Detailed Report on Decision Trees and Random Forests for Loan Prediction

Week 2: Supervised Learning

Task 2.2: Decision Trees and Random Forests

Objective

The objective of this task was to implement tree-based methods for classification using the Loan Prediction Dataset from Kaggle. The goal was to predict loan approval based on attributes such as gender, marital status, income, and loan amount. Specifically, we aimed to:

- Build a decision tree model and visualize the tree.
- Implement a random forest to potentially improve model performance and reduce overfitting.
- Compare the performance of the decision tree and random forest models using accuracy, F1-score, and other relevant metrics.

Dataset

The dataset used for this task is the Loan Prediction Dataset, which can be accessed here. It contains information on loan applicants including their demographics and financial details.

Activities

1. Model Implementation

Data Exploration and Preprocessing:

- Initially, we explored the dataset to understand its structure and characteristics through summary statistics and visualizations.
- Missing values in the dataset were handled using appropriate techniques such as mean imputation for numerical features and mode imputation for categorical features.
- Categorical variables were encoded using `LabelEncoder` to convert them into numerical format suitable for machine learning models.

Decision Tree Model:

- Implemented a Decision Tree Classifier using `DecisionTreeClassifier` from `sklearn.tree`.
- The model was trained on a subset of features (`Credit_History`, `Gender`, `Married`, `Education`) and their impact on `Loan_Status` was evaluated.
- Visualized the decision tree to interpret its structure and decision-making process using `plot_tree`.

Random Forest Model:

- Constructed a Random Forest Classifier using `RandomForestClassifier` from `sklearn.ensemble`.
- Utilized an ensemble of decision trees to improve prediction accuracy and mitigate overfitting.
- Trained the Random Forest model on the same subset of features and evaluated its performance on predicting `Loan_Status`.

2. Performance Comparison

Evaluation Metrics:

- **Accuracy:** Measured the overall accuracy of both models on predicting loan approval.
- **F1-score:** Assessed the harmonic mean of precision and recall to capture model performance across different classes (`Yes` and `No` for loan approval).
- **Confusion Matrix:** Analyzed the distribution of true positives, true negatives, false positives, and false negatives to gain insights into model strengths and weaknesses.

Comparison Results:

- The Random Forest model generally outperformed the Decision Tree model in terms of accuracy and F1-score due to its ability to handle more complex relationships in the data and reduce variance.
- Visual inspection of decision trees from both models provided insights into their respective decision-making processes and feature importance.

Documentation

Model Building Decisions:

- **Choice of Parameters:** Parameters such as `random_state`, `n_estimators` (for Random Forest), and feature subset (`Credit_History`, `Gender`, `Married`, `Education`) were chosen based on initial exploratory analysis and standard practices to balance model complexity and performance.
- **Tree-Based Methods:** Decision Trees and Random Forests were selected for their interpretability, ability to handle categorical data, and capability to capture non-linear relationships in the dataset.

Summary

- This report detailed the implementation of Decision Trees and Random Forests for loan prediction, starting from data preprocessing to model training and evaluation.
- The Random Forest model demonstrated superior performance over the Decision Tree model, showcasing its potential for accurate loan approval prediction.
- Visualizations of decision trees provided insights into feature importance and decision paths, aiding in model interpretability.

Conclusion

Tree-based methods such as Decision Trees and Random Forests offer robust solutions for classification tasks like loan prediction. The choice of Random Forests over Decision Trees in this scenario illustrates their capability to enhance prediction accuracy and handle complex datasets effectively.