Analysis of Student Data Using Clustering and Dimensionality Reduction Techniques

Amit Shchory

Aviad Bloch

Mickey Frankel

1. Data: We have a dataset that includes 4424 records. Each record represents a student and their characteristics, such as grades, parents' education, units per semester, and lastly whether or not they dropped out from university. We got it on the net. It is 521 KB.

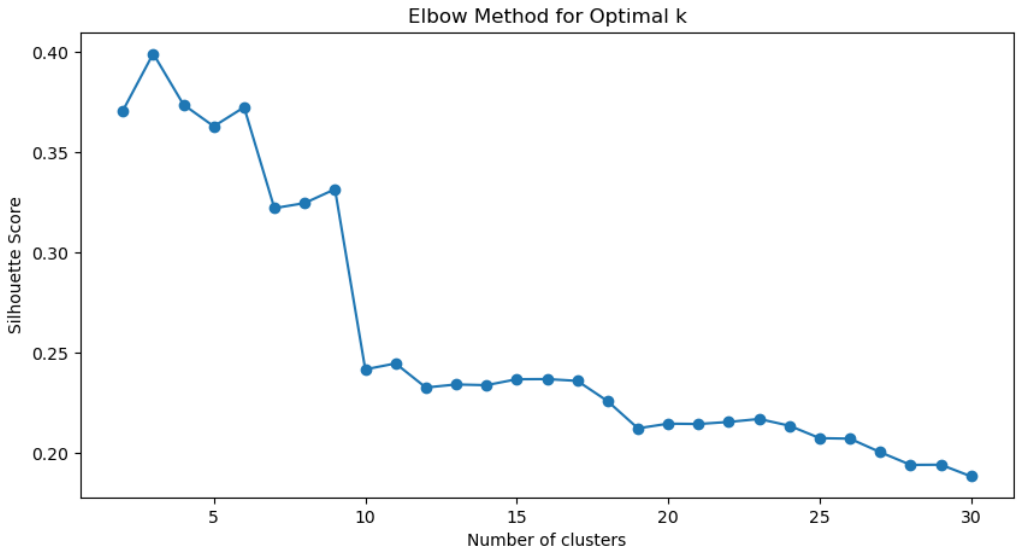
2. Solution: To infer the patterns of the dataset we performed clustering and dimensionality reduction techniques. Our solution consisted of the following key steps: We preprocessed the data and removed the categorical columns because we wanted to focus on the numerical columns. This choice also helped us to ensure compatibility with our chosen analytical methods. Then, we applied standardization to ensure all features contributed equally to the analysis. This helps us dealing with features of varying scales and units. Moreover, we computed a correlation matrix and removed highly correlated features to reduce redundancy in the dataset. This step helps in reducing multicollinearity and can improve the stability of our analyses. In addition, we made a recommendation system for students, which recommends courses and actions that are likely to increase a student's graduation chances, based on their features. The recommendation system has a simple and intuitive UI, making it accessible for all users.

2.2 Dimensionality Reduction: We used the PCA dimensionality reduction technique. PCA is an unsupervised learning technique that transforms the data into a new coordinate system, where the axes (principal components) are ordered by the amount of variance they explain in the data. The PCA technique finds the directions in which the data varies the most. These directions are uncorrelated because they are orthogonal to each other. In addition, the PCA components are ordered by their significance and the percentage of the variance in the data they can represent. The first principal component accounts for the most variance in the data, the second accounts for the second most, and so on. By selecting only the top few principal components, we can reduce the dimensionality of the data while retaining most of the important information. In our analysis, we chose to reduce the data to three principal components, balancing the need for dimensionality reduction with the desire to retain a significant portion of the original variance. This decision allows effective visualization and clustering while still capturing the most important patterns in the data. For each principal component, we identified the top contributing features, providing insights into which student characteristics have the most influence on the variation in the dataset.

2.3 Clustering: To group similar students together into clusters, we used agglomerative clustering, a type of hierarchical clustering algorithm. This algorithm starts with each data point as a single cluster. It then iteratively merges the most similar clusters until a stopping criterion is met. The similarity between clusters is measured using a distance metric and a linkage criterion, which minimizes the variance within clusters. The process creates a dendrogram, a tree-like structure. To determine the optimal number

of clusters, we performed the silhouette score, a measure of how similar an object is to its own cluster compared to other clusters. The silhouette score ranges from -1 to 1, where a higher value indicates better-defined clusters.

This graph shows the results of the Elbow Method for determining the optimal number of clusters (k) in a clustering algorithm, using the Silhouette Score as the evaluation metric.

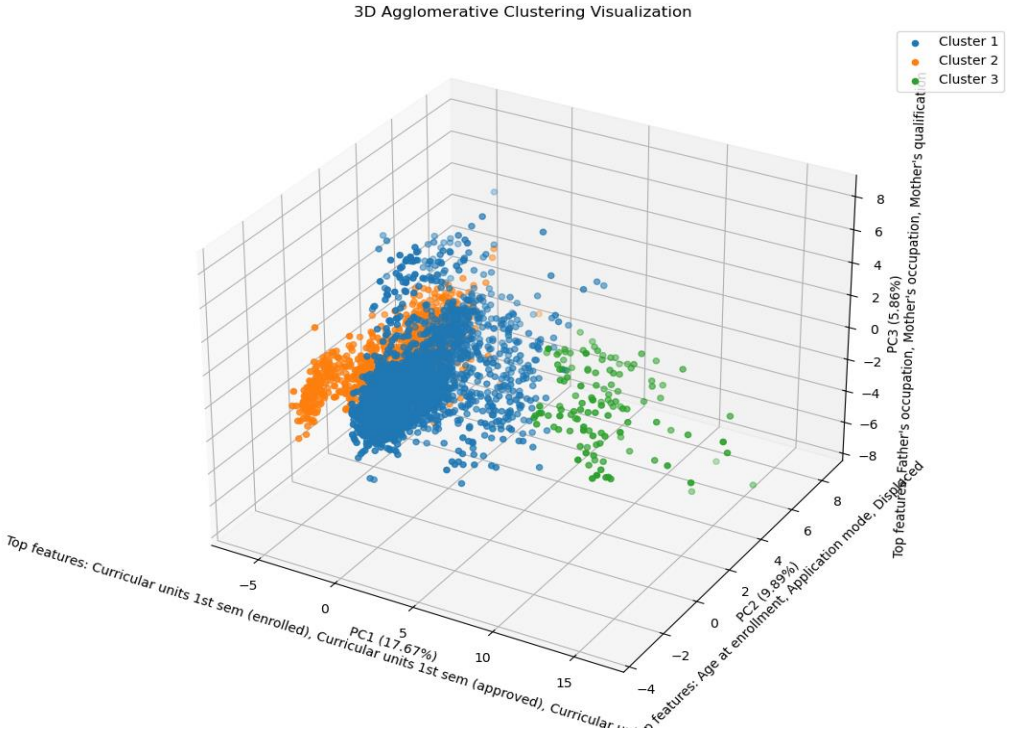


2.4 Statistical Analysis: We conducted a comprehensive statistical analysis to identify the factors influencing student dropout rates, utilizing both traditional statistical methods and regression modeling. These methods provided a robust foundation for interpreting the dataset and drawing actionable insights about which variables most strongly affect student dropout. Our analysis began with **descriptive statistics** for the numerical variables in the dataset. This included calculations for the mean, median, standard deviation, and quartiles, offering a basic understanding of the data's distribution and identifying any potential outliers or anomalies. These descriptive insights helped inform subsequent analyses and guided data cleaning and transformations. Next, we performed a **Chi-Square Test of Independence** to assess the relationship between categorical variables, such as marital status, and dropout rates. The test yielded a chi-square statistic of **53.13** and a p-value of **3.16e-10**, indicating a highly significant association between marital status and dropout rates. This result confirmed that marital status is related to dropout when examined independently, though further analysis was needed to see how this variable interacts with others in a multivariate context. We then employed **logistic regression** to model the probability of student dropout based on several key features, including gender, age at enrollment, and academic performance. The logistic regression model achieved a **Pseudo R-squared** of **0.5191**, suggesting a moderate-to-good fit for the data. The analysis identified statistically significant predictors such as **Gender** (OR: 1.50), **Age at Enrollment** (OR: 1.05), and **Tuition Fees Up to Date** (OR: 0.04), all of which had p-values well below the threshold for significance ($p < 0.05$). In contrast, variables like **Marital Status** and **Nationality** did not show significant effects in the multivariate model. Lastly, we computed **confidence intervals** for the odds ratios to assess the

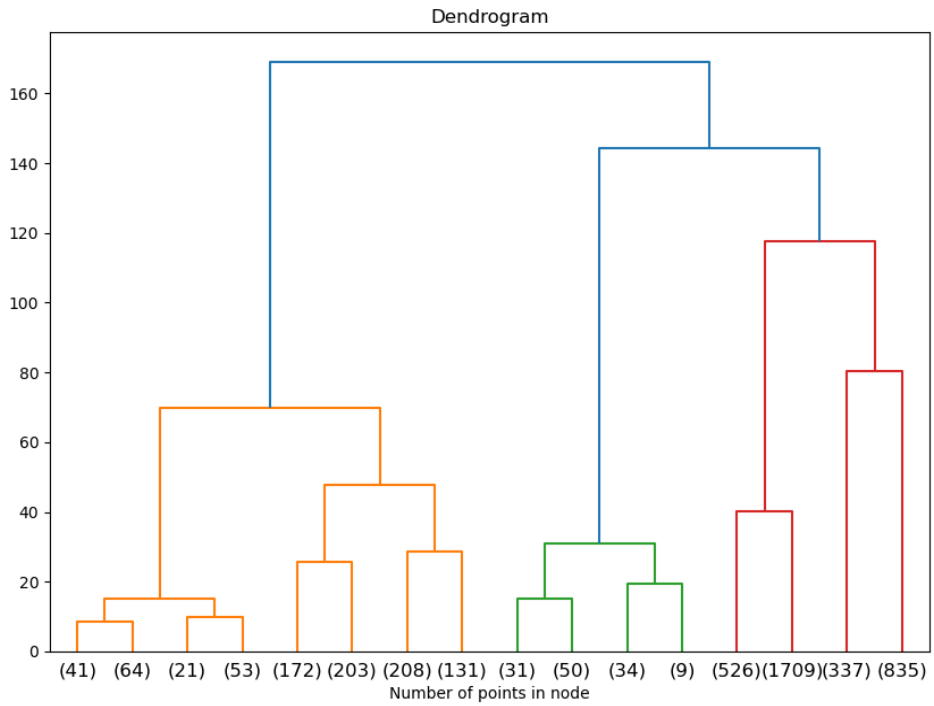
precision of our estimates. Narrower confidence intervals indicated greater precision, particularly for features like **Tuition Fees Up to Date** and **Scholarship Holder**, where the intervals provided clear insight into the strong impact of financial factors on dropout. These statistical methods allowed us to draw clear conclusions about which features most strongly influence student retention, setting the stage for targeted intervention strategies to reduce dropout rates. All this is done in Statistical_Analysis.py.

2.5 Visualization: To make our results interpretable and visually clear, we created two main visualizations:

a) 3D scatter plot: in this visualization, we can see each student represented as a point in a 3D space, with the axes defined by the first three principal components. Each point is colored based on its cluster assignment, allowing us to visually identify how students with similar characteristics are grouped together. For example, we can observe that students in different clusters are generally separated in space, indicating distinct patterns in the data. The blue, orange, and green clusters represent different groups of students based on the characteristics captured by the principal components. In addition, the blue cluster appears to be the largest, suggesting that a majority of students share similar traits. In contrast, the orange and green clusters are smaller but still represent significant portions of the population.
The separation along the axes shows how different features contribute to the clustering. For instance, students in the green cluster may have higher values on certain components, which are related to specific features.



b) Dendrogram: We generated a dendrogram to visualize the hierarchical structure of our clusters. This tree-like diagram shows how clusters are formed and merged at different levels of similarity, providing insight into the relationships between different student groups.



The dendrogram shows how individual data points or small groups of points are merged into larger clusters. The height of each merge represents the distance between clusters-higher merges indicate more distinct clusters.

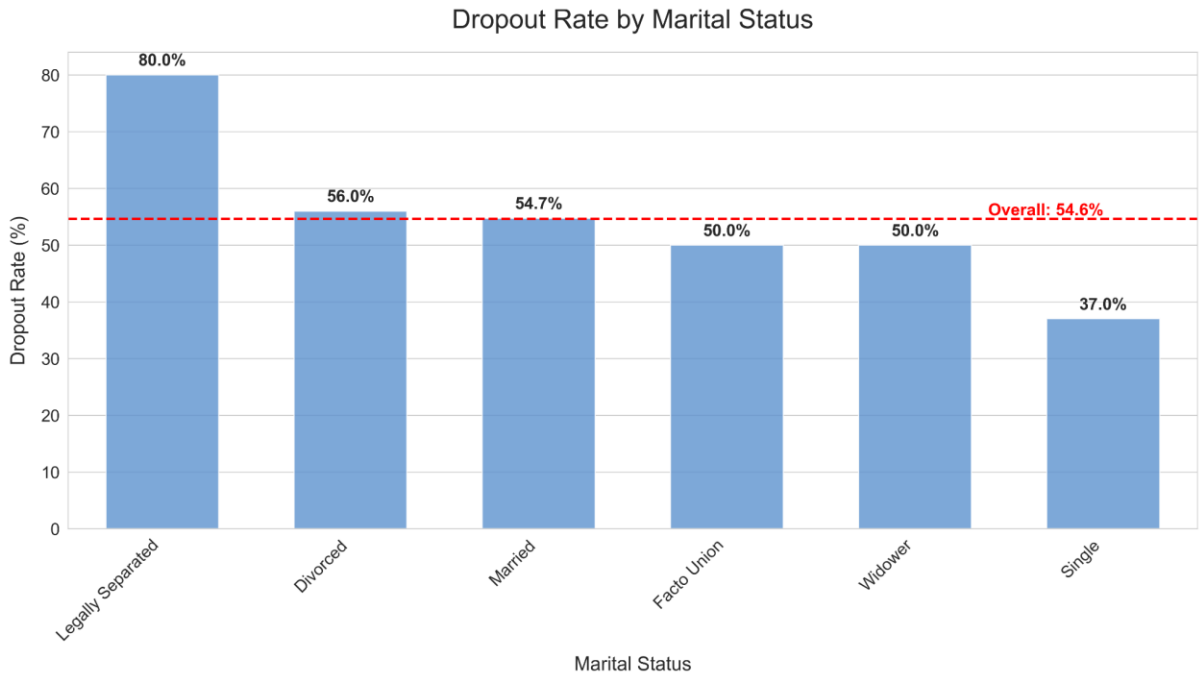
The numbers at the bottom of the dendrogram represent the number of points in each group or cluster at that stage. We can identify a natural cut-off point by looking at the largest vertical distance where clusters merge. This could help determine the number of clusters by "cutting" the dendrogram at a particular height. The large clusters suggest that most students fall into broad categories, while the smaller clusters represent more specialized or distinct profiles. This indicates that the majority of students may share similar patterns, while a few smaller groups stand out with more unique characteristics.

We can infer that the data exhibits a clear hierarchical structure, as shown by the distinct levels in the dendrogram. This suggests that there are subgroups within the data that progressively merge into larger groups, indicating a natural grouping tendency among the students. Moreover, by examining the height at which the clusters merge, we can identify distinct groups in the data. The clusters at lower heights on the dendrogram are more similar to each other, meaning that the data points within these clusters share many common characteristics. These represent closely related student profiles.

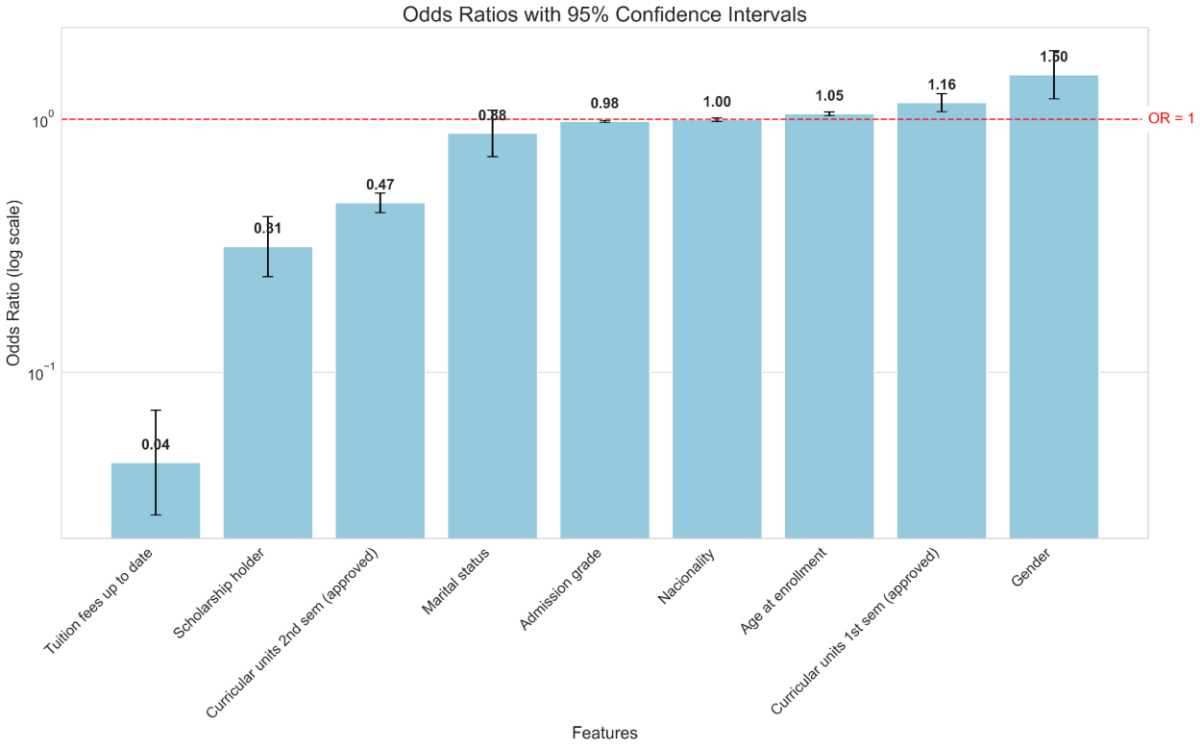
We can observe outliers by the clusters that merge at higher levels may represent more unique or distinct groups of students. These groups could have characteristics that set them apart from the majority of the dataset.

c) Bars-graph: In the **Dropout Rate by Marital Status** bar graph, each marital status category is represented by a bar indicating the dropout rate for that group. The overall dropout rate, chosen due to being the mean value, marked by the red dashed line at 54.6%, serves as a benchmark to compare individual categories. We observe that **Legally Separated** students have the highest dropout rate at **80%**, significantly above the overall average, while **Single** students exhibit the lowest rate at **37%**. This visualization provides a clear comparison

of how different marital statuses are associated with varying dropout rates, highlighting **Legally Separated** and **Divorced** have relatively higher risk for dropping out.



In the **Odds Ratios with 95% Confidence Intervals** graph, we get some twist on our inference; We plot the odds ratios for different features affecting dropout likelihood, with error bars representing the 95% confidence intervals. An odds ratio above 1, such as for **Gender** (1.50), indicates that being male increases the likelihood of dropping out, whereas an odds ratio below 1, such as **Tuition Fees Up to Date** (0.04), suggests that students with up-to-date fees are far less likely to drop out. Thus, we infer that marital status has only a little impact on the odds of students dropping out. As well, we learn that economical features have significant impact on graduating, which can lead to some insights as the recommending preferring male students on getting scholarships. We can interpolate this with the impact of curricular units on semester 2, to condition the receipt of scholarships on some amount of curricular units on that semester.



2.6 Recommendation System: Aiming to increase student graduation chances, we made a recommendation system for students. The system recommends courses and actions that are likely to increase a student's graduation chances, based on their features. Unlike traditional recommendation systems, our system tries to estimate how "beneficial" a certain action is for a student (as in, by how much it is estimated to increase said student's graduation chances), rather than how much the student may "like" it. For the recommendation system, we used collaborative filtering on two utility matrices: one is a utility matrix of actions, and the other is of courses. We built the actions utility matrix as follows: first, we defined the "improvement rate" of a student's action as `new-org` with `org` being the student's original estimated graduation chance and `new` being the student's estimated graduation chance after applying the action (we made the estimations using a simple random forest prediction model). Given this definition, each column of the actions utility matrix represents a possible action of the following: "Prioritize daytime courses", "Prioritize nighttime courses", "Reduce t courses from next semester" for t in [1,2,3] (we assume a course is by average 4 curricular points for this purpose). Each row is for a student, and so in slot i,j we have the estimated improvement rate of student i's graduation chance after applying action j. Similarly, the courses utility matrix follows a similar approach, except there is a column for each available course and slot i,j has the estimated improvement rate of student i's graduation chance if they take course j. Given a new student x, our system makes the recommendation as such: let N be a set of k students most similar to x (we use value 5 for k). Then we make a collaborative filtering recommendation using the approach that we have seen in class:

$$r_{xi} = \frac{\sum_{y \in N} s_{x,y} \cdot r_{yi}}{\sum_{y \in N} s_{xy}} \quad s_{x,y} = \text{sim}(x, y)$$

Meaning, the score an action i gets for student x is a weighted averaged of the k most similar students' score for that action in the utility matrix, weighted by how much similar each

student is to x. Everything said about actions here can also be said about the courses utility matrix – we refer to each course c as the action "take course c".

3. Evaluation: The dataset had clear target labels of "Graduate" and "Dropout", so obtaining ground truth was trivial. The dataset had samples labeled as "Enrolled" which we referred to as "unlabeled" and used for measuring accuracy, as explained below.

We used the silhouette score as our main way to evaluate the quality of our clustering because it gives a clear and easy-to-understand measure, even without knowing the true labels.

From our PCA results, we found that the top three principal components explain 33.5% of the variance in the data. This tells us how much of the original information we've kept after reducing the number of dimensions.

As said, the recommendation system's utility matrices are built using a random forest classifier. The model's performance is measured through accuracy and a classification report generated by `sklearn.classification_report()`. For setup, we split the dataset into train and test sets for training the model to test that it doesn't overfit. For measuring the recommendation system's accuracy, we defined the "recommendation error" of a student sample as following: for each action (or course) recommended by the system, apply action on student and use the base random forest predictor to estimate their new graduation chance. For this recommendation, the error is the absolute difference between the improvement rate estimated by the recommendation system itself and the improvement rate calculated by once again `new-org` (with `new` being the student sample with applied action and `org` being the original estimated graduation chance for that student). We calculate this error for all recommendations of all "unlabeled" samples in the dataset (samples with label "Enrolled"). The recommendation system's accuracy is then defined as one minus the total recommendation error (which is the mean of each such student's recommendation error). In addition, we used a baseline logistic regression model and applied the same evaluations on it, to compare its performance to the one of our final model. All of this is performed in the script "scripts/recsys_eval.py".

To enhance the evaluation process, we added an additional statistical algorithm—**logistic regression**—to complement the existing **random forest** model and **silhouette score** evaluations. Logistic regression was chosen due to its interpretability and suitability for binary classification, which aligns well with the “Graduate” or “Dropout” targets. In this context, logistic regression allows us to examine the direct impact of features like **Gender**, **Admission Grade**, and **Tuition Fees** on dropout probability.

In addition to using the silhouette score to evaluate the quality of clustering, we compared the performance of our logistic regression model with that of the random forest classifier using **accuracy** and a **classification report** generated by `sklearn.classification_report()`. This comparison allows us to assess not just the accuracy of predictions but also key performance metrics such as **precision**, **recall**, and **F1-score**. By applying logistic regression and calculating the odds ratios for significant predictors, we gained valuable insights into which features most strongly influence student outcomes.

4. Future Work: To build upon this initial analysis, several avenues for future work present themselves:

a. Incorporate categorical variables: Develop a method to include categorical data in the analysis.

b. Explore alternative clustering algorithms.

c. Conduct in-depth cluster analysis: Perform a detailed characterization of each cluster to create comprehensive student profiles.

As for the recommendation system, it currently only recommends based on estimated chance of binary labels (graduations). A better system would take into account the final grade, and recommend with the goal of maximizing it. In addition, the current system runs relatively slow; it could use some optimization.

From the statistical analysis perspective,techniques such as **support vector machines (SVM)** or **Naive Bayes** could be explored to improve classification performance. These methods are particularly useful when working with high-dimensional datasets and can offer alternative perspectives on feature importance or separation of classes like **Graduate** and **Dropout**. Moreover, incorporating more sophisticated **regularization techniques** (such as L1 or L2) within the logistic regression framework could help prevent overfitting, especially in scenarios with a large number of features. Similarly, extending the **random forest classifier** with feature selection methods could streamline the model and improve its speed without sacrificing accuracy.

5. Conclusion: In this project, we successfully applied advanced data mining techniques – specifically, agglomerative clustering and principal component analysis – to uncover patterns in a complex dataset of student information. By reducing the dimensionality of the data and identifying optimal clusters, we were able to group students with similar characteristics, potentially revealing insights into factors that influence academic outcomes.

Our statistical analysis further supported this by applying **logistic regression** and **random forest classification** to predict student dropout and assess key factors influencing it. Through the use of odds ratios, we quantified the impact of variables such as **tuition payment status** and **scholarship holding** on dropout risks. The combination of these statistical methods with data mining allowed us to achieve both interpretable insights and robust predictive accuracy.

The visualizations we created, particularly the 3D scatter plot and dendrogram, offer intuitive ways to understand and communicate the complex relationships within the data. These insights have the potential to inform targeted interventions and support strategies, ultimately contributing to improved student outcomes.

Our recommendation system can aid students who seek practical suggestions on how to increase their graduation chances, without needing to know anything about machine learning or data science, using a simple and fairly intuitive UI for recommendations.