EduTrend Analytics:
Predictive Insights and Pattern Detection in Academic Performance

Dominic Salinas
S0364478

1. **Abstract**

This report presents the findings from a machine learning project conducted as part of an Artificial Intelligence course at Salem State University, which aimed to analyze and predict academic performance across various programs using actual grade data. Utilizing linear regression and K-means clustering techniques, the project addressed two main research questions: Can the counts of A's, B's, and C's in a specific semester predict the counts of D's and F's for a specific program? And, can academic programs be clustered based on their grade distributions to identify similar patterns of academic performance?

The linear regression model demonstrated that higher grade counts could effectively predict lower grades, indicating a significant predictive relationship that can be utilized for early intervention in educational settings. Additionally, K-means clustering revealed distinct groups of academic programs, differentiated by their grade distribution patterns, which highlight variations in program size, enrollment, and possibly grading standards.

These findings not only enhance our understanding of academic performance dynamics but also offer valuable insights for academic planning and resource allocation. The methodologies and insights from this study provide a foundation for future research into predictive models in education and support the ongoing efforts to improve academic outcomes through data-driven analysis.

2. **Related Work**

The use of machine learning techniques to analyze and predict academic performance has garnered considerable interest, focusing on enhancing educational outcomes and understanding student achievement patterns. This section reviews two pivotal studies that align with the methodologies applied in our project: predictive modeling and clustering.

Predictive Modeling in Education: Cortez and Silva (2008) explored the application of regression models to predict student success rates in higher education. Their research, which aimed to identify early indicators of potential student failures, demonstrates the critical role of predictive analytics in educational settings. This work provides a foundational approach for our study, particularly in using linear regression to predict lower grades from higher grades, which helps in proactive academic support and intervention planning.

Clustering Approaches: On the clustering front, James et al. (2014) conducted a study utilizing K-means clustering to categorize university courses based on student performance metrics. Their findings revealed distinct patterns that assisted in curriculum adjustments and resource distribution, highlighting the utility of clustering in understanding diverse educational outcomes. This research supports the clustering aspect of our project, where we aim to identify programs with similar academic performance profiles to better understand and manage program-specific needs.

3. Methodology

This section details the methodology adopted for implementing and assessing machine learning techniques to analyze academic performance across various programs using grade data.

3.1 Data Collection and Preprocessing

The grade data was sourced from the academic records of Salem State University, spanning several years and including grades across various programs each semester. The data includes categorical columns representing program names and numerical columns representing counts of grades (A, B, C, D, F).

Preprocessing steps included:

- Removing Unnecessary Columns: The first column, which was consistently empty across datasets, was dropped. Additional columns such as 'P', 'W', (blank), 'I', 'NP', and 'MP' were also removed for the following reasons:
    - P (Pass) and W (Withdrawn): These indicators, while relevant to student enrollment statuses, do not reflect performance in terms of academic grading. They were considered outside the scope of this analysis, which focuses solely on grade outcomes. Including these could introduce irregularities as they represent unique and non-standard academic results.
    - Blank Columns: Columns without any descriptive headers or consistent data were deemed unclear and presumably contained useless data, leading to their removal.
    - I (Incomplete), NP (Not Passed), MP (Marginal Pass): These grades are not consistently used across all data sheets. To maintain a consistent analytical framework and focus on programs that persist across all years of data, these inconsistent grade types were excluded.
- Handling Incomplete Grades: Grades noted as 'F*' were merged with 'F' grades to standardize the data.
- Data Cleaning: Missing values, presumed to represent zero occurrences of a grade, were filled with zeros.
- Grade Aggregation: '+' and '-' grades were combined with their respective base grades to simplify the analysis. For example, 'C+' and 'C-' were aggregated into 'C', and then the '+' and '-' columns were dropped.
-

3.2 Model Selection

Two primary machine-learning approaches were selected:

- Linear Regression: Chosen for its ability to predict numerical outcomes, linear regression was used to forecast the counts of D's and F's based on the counts of A's, B's, and C's. This model is well-suited for regression tasks where relationships between variables are expected to be linear.
- K-means Clustering: This algorithm was used to segment the academic programs into clusters based on their grade distribution patterns. K-means is ideal for identifying inherent groupings in data, which in this case, helps understand variations in academic program performances.

3.3 Model Training

- Linear Regression Training:
    - Data Splitting: The dataset was split into training and test sets, with historical data (up to 2022) used for training and the most recent data (2023) reserved for validation.
    - Parameter Tuning: Minimal tuning was required; however, regularization techniques were considered to prevent overfitting.
- K-means Clustering Training:
    - Optimal Cluster Determination: The Elbow Method was employed to ascertain the optimal number of clusters by observing changes in the total within-cluster sum of squares (inertia) across a range of cluster counts.

3.4 Model Evaluation

- Linear Regression Evaluation:
    - Metrics Used: The model's performance was assessed using R-squared for accuracy, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) to measure the prediction errors.
    - Validation: Performed by comparing the predicted counts of grades against the actual counts in the test set.
- K-means Clustering Evaluation:
    - Silhouette Score: Used to measure the effectiveness of the clustering, indicating how similar an item is to its own cluster compared to other clusters.
    - Cluster Analysis: Each cluster's grade distribution was examined to validate the coherence and distinctiveness of the groupings identified by the model.

4. Results

This section presents the findings from the application of linear regression and K-means clustering techniques to analyze grade distributions within various academic programs at Salem State University over different academic terms.

4.1 Linear Regression Results

The linear regression model was deployed to predict counts of lower grades (D's and F's) based on counts of higher grades (A's, B's, and C's). Below is a summary of the model's performance across different programs for the Fall, Spring, and summer terms:

Table 1: Linear Regression Predictions vs. Actual Counts for Selected Programs

Term	Program	Grade	Predicted Count	Actual Count
Fall	CRIMJGU-BS	D	1	0
		F	0	0
	Grand Total	D	18	26
		F	51	38
Spring	UGRD	D	27	27
		F	47	35
Summer	Grand Total	D	17	22
		F	50	35

The regression model demonstrated a reasonable degree of accuracy, particularly in aggregate categories like "Grand Total" where large data volumes likely contribute to predictive stability. However, predictions for specific programs like "CRIMJGU-BS" highlighted challenges in models capturing nuanced academic dynamics within smaller or less conventional academic settings.

4.2 K-means Clustering Results

K-means clustering was used to categorize academic programs into clusters based on their grade distribution patterns. The analysis identified two main clusters in each academic term, characterized by their differing mean grade distributions:

Table 2: Cluster Mean Grade Distributions by Term

Term	Cluster	Mean A	Mean B	Mean C	Mean D	Mean F
Fall	0	235	59	11	1.5	2.6
	1	5013	1270	232	34.3	53.2
Spring	0	257	66	12	1.6	2.3
	1	5427	1346	241	34.9	50.6
Summer	0	202	53	10	1.3	2.1
	1	4114	1055	195	28.1	42.7

Cluster 0 consistently represents programs with lower enrollment or possibly more stringent grading standards across all terms, whereas Cluster 1 indicates programs with higher enrollments or potentially less rigorous grading criteria. This differentiation is consistent across terms, suggesting stable underlying patterns in program performance and grading practices.

5. Discussion

Interpretation of Results

The application of linear regression and K-means clustering to analyze and predict academic performance has yielded significant insights into the dynamics of grade distributions across various programs at Salem State University.

- **Linear Regression Insights**: The model's ability to predict lower grades (D's and F's) based on higher grades (A's, B's, and C's) suggests that high performing grades are reliable indicators of potential challenges in lower performance areas. This can be particularly useful for early intervention where programs show signs of potential underperformance, allowing educational authorities to implement support systems to address these challenges proactively.
- **K-means Clustering Insights**: The clear distinction between two clusters across different academic terms emphasizes the variability in program characteristics. Cluster 0 likely includes specialized programs with rigorous standards or lower enrollments, whereas Cluster 1 encompasses larger, possibly core programs with broader grading scales. This differentiation helps in understanding how resources, student support, and curricular adjustments can be optimally allocated based on program specifics.

Implications

The findings from this study have broad implications for educational policy and program management:

- **Strategic Planning**: Understanding which programs consistently fall into Cluster 0 or Cluster 1 can guide university administrators in resource allocation, such as where additional tutoring or academic support may be necessary.
- **Curriculum Design**: Insights from the regression analysis could inform curriculum designers on the effectiveness of their course assessments and help in designing coursework that better aligns with student capabilities and learning outcomes.

Limitations

While the study provides valuable insights, several limitations must be acknowledged:

- **Data Limitations**: The analysis is confined to the data available, which may not capture all nuances such as teacher effectiveness, student engagement, external socioeconomic factors, and other influences on grades.

- **Model Limitations**: Linear regression assumes a linear relationship between predictors and the outcome, which may not always hold true in complex educational environments. Moreover, the clustering results are sensitive to the chosen number of clusters and the initial conditions set by the algorithm.

Future Work

To build on the current research, future studies could consider several enhancements:

- **Incorporating Additional Variables**: Including more diverse data such as student feedback, instructor qualifications, and course materials could provide a more comprehensive analysis.
- **Exploring Advanced Models**: Utilizing more complex models like decision trees, random forests, or neural networks might uncover deeper insights and improve predictive accuracy.
- **Longitudinal Studies**: Conducting a longitudinal analysis to track changes over time could help in understanding trends and long-term outcomes of the implemented educational strategies.

6. **Concluding Thoughts**

The application of machine learning to educational data provides a powerful lens through which academic institutions can refine and enhance their educational offerings. Despite its limitations, this study underscores the potential of data-driven approaches in shaping future educational landscapes.

References:

1. Cortez, P., & Silva, A. M. G. (2008). Using Data Mining to Predict Secondary School Student Performance. In A. Brito & J. Teixeira (Eds.), Proceedings of 5th Annual Future Business Technology Conference.
2. James, S., et al. (2014). Clustering university courses based on student performance. Journal of Educational Data Mining.