**Technical Report: Student Graduation Prediction Using Random Forest Classifier**

**1. Introduction**

The goal of this project was to predict whether a student would graduate based on a variety of academic and demographic features. By building a predictive model, educational institutions can proactively identify students at risk of not graduating and offer necessary support to improve outcomes.

The project utilized a Random Forest Classifier, followed by feature selection to optimize model performance. The technical practices adhered to ensure a robust, well-performing model.

**2. Data Collection**

The dataset used in this project was sourced from provided by the organizers, containing information about students’ academic performance, personal background, and other key factors related to their educational journey. The dataset included both categorical and numerical features.

**3. Data Preprocessing**

Data preprocessing steps were applied to clean and prepare the dataset for modeling:

* **Handling Missing Values**: There were no missing values in the dataset.
* **Scaling**: The dataset was scaled using StandardScaler to ensure all features were on the same scale, which is especially important for machine learning algorithms that rely on distance measures.

**4. Initial Model Training with All Features**

The Random Forest Classifier was initially trained on the full feature set to establish a baseline model performance. The Random Forest algorithm was chosen due to its ability to handle both numerical and categorical data and its robustness in managing feature importance.

* **Algorithm Used**: Random Forest Classifier
* **Model Configuration**:
  + Number of estimators: 100
  + Criterion: Gini index
  + Random state: 42 (for reproducibility)

The initial performance of the model was evaluated using various metrics like accuracy, precision, recall, and F1-score.

**5. Feature Selection**

To optimize the model, feature importance scores were generated from the Random Forest model. Based on these scores, the top 10 most important features were selected to train the model again. This step was undertaken to reduce dimensionality, improve model efficiency, and eliminate noise from irrelevant features.

* **Top 10 Features Selected**:
  1. Course
  2. previous qualification (grade)
  3. Admission grade
  4. Tuition fees up to date
  5. Curricular units 1st sem(evaluations)
  6. Curricular units 1st sem (grade)
  7. curricular units 1st sem (approved)
  8. curricular units 2nd sem (grade)
  9. curricular units 2nd sem (approved)
  10. curricular units 2nd sem (evaluations)

**6. Final Model Training**

After feature selection, the Random Forest Classifier was retrained on the top 10 features. The same hyperparameters were retained, and the model’s performance was evaluated and compared to the initial model.

* **Hyperparameters**: Same as the initial model
* **Data Scaling**: The scaled dataset was used to ensure consistency in feature importance.

**7. Model Evaluation**

Both the initial model (trained on all features) and the optimized model (trained on the top 10 features) were evaluated based on the following metrics:

* **Accuracy**: The proportion of correct predictions out of the total predictions.
* **Precision**: The proportion of true positive predictions out of all positive predictions.
* **Recall**: The proportion of true positives out of all actual positives.
* **F1-Score**: The harmonic mean of precision and recall.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Accuracy | Precision | Recall | F1\_score |
| All features | 0.84 | 0.84 | 0.84 | 0.84 |
| Top 10 features | 0.83 | 0.83 | 0.83 | 0.83 |

**8. Conclusion**

This project successfully developed a predictive model to determine whether a student will graduate or not. Feature selection led to improved model efficiency without sacrificing accuracy. By focusing on the most important features, the model offers a streamlined approach that can be easily interpreted by educators and stakeholders.

Future improvements could include hyperparameter tuning using grid search or more advanced techniques, experimenting with other classification algorithms, and adding more features related to student behavior or social factors.