UNITECH SELECTED PAPERS

INTERNATIONAL SCIENTIFIC CONFERENCE
20-22 November 2025, GABROVO

EVALUATION OF MACHINE LEARNING CLASSIFIERS USING WEKA: FINDINGS FROM MULTI-DOMAIN DATA

Fatma BÜYÜKSARAÇOĞLU SAKALLI*, Özlem AYDIN FİDAN

Trakya University, Computer Engineering Department, Edirne, TÜRKİYE
**Corresponding author: fbuyuksaracoglu@trakya.edu.tr*

Abstract

The primary aim of this study is to systematically analyze the performance of classical classification algorithms that are commonly used in machine learning across different datasets and to identify the factors that influence algorithm selection. In the study, Naive Bayes, K-Nearest Neighbor (KNN), Decision Tree (J48), Random Forest (RF), Support Vector Machines (SVM/SMO), and Artificial Neural Networks (ANN) algorithms were implemented using the WEKA software, and four distinct datasets (TURKSTAT Happiness by Gender, Labor, Titanic, and Wine) were examined. The findings revealed that algorithmic performance varies depending on the nature of the dataset. For instance, the Random Forest model achieved the highest accuracy on the Wine Quality dataset, the SMO (SVM) performed best on the Titanic dataset, while Naive Bayes proved to be the most efficient method for the small-scale Labor dataset. Evaluation metrics such as accuracy rate, Kappa statistic, and error measures (MAE, RMSE) enabled a comparative assessment of the models. The main contribution of this study is to present a comprehensive understanding of how classical machine learning algorithms behave across different domain-specific data and to emphasize the importance of data-sensitive algorithm selection. The results are consistent with similar comparative studies in the literature and provide researchers with a methodological framework for model evaluation processes.

Keywords: Classification, Machine learning, WEKA

1. INTRODUCTION

Machine learning (ML) is a subfield of artificial intelligence that enables computers to make predictions or decisions by learning from data without being explicitly programmed. Through this technology, systems can utilize patterns extracted from past data to make intelligent predictions in response to new situations. Machine learning stands out with its ability to automatically extract knowledge from data and to employ this knowledge for predictive purposes. Going beyond traditional software engineering approaches, ML allows machines to train themselves using data, thereby gaining the capability to solve complex problems with minimal human intervention. [1], [2], [6]

WEKA is an open-source data mining software developed at the University of Waikato in New Zealand. With its user-friendly graphical interface, it provides a comprehensive environment for conducting machine learning experiments by integrating a wide range of algorithms within a single platform. [1], [2], [6], [7], [8], [19], [20]

WEKA supports a variety of functions, including data preprocessing, classification, regression, clustering, association rule mining, and data visualization. These algorithms can be applied directly to datasets by the user or integrated through the Java programming language. In addition, the system provides a flexible framework that facilitates the development of new machine learning algorithms.

In this study, the fundamental principles and application domains of various machine learning algorithms were examined, and the scenarios in which each algorithm operates most efficiently were analyzed. Furthermore, the performances of these algorithms were compared, and the conditions under which the most suitable solution can be achieved

doi: www.doi.org/10.70456/.....

with a particular algorithm were discussed. For this purpose, the WEKA software was employed to observe how these algorithms function in practice, and the processes of training and testing machine learning models were experienced. The datasets used in the study were obtained from TURKSTAT, csvbase.com, and the WEKA repository. To support the algorithmic analyses, Python modules compatible with WEKA were developed when necessary, ensuring more accurate and realistic data analysis results. [21], [22], [23]

2. ALGORITHMS

In this study, several widely used classification algorithms that have demonstrated success across different datasets—Naive Bayes, K-Nearest Neighbors (KNN), Decision Tree (J48), Random Forest (RF), Support Vector Machines (SVM), and Artificial Neural Networks (ANN)—were analyzed in detail.

Naive Bayes is a probability-based classification algorithm founded on Bayes' theorem. The term "naive" refers to the assumption that all features are mutually independent of each other. Although this assumption is often unrealistic, the algorithm still achieves high accuracy in many practical scenarios. Its major advantages include effectiveness even with small datasets, robustness in handling missing data, and its simplicity and computational efficiency. However, since it operates under the assumption of feature independence, it cannot model inter-variable relationships and may suffer from the zero-probability problem in certain cases. [18], [15], [10]

K-Nearest Neighbors (KNN) is an instance-based algorithm that classifies data points according to their proximity (distance) to other samples. A new data instance is assigned to the class most common among its k nearest neighbors in the training set. The primary advantages of KNN include its ease of implementation and its non-parametric nature, as it does not require an explicit model to be constructed. However, it tends to be computationally inefficient on large datasets, and its accuracy often decreases as the number of features increases.

Decision Tree (J48) is a hierarchical structure that makes decisions by splitting data into branches. The J48 algorithm implemented in WEKA is an improved version of the ID3 algorithm. At each node, the dataset is partitioned based on a specific attribute, and decisions are made at the leaf nodes. The main advantages of this method include its interpretability, speed, and ability to handle both categorical and numerical data, whereas its main limitation lies in the risk of overfitting when the tree depth becomes excessively large. [5], [11], [12]

Random Forest is an ensemble method that combines multiple decision trees to improve classification performance. Each tree is trained on a different subset of the dataset, and the final classification result is determined through majority voting among the trees. Its main advantages include reducing the risk of overfitting and typically achieving high accuracy. However, since it consists of multiple decision trees, it can be slower in prediction and computationally expensive when working with large datasets, as well as sensitive to noisy data. [5], [11], [12]

Support Vector Machines (SVM) aim to find the optimal hyperplane that best separates different classes within a dataset. Sequential Minimal Optimization (SMO) is a version of SVM implemented in WEKA that identifies the line or hyperplane maximizing the margin between classes. Among its advantages are its effectiveness in high-dimensional data and its strong classification capability when properly tuned. However, the long training time on large datasets and the significant influence of parameter selection (kernel type, C, and gamma values) constitute its main drawbacks. [5], [11], [12]

Artificial Neural Networks (ANN), specifically the Multilayer Perceptron model, are algorithms inspired by the structure of the human brain and composed

doi: www.doi.org/10.70456/.....

of multiple interconnected layers. They consist of an input layer, one or more hidden layers, and an output layer. In WEKA, this model is implemented as Multilayer Perceptron. Their ability to learn complex relationships and to perform effectively with structured data such as images and audio constitutes their primary advantages. However, the long training time and their tendency toward overfitting remain notable disadvantages. [5], [11], [12]

3. MODEL PERFORMANCE METRICS AND THEIR INTERPRETATIONS

The evaluation of models generated through classification algorithms involves the use of specific performance metrics to determine which classifier produces more accurate results. These metrics are generally based on a table structure known as the *confusion matrix*. In machine learning and statistical classification problems, the confusion matrix is a tabular representation developed to visualize the performance of a classifier.

a. **Correctly Classified Instances:** It represents the total number of correctly predicted instances and their proportion within the entire dataset. A higher value indicates greater accuracy of the model. [15], [9]
b. **Kappa Statistic**: It measures how much better the model's predictions are compared to random guessing. A value close to 1 indicates excellent performance, a value around 0 suggests performance similar to random prediction, and a negative value implies poor performance.
c. **Mean Absolute Error**: It represents the average of the absolute differences between the predicted and actual values. This metric indicates the average deviation of the model's predictions from the true values. A lower value is preferred, as it reflects higher prediction accuracy. [5], [11], [12]
d. **Root Mean Squared Error**: It is the square root of the mean of the squared errors. This metric penalizes larger deviations more heavily than smaller ones. A lower value is preferred, as it indicates better model performance.
e. **Relative Absolute Error**: It represents the ratio of the model's absolute error to that of a simple mean prediction model and is expressed as a percentage. A lower value is preferred, indicating that the model performs better than the baseline average prediction.
f. **Root Relative Squared Error**: It is the square root of the ratio between the model's squared error and the squared error of a simple mean prediction model. The result is expressed as a percentage, and a lower value indicates better model performance.
g. **Total Number of Instances:** It refers to the total number of data instances used for training and testing the model. [5], [11], [12]

4. EXPERIMENTAL RESULTS

This dataset, obtained from TURKSTAT, contains data from a study conducted between 2003 and 2024 that examines overall happiness levels by gender.

Table 1. *Performance Results of Classification Algorithms for the "Happiness by Gender" Dataset*

Algorithms	Correctly Classified Instances (%)	Kappa Statistic	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Squared Error (%)	Total Number of Instances
J48	98.64	0.957	0.018	0.1167	5.61	29.15	220
Naive Bayes	100	1	0.0021	0.0176	0.66	4.41	220
Random Forest	100	1	0.0026	0.0198	0.80	4.96	220
KNN	100	1	0.0050	0.0050	1.55	1.25	220
SMO (SVM)	100	1	0	0	0	0	220
ANN	100	1	0.0042	0.0051	1.30	1.27	220

In the classification performed using the SMO algorithm, all 220 instances were correctly classified. The Kappa statistic was calculated as 1, while the mean absolute error, root mean squared error, and relative

doi: www.doi.org/10.70456/.....

error rates were all found to be zero. These results indicate that the model achieved a perfect fit to the dataset.

Table 2. *Performance Results of Classification Algorithms for the Labor Dataset*

Algorithms	Correctly Classified Instances (%)	Kappa Statistic	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Squared Error (%)	Total Number of Instances
J48	73.68	0.4415	0.3192	0.4669	69.77	97.79	57
Naive Bayes	89.47	0.7741	0.1042	0.2637	22.78	55.23	57
Random Forest	89.47	0.7635	0.2294	0.3161	50.16	66.21	57
KNN	82.46	0.6235	0.1876	0.4113	41.01	86.15	57
SMO (SVM)	89.47	0.7635	0.1053	0.3244	23.01	67.95	57
ANN	85.96	0.6919	0.1524	0.3370	33.32	70.58	57

In this dataset, the Naive Bayes algorithm achieved a classification accuracy of 89.47%. With a Kappa value of 0.77, a mean absolute error of 0.104, and a root mean squared error of 0.2637, it was observed that Naive Bayes performed as the most reliable and successful algorithm for this dataset.

The Titanic dataset contains information about passengers aboard the RMS Titanic, which sank in 1912. The objective is to predict whether a passenger survived or not (the Survived variable: 0 = did not survive, 1 = survived).

Dataset Characteristics:
- Total Observations: 891
- Target Variable: Survived

Table 3. *Performance Results of Classification Algorithms for the Titanic Dataset*

Algorithms	Correctly Classified Instances (%)	Kappa Statistic	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Squared Error (%)	Total Number of Instances
J48	72.50	0.0000	0.2875	0.3792	99.80	99.99	891
Naive Bayes	86.87	0.6909	0.1590	0.2660	55.19	70.14	891
Random Forest	90.57	0.7568	0.1399	0.2377	48.57	62.69	891
KNN	80.70	0.4978	0.1526	0.3119	52.97	82.25	891
SMO (SVM)	91.69	0.7899	0.2424	0.3070	84.14	80.98	891

In the Titanic dataset, the SMO algorithm achieved an accuracy rate of 91.69%, demonstrating one of the best performances among the tested models. The Kappa statistic was calculated as 0.79, indicating that the model performed considerably better than random guessing. Although the mean absolute error appeared relatively higher compared to other models, the overall accuracy and Kappa values suggest that SMO is a strong classifier for this dataset. Hence, the model can be considered an effective classifier for the Titanic dataset.

The Wine Quality dataset contains the chemical properties of various wine samples along with their quality ratings. The objective is to predict the quality class of a wine based on its physicochemical characteristics and identification information (the Quality variable typically ranges from 0 to 10).

Dataset Characteristics:
- Total Observations: 1,143
- Target Variable: Quality

Table 4. *Performance Results of Classification Algorithms for the Wine Dataset*

Algorithms	Correctly Classified Instances (%)	Kappa Statistic	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error (%)	Root Relative Squared Error (%)	Total Number of Instances
J48	62.99	0.3883	0.2644	0.4590	65.14	101.89	1143
Naive Bayes	61.07	0.3727	0.3049	0.4223	75.12	93.74	1143
Random Forest	72.18	0.5385	0.2650	0.3547	65.29	78.75	1143
KNN	65.09	0.4283	0.2333	0.4818	57.48	106.95	1143
SMO (SVM)	61.42	0.3209	0.3222	0.4169	79.36	92.55	1143
ANN	63.69	0.3896	0.2913	0.4057	71.75	90.08	1143

The Random Forest model achieved a successful classification with an accuracy rate of 72.18% over a total of 1,143 instances. The Kappa statistic of 0.5385 indicates that the model exhibited a clear agreement with the actual classes. The mean absolute error (0.265) and root mean squared error (0.3547) values were relatively low, suggesting that the model's prediction error

doi: www.doi.org/10.70456/.....

was limited. The relative absolute error (65.29%) and root relative squared error (78.75%) further support the reliability of the model. Overall, the Random Forest model produced superior and consistent results compared to other models applied to this dataset.

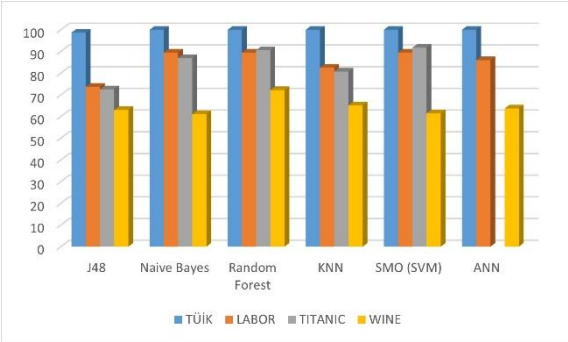

Fig. 1. *Correct Classification Performance (%) of Algorithms Across Different Datasets*

The graph above illustrates a comparison of machine learning algorithms applied to four different datasets in terms of their correct classification percentages. As observed:

• In the Happiness dataset, all algorithms achieved very high accuracy rates ranging between 98.64% and 100%.

• In the Titanic dataset, the most successful algorithms were SMO (SVM) and Random Forest.

• In the Wine Quality dataset, Random Forest outperformed the others, while the remaining algorithms demonstrated performance levels within the 60–65% range.

• In the Labor.arff dataset, Naive Bayes, Random Forest, and SMO exhibited similarly high performance levels.

CONCLUSION

In this study, six fundamental supervised machine learning algorithms — Naive Bayes (NB), K-Nearest Neighbors (KNN), Decision Tree (J48), Random Forest (RF), Support Vector Machines (SVM/SMO), and Artificial Neural Networks (ANN) — were comparatively analyzed on four distinct datasets (TURKSTAT Happiness by Gender, Labor, Titanic, and Wine Quality) using the WEKA platform. The experimental results demonstrate that the performance of classification algorithms is strongly influenced by the intrinsic characteristics of each dataset, such as dimensionality, class imbalance, noise level, and linear separability.

Among the evaluated models, the Random Forest algorithm exhibited the most consistent and reliable performance, particularly in heterogeneous datasets such as Wine Quality. This robustness can be attributed to its ensemble structure and resistance to overfitting. Support Vector Machines (SVM/SMO) achieved the highest accuracy in linearly or semi-linearly separable problems, as observed in the Titanic dataset. The Naive Bayes algorithm performed exceptionally well on small and clean datasets like Labor, distinguished by its simplicity and computational efficiency. The Decision Tree (J48) algorithm offered clear advantages in interpretability, highlighting its suitability for applications where model explainability is essential. In contrast, Artificial Neural Networks produced competitive results on nonlinear and complex data structures but required careful hyperparameter tuning to avoid overfitting. [3], [4], [7], [8], [13], [14], [16], [17], [19], [20]

Overall, this study revealed that there is no universally superior algorithm that performs best across all datasets. Algorithm selection should be made by considering the structure of the data, the characteristics of the target variable, and evaluation metrics such as accuracy, Kappa statistic, and error rates. The findings are consistent with the trends reported in large-scale comparative studies in the literature and further emphasize the importance of model selection and evaluation strategies in machine learning applications.

Future work may enhance prediction accuracy and generalization capability by incorporating deep learning models and

doi: www.doi.org/10.70456/.....

hybrid ensemble approaches. Additionally, employing techniques such as feature selection, dimensionality reduction, and hyperparameter optimization (e.g., grid search or Bayesian optimization) can improve the efficiency and scalability of models. Integrating explainable artificial intelligence (XAI) components into WEKA- or Python-based infrastructures can strengthen the balance between explainability and performance by enhancing the interpretability of model decision processes. [9], [10], [14]

In conclusion, this study not only demonstrates the comparative behavior of classical machine learning classifiers within the WEKA environment but also proposes a methodological framework for algorithm selection and evaluation based on dataset-specific characteristics. The results obtained provide valuable guidance for future academic research and applications in various domains where data-driven decision-making processes are extensively utilized.

REFERENCES

[1] Y. I. Alzoubi, A. E. Topcu, and A. E. Erkaya, "Machine learning-based text classification comparison: Turkish language context," Applied Sciences, vol. 13, no. 16, art. 9428, Aug. 2023.

[2] E. Borandağ, "LSRM: A new method for Turkish text classification," Applied Sciences, vol. 14, no. 23, art. 11143, Dec. 2024.

[3] S. G. Taşkın, "Detection of Turkish fake news in Twitter with machine learning methods," in Proc. Nat. Conf. Artif. Intell. (NCAI), 2021.

[4] Ö. Çelik and D. Kaptan, "Text classification by machine learning algorithms using a new text feature extraction method based on image processing," Turkish Journal of Engineering, vol. 9, no. 4, pp. 712–724, 2025.

[5] I. Sharma and B. Rathodiya, "Bias in machine learning algorithms," Turkish Journal of Computer and Mathematics Education, vol. 10, no. 2, pp. 1158–1161, Sep. 2019.

[6] F. Gürcan, "Multi-class classification of Turkish texts with machine learning algorithms," in Proc. 2nd Int. Symp. Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2018, pp. 260–263.

[7] J. Read, P. Reutemann, B. Pfahringer, and G. Holmes, "MEKA: A multi-label/multi-target extension to Weka," Journal of Machine Learning Research, vol. 17, pp. 1–5, 2016.

[8] J. H. Hayes, G. M. Kapfhammer, and A. A. Zaidman, "Weka meets TraceLab: Toward convenient classification," in in Proc. IEEE International Conference on Automated Software Engineering (ASE), 2014, pp. 900–903.

[9] C. Thornton, F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms," in in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2013, pp. 847–855.

[10] Q. Li, et al., "A survey on text classification: From traditional to deep learning," ACM Transactions on Intelligent Systems and Technology, vol. 13, no. 4, pp. 1–41, 2022.

[11] D. C. Asogwa, S. O. Anigbogu, I. E. Onyenwe, and F. A. Sani, "Text classification using hybrid machine learning algorithms on big data," arXiv:2103.16624, 2021.

[12] M. P. Basgalupp, R. C. Barros, G. L. Pappa, R. G. Mantovani, and A. C. P. L. F. de Carvalho, "An extensive experimental evaluation of automated machine learning methods for recommending classification algorithms," arXiv:2009.07430, 2020.

[13] A. C. Ateş, E. Bostancı, and M. S. Güzel, "Comparative performance of machine learning algorithms in cyberbullying detection: Using Turkish language preprocessing techniques," arXiv:2101.12718, 2021.

[14] K. Taha, R. Aljarah, and H. Alweshah, "A comprehensive survey of text classification techniques," Information Fusion, vol. 105, p. 101918, 2024.

[15] H. Altınbaş and A. Yıldız, "Comparative analysis of machine learning algorithms using WEKA software," Pamukkale University Journal of Engineering Sciences, vol. 29, no. 2, pp. 145–156, 2023.

[16] S. Yılmaz and B. Ermiş, "Comparison of classification performance using support vector machines and artificial neural networks," Fırat University Journal of Science, vol. 34, no. 1, pp. 22–33, 2022.

doi: www.doi.org/10.70456/.....

[17] H. Çetin and M. Güven, "Performance analysis of random forest and decision tree algorithms in machine learning," Gazi University Journal of Science, vol. 40, no. 4, pp. 650–662, 2022.
[18] E. Kara and Z. Küçükkaya, "Performance comparison of Naive Bayes and K-nearest neighbor algorithms on different datasets," Journal of Information Technologies, vol. 16, no. 1, pp. 33–42, 2023.
[19] T. Yaman and E. Aksoy, "Applications of WEKA in data mining: An experimental comparison," Uludağ University Journal of Engineering, vol. 29, no. 3, pp. 451–464, 2024.
[20] Z. Demirtaş, A. Turan, and F. Sert, "Performance analysis of machine learning algorithms in WEKA environment," Selçuk Technical Journal, vol. 21, no. 1, pp. 12–25, 2025.
[21] Türkiye İstatistik Kurumu (TURKSTAT), "Happiness by Gender Dataset," [Online]. Available: https://data.tuik.gov.tr. Accessed: May. 2025.
[22] csvbase.com, "Labor and Titanic Datasets," [Online]. Available: https://csvbase.com. Accessed: May. 2025.
[23] University of Waikato, "WEKA Machine Learning Repository," [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka/. Accessed: May. 2025.

doi: www.doi.org/10.70456/.....