Predicting Students' Final Grades in Machine Learning Course: A Comparative Study of Supervised Learning Approaches

Gayani Liyanage

1. Introduction

This project aims to use different supervised learning approaches which are capable of predicting student grades using Moodle data collected from a nine-week online machine-learning course hosted on the Moodle learning management system. The dataset, while high-dimensional with 48 features, presents a unique challenge due to its relatively small sample size, consisting of data from 107 students. The grading scale ranges from 0 to 5, making this a categorical prediction task. The dataset includes nine grades and 36 logs, each representing various activities conducted by students on the Moodle website, such as content page views, submission updates, and discussion creation. All data in the dataset is numerical, with floating-point grades and integer-based logs. This project aims to use supervised learning algorithms to build a predictive model that will facilitate the categorization of student performance in the course.

2. Data processing

In the initial stages of preparing the dataset for machine learning, several critical data processing steps were undertaken to enhance the dataset's suitability for modelling.

First and foremost, the unique ID column associated with each dataset entry was non-essential for the machine-learning model. Consequently, the ID column was removed, simplifying the dataset while retaining its core information.

To ensure data completeness and reliability, an evaluation was carried out to identify any null values within the dataset. Fortunately, this revealed an absence of missing values, validating the dataset's integrity and suitability for subsequent analysis.

An evaluation was undertaken to identify columns where all values remained uniform. Such columns, characterized by constant values, contribute negligible information to a machine-learning model. As a result, the "Week1_Stat1" column, found to contain uniform values, was logically excluded from the dataset. This decision was crucial, as it guaranteed that the dataset was simplified and free from unnecessary attributes.

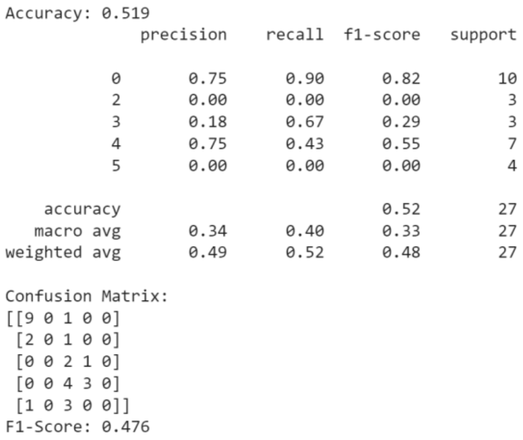
Furthermore, when selecting the target variable for the machine-learning model, the dataset included two sets: "Week8_Total" and "Grade."However, to prevent the model from relying solely on "Week8_Total" for predicting "Grade," it was decided to retain "Grade" as the sole target variable while discarding "Week8_Total."

3. Data analysis

The data was split into two subsets, allocating 75 percentage to the training set and 25 percentage to the testing set. Moreover, features were stored as X and the labels were stored as y.

For this project, three distinct models were used: K-nearest neighbours (KNN), Random Forest, and Decision Tree.

The following results represent the final accuracy outcomes:



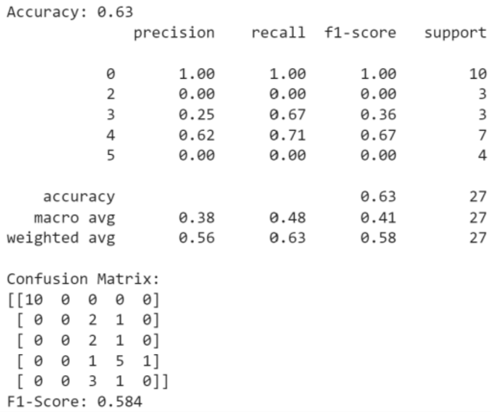
```
Accuracy: 0.519
              precision    recall  f1-score   support

           0       0.75      0.90      0.82        10
           2       0.00      0.00      0.00         3
           3       0.18      0.67      0.29         3
           4       0.75      0.43      0.55         7
           5       0.00      0.00      0.00         4

    accuracy                           0.52        27
   macro avg       0.34      0.40      0.33        27
weighted avg       0.49      0.52      0.48        27

Confusion Matrix:
[[9 0 1 0 0]
 [2 0 1 0 0]
 [0 0 2 1 0]
 [0 0 4 3 0]
 [1 0 3 0 0]]
F1-Score: 0.476
```

Figure 1 – KNN



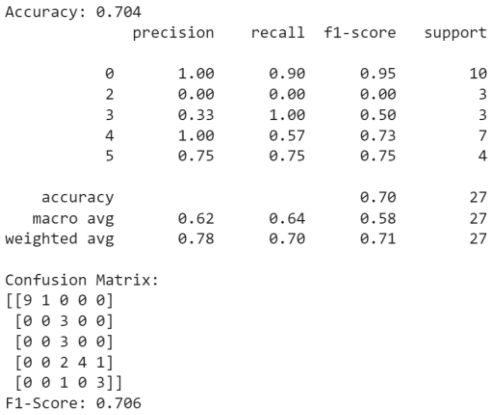
```
Accuracy: 0.63
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        10
           2       0.00      0.00      0.00         3
           3       0.25      0.67      0.36         3
           4       0.62      0.71      0.67         7
           5       0.00      0.00      0.00         4

    accuracy                           0.63        27
   macro avg       0.38      0.48      0.41        27
weighted avg       0.56      0.63      0.58        27

Confusion Matrix:
[[10  0  0  0  0]
 [ 0  0  2  1  0]
 [ 0  0  2  1  0]
 [ 0  0  1  5  1]
 [ 0  0  3  1  0]]
F1-Score: 0.584
```

Figure 2 - Random Forest



```
Accuracy: 0.704
              precision    recall  f1-score   support

           0       1.00      0.90      0.95        10
           2       0.00      0.00      0.00         3
           3       0.33      1.00      0.50         3
           4       1.00      0.57      0.73         7
           5       0.75      0.75      0.75         4

    accuracy                           0.70        27
   macro avg       0.62      0.64      0.58        27
weighted avg       0.78      0.70      0.71        27

Confusion Matrix:
[[9 1 0 0 0]
 [0 0 3 0 0]
 [0 0 3 0 0]
 [0 0 2 4 1]
 [0 0 1 0 3]]
F1-Score: 0.706
```

Figure 3 - Decision tree

When undertaking a comprehensive examination of the dataset, placing significant emphasis on performance indicators, including accuracy, F1-score, and the elucidating confusion matrix for each of the three models: Decision Tree, Random Forest, and K-Nearest Neighbors (KNN).

1. Decision Tree:
   - Accuracy: 0.704
   - F1-Score: 0.706

In the case of the Decision Tree model, it achieves a high accuracy of 70.4%. The F1-Score, which is a measure of the model's balance between precision and recall, is also commendable at 0.706. The precision, recall, and F1-score values for each class indicate that the model performs particularly well for class 0, achieving a high precision and recall.

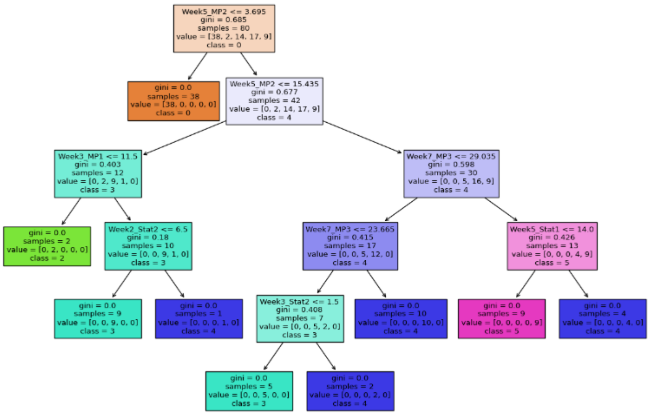


Figure 4 - Plot of Decision tree

The plot of a decision tree typically illustrates the structure and decision-making process of the tree-based model. This plot is a visual representation of how the model makes decisions by branching based on specific features.

2. Random Forest:
   - Accuracy: 0.63
   - F1-Score: 0.584

The Random Forest model exhibits an accuracy of 63%, which is still reasonable but slightly lower compared to the Decision Tree. The F1-Score, at 0.584, reflects a balanced performance in terms of precision and recall. However, it's worth noting that Random Forest may be overfitting the data to some extent.

3. K-Nearest Neighbors (KNN):
    - Accuracy: 0.519
    - F1-Score: 0.476

The KNN model records an accuracy of 51.9%, which is the lowest among the three models. The F1-Score, at 0.476, indicates that this model might not be as effective in achieving a balance between precision and recall for the classes. It appears that the KNN model struggles with certain classes, particularly class 2 and class 5.

The confusion matrices provide additional insights into model performance by showing the true positives, true negatives, false positives, and false negatives for each class. The interpretation can be quite detailed, but it essentially helps identify where the models are making correct and incorrect predictions for each class.

In summary, the Decision Tree model stands out with the highest accuracy and F1-Score, demonstrating strong performance across most classes. The Random Forest model follows, offering a good balance between precision and recall. The KNN model, while still showing potential, may require further optimization to improve accuracy and F1-Score, especially for certain classes.

To evaluate the performance of a Random Forest classifier by varying the number of trees in the ensemble and visualizing how the accuracy of the model changes with different numbers of trees. This process is essential for assessing the impact of the number of trees on the model's performance and helps determine if any changes are needed to improve model accuracy.

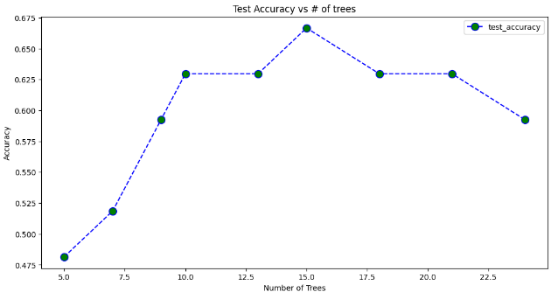


Figure 5 - Random Forest No of Trees

The graph provides a clear representation of how the accuracy of the Random Forest model is influenced by the number of trees. This visualization serves as a valuable aid for model refinement and gaining insights into its behaviour across various setups. Moreover, the graph indicates that 15 trees emerge as the optimal choice, delivering peak accuracy.

As a result, the Python program for the model can be modified to set the `n_estimators` parameter to 15. This adjustment aligns with the optimal number of trees indicated by the graph, further enhancing the model's performance.

The significance of features in predicting students' final grades has been analyzed, revealing that the top three most crucial features are "Week7_MP3"," Week5_MP2", and "Week4_Stat0". This outcome aligns with expectations, as Mini Project 2 and Mini Project 3 are assignments with the greatest impact on the total grade. Additionally, the third most important feature, "Week4_Stat0", underscores that students who engage in reviewing course materials and attending lectures during week 4 exhibit a positive correlation with their "Week5_MP2" grades, rendering it another pivotal feature.

To provide a visual representation of these findings, the following figure showcases the most important features, shedding light on their relative importance in predicting students' final grades.

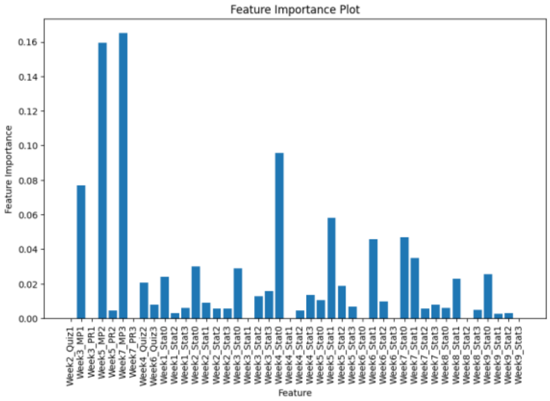


Figure 6 - Feature importance

4. Conclusion

The implementation of the models and the in-depth examination of diverse supervised machine-learning approaches have clarified several crucial insights. The project work emphasizes that a decision tree is the best model for this case. In datasets with numerous features, some less important features can negatively impact the final model when used in decision tree construction. Therefore, there arises an opportunity for further investigation to explore feature combination techniques, such as dimension reduction through aggregation methods like summing log values for a week or computing average log values for distinct log types within each week.

In summary, this study serves as a valuable contribution, offering a comprehensive analysis and the application of three distinct supervised machine learning models, each yielding varying levels of accuracy.

Additionally, it highlights that the problem remains open-ended, offering opportunities for future research to explore the intricate aspects of predictive modelling in the context of predicting student performance.