Project Proposal

Thanh Nguyen, Santiago Ponce, Josh Rubino
tnguy278@charlotte.edu, sponce1@charlotte.edu, jrubino@charlotte.edu.

Title:

Predicting Student Academic Performance Using Machine Learning.

Introduction and Background:

Motivation:

- Academic success is a significant factor in shaping an individual's future career opportunities. As educational institutions aim to improve student outcomes, machine learning provides a powerful tool to identify students at risk of underperforming. This project is particularly interesting because it combines data science with a real-world social impact helping educators and policymakers take proactive measures to support students.

Background:

- Previous research has shown that various factors, including parental education, school resources, and student behavior, affect academic performance. Several publicly available datasets exist that track these variables alongside students' final grades, offering opportunities to build predictive models.

Problem Statement:

- The goal is to develop a machine learning model that can predict a student's final grade based on factors such as past performance, demographic data, and school-related metrics. This predictive model can be used to detect students who may need additional support early in the academic term.

Objectives:

- To explore the relationships between student characteristics and academic performance.
- To build and evaluate classification and regression models to predict students' final grades.
- To identify the most important features influencing academic success.
- To suggest actionable insights for educators based on model findings.

Data:

- Source: UCI Machine Learning Repository - Student Performance Data Set
- Size: ~1,000 student records,
- Format: CSV (comma-separated values).
- Features: Includes 33 features such as school, gender, age, family background, study time, failures, and previous grades (G1, G2).
- Target Variable: Final grade (G3), which can be treated as:
    - Regression: Predict exact grade (0–20).
    - Classification: Classify performance as low (0–9), medium (10–14), high (15–20).

Methodology:

Algorithms/Models:

- Linear Regression (is used to predict a continuous value that is based on one or more input variables, like a student's final grade. It tries to minimize the difference between expected and actual values in order to draw a straight line that best fits the data).

- Logistic Regression (used to predict a class or category, like low, medium, or high performance. It estimates the likelihood that a student would fall into a specific class using a curve, known as a sigmoid function, as opposed to a line, as in linear regression).

- Naive Bayes (a classification algorithm that calculates probability using Bayes' Theorem, because it assumes that all input features are independent, even when they are not, and it is referred to be naive. It often works quite well in spite of this assumption).

Evaluation Metrics:

- For Regression Models:
    - Regression Line (used to visually examine the trend between the input variables and the predicted final grade in order to see how well the linear regression model matches the data).
- For Classification Models:
    - Accuracy (will measure how many students were correctly put into low, medium, or high performance categories).
    - Precision (will show how many students predicted as “high-performing” actually are high-performing).
    - Recall (will show how well the model identifies all students who actually belong to a specific performance group, for example the "at risk" group).
    - F1-score (will balance precision and recall to give an overall description of classification quality, especially when class sizes are not even).
    - Confusion Matrix (will show a visualization on how many students were either correctly or incorrectly classified into each category, helping to show the common incorrect classifications).

Implementation Plan:

Steps:

1. Data Collection (download from UCI repository).
2. Exploratory Data Analysis (EDA).
3. Data Preprocessing:
    a. Handling missing values.
    b. Encoding categorical variables.
    c. Feature scaling.

4. Model Training and Evaluation.
5. Model Tuning.
6. Interpretation of Results.
7. Documentation and Presentation of Findings.

Timeline:

Week	Task
Week 1	Data acquisition and EDA
Week 2	Preprocessing and baseline models
Week 3	Model tuning and evaluation
Week 4	Final analysis, visualization, and report writing

Tools and Libraries:

- Python.
- NumPy, Pandas (data handling).
- Matplotlib, Seaborn (visualization).
- Scikit-learn (ML algorithms).
- Jupyter Notebook (development environment).

Expected Outcomes:

- A functioning model that predicts student performance with high accuracy.
- Insights on the most influential factors affecting grades.
- A user-friendly report or interface (optional) for educators to input new student data and receive predictions.

- Potential extension: apply models to other academic datasets or integrate into school dashboards.

References:

- Cortez, Paulo, and Alice Silva. "Using data mining to predict secondary school student performance." (UCI Repository)
- Scikit-learn documentation: https://scikit-learn.org/stable/

Conclusion:

This project aims to leverage the power of machine learning to predict student academic performance, a challenge with significant implications in the field of education. By analyzing a rich dataset of student characteristics and academic history, we hope to identify patterns that influence final grades and develop predictive models that can assist educators in making data-informed decisions. Our proposed methodology includes a combination of regression and classification techniques, supported by thorough data preprocessing and evaluation. With the right implementation and tuning, the outcomes of this project can offer valuable insights into student learning behaviors and help shape more effective academic interventions. Ultimately, this project not only serves as an opportunity to apply machine learning techniques in a meaningful context but also highlights the potential for technology to enhance educational outcomes and equity.