Predictive Modeling of Student Dropout and Academic Achievement Using Machine Learning Algorithms

Argo Learners - Ruchith Chippari, Mathari Pavani
Department of Mathematics and Statistics
University of West Florida
December 2024

ABSTRACT

The problems of student dropout and academic underachievement are critical issues in education that have major consequences for both individual success and institutional performance. This study seeks to investigate predictive modeling for student performance using machine learning algorithms, predicting dropout rates and academic achievement based on historical data. LR and KNN were chosen because of their complementary strengths: interpretability and probabilistic insights in LR, and flexibility in capturing complex patterns in KNN.

The dataset had features like grades, attendance, parental education, and socio-economic status, preprocessed using feature scaling and encoding. The models were evaluated on accuracy, precision, recall, and F1-score. Logistic Regression had an accuracy of 75.25%, performing the best for dropout and graduation outcome predictions, while KNN obtained an accuracy of 70.17% with high recall for graduates but did not perform well for enrolled students.

The results reflect the trade-offs involved in balancing interpretability with predictive accuracy. Logistic Regression provided practical insights into the methods of intervention, while KNN provided a more granular classification that came with a loss of interpretability. This study has brought into focus the importance of selected features and well-distributed datasets in predicting student performance for the development of an early warning system to support students at risk.

1 Introduction

Student dropout and underachievement are serious problems in educational systems worldwide, with high social and economic costs. Students who do not complete their education often face limited career opportunities, while institutions suffer in terms of reputation and finances. The early identification of at-risk students can allow educators and administrators to take proactive steps, thus establishing support mechanisms that help improve outcomes and reduce dropout rates.

This project discusses the use of machine learning for predicting student performance and student dropout. Based on an analysis of historical data, the study identifies key influencers on student success and works toward models that can give timely warnings for at-risk students. These insights will make informed targeted interventions possible, which can be academic counseling, peer mentoring, or financial support, so that vulnerabilities can be caught and treated at specific levels.

This project incorporates two machine learning algorithms: **Logistic Regression (LR)** and **K-Nearest Neighbors (KNN)**. In that line, the reason for choosing Logistic Regression will have to do with its interpretative power, which allows stakeholders to understand the relationship between different variables like grades, attendance, parental education, and their effects on student performance. The K-Nearest Neighbor was selected because it is versatile in detecting patterns within a multidimensional dataset; hence, this model is efficient in handling complex data structures.

This project evaluates and compares these models based on accuracy, precision, recall, and F1-score. The paper has emphasized the trade-offs between interpretability for Logistic Regression and flexibility with KNN; this therefore brings insights into the suitability of the two algorithms for predictive work in education. Finally, the goal of this research is to highlight the role of machine learning as an effective tool to advance student success and institutional decision-making.

2 Methodology

This project's methodology is based on the development, training, and evaluation of predictive models that will forecast student dropout rates and academic achievement using machine learning algorithms. The project will make use of two models: Logistic Regression and K-Nearest Neighbors. Further, in detail, the steps are given, together with the main formula used in each model and how it was applied to predict student outcomes.

1. Data Collection and Preprocessing

1.1. Data Collection

The dataset used in this project contains historical student records that include various academic, demographic, and socio-economic features. These features include:

- **Academic Performance**: Grades, attendance, and course choices.
- **Demographic Information**: Parental education level, marital status, and application mode.
- **Target Variable:** Student outcome classified into three categories: Dropout, Enrolled, and Graduate.

The dataset consists of 4424 student records with a mix of categorical and continuous variables.

1.2. Data Preprocessing

Before applying machine learning algorithms, several preprocessing steps were undertaken to ensure the data was clean and suitable for modeling:

- **Handling Missing Data:** Missing values were imputed using the mean or median for numerical features and mode for categorical variables.
- **Encoding Categorical Variables:** Categorical features (e.g., marital status, application mode) were transformed using one-hot encoding. For instance, marital status with two categories (Single, Married) is represented as two binary features: [1, 0] for Single and [0, 1] for Married.
- **Feature Scaling:** Continuous features (e.g., grades, attendance) were scaled using standardization (z-score normalization) to ensure all features contribute equally to the model's performance. The formula used for scaling is:

$$z = \frac{X - \mu}{\sigma}$$ Where,

X is the feature value,
μ is the mean, and
σ is the standard deviation

- **Train-Test Split:** The dataset was split into 80% training and 20% testing subsets to assess the model's ability to generalize to unseen data.

2. Model Selection

In this project, two algorithms from machine learning were selected: Logistic Regression (LR) and K-Nearest Neighbors (KNN). Both models have been chosen for their relative merits. First, LR provides interpretability, and second, KNN is a flexible, non-parametric method capable of handling complex relationships.

2.1. Logistic Regression (LR)

Logistic Regression is the most common linear model used for classification. It is widely used for binary and multi-class classification problems, hence suitable for this task since the target variable has three classes: Dropout, Enrolled, Graduate.

Logistic Function: The probability of each class is modeled using the logistic function (sigmoid):

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}}$$ where,

$P(y = 1 \mid X)$: is the probability of the target variable
$X_1, X_2, \ldots, X_n$: are the input features
$\beta_0, \beta_1, \ldots, \beta_n$: are the learned model coefficients.

For multi-class classification, the One-vs-Rest approach was used, which calculates the probability for each class independently and assigns the class with the highest probability.

2.2. K-Nearest Neighbors (KNN)

K-Nearest Neighbors is an instance-based learning algorithm that makes predictions based on the K nearest training samples. It is non-parametric, meaning it does not assume any underlying distribution for the data, which makes it well-suited for complex datasets.

Distance Metric: KNN relies on some distance metric to determine the neighbors closest to a test point. Most common is Euclidean distance:

$$d(x_i, x_j) = \sum_{k=1}^{n} (x_{ik} - x_{jk})^2 \qquad \text{where,}$$

$d(x_i, x_j)$: is the Euclidean distance between two data points x_i and x_j
x_{ik} and x_{jk} : are the feature values of the k^{th} feature of points x_i and x_j
n : is the number of features

Classification: After calculating the distances from all other data points, the algorithm picks the K nearest neighbors. The class label for the test point is decided by majority vote:

$$y_{prediction} = mode(y_1, y_2, \ldots, y_K) \qquad \text{where,}$$

$y_1, y_2, \ldots, y_K$: are the class labels of the K nearest neighbors.

2.3. Hyperparameter Tuning

Hyper parameters were optimized for both models to result in better performance:

- Logistic Regression: The strength of regularization was tuned using cross-validation.
- KNN: Best value for K (number of nearest neighbors) was selected using grid search with cross-validation.

3. Model Training and Evaluation

3.1. Training the Models

Logistic Regression: The model was trained using maximum likelihood estimation to minimize the log-loss function:

$$\text{Log-loss} = -\sum_{i=1}^{N} \sum_{c=1}^{C} y_{ic} \log(p_{ic}) \qquad \text{where,}$$

y_{ic} : is the true label for sample i and class c
p_{ic} : is the predicted probability for sample i and class c

K-Nearest Neighbors: KNN does not require an explicit training phase. It simply stores the training data and classifies new points based on the majority class of their K nearest neighbors.

3.2. Evaluation Metrics

The performance of the models was evaluated using several metrics:

Accuracy: Proportion of correct predictions:

$$Accuracy = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

Precision: Proportion of true positives among predicted positives:

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k}$$

Recall: Proportion of true positives among actual positives:

$$\text{Recall}_k = \frac{TP_k + FN_k}{TP_k}$$

F1-Score: Harmonic mean of precision and recall:

$$F_1\text{-Score}_k = 2 \times \frac{\text{Precision}_k \times \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}$$

The models were evaluated based on these metrics using the test set.

4. Results and Model Comparison

4.1. Logistic Regression Results

- Accuracy: 75.25%
- Precision (Dropout): 0.83, Recall (Dropout): 0.77
- Precision (Enrolled): 0.45, Recall (Enrolled): 0.28
- Precision (Graduate): 0.76, Recall (Graduate): 0.91

Logistic Regression performed well in predicting dropout and graduate categories, although it struggled with predicting enrolled students.

4.2. K-Nearest Neighbors Results

- Accuracy: 70.17%
- Precision (Dropout): 0.81, Recall (Dropout): 0.68
- Precision (Enrolled): 0.37, Recall (Enrolled): 0.28
- Precision (Graduate): 0.72, Recall (Graduate): 0.87

KNN performed reasonably well with graduation predictions, but had lower accuracy overall and struggled with predicting enrolled students.

5. Conclusion

This project focused on the use of machine learning to predict the outcome of students, whether Dropout, Enrolled, or Graduate, based on academic, demographic, and socio-economic

data. By analyzing these patterns, our goal was to provide actionable insights for educators to intervene early and support students more effectively.

What We Found

The results indicated that, on average, Logistic Regression was better, yielding an accuracy of 75.25%, compared to KNN, which yielded an accuracy of 70.17%. Logistic Regression gave very good results for students who would probably Graduate (91% recall) or Dropout (77% recall), though it performed very poorly on the Enrolled group, with only 28% recall. The model also gave insight into important predictors such as grades, parental qualifications, and application mode, which is especially useful when trying to identify risk factors.

While KNN is flexible and powerful in capturing complicated patterns, it was sensitive to imbalanced data and feature scaling. It resulted in slightly lower recalls for Graduate, 87%, and Dropout, 68%, classes, and had the same low recall of 28% for the Enrolled category. In any case, KNN can be useful in some tasks when the structure of the datasets is different, or the interpretability of the model is not as crucial.

Key Challenges

1. **Class Imbalance:** The smaller number of students in the Enrolled category made it difficult for both models to predict this group accurately.
2. **Data Quality:** Missing data had to be imputed, which may have introduced biases into the results.
3. **Model Trade-offs:** While logistic regression presented simplicity and interpretability, it assumed linearity, perhaps compromising its capability to catch complex patterns. KNN was computationally intensive and more sensitive to noisy or irrelevant features.

Implications of the Findings

These findings raise the imperative to leverage data in the early identification of at-risk students. Example:

- Logistic Regression's probabilistic predictions can help schools identify which factors, such as grades or parental education, are most strongly associated with dropout risks and target interventions accordingly, whether through tutoring or mentorship programs.
- KNN may be helpful in detecting non-linear relationships in datasets with less structured patterns, although it requires more preparation and tuning.

Model Outputs in Context

- Logistic Regression Results: This model has correctly predicted Dropout 77% of the time and Graduate 91% of the time; hence, it has an advantage on accuracy with 75.25% and reliability.

- KNN Results: With somewhat less overall accuracy (70.17%), the best balance of precision versus recall was found with the KNN model; however, smaller classes, like the one for Enrolled, present only 28% in recall.

Looking Ahead

To further enhance these models, enabling them to also :

- **Improved Data Collection:** More detailed student data could be added, such as psychological profiles or engagement metrics, which would enhance the predictions of the model.
- **Advanced Algorithms:** Ensemble methods, such as Random Forests or XGBoost, along with handling class imbalance using oversampling techniques, may be explored to refine the performance.
- **Real-Time Applications:** Deploying these models into live systems would enable schools to monitor students dynamically and intervene as issues arise.

Final Thoughts

This project thus shows how machine learning has the ability to make education more inclusive and responsive. Though Logistic Regression was the better performer in this data set, both models contributed something useful. Continuing these approaches and taking care of their limitations may help in reducing dropout rates and assist more students towards success with the help of machine learning. This project marks a further step ahead in combining data with education for bringing about meaningful change.

6. References

1. L. M. Doss P and M. Gunasekaran, "Evasion and Poison attacks on Logistic Regression-based Machine Learning Classification Model," 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2023, pp. 1-8, doi: 10.1109/ICONSTEM56934.2023.10142395. https://ieeexplore.ieee.org/document/10142395

2. S. Sun and R. Huang, "An adaptive k-nearest neighbor algorithm," 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, Yantai, China, 2010, pp. 91-94, doi: 10.1109/FSKD.2010.5569740. https://ieeexplore.ieee.org/document/5569740

3. M. A. Dewi, F. I. Kurniadi, D. F. Murad, S. G. Rabiha and A. Romli, "Machine Learning Algorithms for Early Predicting Dropout Student Online Learning," 2023 IEEE 9th International Conference on Computing, Engineering and Design (ICCED), Kuala Lumpur, Malaysia, 2023, pp. 1-4, doi: 10.1109/ICCED60214.2023.10425359. https://ieeexplore.ieee.org/document/10425359

4. G. P. V, S. Eliyas and S. K. M, "Detecting and Predicting Learner's Dropout Using KNN Algorithm," 2024 OPJU International Technology Conference (OTCON) on Smart

Computing for Innovation and Advancement in Industry 4.0, Raigarh, India, 2024, pp. 1-6, doi: 10.1109/OTCON60325.2024.10688123. https://ieeexplore.ieee.org/document/10688123

5. S. N, R. C, S. S and S. S, "Utilizing Machine Learning to Forecast a Student's Performance," 2024 2nd International Conference on Artificial Intelligence and Machine Learning Applications Theme: Healthcare and Internet of Things (AIMLA), Namakkal, India, 2024, pp. 1-5, doi: 10.1109/AIMLA59606.2024.10531452. https://ieeexplore.ieee.org/document/10531452

I. Appendix

Links:

Data Set: Predict Students Dropout and Academic Success
Code Link: Google Colab Link

Outputs:

```
Logistic Regression Accuracy:
0.752542372881356
              precision    recall  f1-score   support

     Dropout       0.83      0.77      0.80       316
    Enrolled       0.45      0.28      0.34       151
    Graduate       0.76      0.91      0.83       418

    accuracy                           0.75       885
   macro avg       0.68      0.65      0.66       885
weighted avg       0.73      0.75      0.74       885


K-Nearest Neighbors Accuracy:
0.7016949152542373
              precision    recall  f1-score   support

     Dropout       0.81      0.68      0.74       316
    Enrolled       0.37      0.28      0.32       151
    Graduate       0.72      0.87      0.79       418

    accuracy                           0.70       885
   macro avg       0.63      0.61      0.62       885
weighted avg       0.69      0.70      0.69       885
```