Analysis of Student Performance Data to Predict Grades & Determine Common Attributes via Classification & Clustering

https://github.ncsu.edu/bbond/G20-ALDA-Project

Brody Bond
Department of Computer Science
North Carolina University
Raleigh, NC, Wake
bbond@ncsu.edu

Avi Choksi
Department of Computer Science
North Carolina University
Raleigh, NC, Wake
ajchoks2@ncsu.edu

Anant Patel
Department of Computer Science
North Carolina University
Raleigh, NC, Wake
apatel29@ncsu.edu

1 BACKGROUND

1.1 PROBLEM

Identifying the key factors that contribute to student success in academics is a critical challenge in education. However, this can be a challenging task with the number of variables involved with student success. By looking at and understanding the features and patterns of successful students, educators can create new strategies to help and support struggling students and improve educational outcomes. Using machine learning and machine learning techniques such as classifiers and clustering, there is a better and more efficient approach to discovering hidden patterns and gaining insights from student data.

This study aims to leverage various machine-learning techniques to analyze a comprehensive dataset of student data to uncover information that correlates with student success and uncover groups of students that exhibit similar patterns and outcomes.

1.2 RELATED WORK

In recent years, research studies have utilized machine learning techniques to learn more about student performance based on various variables [2, 3]. Researchers have leveraged techniques such as decision trees and neural networks to accomplish this task [1].

Researchers have also leveraged classification trees to help with data mining tasks. In a paper by Chong Ho Yo et al., classification trees were used to predict retention rates for students [2]. During the construction of the tree, the researchers used entropy as the splitting criterion and pruned the classification tree for their final model. Nahar et al. also used multiple classification techniques, such as random forest and decision tree classifiers, to identify what student attributes predict a student's future performance [4].

Methods like clustering have been used to learn about patterns in learning and engagement patterns. For instance, Kabra and Bichkar utilized k-means clustering to identify groups of students with similar learning levels in a web-based educational system [3]. Omar et al. used k-means clustering alongside the elbow method [9] to organize students based on their test scores to profile various students and analyze their performance. Another algorithm, fuzzy clustering, was used by Shahiri and Husain to account for imprecision and uncertainty in their student data [5].

For our study, we plan to use similar methods for creating classification trees and clustering models to help find relevant and useful conclusions.

2 METHODS

2.1 NOVEL ASPECTS

The first novelty aspect chosen was using our dataset to explore the hypothesis of which attributes are a larger factor associated with the success of a student. For this novelty aspect, we plan on analyzing the data using a decision tree classifier to predict the outcomes of various students. We will then look at the resulting tree and analyze the information gained from various factors to determine what the highest contributing factors are to a student's success.

The second novelty factor is creating new knowledge through analytics and pattern-finding. We plan to cluster students by their attributes and analyze the clusters to learn about common features among students. We can look at the most common attributes within clusters and determine relations and patterns between students. Additionally, the clusters can give us a broader, more abstract sense of which attributes may have an association with a student's success. This can help create new conclusions that give a better understanding of the data and how certain features are more impactful.

2.2 APPROACH

For our approach, we have decided to use machine learning techniques, namely decision trees and k-means clustering, to learn more about our data.

Decision trees were the first technique we used in our study. The data had to be encoded into integer values first, followed by splitting it into training and test data. For training and testing, we used a 25% training split. We then tested four different implementations of decision trees - a Gini with a fixed depth, a Gini with a minimum impurity, a gradient-boosted tree with varying learning rates and n-estimates, and a random forest tree with a minimum impurity. All trees used the same seed for testing.

K-means clustering is the second technique that we have included in our study. For this, we have decided to run some preprocessing by converting binary attributes into numeric data and assigning dummy variables to the dataset. After preprocessing, the dataset will be transformed using Principal Component Analysis (PCA) to decrease the number of features in the data while still containing the patterns needed for clustering. To run the clustering, we opted to use a KMeans algorithm on the transformed dataset. Finally, to analyze the attributes in each cluster, the centroids will be run through a PCA inverse transform and analyzed to find the attribute means for each cluster.

2.3 RATIONALE

We chose to use decision trees due to the relatively small size of the dataset we intended to work with. We tested four variations of trees, as mentioned previously. The first two were Gini trees, which utilized minimum impurities and maximum depth; a set depth was used as the basis for our tests. The minimum impurity tree was utilized as it stops its growth if the impurity is too great. The gradient-boosting tree was used because the model corrects errors from previous predictions, but we had to be careful not to overfit the model. A random forest was also used due to its ability to aggregate the output of multiple trees into a better mode, and it also utilized a minimum impurity for the same previously stated reason. The latter three were used because they should, in theory, have a higher prediction accuracy when compared to the first tree (set-depth).

K-means clustering was chosen for a few reasons; the first reason is that it is computationally efficient, making it quicker for datasets that have multiple features like ours. The second reason is that K-means uses distance for clustering. K-means can accurately assign each point to a cluster since our data consists of mainly binary and numeric attributes. The final reason is that K-means adopts the method of hard clustering; by assigning each data point to a singular cluster, we can ensure that we can precisely analyze the attribute values for each cluster without worrying about separate clusters being influenced by similar data points.

3 PLAN & EXPERIMENT

3.1 DATASETS

The dataset we settled on is about student achievement in secondary education of two Portuguese schools with demographic features on students. The data is collected from school reports and questionnaires. The datasets focus on performance on two distinct subjects: Mathematics and Portuguese.

Our dataset contains 30 features. They touch on student information such as type of school, gender, age, address, parents' education, parents' occupation, family size, the reasoning for the school, travel time, study time, number of failures, extra education support, daily educational support, extra paid classes, extra-curricular activities, attended nursery school, wanting to take higher education, internet access at home, in a romantic relationship, quality of family relationships, free time, going out with friends, workday alcohol consumption, weekend alcohol consumption, health status, and absences. The dataset contains 395 students, and each has the features listed above. We are using this data to find out the best features that define student success through decision trees and clustering. We can use this dataset to generate new knowledge through the various techniques that we plan to employ.

For our preprocessing, we went through a few different steps. The first step was to convert any non-numerical binary attributes to their numerical counterparts (0 and 1); the exact transformed values are shown in Table 1 below. The second step was to drop the G1 and G2 grades as they were a breakdown of the final grade, G3, and not factors we could use in our models. We then took the G3 grade and converted it into a binary attribute with a halfway split; this was mainly used to reduce the number of potential labels from 20 as it was causing our classifiers to overfit and produce inaccurate results. Finally, we took any non-numerical, non-binary attributes and replaced them with dummy variables so our algorithms could easily analyze them.

Both the decision trees and k-means clustering algorithms used the previously described preprocessing; however, the k-means clustering algorithm also used principal component analysis (PCA) for additional preprocessing. This was mainly used to procure a more straightforward, and more accurate, visualization of the clusters while maintaining major trends or patterns seen in the dataset.

Feature	0	1
school	GP	MS
sex	M	F
address (urban or rural)	U (urban)	R (rural)
famsize (≤ 3 or > 3)	LE3 (≤3)	GT3 (>3)
Pstatus (parents living together or apart)	A (apart)	T (together)
schoolsup (extra educational support)	yes	no
famsup (family educational support	yes	no
paid (extra paid classes)	yes	no
activities	yes	no
nursery (attended nursery school)	yes	no
higher (plans to go to higher education)	yes	no
internet (access to home internet)	yes	no
romatic (has a romantic relationship)	yes	no

Table 1: Binary Transformations

3.2 HYPOTHESES

One of the biggest questions that we have is what features are the most important to the success of students. Looking at the dataset we have, many features can influence how well a student does. Does something like financial status matter more than their marital status? Are these features not as important?

Using k-means clustering, we can identify distinct groups or clusters of students with similar profiles. These clusters will reveal the key features and factors that differentiate successful students from those who are not as successful academically. We

hypothesize that clusters that correspond with high academic performance will have a combination of features such as high time spent studying, low past failures, and high academic support. We also think that the inverse applies to those who struggle with academics. Those who do not do as well might have less time spent studying compared to those who do better.

Using a decision tree, we can identify what attributes of a student are more important to achieving a better academic performance. The tree can also help create preliminary predictions of how a student will perform, which may require early intervention. We hypothesize that the decision tree will show that school-related features, such as the number of failures, extra education support, and study time, will influence the student's grade more than attributes such as romantic relationships, free time, and going out with friends.

Our hypothesis underlines that there will be natural groupings in the student data, allowing for clusters to be analyzed and provide good insight into factors that drive student success. We also believe that academic features will impact the clusters more. Our hypothesis also states that some attributes will have a large influence on a student's success, and most of the influential features will be school-related.

3.3 EXPERIMENTAL DESIGN

Due to the multiple tree types, each tree had to be manually tuned based on previous test results and educated guesses. Each tree's tests were run on the same five seeds. For the Gini tree with a set depth and n-estimators, values of 2, 3, 4, 5, 7, and 9 were used. For the trees that utilized a minimum impurity or a learning rate, arbitrary values between 0.001 and 0.03 were used. When a roughly optimal solution was found, it was chosen as a basis, and results for plus/minus 0.002 were tested in 0.0005 increments.

Running the experiment for K-means clustering is relatively linear. The first step is analyzing the data and fixing any missing values, encoding non-numeric values, and removing unique values. For the dataset we are using, we preprocessed the columns by converting all the binary outcomes to 0 or 1 and used dummy variables for the nominal features. The preprocessed dataset was then run through PCA due to the large number of attributes the dataset contained; running the PCA gave us two components that we then used to run K-means clustering. A cluster count of 3 was chosen using the elbow method. After clustering the data, we ran an inverse PCA transform on the cluster centroids to analyze the distribution of attributes for each cluster.

4 RESULTS

4.1 EXPERIMENTAL RESULTS

To determine the most effective type of tree, rough estimates of best hyperparameters were found using five set seeds - 7, 1337, 999, 2, and 9. The results for specifically seed 7 can be seen below in Table 2, while the results for the other seeds can be seen in Appendix A. Of note, any combination of values for the gradient boosting tree's learning parameters and n-estimators, ranging from 0.0005 to 0.0045, and 2, 3, 4, 5, 7, and 9, respectively, had no effect on the training and test accuracy. For this reason, we've omitted the parameters from the table but have recorded the training and test accuracy.

Seed 7	Depth	Impurity	Train Accuracy	Test Accuracy
Gini - Depth	3	N/A	0.753	0.747
Gini - Impurity	N/A	0.01	0.767	0.758
Gradient Boost	N/A	N/A	0.649	0.737
Random Forest	N/A	0.0015	0.983	0.747

Table 2: Seed 7 Rough Estimates

Testing showed that a random forest consistently produced the highest accuracy. Thus, our next step was determining what impurity generated the highest accuracy. To determine this, we averaged the test accuracy for seeds 0-199 with all default RandomForestClassifier parameters, 25% testing accuracy, and impurities ranging from 0.0005 - 0.0095. Per Figure 1, there is a general upward trend in accuracy up to the impurity value of 0.0025 with a 70.2% testing accuracy, after which testing accuracy decreases.

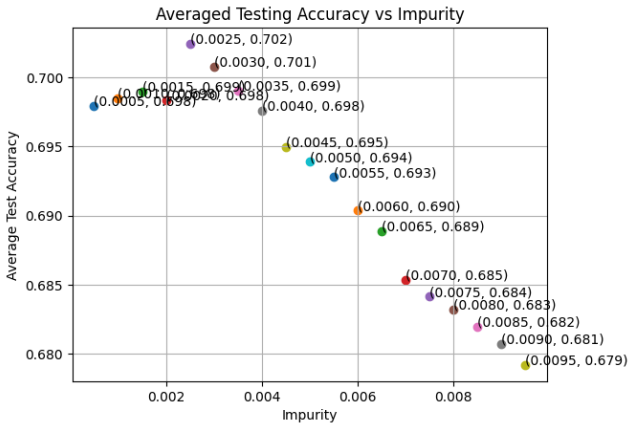

Figure 1: Average Testing Accuracy vs. Impurity with 25% Testing Data

We then computed the average feature importance for random forest trees using the best impurity, 0.0025, over seeds 0-199. We found that the features 'failures' and 'absences' were the two most important features with values of 0.100 and 0.087, respectively, as noted in Table 3. The 'failures' feature indicates the number of failed classes where the values are the number of failed classes up to 4 failed classes; if more than 4 classes are failed, the value is still 4. The 'absences' feature indicates the number of absences, and the values range from 0 to 93. As mentioned in the Preprocessing section, further experimentation was done to test if grouping the values of absences increased accuracy, but when tested over 30 seeds, it generally decreased accuracy by 1-2%.

Feature	Importance
failures	0.100
absences	0.087
goout	0.057
age	0.054
freetime	0.038
health	0.038
Medu	0.037
Fedu	0.036

Table 3: Feature importance

The K-Means clustering graph shown in Figure 2 depicted that three well-defined clusters were present in the dataset.

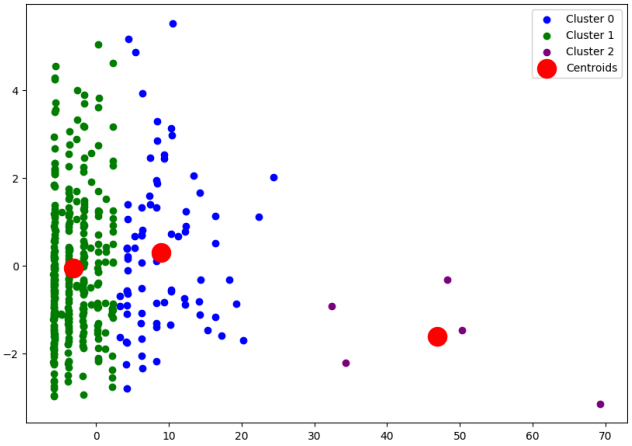

Figure 2: K-Means clustering of dataset

The inverse transformed centroids showed the means of each attribute for the individual clusters. These attributes are seen in Tables 4 and 5, which show how each cluster contains a few attribute values that differentiate it from the others.

Attribute	0	1	2
school	0.097532768	0.12510079132206997	-0.11125975
sex	0.5356808050625392	0.5186843936100518	0.8672754328856636
age	17.04165914201084	16.59150515643772	17.570103231573157
address	0.24651870360311295	0.21633612790086484	0.2368242973561231
famsize	0.6842218597186696	0.7193342205526149	0.6607510424059578
Pstatus	0.8531758516504624	0.9116659949854722	0.6375927131603313
Medu	2.813459294	2.716863598677679	3.7197890440870642
Fedu	2.499753316832868	2.520110715	2.9611540596786936
traveltime	1.4662570427093513	1.446558436	1.2493245845423624
studytime	1.9257381416033668	2.065033306760463	1.9839837482288658
failures	0.4371854409056832	0.30672090491465176	0.36224393360245805
schoolsup	0.8715896189955963	0.8721125150367449	0.7836947430005016
famsup	0.3891505110351564	0.38924343914316883	0.24051718218263493
paid	0.5435127092512883	0.5422365722449802	0.48487394538935363
activities	0.5056910202272785	0.48721698831789784	0.49779559427200376

Table 4: First 15 attribute influences for K-Means clusters

nursery	0.20866392728270106	0.20587950735674163	0.096290823
higher	0.073475003	0.044212459	0.077374974
internet	0.1298992811139733	0.18090254384305132	-0.084145564
romantic	0.5824634064455676	0.6944471605790502	0.24725829179649966
famrel	3.8976113138916757	3.9602972237068825	3.7123282898695327
freetime	3.2353340949787732	3.246501979	2.5537653345227653
goout	3.2829109909825758	3.068113198293176	2.8074462915640317
Dalc	1.7044116294264278	1.4219176726937892	1.514019430815698
Walc	2.6847142158946533	2.1871047072546452	2.344558794
health	3.570649702308098	3.559802319541929	2.959691475
absences	14.662666540665311	2.6033159534336616	52.57987611902166
grade	0.6054971207175014	0.6907980479696272	0.4996272798535206
Mjob_at_home	0.13253767696828503	0.15598628034930312	0.012937507526849384
Mjob_health	0.069373337	0.091370706	0.029442275
Mjob_other	0.38436928463286774	0.34822119961064435	0.45314745300972203
Mjob_services	0.2730083235779633	0.25666248660171953	0.3155234860507286
Mjob_teacher	0.14071137748007695	0.14775932697233307	0.188949278
Fjob_at_home	0.041430004	0.05354643	0.019664543
Fjob_health	0.039914109	0.046757535	0.063775747
Fjob_other	0.5660204277988786	0.5448651830312933	0.5578007583242469
Fjob_services	0.2931733808039003	0.2777090752703345	0.2881703792701424
Fjob_teacher	0.059462078	0.077121776	0.070588573
reason_course	0.2880988537549581	0.39632979882753566	-0.160382998
reason_home	0.33107872458176213	0.2566954201313357	0.5727476976589083
reason_other	0.09510339	0.090616043	0.059253643
reason_reputation	0.2857190313801814	0.25635873841591217	0.5283816575376854
guardian_father	0.17115328452020656	0.2473293724765613	-0.057638428
guardian_mother	0.6989599232883451	0.6876936057584428	0.7773844068570424
guardian_other	0.12988679219144839	0.064977022	0.280254021

Table 5: Last 29 attribute influences for K-Means clusters

4.2 DISCUSSION

Applying the decision tree classifier and k-means clustering, we have gained some valuable insights into important metrics of student performance. The insights we have gained have provided some answers along with connections to prior works that were done.

Decision Trees

For our decision trees, we tested multiple types of decision trees (Gini utilizing a set depth, Gini utilizing an impurity, gradient boosting with learning rates and estimators, and a random forest with an impurity) with various seeds. We wanted to optimize the performance of our tree with the chosen dataset.

The Gini tree decision tree was the first decision tree we worked with. Gini trees let us work with the Gini index to give us the option to control where there are splits within the decision trees. This customizability gives us better control, allowing for the optimization of the tree. An added benefit is that it is easy for us to observe the tree and tune the impurity to our needs. Given various seeds, the impurity that was used changed.

Based on the collected data, as seen in Table 3, our decision tree classifier has determined some important features for students that determine their academic performance. The two biggest features we observed were ‘failures’ and ‘absences.’ The feature ‘failures’ determined the number of past class failures that were provided in a numeric form. The decision tree determined that ‘failures’ were the number one determining factor for student success. This is reasonable as a high number of failures indicates student troubles in some way which contributes to success or a lack thereof in classes. The second most important feature was ‘absences’; this

aligns with our initial hypothesis and earlier works, which looked at academic preparedness (time spent studying) and engagement as critical factors of student success.

Interestingly, some non-academic factors, such as going out with friends (goout) and age, exhibited some importance in the classification of student performance. While they are not the top predictors for student success, it showed that their personal lives, social interactions, and maturity of students contributed to their academic performance. However, factors such as the size of their family or activities seemed to have little impact on their performance.

K-Means Clustering

Looking at Figure 2, we can see that the K-Means clustering algorithm depicts 3 clusters in which the first 2 cluster centroids are closer than the 3rd cluster centroid. This showcases that the dataset and the students it depicts have relatively similar attributes except for the 5 outliers that were assigned the the 3rd cluster.

Looking a bit deeper, we can analyze the inverse-transformed centroids to see which attributes have the most influence. One important thing to note is that the relevance of an attribute is not just based on the numerical difference but is also decided by the attribute's potential values. For example, a difference of 1 in age isn't big given that a student's age can be any integer between 15-22, but a difference of 0.2 in Pstatus is big as the set of values for that attribute is {0, 1}.

Analyzing the attributes shows that address, age, family size, travel time, and many others have similar means between clusters, which showcases that they have little to no influence on the cluster to which each point was assigned.

On the other hand, attributes such as grade, absences, internet, nursery, and many others have relatively large differences for their respective set of values. Attributes with these large differences also have a large influence on the cluster to which each point was assigned.

Using the previously explained process to analyze the entire table results in two overarching conclusions. The first is that attributes that directly relate to school, studying, and homework have the largest influence on defining each cluster; the second is that attributes relating to the student's guardian or other non-academic relations have little to no influence on each cluster.

Analyzing the individual clusters also reveals some patterns. For this analysis, we mainly compared the 3rd cluster to the 1st and 2nd clusters due to the small inter-cluster distance between the two. Looking at the table, we see that the 3rd cluster has the lowest grades on average combined with the highest absences, higher amount of romantic relationships, higher amount of parents living apart, and choosing to join the school due to reputation and being close to home. While this doesn't give us a precise view into the reasons behind having less success in school, it does give us an approximate conclusion that focusing less on school, due to various reasons, and not attending school can cause a decrease in success. This conclusion aligns well with our hypothesis that academic features have the largest influence on student success. However, it is important to note that these clusters were made to predict patterns in students and represent attributes that differentiate groups found in the dataset, the clusters are in no way a precise representation of which attributes have the largest effect on another attribute or the target attribute.

Prior Work

Our findings reveal some connections between our conclusions and the conclusions of prior works. One example is in a paper authored by M.N. Berg and W.H. Hofman, in which they analyzed how student and faculty factors affected student success in universities; through their experiments, they were able to conclude that 95% of the variance in the data they analyzed was due to student factors [1]. Our clustering results shared a similar conclusion with the attributes exerting the most influence on each cluster having a strong relation to student and academic factors; in other words, our k-means cluster gave us a conclusion similar to what Berg and Hofman found in that the most influential attributes (those who explain the most variance) are those relating to academia and student life.

The second prior work that our conclusions can also relate to is a paper by Nahar et al., in which they leveraged data mining to predict student performance; through their analysis, they deduced that the two leading factors contributing to student success were the student's performance in prerequisite courses and their performance till their midterm exam [4]. Our decision tree showed a similar conclusion in that the two main factors determining student success were past failures and absences, both of which can be linked to prior and current student performance.

Limitations of this Study

Despite the insights that we have gained from the study, some limitations come with the information that we used. Firstly, our dataset, comprehensive as it is, only really looks at one demographic and one specific educational context. With the use of this data, there is an inherent bias based on the education and socioeconomic status of this specific population/region. Using data from multiple regions around the world would give us more data to work with that can be better generalized. More variety in our data would also bring in more concrete answers and ideas, allowing for a more in-depth understanding of the various features affecting academic performance.

5 CONCLUSIONS

5.1 LESSONS LEARNED

The first dataset we used contained 145 students with various features [8]. This data had many features that were partly explained through labels. When creating the decision trees, we learned we were achieving poor results for training and testing accuracies due to overfitting. Originally, we were getting testing accuracies around 20-35%, which was not ideal. The reason for the data overfitting was due to the numerous features of the dataset and the low number of students in the data. We also

learned that the first split would be on the Course ID when making our decision tree. The dataset information card did not provide any description of the Course ID, and we could not analyze the resulting classification accurately or provide any worthwhile inferences. Thus, we determined that the dataset was not sufficient and moved to a new dataset. With the new dataset, we were able to achieve much better results with preprocessing and had more data to work with. The new dataset contained 395 entries versus 145 in the original, a little over double. Additionally, the features were well described, meaning that we knew more about the data and could make meaningful conclusions about the results of our study.

One of the attributes in the dataset was 'absences,' which ranged from 0-93. We tested three ways of preprocessing the wide-spanning values of 'absences' with the assumption that grouping the data would increase accuracy. The first was not modifying the grouping and leaving the values as 0-93, the second was grouping into groupings with values of 0-2, 3-5, 6-8, 9-11, and 12+, and the third was maintaining any values less than or equal to 8 otherwise they were assigned the value of 8. However, upon testing, we discovered that leaving the values in their default range provided slightly better accuracy; upon reflection, this makes sense as the grouping was manually done, whereas the tree could split on its own where it determined the value was best.

Looking at our K-Means clustering, we found that there were only 5 outliers in the data, while the other points were relatively similar. Because of this imbalance, we weren't able to accurately evaluate the patterns within student groups without defining the outliers as a separate cluster. In the future, it would be better to find a dataset with a larger variety of data while staying balanced.

5.2 BROADER IMPLICATIONS

Initially, we hypothesized that we could expect some favorable traits related to academics, like studying, would have the greatest influence on student performance. By identifying the key factors associated with positive academic performance, educational institutions, administrators, and teachers can make more informed decisions regarding curriculum design, resource allocation, and targeted strategies.

Our study has shown positive correlations between a student's academic performance and certain factors like the number of absences and failures a student incurs. The clustering of the data shows a strong correlation between students who attend classes regularly and pass classes to be academically successful compared to those who do not. However, we did not expect some non-academically related features to be as prominent such as going out with friends and their age. Knowing how students act, solutions can be implemented to help with their academic success. Obvious factors like attending class regularly and passing courses have a good indication of success. However, with additional features like going out with friends, age, and other things like health, it is vital to keep students' personal lives healthy. There seems to be a good balance between academics and having personal time.

From this, we could see a couple of solutions that schools and teachers could implement to help students succeed. Firstly, it is a good idea to implement ways where students can be encouraged to attend class more often. This could be done through providing incentives to come to class or setting up a more interactive environment. Additionally, teachers could consider student's personal time when assigning work and planning their schedules for their classes. Giving students some time to focus on themselves or other work they might have may have a positive effect on their performance in all classes.

Beyond educational instances, this study also contributes to the methodologies and approaches used in machine learning, especially those related to tasks where education is being observed. By looking at the specific challenges of this study, future research can look to improve and learn from our findings to ultimately advance our understanding of the learning process and the dynamics that shape academic achievement.

6 SCHEDULE

These are the times for the team meetings to discuss the status of the project and report. Along with the meetings, each member's attendance is recorded.

Meeting #1: April 4, 2024, 5:30 PM - 6:30 PM.
Attendance: Brody Bond, Avi Choksi, Anant Patel

Meeting #2: April 9, 2024, 2:30 PM - 3:00 PM
Attendance: Brody Bond, Avi Choksi, Anant Patel

Meeting #3: April 13, 2024, 4:00 PM - 5:00 PM
Attendance: Brody Bond, Avi Choksi, Anant Patel

Meeting #4: April 14, 2024, 4:00 PM - 5:00 PM
Attendance: Brody Bond, Avi Choksi, Anant Patel

REFERENCES

[1] M.N. Berg and W.H. Hofman. 2005. Student Success In University Education: A multi-measurement study of the impact of student and faculty factors on Study Progress. *Higher Education* 50, 3 (October 2005), 413–446. DOI:http://dx.doi.org/10.1007/s10734-004-6361-1

[2] Chong Ho Yu, Samuel DiGangi, Angel Jannasch-Pennell, and Charles Kaprolet. 2021. A data mining approach for identifying predictors of student retention from sophomore to Junior Year. Journal of Data Science 8, 2 (July 2021), 307–325. DOI:http://dx.doi.org/10.6339/jds.2010.08(2).574

[3] Kabra, R. R., & Bichkar, R. S. (2011). Performance prediction of engineering students using decision trees. International Journal of Computer Applications, 36(11), 8-12.

[4] Khaledun Nahar, Boishakhe Islam Shova, Tahmina Ria, Humayara Binte Rashid, & A. H. M. Saiful Islam. (2021). Mining educational data to analyze students' academic performance. Education and Information Technologies, 26(3), 2977-2994. DOI:https://doi.org/10.1007/s10639-021-10575-3

[5] Amirah Mohamed Shahiri, Wahidah Husain, and Nur'aini Abdul Rashid. 2015. A review on predicting student's performance using data mining techniques. *Procedia Computer Science* 72 (December 2015), 414–422. DOI:http://dx.doi.org/10.1016/j.procs.2015.12.157

[6] Dev Ansodariya. 2022. Student performance dataset. (May 2022). Retrieved March 29, 2024 from https://www.kaggle.com/datasets/devansodariya/student-performance-data.

[7] Aman Chauhan. 2022. Student performance. (October 2022). Retrieved March 29, 2024 from https://www.kaggle.com/datasets/whenamancodes/student-performance.
[8] Joakim Arvidsson. 2023. Students Performance. (October 2023). Retrieved April 12, 2024 from https://www.kaggle.com/datasets/joebeachcapital/students-performance/data.
[9] Tallal Omar, Abdullah Alzahrani, and Mohamed Zohdy. 2020. Clustering approach for analyzing the student's efficiency and performance based on Data. Journal of Data Analysis and Information Processing 08, 03 (2020), 171–182. DOI:http://dx.doi.org/10.4236/jdaip.2020.83010.

APPENDIX

Highest Accuracy with Various Hyperparameters using 25% Testing Data. As previously noted, all Gradient Boosting Trees stated the same testing and training accuracy, regardless of the values (within the ranges we specified), as being equally valid so we have omitted the learning rate and estimators from this table.

Seed 1337	**Depth**	**Impurity**	**Train Accuracy**	**Test Accuracy**
Gini - Depth	3	N/A	0.753	0.687
Gini - Impurity	N/A	0.0105	0.743	0.687
Gradient Boost	N/A	N/A	0.676	0.657
Random Forest	N/A	0.002	0.986	0.717

Table 6: Seed 1337 Rough Estimates

Seed 999	**Depth**	**Impurity**	**Train Accuracy**	**Test Accuracy**
Gini - Depth	4	N/A	0.804	0.697
Gini - Impurity	N/A	0.0105	0.750	0.687
Gradient Boost	N/A	N/A	0.666	0.687
Random Forest	N/A	0.0045	0.824	0.717

Table 7: Seed 999 Rough Estimates

Seed 2	**Depth**	**Impurity**	**Train Accuracy**	**Test Accuracy**
Gini - Depth	4	N/A	0.791	0.758
Gini - Impurity	N/A	0.008	0.791	0.768
Gradient Boost	N/A	N/A	0.662	0.697
Random Forest	N/A	0.0015	0.993	0.727

Table 8: Seed 2 Rough Estimates

Seed 9	**Depth**	**Impurity**	**Train Accuracy**	**Test Accuracy**
Gini - Depth	3	N/A	0.747	0.778
Gini - Impurity	N/A	0.0105	0.743	0.778
Gradient Boost	N/A	N/A	0.649	0.737
Random Forest	N/A	0.001	1.0	0.808

Table 9: Seed 9 Rough Estimates