**HOMEWORK-3**

1. Preprocessing
2. Remove instances where categorical attributes have missing values. For numeric attributes, fill missing values using the mean value.

Preprocessing

After loading the dataset, I applied preprocessing to clean null values from categorical columns and imputed the mean for numerical columns. I used StringIndexer to convert categorical values into numerical format for Spark ML. Then, I applied VectorAssembler to combine all features into a single column, making the data suitable for modeling. Finally, I split the dataset into 80% training and 20% testing to ensure a balanced evaluation.

A screen shot of a computer

AI-generated content may be incorrect.

1. Classification: Split the data after preprocessing from above step to 80% training and 20% test using randomized split.
2. Implement classification using decision tree to classify the target (label) attribute income using Spark ML on databricks community edition.
3. Show the evaluation metrics such as: accuracy, precision and recall of the classifier.
4. Implement classification using a different algorithm and compare it with decision tree using the evaluation metrics from above.

**Decision Tree Classification**

After fitting the Decision Tree model, I evaluated its performance using accuracy, precision, and recall. The accuracy of the Decision Tree model is **0.81**, precision is **0.79**, and recall is **0.81**. These metrics indicate how well the model classified income levels based on the available features.

A screenshot of a computer

AI-generated content may be incorrect.

**Random Forest Classification**

After training the Random Forest model with 100 trees, I evaluated its performance using the same metrics. The accuracy is **0.80**, precision is **0.79**, and recall is **0.80**. This shows an improvement in classification performance compared to the Decision Tree.

**Comparison**

The Decision Tree performed better than the Random Forest in all evaluation metrics.

A screenshot of a computer

AI-generated content may be incorrect.