# Introduction

This report presents a comprehensive analysis of a dataset collected from FPT Polytechnic College, specifically from the Hanoi campus. The dataset encompasses academic performance records for students across three semesters, detailing their attendance rates, average scores, and other related metrics. The primary objective of this analysis is to provide insights into the dataset structure, which is intended for predicting student behavior, specifically to forecast whether students are likely to drop out in the third semester. The data spans from Summer 2018 to Spring 2024.

# Dataset Overview

The dataset contains many entries and multiple columns. Each entry corresponds to a student's performance data across three semesters. Below is a summary of the dataset structure:

* Total Entries: 6840
* Total Columns: 11
* Total 524 records are students who dropped out (THO), 6316 records are students who continue studying (HDI)

# Column Descriptions

1. **student\_code**: (String) Unique identifier for each student.
2. **semester\_1**: (Object) JSON-like string containing detailed information for each subject in the first semester. This includes:
   * **full\_subject\_code**: Full code of the subject.
   * **attendance\_rate**: Attendance rate for the subject.
   * **passed**: Whether the student passed or failed the subject.
   * **average\_score**: Average score for the subject.
   * **number\_of\_credit**: Number of credits for the subject.
   * **learnt\_times**: Number of times the subject was taken (default is 1).
3. **semester\_2**: (Object) Similar to semester\_1 but for the second semester.
4. **semester\_3**: (Object) Similar to semester\_1 but for the third semester.
5. **semester\_3\_status**: (String) Status of the student at the end of the third semester, which includes indicators such as 'HDI' (possibly indicating 'High Distinction') or dropout status (THO).
6. **semester\_1\_attendance\_rate**: (Float64) Overall attendance rate for the first semester.
7. **semester\_1\_average\_score**: (Float64) Overall average score for the first semester.
8. **semester\_2\_attendance\_rate**: (Float64) Overall attendance rate for the second semester.
9. **semester\_2\_average\_score**: (Float64) Overall average score for the second semester.
10. **semester\_3\_attendance\_rate**: (Float64) Overall attendance rate for the third semester.
11. **semester\_3\_average\_score**: (Float64) Overall average score for the third semester.

# Data Characteristics

The data reveals the following characteristics:

* **Student Code**: Each student has a unique identifier.
* **Semester Data**: The semester columns contain JSON-like strings, providing detailed attendance and performance data for multiple subjects. Each subject entry includes the subject code, attendance rate, pass/fail status, average score, number of credits, and number of attempts.
* **Attendance and Scores**: Attendance rates and average scores are represented as floating-point numbers.
* **Completion Status**: The semester\_3\_status column indicates the completion status, which includes potential dropout information.

# Sample Data

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| student\_code | semester\_1 | Semester\_2 | Semester\_3 | semester\_1\_attendance\_rate | semester\_1\_average\_score | semester\_2\_attendance\_rate | semester\_2\_average\_score | semester\_3\_attendance\_rate | semester\_3\_average\_score | semester\_3\_status |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

# Data Visualization and Analysis

To gain deeper insights into the dataset, we plotted various charts and analyzed them to understand the patterns and trends within the data. Below are the visualizations and their analyses:

## Boxplot of Average Scores for Different Statuses

A graph of a box diagram

Description automatically generated with medium confidence

Figure 1 Boxplot comparing the average scores of students who continued (HDI) and those who dropped out (THO).

**Analysis**:

* The median average score for students who continued their studies (HDI) is approximately 7, while for those who dropped out (THO), it is around 5.
* The lower whisker for HDI students extends to around 4, with numerous outliers below this point, suggesting a small subset of continuing students with significantly lower scores. For THO students, the lower whisker extends to 0.
* The interquartile range (IQR) for HDI students is relatively tight, indicating consistent and high scores. In contrast, the IQR for THO students is wider, reflecting more variability in scores.
* The boxplot highlights that higher average scores are strongly associated with the likelihood of students continuing their studies, whereas lower scores are more common among those who drop out.

## Boxplots of Average Scores by Semester and Student Status

A diagram of a graph

Description automatically generated with medium confidence

Figure 2 Boxplots comparing the average scores of students who continued (HDI) and those who dropped out (THO) for 3 semesters

**Analysis:**

* Higher Scores: Students who continued their studies (HDI) consistently show higher median scores (≈ 8) across all three semesters compared to those who dropped out (THO, ≈ 7).
* Consistency: The IQR for HDI students (≈ 7 to 8.8) indicates high and consistent scores, while the IQR for THO students (≈ 6 to 8) shows more variability.
* Early Identification: The lower whiskers for THO students extend to ≈ 3, indicating the presence of students with very low scores who may benefit from early intervention and support.

## Attendance Rates and Student Status

A diagram of a group of people

Description automatically generated with medium confidence

Figure 3 Boxplot comparing the attendance percentages of students HDI and THO

Analysis:

* **Higher Attendance:** HDI students have a higher median attendance rate (≈ 80%) compared to THO students (≈ 65%).
* **Consistency:** The IQR for HDI students indicates more consistent attendance (75% to 85%), while the IQR for THO students shows more variability (55% to 75%).
* **Early Identification and Support:** The presence of lower outliers in THO students' attendance rates highlights the need for early identification and targeted support to prevent dropouts. High attendance rates are strongly associated with student retention, emphasizing the importance of encouraging regular class attendance to improve academic outcomes.

## Boxplots of Attendance Rates by Semester and Student Status

A diagram of a graph

Description automatically generated with medium confidence

Figure 4 Boxplots comparing the attendance rates of HDI and THO for 3 semesters.

Analysis:

* **Consistently Higher Attendance:** HDI students have a consistently higher median attendance rate (≈ 80% in the second and the third semester but 70% in the first semester) compared to THO students (≈ 70-75% but much lower in the third semester ≈ 40%) across all three semesters.
* **Increasing Variability:** The IQR for both HDI and THO students widens from Semester 1 to Semester 3. For HDI students, the IQR increases from ≈ 65-70% in Semester 1 to ≈ 70-90% in Semester 3. For THO students, the IQR shows a more dramatic increase, from ≈ 60-70% in Semester 1 to ≈ 25-65% in Semester 3.
* **Greater Variability for THO:** The increase in IQR is more pronounced for THO students, indicating greater variability in attendance rates as semesters progress. This widening gap suggests that attendance becomes more inconsistent among students who eventually drop out.
* **Early Identification:** The presence of lower whiskers extending to as low as 0% for THO students in Semester 3 highlights the need for early identification and targeted support to improve attendance and reduce dropout rates. Regular and high attendance rates are strongly associated with student retention, emphasizing the importance of maintaining good attendance to achieve academic success.

## Attendance Rate And Average Score By Semester

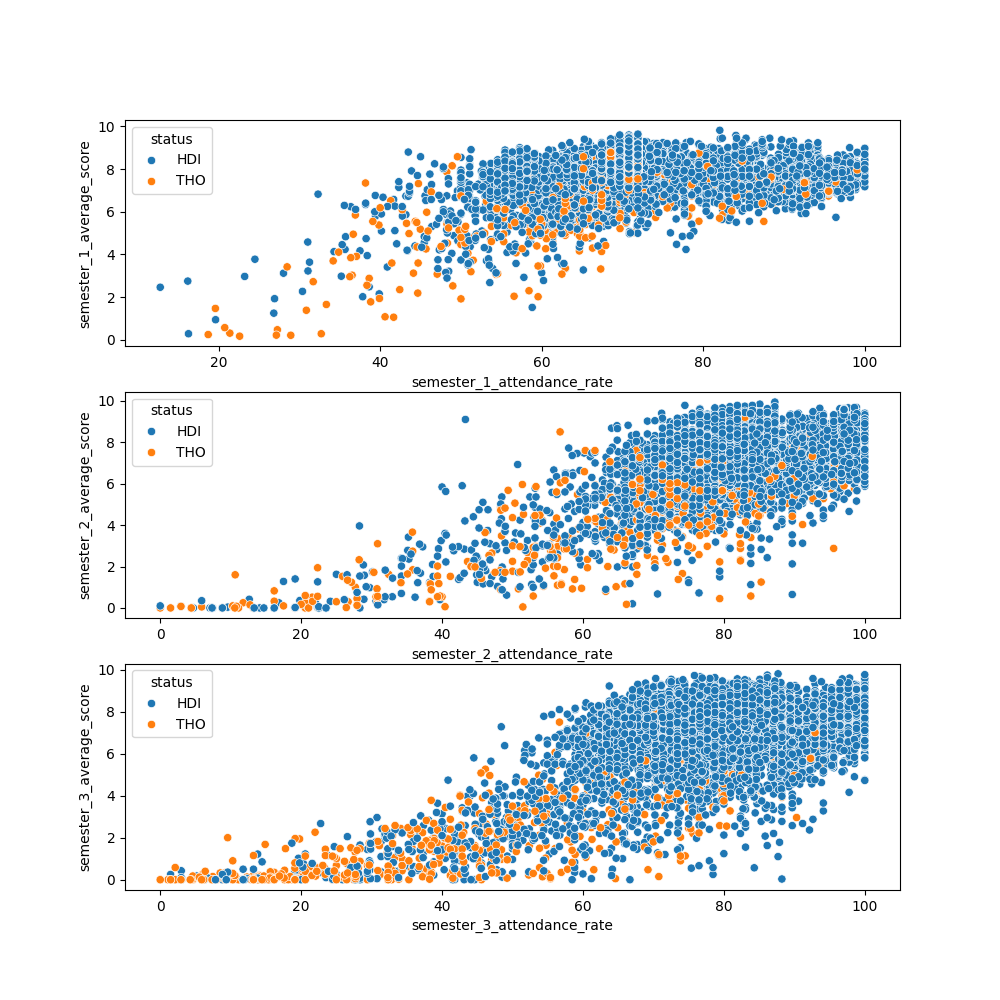


Figure 5 Scatterplot attendance rate and average score by semester

**Analysis**:

All three scatterplots show a positive correlation between attendance rate and average score. This means that students who attend class more often tend to have higher average scores. However, the strength of the correlation varies from semester to semester. Semester 1 has the strongest correlation, followed by semester 2, and then semester 3.

Across all three semesters, there is a clear positive relationship between attendance rates and average scores. Students who continue their studies (HDI) generally achieve higher academic scores and attendance rate compared to those who have dropped out (THO). On the third semester the positive relationship of attendance rate and average score seem to be weaker with the group of students who dropped out.

## Relationship Between Subject Average Score And The Study Status

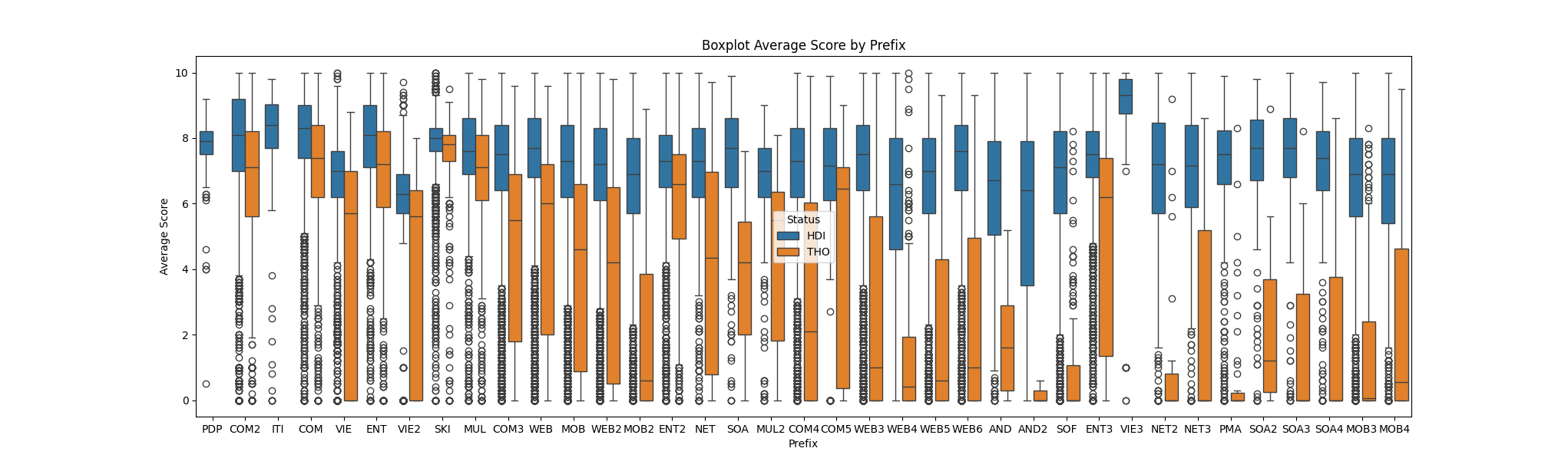


Figure 6 Boxplot Subject Average Score And The Study Status

**Analysis:**

1. **General Observation:**
   * Across most subjects, students who continued their studies (HDI, blue) generally have higher average scores compared to those who dropped out (THO, orange).
   * The interquartile ranges (IQRs) for HDI students are typically higher, indicating better overall performance and less variability in scores compared to THO students.
2. **Subjects with Significant Differences:**
   * **COM, WEB, SOA, and MOB prefixes:** These subjects show a clear distinction between HDI and THO students, with HDI students having notably higher scores.
   * **ENT and VIE prefixes:** These subjects also display significant differences, suggesting that performance in these areas is strongly associated with student retention.
3. **Subjects with Less Pronounced Differences:**
   * **ITI and COM2 prefixes:** These subjects show less pronounced differences between HDI and THO students, suggesting that performance in these areas may be less indicative of dropout risk.
4. **Outliers:**
   * There are several outliers present in both HDI and THO groups across various subjects, indicating individual students with unusually high or low scores.
   * The presence of outliers, especially in the THO group, may suggest particular cases where students either excelled or struggled significantly in specific subjects before dropping out.

The differences in scores across subjects highlight the importance of including subject-specific performance as features in the predictive model.Subjects with significant differences between HDI and THO students (e.g., COM, WEB, SOA, MOB) should be given higher importance in the model.

## Correlation Heatmap of (Study Status, Attendance Rate, Average Score)

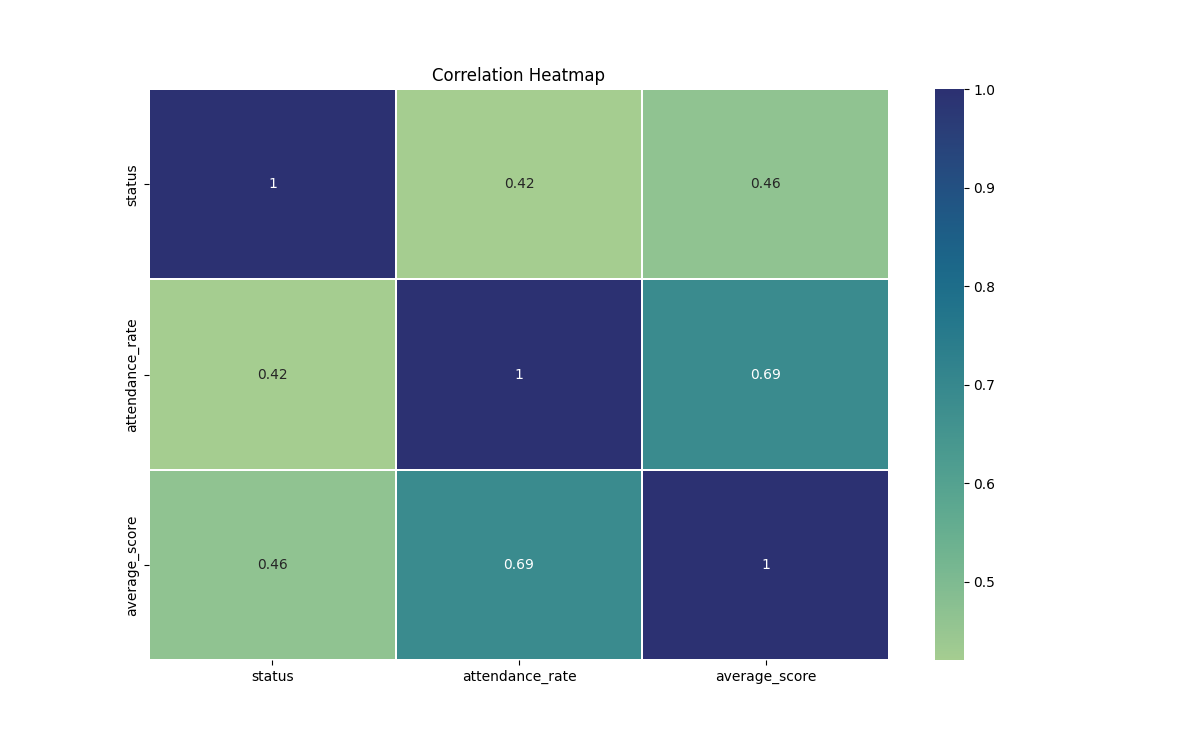


Figure 7 Correlation Heatmap of (Study Status, Attendance Rate, Average Score)

**Analysis:**

1. **Status and Attendance Rate (0.42):**
   * This indicates a moderate positive correlation, suggesting that students who have higher attendance rates are somewhat more likely to continue their studies (HDI). While not a strong predictor on its own, it still provides useful information about student behavior.
2. **Status and Average Score (0.46):**
   * Similarly, a moderate positive correlation between status and average score indicates that students who achieve higher average scores are somewhat more likely to continue their studies. This relationship is not overwhelmingly strong but is significant enough to be a useful predictor.
3. **Attendance Rate and Average Score (0.69):**
   * This stronger correlation shows that higher attendance rates are strongly associated with higher average scores. This relationship underscores the importance of attendance in academic performance.

While the moderate correlations (0.42 and 0.46) suggest that attendance rate and average score are useful but not strongly predictive on their own, the class imbalance likely exacerbates this issue, with only 524 out of 6840 students (approximately 7.7%) dropped out.