# Enhancing Academic Trajectories: A Machine Learning Framework for Optimized Student Placement

Yashpreet Malhotra

yashmalhotra9323@gmail.com

**Abstract.** In the context of increasing enrollments and concerns over student retention in higher education, this study introduces a machine learning framework designed to optimize student placement in academic programs. Addressing the challenges posed by the surge in student numbers and the complexities of matching student profiles to suitable programs, the proposed methodology leverages data analytics to predict student success and mitigate dropout rates. The framework facilitates the creation of student profiles and employs machine learning techniques to align incoming students with optimal academic paths, with the goal of fostering a more effective and personalized educational environment.

**Keywords:** Student retention · Machine learning · Academic placement · Higher education · Student profiling · Dropout prediction · Educational data mining

## 1 Introduction and Motivation

The landscape of higher education is rapidly evolving, driven by increased global enrollment and the complex challenges of student retention. According to UNESCO, global tertiary enrollment rose from 28.5 million in 1970 to an estimated 235 million by 2023 [1]. For instance, in France alone, 2.97 million students registered for higher education in the 2021–2022 academic year, marking a 2.5% increase from the previous year [2].

However, this growth has not been accompanied by improved success rates. Data from the 2010–2011 Bachelor's cohort in France revealed that only 39.8% of students graduated within four years [3]. This mismatch between student potential and academic program suitability is a key concern and has prompted systemic reforms, such as the introduction of the Parcoursup platform [4]. Despite efforts to improve transparency and fairness, critics argue that such platforms often fail to capture the nuanced profiles of students, thereby limiting their effectiveness.

Success in higher education is often narrowly defined by graduation rates [5]. Yet, this binary view neglects qualitative dimensions such as student engagement, intellectual development, and alignment between personal interests and academic paths. In this study, we define an "excellent student" as one who not

only excels academically but also demonstrates genuine interest and conceptual understanding in their chosen field.

This research aims to develop a machine learning framework that enhances the student admission process by predicting student success and aligning student profiles with suitable academic programs. The methodology leverages prior data to create three distinct student clusters—Bad, Average, and Excellent—and uses these profiles to predict the most suitable placement for incoming students.

While the experiments in this paper are grounded in the French academic system, the proposed framework is designed to be adaptable across international education systems. Cultural and institutional differences are acknowledged, but universal factors like academic background, motivation, and social integration form the foundation of this study. Importantly, the objective is not to create an elitist system but to reduce academic mismatches and foster equitable opportunities for all students.

## 2 State of the Art

The optimization of student placement using machine learning (ML) has attracted significant research attention. Numerous studies focus on predicting student dropout as a proxy for academic risk, while others attempt to forecast academic success by analyzing historical and behavioral data. This section reviews existing literature under two major lenses: analytical approaches (focused on human and sociological factors) and predictive approaches (centered on algorithmic modeling).

### 2.1 Predicting Student Dropout

Student dropout has been studied extensively across sociological, psychological, and computational domains. Research shows that dropout is influenced by factors such as family background, social integration, institutional commitment, and satisfaction with the chosen program [6,**?**,**?**]. Analytical studies suggest that dropout is not merely an academic failure but a social phenomenon often tied to student support networks and mental well-being [7,**?**].

Advanced machine learning techniques such as Decision Trees (DT), Random Forests (RF), k-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Neural Networks (NN) have been widely applied to model dropout prediction with accuracies often exceeding 80% [8,**?**,**?**]. However, most of these models focus on identifying at-risk students after enrollment, offering limited utility during the admission phase.

### 2.2 Predicting Student Success: A Human-Centric Challenge

Defining and predicting academic success is inherently complex. While institutions often equate success with graduation and grades, students may define it

through engagement, understanding, and satisfaction [5]. This divergence necessitates a holistic modeling approach that incorporates socio-psychological factors like motivation, peer relationships, and adaptability.

Multiple studies emphasize that success cannot be universally defined; it varies across time, culture, and institutional contexts. Consequently, predictive models must be flexible enough to adapt to localized definitions of excellence while ensuring fairness and transparency [9].

### 2.3 Analytical Approaches

Analytical literature underscores the significance of pre-admission indicators such as prior academic performance, institutional fit, and family support [10,**?**,**?**]. These variables have been incorporated into persistence models that help identify students likely to succeed or struggle. Particularly, composite models such as Tinto's and Bean's theories of student departure have inspired multifactor frameworks for dropout analysis [11,**?**].

### 2.4 Predictive Approaches Using Machine Learning

Studies employing ML for student modeling have shown promising results. Decision Trees and Random Forests offer interpretability, while Neural Networks and SVMs provide high predictive accuracy for complex, nonlinear data. The Synthetic Minority Over-sampling Technique (SMOTE) has been applied to mitigate class imbalance, especially when identifying rare but valuable categories like "excellent" students [12].

Some frameworks even utilize Natural Language Processing (NLP) to analyze qualitative inputs such as motivation letters, enabling a richer understanding of student intent and capacity [13]. Table 1 shows the frequency of ML models used across reviewed studies.

**Table 1.** Common Machine Learning Models Used in Literature

| **Technique** | **Frequency** |
|---|---|
| Decision Tree (DT) | 49 |
| Neural Networks (NN) | 29 |
| Logistic Regression | 25 |
| k-Nearest Neighbors (KNN) | 9 |

These models each have strengths and trade-offs. For example, DTs offer explainability, while NNs perform well with high-dimensional data but lack transparency. The best model for deployment must be chosen based on institutional context and available data.

# 3 Framework and Methodology

This section outlines the architecture of the proposed machine learning (ML) framework designed to optimize student placement during the admission process. Our approach integrates both supervised and unsupervised ML techniques to classify students into three categories: **Excellent**, **Average**, and **At Risk**. The workflow is composed of data preprocessing, clustering, outlier detection, and prediction layers, each powered by distinct algorithms suited for its task.

## 3.1 Algorithm Selection

Based on our literature review, the following algorithms were selected for their complementary strengths:

- **Neural Networks (NN)**: Used to process unstructured data such as motivation letters and convert it into numerical embeddings.
- **Principal Component Analysis (PCA)**: Applied for dimensionality reduction and improved efficiency.
- **K-Means Clustering**: Groups students based on feature similarity into predefined clusters (Excellent, Average, At Risk).
- **Isolation Forest (IF)**: Identifies anomalous profiles that deviate from the majority—typically outliers such as exceptionally strong or weak candidates.
- **Lasso Regression (LR)**: Performs feature selection and highlights variables that most influence classification.

## 3.2 Workflow Architecture

The data pipeline is structured into modular blocks, as shown in Figure 1, enabling transparency and flexibility.

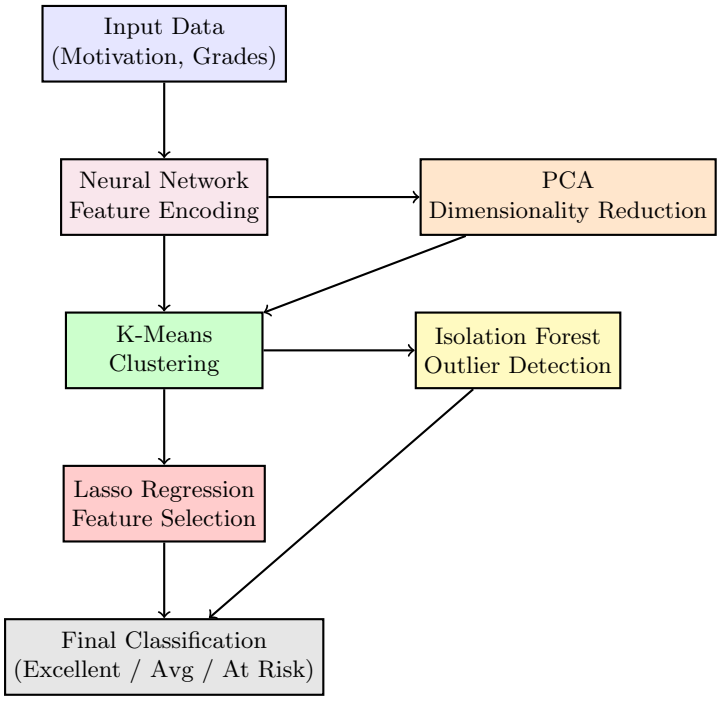


**Fig. 1.** Algorithmic workflow for student success prediction.

### 3.3 Input Data Types

The system is designed to ingest three categories of data:

- **Motivation Letters:** Parsed with Natural Language Processing (NLP) and converted to vectorized embeddings using NN.
- **Academic Performance:** Includes prior grades, degree levels, and education system types.
- **Program Metadata:** Institutional data describing formation difficulty, competitiveness, and learning model.

### 3.4 Cluster Assignment and Labeling

Each student is assigned to a cluster through K-Means based on similarity to historical profiles. Isolation Forest identifies extreme deviations, and Lasso Regression interprets the importance of features, helping in both validation and refinement of clusters.

### 3.5 Modularity and Reusability

The framework is intentionally modular. Depending on institutional goals or regional constraints, models can be re-trained with local data. Moreover, the

system allows toggling of sensitive inputs (e.g., socio-demographic details) to remain compliant with ethical guidelines.

## 4 Conceptual Proposal and Validation

To translate the proposed machine learning framework into a deployable system, we conceptualize a three-part pipeline that integrates with existing university registration infrastructures. This section outlines the data ingestion requirements, modular design, and validation approach, as well as ethical considerations for real-world application.

### 4.1 System Overview

The system is designed to operate at the intersection of institutional registration data, program metadata, and applicant information. The goal is to output a recommendation indicating the applicant's suitability for a program based on past data and clustering logic.

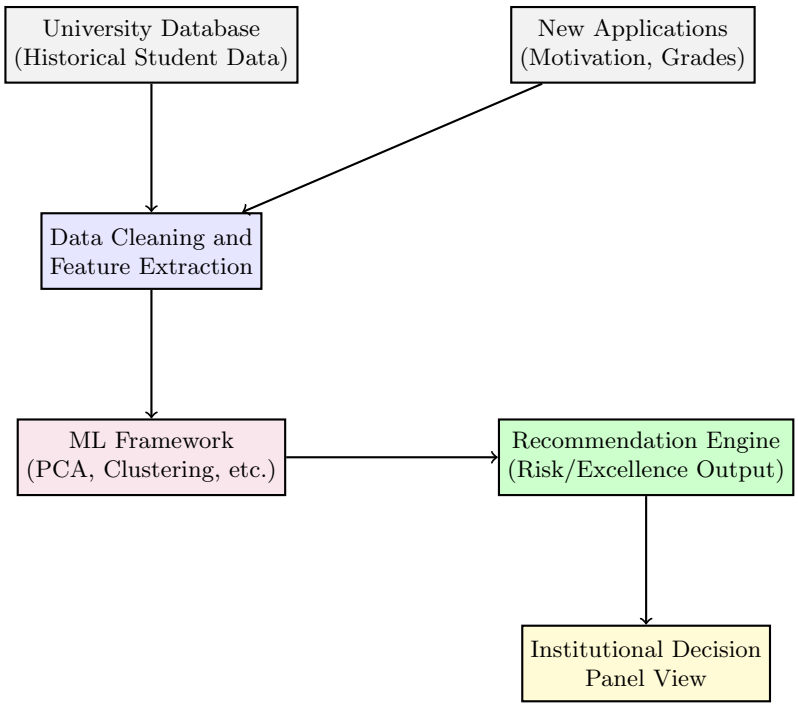


**Fig. 2.** Conceptual deployment of the student placement framework.

### 4.2 Feeding the Model

The system requires three types of data inputs for optimal performance:

- **Historical Student Profiles:** Past students' academic performance and outcomes for supervised learning.
- **Application Materials:** Including motivation letters and transcripts from incoming students.
- **Program-Level Information:** Metadata such as dropout rates, GPA thresholds, and pedagogical style.

### 4.3 Validation Strategy

To assess the reliability of the framework, a two-phase validation process is proposed:

1. **Retrospective Validation:** Apply the model to historical datasets where student outcomes are known. This phase verifies if the framework can classify students accurately into their eventual performance groups.
2. **Prospective Validation:** Use the model during ongoing admissions and monitor how classified "excellent" students perform over time.

Performance metrics such as accuracy, precision, recall, and ROC-AUC scores will be calculated and monitored continuously to refine model thresholds.

### 4.4 Ethical and Operational Considerations

To ensure fair deployment:

- No personal identifiers (name, nationality, gender) are used during model training or inference.
- Admissions committees must not be exposed to feature weights or decision rationales to prevent bias.
- Recommendations are advisory, not mandatory. Human discretion remains in the loop.

### 4.5 Scalability and Generalization

Although developed with the French university system in mind, this framework can be generalized to any academic institution by training on localized data and refining definitions of success. Institutions may customize success labels and program difficulty indicators to match their strategic priorities.

## 5 Implementation and Experimental Results

This section presents the practical implementation of the proposed framework using real-world data, evaluates its clustering capabilities, and discusses early findings from the experiment. The experimental setup was developed using RapidMiner to simulate data ingestion, processing, clustering, and classification workflows.

### 5.1 Dataset Description

The dataset used for the experiment was obtained from the Université de Pau et des Pays de l'Adour (UPPA), comprising student application data for the SIGLIS Master's program between 2019 and 2024. The dataset includes:

- Academic history (grades, diplomas)
- Admission year
- Age and year of birth
- Final mention (e.g., Très Bien, Bien, Passable)

The dataset contains 90 entries, which were cleaned and anonymized before being processed.

### 5.2 Experimental Setup

Due to limited data availability, the first implementation focused on clustering students into three groups: **Excellent**, **Average**, and **At Risk**. Motivation letter analysis (NN component) was omitted in this pilot due to the lack of text input. The working pipeline is shown in Figure 3.

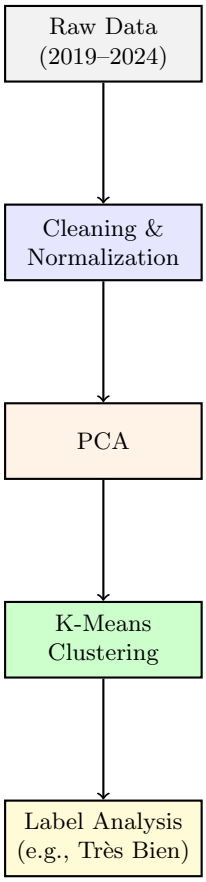


**Fig. 3.** Simplified pipeline implemented in the experiment.

### 5.3 Clustering Results

Students were clustered into three groups using K-Means. Clusters were interpreted based on their relationship to academic mentions:

- **Cluster 0:** Mostly average performance, with some missing mention data.
- **Cluster 1:** Small group with no mention and possibly incomplete records.
- **Cluster 2:** Dominated by students with a *Très Bien* (very good) mention—classified as "Excellent".

Although the dataset is small (only 90 entries), the model showed clear separation in mention quality across clusters.

### 5.4 Quantitative Summary

**Table 2.** Number of Students per Academic Year

| Academic Year | Number of Students |
|---|---|
| 2018–2019 | 14 |
| 2019–2020 | 13 |
| 2020–2021 | 13 |
| 2021–2022 | 17 |
| 2022–2023 | 32 |

The year 2022–2023 had the highest proportion of students with excellent mentions, reinforcing the cluster validity.

### 5.5 Limitations

The following limitations were encountered during this phase:

- Small dataset size hindered generalization.
- No access to textual inputs (e.g., motivation letters) limited testing of NLP components.
- Lack of final performance outcomes (e.g., GPA) made success classification dependent solely on mentions.

### 5.6 Early Conclusions

Despite limitations, this pilot confirms that historical student data can be clustered meaningfully. The clustering component successfully identified profiles resembling high-performance students. With richer datasets (e.g., including motivation letters, interviews, or assessments), the full pipeline could be tested and validated for deployment.

## 6 Conclusion and Future Work

This research proposed and implemented a machine learning framework for optimizing student placement in higher education. Through a multi-layered approach incorporating clustering, dimensionality reduction, and predictive modeling, the system aims to assist institutions in identifying students most suited to specific academic programs—thus improving both success rates and institutional efficiency.

### 6.1 Summary of Contributions

The key contributions of this study include:

- A literature-informed design of a modular ML framework integrating PCA, K-Means, Isolation Forest, and Lasso Regression for student classification.
- A conceptual deployment model for institutional use, preserving fairness and transparency.
- A preliminary implementation using real data from UPPA, which demonstrated promising clustering behavior based on student mentions.

The framework is built to be adaptable across institutions and cultural contexts, with flexibility to redefine "success" based on localized educational goals. Importantly, ethical safeguards are embedded in the design to avoid misuse or algorithmic bias.

### 6.2 Addressing Research Gaps

While the literature on dropout prediction is vast, few studies address proactive placement at the time of application. This study fills that gap by shifting the focus from detecting failure to enabling success from the outset. Furthermore, the inclusion of socio-academic indicators and modular interpretability enhances the practical utility of the system.

### 6.3 Limitations

Despite its promise, the study is subject to several limitations:

- The dataset was relatively small (90 students) and specific to one master's program.
- Motivation letter processing was not tested due to the absence of textual data.
- Success was defined purely based on mention or final grade, which may oversimplify performance.

These limitations restrict the generalizability of the experimental findings but do not undermine the conceptual value of the framework.

### 6.4 Future Directions

Several paths for improvement and expansion are proposed:

1. **Expanded Dataset:** Apply the framework to larger and more diverse datasets, including undergraduate programs, other universities, and international student pools.
2. **NLP Integration:** Incorporate neural network models to analyze motivation letters, personal statements, or interview transcripts for more nuanced profiling.

3. **Longitudinal Tracking:** Implement a follow-up system to track students over time and refine predictive labels using real performance data.
4. **Policy Integration:** Collaborate with education ministries or accreditation bodies to align the framework with institutional regulations and national goals.

### 6.5 Concluding Remarks

As higher education systems grow in scale and complexity, personalized and data-driven tools will become increasingly necessary to match students with academic pathways that maximize their potential. The framework presented in this work offers a structured, ethical, and modular solution that institutions can tailor to their specific contexts. With further validation and development, this approach holds the potential to reshape the admissions landscape and support more equitable, efficient, and successful academic journeys.

## References

1. UNESCO, "Higher education — articles," 2023, accessed: 2024-06-12. [Online]. Available: https://www.unesco.org/en/higher-education
2. Ministère de l'Enseignement Supérieur et de la Recherche, "Les effectifs d'étudiants dans le supérieur continuent leur progression en 2021-2022," 2022, accessed: 2024-06-12. [Online]. Available: https://www.enseignementsup-recherche.gouv.fr/fr/les-effectifs-d-etudiants-dans-le-superieur-continuent-leur-progression-en-2021-2022-88609
3. I. Kabla-Langlois, "Les jeunes et le système éducatif en france," *INSEE Éclairage*, 2014.
4. M.-P. Couto, F. Bugeja-Bloch, and L. Frouillou, "Parcoursup : les prémices d'un accroissement de la stratification sociale et scolaire des formations du supérieur," *Agora débats/jeunesses*, vol. 89, no. 3, pp. 23–38, 2021. [Online]. Available: https://www.cairn.info/revue-agora-debats-jeunesses-2021-3-page-23.htm
5. M. Weatherton and E. E. Schussler, "Success for all? a call to re-examine how student success is defined in higher education," *CBE—Life Sciences Education*, vol. 20, no. 1, p. es3, 2021. [Online]. Available: https://www.lifescied.org/doi/full/10.1187/cbe.20-09-0223
6. W. G. Spady, "Dropouts from higher education: An interdisciplinary review and synthesis," *Interchange*, vol. 1, no. 1, pp. 64–85, 1970.
7. E. Durkheim, *Suicide: A Study in Sociology*. Free Press, 1951.
8. D. Opazo, S. Moreno, E. Álvarez Miranda, and J. Pereira, "Analysis of first-year university student dropout through machine learning models: A comparison between universities," *Sustainability*, vol. 9, no. 20, p. 2599, 2022.
9. G. D. Kuh, J. Kinzie, and J. A. Buckley, "What matters to student success: A review of the literature," 2006.
10. D. Rowntree, "Teaching and learning online: A correspondence education perspective," 1995.
11. V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Review of Educational Research*, vol. 45, no. 1, pp. 89–125, 1993.

12. E. R. Sinchi and G. P. G. Ceballos, "Acceso y deserción en las universidades. alternativas de financiamiento," *Alteridad*, vol. 13, no. 2, pp. 274–287, 2018. [Online]. Available: https://revistas.ups.edu.ec/index.php/alteridad/article/view/2.2018.10
13. J. Caspersen, "Teachers' learning activities in the workplace: How does teacher education matter?" *Creative Education*, vol. 6, no. 1, pp. 46–63, 2014.