Using Azure AutoML to Analyze the Effect of Attendance and Seat Choice on Student Grades

Ac Hybl
School of Computing
Southern Adventist University
Email: matoush@southern.edu

***Abstract*—The students at Southern Adventist University submit valuable attendance data daily through an attendance-tracking system once used for COVID-19 contact tracing. This study organizes some of this data and employs machine learning to analyze the claim that class attendance and sitting at the front of a classroom improve student grades. We perform a correlation analysis in Azure's Machine Learning workspace by training regression models. No correlation is found between student attendance and seat choice and final course grades. Next we use the K-means clustering algorithm to train clustering models in Azure. At $k = 2$ clusters, a cluster with perfect attendance shows a higher average grade than a cluster with a late attendance average. Seat choice within the classroom does not prove important to the clustering models. Project improvements and future work are discussed at the end.**

***Index Terms*—Microsoft Azure Machine Learning Workspace, Regression, K-means clustering, Class attendance, Seat selection, Grade prediction**

I. Introduction

In 2020, the rapid spread of the SARS-CoV-2 virus, which causes the disease commonly known as COVID-19, changed institutions and systems all around the Earth. While some of these changes made daily life more difficult, others presented unforeseen opportunities.

Southern Adventist University (SAU), like many other educational institutions, implemented strict quarantines and contact tracing to allow their students to attend classes in person during the pandemic [1], [2]. Using the web interface shown in Figure 1, SAU required thousands of students to select their seat in every class they attended all semester. This data allowed for digital contact tracing (DCT) whenever a new case of COVID-19 was identified.

As campus activity returned to normal and concern over Covid subsided over the next few semesters, the new attendance system remained. Many professors simply found the system much more convenient than manually taking note of absent or late students. Not only did the new system allow professors to discuss course material sooner, but it continued collecting attendance and seating data which could be useful for more than just contact tracing.

Upon identifying the opportunity this data held, faculty and students at the campus asked two primary questions:

1) Does class attendance and punctuality foreshadow higher course grades?
2) Do students that sit in the front of class receive higher marks than those that choose to sit near the back?

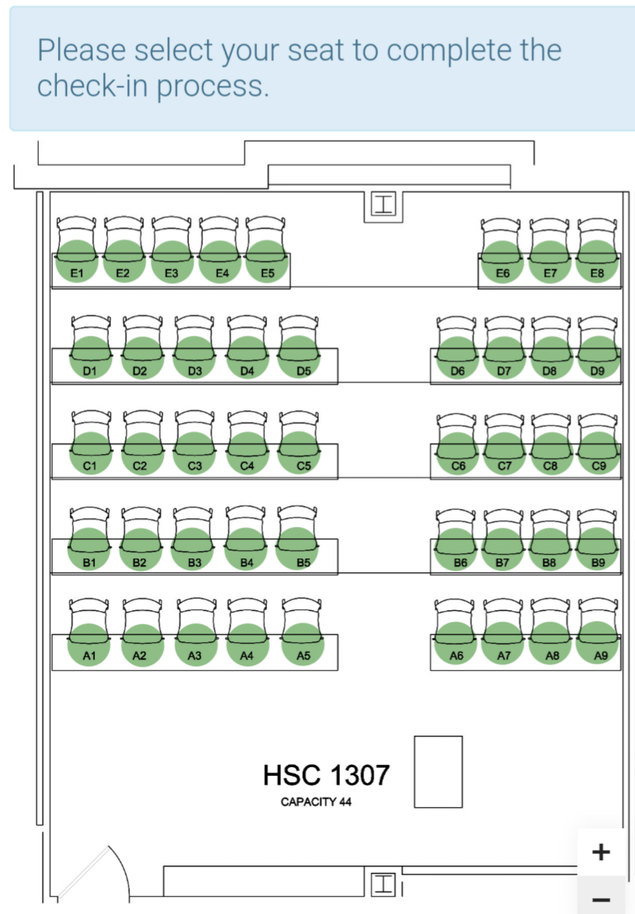

Fig. 1. The ATS interface in one of the many classrooms on campus.

To address these questions, many researchers might perform mathematically-intensive statistical studies. However, because this project was part of machine learning and database courses, we applied two data science techniques instead - regression and clustering.

This project's aim is not to prove or disprove causation between attendance or seat choice and student grades. Rather, it is a case study of how these experiments can be performed in Azure Machine Learning Studio and whether they support the commonly held claims given the current available data.

II. Theoretical Framework

A. Concepts

School websites arguing for the importance of class attendance can be found quite easily.[1, 2] According to academic research, regular attendance promises many benefits for students. Some schools even quote Woody Allen, arguing that "80% of success is just showing up."[3]

The National Center for Education Statistics explains that, starting in kindergarten and progressing through high school, commonly absent students miss out on learning opportunities [3]. As a result, even the best teacher's ability to enable student success is limited. Moreover, after leaving school, absentees "exhibit a history of negative behaviors."

Moreover, during the data collection phase of this project, one of the attendance system managers recalled a high school course in which their instructor announced that students would be graded based on where they sat. Students in the front rows would receive higher marks than those who chose to sit further away. In this way, the instructor was hastening what he assumed to be an inevitable outcome. Several studies, including our project, test this hypothesis.

B. Similar Studies

Researchers and scientists in various departments have conducted studies concerning some aspect of student attendance and course performance. In their 2015 study of first-year psychology courses, Alexander and Hicks analyzed whether class attendance was linked to increased student performance in modern classrooms with online lectures [4]. Their results featured significant ($p < 0.001$ and $p < 0.05$) correlations between student attendance and performance on assignments.

Furthermore, several studies have been done concerning seat choice and student grades. In a 1973 issue of *Sociometry*, Becker et al. demonstrate that students sitting nearer to their instructor not only received higher grades than those further away, but also liked their professor more ($p < 0.01$) [5]. However, other studies present contradictory conclusions [6].

Others have conducted experiments using machine learning to predict grades or analyze groups of students in classrooms. For example, Zabriskie et al. studied which pieces of information best predicted a student's grade in physics courses as the semester progressed [7]. At first, a student's GPA was the strongest predicting factor, but eventually the first test grade surpassed this measure with homework performance in second place.

This project combines the interest in student attendance, seating arrangement, and machine learning algorithms to predict student grades.

III. Methodology

The first step to performing successful data science experiments is obtaining good data. Therefore, this section first describes where data was obtained, how it was organized, and what precautions were practiced to avoid potential problems. Next the tools used are introduced along with the procedures and experiments performed.

A. Data Collection

For this study, all data was gathered from one institution, Southern Adventist University. Because the driving questions concern a typical classroom setting, the data collected needed to reflect only this setting.

Some courses at SAU only contained a handful of students per a semester. Additionally, several classrooms had movable desks. Because stickers on these desks indicated the seat row and column (as shown in Figure 1), adjustable desks could have introduced systematically flawed data in a study where seat placement is critical. To eliminate both of these issues, only physically large classrooms with bolted desks were selected.

Some faculty raised further concerns that the attendance system had changed as the university relaxed its COVID-19 restrictions. For instance, during its first semester of use, the attendance tracking system (ATS) only permitted students to sit in every other seat, thus restricting student seat choice. However, in following semesters, ATS allowed students to sit in any seat in a classroom. In favor of consistency, only the latter of these systems was used, resulting in two semesters of data (Fall 2021 and Winter 2022).

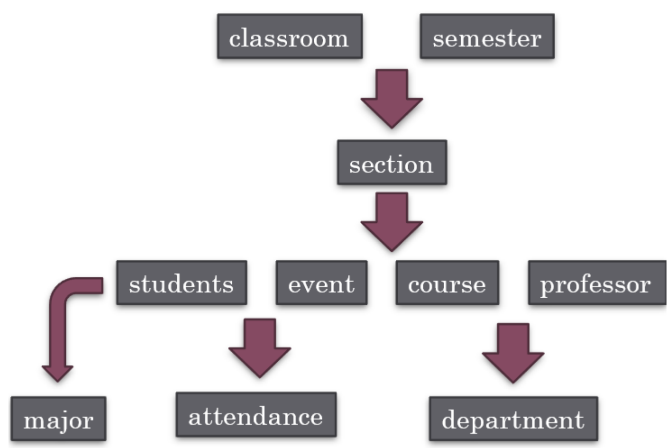

Fig. 2. The natural progression of entities in data collection.

Once locations and times were chosen for data collection, all other entities followed naturally as shown in Figure 2. 159 course sections with twenty or more students enrolled were found using the seventeen chosen classrooms over the two semesters. 2067 students were enrolled in one or more of these sections (see Figure 3 for demographics) which were taught by sixty-three professors representing thirteen departments.

Each section had associated events that represented one class period. The ATS stored data for each student that was present, but it did not always specify if a student was absent from a class. Thus, using several structured query language

[1] fondafultonvilleschools.org
[2] egcsd.org
[3] marktomforde.com

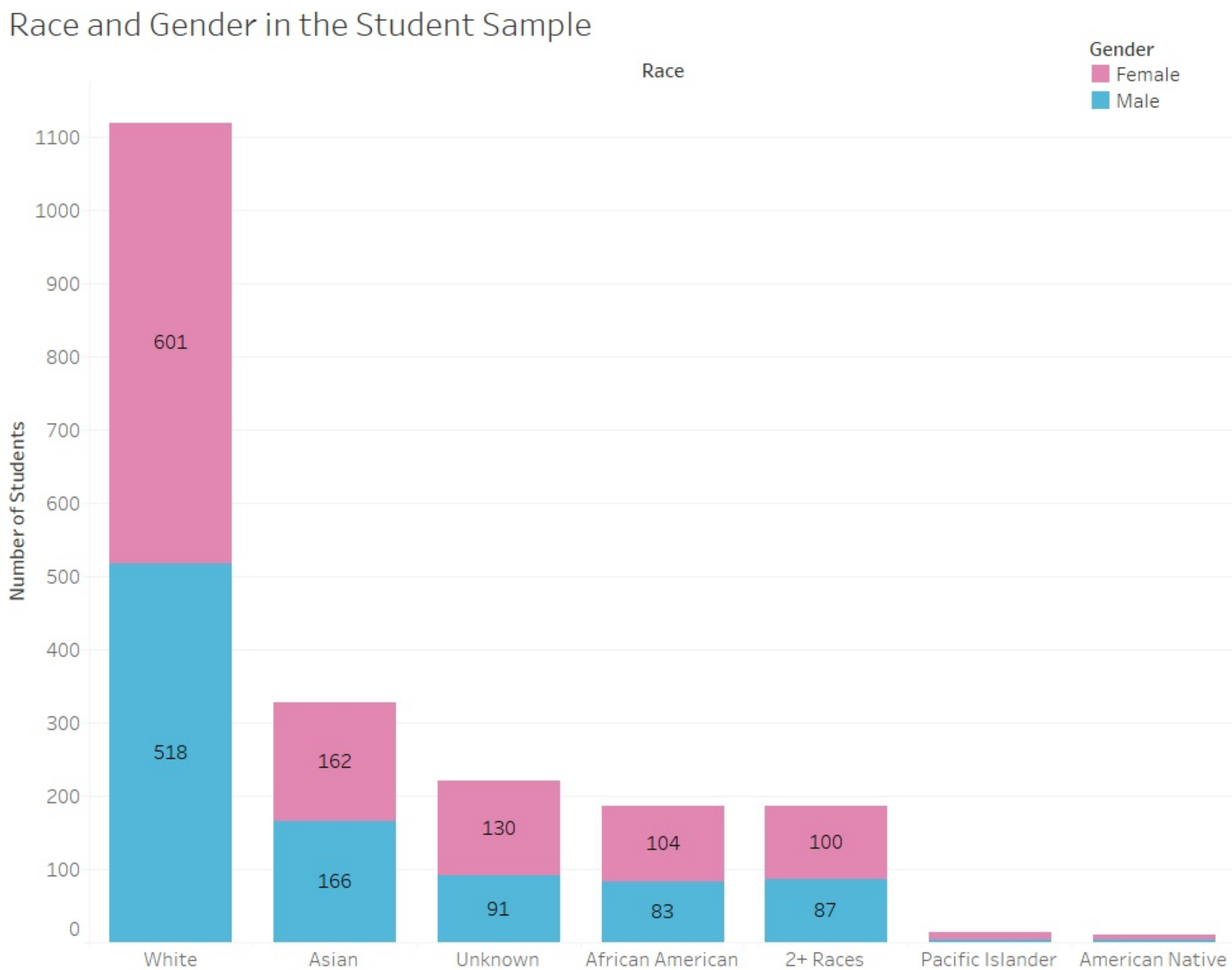

Fig. 3. Demographics of the students in our data. Graph created in Tableau.

(SQL) scripts, this data was imputed. If a student was enrolled in a section but had no record of attending any one of its events, they were assumed absent.

B. Data Organization and Tools

Separating the above entities into relations was the most natural way of storing and organizing the data. Because the data was pulled from a data warehouse partially external to the university, it was not separated into the various entities described above. Thus, Tableau Prep Builder was used to segregate and clean the data.[4]

Because Microsoft Azure was to be used for the data science experiments, it was also chosen to store and serve the data. Using an Azure SQL Database on an Azure SQL Server, we constructed a relational database from the outputs of Tableau Prep Builder. After resolving all the bugs encountered during data collection, the cleaned CSV files provided by Tableau Prep Builder were simply imported as tables into the database using Microsoft SQL Server Management Studio (SSMS). Primary and foreign keys were also configured in SSMS. The resulting schema for this database can be seen in Figure 4. All entities are transitively related using primary/foreign keys.

The initial reasons for using Microsoft Azure were its advertized ease of use and Automated Machine Learning workspace (AutoML). AutoML is an emerging technology that offers automatic training of various machine learning models without the need for code.

In a typical solution, data is first supplied to a model training component. The component trains a prediction model, often even automatically choosing which algorithm is best for solving a given classification, regression, clustering, or forecasting problem. It may also tune the model's hyperparameters.

C. Experiment Plan

The general plan for regression and clustering experiments was to connect the Azure AutoML workspace to the Azure SQL Database so that the AutoML components could automat-

[4] tableau.com/products/prep

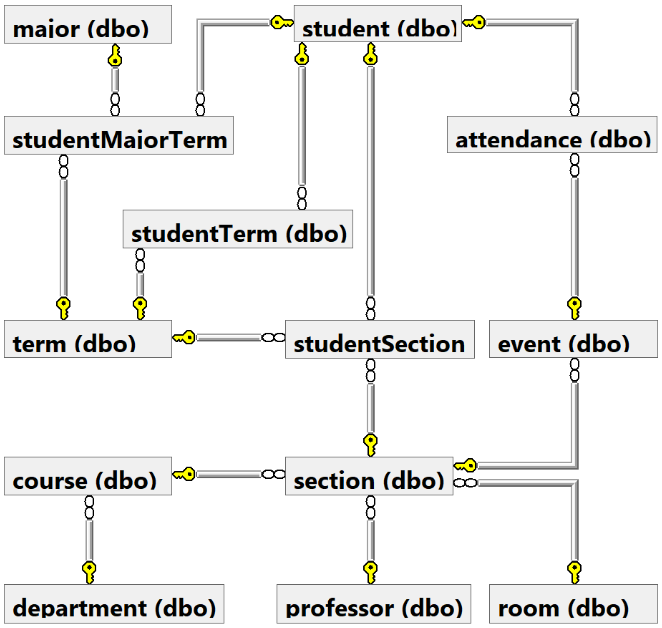

Fig. 4. An abbreviated form of the database schema as visualized by Microsoft SQL Server Management Studio.

ically extract the latest data from the database. Using these two platforms, we created automatic pipelines that built machine learning models for student grades. Finally, various subsets of attributes were provided to the AutoML pipeline to run experiments on.

After producing a model, Azure supplied various metrics associated with that model. For regression, these included error scores and correlation coefficients, allowing for a simple correlation analysis of any subset of attributes. For clustering models, metrics included cluster densities and diameters as well as each record's cluster assignment. Using this information, the model could provide the average values of each attribute in each cluster, which could give insight into how the algorithm naturally organized the data.

To perform logistic regression and data clustering, class data first needed to be transformed into a numerical format. The twelve grade categories "A-F" were converted to the numbers 1-12, respectively. Also, "I" (incomplete) and "IP" (incomplete passing) were assigned values of 13 and 14.[5]

Other categorical variables were converted in a similar manner. For example, there were five categories for attendance status. The labels "Present," "Online," "Late," "Excused," and "Absent" were assigned the values 0-4 respectively.

Further, as shown in the ATS interface in Figure 1, students selected their seat using a numerical column and a *row letter*. Most training models would perform better with a *row number* rather than a letter. Also, the number of rows and spacing between those rows in each classroom varied, rendering any categorical row data inconsistent. To provide the most useful data to the algorithms that would train our models, the row letters were extracted, aggregated for each classroom, and converted to a normalized distance from the front of the classroom. This new attribute, called "distanceToFront," measured how far a student's chosen seat was from the front of the classroom. Values closer to "0" indicate seats closer to the front row of a classroom while those closer to "1" represent seats at the back of a classroom.

The final query fetches the attributes of interest for this project (shown below).[6] These attributes included student demographic information, credit load, hours worked during the semester, distance to the front of the classroom, attendance status, and final grade. This query was run against the Azure SQL database and the resulting data was used as a starting point for Azure's AutoML experiments.

```
select s.isFemale, s.isHispanic, s.race,
  st.housing, st.gradeLevel, st.credits,
  st.tensOfHoursWorked,
  a.distanceToFront, a.seatColumn, a.statusCode,
  sn.finalGradeCode
from attendance a
  join student s on a.studentID = s.studentID
  join studentTerm st on s.studentID = st.studentID
  join studentSection sn on s.studentID = sn.
    studentID
    and sn.termID = st.termID
  join section n on sn.sectionID = n.sectionID
  join event e on n.sectionID = e.sectionID and e.
    eventID = a.eventID
  join course c on n.courseID = c.courseID
where c.departmentID != 'NRSG'
  and c.departmentID != 'PEAC';
```

1) Correlation Analysis: Regression experiments were configured as "jobs" and started in Azure's Machine Learning Workspace. To analyze correlation across different groups of attributes, the experiment was run multiple times with different subsets of the columns shown in the query above. The experiments and their results are presented in the "Results" section.

2) Clustering: While regression can show correlation between sets of independent and dependent variables, it does not allow an algorithm to make its own conclusions about the data. However, clustering algorithms can take data and learn to organize it. This is a classic example of unsupervised learning. Rather than telling the algorithm how the data should be fit, unsupervised learning allows the model to try to group the data based on all the given attributes. After the algorithm has formed clusters of data that minimize some cost function, the clusters can be assessed to form conclusions.

Azure's AutoML platform can perform popular clustering algorithms such as K-means. Perfecting these experiments was the longest part of this project. However, after many cycles of trial, error, research, and reconfiguration, we arrived at an abstract, streamlined clustering solution.

[5]This assumes that not completing a class is a less favorable outcome than failing it. Also "Incomplete Passing" is marked as lower than "Incomplete," but it is not a cause for concern as this represents less than 0.1% of the data.

[6]Notice that courses in the Nursing (NRSG) and Physical Education (PEAC) departments were filtered out. Nursing courses were removed because students were often assigned seats in these courses, thus removing the student's ability to choose their seat. Physical education courses, on the other hand, were not considered a "normal classroom setting."

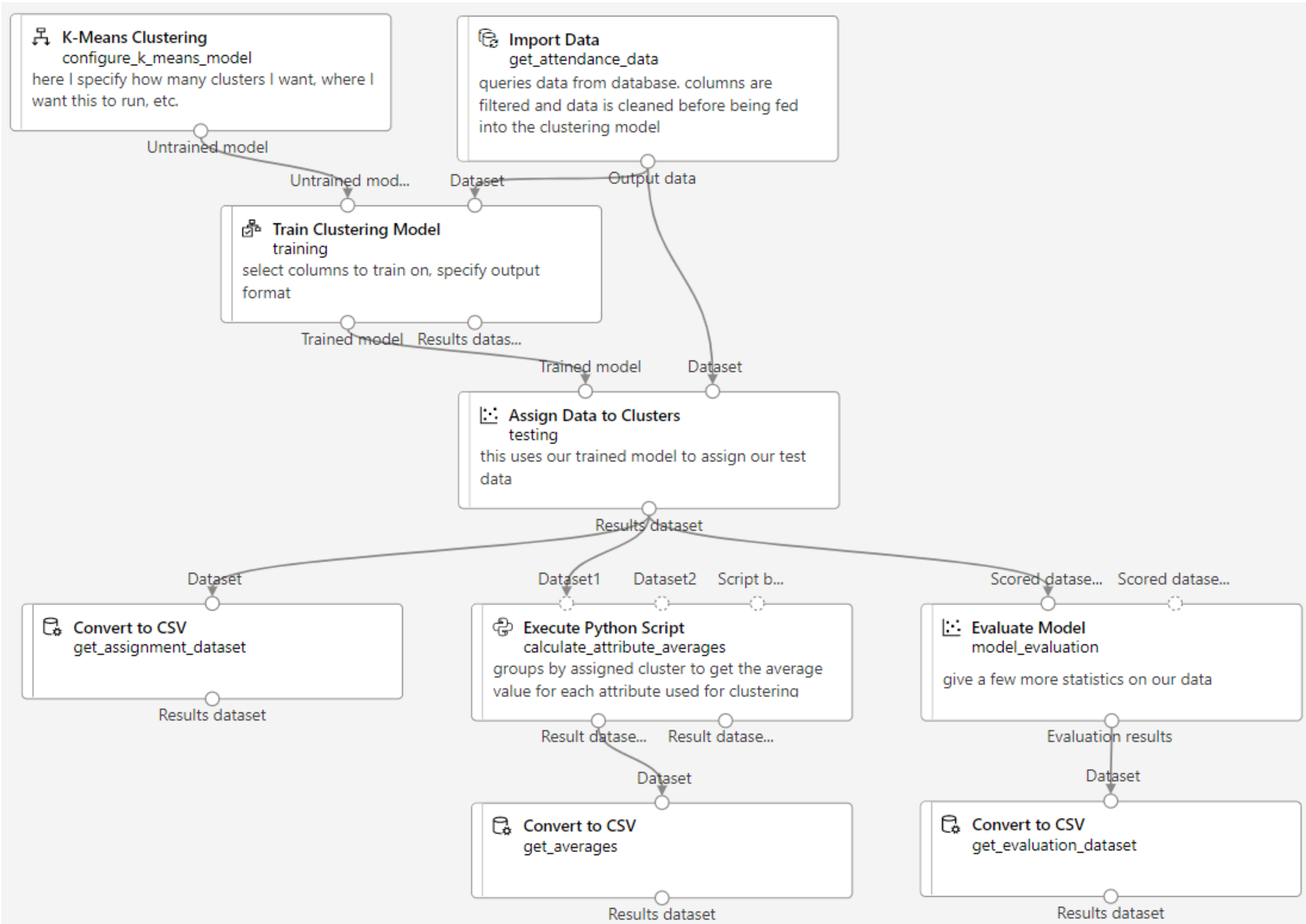

Fig. 5. The clustering pipeline used in this project. Note that to make this image fit better on paper, a data cleaning step and a column selection step were removed from the pipeline. The cleaning step simply imputed missing values using column means.

This solution was created as a pipeline. In a pipeline, several components can be "wired together" to create a single process. This process is capable of gathering data, processing it, training machine learning models, testing those models, and generating result data in automatic succession. The machine learning pipeline shown in Figure 5 was used to perform clustering for this project.

In the **Import Data** component (top right), the pipeline fetches data dynamically from the Azure SQL Database. Different attributes can be selected within or after the query that this component performs. Initially, numerical, textual, and categorical data was being returned. However, after several experiments and database adjustments, only numerical data was fetched using query shown above.

On the top left, the **K-Means Clustering** component is used to initialize and configure an untrained clustering model. Several parameters can be set here: the number of clusters desired in the output, the feature normalization option, a model weight initialization algorithm, and a multi-dimensional distance metric.[7] The **Train Clustering Model** component then trains this model using the imported data.

Assign Data to Clusters takes all the provided data and assigns it to a cluster using the trained model. Using this data, the model can be evaluated in the **Evaluate Model** block which measures average distances between all clusters. It may be possible to generate more evaluation metrics, but we were not able to achieve these results because documentation for this component was inconsistent and generally lacking.

Luckily, Azure provides a mechanism for executing Python code anywhere[8] in a pipeline. A custom Python script only needs to contain an `azureml_main()` function that receives and returns up to two dataframes.[9] The rest is up to the pipeline designer. To generate the necessary data, **Execute Python Script** receives all the records along with their cluster

[7] In our experiments, the number of clusters varied from two to fourteen, normalization was enabled, weights were initialized with the "K-Means++" algorithm, and a Euclidean distance metric was used.

[8] anywhere that the proper input and outputs can be wired, that is

[9] A dataframe is just a specific format for a table of data usually managed by the `pandas` library.

assignments and calculates the average value of every feature for each cluster (see code below).

```
import pandas as pd
def azureml_main(dataset, optional_data = None):
    return dataset.groupby("Assignments").mean()
```

Finally, all the data generated by these components is converted into comma-separated value (CSV) format. The results are discussed below.

IV. Results

A. Correlation Analysis

In the first experiment, all of the interesting attributes mentioned above were fed into the regression model.[10] Azure provided many different performance metrics. The Root Mean Squared Error (RMSE), Explained Variance (EV), Spearman Correlation Coefficient, and R^2 Score are presented in Table I.

This configuration provided a fairly accurate regression model. The Spearman correlation coefficient was over 0.9, suggesting a high correlation between all the input and target attributes. As a visual representation of its accuracy, the regression graph is shown in Figure 6.

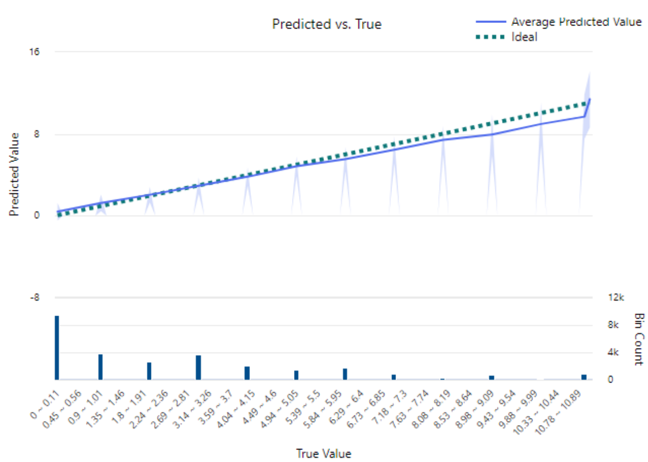

Fig. 6. The predicted values approximate the true values very well.

However, Azure also provided insight into which attributes were most important to the accuracy of the model. As shown in Figure 7, the model relied heavily on class standing, term information, and demographic information. The distance to the front of the classroom falls in fifth place and attendance status in seventh.

Based on the composition of other highly-accurate models, we suspect that the automated machine learning algorithm may have trained the regression model to recognize individual students and predict grades based on that student's historical performance.

To avoid this issue, the following experiment only retained two attributes, attendance status (statusCode) and distance to the front of the classroom (distanceToFront). After trying

[10]This included demographics, course and work amounts, attendance information, and finalGradeCode - the target variable.

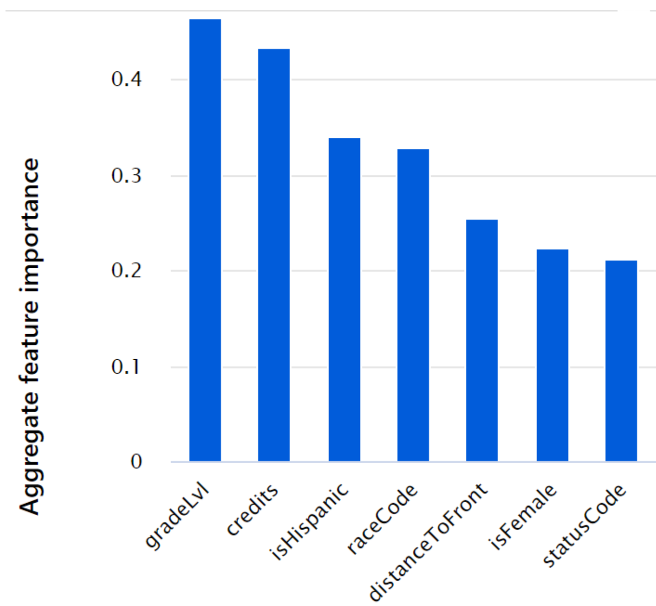

Fig. 7. The top seven attributes used by the first regression model, ranked in order of predictive importance.

to correlate these attributes with the students' final grades for forty-eight minutes, the best model that Azure trained offered a mean absolute error that was nearly a fifth of the entire range. Very little of this error was explained by the variance in attendance data. Additionally, both correlation metrics suggested no correlation between the attendance data and student grades. In fact, the R^2 score suggests that there is a 95.7% chance of getting this correlation from unrelated attributes.

TABLE I
Regression Performance Metrics

Attributes Used	RMSE	EV	Spearman	R^2 Score
All shown in SQL query	1.110	0.859	0.906	0.859
statusCode, distanceToFront	2.314	0.043	0.186	0.043
statusCode	2.923	0.020	0.107	0.020
distanceToFront	2.922	0.020	0.140	0.020

Isolating the attendance status and seat row attributes even further did not improve results. Both prediction models produced separately by these two attributes yielded even lower correlation values.

Overall, the results of the correlation analysis showed that attendance and seat choice could not be used to accurately or precisely predict student grades in the data obtained from Southern Adventist University.

B. Clustering

Before other hyperparameters were tuned, the number of clusters, k, had to be decided. One approach, commonly called the "Elbow Method," attempts to utilize the law of diminishing returns. As k increases, the average size of each cluster naturally decreases. However, the number of clusters cannot be allowed to grow indefinitely. Thus, if several values for k are tested and their average distance between a point and its assigned cluster is graphed, the inflection point can be selected as the best value for k. This method ensures that the average cluster size is relatively small while k is also minimized.

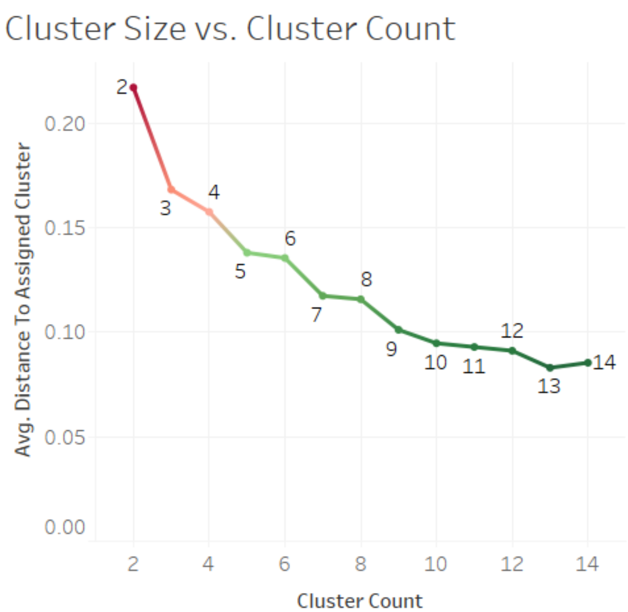

Fig. 8. The average distance to a cluster compared across various number of clusters.

As shown in Figure 8, the inflection point is slightly unclear. Thus, experiments were first performed with $k = 5$ clusters. The first experiment used all the queried columns. However, the clustering algorithm grouped data primarily using demographic information.[11]

After limiting the number of attributes available to the clustering algorithm, k-means grouped the data primarily based on attendance status and distance to the front of the classroom. However, five clusters proved too many for this set of attributes so the count was reduced to $k = 3$.

The averages in Table II were obtained using three clusters and two input attributes. Data placed in Cluster 0 represented a mostly "Present" attendance status and a seat choice roughly halfway from the front of the classroom. Nearly all of the attendance records in this cluster had perfect scores of "A". Cluster 1 has similar averages, but with grades much closer to "C+" and "C". Finally, the cluster with chronically late attendance and a seating preference slightly beyond the halfway point has an average grade around a "B+" or "B".

The experiment was repeated with three clusters and no results wavered. Although Cluster 2 does show a group of students that is often late and performs worse than average, the three clusters together are inconclusive.

[11]For example, four of the five clusters would be entirely based on a student's gender and hispanic status.

[12]This feature was originally named "Average Distance to Cluster Center" and has been scaled so that the largest value is 1.

TABLE II
AVERAGE ATTRIBUTE VALUES IN THREE CLUSTERS

	status	distanceToFront	grade	radius[12]	point count
Cluster 0	0.020	0.575	1.004	0.404	104833
Cluster 1	0.030	0.613	6.384	0.569	42212
Cluster 2	2.922	0.610	3.618	1.000	18143
Total	0.341	0.589	2.666	0.512	165188

Finally, the experiments were repeated with only two clusters. The results are displayed in Table III. As seen before, distance to the front of the classroom is nearly the same for both clusters. However, Cluster 1 specifically represents records where a student was, on average, late to class. This cluster has an average grade between "B+" and "B". Cluster 0, on the other hand, displays much better attendance and grades on average between "A-" and "B+".

TABLE III
AVERAGE ATTRIBUTE VALUES IN TWO CLUSTERS

	status	distanceToFront	grade	radius[13]	point count
Cluster 0	0.022	0.586	2.548	0.689	147033
Cluster 1	2.921	0.610	3.624	1.000	18155
Total	0.341	0.589	2.666	0.651	165188

Though interesting, these clusters are still far from ideal. One major issue is that they are unbalanced. One includes 89% of the attendance data and the other only 11%. A more balanced dataset is desirable.

To further validate Azure's automated clustering algorithm, clustering was also performed in Weka 3.[14] The results from this experiment were identical to those obtained from Azure's k-means clustering.

V. CONCLUSION

The correlation analysis did not support any correlation between a student's attendance and their performance in class. With correlation coefficients as low as $R^2 = 0.020$, the lack of any relationship between the independent and dependent variables is easier to argue than a correlation or causation between them.

Clustering provided more insight. A cluster of data was identified that represented tardiness or absence along with lower grades. With further work and more balanced data, this Machine Learning approach holds the most promise.

No experiments showed that sitting nearer to the front of the classroom positively impacted a student's grades. Instead, this attribute seemed generally unimportant. It only seemed useful for identifying particular students (assuming students often sat in the same seat throughout the semester).

Overall, this study does not conclusively prove correlation between student attendance and grades, nor does it prove

[13]See above.

[14]cs.waikato.ac.nz/ml/weka

that student seating is irrelevant to grades. It does present methods that have and have not worked well for analyzing student attendance with machine learning. Clustering, when tuned properly, could provide key insights in future studies.

VI. Future Work

A. Platform

We faced several issues during this project. Primary of these is that Azure's AutoML Workspace was likely not the best tool for the job. The platform promised an easy-to-use automated machine learning process, but proved very difficult to navigate and configure. This is probably because this platform does not target scientific research, but rather automatic model and business intelligence development and their deployment.

Thus, in future work, a different tool would be used. Weka 3, a software for machine learning written in Java, shows more promise with the added benefit that it is free to use. Moreover, it offers many types of clustering algorithms while Azure AutoML only offers one. Thus, any future clustering experiments will probably be performed in Weka 3.

B. Data Integrity

Several aspects of and incorrect assumptions about the original data may have caused issues in this project. The experiments might benefit from having data that has been balanced to include the similar amounts of high and low scores. More data with final grades below "B" should be collected and combined with the current dataset. Also, a better balance is needed between the different attendance categories. Currently, "Late," "Absent," and "Excused" attendance records represent only 11.6% of attendance data.

Other assumptions remain untested. For instance, "distanceToFront" was relative for each classroom. Thus, some students in the back of one classroom may be farther away from their instructor than students in the back of another classroom that contains fewer rows. If physical distance has a significant impact, that factor is lost in this process, as both seats in the rear of any classroom would be assigned values of "1". Also, data about the horizontal placement of a student within a classroom was not processed or converted into a numerical format.

Moreover, correct data entry cannot be guaranteed unless researchers or their representatives ensure that a student's self-reported attendance matches reality. For example, if one student accidentally selects the wrong seat in the ATS, they cannot change their selection. Even worse, in the event of such an accident other students cannot select that seat even if they are actually sitting in it. This often triggers a domino effect of incorrect data as students select random seats in the classroom just to satisfy class attendance requirements before they are marked as "Late."

C. Approach

In this project, models were trained on records of every piece of attendance information with a final grade attached. This results in a system where courses that meet many times during a week have more influence on machine learning models than courses that meet fewer times. Classes that meet four times a week should not have this type of privilege over courses that meet once or twice a week for the same total hours.

A better method might calculate average metrics such as "distanceToFront" and attendance for each student in each course. This data could then be provided to the AutoML platform, avoiding the "meeting times" issue. It could even be systematically scaled by the number of credits offered in the course to further even out the weight of the information.

Other variables, such as a professor's teaching style, adjustable tardiness thresholds,[15] and time of day were not controlled. To mitigate this, a more structured study needs to be planned and documented.

In summary, many areas require improvement and further research. These studies and their experiments should take all the above factors into account.

References

[1] H.-C. Sun, X.-F. Liu, Z.-W. Du, X.-K. Xu, and Y. Wu, "Mitigating covid-19 transmission in schools with digital contact tracing," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 6, pp. 1302–1310, 2021.

[2] D. McCauley, "Covid-19 forced rapid changes in education, but which changes should we keep?" *CSA News*, vol. 66, no. 10, p. 44, 2021.

[3] "Every school day counts: The forum guide to collecting and using attendance data," Feb 2009. [Online]. Available: https://nces.ed.gov/pubs2009/attendancedata/chapter1a.asp

[4] V. Alexander and R. E. Hicks, "Does class attendance predict academic performance in first year psychology tutorials?" *International Journal of Psychological Studies*, vol. 8, no. 1, 2016.

[5] F. D. Becker, R. Sommer, J. Bee, and B. Oxley, "College classroom ecology," *Sociometry*, pp. 514–525, 1973.

[6] S. Kalinowski and M. L. Toper, "The effect of seat location on exam grades and student perceptions in an introductory biology class." *Journal of College Science Teaching*, vol. 36, no. 4, 2007.

[7] C. Zabriskie, J. Yang, S. DeVore, and J. Stewart, "Using machine learning to predict physics course outcomes," *Physical Review Physics Education Research*, vol. 15, no. 2, p. 020120, 2019.

[15] Professors were able to set what counted as "Late" in the ATS for each class. Thus, in some classes selecting a seat five minutes after the start of the event marked the student as "Late" while in others the system would record that they were "Present."