Abstract

This report evaluates the performance of three machine learning models- Random Forest, Decision Tree, and Logistic Regression- on three datasets from the UCI Machine Learning Repository: Student Performance, Car Evaluation, Obesity Levels. Each dataset presents unique challenges due to its structure and feature types, ranging from entirely categorical data to mixed numerical and categorical data/ The study employs preprocessing techniques, hyperparameter tuning, and cross-validation to optimize model performance. Multiple train-test splits (80/20, 50/50, 20/80) are used to assess generalization capabilities. Results indicate that Random Forest outperformed other models in most cases, while Logistic Regression showed stable but lower accuracy. Decision Tree provided moderate accuracy with sensitivity to hyperparameters.

Introduction

Machine learning models are widely used for classification tasks in various domains, but their performance can vary significantly depending on the dataset's characteristics. This study evaluated a subset of these models on diverse datasets to understand their generalization capabilities and robustness. This study focuses on three datasets with distinct characteristics:

1. Student Performance: A mixed dataset with both categorical and numerical features predicting student grades
2. Car Evaluation: A categorical dataset for classifying car acceptability based on features like buying price, maintenance cost, and safety
3. Obesity Levels Dataset: A mixed dataset predicting obesity levels based on eating habits and physical condition

The goal is to compare the performance of Random Forest, Decision Tree, and Logistic Regression models across these datasets. We aim to analyze how data characteristics and train-test splits influence model accuracy.

Method

Data Preprocessing

For all the datasets, categorical variables were one-hot encoded and numerical variables were standardized with "StandardScaler", whenever relevant. Pipelines were used to streamline preprocessing and model training.

Models

Three models were evaluated and compared, beginning with Random Forest, a tree-based ensemble model known for its robustness and ability to handle categorical data. The next model that was evaluated was Decision Tree, another tree-based model that has the drawback of overfitting regularly. The final model was the logistic regression, a linear model suitable for binary or multiclass classification.

Hyperparameter Tuning

To find the best hyperparameters, GridSearchCV was used. The hyperparameters that were tuned for the Random Forest model were ‘n_estimators’ and ‘max_depth’. For the Decision Tree model ‘min_samples_split’ along with ‘max_depth’ were optimized. For the Logistic Regression model, only regularization strength (‘C’) was optimized.

Evaluation Metrics

The primary evaluation metric was accuracy. Models were trained and tested on three train-test splits (80/20, 50/50, 20/80) in order to analyze how different amounts of training data would impact model performance.

Experiment

Student Performance Dataset

This dataset includes numerical and categorical features. The label is the G3 (Final grade) feature, making this a classification problem. The Random Forest model outperformed the other models by effectively handling mixed data types and capturing complex interactions between features, while the Decision Tree model provided reasonable accuracy but was sensitive to hyperparameter settings such as maximum depth. The Logistic Regression model underperformed due to the dataset’s non-linear relationships, which it could not model effectively.

Car Evaluation Dataset

This dataset is entirely composed of categorical variables (buying price, maintenance cost, safety cost, etc.) This dataset is appropriate for tree-based models like Random Forest and Decision Tree due to their ability to handle categorical variables. The Random Forest model achieved high accuracy across all splits due to its ensemble nature and ability to capture complex relationships in categorical data. The Decision Tree model performed moderately well but showed signs of overfitting with smaller training sets. Expectedly, the Logistic Regression model struggled due to its linear nature, which bottlenecked the capability to capture non-linear relationships that were in the dataset.

Obesity Levels Dataset

This dataset predicts obesity levels based on eating habits, physical activity, and demographic information. It contains a combination of numerical and categorical variables. The Random Forest demonstrated strong performance across all splits, leveraging its ensemble nature to handle both numerical and categorical features effectively. The Decision Tree performed reasonably well but overfitted with larger training sets. Characteristically, the Logistic Regression delivered stable but lower accuracy compared to tree-based models due to its linear assumptions.

Results

Train 80/Test 20 Split

In this partition, Random Forest achieved the highest accuracy across all datasets due to its ability to leverage a large training set effectively. The Decision Tree showed moderate performance but was prone to overfitting with deeper trees. Logistic Regression demonstrated stable but lower accuracy due to its linear assumptions.

Train 50/Test 50 Split

With an equal split between training and test data, Random Forest maintained its superior performance across all datasets. Decision Tree's performance remained consistent but slightly declined compared to the 80/20 split. Logistic Regression's accuracy remained stable but continued to lag behind tree-based models.

Train 20/Test 80 Split

In this partition, Random Forest still outperformed other models but showed a slight decline in accuracy due to limited training data. The Decision Tree exhibited noticeable overfitting tendencies with smaller training sets. Logistic Regression's performance remained consistent but low across all datasets.

Conclusion

The analysis of model performance across different datasets reveals several key insights. Random Forest consistently demonstrates superior accuracy, leveraging its ensemble approach to effectively handle both categorical and mixed data types. This robustness is evident in its ability to generalize well across various train-test splits, maintaining high accuracy even with limited training data. The model's strength lies in its capacity to capture complex patterns and interactions within the data, making it a versatile choice for diverse datasets.

Decision Tree models offer moderate performance, with their simplicity and interpretability being significant advantages. However, they exhibit sensitivity to hyperparameters, particularly the maximum depth, which can lead to overfitting when not carefully tuned. Despite these challenges, Decision Trees provide a reasonable balance between complexity and accuracy, performing adequately across different data splits.

Logistic Regression, while stable, generally underperforms compared to tree-based models. Its linear nature limits its ability to model non-linear relationships and complex feature interactions present in the datasets. This limitation is particularly pronounced in datasets with intricate categorical features or where non-linear patterns are prevalent.

Overall, the findings underscore the importance of selecting models based on dataset characteristics. Random Forest emerges as a robust choice for tasks requiring high accuracy and the ability to handle complex data structures. Decision Trees offer a simpler alternative with moderate accuracy, while Logistic Regression serves as a baseline model with stable but lower performance. Future research could explore integrating additional feature engineering techniques or employing advanced algorithms like Gradient Boosting Machines or Neural Networks to further enhance model performance across these datasets.

References

1. UCI Machine Learning Repository: Car Evaluation Dataset
2. UCI Machine Learning Repository: Student Performance Dataset
3. UCI Machine Learning Repository: Obesity Levels Dataset
4. Breiman, L., “Random forests,” Machine Learning 45(1), 5–32 (2001).
5. Quinlan, J.R., “Induction of decision trees,” Machine Learning 1(1), 81–106 (1986).
6. Cox, D.R., “The regression analysis of binary sequences,” Journal of the Royal Statistical Society: Series B (1958).