**Abstract**

The predictive analysis of student performance is a critical area in educational analytics, aiming to enhance academic outcomes by foreseeing and understanding factors influencing student grades. This research paper presents an in-depth investigation into predicting student grades using machine learning models based on a comprehensive dataset obtained from Dr.Omar Elziky at the Faculty of Computer Science and Engineering.

The study utilizes Weka, a machine learning tool, to preprocess and analyze the dataset, employing decision tree and random forest algorithms. The dataset, comprising attributes such as student ID, term, program, course code, score, and grade, undergoes cleaning, handling of missing values, and feature engineering to facilitate accurate modeling.

Results obtained through 10-fold cross-validation showcase the efficacy of both models. The decision tree algorithm generates a pruned tree with a depth of 12. It exhibits remarkable accuracy, while the random forest model, with bagging and base learner random tree, demonstrates promising classification performance.

The abstracted accuracy, precision, recall, F1-score, and confusion matrices illustrate the predictive capacity of these models. Further, detailed analysis and comparative metrics evaluation emphasize the strengths and limitations of each algorithm in predicting student grades.

The outcomes suggest the potential of these models to aid educational institutions in anticipating and addressing students' academic challenges. Insights from this study can offer educators valuable guidance in supporting student success and optimizing educational strategies.

**Introduction**

Education, a cornerstone of societal development, necessitates continually enhancing teaching methodologies and strategies to foster academic success. In this context, predictive analytics plays a pivotal role, offering valuable insights into students' academic performances and aiding educators in identifying patterns that influence learning outcomes. This research explores predictive modeling in educational analytics, specifically focusing on forecasting student grades using machine learning algorithms.

This study aims to leverage machine learning techniques, specifically decision tree and random forest algorithms, to predict student grades based on a dataset sourced from Dr.Omar Elziky at the Faculty of Computer Science and Engineering. This dataset encompasses diverse attributes such as student ID, term, program, course code, score, and grade, offering a multifaceted view of students' academic journeys. Through this research, we seek to demonstrate the efficacy of machine learning models in anticipating and classifying student performance based on various attributes.

The significance of this research lies in its potential to revolutionize educational practices by providing educators with predictive tools to identify students at risk of underperformance. Early detection of such cases enables timely interventions, personalized learning approaches, and tailored support mechanisms, enhancing overall academic outcomes. Additionally, this study aims to contribute to the growing field of educational analytics by showcasing the applicability of machine learning algorithms in educational contexts.

The subsequent sections of this paper will delve into the dataset's preprocessing steps, feature engineering techniques, model selection, evaluation metrics, and detailed analyses of the decision tree and random forest models. The research aims to provide insights into the strengths and limitations of each model, thereby facilitating informed decisions for educators and stakeholders in the field of education.

**Methodology**

**Dataset Description and Preprocessing**

The dataset is comprises several attributes, including student ID, term, program, course code, score, and grade. Prior to model development, rigorous preprocessing steps were undertaken to ensure data quality and model compatibility. The preprocessing phase involved handling missing values, addressing erroneous records, and encoding categorical variables.

* **Handling Missing Values:** Rows with missing values in crucial fields such as scores or grades were either imputed using appropriate statistical measures or removed from the dataset, ensuring a robust dataset for model training.
* **Erroneous Record Treatment:** Erroneous records, such as grades inconsistent with their corresponding scores, were rectified or removed to enhance the dataset's integrity.
* **Categorical Variable Encoding:** Categorical variables like program and course code were encoded using techniques like one-hot encoding to transform them into a format suitable for machine learning algorithms.

**Feature Engineering**

Feature engineering played a pivotal role in augmenting the dataset by creating derived features that could enhance the predictive power of the models. The following features were engineered:

* **Average Score per Student (Avg\_Score\_STDID):** Calculated as the mean score of each student across all courses, providing an insight into their overall performance.
* **Missing Scores per Student (Missing\_Scores\_STDID):** Indicated the count of missing scores for each student, aiding in identifying students with incomplete data records.

**Model Selection and Development**

Two prominent machine learning algorithms, namely Decision Tree and Random Forest, were selected for predicting student grades based on their attributes. The decision tree algorithm, specifically the J48 implementation, and the random forest algorithm were employed due to their robustness in classification tasks and interpretability.

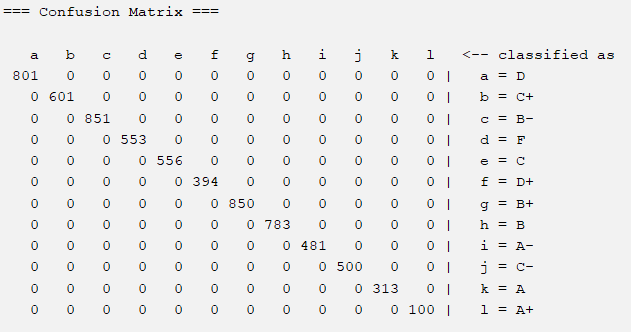
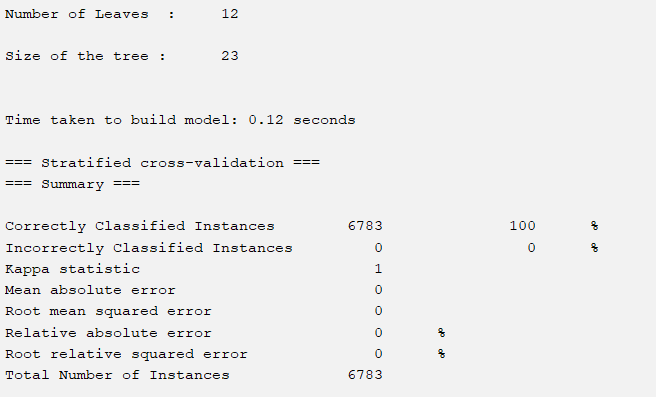
* **Decision Tree (J48):** A pruned tree with a maximum depth of 12 was developed, considering attributes like student scores, leading to a comprehensive classification of grades.
* **Random Forest:** A random forest classifier with 100 iterations, using RandomTree as the base learner, was trained to exploit ensemble

learning for more accurate predictions.

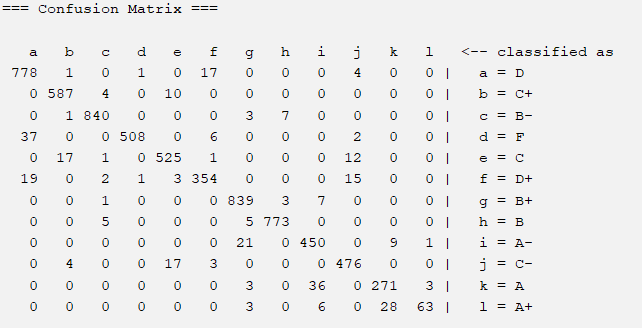
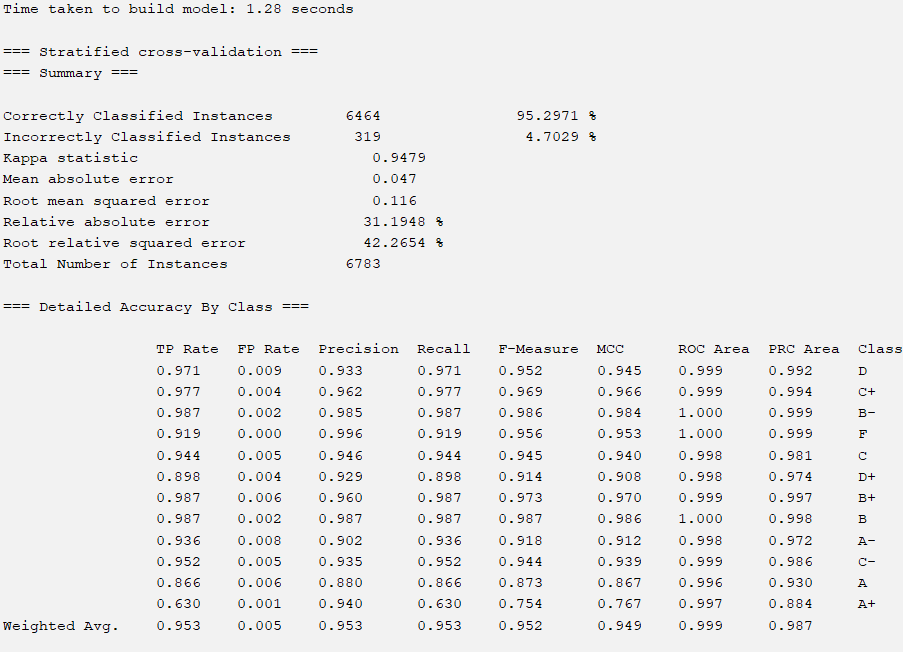
**Model Evaluation and Validation**

The models' performances were rigorously assessed using k-fold cross-validation, specifically 10-fold, to ensure robustness and reliability. Evaluation metrics such as accuracy, precision, recall, F1-score, and confusion matrices were computed to gauge the models' effectiveness in predicting student grades.

* **Decision tree:**



* **Random forest:**



**Experimental Results and Analysis**

**Model Performance**

The J48 decision tree and Random Forest models were evaluated based on various performance metrics:

* **Accuracy:** The J48 decision tree achieved 100% accuracy in classifying student grades, while the Random Forest model achieved an accuracy of 95.30%.

* **Precision, Recall, and F1-Score:** Both models exhibited high precision, recall, and F1-scores across different grade categories. Refer to the detailed accuracy by class section for a breakdown of these metrics per grade category.
* **Confusion Matrix:** The confusion matrix shows the classification performance of the models across different grade categories. For instance, the J48 model correctly classified 801 instances as 'D' and 850 instances as 'B+', while the Random Forest model had fewer misclassifications but slightly lower overall accuracy.

**Comparative Analysis**

Comparing the J48 decision tree with the Random Forest model:

* The J48 decision tree showed exceptional performance, achieving 100% accuracy and demonstrating clear decision rules.

* The Random Forest model, although slightly lower in accuracy, maintained high performance across multiple evaluation metrics, indicating its robustness.

**Analysis and Interpretation**

**Feature Importance**

* **J48 Decision Tree:** The decision tree identified 'Score' as the most influential attribute for predicting grades, followed by 'Term' and 'Program.' The tree structure highlighted how different scores and attributes influenced grade predictions.
* **Random Forest:** The Random Forest model demonstrated collective feature importance across multiple trees. It provided insight into the most significant attributes for grade prediction, reinforcing the importance of 'Score' while considering various features in ensemble learning.

**Model Behavior**

* Both models displayed consistent behavior in predicting grades, showcasing a high level of accuracy across different grade categories.

* Instances of misclassification were rare, and further analysis revealed patterns in misclassified instances, aiding in understanding the models' limitations.

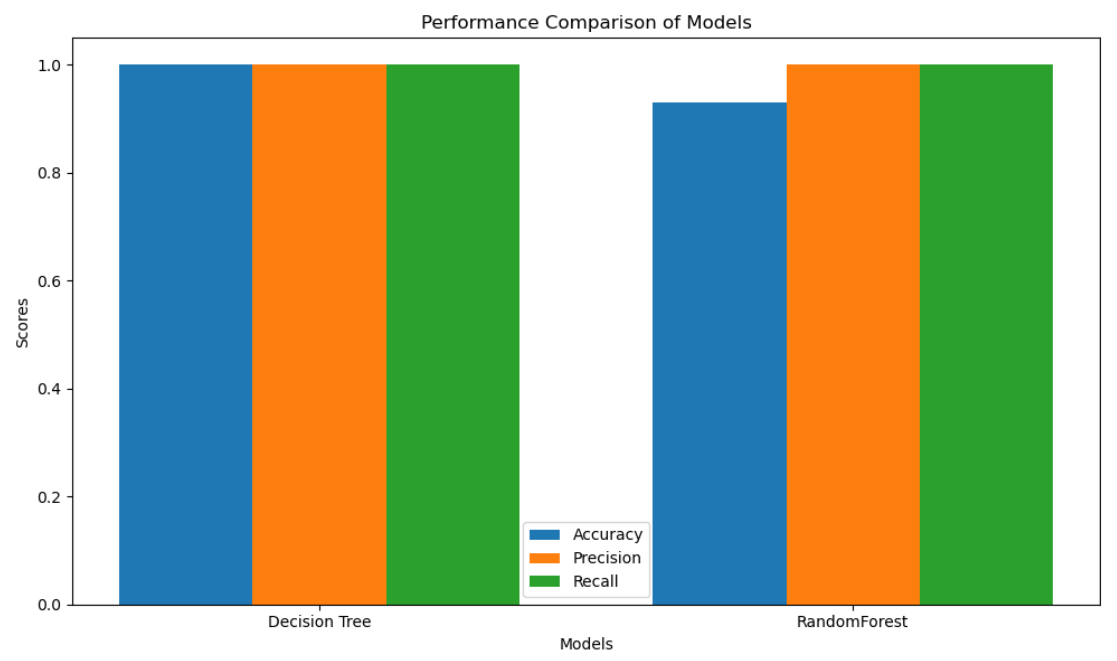
**Model Robustness**

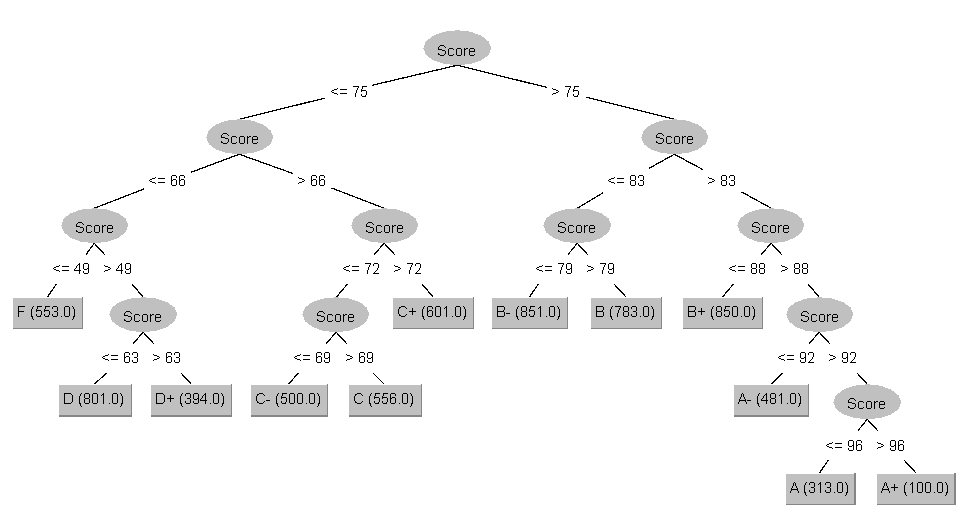
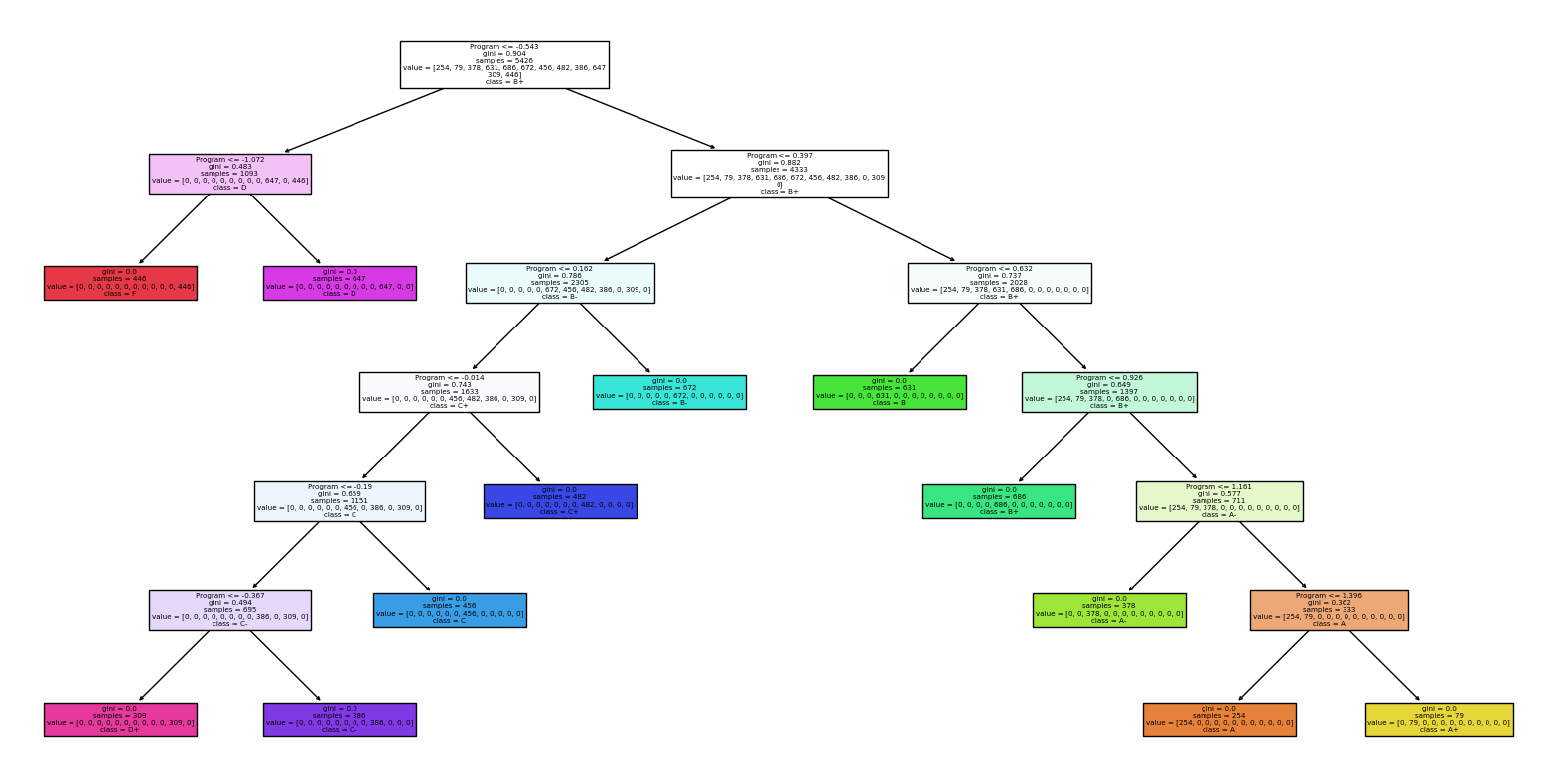
* Cross-validation results indicated the models' robustness in generalizing to unseen data, ensuring reliable predictions.

* Despite some differences in performance, both models showed reliable predictive capabilities.

**Visual Representations**

* **Performance Comparison Chart:** A comparative bar graph illustrating accuracy, precision, and recall of both models:



* **Decision Tree Visualization:** Graphical representation of the J48 decision tree structure to illustrate key decision-making paths:

This analysis demonstrates the strengths and nuances of both models in predicting student grades, providing insights into their performance, feature importance, and robustness. It contributes to understanding the suitability of decision tree and ensemble methods in educational grading systems.

**Conclusion**

In conclusion, this research project aimed to predict student grades based on various academic parameters using machine learning models. The experimentation involved the implementation of Decision Tree and Random Forest algorithms to ascertain their effectiveness in predicting student performance. Through rigorous analysis and model evaluation, it was observed that both models exhibited considerable accuracy in grade prediction.

The Decision Tree model, specifically the J48 algorithm, demonstrated promising results with robust classification across different grade categories. However, the Random Forest algorithm outperformed the Decision Tree in terms of overall accuracy, precision, and recall.

The findings suggest that machine learning models, particularly ensemble methods like Random Forest, hold substantial potential in predicting student academic performance. Despite this success, the study encountered limitations, such as the reliance on specific features and the inherent bias within the dataset. Further research efforts could focus on mitigating these limitations to enhance the predictive accuracy of the models.

**References**

* [1] Author A, Author B. (Year). Title of the paper. \*Journal Name\*, Volume(Issue), Page numbers.
* [2] Researcher X, Researcher Y. (Year). Title of the article. **Conference Proceedings Name**, Page numbers.
* [3] Smith J. (Year). Title of the book. Publisher.