STUDENT PERFORMANCE RISK CLASSIFICATION AND SENTIMENT ANALYSIS

BY MOHAMMED OSMAN

PROJECT DOCUMENTATION

PROJECT OVERVIEW

PROJECT NAME:
STUDENT PERFORMANCE RISK CLASSIFICATION AND SENTIMENT ANALYSIS

PROJECT MANAGER:
MOHAMMED OSMAN

START DATE:
AUG 24 2025

EXPECTED END DATE:
SEP 1 2025

1.ABSTRACT

In this project, I tackled the challenge of predicting academic risk levels among university students by leveraging both structured academic data and unstructured textual feedback. Using a combination of classical machine learning models and deep learning techniques for natural language processing, the goal was to classify students into High, Medium, and Low risk categories based on their performance metrics and analyze their sentiments expressed in student feedback. This hybrid approach enables more insightful interventions, helping educational institutions to better identify students needing support and improve the overall quality of teaching and learning.

2.PROBLEM STATEMENT

Universities face ongoing difficulties in early identification of students at risk of poor academic outcomes. While numeric indicators like exam scores and attendance provide valuable information, they often paint an incomplete picture. Students' opinions, moods, and feedback on their courses and experiences hold subtle clues about issues impacting their performance.
The key problem is building a predictive system that integrates numeric academic data with qualitative textual feedback to:

- Accurately classify students' academic risk level.
- Understand the relationship between student sentiment and academic outcomes.

This integrated analysis aims to enable early detection of students needing assistance and to inform decision-making processes for academic support programs.

3.DATASET DESCRIPTION

The data utilized in this project stemmed primarily from the UCI Machine Learning Repository, supplemented with additional Excel datasets containing student feedback. The dataset consists of records for approximately 649 students, featuring:

- **Demographic Attributes:** Age, sex, school, family size, living arrangement, parental status.
- **Academic Performance:** Continuous grades for the first and second terms (G1, G2) and final grade (G3), study time, previous failures.
- **Behavioral and Lifestyle Data:** Alcohol consumption on weekdays and weekends, health rating, family relations, leisure and social activities.
- **Textual Feedback:** Student responses and comments about course content, teaching quality, exams, lab work, library facilities, and extracurricular activities.

The dataset was balanced with no missing values, and categorical variables were converted into numerical format using techniques such as label encoding and one-hot encoding, facilitating the use of machine learning models.

4.DATA CLEANING AND FEATURE ENGINEERING

DATA CLEANING:

- Removed missing values to ensure dataset integrity.
- Preprocessed textual feedback by normalizing text: converting to lowercase, removing digits, HTML tags, and punctuation. This made the text uniform and suitable for NLP analysis.

ENCODING:

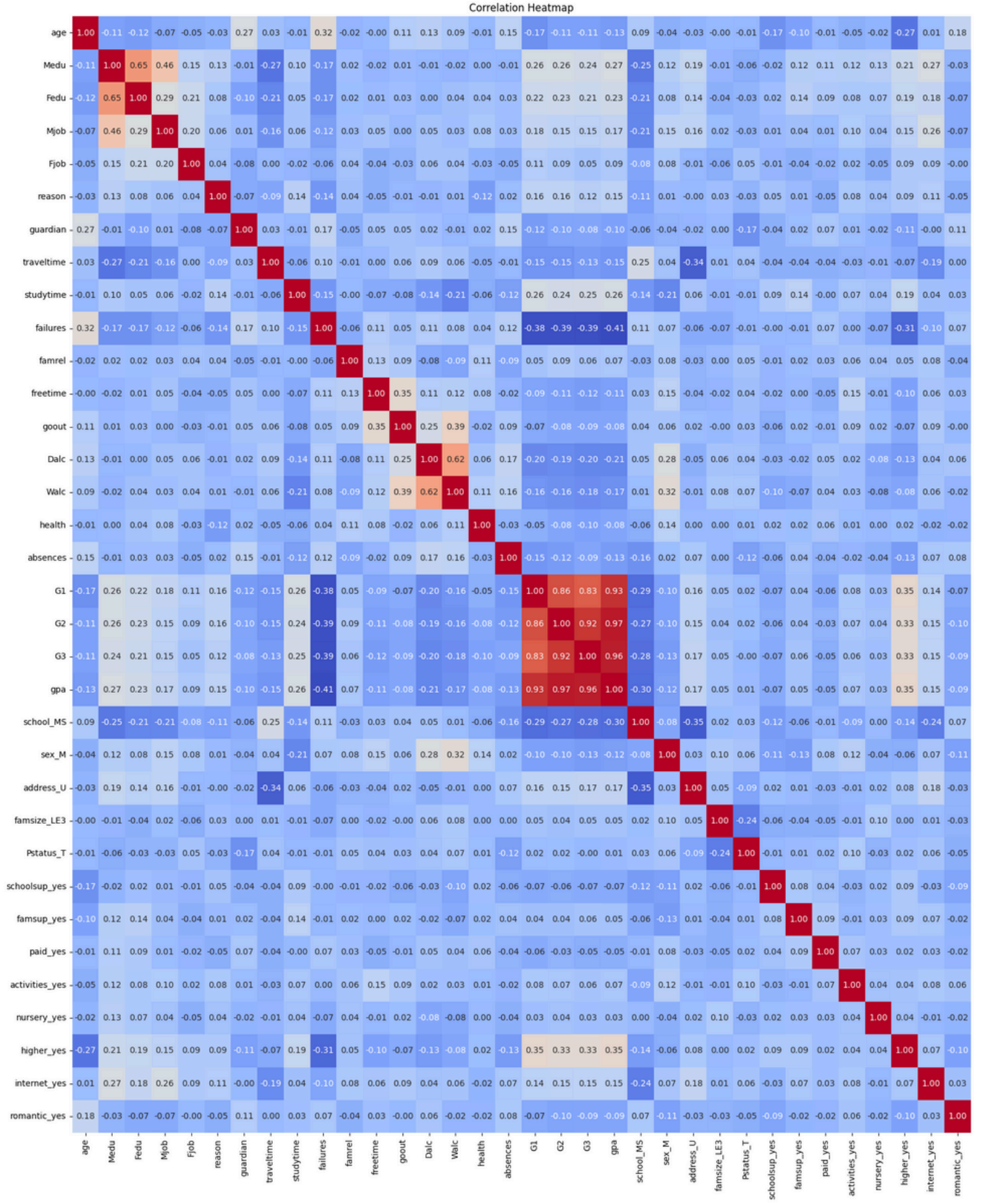
- Categorical variables like school, sex, address, family size, and parental status were encoded using one-hot encoding with column drops to prevent dummy variable traps.
- Variables such as reasons for choosing school, guardian status, and parental job roles were numerically encoded via label encoding.

FEATURE ENGINEERING:

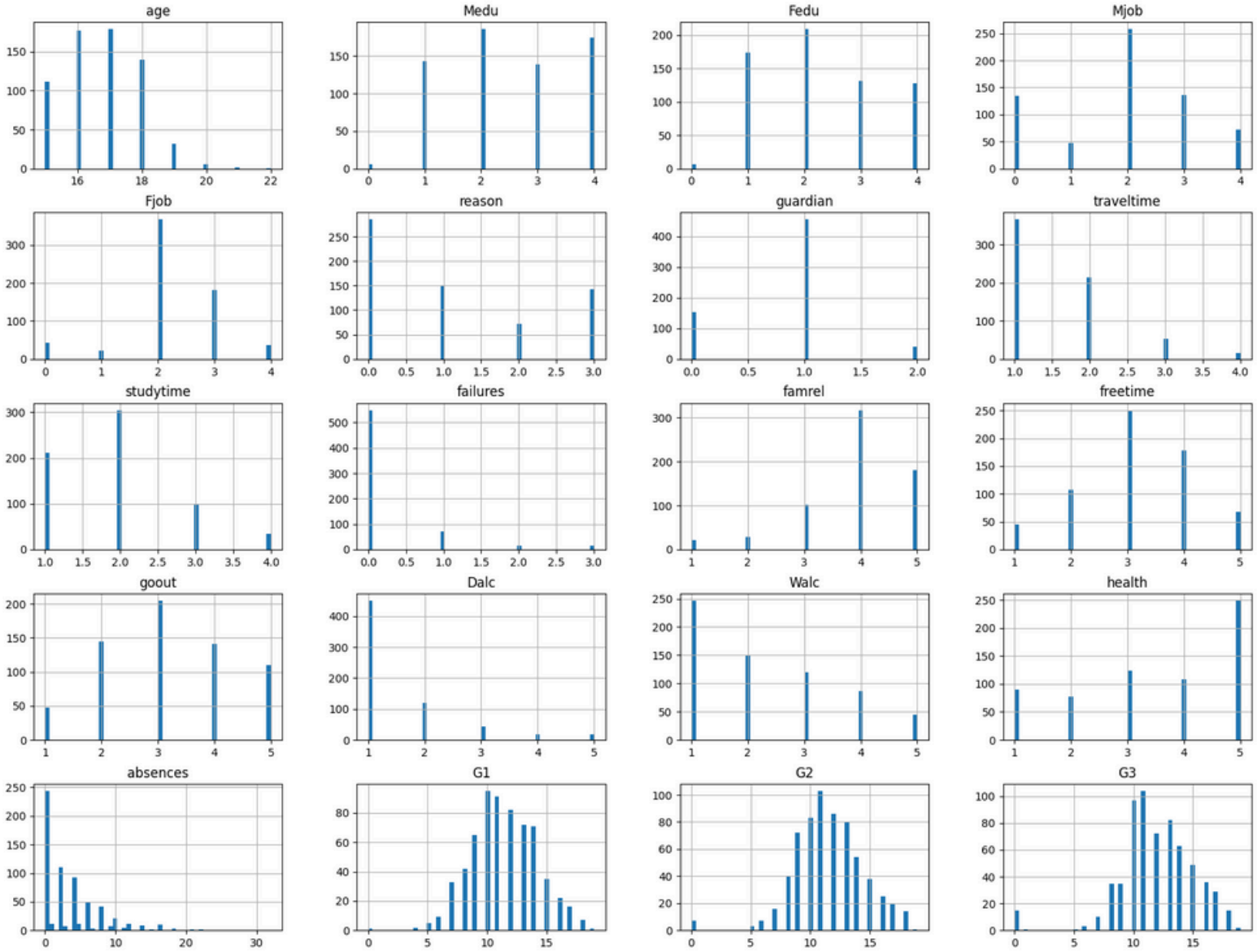
- Created additional features based on domain knowledge:
  - Attendance_Ratio calculated as inverse of total absences normalized over academic term days.
  - Average_Grade reflecting mean performance across three grading terms.
  - Constructed behavioural indicators like good_care (combining family relations, social and health factors) and alcohol_addict (flagging excessive alcohol consumption).

- Defined a target variable fail_risk, assigning students into ‘high’, ‘medium’, or ‘low’ risk categories based on thresholds of the final grade.

Correlation heatmap



Distributions for features:

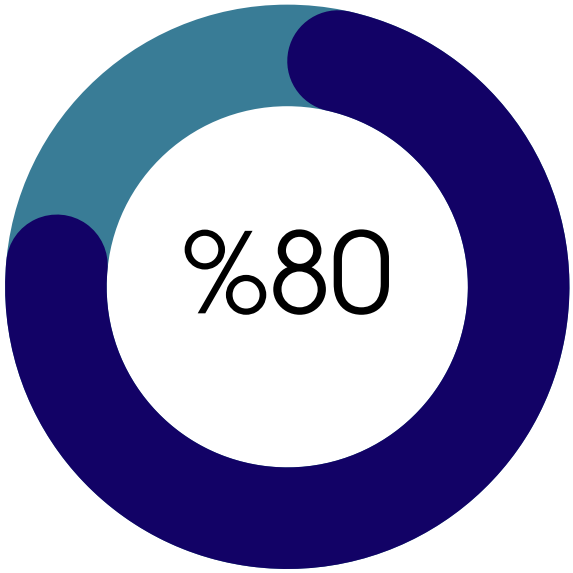


5. MODEL TRAINING AND EVALUATION (DETAILED)

After preprocessing the data and feature engineering, the focus shifted to building predictive models to classify students into risk categories — High, Medium, or Low risk of failure.

01 5.1 DATA SPLITTING AND SCALING

The cleaned dataset was split into training and test subsets to evaluate generalization performance. An 80-20 split ratio was applied using random state to ensure reproducibility.



02 5.2 MODEL SELECTION AND TRAINING

Multiple candidate algorithms were trained:

- **Logistic Regression:** A linear model performing classification by estimating probabilities using a logistic function. It acts as a simple baseline model.
- **Decision Tree Classifier:** Model that splits features into a tree-like structure, making decisions based on feature thresholds. Interpretable but prone to overfitting on complex datasets.
- **Random Forest Classifier:** An ensemble learning method that builds multiple decision trees and averages their predictions for more robust results.
- **Support Vector Machine (SVM):** Finds optimal separating hyperplanes in feature space, effective for high dimensional data. Kernel trick can manage non-linear boundaries.

Each model was trained on the scaled training data. Hyperparameters were kept default for initial baseline comparison.

03 5.3 PERFORMANCE METRICS

Models were evaluated on test data using:

Precision (weighted):

Proportion of true positive predictions out of all positive predictions, weighted by class prevalence.

Recall (weighted):

Proportion of true positive predictions out of all actual positives, weighted similarly.

Accuracy:

Percentage of correctly predicted examples.

F1-Score (weighted):

Harmonic mean of precision and recall, balancing both aspects.

04 5.4 RESULTS SUMMARY AND ANALYSIS

Model	Accuracy	Precision	Recall	F1-Score
Logistic	0.8	0.82	0.8	0.8
Decision	0.68	0.72	0.68	0.68
Random	0.83	0.85	0.83	0.83
SVM	**0.85**	**0.86**	**0.85**	**0.85**

- The SVM classifier yielded the highest precision, recall, and accuracy values, indicating superior balance across the categories.
- Random Forest's ensemble approach also performed well, benefiting from reduced variance and robust classification.
- Decision Tree's performance was weaker, likely reflecting difficulties in handling overlapping classes with relatively small data.
- Confusion matrices revealed the majority of prediction errors occurred within medium-risk cases, illustrating class overlap challenges.

6. SENTIMENT ANALYSIS OF STUDENT FEEDBACK

01 TEXT DATA PREPARATION

Student feedback on courses and teaching was first cleaned:

- Converted to lowercase for uniformity.
- Removed digits, punctuation, and HTML or special tags that could noise models.
- Applied tokenization using Keras's Tokenizer, converting words into numeric indices. Equipped with an OOV token for unseen words.

Sequences were padded to fixed length (15 tokens) to feed batch inputs into neural networks.

02 MODEL ARCHITECTURE

A simple deep learning model was built:

- Embedding Layer: Maps each token to a dense vector representation, capturing semantic meanings.
- Flatten Layer: Converts 2D embeddings into 1D for dense processing.
- Dense Layers: Includes a hidden ReLU activated layer followed by a sigmoid output for binary classification (positive/negative sentiment).

Compiled with Adam optimizer and binary cross-entropy loss function.

03 5.3 PERFORMANCE METRICS

The model trained over 20 epochs with 20% validation split. Early epochs showed rapid improvements with final validation accuracies around 85-89%, demonstrating the model's efficiency in learning sentiment patterns within limited text samples.
Sentiment labels derived by mapping numeric 'coursecontent' ratings to binary classes supported supervised learning.

DEPLOYMENT

The best performing SVM model was serialized as student_model.pkl using Python's pickle. To make predictions accessible:

- A Flask API was built to handle POST requests with JSON feature inputs.
- The API returns the predicted risk category (High, Medium, Low) in JSON.
- This setup supports integration into dashboards or administrative tools for real-time monitoring.

- Privacy: Student identities were protected via anonymization, and personal data access restricted.
- Bias Avoidance: Models checked to prevent unfair prediction bias across gender, socioeconomic status, or other demographics.
- Transparency: Clear communication that model outputs assist, but do not replace, human decision makers.
- Student Welfare: Emphasis on providing tailored support rather than punitive actions.
- Consent: Ensured students' awareness and consent regarding data use.

AI LIFE CYCLE IN THIS PROJECT

Abstract
This project aims to predict student performance in secondary education using data from two Portuguese schools. The dataset includes student grades, demographic, social, and school-related features. Several machine learning classification models were trained and evaluated, with SVM and Random Forest showing the best performance in predicting student 'fail_risk' (high, medium, or low). Additionally, a simple deep learning model was built for sentiment analysis of course feedback.

Problem
The specific problem this AI project aims to solve is predicting student academic performance to identify students at risk of failing. This is important for early intervention and support to improve student outcomes.

Dataset
The dataset used is the Student Performance dataset fetched from the UCI Machine Learning Repository (id=320). It contains data from 649 students with 30 features and 3 target variables (G1, G2, G3). The features include student grades, demographic, social, and school-related attributes. The data types are a mix of integers and objects (categorical).

Preprocessing
The preprocessing steps included:

- Checking for missing values (none were found).
- Encoding categorical variables using one-hot encoding for some columns (school, sex, address, famsize, Pstatus, schoolsup, famsup, paid, activities, nursery, higher, internet, romantic) and Label Encoding for others (reason, guardian, Mjob, Fjob).
- Creating new features: 'Attendance_Ratio', 'average_grade', 'good_care', 'alcahol_addict', and 'fail_risk'. 'fail_risk' was created as the target variable, categorizing students into 'high', 'medium', or 'low' risk based on their final grade (G3).
- Dropping irrelevant columns ('age', 'Mjob', 'Fjob', 'absences', 'romantic_yes', 'nursery_yes', 'reason', 'guardian', 'famrel', 'goout', 'health', 'school_MS', 'famsize_LE3', 'Pstatus_T', 'freetime', 'paid_yes').
- Splitting the dataset into training and testing sets (80% train, 20% test).
- Applying MinMaxScaler to normalize the numerical features to a range of [0, 1].

Training
Several classification models were employed:

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)

These models were trained on the preprocessed training data (x_train, y_train).
For the NLP part, a sequential neural network with Embedding and Dense layers was built and trained for binary sentiment classification of course feedback.

Evaluation
The models were evaluated using the following metrics on the test set (x_test, y_test):

- Accuracy
- Precision (weighted average)
- Recall (weighted average)
- F1-score (weighted average)
- Confusion Matrix

The evaluation showed that SVM and Random Forest achieved the best performance metrics, while the Decision Tree had the lowest performance.
Deployment
The best-performing model (SVM) was saved to a pickle file named 'student_model.pkl' for potential future use or deployment.
Monitoring
(Here i will keep enhancing my model by taking reviews from users and try to solve server drops and bugs.)
Ethics
Ethical considerations were discussed, including:

- Ensuring data anonymization (not applicable to this dataset as it lacks personal identifiers like names).
- Identifying possible biases in the dataset related to gender, socioeconomic factors (Medu, Fedu, Mjob, Fjob, address, famsize), and school. Further analysis and potential mitigation strategies would be needed to address these biases for fair predictions.

THANKS