Machine Learning Model for Student Drop-out Prediction Based on the Student Engagement

Lucija Brezočnik[1], Giacomo Nalli[2], Renato De Leone[2], Sonia Val Blasco[3], Vili Podgorelec[1], and Sašo Karakatič[1]

Abstract Nowadays, the issue of student drop-out is not only addressed through the prism of pedagogy but also by technological practices. In this paper, we demonstrate how a student drop-out could be predicted through a student's performance using different machine learning techniques, i.e., supervised learning, unsupervised learning, and clustering. The results show that various types of student engagement are essential factors in predicting drop-out and the final ECTS points achievements.

Key words: Machine Learning, Student Drop-out, Academic Drop-out, Student Engagement, Student Drop-out Prediction

1 Introduction

The problem of student drop-out has been increasingly raising concern because of the complexity of the issue [1]. It is relevant not only for the professors who want to minimize the number of students that do not finish their studies but also to the tutors who work with students and, nevertheless, to the students themselves.

Many papers were written in the mentioned problem domain, but mainly from the pedagogical point of view [2–4]. For this research, one of the most meaningful results from their studies was the proven correlation between overall student engagement (behavioral, emotional, and cognitive) and academic achievement [5]. It is vital to emphasise, that some student engagements can be easily tracked, e.g., demographic and academic background, but behavior ones can be trickier. Usually,

Intelligent Systems Laboratory, Faculty of Electrical Engineering and Computer Science, University of Maribor, Slovenia
e-mail: lucija.brezocnik@um.si, saso.karakatic@um.si
· School of Science and Technology, University of Camerino, Italy
e-mail: giacomo.nalli@unicam.it, renato.deleone@unicam.it
· School of Engineering and Architecture, University of Zaragoza, Spain
e-mail: sonia@unizar.es

they are being tracked by faculties student ID cards, but the faculty or university must provide them. Hence, it is not a norm around EU Universities.

A few attempts at applying different Machine Learning (ML) techniques to the student drop-out prevention have been made [6–10]. Usually, the used dataset comprised a small number of features being collected from one faculty. Because of that, the main missions of this paper are:

- to define a set of features relevant to be collected for the later student drop-out identification;
- to define ML models able to predict student drop-out based on the student performance;
- to provide a list of the most informative features for student drop-out.

2 Machine Learning Algorithms

The prediction of students' drop-out based on their performance can be tackled through different approaches. For this reason, sections 2.1 and 2.2 present a brief overview of the most prominent learning approaches that were also used in our proposed method.

2.1 Supervised Learning

In Supervised Learning [11], the training set is made of P input vectors x with corresponding P output vectors y (labels). Therefore, data and their corresponding "correct" answers are available in this paradigm. The aim is to learn a rule linking the inputs to their corresponding output values (see Eq. 1).

$$f : x \in \mathbb{R}^N \rightarrow y \in \mathbb{R}^M \tag{1}$$

Moreover, the machine must then be able to predict the output for new input values. The two main problems that fall into this category are *Classification problems* and *Regression problems*. In Classification problems, the goal is to classify data into a finite number of categories. In general, unless some specific encoding is utilized, $M = 1$ (there is only a single output value) and $y_p \in \{1, \ldots, K\}$. A special case is a binary classification where it is customary to have $y_p \in \{0, 1\}$ or, even more often, $y_p \in \{-1, +1\}$.

In Regression problems, the output value is a real value (in this case $M = 1$), that is $y_p \in \mathbb{R}$. Again, the aim is to determine the function presented in Eq. 1 that, for given x, predicts the corresponding value y.

2.2 Unsupervised Learning

In Unsupervised Learning [12], the training set is made of only the P input vectors with no corresponding labels:

$$x^p = \begin{bmatrix} x_1^p \\ x_2^p \\ \vdots \\ x_N^p \end{bmatrix} \in \mathbb{R}^N, p = 1, \ldots, P \tag{2}$$

Two main problems are being tackled in this domain: *Clustering* and *Dimensionality reduction*. In Clustering problems [13, 14], the aim is to identify similarities among the elements in the training set. Objects must be organized in clusters so that objects in a particular cluster are as similar as possible and clusters themselves as different as possible. Therefore, the main aim is to maximize the similarity within clusters and the dissimilarity between clusters. Usually, clustering is conducted based on similarity measures (e.g., Euclidean distance) [15], which deals with finding a structure in a collection of unlabeled data. Most clustering algorithms are based on two popular techniques known as Hierarchical and Partitioned clustering [14].

In Dimensionality reduction problems [16], the goal is to transform the data from a high-dimensional space into a low-dimensional space. However, such low-dimensional representation must retain all the informative properties of the original data. The Principal Component Analysis (PCA) is the most utilized technique for this problem.

3 Student Engagement Data

Data is valuable but cannot be used if it is unrefined. Therefore, we first analyzed all contributed student engagement data from the universities in Italy, Slovenia, Spain, Cyprus, and Lithuania. After a quick review of the shared data, we realized that most universities are very sparse in their data collection. Not only that, data is usually too anonymized and thus not usable. Both reasons could be mainly attributed to the GDPR law. On the other side, sometimes the data is well collected, i.e., many parameters are being tracked per student but are stored in multiple separated systems which cannot be linked. Lastly, even if universities collect data, they are very reluctant to share or use this data for any research purposes.

In our research, we tackled the following challenges while trying to make data fusion:

- heterogeneous data,
- unstructured data,
- insufficient amount of collected parameters,

- collected data differ across different universities.

In the preprocessing phase, we defined the final set of features, which can be grouped into four main areas that play a vital role in a student's academic career. The main areas are:

- Demographic characteristics (e.g., age, gender, and distance from the university);
- Financial aspects (e.g., income, presence of a scholarship, and assigned free accommodation);
- Cognitive and academic aspects (e.g., educational background, the study progress in the university, and academic results);
- Engagement level in university life (e.g., use of the students' services and facilities).

The final dataset comprises the following list of features: Numeric ID, Student's gender; Name of the degree course; A binary value that determines if the student is active or not; First year of enrollment in the University; Actual year of enrollment (first, second or third); The status of the enrollment (if the student has to repeat the year); a binary value that indicates the student's room in campus; Number of meals per student at the University Canteen; A binary value that indicates if the student earned the scholarship; A binary value that determines if the student filled out the survey; ID that indicates the interest of the student; A binary value that indicates if a student attended less than 50% of the total lectures; A binary value that indicates if a student never attended lectures; A binary value that indicates if a student attended more than 50% of the total lectures; Number of acquired ECTS per student in the current year; and Number of acquired ECTS per student from the first year degree.

Those features have been collected for 412 students, 182 of them women and 230 men.

4 Proposed Model

Prior to defining the most suitable ML model, we first had to address the main challenge: *there is no direct measurement of student engagement*. The solution was found in the student's final grade feature, which we took as a proxy.

Prior to deciding on the best model, we conducted multiple experiments. Firstly, we performed regression analysis and treated the dependent variable (student's final ECTS grade) as a ratio. Eq. 3 presents calculation of the dependent variable denoted by Y.

$$Y = \frac{total\_credits}{number\_of\_study\_years \times 60} \tag{3}$$

With this approach, we predicted the student's final score, i.e., the final ECTS points, based on student engagement.

Similarly, we conducted classification analysis, but we treated the dependent variable as nominal this time. Accordingly, we performed the discretization step. The latter converts a continuous range of Y according to the equation 4.

$$student\_group = \begin{cases} 1; & \text{if } Y >= 1, \\ 0; & \text{if } Y < 1 \end{cases} \tag{4}$$

The $student\_group$ 1 stands for selected student passing all obligations, and 0 represents the opposite.

For Supervised learning, the following regression and classification algorithms were utilized:

- CART decision trees;
- Random Forest ensembles;
- different Gradient Boosting ensembles;
- Support Vector Machines.

In order to obtain the best clustering algorithm, an experiment was carried out in which we selected the algorithm that best determined the students' different levels of engagement. The following four clustering algorithms were tested:

- K-means;
- Agglomerative Cluster;
- Density-based spatial clustering of applications with noise (DBScan);
- Gaussian Mixture Models Clustering.

5 Results

All tested methods were implemented in the Python programming language. The experiment was run on a computer with a Windows operating system and an Intel Core i7 processor with 16 GB of RAM. The Regression, Classification, and Clustering results are presented in the following sections.

5.1 Regression Results

For regression, we utilized five regressors mentioned in section 4 and used the following performance evaluation metrics: Mean squared error (MSE), Mean absolute error (MAE), Mean absolute percentage error (MAPE), and Explained variance score (EVAR). Table 1 demonstrates summarized results.

The decision tree performed the worst out of all regressors, with the highest errors and the lowest explainer variance. In mean absolute percentage error, it is evident that support vector regressors (SVR) are slightly worse than others. In general, we

Table 1: Results of the regression analysis.

Regressor	**MSE**	**MAE**	**MAPE**	**EVAR**
Decision Tree Regressor	0.108	0.218	9.03×10^{13}	0.277
Random Forest Regressor	0.045	0.156	9.06×10^{13}	0.703
Gradient Boosting Regressor	0.041	0.157	9.98×10^{13}	0.726
Histogram-based Gradient Boosting Regression	0.045	0.167	9.37×10^{13}	0.701
SVR	0.045	0.159	7.07×10^{13}	0.706

can observe that Gradient boosting, Histogram based Gradient boosting, and Random Forest perform the best.

Figure 1 shows the features' importance of the Random Forest regression model. The feature *surveys_completed*, i.e., percentages of completed final subjects surveys, is the most important feature in the prediction made with Random Forest. This is followed by features *att_more*50 (more than 50% lecture attendance), *level_topics_interest* (interest in the topic of the lectures), and *graduation_year_*2018 (successfully graduating in the previous academic year). It is encouraging that gender and courses taken do not play an essential role in the predictions. Otherwise, the model would show the biasness of predictions to some gender or courses and, therefore, present a severe case of the unfairness of the decisions.

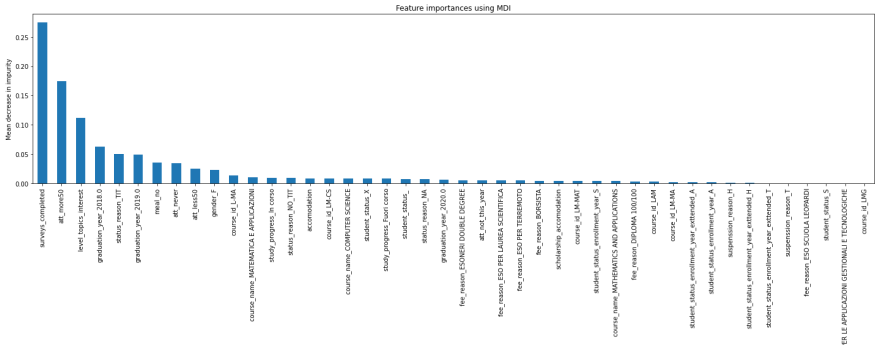

Fig. 1: Features' importance by Random Forest Regressor.

Similar results of the most important features are also demonstrated in Fig. 2 with the SHAP value horizontal scatter plot. The latter plots each student (one dot) either more to the left, i.e., when the instance influenced the prediction in predicting the lesser values, or right, i.e., when the instance influenced the prediction in predicting the higher values. The color of the dots/instances presents the feature values.

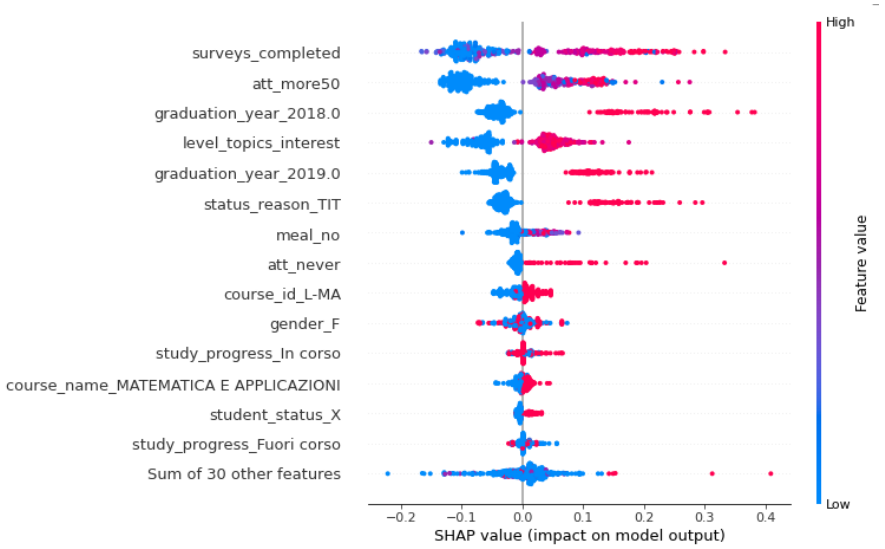

Fig. 2: SHAP results for Random Forest Regressor.

5.2 Classification Results

Seven classifiers presented in section 4 were evaluated using the following performance metrics: Classification Accuracy (ACC), F1-score, Precision, and Recall. Table 2 demonstrates summarized results.

Table 2: Results of the classification analysis.

Classifier	**ACC**	**F1-score**	**Precision**	**Recall**
DecisionTreeClassifier	0.897	0.794	0.818	0.771
RandomForestClassifier	0.904	0.787	0.923	0.686
GradientBoostingClassifier	0.919	0.825	0.929	0.743
HistGradientBoostingClassifier	0.934	0.862	0.933	0.800
SVC	0.882	0.714	0.952	0.571
SGDClassifier	0.875	0.730	0.821	0.657
LGBMClassifier	0.956	0.912	0.939	0.886

The highest classification accuracy obtained LGBM classifier (96%), followed by HistGradientBoostingClassifier (93%), GradientBoostingClassifier (92%), and RandomForestClassifier (90%). Similar results were also obtained for the F1-score.

In order to compare results with the regression analysis, we similarly first examined in detail the Random Forest Classifier results. Figure 3 shows the most important features, which are *surveys_completed*, *att_more*50, *level_topics_interest*,

*graduation_year_*2018, *meal_no*, and *status_reason_TIT*. It is clear that both algorithms defined practically the same informative features. The latter is also visible in Fig. 4.

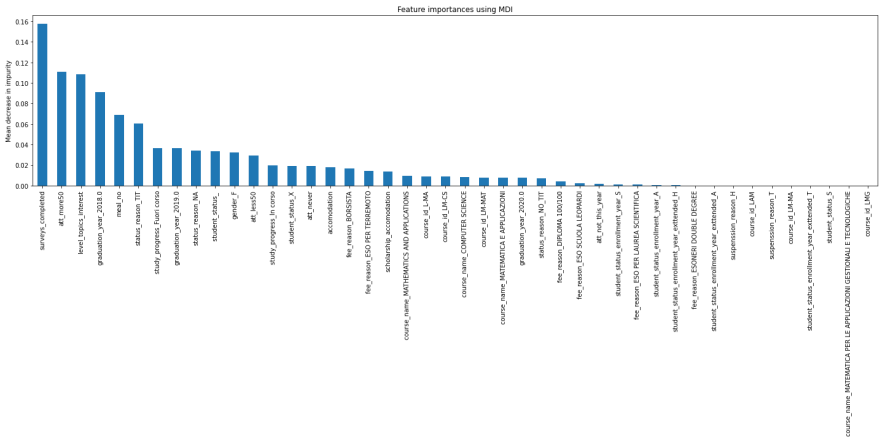

Fig. 3: Features importance by Random Forest Classifier.

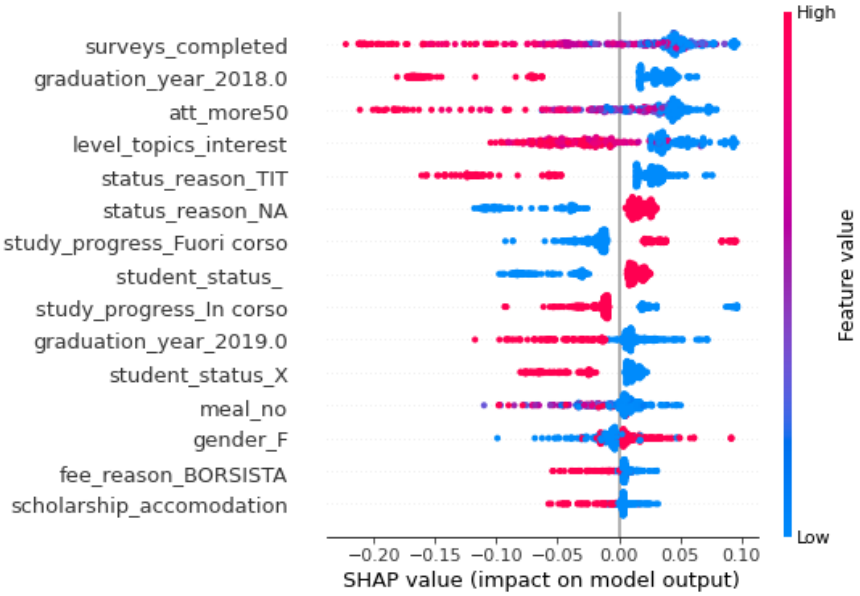

Fig. 4: SHAP results for Random Forest Classifier.

5.3 Clustering Results

After applying different ML algorithms, the next step was the cluster performance evaluation using Silhouette Analysis [17], which provides insight into how the clusters change depending on the algorithm and returns the natural trend of the grouped data. Silhouette Analysis was applied to each algorithm to interpret and validate the consistency within the data.

The range of the silhouette value $S(i)$ is between $[-1, 1]$.

- If $S(i)$ is close to 1, the sample is far away from the neighboring clusters. Hence, the sample is well-clustered and already assigned to a very appropriate cluster.
- If $S(i)$ is around 0, the sample is very close to the neighboring clusters and could be assigned to another closest cluster. Moreover, this indicates an overlapping cluster.
- If $S(i)$ is close to –1, the sample is assigned to the wrong cluster and placed somewhere between the clusters.

Therefore, we need the coefficients to be as high as possible in order to have good clusters. Comparison of different algorithms with calculated silhouette values demonstrated that Agglomerative Cluster was the best algorithm compared to the others and allowed the most optimal cluster realization, as Fig. 5 also shows.

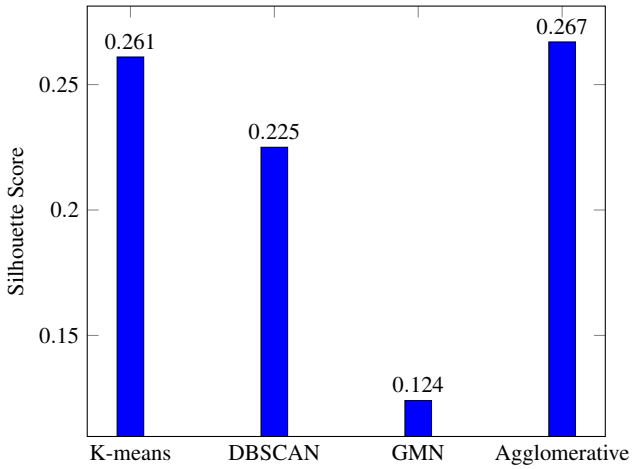

Fig. 5: Silhouette values of clustering algorithms.

Table 1 depicts differences between clusters of students' profiles in terms of engagement. In particular, Cluster 0 represents students with meager participation in University activities, i.e., low attendance at lectures, a high number of repeating students (57%), low value of ate meals at University, and a low level of scholarship won (4%). Even if Cluster 1 has no students with scholarships, it comprises

an average level of engagement, with an improvement in all categories. For example, attendance at lectures has a value of 142, which is also reflected in the number of repeating students (0%), and more ate meals at University. Cluster 2 represents highly engaged students at the University. These students have a high value for each feature, starting from the value of the high lecture attendance (349) that is related to the high average number of ate meals (178). However, such behavior may reflect the students' presence during the year at the University's campus, not specifically to attend lectures. The high level of attendance benefits students, highlighted by the number of scholarships obtained (100%) and the 0% of repeated students.

Table 3: Differences between clusters.

	Cluster 0	Cluster 1	Cluster 2
N° Students	176	172	68
N°Men	100 (57%)	88 (51%)	46 (68%)
N° Women	76 (43%)	84 (49%)	22 (32%)
N° Repeating student	101 (57%)	0 (0%)	0 (0%)
Attendance to lectures >50%	0	142	349
Scholarship	8 (4%)	0 (0%)	68 (100%)
Average Meals for students	5	25	178

There is also a notable difference between clusters based on the obtained ECTS points. The level of students' engagement reflects the average ECTS points achieved. Thus, Cluster 0 represents the students most at risk, not only for the low value related to ECTS points achieved (average of 67 points) but also for the number of students that did not get any credits for a year. To the latter group falls 25% of students, meaning that they found some difficulties in the study that affected their learning process. On the contrary, Cluster 1 and Cluster 2, characterized by good social interactions in the University, comprised a meager percentage of students that did not get credits (2% and 0% per Cluster 1 and 2, respectively) and higher ECTS points achieved, i.e., 79 and 85, respectively.

6 Conclusion

This paper shows that Machine Learning algorithms can be used to predict students' academic performance. The latter plays a vital role in the educational system because analyzing the students' status helps to improve their services and, consequently, academic performances, preventing drop-out and increasing motivation. Creating a robust model considering students' demographic, family, and social aspects, along with academic attributes and behaviors, is a very challenging task. Making predictions needs a suitable source of information that can be utilized in multiple ways to improve the quality of education and services. The prediction of students' engagement from academic data and personal habits, along with other features, is

a valuable application in defining strategies for improving students' services like tutoring, career guidance, didactics, and others. We must primarily join forces to achieve this goal by standardizing features collected among different institutions and countries.

Acknowledgements The authors acknowledge the financial support from the Slovenian Research Agency (Research Core Funding No. P2-0057), and European Commission (Project Code 2020-1-ES01-KA203-082090).

References

1. C. Truta, L. Parv, and I. Topala, "Academic engagement and intention to drop out: Levers for sustainability in higher education," *Sustainability*, vol. 10, no. 12, p. 4637, 2018.
2. N. Ruiz and M. Fandos, "The role of tutoring in higher education: improving the student's academic success and professional goals," *Revista Internacional de Organizaciones*, no. 12, pp. 89–100, 2014.
3. A. Hellas, P. Ihantola, A. Petersen, V. V. Ajanovski, M. Gutica, T. Hynninen, A. Knutas, J. Leinonen, C. Messom, and S. N. Liao, "Predicting academic performance: a systematic literature review," in *Proceedings companion of the 23rd annual ACM conference on innovation and technology in computer science education*, pp. 175–199, 2018.
4. M. d. C. Nicoletti, "Revisiting the tinto's theoretical dropout model.," *Higher Education Studies*, vol. 9, no. 3, pp. 52–64, 2019.
5. H. Lei, Y. Cui, and W. Zhou, "Relationships between student engagement and academic achievement: A meta-analysis," *Social Behavior and Personality: an international journal*, vol. 46, no. 3, pp. 517–528, 2018.
6. G. Nalli, D. Amendola, and S. Smith, "Artificial intelligence to improve learning outcomes through online collaborative activities," in *European Conference on e-Learning*, vol. 21, pp. 475–479, 2022.
7. S. Lee and J. Y. Chung, "The machine learning-based dropout early warning system for improving the performance of dropout prediction," *Applied Sciences*, vol. 9, no. 15, p. 3093, 2019.
8. C. Burgos, M. L. Campanario, D. de la Peña, J. A. Lara, D. Lizcano, and M. A. Martínez, "Data mining for modeling students' performance: A tutoring action plan to prevent academic dropout," *Computers & Electrical Engineering*, vol. 66, pp. 541–556, 2018.
9. N. Bedregal-Alpaca, V. Cornejo-Aparicio, J. Zárate-Valderrama, and P. Yanque-Churo, "Classification models for determining types of academic risk and predicting dropout in university students," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 1, 2020.
10. S. Oloruntoba and J. Akinode, "Student academic performance prediction using support vector machine," *International Journal of Engineering Sciences and Research Technology*, vol. 6, no. 12, pp. 588–597, 2017.
11. P. Cunningham, M. Cord, and S. J. Delany, "Supervised learning," in *Machine learning techniques for multimedia*, pp. 21–49, Springer, 2008.
12. H. B. Barlow, "Unsupervised learning," *Neural computation*, vol. 1, no. 3, pp. 295–311, 1989.
13. S. Äyrämö and T. Kärkkäinen, "Introduction to partitioning-based clustering methods with a robust example," *Reports of the Department of Mathematical Information Technology. Series C, Software engineering and computational intelligence*, no. 1/2006, 2006.
14. Y. Leung, J.-S. Zhang, and Z.-B. Xu, "Clustering by scale-space filtering," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 12, pp. 1396–1410, 2000.

15. A. K. Jain, R. P. W. Duin, and J. Mao, "Statistical pattern recognition: A review," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 1, pp. 4–37, 2000.
16. L. Brezočnik, I. Fister, and V. Podgorelec, "Swarm intelligence algorithms for feature selection: A review," *Applied Sciences*, vol. 8, no. 9, 2018.
17. M. Shutaywi and N. N. Kachouie, "Silhouette analysis for performance evaluation in machine learning with applications to clustering," *Entropy*, vol. 23, no. 6, p. 759, 2021.