MAST90106 - Data Science Project Part 1
Group 25

Proactive Student Support And Next Best Action

Client - Mr Milad Chenaghlou and Prof. Kate Smith-Miles,
The University of Melbourne

Supervisor - Mr Milad Chenaghlou

Student Name	Student ID
Haopeng Yan	962332
Harshal Shah	1020849
Sudheer Kumar Kolla	1051717
Xiaowen Jin	1023499
Yiming Xu	863672

Contents

1. Introduction

One of the perennial challenges faced by educational institutions is to identify the students who are about to fail a particular course and the students who are progressing smoothly. Studies have shown that not completing the course within a stipulated amount of time and with grades above a certain threshold impacts a student's future career prospects tremendously [Lakkaraju et al., 2015]. A multitude of reasons ranging from mental health issues, study level, personal and family issues, etc. can affect a student's academic growth. To handle these issues and support the students, the educational institutions are putting their best efforts in finding the solution to these problems by developing a successful pipeline that predicts the students at risk and provides them the relevant support.

What is student success? What does it exactly mean? In [Kuh et al., 2006], a definition of student success is synthesized from the literature as *"Student success is defined as academic achievement, engagement in educationally purposeful activities, satisfaction, acquisition of desired knowledge, skills and competencies, persistence, attainment of educational outcomes, and post-college performance".*

"Students are the most essential Asset for any Institute". Thus, the success of the educational organizations depends on the performance of the students and the services they provide to them. The most important factors here are the grades of the students. The process of analyzing the student data and generating meaningful insights from it is known as "Educational Data Mining". These results would help the instructors to change their teaching methodology if needed and the university to improve its services in the desired sectors. As the number of students being admitted to Universities is increasing day by day, it is very difficult to analyze the student's data. The solution to this problem is much needed in today's stressful world to help students to complete their courses on time. It will also help the university to maintain its standard as well as improve its student support services.

The growth of a student is affected by numerous factors. They are school-related, social, economic, psychological, environmental, and personal factors. The factors vary from student to student and the region where the student belongs to. For some students, the economic condition may be quite stable but the student may be going through some family issues. To understand each of the components in detail, a substantial amount of student data is required. From the data perspective, this is a major issue as it deals with confidentiality. Further, understanding the data, handling the dirty data, and dealing with a lack of data are the major issues in this context.

2. Literature Review

Various researchers have directed broad research on strategies for identifying the students at risk in academic context by predicting their performance. [Mushtaq and Khan, 2012] demonstrated that several factors affect the performance of the student in the examinations (CGPA). They are communication, learning facilities, proper guidance, and family stress. They concluded that the factor "Family stress" has the most adverse effect on the student's success. Along the same lines, [Majeed and Naaz, 2018] focused majorly on the grades of internal and external assessments as these two variables contributed 98% to the final accuracy. The final accuracy of the entire model with all features including demographics, grades, social, etc. was 91% using the decision tree algorithm. They mentioned that the result of the above techniques can be improved using different data and better models. Therefore, [Heuer and Breiter, 2018] used 4 different techniques to predict the performance of students by exploring their daily activities in a virtual learning environment. As per the scores, SVM and Logistic Regression were the best 2 models that dealt with the information of the students. This data included two highly important factors which are " Clicks per day" and "Interactions per day" as the system was made for online environments only. The target variable, in this case, was the final grade which is either "pass" or "fail".

[Na and Tasir, 2017] made some improvements in identifying at-risk students in online learning. The data used in this computational model was based on behavior, academics, networking in online courses, emotional balance, and the personal information of the students. [Na and Tasir, 2017] implemented the four different methodologies to predict the performance and provide relevant support to the students in need. The most common strategies suggested are improving study resources, providing extra support, and mentoring. In the same way, [Lane et al., 2019] mentioned some extra dimensions for support to the students. These include networking skills, collaboration, mindset, time management, academic capabilities, professional career, and goals, etc. These dimensions of support for learning were determined after the authors analyzed the data which they collected by conducting consulting sessions, questionnaires, surveys of the students, teaching staff, and all other stakeholders.

To get a deeper understanding of all the existing systems, [Rastrollo-Guerrero et al., 2020] explored 70 different papers. They found out that the majority of the work has been done at the university level and the implementations can be improved if done at the school level as well. The most commonly used algorithm in this student prediction domain is "Support Vector Machine" for linearly separable data. Other algorithms that are used frequently in all existing systems are Logistic Regression, Decision Tree, Naive Bayesian, and Random Forest classifiers. According to them, to

improve the prediction accuracy and to build more scalable and long-lasting systems, neural networks and collaborative filtering techniques should be used. Thus, [Hung et al., 2019] used a novel based predictive modeling system which was based on several methodologies. These include the Support Vector Machines, Random Forest, and Deep Neural Networks. In this case, the final grade was used to decide the "risk" status of the student. The final dataset used consisted of offline and online course information. After modeling, the best results were obtained using the Deep Neural network framework that had an accuracy of approximately 95%. The predictions were perfect according to the information provided by the students and thus it was the most reliable method over the basic machine learning algorithms.

Currently, many researchers believe that there are many different approaches to solve the problem as far as the domain of Educational Data Mining is considered. [Er, 2012], used the "time-variant data set" which only had the factors that would change concerning time. These features do not include age, gender, and other static features. Here, the authors concluded that using the combination of different models was the best way to solve the problem. If all algorithms give the same result then it's a "yes" or else it's a "no". In this context, 'yes' and 'no' were referring to whether the performance was worse, poor, average or not. [Lakkaraju et al., 2015] added some information to improve the results obtained from the previous model. The authors mentioned that using "Information Gain" and "Gini Index" can improve the results obtained from the model. In this context, the Random Forest classifier gave the best results in predicting whether the student will graduate on time or not, based on the "GPA" of the students.

[Sarra et al., 2019] focused more on considering student's performance, motivation, resilience to understand the profile of every student in detail and thereby getting more accurate and finely grained data about the students who are at risk of academic failure. The student profile was assessed using the Bayesian profile regression. The model suggested that the students will have less chance of failing if the instructors make the necessary things clear about what is expected and required to do for the semester. Creating more supportive environments would also help the students to be successful academically. [Jackson and Read, 2012] from the Edith Cowan University mentioned that only three Universities across Australia and New Zealand have an up and running student proactive support system. The ECU's student support system flags the students who are at risk and then helps them proactively to make sure that the students succeed in their academics. The system was named "Connect For Success" and was built based on statistical learning methods that used the demographic and grade data of the students.

In conclusion, both the time-variant and time-invariant data used with Deep Learning methods, Decision Tree, Logistic Regression and Support Vector Machine give the best accuracy and are effective for identifying the students at risk and provide them with the relevant support to achieve their goals.

3. Data

3.1 Data Description

The data sets were obtained from the UCI Machine Learning Repository [Cortez and Silva, 2008]. It contains information about the student grades in two Portuguese schools. In total, there are 649 instances for Portuguese and 395 instances for Maths with 33 different data attributes each which are mentioned along with their type in Tables 3.1, 3.2, and 3.3 respectively. All the data exploration and analysis parts have been performed on the merged data which consists of 382 different instances and their respective data attributes. The column name with suffix 'x' and suffix 'y' are the attributes for Mathematics and Portuguese language subjects respectively.

Table 3.1: Binary data attributes

Binary Data Attribute	**Values**
School	GP, MS
Sex	Male, Female
Address	Urban, Rural
Family Size	<=3 , >3
Parent's Cohabitation Status	Living Together, Apart
School Educational Support (x,y)	Yes, No
Family Educational Support (x,y)	Yes, No
Extra Paid Classes (x,y)	Yes, No
Extra Curricular Activities (x,y)	Yes, No
Attended Nursery School	Yes, No
Planning for Higher Education (x,y)	Yes, No
Internet Access at Home	Yes, No
Romantic Relationship (x,y)	Yes, No

Table 3.2: Nominal data attributes

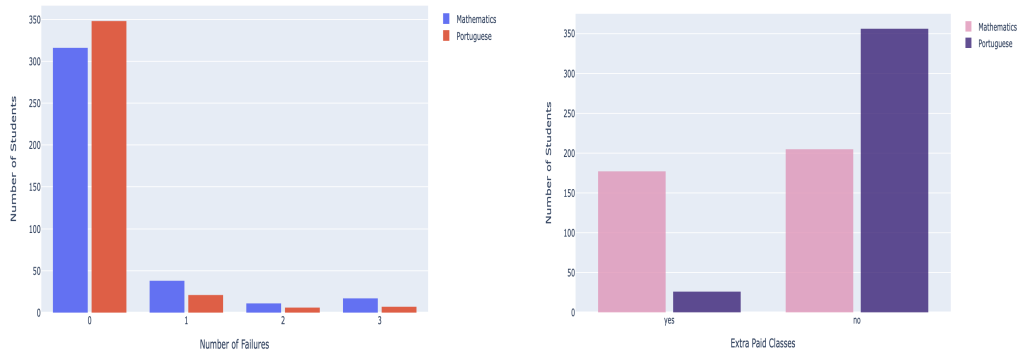
Nominal Data Attribute	**Values**
Mother's job, Father's Job	Teacher,Health Care,Civil Services,At Home,Other
Reason To Choose this School	Close to Home, School Reputation,Course Preference,Other
Student Guardian (x,y)	Mother or Father

Table 3.3: Numeric data attributes

Numeric Data Attribute	**Values**
Age	15 - 22
Mother's Education, Father's Education	0 - Not Educated 1 - Till 4th Grade 2 - 5th to 9th Grade 3 - Secondary Education 4 - Higher Education
Travel Time to School (x,y)	1 - <15 min 2 - 15 to 30 min 3 - 30 min. to 1 hour 4 - >1 hour
Weekly Study Time (x,y)	1 - Less than 2 hours 2 - 2 to 5 hours 3 - 5 to 10 hours 4 - Greater than 10 hours
Number of Past Failures (x,y)	n: if 1<=n<3, 4 otherwise
Quality of Family Relationships (x,y)	1 - Very Bad to 5 - Excellent
Free Time After School(x,y)	1 - Very Low to 5 - Very High
Going Out with Friends(x,y)	1 - Very Low to 5 - Very High
Workday Alcohol Consumption(x,y)	1 - Very Low to 5 - Very High
Weekend Alcohol Consumption (x,y)	1 - Very Low to 5 - Very High
Current Health Status (x,y)	1 - Very Bad to 5 - Very Good
Number of School Absences (x,y)	0 - 93
First period Grade(G1) [x,y]	0-20
Second period Grade(G2) [x,y]	0-20
Final Grade(G3) [x,y]	0-20

3.2 Data Exploration

As the data was collected with the help of questionnaires and surveys, some of the variables consisted of the qualitative data which varies from student to student, which is also known as the time-variant data. After the initial exploration, it was found that there were no missing values and duplicate values in the dataset. All the attributes of the data were understood in detail and basic analysis was performed. The results for some of the attributes are mentioned below. From Figure 3.1(a), it is observed that a lot of students find it more difficult to clear the Mathematics Course as compared to Portuguese. The number of students who prefer going to extra paid classes is dependent on the subject they choose. As it is observed from Figure 3.1(b), the students who prefer paying extra fees, and attending classes are more for the mathematics subjects. For Mathematics, the number of students failing is more and therefore they prefer going to extra classes.

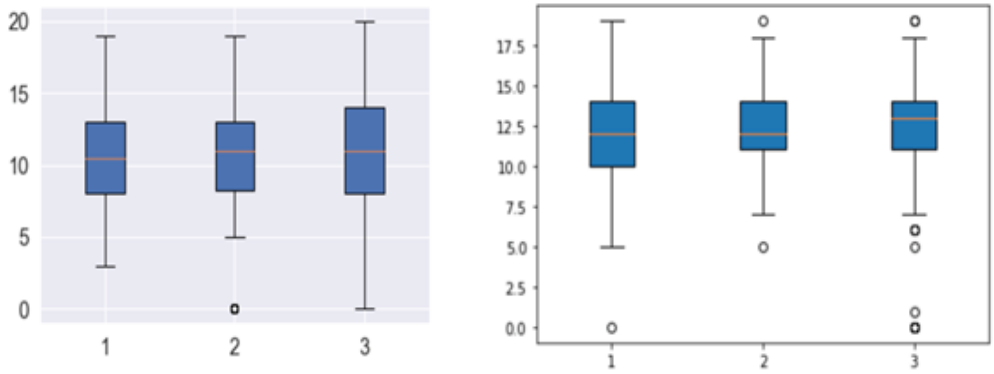


(a) Number of Failures (b) Students Taking Extra Paid Classes

Figure 3.1: Count for Number of failures and Extra Classes

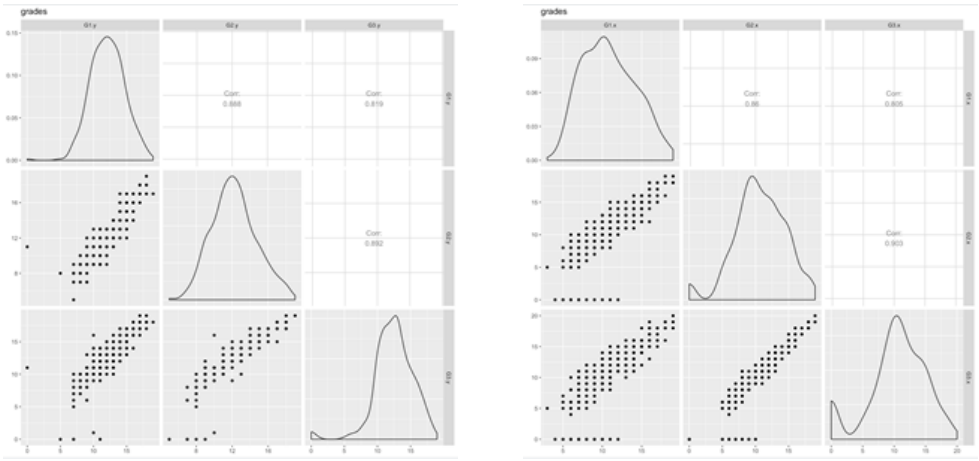
The final target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final grade, while G1 and G2 correspond to the 1st and 2nd-period grades. Figures 3.2(a) and 3.2(b) below show the outliers present in the grades for both the subjects. It can be observed that the average score for Mathematics subjects in all the three periods is approximately between 8 and 14. On the other hand, the grades for the Portuguese subject are gathered around 10-15 points but have more outliers as compared to the scores of Mathematics. The overall results are better for the second subject. There is a high positive correlation between grades G1, G2, and G3 for both the subjects. As G1 scores increase, G2 also goes up. This means that G1 and G2 are linearly proportional. The same hypothesis is valid for G2 and G3 scores as well. Although there are some outliers, it may be because students did not perform well in the exams. Also, the data is normally distributed for the grades and this can be observed along the diagonal in Figures 3.3(a) and 3.3(b)

respectively.



(a) Mathematics subject (b) Portuguese subject

Figure 3.2: Box Plot for all Three Grades in Both the Subjects



(a) Mathematics Subject (b) Portuguese Subject

Figure 3.3: Correlation between all Three Grades

3.3 Feature Engineering

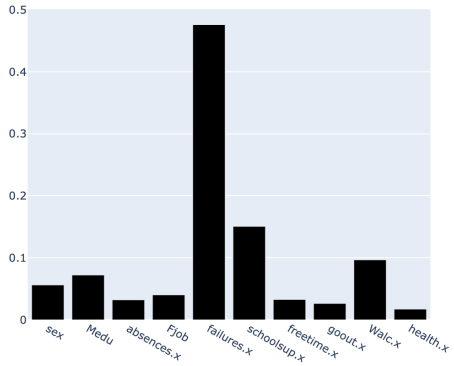
Feature selection is the set of techniques that could be used for identifying the important features for a given problem which can further help in the prediction of the final target variable. The feature importance technique returns scores for each of the features after fitting the input data in some models and that score determines which feature has the highest weightage in the prediction of the target. Amongst the 30 different features (except for G1, G2, and G3), the ones which had more impact on the grades were found out with the help of feature engineering using two different techniques viz. Feature importance using Decision Tree Regressor and Lasso Regularisation. In the first method, all features which have a score of more than 0.00 were considered. The score was very less because the number of attributes are more and some of them are not very relevant.

From the correlation matrix, it can be inferred that the Grade 1,2 and 3 exam scores are highly correlated to each other. In this context, G1 grade should be considered the most important factor for the prediction of G2 score, irrespective of the subject, and other features. In the similar lines, the best features for predicting Grade 3 exam scores are the Grade 1 and Grade 2 scores.

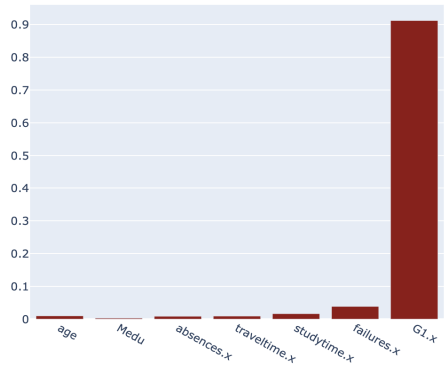
The Figure 3.4(a), 3.4(b), 3.4(c) show the features for predicting Grade 1 exam, Grade 2 exam, and Grade 3 exam scores that have positive scores for Mathematics. As shown in the three graphs below, the features important for the prediction of G1 are Number of failures, school support, Mother's education, health, etc. All the features selected have a good score and help in the prediction of grade 1 score successfully. As mentioned before, the grade 1 score has a great impact on grade 2 score, and similar is the case for Grade 3 score as well. It can be clearly understood from the graphs shown below. For G2 and G3, some more important features apart than G1 and G2 are the Number of absences, Travel time to school, Weekend alcohol, etc. Similar results were obtained for Portuguese language as well.

The second method for feature selection is the Lasso Regularisation. The lasso regularisation heavily penalizes the features or columns that are not relevant to the problem by pushing the value of the coefficient to zero. The amount of penalization depends on the alpha value, higher the alpha value higher the penalties given to the columns, and the number of relevant features obtained would be lesser. In this case, an alpha value of 0.2 was chosen which would ensure that there is a decent amount of penalization on the data, and at the same time, it would ensure that there are quite a few features that could be used in the model. So after applying lasso regularization over the data, the features with higher coefficient value were more relevant in predicting the value of the response variable.

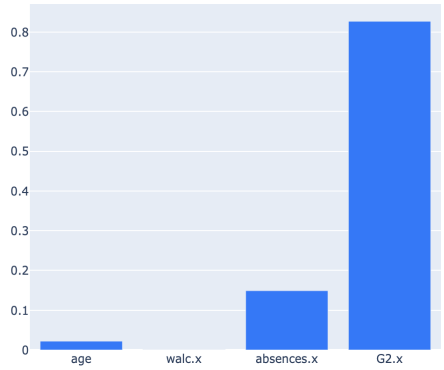
The main difference between the Decision Tree Regressor method and Lasso Regression method is that in this case, the number of features obtained were less but were more beneficial. The reason behind



(a) Grade 1



(b) Grade 2

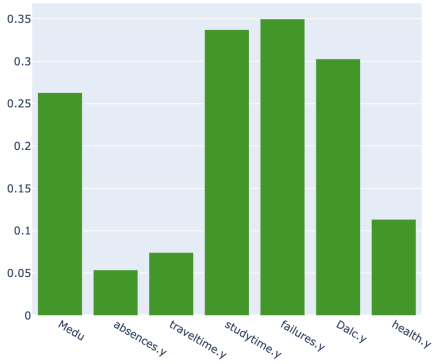


(c) Grade 3

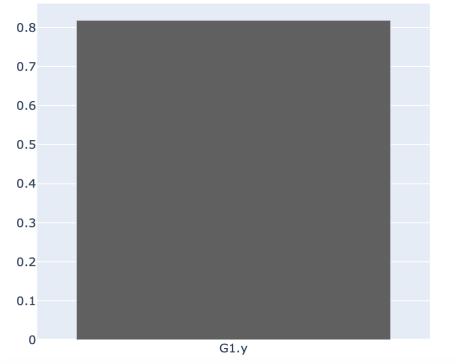
Figure 3.4: Important features for predicting grades for Mathematics using Decision Tree Regressor

this is the penalization concept. Some new features obtained in this case are the father's education, quality of family relation, etc. but the main features remain the same, i.e G1 and G2 are the main features for the prediction of G2 and G3 scores respectively. The most important and common feature between both the techniques for the prediction of G1 score is "the number of past failures" for the student. The graphs Figure 3.5(a), 3.5(b), 3.5(c) are the best features for predicting the grades of Portuguese Language using Lasso and nearly similar results were obtained for Mathematics.

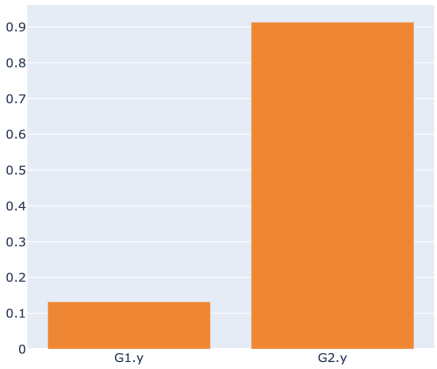
From this step, it was concluded that the Lasso gives fewer features but with more importance. On the contrary, Decision Tree gives more features. To build a basic model, a Decision Tree Regressor was used so that any of the features are not ignored or penalized and can be used to check its effect on the basic model.



(a) Grade 1



(b) Grade 2



(c) Grade 3

Figure 3.5: Important features for predicting grades for Portuguese using Lasso Regression

4. Initial Prototype & Results

For building the initial model, the dataset for the Mathematics subject has been used. Since the aim of the project is to predict where a student is at risk or not, it is quite reasonable to group the final marks (G1, G2, and G3) into groups. Two different groups were formed based on the number of categories which were decided based on the range of scores. Some numerical features in the dataset were also converted into nominal features as they were having more than 2 levels. The converted features are Medu, Fedu, famrel, freetime, goout, Dalc, Walc, health.

Table 4.1: Grouping 1

Group Name / Categories	**A**	**B**	**C**	**D**
Grade scores	0-5	6-10	11-15	16-20

Table 4.2: Grouping 2

Group Name / Categories	**A**	**B**	**C**
Grade scores	0-5	6-10	>10

The Decision Tree Regressor approach was used for selecting important features which were further used by the model for the final predictions. Decision trees were built for each grouping and each period grades (G1, G2, G3). The reason behind the difference in the features obtained in the Feature Engineering step and the ones obtained here despite using the same method is that some of the variables are converted from categorical to Nominal, as mentioned above. Features that were obtained and used in prediction of each period's grades are shown in the table below. These are stored in decreasing order based on their information content scores obtained by the Decision Tree.

Table 4.3: Features for Grouping 1

Period Grades	**Features**
G1	Dalc, failures, freetime, goout, health, Mjob, reason, sex
G2	G1, Walc
G3	absences, Fedu, G2

With most of the features to be nominal and more than 2 levels, a multinomial logistic regression model was chosen as the initial model. Logit model restricts the probability values of predictions to be

Table 4.4: Features for Grouping 2

Period Grades	**Features**
G1	absences, failures, Fedu, Fjob, freetime, Health Medu , Mjob, Paid, Reason, Schoolsup
G2	G1
G3	absences, Fedu, G2

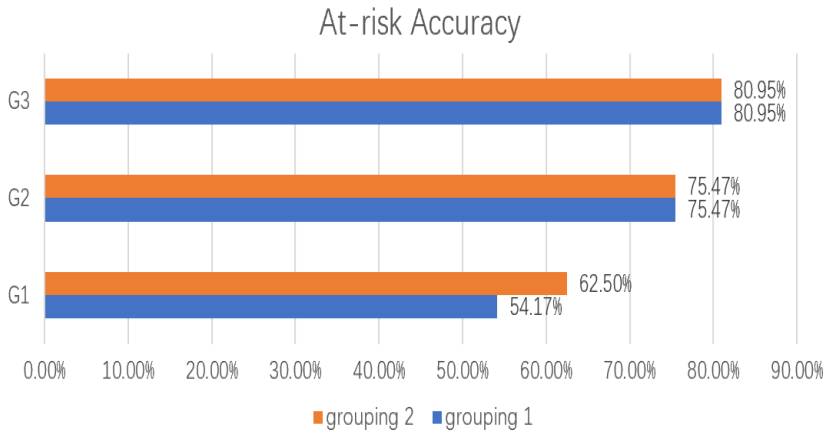
in range 0 to 1, and it is also assumed that the distribution of each period grade follows a multinomial distribution. Dataset was split into 2 parts where 70% used as a training set and 30% used as a test set.

The multinomial logistic regression model was trained initially using the traning set for each period grade with its respective features obtained. As the order of the features matters when it comes to fitting the model, the features were sent as an input to the model in the same way as mentioned in the table. This helped the model to learn the dataset in a better way. Further, the test set was used to evaluate the model accuracy. Accuracy and At Risk accuracy were the two important measures of Scoring and for the performance evaluation of the model. The accuracy was based on overall performance of the model while at-risk accuracy represents the performance of the model particularly on predicting students who did not clear the course. The following two graphs, Figure 4.1(a) and 4.1(b) show the results of the model on two different groupings.

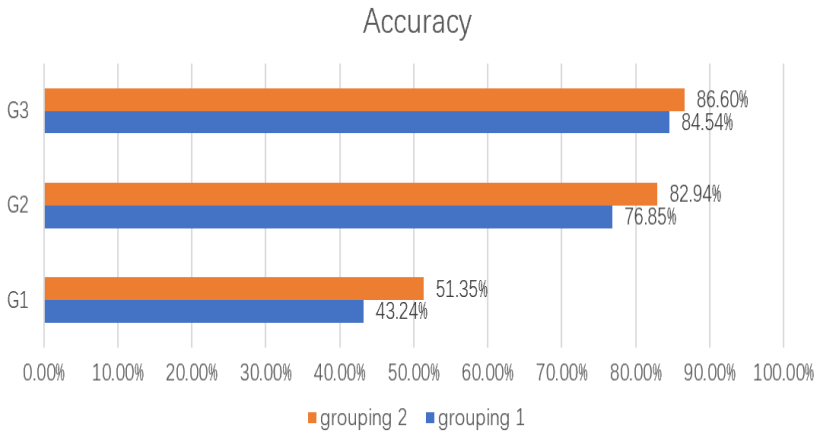
$$At - Risk - Accuracy = \frac{\text{total number of correct predictions of group A and B}}{\text{total number of students with grades in group A and B}}$$

$$Accuracy = \frac{\text{total number of correct predictions}}{\text{total number of predictions}}$$

The accuracy of grouping 1 for the prediction of all period grades is lower than that for grouping 2. For grouping 2, the accuracy for the prediction of G1 scores is 51.35%, which is lower than G2 and G3 which are 82.94%, and 86.60% respectively. From the results, it can be inferred that G1 and G2 are highly correlated to G2 and G3 respectively. The at-risk accuracy of grouping 2 for the first-period grade prediction was higher than grouping 1, while the at-risk accuracies for the second and third-period grades were the same for both groupings, representing the models' ability to predict students at risk for the second and third period with equal probability. From this result, it can be understood that the at-risk prediction worked well for the G1 score.



(a) At Risk Accuracy



(b) Accuracy

Figure 4.1: Accuracy for both Groupings based on prediction of all three grades for Mathematics

5. Timeline and Further Work

The further plan is to work on the initial baseline model and improve its accuracy by using different approaches. The first one is to use the features obtained by the Lasso Regression method and check if there is any improvement. The next step is to collect more data from various sources. This could be the data from the library, the student service desk, etc and find more meaningful information. As it was seen from the literature review that most authors preferred deep learning methods for final processing, our plan is to do the same along with providing the relevant support to the required students once the appropriate dimensions of support to the students are figured out. Once everything works perfectly fine then various combinations of features can be used as the input to the model and, prediction of G3 without G1 and G2 is a new story to explore! At the last, proper documentation along with the desired presentation will be delivered along with the project before the deadline and best efforts will be made to meet all the expectations of the project.

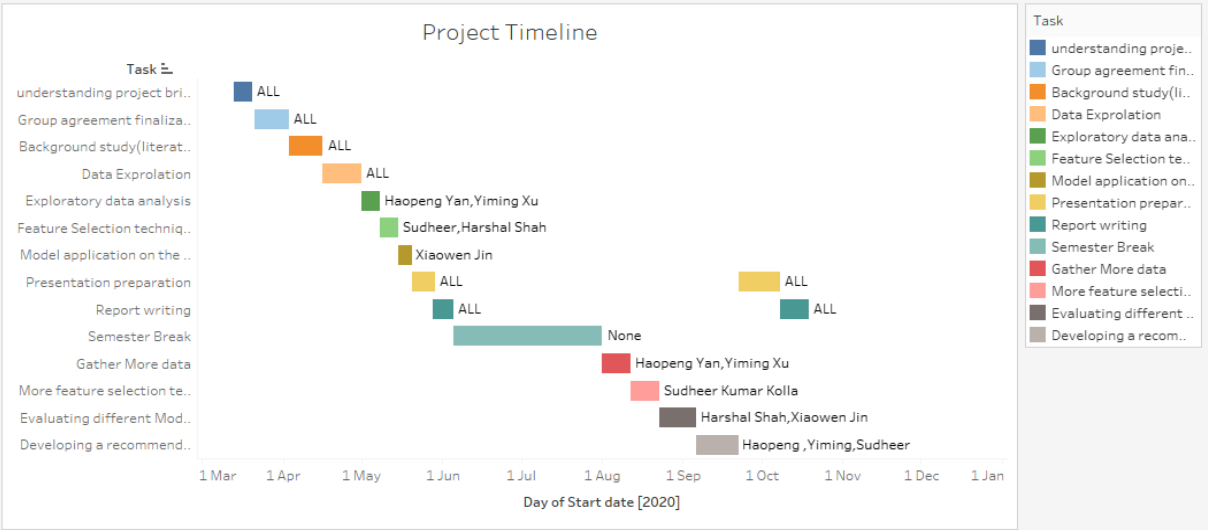


Figure 5.1: Project Timeline

Bibliography

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

[Er, 2012] Er, E. (2012). Identifying at-risk students using machine learning techniques: A case study with is 100. *International Journal of Machine Learning and Computing*, 2(4):476.

[Heuer and Breiter, 2018] Heuer, H. and Breiter, A. (2018). Student success prediction and the trade-off between big data and data minimization. *DeLFI 2018-Die 16. E-Learning Fachtagung Informatik*.

[Hung et al., 2019] Hung, J.-L., Shelton, B. E., Yang, J., and Du, X. (2019). Improving predictive modeling for at-risk student identification: a multistage approach. *IEEE Transactions on Learning Technologies*, 12(2):148–157.

[Jackson and Read, 2012] Jackson, G. and Read, M. (2012). Connect 4 success: a proactive student identification and support program. *ECU: Joondalup, Australia*.

[Kuh et al., 2006] Kuh, G. D., Kinzie, J. L., Buckley, J. A., Bridges, B. K., and Hayek, J. C. (2006). *What matters to student success: A review of the literature*, volume 8. National Postsecondary Education Cooperative Washington, DC.

[Lakkaraju et al., 2015] Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., and Addison, K. L. (2015). A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1909–1918.

[Lane et al., 2019] Lane, M., Moore, A., Hooper, L., Menzies, V., Cooper, B., Shaw, N., and Rueckert, C. (2019). Dimensions of student success: a framework for defining and evaluating support for learning in higher education. *Higher Education Research & Development*, 38(5):954–968.

[Majeed and Naaz, 2018] Majeed, I. and Naaz, S. (2018). Current state of art of academic data mining and future vision. *Indian Journal of Computer Science and Engineering*, 9(2):49–56.

[Mushtaq and Khan, 2012] Mushtaq, I. and Khan, S. N. (2012). Factors affecting studentsâ^{TM} academic performance. *Global journal of management and business research*, 12(9).

[Na and Tasir, 2017] Na, K. S. and Tasir, Z. (2017). Identifying at-risk students in online learning by analysing learning behaviour: A systematic review. In *2017 IEEE Conference on Big Data and Analytics (ICBDA)*, pages 118–123. IEEE.

[Rastrollo-Guerrero et al., 2020] Rastrollo-Guerrero, J. L., Gómez-Pulido, J. A., and Durán-Domínguez, A. (2020). Analyzing and predicting students' performance by means of machine learning: A review. *Applied Sciences*, 10(3):1042.

[Sarra et al., 2019] Sarra, A., Fontanella, L., and Di Zio, S. (2019). Identifying students at risk of academic failure within the educational data mining framework. *Social Indicators Research*, 146(1-2):41–60.