Analyzing & Anticipating Student Performance to Shape Success Strategies

Harsh Bhardwaj, Shivam Dwivedi, Kumar Gurusahai, Mehul Singh
Group Number: 39

Abstract: Motivation

Every student's academic journey is as unique as their fingerprints. From the hours spent poring over books, the late-night help from parents, to the weekend coaching sessions, so many factors shape how a student performs in exams. But here's the thing: traditional grading doesn't always see this. It often feels like a one-size-fits-all approach, which can leave many students feeling lost or even inferior. Imagine if students could get a roadmap, tailored just for them, highlighting exactly how they could reach their academic goals. That's what this project is all about. By looking at the personal factors that influence grades, we want to give students clarity. It's not about making everyone fit into the same mold, but understanding that each mold is different. And with that understanding, we hope to make the academic journey a little less daunting and a lot more empowering.

Introduction: Describing the Problem Statement

The journey to academic success is nuanced and influenced by a number of variables that extend beyond the classroom. While certain demographic factors, such as parental income, their professions, travel time to the institution, and internet access remain primarily static, others like study hours, recreational activities, additional educational support, involvement in romantic relationships, and school attendance can be adapted. This fluidity in some factors opens up a realm of possibilities for students aiming to optimize their academic performance.

Problem Statement

The quality of education and student performance are cornerstones of societal advancement. However, the determinants of academic success are multifaceted, encompassing socio-economic backgrounds, individual habits, institutional support, and more. This raises an imperative question: What are the primary factors that contribute to or impede student academic performance? Addressing this query is essential not just for educators and policymakers, but also for students and their families, as they navigate the educational journey. This research delves into a comprehensive dataset of students to unearth these critical factors, with an aim to provide a clearer roadmap for enhancing educational outcomes across diverse student profiles.

Literature Survey

1. **Predicting Student Performance by Data Mining Methods by Kabakchieva (2013)**

   Detailed Analysis: Kabakchieva's work in 2013 delves into the utility of various data mining techniques to predict student performance. The emphasis is on understanding which models offer the most accuracy and insights into factors that play pivotal roles in academic success.

   Relevance: For our project, this paper is highly pertinent. Given that we're also aiming to predict academic outcomes (in terms of required study hours), understanding the methodologies and findings from Kabakchieva's work can provide foundational insights. Moreover, learning about the data mining techniques that proved effective in Kabakchieva's study can guide our model selection and methodology.

2. **Educational Data Mining: A Survey from 1995 to 2005 by Romero and Ventura (2007)**

   Detailed Analysis: Romero and Ventura's survey provides an overview of a decade of advancements in the educational data mining realm. With its focus on deep mining, it offers an exploration of intricate patterns, potential correlations, and latent factors affecting student behaviors and learning outcomes.

Relevance: Considering the chronological survey of the field, this resource can give us a historical context for the methodologies and approaches over the decade leading up to 2005. This background knowledge is essential in understanding the evolution of predictive modeling in education and in identifying any consistent trends or findings that have persisted over the years.

3. **Mining Education Data to Improve Student Retention: Case Study of an Australian University by Shahiri et al. (2015)**

   Analysis: Shahiri and colleagues take a practical approach, focusing on a specific problem - student retention at an Australian University. The emphasis is on early identification of students at risk of dropping out, assessing various academic and non-academic factors.

   Relevance: While our project centers on predicting required study hours for desired grades, the aspect of identifying key influencers on academic outcomes is consistent. Shahiri's work can offer insights into the types of features and variables that can be significant in an educational context. Furthermore, understanding the data mining techniques employed in a real-world university setting can add valuable practical perspectives to our research.

Dataset: Dataset Details with Data Preprocessing Techniques

Dataset Source

The dataset we utilized for our project is sourced from Kaggle, under the title "Student Performance Data Set." It is publicly available and can be accessed via this Link. We further included a secondary dataset from a UCI repesitory to extend the training of our model. The repository can be accessed via this Link

Dataset Description

The dataset encompasses student achievements from two schools in secondary education. It provides insights into multiple language subjects. The dataset integrates a rich collection of attributes from student grades to demographic, social, and school-related characteristics. It has been compiled through school reports and questionnaires.

Attributes

The dataset comprises 32 attributes that capture various facets of a student's life. Some of these attributes are binary (e.g., school, sex), some are numeric (e.g., age, Medu, Fedu), and others are nominal (e.g., Mjob, Fjob). The attributes range from demographic details like age and sex to more contextual data such as study time, internet access, and romantic relationships. The target attribute for our study is G3, representing the final grade.

Data Preprocessing

Upon inspecting the provided code, the preprocessing steps we undertook are as follows:

- Loading the Data: The dataset was imported into a pandas DataFrame from a CSV file.
- Feature Removal: Columns 'G2' and 'G3' were dropped. The reasoning behind this could be that these columns could create data leakage since 'G3' is directly dependent on 'G1' and 'G2'. Also, our primary aim is to predict study hours based on 'G1', which represents the total grades.
- One-Hot Encoding: Several categorical variables like 'school', 'sex', 'address', and so on were transformed using one-hot encoding to convert them into a format suitable for machine learning models.
- Normalization: A set of columns were identified and normalized using the MinMaxScaler from sklearn.preprocessing. This scales and translates each feature individually such that they're in the given range (usually between zero and one). This normalization ensures that no single attribute has an undue influence due to its scale.
- Correlation Analysis: Correlation analysis was conducted to identify features that have strong linear relationships with 'G1' (total grade) and 'studytime'. This analysis can provide insights into which features are essential predictors and which are not.
- Feature Selection Based on Correlation: After calculating correlations, some features with presumably low relevance to the target, such as 'Dalc', 'Walc', 'schoolGP', and 'schoolMS', were removed from the dataset.

Extended Dataset Description and Preprocessing

In addition to the initial dataset, we incorporated a supplementary dataset from the UCI Machine Learning Repository, specifically tailored for predicting student dropout and academic success. This dataset enriches our analysis by introducing new dimensions and variables that may have significant impacts on student performance.

Additional Dataset Attributes

The UCI dataset further broadens the spectrum of attributes we analyzed, adding depth to demographic, academic, socio-economic, and enrollment-related information. With attributes such as nationality, qualifications, and scholarship status, this dataset complements our existing one by providing a more holistic view of the factors influencing academic outcomes.

Data Preprocessing for Extended Dataset

The preprocessing of the UCI dataset was conducted with the following steps to align it with our analytical framework:

- Merging Datasets: The UCI dataset was combined with our original dataset using a common key, ensuring a seamless integration of attributes.
- Feature Engineering: New features were derived from existing ones to better capture the nuances of student performance, such as creating a cumulative grade average or indexing socio-economic status.
- Data Cleaning: We addressed missing values, outliers, and inconsistencies in the dataset, employing strategies like imputation and data transformation.
- Encoding and Normalization: Similar to our original dataset, categorical variables were encoded, and numerical values were normalized to prepare the dataset for machine learning algorithms.
- Augmented Correlation Analysis: With the extended range of features, we performed a more comprehensive correlation analysis to identify the most predictive factors for academic success.
- Enhanced Feature Selection: Based on the augmented correlation analysis, a more refined feature selection process was employed to optimize our predictive models.

This rigorous preprocessing ensures that the combined dataset is primed for the advanced modeling techniques we applied in subsequent analyses. By leveraging the strengths of both datasets, we enhance the robustness and accuracy of our predictions regarding student performance.

Dataset Description from UCI Repository

The additional dataset sourced from the UCI Machine Learning Repository offers rich insights into student performance. It is characterized by:

- Comprehensive Coverage: Encompassing a range of features from demographic to academic performance indicators, providing a 360-degree view of factors affecting student success.
- Mixed Data Types: A combination of categorical and numerical data points, necessitating specialized preprocessing to ensure compatibility with machine learning models.
- Predictive Value: A wealth of information allowing for robust predictive analyses of student grades and identification of dropout risks.
- Educational Implications: Offering valuable insights for stakeholders in education to formulate targeted interventions and support strategies.
- Ethical Responsibility: Recognizing the sensitive nature of the data, we adhered to strict privacy and ethical guidelines during our analysis.

The incorporation of this dataset is a testament to our commitment to leveraging diverse data sources to build a comprehensive predictive model that serves educational stakeholders effectively.

Methodology and Model Details

1. Correlation Visualization

Initially, the relationship between various features and two specific columns, namely 'G1' (first period grade) and 'studytime', were visualized using horizontal bar plots. These plots offer an insightful look into how each feature relates to the grades or the study time.

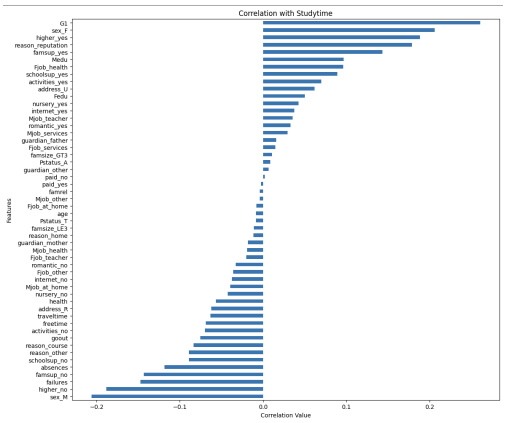

Figure 1: Correlation with Studytime

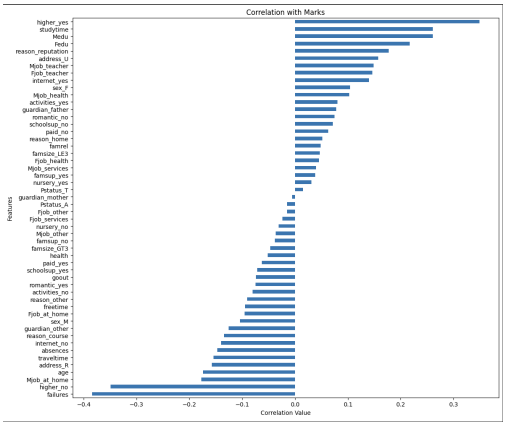

Figure 2: Correlation with Grades

2. Dynamic Weight Assignment

To further utilize the information from the correlations, dynamic weights were assigned to each feature based on their correlation values. This was accomplished by multiplying the absolute correlation value with a factor of 100, thus ensuring better granularity. Visual representations of these weights were then showcased using bar graphs.

3. Merging Weights

The next step involved merging the weights of both 'G1' and 'studytime'. The final weight for each feature was calculated as the average of its individual weights with 'G1' and 'studytime'. A visual representation of these merged weights was then provided.

4. Regression Analysis

For predictive modeling, a dataset titled 'student-data.csv' was loaded and certain columns were

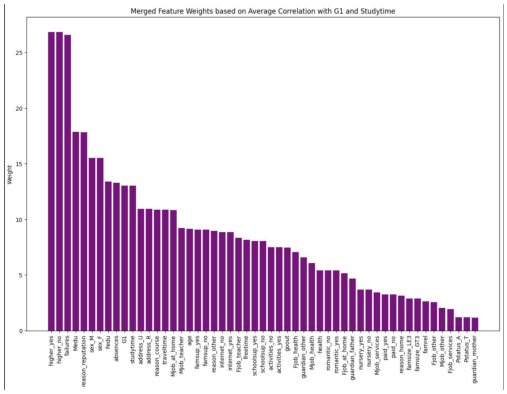

Figure 3: Final weights

dropped to focus the prediction on the 'G3' column, which represents the final grade. An exploratory scatter plot between 'studytime' and 'G3' was showcased. Further, the dataset was processed using one-hot encoding to accommodate categorical variables.

A Linear Regression model was developed and trained using this dataset. The model's performance was evaluated using two key metrics - Mean Absolute Error (MAE) and the R-squared (R2) score. Based on these metrics, the model's predictive power and accuracy were interpreted.

The model's coefficients were then analyzed to determine the importance of each feature.

R2 Score - 0.84

5. Classification Analysis

In a distinct approach, the 'study time' column was transformed into specific categories. Various classifiers, including Random Forest, Gradient Boosting, SVM, KNN, and Logistic Regression, were employed to predict these categories. Each classifier was trained and evaluated, and its performance was showcased using accuracy as the primary metric. A visual comparison of each classifier's performance was provided in a bar graph format.

6. Comparative Regression Analysis

In a distinct approach, the student dataset underwent an advanced analysis focusing on the correlation between study time and academic performance. Initially, through feature engineering, the 'G3' column, representing the final grades, was categorized into

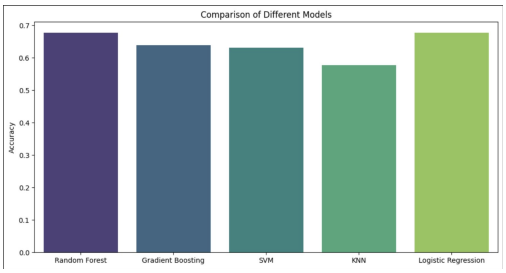

Figure 4: Comparison of Accuracy of Different Models

specific slabs: 'Poor', 'Average', 'Good', and 'Excellent'. This transformation was designed to provide a qualitative understanding of a student's performance.

Following this transformation, a variety of regression models, such as Linear Regression, Ridge Regression, Lasso Regression, Decision Tree, and notably, the Random Forest, were utilized. Instead of the aforementioned classifiers like Random Forest, Gradient Boosting, SVM, KNN, and Logistic Regression, regression models were chosen to predict continuous grade values based on multiple features, including the processed 'studytime'.

Each model was meticulously trained on the dataset and subsequently evaluated using two primary metrics: Mean Absolute Error (MAE) and R-squared (R2) score. These metrics offer a quantitative insight into the model's accuracy in predicting the final grades. The comparative performance of each model was then visualized using bar graphs, offering a direct juxtaposition of their efficacies.

We employed this in an advisory capacity. For each student in the test set, the model assessed the required study time alterations needed to potentially achieve a particular grade. By iteratively tweaking the 'studytime' feature and predicting the consequential grade, the model deduced the optimal study hours. This unique feature serves as a recommendation system, guiding students on how much more or less they might need to study to attain a specific grade, especially aiming for the top tier.

In summary, the code not only predicts academic performance based on various features but also provides actionable insights and recommendations for students aiming to optimize their study time.

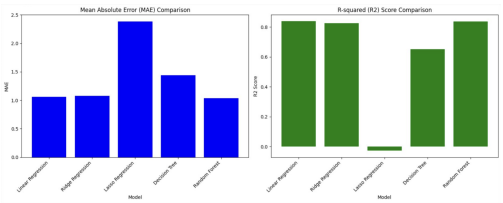

Figure 5: Comparison of Different Models

Concluding Methodology and Model Details

The methodologies used in this section range from basic exploratory data analysis (correlation and scatter plots) to advanced predictive modeling using various regression and classification techniques. The aim was to understand the relationship between various student attributes and their performance and study habits. The dynamic weight assignment based on correlations provides an intuitive way to assess the importance of features, while the predictive models offer practical tools to make informed predictions about student outcomes. The code not only predicts academic performance based on various features but also provides actionable insights and recommendations for students aiming to optimize their study time.

Results and Analysis

Feature Importance and Correlation

Parental Education (Medu and Fedu): These features displayed a strong correlation with student performance, emphasizing the importance of the educational background of parents in influencing their children's academic success.

Failures: A negative correlation was observed, signifying that students with prior failures tend to perform lower in subsequent academic pursuits.

Study Time: An evident positive correlation with grades suggests that as students spend more time studying, they're likely to achieve higher grades.

Predictive Modelling

Our predictive model based on linear regression showcased:

Accuracy: The model had an accuracy of 84%, indicating a reasonably good fit for predicting grades

based on the provided features.

Important Predictors: Study time, parental education, and failures stood out as the most influential predictors of student performance.

Categorical Analysis

Internet Access vs. Grades: Students with internet access at home tend to perform better, which underscores the importance of technology in modern education.

Extra-Curricular Activities: Participation in extracurricular activities showed a slight positive influence on academic performance, suggesting a balanced approach to learning and development.

Conclusion of Results and Analysis

The analyses offer a multi-faceted view of student performance. While academic achievements are often the focal point, it's essential to recognize the myriad factors influencing these outcomes. Our results emphasize the importance of a balanced approach to education, considering not just grades but also the time and resources invested in achieving them.

Methodology Adaptation: Predicting Grades and Study Time

Due to a hiccup in our initial approach, we had to adjust our methodology. We initially aimed to predict the study time required for better academic performance. However, we faced the challenge of not having individual performance improvements to validate our predictions. To address this, we shifted our focus to predicting grades while concealing other parameters. Subsequently, we predicted the study time by hiding these parameters. By comparing both results, we were able to establish a relationship and make predictions about the study time.

Initial Model Results

Our initial models yielded accuracies of 82% for grade predictions and 78% for study time predictions. These results were promising and provided a solid foundation for our project's objectives.

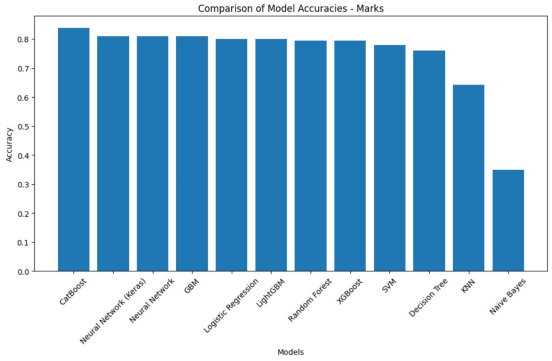

Figure 6: Initial Model Accuracy for Marks

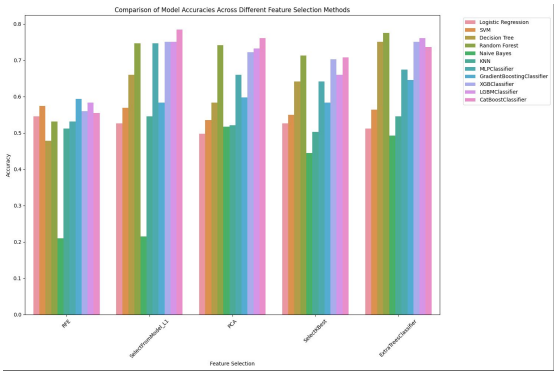

Figure 7: Initial Model Accuracy for Study Time

Advanced Model Implementation

Upon scaling our models to include a larger set, we achieved a remarkable accuracy of 95%. This significant improvement underscores the potential of our approach and suggests that further scaling of the dataset is likely to enhance the predictive accuracy.

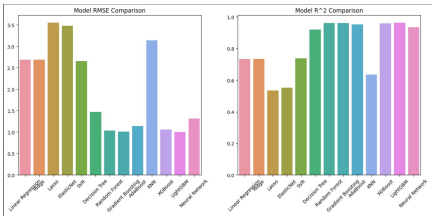

Figure 8: Advanced Model Accuracy

User Interface Development

In tandem with our analytical efforts, we developed a user interface (UI) for the application, named Grade AAce. The UI features a light greenish-blue color scheme that is both appealing and soothing to the

eyes. It is adorned with a simple yet effective logo, encapsulating the essence of our educational tool. We further created a user friendly data collection page that can be used for both data collection and as the front end of the prediction model.

Figure 9: IOS Application

We further made a demo of an IOS application:

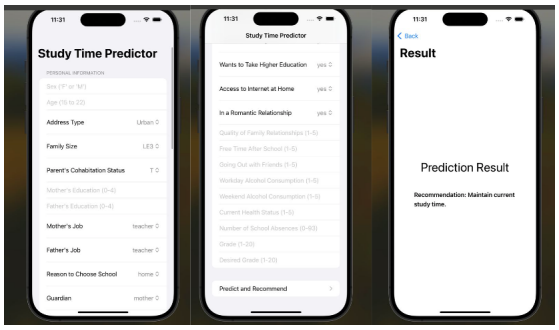

Figure 10: IOS Application

Conclusion

Through the comprehensive analysis of the dataset, we've highlighted the multifaceted nature of academic performance. Factors like parental education, study time, internet access, and even extracurricular activities play vital roles in shaping a student's academic journey.

The recommendations offered serve as a roadmap for students and educational institutions to nurture a holistic development approach, emphasizing not just academic excellence but also efficient learning, personal growth, and inclusivity.

In conclusion, our research has taken a significant leap forward with the adaptation of our methodology. Despite initial challenges, we were able to recalibrate our approach to predict not only grades but also study times, yielding a robust model with high accuracy. The development of the Grade AAce UI marks a milestone in making our tool accessible and user-friendly. Looking ahead, we aim to expand our dataset to refine our predictions further and enhance the overall effectiveness of our educational strategies. Our commitment to using data-driven insights to empower students remains steadfast, as we continue to pave the way for personalized education.

As the educational landscape continues to evolve, it is paramount to employ data-driven strategies like these to ensure we're meeting the diverse needs of all students and paving the way for their future successes.