National College of Ireland

Project Submission Sheet – 2020/2021

School of Computing

Student Name:	KENECHUKWU OTITO AJUFO		
Student ID:	x19190174		
Programme:	MSc Data Analytics	**Year:**	2020
Module:	DOMAIN APPLICATION OF PREDICTIVE ANALYTICS		
Lecturer:	VIKAS SAHNI		
Submission Due Date:	SUNDAY 23rd AUGUST, 2020		
Project Title:	**Predicting Student Academic Performance in Portugal using Machine Learning**		
Word Count:	3465		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are encouraged to use the Harvard Referencing Standard supplied by the Library. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action. Students may be required to undergo a viva (oral examination) if there is suspicion about the validity of their submitted work.

Signature: KOAjufo

Date:19/08/2020

PLEASE READ THE FOLLOWING INSTRUCTIONS:

1. Please attach a completed copy of this sheet to each project (including multiple copies).
2. **You must ensure that you retain a HARD COPY of ALL projects**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting Student Academic Performance in Portugal using Machine Learning

Kenechukwu Otito Ajufo
School of Computing
National College of Ireland
Dublin, Ireland
x19190174

***Abstract*—Education is a necessity for every human being. It helps to mold functional individuals for the society. Educational Data Mining (EDM) is centred on the application of scientific and data-driven methods such as Machine Learning to process and analyze big educational data. It provides appropriate information for stakeholders to make decisions which can be used to improve the educational experience. The objective of this research study was to identify the factors which can predict student academic performance in two secondary schools in Portugal. This was accomplished by using a C5.0 Decision Tree model to perform a binary classification predicting the student's likelihood of passing. The model's performance was measured using metrics like accuracy, precision, and recall. The results showed that the student's prior performance is crucial in predicting the student's performance in a subsequent examination. The model can, therefore, be used to identify students who are at risk of failing.**

***Index Terms*—educational data mining, machine learning, c5.0 decision tree, accuracy, precision, recall**

I. Introduction

Educational Data Mining (EDM) is a field concerned with the use of data mining, machine learning and statistics to understand, manage, exploit and use different types of educational data. The use of EDM in solving research problems in the Education field is shown in [1]. It discusses how EDM assists in understanding how learning occurs in educational institutions, in predicting failure or success of students, and in establishing the factors that are responsible for performance of students.

It is worthy of note that Learning Analytics (LA) and Educational Data Mining (EDM), two relatively new and increasingly popular fields of research concerned with the collection, analysis, and interpretation of educational data, intersect as both research communities share an interest in improving the educational practice and make use of data-inventive methods on education data [2] given the overlaps in their research interests, goals and approaches.

In [3], the authors discuss the five fundamental differences between the two fields and concluded that the major differences between the two is that while EDM has a considerably greater focus on automated discovery, LA focuses greatly on human judgement as a central point.

This research work attempts to use machine learning to predict student academic performance in Portugal. A C5.0 Decision tree is used to predict and classify the factors which can be used in predicting academic performance.

The remainder of this report is structured as follows: Part II will present related works; it will examine the applicable techniques for achieving the research objective. Part III will discuss the proposed methodology to be used in the work. Part IV will show the results and key findings, and finally Section V will conclude and describe future areas which can be explored.

II. Literature Survey

Educational Data Mining and Learning Analytics are emerging fields. They use the power of Data Mining to unearth and generate knowledge, and make recommendations for improvements or changes in the Education domain based on the inferences from such knowledge.

Researchers in the Education domain have used numerous machine learning techniques to predict student academic performance. In this chapter the current literature on the baseline technique that would be employed to analyze the data in our study is reviewed.

Three learning approaches were deployed in [4] to model students' Math and Portuguese performance in a secondary school in Portugal. In the work a comparative analysis of five machine learning techniques was done using Regression, a binary classification (Pass/Fail), and a 5-Level Classification based on the *EuRopean Community Action Scheme for the Mobility of University Students* (ERAMUS) grading system. The result obtained showed that students' past performance can and does accurately predict their performance.

In [5], five predictive classification models were built to predict the prospects of students on scholarships finishing their degree program in Pakistan. Features and characteristics of the students' families such as were collected and categorized. The aim was to ascertain the influence of these features on the student's academic performance. F1 Measure was used to evaluate the models and Support Vector Machines model was the best performer. The results revealed that Family Expenditure and Personal Information were the most influential characteristics in determining students' performance.

Researchers in [6] examined the predictive factors which cause students to excel in their English exit examinations. They described the difficulties the universities in non-native

English-speaking countries face in trying to boost their students' English speaking abilities. A C4.5 Decision Tree model was used to classify the data. The classification model produced an outcome that showed the most important determinant for a student being successful in the English exit examination is his or her performance in the English placement test.

The authors in [7] centred their work on building classification models for teachers to use in predicting the odds of students qualifying for the Licensure Examination. Of the various measures used to evaluate the models - Accuracy, F1 Score and Area Under the Curve, e.t.c. -the C4.5 Decision Tree model was reported to be the best model, with the Neural Network model following closely.

Classification models were utilized in [8] to predict the likelihood of first-year Economics majors of a University passing a module in the first-year curriculum. The data consisted of socioeconomic factors and results obtained in high school. The study revealed that the Naïve Bayes classifier outperformed the C4.5 Decision Trees and Neural Network. They, therefore, proposed it as the best classifier that teachers can use in improving students' academic performance.

Using feature selection, the authors in [9] attempted to build several machine learning models to predict the pass rates of students who engage in online education. The models were evaluated by comparing the recall, precision and F1 scores. The Decision Tree model, according to the report, is the best method in predicting student pass rate.

The work in [10] focuses on building predictive classification models to predict high school students' performance. The purpose here is to assist educators in the early identification of students who are at risk of failing. Support Vector Machines (SVM), K Nearest Neighbors (kNN) and Random Forest classifiers models were created. The comparative analysis revealed the SVM model surpassed the other two models.

To classify and predict student performance in [11], the authors explored two datasets. They conducted several experiments with the datasets utilizing four machine learning techniques and showed the best model to be one with the lowest Mean Square, and highest R square values.

The researchers in [12] proposed a predictive academic performance framework following the investigation of the impacts of students' personality traits on students' performance. This framework was compared with different machine learning algorithms like Decision Trees, Random Forest. They showed that the framework outperformed the machine learning algorithms with the highest Spearman Correlation values.

An approach to predict student performance is presented in [13]. Data collected from sixty-five university students using white box machine learning classification techniques was utilized in building models to predict their performance. The Partial Decision Tree model had the best prediction accuracy, followed closely by the C4.5 Decision Tree.

The use of Emotional Quotient (EQ) and Intelligent Quotient (IQ) to predict student academic performance is proposed in [14]. They argue that both the EQ and IQ vary in humans and have an influence in determining academic performance. The results showed that Logistic Function is the best approach in predicting students' performance, and they concluded that EQ and IQ have equal significance in predicting same.

The study in [15] discussed the analysis conducted to predict student performance using supervised machine learning algorithms. The analysis was validated by comparing the Recall, Precision and F-Measure. J48 Decision Tree was found to be the best performing algorithm.

The authors in [16] offered a method of predicting student performance using K-Means Clustering Algorithm on the dataset comprising computer science students' grades. They suggested the use of the clustering algorithm as a standard to observe the students' performance. The importance of data mining techniques such as decision trees in predicting student academic performance was elucidated in [17]. A J48 decision tree was fitted, and the results obtained showed that poor performance in prior tests will lead to failure in the final examination.

A Feedforward Neural Network was created to predict student academic performance in [18]. The students' demographic profile, activities and grades were used as inputs to the network. The model yielded a low accuracy and the researchers attributed this to the size of the sample as Deep Learning performs better with large sample sizes.

In [19] to leverage the power that deep learning has over unlabeled data, Deep Learning techniques were applied to predict student performance. The report lamented the lack of a benchmark for prediction in the educational domain, and compared their proposed model with three classification models. Their model obtained the best accuracy.

In conclusion, after careful review of the related work, to achieve the research objective C5.0 Decision Trees will be implemented. This is because of the prevalence of Decision Trees in the literature and its overall performance against other algorithms in the literature.

III. Data Mining Methodology

For this research, the Knowledge Discovery in Database (KDD) approach is employed. Knowledge discovery in databases (KDD) is the process of unearthing useful knowledge from a collection of data.

This process includes preparation and selection of data, data cleaning etc. This research was conducted in the following stages:

A. Data Selection

The data used for this research is from the University of California Irvine (UCI) Machine Learning Repository. It is available at [20] and contains information on the academic performance of secondary students of two schools in Portugal. The data set has 33 attributes and 649 instances. Table I gives a description of some of the attributes in the data set.

TABLE I
DESCRIPTION OF SOME OF THE ATTRIBUTES IN THE DATA SET

Attribute Name	Attribute Type	Description and Value
school	Binary	Student's school: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira
sex	Binary	Student's Sex: Male or Female
age	Numeric	Student's Age: 15-22 yo
address	Binary	Student's Address: 'U' - Urban or 'R' - Rural
famsize	Binary	Student's Family Size: 'LE3' - less or equal to 3 or 'GT3' - greater than 3
famrel	Numeric	Quality of family relationships: 1 - very bad to 5 - excellent
pstatus	Binary	Student's parent's cohabitation status: Binary: 'T' - Living together or 'A' - Apart
mjob	Nominal	Student's mother's job: teacher, health, civil services, at home or other
fjob	Nominal	Student's father's job: teacher, health, civil services, at home or other
traveltime	Numeric	Home to school travel time: 1 - $<$ 15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 -> 1 hour
studytime	Numeric	Student's weekly study time: 1 - $<$ 2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - $>$ 10 hours
goout	Numeric	Student going out with friends: 1 - very low to 5 - very high
freetime	Numeric	Student's free time after school: 1 - very low to 5 - very high
dalc	Numeric	workday alcohol consumption: 1 - very low to 5 - very high
higher	Binary	Student's intention to get higher education: yes or no
internet	Binary	Access to internet at home: yes or no
health	Numeric	Student's current health status: from 1 - very bad to 5 - very good
activities	Binary	Student's Extra-curricular activities: yes or no
G1	Numeric	First Period Grade: 0-20
G2	Numeric	Second Period Grade: 0-20
G3	Numeric	Final Grade: 0-20

B. Data Preparation, Processing and Transformation

At the stage, the data is prepared for analysis. First, the data set is checked for null values. There were no null values in the data set hence no need for data cleaning.

Following this, exploratory data analysis is conducted to reveal insights in the data set, details are in the Project Design document.

The data set is then split into two, 80 percent for training and 20 percent for test and validation. In addition, the target variable, 'G3' i.e. Final Grade is transformed to satisfy the requirement for a Binary classification. The data transformation is done according to [4], if G3 is greater than or equal to 10, the student passed. The data transformation is done programmatically using R.

C. Data Mining

The data mining technique selected for the analysis is the C5.0 Decision Tree, this is based on the literature review done. The Accuracy, Precision, and Recall are used to evaluate the chosen technique. C5.0 decision trees are based on C4.5 decision trees, having all the functionalities but have more memory efficiency with the introduction of boosting [21].

D. Experiment Settings

This section is used to describe the configuration of the system the experiment was executed on. The details are in the table below.

TABLE II
EXPERIMENT ENVIRONMENT

Operating System	Windows 10
Memory	8 GB
CPU	Intel® Core™ i5-10210U processor
Software	R & R Studio

IV. RESULTS

In this section the results from the model will be discussed extensively. One of the objectives of this research was to discover what factors can predict student academic performance, the summary of the trained C5.0 Decision tree model contains the factors that are most important for predicting the student academic performance and their predictive power. Details are listed in Table V

A. Discussion and Interpretation

The decision tree plot is shown in the Figure 1 below. It shows that the data set is split into two child nodes at the parent node. If ‘G2’ i.e. second period grade is greater than 9 and ’G1’ is greater than 10 the student passes, if ’G2’ is less than or equal to 7 the student fails, the tree is split further into child nodes till it gets to the leaves nodes. Details of the are shown in Figure 2

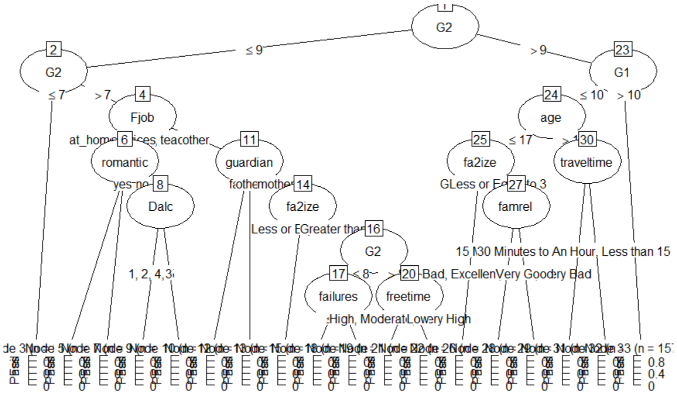
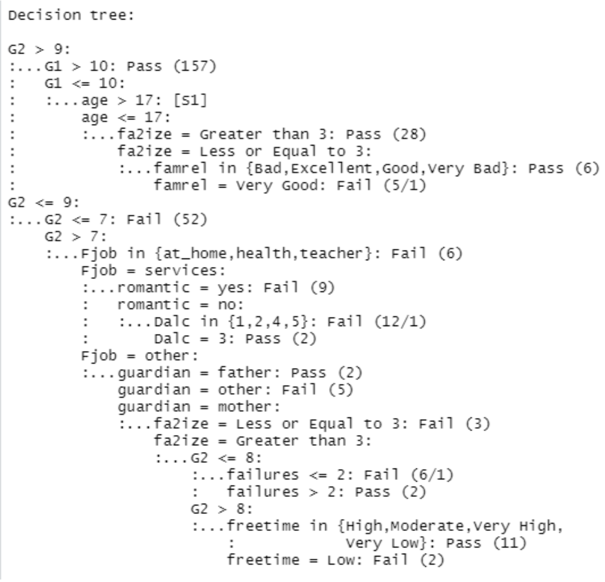


Fig. 1. C5.0 Decision Tree



```
Decision tree:

G2 > 9:
:...G1 > 10: Pass (157)
:   G1 <= 10:
:   :...age > 17: [S1]
:       age <= 17:
:       :...fa2ize = Greater than 3: Pass (28)
:           fa2ize = Less or Equal to 3:
:           :...famrel in {Bad,Excellent,Good,Very Bad}: Pass (6)
:               famrel = Very Good: Fail (5/1)
G2 <= 9:
:...G2 <= 7: Fail (52)
    G2 > 7:
    :...Fjob in {at_home,health,teacher}: Fail (6)
        Fjob = services:
        :...romantic = yes: Fail (9)
        :   romantic = no:
        :   :...Dalc in {1,2,4,5}: Fail (12/1)
        :       Dalc = 3: Pass (2)
        Fjob = other:
        :...guardian = father: Pass (2)
            guardian = other: Fail (5)
            guardian = mother:
            :...fa2ize = Less or Equal to 3: Fail (3)
                fa2ize = Greater than 3:
                :...G2 <= 8:
                    :...failures <= 2: Fail (6/1)
                    :   failures > 2: Pass (2)
                    G2 > 8:
                    :...freetime in {High,Moderate,Very High,
                        :               Very Low}: Pass (11)
                        freetime = Low: Fail (2)
```

Fig. 2. C5.0 Decision Tree Breakdown

The performance of the C5.0 Decision Tree algorithm on the test data set is shown in the confusion matrix in Table III.

Table IV shows a summary of the model’s metrics. The model has an overall accuracy of 90%. This means that the model can predict a student passing the subsequent examination 90 percent of the time. However, this is not quite optimal, a deeper understanding of the model’s performance can be gained from the Precision and Recall of the model.

The Precision of a model is a measure of the quality of the predictions the model makes. It quantifies the amount of correct predictions that are actually true. In simpler terms, the model precision shows how often the model is correct, i.e., the percentage of times the model predicts the student passes correctly. The precision of the model is 86%.

On the other hand, the Recall of a model measures the quality of the model in respect of the mistakes the model makes. This is so important to identify because false claims can be very costly to a business venture. A false claim in this case will be predicting a student passing and then the student fails. The recall of the model is 86%.

TABLE III
C5.0 Decision Tree Model Confusion Matrix

	Actual Fail	Actual Pass
Predicted Fail	25	4
Predicted Pass	4	46

TABLE IV
C5.0 Decision Tree Model Metrics

Accuracy	Precision	Recall
0.8987	0.8621	0.8621

TABLE V
Factors that predict academic student performance

Attribute	Attribute’s Usage (%)
G2	100.00
G1	64.56
famsize	19.94
Fjob	18.99
age	14.87
guardian	9.81
romantic	7.28
dalc	4.43
freetime	4.11
famrel	3.48
traveltime	2.53
failures	2.53

V. Conclusion and Future Work

Education is a critical aspect of human life as it provides human beings with knowledge, shapes our perspectives of life and gives us critical thinking skills. This research attempted to identify the factors which affect a student’s academic performance, and then used these factors to predict the student’s performance in subsequent examination. A C5.0 Decision Tree model was used to perform a binary classification to achieve the objective.

The model achieved an accuracy of 90%, with precision of 86% and recall of 86%. The results show that the most important factor that predicts student performance is their previous academic performances, Second Period Grade and First Period Grade.

For future consideration, Clustering, another machine learning technique, can help researchers to group students that pass or fail into clusters showing the factors the clustering algorithm used to group them. Models of Deep Learning, a promising area of Machine Learning, can also be utilized to explore the dataset further.

References

[1] C. Romero and S. Ventura, "Educational data mining: a review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 40, no. 6, pp. 601–618, 2010.

[2] R. S. Baker and P. S. Inventado, "Educational data mining and learning analytics," in *Learning analytics*. Springer, 2014, pp. 61–75.

[3] G. Siemens and R. S. d. Baker, "Learning analytics and educational data mining: towards communication and collaboration," in *Proceedings of the 2nd international conference on learning analytics and knowledge*, 2012, pp. 252–254.

[4] P. Cortez and A. M. G. Silva, "Using data mining to predict secondary school student performance," 2008.

[5] A. Daud, N. R. Aljohani, R. A. Abbasi, M. D. Lytras, F. Abbas, and J. S. Alowibdi, "Predicting student performance using advanced learning analytics," in *Proceedings of the 26th international conference on world wide web companion*, 2017, pp. 415–421.

[6] W. Puarungroj, N. Boonsirisumpun, P. Pongpatrakant, and S. Phromkhot, "Application of data mining techniques for predicting student success in english exit exam," in *Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication*, 2018, pp. 1–6.

[7] R. A. Rustia, M. M. A. Cruz, M. A. P. Burac, and T. D. Palaoag, "Predicting student's board examination performance using classification algorithms," in *Proceedings of the 2018 7th International Conference on Software and Computer Applications*, 2018, pp. 233–237.

[8] E. Osmanbegovic and M. Suljic, "Data mining approach for predicting student performance," *Economic Review: Journal of Economics and Business*, vol. 10, no. 1, pp. 3–12, 2012.

[9] X. Ma, Y. Yang, and Z. Zhou, "Using machine learning algorithm to predict student pass rates in online education," in *Proceedings of the 3rd International Conference on Multimedia Systems and Signal Processing*, 2018, pp. 156–161.

[10] C. I. P. Benablo, E. T. Sarte, J. M. D. Dormido, and T. Palaoag, "Higher education student's academic performance analysis through predictive analytics," in *Proceedings of the 2018 7th International Conference on Software and Computer Applications*, 2018, pp. 238–242.

[11] B. Sekeroglu, K. Dimililer, and K. Tuncal, "Student performance prediction and classification using machine learning algorithms," in *Proceedings of the 2019 8th International Conference on Educational and Information Technology*, 2019, pp. 7–11.

[12] H. Yao, D. Lian, Y. Cao, Y. Wu, and T. Zhou, "Predicting academic performance for college students: A campus behavior perspective," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 3, pp. 1–21, 2019.

[13] W. Chango, R. Cerezo, and C. Romero, "Predicting academic performance of university students from multi-sources data in blended learning," in *Proceedings of the Second International Conference on Data Science, E-Learning and Information Systems*, 2019, pp. 1–5.

[14] J. Denny, M. Rubeena, and J. K. Denny, "A noval approach for predicting the academic performance of student," in *2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*. IEEE, 2019, pp. 1–5.

[15] C.-C. Kiu, "Data mining analysis on student's academic performance through exploration of student's background and social activities," in *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*. IEEE, 2018, pp. 1–5.

[16] J. Jamesmanoharan, S. H. Ganesh, M. L. P. Felciah, and A. Shafreenbanu, "Discovering students' academic performance based on gpa using k-means clustering algorithm," in *2014 World Congress on Computing and Communication Technologies*. IEEE, 2014, pp. 200–202.

[17] P. Guleria, N. Thakur, and M. Sood, "Predicting student performance using decision tree classifiers and information gain," in *2014 International Conference on Parallel, Distributed and Grid Computing*. IEEE, 2014, pp. 126–129.

[18] A. Yunita, H. B. Santoso, and Z. A. Hasibuan, "Deep learning for predicting students' academic performance," in *2019 Fourth International Conference on Informatics and Computing (ICIC)*. IEEE, 2019, pp. 1–6.

[19] B. Guo, R. Zhang, G. Xu, C. Shi, and L. Yang, "Predicting students performance in educational data mining," in *2015 International Symposium on Educational Technology (ISET)*. IEEE, 2015, pp. 125–128.

[20] U. of California Irvine Machine Learning Repository. Uci machine learning repository: Student performance data set. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Student+Performance

[21] S.-l. Pang and J.-z. Gong, "C5. 0 classification algorithm and application on individual credit evaluation of banks," *Systems Engineering-Theory & Practice*, vol. 29, no. 12, pp. 94–104, 2009.