[1]Rachid Ed-Daoudi
[2]Mourad Azhari
[3]Badia Ettaki
[4]Jamal Zerouaoui

Academic Performance Prediction in Virtual Environments Using Big Data and Machine Learning

Journal of Electrical Systems

Abstract: - The integration of technology in education has transformed how students learn and how educators manage the teaching process. This study focuses on predicting student academic performance on the Moodle platform using Big Data and Machine Learning techniques. By analyzing interaction Data from 16,113 students across 211 courses, a total of 83,455 records were processed to develop predictive models. The methodology, based on the CRISP-DM framework, involved Data extraction, variable definition, and model training with algorithms like Random Forest (RF), CatBoost, and XGBoost. To address class imbalance, undersampling and oversampling techniques were applied, creating balanced Datasets that improved model performance. Among the tested algorithms, RF demonstrated the highest precision at 93.2%, making it a reliable choice for predicting academic outcomes. The findings, visualized through Power BI, offer valuable insights into student engagement and performance, supporting informed decision-making in virtual learning environments. This research underscores the potential of Big Data and Machine Learning to enhance educational processes, paving the way for future studies to expand and refine predictive models for educational Data analytics.

Keywords: Academic performance prediction, Big Data, Machine Learning, Moodle platform, Educational Data analytics.

I. INTRODUCTION

The integration of technology into education has revolutionized how students access knowledge and how educators manage the teaching and learning process. In this context, online learning platforms such as Moodle have gained wide adoption in educational institutions globally [1]. Moodle has become one of the most used educational platforms, offering a virtual learning environment that promotes collaboration, communication, and student assessment [2]. These platforms facilitate student-teacher interaction, easy access to materials, and performance evaluation tools [3]. However, despite their advantages, predicting student academic performance within these platforms remains a challenge [4].

Analyzing large Datasets, known as Big Data, has emerged as a powerful tool for understanding and forecasting various phenomena. In education, Big Data has shown promising potential in predicting student performance. The application of Big Data techniques in education can help teachers identify hidden Data patterns and make informed decisions to improve students' academic outcomes [5]. With Moodle capturing a wealth of student interaction Data, it serves as a valuable resource for implementing Big Data methods to improve decision-making [6].

This article addresses the prediction of university students' academic performance on the Moodle platform through Big Data. Various techniques and approaches to predicting academic performance were examined, alongside the predictive variables used in these models [7]. Selecting appropriate predictor variables is fundamental to forecasting academic performance [8]. In fact, students' demographic Data, previous academic records, platform activity, and patterns of accessing educational resources are among the most used predictors of academic performance in virtual learning environments, as supported by several authors [9]. Combining these variables with Machine Learning algorithms and Data mining techniques allows for predictive models that can identify patterns and trends in academic performance [10].

Predicting student performance on the Moodle platform is beneficial not only for teachers but also for the students themselves. Academic performance prediction can help students identify areas for improvement and adopt more

[1] *Corresponding author: LyRICA: Laboratory of Research in Informatics, Data Sciences and Artificial Intelligence, School of Information Sciences, B.P. 604, Rabat-Instituts, Rabat, Morocco

[2] Center of Guidance and Educational Planning, B.P. 6222, Rabat-Instituts, Rabat, Morocco

[3] LyRICA: Laboratory of Research in Computer Science, Data Sciences and Artificial Intelligence, School of Information Sciences Rabat, Morocco, B.P. 604, Rabat-Instituts, Rabat, Morocco

[4] Laboratory of Engineering Sciences and Modeling, Faculty of Sciences-Ibn Tofail University, B.P 133, University Campus, Kenitra, Morocco

effective study strategies, contributing to improved academic success [11]. Additionally, educational institutions can use these predictive results to provide personalized interventions and improve the quality of higher education.

The Faculty of Sciences and Techniques of Gueliz-Marrakech is a higher education institution that began exploring virtual education in 2014. Since 2020, the use of technology and systems at the faculty has grown significantly due to the global COVID-19 pandemic, prompting faculty members to transition traditional teaching models to a virtual format. This shift was essential to meet the needs of approximately 16,113 students enrolled in various programs.

The move to virtual education has generated a substantial volume of Data from user interactions with the online platform. However, this Data remains largely unexploited, as there are no Data analysis mechanisms within the platform to monitor student behavior in virtual settings or predict their academic performance.

Analyzing the activity logs recorded by Moodle could offer teachers and administrators valuable information about student engagement in virtual classrooms, help identify different types of student behavior and predict performance outcomes.

Consequently, implementing Big Data mechanisms within Moodle is essential, as there are no existing procedures to analyze the large Datasets. Additionally, examining key algorithms used in similar studies and applying them could allow the development of a predictive model tailored to the faculty's context, thus improving the teaching and learning process in a virtual environment.

II. MATERIALS AND METHODS

The methodology used in this research is based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework [12], which has been adapted to address the specific needs and requirements for handling large Data volumes. The first phase involves Data analysis, extraction, and storage. It is followed by a second phase where variables are defined, and Data is filtered. The final phase focuses on prediction, as illustrated in Figure 1.

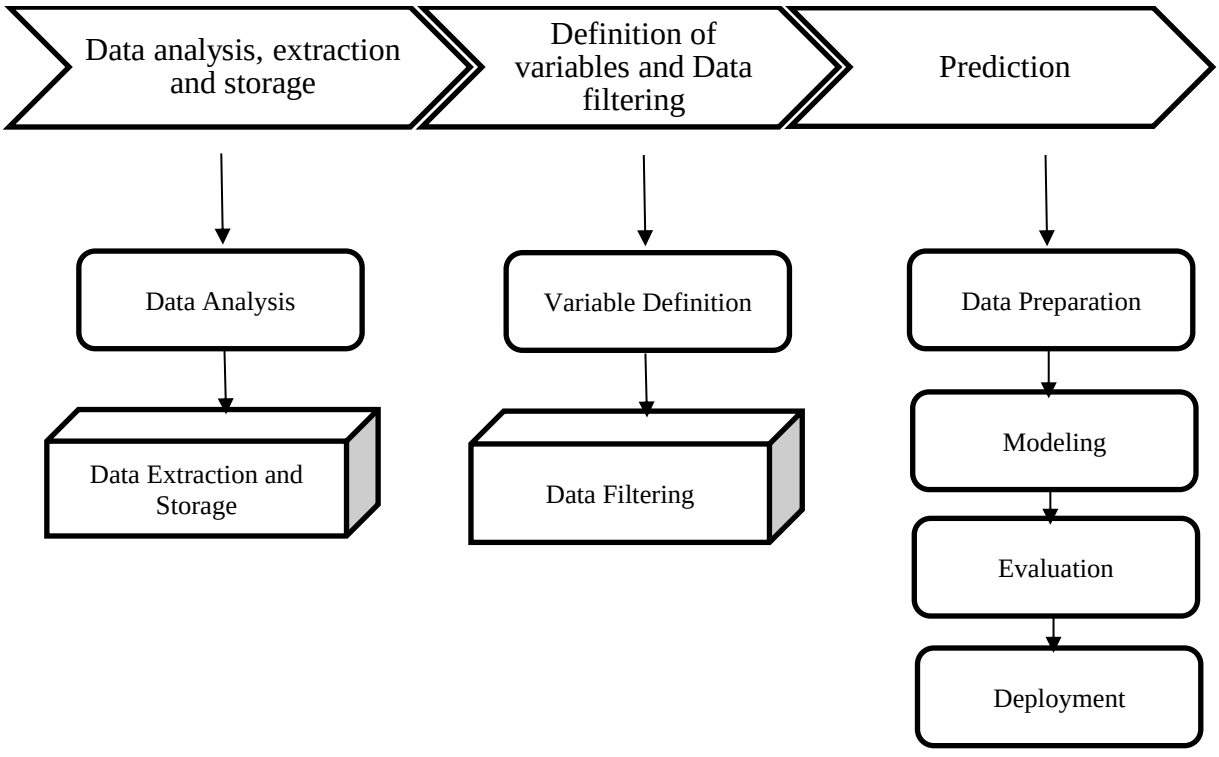

Fig 1. Stages of the research methodology used

The development of this methodology contributes to analyzing student behavior on the Moodle platform and predicting academic performance, incorporating the interaction of the tools employed in the process, as shown in Figure 2.

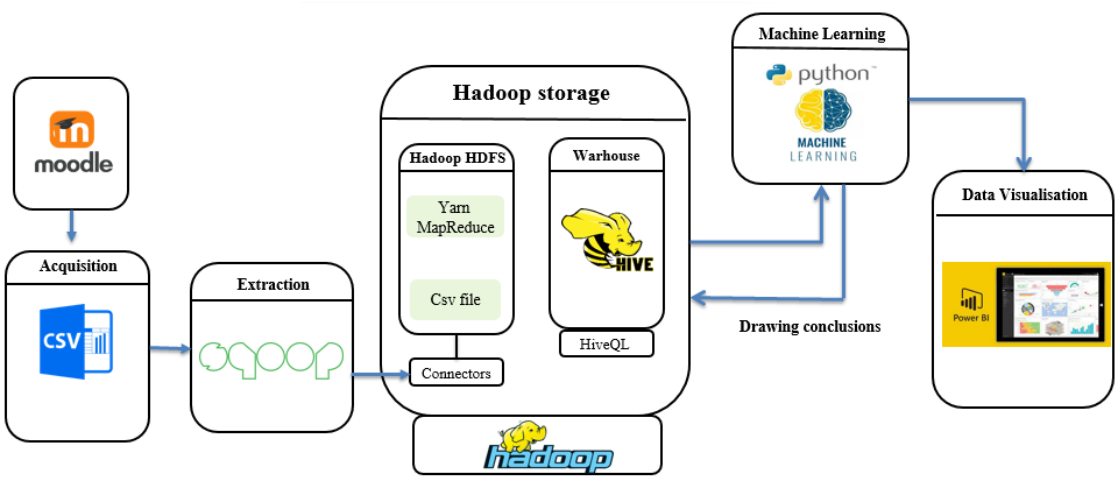

Fig 2. Interaction of various tools.

In the first phase, an exploratory analysis was conducted on Moodle's relational Database, which contains 462 tables with information on various courses, to understand the Data model used by the platform. Relevant tables for the study were identified, as presented in Table 1.

Table 1. Identification of the most relevant Moodle tables for the study

Moodle Tables	
mdl_logstore_standard_log	
Set of Tables I	mdl_user mdl_assign, mdl_assign_submission, mdl_config, mdl_context, mdl_course, mdl_course_modules, mdl_forum, mdl_forum_discussions, mdl_forum_posts, mdl_grade_grades, mdl_grade_items, mdl_groups, mdl_groups_member, mdl_modules, mdl_quiz, mdl_quiz_attempts, mdl_role_assignments, mdl_user_enrolments
Set of Tables II	mdl_feedback, mdl_chat, mdl_assignment, mdl_book, mdl_Data, mdl_folder, mdl_glossary, mdl_label, mdl_lesson, mdl_lti, mdl_page, mdl_quest, mdl_resource, mdl_scorm, mdl_survey, mdl_url, mdl_wiki, mdl_workshop

Similarly, in this phase, after identifying the potential tables for the study, the Big Data infrastructure was prepared. VirtualBox, a virtualization platform offering extensive functionalities for managing virtual machines, was chosen. A new cluster was created to which the virtualized Apache Hadoop Cloudera service was exported, including tools such as Sqoop, HDFS, and Hive.

In the second phase, based on the Data stored in the Big Data infrastructure, new variables were defined as indicated in Table 2. This process requires filtering the Data to select information that is essential for knowledge extraction and discard any unnecessary Data. Techniques such as HiveQL queries and statistical analysis were applied to the Data stored in Hive, with the results stored in a new Database.

Table 2. Variables for generating the prediction model.

Variable	**Description**
id_course	Short name of the course
student	Student identification
num_accesses	Total number of views by a student in the course

num_unique_accesses	Number of unique accesses made per day
num_time_course	Total number of minutes a student viewed the course
num_resource_visits	Total number of visits to a resource module (File, Folder, Page, URL)
num_url_visits	Number of visits to the URL resource
num_quizzes_completed	Total number of quizzes completed by a student in a course
num_forum_posts	Total number of forums viewed by a student
num_discussion_replies	Total number of replies made by a student in a course forum
num_assignments_submitted	Number of assignments submitted
num_night_views	Total number of course views by a student during the night
num_weekend_views	Total number of course views by a student during weekends (Friday, Saturday, and Sunday)
num_forum_posts	Number of forum participations
num_assigned_activities	Total assigned activities in the course (e.g., assignments, quizzes, forums, external assignments)
num_completed_activities	Number of completed activities in the course
num_quizzes_completed	Number of quizzes completed
num_external_assignments	Total assignments submitted using the assignment module and external tools like Google
course_completion_percentage	Percentage ratio of submitted assignments to the total assignments in the course
num_views_monday	Total number of course views by a student on Monday
num_views_tuesday	Total number of course views by a student on Tuesday
num_views_wednesday	Total number of course views by a student on Wednesday
num_views_thursday	Total number of course views by a student on Thursday
num_views_friday	Total number of course views by a student on Friday
num_views_saturday	Total number of course views by a student on Saturday
num_views_sunday	Total number of course views by a student on Sunday
final_grade	Final grade of a student in a course, either A (Passed) or P (Failed)

Finally, in the third phase, the CRISP-DM methodology is applied, detailing the tasks and activities needed to analyze student behavior and make predictions. These results support the improvement of the virtual teaching and learning process.

Data from 211 courses, involving 16,113 students, was processed. Based on the established variables, a total of 83,455 records were obtained. The records correspond to two academic periods in 2019/2020 and 2020/2021 for the in-person modality, which, due to the COVID-19 pandemic, were conducted virtually. Therefore, all subjects taught by program at the faculty were considered.

After obtaining the 83,455 records for modeling, the Data was divided, allocating 80% for training and 20% for evaluation.

As part of the Data preparation phase, variables with outliers were identified and appropriately treated. Additionally, an analysis of the academic performance variable revealed an imbalance in the Data, with 26% of the training records representing failed students. This necessitated the use of undersampling and oversampling balancing techniques:

1. Undersampling reduces the number of majority class samples (students who passed) to match the minority class (failed), balancing the classes but reducing the Dataset size.

2. Oversampling increases the minority class by generating synthetic samples, using a technique like Synthetic Minority Over-sampling Technique (SMOTE), to match the majority class, preserving all Data but potentially increasing redundancy.

In this case, the RandomUnderSampler for undersampling and SMOTE for oversampling, both available in the imbalanced-learn Python library, are used as demonstrated in the following code:

```
from imblearn.under_sampling import RandomUnderSampler
from imblearn.over_sampling import SMOTE
from sklearn.model_selection import train_test_split

# Split Data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Apply undersampling
undersampler = RandomUnderSampler(random_state=42)
X_train_under, y_train_under = undersampler.fit_resample(X_train, y_train)

# Apply oversampling with SMOTE
smote = SMOTE(random_state=42)
X_train_over, y_train_over = smote.fit_resample(X_train, y_train)
```

After applying these techniques, three Datasets were created for the application of various models, as presented in Table 3.

Table 3. Dataset for applying various models

Data Type	**Number of Records**	**Description**
Original	17 359	Records with both classes unbalanced
Undersampling	7 566	Records with balanced minority class - failed
Oversampling	27 151	Records with balanced majority class - passed

Using these balanced Datasets helps prevent model bias towards the majority class and supports better predictive performance across all classes.

III. RESULTS AND DISCUSSION

Executing each phase of the proposed methodology allowed for the establishment of the Big Data infrastructure needed for storing and processing large Data volumes. Following the analysis of the Moodle Database, the primary variables defining student behavior were identified, which were then used to create a predictive model of academic performance utilizing supervised algorithms. Power BI was employed to implement and deploy the proposed models, providing teachers and administration with a tool for visualizing results. This allows them to see, in summary, the main interactions on the platform, facilitating decision-making to improve the educational process in virtual environments.

After dividing and processing the Data and selecting the most relevant variables for the models, classification algorithms were applied to the different Datasets to identify which Dataset demonstrated the best performance. Eight supervised algorithms from Python's scikit-learn library were analyzed. After training, the models were evaluated using Machine Learning evaluation metrics, including precision, recall, accuracy, and F1 score, for each Dataset, as outlined in Table 4.

Table 4. Trained models evaluation results.

Algorithms	**Data Type**	**Precision**	**Recall**	**Accuracy**	**F1 Score**
XGBoost	Original	0.875	0.968	0.860	0.920
	Undersampling	0.915	0.860	0.820	0.880
	Oversampling	0.912	0.860	0.825	0.885
Catboost	Original	0.895	0.970	0.885	0.930
	Undersampling	0.928	0.865	0.840	0.895
	Oversampling	0.935	0.882	0.860	0.910
RF	Original	0.895	0.969	0.884	0.928
	Undersampling	0.932	0.860	0.835	0.895
	Oversampling	0.900	0.960	0.887	0.930
Decision Tree	Original	0.880	0.885	0.828	0.890
	Undersampling	0.917	0.745	0.740	0.820
	Oversampling	0.887	0.890	0.828	0.890
SVC - Linear Kernel	Original	0.822	0.985	0.820	0.895
	Undersampling	0.860	0.850	0.770	0.850
	Oversampling	0.870	0.855	0.780	0.860
k-Nearest Neighbors	Original	0.835	0.950	0.810	0.885
	Undersampling	0.855	0.680	0.655	0.760
	Oversampling	0.850	0.695	0.665	0.765

Naive Bayes	Original	0.860	0.870	0.780	0.865
	Undersampling	0.880	0.730	0.710	0.800
	Oversampling	0.878	0.740	0.715	0.805
Logistic Regression	Original	0.820	0.985	0.820	0.900
	Undersampling	0.860	0.825	0.750	0.840
	Oversampling	0.865	0.805	0.750	0.835

In the results obtained for each algorithm and Dataset, considering the precision metric, the models perform better with balanced Data using the undersampling technique. For the recall metric, the models perform better with the unbalanced Dataset. When evaluating the algorithms with the F1 score, which combines the performance of precision and recall, the best results are achieved with the original Data.

Considering the analysis of the results and the research objective, the precision metric was selected for choosing the algorithm that provides the best quality in predictions. According to the results presented in Table 4, the algorithm with the best performance is RF with 93.2% precision and 86% recall, followed by CatBoost with 92.8% precision and XGBoost with 91.5%.

After selecting the RF model, the confusion matrix was analyzed, showing the performance of the classification algorithm. The matrix displays 703 true negatives and 1476 true positives, as shown in Figure 3.

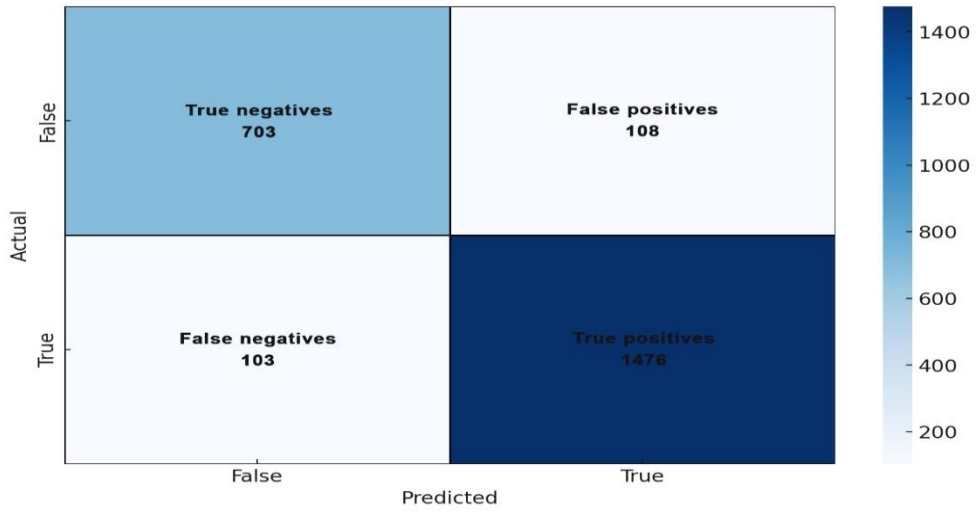

Fig 3. Confusion Matrix for the RF model

Based on the confusion matrix, the precision of 0.935 for the RF model is determined using the following formula:

$$Precision = \frac{True\ positives}{True\ positives + False\ positives}$$

In this case:

$$Precision = \frac{1476}{1476 + 108} \approx 0{,}932$$

Another way to evaluate the algorithms is to compare them graphically by plotting the top-performing ones. For this purpose, the ROC curve was used, which shows the false positive rate and false negative rate, revealing again

that RF achieves an AUC of 0.820, higher than CatBoost and XGBoost. Similarly, in the Precision-Recall curve, RF demonstrates higher performance, as outlined in Figure 4.

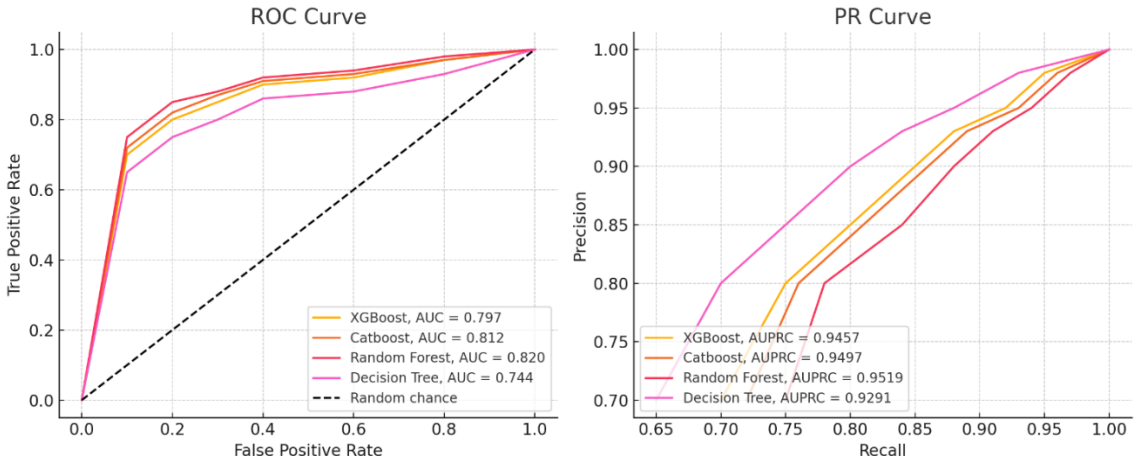

Fig 4. ROC and PR Curve.

For deployment and visualization of the results, Power BI is used, as it can handle large volumes of Data. Additionally, the RF model script was executed within Power BI, enabling predictions based on a new Dataset provided according to the established Data structure. This setup generates reports predicting student academic performance, as shown in Figures 5 and 6.

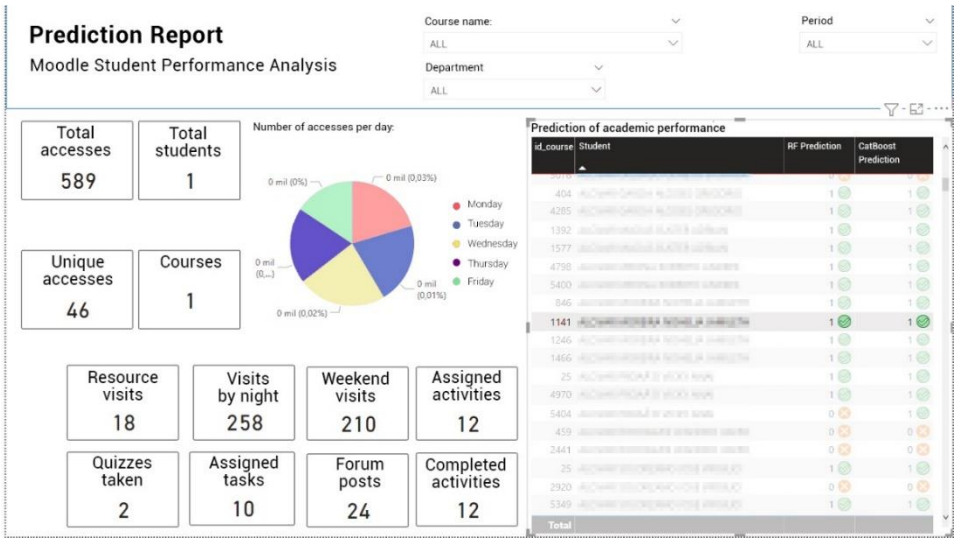

Fig 5. Prediction of a student who passes a course.

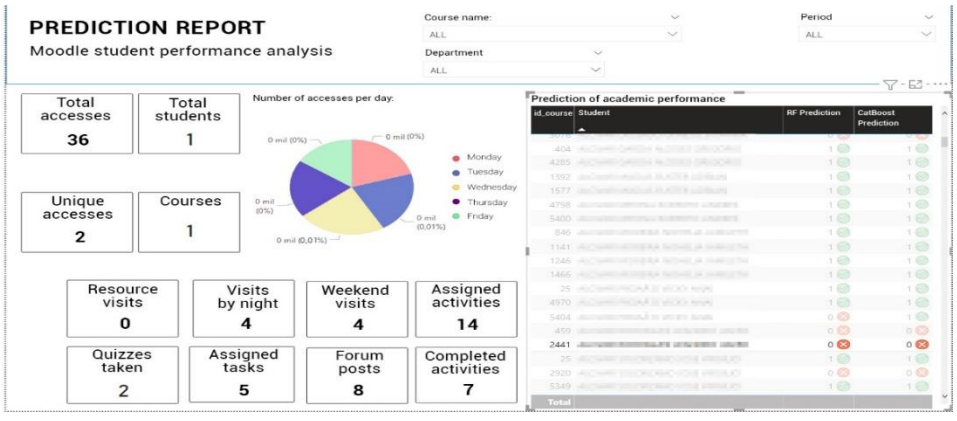

Fig 6. Prediction of a student who fails a course.

This study revealed that the use of Big Data and technologies like Hadoop, Machine Learning, and Power BI is essential for storing, processing, and analyzing large volumes of academic Data. These findings align with previous research that highlights the benefits of Big Data techniques and tools in university settings [13]. Predictor variables are crucial for forecasting student academic performance, such as the number of completed activities, platform access, forum participation, and grades. These findings are supported by similar studies using comparable variables, demonstrating the importance of student interaction with the Moodle platform [14] [15]. The development of predictive models using supervised algorithms, specifically the top three models, achieved over 90% performance accuracy in predicting student academic outcomes, surpassing results in some other studies [16] [17]. This demonstrates the potential for obtaining reliable predictions from large Datasets and provides evidence of performance comparable to previous research in the field.

IV. CONCLUSION

Analyzing the information generated by students' interactions on the Moodle platform through Big Data mechanisms allows insights into user behavior in virtual courses and how their actions impact learning. By training and evaluating various Machine Learning algorithms, a predictive model of academic performance can be developed. Based on the study and considering the large volume of Data generated on Moodle, Hadoop Cloudera is identified as a key tool for storing, processing, and analyzing academic Data, integrating technologies such as HDFS, Sqoop, and Hive. The application of Machine Learning classification techniques for academic performance prediction, alongside Power BI for presenting the results, is emphasized.

The findings demonstrate that predictor variables, such as the number of completed activities, platform access, and forum participation, are crucial in understanding and predicting academic outcomes. Among the models tested, RF consistently showed superior performance, achieving a precision of 93.2% and an AUC of 0.820. These results underscore the model's capability to provide accurate predictions, allowing for important information about student behavior and performance.

Future research can explore additional Machine Learning models and deep learning techniques to further improve the accuracy of academic performance predictions. Integrating real-time Data processing capabilities could enable continuous monitoring of student engagement, providing timely interventions for at-risk students. Expanding this framework to include more complex predictor variables, such as emotional or social factors, may yield a more comprehensive understanding of student success. Additionally, applying this approach across different institutions and learning platforms would help validate the model's adaptability and generalizability, contributing to a broader impact on educational analytics.

REFERENCES

[1] M. Zabolotniaia, Z. Cheng, E. Dorozhkin, and A. Lyzhin, "Use of the LMS Moodle for an effective implementation of an innovative policy in higher educational institutions," International Journal of Emerging Technologies in Learning, vol. 15, issue 13, 2020.

[2] A. S. Mustafa and N. Ali, "The adoption and use of Moodle in online learning: A systematic review," Information Sciences Letters, vol. 12, issue 1, pp. 341-351, 2023.

[3] S. Blagoeva-Karamfilova and S. Parusheva, "Adoption of LMS Moodle tools in student learning – in line with teaching practices," Pedagogy, vol. 94, issue 8, 2022.

[4] H. Abuhassna, W. M. Al-Rahmi, N. Yahya, M. A. Z. M. Zakaria, A. B. M. Kosnin, and M. Darwish, "Development of a new model on utilizing online learning platforms to improve students' academic achievements and satisfaction," International Journal of Educational Technology in Higher Education, vol. 17, pp. 1-23, 2020.

[5] C. Fischer, Z. A. Pardos, R. S. Baker, J. J. Williams, P. Smyth, R. Yu, and M. Warschauer, "Mining big data in education: Affordances and challenges," Review of Research in Education, vol. 44, issue 1, pp. 130-160, 2020.

[6] P. S. Aithal and S. Aithal, "Stakeholders' Analysis of the Effect of Ubiquitous Education Technologies on Higher Education," International Journal of Applied Engineering and Management Letters (IJAEML), vol. 7, issue 2, pp. 102-133, 2023.

[7] M. Shoaib, N. Sayed, J. Singh, J. Shafi, S. Khan, and F. Ali, "AI student success predictor: Enhancing personalized learning in campus management systems," Computers in Human Behavior, vol. 158, pp. 108301, 2024.

[8] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," Smart Learning Environments, vol. 9, issue 1, pp. 11, 2022.

[9] A. Namoun and A. Alshanqiti, "Predicting student performance using data mining and learning analytics techniques: A systematic literature review," Applied Sciences, vol. 11, issue 1, pp. 237, 2020.

[10] B. Albreiki, N. Zaki, and H. Alashwal, “A systematic literature review of student performance prediction using machine learning techniques,” Education Sciences, vol. 11, issue 9, pp. 552, 2021.
[11] D. H. Tong, B. P. Uyen, and L. K. Ngan, “The effectiveness of blended learning on students' academic achievement, self-study skills and learning attitudes: A quasi-experiment study in teaching the conventions for coordinates in the plane,” Heliyon, vol. 8, issue 12, 2022.
[12] R. Wirth and J. Hipp, “CRISP-DM: Towards a standard process model for data mining,” in Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, vol. 1, pp. 29-39, April 2000.
[13] K. L. M. Ang, F. L. Ge, and K. P. Seng, “Big educational data & analytics: Survey, architecture and challenges,” IEEE Access, vol. 8, pp. 116392-116414, 2020.
[14] Y. Gudkova, S. Reznikova, M. Samoletova, and E. Sytnikova, “Effectiveness of Moodle in student’s independent work,” in E3S Web of Conferences, vol. 273, p. 12084, EDP Sciences, 2021.
[15] S. H. Gamage, J. R. Ayres, and M. B. Behrend, “A systematic review on trends in using Moodle for teaching and learning,” International Journal of STEM Education, vol. 9, issue 1, pp. 9, 2022.
[16] S. Batool, J. Rashid, M. W. Nisar, J. Kim, H. Y. Kwon, and A. Hussain, “Educational data mining to predict students' academic performance: A survey study,” Education and Information Technologies, vol. 28, issue 1, pp. 905-971, 2023.
[17] S. A. Alwarthan, N. Aslam, and I. U. Khan, “Predicting student academic performance at higher education using data mining: A systematic review,” Applied Computational Intelligence and Soft Computing, vol. 2022, issue 1, pp. 8924028, 2022.