Feature Importance Analysis for Predicting University Students' GPA

Shaikh Abdul Rafay
k21 3051

Rayyan Ahmed
k21 3079

Minal Alam
k21 3072

December 13, 2023

Abstract

This project investigates the multifaceted factors influencing students' academic performance, employing a comprehensive data-driven approach. Data was collected through a Google Form distributed to participants, capturing insights into their university experience, lifestyle, and family background. The study focuses on preprocessing raw data, conducting exploratory data analysis (EDA), and utilizing machine learning models for predictive analysis. Challenges such as imbalanced data classes were addressed through an alternative approach. The models chosen for this analysis include RandomForestClassifier, DecisionTreeClassifier, KNeighborsClassifier, AdaBoostClassifier, GradientBoostingClassifier, XGBClassifier, SVC, Perceptron, VotingClassifier and Bagging Classifier. The report outlines the methodology, key findings, and insights gained from the evaluation of these models.

1 Introduction

The academic performance of students is a crucial aspect of higher education institutions, and understanding the factors influencing it can inform educational strategies and interventions. This project aims to explore the intricate interplay of various elements, ranging from attendance and study habits to family support and extracurricular activities, in determining students' Grade Point Averages (GPA). The data for this study was collected through a Google Form survey, enabling the acquisition of insights from participants across different universities and degree programs.

2 Literature Review

The existing body of literature highlights several crucial factors that have been identified as influencing student academic performance. A comprehensive understanding of these factors is essential for contextualizing the current study.

Academic performance determinants have been extensively explored in prior research, with a focus on attendance, study habits, extracurricular activities, and family support (Tinto, 1997; Pascarella Terenzini, 1980). Studies consistently emphasize the multifaceted nature of these determinants, suggesting that no single factor operates in isolation.

Attendance emerges as a significant predictor of academic success, with higher attendance rates associated with improved grades and overall academic achievement (Credé et al., 2010). Moreover, effective study habits and time management are critical contributors to academic success, impacting students' ability to allocate dedicated time for study and employ effective learning strategies (Hartwig Dunlosky, 2012).

Participation in extracurricular activities has been linked to enhanced academic performance, fostering skills that positively impact overall achievement (Fredricks Eccles, 2006). Additionally, family support, encompassing emotional and financial assistance, plays a pivotal role in academic success, and socioeconomic factors, such as household size and parental education, are correlated with academic achievement (Davis-Kean, 2005; Sirin, 2005).

Adequate sleep is increasingly recognized for its impact on cognitive functioning and academic performance (Walker, 2017). Research suggests a positive correlation between sufficient sleep and higher GPA. Engaging in physical activities, such as regular workouts, has also been associated with improved cognitive function and overall well-being, potentially influencing academic outcomes (Hillman et al., 2008).

Membership in university societies is explored as a potential factor, with research suggesting that involvement in such groups can contribute to a sense of community and positively impact academic engagement (Astin, 1993). Finally, parental education levels are recognized as influential in shaping students' academic aspirations and achievements, with higher parental edu-

cation often correlated with higher academic attainment in their children (Kuncel et al., 2005).

The utilization of GPA as an outcome measure is common in academic research, providing a quantitative representation of a student's overall academic performance. This metric serves as a reliable measure for assessing the effectiveness of various interventions (Hattie Timperley, 2007).

3 Methodology

3.1 Data Collection

The success of any data-driven project hinges on the quality and relevance of the collected data. In this section, we outline the meticulous process of data collection undertaken for this project, emphasizing the transparency and reliability of the dataset.

3.1.1 Survey Instrument

To gather information on factors influencing student academic performance, a Google Form survey was designed and distributed to participants. The survey comprised a thoughtful selection of attributes, including:

University Name: Identifying the institution the participant is affiliated with.

Degree Name: Capturing the academic program the participant is enrolled in.

Average Attendance: Quantifying the regularity of attendance, a crucial factor in academic success.

Average Study Time per Day: Exploring the time allocated to study activities on a daily basis.

Extracurricular Activities: Investigating participation in activities beyond the academic curriculum.

Average Sleep Time: Understanding the role of sleep patterns in academic performance.

Household Size: Exploring the potential influence of family structure.

Workout: Investigating the impact of physical activity on academic outcomes.

Free Time Activity: Capturing the nature of activities during leisure time.

University Society Membership: Identifying whether the participant is a member of any university society.

Parents' Highest Qualification: Gauging the educational background of the participant's parents.

GPA: The primary outcome variable, representing the academic performance of the participant.

3.1.2 Participant Recruitment

Participants were recruited from diverse academic backgrounds and universities to ensure a representative dataset. The recruitment strategy involved distributing the survey link through university channels, student forums, and social media platforms. Clear instructions were provided, emphasizing the voluntary nature of participation and the importance of honest responses.

3.1.3 Ethical Considerations

This data collection adhered to ethical standards, ensuring participant anonymity and confidentiality. Informed consent was obtained at the beginning of the survey, outlining the purpose of the study and the usage of the collected data. Participants were assured that their responses would be used solely for research purposes and would remain confidential.

3.1.4 Data Completeness and Quality Control

To enhance the reliability of the dataset, measures were implemented to ensure completeness and quality. Mandatory fields in the survey minimized missing data, and ranges and constraints were applied to certain fields to prevent outliers. Regular checks were conducted to identify and rectify any anomalies or inconsistencies in the collected data.

3.1.5 Conclusion

The data collection phase of this project reflects a comprehensive and systematic approach to gathering information on factors influencing student academic performance. The diverse and representative dataset obtained through the Google Form survey lays the foundation for meaningful analysis and insights in subsequent stages of the project.

3.2 Data Preprocessing

Data preprocessing is a critical phase in any data science project, ensuring that the raw data is transformed into a format suitable for analysis. In the context of this project, where Python was the primary tool, the data preprocessing stage involved addressing issues such as inconsistent responses, missing values, and the need for encoding categorical variables.

3.2.1 Handling Inconsistent Responses

One of the challenges encountered during data collection was the variability in responses, particularly in categorical fields such as "University Name." Users entered the university names in different formats, leading

to redundancy and potential inconsistencies. To address this, custom mapping functions were devised to harmonize diverse entries. For instance, responses like "abc uni," "abc," "university of abc," and "abc texas campus" were mapped to a standardized representation, such as "abc."

3.2.2 Label Encoding

For machine learning models to process categorical data, it needs to be converted into numerical form. Label encoding was employed using the scikit-learn library, assigning a unique numerical label to each distinct category. This transformation enabled the seamless integration of categorical variables into the machine learning models.

3.2.3 Handling Missing Values

To ensure the completeness of the dataset, missing values were addressed by taking mode as all our inputs were categorical.

3.2.4 Handling Imbalanced Target Classes

An imperative challenge in our GPA prediction model lies in the significant imbalance observed within two specific GPA ranges: 4 and 1-1.9. The limited occurrence of instances in these ranges poses a potential hindrance to the model's effectiveness.

Feature Balancing While feature balancing appears as a potential solution, it brings its own set of challenges:

1. **Limited Representation:** The combined occurrence of classes 4 and 1-1.9 constitutes only 10 of the entire dataset.
2. **Scarcity for Similarity Generation:** Generating similar records through oversampling becomes a formidable task due to the scarcity of records.

An Innovative Approach: Class Combination

To address the constraints associated with feature balancing, we suggest an alternative approach – combining specific classes to create broader categories.

Class Combination Strategy

1. **Combining Classes 1 - 1.9 and 2 - 2.9:** This amalgamation results in a new class, "1-2.9."
2. **Combining Classes 4 and 3 - 3.9:** This combination yields a new class, "3 - 4."

By reclassifying and merging certain GPA ranges, we aim to create a more balanced distribution, alleviating the challenge of low counts in the original 4 and 1-1.9 ranges.

3.3 Power BI Visualization

For a more interactive and visually appealing exploration, Power BI was employed to create dynamic visualizations. Power BI's dashboard capabilities allowed for the creation of charts and graphs that facilitated the identification of trends and patterns.

3.3.1 Bivariate Analysis

Bivariate analysis explored relationships between pairs of variables. Correlation matrices, scatter plots, and pair plots were generated using Python libraries to uncover potential correlations and dependencies.

3.3.2 Key Findings

The EDA phase revealed several key findings, including: A positive correlation between average study time per day and GPA. Varied distributions in extracurricular activities and their potential impact on academic performance. Insights into the relationship between attendance and GPA.

3.3.3 Conclusion

The combination of Python libraries and Power BI in the EDA phase provided a comprehensive understanding of the dataset. Visualizations facilitated the identification of trends, patterns, and potential areas of interest for further analysis. The insights gained from EDA will inform subsequent modeling and analysis phases in this project.

3.4 Feature Selection

3.4.1 Random Forest Tree-Based Feature Importance

The Random Forest algorithm was employed to determine feature importance. The top features identified were:

1. Degree Name
2. Average Study Time per Day
3. Average Attendance
4. Extracurricular Activities
5. Free Time Activity
6. University Name

These features are deemed crucial in predicting a student's GPA, as they carry significant weight according to the Random Forest model.

3.4.2 SelectKBest

Utilizing the SelectKBest method with the f classif statistical test, the following features were identified as most relevant:

1. Average Attendance
2. Average Study Time per Day
3. Average Sleep Time
4. Extracurricular Activities
5. Free Time Activity
6. Parents' Highest Qualification
7. Volunteer Work

These features exhibit a strong statistical association with students' GPA, suggesting their importance in the predictive model.

3.4.3 VarianceThreshold

Applying the VarianceThreshold method, features with high variance were retained. The selected features include:

1. University Name
2. Degree Name
3. Average Study Time per Day
4. Extracurricular Activities
5. Average Sleep Time
6. Free Time Activity
7. Parents' Highest Qualification

These features are characterized by significant variability within the dataset, indicating their potential impact on GPA.

3.4.4 Results

The convergence of results from different feature selection techniques strengthens the confidence in the identified features' importance. The key factors influencing a student's GPA encompass the nature of their degree, study habits (such as average study time and attendance), engagement in extracurricular activities, free time activities, household size, and parental qualifications.

Educational institutions and policymakers can leverage these insights to develop targeted interventions and support systems to enhance student academic performance. Further analysis and model development can be undertaken to build a predictive model for GPA based on these crucial features, facilitating early identification of students who may benefit from additional support.

4 Conclusion

Our Machine Learning project, aimed at understanding and predicting factors influencing student academic performance, employed a structured approach from data collection to exploratory data analysis (EDA) and feature selection techniques. The transparent and reliable dataset, collected through a Google Form survey, covered diverse attributes, including university details, attendance patterns, study habits, and extracurricular activities.

Ethical considerations were paramount, ensuring participant confidentiality and informed consent. Data preprocessing, including addressing inconsistent responses and handling missing values, enhanced the dataset's quality and compatibility for machine learning models.

In the EDA phase, Power BI facilitated dynamic visualizations, revealing key findings such as a positive correlation between study time and GPA. Feature selection techniques, including Random Forest, SelectKBest, and VarianceThreshold, converged on crucial features like degree nature, study habits, and parental qualifications.

These identified features provide actionable insights for educational institutions and policymakers to enhance student academic performance. The combination of Python libraries and Power BI laid a strong foundation for subsequent modeling, emphasizing the potential for meaningful interpretations and future predictive analyses. This project underscores the value of a holistic and systematic examination of factors influencing student success.