Predicting Employee Attrition in a Company

You work in the HR department of a large corporation, and you're tasked with understanding and predicting employee attrition (whether employees leave the company). Your goal is to develop a model that can help identify factors contributing to attrition and predict which employees are at risk of leaving.

Dataset: You have access to a comprehensive dataset containing employee information with the following columns:

- 'status': Current employment status (Employed / Left)
- 'department': Department employees belong(ed) to
- 'salary': Salary level relative to the rest of their department
- 'tenure': Number of years at the company
- 'recently_promoted': Was the employee promoted in the last 3 years?
- 'n_projects': Number of projects the employee is staffed on
- 'avg_monthly_hrs': Average number of hours worked per month
- 'satisfaction': Score for the employee's satisfaction with the company (higher is better)
- 'last_evaluation': Score for the most recent evaluation of the employee (higher is better)
- 'filed_complaint': Has the employee filed a formal complaint in the last 3 years?

1. Data Exploration and Visualization:

- Conduct an exploratory data analysis (EDA) to understand the dataset. Visualize the distribution of employee tenure using a histogram. Are there any noticeable trends in tenure among employees who have left compared to those who are still employed?

- Create a boxplot to examine the distribution of job satisfaction scores among employees who left the company and those who are still employed.

- Generate a pie chart to show the distribution of attrition (employees who left vs. employees who are still employed) in the dataset. What percentage of employees have left the company?

- Create a scatter plot to explore the relationship between employee satisfaction scores and their last evaluation scores. Is there a correlation between these two variables for employees who left?

2. Decision Tree Modeling with Tree Pruning and Split Criteria:

- Split the dataset into a training set and a testing set (e.g., 80% training, 20% testing).

- Build a decision tree classifier to predict employee attrition based on selected features (e.g., tenure, satisfaction, number of projects).

- Utilize both Gini impurity and entropy as criteria for finding the best splits in the decision tree. Experiment with different criteria to determine which one results in a more effective model.

- Visualize the decision tree structure. How deep is the tree, and what are the most influential features for predicting attrition?

- Apply post-pruning techniques to control the complexity of the tree and prevent overfitting. Experiment with different pruning strategies, such as minimum leaf size or maximum depth, to find the optimal tree size.

3. Model Evaluation:

- Evaluate the pruned decision tree model using appropriate metrics such as accuracy, precision, recall, and F1-score. How well does the pruned model perform in predicting employee attrition for both Gini impurity and entropy criteria?

- Create a diverging bar chart to display the confusion matrix, showing true positives, true negatives, false positives, and false negatives for both criteria.

4. Feature Importance Visualization:

- Generate a bar chart to visualize the importance of each feature in the pruned decision tree model for both Gini impurity and entropy criteria. Which features are most crucial for predicting attrition under each criterion?

By incorporating both Gini impurity and entropy as criteria for finding the best splits in the decision tree, you can compare the performance of the model and assess which criterion results in a more effective predictor of employee attrition. Adjust the pruning parameters as needed to optimize the model's performance under both criteria.

Evaluation Rubrics

Preprocessing:2 Marks
Implementation:3 Marks
Tree Pruning:3 Marks
Model Evaluation:2 Marks
Total:10 Marks