An Effective Decision-Making Support for Student Academic Path Selection using Machine Learning

Pélagie HOUNGUE
Institut de Mathématiques et de Sciences Physiques
Université d'Abomey-Calavi
Dangbo, Bénin

Michel HOUNTONDJI
Institut de Mathématiques et de Sciences Physiques
Université d'Abomey-Calavi
Dangbo, Bénin

Théophile DAGBA
Ecole Nationale d'Économie Appliquée et de Management
Université d'Abomey-Calavi
Cotonou, Bénin

***Abstract*—In Benin, after the GCSE (General Certificate of Secondary Education), learners can either enroll in a Technical and Vocational Education and Training (TVET), or further their studies in the general education. Majority of those who take the latter path enroll in Senior High School by choosing the Biology stream or field of study. However, most of them do not have the abilities required to succeed in this field. For instance, for the last edition of the Senior Secondary Education Certificate (French baccalaureate) held in June 2022 in Benin, the Biology field of study had a low success rate of 42%. Therefore, one may consider that there is a problem in the orientation of the students. In recent years, Machine Learning has been used in almost every field to optimize processes or to assist in decision-making. Improving academic performance has always been of general interest. And, good academic performance implies good academic orientation. The goal of this study is to optimally help learners who have just obtained their GCSE to select their field of study. For this purpose, two major elements are predicted: i) Scientific or Literary ability of students, ii) Literature or Mathematics and Physical Sciences (MPS) or Biology stream of learners. More precisely, the average marks in Mathematics, Physics and Chemistry Technology (PCT) and Biology from 6^{th} to 9^{th} grade for 325 students are used. Machine Learning algorithms such as Decision Tree, Random Forest, Linear Support Vector Classifier (SVC), K-Nearest Neighbors (KNN), and Logistic Regression are used to predict learners' ability and the stream. As a result, for learners' ability prediction, we obtained the best accuracy of 99% with the random forest algorithm for a split that reserved around 21% of the dataset for testing. As for the learners' stream prediction, we obtained the best accuracy of 95% with the Linear SVC algorithm for a split that reserved around 20% of the dataset for testing. This study contributes to Educational Data Mining (EDM) by performing academic data exploration using numerous methods. Furthermore, it provides a tool to ease students academic path selection, which may be used by educational institutes to ensure student performance. This paper presents the steps and the outputs of the study, we performed with some recommendations for future research.**

***Keywords*—Academic path; academic performance; machine learning; educational data mining**

I. Introduction

The improvement of academic performance has always been a concern for the educational system's actors [1], because the supposed performances in schools no longer satisfy everyone expectations. This situation requires a formula that could invert the trend. The education system is challenged to find a scientific instrument to overcome practices which continue to promote this deleterious situation [2].

According to [3], factors that impact field of study choice include: ability, experience, habit, program, instructor's role, university/school atmosphere and study culture. Authors in [4] also acknowledged that the field of study chosen by a student is in relation with prior knowledge and judgment of his/her own competence. Students often engage with peers and educational institutes through social networking to gather information about the university/school's fields of study or forthcoming courses. Therefore, deciding on a academic path can be stressful for secondary education students. Authors in [5] think that students need group guidance for fields of study and major choices.

If good or bad academic results have brought praise to some schools or tarnished the image of others, it is because of the lack of proper use of the data to direct students in the different fields of study. In fact, in Benin's high schools, students have to further their studies in general education track (Literature, Mathematics and Physical Sciences, Biology) or in Technical and Vocational Education and Training (TVET). Orientation in the TVET is not systematic because students must take an entrance exam before being admitted. Therefore, orientation problem is more acute in general education because it is done at best, on the basis of the marks obtained at the General Certificate of Secondary Education (GCSE). Otherwise, learners are oriented according to their parents' choice.

The use of Artificial Intelligence (AI) in general and more particularly, Machine Learning, in almost all fields, allows nowadays, to predict from the available data, a number of interesting elements for decision-making [6][7]. Thus, it is easy to understand the importance of using Machine Learning to improve the quality of academic performance [8][9]. This study contributes to research in Educational Data Mining (EDM) by developing a prediction model for academic orientations [10][11], in high schools in Benin. The study is articulated in two parts. We predict: (1) learner's aptitude or ability (literary or scientific), (2) learner's fields of study (Literature, Mathematics and Physical Sciences-MPS, Biology).

In the remaining of the paper, we address the background concepts of the performed study. Then, a state-of-the-art analysis is performed. Subsequently, the proposed prediction model' architecture is released, followed by the performance evaluation results. Furthermore, we highlight discussions and give an overview of the application that shows a concrete use of the optimal model. Finally, we summarize contributions and limitations of the proposed model and give the conclusions.

II. Background

In this section, we briefly discuss the concept of Educational Data Mining (EDM) and describe the beninese education system.

A. Educational Data Mining

Nowadays, many research activities are interested in data mining, and EDM has become a promising field of research [12], [13]. EDM uses several algorithms to improve educational results and account for educational procedures in future decision-making. EDM can be defined as the techniques for finding the specific types of data coming from the education system and implementing these techniques to better understand students and the system [14] [15]. Some applications of EDM can likely be a recommender system for students and prediction of their performance.

B. The Beninese Education System

After two years in Preschool (optional), students must complete six years in Primary school to achieve and obtain the Primary School Certificate (PSC) [16]. Primary school years are: Year 1 (Introductory Courses - IC), Year 2 (Preparatory Courses - PC), Year 3 (Elementary Courses 1^{st} year - EC1), Year 4 (Elementary Courses 2^{nd} year - EC2), Year 5 (Middle Courses 1^{st} year - MC1) and Year 6 (Middle Courses 2^{nd} year - MC2). Then, it takes seven years to complete Secondary school. Indeed, Secondary school in Benin, is made up of two levels: the first known as General Junior Secondary and the second is divided in General Senior Secondary and TVET. The General Junior Secondary grades are: 6^{th} grade, 7^{th} grade, 8^{th} grade and 9^{th} grade. Thus, at the end of General Junior Secondary school, learners must pass the exam of the GCSE. Afterward, grades of General Senior Secondary or TVET are: 10^{th} grade, 11^{th} grade and 12^{th}. In addition, to completing the 10^{th} grade, learners must enroll in a specific domain of study such as: General Senior Secondary (Literature, MPS or Biology stream for example) and TVET (Mechanical Science and Technology or Electrical Science and Technology stream for example). At the end of the General Senior Secondary or Technical Secondary, learner should take the Senior Secondary Education Certificate, equivalent of the French Baccalaureate. After that, they can enroll in the Tertiary Education in the existing courses and complete the years required to obtain a bachelor degree, a master or a doctorate [17]. Table 1 presents the main subjects taught in three streams, selected for the performed research. Fig. 1 illustrates the Beninese education system.

TABLE I. Main Subjects

Stream	Main subjects
Literature stream	French, English, German, Spanish, Philosophy
MPS stream	Math, PCT
Biology stream	Math, PCT, Biology

III. Related Work and Discussions

Hereby, are an overview of existing work and their shortcomings compared to the model proposed in this study.

A. Related Work

In literature, several ML algorithms are used for students academic orientation in high school. Whatever the educational system, at a given moment in his/her academic career, student at junior secondary level is required to make a choice of stream.

Therefore, to better orient learners, many research studies have been undertaken. In [18], authors predicted the performance of students in bachelor's and master's degrees in computer science and telecommunications.

Some models are based on real data. Usually, authors used tools such as a chatbot to collect data [19]. In [20], authors noticed that the increase in data did not significantly improve the obtained results.

Individuals in a database are characterized by a number of variables and all of them are not necessarily relevant for learner orientation. In [21], authors insist on correctly detecting the relevant variables involved in the process and their relationships with each other. In [22], authors used learners' scores to make predictions. In addition to grades, they used the number of absence per subject of the student.

In order to have a maximum prediction accuracy, in [23], authors compared several ML algorithms such as Support Vector Machine (SVM), neural networks, regressions, random forest, k-nearest neighbors, Naive Bayes', decision trees, etc.

Several performance measures exist to determine the degree of reliability of a prediction model. Following the example of [24], in which authors used three performance scores to validate their model (accuracy, Cohen's kappa, and the ROC curve), other authors used only the accuracy to assess the performance of their models. They used data that is not very large in size (a size that varies between 100 and 250 individuals). They obtained an accuracy of 94% with the random forest. In [25], authors were faced with an explosion of data and they obtained the best accuracy of 97% with Bayes' naive.

In addition, the authors of [26] implemented a framework which predicts academic orientation using supervised machine learning. They had a dataset of 350 individuals and compared the performance of decision tree, KNN, SVM and logistic regression by cross-validation and by a split that reserved 30% of the data for testing. Their proposal is mainly based on personality types such as Realistic, Investigative, Artistic, Social, Enterprising, and Conventional. They obtained the best AUC (Area Under Curve ROC) with the decision tree which was 0.8.

B. Discussions

From all above, we can notice three types of categories to classify the models aiming at providing academic or professional guidance for learners. There are the size of the data, the nature of the data and the nature of the prediction.

As far as data size is concerned, we can distinguish models that are based on a huge data size [25] and those based on a relatively small data size [24].

For the nature of data, some models are designed based solely on learners' grades and others are designed based

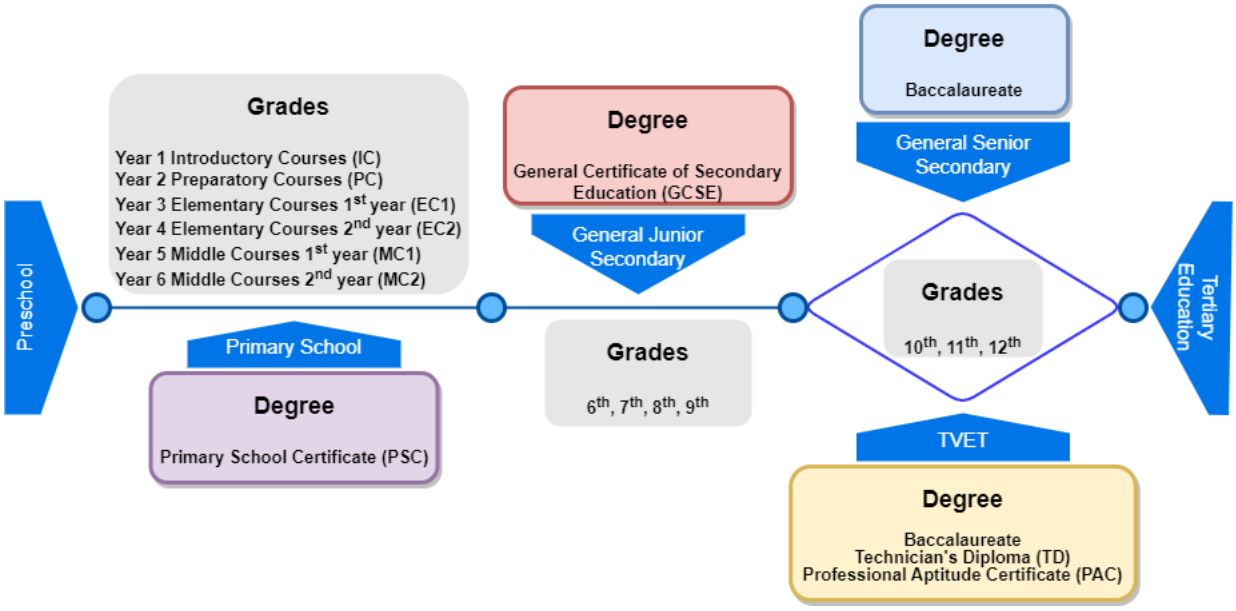

Fig. 1. The Beninese Education System.

not only on grades but also on learners' social environment variables [22].

Finally, different models are not intended to predict the same things. One set is designed to predict school dropout [21], other to predict the stream or the aptitude/ability in a given track and another to predict a score or the average mark [23].

Based on related work study, we can conclude that for academic orientation prediction, Bayes naive is the most suitable, when dealing with massive data, while random forest is the most suitable when data size is relatively small. It should also be noted that the combination of random forest and regression does not really provide good prediction accuracy.

Unlike previous works, current study focuses at first stage on learners' grades for guidance. Our educational context is different from those found in the literature and not all subjects have an impact on the learners' streams choice. Since a database of learners' digitized grades is not available, we started by collecting these grades using the students' transcripts of records. We trained the models on several ML algorithms with the scope of having maximum accuracy.

IV. Our Prediction Model

To achieve academic orientation prediction for Benin's high schools, we propose the architecture illustrated in Fig. 2. Our architecture includes several stages: preprocessing, model creation, model evaluation, and model optimization. However, the first three are compulsory for any prediction model and are described in this section.

A. The Dataset

A learner in the MPS stream, must have a basic knowledge of mathematics and PCT, and a learner in the Biology stream must have a basic knowledge of mathematics and PCT, as well as Biology. We can conclude that these subjects make it possible to distinguish scientific learners from literary learners. Moreover, considering subjects such as French, English, Philosophy, History and Geography, would not be optimal because they are cross-cutting subjects. One can be in Literature, MPS or Biology stream and be excellent in these subjects. In Benin, evaluations are done in secondary schools on a semester or quarterly basis. A semester or quarterly average in a given subject does not reflect the actual performance of learners in that subject over the course of a year. For this purpose, we used available transcripts of records to calculate annual averages in mathematics, PCT and Biology.

The dataset used in this study contains 325 learners' instances with 13 variables. Table II shows a description of all variables of the dataset.

TABLE II. Description of Variables

Symbol	Meaning
Mm6	Annual average mark of the 6^{th} grade in Mathematics
Mp6	Annual average mark of the 6^{th} grade in PCT
Ms6	Annual average mark of the 6^{th} grade in Biology
Mm5	Annual average mark of the 7^{th} grade in Mathematics
Mp5	Annual average mark of the 7^{th} grade in PCT
Ms5	Annual average mark of the 7^{th} grade in Biology
Mm4	Annual average mark of the 8^{th} grade in Mathematics
Mp4	Annual average mark of the 8^{th} grade in PCT
Ms4	Annual average mark of the 8^{th} grade in Biology
Mm3	Annual average mark of the 9^{th} grade in Mathematics
Mp3	Annual average mark of the 9^{th} grade in PCT
Ms3	Annual average mark of the 9^{th} grade in Biology
S	The stream of the learners

B. Preprocessing

Since the data collection stage has been performed manually, the probability to push up some shortcomings is high.

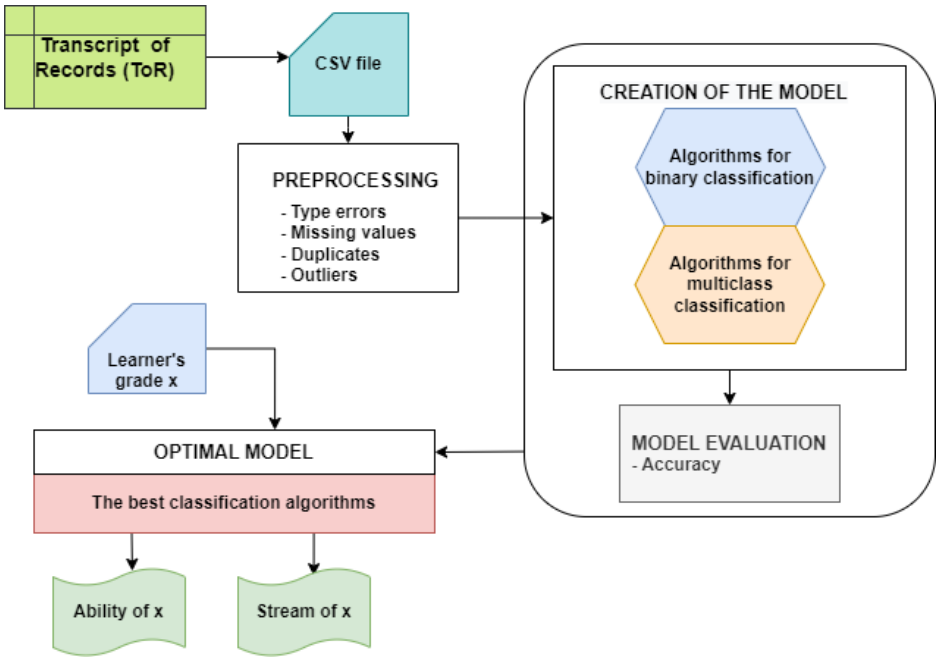

Fig. 2. Machine Learning-Based Architecture for Academic Orientation.

The most common are:

- Outliers Values
  Generally, the average mark is between 0 and 20. A typo error can lead to enter a value outside this range.
- Duplicates Values
  Two learners may have the same average marks in all considered subjects. Also, the same average marks can be entered twice by mistake.
- Missing Values
  Students may not have average marks in some subjects, for a variety of acceptable and unacceptable reasons.

The preprocessing stage allows us to clean the collected data (the average marks) to deal with missing, duplicates, and outliers values [27]. Then, the exploration of the average marks was done and allowed to notice that all the average marks of the learners will not be relevant for the prediction of their orientation. Indeed, we want to predict the optimal study path for this mass of general education learners who rush to the Biology stream because of the several opportunities it offers. Among this batch of learners who enroll in the Biology stream, some are more likely to take literature and others to choose TVET.

Taking the MPS or Biology stream, requires for learners to have good skills in Mathematics, Physics, Chemistry and Technology (PCT) and Biology. Therefore, we collected learners' yearly average marks in Mathematics, PCT and Biology from the 6^{th} grade to 10^{th} grade and the yearly average of the 10^{th} grade. Fig. 3 provides an overview of the data.

It can occur that some students choose Biology stream and fail, probably because they underperformed or were not proficient. Our concern here is to orient the learners in the best possible way, that is finding the right stream. Then, we proceeded to label the data using the following assumptions:

- If the annual yearly average as well as the yearly averages in PCT, mathematics and Biology in the 10^{th} grade are greater than or equal to 11, this instance is labeled as Biology stream.
- If an average mark of a given instance of data, in the 10^{th} grade is greater than or equal to 14 and the annual average marks in mathematics and PCT in the 10^{th} are greater than or equal to 15, then, the instance is labeled as Mathematics and Physical Sciences (MPS) stream.
- Otherwise, the instance is labeled as Literature stream.

C. Model Design

During this stage, ML algorithms allowed us to create prediction models. For this purpose, we use five Machine Learning algorithms and design two types of models:

- The first category of model is used to predict a learner's literary or scientific aptitude. It also allows the orientation of some learners towards TVET, since scientific aptitudes are compulsory for some streams such as Mechanical Science and Technology or Electrical Science and Technology for example.
- The second category of model predicts Literature, Biology or MPS stream of the learner.

D. Model Evaluation

In the model evaluation stage, we mainly use accuracy metric.

	Mm6	Mp6	Ms6	Mm5	Mp5	Ms5	Mm4	Mp4	Ms4	Mm3	Mp3	Ms3	MS
count	325.000000	325.000000	325.000000	325.000000	325.000000	325.000000	325.000000	325.000000	325.000000	325.000000	325.000000	325.000000	325.000000
mean	13.763077	14.132308	13.907692	12.680000	13.960000	13.590769	12.021538	12.926154	12.683077	10.316923	10.907692	11.747692	12.006585
std	2.722084	2.190968	2.602961	2.985934	2.622599	2.812325	3.241727	2.876162	2.846027	3.853152	3.434228	3.166818	2.225005
min	6.000000	7.000000	6.000000	5.000000	7.000000	7.000000	4.000000	5.000000	6.000000	2.000000	3.000000	3.000000	5.730000
25%	12.000000	13.000000	12.000000	11.000000	12.000000	12.000000	10.000000	11.000000	11.000000	7.000000	8.000000	9.000000	10.390000
50%	14.000000	14.000000	14.000000	12.000000	14.000000	14.000000	12.000000	13.000000	12.000000	10.000000	11.000000	11.000000	11.710000
75%	16.000000	16.000000	16.000000	15.000000	16.000000	16.000000	14.000000	15.000000	15.000000	13.000000	13.000000	14.000000	13.570000
max	20.000000	19.000000	19.000000	19.000000	20.000000	20.000000	20.000000	20.000000	20.000000	19.000000	19.000000	19.000000	18.270000

Fig. 3. Data Overview.

The accuracy is the metric that is often used to evaluate the performance of a classification model. It is the rate of good prediction. Therefore, the closer the accuracy is to 1, the better the model is performing.

V. Performance Evaluation

This section focuses on presenting the outcomes of the study in terms of model performance. Indeed, performance evaluation is performed in two steps: the first one uses 10 folds cross-validation and the second performs a specific split. The implementation code of the proposed models is available online (https://github.com/Jomamer/StudentAcademicPathSelection).

A. Overview of Labels

Before presenting the outcomes of the performed study, we present here the labels. In the dataset, there are 41% (133 instances) literary learners and 59% (192 instances) scientific learners.

Furthermore, there are 41% (133 instances) of Literature learners, 11% (37 instances) of MPS learners and 48% (192 instances) of Biology learners.

B. Cross-Validation Performances

The model is evaluated by performing a cross-validation of 10 folds.

Fig. 4 shows the mean accuracy and the std (standard deviation) of each algorithm for predicting learners' scientific or literary ability. Random Forest has the highest mean accuracy of 0.94 and the third lowest std of 0.08. It is followed by Linear SVC which gets 0.92 as mean accuracy and the second lowest std of 0.07. Logistic Regression gets the lowest mean accuracy of 0.87 and the highest std of 0.13.

Fig. 5 shows the mean accuracy and the std of each algorithm for predicting learner' Literature, MPS or Biology stream. Linear SVC and Random Forest obtain the best mean accuracy which is 0.90. They have the second and third lowest std of 0.08 and 0.1 respectively. Logistic Regression gets the lowest mean accuracy of 0.71 and the highest std of 0.15.

C. A Specific Split Performances

At this stage, we have booked for the prediction of each element, at least 20% of the dataset for validation.

In order to predict learners' scientific or literary ability, the size used for the test set by this split is 71 (more than 21% of the dataset). Table III presents the performances achieved. Random Forest has the best performance, followed by Linear SVC. KNN and Logistic Regression obtain the lowest performances.

For predicting Literature, MPS or Biology stream of the learners, the size dedicated to the test set, by this split is 66 (more than 20% of the dataset). Table IV presents the performances obtained. Linear SVC has the best accuracy.

TABLE III. A Specific Split Performances for Ability Prediction

Algorithms	Accuracy	Recall	f1_score
Decision Tree	0.97	0.96	0.98
Random Forest	0.99	0.98	0.99
Linear SVC	0.97	0.98	0.98
KNN	0.92	0.87	0.93
Logistic Regression	0.9	0.96	0.92

TABLE IV. A Specific Split Performances for Stream Prediction

Algorithms	Accuracy
Decision Tree	0.86
Random Forest	0.92
Linear SVC	0.95
KNN	0.94
Logistic Regression	0.82

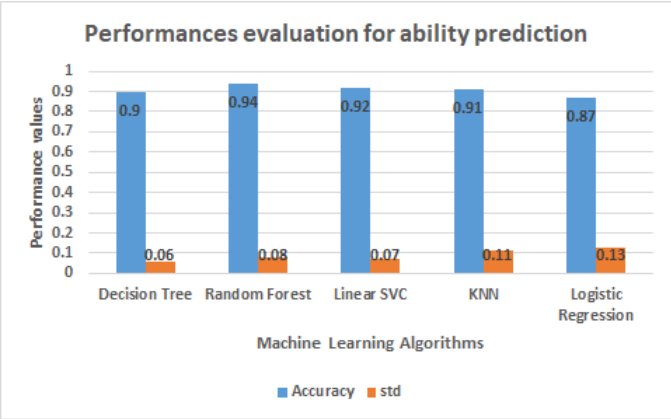

Fig. 4. Performances for Cross-Validation based Ability Prediction.

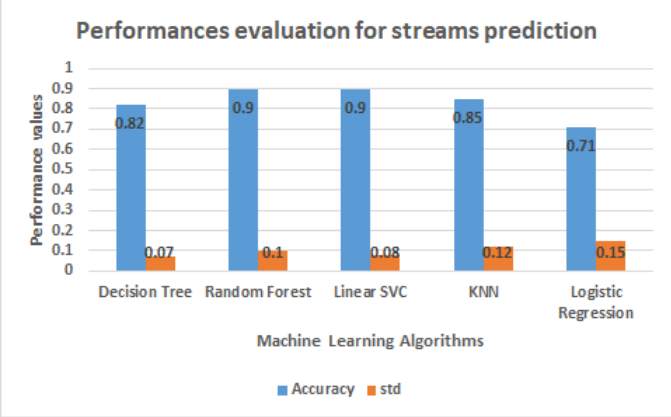

Fig. 5. Performances for Cross-Validation based Streams Prediction.

Fig. 6 and Fig. 7 represent respectively the best confusion matrices for ability prediction and stream prediction. Indeed, Random Forest presents the best confusion matrix for ability prediction. However, it raised one error. In fact, the algorithm predicts a scientific ability for the learner instead of the literary one.

On the other hand, Linear SVC gives the best confusion matrix for Literature, MPS or Biology stream prediction. It correctly predicts the learners in MPS stream. However, it made an error, which is to predict two learners for Biology stream, whereas they are actually in Literature stream. Another prediction error, less serious than the previous one, is the classification of a learner in Literature study rather than Biology stream. This error seems to be less severe because, usually a learner who is able to attend Biology stream can also attend Literature stream.

D. Optimization of the Model

The tuning of algorithm parameters allowed us to reach the optimal model. To achieve it, we used the Gridsearch method to make cross-validation of K parameter of the KNN, the max-depth for the trees decision and the n-estimators for the random forest on the training set. A higher number of trees gives better performance but slows down the code.

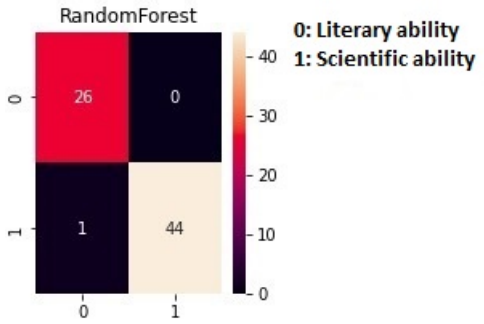

Fig. 6. Confusion Matrix for Ability Prediction.

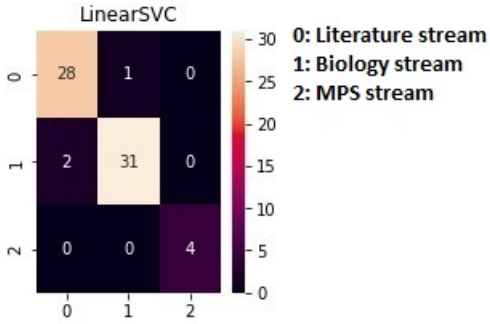

Fig. 7. Confusion Matrix for Stream Prediction.

In order to get better results, we tried to monitor the importance of the characteristics with Random Forest for the learners' scientific or literary ability, and the learners' stream. Taking into account the most important features, could improve the results. Fig. 8 gives an overview of the importance of the characteristics with Random Forest for the scientific or literary aptitude of the learners as well as the learners' stream.

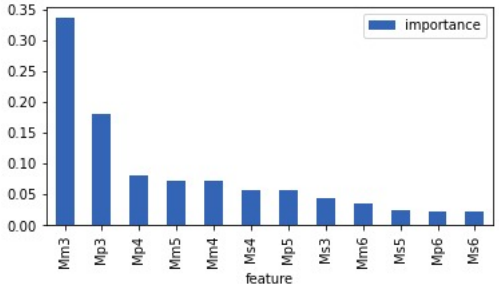

Fig. 8. Importance of Characteristics for Learner Ability and Stream with Random Forest.

Overall, the yearly average mark of mathematics and PCT of learner in 9^{th} grade are very important. Average mark of 6^{th} grade are less important. We tried to improve the performance obtained by considering only the characteristics that were having at least a given importance but it was not the case. We can conclude that average mark of the 9^{th} grade are certainly more important but they are not enough to guide the learners.

The optimal model is the one that offers the best performance. Two models are chosen at the end of the evaluation stage:

- a model to predict the scientific or literary ability of learner.
- a model to predict the Literature, MPS or Biology stream of learner.

VI. Discussion

This discussion is conducted along two axes: starting with the prediction of the learners' scientific or literary ability, then approaching the prediction of the Literature, MPS or Biology stream of the learners.

A. Case 1: Predicting Learners' Scientific or Literary Ability

Random Forest has the best mean accuracy of 0.94 and the third lowest std of 0.08. Linear SVC is not far behind with 0.92 for mean accuracy and 0.07 for std. With the specific split, which reserves more than 21% of the dataset for testing, Random Forest presents the best performances, which are 0.99 for accuracy, 0.99 for f1-score and 0.98 for recall. With the confusion matrix obtained, we noticed that, it made only one error. It predicted that a learner is in a scientific stream when in fact he is in literary stream.

From the above, we can deduce that for the prediction of scientific or literary aptitude of learners, Random Forest is the most suitable. Linear SVC can also be used as its performance is very close to the Random Forest one. It even has a lower std than Random Forest in the cross-validation comparison.

B. Case 2: Predicting Learners' Literature, MPS or Biology Stream

Linear SVC and Random Forest had the best mean accuracy of 0.9 and the second and third lowest std of 0.08 and 0.1 respectively. With the specific split, reserving more than 20% of the dataset for testing, Linear SVC has the best accuracy of 0.95. With the confusion matrix obtained, we noticed that it correctly predicts the learners of the MPS stream. However, it makes one error, in the sense that it predicts two learners from the Biology stream, when they are actually from Literature stream. There is also another prediction error that is less serious than the previous one. It predicts that one learner is from Literature stream when he is actually from Biology stream.

After all, we can conclude that for predicting Literature, MPS or Biology stream of the learners, Linear SVC is the most suitable. Random Forest can also be used as its performance is very close to that of Linear SVC. It even obtains the same average accuracy as Linear SVC in the cross-validation comparison. It is true that KNN has the second best accuracy in the specific split comparison, but we do not focus on it, because it has the second lowest performance in the cross-validation comparison.

In general, performance obtained when predicting learners' scientific or literary ability is better than those obtained when predicting learners' streams. This may be related to the fact that, usually, a learner who is able to attend MPS stream can also attend Biology stream. So, if the proposed model predicts the Biology stream for a learner who is actually enrolled in MPS stream, this is not an error.

As a matter of fact, it should be recognized that a learner's average mark does not only depend on the previous performance of the learners. It could rely on many factors such as: i) *How does the teacher explain the lessons?* ii) *Are the classmates much motivated to outperform?* iii) *Did the learner change school?* iv) *Is the teacher proficient with the subject?*

VII. Application of the Optimal Model

A Web page in https://jomamer-orientation-streamlit-app-rl3b0z.streamlitapp.com/ gives an overview of the framework that we designed to allow school's authorities to use the model. As a matter of fact, to make the proposed model usable, we designed a simple ML web application with Python and Streamlit. In particular, Streamlit is a Python package compatible with most of the Python libraries that are used for ML (scikit learn, keras, seanborn, matplotlib, numpy, pandas, tensorflow, etc.). It is an open source application framework in Python language that allows to create web application for data science and machine learning in a short time.

Notebook Jupyter allowed us to pre-test a number of things in order to deduce optimal models for predicting scientific or literary ability of the learners and Literature, MPS or Biology stream of the learners. After importing Streamlit, the useful libraries have been called. Then, we imported the entire dataset and trained the identified optimal models. As a matter of fact, the split allows us to reserve a part of the dataset to evaluate the model performance. In production, it was no longer a question of testing the model, but rather of using it. Therefore, we have used the entire dataset so that the model learns more cases.

Finally, the obtained results are quite satisfactory. Indeed, the strength of the proposed model relies on two predicted indicators which are complementary and play absolutely, a major role in the effectiveness of learner's orientation.

VIII. Contributions and Limitations

In this work, we developed a model for predicting academic orientations using ML techniques. To achieve this, we predicted the learners' scientific or literary ability and Literature, MPS or Biology stream. We used mathematics, PCT, and Biology average marks from the 6^{th} grade to the 9^{th} grade and the size of the dataset is 325. Our results are quite satisfactory for both models. The two models, together, allow achieving an optimal orientation for the learners.

However, it is worth to emphasize that in the paper [26] published in June 2022, authors went in the same direction. But, the major difference between both proposed approaches lies in the fact that they based themselves only on personality variables whereas the study performed in this paper is based on learner' average marks. Overall, the obtained performance results are better. It should be noted that both approaches may be complementary and future studies could look at how to combine them for optimal learner guidance about academic path selection.

On the other hand, with respect to the current study, authors of [24] used the same algorithms. Random forest has given an

accuracy of 94%, Decision tree 93% and Logistic regression 85%. They had bootstrapped a dataset of 101 records. One can notice that, the performances we obtained are better. Moreover, our data are real and in addition, we have also used the cross-validation which is recommended when dealing with a small size of data. It should be noted that authors of [24] have also taken into account characteristics such as age, gender and geographical area.

As a limitation, the current size of the dataset, can be underlined. Moreover, academic performance does not depend solely on the average marks obtained by learners. Several variables such as social environment factors could be taken into account.

IX. Conclusions

Everyone, no matter the position, tries to find a way to improve student's academic performances. In this work, we brought our support with a model, using Machine Learning which is very useful in almost every field nowadays. We compared the performance of five ML algorithms using a cross-validation and a specific split to predict learner's scientific or literary ability and the Literature, MPS or Biology stream of the learners. For this purpose, the Mathematics, PCT and Biology average marks from 6^{th} grade to 9^{th} grade of 325 instances of data are used. The best performance is pointed out by Random Forest algorithm with 99% of accuracy. Most probably, a larger data size could allow improving performance results even for the other algorithms.

The main limitation of this study resides on only considering the average marks obtained by the learners to perform the prediction. Thus, other parameters such as social environment factors may improve the performance results. The dataset size may also be increased. It would be interesting to look in these directions for future studies.

Acknowledgment

We thank the African Center of Excellence in Mathematical Sciences and Applications (CEA-SMIA) for funding this work.

References

[1] C. Bellei, X. Vanni, J. P. Valenzuela, and D. Contreras, "School improvement trajectories: an empirical typology," *School Effectiveness and School Improvement*, vol. 27, no. 3, pp. 275–292, 2016.

[2] S. Dhawan, "Online learning: A panacea in the time of covid-19 crisis," *Journal of educational technology systems*, vol. 49, no. 1, pp. 5–22, 2020.

[3] C. E. Garcia and C. W. Yao, "The role of an online first-year seminar in higher education doctoral students' scholarly development," *The Internet and Higher Education*, vol. 42, pp. 44–52, 2019.

[4] A. Dirin, M. Nieminen, and A. Alamäki, "Social media and social bonding in students' decision-making regarding their study path," *International Journal of Information and Communication Technology Education (IJICTE)*, vol. 17, no. 1, pp. 88–104, 2021.

[5] D. Park and Y. Kim, "A study on the effects of paramedic students' major selection motivation and occupational values on employment preparation behavior," *Journal of Digital Convergence*, vol. 18, no. 8, pp. 263–270, 2020.

[6] P. Kaur and R. K. Singh, "Feature selection pipeline based on hybrid optimization approach with aggregated medical data," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022.

[7] I. Paryudi, A. Ashari, and K. Mustofa, "The performance of personality-based recommender system for fashion with demographic data-based personality prediction," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022.

[8] B. Meriem, H. Benlahmar, M. A. Naji, E. Sanaa, and K. Wijdane, "Determine the level of concentration of students in real time from their facial expressions," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022.

[9] A. S. A. Osman, "Assessing the quality of educational websites in sudan using quality model criteria through an electronic tool," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022.

[10] A. Abdelhadi, S. Zainudin, and N. S. Sani, "A regression model to predict key performance indicators in higher education enrollments," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022.

[11] A. A. Saa, "Educational data mining & students' performance prediction," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 5, 2016.

[12] J. Calderon-Valenzuela, K. Payihuanca-Mamani, and N. Bedregal-Alpaca, "Educational data mining to identify the patterns of use made by the university professors of the moodle platform," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 1, 2022.

[13] W. Al Madhoun, "Predictive modelling of student academic performance–the case of higher education in middle east," Ph.D. dissertation, University of East London, 2020.

[14] R. S. Baker, T. Martin, and L. M. Rossi, "Educational data mining and learning analytics," *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*, pp. 379–396, 2016.

[15] C. Romero and S. Ventura, "Educational data mining and learning analytics: An updated survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 3, p. e1355, 2020.

[16] J. Alladatin, J. Bernachez, D. Bergeron *et al.*, "Overview of primary school principals' educational level and training in benin: The challenges related to the expected competencies and skills," *Annals of the University of Craiova, Series Psychology, Pedagogy*, vol. 43, no. 2, pp. 145–162, 2021.

[17] A. Assogbadjo, R. Idohou, and B. Sinsin, "Review of the higher education system in benin: Status, challenges, opportunities and strategies for improvement," *African Journal of Rural Development (AFJRD)*, vol. 1, no. 1978-2017-2051, pp. 139–149, 2016.

[18] H. Hamsa, S. Indiradevi, and J. J. Kizhakkethottam, "Student academic performance prediction model using decision tree and fuzzy genetic algorithm," *Procedia Technology*, vol. 25, pp. 326–332, 2016.

[19] O. Zahour, A. Eddaoui, H. Ouchra, O. Hourrane *et al.*, "A system for educational and vocational guidance in morocco: Chatbot e-orientation," *Procedia Computer Science*, vol. 175, pp. 554–559, 2020.

[20] R. Bertolini, S. J. Finch, and R. H. Nehm, "Testing the impact of novel assessment sources and machine learning methods on predictive outcome modeling in undergraduate biology," *Journal of Science Education and Technology*, vol. 30, no. 2, pp. 193–209, 2021.

[21] M. d. C. Nicoletti and O. L. de Oliveira, "A machine learning-based computational system proposal aiming at higher education dropout prediction." *Higher Education Studies*, vol. 10, no. 4, pp. 12–24, 2020.

[22] F. Ouatik, M. Erritali, F. Ouatik, and M. Jourhmane, "Students' orientation using machine learning and big data," 2021.

[23] A. Tarik, H. Aissa, and F. Yousef, "Artificial intelligence and machine learning to predict student performance during the covid-19," *Procedia Computer Science*, vol. 184, pp. 835–840, 2021.

[24] A. Dirin and C. A. Saballe, "Machine learning models to predict students' study path selection," *iJIM*, vol. 16, no. 01, p. 159, 2022.

[25] F. Ouatik, M. Erritali, and M. Jourhmane, "Student orientation using machine learning under mapreduce with hadoop." *J. Ubiquitous Syst. Pervasive Networks*, vol. 13, no. 1, pp. 21–26, 2020.

[26] H. El Mrabet and A. Ait Moussa, "A framework for predicting academic orientation using supervised machine learning," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–11, 2022.

[27] B. Malley, D. Ramazzotti, and J. T.-y. Wu, "Data pre-processing," *Secondary analysis of electronic health records*, pp. 115–141, 2016.