Education Challenge

Suhaib Aburaidah
dept. Mechanical Engineering
Jordan University of Science and Technology
Ramallah, Palestine
suhaib_aburaidah@hotmail.com

Yousef Qudeisat
dept. Computer Engineering
Jordan University of Science and Technology
irbid, Jordan
qudeisat@gmail.com

I. Introduction

Educational system is one of the most important inventions in the history of human civilization. People throughout their history and from different civilizations have been concerned about improving the educational process, where a better educational system means a more educated society and therefore a more developed nation. Improving educational system includes different aspects, such as: improving the facilities and equipment, improving methods of explanation and demonstration, and improving methods of conveying information to students. Another aspect for the improvement of educational system is the assessment and evaluation process. Assessment and evaluation process is very important, since a good evaluation of the students is very helpful in determining what skills, the students have mastered and what skills need to be improved.

A lot of experiments and research have been done to study students' way of thinking and how to evaluate students depending on their answers and performance in carefully prepared tests. Analyzing students' answers has become the new trend in the process of education development. What encouraged such a trend is the technological advancement, where it became possible to do a lot of electronic tests for a lot of students from different regions in the world and all the data can be easily collected and saved for future analysis. A huge database is being collected, and all that it needs is a powerful tool for analysis. Out of all the tools developed or being developed, machine learning techniques proved to be the most powerful and helpful tool for the job.

Machine learning participation in the education sector has increased significantly in the recent years. It tries to give each student a personalized experience depending on his skills, background, and way of thinking. Machine learning has many applications in the education sector. It includes adaptive learning which analyzes students' performance and adjust the teaching methodology depending on the data. It also includes predictive analytics which predicts students' future performance depending on their current situation, this may help students to know their needs and what they should do to get better results. Machine learning also includes efficiency increasing which compares different curriculums and tries to find the best of them and the most suited for given students.

In this paper a machine learning model is going to be done to solve the NeurIPS 2020 Education Challenge, the challenge intends to learn from the answers of a big sample of students depending on some features they have, so that a model can predict future answers for different students or different questions. Students from different backgrounds and skills were tested to hundreds of multiple-choice questions. The students' features and all the questions and answers were saved for analysis. To find the best solution for the challenge, many different machine learning algorithms will be utilized, such as: SVM, regression, decision tress, NB, random forest, etc. The results for all of these algorithms will be shown including details about the parameters and functions used in each algorithm, and finally, a comparison is done to find the best algorithm and the best model for the solving the challenge.

II. Background

Machine learning is a part of artificial intelligence. It is a data analysis method that learns from data, classify objects, and find patterns to make future decisions about a given subject. Machine learning started in 1950s [2].

Machine learning can be classified into four categories. The first category is supervised learning which learns from labeled examples of data to apply what it learned on unlabeled data and predict their labels. The second category is unsupervised learning which tries to find patterns and classifications in given data that are not labeled or classified. The third category of data is semi supervised learning which takes two sets of data, a small amount labeled data and a big amount unlabeled data. The fourth category is reinforcement learning which interacts with its environment by producing actions and discovers errors or rewards by trial and error.

Machine learning types of prediction are two types. Classification and regression. Classification is the prediction of a sample label among limited number of labels. It tries to divide training samples into number of groups and gives each group a certain label, then predicts the label of a future sample or in other words, to what group it belongs. Such as: classifying products in a production line depending on their shapes or classifying E-mails into spam or not spam. Regression in the other hand, is the prediction of a sample label which is usually a quantity that belongs to a continuous interval with infinite number of values. Such as predicting future prices of a given

Identify applicable funding agency here. If none, delete this.

product or predicting future values of some physical properties such as: temperature or pressure.

Many machine learning algorithms have evolved through years. Some of the most prominent algorithms will be discussed. First, linear regression algorithm. The purpose of this algorithm is to estimate real values of a continuous independent variable depending on some labeled data of that variable, the algorithm tries to find the best line that fits the labeled data. Second, logistic regression algorithm. This algorithm tries to fit an s-shaped line to a sample of data that have binary targets (True or False). Third, SVM (support vector machine) algorithm. This algorithm is a classifier that tries to find the best splitting lines between different categories of data. Fourth, NB (naive bayes) algorithm. This algorithm depends on the bayes theorem to predict the probability of an event given some features with the assumption of independency between features. Fifth, decision tree algorithm. This algorithm is a classification algorithm that is used to classify data into subgroups that are divided into other subgroups and so on until we make as distinguished groups as possible. Sixth, random forest algorithm. Random forest is an ensemble of decision tree. It makes a collection of decision trees, the most frequent output resulting from all the decision trees is the correct output.

III. Literature Review

In [1] the performance of university students is predicted using students' marks in 10th, 12th, and previous semester marks. The study is done using different algorithms including logical regression, decision tree, entropy and KNN classifier. The purpose of this paper is to help students expect their final grade and increase their effort accordingly if needed. However, in our research we are going to predict and analyze students' performance using their answers to a collection of questions not their previous marks. We are also going to use other machine learning algorithms than the mentioned in [1].

In [2] the performance of postgraduate students using machine learning is studied. The main purpose of the article is to predict the final grade of postgraduate students at Ionian University. The study includes five academic courses with their own datasets. Six machine learning classifier algorithms were used, decision tree, two k-nearest neighbor algorithms, naïve bayes, random forest and support vector machine (SVM). The main challenge of the research is the small size of datasets and the imbalance of the distribution. Results of the research show that Naïve bayes and 1-NN algorithms are the best algorithms for the prediction. [2] is good research, it's very informative, however it restricts the research for postgraduate students only.

Lincke et al. [3] conducted a research article on students' answers prediction using machine learning approaches. The purpose of the study is to make the teaching process more adaptive and personalized. The study shows that the difficulty level of the question and the incorrectly answered previous questions are very useful features for the predicting process. Results of the study also shows that gradient boosted tree and XG-Boost algorithms are the best algorithms for the predicting process of the study.

In [4] a solution for task 2 of the NeurIPS 2020 education challenge is submitted. The paper uses CNNs machine learning algorithm and order-aware cognitive diagnosis model (OCD model). The accuracy achieved in the prediction is about 0.68, which is the highest performance in the task. The research has four sections. In section 2 the structure of OCD model is represented. In section 3 the data is processed, trained, and experimented. In section 4 discussion and a brief summery in done. This research is very good, specially that it got the best performance in the challenge, however, it is based on the OCD model which is out of the scope of our research.

Zhang et al. [5] presented a solution for task 1 and task 2 of the NeurIPS 2020 education challenge. The paper uses three different machine learning algorithms to solve the question. GBDT algorithm, multi-head attention-based network algorithm and a transformer-based network algorithm. In the paper the two tasks were treated as a recommender systems (RS) problem. The solution in the paper has got the second and fourth places in task 1 and task 2 respectively. Results show that GBDT got a prediction accuracy of 0.7581 while multi-head attention got 0.7594 and Transformer-based model got 0.7661. This means that multi-head attention algorithm got the highest performance.

Pallathadka et al. [6] conducted a research classification and prediction of students' performance using machine learning algorithms. The purpose of the research is to help students concentrate their efforts in a specific area depending on advice from the machine learning model and therefore reduces failures rate in educational institution. Different machine learning algorithms are used such as: Naïve bayes, SVM, ID3 and C4.5. the dataset used in the research is UCI machinery student performance dataset. Results show that SVM got the highest accuracy with of about 85 percent. The second highest performance algorithm with an accuracy of about 75 percent.

In [7] machine learning algorithms are used to predict students' performance in distance learning. The purpose of the paper is to develop and enhance the process of distance learning by applying artificial intelligence and machine learning techniques. The dataset used is provided by Hellenic Open University. The paper conducts two experiments with six different machine learning algorithms. Results show that Naïve bayes algorithm got the highest performance with a predictive accuracy of about 70.5 percent, however, for Naïve bayes algorithm to have a good accuracy, the minimum number of instances should exceed 70 instances.

Oku and Sato [8] conducted a research paper for predicting the Student Performance Using Machine Learning in fNIRS Data. Functional near-infrared spectroscopy (fNIRS) is a noninvasive optical imaging technique that measures changes in hemoglobin (Hb) concentrations within the brain by means of the characteristic absorption spectra of Hb in the near-infrared range [9]. The paper uses random forest and penalized logistic regression algorithms to classify correct answers. Both algorithms got 0.67 and 0.65 area of the ROC curve, respectively.

IV. Data Description

The data-set in this project is provided by the NeurIPS competition under the title ”Education Challenge”. The data-set consists of a group of students who are subjected to a group of multiple-choice questions. Each student is subjected to a subset of the questions that exceeds 50 questions. The challenge includes four tasks. The first task is to predict whether the student’s answer is correct or not. The second task is to predict student’s answer. The third task is design a metric to measure question’s quality. The fourth task is to acquire a limited set of answers from students for student performance prediction on unseen questions. In this paper focus is going to be on the first task only.

The data contains the train_task_1_2 data file which consists of the student’s Id number, question Id number, answer Id number, student’s answer, the correct answer and the validity of the answer. Data also contains the test_public_answers_task_1 file and test_private_answers_task_1 file which contain student’s Id number, question’s Id number, answer’s Id number and the validity of the answer (whether the answer is correct or not).

The data-set contains a student meta-data file that contains the information or features about each student. the features include: student’s Id number, Student’s gender, student’s date of birth and student’s premium pupil advantage. The data-set also contains a questions meta-data file that contains information or features about each question in the data-set. the questions meta-data includes: question’s Id number and question’s subject Id. Subject Id contains all the subjects of mathematics that a question involves. The data-set contains also an answer meta-data file, which contains information or features about each answer. The answer meta-data file includes: Answer’s Id number, date of answer, student’s confidence of the answer, group Id which is the group in which the student was assigned the question, and finally quiz Id which is the quiz that contains the question.

In order to train the model, the train data file is merged with meta-data files so that the features of the students, questions and answers are all know for each sample in the train data file. The same thing is done for the test data file. Not all the features are used, some of them are omitted because they are whether unimportant or difficult to deal with. The features chosen are: QuestionId, UserId, student’s gender, student’s age, student’s groupId, QuizId, subject 2 and subject 3. It is important to mention that date of birth was treated so that the year is only thing taken into consideration, while other information such as: hour, day, or even month are all neglected. This makes sense, where hour, day and month usually have insignificant effect on the academic development of a student. A table that includes all the features is shown below, the table shows the first 20 samples from above. Note that the target is omitted from the table for convenience. The target is validity of the answer (whether it is correct or not).

Gender	Age	GroupId	QuizId	Sub2	Sub3
1	14	4975	17226	42	211
0	12	3874	17226	42	211
1	13	10399	17226	42	211
2	14	10035	17226	42	211
1	33	6005	17226	42	211
1	33	10318	17226	42	211
0	12	8519	17226	42	211
1	33	7742	17226	42	211
1	13	1238	17226	42	211
2	13	4957	17226	42	211
1	12	6427	17226	42	211
1	15	11573	17226	42	211
0	12	2580	2210	42	211
2	12	1915	17226	42	211
0	12	10732	17226	42	211
1	12	7949	17226	42	211
0	12	8519	17226	42	211
0	12	8508	17226	42	211
1	15	4576	17226	42	211
1	11	7949	17226	42	211

Since the data-set is very large and hard to deal with using normal personal computer, the data was reduced by omitting each sample or row that doesn’t contain subjectId 32. This helped in two ways, first it reduced the size of the data and made it easier to process, second it made the training process simpler by omitting questions that do not relate to numbers.

V. Architecture of the successful model

Out of all the models that were used, the random forests technique had the best results. Random forests is a part of the ensembling techniques. Random forests combines several decision tree models each trained on a subset of the features of the data-set and then it makes a prediction depending on the concept of majority voting of the decision tree models. An illustration of the working principle of the random forest classifier is show in Fig. 1.

The random forests model used in this paper was applied on decision tree classifiers. The decision tree classifier was done using the default hyper-parameters: criterion is gini and max depth is none (unlimited). The hyper-parameters of the random forests model it self are: number of estimators are 20 and n-jobs is -1. The hyper-parameters of the random forests model are summarized below:

criterion	max depth	# estimators	n-jobs
Gini	none	20	-1

Although the random forests model got the best performance, it didn’t differ so much from other models. This indicates that the system doesn’t get affected a lot with changing model. The performance of the random forests model was not excellent, however, compared with the complexity of the data and the difficulty to find a pattern in it, the performance is acceptable. The system got a an accuracy score of 69.2%. The f1-score is about 77.6%. The precision is about

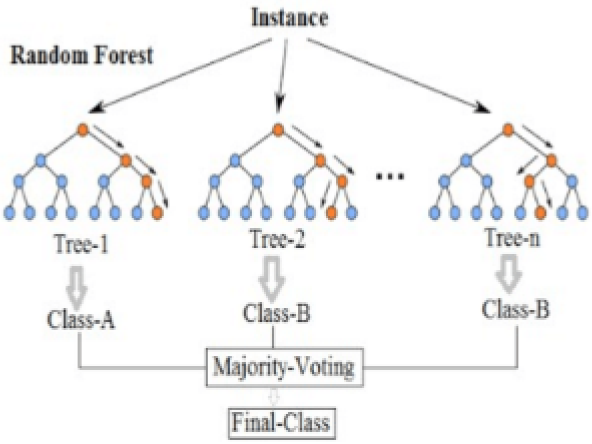

Fig. 1. Random forests classifier illustration

75.5% and the recall is about 79.7%. The performance scores are all summarized in the below table.

Accuracy	f1_score	precision	recall
69.2%	77.6%	75.5%	79.7%

The data trained in the random forests model is the same data described in the data description section. The features used are eight features. Gender, age, groupId, quizId, sub2, sub3, questionId and userId.

VI. DISCUSSION OF RESULTS AND MODELS EVALUATION

In order to give a comprehensive view of the problem, we have started with exploring the dataset. After downloading the training dataset from Kaggle, we have split the data into train and target dataset. Then, we have ranked the features and selected the important ones for our experiment. Therefore, the underlying frequency distribution of the features has been studied, and the correlation matrix between the features has been calculated. As a result, we have found that there is a small correlation between features that means the features are mostly independent from each other. For this reason, a feature selection technique has been applied in order to select the most important features and drop the rest.

After exploring the given dataset and preparing it to become compatible with machine learning algorithms, we studied different classification models and results as follow:

1. SVM didn't work Due to large Data and Processing limitations.

2. GaussianNB, with no parameters

Accuracy	f1_score	precision	recall
66.3%	79.6%	66.7%	98.2%

3. LogisticRegression, solver="lbfgs" with no parameters

Accuracy	f1_score	precision	recall
66.8%	80.1%	66.8%	1%

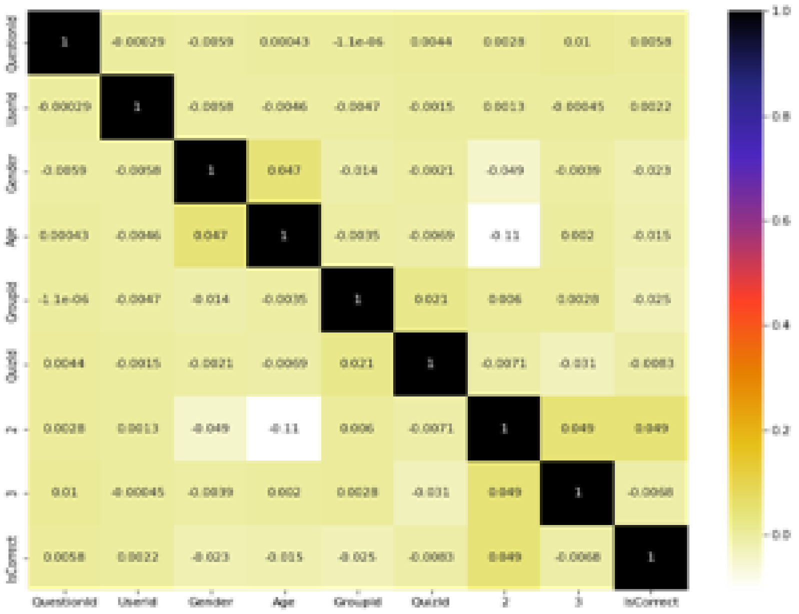

Fig. 2. Features correlation illustration

4. DecisionTreeClassifier we have tried 3 models and there was an improvement as follow:

a. first one with no parameters

Accuracy	f1_score	precision	recall
65.9%	74.4%	74.7%	74.1%

b. second one with max_depth=2,min_samples split=500000

Accuracy	f1_score	precision	recall
66.8%	80.1%	66.8%	1%

c. Third one with max_depth=10,min_samples split=200000

Accuracy	f1_score	precision	recall
66.8%	80.1%	66.8%	1%

5. RandomForestClassifier we have tried 3 models with different parameters and there was an improvement as follow: a. n_estimators=3

Accuracy	f1_score	precision	recall
66.9%	75.8%	74.2%	77.3%

b. n_estimators=10

Accuracy	f1_score	precision	recall
68.2%	76.4%	75.9%	76.9%

c. n_estimators=20

Accuracy	f1_score	precision	recall
69.2%	77.6%	75.5%	79.7%

d. but with GridSearchCV improvement was different:

The best parameters were: criterion: Gini, max_depth: 2, max_features: auto, n_estimators: 20

Accuracy	f1_score	precision	recall
66.8%	80.1%	66.8%	1%

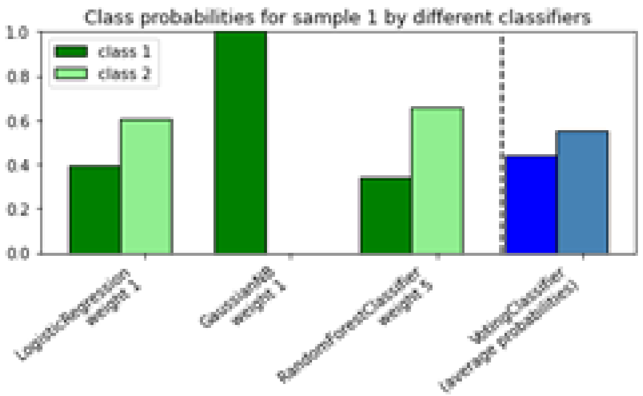

Fig. 3. Ensemble results illustration

6. Gradient boosting classifier, it was done using grid search. The best parameters were: min_samples_leaf: 100000, min_samples_split: 500000

Accuracy	f1_score	precision	recall
66.8%	80.1%	66.8%	1%

7. Bagging classifier with two different classifiers:

a. Base_estimator = decision tree classifier(max_depth=5), n_estimators=20, bootstrap=True ,n_jobs=-1

Accuracy	f1_score	precision	recall
66.8%	80.1%	66.8%	1%

b. Base_estimator = GaussianNB classifier(), n_estimators=100, bootstrap=True ,n_jobs=-1

Accuracy	f1_score	precision	recall
66.8%	80.1%	66.8%	1%

8. Ada-boost classifier with search_grid:

Best parameters: learning_rate: 0.05, n_estimators: 10

Accuracy	f1_score	precision	recall
66.8%	80.1%	66.8%	1%

9. We tried ensemble with logistic regression, decision tree classifier and Gaussian Naive Bayes, with soft and weighted voting.

The results were as shown in Fig.3:

VII. Conclusion

Improving the educational system is one of the most important targets for all humanity, however it is not an easy problem. It is very complicated and includes so many parameters to consider. In this paper huge efforts were exerted to solve the educational problem in which the correctness of the student's answers is to be expected.

Many different algorithms were used to build a model that predicts the correctness of each student's answer. However, it was found that the best model is random forests with number of estimators equal to 20 and n_jobs equal to -1. The model got an accuracy of 69.2 percent, f1-score equals to 77.5 percent, precision equals to 75.5 percent and recall equals to 79.7 percent.

The results seem to be not so good, where the best accuracy was 69.2%. However, these results are not bad and considered very good compared to the complication of the problem since it is not easy at all to predict the correctness of the student's answers with only some attributes related to the students such as their gender, age and the group they belong to.

References

[1] J. Dhilipan, N. Vijayalakshmi, S. Suriya and A. Christopher4, ”Prediction of Students Performance using Machine learning,” IOP Conference Series:, vol. 1055, 2021.

[2] M. Koutina and K. L. Kermanidis, ”Predicting Postgraduate Students' Performance Using Machine Learning Techniques,” in Artificial Intelligence Applications and Innovations, Corfu, Greece, International Conference on Engineering Applications of Neural Networks, 2011, pp. 159-168.

[3] A. Lincke, . M. Jansen, . M. Milrad and E. Berge, ”The performance of some machine learning approaches and a rich context model in student answer prediction,” Springer, 2021.

[4] Shen, Liu, Chen, Tong, Huang, Tong, Su and Wang, ”Which to Choose? An Order-aware Cognitive Diagnosis Model for Predicting the Multiple-choice Answer of Students,” Big Data Analysis and Application, 2020.

[5] Zhang, Qin, Zou, Zhu, Liu, Liang and Zhang, ”How to Predict Students' Interactions with Diagnostic Questions: from A Perspective of Recommender System,” NetEase, Inc., 2020.

[6] Pallathdka, Wenda, Ramirez-Asis, Asis-Lopez, Phasinum and Flores-Albornoz, ”Classification and prediction of student performance data using various machine learning algorithms,” Materialstoday, 2021.

[7] Kotsiantis, Pierrakeas and Pintelas, ”Predicting students' performance in distance learning using machine learning techniques,” Applied Artificial Intelligence, pp. 411-426, 2010.

[8] Oku and Sato, ”Predicting Student Performance Using Machine Learning in fNIRS Data,” National Library of Medicine, 2021.

[9] Di Domenico and Ruocco, ”Functional near Infrared Spectroscopy,” in Neuroergonomics, 2018.