Mini Project II – Students Performance

Naimur Rahman

I. Introduction

The evaluation and maintenance of performance standards in organizations play a vital role in ensuring the ongoing excellence and sustainability of institutions. At the heart of this evaluation are the grades attained by students and their engagement in course-related activities, acting as fundamental indicators for assessing a student's performance in a particular course. Identifying students who are encountering difficulties in the course or are at risk of discontinuing their studies is essential. This enables instructors to offer the necessary support required to enhance these students' performance.

Furthermore, the analysis of activities provides insights that can be used to fine-tune the structure of the course, ensuring it remains effective and responsive to the evolving needs of the students.

To confront this challenge head-on, I have harnessed the power of data collected from student grades and course activities or features. This data serves as the foundation upon which machine learning algorithms are trained, which in turn, predict the overall performance of students. This predictive capability equips instructors and student tutors with the means to focus their efforts on students who are projected to attain lower grades, thus, proactively mitigating course dropout rates.

In addition to performance prediction, this project also aims to unravel the activities that wield the most influence on grading. Identifying these pivotal activities offers the potential to inform the equitable distribution of educational resources and curriculum emphasis, creating a more balanced and effective course structure.

In this project, two distinct machine learning models were harnessed to achieve our goals:

1) **Random Forest Classifier (RF):** This model is renowned for its capacity to handle complex relationships within data and provide robust predictions by aggregating the outputs of multiple decision trees.
2) **Decision Tree Classifier (DT):** Decision trees offer transparency in decision-making processes and are particularly effective in identifying key factors influencing outcomes.

Our implementation involved the creation of a well-defined pipeline encompassing the following steps:

- **Preprocessing:** Data preprocessing was undertaken to cleanse and prepare the data for subsequent analysis and modeling.
- **Feature Selection:** This step involved identifying and selecting the most relevant features, thereby reducing the dimensionality of the data and enhancing the models' efficiency.
- **Training:** The selected models, RF and DT, were trained using the preprocessed data to build predictive models.
- **Inference:** Following training, these models were deployed to make predictions on student performance.
- **Performance Evaluation:** Rigorous assessment of model performance was conducted to gauge their accuracy and effectiveness in predicting student grades.

II. Tools Utilized

To implement this multifaceted project, we relied on a suite of powerful and widely-recognized tools, including:

- **Scikit-learn:** An essential library for machine learning and data analysis, providing a wide array of tools for model building and evaluation.
- **Pandas:** This library was instrumental in data manipulation, data analysis, and structuring.
- **NumPy:** A fundamental library for numerical and matrix operations, contributing to data handling and analysis.
- **Seaborn:** Used for data visualization, Seaborn facilitated the graphical representation of insights and trends within the data.
- **Matplotlib:** This library enabled the creation of customizable plots and visualizations to convey our findings effectively.

Collectively, these tools and techniques were instrumental in our quest to assess student performance, predict outcomes, and improve the educational experience within our institution.

III. Data Description

In this dataset, we have gathered anonymized information from 107 students who participated in an educational course. The dataset offers valuable insights into student performance and behavior throughout the course. It comprises data on several aspects, including:

- **Grades:** The dataset contains data on student grades, encompassing three mini-projects, three quizzes, three peer reviews, and the final overall grade. These assessments are distributed across the course's timeline, with specific deadlines for mini-projects occurring in weeks 3, 5, and 8, and quiz deadlines in weeks 2, 4, and 8.
- **Course Logs:** The dataset provides a comprehensive log of student interactions with course materials, classified into four categories:
  - **Status 0 (course/lectures/content-related):** This category records activities related to course modules, content viewing, lesson engagement (start, resume, restart, end), and other course-related interactions.

  - **Status 1 (assignment-related):** These logs capture students' interactions with assignments, quizzes, questions, and submissions. Activities such as quiz attempts, submissions, assessments, and reviews are documented here.
  - **Status 2 (grade-related):** This category includes records related to the grading process, such as user grade reports, grade overview reports, grade deletions, user profile views, and other grading-related activities.
  - **Status 3 (forum-related):** Logs in this category track interactions within course forums, including posts, discussions, content postings, and views.
- **Grades Data:** The dataset contains information on student performance in the form of nine grades, covering various assessments and assignments, including Week2_Quiz1, Week3_MP1, and so forth. These grades offer insights into student achievements and progress within the course.
- **Logs Data:** The dataset also includes detailed records of student activity logs, encompassing 36 unique logs for each week. These logs are labelled as Week1_Stat0, Week1_Stat1, Week1_Stat2, Week1_Stat3, and so on, up to Week9. They provide a comprehensive account of student interactions and engagement with course content, assignments, grading activities, and forum discussions, among other aspects.

In summary, this dataset serves as a valuable resource for researchers and educators interested in gaining a nuanced understanding of student engagement, assessment outcomes, and course interaction patterns within an educational context. It offers a wealth of data for in-depth analysis and insights into student behaviour and performance throughout the course.

IV. Data Analysis and Processing

Before utilizing the data for model training and prediction, I undertook a data preprocessing phase. The dataset, initially in CSV format, was imported from a local source. Most of the dataset primarily consisted of integer and float data types, eliminating the need for any further numerical conversion. I diligently examined the dataset for any missing values, and fortunately, no null values were identified. Additionally, I ensured the absence of duplicate records, maintaining data integrity.

In the context of feature selection, I recognized that not all features are equally crucial for effective model training. Achieving a well-performing model necessitates the identification of features that offer the most pertinent information for model learning. The process of pinpointing these significant features and reducing the feature set is vital, as it diminishes the influence of extraneous variables and enhances model performance. Therefore, our approach involved selecting features with a strong correlation to the target variable. Specifically, I retained features with a correlation score greater than or equal to 0.50 and less than or equal to -0.50. This selective approach yielded the top 6 pertinent features concerning the "Grade" column, thereby streamlining our dataset for optimal model training and prediction (fig. 1).

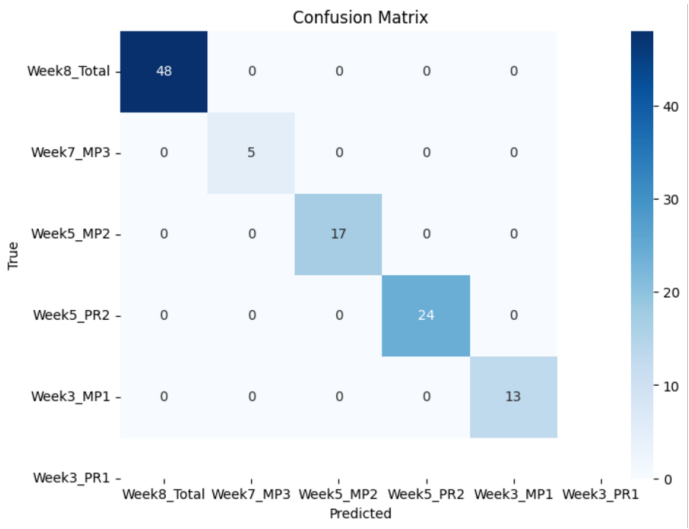

Fig. 1: Confusion matrix for top 6 features

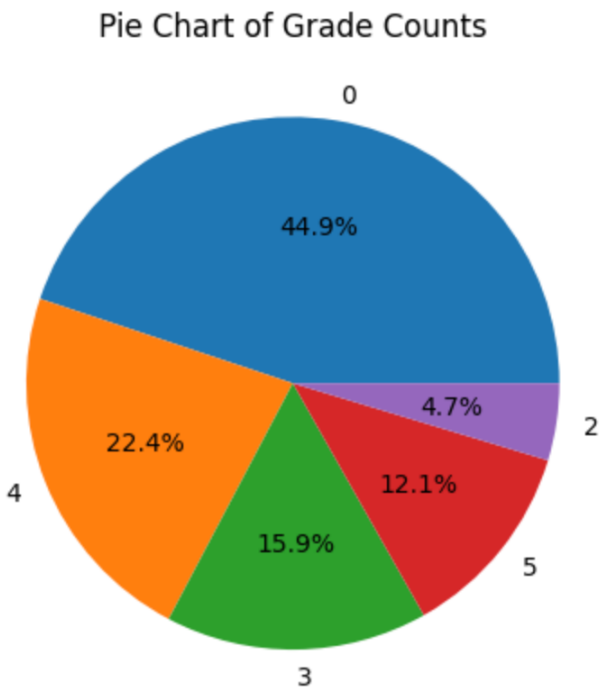

Fig. 2: Pie chart to show the percentage for individual grades

One interesting thing I have got is the dataset is imbalanced. However, I do not have enough records, I decided to not truncate or add anymore synthetic data, as the model might be overfitted (fig. 2).

So after doing all the pre-processing, the features that remained are given in fig 3.

V. Model Training

The dataset was split into an 80:20 ratio, leading to the creation of a training dataset containing 91 rows and a test dataset containing 16 rows. I employed the chosen essential features to train both the Random Forest classifier and the Decision Tree Classifier. To ensure consistency in our results,

Grade	1.000000
Week8_Total	0.972348
Week7_MP3	0.968130
Week5_MP2	0.953488
Week5_PR2	0.907837
Week3_MP1	0.901788
Week3_PR1	0.887352
Week7_PR3	0.865616
Week6_Quiz3	0.849920
Week4_Quiz2	0.810920
Week6_Stat1	0.771988
Week2_Quiz1	0.689783
Week4_Stat1	0.662946
Week3_Stat0	0.643789
Week6_Stat0	0.635807
Week4_Stat0	0.625359
Week3_Stat1	0.596824
Week5_Stat0	0.590146
Week8_Stat1	0.584425
Week9_Stat0	0.545532
Week9_Stat1	0.496753
Week5_Stat1	0.484030
Week8_Stat0	0.450807
Week7_Stat3	0.439733
Week7_Stat1	0.424807
Week2_Stat1	0.406120

Fig. 3: Filtered relevant features for Grades

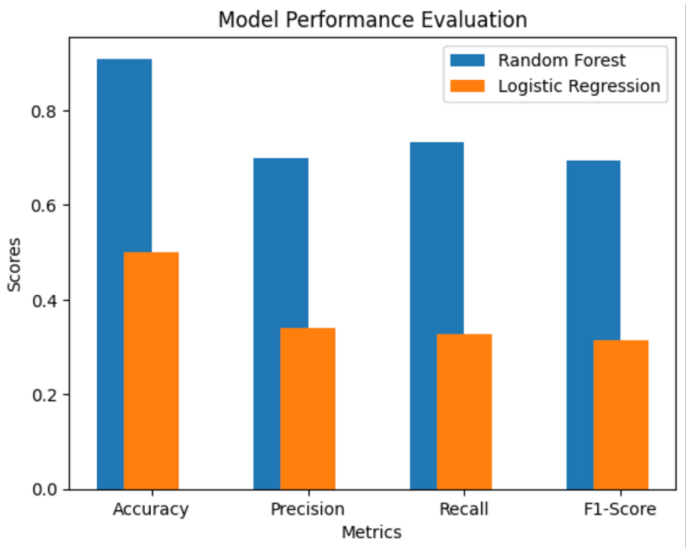

Fig. 4: Random Forest and Decision Tree classifier for different metrics

	precision	recall	f1-score	support
0	1.00	1.00	1.00	10
2	0.00	0.00	0.00	1
3	0.50	1.00	0.67	2
4	1.00	1.00	1.00	6
5	1.00	0.67	0.80	3
accuracy			0.91	22
macro avg	0.70	0.73	0.69	22
weighted avg	0.91	0.91	0.90	22

Fig. 5: Different metrics for the prediction with Random Forest

a specific random state was set. Subsequently, the two trained models were utilized to predict grades for the test dataset, and these predictions were stored in the test data frame for further analysis.

VI. Performance Evaluation

Performance evaluation is a very crucial part for data analysis. In every section, Random Forest gave better accuracy, which can be seen in fig. 4.

Hence, I recommend the adoption of the Random Forest model for this specific dataset. In-depth details regarding the metrics for the Random Forest model are provided in fig. 5.

Top 3 features those were mostly responsible for giving the predictions are given in fig. 6.

VII. Conclusion

In conclusion, the project revolves around the critical task of assessing and maintaining performance benchmarks within educational organizations. The primary metrics for this assessment are students' grades and their engagement in course activities. Identifying students at risk of underperformance and dropout and fine-tuning the course structure are central objectives.

To address this challenge, machine learning algorithms were employed, utilizing data on student grades and course activities. This predictive capability enables instructors to proactively support struggling students and reduce course dropout rates. Additionally, the project aimed to identify activities with the most influence on grading, optimizing the course structure accordingly.

Despite the project's success, certain bottlenecks were encountered. Notably, the dataset's size was limited, comprising data for 107 students, which may not be sufficient for creating a universally applicable model. The dataset also suffered from imbalance issues, potentially affecting model performance.

	Random Forest	Decision Tree
0	Week8_Total	Week8_Total
1	Week7_MP3	Week7_MP3
2	Week5_MP2	Week5_MP2

Fig. 6: Three top-most essential features

Furthermore, it was discerned that not all features contributed equally to the predictive power, with some features exhibiting high correlations with outcomes, while others played a more peripheral role.

To mitigate these challenges and enhance the project's utility, several strategies can be employed. Collecting a larger dataset with diverse student profiles could lead to more generalized and robust models. Addressing class imbalance through data augmentation or resampling techniques would improve model performance. Additionally, feature selection techniques can be utilized to identify and prioritize essential features for model training, potentially increasing predictive accuracy.

In sum, this project offers a foundation for leveraging machine learning to enhance educational outcomes. By addressing the bottlenecks and refining the methodology, it has the potential to contribute significantly to improving student performance and the overall educational experience within our institution.