Machine Learning in The Education Process
Capstone Project

Project Overview

Improving the performance of education has a significant impact on ensuring the nations' economic prosperity and represents a central focus of the government when making education policies. During the last years, machine learning techniques achieve this goal in education by developing methods of exploring data from computational educational settings and discovering meaningful patterns (Baker and Yacef 2009). This project is to model student performance which is an important tool for both educators and students since it can help a better understanding of this phenomenon and ultimately improve it in different educational stages. The main target of our work is to show how a student's family affect his educational performance.

Problem statement

Predicting student's performance is an important task in educational environments. There are several machine learning methods such as Support Vector Machine (SVM), Artificial Neural Networks (ANN) and Naive Bayes (NB) had been applied to model student's performance. The goal of this project is to build student's performance prediction model based on student's family data. With different two datasets, I will be applying several machine learning methods on both data sets. the structure of both data sets is different so I will used them individually.

We are going to answer the following two questions:

- How a student's family affect his educational performance in the secondary stage?
- Can students rely on themselves to study online without family control?

Evaluation Metrics

In order to evaluate the effectiveness of a prediction model, predicted values must be compared with actual values. The matrix that shows the possible prediction results is called a confusion matrix (Powers, 2011).
In the first data set target values are continues so, I will use Root Mean Squared Error (RMSE) to compare model's performance -a regressor should present a low global error (i.e. RMSE close to zero)- as used in (Paulo et al.,2008)

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

However, in second dataset the target is classified in to three balanced categories so predictions made by these models are compared using common evaluation criteria, such as
Accuracy is basically the ratio of correct predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$

Precision and recall are used together to make a better evaluation.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

F-measure is the final evaluation criteria for comparisons in this project.
Which used in (Amrieh et al., 2016)

$$Fc = 2\frac{Precisionc * Recall\ c}{Precisionc + Recall\ c}$$

Data Exploration

This project aims at exploring and analyzing student performance through Two different data sets containing records of student information by applying machine learning methods to both.

1.Exploring and Visualizing First Data

The First data set was originally used in a research done at the University of Minho, Portugal (Cortez and Silva, 2008). It contains information about 395 students has 33 different variables see table 1.
I will use these features to predicate G3 - final grade (numeric: from 0 to 20) as target value

Table 1. Feature descriptions for the first data set.

Features Category	Feature	Description	Type	Family feature
Demographical Features	School	Name of student's school	Nominal	
	Sex	Gender of student	Nominal	√
	Age	Age of student	Quantitative	√
	Address	Whether the student lives in urban or rural area	Nominal	√
	Famsize	Student's family size	Nominal	√
	Pstatus	Whether the parents are living together or apart	Nominal	√
	Medu	Mother's education	Quantitative	√
	Fedu	Father's education	Quantitative	√
	Mjob	Mother's job	Nominal	√
	Fjob	Father's job	Nominal	√
	Reason	Reason to choose the school	Nominal	√
	Guardian	Student's guardian	Nominal	√
	Traveltime	Travel time between home and school	Quantitative	√
Behavioral Features (academic)	Studytime	Study time in a week	Quantitative	
	Failures	Number of times student failed in past	Quantitative	
	Schoolsup	Educational support from school	Nominal	
	Famsup	Educational support from family	Nominal	√
	Paid	Extra paid classes	Nominal	√
	Absences	Number of times student was absent	Quantitative	
Behavioral Features (community)	Activites	Extra activities	Nominal	
	Nursery	Attended nursery school	Nominal	
	Higher	If the student wants to pursue higher education	Nominal	
	Internet	If the student has internet at home	Nominal	√
	Romantic	Does the student have a relationship	Nominal	
	Famrel	Family relations quality	Quantitative	√
	Freetime	Student's amount of free time	Quantitative	
	Goout	Going out with friends	Quantitative	√
	Dalc	Alcohol take during weekdays	Quantitative	
	Walc	Alcohol take during weekends	Quantitative	
	Health	Student's health	Quantitative	

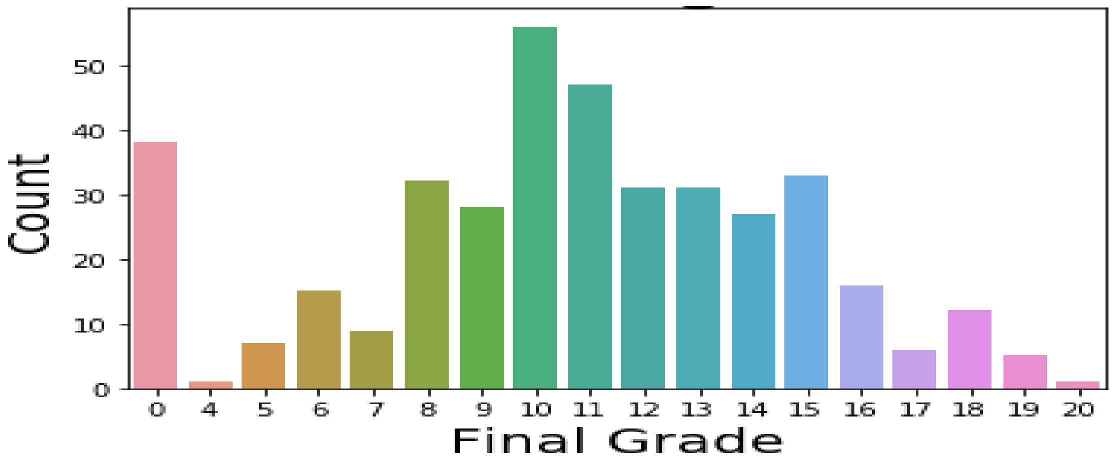


Figure 1: Histogram of students' final grad

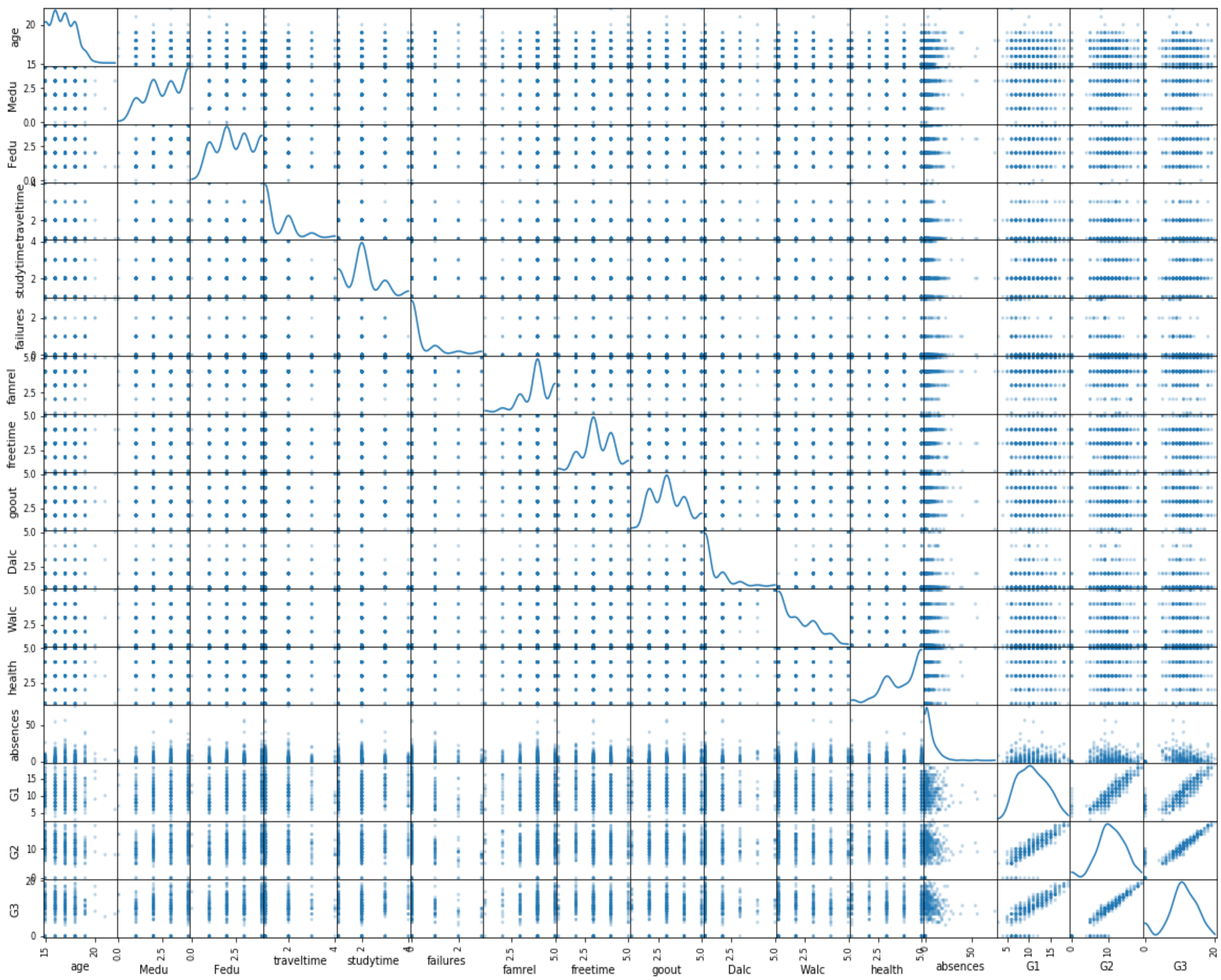


Figure 2: Scatter plot of first data -only between numeric values-

Table 2. Correlation between various features each other.

	age	Medu	Fedu	traveltime	studytime	failures	famrel	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3
age	1	-0.163658	-0.163438	0.0706407	-0.00414004	0.243665	0.0539401	0.0164344	0.126964	0.131125	0.117276	-0.0621874	0.17523	-0.0640815	-0.143474	-0.161579
Medu	-0.163658	1	0.623455	-0.171639	0.0649441	-0.23668	-0.00391446	0.0308909	0.0640944	0.0198341	-0.0471235	-0.0468778	0.100285	0.205341	0.215527	0.217147
Fedu	-0.163438	0.623455	1	-0.158194	-0.00917464	-0.250408	-0.00136973	-0.0128455	0.0431047	0.00238643	-0.012631	0.0147415	0.0244729	0.19027	0.164893	0.152457
traveltime	0.0706407	-0.171639	-0.158194	1	-0.100909	0.0922387	-0.016808	-0.0170249	0.0285397	0.138325	0.134116	0.00750061	-0.0129438	-0.09304	-0.153198	-0.117142
studytime	-0.00414004	0.0649441	-0.00917464	-0.100909	1	-0.173563	0.0397307	-0.143198	-0.0639037	-0.196019	-0.253785	-0.0756159	-0.0627002	0.160612	0.13588	0.0978197
failures	0.243665	-0.23668	-0.250408	0.0922387	-0.173563	1	-0.0443366	0.0919875	0.124561	0.136047	0.141962	0.0658273	0.0637258	-0.354718	-0.355896	-0.360415
famrel	0.0539401	-0.00391446	-0.00136973	-0.016808	0.0397307	-0.0443366	1	0.150701	0.0645684	-0.0775944	-0.113397	0.0940557	-0.0443541	0.0221683	-0.0182813	0.0513634
freetime	0.0164344	0.0308909	-0.0128455	-0.0170249	-0.143198	0.0919875	0.150701	1	0.285019	0.209001	0.147822	0.0757334	-0.0580779	0.0126129	-0.0137771	0.0113072
goout	0.126964	0.0640944	0.0431047	0.0285397	-0.0639037	0.124561	0.0645684	0.285019	1	0.266994	0.420386	-0.00957725	0.0443022	-0.149104	-0.16225	-0.132791
Dalc	0.131125	0.0198341	0.00238643	0.138325	-0.196019	0.136047	-0.0775944	0.209001	0.266994	1	0.647544	0.0771796	0.111908	-0.0941588	-0.0641202	-0.05466
Walc	0.117276	-0.0471235	-0.012631	0.134116	-0.253785	0.141962	-0.113397	0.147822	0.420386	0.647544	1	0.0924763	0.136291	-0.126179	-0.0849274	-0.0519393
health	-0.0621874	-0.0468778	0.0147415	0.00750061	-0.0756159	0.0658273	0.0940557	0.0757334	-0.00957725	0.0771796	0.0924763	1	-0.0299367	-0.0731721	-0.0977199	-0.0613346
absences	0.17523	0.100285	0.0244729	-0.0129438	-0.0627002	0.0637258	-0.0443541	-0.0580779	0.0443022	0.111908	0.136291	-0.0299367	1	-0.0310029	-0.0317767	0.0342473
G1	-0.0640815	0.205341	0.19027	-0.09304	0.160612	-0.354718	0.0221683	0.0126129	-0.149104	-0.0941588	-0.126179	-0.0731721	-0.0310029	1	0.852118	0.801468
G2	-0.143474	0.215527	0.164893	-0.153198	0.13588	-0.355896	-0.0182813	-0.0137771	-0.16225	-0.0641202	-0.0849274	-0.0977199	-0.0317767	0.852118	1	0.904868
G3	-0.161579	0.217147	0.152457	-0.117142	0.0978197	-0.360415	0.0513634	0.0113072	-0.132791	-0.05466	-0.0519393	-0.0613346	0.0342473	0.801468	0.904868	1

Histogram of students' final grad in figure 1 shows that a part from the high number of students scoring 0, but after checking for null values in the data set, we conclude that the distribution is normal as expected.

we can see there are several features for each record that are non-numeric. Typically, learning algorithms expect input to be numeric, which requires that non-numeric features (called *categorical variables*) be converted. G1 and G2 are highly correlated to the final grade G3. Mother's education and father's education as family features are among the most important family factors influencing a student's G3.see figure 2 and table 2.

From the distribution of ages figure 3, in the data we find that the majority between the ages of 15 and 19 who are still under the supervision of their family.

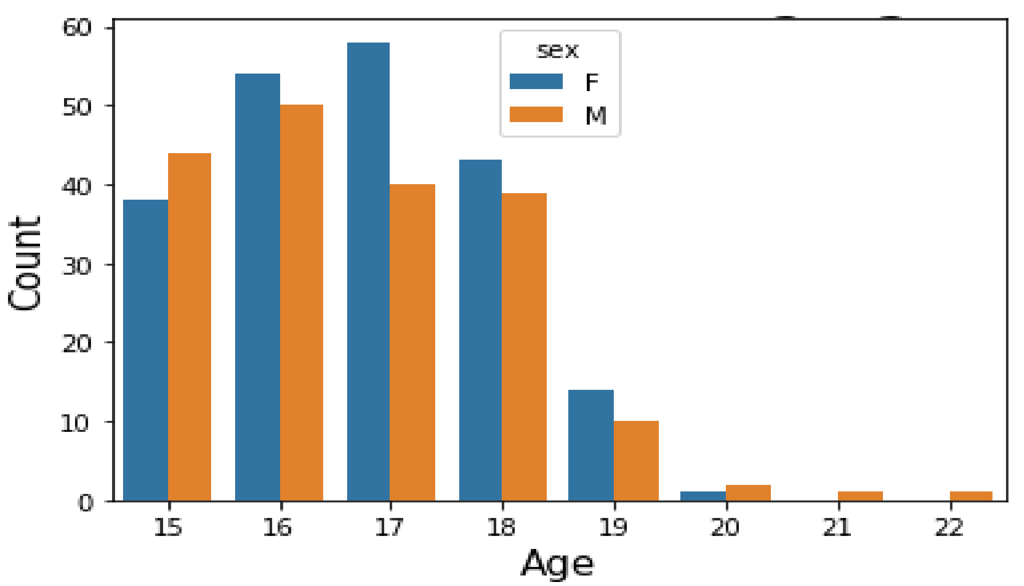


Figure 3: Student age and sex distribution.

2.Exploring and Visualizing Second Data

The second data set was originally used in research made at the University of Jordan. It contains information about 480 students from various countries, mostly in the Middle East. The data has a total of 16 variables. The features are classified into three main categories: (1) Demographic features. (2) Academic background features. (3) Behavioral features. See table 2 (Amrieh et al., 2016)
I will use these features to predicate Class - final grade ("H","L","M" represents "High", "Medium" and "Low" in student's academic performance, with balanced distribution) as target value.

Table 3. Feature descriptions for the second data set.

Features Category	Feature	Description	Type	Family Feature
Demographical Features	Nationality	Student nationality	Nominal	
	Gender	The gender of the student	Nominal	
	Place of Birth	Place of birth for the student	Nominal	
	Parent responsible for student	Student’s parent	Nominal	√
Academic Background Features	Educational Stages (school levels)	Stage student belongs	Nominal	
	Grade Levels	Grade student belongs	Nominal	
	Section ID	Classroom student belongs	Nominal	
	Semester	School year semester	Nominal	
	Topic	Course topic	Nominal	
Parents Participation on learning process	Parent Answering Survey	Parent is answering the surveys that provided from school or not.	Nominal	√
	Parent School Satisfaction	This feature obtains the Degree of parent satisfaction from school	Nominal	√
Behavioral Features	Discussion groups	Student Behavior e-learning system.	Quantitative	
	Visited resources		Quantitative	
	Raised hand on class		Quantitative	
	Viewing announcements		Quantitative	
	Student Absence Days		Quantitative	

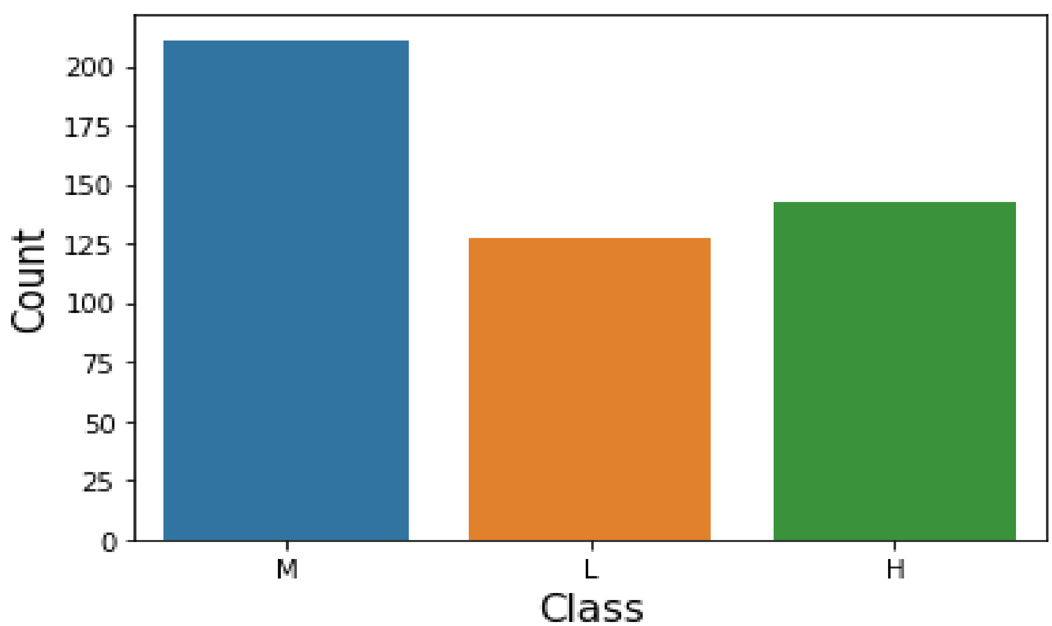


Figure 4: Student Class distribution.

Figure 5: Parent responsible for student.

From this **Figure** we find that the care of the mother of the student more impact on the performance of the student than the care of the father.

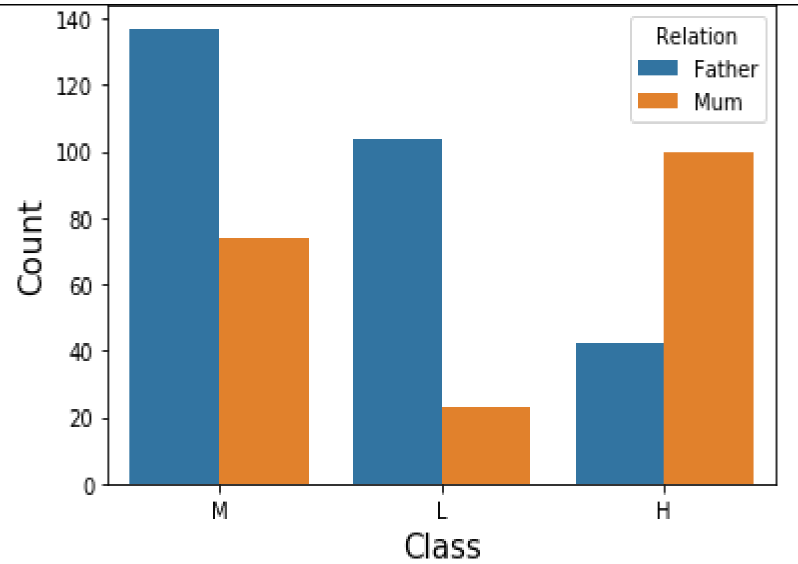


Figure 6: The Degree of parent satisfaction from school

When the degree of parent satisfaction from school is good, this is reflected positively on his performance.

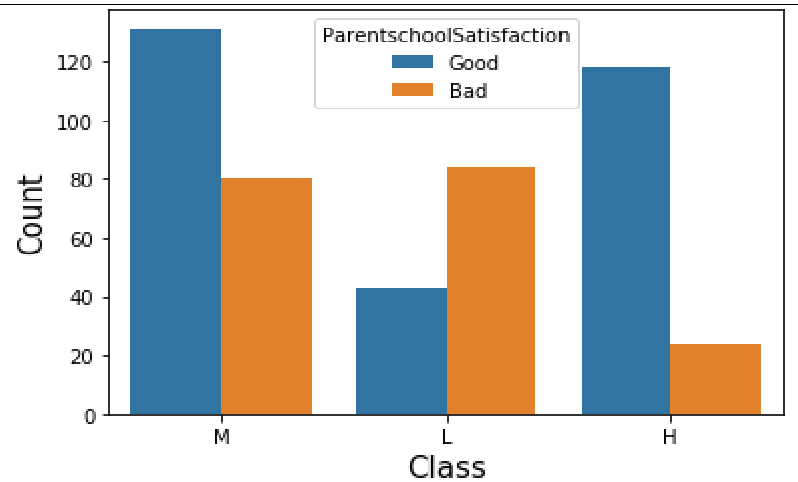
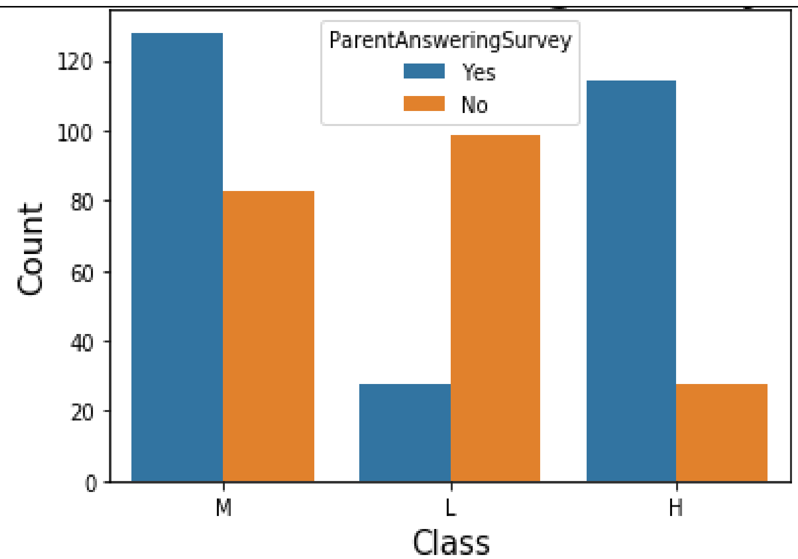


Figure 7: Parent is answering the surveys that provided from school or not.

I note that this distribution is close to the distribution in the figure 6 of parental satisfaction.



Benchmark Model

(Paulo et al.,2008) used first data to addressed the prediction of secondary student grades of two core classes (Mathematics and Portuguese) by using past school grades (first and second periods), demographic, social and other school related data. Four ML methods, i.e. Decision Trees (DT), Random Forests (RF), Neural Networks (NN) and Support Vector Machines (SVM), were tested. See table 3

Table 4. RMSE values of prediction models applied in mathematics dataset

	NV	ANN	SVM	DT	RF
RMSE values	4.59	4.41	4.37	4.46	3.90

For second dataset (Amrieh et al., 2016) measured the impact of behavioral features on student's academic performance using different classification techniques such as (DT, ANN and NB) and show how the accuracy of the proposed model using behavioral features achieved up to 22.1% improvement comparing to the results when removing such features. See table 4

Table 5. Classification Method Results with Behavioral Features (BF) and Results without behavioral features (WBF)

Evaluation Measure	**DT**		**ANN**		**NB**	
Behavioral features existence	**BF**	**WBF**	**BF**	**WBF**	**BF**	**WBF**
Accuracy	75.8	55.6	79.1	57.0	67.7	46.4
Recall	75.8	55.6	79.2	57.1	67.7	46.5
Precision	76.0	56.0	79.1	57.2	67.5	46.8
F-Measure	75.9	55.7	79.1	57.1	67.1	46.4

Algorithms and Techniques

1. Regression

Regression is the supervised learning task for modeling and predicting continuous, numeric variables. Examples include predicting real-estate prices, stock price movements, or student test scores. Regression tasks are characterized by labeled datasets that have a numeric target variable.

For our target (G3) in first dataset we compared between the following supervised regression techniques:
Linear Regression, Elastic Net Regression, Random Forest, Extra Trees, Gradient Boosted, Gaussian NB, SVM and Neural Network

2. Classification
Classification is the supervised learning task for modeling and predicting **categorical** variables. Examples include predicting employee churn, email spam, financial fraud, or student letter grades.

For our target (Class) in second dataset we compared between the following supervised Classification techniques:
Gaussian NB, SVM, Neural Network, Random Forest Gradient Boosting and Logistic Regression.

Methodology

The aim of the project was to compare different machine learning methods with two different datasets in the student performance prediction. Because of difference of our target in these two data sets we apply supervised regression techniques on first data set and apply supervised Classification techniques on first data set. The prediction models were created using the python language.

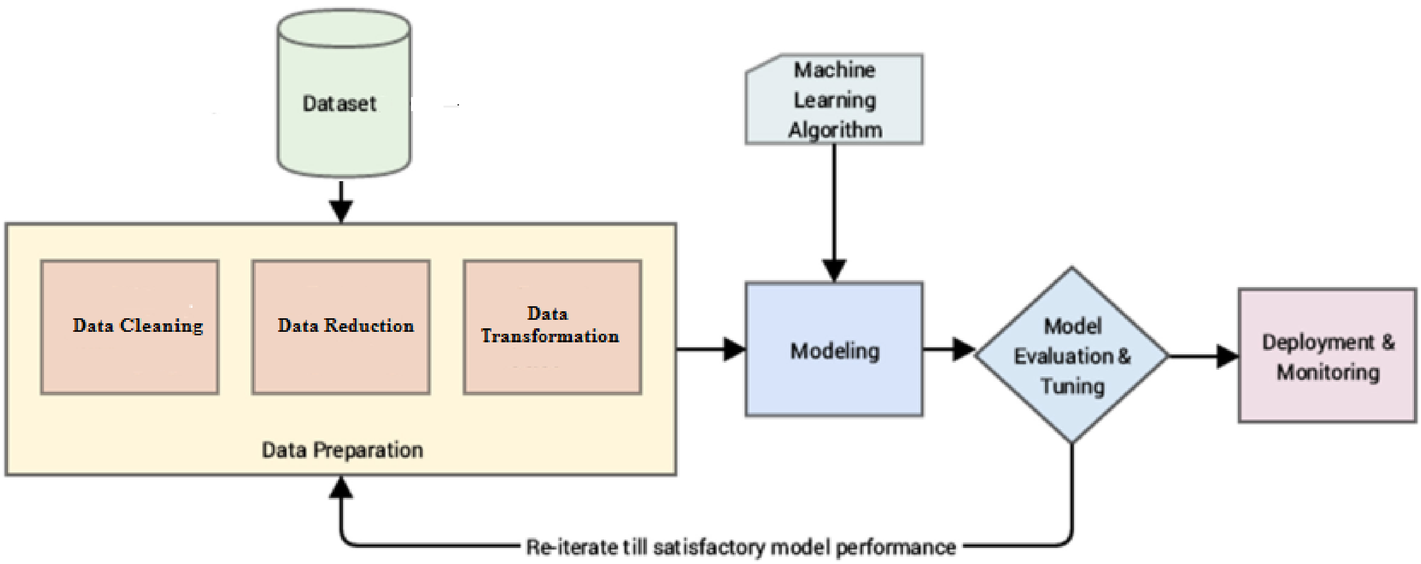


Figure 8: Project work flow

This project can be decomposed into several stages:

1. **Preparation Data**

pre-processing is considered an important step in the knowledge discovery process, which includes

Data Cleaning: removing irrelevant items and missing values

From data exploration in both datasets we didn't found null values or outlier

Data Transformation: converting non-numeric features into numeric

Transformation From the tables in **Exploring the Data** above, we can see there are several features for each record that are non-numeric. Typically, learning algorithms expect input to be numeric, which requires that non-numeric features (called *categorical variables*) be converted. One popular way to convert categorical variables is by using the **one-hot encoding** scheme. One-hot encoding creates a *"dummy"* variable for each possible category of each non-numeric feature.

Feature Selection: select an appropriate subset of features which can efficiently describe the input data

In first data set we apply techniques with all features as input, then compare its performance when we make inputs are family features only.as in table 1

In second data set we apply techniques with all features as input, then compare its performance when we remove family features from inputs.as in table 3

2. **Model Evaluation & Tuning**

In this stage I will evaluate every machine learning algorithm performance with different hyper-parameter to be more efficient according to Evaluation Metrics

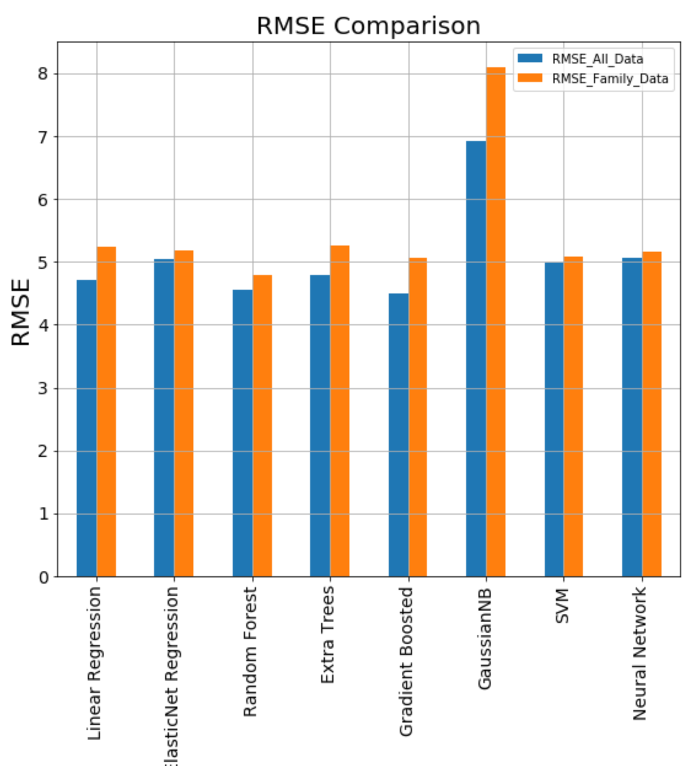
Results of model’s evaluations

1.The first data set

As shown in in the Figure 9 and Table 5, we can see Gradient Boosted and Random Forest are the best in predicting student performance with lower RMSE.
The important notice in this section that is when we depend on family features only RMSE change a little which main that family features are good for predicate student performance.

Table 5. Comparison of different ML techniques for first dataset

	All features		**Family features only**	
	RMSE	**Time Period**	**RMSE**	**Time Period**
Linear Regression	4.712	0.016	5.25	0.05
Elastic Net Regression	5.051	0.002	5.191	0.002
Random Forest	4.553	0.247	4.799	0.21
Extra Trees	4.788	0.264	5.258	0.21
Gradient Boosted	4.5	0.034	5.066	0.028
Gaussian NB	6.92	0.003	8.098	0.003
SVM	4.996	0.009	5.081	0.008
Neural Network	5.063	0.466	5.171	0.391



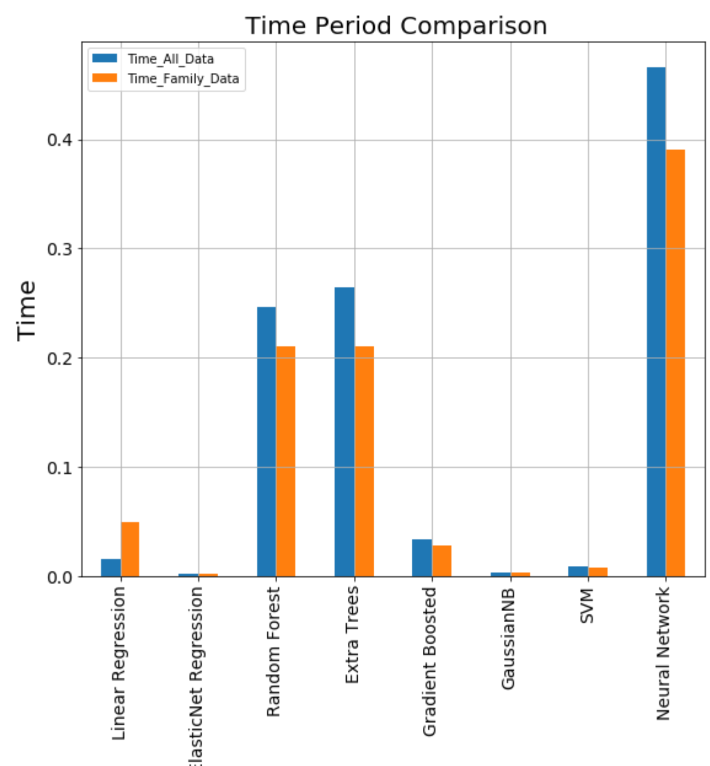


Figure 9: RMSE and Time period comparison

2.The second data set

As shown in in the Table 6 and Figures (10,11 and 12) we can see Gaussian NB has the worst scores in predicting student performance with lower measurements. Gradient Boosted and Random Forest are the best in classified student performance with higher scores.
The important notice in this section that is when we remove family features from original dataset values of Accuracy,F1score, Precision Score and Recall Score for testing data are change down a little but still Gradient Boosted and Random Forest having best evaluations which main that family features in e-learning process can be neglected for classifying student performance.

Table 6. Classification Method Results with All features and Results without family features

	All features						Without family features					
	Acc Train	Acc Test	F1 Score Train	F1 Score Test	Precision Score	Recall Score	Acc Train	Acc Test	F1 Score Train	F1 Score Test	Precision Score	Recall Score
Gaussian NB	0.661	0.583	0.647	0.575	0.603	0.583	0.627	0.479	0.609	0.459	0.491	0.479
SVM	0.841	0.75	0.841	0.747	0.746	0.75	0.812	0.699	0.813	0.69	0.691	0.698
Neural Network	0.940	0.76	0.94	0.755	0.76	0.76	0.906	0.677	0.906	0.672	0.67	0.677
Random Forest	0.994	0.76	0.995	0.756	0.771	0.76	0.987	0.677	0.987	0.669	0.685	0.677
Gradient Boosting	0.994	0.812	0.995	0.81	0.81	0.812	0.98	0.76	0.982	0.755	0.755	0.76
Logistic Regression	0.828	0.687	0.827	0.674	0.674	0.687	0.768	0.646	0.767	0.629	0.626	0.646

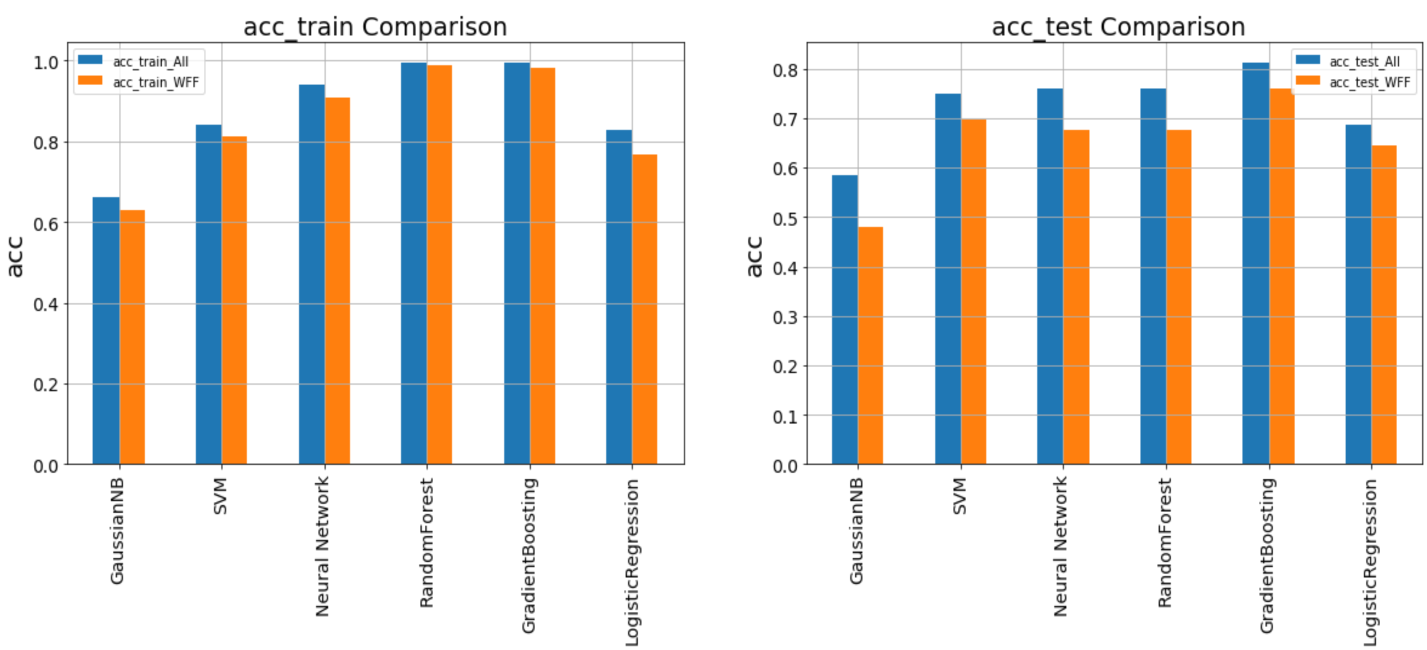


Figure 10: Accuracy for training and testing comparison

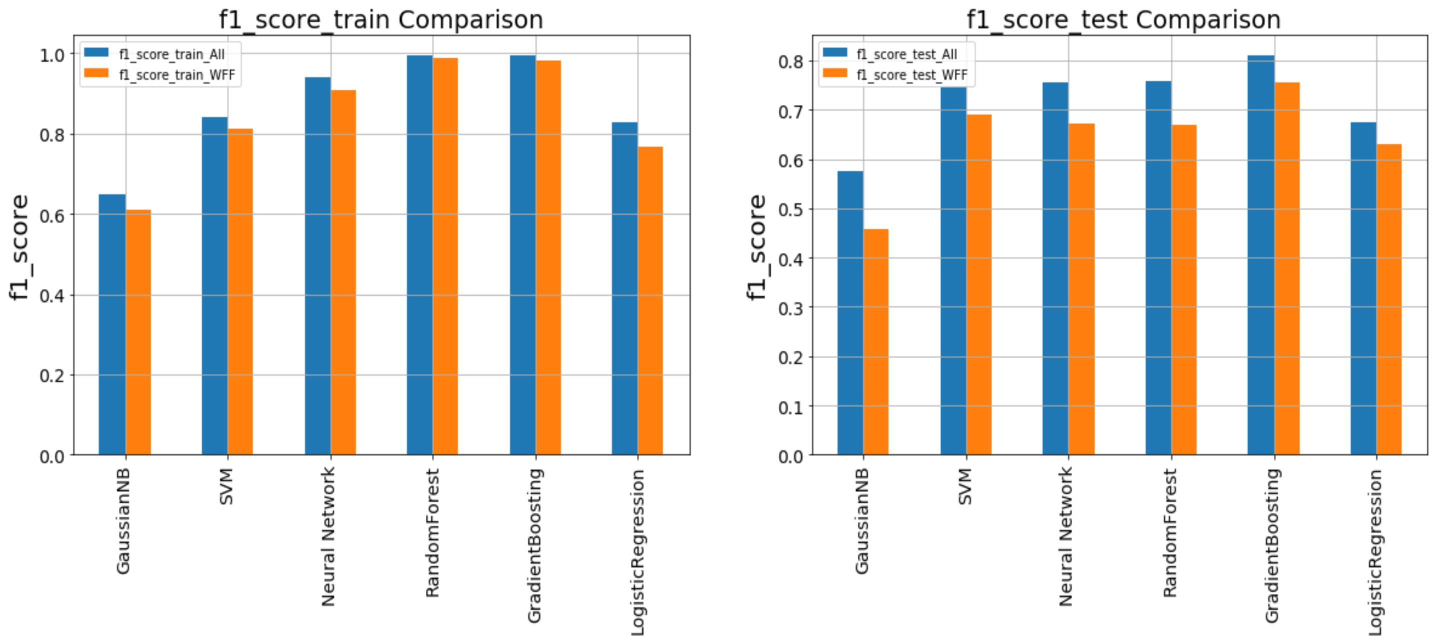


Figure 11: F1 Score for training and testing comparison

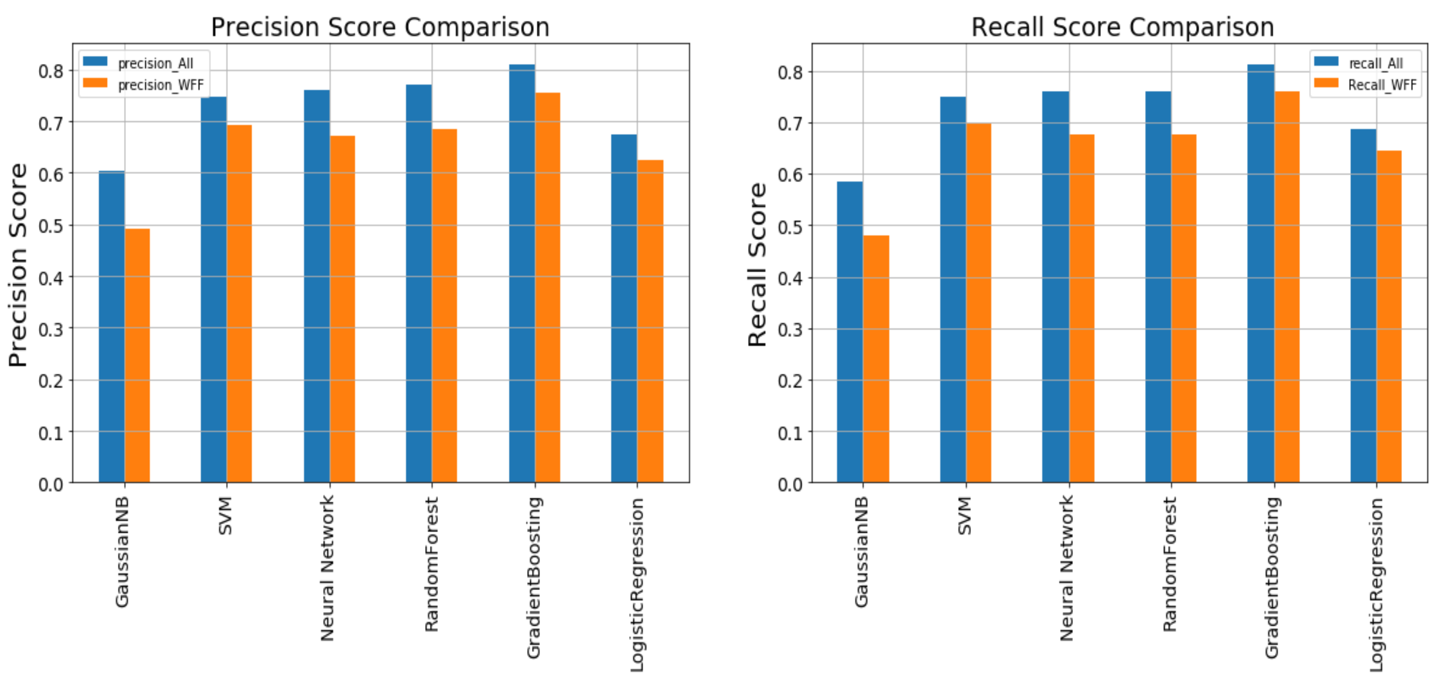


Figure 12: Precision Score and Recall Score comparison

Conclusion

Improving the performance of education has a significant impact on ensuring the nations' economic prosperity and represents a central focus of the government when making education policies. During the last years, machine learning techniques achieve this goal in education by developing methods of exploring data from computational educational settings and discovering meaningful patterns. The aim of this project is to discuss student's performance prediction model based on student's family data. With different two datasets, I applied several machine learning methods on both data sets. the structure of both data sets is different so I used them individually.

In the first data set (secondary school), shows that students with ages 15-19 who are under their family supervision their studies are directly influenced by their families. Although the family variables are poorly correlated with the final estimate, we were able to find a model using Random Forest based on those variables only as inputs to predict the final estimate.

In the second data set (e-learning process), family features in e-learning process can be neglected for classifying student performance. students can rely on themselves to study online without family control

Reflection

What led me to think about this project was to try to contribute to improving the status of education in my country. And to study how a student's family affects the student's educational performance. I have just listed the previous work that we can compare with data within our country. I found two different databases and that's what strengthened my idea for the project to work on two separate databases. The results obtained from them were expected, which will motivate me in the future to work on the application of these models to cases within my country.

Improvement

In this work, for the many variables that exist. if we can collect data with more variables or at least that existed in the previous studies have been integrated to improve the performance of the proposed model I think that would be better.

References:

Amrieh, Elaf Abu, Thair Hamtini, and Ibrahim Aljarah. "Mining educational data to predict student's academic performance using ensemble methods." International Journal of Database Theory and Application 9.8 (2016): 119-136.

Baker, Ryan SJD, and Kalina Yacef. "The state of educational data mining in 2009: A review and future visions." JEDM| Journal of Educational Data Mining 1.1 (2009): 3-17.

Powers, David Martin. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation." journal of Machine Learning Technologies. (2011):2(1), 37-63

Cortez, Paulo, and Alice Maria Gonçalves Silva. "Using data mining to predict secondary school student performance." In: Proceedings of 5th Annual Future Business Technology Conference, Porto, 5-12. (2008).

Data Source:

student-Grade-Prediction Source: **https://www.kaggle.com/dipam7/student-grade-prediction**

Students' Academic Performance Dataset Source: **https://www.kaggle.com/aljarah/xAPI-Edu-Data**