1. **Introducere**- din fiecare articol parcurs ar trebui să extrageți scopul, abordarea, concluziile și probleme rămase nerezolvate. Puteți structura articolele pe categorii.

[1]

Scop:

This paper proposes a more flexible framework to predict the students’ academic performance. In this framework, the raw data is used directly to construct the prediction model without the feature engineering step.

Abordare:

The feature selection is instead based on model interpretability. The framework is applied to the open university learning analytics dataset (OULAD) with two different type of classifiers: random forest and artificial neural networks.

This study details three main contributions to improve the flexibility and automation of prediction models: 1) a framework that reduces the complexity of the approaches available in the literature by eliminating the feature engineering step is proposed; 2) model interpretation methods to automate the feature selection for different prediction scenarios is introduced; 3) the proposed framework is verified with two popular prediction models available in the literature: artificial neural networks and random forests. This work uses the manylearners’ records that are available in the Open University Learning Analytics Dataset (OULAD) (Kuzilek, Hlosta, and Zdrahal (2017)). These records contain students’ demographics and their interactions with a virtual learning environment.

In this work, the Open UniversityLearning Analytics dataset (OULAD) (Kuzilek et al. (2017)) is selected as it contains a mixture of students’ interactions with the virtual learning environments and their demographics. Also, a sufficient number of research results were published using this dataset, making it an excellent choice for comparing our work with state of the art methods.

The target of predictions in this work is the students ’ academic performance. In OULAD, this target is expressed by four different categories: Distinction, Pass, Fail, and Withdrawn. The distribution of the 32,593 students based on the final result is the following: Distinction (3,024), Pass (12,361), Fail (7,052), and Withdrawn (10,156). Following the approach of (Waheed et al. (2020)), the given students’ academic performances are split into four different categories, where each category is a binary classification problem.

In (Zheng and Casari (2018)), feature engineering is described as “the act of extracting features from raw data and transforming them into formats that are suitable for the machine learning model”. This work ’s main contribution introduces eliminating the feature engineering steps contrary to the other researchers’ approaches. For example, in (Waheed et al. (2020) ), 54 features were engineered and fixed in the pre-processing stage. Our work only considers the features present in the raw data of OULAD. These features can be categorized into three types: students’ information and demographics, courses information, and interactions with the virtual learning environment, described in Table 1, Table 2, and Table 3 respectively. All the features sum up to a total of 32. Regarding demographic features, the work of (Al-Zawqari and Van dersteen (2022)) investigated the role of students ’ background information in the performance of prediction models. However, to draw a fair comparison between the flexible feature selection approach and other work in the literature, we are considering all the raw data.

Still, for simplicity, it is decided to divide the course timeline into four moments of pre diction distributed evenly over the course duration. The length of the courses varies between 234 and 269 days. So, each class has different moments of predictions. The first quarter contains all the student interactions until the last day of the quarter. These interactions alongside the static featuresare used to constructthe prediction model. The second quarter data adds the newlyrecorded interactions until the middleof the course to what we had in the first quarter.The same reasoningis applied for the third and the fourth quarters. As no one is interested in having prediction moments on the last day of the course, the end of the fourth quarter is shifted two weeks before the course’s last day. In other words, the fourth quarter has the recordedstudent’s interactions available from the first quarter, second quarter, third quarter, and the recorded interactions until 14 days before the last day of the course.

The main motivation for this work is the observation that con structing a fixed set of features to predict the students’ academic performance is not straightforward. In addition, feature engineering adds a series of pre-processing steps and prevents the framework from being adaptable based on the changes in students ’ records at different moments of prediction. This section rationalizes the reason for removing the feature engineering step in the proposed framework, followed by a description of two different types of prediction models used in the experiments, namely, artificial neural networks and random forests. The choice of these prediction models is based on two criteria: 1) their frequent use in the literature; 2) the fact that they originate from two distinct families of machine learning algorithms. Finally, the possibility of linking a flexible feature selection/reduction step to the proposed model is explained.

From CHPATER 4.1 ALMOST ALL

Concluzii:

Obtained results show that the feature engineering step can be abandoned without affecting the models’ prediction performance. The prediction results of the flexible feature selection framework either outperform or have a difference of less than 1% accuracy compared to other work in the literature that relies on a manual feature engineering step. Both random forest and artificial neural networks without feature engineering accomplish a high prediction accuracy for the case of students at risk of failing with 86% and 88% compared to all students with pass grades and students with distinction grades, respectively. Also, the prediction models have the highest accuracy rate of 93% in predicting drop-out students. Yet, the prediction models in the proposed framework and previous research work perform poorly in predicting high achieving students with maximum accuracy of 81%, a precision of 69%, and a recall of 57%.

Each model has been investigated against 16 prediction moments, four quarters for each of the four binary classification problems. The results obtained from the prediction models show three important conclusions: 1) feature engineering steps do not add significantly to the accuracy of the prediction models so that they can be eliminated;2) a randomforest classifier is as competent as a deep learning classifier for predicting students’ academic performance; 3) random forest or other tree-based classifiers have the advantage of explainability when compared to deep learning classifiers. The prediction models in the proposed framework are evaluated at the fourth quarter with an accuracy of 85.83%, 88.15%, 81.26%, and 92.91%, for Pass-Fail, Distinction-Fail, Distinction-Pass, and Withdrawn-Pass, respectively. These results either outperform or have a difference of less than 1% compared with the state of the art accuracy,which relies on a manual feature engineering step. Our study also reveals that both the proposed framework and the available literature perform poorly in classifying students with distinction grades against students with pass grades.

Probleme ramase nerezolvate:

Future work will focus on solving this problem in addition to developing an algorithm/framework that can deal with a small dataset. Moreover, the students ’ demographic information helps predict their performance, but their misuse -e.g., using them for profiling-can lead to unexpected negative consequences. Therefore, future work will additionally focus on avoiding this demographic information as it might have a historical bias. Furthermore, an investigation of the proposed framework in different educational set-ups is needed, e.g., using this framework with a dataset generated from a physical classroom.

[2]

Scop:

In the present paper, five single supervised machine learning techniques have been used, including Decision Tree, Naïve Bayes, k-Nearest-Neighbor, Support Vector Machine, and Logistic Regression. To analyze the effect of an imbalanced dataset, the performance of these algorithms has been checked with and without various resampling methods such as Synthetic Minority Oversampling Technique (SMOTE), Borderline SMOTE, SVM-SMOTE, and Adaptive Synthetic (ADASYN). The Random hold-out method and GridSearchCV were used as model validation techniques and hyper-parameter tuning respectively.

Thus, it is quite important to predict low-performing students at an early stage with higher accuracy, along with the important factors that may affect their performance. To achieve this goal, the present study has three important research objectives: (i) to identify the influential features by using a filter-based feature selection technique. (ii) to identify the best performing classifier by comparing various singlesupervised machine learning techniques, viz., decision trees, Naïve Bayes, k-Nearest Neighbor, Logistic Regression, and Support Vector Machine with various resampling techniques such as random oversampling, SMOTE, Borderline SMOTE, SVM-SMOTE, and ADASYN. (iii) to enhance the prediction rate of the students at-risk by using an ensemble model that integrates the most suitable data mining technique.

Although there are several studies to predict the students‟ academic performance, the study which considers all categories of variables, i.e., background, academic, social, and psychological, and predicts students at-risk at an early stage with adequate accuracy is lacking. Also, a single classifierbased prediction is not suitable from one perspective to another. Moreover, a classifier giving the highest prediction accuracy for a particular dataset may not be valid for a different dataset. Thus, the aim of the present study is to identify low performers at an early stage with a higher prediction rate by using a scalable approach.

Abordare:

To make the data versatile, it is collected from the two different engineering colleges situated in different regions (the north and south of India). In the present paper, the sample size comprises 550 engineering students from two different engineering colleges in India, i.e., Bipin Tripathi Kumaon Institute of Technology, Dwarahat, Uttarakhand, and Cochin University of Science & Technology, Trivandrum, Kerala. The dataset includes information regarding background, past academic, social, and psychological factors with 30 different attributes, of which three attributes (roll-number, name, and branch) are used for identification purposes only and do not play any role in the prediction of low-performers. So, only 27 attributes were used for the present work, with first semester GPA as the output variable. For these attributes, data was collected online with the help of a multiple choice questionnaire created via outsourced technology, i.e., Google Form. As the aim of the paper is to identify the students having the highest risk of dropping out of college, the information about the output attribute for the dataset is divided only into two categories, i.e., low performers and high performers, based on the first-semester grade point of the students.

Feature selection: Feature selection is an important part of the students‟ performance prediction model for two main reasons:  The main purpose of the prediction of students‟ academic performance is to provide timely support to the low-performing students in the area where they are lacking. Only after identifying the attributes that have a significant impact on the output variable, i.e., students‟ academic performance, suitable corrective measures may be taken to provide support to the lowperforming students.  With the help of feature selection, irrelevant attributes may be removed from the data without losing reliability in classification. Thus, the dimensionality reduction raises the processing speed, and hence the classifier can learn faster. There are three main feature selection techniques: manual selection based on pedagogical theories or expert experience; filter-based selection; and wrapper feature selection [19]. In the present study, as all the attributes were categorical, a filterbased feature selection technique, namely “chi-square”, was used by which p-values were calculated for each attribute [8]. The attributes having a p-value of less than 0.01 show a highly significant correlation with the student's grades.

There are different types of classification machine learning models that may be used to predict the students‟ academic performance. In the present study, five single supervised machine learning models have been applied, including Decision Tree [25], Naïve Bayes [9, 26], k-Nearest-Neighbor [27], Support Vector Machine [28], and Logistic Regression [29]. To achieve the best performance of these machine learning models, the passing parameters for these models were set with the help of an algorithm called "GridSearchCV" which gives the best combination of passing parameters [30]. These combinations of passing parameters are listed in Table I.

Concluzii:

From this table, it is depicted that after applying the feature selection technique, the following 11 features are selected as influential features that affect students‟ academic performance: percentage in 10th standard, percentage in 12th standard, confidence, mathematics % in 12th standard, punctuality, curiosity, medium/language of previous study, category, father‟s highest qualification, mother‟s highest qualification, and mental stress. After selecting the most influential attributes, Decision Tree, Naïve Bayes, k-Nearest-Neighbor, Support Vector Machine, and Logistic Regression algorithms have been applied to the dataset, which contains only the 11 selected most influential attributes. The results obtained for accuracy, precision, recall, and f1-score of these algorithms are represented in Table III.

From the present work, it may be concluded that students‟ past academic performance (10th standard %, 12th standard %, and Math‟s % in the 12th standard), their background (category, parents‟ qualification, and medium of the previous study), and their psychological features (mental stress, confidence, curiosity, and punctuality) were the relevant attributes. Thus, to increase the academic performance of the students, these factors may be considered as the focus points. In the present study, all the used classifiers were able to predict students‟ outcomes with reasonable accuracy of more than 80%. Among all the used classifiers, Logistic Regression was the best performing algorithm with a balanced as well as an imbalanced dataset. Further, the accuracy and prediction rate for identifying low performers as well as for high performers were improved when the Logistic Regression was applied to the balanced dataset. The prediction accuracy was further enhanced with the use of an ensemble classifier in which three Logistic Regression classifiers (because of its highest performance) were integrated with the help of bootstrap aggregation. The proposed integrated model has achieved the highest accuracy of 95.45% and the highest precision and recall for low performers with the balanced dataset formulated with the help of the resampling technique SMOTE. It should be noted that with different datasets, the different classifiers may give the highest prediction accuracy, and hence there is a need for the methodology to be scalable for every situation. Thus, the main advantage of the present approach is its scalability for different datasets.

The results of the present study indicated that Logistic Regression is the best performing classifier with every balanced dataset generated using all of the four resampling techniques and also achieved the highest accuracy of 94.54% with SMOTE. Furthermore, to improve the prediction results and to make the model scalable, the most suitable classifier was integrated with the help of bagging, and a well-accepted accuracy of 95.45% was achieved.

Probleme ramase nerezolvate:

Further, this study may also be applied to the different domains of data mining and machine learning applications for enhancing prediction accuracy. The limitation of the present study is that the examined dataset has a small sample size and slightly imbalanced data, so in the future, the proposed methodology should be used with large sample sizes and highly imbalanced data for the prediction of students‟ academic performance.

[3]

Scop:

The objective of this paper is to propose an enhanced predictive model for students’ performance prediction. Selecting the most important features is a crucial indicator for the academic institutions to make an appropriate intervention to help students with poor performance and the top influencing features were selected in feature selection step besides the dimensionality reduction and build an efficient predictive model. DB-Scan clustering technique is applied to enhance the proposed predictive model performance in the preprocessing step. Various classification techniques are used such as Decision Tree, Logistic regression, Naive Bayes, Random Forest, and Multilayer Perceptron. Moreover ensemble method is used to solve the trade-off between the bias and the variance and there are two proposed ensemble methods through the experiments to be compared.

Therefore, the objective of this paper is to propose an enhanced predictive model for accurate prediction of students’ performance. Various machine learning techniques were experimented. For enhancing the predictive model, we applied DB-Scan clustering technique and feature selection approach. Experimental results proved the effectiveness of the proposed model.

Abordare:

The purpose of this paper is to propose a predictive model for students’ performance prediction. This is achieved by exploring various classification techniques, besides the ensemble method that solves the trade-off between the bias and variance, to investigate which one would achieve the best performance. Moreover, DB-Scan was used in preprocessing for outlier detection and features selection was used to enhance the predictive model of students’ performance.

Pre-processing is an essential process for any data set. It includes data cleaning and transformation. We pre-processed the dataset through three steps. Firstly, data was converted from nominal to numerical values. Secondly, some features were reshaped to be within a certain range using standardization method. Finally, DB-Scan clustering methodology was applied for outlier detection, as it had a great efficiency in [16]. B. Feature Selection Feature selection aims to select the most important and influencing features in the dataset. Also, it is very important for dimensions’ reduction before implementing the prediction and classification methods. It works by selecting the best features that contribute most to the target variable based on univariate statistical tests. We used the SelectKBest technique [15]. SelectKBest technique selects the first k features with the highest score values based on the Chi-Square test, for comparing the actual and predicted results, as a score function [14] using equation (1). X2 = P(Oi–Ei)2/Ei (1) where Oi and Ei are the actual and expected values, respectively. C. Classification For classification, we designed two different Ensemble models. One consists of Multilayer Perceptron (MLP), Random Forest (RF) and Decision Tree (DT), the other consists of RF, Logistic regression (LR) and DT.

The dataset is collected from a Learning Management System (LMS) [24]. It contains 480 records of students’ data in various educational levels with 16 features. The features were categorized as follows: • Academic features: Section id, Semester, Educational stages, Viewing announcements, Grade levels, Topic, Discussion groups, Visited resources, Raised hand, and Student absence days. • Personal features: Gender, Parent responsible for student, Parent answering survey, and Parent school satisfaction. • Demographic features: Nationality and Place of birth. We pre-processed the dataset through two steps. Firstly, we converted the data from nominal to numerical values for the features: Section Id, Semester, Educational stages, Grade levels, Topic, Discussion groups, Gender, Parent responsible for student, Parent answering survey, Parent school satisfaction, Nationality, Place of birth, and Student absence days. Secondly, the following features: Viewing announcements, Discussion, Visited resources, and Raised hand, were reshaped to be within a certain range using standardization method. Fig. 2 shows a sample of the dataset’s features and instances.

FROM C EVALUATION MEASURE -> ALMOST EVERYTHING

Concluzii:

The proposed model is an ensemble classifier of Multilayer Perceptron, Decision Tree, and Random Forest classifiers. The proposed model achieves an accuracy of 83.16%.

In this paper, we proposed an enhanced predictive model for students’ performance to improve the prediction accuracy. We applied various machine learning techniques for predicting the students’ performance. Additionally, DB-Scan clustering algorithm and feature selection steps have been exploited, for choosing the significant features. Our first ensemble method has achieved an accuracy of 83.16%, 78.95%, and 65.26% using all the features, the top 10 influencing features, and the top 5 influencing features, respectively. The proposed predictive model outperformed previous work using the same dataset from the learning management system. Applying DB-Scan clustering technique as a preprocessing step has a great effect on enhancing the predictive model performance and the distribution of results as seen in the confusion matrix of each predictive model.

Probleme ramase nerezolvate:

For future work, we intend to apply the proposed predictive approach to various datasets, experiment different feature selection techniques, and implement alternatives for DB-Scan clustering technique.

[4]

Scop:

Abordare:

Concluzii:

Probleme ramase nerezolvate:

[5]

Scop:

Abordare:

Concluzii:

Probleme ramase nerezolvate:

[6]

Scop:

Abordare:

Concluzii:

Probleme ramase nerezolvate:

[7]

Scop:

Abordare:

Concluzii:

Probleme ramase nerezolvate:

[8]

Scop:

Abordare:

Concluzii:

Probleme ramase nerezolvate:

[9]

Scop:

Abordare:

Concluzii:

Probleme ramase nerezolvate:

[10]

Scop:

Abordare:

Concluzii:

Probleme ramase nerezolvate:

[11]

Scop:

Abordare:

Concluzii:

Probleme ramase nerezolvate: