Integrated Analysis of Student Surveys with Machine-Learning

Neli Arabadzhieva-Kalcheva[a], Maya Todorova[b], Evgeniya Rakitina-Qureshi[c], Ivo Rakitin[d], Hristo Nenov[e] and Dimitrichka Nikolaeva[f]
Faculty of Computing and Automation, Technical University of Varna, Varna, Bulgaria

Keywords: Decision Tree, Gaussian Naïve Bayes, Gradient Boosting, Histogram-Based Gradient Boosting, K-Nearest Neighbours, Logistic Regression, Machine Learning, Neural Networks, Random Forest, Support Vector Machine.

Abstract: This paper presents an integrated analysis of student survey data using machine learning methods. The main objective is to evaluate and compare the effectiveness of different algorithms in multi-class classification tasks related to student satisfaction. Ten widely used algorithms were applied: Decision Tree, Random Forest, Gradient Boosting, Histogram-based Gradient Boosting, Logistic Regression, Linear Support Vector Machine, Support Vector Machine with RBF kernel, k-Nearest Neighbours, Gaussian Naïve Bayes, and Multilayer Perceptron. The evaluation was conducted using 5-fold cross-validation and four standard performance metrics: accuracy, precision, recall, and F1-score. Experimental results show that ensemble tree-based methods, particularly Gradient Boosting, Random Forest, and Histogram-based Gradient Boosting, achieved the highest performance across all metrics.

1 INTRODUCTION

In modern higher education, student feedback plays a key role in evaluating and improving the learning process. Survey studies are not only a means of measuring satisfaction but also a foundation for strategic decision-making in education management (Atanasova, Filipova, Sulova, & Alexandrova, 2019; Gaftandzhieva, Doneva, & Bandeva, 2019).

With advances in artificial intelligence and machine learning, new techniques such as correlation analysis, decision trees, and Random Forest enable deeper knowledge extraction from survey data (Ivanov, 2020; Villegas-Ch, Román-Cañizares, & Palacios-Pacheco, 2020). This, in turn, facilitates the improvement of the content and organization of the learning process, personalization of teaching, and sustainable management of the educational environment (Nawaz, Sun, Shardlow, & Kontonatsios, 2022; Katragadda, Ravi, & Kumar, 2020).

Recent research has emphasized the analysis of both textual and quantitative data from student surveys using classification, clustering, and sentiment analysis algorithms (Anderson, Dryden, & Variava, 2018; Kastrati, Dalipi, Imran, Nucci, & Vanni, 2021; Dervenis, Kanakis, & Fitsilis, 2024). These methods not only identify key factors but also help in the early detection of problems, increasing student motivation, and enhancing communication between instructors and learners (Abdi, Sedrakyan, Veldkamp, & van Hillegersberg, 2023).

[a] https://orcid.org/0000-0002-9277-2803
[b] https://orcid.org/0000-0002-0266-9723
[c] https://orcid.org/0000-0002-8612-0876
[d] https://orcid.org/0000-0001-7960-1295
[e] https://orcid.org/0000-0003-1813-8406
[f] https://orcid.org/0000-0002-2980-581X

Arabadzhieva-Kalcheva, N., Todorova, M., Rakitina-Qureshi, E., Rakitin, I., Nenov, H. and Nikolaeva, D.
Integrated Analysis of Student Surveys with Machine-Learning.
DOI: 10.5220/0014328300004848

In *Proceedings of the 2nd International Conference on Advances in Electrical, Electronics, Energy, and Computer Sciences (ICEEECS 2025)*, pages 266-270
ISBN: 978-989-758-783-2

2 INTEGRATED ANALYSIS

2.1 Machine Learning Algorithms

Decision Tree – A non-parametric model that recursively splits the data into homogeneous leaves; easy to interpret but prone to overfitting on complex datasets (Quinlan, 1993).

Random Forest – An ensemble of trees trained on bootstrap samples with random feature selection at each split; reduces variance and is generally more accurate and robust than a single tree (Breiman, 2001).

Gradient Boosting – Builds a sequence of weak learners where each model corrects the residuals of the previous one; offers strong predictive power but requires careful regularization (Friedman, 2001).

Histogram-based Gradient Boosting – A variant of GB that uses histogram approximation of features, improving training speed and memory efficiency, especially for large datasets (Ke, Meng, Finley, et al., 2017).

Logistic Regression – A linear probabilistic classifier based on the logistic function; interpretable and efficient, often used as a strong baseline (Hosmer, Lemeshow, & Sturdivant, 2013).

Linear Support Vector Machine – A linear SVM optimized for large-scale and high-dimensional datasets; performs well when classes are approximately linearly separable (Cortes & Vapnik, 1995).

Support Vector Machine with RBF kernel – A nonlinear SVM using the radial basis function kernel; in many applications, such as text classification, the RBF kernel outperforms linear and polynomial kernels (Cortes & Vapnik, 1995; Kalcheva, Karova, & Penev, 2020).

k-Nearest Neighbours – Classifies samples according to the majority class among the k closest neighbors; sensitive to distance metrics and scaling (Cover & Hart, 1967; Kalcheva, Todorova, & Penev, 2023).

Gaussian Naïve Bayes – A Naïve Bayes classifier assuming Gaussian distributions for continuous features; computationally efficient but less accurate when normality assumptions are violated (Zhang, 2004).

Multilayer Perceptron – A feed-forward neural network with one or more hidden layers and nonlinear activations; capable of capturing complex interactions but computationally demanding and sensitive to hyperparameter tuning (Haykin, 2009).

2.2 Methodology

The study was conducted using a publicly available dataset of student satisfaction surveys from Kaggle (Kaggle, 2025). The dataset contains multiple attributes describing the demographic profile of students, their perceptions of teaching quality, the learning environment, and available academic resources, together with an overall satisfaction rating.

Prior to analysis, the data underwent preprocessing, including removal of missing values and normalization of selected numerical attributes to ensure comparability across models.

For the comparative analysis, ten machine learning algorithms were selected: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Gradient Boosting (GB), Histogram-based Gradient Boosting (HGB), Linear Support Vector Machine (SVM_linear), Support Vector Machine with RBF kernel (SVM_rbf), k-Nearest Neighbours (KNN), Gaussian Naïve Bayes (GNB), and Multilayer Perceptron (MLP). The selection was motivated by their wide application in survey analysis, their ability to capture both linear and nonlinear relationships, and their robustness when handling categorical and numerical data (Petrova & Bozhikova, 2022).

Model performance was evaluated using 5-fold cross-validation and four standard metrics: accuracy, precision, recall, and F1-score. All experiments were implemented in Python using the scikit-learn library in a consistent computational environment to ensure reproducibility and fairness in comparison.

2.3 Analysis of Results

2.3.1 Accuracy

Figure 1 illustrates the accuracy scores obtained by the evaluated algorithms on the public dataset. The ensemble-based models achieved the highest results, with Gradient Boosting (0.978), Random Forest (0.976), and Histogram-based Gradient Boosting (0.972) slightly outperforming all other approaches. This confirms the strong generalization ability of tree-based ensemble methods. A single Decision Tree reached 0.941 accuracy, which, although relatively high, remains significantly lower than its ensemble counterparts, highlighting the benefit of aggregation in reducing variance.

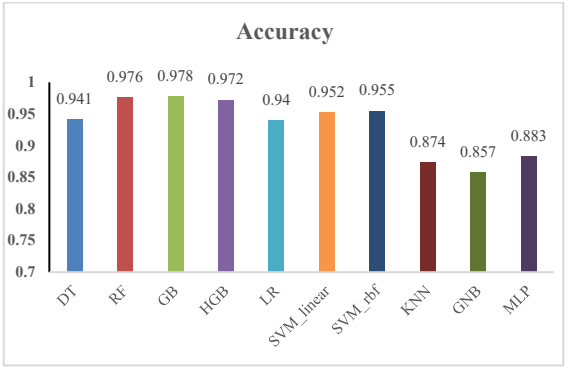

Figure 1: Accuracy of the evaluated algorithms.

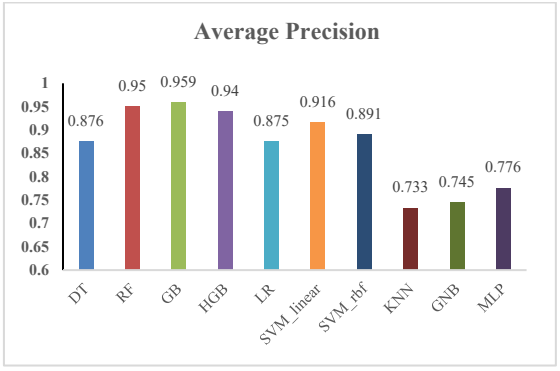

Figure 2: Average precision of the evaluated algorithms.

Support Vector Machines also performed well, with Linear Support Vector Machine (0.952) and Support Vector Machine with RBF kernel (0.955) exceeding the performance of Logistic Regression (0.940). These results suggest that margin-based separation and nonlinear kernels provide more robust classification boundaries compared to linear probability-based models.

The lowest accuracies were recorded by Gaussian Naïve Bayes (0.857), k-Nearest Neighbours (0.874), and Multilayer Perceptron (0.883), indicating their limited suitability for this dataset. Their weaker performance may be attributed to sensitivity to feature distribution assumptions (GNB), neighbourhood parameterization (KNN), and suboptimal hyperparameter tuning (MLP). Overall, the accuracy analysis confirms the superiority of ensemble classifiers, with SVMs as strong alternatives.

2.3.2 Average Precision

Figure 2 presents the average precision values for the ten evaluated algorithms on the public dataset. The ensemble-based methods achieved the highest precision, with Gradient Boosting (0.959), Random Forest (0.950), and Histogram-based Gradient Boosting (0.940) outperforming all other approaches. These results confirm the advantage of variance reduction and aggregation in tree-based ensembles, which provide more reliable decision boundaries compared to a single Decision Tree (0.876).

Among the linear models, Linear Support Vector Machine (0.916) showed higher precision than Logistic Regression (0.875), suggesting that margin-based separation is more effective than probability-based classification in this case. The Support Vector Machine with RBF kernel achieved 0.891, demonstrating the benefits of nonlinear kernels, although still below the ensemble methods.

The lowest-performing models were k-Nearest Neighbours (0.733), Gaussian Naïve Bayes (0.745), and Multilayer Perceptron (0.776), which appear less suited for this dataset. Their weaker performance may be attributed to sensitivity to feature scaling (KNN), violated distributional assumptions (GNB), and insufficient hyperparameter optimization (MLP). Overall, the results highlight the clear superiority of ensemble classifiers in terms of precision.

2.3.3 Average Recall

Figure 3 shows the average recall values obtained by the evaluated algorithms on the public dataset. The highest recall was achieved by Gradient Boosting (0.961), Random Forest (0.956), and Histogram-based Gradient Boosting (0.954), confirming again the superior ability of ensemble methods to correctly identify positive cases. The Support Vector Machine with RBF kernel also reached a competitive score of 0.933, making it a strong alternative where minimizing false negatives is critical.

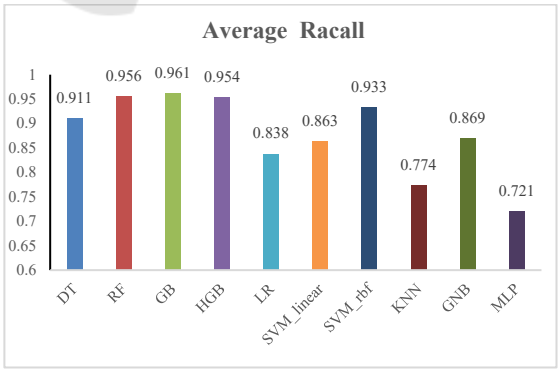

Figure 3: Average recall of the evaluated algorithms.

Decision Tree performed reasonably well with 0.911, but it was clearly outperformed by its ensemble variants. Among linear models, Logistic

Regression (0.838) and Linear Support Vector Machine (0.863) showed lower recall, indicating that they missed a larger proportion of positive instances compared to more advanced approaches.

The lowest recall values were observed for KNN (0.774), Gaussian Naïve Bayes (0.869), and MLP (0.721). These results highlight the limitations of these methods in handling the dataset, particularly the sensitivity of k-Nearest Neighbours to neighbourhood selection, the distributional assumptions of GNB, and the optimization challenges of MLP. Overall, the recall analysis demonstrates that ensemble tree-based classifiers are the most effective in capturing positive examples, with Support Vector Machine with RBF kernel serving as a strong nonlinear alternative.

2.3.4 Average F1-Score

Figure 4 presents the average F1-scores of the evaluated algorithms on the public dataset. Gradient Boosting achieved the highest score (0.960), closely followed by Random Forest (0.953) and Histogram-based Gradient Boosting (0.947). These results confirm that ensemble-based classifiers not only balance precision and recall effectively but also deliver the strongest overall performance.

The Support Vector Machine with RBF kernel reached 0.910, indicating a robust trade-off between precision and recall, while Linear Support Vector Machine (0.878) and Logistic Regression (0.849) performed moderately, reflecting their limitations in capturing nonlinear decision boundaries. The single Decision Tree obtained 0.892, which is significantly lower than its ensemble counterparts, but still stronger than the weakest models.

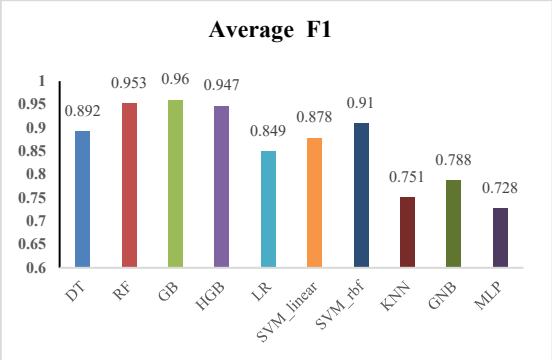

Figure 4: Average F1-score of the evaluated algorithms.

The lowest F1-scores were recorded by k-Nearest Neighbours (0.751), Gaussian Naïve Bayes (0.788), and Multilayer Perceptron (0.728), emphasizing their reduced effectiveness on this dataset. These findings underline that the combination of variance reduction and nonlinearity in ensemble methods ensures superior classification performance compared to both linear and simpler nonlinear approaches.

3 CONCLUSIONS

This study presented a comparative analysis of ten machine learning algorithms applied to a three-class student survey dataset. Evaluation was performed using 5-fold cross-validation and four key metrics: accuracy, precision, recall, and F1-score. Across all metrics, ensemble-based classifiers, particularly Gradient Boosting, Random Forest, and Histogram-based Gradient Boosting, consistently achieved the highest performance. These methods demonstrated strong generalization ability and robustness, confirming their superiority for this type of classification task.

Support Vector Machines with RBF also performed competitively, with the RBF kernel variant achieving high recall and balanced F1-scores, making it a strong alternative when minimizing false negatives is a priority. Linear models such as Logistic Regression and Linear Support Vector Machine showed moderate results and can be considered as stable baselines with lower computational cost.

By contrast, k-Nearest Neighbours, Gaussian Naïve Bayes, and Multilayer Perceptron recorded the lowest performance across the evaluation metrics, highlighting their limited suitability for this dataset without further preprocessing or hyperparameter optimization.

In summary, the findings confirm that ensemble tree-based methods are the most effective for multi-class classification of survey data, while SVMs offer a promising nonlinear alternative, depending on application-specific requirements.

Future research may extend this approach to larger and multi-institutional datasets or explore hybrid deep learning methods.

ACKNOWLEDGEMENTS

This paper is supported by the Scientific Project “Integrated Approach for Analysis of Student Surveys through Statistics and Machine Learning”, Technical University - Varna, 2025, financed by the Ministry of Education and Science.

REFERENCES

Abdi, A., Sedrakyan, G., Veldkamp, B., & van Hillegersberg, J. (2023). *A deep learning and language-knowledge-based model for analyzing student feedback in intelligent educational systems*. *Soft Computing*.

Anderson, E., Dryden, K., & Variava, K. (2018). *Applying machine learning to student feedback using clustering and sentiment analysis*. In *Proceedings of the Canadian Engineering Education Association (CEEA)*. Toronto, Canada.

Atanasova, Ts., Filipova, N., Sulova, S., & Alexandrova, Y. (2019). *Intelligent data analysis for students*. University of Economics – Varna.

Breiman, L. (2001). *Random forests*. *Machine Learning, 45*(1), 5–32.

Cover, T., & Hart, P. (1967). *Nearest neighbor pattern classification*. *IEEE Transactions on Information Theory, 13*(1), 21–27.

Cortes, C., & Vapnik, V. (1995). *Support-vector networks*. *Machine Learning, 20*(3), 273–297.

Dervenis, K., Kanakis, G., & Fitsilis, P. (2024). *Sentiment analysis of student feedback: A comparative study using lexicon-based and machine learning techniques*. *Information Processing & Management, 61*(3).

Friedman, J. H. (2001). *Greedy function approximation: A gradient boosting machine*. *Annals of Statistics, 29*(5), 1189–1232.

Gaftandzhieva, S., Doneva, R., & Bandeva, S. (2019). *Intelligent data analysis for improving learning outcomes*. In *Proceedings of ERIS 2019*. Plovdiv, Bulgaria.

Haykin, S. (2009). *Neural networks and learning machines* (3rd ed.). Pearson.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. (2013). *Applied logistic regression* (3rd ed.). Wiley.

Ivanov, G. (2020). *Modern methods for intelligent data analysis*. New Bulgarian University.

Kaggle. (2025). *Engineering student journey*. Kaggle Datasets. Retrieved August 25, 2025, from *https://www.kaggle.com/datasets/prasad22/student-satisfaction-survey*

Kalcheva, N., Karova, M., & Penev, I. (2020). *Comparison of the accuracy of SVM kernel functions in text classification*. In *Proceedings of the International Conference on Big Data, IoT and Artificial Intelligence (BIA 2020)* (pp. 141–145).

Kalcheva, N., Todorova, M., & Penev, I. (2023). *Study of the k-nearest neighbors method with various features for text classification in machine learning*. In *Proceedings of the International Conference on Artificial Intelligence (ICAI 2023)* (pp. 37–40).

Kastrati, Z., Dalipi, F., Imran, A. S., Nucci, K. P., & Vanni, M. A. (2021). *Sentiment analysis of student feedback using NLP and deep learning: A systematic mapping study*. *Applied Sciences, 11*(9), 3986.

Ke, G., Meng, Q., Finley, T., et al. (2017). *LightGBM: A highly efficient gradient boosting decision tree*. In *Advances in Neural Information Processing Systems (NeurIPS)* (pp. 3146–3154).

Katragadda, S., Ravi, V., & Kumar, P. (2020). *Analyzing student feedback using machine learning algorithms*. In *Proceedings of the IEEE International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*. Vellore, India.

Petrova, D., & Bozhikova, V. (2022). *Random forest and recurrent neural network for sentiment analysis on texts in Bulgarian language*. In *Proceedings of the International Conference on Biomedical Innovations and Applications (BIA 2022)* (pp. 66–69). Varna, Bulgaria.

Quinlan, J. R. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.

Villegas-Ch, W., Román-Cañizares, M., & Palacios-Pacheco, X. (2020). *Improving online education model with machine learning and data analytics in LMS*. *Applied Sciences, 10*(15), 5371.

Zhang, H. (2004). *The optimality of naive Bayes*. In *Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference (FLAIRS)* (pp. 562–567).