanations method, and to demonstrate whether the obtained person-based models are more valid and applicable than overall models. **Material and Methods:** Simulated datasets encompassed 13 independent and 1 dependent variable, across sample sizes of 250, 500, and 1,000, while the real dataset contained 826 patient records with 11 variables. "bindata", "shaper" and "RWeka" packages in the R (version 4.1.2) programming language were used. Extreme Gradient Boosting, Bagging, Random Forest, Support Vector Machine and Logistic Regression were used as classification methods. The assessment employed 10-fold cross-validation, repetaed 1,000 times. **Results:** Accuracy values of the overall model in the datasets with 250, 500, and 1,000 samples were found to be 0.856, 0.886, and 0.891, respectively. In these samples, the person-based accuracy values were found to be 0.886, 0.964, and 0.962 for those with "yes" prediction results, and 0.930, 0.961, and 0.961 for those with "no" prediction results, respectively. In the real dataset, the accuracy of the overall model was found to be 0.736. The person-based accuracy values were found to be 0.783 in the patient who was predicted with stroke, and 0.868 in the patient who was predicted without stroke. **Conclusion:** Person-based predictions consistently outperformed model-based results across datasets due to real-life individual heterogeneity, emphasizing the need for attention. Considering this diversity, person-based modeling is expected to produce a more realistic and clinically applicable model.

Keywords: Prediction; person-based prediction models; SHapley Additive exPlanations

ÖZET Amaç: Bu çalışmanın amacı, "SHapley Additive exPlanations" yöntemi ile simüle ve gerçek veri setleri için ayrı ayrı kişi temelli tahmin modelleri oluşturmak ve elde edilen kişi temelli modellerin genel modellere göre daha geçerli ve uygulanabilir olup olmadığını göstermektir. **Gereç ve Yöntemler:** Simüle veri setleri sırasıyla 250, 500 ve 1.000 örneklem büyüklükleriyle 13 bağımsız ve 1 bağımlı değişken içerirken, gerçek veri seti 11 değişkenden oluşmakta olup, 826 hasta verisi içermektedir. Analizler için R (versiyon 4.1.2) programlama dilindeki "bindata", "shapper" ve "RWeka" paketleri kullanılmıştır. Sınıflandırma yöntemleri olarak "Extreme Gradient Boosting", Bagging, Rastgele Orman, Destek Vektör Makinesi ve Lojistik Regresyon kullanılmıştır. Veri seti 10-kat çapraz doğrulama kullanılarak değerlendirilmiş ve analizler 1.000 kez tekrarlanmıştır. **Bulgular:** 250, 500 ve 1.000 örneklem büyüklüğüne sahip veri setlerinde genel modelin doğruluk değerleri sırasıyla 0,856, 0,886 ve 0,891 olarak bulunmuştur. Bu örneklem büyüklüklerinde kişi temelli doğruluk değerleri "evet" tahmin sonucuna sahip olanlar için sırasıyla 0,886, 0,964 ve 0,962; "hayır" tahmin sonucuna sahip olanlar için ise sırasıyla 0,930, 0,961 ve 0,961 olarak bulunmuştur. Gerçek veri setinde, genel modelin doğruluğu 0,736 olarak bulunmuştur. Kişi temelli doğruluk değerleri ise inme tahmini yapılan hastada 0,783, inme tahmini olmayan hastada ise 0,868 olarak bulunmuştur. **Sonuç:** Tüm veri setlerinde kişi temelli tahmin sonuçları, model bazlı sonuçlardan daha yüksek bulunmuştur. Bu gerçek hayatta kişiler arası heterojenite nedeniyle göz ardı edilmemesi gereken bir durumdur. Bu farklılık göz önünde bulundurularak, kişi temelli modelleme yapıldığında, modelin daha gerçekçi olacağı ve klinik kullanıma daha uygun hâle geleceği düşünülmektedir.

Anahtar kelimeler: Tahmin; kişi temelli tahmin modelleri; SHapley Additive exPlanations

Correspondence: Batuhan BAKIRARAR
Department of Biostatistics, Ankara University Faculty of Medicine, Ankara, Türkiye
E-mail: batuhan_bakirarar@hotmail.com

Peer review under responsibility of Turkiye Klinikleri Journal of Biostatistics.

Received: 11 Aug 2023 ***Received in revised form:*** 02 Oct 2023 ***Accepted:*** 04 Oct 2023 ***Available online:*** 17 Oct 2023

Nowadays, predictive modeling is one of the most popular and important methodologies used in clinical and health research.[1] Risk assessment has been successfully applied in many areas, including diagnosis, prognosis, and the effect of treatment.[2,3] The conventional method of modeling clinical outcomes involves constructing a singular predictive model using a dataset of individuals with known outcomes, followed by applying this model to forecast outcomes for upcoming individuals.[4] This type of model is referred to as a population model, as its purpose is to be employed across an entire future population of individuals, aiming to exhibit strong average predictive performance across the entire population.[4,5] Recent research in the field of personalized medicine indicates that the patient population displays heterogeneity, with each patient possessing distinct attributes, underscoring the significance of tailored, patient-specific predictions, suggestions, and interventions.[4]

Person-based prediction models are customized for an individual patient and trained using information on characteristics from similar patients. Compared with overall models trained on all patients, it can produce more accurate risk scores and identify more relevant risk factors for each patient.[4] The overall model captures information that is important to the entire patient population, but may miss information that is important to individual patients and less important to the general population. Another option is to establish a patient-specific or "personalized" predictive model for each patient.[5] The model is individualized for each patient, utilizing data from the patient themselves or those sharing clinically similar attributes. Given their dynamic training tailored to specific patients, personalized predictive models can harness the most pertinent patient data and effectively discern crucial patient-specific risk factors.[4] The most important step for person-based prediction is to identify the most important variables specific to that person. This type of approach is called person-based variable importance, which has recently been used frequently, especially in the field of health recently.[1,4] With the importance of person-based variables, separate models can be created for individuals with the same characteristics, and better prediction results can be obtained with these models.[4,5] The number of R packages created to solve the person-based variable importance problem in the literature is quite low. The most commonly used R packages are shapper and LIME. In the study, the shapper package was preferred because it can be applied to both qualitative and quantitative data and offers more advanced features.

This study aimed to build person-based prediction models for each of the datasets generated by the SHapley Additive exPlanations (SHAP) method and a real dataset and to show that the obtained person-based prediction models are more valid and applicable than overall models.

MATERIAL AND METHODS

In this study, which was planned as a methodological study, three simulated datasets with different sample sizes and a real dataset from an open access database were used. Samples were randomly selected for person-based analyses.

SIMULATED DATASETS

Correlation coefficients between dependent and independent variables were defined as (0.30-0.70) for all scenarios. Correlation coefficients between independent variables were designed in the range of (0.30-0.60). The datasets consist of 14 variables, with 13 independent variables and one dependent variable. The sample sizes were generated as 250, 500 and 1,000, respectively.

REAL DATASET

The original dataset was accessed from the Kaggle database and included 42,617 patients considered at risk of stroke.[6] The dataset consists of 11 variables, including 10 independent [age, gender, body mass index (BMI), smoking status, hypertension, marital status, heart disease, average glucose level, work type, and residence type] and one dependent (presence of stroke) (Table 1). After removing the missing and incor-

rectly entered data in the dataset, a sample of 826 (2% of the original data) was randomly selected from the remaining data, so as not to disturb the distribution of the original dataset, and the study was carried out using this sample.

TABLE 1: Variables for stroke dataset.

Variables	Values
Gender	Female/male
Age	19-90
Hypertension	No/yes
Heart disease	No/yes
Marital status	No/yes
Work type	Never worked/goverment/self-employed/private sector
Residence type	Urban/rural
Average glucose level	60.06-250.89
BMI	17.1-44.8
Smoking status	No/yes
Stroke	No/yes

BMI: Body mass index.

SHAP

SHAP, created by Lundberg and Lee, is a method for describing individual predictions and is a game-theoretic approach. SHAP is based on the Shapley values in game theory. Shapley values correspond to the contribution measures of each feature in a machine learning model. SHAP can be implemented in Python (shap library) and R (DALEX and shapr packages) programming languages.[7]

STATISTICAL ANALYSIS

In the analysis, bindata, shapper, DALEX and RWeka packages in the R (ver 4.1.2) programming language were used.[8-11] InfoGain and Logistic Regression (LR) methods were used for variable importance. Extreme Gradient Boosting (XGBoost), Bagging, Random Forest (RF), Support Vector Machine (SVM), and LR were used as classification methods. The dataset was evaluated using 10-fold cross validation, and all analyzes were repeated 1,000 times.

RESULTS

XGBoost, Bagging, RF, SVM and LR were used to evaluate the overall model performance for all datasets, and the best performance was achieved with the LR method. Therefore, the LR results for the overall model are presented.

SIMULATED DATASETS

In Figure 1, the variable importance was examined in a simulated dataset with 250 samples by using the variable importance tests of InfoGain and LR. The overall model consisted of x9, x1, x8, x3, x12, x13, x2 and x10 variables according to the results of the variable importance test performed with the InfoGain method and clinical importance. As a result, 9 variables (8 independent, 1 dependent variable) were included in the study for the overall model and machine learning analyzes were performed using these variables (Figure 1A). The results of the variable importance test for each of the data sets including patients classified as yes and no are given in Figure 1B and Figure 1C. These results differ from the variable importance results for the overall model presented in Figure 1A.

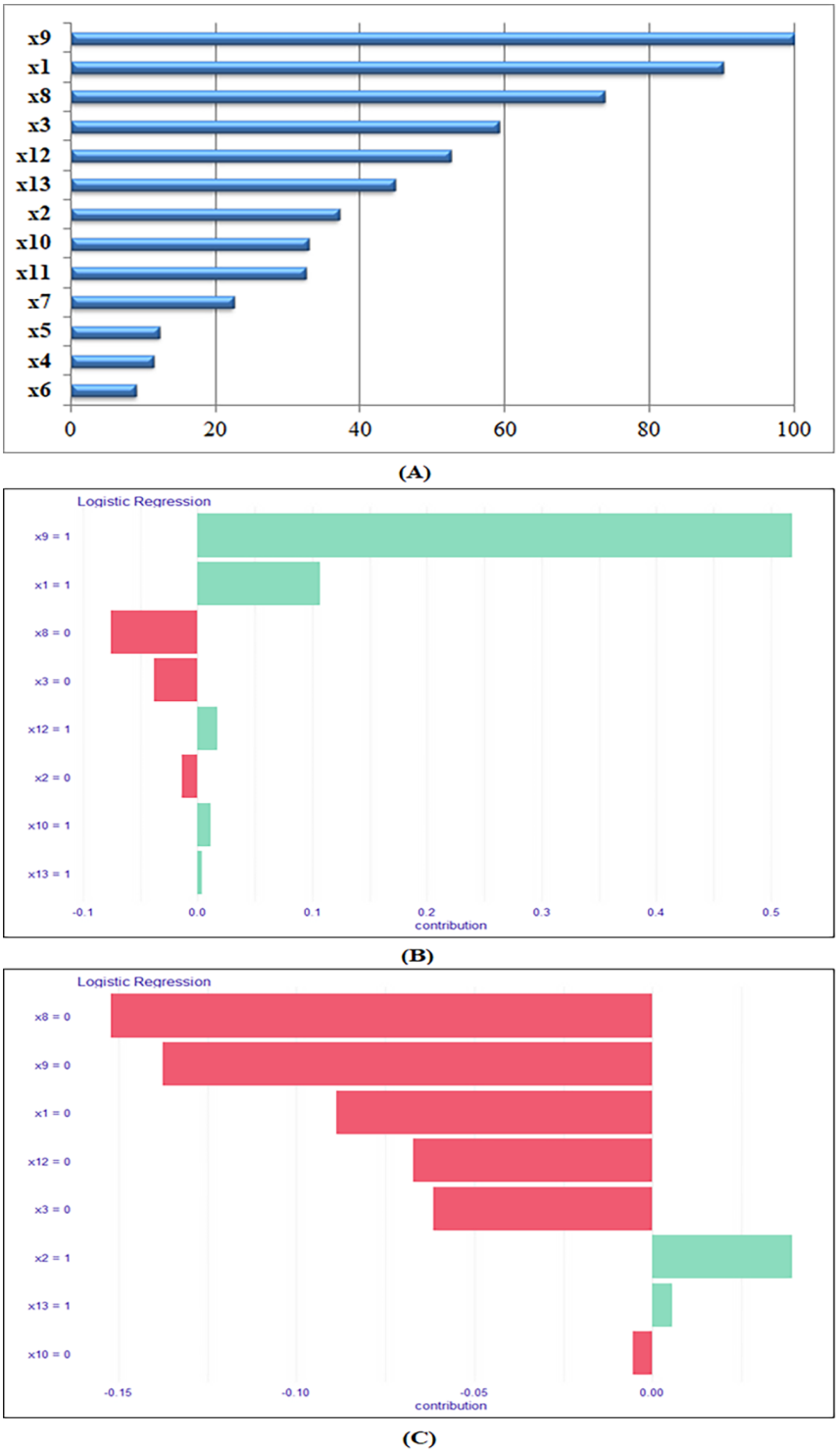

FIGURE 1: Variable importance for 250 samples. A) Overall model, B) Patient classified as "yes", C) Patient classified as "no".

For the overall model with 250 samples, the accuracy value was found to be 0.847 in patients with the outcome variable yes (classified as yes) and 0.864 in patients with the outcome variable no (classified as no). The accuracy value was 0.886 in the sample patient with the outcome variable yes, and 0.930 in the sample patient with the outcome variable no. As a result, the accuracy values of the sample patients were found to be higher than the values of the overall model (Table 2).

In Figure 2, the variable importance was examined in a simulated dataset with 500 samples by using InfoGain and LR methods. The overall model consisted of x9, x1, x3, x8, x13, x12, x11 and x10 variables according to the results of the variable importance test performed with the InfoGain method and clinical importance. As a result, 9 variables (8 independent, 1 dependent variable) were included in the study for the overall model and machine learning analyzes were performed using these variables (Figure 2A). The results of the variable importance test for each of the data sets including patients classified as yes and no are given in Figure 2B and Figure 2C. These results differ from the variable importance results for the overall model presented in Figure 2A.

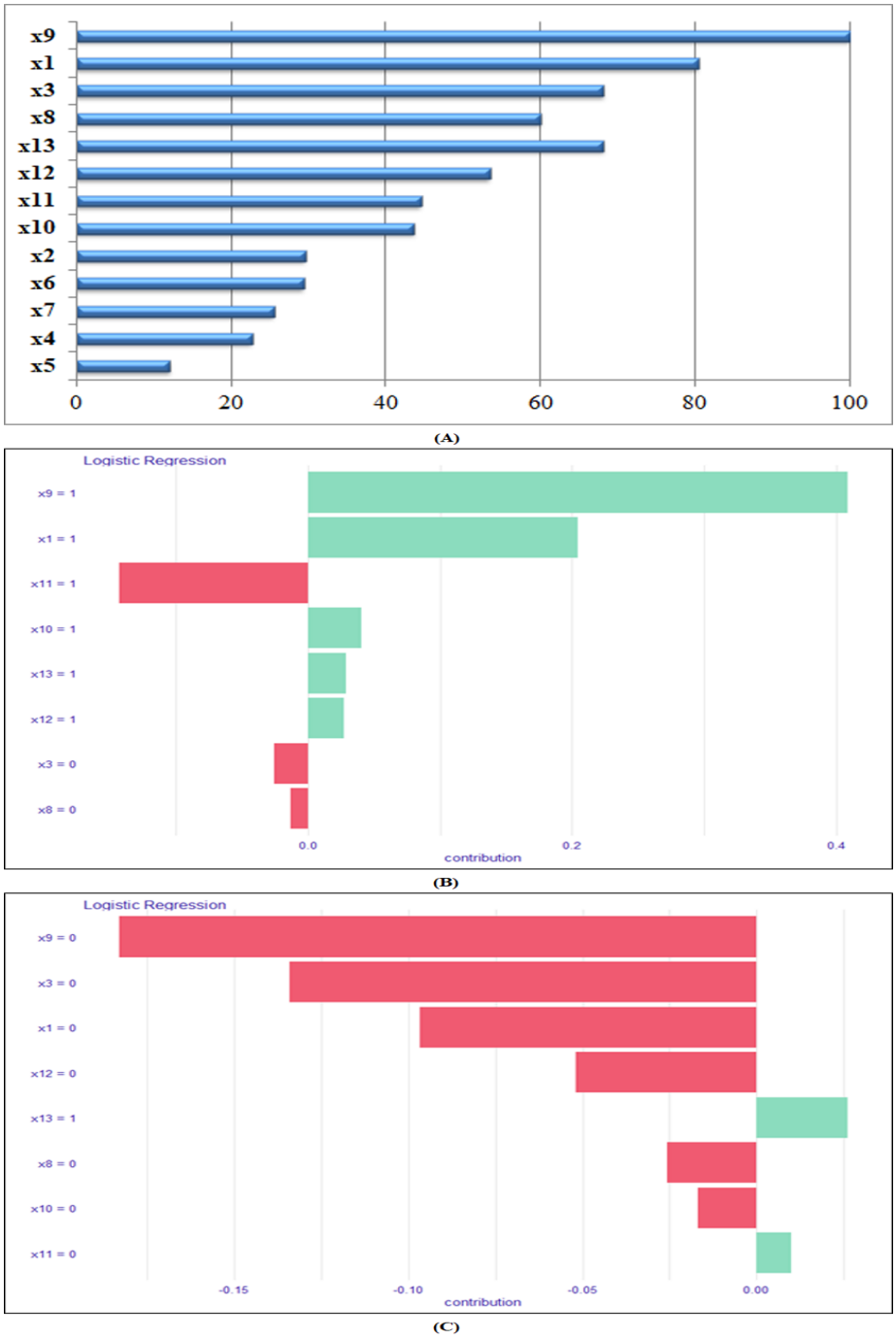

FIGURE 2: Variable importance for 500 samples. A) Overall model, B) Patient classified as "yes", C) Patient classified as "no".

For the overall model with 500 samples, the accuracy value was found to be 0.866 for patients with the outcome variable yes (classified as yes) and 0.906 for patients with the outcome variable no (classified as no). The accuracy value was 0.964 in the sample patient with the outcome variable yes, and 0.961 in the sample patient with the outcome variable no. As a result, the accuracy values of the sample patients were found to be higher than the values of the overall model (Table 2).

TABLE 2: Performance measures for simulated datasets.

Datasets	Model		Accuracy	F-measure	MCC	ROC area
250 sample	All samples	No	0.864	0.864	0.711	0.924
		Yes	0.847	0.847		
		Overall	0.856	0.856		
	Patient 1	No	0.114			
		Yes	0.886			
	Patient 2	No	0.930			
		Yes	0.070			
500 sample	All samples	No	0.906	0.890	0.772	0.949
		Yes	0.866	0.882		
		Overall	0.886	0.886		
	Patient 1	No	0.036			
		Yes	0.964			
	Patient 2	No	0.961			
		Yes	0.039			
1,000 sample	All samples	No	0.900	0.892	0.782	0.955
		Yes	0.882	0.890		
		Overall	0.891	0.891		
	Patient 1	No	0.038			
		Yes	0.962			
	Patient 2	No	0.961			
		Yes	0.039			

MCC: Matthew's correlation coefficient; ROC: Receiver operating characteristic.

In Figure 3, the variable importance was examined in a simulated dataset with 1,000 samples by using InfoGain and LR methods. The overall model consisted of x9, x1, x3, x12, x8, x13, x10 and x11 variables according to the results of the variable importance test performed with the InfoGain method and clinical importance. As a result, 9 variables (8 independent, 1 dependent variable) were included in the study for the overall model, and machine learning analyzes were performed using these variables (Figure 3A). The results of the variable importance test for each of the data sets including patients classified as yes and no are given in Figure 3B and Figure 3C. These results differ from the variable importance results for the overall model presented in Figure 3A.

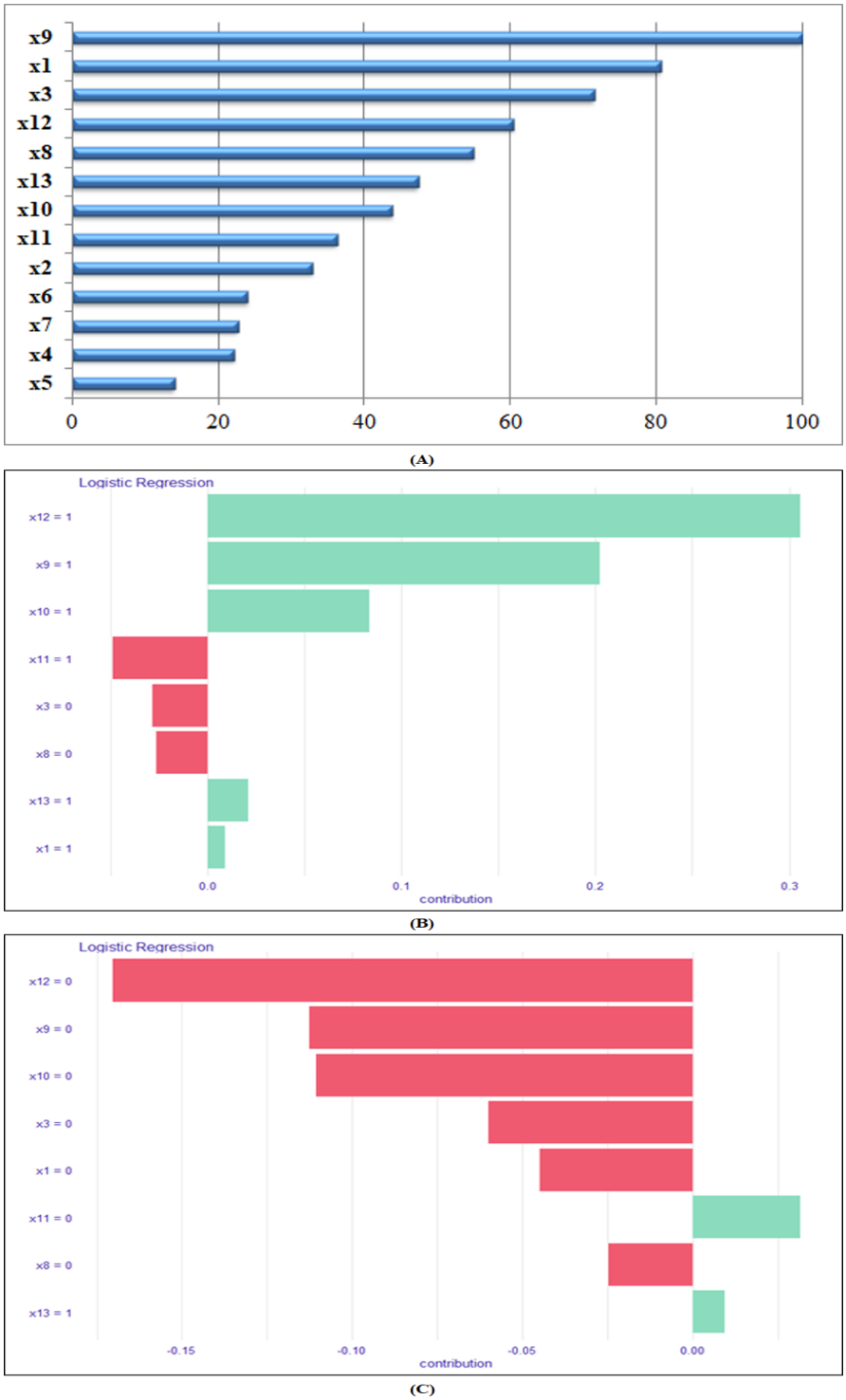

FIGURE 3: Variable importance for 1,000 samples. A) Overall model, B) Patient classified as "yes", C) Patient classified as "no".

For the overall model with 1,000 samples, the accuracy value was 0.882 in patients with the outcome variable yes (classified as yes) and 0.900 in patients with the outcome variable no (classified as no). The accuracy value was 0.962 in the sample patient with the outcome variable yes, and 0.961 in the sample patient with the outcome variable no. As a result, the accuracy values of the sample patients were found to be higher than the values of the overall model (Table 2).

STROKE DATASET

In Figure 4, the variable importance was examined in the stroke dataset by using InfoGain and LR methods. As a result of the variable importance test performed with the InfoGain method and according to clinical importance, the overall model consisted of age, hypertension, average glucose level, heart disease, marital status, gender, work type and BMI variables. As a result, 9 variables (8 independent, 1 dependent variable) were included in the study for the overall model and machine learning methods were applied (Figure 4A). The results of the variable importance test for each of the data sets including patients classified as stroke and non-stroke are given in Figure 4B and Figure 4C. These results differ from the variable importance results for the overall model presented in Figure 4A.

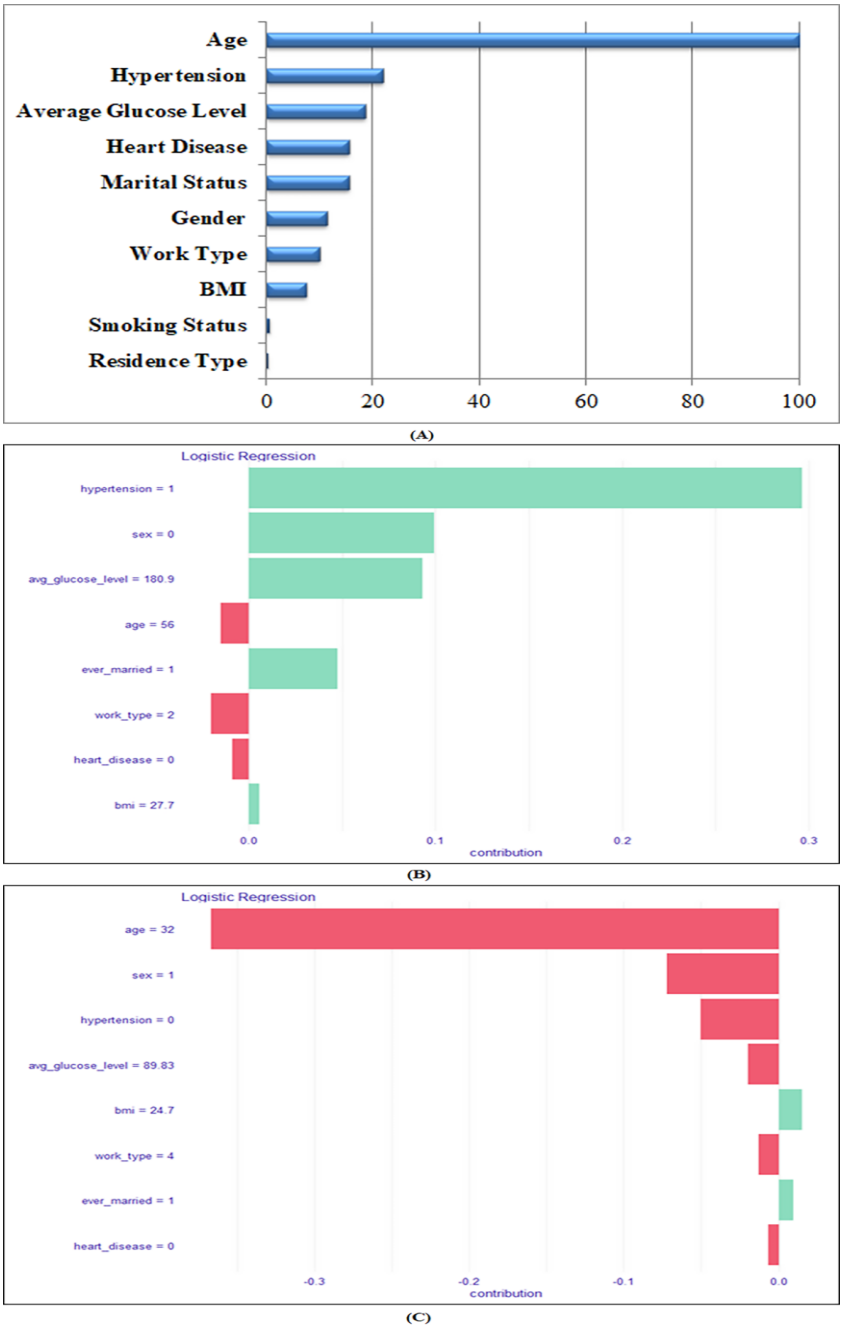

FIGURE 4: Variable importance for the stroke dataset. A) Overall model, B) Patient classified as "yes", C) Patient classified as "no". BMI: Body mass index.

In the stroke dataset of 826 patients, the accuracy value for the overall model was 0.733 in patients with stroke and 0.740 in patients without stroke. The accuracy value was 0.783 in the sample patient with stroke, and 0.868 in the sample patient without stroke. As a result, the accuracy values of the sample patients were found to be higher than the values of the overall model (Table 3).

TABLE 3: Performance measures for stroke dataset.

Datasets	Model		Accuracy	F-measure	MCC	ROC area
Stroke dataset	Model	No	0.740	0.736	0.472	0.805
		Yes	0.733	0.736		
		Overall	0.736	0.736		
	Patient 1	No	0.217			
		Yes	0.783			
	Patient 2	No	0.868			
		Yes	0.132			

MCC: Matthew's correlation coefficient; ROC: Receiver operating characteristic.

DISCUSSION

The conducted analyses led to the observation of disparities in both variable importance and performance outcomes between the overall model and the person-based results. It was emphasized that the importance of person-based variables and classification are crucial for improving the prediction results, and it was seen that the results were more accurate when models were created according to person-based characteristics while predicting.

In the literature, the LIME method is generally used for person-based variable importance. It is thought that the reasons for this are that it has more resources, ready-to-use codes and its codes are more understandable. The rationale for incorporating the SHAP method in our study resides in its versatile features, compatibility across various data types, and our objective of providing an illustrative exemplar to enrich the existing literature.

Pan et al. used LR, AdaBoost (Adaptive Boosting), Gradient Boosting Decision Tree (GBDT) and CatBoost methods for the overall prediction results in their study with 123 coronavirus disease-2019 patients and achieved the highest performance with the XGBoost method. In the same study, when the person-based prediction results with the LIME method were examined on the sample 4 patient data, the accuracy values of the 2 patients who were predicted to survive were found to be 0.980 and 0.580, respectively, and the accuracy values of the 2 patients who were predicted to die were found to be 0.850 and 0.680, respectively.[12] Hong et al., in their study on 1,585 heart patient data, evaluated the prediction results of the overall model with LR, SVM, Decision Tree, RF, XGBoost and Light Gradient Boosting (LGBoost) methods and achieved the best prediction results with the RF method. In the study, the person-based variable importance of two patients was also examined using the LIME method, and the predicted accuracy values were found to be 0.680 and 0.920, respectively.[13] Lin et al. evaluated the performance of the overall model with LR, RF, SVM, GBDT and Deep Neural Network (DNN) methods in their study on 372 patients to predict the risk of recurrence after endovascular treatment, and reported that the model with the best performance was obtained with the GBDT method. They reported that with the LIME method, they found the person-based prediction values of two patients to be 0.680 and 0.640.[14] In the study of An et al. to predict the risk of recurrence in 1,574 cancer patients, XGBoost, SVM, RF and LR methods were used for one and two-year prediction, and the best result was obtained with the XGBoost method. They found the person-based prediction results with the LIME method to be 0.810 and 0.720, respectively.[15] Alabi et al. used the Voting Ensemble, LGBoost, XGBoost, RF, and Extreme Random Trees methods in their study with 3,164 oropharyngeal cancer patients and reported that they achieved the best performance with the Voting Ensemble method. They also reported

that they evaluated the person-based variable importance result of a patient with the LIME method and found the accuracy value as 100.0.[16] Chan et al. used XGBoost, RF, and LR methods to predict 30-day, 90-day, and 1-year mortality in a study involving 6,994 critically ill ventilated patients, and reported that they achieved the highest accuracy value with the XGBoost method (0.858, 0.839, and 0.816, respectively). In the study, when the person-based variable importance for 2 patients was examined by the SHAP method, the accuracy values for mortality were found to be 0.230 and 0.710, respectively.[17] In a multicenter study by Yin et al., in which 1,012 patients with acute pancreatitis were included, Gradient Boosting, XGBoost, RF, Generalized Linear Models, DNN, and LR methods were used for the overall model, and the best performance measures were achieved with the XGBoost method. When the person-based prediction results for 6 patients were examined by the LIME method, the accuracy values were found to be 0.980, 0.980, 0.960, 0.840, 0.950, and 0.830, respectively.[18] Zheng et al. used XGBoost, RF, Neural Network, LR, Gaussian Naive Bayes, and k-nearest Neighbors methods in their study on 10,476 patients with ischemic stroke, and the highest accuracy value was obtained with XGBoost and RF methods (Accuracy: 0.840). In the same study, the person-based variable importance for four patients was examined with the LIME method, and the accuracy values for 2 patients who were predicted to have ischemic stroke were 0.980 and 0.650, respectively, and the accuracy values for 2 patients who were predicted to have no ischemic stroke were found to be 0.930 and 0.610, respectively.[19] In the scope of our study, the accuracy metrics derived from the overall model aligned comparably with the established literature and demonstrated commendable classification proficiency. Diverging from conventional methodologies, our inquiry delved into person-based prediction outcomes via the SHAP method, yielding superior predictive outcomes in contrast to both the overall model and extant literature findings.

CONCLUSION

Discrepancies exist between person-based variable importance and classification outcomes as compared to model-based results. This disparity underscores the necessity of acknowledging the heterogeneous interpersonal framework inherent in real-world scenarios. The adoption of person-based variables is postulated to confer enhanced realism and heightened applicability for clinical contexts. Contrasting the person-based models with the overall model outcomes, it becomes evident that the former yield more nuanced and elevated classification performance.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and/or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *İrem Kar;* ***Design:*** *İrem Kar, Batuhan Bakırarar;* ***Control/Supervision:*** *İrem Kar, Batuhan Bakırarar;* ***Data Collection and/or Processing:*** *İrem Kar, Batuhan Bakırarar;* ***Analysis and/or Interpretation:*** *İrem Kar, Batuhan Bakırarar;* ***Literature Review:*** *İrem Kar;* ***Writing the Article:*** *İrem Kar;* ***Critical Review:*** *İrem Kar, Batuhan Bakırarar.*

REFERENCES

1. Kantardzic M. Data Mining: Concepts, Models, Methods, and Algorithms. 2nd ed. New Jersey: John Wiley & Sons; 2011. [Crossref] [PubMed]
2. Brownlee J. Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python. 1st ed. Machine Learning Mastery; 2020.
3. Kuhn M, Johnson K. Applied Predictive Modeling. 1st ed. New York: Springer; 2013. [Crossref] [PMC]
4. Ng K, Sun J, Hu J, Wang F. Personalized predictive modeling and risk factor identification using patient similarity. AMIA Jt Summits Transl Sci Proc. 2015;2015:132-6. [PubMed] [PMC]
5. Visweswaran S, Ferreira A, Ribeiro GA, Oliveira AC, Cooper GF. Personalized modeling for prediction with decision-path models. PLoS One. 2015;10(6):e0131022. [Crossref] [PubMed] [PMC]
6. Liu T, Fan W, Wu C. A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. Artif Intell Med. 2019;101:101723. [Crossref] [PubMed]
7. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017. [Link]
8. Leisch F, Weingessel A, Hornik K. bindata: Generation of Artificial Binary Data. 2021. R package version 0.9-20. [Link]
9. Maksymiuk S, Gosiewska A, Biecek P. shapper: Wrapper of Python Library 'shap'. 2020. R package version 0.1.3. [Link]
10. Biecek P. "DALEX: explainers for complex predictive models in R." J Mach Learn Res. 2018;19(1):3245-9. [Link]
11. Hornik K, Buchta C, Zeileis A. Open-Source Machine Learning: R Meets Weka. Comput Stat. 2009;24(2):225-32. [Crossref]
12. Pan P, Li Y, Xiao Y, Han B, Su L, Su M, et al. Prognostic assessment of COVID-19 in the intensive care unit by machine learning methods: model development and validation. J Med Internet Res. 2020;22(11):e23128. [Crossref] [PubMed] [PMC]
13. Hong L, Xu H, Ge C, Tao H, Shen X, Song X, et al. Prediction of low cardiac output syndrome in patients following cardiac surgery using machine learning. Front Med (Lausanne). 2022;9:973147. [Crossref] [PubMed] [PMC]
14. Lin S, Zou Y, Hu J, Xiang L, Guo L, Lin X, et al. Development and assessment of machine learning models for predicting recurrence risk after endovascular treatment in patients with intracranial aneurysms. Neurosurg Rev. 2022;45(2):1521-31. [Crossref] [PubMed]
15. An C, Yang H, Yu X, Han ZY, Cheng Z, Liu F, et al. A machine learning model based on health records for predicting recurrence after microwave ablation of hepatocellular carcinoma. J Hepatocell Carcinoma. 2022;9:671-84. [Crossref] [PubMed] [PMC]
16. Alabi RO, Almangush A, Elmusrati M, Leivo I, Mäkitie AA. An interpretable machine learning prognostic system for risk stratification in oropharyngeal cancer. Int J Med Inform. 2022;168:104896. [Crossref] [PubMed]
17. Chan MC, Pai KC, Su SA, Wang MS, Wu CL, Chao WC. Explainable machine learning to predict long-term mortality in critically ill ventilated patients: a retrospective study in central Taiwan. BMC Med Inform Decis Mak. 2022;22(1):75. [Crossref] [PubMed] [PMC]
18. Yin M, Zhang R, Zhou Z, Liu L, Gao J, Xu W, et al. Automated machine learning for the early prediction of the severity of acute pancreatitis in hospitals. Front Cell Infect Microbiol. 2022;12:886935. [Crossref] [PubMed] [PMC]
19. Zheng Y, Guo Z, Zhang Y, Shang J, Yu L, Fu P, et al; Global Health Epidemiology Reference Group (GHERG). Rapid triage for ischemic stroke: a machine learning-driven approach in the context of predictive, preventive and personalised medicine. EPMA J. 2022;13(2):285-98. [Crossref] [PubMed] [PMC]