Midterm Exam Report: Airbnb Dataset

Created by: Solis, Alexander Jon S. & Virata, Sean Maverick A. & Bolante, Shekinah D.

Link to our files: Finals Project

This project uses the Student Performance & Behavior Dataset made and collected by Mahmoud Elhemaly from a private learning provider. This dataset contains 5000 records and includes key attributes necessary for exploring patterns, correlations, and insights related to academic performance.

These are the columns from the original dataset before data cleaning:

- Student_ID: Unique identifier for each student.
- First_Name: Student's first name.
- Last_Name: Student's last name.
- Email: Contact email (can be anonymized).
- Gender: Male, Female, Other.
- Age: The age of the student.
- Department: Student's department (e.g., CS, Engineering, Business).
- Attendance (%): Attendance percentage (0-100%).
- Midterm_Score: Midterm exam score (out of 100).
- Final_Score: Final exam score (out of 100).
- Assignments_Avg: Average score of all assignments (out of 100).
- Quizzes_Avg: Average quiz scores (out of 100).
- Participation_Score: Score based on class participation (0-10).
- Projects_Score: Project evaluation score (out of 100).
- Total_Score: Weighted sum of all grades.
- Grade: Letter grade (A, B, C, D, F).
- Study_Hours_per_Week: Average study hours per week.
- Extracurricular_Activities: Whether the student participates in extracurriculars (Yes/No).
- Internet_Access_at_Home: Does the student have access to the internet at home? (Yes/No).
- Parent_Education_Level: Highest education level of parents (None, High School, Bachelor's, Master's, PhD).
- Family_Income_Level: Low, Medium, High.
- Stress_Level (1-10): Self-reported stress level (1: Low, 10: High).
- Sleep_Hours_per_Night: Average hours of sleep per night.

The dataset underwent several cleaning and data transformation using python to ensure quality and consistency:

1. **Handling missing values**
   To handle the missing values, we used the pandas fillna() method and filled the missing values with the median values since there is a large spread between the data points on

the Assignments_Avg and Attendance(%) which may affect the mean if we used it which may results to overskewed data.

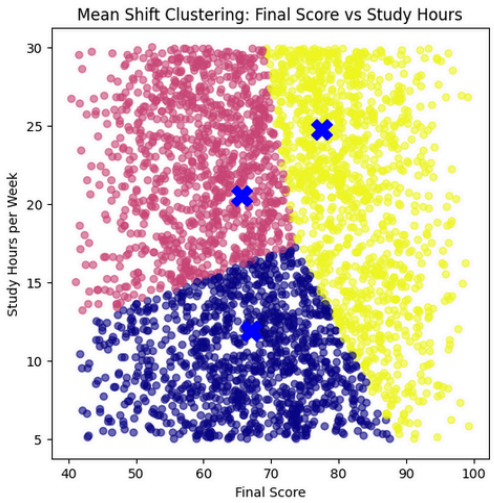
2. **Data Removal**
   We've utilized both the .dropna() and .drop() pandas functions in order to drop rows remained null after our initial cleaning and we dropped the columns: Student_ID, First_Name, Last_Name and Email which are unnecessary data columns for our analysis and machine learning.

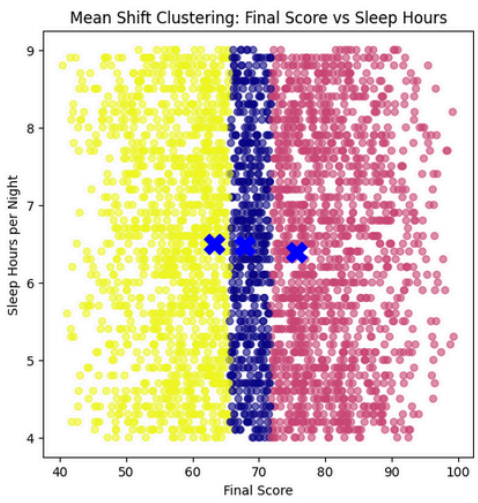
Python Machine Learning:

For this project, we've used two types of machine learning algorithms which are the Meanshift clustering algorithm and the Random Forest classifier algorithm. We used two types of algorithms since the task given to us requires us to cluster students by their performance and use classify which students are at academic risk.

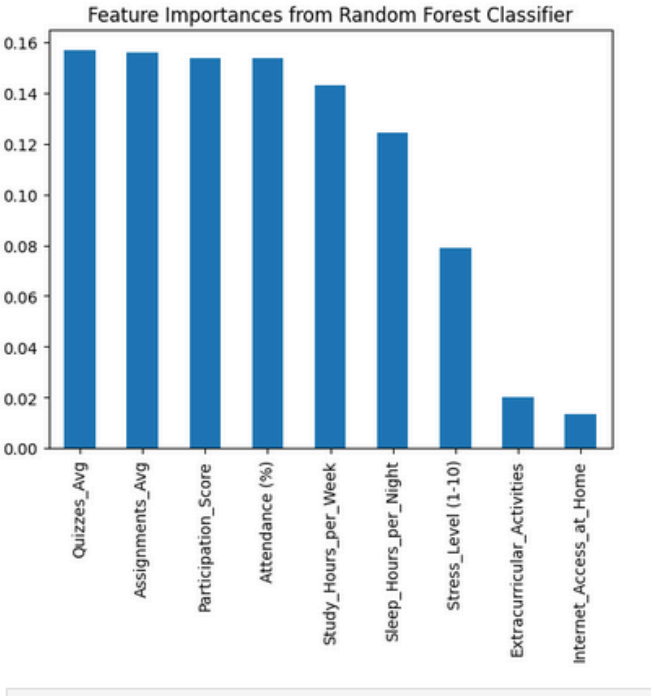
We chose the Meanshift clustering algorithm alongside the K-Means clustering we did on Rapidminer is because they may be similar in terms of how they cluster students, what's different between them is that K-Means clustering requires a predefined number of clusters before training which may force the data into a fixed number of cluster that may not exist. Meanwhile, Mean Shift automatically determines the number of clusters by detecting dense areas in the data which makes it suited for discovering, natural or unevenly sized student clusters.

Towards the results, we clustered the students using two parameters that may affect their final grade which are their study hours and their sleep hours. As we can see on the results visualized by the scatterplot, it divided the students into three groups for each category. It showed that on the pink dots there are students who give more time on studying yet still have a lower to average score, then the blue dots that are students who gave an average study hours and does have average results and lastly, on the yellow dots there are many students who gave a higher study time which resulted to higher final scores. On the other hand, the results towards the sleep hours show that sleep is not a big factor that may affect their final grade as we can see with the vertical alignment. Which means that no matter how a student may sleep, their score mostly depends on their study habits.

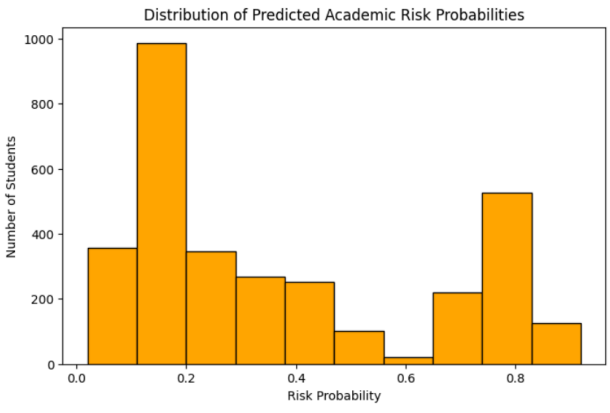




For our classification machine learning model, we used the Random Forest Classifier as it can handle non-linear patterns in data while reducing overfitting by combining multiple decision trees. The model used nine features which are Assignments_Avg, Quizzes_Avg, Study_Hours_per_Week, Sleep_Hours_per_Night, Attendance, Extracurricular_Activities, Internet_Access_at_Home, Stress_Level, Participation_Score. These features are different factors that may help identify which students are at Academic Risk. As shown by the bar plot, the machine learning algorithm prioritized the Quizzes_Avg and the Assignments_Avg features as they directly affect the academic performance which can be a big sign for academic risk if the students' grades are low, The next ones which are the Participation_Score, Attendance, Study_Hours_per_Week and Sleep_Hours_per_Night tells us that being able to form a good study habit is important in identifying students that are academic risks. Meanwhile, Extracurricular_Activities, Internet_Access_at_Home and Stress_Level are not given much importance like the others as they may not directly impact academic performance even though Stress_Level plays a much higher role than the other two in the bottom.



The distribution of students' estimated academic risk probability predicted by the Random Forest Algorithm is shown by the histogram. The majority of students are in the low-risk category, which is between 0.0 and 0.2. Around 900 to 1,000 students have a 20% chance of being at academic risk. There are fewer students in this middle-risk category as the chance increases approaching the medium range (0.2 to 0.5). About 500 kids make up a smaller but noticeable group that falls into the high-risk range (0.7 to 0.8), indicating that they are much more likely to experience academic issues. Lastly, only 100 to 150 students are required to get academic assistance as they are classified into the very high-risk group (0.9 to 1.0). Overall, the results suggest that the majority of students have a low probability of academic risk, while only a minority exhibit high or critical risk levels.



Power BI Dashboard:

In this Project, we created this Power BI Dashboard to provide a data-driven understanding of student performance through clustering and risk-level classification. The goal for this analysis is to be able to group students according to their said academic results and demographic profiles, this can be extremely valuable for school educators and administrators to be able to identify patterns that influence students' academic outcomes and determine varying degrees of academic risk. The dataset we used contains student records, including metrics such as total score, attendance rate, stress levels, gender, socioeconomic information, and other performance indicators.

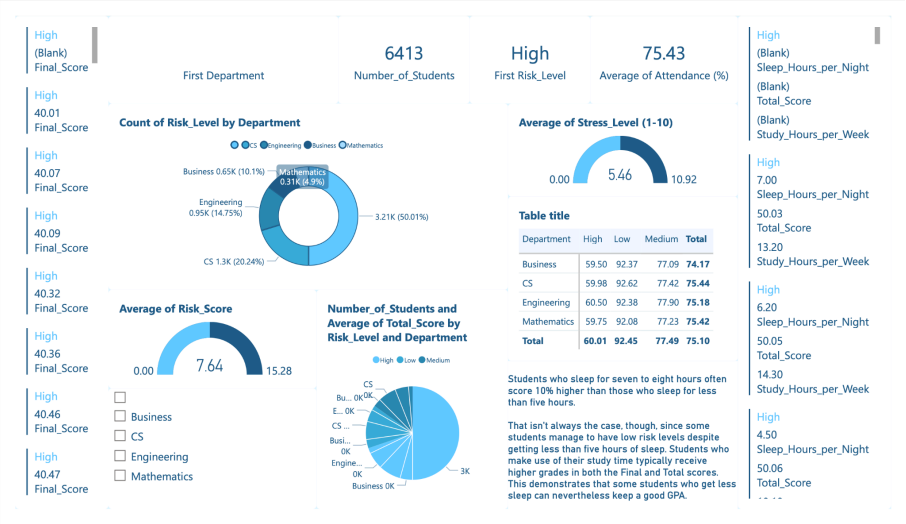
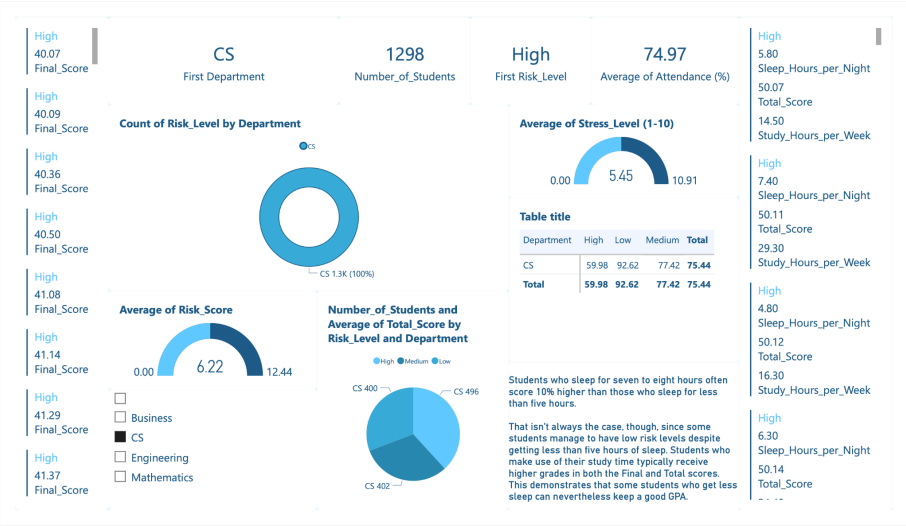


Figure 1 here represents the full overview of all the departments combined. At the very top, it is visible, the three main KPI Cards we used to summarize the core metrics: the total number of students being (6,413), the predominant risk level being (High), while the average attendance rate is (75.43%). These KPI Cards serve as a key indicator of overall academic health within the student population. The "Count of Risk Level by Department" Donut chart visualizes how academic risk levels of students are distributed among departments.



The Computer Science (CS) Department constitutes the largest portion (50%), followed by Engineering, Business, and Mathematics. This Dashboard provides a clear view of how academic challenges completely differ by the field of study.

To complement this, the “Average of Stress Level” gauge chart was able to reveal an average value of 5.46 on a 1-10 scale, suggesting that while stress levels are moderate, they may still impact performance. The “Average of Risk Score” gauge shows a value of 7.64, indicating that the student population overall falls on the higher end of the academic risk spectrum.

On Figure 2, The updated KPI Cards show 1,298 students, maintaining a HIgh Risk level with an Average Attendance rate of 74.97% which is slightly below the overall average, suggesting attendance may contribute to the higher risk category. The Data Table and Pie Chart provide further clarity on the internal segmentation of CS students. The table was able to reveal a relatively consistent score distribution across High, Medium, and Low risk categories, while the pie chart highlights how the department’s performance highly remains steady despite differing risk levels This insight suggests that CS students were able to maintain balanced performance even under moderate academic stress. Lastly, as shown on the behavioral note, the deep correlation between sleep and performance persists in this filtered view: students who maintain consistent study hours even with varying sleep patterns can achieve competitive total scores. This highlights the finding that structured study habits may mitigate the probability of an academic risk more effectively than sleep duration alone.

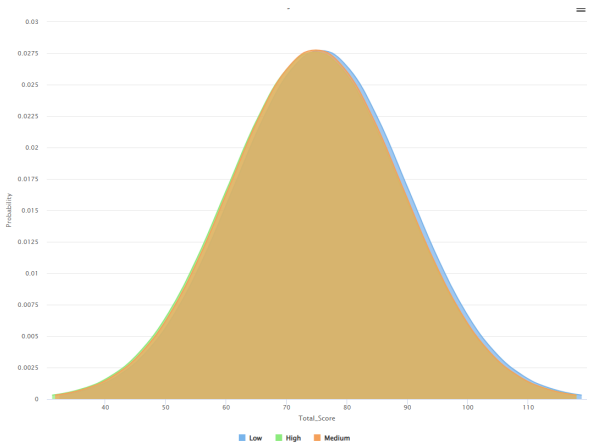
Altair AI Studio:

This project also utilized Altair AI Studio, formerly known as RapidMiner, to conduct K-means clustering to the student database to determine clusters with common patterns and traits related to the student’s academic risk. Unlike in Python, AI Studio does not support Mean Shift clustering. In this part of the project, K-means clustering is the sole clustering method used to identify student clusters. Figure 3 shows the table that presents 24/3,206 examples of the dataset their key corresponding attributes such as age, total score, gender, weekly study hours, daily sleep hours, academic income, and academic risk.

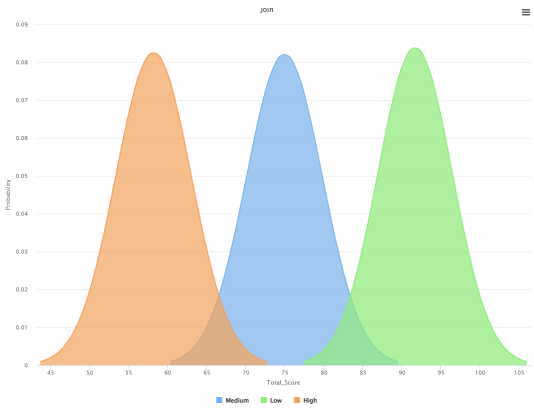
Row No.	id	cluster	Age	Total_Score	Study_Hour...	Sleep_Hour...	Academic_r...	Gender	Family_Inco...
1	1	cluster_0	24	70.300	20.700	6.200	Medium	Male	Low
2	2	cluster_2	22	92.880	15.100	5.100	Low	Female	Low
3	3	cluster_1	23	66.130	15.400	7.100	High	Female	High
4	4	cluster_1	21	62.080	8.500	5	High	Male	High
5	5	cluster_2	21	95.620	8	8.800	Low	Female	Low
6	6	cluster_1	23	58.250	13.200	4.100	High	Female	Low
7	7	cluster_1	22	61.340	28.900	6.600	High	Male	High
8	8	cluster_1	21	59.060	18.100	4.500	High	Male	High
9	9	cluster_1	19	50.690	19.700	4.800	High	Female	Low
10	10	cluster_1	23	55.250	6.200	7.600	High	Female	Medium
11	11	cluster_0	20	66.940	23.500	4.400	Medium	Male	Low
12	12	cluster_2	18	88.470	16.900	8.100	Low	Male	Medium
13	13	cluster_1	19	61.810	24.900	6	High	Female	Low
14	14	cluster_2	21	87.350	21.600	6.200	Low	Female	Medium
15	15	cluster_2	18	90.240	23.500	5.600	Low	Male	High
16	16	cluster_0	21	72.200	11.200	4	Medium	Female	Low
17	17	cluster_0	23	68.530	7	6.700	Medium	Female	High
18	18	cluster_2	22	85.100	28.300	4.100	Low	Female	Medium
19	19	cluster_1	22	52.910	19.100	6.300	High	Male	Low
20	20	cluster_0	22	82.050	16.800	6	Medium	Female	High
21	21	cluster_2	21	86.480	5.800	5.600	Low	Female	Low
22	22	cluster_0	18	81.780	12.900	8.800	Medium	Female	Low
23	23	cluster_1	19	52.860	28.300	4.600	High	Female	Medium
24	24	cluster_2	22	97.630	15.400	4	Low	Female	Low

The K-means method managed to identify three clusters that are in line in low, medium, and high academic risk groups. Figure 4 shows the scatterplot of academic risk clusters relative to their Total_score with two different colors indicating the students’ gender. The distribution of male and female students among three different risk clusters are even and indicates that neither gender exhibit increased risk over the other.

When looking over family income level, total score, and academic risk, all three income levels exhibit a similar distribution of low, medium, and high academic risk students as shown in Figures 5 & 6 which are scatter plot and bell curve, respectively. However, it is evident that high income level families appear relatively less than low and medium income equivalents as shown in their visual density in figure 5.



Lastly, figure 7 shows the most common scores for each academic risk clusters. The three peaks of the bell curve indicating three major averages among their group. Majority of low academic risk students tend to have scores close to 91. Meanwhile, medium risk students scores around 75. While high risk students mostly tend to have total scores close to 58.



Simply, the k-clustering method in AI Studio have managed to identify the three distinct academic risk groups based on their total scores. It is worth mentioning that certain income levels have a slightly greater population than the other yet it does not affect the results of the project. The visualization methods show that there are no major factors other than total scores that can be correlated to a student's academic risk of failure. Therefore, a student's academic risk is independent of their age, gender, family income, weekly study hours, and daily sleep hours.

Conclusion:

Overall, The analysis truly demonstrates that academic performance and risk levels are all influenced by a mere combination of factors, such as attendance, stress, and study habits.

Department such CS and Mathematics show relatively higher averages despite elevated risk levels, implying that internal motivation and study efficiency play in a students' significant roles. The use of clustering allows administrators and educators to easily identify which segments of students may require intervention or targeted support.