Predicting Student Performance Using Machine Learning Analytics

Michael O Sullivan

R00077764
DCOM4
Cork Institute of Technology
Cork

Abstract Machine Learning methods can be implemented to predict the outcome using previously gathered information against new information. To display these methods, this research paper will demonstrate how they are used with a dataset named Student Performance which is a dataset of student achievements in two Portuguese secondary schools. The report will display analytics of machine learning algorithms.

2 Introduction

This dataset consists of two datasets, a dataset for mathematic achievements and a dataset Portuguese achievements. The data was collected during 2005-2006. The dataset consists of 32 attributes such as age, parent's education, study time, internet access, romance, free time, alcohol consumption, number of absenteeism and final grades. The schools uses a 20-point grading scale where 0 is the lowest grade and 20 is the highest grade. These results are collected three times throughout the year, G1 is the first period, G2 is the second period and G3 is the final grade. The information was collected using a questionnaire sheet with closed questions such as mother's education, family income, social questions (e.g. alcohol consumption), relationship status and school related questions such as number of failed classes in the past.

The main objective of this project is to use classifications and regression on the final grade to predict a student's performance. The grade is represented in values ranging from 0 – 20, with 20 being the highest grade. First all change this value using the average final grade achieved by all students. This can in turn make this column into a binary column by comparing the student's grade to the average. If the students grade is greater than the average they get a binary value of 1, otherwise they get a value of 0.

Some machine learning algorithm that will be used for the Student Performance dataset:
Decision Trees
Support Vector Classification (SVC)
Random Forrest
Gaussian Naive Bayes

As for this research paper all code be will be written using python and libraries such as Sklearn. The information will have to read in from the dataset and find the valid features. By using feature selection such as greddy feature selections and univartes which are part of the Sklearn library to remove features that have no significant value to the dataset. Once I have the valid features I will implement the classification algorithms that are mentioned above

3 Related Research

A research paper based on the Student Performance dataset has been published by Paulo Cortez and Alice Silva under the heading Using Data Mining To Predict Secondary School Student Performance (Silva). The finding of this research paper are using a 5 level classification to perdict students final grades using four macine learning algorithm Decision Trees (DT), Random Forests (RF),
Neural Networks (NN) and Support Vector Machines(SVM). In the paper, they were able to achieve a hight accuracy as long as first and/or second school period grades are known.

4 Algorithm/ Model Details

Attributes	Description (Domain)	Values
school	student's school	'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira
sex	student's sex	F - female or M - male
age	student's age	numeric: from 15 to 22
address	student's address type	'U' - urban or 'R' - rural
famsize	family size	LE3' - less or equal to 3 or 'GT3' - greater than 3
Medu	mother's education	0 - none,1 -primary education, 2 5th to 9th grade, 3secondary education 4 higher education
Fedu	father's education	0 - none,1 -primary education, 2 5th to 9th grade, 3secondary education 4 higher education
Mjob	mother's job	teacher, health care, civil services, at home or other
Fjob	father's job	teacher, health care, civil services, at home or other
reason	reason to choose this school	close to 'home', school 'reputation', 'course' preference or 'other'
guardian	student's guardian	mother, father or other
traveltime	home to school travel time	1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour
studytime	weekly study time	1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours
failures	number of past class failures	1 if <3, else 4
schoolsup	extra educational support	yes or no
famsup	family educational support	yes or no
paid	extra paid classes	yes or no
activities	extra-curricular activities	yes or no
nursery	attended nursery school	yes or no
higher	wants to take higher education	yes or no
internet	Internet access at home	yes or no
romantic	with a romantic relationship	yes or no
famrel	quality of family relationships	from 1 - very bad to 5 - excellent
freetime	free time after school	1 - very low to 5 - very high
goout	going out with friends	1 - very low to 5 - very high
Dalc	workday alcohol consumption	1 - very low to 5 - very high
Walc	weekend alcohol consumption	1 - very low to 5 - very high
health	current health status	1 - very bad to 5 - very good
absences	number of school absences	from 0 to 93
G1	first period grade	from 0 to 20
G2	second period grade	from 0 to 20
G3	final grade	from 0 to 20, output target

4.1 Algorithms

There will be 5 supervised classification algorithm used on this data set.

Decision Tree Classifier – The decision tree classifier uses a series of questions to and follow up question until it reaches a class label. The decision tree is given data a and a target.

Support Vector Classification – This is supervised learning algorithm, this algorithm can be used for bot classification and regression implements.

5 Empirical Evaluation

Experimental methodology, results depicted in graphs, interpretation of results.

The following results all have a start point, this will act as a base for when the greedy selection algorithm start removing feature and I then compare results to these. This will help with selecting which feature to remove. The after column is the value after most of the invaluable features are removed.

The greedy feature selection loops through each feature and remove the worst feature, this will keep looping until there is only one feature left. Each value that is in bold and underlined is the best algorithm for that subject. Just bold figures is the best algorithm without any feature selection.

5.1 Feature Selection

The column G3 was converted to binary, the average final grade was calculated. If the student scored over the average they would receive a value 1 and if lower receive a value 0.

	With Standardization				Without Standardization			
Subject	**Math**		**Portuguese**		**Math**		**Portuguese**	
	Before	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**
Decision Tree	85.04 %	**91.83 %**	**91.69 %**	**93.82** %	85.80 %	**91.83 %**	91.85 %	**93.82 %**
SVM	83.27 %	**91.83 %**	87.23 %	**93.82** %	85.81 %	88.08 %	91.23 %	**93.82 %**
NNeighbour	75.67 %	78.67 %	77.53 %	93.05 %	85.05 %	78.67 %	87.84 %	93.05 %
RForest	**87.80 %**	**91.83 %**	90.74 %	**93.82** %	**87.81 %**	**91.83 %**	**91.98 %**	**93.82 %**
Naive Bayes	82.23 %	85.08 %	84.64 %	**93.82 %**	82.23 %	85.08 %	84.64 %	**93.82 %**
Best Feature	G2		G2		G2		G2	
Worst Feature	freetime		Fjob		freetime		Fjob	

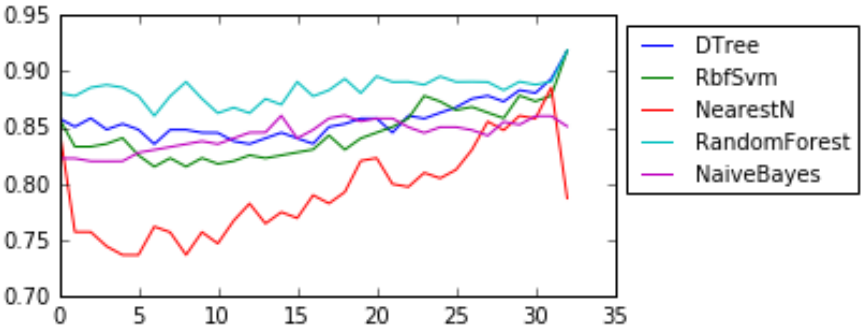

Figure 1 Math with Standardization

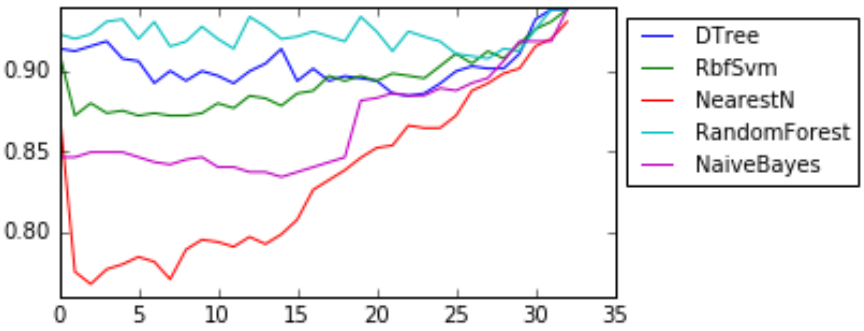

Figure 2 Portuguese with Standardization

These mathplt diagrams show how the algorithm increase in accuracy or decrease in accuracy when a feature is removed. As you can see the more feature removed the more accurate the algorithm get.

I now change the G3 column to a 5 level classification instead of a binary classification as stated previously. This is how the G3 results are broken up:

Value	1	2	3	4	5
Grade	16-20	14-15	12-13	10-11	0 - 9

Before running the greedy selection I ran the dataset through the 5 classifiers again resulting in the following:

	With Standardization				Without Standardization			
Subject	**Math**		**Portuguese**		**Math**		**Portuguese**	
	Before	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**
Decision Tree	**64.84 %**	**78.48 %**	**65.39 %**	**76.69 %**	62.81 %	**78.48 %**	64.44 %	**76.69 %**
SVM	51.98 %	**78.48 %**	56.89 %	**76.69 %**	**67.66 %**	**78.48 %**	**67.93 %**	**76.69 %**
NNeighbour	43.95 %	64.14 %	41.22 %	68.31 %	57.31 %	64.14 %	61.39 %	68.31 %
RForest	63.05 %	**78.48 %**	63.14 %	76.23 %	59.40 %	**78.48 %**	67.05 %	**76.69 %**
Naive Bayes	51.52 %	75.09 %	45.36 %	**76.69 %**	54.04 %	75.09 %	49.52 %	**76.69 %**
Best Feature	G2		G2		G2		G2	
Worst Feature	higher		Fedu		health		Fedu	

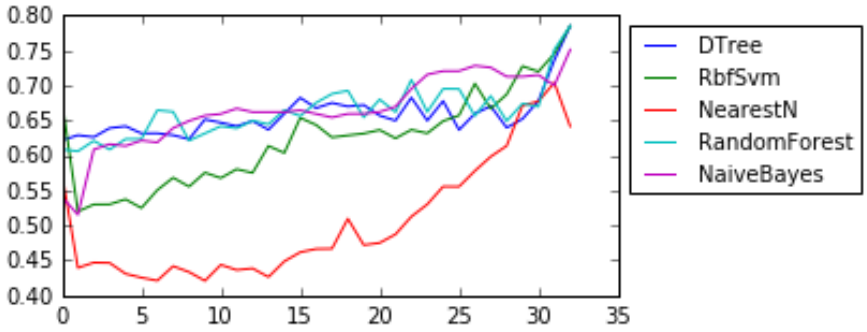

Figure 3 Math with Standardization

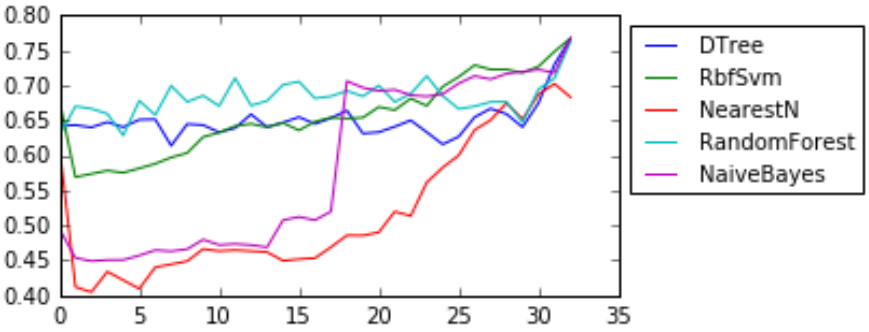

Figure 4 Portuguese with Standardization

Instead of using the 5 level classification as stated above, for this test the actual final grades are used which range from (0-20).

	With Standardization				Without Standardization			
Subject	**Math**		**Portuguese**		**Math**		**Portuguese**	
	Before	**After**	**Before**	**After**	**Before**	**After**	**Before**	**After**
Decision Tree	**34.12 %**	49.89 %	**38.99 %**	46.05 %	34.47 %	49.89 %	36.63 %	46.05 %
SVM	24.11 %	45.18 %	25.56 %	**46.56 %**	**40.83 %**	49.42 %	**40.71 %**	**47.12 %**
NNeighbour	16.82 %	40.16 %	17.44 %	42.12 %	32.70 %	40.16 %	31.48 %	42.12 %
RForest	31.83 %	47.58 %	34.84 %	45.67 %	30.79 %	48.97 %	33.02 %	46.23 %
Naive Bayes	20.30 %	**49.91 %**	13.43 %	40.59 %	23.61 %	**49.91 %**	17.50 %	40.59 %
Best Feature	G2		G2		G2		G2	
Worst Feature	traveltime		school		absences		freetime	

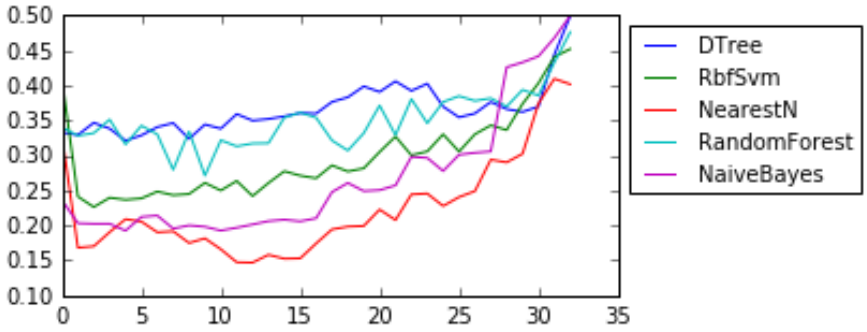

Figure 5 Math with Standardization

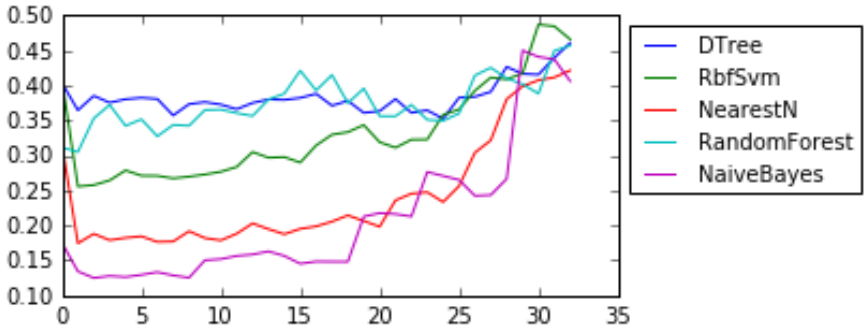

Figure 6 Portuguese with Standardization

To help increase accuracy, I decided to run the last test again using the actual student results but concatenating the Portuguese dataset and the Math dataset. With more records the accuracy hopefully will increase. To compare these, I got the average before and after from above ie.

DT – math before = 34.12 %
Dt - **Portuguese = 38.99 %**
Is an average of 36.56 %

	With Standardization		**Without Standardization**	
Subject	**Concatenated Datasets**			
	Before	**After**	**Before**	**After**
Decision Tree	**34.06 %(-2.5%)**	**49.53 %(+1.56)**	34.78 %(-.77%)	**49.53 %(+1.56** %)**)**
SVM	26.28 % (+1.44%)	49.07 %(+3.2%)	**39.08 %(-1.69%)**	49.46 %(+1.19 %))
NNeighbour	15.61 %(-1.52 %)	40.28 %(-.86%)	32.87 %(+.78%)	40.28 %(-.86 %))
RForest	31.79 %(-1.55 %)	48.65 %(+2.02 %))	31.39 %(0)	48.03 %(-.43 %))
Naive Bayes	19.99 %(+3.12 %)	41.71 % (-3.54 %)	22.67 %(+2.11%)	41.71 %(-3.54 %))
Best Feature	G2		G2	
Worst Feature	Higher		famrel	

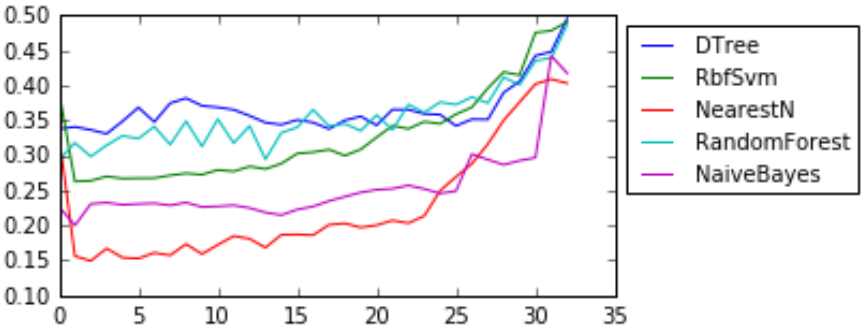

Figure 7 Concatenated Datasets with Standardization

As the results from above show how dominant G2 is as a feature, the following results are with G2 removed. As from the result show from above that the results are slightly better without standardization

	Binary Classification					
Subject	**Math**			**Portuguese**		
	Before	**G2 Removed**	**G1 and G2 Removed**	**Before**	**G2 Removed**	**G1 and G2 Removed**
Decision Tree	85.80 %	76.60 %	62.21 %	91.85 %	80.77 %	64.10 %
SVM	85.81 %	**81.95** %	68.09 %	91.23 %	**84.92 %**	68.12 %
NNeighbour	85.05 %	78.91 %	66.81 %	87.84 %	80.63 %	63.18 %
RForest	**87.81 %**	77.38 %	61.46 %	**92.61%**	82.61 %	64.38 %
Naive Bayes	82.23 %	74.09 %	**70.05 %**	84.64 %	79.56 %	**72.13 %**

	5 Level Classification					
Subject	**Math**			**Portuguese**		
	Before	**G2 Removed**	**G1 and G2 Removed**	**Before**	**G2 Removed**	**G1 and G2 Removed**
Decision Tree	62.81 %	53.99 %	27.88 %	64.44 %	45.39 %	28.50 %
SVM	**67.66 %**	**55.24 %**	**32.96 %**	**67.93 %**	**55.04 %**	**34.70 %**
NNeighbour	57.31 %	45.91 %	26.59 %	61.39 %	48.14 %	28.39 %
RForest	59.40 %	44.38 %	26.36 %	67.05 %	45.48 %	29.44 %
Naive Bayes	54.04 %	35.77 %	27.86 %	49.52 %	33.80 %	24.94 %

	Regression G3 (0-20)					
Subject	**Math**			**Portuguese**		
	Before	**G2 Removed**	**G1 and G2 Removed**	**Before**	**G2 Removed**	**G1 and G2 Removed**
Decision Tree	34.47 %	21.50 %	10.90 %	36.63 %	23.50 %	13.11 %
SVM	**40.83 %**	**30.78 %**	**18.40 %**	**40.71 %**	**28.39 %**	**19.76 %**
NNeighbour	32.70 %	24.21 %	13.46 %	31.48 %	22.81 %	13.91 %
RForest	30.79 %	23.12 %	13.63 %	33.02 %	21.23 %	15.29 %
Naive Bayes	23.61 %	13.07 %	07.39 %	17.50 %	09.43 %	07.43 %

Hyper-Parameter Optimization

Using hyper-parameter optimization helps choose which learning algorithm to use. For the examples below, GridSearch is used to exhaustively builds and evaluates models. For the SVM method it return the mean result(%), return the best kernel either 'rbf', 'poly', 'linear'.

Values list below with underscores are the best models per subject.

	Binary					
Subject	**Math**			**Portuguese**		
	Before	**G2 Removed**	**G1 and G2 Removed**	**Before**	**G2 Removed**	**G1 and G2 Removed**
SVM	**89.57%(linear)(13)**	82.20%(linear)(1)	70.58%(linear)(2)	**93.38%(linear)(4)**	84.92%(rbf)(1)	71.80%(linear)(1)
KNeighbour	84.05 %(3)(2)	8126(6)(4)	67.84 %(26)(1)	91.07 %((56)(3)	86.00 %(28)(3)	68.56 %(26)(4)

	5 Level Classification					
Subject	**Math**			**Portuguese**		
	Before	**G2 Removed**	**G1 and G2 Removed**	**Before**	**G2 Removed**	**G1 and G2 Removed**
SVM	**70.63%(linear)(1)**	55.24%(rbf)(1)	32.96 %(rbf)(1)	**69.81%(rbf)(2)**	55.04 %(rbf)(1)	35.31 % (rbf)(2)
KNeighbour	64.82 %(16)(1)	52.88%(37)(5)	35.47 %(71)(2)	66.29%(44)(2)	55.02%(14)(2)	36.84 %(44)(2)

	Regression G3(0-20)					
Subject	**Math**			**Portuguese**		
	Before	**G2 Removed**	**G1 and G2 Removed**	**Before**	**G2 Removed**	**G1 and G2 Removed**
SVM	**42.89%(poly)(1)**	30.78 %(rbf)(1)	23.36%(rbf)(1)	**42.13%(rbf)(2)**	29.83%(rbf)(3)	19.76% (rbf)(1)
KNeighbour	38.67%(29)(1)	31.20%(34)(2)	18.40%(73)(1)	34.98%(10)(1)	28.11%(21)(5)	1975%(31)(4)

Conclusion

The results of this paper show how importance of the first period grade and the second period grade to these datasets. If the students do well in these periods it will reflect on their final grade.

Using standardization for this dataset on these dataset is unnecessary as it lowers the accuracy in places.

My results follow up one on (Silva) who mentioned that you can achieve a high accuracy knowing the first period and second period result. To test it even further and I removed theses grades to see how the accuracy would be affected. As the figures show above, it still relatively high when using binary but once the 5 level classification and the actual final grades were used there was quite a reduction in accuracy.

Bibliography

Silva, P. C. (n.d.). USING DATA MINING TO PREDICT SECONDARY SCHOOL STUDENT PERFORMANCE. http://www3.dsi.uminho.pt/pcortez/student.pdf