**CEN 352 Artificial Intelligence**

**PROJECT  
  
Topic:** Student performance prediction

***Students Name:***

**1. Atea Caslli**

**2. Daniela Shahini**

**3. Enia Çerri**

**4. Irvi Rrika**

**5.Rosela Berberi**

**6.Thanas Papa**

**Study Program:** Software Engineer 3

**Group:** 3

**Date of submission**: 11/19/2024

Table of Contents

1. [**Introduction**](#Introduction)
2. [**Methodology**](#Methodology)
   * [Data Preprocessing](#DataPreprocessing)
   * [Feature Description](#FeatureDescription)
3. [**Dataset Overview**](#DatasetOverview)
   * 5-Level Classification
   * Binary Classification
   * Continuous Distribution of Final Grades
4. [**Data Processing**](#DataProcessing)
5. [**Results**](#Results)
   * Five-Level Classification Results
   * Binary Classification Results
   * Regression Results
6. [**Comparison**](#Comparison)
7. [**Conclusion**](#Conclusion)
8. [**Supplementary Materials**](#Repo)

**Introduction**  
 Predicting student performance is a vital area of education research, aimed at understanding the factors that influence academic success and finding ways to provide timely support to students. In this study, we utilize the UCI Student Performance dataset to predict students' final grades (G3) under three different scenarios: using all available features, excluding the second-period grade (G2), and further excluding the first-period grade (G1). The analysis is conducted across three prediction types—binary classification (pass or fail), 5-level categorical classification, and regression (predicting continuous grades)—to evaluate and compare the effectiveness of different predictive models.  
  
 The results indicate that high predictive accuracy can be achieved if the grades from the first and/or second school periods are available. While previous grades strongly influence student performance, an analysis also revealed other important factors, such as the number of absences, parents' jobs and education levels, and alcohol consumption. This research can lead to the development of better tools for predicting student performance, helping to improve education quality and optimize the use of school resources.  
  
**Methodology**  
 The study used the UCI Student Performance dataset, which contains information about students' demographic, social, and academic characteristics. Data preprocessing was an essential first step, involving cleaning to handle missing values, encoding categorical variables into numerical formats, and normalizing continuous variables. This prepared the dataset for effective analysis.

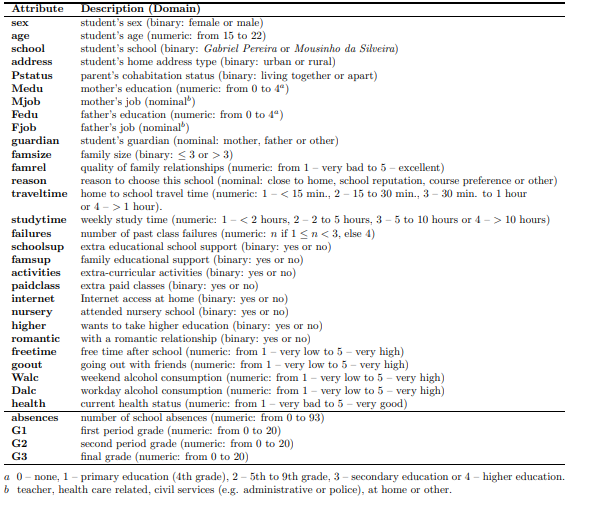
The dataset was divided into three scenarios to explore the impact of various features. The first scenario used all available features to predict the final grade (G3). In the second scenario, the second-period grade (G2) was removed to evaluate the effect of excluding this important feature. In the third scenario, both the first-period grade (G1) and the second-period grade (G2) were excluded, further challenging the models to predict G3 based on the remaining attributes.

Three types of prediction tasks were performed in each scenario. The first task was binary classification, which aimed to determine whether a student passed or failed based on their final grade. The second task involved a 5-level classification, assigning grades into categories ranging from 0 (lowest) to 5 (highest). The third task used regression techniques to predict the exact numerical value of the final grade.

To achieve these predictions, several machine learning and deep learning models were applied, including Neural Networks, Support Vector Machines (SVM), Decision Trees, Random Trees, Naive Bayes, Logistic Regression, and advanced Deep Learning architectures. The performance of these models was evaluated using appropriate metrics for each task. For classification tasks, metrics such as accuracy, precision, recall, and F1-score were employed. For regression tasks, performance was assessed using metrics like Root Mean Squared Error.

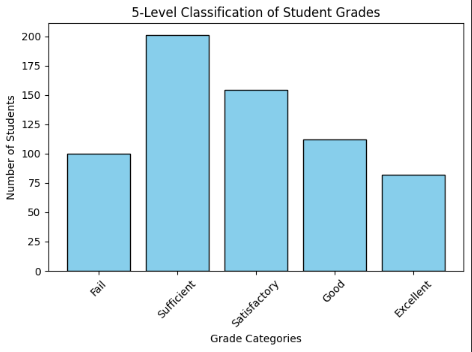
Cross-validation techniques were used to ensure the robustness of the evaluation process. By splitting the dataset into training and testing sets multiple times, the models' performance was validated across different data partitions. This helped provide a reliable comparison of how well each model performed in the various scenarios and prediction tasks.

Finally, the results were analysed to identify which models and features contributed most to accurate predictions. Insights gained from this analysis were used to conclude the key factors affecting student performance and the practical implications for educational systems.

**Feature Description  
  
FEATURES:** 30 **TARGETS:** 3  **INSTANCES:** 649 **MISSING VALUES:** No

**Dataset Overview** In this study, the academic performance of 650 students was analysed using three distinct classification methods: a 5-level classification, a binary pass/fail classification, and a continuous distribution of final grades. Each approach provides valuable insights into the students’ overall performance, and the histograms derived from these classifications reveal important trends and patterns.

**1. 5-Level Classification (Very Poor to Excellent)**

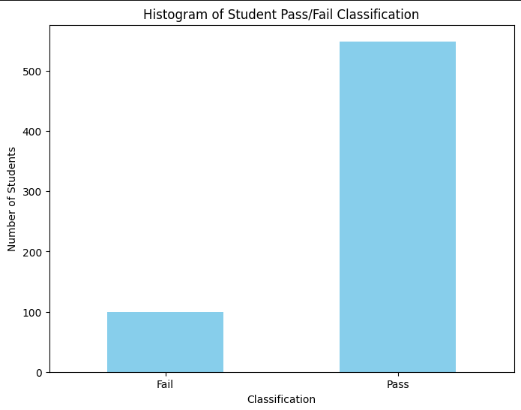
 The first classification divides the students' final grades (G3) into five levels of performance:

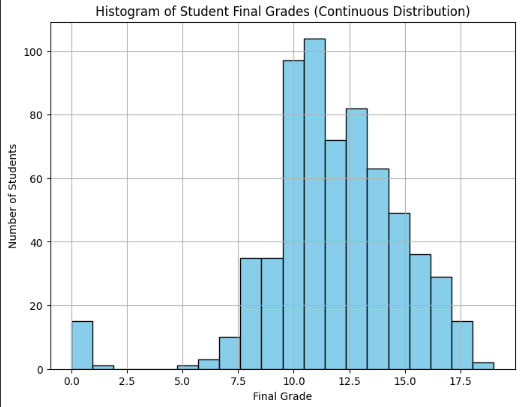
* **Level 1: Fail (0-9)**
* **Level 2: Sufficient (10-11)**
* **Level 3: Satisfactory (12-13)**
* **Level 4: Good (14-15)**
* **Level 5: Excellent (16-20)**

This approach allows for a detailed analysis of student performance across a broad spectrum of grades. The histogram of this classification shows that most students tend to cluster around the middle levels, with **Level 3 (Average)** and **Level 4 (Good)** having the highest frequency of students. This suggests that the overall performance of students is moderately distributed, with fewer students at the extremes (very poor or excellent). The distribution is somewhat skewed, with a notable concentration of students in the **Average** to **Good** range, indicating that most students perform at or above average.

**2. Binary Classification (Pass/Fail)**

The binary classification splits the students into two categories based on whether their final grade is below or above the passing threshold:

* **Fail (0-9)**: Students who did not meet the passing grade.
* **Pass (10-20)**: Students who successfully passed the course.

**** The histogram for the pass/fail classification reveals that a larger proportion of students fall into the **Pass** category, with only a small percentage failing. This reflects a generally high success rate among students, suggesting that the educational approach employed may be effective in helping the majority of students achieve a passing grade. The **Pass** category is significantly more populated, while the **Fail** category is relatively sparse, indicating that most students in this cohort are capable of meeting the minimum grade requirements.  
  
**3. Continuous Distribution of Final Grades** The third method presents the students' final grades as a continuous distribution, which provides a more detailed view of how grades are spread across the entire scale (0-20). The histogram for this continuous distribution shows that the grades are not evenly distributed. There is a concentration of students around the **Average** to **Good** ranges, with the distribution tailing off towards both the **Very Poor** and **Excellent** extremes. The grade range appears to be right-skewed, as the bulk of students are clustered in the middle to upper-middle part of the grading scale, and fewer students achieve the highest possible grades.**Data preprocessing** In this study, various machine learning models were employed to predict student performance based on their features. The models used include Logistic Regression, Naive Bayes, Random Forest, Decision Tree, Support Vector Machine (SVM), Neural Network, and Deep Learning. Each model was subjected to different data preprocessing techniques, which included encoding, normalization, and dimensionality reduction (PCA). The goal was to find the best-performing model for predicting student grades while evaluating the impact of different preprocessing methods on model accuracy.

**1.** **Data Encoding**For categorical variables, two types of encoding techniques were applied:

* **One-Hot Encoding:** This encoding technique was used for variables where the categorical values were nominal (without any intrinsic order).
* **Label Encoding:** This method was applied to ordinal categorical variables where there is a clear ranking or order.

These encoding techniques were chosen to ensure that the categorical features could be used effectively in the machine learning models.

**2.** **Normalization**

To ensure that all features contributed equally to the models, **Standardization** (using StandardScaler) was applied for most of the models. This step transformed the data so that each feature had a mean of zero and a standard deviation of one, which is particularly important for models sensitive to the scale of input features, such as Logistic Regression, Naive Bayes, SVM, and Neural Networks. In particular cases (for testing reasons), such as Naive Bayes, Z-Score normalization showed better results than Min-Max normalization.  
 However, two models did not require normalization. **Random Forest** and **Decision Tree** models were not normalized as they are not sensitive to the scale of the data. These models inherently handle different scales well by making splits based on feature values.

**3. Principal Component Analysis (****PCA)**

PCA was used as a dimensionality reduction technique to reduce the number of features while preserving as much variance as possible in the data. The optimal number of principal components was determined using the **Elbow Rule**, which indicated that 23 components should be retained for further analysis.

* **Impact of PCA on Accuracy:**
  + **Classifiers (e.g., Logistic Regression, Naive Bayes, Decision Tree, Random Forest, SVM):** For these models, PCA generally **reduced accuracy** because the reduced number of features sometimes caused the model to lose valuable information. This was especially noticeable in models like **Logistic Regression** and **Naive Bayes**.
  + **Regressors (e.g., SVM):** On the other hand, PCA **improved the accuracy** for models like **SVM**, where reducing dimensionality helped the model focus on the most significant features, thereby enhancing performance.

**Data Processing**  
 Based on the three different prediction tasks (binary classification, 5-level classification, and regression), the dataset was divided into three scenarios:  
 *Scenario A:* All features were used to predict the final grade (G3).  
 *Scenario B:* The second-period grade (G2) was removed to assess the effect of excluding this important feature.  
 *Scenario C:* Both the first-period grade (G1) and second-period grade (G2) were excluded, challenging the models to predict the final grade based on the remaining attributes.

**Results**  
Binary Classification Results (**Accuracy,** in % – the left-most column of each model, **Recall,** in % **–** the right-most column of each model, underline- best model, **bold-** best within input setup)

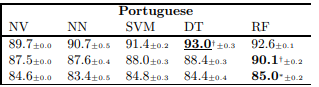
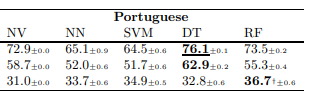
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | NB | RF | DT | SVM | LR | NN | DL | HYBRID |
| A | 93 | 92 | 95 | 91 | 90 | 96 | 90 | **97** |
| B | 91 | 90 | 92 | 89 | 91 | 94 | 88 | **95** |
| C | 87 | 87 | 88 | 89 | 88 | 90 | 86 | **92** |

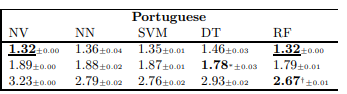
Five-Level Classification Results (**Accuracy,** in % – the left-most column of each model, **Recall,** in % **–** the right-most column of each model, underline- best model, **bold-** best within input setup)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | NB | RF | DT | SVM | LR | NN | DL | HYBRID |
| A | 61 | 75 | 75 | 66 | 63 | 70 | 60 | **80** |
| B | 54 | 51 | 62 | 49 | 53 | 64 | 49 | **66** |
| C | 49 | 39 | 34 | 42 | 30 | **50** | 36 | 45 |

Regression results (RMSE values, underline- best model, **bold-** best within input setup)

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | NB | RF | DT | SVM | NN | DL | HYBRID |
| A | 2.49 | 1.32 | 1.23 | 1.81 | 1.28 | 1.59 | **1.16** |
| B | 2.96 | 1.70 | **1.63** | 2.51 | 1.78 | 1.73 | 1.69 |
| C | 3.92 | 2.82 | 2.92 | 2.93 | 2.93 | 2.86 | **2.79** |

**Comparison**The tables below show the results conducted by a study taken in 2008 in Portugal(“[**Using data mining to predict secondary school student performance**](https://www.semanticscholar.org/paper/61d468d5254730bbecf822c6b60d7d6595d9889c)”, P. Cortez, A. M. G. Silva. 2008) high schools using the same dataset (except they also conducted research and tried to predict grades of students in Mathematics which had a size sample of 350 students).   
  
Binary Classification Results (**Accuracy,** in % – the left-most column of each model, **Recall,** in % **–** the right-most column of each model, underline- best model, **bold-** best within input setup)  
  
Five-Level Classification Results (**Accuracy,** in % – the left-most column of each model, **Recall,** in % **–** the right-most column of each model, underline- best model, **bold-** best within input setup)****Regression results (RMSE values, underline- best model, **bold-** best within input setup)

****Here is what head-to-head comparison of the best models looks like:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Binary Classification | | 5 level Classification | | Regression | |
| Year | **2025** | **2008** | **2025** | **2008** | **2025** | **2008** |
| A | **NN-**96 | **DT-**93 | **RF,DT-**75 | **DT-**76 | **NN-**1.28 | **NV,RF-**1.32 |
| B | **NN-**94 | **RF-**90.1 | **DT-**62 | **DT-**63 | **DT-**1.63 | **DF-**1.78 |
| C | **NN-**90 | **RF-**85 | **NN-**50 | **RF-**37 | **RF-**2.82 | **RF-**2.67 |

**Conclusion**  
 This study explored predicting student performance using the UCI Student Performance dataset. By employing advanced machine learning and deep learning models, it examined binary classification, five-level classification, and regression tasks across multiple scenarios. The findings demonstrated that prior grades (G1 and G2) significantly influence prediction accuracy, but even without these features, the models were able to extract valuable insights from other attributes such as absences, parental education, and lifestyle habits. This analysis contributes to identifying the key factors affecting academic success and provides educators with tools to enhance interventions and resource allocation.  
  
 The results highlight the effectiveness of neural networks, which achieved the highest accuracy (96%) for binary classification and competitive performance across other tasks (except hybrid model). Compared to prior research in 2008, which relied on decision trees and random forests as top-performing models, our study demonstrates improved predictive capabilities, especially with regression tasks where RMSE values were consistently lower. For example, the hybrid model achieved an RMSE of 1.16 in Scenario A compared to 1.32 in prior research. These advancements emphasize the potential of deep learning techniques in predictive analytics, offering superior performance over traditional methods.

**Supplementary Materials**The complete implementation of the models, preprocessing steps, full codes, analysis and visualizations performed in this study can be found in our GitHub repository:[**Student Performance Prediction - GitHub Repository**](https://github.com/Slowbro213/AI-Project/tree/master)