Predicting the Academic Performance of Undergraduate Computer Science Students Using Data Mining

Faiza Khan
Research Advisors: Gary M. Weiss and Daniel D. Leeds
Dept. of Computer and Information Sciences, Fordham University
Bronx, NY 10458, USA

***Abstract*—There are myriad factors which can affect a student's academic performance as measured by Grade Point Average (GPA). Identifying characteristics of students with high GPA can help more students understand how to achieve the best grades possible. In this paper, a variety of data mining algorithms are used to predict the GPA of undergraduate students majoring in Computer Science based on survey questions. The results demonstrate that the number of hours of sleep per night, the frequency of illicit drug use, the number of hours spent studying per week, and the number of hours spent on social media platforms per week are important factors that can be used to classify student GPA. The Random Forest data mining algorithm performed the best and was able to achieve a predictive accuracy of 95% when placing students into one of four academic performance groupings.**

***Index Terms*—educational data mining, data mining, classification**

I. INTRODUCTION

The field of data mining is concerned with finding relevant and meaningful patterns within a dataset. A dataset contains instances of data from various attributes, or factors. An instance contains the values of the attributes for one student, therefore, there were 82 instances in this dataset as 82 computer science students at Fordham University submitted the anonymous survey. The survey contained 23 questions related to the student's demographics, lifestyle, etc., and also asked for the student's GPA. Predictive models were built using a variety of data mining methods to identify the key features that impact student performance. By observing meaningful patterns within this student GPA dataset, we can determine which of the 23 factors are most useful in predicting student GPA and differentiating between below average, average, good, and great students. The design of the survey was influenced by the related work in this field, which suggested some factors that can impact student performance.

Sleep is one of the most influential factors of student GPA. Suffering from sleep deprivation and poor sleep quality negative impacts the academic performance of students [1,2,3,4,5]. A study shows that school-age students who were not sleep deprived were healthier and had higher IQ and perceptual reasoning overall as measured by the WISC-IV, a test to measure intellectual ability of children [2]. In another study, among 144 medical students, sleep quality of the students prior to the pre-clinical examination affected their result and final GPA [3]. Another study has results which indicates that students with the highest GPA had earlier bedtimes and earlier wake times [4]. Thus, the number of hours of sleep a student receives per night as well as the time that they typically go to sleep can be good indicators in predicting student GPA.

Moreover, sleep deprivation most prominently affects functional connectivity involving prefrontal areas [5]. Because the prefrontal cortex is responsible for decision making, students who are sleep deprived have less activity in the prefrontal cortex, which can make it harder for them to make decisions during tests and obtain a decent GPA. Furthermore, other factors may also affect one's sleep quality, such as their outlook on life [6]. People who are optimistic may fall asleep faster than those who are pessimistic. Because the studies that were previously mentioned showed that good sleep quality resulted in higher GPA and people who are optimistic tend to have better sleep quality, it can be concluded that students who are optimistic generally have higher GPA than students who are pessimistic.

Drug use can also be detrimental to academic performance. Research shows that alcohol and drug use among college students resulted in lower GPA's than those who used such drugs minimally or refrained from drug use [8]. The prefrontal cortex is also affected by drug use, especially among teens and young adults since this region of the brain develops until the mid 20s. Damage to the prefrontal cortex from substance abuse, therefore, can result in poor performance.

The number of hours spent studying, social media use, and the student's personality may also influence their GPA. Students who study for longer periods of time are more likely to be prepared for exams than students who do not dedicate as much time to their studies. Because the overall grades for undergraduate computer science courses are primarily determined by midterm and final examination scores, performing well on exams results in a higher GPA among undergraduate students. Moreover, increased use of social media may result in less time dedicated towards studying. In turn, this can lead to a lower GPA. Personality is another factor which can affect GPA since studies show that introverts tend to be better listeners than extroverts, [9] which signifies why this feature was another good indicator determining student GPA on some classifiers. Undergraduate courses are typically lecture-based, thus listening attentively and taking good notes can help students better prepare for exams, and in turn, earn a higher GPA.

The next section, section II, covers experimental methodology which explains the approach to this classification task and

how the performance of each of the classifiers were evaluated. The results and analysis the performance of the classifiers on this student GPA dataset are covered in section III. Section IV mentions related work in this field and the concepts that future studies can focus more on are discussed in section V.

II. Experiment Methodology

A. Attributes Used In the Survey

82 undergraduate computer science students at Fordham University were given a survey which they filled out anonymously. The survey contains 23 questions, and the responses were used to predict GPA.

Table 1. Name and Description of Attributes

	Feature Name	Description
1	gender	male, female, or other
2	age	student's age
3	year	year in school
4	major	all students were Computer Science majors
5	race	student's race
6	ethnicity	Hispanic or Not Hispanic
7	sleep	hours of sleep per night
8	socialMedia	hours of social media use per week
9	studying	hours of studying per week
10	ResComm	campus resident or commuter
11	studIncome	student's annual income
12	parIncome	total household annual income
13	job	whether or not the student currently has a job
14	drugs	illicit drug use on a scale of 1-5
15	alcohol	alcohol use on a scale of 1-5
16	physical	hours of physical activity per day
17	notes	prefer taking notes with notebook or laptop
18	OlderSiblings	number of older siblings
19	YoungerSiblings	number of younger siblings
20	outlook	pessimistic, optimistic, or neutral outlook on life
21	personality	introverted or extroverted
22	satisfied	satisfied with academic grades
23	classGrades	GPA (below average, average, good or great)

The number of hours spent studying per week, attribute 9, was a multiple choice question for students in which 1 represents studying for 0-5 hours per week, 2 represents studying for 6-10 hours per week, 3 represents studying for 11-15 hours per week, 4 represents studying for 16-20 hours per week, and 5 represents studying for more than 20 hours per week.

Illicit drug use and alcohol use, attribute 14 and 15 respectively, were based on a scale of 1-5 in which 1 signifies never used, 2 signifies rarely used, 3 signifies used every other week, 4 signifies used every week, and 5 signifies everyday use.

Total annual household income, attribute 12, was a multiple choice attribute for students in which 1 represents $\leq$ \$30k, 2 represents between \$30k and \$50k, 3 represents between \$50k and \$70k, 4 represents between \$70k and \$100k, 5 represents between \$100k and \$350k, and 6 represents $\geq$ \$350k.

Below are attribute distributions of some of the factors which will be analyzed in section III. Table 2 shows the attribute distribution of illicit drug use and Table 3 shows the attribute distribution of the number of hours spent on social media per week. Table 2 indicates that almost two-thirds of the students in this dataset reported never using illicit drugs. Table 3 shows that more than one-third of the students in the dataset use social media for less than five hours per week, while the rest of the students use social media for longer periods of time.

Table 2. Illicit Drug Use Attribute Value Distribution

Illicit Drug Use Frequency	Number of Students
1 (Never used)	51
2 (Once a month/rarely)	16
3 (Sometimes)	5
4 (Every week)	8
5 (Almost everyday)	2

Table 3. Hours Spent on Social Media Per Week

Social Media Use	Number of Students
0-5 hours per week	32
5-10 hours per week	20
10-15 hours per week	13
15-20 hours per week	7
20+ hours per week	10

Table 4 shows the GPA class distribution. Students at Fordham University have a GPA between 0.0 and 4.0. In order to convert this regression task into a classification task, the corresponding class values were assigned to the following GPA scores: poor (0.0-2.49), below average (2.50-2.99), average (3.0-3.32), good (3.33-3.66) and great (3.67-4.0). A 4.0 is the highest possible GPA a student can obtain at Fordham University. No student reported a GPA under 2.50, hence the "poor" category does not appear in this dataset. There were 82 total instances in the dataset, however, 2 of those instances do not contain the class value. Hence, the performance of the algorithms was determined by the 80 instances which have the class value (student GPA).

Table 4. GPA Class Distribution

Academic Performance Category	Number of Students
Below Average	5
Average	14
Good	26
Great	35

B. Data Mining Algorithms Used

The classifiers used on this student GPA dataset include Nearest Neighbor, Decision Trees, Random Forest, Random Trees, and Multilayer Perceptron. The performance of these data mining algorithms were compared against the performance of ZeroR, the baseline classifier which always predicts the majority class value of a dataset. For all algorithms, the default parameters on the WEKA data mining application are used to ensure fairness in performance on this dataset.

1) IBk Nearest Neighbor: Nearest Neighbor is an instance-based classifier which classifies an unseen instance based on its distance to other training records. The distance to the training record determines the level of similarity, thus an unseen instance is classified into the same category as its closest trained instance due to their similar attribute values. The default parameter uses one training record as a Nearest Neighbor (also known as 1-Nearest Neighbor), hence the class label of the closest training instance determines the class value of that unseen instance.

2) J48 Decision Tree: A Decision Tree classifier splits the data based on the attributes that result in the lowest entropy in an effort to maintain homogeneity. Each leaf node in the Decision Tree corresponds to a combination of rules, or factors, which resulted in that classification.

3) Random Forest: Random Forest is a computationally fast algorithm which creates an ensemble of Decision Trees. Ensembles have more expressive power and can assist with bias and variance by averaging over multiple runs. The final decision is made via majority voting of the trees in the forest in the Random Forest method.

4) Random Tree: Similar to the Decision Tree algorithm, the Random Tree algorithm generates an output in the structure of a tree that is easy to understand and justify. It also employs the method of bagging to produce a random set of data for constructing a Decision Tree and generates many individual learners.

5) Multilayer Perceptron: Multilayer Perceptron is a class of artificial neural networks. This algorithm leverages the use of hidden layers for inputs, and it assigns weights to the attributes based on usefulness in classification of instances. These weights are typically not easy to justify or explain. The weights of the same attribute can vary among the different sigmoid nodes in this method. Sigmoid nodes are used in backpropagation with the associated data in the Multilayer Perceptron algorithm. Each sigmoid node contains the attributes and their associated weight.

C. Evaluation Metrics

A training set contains instances which serves as input to the data mining algorithm while the test set contains instances which must be classified by the algorithm. Classifiers build a model on the training set and they are evaluated based on their performance on the test set. 10-fold cross-validation, the partitioning method used in this experiment, is a type of cross-validation which entails partitioning a dataset into 10 partitions, training on 9 partitions and testing on the remaining section, iterating for 10 times. To generate a test set, the 10-fold cross-validation method is used so that every instance in the training set can be a part of the test set. The results of the 10-fold cross-validation on the different algorithms will be analyzed in section III.

There is also minimal preprocessing involved for this experiment. The preprocessing only consists of the removal of ID (an attribute which was assigned to each of the surveys to identify the instances). Because many of the attributes in combination with one another proved to be useful in predicting student GPA, no additional attributes were removed in the preprocessing stage.

To emphasize the importance of a large dataset and to analyze how the predictive accuracy changes based on a smaller sample size, the algorithms are also tested using 10-fold cross-validation on 25%, 50%, and 75% of the data. As the size of the dataset decreased, the algorithms generally performed worse.

III. Results

In this section, the performance of the following algorithms using 10-fold cross-validation are discussed. Random Forest had the highest predictive accuracy on this dataset.

Table 5. Accuracy Among Different Classifiers

Classifier	Nearest Neighbor	Decision Tree	Random Forest	Random Tree	Multilayer Perceptron	ZeroR (baseline)
Accuracy	91.25%	75.0%	95.0%	82.5%	90.0%	43.75%

The ZeroR (baseline) classifier performs the worst, as depicted in the table. As mentioned previously, this classifier predicts the majority class value. Because the majority of the students in the dataset reported having a great GPA (35 out of 80 students), the baseline always predicts that the students have a great GPA. The table indicates that Nearest Neighbor, Decision Trees, Random Forest, Random Tree, and Multilayer Perceptron were able to learn the dataset well since their predictive accuracy was significantly higher than the performance of the baseline classifier. The performance of those algorithms yields useful information regarding the GPA classification of undergraduate computer science students.

A. Analysis of the Performance of IBk Nearest Neighbor

The accuracy rate for Nearest Neighbor algorithm on this dataset is 91.25%, which shows that there are similar characteristics among students with the same GPA classification. Instances that are a part of the training set may have similar attribute values to unseen instances in the test set. The high accuracy rate indicates that the Nearest Neighbor algorithm learned this dataset well. Using 1-Nearest Neighbor (the default parameter of Nearest Neighbor) shows that students within the same GPA classification typically shared similar attribute values.

B. Analysis of the Performance of J48 Decision Tree

The accuracy rate for Decision Tree on this dataset is 75.0% and it was the poorest performing algorithm among the 5 classifiers discussed in this section. Decision Trees typically handle irrelevant features well, but perhaps the attribute values used to split the data did not yield the lowest entropy. Complex rules cannot be expressed well with Decision Trees, and because there are several factors which can be used to predict student GPA in combination with one another, the Decision Tree was not able to perform very well on this dataset.

The output of the Decision Tree suggests that the number of hours of sleep and illicit drug use are the two most important factors for classification of student GPA in this dataset since these are the first two splits in the tree. Sleep is the first split, and students who reported sleeping less than 4 hours per night were automatically categorized as having a "below average" GPA. This signifies that sleeping more hours per night may increase the likelihood of a student performing well academically. Among students who sleep more than 4 hours per night, those who use illicit drugs generally have a low GPA classification whereas students who reported never using

```
sleep <= 4: belowAverage (5.0/1.0)
sleep > 4
|   drugs <= 1
|   |   studying <= 3
|   |   |   personality = introverted
|   |   |   |   race = AmericanIndian: great (0.0)
|   |   |   |   race = Asian: great (6.0)
|   |   |   |   race = AfricanAmerican: good (2.0)
|   |   |   |   race = NativeHawaiian: great (0.0)
|   |   |   |   race = White
|   |   |   |   |   gender = M
|   |   |   |   |   |   studIncome <= 7000: good (5.0)
|   |   |   |   |   |   studIncome > 7000: great (2.0)
|   |   |   |   |   gender = F: great (7.0)
|   |   |   |   |   gender = O: great (0.0)
|   |   |   |   race = Other: good (1.0)
|   |   |   personality = extroverted
|   |   |   |   ResComm = resident
|   |   |   |   |   alcohol <= 2: great (3.0)
|   |   |   |   |   alcohol > 2: good (2.0)
|   |   |   |   ResComm = commuter: good (8.0)
|   |   studying > 3: great (13.0)

|   drugs > 1
|   |   age <= 20: average (12.0/1.0)
|   |   age > 20
|   |   |   alcohol <= 3: average (2.0)
|   |   |   alcohol > 3
|   |   |   |   ResComm = resident
|   |   |   |   |   year = freshman: great (0.0)
|   |   |   |   |   year = sophomore: great (0.0)
|   |   |   |   |   year = junior: good (2.0)
|   |   |   |   |   year = senior: great (4.0)
|   |   |   |   ResComm = commuter: good (6.0)
```

Fig. 1. Decision Tree Output

drugs are classified as having a good or great GPA. Students who use drugs are not very likely to have a great GPA in this dataset since only one leaf node in the drugs greater than 1 subtree contains instances classified as "great."

The third split in the Decision Tree is either the number of hours spent studying or student age, depending on which subtree of the drugs feature the student belongs to. For students who never used illicit drugs, the next feature that the Decision Tree splits on is the number of hours spent studying. As shown in the Decision Tree Output, students who sleep more than 4 hours per night, have never used drugs, and study more than 15 hours per week were all classified as having a great GPA. This particular rule applied to 13 instances, signifying that more than one-third of the students who had a great GPA were classified as "great" based on the rule which combines these three factors. For students who have used illicit drugs, the Decision Tree splits on age. All students who sleep more than 4 hours per night, use drugs, and are 20 years old or younger are classified as having an average GPA, which is the second lowest classification in this dataset.

```
===Confusion Matrix===
a     b     c     d     ← classified as
4     1     0     0     |     a = belowAverage
1     11    2     0     |     b = average
0     0     16    10    |     c = good
1     0     5     29    |     d = great
```

Fig. 2. Confusion Matrix for Decision Tree

One student who had great GPA was classified as having a below average GPA based on the Decision Tree confusion matrix, which demonstrates that this algorithm was off by three classes for only one particular instance. For all other instances, the Decision Tree either classified the GPA correctly, or misclassified by only one class value. Additionally, some students with good GPA were classified as "great" and some students with great GPA were classified as "good." This means that the attribute values for students with a GPA of 3.33 or above were more similar than the attribute values for students with significantly lower GPA. Because some of the responses for these two groups of students were similar, some attributes of the Decision Tree split did not provide low entropy for the classification of these student GPA's.

C. Analysis of the Performance of Random Forest

The Random Forest classifier used an ensemble of Decision Trees in order to classify student GPA, and it resulted in the highest accuracy rate among all of the algorithms discussed in this paper. The accuracy rate for Random Forest is 95.0%, which is significantly higher than the accuracy rate for Decision Tree. This suggests that the Random Forest algorithm was able to learn the dataset extremely well.

```
===Confusion Matrix===
a     b     c     d     ← classified as
4     0     0     1     |     a = belowAverage
0     14    0     0     |     b = average
0     0     24    2     |     c = good
0     0     1     34    |     d = great
```

Fig. 3. Confusion Matrix for Random Forest

While this algorithm performed the best among all classifiers, it categorized one student with a below average GPA as "great". These two classes are on opposite ends of the spectrum as the lowest GPA classification is below average (2.50-2.99) and the highest GPA classification is great (3.67-4.0) in this dataset. However, the classification accuracy is significantly higher for this classifier since only 4 instances were categorized inaccurately.

D. Analysis of the Performance of Random Tree

The accuracy rate for Random Tree on this dataset is 82.5%. Because Random Tree performed better than Decision Tree for this dataset, the attribute values which were used to split the data in Random Tree most likely resulted in lower entropy for more accurate classification.

The output of the Random Tree classifier suggests that the number of hours spent studying per week, total annual household income, and the number of hours of physical activity per day are the three most important factors in the classification of student GPA since these are the first three splits in the tree. The Random Tree first splits on the number of hours spent studying per week. Students who studied less than 15 hours a week typically had lower GPA classifications, depending on other factors. Next, the Random Tree splits on either total household income or the number of hours of physical activity per day, depending on which subtree the student belongs to. Students who studied more than 15 hours per week but exercised for less than three-quarters of an hour (45 minutes) per day typically had lower GPA classifications than students who studied more than 15 hours per week and exercised at least 45 minutes

per day. Thus, exercising regularly can improve the academic performance of students. More than one-third of the students who were classified as having a great GPA reported studying more than 15 hours per week and exercising at least 45 minutes per day.

Additionally, for students who study less than 15 hours per week, the next attribute that the Decision Tree splits on is the income attribute. If the total household annual income exceeds $100k (value of 4 or greater on the scale of 1-6), then the student is more likely to achieve either a good or great GPA if they use social media for less than 17.5 hours per week and use illicit drugs either rarely or never. In contrast, students who reported having a household income greater than $100k but used social media for more than 17.5 hours per week generally had lower GPA classifications. In fact, while many of the leaves in the income greater than $100k subtree correspond to having a great or good GPA, students who spent more than 17.5 hours on social media per week and slept for less than 5.5 hours per night were all classified as having below average GPA, which is the lowest classification. Therefore, an increased number of hours spent on social media, combined with other factors, can negatively impact GPA for undergraduate computer science students. Furthermore, students can still be classified as having a great GPA with a household income lower than $100k if they are optimistic and use social media platforms for less than 4.5 hours per week. This suggests that while household income can be an important factor for predicting academic performance, other factors such as social media use and drug use can determine the final classification of student GPA.

The split on the outlook attribute is interesting because some leaf nodes corresponded to a great GPA for optimistic and neutral outlooks, but no leaf nodes corresponded to a great GPA for a pessimistic outlook. This shows that a positive or neutral outlook on life can help undergraduate computer science students obtain a higher GPA.

Additionally, the frequency of drug use also impacted GPA. All 5 students who were in the lowest GPA classification (below average) reported moderate to high frequency of illicit drug use (attribute value was 3 or higher on the scale of 1-5). Therefore, drug use negatively impacts student GPA.

It is important to note that the Random Tree and Decision Tree output generates some rules which may not be outputs of a larger student GPA dataset since some factors may not be causally linked and may simply be co-occurrences of one another. Additionally, the Random Tree also splits on the same attribute within its subtrees sometimes, unlike the Decision Tree output which splits on each attribute only once in each subtree. For instance, there were several splits in the Random Tree on attributes such as the number of hours spent studying per week and number of hours spent on social media platforms. This suggests that even minor increases in the number of hours spent on social media can negatively impact undergraduate student GPA.

E. Analysis of the Performance of Multilayer Perceptron

The Multilayer Perceptron algorithm had an accuracy rate of 90.0% on the classification of student GPA for this dataset. It may be difficult to justify the weights of the Multilayer Perceptron algorithm, however, there were 25 sigmoid nodes (which are nodes with different weights on features) in which the highest weights were usually assigned to the following attributes: studying, drugs and parental income. This may signify that these values were the most useful in predicting student GPA for this classifier.

F. General Analysis

Below are some more general observations regarding the output of these algorithms:

- Students who slept less than 4 hours per night have lower GPA as suggested by the J48 Decision Tree Algorithm. Additionally, students who slept more than 8 hours (in combination with other factors) were classified as having great GPA by the Random Tree output.
- Students who studied more than 15 hours per week, have never used illicit drugs, and typically slept for more than 4 hours per night were classified as having a great GPA.
- Undergraduate seniors can be classified as having a great GPA in the Decision Tree even if they drink alcohol at least once a week and use illicit drugs. This suggests that those who have reached the legal age of drinking might be impacted less severely than those who are younger.
- Students who are older tended to perform better academically than younger students. This may occur because computer science students can enhance their knowledge as they gain more experience in their field of study. As a result, upperclassmen may perform better in their courses and obtain a higher GPA than freshmen and sophomores.
- Students who spent 17.5 or more hours per week on social media tended to study less than 15 hours per week, and as a result, had lower GPA's than those who dedicated more time to their studies.

G. Relationship Between Features and the GPA Class Value

Below are the graphs which depict the relationship between the frequency of illicit drug use and GPA (Figure 4(a)), the number of hours spent on social media platforms per week and GPA (Figure 4(b)), and the number of hours of studying per week and GPA (Figure 4(c)). The purple dots represent students belonging to the "below average" GPA class, the red dots represent students belonging to the "average" GPA class, the green dots represent students belonging to the "good" GPA class, and the blue dots represent students belonging to the "great" GPA class.

Figure 4(a) shows that as the frequency of drug use increases, students typically have lower GPA's. Most of the students with good or great GPA reported never using illicit drugs. Figure 4(b) shows that while it is possible to obtain a great GPA when using social media for more than 12.5 hours per week, most students with great GPA use social media platforms for fewer hours. Figure 4(c) shows that

as the number of hours spent towards studying per week increases, students are generally able to achieve a higher GPA classification. Most students with a great GPA studied for more than 15 hours per week (indicated by 3 or higher as shown in the graph).

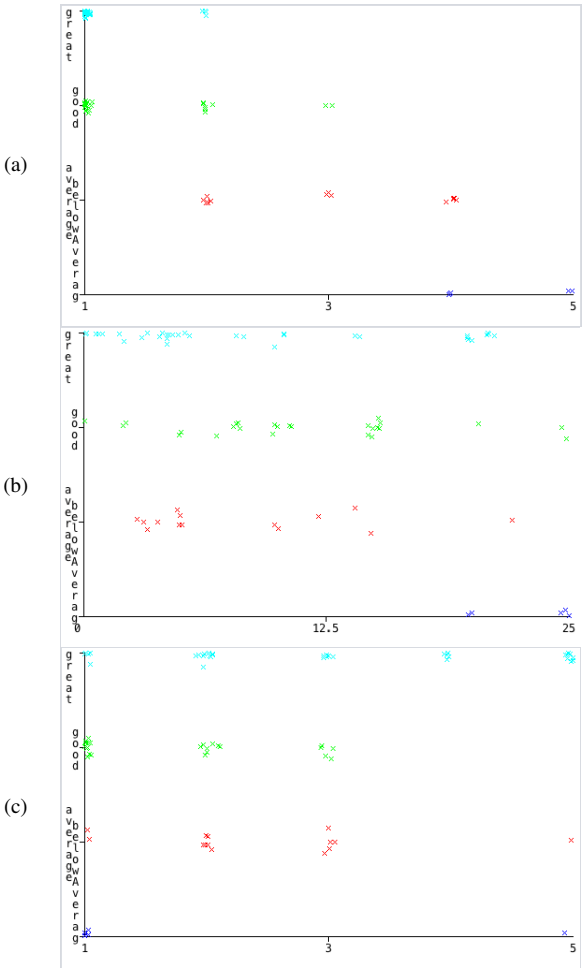

Fig. 4. Relationship Between Feature and GPA. (a) Illicit Drug Use Frequency. (b) Hours of Social Media Use Per Week. (c) Hours Spent Studying Per Week.

H. Results Obtained by Reducing the Size of the Dataset

The size of the datset was reduced to observe how the algorithms perform using 75% of the dataset, 50% of the dataset, and 25% of the dataset. As the number of instances decrease, the algorithms generally perform worse. The original, full dataset has the highest accuracy rates since the algorithms had enough instances to learn the data and provide useful information regarding classification of student GPA.

Table 6. Accuracy of Reduced Size Datasets on 10-Fold Cross-Validation

Training Size	Nearest Neighbor	J48 Decision Tree	Random Forest	Random Tree	Multilayer Perceptron
75% of dataset	78.3%	63.3%	83.3%	73.3%	80.0%
50% of dataset	52.5%	45.0%	57.5%	40.0%	62.5%
25% of dataset	40.0%	45.0%	50.0%	45.0%	55.0%

The algorithms were not able to learn the data well when the number of instances were reduced. More training data generally improves classifier performance. Perhaps if the original, full dataset contained more instances, the accuracy rates would have been much higher since the algorithms would have more learning data. Additionally, even if the accuracy rates were higher after reducing the size of the dataset, we cannot generalize such information as it may not apply to a larger population of undergraduate students. The more instances in the dataset, the less generalization will occur.

IV. Related Work

This section contains discussion regarding some related works of researchers who studied the classification of student GPA using other datasets.

A study aimed to predict the students' final GPA based on the dataset performance on Decision Trees [7]. Based on student grades on previous courses, such as computer architecture, computer ethics and software engineering, the researchers predicted the student final GPA as average, good, very good, and excellent. The predictive accuracy of the models were not presented in this paper, but the researchers demonstrate how GPA results from previous courses can be used to predict student GPA for future semesters. The attribute values used for this classification task include student grades from introductory courses such as software engineering, computer architecture, Java1, etc.

Another study uses feature extraction to classify students based on their academic performance [10]. The researchers primarily focused on the classification of students who have poor academic performance. They extracted features from historical grading data in order to test different simple and sophisticated classification methods based on big data approaches. Gradient Boosting and Random Forest performed the best for this experiment. To classify level A students, the highest accuracy was obtained when using course background and prerequisite attributes. To classify students who have failing GPA, the researchers examined specific reasons which are related to the individuals themselves, not the class background, prerequisite, or similar courses. Area under the receiver operating characteristic (ROC) curve for Gradient Boosting was the highest with a value of 0.877.

Alcohol and drug use can also be significant factors in predicting student GPA as shown in various studies. A study shows that high drug use leads to more absence from school, which affects students' overall academic performance [11]. The results from the dataset demonstrate that girls were more likely than boys to try alcohol and boys were more likely to try illicit drugs. Both illicit drugs and alcohol affect student GPA according to the researchers as increasing levels of drug and alcohol consumption were associated with lower GPA and a higher number of days and hours missed from school [11]. Logistic Regression was the data mining algorithm that was used by these researchers, but the accuracy results of this method are not stated. According to the researchers, some attributes were not helpful in predicting student GPA since

girls typically had higher GPA than boys, but also had more missed days from school on average. Thus, days missed from school was not as important as the student's alcohol and drug use for this dataset.

V. Conclusion

In this paper, the features which were most important in predicting the GPA of undergraduate computer science students in this dataset were analyzed. The output of the data mining algorithms suggest that a combination of factors such as the number of hours of sleep per night, the frequency of illicit drug use, the number of hours spent studying per week, and the number of hours of social media use per week provide useful information that can be used to successfully classify the GPA of computer science majors. Students who slept for more hours per night, studied for more than 15 hours per week, spent less time using social media per week, and never used illicit drugs tended to have the highest classification of GPA. The Random Forest algorithm performed the best among all classifiers with a 95.0% accuracy rate, but the other algorithms still performed much better than the baseline classifier. Therefore, students with the same classification of GPA tend to have similar characteristics.

Future studies can integrate some concepts from the attributes that were used in this dataset and also focus on how specific kinds of drugs affect GPA since some drugs might be more detrimental to academic performance than others. Researchers can investigate how other factors impact GPA, such as the number of hours per week spent towards self-care or meditation. Future studies can also increase the sample size as doing so can result in higher accuracy rate and yield useful information pertaining to student performance. One limitation of this dataset is that there might not be enough instances to generalize the findings to a larger population of undergraduate computer science students.

References

[1] Curcio G, Ferrara M, De Gennaro L. Sleep loss, learning capacity and academic performance. Sleep Med Rev. 2006;10(5):323–337

[2] Gruber R, Laviolette R, Deluca P, Monson E, Cornish K, Carrier J. Short sleep duration is associated with poor performance on IQ measures in healthy school-age children. Sleep Med. 2010;11(3):289–294.

[3] Ahrberg K, Dresler M, Niedermaier S, Steiger A, Genzel L. The interaction between sleep quality and academic performance. J Psychiatr Res. 2012;46(12):1618–1622.

[4] Eliasson, A. H., Lettieri, C. J. and Eliasson, A. H. Early to bed, early to rise! Sleep habits and academic performance in college students. Sleep Breath., 2010, 14: 71– 75.

[5] Verweij IM, Romeijn N, Smit DJ, et al. Sleep deprivation leads to a loss of functional connectivity in frontal brain regions. BMC Neurosci 2014;15:88

[6] Harvard Health Publishing, Harvard Medical School. Positive outlook may mean better sleep https://www.health.harvard.edu/staying-healthy/positive-outlook-may-mean-better-sleep

[7] Al-Barrak MA, Al-Razgan M. Predicting Students Final GPA Using Decision Trees: A Case Study. International Journal of Information and Education Technology. 2016 Jul; 6(7):528–33.

[8] PLOS Research News. Alcohol and marijuana use associated with lower GPA in college. https://researchnews.plos.org/2017/03/08/alcohol-and-marijuana-use-associated-with-lower-gpa-in-college/

[9] Science Daily. Who learns foreign language better, introverts or extroverts? https://www.sciencedaily.com/releases/2017/07/170721104246.htm

[10] Agoritsa Polyzou and George Karypis. 2018. Feature extraction for classifying students based on their academic performance. In Proceedings of the 11th EDM Conference.

[11] Heradstveit, O., Skogen, J. C., Hetland, J., and Hysing, M. (2017). Alcohol and illicit drug use are important factors for school-related problems among adolescents. Front. Psychol. 8:1023. doi: 10.3389/fpsyg.2017.01023 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5476929/