**Name : Ayman Abdelnasser Gamil**

**Project : Students performance**

**Abstract**

**This project investigates the prediction of student performance in two subjects: Mathematics and Portuguese. The core objective is to build machine learning models to predict outcomes based on a rich set of demographic, social, and academic features. The data from both subjects were meticulously merged into a single DataFrame to create a more robust and generalized model. The preprocessing pipeline involved encoding categorical variables and applying a robust feature engineering process to identify and retain only the most highly correlated features.**

**A variety of machine learning models were trained and rigorously compared, with their accuracies ranging between 80% and 86%. The Logistic Regression model emerged as the top performer, achieving an impressive accuracy of 0.876 after a meticulous process of feature selection and hyperparameter tuning. The project also delved into advanced techniques, including a Deep Learning approach and a Sentiment Analysis experiment, to provide a comprehensive and forward-looking analysis of student performance factors.**

**1. Problem Statement**

**Student performance prediction is a critical and transformative task in the field of educational data mining. It serves as a powerful tool for educators to proactively identify students who are at risk of academic failure and to provide timely, targeted interventions. The central problem addressed in this work is how to effectively predict student outcomes by leveraging their personal, academic, and social attributes. By solving this problem, the project provides a data-driven framework that enables educators to make informed decisions, improve learning outcomes, and ultimately, help students reach their full potential. This initiative is not merely about forecasting grades; it's about empowering institutions to foster a supportive and effective learning environment for every student.**

**2. Dataset and Data Description**

**The foundation of this project is built upon a unique dataset derived from two separate sources: one for student performance in Mathematics and another for Portuguese. By combining these two datasets, a single, comprehensive DataFrame was created. This strategic decision was crucial for building a more generalized predictive model that is not confined to a single subject, while also significantly increasing the volume of data available for training.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**The dataset includes an extensive range of features that can be broadly classified into three categories:**

**Demographic and Personal Features: This category includes fundamental information such as the student's age (age), gender (sex), residential address (address), family size (famsize), and number of absences (absences). These variables provide a basic but essential understanding of the student's background.**

**Social and Family Features: This group of features delves into the student's family environment, capturing important details like the educational level of their mother (Medu) and father (Fedu), their respective jobs (Mjob, Fjob), the presence of family support (famsup), and the family's cohabitation status (Pstatus). These are often powerful indicators of the level of educational encouragement student receives at home.**

**Academic and Behavioral Features: This category is directly linked to the student's school life. It includes vital variables such as weekly study time (studytime), the number of past class failures (failures), whether they receive extra school support (schoolsup), and their ambition for higher education (higher). Most importantly, the dataset includes previous grades (G1, G2), which are naturally the strongest predictors of final performance. The data also includes behavioral attributes like weekday (Dalc) and weekend (Walc) alcohol consumption.**

**The target variable for this project is the student's final grade, which I transformed from a continuous numerical value into a categorical variable with three distinct risk levels:**

**A screenshot of a computer

AI-generated content may be incorrect.**

**High Risk (Fail < 10)**

**Medium Risk (Pass 10–13)**

**Low Risk (Pass ≥ 14)**

**3. Methodology**

**The methodology employed in this project follows a structured and comprehensive pipeline, starting from data preprocessing and moving through model training, tuning, and evaluation.**

**3.1. Preprocessing Steps**

**Before any models were trained, the raw data underwent a series of critical preprocessing steps to prepare it for analysis:**

**Data Merging: The Math and Portuguese datasets were merged to create a single DataFrame. A new feature, Subject, was added to differentiate between the two subjects, which proved to be a valuable predictor.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Categorical Encoding: All categorical features (sex, address, reason) were transformed into numerical representations. This was achieved using LabelEncoder, which assigns a unique integer to each category, making the data readable for machine learning algorithms.**

**A black rectangular object with white lines

AI-generated content may be incorrect.**

**Feature Engineering: A new feature, Averge\_rate, was engineered by calculating the average of the first (G1), second (G2), and final (G3) grades. This feature provided a holistic view of the student's academic progression.**

**A black screen with green text

AI-generated content may be incorrect.**

**Feature Selection: A feature correlation analysis was performed to identify the most influential attributes. The f\_classif method was used to select the top-performing features, resulting in a refined list of 14 key variables for model training: ['school', 'Medu', 'Fedu', 'reason', 'studytime', 'failures', 'schoolsup', 'higher', 'Dalc', 'Walc', 'G1', 'G2', 'Subject', 'Averge\_rate'].**

**A graph showing a number of different colored bars

AI-generated content may be incorrect.**

**3.2. Machine Learning Models**

**A variety of classical machine learning models were trained and their performance was compared:**

**k-Nearest Neighbors (KNN): A non-parametric classification method that classifies a new data point based on the majority class of its nearest neighbors.**

**Support Vector Machine (SVM): A powerful model that finds the optimal hyperplane to separate data points into different classes.**

**Logistic Regression: A linear model used for classification. Despite its simplicity, it often performs exceptionally well, as demonstrated in this project.**

**Random Forest: An ensemble method that constructs multiple decision trees and merges their predictions to improve accuracy and control for overfitting.**

**Decision Tree: A model that uses a flowchart-like structure to classify data points based on a series of feature-based decisions.**

**3.3. Model Tuning**

**To maximize performance, a hyperparameter tuning process was applied to the models. This involves systematically testing different parameter combinations to find the optimal set for a given model and dataset. For the Logistic Regression model, the best-performing parameters were found to be {C: 10, penalty: 'l2', solver: 'lbfgs'}. This fine-tuning was crucial for achieving the model's high accuracy.**

**And my best model I Logistic Regression**

**A screenshot of a computer

AI-generated content may be incorrect.**

**3.4. Deep Learning Approach**

**A deep learning model was implemented using the popular Keras and TensorFlow libraries. The model consisted of a multi-layer perceptron architecture, which was trained over multiple epochs. While this approach is highly effective for large datasets, the model achieved a lower accuracy of approximately 0.82 due to the relatively small size of the dataset. This highlights a common limitation of deep learning, which often requires a massive amount of data to outperform classical machine learning models.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**A graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of a graph of

AI-generated content may be incorrect.**

**3.5. Sentiment Simulation (Natural Language Processing)**

**In an innovative step, the project explored the potential impact of sentiment on student performance by analyzing the reason feature (why a student chose a specific school). This task falls under the domain of Natural Language Processing (NLP).**

**An initial attempt was made using the TextBlob library, which resulted in errors.**

**The project successfully pivoted to using BERT (Bidirectional Encoder Representations from Transformers), a state-of-the-art NLP model.**

**The analysis successfully showed a strong correlation between positive sentiment and higher student performance, suggesting that a student's attitude or motivation is a significant, measurable factor that can influence their academic outcomes.**

**4. Results**

**4.1. Machine Learning Models**

**The trained machine learning models achieved accuracies ranging between 0.80 and 0.86.**

**The Logistic Regression model, after comprehensive tuning, demonstrated the best performance features with an accuracy of 0.876.**

**4.2. Classification Report and Confusion Matrix (Logistic Regression)**

**The performance of the best model was further evaluated using a classification report and a confusion matrix.**

**Classification Report:**

**Precision: The model achieved a high precision of 0.89 for both the High Risk and Low Risk categories, indicating its reliability in correctly identifying these students.**

**Recall: The recall scores ranged from 0.85 to 0.89 across all classes, showing that the model was effective at identifying a high percentage of students in each category.**

**F1-score: The F1-score was approximately 0.88, a strong indicator of the model's balance between precision and recall.**

**Confusion Matrix: [[39 0 7], [0 51 8], [5 6 93]] The confusion matrix provides a clear visualization of the model's performance. The high values on the diagonal (39, 51, 93) confirm that the model correctly classified a large number of students in each risk category. The low numbers off the diagonal indicate a minimal number of misclassifications, demonstrating the model's high predictive power**

**4.3. Deep Learning and NLP Results**

**Deep Learning: The neural network achieved an accuracy of 0.82. This confirms the hypothesis that the limited dataset size was a bottleneck for the deep learning approach.**

**Sentiment Simulation: The NLP analysis successfully demonstrated a positive correlation between student sentiment (as captured from the reason variable) and their academic performance, opening new avenues for future research into non-academic factors.**

**5. Ethics Considerations**

**Throughout the project, a strong commitment to ethical data practices was maintained.**

**Anonymization: All student names and personal identifiers were removed from the datasets to ensure complete privacy.**

**Fairness: The features were carefully selected to avoid any inherent biases that could unfairly target specific groups of students. The project explicitly states that no features were used that could unfairly bias results against specific groups.**

**Use of Results: The project's findings are intended to be a supportive tool for educators to improve academic outcomes, not for punitive measures or to stigmatize students.**

**Transparency: The limitations of the project, such as the dataset size being a constraint for deep learning models, are openly reported to maintain full transparency.**

**6. Conclusion**

**This project successfully demonstrates that student performance prediction is a feasible and highly impactful application of machine learning. The Logistic Regression model, with its optimal feature selection and hyperparameter tuning, proved to be the most reliable and accurate solution, achieving a stellar 0.876 accuracy. While the deep learning model performed less effectively due to data limitations, the innovative sentiment analysis experiment provided valuable insights into the influence of external factors.**

**The project not only provides a powerful predictive tool but also highlights the critical importance of key factors like parental education, study habits, and even personal sentiment in determining academic success. Future work should focus on expanding the dataset, exploring more advanced deep learning architectures for larger data, and deploying the model in real-world educational settings to provide real-time, actionable insights to teachers and administrators.**

**Application Flask**

**A screenshot of a computer

AI-generated content may be incorrect.**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Relation between positive comments and degrees ?**

**A graph showing a line graph

AI-generated content may be incorrect.**

**Positive Relation**

**As positive comments As Higher Degree you Gain**

**Streamlit App**

**A screenshot of a computer

AI-generated content may be incorrect.**

A screenshot of a computer

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.