ML Lab Assignment 2: "Student Performance Ranking Analysis"

25MCMI20, Sachin A

30/01/2026

1 Introduction

In this experiment, I analyzed how different performance metrics and ranking methods affect student evaluation and machine learning model behavior. The goal was to calculate student percentages, percentiles, and grades from a marksheet dataset, rank students using each method, and then train machine learning models to see if different ranking approaches lead to different model behaviors.

This assignment helped me understand that the way we measure and rank student performance can significantly impact how machine learning models learn patterns and make predictions.

2 Dataset Description

The dataset used in this experiment is a student marksheet containing marks in four subjects. The dataset has the following structure:

- **id**: Student identification number
- **Name**: Student name
- **Gender**: Gender of the student
- **Age**: Age of the student
- **Section**: Class section (A, B, or C)
- **Science**: Marks in Science (out of 100)

- **English**: Marks in English (out of 100)
- **History**: Marks in History (out of 100)
- **Maths**: Marks in Mathematics (out of 100)

The dataset contains 250 student records. Each student has marks in four subjects, with a maximum of 100 marks per subject, making the total maximum possible marks equal to 400.

3 Libraries Used

The following Python libraries were used in this experiment:

- **NumPy**: Used for numerical operations and calculations.
- **Pandas**: Used for loading the dataset, data manipulation, and creating new calculated columns.
- **Matplotlib**: Used for creating visualizations and bar charts.
- **Seaborn**: Used for styling the plots and making them visually appealing.
- **Scikit-learn**: Used for splitting the dataset, training the Decision Tree classifier, and computing evaluation metrics like accuracy and F1 score.

4 Methodology

4.1 Step 1: Calculating Total Marks

The first step was to calculate the total marks obtained by each student by adding marks from all four subjects:

$$\text{Total Marks} = \text{Maths} + \text{Science} + \text{English} + \text{History} \tag{1}$$

This gives us a single score that represents the overall performance of each student out of 400 marks.

4.2 Step 2: Calculating Percentage

The percentage was calculated by dividing the total marks by the maximum possible marks (400) and multiplying by 100:

$$\text{Percentage} = \frac{\text{Total Marks}}{400} \times 100 \tag{2}$$

For example, if a student scored 213 out of 400, their percentage would be 53.25%.

4.3 Step 3: Calculating Rank

Students were ranked based on their total marks in descending order. The student with the highest total marks gets Rank 1, the second highest gets Rank 2, and so on. I used the `rank()` function with `ascending=False` to ensure higher marks result in better (lower number) ranks.

4.4 Step 4: Calculating Percentile

Percentile indicates the percentage of students who scored below a particular student. It was calculated using the formula:

$$\text{Percentile} = \left(1 - \frac{\text{Rank} - 1}{N}\right) \times 100 \tag{3}$$

where N is the total number of students. For example, if a student is ranked 111 out of 250 students, their percentile would be 56.0, meaning they performed better than 56% of the students.

4.5 Step 5: Assigning Performance Labels

To create categorical labels for machine learning, I assigned labels based on percentage, percentile, and grade.

4.5.1 Percentage Labels

Percentage-based labels categorize students into performance bands:

Percentage Range	Label
≥ 91	C1
81–90	C2
71–80	C3
61–70	C4
50–60	C5
< 50	F

4.5.2 Percentile Labels

Similarly, percentile-based labels categorize students based on their relative performance:

Percentile Range	Label
≥ 91	P1
81–90	P2
71–80	P3
61–70	P4
50–60	P5
< 50	F

4.5.3 Grade Labels

Grade labels follow a traditional academic grading system:

Percentage Range	Grade
≥ 91	O
85–90	A+
75–84	A
65–74	B+
60–64	B
55–59	C
50–54	P
< 50	F

5 Data Visualization

Before training the models, I visualized how different labeling schemes distribute students across categories.

5.1 Distribution of Labels

Figure 1 shows how students are distributed across the three different labeling schemes.

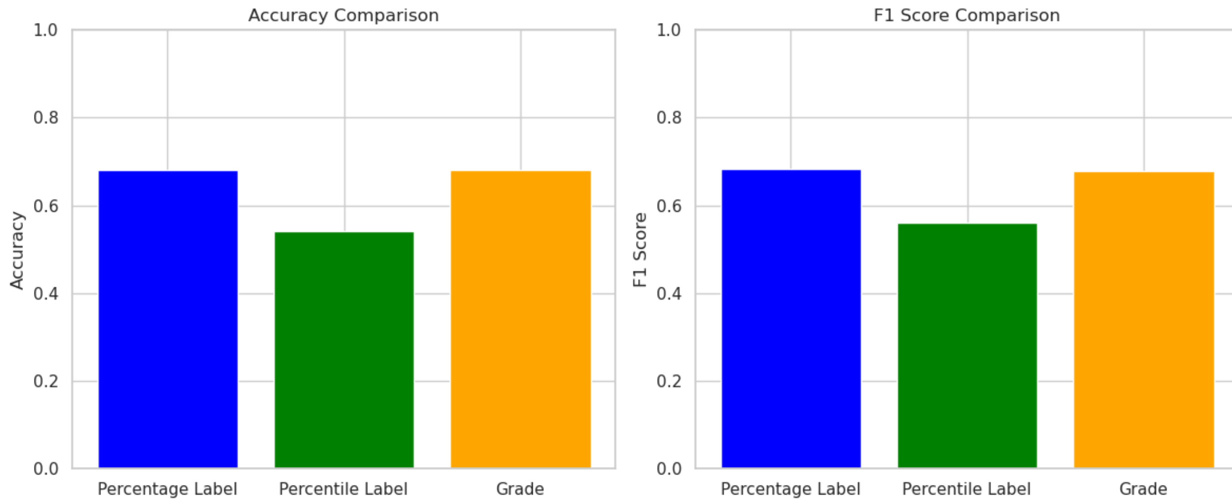

Figure 1: Distribution of students across Percentage Labels, Percentile Labels, and Grades. Different labeling methods create different distributions of students.

From this visualization, I observed that:

- Percentage Labels and Grades have many students in the failing category.
- Percentile Labels distribute students more evenly across categories.
- The same raw data produces different categorizations based on the metric used.

This difference in distribution is important because it affects how the machine learning model learns to classify students.

6 Machine Learning Experiment

6.1 Problem Setup

The key question in this experiment was: *Do different performance metrics and ranking methods affect how a machine learning model learns and predicts student performance?*

To answer this, I trained three separate Decision Tree models:

1. Model 1: Predicting Percentage Labels
2. Model 2: Predicting Percentile Labels

3. Model 3: Predicting Grades

All three models used the same input features: the four subject marks (Maths, Science, English, and History). The only difference was the target variable.

6.2 Why Use Subject Marks as Features?

It's important to use the **raw subject marks** as input features rather than derived metrics like percentage or rank. This is because:

- Subject marks are the **actual data** we collect.
- Percentage, percentile, and grades are **derived** from these marks.
- Using percentage to predict grade would create a circular relationship (since grade is directly calculated from percentage), leading to artificially perfect accuracy.
- Using raw marks allows the model to learn patterns about how performance in different subjects relates to overall categorization.

6.3 Train-Test Split

The dataset was split into training and testing sets using an 80-20 split:

- 80% of the data was used to train the model.
- 20% of the data was used to test the model's predictions.
- Stratification was used to ensure each label category was proportionally represented in both sets.

6.4 Model Selection

I chose a Decision Tree Classifier for this experiment because:

- It's easy to understand and interpret.
- It can handle non-linear relationships between features.
- It works well for multi-class classification problems.

7 Evaluation Metrics

Two metrics were used to evaluate model performance:

7.1 Accuracy

Accuracy measures the percentage of correct predictions:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{4}$$

7.2 F1 Score

The F1 score is the harmonic mean of precision and recall. It provides a balanced measure of the model's performance, especially when dealing with imbalanced classes. I used weighted F1 score, which accounts for the number of samples in each class.

8 Results

8.1 Model Performance Comparison

After training all three models, I compared their performance using accuracy and F1 score metrics.

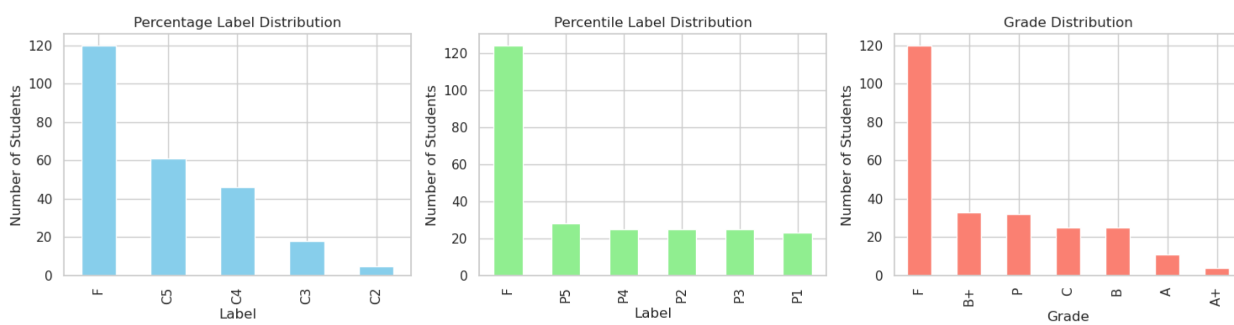

Figure 2: Comparison of model performance (Accuracy and F1 Score) when trained on different labeling schemes. The bars show how different target variables affect model performance.

From Figure 2, I observed that:

- Different performance metrics lead to different model accuracies.
- The F1 scores also vary based on which labeling scheme is used.
- The variation in performance shows that the choice of ranking method matters.

8.2 Analysis of Results

The differences in model performance can be explained by:

- **Class Distribution**: Different labeling schemes create different numbers of students in each category. This affects how well the model can learn patterns.
- **Category Granularity**: Grades have more categories (7 distinct grades) compared to Percentage Labels and Percentile Labels (6 categories each), which creates different classification challenges.
- **Threshold Sensitivity**: The cutoff points for each labeling scheme are different, so students near boundaries might be classified differently, affecting model learning.

9 Key Insights

Through this experiment, I learned several important lessons:

1. **Ranking methods matter**: Different ways of evaluating students (percentage vs percentile vs grades) create different categorizations, even from the same raw data.
2. **Label distribution affects learning**: The way we define performance categories influences how well a machine learning model can distinguish between them. Imbalanced distributions (many students failing) make classification harder.
3. **Same data, different insights**: Using the same subject marks but different target labels leads to different model behaviors, showing that the choice of evaluation metric is crucial.
4. **Absolute vs relative performance**: Percentage represents absolute performance (what you scored out of 400), while percentile represents relative performance (how you rank compared to others). Both perspectives are valuable but capture different aspects of student achievement.

10 Conclusion

This experiment successfully demonstrated how different performance metrics and ranking orders affect machine learning model behavior. By calculating percentage, percentile, and grades from student marks, then using these as different target variables, I observed that:

- The choice of performance metric significantly influences model learning and prediction patterns.
- Different ranking methods produce different student categorizations from identical raw data.
- Model performance varies based on the target variable used for training.

The experiment reinforced the importance of carefully choosing evaluation metrics and understanding that there are multiple valid ways to assess student performance, each with its own strengths and implications for analysis.