Machine Learning in The Education Process
Capstone Proposal

Project Background

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). The aim of 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 (Amrieh *et al*., 2016).

Problem statement

Predicting student's performance is an important task in educational environments. There are several machine learning methods such as Decision Tree (DT), 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.

Datasets

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.

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
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

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 2. Feature descriptions for the second data set.

Features Category	Feature	Description	Type
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

Our goal will be to predict the final grade of the student based on the available data to us in both datasets. each data set is divided into training validation and testing (70%, 15%and 15%).

Solution Statement

First, different algorithms are applied to a data set to build prediction models. Then, predictions made by these models are compared using common evaluation criteria, such as accuracy, precision, recall and F-measure. With these evaluation measurements I will how student's family data effect on student's performance in education process.

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 3. 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 4. 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

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)

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.

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

F-measure is the final evaluation criteria for comparisons in this project.

Which used in (Amrieh et al., 2016)

Project Design

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

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

   Data Transformation: converting non-numeric features into numeric

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

3. Deployment & Monitoring

   Deployment of machine learning models is the process for making your models available in production environments.

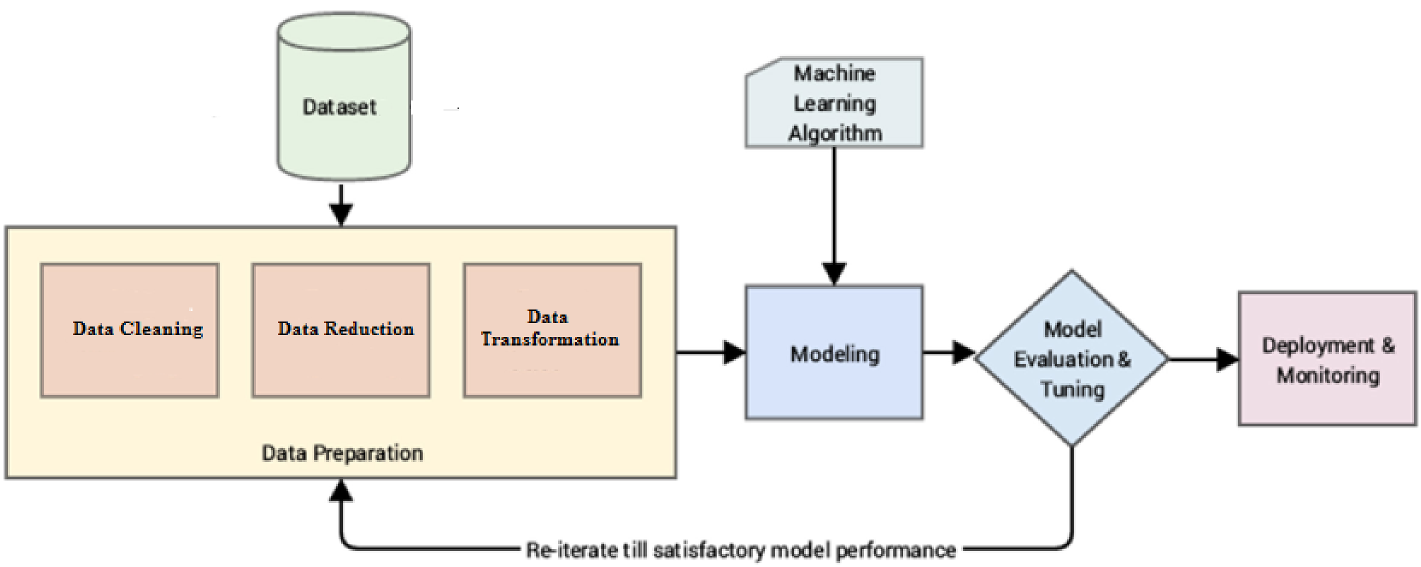

Figure 1: Project work flow

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**