COMP-381-AB1  
 Ismail El Sayad  
 Sahijdeep Waraich  
 Adakole Jumbo-Ochigbo  
 Gurveer Chahal  
 June 20, 2025

**Predicting University Admission Status**

**Introduction**

In today’s competitive academic environment, university admissions have become increasingly selective, with institutions relying on various academic and extracurricular factors to make informed decisions. Predicting whether a student will be accepted, rejected, or waitlisted is a real-world challenge faced by both applicants and admissions officers. This project aims to use machine learning to model and predict admission outcomes based on student profile data.

We selected this topic because of its relevance to students and educational institutions. Understanding the patterns that influence admission decisions can help students better prepare their applications and assist universities in streamlining their review processes. The use of data-driven models offers an objective way to identify which features—such as GPA, SAT scores, and extracurricular involvement—have the greatest impact on admission status.

To carry out our analysis, we utilized a dataset originally sourced from Kaggle, which we enhanced by engineering additional features such as GPA category, activity level, SAT percentile, and academic score. These new features were designed to better capture the complexity of student profiles and improve model performance.

Our project applies three supervised machine learning algorithms—Logistic Regression, Linear Discriminant Analysis (LDA), and K-Nearest Neighbors (KNN)—to classify admission status into one of three categories: Accepted, Rejected, or Waitlisted. The effectiveness of each model was evaluated using key metrics such as accuracy, precision, recall, F1 score, and confusion matrices.

In the sections that follow, we describe the dataset and preprocessing steps, explain our model selection and evaluation criteria, present and compare the results, and conclude with observations on model performance and future improvements.

**Dataset Description**

The dataset used in this project was originally obtained from Kaggle (Kumar, 2023) and was enhanced to improve prediction accuracy. It consists of student application records and includes both academic and non-academic features that influence university admission decisions. The dataset contains a total of 75 entries and the following key features:

* **GPA**: The student's Grade Point Average on a 4.0 scale.
* **SAT\_Score**: The student’s total SAT score, ranging from 400 to 1600.
* **Extracurricular\_Activities**: A score from 0 to 5 that reflects the student’s involvement in extracurricular activities.
* **GPA\_Category**: A categorical feature derived from GPA, classifying it as Low, Medium, or High.
* **Activity\_Level**: A categorical representation of extracurricular involvement, labeled as Low, Moderate, or High.
* **SAT\_Percentile**: A derived score indicating the student's SAT score as a percentile relative to the dataset.
* **Academic\_Score**: A calculated metric combining GPA and SAT score to represent overall academic strength. This was computed using a formula:  
  Academic\_Score = (GPA × SAT\_Score) / 2
* **Admission\_Label**: A numerical encoding of the target variable, where 0 = Rejected, 1 = Accepted, and 2 = Waitlisted.
* **Admission\_Status**: The actual target class label corresponding to the outcome of the admission decision.

Before training the models, we performed light preprocessing including feature scaling where necessary and encoding categorical features where applicable. The additional engineered features (e.g., SAT Percentile and Academic Score) were introduced to provide the models with more nuanced insights and improve classification performance.

Overall, the dataset was structured to support multi-class classification and allowed us to explore how different academic and non-academic indicators contribute to university admission outcomes.

**Algorithms Used and Evaluation Metrics**

To predict admission outcomes, we implemented three supervised machine learning algorithms introduced during the course: **Logistic Regression**, **Linear Discriminant Analysis (LDA)**, and **K-Nearest Neighbors (KNN)**. These algorithms were selected for their interpretability, effectiveness in classification tasks, and compatibility with multi-class problems.

Each model was trained using the same dataset and evaluated using a consistent set of performance metrics. The following evaluation metrics were used to assess and compare model performance:

* **Confusion Matrix**: A summary of prediction results on the test set, showing the number of correct and incorrect predictions broken down by class. This helps us understand how well the model performs across all categories—Accepted, Rejected, and Waitlisted.
* **Accuracy**: The overall percentage of correctly predicted cases out of the total predictions made.
* **Precision**: The proportion of correct positive predictions for each class. It measures how reliable the model’s positive classifications are.
* **Recall**: The proportion of actual positives correctly identified. It assesses how well the model captures each class.
* **F1 Score**: The harmonic mean of precision and recall. This single metric balances both concerns, especially when classes are imbalanced.

By applying these metrics across all three models, we were able to compare their ability to handle multi-class classification effectively. The confusion matrix provided a detailed view of specific areas of strength or confusion, while the other metrics offered a high-level summary of each model’s reliability, sensitivity, and overall performance.

**Results and Comparison**

For Linear Discriminant Analysis (LDA), the accuracy was 98.7%, so almost all predictions matched the actual admission status. The precision was 98.7%, so prediction of a class was right 98.7% of the time. The recall was 98.7%, so the model correctly identified 98.7% of the real students in each class. The F1 score was 98.7%, so the model performed well in both predicting and identifying students. The confusion matrix showed that LDA correctly predicted 21 Accepted students, 21 Rejected students, and 32 Waitlisted students, with only one Waitlisted student misclassified.

For Logistic Regression, the accuracy was 97.3%, slightly lower than LDA. The precision was 97.5%, so the model was right 97.5% of the time when making predictions. The recall was 97.3%, it correctly identified 97.3% of the real students in each class. The F1 score was 97.3%, showing strong overall performance. The confusion matrix indicated that Logistic Regression correctly predicted 21 Accepted students, 21 Rejected students, and 31 Waitlisted students, with two Waitlisted students misclassified.

For K-Nearest Neighbors (KNN), the accuracy was 96.0%, slightly below the other models. The precision was 96.2%, so predictions were correct 96.2% of the time. The recall was 96.0%, so it correctly identified 96% of the actual students in each class. The F1 score was 96.0%, showing good but slightly lower overall performance compared to the other two models. The confusion matrix showed that KNN correctly predicted 20 Accepted students, 21 Rejected students, and 31 Waitlisted students, with 1 Accepted student misclassified and 2 Waitlist students.

**Summary of Comparison**

The accuracy was high for all of the models, ranging from 96.0% to 98.7%, which shows that the features were effective in predicting admission status. Linear Discriminant Analysis performed the best, followed closely by Logistic Regression, with K-Nearest Neighbors performing slightly worse but still very well. The evaluation metrics indicate that all models were able to distinguish between Accepted, Rejected, and Waitlisted students with only some confusion, mainly in the Waitlisted category.

A graph of different colored bars

AI-generated content may be incorrect.

**Conclusion**

This project demonstrated how machine learning techniques can be applied to predict university admission outcomes using real-world student profile data. By exploring Logistic Regression, Linear Discriminant Analysis (LDA), and K-Nearest Neighbors (KNN), we were able to build multi-class classification models that predicted whether a student would be accepted, rejected, or waitlisted based on academic and extracurricular features.

Among the three models, LDA achieved the highest performance across all evaluation metrics, including accuracy, precision, recall, and F1 score. Logistic Regression also performed strongly, closely trailing LDA. KNN showed slightly lower but still respectable results. These outcomes suggest that the features we selected and engineered—such as SAT percentile and academic score—played a meaningful role in improving the model's ability to distinguish between the different admission categories.

While the results are promising, this project has some limitations. The dataset was relatively small, which could limit the generalizability of the models. Additionally, class imbalance (especially in the Waitlisted category) posed challenges for perfect classification. Future improvements could include collecting a larger and more diverse dataset, experimenting with other algorithms such as decision trees or ensemble methods, and tuning hyperparameters more extensively.

Overall, this project illustrates the potential of machine learning in educational decision-making and highlights how data science can be used to uncover patterns in complex, real-world problems.

**References**

* Kumar, A. (2023, September 6). Student admission dataset. Kaggle. <https://www.kaggle.com/datasets/amanace/student-admission-dataset>
* Matplotlib. (2023). Bar chart with multiple series. https://matplotlib.org/stable/gallery/lines\_bars\_and\_markers/bar\_stacked.html