# ISL Project Fall 24 **TASK-2**

# Data Description

**Dataset Overview:**  
The dataset comprises **4,424 samples** with **36 features**. These features represent a combination of **categorical, integer, and real values**. They provide comprehensive information about students' demographics, socio-economic backgrounds, and academic details, aiming to predict outcomes such as dropout, continued enrollment, or graduation.

**Feature Categories:**

1. **Demographics:** Includes gender, marital status, and age.
2. **Socio-economic Factors:** Encompasses financial status, debt status, and tuition payment records.
3. **Academic Details:** Covers attendance type (daytime or evening), previous qualifications, and academic achievements.

**Target Variable:**  
The dataset focuses on predicting student outcomes with three possible labels:

* **Dropout**
* **Enrollment Continuation**
* **Graduation**

**Additional Notes:**

* The data is imbalanced across the target categories.
* Preprocessing steps have addressed outliers and anomalies to enhance data reliability.

Data Preprocessing:

**Null Values:**

* **The dataset contained missing values in several columns. Notably:**
  + Target: 5 missing values
  + Application order: 1 missing value
  + Course: 2 missing values
  + Other columns, such as Previous qualification, Debtor, Tuition fees up to date, and Gender, also had missing values.
* **To address missing values:**
  + **Target**: Dropped rows with missing values to avoid biased results during model training, as this column is crucial for the outcome.
  + **Other Features**: Filled numerical columns with the mean because it is a common and effective method for dealing with missing continuous data without introducing bias.

Exploratory data analysis:

**1. Distribution of Target Variable**

The distribution of the 'Target' variable reveals the overall outcomes for students, categorized as dropout, enroll, and graduate. This provides a baseline for understanding how different characteristics correlate with these outcomes.

**2. Impact of Gender on Target Variable**

* Female students outnumber males across all categories: dropout, enroll, and graduate.
* However, male students are more likely to drop out than graduate, indicating a potential gender disparity in academic outcomes.
* Males also show a lower tendency to enroll compared to females.

**3. Impact of Marital Status on Target Variable**

* The majority of students are single, making up a significant portion of the data.
* Married individuals form a smaller group, while divorced students constitute only a negligible proportion.
* These differences suggest that single students dominate the dataset, making it difficult to draw strong conclusions for other marital statuses.

**4. Daytime/Evening Attendance vs. Target**

* A clear preference is observed for daytime attendance over evening classes.
* The success rates (graduation and enrollment) are higher among daytime students, indicating that attending during the day may offer advantages in achieving academic success.

**5. Relationship between Previous Qualification and Target**

* Most students have either secondary education or higher education (bachelor's degree) as their prior qualifications.
* For students with secondary education, the number of graduates is notably high, suggesting a strong foundation for academic success.
* However, among students with higher education qualifications, dropout rates exceed graduation rates, potentially highlighting challenges faced by this group.

**6. Debtor vs. Target**

* Students who are not debtors are more likely to enroll and graduate.
* This indicates that financial stability plays a critical role in ensuring academic success, as debt appears to correlate with higher dropout rates.

**7. Tuition Fees Up to Date vs. Target**

* Students whose tuition fees are not up to date have a higher dropout count, nearly matching their enrollment and graduation numbers.
* This suggests that financial challenges, reflected by delayed tuition payments, significantly impact academic outcomes.

**8. International vs. Target**

* The dataset predominantly consists of local students.
* International students are less likely to drop out within their own group, indicating better retention rates compared to their domestic counterparts.

**Summary of Findings**

These visualizations reveal critical insights into the factors influencing student outcomes. The analysis highlights that gender, financial stability (debts and tuition fee payments), and previous education significantly impact success and dropout rates. Additionally, the distribution patterns emphasize the importance of addressing the unique challenges faced by certain groups, such as male students and those with higher education qualifications, to improve overall academic outcomes.

This report provides a solid foundation for designing interventions tailored to specific student needs, enhancing retention and success rates across different demographic groups.

Model Development: Hyperparameters and Selection:

**Logistic Regression:**

For the **Logistic Regression** model, the main hyperparameter selected was the maximum number of iterations (maxiter=100) for convergence during model fitting. This ensures the model converges within a reasonable time frame. A constant was added to the predictor variables to model the intercept, which is essential for logistic regression models to capture the baseline effects of the predictors.

**K-Nearest Neighbors (KNN):**

For **KNN**, the primary hyperparameter is the number of neighbors (n\_neighbors). We evaluated multiple values for k (from 1 to 19, with a step of 2) to identify the optimal number of neighbors for our model. Based on the accuracy score, the best-performing value of k was 15, with an accuracy of 0.6086.

**Naive Bayes:**

* For **Naive Bayes**, no hyperparameters were specifically tuned, as the default model was used. The accuracy achieved was 0.6674, providing a baseline for comparison with other models.
* Further experimentation with **Decision Trees** and **Random Forest** models will provide additional insights into optimal hyperparameters and model performance.

**Decision Tree:**

For the **Decision Tree** model, the hyperparameter criterion was selected as 'gini', which is the default for splitting nodes. This choice helps minimize the Gini impurity during the model's learning process. Additionally, pruning was applied using the ccp\_alpha hyperparameter, which helps avoid overfitting by controlling the size of the tree. The optimal ccp\_alpha was found to be 0.0011, which resulted in a test accuracy of 0.7591. The tree had 47 leaves, ensuring an optimal balance between accuracy and complexity.

**Random Forest:**

For **Random Forest**, the primary hyperparameters used were max\_features='sqrt' and n\_estimators=100. The choice of max\_features='sqrt' allows the model to evaluate only a subset of features at each split, reducing overfitting and improving generalization. The n\_estimators value of 100 was chosen to ensure enough trees for a robust ensemble while balancing computational efficiency. The Random Forest model achieved a test accuracy of 0.7591, which is consistent with the decision tree's performance, showing the power of ensemble methods. Feature importance was also visualized, helping identify key predictors.

Performance Evaluation: Model Results

**Logistic Regression:**

* + Accuracy: 75.91%
  + Precision (Class 0): 0.80, Recall (Class 0): 0.77
  + Precision (Class 1): 0.51, Recall (Class 1): 0.31
  + Precision (Class 2): 0.78, Recall (Class 2): 0.92
  + **Confusion Matrix: [[229 27 41]**

**[ 42 50 67]**

**[ 14 22 392]]**

**K-Nearest Neighbors (K=15):**

* + Accuracy: 60.86%
  + Precision (Class 0): 0.60, Recall (Class 0): 0.50
  + Precision (Class 1): 0.31, Recall (Class 1): 0.35
  + Precision (Class 2): 0.65, Recall (Class 2): 0.70

**Naive Bayes:**

* + Accuracy: 66.74%
  + Precision (Class 0): 0.75, Recall (Class 0): 0.68
  + Precision (Class 1): 0.35, Recall (Class 1): 0.19
  + Precision (Class 2): 0.68, Recall (Class 2): 0.83
  + **Confusion Matrix: [[203 18 76]**

**[ 36 30 93]**

**[ 33 38 357]]**

**Decision Tree:**

* + Accuracy: 75.91%
  + Mean Squared Error (MSE): 0.5249
  + Best Pruned Tree: 47 leaves
  + Test Accuracy after pruning: 75.91%

**Random Forest:**

* + Accuracy: 75.91%
  + Strong ensemble performance was similar to the decision tree.

Interpretation: why a model is performing better/worse?

**Logistic Regression**: Performs well due to the ability to handle multi-class classification. It likely performs better for students with higher attendance and prior qualifications. Features such as Previous Qualification, Tuition Fees, and Age influence the likelihood of success, explaining why the model classifies students more accurately in certain categories.

**KNN**: The performance is relatively weaker, especially for lower values of k. It struggles with distinguishing students in borderline categories, particularly in the dropout class. The model is sensitive to the distance metric, which impacts performance, especially in sparse data.

**Naive Bayes**: Performs moderately well by assuming feature independence, making it suitable for simpler relationships. Key features like Age and Prior Academic Performance play significant roles in predicting success or failure, but it’s less effective when features are correlated.

**Decision Tree**: The pruning process helps by focusing on critical splits, improving accuracy. Key features such as Tuition Fees and Attendance contribute to the classification, but the model tends to overfit in certain cases, as indicated by the drop in performance with pruning.

**Random Forest**: This shows a similar performance to the decision tree but is more robust due to the ensemble approach. It effectively handles the complex interactions between features like Age, Prior Academic Performance, and Tuition Fees. Feature importance analysis reveals that these variables contribute most to the predictions.

**General Analysis:**

* **Feature Importance**: Features such as Previous Qualification, Age, Tuition Fees, and Attendance are key in classifying students. Students with a solid academic background and regular attendance are more likely to succeed.
* **Performance Differences**: Logistic regression and random forest achieve better performance due to their ability to handle complex relationships and overfitting (in the case of random forest). Models like Naive Bayes and KNN show weaker performance, especially with misclassifications in critical categories like dropout.
* **Cluster Classification**: Models classify students into clusters based on their academic history, socioeconomic status, and engagement (e.g., attendance). Students in the 0 category (likely high-performing) show clear distinctions, while those in the 1 category (possibly dropouts) tend to have weaker performance across all models.

Conclusion:

This individual project aimed to predict student outcomes such as dropout, continued enrollment, and graduation using a dataset containing demographic, socio-economic, and academic factors. Various machine learning models, including Logistic Regression, K-Nearest Neighbors, Naive Bayes, Decision Tree, and Random Forest, were implemented.

While Logistic Regression, Decision Tree, and Random Forest performed similarly with a test accuracy of around 75.91%, KNN, and Naive Bayes showed comparatively lower performance. Key factors like previous qualifications, tuition fees, and attendance were crucial in influencing predictions. This analysis highlights the importance of addressing financial and academic engagement factors to improve student retention and success.

References:

<https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success>

<https://scikit-learn.org/dev/modules/generated/sklearn.linear_model.LogisticRegression.html>

<https://scikit-learn.org/1.5/modules/naive_bayes.html>

<https://graphviz.org/>