Project E9: KAGGLE - STUDENTS' PERFORMANCE ANALYSIS

Team members : Annabel Aleksius, Mattias Minejev, Tõnis Tõnissoo

Business Understanding Report

Identifying Your Business Goals

Background

High school academic performance forms the basis of opportunities that students will have in higher education and beyond. However, a student's grades are heavily influenced by several factors, including study habits, parental involvement, and extracurricular activities. In analyzing this dataset, we seek to gain insight into these factors and further develop a predictive model to identify students who might be at risk of poor performance, enabling timely intervention.

Business Goals

- Identify key factors that influence students' academic performance.
- Develop a predictive model that categorizes the students' grades into the following categories: 'A', 'B', 'C', 'D', 'F'.

Business Success Criteria

- The model must achieve an accuracy of at least 85%.
- Provide interpretable results highlighting the top factors affecting performance.
- Recommend actionable strategies to improve academic outcomes for students in lower grades ('C', 'D', 'F') by at least 10%.

Assessing Your Situation

Resources Inventory

- Dataset: A synthetic dataset of 2,392 high school students with attributes such as demographics, study habits, parental involvement, extracurricular activities, and GPA.
- Tools: Python (Jupyter Notebook) for data analysis, modeling, and visualization.

- Infrastructure: Cloud-based resources for data storage and model training.

Requirements, Assumptions, and Constraints

- Requirements:
    - Access to all features in the dataset, particularly GPA, GradeClass, and other influencing factors like StudyTimeWeekly and ParentalSupport.
- Assumptions:
    - The dataset accurately represents typical high school students' demographics and behaviors.
    - External factors like curriculum changes or policy variations do not significantly alter outcomes during the study.
- Constraints:
    - Limited time (around one month) to complete analysis and modeling.
    - The synthetic nature of the dataset may limit generalizability to real-world applications.

Risks and Contingencies

- Risk: Model Overfitting
    - Contingency: Use cross-validation to test the model on different subsets of the data. Keep the model straightforward and apply techniques like regularization to prevent it from being too tailored to the training data. Test the model on separate data to ensure it works well in practice.
- Risk: Resistance to Recommendations
    - Contingency: Present findings clearly using easy-to-understand visuals and practical suggestions. Involve teachers and parents early to ensure their input is considered. Start with small-scale trials to show how the recommendations can work effectively.

Terminology

- GradeClass: The classification of the grades of students based on GPA.
- Parental Support: The involvement and assistance provided by parents in the student's academic life.

- Extracurricular Activities: Non-academic engagements like sports, music, and volunteering, which may influence performance.

Costs and Benefits

- Costs:
    - Time spent cleaning and analyzing the dataset.
    - Computational resources for running models.
- Benefits:
    - Improved academic performance and reduced failure rates.
    - Enhanced parental awareness and engagement in education.
    - Better decision-making tools for educators.

Defining Your Data-Mining Goals

Data-Mining Goals

- Build a predictive model to classify students into GradeClass categories.
- Identify the most relevant predictors of academic performance.
- Offer actionable insights to better support students in the lower categories of performance.

Data-Mining Success Criteria

- Model achieves at least 85% accuracy on test data.
- Feature importance analysis identifies top factors influencing grades.
- Simulation or case study demonstrates potential improvements based on recommendations.

https://github.com/mattiasminejev/andmeteadus.git