A Comparative Analysis of Decision Tree and K-Nearest Neighbors for Classification Tasks

Data Preprocessing

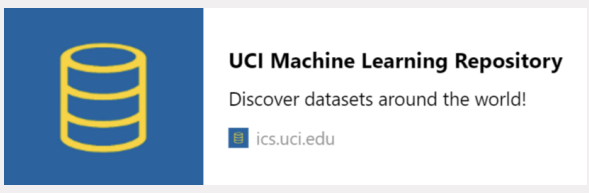
Data Cleaning
- Remove rows with unknown values
- Remove duplicate rows
- Remove similar/redundant field

Data Transformation
- Transform fields from categorical data to binary data
- Standardise data

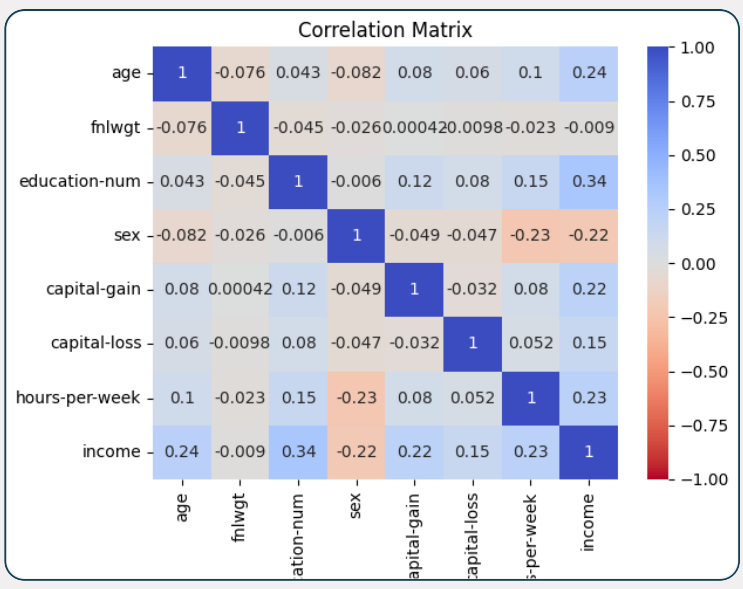
Feature Selection
- Check correlation with label
- Remove feature(s) with low correlation with label

Feature Engineering
- One-hot encoding for categorical features

Sampling
- Oversample minority class



Dataset:
https://archive.ics.uci.edu/ml/datasets/Census+Income



Implementation of Classifiers

Decision Tree

Grid parameters (best in bold):

"criterion": [**"gini"**,"entropy"],
"max_depth": [5,10,**15**],
"min_samples_split": [**2**,5,10],
"min_impurity_decrease": [**0.0**,0.1]

Time taken for:

hyperparameter tuning: **51.6s**
model training: **0.7s**
prediction: **0.0s (51ms)**

K-Nearest Neighbors

Grid parameters (best in bold):

"n_neighbors": [**3**,5,7],
"weights": ["uniform",**"distance"**],
"leaf_size": [**20**,30,40],
"p": [**1**,2]

Time taken for:

hyperparameter tuning: **20m 44.2s**
model training: **0.0s (68ms)**
prediction: **16.8s**

Performance Evaluation

To ensure a **reliable estimate of performance**, these approaches were used:
- Stratified Sampling - Use SMOTE RandomOverSample for oversampling minority class
- Cross Validation - 5-fold cv in GridSearchCV

Performance Evaluation Metrics
- Accuracy
- Precision
- Recall
- F1
- ROC-AUC

In this scenario,
Decision Tree
is a better classifier than
K-Nearest Neighbors.