Machine Learning based Students Performance Evaluation for Decision Making of Dropout or Course Continuation

Chih-Yu Wen and Yen-Chieh Ouyang

Abstract---Higher institutions are setup to provide quality education capable of transforming the level of awareness, knowledge, and the capacity of the human mind. In India, the educational sector has suffered many setbacks due to under development, economic hardship, insufficient budget and corruption. The academic performance of engineering students from their first year to the third year is very vital in terms of acquisition of foundational knowledge, and its impact on their final graduation Cumulative Grade Point Average (CGPA). It is often said that beyond the third year it is very challenging for a student to move from the current class of grade. Six data mining algorithms were considered, and a maximum accuracy of 91% was achieved. The result was verified using both linear and pure quadratic regression models. This creates an opportunity for identifying students that may graduate with poor results or may not graduate at all, so that early intervention may be deployed.

Keywords—Academic Performance Indicators, Machine Learning, Dropout Classification.

I. Introduction

IN recent years Educational Data Mining has emerged as a new field of research due to the development of several statistical approaches to explore data in educational context. One such application of EDM is the early prediction of student results. This is necessary in higher education for identifying the weak students so that some form of remediation may be organized for them. The ability to choose the right option depending on the problem it has, is the matter of decision making. Decision making helps the student to be motivated and to choose the right path. When a situation rises, the individual must be able to take a good decision rather than wavering this is accomplished by the complete support of a teacher (mentor).Teacher could determine which option will suit that particular situation.

Many studies have investigated the ways of applying machine learning techniques in the learning field for various educational purposes. One of the focuses of these studies is to identify high-risk students, as well as to identify features which affect the performance of students. A number of secondary school students fails the course in their 10th grade. Based on review of the literature, various reasons were identified for measuring „relative importance‟ of student drop out. These reasons were then broadly grouped in to three basic categories:

? Individual factors
? Schooling factors
? Family factors

The concept of machine learning is used to analyze digital data and to find ways that is too complex for a human to do. The basic idea of machine learning is that a computer cans performance. Depending on the type of input data, machine learning algorithms can be divided into supervised and unsupervised learning. In supervised learning, input data comes with a known class structure [1] this input data is known as training data. In unsupervised learning, input data does not have a known class structure, and the task of the algorithm is to reveal a structure in the data [2]

General applications of Support Vector Machine (SVM) are text categorization such as email classification and web searching, image classification, bioinformatics, hand-written character recognition etc. Support vector machine can do well when with a small set as long as the number of data points are larger than the number of features being considered.

Here we used anaconda python for classification. It is the easiest way to perform Python and machine learning on Windows. Python is a widely used high-level, general-purpose, dynamic programming language on its own, but with the help of a few popular libraries such as Numpy, matplotlib etc.

II. Architecture

Our objective is to build a model that would predict whether or not a student would fail their 10th grade. We focused on failure rates and identified those students who might need early intervention and interact with them using a user friendly chatbot.

Schooling factors	Individual factors	Family factors	Data Collection
Prediction	Support on vector	Classification	Data processing

Fig. 1. Block diagram of student performance prediction system

Manuscript received on August 05, 2019, review completed on August 06, 2019 and revised on August 12, 2019.

Chih-Yu Wen is with the Department of Information Systems Engineering, Near East University, NEU Boulevard, Nicosia, TRNC, Mersin, Turkey.

Yen-Chieh Ouyang is with the Department of Information Systems Engineering, Near East University, NEU Boulevard, Nicosia, TRNC, Mersin, Turkey.

Digital Object Identifier: AIML082019004.

A. *Schooling factors*

Factors such as classrooms, textbooks, equipment, school supplies, and other instructional materials also affect the efficiency of learning in school factors such as attendance percentage, previous failures, extracurricular activities, study time and school support influences classroom level factors. This is an important predictor of school effectiveness. Student academic performance depend on both internal and external features. Schooling factors has a major role in the prediction of student performance. School factors influence the classroom-level factors, especially the teaching practice

B. *Individual factors*

As a unique individual, each individual has their own special profile. It includes things like cultural background, sex, age, address and travel time. Their physical health, mental health and attitude also matters. They play a major role in the behavior of an individual. Hence their studies. Personal factors, such as instincts and emotions are directly related to a complex psychology of motivation. The process of learning has been declared as an individual's effort. Therefore, developing and evaluating learning efforts of a student is not an easy task. Student performance evaluation system can help in targeting the right students. It is the most frequent and the most significant factors related to students‟ academic achievement and motivation for learning. Grouping all these individual factors we lead to self-efficacy of a student. Self-efficacy is how people feel about themselves and how much they like themselves and that leads to a bright future.

C. *Family factors*

Family-related factors play a critical role in a student's academic performance. Family factors can affect a child's behavior and his or her ability to perform in the classroom. These include parent's attitude towards education, their economic stability because children from lower income experience lack of consistency and supervision. Family size, parent's status, mother's education, father's education, mother's job, father's job and family support. When a child fails or misbehaves to meet expectations at school, the child's home and family life should be considered. It includes changes in family relationships, parental attitudes toward education and economic stability. Family relationship such as divorce has long been linked to behavior problems, anxiety and depression in children. Children from lower-income homes often experience a lack of supervision, poor nutrition and poor role-modeling.

D. *Data Collection*

This is an educational data set which is collected from a Secondary school. The dataset consists of 791 student records and 18features. The features are classified into three major categories: (1) Individual factor (2) Schooling factors and (3) Family factors.

TABLE 1
DATA DESCRIPTION

Features	Data Vectors Sex
Male-0 Female-1 Age	15yrs-1,16yrs-2,17yrs- 3,18yrs-4,19yrs-5
Address FamSize Pstatus Medu, Fedu Mjob, Fjob Travel time Study time Failures Famtype Famsup Paid Activities Attendance percentage Internet passed	Urban-1 Rural-0 ET3-0, GT3-1, LT3-2 Together-0 Apart-1 None-0, primary education (4th grade)-1, 5th to 9th grade-2, secondary education -3 or higher education-4 None-0, primary education (4th grade)-1, 5th to 9th grade-2, secondary education -3 or higher education-4 <15 min -1, 15 to 30 min -2, 30 min to 1 hour-3, or >1 hour -4 <2 hours -1, 2 to 5 hours -2, 5 to 10 hours -3, or >10 hours -4 n if 1<=n<3, else 4 Joint-1,Neutral-0 Yes-1,No-0 Yes-1,No-0 Yes-1,No-0 Greater than 75%-1,Less than 75%-0 LT1 hr-0, 1 hr-1, 2 hr-2, 3 hr-3, 4 hr-4,GT4 hr-5 Yes-1,No-0

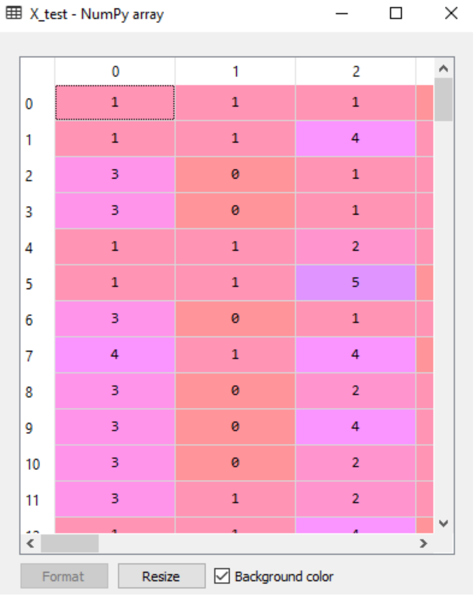
The dataset consists of 374 males and 416 females. The students come from different origins including Urban and Rural areas. The data of students such as previous failures is also collected from their previous class. The data set includes the school absence feature, features which show their age, address, family size, father's and mother's education, their job, financial status, travel time, study time, family support, access to internet etc.

E. *Data preprocessing*

In this section, the data is prepared for modelling, training and testing. There are several ways of data portioning in Machine learning. Training/test partitioning or cross-validation are the most popular methods. Testing the parameters of a prediction function and testing it on the same data causes over fitting. Repeating the labels of the samples might give a perfect score but would fail to predict anything useful on yet-unseen data. Hence to avoid it, it is a common practice when performing a (supervised) machine learning experiment to hold out part of the available data as a test set X_test, y_test[3].

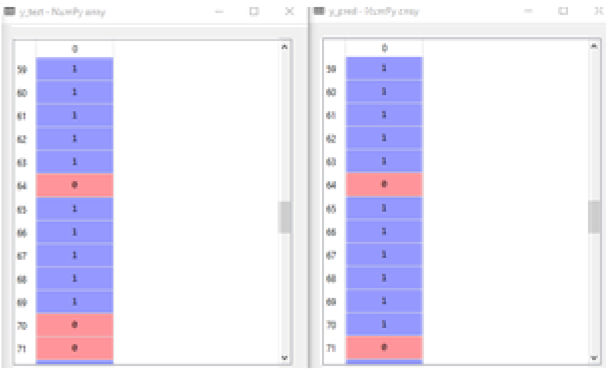
As most machine learning algorithms expect numeric data to perform computations we changed columns with non-numeric features to numeric features. Many of them were converted into 1/0 (binary) values. Whereas other columns, like Mjob and Fjob, have more than two values. Such variables are known as *categorical variables.* One way to handle such a column is to assign numbers to each of the categorical variable. First, convert all *categorical* features into numeric values and split the data into training and testing set.

We used 600 training points (approximately 75%) and 190 testing points (approximately 25%).



	0	1	2
0	1	1	1
1	1	1	4
2	3	0	1
3	3	0	1
4	1	1	2
5	1	1	5
6	3	0	1
7	4	1	4
8	3	0	2
9	3	0	4
10	3	0	2
11	3	1	2

Fig .2. Test set of x



	0
59	1
60	1
61	1
62	1
63	1
64	0
65	1
66	1
67	1
68	1
69	1
70	0
71	0

	0
59	1
60	1
61	1
62	1
63	1
64	0
65	1
66	1
67	1
68	1
69	1
70	1
71	0

Fig.3. Test set of y and predicted set of y

F. Classification

This is a classification problem because the output that we are looking for will be used to determine whether a given student will fail or continue the course. Many machine learning algorithms can be used for the process of classification and regression. In the proposed method, Comparison was made between the algorithms like Support Vector Machine (SVM) and K –nearest neighbors using the scikit tool. Based on the analysis carried out Multiclass SVM showed prominent accuracy. The Multiclass SVM is implemented on the basis of one-against rest strategy use of class labels which is primarily an extension of linear SVM. The reason for selecting SVM over k-nearest neighbor is that it can do well with a small set as long as the number of data points is larger than the number of features being considered. SVM can avoid localization problems and can uses the kernel trick while using K- nearest neighbor algorithm results may change over time as the algorithm is query based. These are considered as disadvantages for K- nearest neighbor and therefore Support Vector Machine is selected for further classification and prediction.

Support Vector Machines can be used for Classification (labeling) or Regression (numeric) predications. Since we are trying to find out if an intervention is necessary or not, we will make use of specific type of SVMs called Support Vector Classifier (SVC) to get the results we are looking for. SVM takes information about past students (age, gender, family, etc.), and use that information to create predictions about new students. These predictions are made by creating a function that draws a boundary between the students who graduated and those who did not. The boundary should be drawn so as to maximize the space between itself and each of the classifications (graduation results), this space is called a margin. We are classifying the sample data set into two major classifications either (“yes” or “no”) or (“1” and “0”). So when the dimension of the objects (students) was increased the SVM would go for Maximum Marginal Hyperplane (MMH). [4] There are infinite number of separation lines among them, we choose the best separator with minimum classification error.

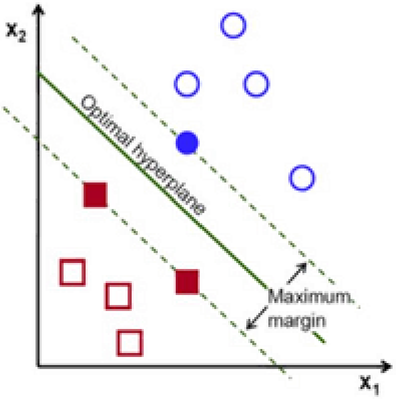
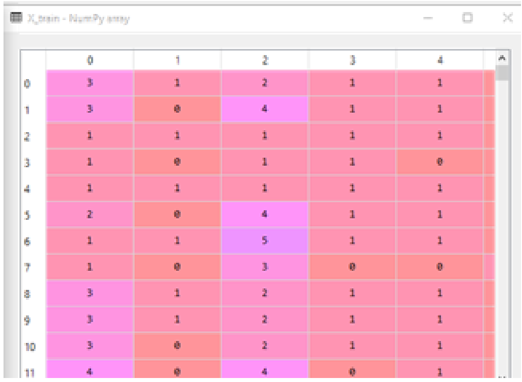


Fig. 4. Support Vector Machine

SVM separates the passing and failing students by turning the dimensions working with from a plane into a higher dimension. Once the SVM changes its way of looking at the data, it uses a plane to separate the data instead of just a line.

We use the tool scikit learn for classification .It is an open source library used to perform machine learning in Python and provides a range of supervised and unsupervised learning algorithms in Python. This tool provides algorithms and libraries. Apart from that, it also contains packages like NumPy, Matplotlib, panda etc. We have to import this packages to implement them.

Usually, support vector machine is a global solution for classification. But other approaches like neural networks, K –nearest neighbor, AdaBoost classifier produce local minima in their solution. Another important advantage of support vector machine is that it is well-suited for multi-class case, where the classification has to obtain the result with more than two classes.



X_train - NumPy array

	0	1	2	3	4
0	3	1	2	1	1
1	3	0	4	1	1
2	1	1	1	1	1
3	1	0	1	1	0
4	1	1	1	1	1
5	2	0	4	1	1
6	1	1	5	1	1
7	1	0	3	0	0
8	3	1	2	1	1
9	3	1	2	1	1
10	3	0	2	1	1
11	4	0	4	0	1

Fig. 5. Training set of x

G. *Prediction*

After considering all the features of a student we predict whether the student will pass or not. The SVM takes data about students such as age, family, gender etc. and uses them to create predictions about new student case. These predictions are made by creating a function that draws a boundary between the students who graduated and those who did not. The boundary should be drawn so as to maximize the space between itself and each of the graduation results. SVM separates the passing and failing students by turning the dimensions we're working with from a plane into a higher dimension such as a cube. Once the SVM changes its way of looking at the data, it can then use a plane to separate the data instead of just a line. SVM algorithm predicts if a student will graduate or not with an accuracy of 80% or more**.**

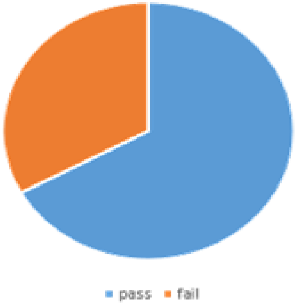


Fig.6. Graphical representation of pass percentage of students

III. Tools Used

The aim of the research was to apply machine learning methods and feature engineering in the student performance prediction. The prediction model was created using the Python language.. The code written in python language is run on an application called Anaconda Spyder. Python is a widely used high-level, general-purpose, dynamic programming language on its own, but with the help of a few popular libraries such as Numpy, matplotlib etc. NumPy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python. It contains various features including these important ones

? A powerful N-dimensional array object

? Sophisticated (broadcasting) functions

? Tools for integrating C/C++ and Fortran code

? Useful linear algebra, Fourier transform, and random number capabilities

Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined using Numpy which allows Numpy to seamlessly and speedily integrate with a wide variety of databases. [5]

Pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the python programming language [6].

IV. Decision Making and Interactive Website

Decision making helps the student to be motivated and to choose the right path. When a situation rises, the individual must be able to take a good decision rather than wavering this is accomplished by the complete support of a teacher (mentor). The ability to choose the right option depending on the problem it has, is the matter of decision making. Mentor must be able to identify the problem in the right time to improve the student performance by taking appropriate steps at right time to improve the quality of learning and to reduce failing ratio. This is achieved using a chatbot where the teacher and student can interact freely and friendly. This helps to identify students who might need early intervention before they fail to graduate and interact with them using a chatbot. This is achieved using a website called “Mentoring Minds”. There will be a student and teacher login. Using these logins teacher and student can interact with each other. Individual, school and family factors that affect the student’s academic performance are collected. Using these data we will be able to predict the problems that can be faced by the respective students. When a student logins to his/her account a pop up message will be shown which indicates the question (i.e. the problem) faced by that student. The student can answer (yes/no) to the question. The answers will be sent back to the respective class teacher (mentor). Mentor notices the problem and takes appropriate steps.

V. Evaluation And Results

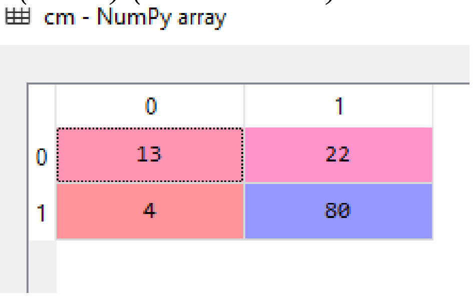
In order to evaluate the effectiveness of a prediction model, predicted values must be compared with actual values. There are multiple criteria for prediction effectiveness Table 2 shows the possible results of prediction for binary values.

Table 2
Possible Prediction Results

	Predicted as True	Predicted as False
Actually True	True Positive	False Negative
Actually False	False Positive	True Negative

The matrix shown in Table 2 is called a confusion matrix which shows the possible prediction results. There are different evaluation criteria that can be obtained from these values. One is accuracy, defined as (Powers, 2011):

Accuracy= (TP+TN) /(TP+TN+FP+FN)



	0	1
0	13	22
1	4	80

Fig .7. Confusion matrix

Fig 7 shows the confusion matrix of our prediction model. The accuracy of our model varies between 70 and 80.

VI. CONCLUSION

The performance of students in higher education in India isa turning point in the academics for all students for their brightest career, by predicting student at risk and give them better training to improve their performance will surely be beneficial for their individual results and also for academic institution profile. However, it can be concluded that our methodology can be used to help students and teachers to improve the student performance by taking appropriate steps at right time to improve the quality of learning and to reduce failing ratio.

REFERENCES

[1] Mohri et al., 2012; Mitchell, 1997

[2] Sugiyama, 2015; Mitchell, 1997

[3] GithubAvailable:https://scipy-lectures.org/packages/scikit-learn/index.html

[4] International Journal of Mechanical Engineering and Technology (IJMET) Volume 8, Issue 11, November pp. 649–662, Article ID: IJMET_08_11_066

[5] https://www.geeksforgeeks.org/numpy-in-python-set-1-introduction

[6] https://pandas.pydata.org/pandas-docs/stable/

[7] International Journal of Computer Applications (0975– Volume 107 – No. 1, December 2014

[8] International Journal of Computer Applications (0975–8887) Volume 107 – No. 1, December 2014

[9] Oloruntoba S.A* et al., 6(12): December, 2017

[10] Ankita Katare and Shubha Dubey2017, "A Study of various Techniques for Predicting student Performance under Educational Data Mining " International Journal of Electrical, Electronics ISSN No. (Online): 2277-2626 and Computer Engineering 6(1): 24-28(2017)

[11] G. Gray, C. McGuinness, P. Owende, 2014,"An application of classification models to predict learner progression in tertiary education", in: Advance Computing Conference (IACC), 2014 IEEE International, IEEE, 2014, pp. 549–554

[12] Ramaswami, M., Bhaskaran, R., 2010, "A CHAID Based Performance Prediction Model in Educational Data Mining", International Journal of Computer Science Issues, Vol. 7, Issue 1, No. 1.