# PREDICTING STUDENT PERFORMANCE USING MACHINE LEARNING TECHNIQUES

Noha Alawwad

17214080

Submitted as a requirement for the course CA684P: Machine Learning. MSc in Data Analytics. Spring 2018

# Table of Contents

1. Introduction ... 1
2. Problem definition ... 1
3. Related work ... 1
4. Proposed methods ... 2
4.1. Intuition ... 3
4.2. Description of the algorithms ... 3
4.2.1 Linear and Logistic Regression ... 3
4.2.2 Naive Bayes (NB) ... 3
4.2.3 Artificial Neural Network (ANN) ... 4
4.2.4 Random Forest (RF) ... 4
4.2.5 k-Nearest Neighbors (KNN) ... 4
4.2.6 Support Vector Machine (SVM) ... 4
4.2.7 Decision Tree (DT) ... 4
5. Experiments ... 4
5.1. The data set ... 4
5.2. Testbed ... 6
5.3 Evaluation Methods ... 6
5.4. Models construction ... 7
5.4.1. Preprocessing ... 7
5.4.2. Features selection ... 8
5.4.3 Training and Testing ... 10
6. Results and Analysis ... 10
6.1 Classification models ... 10
6.2 Regression models ... 11
7. Recommendations ... 12
8. Conclusion ... 13
References ... 13

## 1. Introduction

Education is the most significant factor in the development of countries. With a complete education system, starting from kindergarten to high schools, until colleges and universities, people are prepared to actively contribute to the development of their countries. Student performance measurement is a key element of an education process to achieve a long-term economic progress. Educational research has found that there are several social, and economic factors that take part in academic failure in many countries. The motivation for analyzing the student performance is to investigate any factors that impact students' failure and provide opportunities for learners to demonstrate their weakness and strengths and encourage them to take responsibility for their own learning. In addition, help the teachers to improve their teaching methods.

In this report, several machine learning (ML) algorithms have been applied to students' performance datasets. The analysis process involved various evaluation methods that measure the prediction accuracy of each model. The goal is to identify the main factors associated with students and their environment that may impact their performance.

## 2. Problem Definition

Talking about the Portuguese educational, in particular, the educational level has improved clearly in the last years. Nevertheless, surveys of education in Europe are still putting Portugal at the tail of the list because of high students' failure rates, especially in the core classes of Mathematics and Portuguese Language [1]. In order to investigate and extract high knowledge from raw educational data, Machine Learning (ML) algorithms can be utilized in different ways to aid the education experts in discovering any potential reasons or factors affecting the student's performance.

## 3. Related Work

M.Ramaswami and R.Bhaskaran [2] analyzed the relations between many variables of a higher secondary school education to predict the student's performance. They used Chi-square Automatic Interaction Detector (CHAID) prediction model with seven features and found that the strongest features affecting the prediction are: marks, school location, living area and education type.

Another study [3] have used two ML methods to enhance the outcome of academic performance prediction. For small datasets, they applied Support Vector Machine (SVM), and with big datasets, they used Decision Tree, and both gave satisfactory results.

Arockiam et al. [4] tried to perform a comparison of the programming skills between urban and rural students by using two techniques. He applied Frequent Pattern Tree (FP Tree) association method to discover common patterns. He also used K-means clustering algorithm to determine the students programming skills. The result shows that urban students are better in programming skills than rural students.

Cortez and Silva [5] used datasets of two secondary school students in order to predict the student performance in two core classes. In this study, four data mining methods are used to build the classification models: Decision Trees (DT), Random Forests (RF), Artificial Neural Networks (ANN) and Support Vector Machines (SVM). The result shows the higher prediction accuracy of DT and ANN, with 93% and 91% respectively.

Ramesh, Parkavi and K. Ramar [6] aimed to identify the factors that influence the performance and grades of students in final examinations. In this study, five different classification algorithms (Naive Bayes, Multilayer Perceptron (MLP), Social Media Optimization (SMO), J48, REP Tree) were used to classify the data set. After a comparison, it found that the MLP algorithm was the best method to predict the performance of the students, with an accuracy of 72%. The results of this report will be compared later with the results of Parkavi and K. Ramar as they used the same data set [1].

## 4. Proposed Methods

The present work intends to approach students' performance in two Portuguese secondary schools in the classes of Mathematics and Portuguese language by applying some ML techniques on real data sets. The goal is to predict student’s results, along with the key variables affecting the outcomes. The two core classes data sets are used to build two models:

1- Binary classification (pass/fail).

2- Regression, with a numeric output (from 0 to 20).

### 4.1. Intuition

Parkavi and K. Ramar paper [6] applied classification and regression methods for building the models without feature selection approach. They considered all attributes when building their models. This research in constraint applies feature selection in order to obtain a better result and consider the most relevant features affecting the prediction of student performance. Feature selection is the procedure of automatically choosing the attributes with a high contribution in predicting the needed classes or values. Selecting some of the features to build the model rather than considering the whole data could increase the performance accuracy.

Moreover, reducing the redundant data could help in overcoming the overfitting problem and lower the probability of providing decisions based on noise. Furthermore, training a subset of data is always consuming less time [8]. In addition, making some preprocessing on the data (e.g., dealing with outliers) can also enhance the overall results.

### 4.2. Description of The Algorithms

In this paper different types of algorithms have been used to measure students' performance and determine the influencing factors. Generally, two methods have been applied to predict the final grade (G3). First, regression to predict continues values of G3. Second, classification with two G3's binary values: 0 represents fail and 1 represents pass. Under each of the two methods, the following seven ML algorithms have been applied:

#### 4.2.1 Linear and Logistic Regression

Linear Regression used to predict and estimate a continuous output value for a set of input values [9]. The report uses this algorithm to predict final Grade values depend on a set of students' attributes.

Logistic Regression is a method for binary classification problems. That is, the output takes only two values, "0" or "1". In the report, the G3 has been converted into binary format.

#### 4.2.2 Naive Bayes (NB)

Naive Bayes (NB) algorithm is based on Bayes' theory that assumes independence relation between the variables to calculate the probabilities of hypothesis [7]. NB can be applied on continuous values beside binary values. Gaussian NB has been applied for both binary and continuous values.

#### 4.2.3 Artificial Neural Network (ANN)

ANN works by testing each input attribute against each predictable attribute [11]. In the report, the ANN has been applied in binary and continuous values.

#### 4.2.4 Random Forest (RF)

It is an algorithm for classification and regression that operate by constructing a multitude of decision trees and classify the output in classification or mean prediction in regression.

#### 4.2.5 k-Nearest Neighbors (KNN)

It is an algorithm for classification and regression. The input consists of the k closest objects in the feature space. The output in classification is classified by popular vote of its neighbors. On the other hand, regression output is mean of the values of its k nearest neighbors [9].

#### 4.2.6 Support Vector Machine (SVM)

It is also an algorithm for classification and regression. This algorithm builds a model that assigns new objects to one category and making it a non-probabilistic binary linear classifier [7]. SVM used in this report for both continuous and binary values.

#### 4.2.7 Decision Tree (DT)

Is classification approach to identify the best model fits the relationship between the attributes and class label. Decision Tree Classifier poses a series of questions about the attributes of the test record. Each time it receives an answer, a follow-up question is asked until reached to final a class label [11]. In this report, DT has been applied to both binary and continuous G3 values.

## 5. Experiments

All the details of data, implementation environment, and models' construction and evaluation are shown in the following subsections.

### 5.1. The Dataset

The students' performance dataset provides information related to student's achievement in Mathematics and Portuguese language subjects for two Portuguese schools. The data was collected from the schools' reports and questionnaires. The data was published in two datasets: Mathematics that has 395 instances, and the Portuguese language with 649 instances. Table 5.1 lists all attributes and their descriptions. Figure 5.1 shows histogram charts of some important attributes.

**Table 5.1: Attributes and their descriptions**

| **Attribute** | **Description (Domain)** |
|---|---|
| Sex | Student's sex (binary: female or male) |
| Age | Student's age (numeric: from 15 to 22) |
| School | Student's school (binary: Gabriel Pereira or Mousinho da Silveira) |
| Address | Student's home address type (binary: urban or rural) |
| P status | Parent's cohabitation status (binary: living together or apart) |
| Medu | mother's education (numeric: from 0 to 4) |
| Mjob | mother's job (nominal) |
| Fedu | father's education (numeric: from 0 to 4) |
| Fjob | father's job (nominal) |
| Guardian | student's guardian (nominal: mother, father or other) |
| Famsize | family size (binary: ≤ 3 or > 3) |
| Famrel | quality of family relationships (numeric: from 1 – very bad to 5 – excellent) |
| Reason | reason to choose this school (nominal: close to home, school reputation, course preference or other) |
| Travel time | home to school travel time (numeric: 1–<15 m., 2–15 to 30 m., 3–30 m. to 1 hour or 4–>1 hour). |
| Studytime | weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours) |
| Failures | number of past class failures (numeric: n if 1 ≤ n < 3, else 4) |
| Schoolsup | extra educational school support (binary: yes or no) |
| Famsup | family educational support (binary: yes or no) |
| Activities | extra-curricular activities (binary: yes or no) |
| Paid | extra paid classes (binary: yes or no) |
| Internet | Internet access at home (binary: yes or no) |
| Nursery | attended nursery school (binary: yes or no) |
| Higher | wants to take higher education (binary: yes or no) |
| Romantic | with a romantic relationship (binary: yes or no) |
| Freetime | free time after school (numeric: from 1 – very low to 5 – very high) |
| Goout | going out with friends (numeric: from 1 – very low to 5 – very high) |
| Walc | weekend alcohol consumption (numeric: from 1 – very low to 5 – very high) |
| Dalc | workday alcohol consumption (numeric: from 1 – very low to 5 – very high) |
| Health | current health status (numeric: from 1 – very bad to 5 – very good) |
| Absences | number of school absences (numeric: from 0 to 93) |
| G1 | first-period grade (numeric: from 0 to 20) |
| G2 | second-period grade (numeric: from 0 to 20) |
| G3 | final grade (numeric: from 0 to 20) |

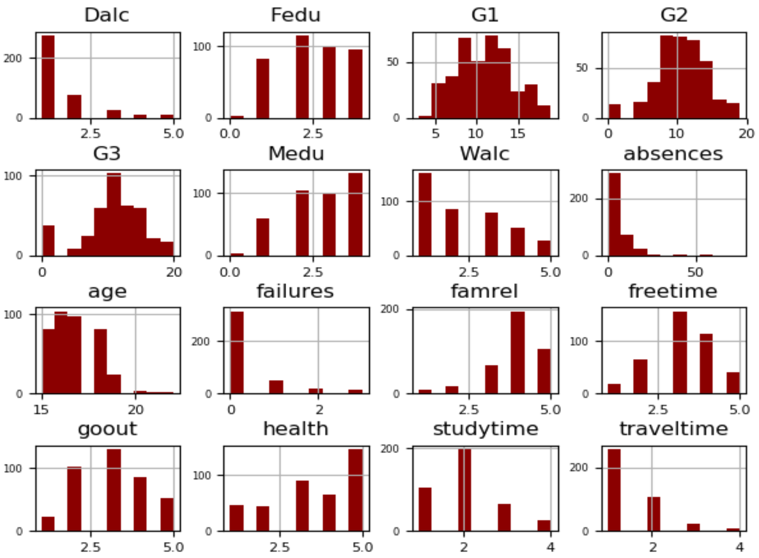


**Figure 5.1: Histograms for some of the attributes.**

## 5.2. Testbed

In this project, *Weka* 3.6 software has been used in feature selection process. Also, *Anaconda* platform to lunch python 3.6. Many libraries have been imported in python platform to accomplish the analysis process.

## 5.3 Evaluation Methods

There are various methods and metrics to evaluate the performance of classification and regression models. We used K-fold cross-validation to estimate the performance of a machine learning algorithms. In the beginning K-fold cross-validation was applied on the whole data, the consequence was with the large number of k fold, as the bias of the error rate will be small which will not reflect the exact accuracy level. In addithion, the time of computation will be high [9].

Based on the mintioned issues with K-fold cross-validation, Hold-out cross validation method is applied then and considered as the main evaluation method for this research. Hold-out cross validation method works with test and train split method where the cross validation with 10 folds is applied on training set to estimate the model parameters. Test data used to examine the validity of the model by comparing the predicted values with actual values. Prediction accuracy of the model will be more accurate in Hold-out method than as the test part is isolated from training part as a hold-out set [10]. In addition, there are some metrics for evaluation based on confusion matrix: accuracy, recall, sensitivity, specificity, and F-measure. Furthermore, Root Mean Squared

Error (RMSE) metric and Mean Squared Error (MSE) can be used to measure regression performance [7]. In this paper, two evaluation methods have been used as follows:

- Classification method: In classification, algorithms accuracy have been evaluated using Percentage of Correct Classifications (PCC). The confusion matrices has been applied to show the number of correct and incorrect prediction in comparison with actual outcome in the dataset.
- Regression method: The common evaluation metric for regression models is root mean squared error (RMSE), Where RMSE value closer to 0 indicate a higher accuracy of prediction.

## 5.4. Models Construction

### 5.4.1. Preprocessing

Preprocessing and transforming data are always essential steps before introducing the problem's structure to the ML algorithms. Preprocessing is applied to both datasets as follows:

- Encoding categorical features such as father job attribute(Fjob), where {1:at_home, 2: health, 3: services, 4: teacher, 5: other}.
- Transforming all yes/no columns to binary values through *get_dummies()* function.
- Converting final grade attribute (G3) from continues to categorical, where {pass: G3 >= 10, fail: G3 < 10}. Then transform it to binary in order to work with classification models.
- Dropping-off instances with 0 value in final grade (G3) column to enhance the results, assuming that they are outliers and students were absent in the final exam. Figures 5.2-a and 5.2-b show the scatter plots for the G3 with and without outliers.

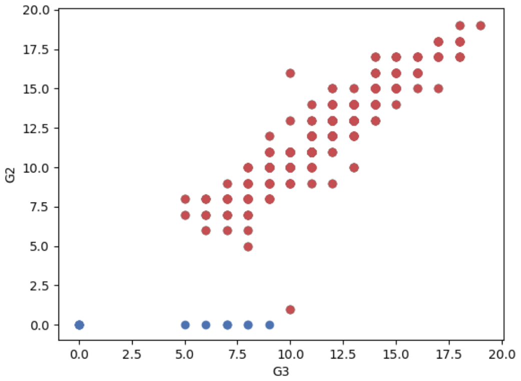


**Figure 5.2-a: Scatter plot for the original G3**

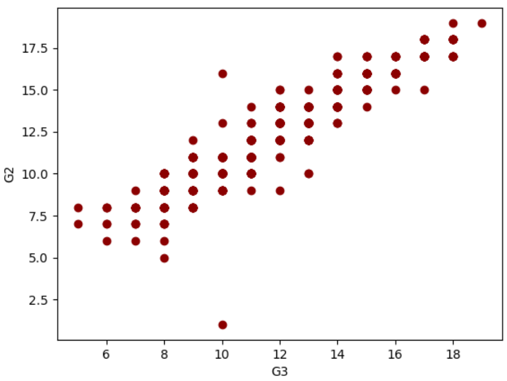


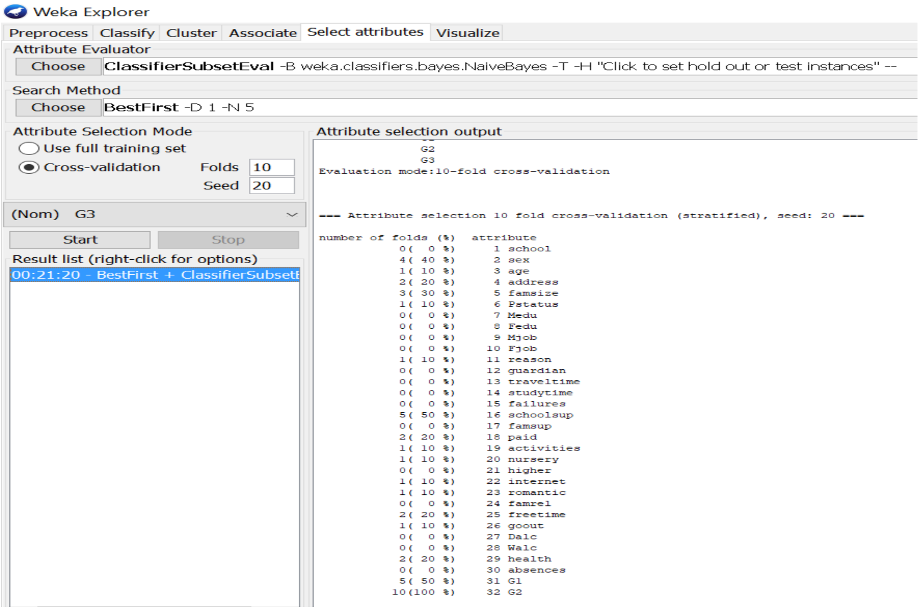
**Figure 5.2-b: Scatter plot for G3 after removing the outliers.**

- Normalizing data using the standardization method. This method is widely used for optimizing the machine learning algorithms (e.g. we applied it with ANN, KNN, linear regression, and logistic regression models and the results were much better). The general method of calculation is giving in the following equation:

$$x_{new} = \frac{x - \mu}{\sigma}$$

### 5.4.2. Features Selection

As recommended in Parkavi and K. Ramar paper [6], features selection process was performed automatically using the wrapper function called Classifier Subset Evaluator in Weka [8]. The below figure shows a sample of the feature selection result with NB algorithm in Portuguese's dataset.



Attribute Evaluator
ClassifierSubsetEval -B weka.classifiers.bayes.NaiveBayes -T -H "Click to set hold out or test instances" --

Search Method
BestFirst -D 1 -N 5

Attribute Selection Mode
○ Use full training set
● Cross-validation Folds 10 Seed 20

(Nom) G3

Attribute selection output

```
                G2
                G3
Evaluation mode:10-fold cross-validation



=== Attribute selection 10 fold cross-validation (stratified), seed: 20 ===

number of folds (%)  attribute
           0(  0 %)     1 school
           4( 40 %)     2 sex
           1( 10 %)     3 age
           2( 20 %)     4 address
           3( 30 %)     5 famsize
           1( 10 %)     6 Pstatus
           0(  0 %)     7 Medu
           0(  0 %)     8 Fedu
           0(  0 %)     9 Mjob
           0(  0 %)    10 Fjob
           1( 10 %)    11 reason
           0(  0 %)    12 guardian
           0(  0 %)    13 traveltime
           0(  0 %)    14 studytime
           0(  0 %)    15 failures
           5( 50 %)    16 schoolsup
           0(  0 %)    17 famsup
           2( 20 %)    18 paid
           1( 10 %)    19 activities
           1( 10 %)    20 nursery
           0(  0 %)    21 higher
           1( 10 %)    22 internet
           1( 10 %)    23 romantic
           0(  0 %)    24 famrel
           2( 20 %)    25 freetime
           1( 10 %)    26 goout
           0(  0 %)    27 Dalc
           0(  0 %)    28 Walc
           2( 20 %)    29 health
           0(  0 %)    30 absences
           5( 50 %)    31 G1
          10(100 %)    32 G2
```

**Figure 5.3: Feature selection in Weka for NB algorithm.**

In addition, different classifier where chosen each time with each feature selection process based on the ML algorithm that will be applied. The following table shows the classifiers used for each algorithm in Weka.

**Table 5.2: The selected classifiers per algorithm**

| Algorithms | Classifier |
|---|---|
| linear regression | Linear regression |
| Logistic regression | Logistic |
| SVM | LibSVM |
| NV | Naive Bayes |
| RF | Random Forest |
| DDT | J48 |
| KNN | IBK |
| ANN | MultilayerPercepton |

The significant features for each model as a result of the feature selection process are shown in following table.

**Table 5.3: the selected feature per algorithm**

| Attributes/ Algorithms | Linear | Logistic | NB | RF | SVM | DT | KNN | KNN |
|---|---|---|---|---|---|---|---|---|
| Sex | | | ✓ | | | | | |
| Age | | ✓ | | ✓ | ✓ | | ✓ | |
| School | | | | | | | | ✓ |
| Address | ✓ | | | | | | | |
| Pstatus | | | | | | | | |
| Medu | | | | ✓ | | | | |
| Mjob | | ✓ | | | | | | ✓ |
| Fedu | | | | | | ✓ | | ✓ |
| Fjob | ✓ | ✓ | | | | | | ✓ |
| guardian | | | | | | | | |
| famsize | | | | | | | | |
| famrel | | ✓ | | ✓ | | ✓ | ✓ | ✓ |
| reason | ✓ | | | | | | | ✓ |
| traveltime | ✓ | | | | | | | ✓ |
| studytime | | | | | | | | |
| failures | ✓ | | | | | | | |
| schoolsup | | | ✓ | | | | | |
| famsup | | | | | | | | |
| activities | | | | | | | | |
| paid | | ✓ | | | ✓ | | | |
| internet | | | | | | | | |
| nursery | | | | | ✓ | | | |
| higher | | | | | | | | |
| romantic | | | | | | | | |
| freetime | | | | | | | ✓ | |
| goout | | | | ✓ | | | | ✓ |
| Walc | | | | | | | | ✓ |
| Dalc | ✓ | | | | | | | |
| health | ✓ | | | | | | | |
| absences | | | | ✓ | | | ✓ | |
| G1 | ✓ | | ✓ | | ✓ | | ✓ | |
| G2 | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

### 5.4.3 Training and Testing

The complete dataset has been divided into two sets: 80% for training dataset, 20% for testing dataset. In the training dataset, cross-validation method has been used to train the model and estimate the parameters. After training, the testing dataset has been used to evaluate the final models' performance by comparing the predicted values with actual values. Details of the evaluation methods are shown is section 5.3.

## 6. Results and Analysis

As a result of considering part of the data using feature selection for each prediction model in this report, the performance has changed compared to the results of the previous work of Parkavi and K. Ramar [6] with considering all the data. In the following two subsections, current work's results have been analyzed, then compared with the previous work for both classification and regression models in both Portuguese and Mathematics datasets.

### 6.1 Classification Models

Table 6.1 shows the results of all the applied classification models. In Portuguese dataset, NB was the best model with the percentage of correct classification equal to 92.91%, where RF, SVM and ANN all had the same accuracy equal to 91.33%. The worst model was the DDT with an accuracy of 89.76%. For Mathematics dataset, Logistic regression, RF and DT models present the best results with accuracy equal to 93.05%, where NB showed the worst model.

For comparison, in the previous work, the best model applied on Portuguese dataset was DT with accuracy of 93%, which is better by only 0.1% comparing with NB accuracy result in the current work. In contrast, for Mathematics dataset, current work's RF model resulted in 93.05%, which is better than the accuracy percentage of the same model in the previous work (91.2%).

**Table 6.1: Binary classification results (PCC values in %, Best result in bold)**

| | Portuguese | | | | | | | Mathematics | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Logistic | **NB** | RF | SVM | DT | KNN | ANN | **Logistic** | NB | **RF** | SVM | **DT** | KNN | ANN |
| **Present work** | 90.55 | **92.91** | 91.33 | 91.33 | 89.76 | 90.55 | 91.33 | **93.05** | 84.72 | **93.05** | 87.5 | **93.05** | 87.5 | 90.27 |
| **Previous work** | - | **-** | 92.6 | 91.4 | **93.0** | - | 90.7 | **-** | - | **91.2** | 86.3 | 90.7 | - | 88.3 |

In Mathematics dataset, the confusion matrix was applied on the test part for all the algorithms, as a result RF and DDT models has 67 correct classifications out of 72 and 5 misclassifications on mathematics dataset which is the best. The full results of confusion matrices for classification models on Mathematics dataset are shown from table 6.2 to table 6.7.

**Table 6.2: KNN CM**

| | PASS | FAIL |
|---|---|---|
| PASS | 13 | 4 |
| FAIL | 5 | 50 |

**Table 6.3: NB CM**

| | PASS | FAIL |
|---|---|---|
| PASS | 15 | 2 |
| FAIL | 9 | 46 |

**Table 6.4: SVM CM**

| | PASS | FAIL |
|---|---|---|
| PASS | 12 | 5 |
| FAIL | 4 | 51 |

**Table 6.5: ANN CM**

| | PASS | FAIL |
|---|---|---|
| PASS | 20 | 3 |
| FAIL | 4 | 45 |

**Table 6.6: DDT CM**

| | PASS | FAIL |
|---|---|---|
| PASS | 16 | 1 |
| FAIL | 4 | 51 |

**Table 6.7: RF CM**

| | PASS | FAIL |
|---|---|---|
| PASS | 17 | 0 |
| FAIL | 5 | 50 |

In contrast, confusion matrices for Portuguese dataset show that the best model was the NB model with 118 correct classifications out of 217 and 11 misclassifications. Below are the confusion matrices results for Portuguese dataset as shown in table 6.8 to table 6.13.

**Table 6.8: KNN CM**

| | PASS | FAIL |
|---|---|---|
| PASS | 8 | 11 |
| FAIL | 1 | 107 |

**Table 6.9: NB CM**

| | PASS | FAIL |
|---|---|---|
| PASS | 15 | 4 |
| FAIL | 5 | 103 |

**Table 6.10: SVM CM**

| | PASS | FAIL |
|---|---|---|
| PASS | 10 | 9 |
| FAIL | 2 | 106 |

**Table 6.11: ANN CM**

| | PASS | FAIL |
|---|---|---|
| PASS | 13 | 6 |
| FAIL | 5 | 103 |

**Table 6.12: DDT CM**

| | PASS | FAIL |
|---|---|---|
| PASS | 11 | 8 |
| FAIL | 5 | 103 |

**Table 6.13: RF CM**

| | PASS | FAIL |
|---|---|---|
| PASS | 11 | 8 |
| FAIL | 4 | 104 |

## 6.2 Regression Models

The regression results of ML models indicate that the linear regression has the best RMSE value in both Portuguese and Mathematics datasets with 0.31, and 0.24 respectively, followed by ANN. While the worst result is shown with NB in both datasets with 2.46 in Portuguese dataset and 1.13 in Mathematics dataset. By looking at the following table, RMSE results of the models applied in the previous work dramatically enhanced in the current work considering only the highly contributed features.

**Table 6.14: Regression results (RMSE values, Best result in bold)**

| | Portuguese | | | | | | | Mathematics | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Linear** | NB | RF | SVM | DT | KNN | ANN | **Linear** | NB | RF | SVM | DT | KNN | ANN |
| **Present work** | **0.31** | 2.46 | 1.02 | 1.01 | 1.07 | 0.42 | 0.34 | **0.24** | 1.13 | 0.90 | 1.04 | 0.25 | 0.31 | 0.26 |
| **Previous work** | **-** | - | **1.32** | 1.35 | 1.46 | - | 1.36 | **-** | - | **1.75** | 2.09 | 1.94 | - | 2.05 |

For comparison, the previous work's best model was RF with RMSE equal to 1.32 in Portuguese dataset and 1.75 in Mathematics dataset. Liner regression is the best model in this work with 0.31 and 0.24 RMSE in Portuguese and Mathematics datasets, respectively.

Fig 6.1-a and Fig 6.1-b show the scatter plot of linear model in regression results for Portuguese and Mathematics datasets.

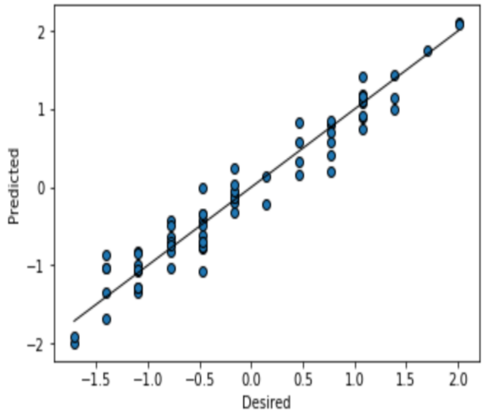


**Figure 6.1-a: Linear scatter plot for Math**

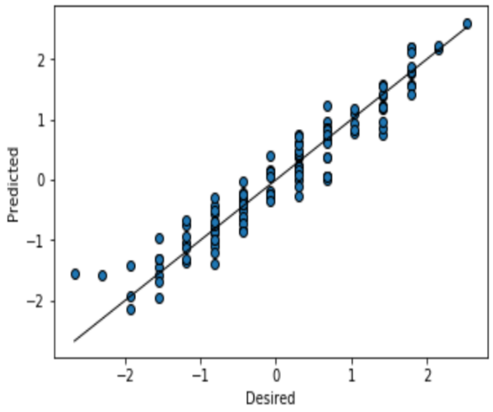


**Figure 6.1-b: Linear scatter plot for Portuguese**

## 7. Recommendations

For future work, the analysis process can be extended with other datasets that cover more regions. Furthermore, more sociological studies can be done by the schools to the students to get other related attributes such as number of the family members, and the income source to get more effective results and a broader approach. Also, experiments by using the average score of the first period and the second period scores (G1, G2) could help in improving the prediction accuracy results.

## 8. Conclusion

Education is a key element in our society. Measuring students' performance could enhance the academic achievements. In this paper, we have explored the factors that have significant impact in predicting student's achievement of secondary students in Mathematics and Portuguese subjects. Moreover, features selection method with wrapper function was used to identify the most related factors. Many attributes have been considered, such as grades of first and second periods, ethical and social issues. Seven different machine learning algorithms, for instance, Artificial Neural Networks (ANN), Support Vector Machine (SVM) , linear and logistic regression. Finally, two evaluation methods such as accuracy for binary classification and Root Mean Squared Error (RMSE) for regression were tested.

## References

[1] "UCI Machine Learning Repository: Student Performance Data Set", Archive.ics.uci.edu, 2018. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Student+Performance. [Accessed: 07- Feb-2018].

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

[3] Nguyen Thai-Nghe, Andre Busche, and Lars Schmidt-Thieme, “Improving Academic Performance Prediction by Dealing with Class Imbalance”, 2009 Ninth International Conference on Intelligent Systems Design and Applications.

[4] L.Arockiam, S.Charles, I.Carol, P.Bastin Thiyagaraj, S. Yosuva, V. Arulkumar, “Deriving Association between Urban and Rural Students Programming Skills”, International Journal on Computer Science and Engineering Vol. 02, No. 03, 2010, 687-690

[5] P. Cortez, and A. Silva, “Using Data Mining To Predict Secondary School Student Performance”, In EUROSIS, A. Brito and J. Teixeira (Eds.), 2008, pp.5-12.

[6] V. Ramesh, P. Parkavi, and K. Ramar, “Predicting Student Performance : A Statistical and Data Mining Approach,” Int. J. Comput. Appl., vol. 63, no. 8, pp. 35–39, 2013.

[7] J. Brownlee, Machine learning mastery with python, Australia: Machine Learning Mastery Pty Ltd, 2018.

[8] J. Brownlee, Machine_learning_mastery_with_weka, Australia: Machine Learning Mastery Pty Ltd, 2018.

[9] A. C. M. Guido, Introduction to Machine Learning with Python: A Guide for Data Scientists, United States: o'reilly media, 2017.

[10] M. B. D. v. A. M. S. Dongarra, Computational Science — ICCS 2004: 4th International Conference, New York: Springer, 2004.

[11] J. Brownlee, Discover How They Work and Implemment Them From Scratch, Australia: Machine Learning Mastery Pty Ltd, 2018.