An Early Warning Model for School Dropout: a Case Study in E-learning Class

Felipe Neves, Fernanda Campos, Victor Ströele, Mário Dantas

[1]Computer Science Postgraduate Program – Federal University of Juiz de Fora
36036-900– Juiz de Fora– MG– Brazil

{felipe.neves.braz,victor.stroele,mario.dantas}@ice.ufjf.br,

fernanda.campos@ufjf.edu.br

Abstract. *Dropping out of school is a real challenge for educational specialists. Considering distance education classes, we have to deal with a huge number of students' disengagement with social and economic consequences. In order to solve the early drop out problem, this paper proposes the use of an Early Warning System capable of predicting the disengagement of students along the class and notify teachers about this behavior, enabling them to intervene in an effective way and make student's success possible. In order to evaluate our proposal, we carried out a case study which showed the feasibility of the proposal and the use of its technologies. The results pointed out a significant increase of gain in accuracy along the course, reaching 93% of precision at the end.*

1. Introduction

According to [Bagheri and Movahed 2016], Education has changed from a knowledge-transfer model to an active collaborative self-directed model by disruptive influence of technology in today educational institutions. Internet, Learning and Social Media technologies have influenced many aspects of education, from teacher role to student engagement, from innovation to student assessment, from personalized and unique interaction to security and privacy concerns.

E-learning systems allow teachers and students to interact in a challenging way, and provide an exponentially increasing amount of educational data. Student's behavioural characteristics, related to the learner experience during the educational process, is an important feature to predict student's performance. In the context of smart classrooms, the greatest challenge is not only sent students' recommendations and academic topics but predict learning issues, sending alerts to teachers, administrators, students, and families.

Dropping out of school comes from a long-term process of disengagement from school and classes, and has profound social and economic consequences for students, their families, and their communities [Márquez-Vera et al. 2016]. Behavioural, cognitive and demographic factors may be associated with early school dropout. Being able to predict this behaviour early could improve students' performance, as well as minimize their failures and disengagement.

For all available educational data related to courses, classes, students, resources usage, and interactions, different data mining and machine learning techniques can be used to extract useful knowledge that helps to improve e-learning systems

DOI: 10.5753/cbie.sbie.2019.1441

[Kumari et al. 2018]. An alternative utilization of these data, aggregated with intelligent techniques, could be Early Warning Systems (EWS).

According to [Grasso and Singh 2011], EWS is any system that is designed to alert decision-makers of potential dangers and helps to reduce economic losses and mitigates the number of damage in a context. These systems have been enhanced with different methodologies for predicting potential dangers, aiming to empower people to take action when a disaster is about to happen.

An early prediction of students' performance is a recurrent problem in the educational system, especially in E-learning context. Several works have been done in this field, aiming to identify student's interaction with the e-learning system and to predict his/her performance [Kumari et al. 2018].

Educational Data Mining (EDM) and Machine Learning have techniques, such as Classification, Regression, Clustering, and Relationship Mining [Kumari et al. 2018], capable to predict factors that have influenced in students' dropout index. Higher educational institutions, mainly the online classes, can get the advantage of early prediction of student's performance. By using more than one technique, in a combined way, we could improve prediction accuracy as all of them are considered in data analysis. This approach is named *ensemble models*.

Ensemble methods provide classification accuracy by aggregating the prediction of multiple classifiers [Kumari et al. 2018]. These methods construct a set of base classifiers from training data and perform classification by taking the vote on the predictions made by each of them.

The use of these systems in Education context is of great relevance since they can perform a prior diagnosis of student's disengagement or early school dropout possibility and notify teachers of this event. Different techniques and metrics can be applied to reduce the processing time and increase the certainty of notifications. Aiming to solve the aforementioned problem, the main contributions of this paper are:

- A proposal for an autonomous ensemble predictive model to an EWS;
- Evaluation in Educational context, specifically the discovery of students with disengagement or early school dropout possibility in a specific class;
- A comparison of prediction models to identify students' performance and the possibility of class dropout.

Beyond this introduction, the paper is organized as follows: Section 2 presents related works that implement models to predict students' performance and dropout possibility. Section 3 describes the proposed model. Section 4 presents the development of a Case Study, followed by evaluation and results in Section 4.1. Finally, in Section 5 we summarize our contributions and present further research directions.

2. Related Works

Numerous studies have been done in e-learning Systems which focused on the students' behaviour. This section will discuss some related works that implement models to student's disengagement based on his/her learning activities.

In [Kumari et al. 2018], the authors proposed a model to evaluate the impact of student's learning behavioural features based on his/her academic performance. The performance analysis task is performed by using Classification as a data mining technique. Besides that four classifiers were used: ID3, Naive Bayes, K-Nearest Neighbour (KNN) and Support Vector Machine (SVM). For improving classifiers performance and the accuracy of the student's performance model, the authors used the ensemble methods Bagging, Boosting and Voting. The last one was used for improving the classification accuracy.

Other work that aims to predict students' dropout possibility through educational data was presented by [Barbosa et al. 2017]. The authors proposed a student dropout prediction strategy based on classification with the reject option paradigm. They designed a classifier with a reject option where a Feed-forward Neural Network with Random Weights was used as a base learner. The classifications were divided into three groups: the ones who will certainly get approval; the ones who will certainly drop out of school and the ones whom the authors are not sure about their future. The evaluation was pursued with the Rejection curve.

In the work developed by [Cerezo et al. 2019], the authors proposed an algorithm to discover students' self-regulated learning during an e-Learning course by using Process Mining Techniques. The authors applied a new algorithm in the educational domain called Inductive Miner on Moodle platform. The technique used was capable of discovering optimal models in terms of fitness for both Pass and Fail students, as well as models at a certain level of granularity [Cerezo et al. 2019].

[Márquez-Vera et al. 2016] present a methodology to predict student's dropout earlier, using rules to define the student dropout probability and considering time bands throughout the course to predict this probability. The authors point out that there is no need to wait until the end of the course to predict and make a decision to react and provide specific help to students that are presenting a disengagement and have a risk to drop out.

Our proposal, motivated by these related works and our research group previous results in Educational Recommender Systems and E-learning assistance [Pereira et al. 2018, Neves et al. 2019] tackles this perspective of prevention student dropout. We use six classic machine learning techniques in ensemble form, some of them are used by [Kumari et al. 2018, Márquez-Vera et al. 2016], but we evaluate the proposal from another perspective aiming to assure a higher level of certainty.

3. Proposed Dropout Prediction Model

The main goal of this paper is to present an Ensemble Model used as a core by an Early Warning System to predict student's dropout. The proposed model of this research is illustrated in Figure 1. The model can predict if a student presents a risk to drop out of class and the system can notify teachers about this possibility. The model is dependent on a data set in the class context; it can be used in different classes, requiring only a configuration compatible with the class model, such as the total number of tasks, the assigned values and so on.

The EWS is composed of six main layers:

(I) **Extraction Layer**: responsible to extract all student's information needed to compose his activity profile. In Educational context, information as student's actions and rating notes are important to understand the student's activity, as well as his/her performance in the class.

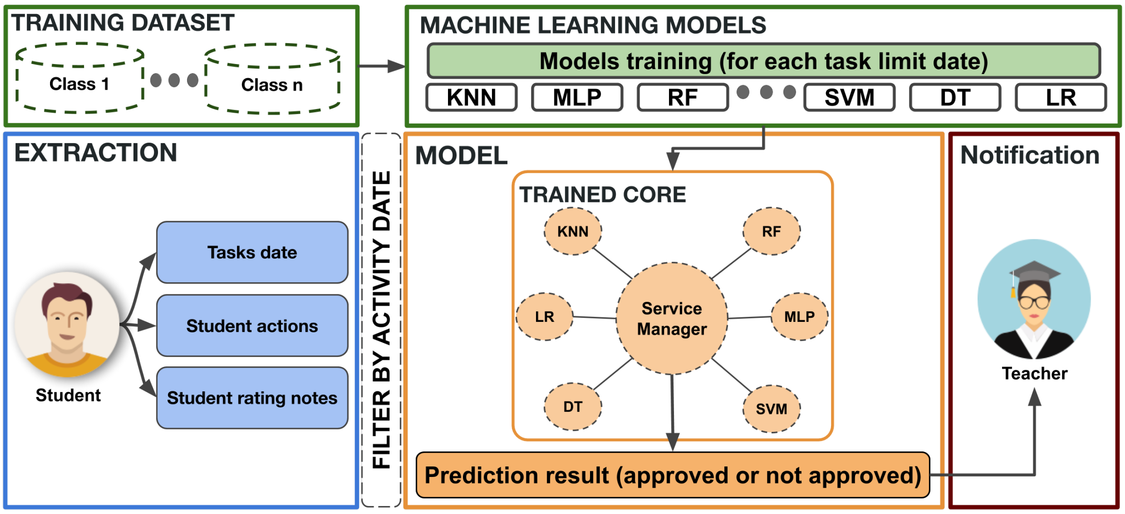

Figure 1. Proposed Earlier Warning System

(II) **Filter Layer**: As we want to predict the student's dropout probability, it's necessary to filter his/her performance throughout the course. This layer filters all students' activities by limit date, pre-established by the teacher.

(III) **Ensemble Model**: The core model combines different classic Machine Learning models, offering a more accurate result with a higher certainty. Each model is intended to be an autonomous service capable to deal with requests and predict the student's dropout probability. In the core model, we have defined six different machine learning techniques to compose the main autonomous services. To synchronize and combine them, we use a seventh service, that works as coordinator of the proposed model. All the adopted models in this proposal are supervised because the issue is a classification problem, which consists in indicating if a student has a possibility to drop out of class. The chosen ensemble method was the Voting Ensemble Method by averaging of positive predictions. This approach aims to minimize the difference between models prediction and maximize the model assertiveness. The averaging process takes into account only the prediction classes that are related to the notification trigger, which is the classification that intended to be notified. If for a set of student's characteristics the result presents that the student will not be approved, the system must notify.

(IV) **Notification layer**: a notification is sent to the teacher responsible for the class, where the teacher can intervene to prevent the student from dropping out of the class.

The other two layers (**Training Dataset Layer** and **Machine Learning Models Layer**) are responsible for the machine learning models training with past data from another previous class, which shares the same structure of evaluating student performance.

3.1. Data Description

We had permission of a professor in Federal University of Juiz de Fora to collect data produced by her Learning Management System (LMS) Moodle classes. Each class schedule includes various events, such as *"some content was published", "comment created", "course has seen", "sent submission" or "a file was submitted"* that are triggered by the students and their rating notes. The class data was collected and it was treated to offer a better environment to the experiment.

The treatment of data was made by setting the missing ones with zero value (0) and replace all non-English characters to English ones. After the cleaning process, we split the past data by date, generating a file for each limit task date, classifying the features in *active*, which represents the student approval and *inactive* otherwise. The prediction process considers student's activities in Moodle since their first access until the last task made. We consider that each time a student does not make an activity, higher is the chance to drop out of class.

Each generated file is responsible for training the proposed model with respective data for each task to be delivered. The whole notification process will be presented in Section 4.1. All data and code used are available and can be accessed at GitHub[1].

3.2. Machine Learning Models

Seeking to compose the proposed ensemble model with the most adherent models for this supervised classification problem, we have identified in literature those that are classically used. The chosen techniques are showing as follows:

Decision Tree: is an abstract structure that is characterized by a tree, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions [Han et al. 2011].

KNN: K-Nearest Neighbors is a classical and lazy learner model. They are classifiers based on learning by analogy, which measures the distance between a given test tuple with training tuples and compares if they are similar [Han et al. 2011].

SVM: is a classification algorithm that works with linear and nonlinear data [Han et al. 2011]. This model transforms the original data in a higher dimension, from where it can search and find a hyperplane for the linear optimal data using training tuples called support vectors.

Random Forest: is a classifier consisting of a collection of tree-structured classifiers which fits some classifying decision trees to various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control overfitting [Breiman 2001].

Logistic Regression: is a generalized linear approach that models the probability of some event to occur and is modelled as a linear function of a set of predictor variables [Han et al. 2011].

Multi-Layer Perceptron: This model is represented by a neural network that contains neurons to pass data through it. It can learn a nonlinear function approximator for either classification or regression. We used 3 hidden layers with 6 nodes each.

[1] https://github.com/FelipeNb/Early-Warning-System

4. Case Study

To understand the solution applicability in a real educational context, a case study was carried out, where we instantiate the proposed model and use real data from a Computer Science curriculum class with its past and current data. The Fundamentals of Information Systems class is offered to most Computer Science courses and its main goal is to prepare the students to recognize the importance of information systems in different organizations and identify different possibilities for their implementation.

Case Study is an empirical investigation, which is based on different sources of evidence, used when the object of the study is a contemporary phenomenon difficult to be studied in isolation [Wohlin et al. 2012, Runeson et al. 2012]. According to [Wohlin et al. 2012], the main advantage of a case study is the ease of planning it and also the characteristic of being more realistic, while at the disadvantage the authors present the difficulty of generalizing and interpreting the obtained results.

We adopted the methodology followed by [Márquez-Vera et al. 2016]. The process consists in to predict student's dropout possibility throughout the course or class. Thus, for each task delivered, a prediction can be made and different instances of the proposed model can be trained to act at different times throughout the class; thus making predictions of potential student dropouts.

Empirically defining parameters for Machine Learning models is not recommended, since those choices may not be the best, hampering the algorithm performance. Research works as the developed by [Denil et al. 2013] and [Hamers et al. 2003] aim to reduce the number of parameters in models. This problem is recurrent and is a challenge, mainly in ensemble models that have to combine parameters, given an increase in prediction potential of the entire model.

In this way, aiming to parametrize our dropout prediction model, we combined different model's configurations and generated a CSV file with a total of 36.000 combinations. The goal was to generate candidate configurations from possible input parameters and find the best ones to be used in the ensemble model. After that, we trained the data for each combination until getting high accuracy in each one. This process generated a total of 37 different configurations. These configurations are available and can be accessed[2]. We set each model with a configuration that was more representative to the ensemble model and more accurate as well.

For this Case Study, we selected two classes from the online Computer Science Teaching course, 2018 and 2019. The course uses the Learning Management System Moodle and teacher's and students' interactions are based on messages, forum, chat, wiki and choose the group members.

In Machine Learning solution it is necessary to define the training set, used in the proposed model, and the test set, used to evaluate model performance, i.e, test sample is responsible to show the results of predictions when the model is presented for unknown data. So, we used data from 2018 as our training set and the current class data (2019) was used as a test set.

To measure the proposed model quality, we chose classical statistical methods

[2] https://github.com/FelipeNb/Early-Warning-System/tree/master/modules/Prediction

RMSE, MAE to evaluate error and Precision, Recall, and F-measure measures to evaluate accuracy. We compared the predictive models adopted together and separately.

In Equations 1, 2 and 3, *TP* represents *true positive* classifications, which means that the prediction has a positive result and the classification was positive [Sokolova and Lapalme 2009]. *FP* represents *false positive* classifications, which means that the result was positive, but the correct classification was negative, and *FN* represents *false negative* for otherwise. RMSE and MAE evaluation utilize the available functions in the Sklearn framework. These methods of evaluation are more adherent to supervised learning, which is the type used in our solution.

$$Precision = \frac{TP}{TP+FP} \quad (1) \quad Recall = \frac{TP}{TP+FN} \quad (2) \quad F-measure = 2\cdot\frac{precision \cdot recall}{precision + recall} \quad (3)$$

4.1. Results and Discussion

Class dropout is a real challenge, and considering the distance education, we have to deal with student's disengagement since the beginning, as most of the time, they even make any task. With separated data and the models trained, we executed the experiment following the methodology presented in Section 4 . The complete process is represented in Figure 2. The student's data was collected to predict his/her performance in the class and the same data structure was maintained.

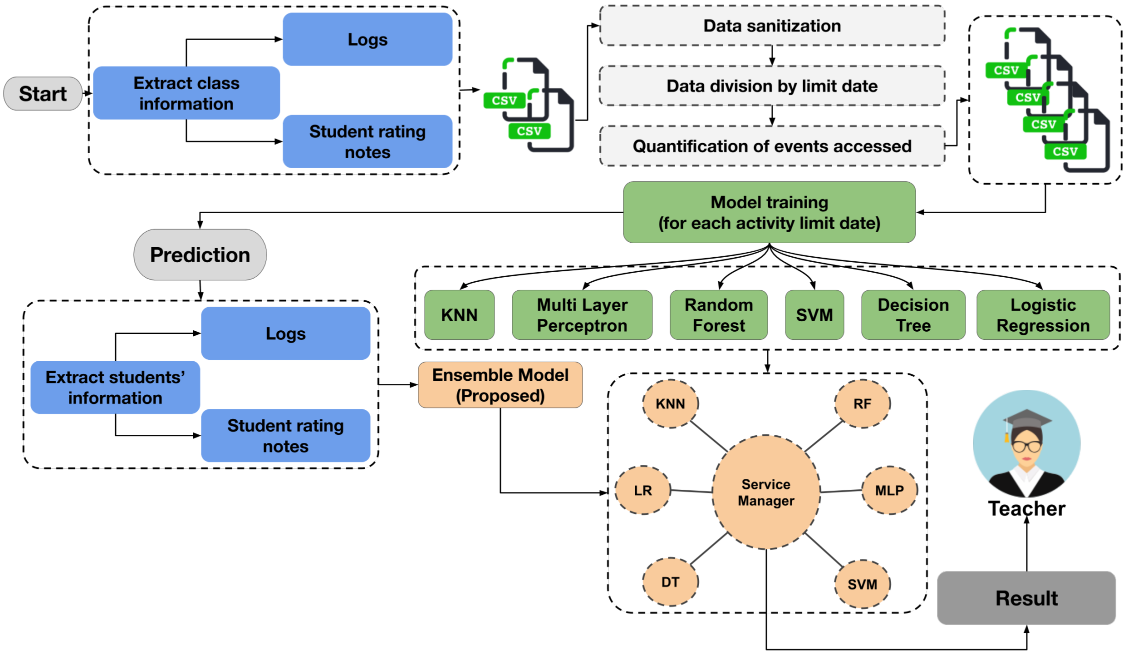

Figure 2. Complete notification process.

The 2019 class had 37 students, from 12 different cities. The main evaluation activities include participation in two forums, two individual tasks, and a group final task. The second task includes a visit to a school to identify its Information System features and the group activity proposes a School Information System component. The groups were composed of one, two or three members. The class evaluation process also includes a peer review of the final task, a presentation.

We ran the experiment and for each task delivered in the class, a prediction was done. When the results output the *inactive* class, a notification is triggered and it is sent to the teacher. We evaluated the obtained results using Equations 1, 2 and 3, and we compared them to visualize and understand the differences between models. The main accuracy results presented in Figure 3 show that individually, the models present a linear structure where the observed accuracy has not increased along time in ascending form, presenting variations. However, in our solution, which combines the models using the voting ensemble method, the accuracy increases gradually throughout time as the class goes on and students have more activities in their schedule.

Due to the limited space in this paper, we chose to highlight the measure that addresses the predictions' accuracy more completely. The other metrics can be viewed, individually, in the following link[3].

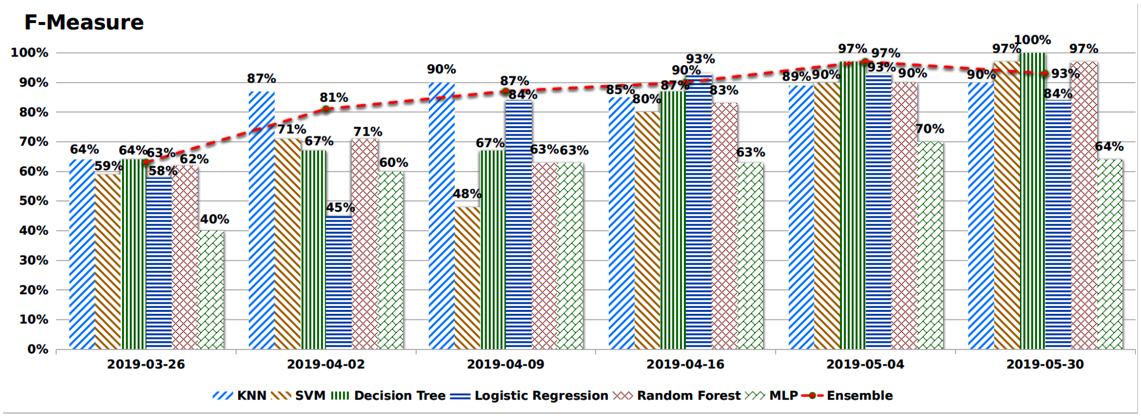

Figure 3. F-Measure metric.

Individually, some models present a better performance in specific metrics, such as Decision Tree model in F-Measure. However, the overall result shows a gain of performance and certainty in predictions through our ensemble model.

4.2. Observed Evidences

The 2019 class final results presented that, 29.72% of the students didn't make any task. 16.21% made one, two or three tasks. 8.10% had made 4 of 6 tasks and one of them failed. 21,62% made 5 of 6 tasks. Finally, 24.32% made all the tasks, but one of them failed. Considering the results we can say that at least 43.24% should have had special attention and motivation for not making the tasks and at least 5.40% could be approved finishing one or two more tasks.

As the results of the proposed model are based on student's activities at Moodle, we can infer that if the teacher received a notification from the system it would probably be helpful to avoid so many dropouts in the class. The first results enable us to observe that the system would have the following behaviour:

- All students that didn't make any task would be notified after all tasks deadline;
- 80% of the students that made 1, 2 or 3 tasks would be notified after all tasks deadline;

[3] https://github.com/FelipeNb/Early-Warning-System/tree/master/Graphs

- All students that made at least one task would be notified at least once;
- All approved students would not be notified by the system.

From Educational perspective, the Case Study demonstrated the feasibility of the proposal, the involved concepts and technologies, the use of machine learning techniques, the ensemble model definition, and its use in the context of an EWS system. It was observed that the ensemble model presents a good performance in the identification of students with high disengagement, which is dropout indicative. A detailed analysis of the results obtained for the current 2019 class shows that the system would be able to identify and alert the teacher about students with engagement problems. It was also observed that some students only sent the resolution of the tasks but did not actively participate in the class and in this case, the system would notify the teacher, even if the student did not effectively drop out the class.

5. Final Remarks

In this paper, we constructed an ensemble model to predict student's performance to avoid dropout to class. This approach can provide different advantages in the field of education, allowing students, know their performance, notifying them and allowing them to improve their performance in the future. And for teachers it allows them to understand their students to perceive their deficiencies and needs, their way of learning and making possible the improvement of their teaching didactic and pedagogical theory.

A huge amount of educational data has the potential to become new knowledge to improve all instances of e-Learning. Data Mining processes can explore that data and predict student's performance. Some limitations were observed by the obtained results, such as students that do not enter at least once in the platform, were never notified, as there were no events for these students.

In future works, we have the intention to solve these limitations by understanding student's behaviour beyond the events emission and making content recommendations that could prevent student's dropout. It may improve their capabilities by motivating them to engage in the class and help to get a better result at the end.

6. Acknowledgements

This work was carried out with the support of the Coordination of Improvement of Higher Education Personnel - Brazil (CAPES) - Financing Code 001, Federal University of Juiz de Fora (UFJF), CNPq and FAPEMIG.

References

Bagheri, M. and Movahed, S. H. (2016). The effect of the internet of things (iot) on education business model. In *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 435–441. IEEE.

Barbosa, A., Santos, E., and Pordeus, J. P. (2017). A machine learning approach to identify and prioritize college students at risk of dropping out. In *Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação-SBIE)*, volume 28, page 1497.

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

Cerezo, R., Bogarín, A., Esteban, M., and Romero, C. (2019). Process mining for self-regulated learning assessment in e-learning. *Journal of Computing in Higher Education*, pages 1–15.

Denil, M., Shakibi, B., Dinh, L., De Freitas, N., et al. (2013). Predicting parameters in deep learning. In *Advances in neural information processing systems*, pages 2148–2156.

Grasso, V. F. and Singh, A. (2011). Early warning systems: State-of-art analysis and future directions. *Draft report, UNEP*, 1.

Hamers, B., Suykens, J., Leemans, V., and De Moor, B. (2003). Ensemble learning of coupled parameterized kernel models. In *Supplementary Proc. of the International Conference on Artificial Neural Networks and International Conference on Neural Information Processing (ICANN/ICONIP)*, pages 130–133.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

Kumari, P., Jain, P. K., and Pamula, R. (2018). An efficient use of ensemble methods to predict students academic performance. In *2018 4th International Conference on Recent Advances in Information Technology (RAIT)*, pages 1–6. IEEE.

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., and Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, 33(1):107–124.

Neves, F., Ströele, V., and Campos, F. (2019). Information diffusion in social networks: a recommendation model in the educational context. In *Proceedings of the XV Brazilian Symposium on Information Systems*, page 25. ACM.

Pereira, C. K., Campos, F., Ströele, V., David, J. M. N., and Braga, R. (2018). Broad-rsi–educational recommender system using social networks interactions and linked data. *Journal of Internet Services and Applications*, 9(1):7.

Runeson, P., Host, M., Rainer, A., and Regnell, B. (2012). *Case study research in software engineering: Guidelines and examples*. John Wiley & Sons.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in software engineering*. Springer Science & Business Media.