Predicting Student Performance using Machine Learning Techniques ECS784P

Written by Thomas Jackson (200835525)

***Abstract*—**This paper contains findings on student performance to examine the level of educational attainment within Portuguese secondary students. The machine learning methods used to analyze the data are Random Forests, Support Vector Classification and Logistic Regression. The dataset comes from two schools in Portugal alongside a series of possible factors which were considered to have an impact on student performance. In this report I discuss data transformation and data cleansing. The models themselves are contextualized to allow clarity and the data is examined with the aim of allowing individuals unfamiliar with data concepts to understand the underlying process. The results of the study are then displayed to try and highlight the most important factors in relation to student performance. A literature review is present to give an insight into the major areas of interest that surround this dataset. The findings and many parts of the data are displayed using tables and graphs to better visualize concepts for the reader.

Keywords—Random Forests, ROC, Support Vector classification, Machine Learning, Logistic regression, Student Performance.

Introduction

This paper will show how to use three different data analysis models to process a classification problem relating to student performance. The three models chosen are a Support Vector machine performing a classification analysis, a Random Forest decision tree aiming to provide insight into the impact of variables on classification and Logistical Regression to provide a counterbalance and increase accuracy.

The paper uses a dataset based on educational attainment at two difference schools in Portugal looking at performance in literacy (Portuguese). This is a worth while area to investigate as language skills can be vital to student's further development, not only are they necessary to understand other subjects they are also used in everyday life. The aim of this study then is to find a way to help educators predict how well a student will do in formative assessment so that resources within a classroom can be optimally assigned.

Literature review

Education Data Mining (EDM) is a discipline which aims to use data mining and data analysis techniques to focus on problems specific to educational institutions and the students that attend them. In this endeavor the EDM community has grown over the last two decades, with dedicated Journals appearing to directly support the free flow of information within this specialized field. A good starting point to understand EDM is to look at the 2009 review conducted by *Baker and Yacef* [1]. In this review they argue there are five primary classifications of the work of data mining: Prediction, Clustering, Relationship mining, Distillation of data for human judgement and Discovery with models. Of these the first three are staples of all forms of data mining, whereas the final two speak to the unique need for EDM to provide a result which can be used proactively in making educational assessments. With many research papers being written around student performance, it is important to make sure that models and researchers utilize both strong methodology and appropriate domain knowledge.

An example of applied domain knowledge can be found in *Salikutluk and Heyne*, where they investigate the impact of gender on student performance. "The effect of a student's own gender role seems to be only weak for boys and unrelated to girls' performance in maths" [2]. These findings that the students on gender have little to no impact on student performance is important as it means that other factors which might affect girls and boys differently will be a more direct cause of grade disparities. As will be seen later within my data set, the gender of the student was one of the least used variables as part of Random Forrest classification, in part due to this weak connection between performance and gender.

Separately, *Jensen* [3] uses data analytics to investigate the impact of additional studying time on literacy and maths scores. Their findings indicated "that students are more sensitive to changes in maths than in literacy teaching because literacy teaching also takes place at home." This quote from *Jensen* highlights a phenomenon observed within my data, as will be shown later many of the home life variables had little to no impact on final grades. It is worth noting that the subject type can have an impact on the importance of certain data points within an analytical study. Though many students have different levels of attainment in different subjects it is necessary to calibrate any predictive model at least in part around the core subject being assessed.

The dataset used in this study original comes from Kraggle and was part of a previous study by *Cortez and Silva* [4]. I have used a different methodology from this original study to attempt to generate a new understanding of the same source data. It is worth understanding Cortez and Silva's reasoning for collecting the data in its raw form. They chose to focus on Literacy and Mathematics and correlate these scores against plausible outside variables. "In particular, failure in the core classes of Mathematics and Portuguese (the native language) is extremely serious, since they provide fundamental knowledge for the success in the remaining school subjects (e.g. physics or history)." By understanding the variables which most impact student performance in core subjects we can then focus better on reforms which will be more statistically impactful. This has the benefit not only of improving understanding in the core subject but also the students' performance in the remaining subjects who rely on skills learned through core disciplines.

I will know assess several pieces of literature which helped to guide me as to the best data mining methods for this report. I have chosen to use three different classification techniques to test my dataset. Logistical Regression, Support Vector Classification and Random Forests where each used to test the data. As will later be shown the results for all three where strongly consistent with each other. Before breaking down each technique in detail I feel it is important to justify the major change I made from the original dataset.

In the original study [4], the researchers chose to train their model to predict the end of year grade in Portuguese and Maths. This study only had access to the Portuguese data and so could not do comparative analysis. In the Portuguese educational system grades are awarded on a scale between 0-20 with higher numbers resulting in higher grades. This meant that their model had to accurately predict which integer between 0-20 the final grade (G3 in the dataset) would be on a per student basis. I felt that for my study a different approach would be taken, based on the work conducted by Kabakchieva [5] I chose to follow a binary classification method. "Another important operation during the pre-processing phase is also the transformation of some variables from numeric to nominal because they are much more informative when interpreted with their nominal values." Therefor the option existed to create a binary categorization of my data, which would have split the students into those who are Strong (a predicted G3 score equal to or above 11) and those who are Weak (a predicted G3 score below 11). A G3 score of 11.2 was the mean result from the raw data and so was used to find the cut off point for my analysis. This classification system is similar to the system used to study Bulgarian students, with the strong/weak cut-off being determined by the standard deviation of the scores. This Strong/Weak analysis would make the model far more streamlined and it better within the scope of this paper, it will also make result analysis more feasible.

Random Forests was used on this dataset after consulting *Breiman* [6], *Belgiu and Dragut*[7] and *Pal* [8]. Random Forests work by creating a series of different variations which take a random sample of data and pick the variable which best splits the data into its next logical series of sub parts. This means that as the Tree is created and filled out "this classifier can be successfully used to select and rank those variables with the greatest ability to discriminate between the target classes" [7]. This makes Random Forests very useful for giving a strong indication as to the validity of the data. *Breiman* argues that "the random forest classifier is almost insensitive to overfitting for this dataset" [8]. This means that with random forests we can believe that the more trees that are created the better, as this will naturally prevent overfitting data. This is due to the impact of the "Strong Law of Large Numbers (which) shows that they always converge" [6], this tendency to convergence helps to mitigate overfitting.

The Support Vector Classification method is a type of support vector machine which is a supervised learning model. Support Vector models can do either regression or classification, with the classification version using a linear classifier to separate the data to creating strong generalizations about the data. According to *Novakovic and Veljovic* "SVM constructs a hyperplane or set of hyperplanes in a high dimensional space, which can be used for classification" [9]. Once this set of hyperplanes is created, the researcher must then attempt to find the most optimal hyperplane to reduce errors. "The best hyperplane is the one that represents the largest separation, or margin, between the two classes. The larger the margin it is the lower the generalization error of the classifier" [9]. As *Gunn* [10] points out though it is necessary when using a SVM to account for "the curse of dimensionality".

The final type of analysis used is Logistical Regression. Logistical Regression was chosen in part due to it being relatively easy to apply to my data while still providing valuable outputs. Logistical Regression can have issues with overfitting which will hopefully be mitigated by the cross validation with Random Forests. As stated in *Dreiseitl and Ohno-Machado* “the relationship between *x* and *y* has to be described more generally by a probability distribution *P*(*x*,*y*)” [11] . This in turn means that a Logistical regression model will be able to provide not only the class label for a data point but also the probability of that label. In context, it is assumable that a correctly fitted and modelled logistical regression would be able to predict the probability (*P*) of any student (*x*) achieving a given G3 score (*y*).

Data Management

This data originally comes from Kaggle [12] and was taken from an original data analytics study on Portuguese education. This meant that it had already been normalized and contained no null values. The data had also already seen a level of transformation as all the non-numeric data had been turned into either a binary output or had been turned into a numeric scale. This meant that my primary piece of data management was to decide which variables to drop to create a reasonable scope for my project. Many of the variables that I dropped where because they focused on student health which was a mitigating factor on performance that I have no domain knowledge of and so felt unable to adequately understand its impact or results. I also dropped the few data sets which still contained alphanumeric data, for instance parent job which had 8 different types of job. To make sure that the impact of parent’s where not undervalued I decided to keep Parent Education data even though it was not particularly well correlated.

Attribute	**Description**
School	student’s school either Gabriel Pereira or Mousinho da Silveira (a)
Gender	Student’s Gender (a)
Age	Students Age, from 15-22
Address	Students home address as either rural or urban (a)
PStatus	Parent’s cohabitation status (a)
Medu	Mother’s Education 0-4
Fedu	Fathers Education 0-4
Guardian	Students’ primary guardian either; Mother, Father or Other
Famsize	Family Size, either greater than or less than 3 living in the household (a)
Famrel	Students opinion on quality of family relationships, expressed as a number between 1 - 5
StudyTime	weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours)
Schoolsup	extra educational school support (a)
Famsup	family educational support (a)
Activities	extra-curricular activities (a)
paidclass	extra paid classes (a)
Internet	Internet access at home (a)
Nursery	attended nursery school (a)
Higher	wants to take higher education (a)
Romantic	Currently in a romantic relationship (a)
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)
Absences	Number of times the student missed class as a whole interger.
G1	First period of assessment (numeric: from 0 -20)
G2	Second period of assessment (numeric: from 0 -20)
G3	Third period of assessment (numeric: from 0 -20)
Pass_G3	Did the student achieve over the mean average.

This table shows all the features of my dataset, these variables where selected to show possible impacts on student performance. They can be broken down into three sections.

The first section are attributes about the individual student and there parents. This includes which of the two schools they attended alongside their age and gender. All three of these are useful baseline analytics to make sure that the data is not skewed due to the specific sample taken. The dataset then attempts to account for how home life might affect the student detailing the parents marital status, parents education and whether they live in an urban or rural environment.

The second section looks at the student's time and attitudes towards education. This is where factors like Study Time and level of in school and familial support are accounted for. As we know from the literature review the amount of time spent studying can impact children's grades and so we would expect it to impact G3 performance. This section also contains data on the student's interest in higher education (which would necessitate higher grades) and also how free they are to pursue non-educational activities.

The third section details the specific data on performance. This data is where we see the results of the impact of the other input variables. The data here is a score between 1-20 based on the students scores in three separate tests across the year. G1 is the first set of data from the first testing period, G2 the results of the middle test and G3 is the final grade in the subject. The G3 grade was the result that I was trying to predict and so was not included in the model. During implementation it was decided to add an additional datapoint to the original source. This was the attribute Pass_G3 which turned the original numeric data into a binary classification. For pass_G3 I chose to use the mean result for this data set to decide if a student passed or failed. As the mean was 11.2 all students who scored 10 or less where counted as "fails" and all who scored 11 or more as "passes".

Methodology

For this paper I used Python as my primary coding language. I used the following libraries to provide useful functions for data manipulation, transformation and analysis.

Pandas: Used to provide a framework for interacting with the dataset.

Matplotlib: A Library for data visualization, useful for plotting variables and best fit correlations.

Seaborn: Seaborn is a library based on matplotlib that provides additional forms of data display.

NumPy: Library which allows for the manipulation of data especially in creating 3D arrays of data.

Scikit-learn: A machine learning python library that provided the learning algorithms and many testing algorithms.

Graphviz: a library used to export graph data from the python object, specifically to create a visual representation of my Random Forest.

OS: a library which allows functionality to run Operating system commands within python.

The approach taken to predict student performance was to use three different data analytics techniques to classify student grades. The three methods used are Random Forest, Logistical Regression and Support Vector Classification.

I chose to do a classification study because I felt it would be a better way of understanding the data. Random Forest was an obvious choice as it is both widely used and well respected for this task. The random forest would allow me to overcome both the tendency of new researchers to overfit data and reduce the chance for sampling errors within my data. I also feel that Random Forest produces a very clear result that can be easily interrupted visually.

I chose to use logistical regression within this project in part because I understood how to correctly approach this technique. As the technique is based on finding the decision boundary to identify the point at which the probability switches from a negative to a positive class. This means we can find a graphical point at which the independent variables point towards a classification result even if no singular variable can on its own predict the outcome. For a classification test like this where the end value is Bayesian (either pass or fail) logistic regression is viable.

The use of SVM as my final data method was picked in part due to it being so well supported within Scikit-learn. SVM as a supervised machine learning system presented a different data analysis challenge which justified inclusion on this merit alone. The support vector would also allow me to better understand the margin of difference between my two classification outcomes.

Exploratory Data Analysis

To better help the reader to understand the dataset being used for this paper an exploration of the data through graphs has been used. This section contains firstly several graphs showing the ratio and distribution of certain important variables. Then there is a section showing the ranking of all the features included in the study and final a Correlation Matrix is included. The correlation matrix is very helpful in showing how different variables might impact each other. This is useful in allowing both researchers and educators to understand how improvements in one area might have a knock-on effect on other important factors.

This section contains some but not all the graphs produced to help readers visualize this data set. In the appendix attached you will find additional useful graphs along with full size versions of all the following.

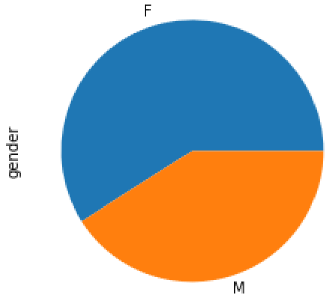


Figure 1: Student Gender Pie Chart

This pie chart shows the split within the sample data between male and female students. We can see that there are more female than male students but male students are still plentiful enough as to not significantly bias the results of the study.

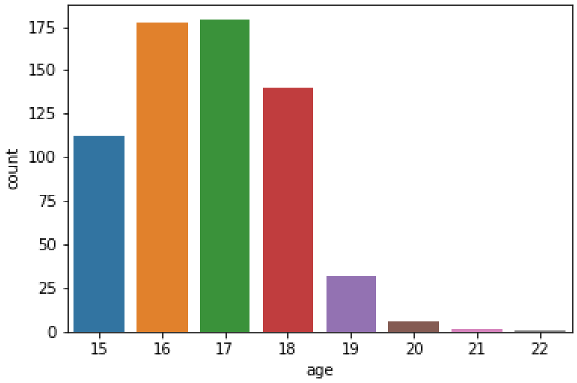


Figure 2: Student Age Bar Graph

This graph shows the breakdown of the ages of the students who took part in the study. Though most of the students where aged 16-18 there were several outliers. Of note to my study is the number of mature students, these students we would expect to have stronger speaking skills and so are likely to do better when tested for a linguistic subject. We do not however have any data to explain why these students where still in school at this higher age and so confounding variables maybe impacting results.

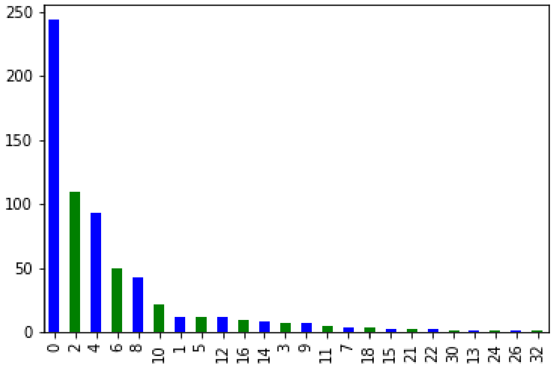


Figure 3: Number of days missed ordered by count.

This graph shows the number of absences per student across the school year displayed. The graph is ordered quite deliberately to show that the vast majority of students had either 0 or very few days of absence over the year. Though some students had over 20 days of absence, this group where in total only 15 students. This means that the fact that student absences appear as important feature be rank to indicate how much this group of outliers impacted the data.

Below are three tables each showing the score distribution of the three grading periods. The scores are broken into their three periods G1, G2 and G3. As we can see the majority of students cluster around the mean with G1 and G2 appearing very similar to a standard distribution. However, we see a more defined bump in the number of 0 scores in G3 along with a sharper decline from the mean score of 11. This indicates that the final test might have been harder than the previous two causing students on the boundary to be pushed more definitively into one grade or another.

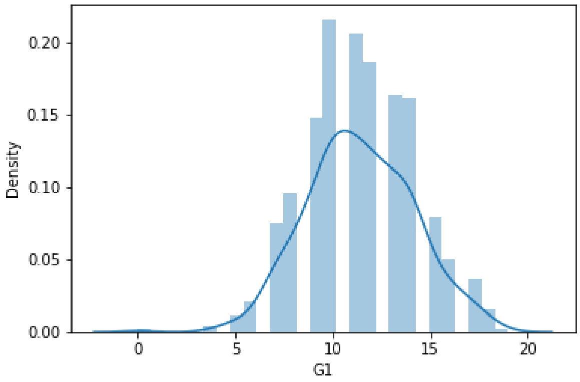


Figure 4: Student G1 Score Distribution

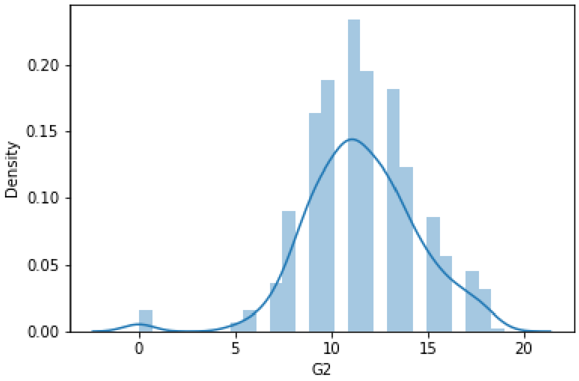


Figure 5: Student G2 Score Distribution

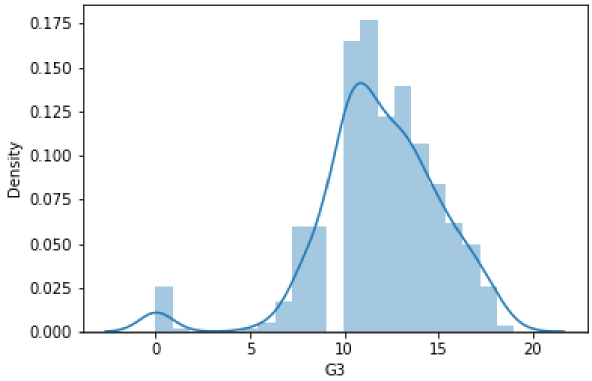


Figure 6: Student G3 Score Distribution

Next, I have a very interesting graph which shows the number of students who achieved above the mean grade on their G3 test broken down by school. This graph clearly shows that one of the schools (GP) had a far higher success rate than the other and that we can assume this correlation to be causal as it goes far beyond a reasonable coincidence. This graph also shows that as a student who achieved the mean was considered to have "passed" that there where inevitably going to be more passes then failures.

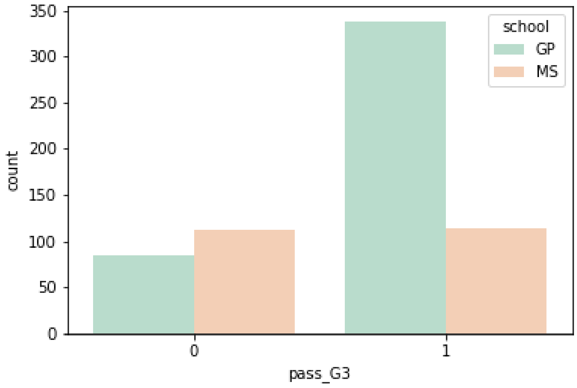
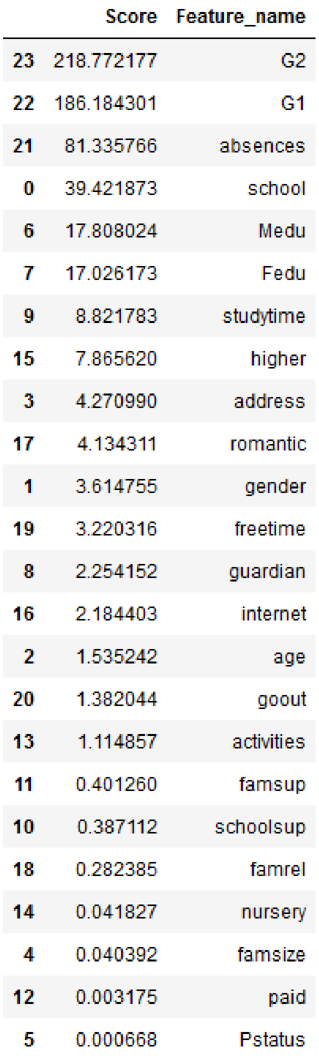


Figure 7: Pass rate of G3 by school 0=fail, 1=Pass

Using the KBest feature score function within sklearn I was able to assess the feature's statistical value within my data set. This used the Chi-squared algorithm to give a probability-based score for each feature. This score is not able to determine the actual impact of a feature on data. However, it can give us an understanding of the confidence level we should have that the data selected will be statistically significant. We can say for instance that features like G2, G1, school, absences and parent education (Medu, Fedu) are likely to be relevant to the model's classification. This also shows us that non-school based variables seem to have very little impact on performance. This is slightly counter-intuitive as the Kbest score implies that private tuition (paid) has little to no impact at all, whereas most teachers would say it has a large impact.



	Score	Feature_name
23	218.772177	G2
22	186.184301	G1
21	81.335766	absences
0	39.421873	school
6	17.808024	Medu
7	17.026173	Fedu
9	8.821783	studytime
15	7.865620	higher
3	4.270990	address
17	4.134311	romantic
1	3.614755	gender
19	3.220316	freetime
8	2.254152	guardian
16	2.184403	internet
2	1.535242	age
20	1.382044	goout
13	1.114857	activities
11	0.401260	famsup
10	0.387112	schoolsup
18	0.282385	famrel
14	0.041827	nursery
4	0.040392	famsize
12	0.003175	paid
5	0.000668	Pstatus

Figure 8: Data Feature ranked by Score

Finally, within the exploratory data analysis we have the correlation matrix which is a table which computes the correlation coefficients for the different variables. With a correlation matrix a score closer to 1 is considered to have very high correlation between the two features whereas a score close to 0 is considered to have no correlation. It is possible to have a negative correlation where the lack of one feature increases the result of the other.

To pick out a few the more important correlations within my dataset I would like to first highlight the very strong correlation between Mother's education level and Father's education level. This correlates at 0.65 which shows that households tend to be built around similar levels of educational attainment. We could therefore theorize that children will be impacted both positively and negatively by the education level of their parents reinforcing views on educational value.

Next it is worth highlighting those features that correlate above average with all three grade scores. These are; Medu, Fedu, study time, higher (the students interest in higher educational study) and most obviously the grade scores themselves. Of these Higher and study time have the highest correlation which indicates a connection between attitude towards studying and results. This correlation is not surprising, however the fact that Pass_G3 and G3 do not correlate as highly as G2 and G3 might surprise some readers. The reason for G2 correlating so much higher is due to it being measured on a scale of 0-20 the same as G3 where as Pass_G3 is measured on a binary scale and so the nuances of the different student's grade are missed. A student might score a 12 or a 20 and still have the same "pass" result.

Finally, it is worth discussing a number of the negative correlations within the dataset. Specifically the negative correlation that both school and age have with all 3 grade scores. The negative correlation of age ranges from -0.11 to -0.17 and so indicates a slight tendency towards older students performing worse. This we can assume is due to the underlying reason which has caused the student to be taking this exam at an older age. These could be anything from health issues through to family disruption or even having migrated from a different country. We see a much stronger negative correlation between school and the three grade scores. This is at its most extreme with the -0.31 correlation between pass_G3 and school. I do not fully understand why this is but it could be down to the normalization of the data. It is possible that if the school normalized as 0 performs better than the school normalized as 1 that this accounts for the negative inversion. As we have seen from the count graph in this section it is clear there is a correlation and future study may help to examine this apparent contradiction.

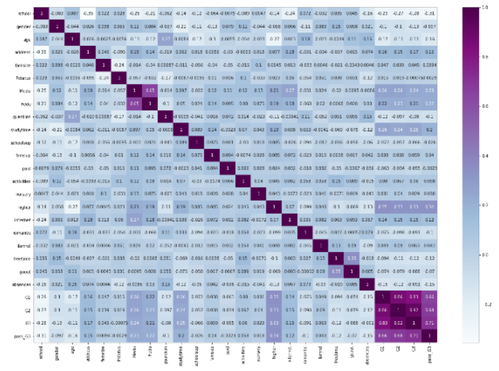


Figure 9: Correlation Matrix, please see appendix for full version

Testing and Results

For a classification problem such as this the aim is to train the machine learning algorithm to have perfect sensitivity and specificity to the data. In the real world very, few algorithms achieve both at the same time. Sensitivity to data is how likely the algorithm is to correctly predict a True Positive result. Specificity to data is how likely the algorithm is to correctly predict a True Negative. For my testing I used the recommended test to train split, 20:80, and then applied a model fit before running the analysis through a cross validation metric.

The first attempt at data analysis yielded very poor accuracy rates and so the default algorithm's for both logistic regression and SVC where changed. I chose to use the Liblinear algorithm for Logistic Regression which yielded an 86.9 accuracy. For SVC I used the linear kernel which produced a similar 86.15 accuracy. I chose to use the decision tree classifier which also produced an accuracy of 86.9.

The full decision tree was generated and can be seen within the attached appendix. At the top of the tree was the G2 result which is to be expected, especially as the next node was G1 result. This reinforces the clear correlation found in the exploratory analysis that past grades are excellent indicators for future results. This can be assumed to be in part due to them being points where unobservable variables (e.g. exam related anxiety) can be measured within the dataset.

I now present the Confusion Matrix heatmap of each algorithm, with SVC first followed by the decision Tree and finally logistical regression. The heatmap shows the rate at which the model achieved True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN).

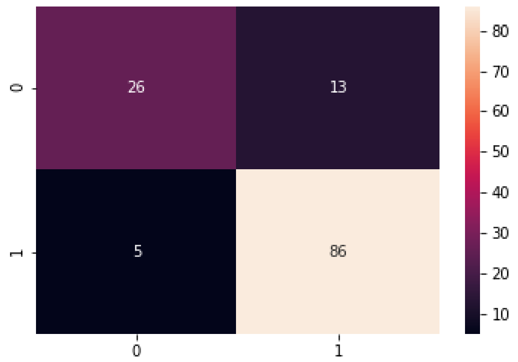


Figure 10: SVC Confusion Matrix

This matrix for the SVC algorithm shows that of the 39 negatives in the sample of data the algorithm predicted 26 correctly as TN with 13 incorrect as FP. This gave the model a true negative rate of 66%. The sample also had 91 positives of which the model predicted 86 as TP with only 5 FN. This gave the model a true positive rate of 94%. This means the model was much better at predicting Positive results although this is almost certainly due to the much larger number of positives within the sample. This issue of oversampling positives can be found across all three matrix's and so means the models can be fairly compared as all three have the same complicating factor.

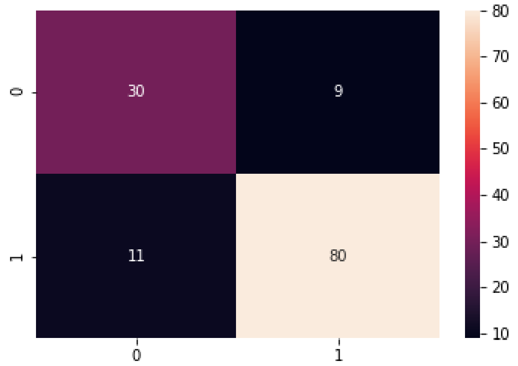


Figure 11: Tree Confusion Matrix

This matrix for the Decision Tree was based on the same sample as the previous matrix. In this sample of 39 real negatives, it predicted 30 TN and 9 FP. This is a true negative rate of 76%, this is 10% better than the results found in SVC testing. The Decision Tree however was worse at predicting true positives as out of a possible 91 positives it predicted 80 TP and 11 FN. This gave the model a True Positive rate of 87%, which is a 7% dip in accuracy compared with SVC. It is plausible to assume that this model was less affected by over sampling of positives due to the much closer accuracy between true negatives and true positives.

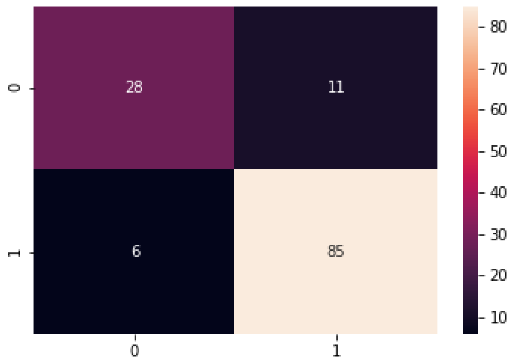


Figure 12: Logistical Regression Confusion Matrix

The logistical Regression model is the most accurate of the three. It has a True positive rate of 93%, after predicting 85 TP results to only 6 FN. The model also did well in predicting negatives with 28 TN and 11 FP giving it a True negative rate of 71%. With the true positive rate being only 1% worse then SVC and the true negative rate only 5% worse then the decision tree, it is possible to state that logistical regression had a better accuracy overall even though it was not the best at anyone classification task.

To better visualize this difference between the models in terms of accuracy I have used am ROC curve. With this graphical technique we can show how well a model is able to fit its data to the truth based on the expected probability of an outcome. A perfect algorithm would have an AUC or area-under-curve of 1 as it would have only true positives and true negatives. An algorithm that was picking at random would be expected to have an AUC of 0.5, assuming it is randomly correct 50% of the time.

Below are all three ROC curves in the same order (SVC, Tree, Log):

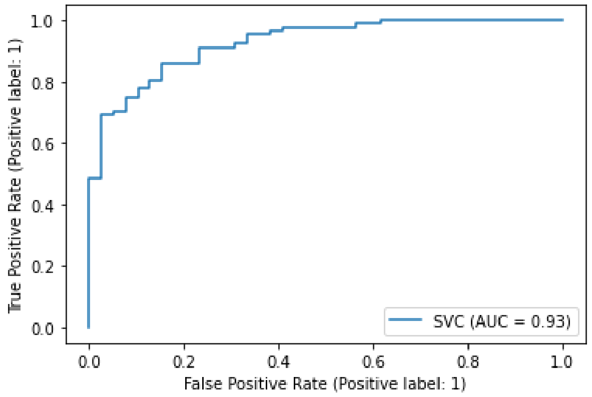


Figure 13: SVC ROC

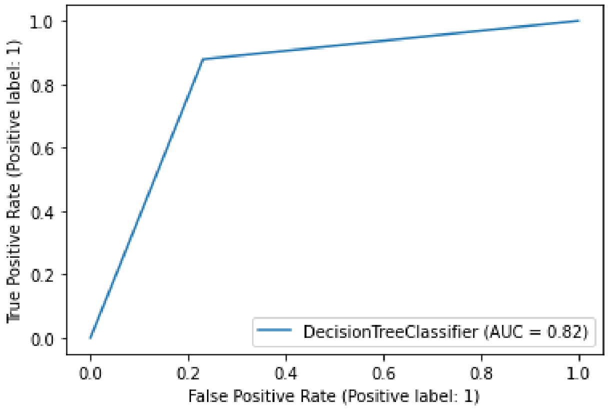


Figure 14: Decision Tree ROC

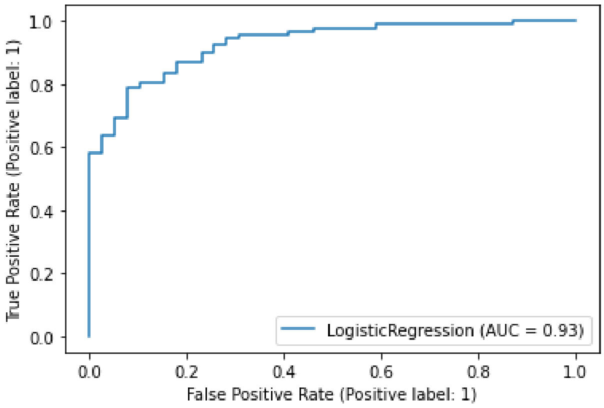


Figure 15: Logistical Regression ROC

As is shown in the above graphs, the area-under-curve result of both the SVC and the Logistic Regression are identical at 0.93. This is a very high AUC curve and shows just how well these algorithms can identify True Positive results. It is not surprising that the Decision Tree model does slightly less well with an AUC of 0.82. This result is still very good and shows the accuracy of the model but demonstrates that is perhaps is less good according to this analysis metric.

Conclusions

To conclude I feel this project has delivered its aim of providing a context for predicting student performance. As shown above all three of the models tested where able to have a very high accuracy in correctly predicting whether a student would get a pass or a fail grade in their final assessment. As the decision tree shows there are several categories which have a large impact on a correct prediction. The most important categories are G2 and G1 results which underlines the direct link between past good performance and future success. Also, the decision tree showed a strong preference to using the data on student absentees, this combined with a high feature rank score and be a good indicator of the importance of time spent in school. The mean number of absences was 3 which indicates that the impact of this feature was for students with very high number of missed school days.

I feel that off the three models examined it is possible to draw a few conclusions. Firstly, the decision tree model has far greater clarity in terms of why it has made its decisions and so would be easier to use to explain machine learning results to a non-computer scientist. The ability to easily point to each node in the decision process reduces the need to rely on purely numeric rankings. Secondly, the support vector model and the logistical regression had very close results in their ability to predict both True Negatives and True Positives. This high accuracy that both models showed is impressive and very useful in making sure that my results can be viewed as valuable.

Thirdly, I feel it is worth examining some of the short falls with this project. It would have made for a better level of analysis if I had been able to conduct primary research independently and therefore had some way of verifying the integrity of the data. Conduct this paper using firsthand data would have also opened the possibility of conducting follow analysis of the longer-term impacts of these results. Specifically, I would like to firstly if the model could predict the end of year results for a future class based on this data and to see how closely G3 results correlate to higher educational success.

References

[1] Baker, R. S., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A Review and Future Visions. *Journal of Educational Data Mining*, *1*(1), 3-17.

[2] Salikutluk, Z. and Heyne, S., 2017. Do Gender Roles and Norms Affect Performance in Maths? The Impact of Adolescents' and their Peers' Gender Conceptions on Maths Grades. *European Sociological Review*, 33(3), pp.368-381.

[3] Jensen, V., 2013. Working longer makes students stronger? The effects of ninth grade classroom hours on ninth grade student performance. *Educational Research*, 55(2), pp.180-194.

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

[5] Kabakchieva, D., 2013. Predicting Student Performance by Using Data Mining Methods for

Classification. *Cybernetics and Information Technologies*, 13(1), pp.61-72.

[6] Breiman, L., 2001. *Machine Learning*, 45(1), pp.5-32.

[7] Belgiu, M. and Drăguţ, L., 2016. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, pp.24-31.

[8] Pal, M., 2005. Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1), pp.217-222.

[9] Novakovic, J. and Veljovic, A., 2011. C-Support Vector Classification: Selection of kernel and parameters in medical diagnosis. *2011 IEEE 9th International Symposium on Intelligent Systems and Informatics*,.

[10] University of Southampton, 1998. *Support Vector Machines for Classification and Regression by Steve R. Gunn*. Technical Report. Southampton: University of Southampton.

[11] Dreiseitl, S. and Ohno-Machado, L., 2002. Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, 35(5-6), pp.352-359.

[12] https://www.kaggle.com/larsen0966/student-performance-data-set

Appendices

Github Repository: https://github.com/tom-jackson1995/expert-eureka.git

This Repository contains not only the jupyter notebook file used to create the code but also the CSV file with the data set used.

I have also attached copies of the diagrams used in this paper in case any images are difficult to understand due to compression.