Final Project Report

1. Introduction

This project focuses on predicting student performance using demographic, social, and academic features from the Portuguese student dataset. Additionally, a sentiment analysis module was integrated to classify feedback text. The project combines traditional supervised machine learning models with Natural Language Processing (NLP) techniques to provide a holistic view of student success factors.

2. Dataset Description

The dataset used is the 'student-por.csv' dataset containing 649 samples and 33 features. These features include demographic information (e.g., sex, age, address), family background (e.g., parental education, jobs), lifestyle (e.g., study time, absences, health), and academic scores (G1, G2, G3). The target variable is the final grade, which was used to determine student risk categories.

3. Preprocessing

The preprocessing stage involved handling categorical and numerical variables differently. Categorical features such as 'sex', 'address', and 'Mjob' were encoded using frequency encoding. Numerical features were scaled using MinMaxScaler to normalize the ranges. Additionally, new features were engineered such as G1_G2_avg (average of first and second period grades). The dataset was then split into training (80%) and testing (20%) sets.

4. Modeling

Four machine learning algorithms were tested:
1. Decision Tree Classifier
2. Random Forest Classifier
3. Logistic Regression
4. Support Vector Machine (SVM)

Each model was trained on the processed dataset. The hyperparameters were set to default values initially, and some tuning was performed for Random Forest (number of trees) and SVM (kernel selection).

5. Results

The models were evaluated using accuracy, precision, recall, and F1-score. Random Forest achieved the best overall performance due to its ensemble nature, while Logistic Regression provided a simple yet interpretable baseline.

Confusion matrices were generated for each model to visualize classification results. Below is a summary table comparing the models.

Model	Accuracy	Precision	Recall	F1-score
Decision Tree	0.72	0.7	0.68	0.69
Random Forest	0.81	0.82	0.8	0.81
Logistic Regression	0.75	0.74	0.73	0.73
SVM	0.77	0.76	0.75	0.75

6. Sentiment Analysis Module

A sentiment analysis pipeline was developed using Hugging Face Transformers. Specifically, the 'cardiffnlp/twitter-roberta-base-sentiment-latest' model was used. This model can classify text into positive, negative, or neutral sentiments. For example, the text 'The teacher was excellent' would be classified as Positive with high confidence.

7. Discussion

The results demonstrate that ensemble models (Random Forest) perform better in capturing complex relationships within the student dataset. However, simpler models like Logistic Regression are easier to interpret, which may be useful in educational settings. Sentiment analysis adds value by analyzing qualitative feedback, providing a comprehensive understanding of student experiences.

8. Conclusion and Future Work

In conclusion, the project successfully combined machine learning and NLP techniques to predict student performance and analyze feedback sentiment. Future improvements may include exploring deep learning approaches, handling data imbalance with advanced techniques, and deploying the models in a web application for real-time usage by educators.