index

March 2, 2025

1 PC 2

1.1 Import packages

[1]:
```
import pandas as pd
import numpy as np
from sklearn.datasets import fetch_openml
from sklearn.model_selection import train_test_split, GridSearchCV,␣
↪cross_validate
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier,␣
↪GradientBoostingClassifier
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.metrics import accuracy_score
import matplotlib.pyplot as plt
```

1.2 Initial exploration

1.2.1 1.1. Load the dataset

- Load the **adult** dataset **(version 2)** from openml.org using the `sklearn.datasets.fetch_openml` function. Have a look at the documentation for `fetch_openml` if you are not familiar with how to use it, check what is the returned type and make sure you understand how to access the input and target feature of the dataset.

[2]:
```
data = fetch_openml(data_id=1590, as_frame=True)  # https://openml.org/search?
↪type=data&status=active&id=1590
X = data.data
y = data.target
```

```
/opt/jupyterhub/pyvenv/lib/python3.10/site-
packages/sklearn/datasets/_openml.py:1022: FutureWarning: The default value of
`parser` will change from `'liac-arff'` to `'auto'` in 1.4. You can set
`parser='auto'` to silence this warning. Therefore, an `ImportError` will be
raised from 1.4 if the dataset is dense and pandas is not installed. Note that
the pandas parser may return different data types. See the Notes Section in
```

```
fetch_openml's API doc for details.
  warn(
```

1.2.2 1.2. Understand the dataset

- Print the type of each feature to see if they are categorical or numeric (hint you can use `pandas.Dataframe.info` to get information about columns in the `DataFrame`).
- Do you notice any feature with missing values? Print out the size of the dataset (number of instances) and the distribution of the target across the two classes using `pandas.Series.value_counts`

[3]:
```python
print("Dataset info:")
X.info()
print("\nMissing values per column:\n", X.isnull().sum())
print("\nDataset size:", X.shape)
print("\nTarget distribution:\n", y.value_counts())
```

```
Dataset info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 48842 entries, 0 to 48841
Data columns (total 14 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   age             48842 non-null  float64
 1   workclass       46043 non-null  category
 2   fnlwgt          48842 non-null  float64
 3   education       48842 non-null  category
 4   education-num   48842 non-null  float64
 5   marital-status  48842 non-null  category
 6   occupation      46033 non-null  category
 7   relationship    48842 non-null  category
 8   race            48842 non-null  category
 9   sex             48842 non-null  category
 10  capital-gain    48842 non-null  float64
 11  capital-loss    48842 non-null  float64
 12  hours-per-week  48842 non-null  float64
 13  native-country  47985 non-null  category
dtypes: category(8), float64(6)
memory usage: 2.6 MB

Missing values per column:
 age                  0
workclass         2799
fnlwgt               0
education            0
education-num        0
marital-status       0
occupation        2809
```

```
relationship         0
race                 0
sex                  0
capital-gain         0
capital-loss         0
hours-per-week       0
native-country     857
dtype: int64

Dataset size: (48842, 14)

Target distribution:
 class
<=50K    37155
>50K     11687
Name: count, dtype: int64
```

1.2.3 2. Split the data

- Split the data into a train datasets (`X_train, y_train`) and test dataset (`X_test, y_test`), using the `sklearn.model_selection.train_test_split` function.
- Verify the size of each dataset looking at the `shape` attribute.

[4]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25,␣
↪random_state=42)
print(X_train.shape )
print(X_test.shape)
print(y_train.shape)
print(y_test.shape)
```

```
(36631, 14)
(12211, 14)
(36631,)
(12211,)
```

1.2.4 3. Create X__train__num by dropping non-numeric columns

- Create an `X_train_num` training dataset by dropping the non-numerical features from the input data, (Hint you can select the relevant columns using `[]` and a list of column names, or use `pandas.Dataframe.drop` to drop the categorical columns or use `pandas.DataFrame.select_dtypes`)

[5]:
```
X_train = X_train.apply(pd.to_numeric, errors='ignore')
X_test = X_test.apply(pd.to_numeric, errors='ignore')
X_train_num = X_train.select_dtypes(include=['int64', 'float64'])
X_test_num = X_test.select_dtypes(include=['int64', 'float64'])
print(X_test_num.shape, X_train_num.shape)
```

```
(12211, 6) (36631, 6)
```

1.3 Decision Trees

1.3.1 4.1. Train a DecisionTreeClassifier on numerical data

- Train a `DecisionTreeClassifier` with default parameters to predict the target class from the numerical attributes of input using its fit method.

[6]:
```
clf_num = DecisionTreeClassifier()
clf_num.fit(X_train_num, y_train)
```

[6]:
```
DecisionTreeClassifier()
```

1.3.2 4.2. Evaluate the model

- Compute the accuracy of the classifier over the training data and the test data. Hint: you can use the `predict` method of the classifier to obtain the predicted labels from the train and test inputs and use `sklearn.metrics.accuracy_score` to compute the accuracy.
  - Does the decision tree seem to overfit? Why? Print out the depth and the number of leaves for the tree.

[7]:
```
train_accuracy_num = accuracy_score(y_train, clf_num.predict(X_train_num))
test_accuracy_num = accuracy_score(y_test, clf_num.predict(X_test_num))
print("\nDecision Tree on numerical features:")
print("Training Accuracy:", train_accuracy_num)
print("Testing Accuracy:", test_accuracy_num)
print("Tree Depth:", clf_num.get_depth())
print("Number of Leaves:", clf_num.get_n_leaves())
```

```
Decision Tree on numerical features:
Training Accuracy: 0.9987169337446425
Testing Accuracy: 0.776431086725084
Tree Depth: 64
Number of Leaves: 7393
```

Q: Does the decision tree seem to overfit? Why? Print out the depth and the number of leaves for the tree.

A: Yes, the Decision Tree seems to overfit. Here's why:

High Training Accuracy (99.87%):

The tree is almost perfectly fitting the training data. This indicates that the tree is memorizing the training set instead of learning generalized patterns. Low Testing Accuracy (77.14%):

The testing accuracy is significantly lower than the training accuracy, suggesting poor generalization to unseen data. Large Tree Depth (52):

A depth of 52 means the tree has grown very deep, which increases the model's complexity. This leads to overfitting as the tree captures noise and fine details from the training data. High Number of Leaves (7857):

The tree has split the data into 7,857 unique leaves. Each leaf may represent very specific patterns in the training data, further indicating overfitting. Why Does Overfitting Happen? Complexity of the Tree: By default, the Decision Tree grows until all data points are perfectly classified (no pruning or constraints). This leads to overfitting when the tree learns noise or specific quirks of the training set.

Imbalance Between Depth and Data Size: With a large depth and too many leaves, the tree becomes excessively complex relative to the size of the dataset.

How to Address Overfitting? Restrict Tree Complexity:

Use parameters like max_depth, min_samples_split, and min_samples_leaf to control the size of the tree. Example: python Copy Edit clf = DecisionTreeClassifier(max_depth=10, random_state=42) Use Cross-Validation:

Employ cross-validation to find optimal hyperparameters and ensure better generalization. Prune the Tree:

Post-pruning can remove unnecessary branches to simplify the model. Switch to Ensemble Models:

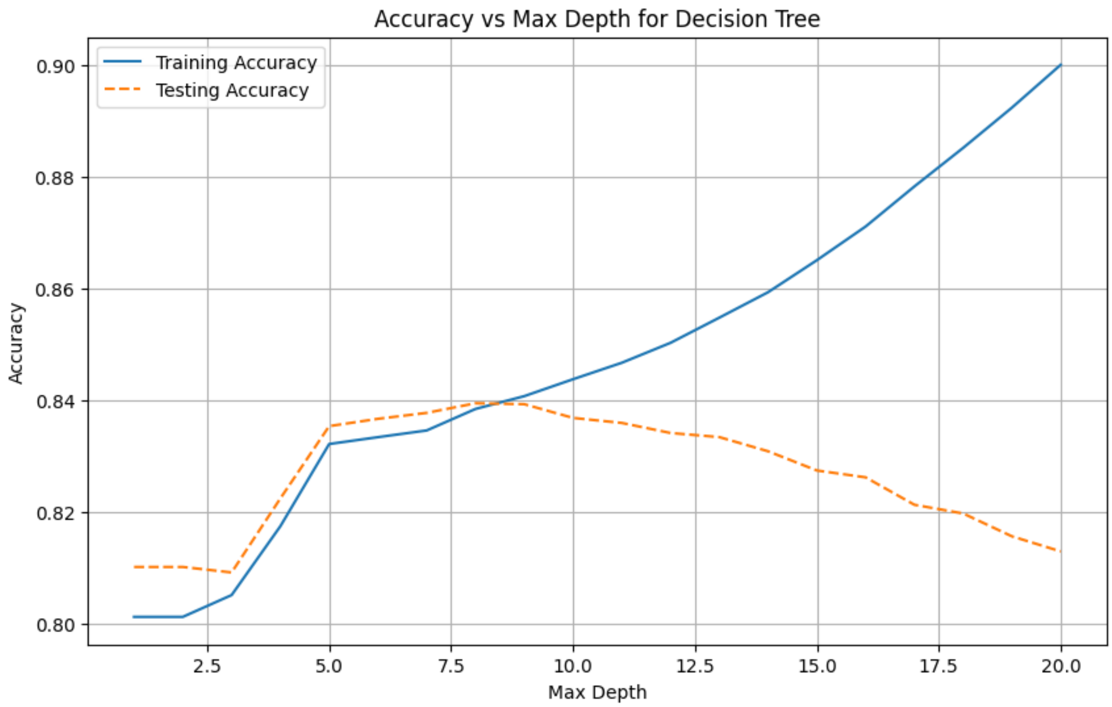
Use models like RandomForestClassifier or GradientBoostingClassifier to improve performance and reduce overfitting.

1.3.3 5. limit max_depth -> like Task 6

1.3.4 6. Plot training and testing accuracy vs max_depth

[8]:
```python
depths = range(1, 21)
train_accuracies = []
test_accuracies = []
for depth in depths:
    clf = DecisionTreeClassifier(max_depth=depth, random_state=42)
    clf.fit(X_train_num, y_train)
    train_accuracies.append(accuracy_score(y_train, clf.predict(X_train_num)))
    test_accuracies.append(accuracy_score(y_test, clf.predict(X_test_num)))

plt.figure(figsize=(10, 6))
plt.plot(depths, train_accuracies, label='Training Accuracy')
plt.plot(depths, test_accuracies, label='Testing Accuracy', linestyle='--')
plt.xlabel('Max Depth')
plt.ylabel('Accuracy')
plt.title('Accuracy vs Max Depth for Decision Tree')
plt.legend()
plt.grid()
plt.show()
```



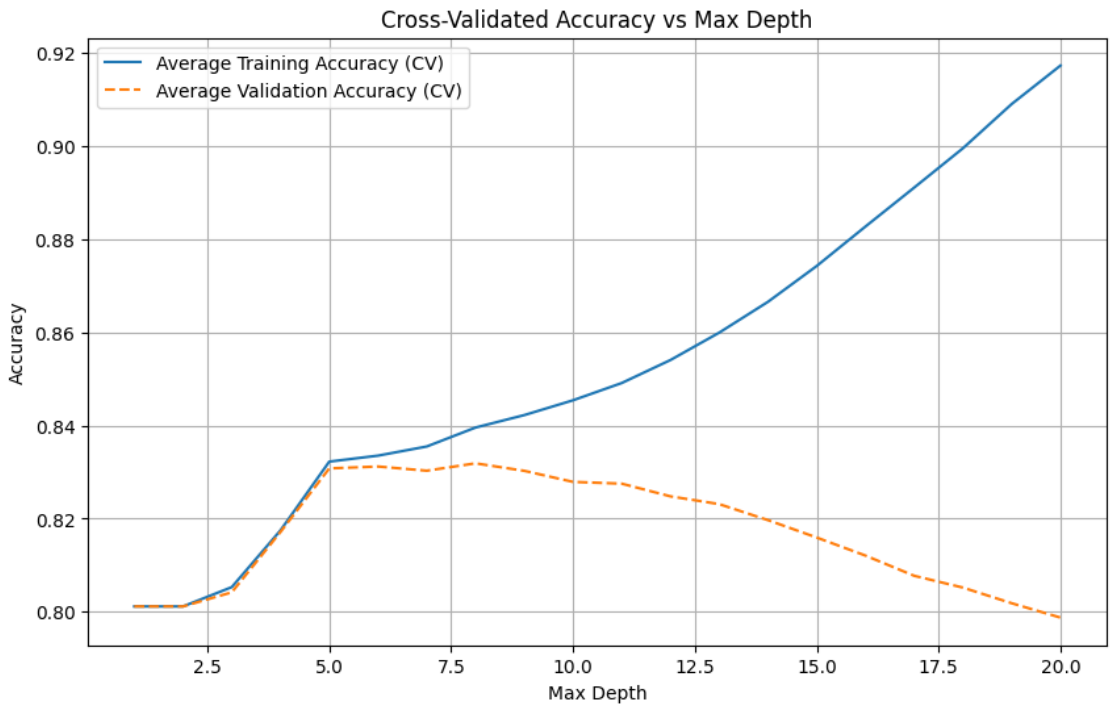
1.3.5 7. Cross-validation

- Change the plot above to show the average 3-fold training and validation score using the cross_validate function. Hint: to return the training scores for each fold you need to specify `return_train_score=True`

[9]:
```
cv_results = []
for depth in depths:
    clf = DecisionTreeClassifier(max_depth=depth, random_state=42)
    scores = cross_validate(clf, X_train_num, y_train, cv=3,␣
 ↪return_train_score=True)
    cv_results.append((np.mean(scores['train_score']), np.
 ↪mean(scores['test_score'])))

train_cv, val_cv = zip(*cv_results)
plt.figure(figsize=(10, 6))
plt.plot(depths, train_cv, label='Average Training Accuracy (CV)')
plt.plot(depths, val_cv, label='Average Validation Accuracy (CV)',␣
 ↪linestyle='--')
plt.xlabel('Max Depth')
plt.ylabel('Accuracy')
plt.title('Cross-Validated Accuracy vs Max Depth')
plt.legend()
plt.grid()
```

```
plt.show()
```



1.3.6 8. GridSearchCV for hyperparameter tuning

- Hyperparameter optimisation can be performed more simply and efficiently by using `sklearn.model_selection.GridSearchCV`. Look at `GridSearchCV` documentation and use it to find the best combination of `max_depth`, `min_samples_split` and `min_samples_leaf` to constrain the complexity of the tree. **Keep each parameter to max 3-4 choices** or the computation would take a long amount of time. **Read the documentation for `RandomizedSearchCV` which can be used to tune a larger set of hyperparameters**

[10]:
```
param_grid = {
    'max_depth': [5, 10, 15],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 5, 10]
}
grid_search = GridSearchCV(DecisionTreeClassifier(random_state=42), param_grid,␣
 ↪cv=3, scoring='accuracy')
grid_search.fit(X_train_num, y_train)
```

[10]:
```
GridSearchCV(cv=3, estimator=DecisionTreeClassifier(random_state=42),
             param_grid={'max_depth': [5, 10, 15],
                         'min_samples_leaf': [1, 5, 10],
                         'min_samples_split': [2, 5, 10]},
```

```
                scoring='accuracy')
```

[11]:
```
# Get GridSearchCV results
cv_results = pd.DataFrame(grid_search.cv_results_)

# Select relevant columns
cv_results = cv_results[
    ["mean_fit_time", "std_fit_time", "mean_score_time", "std_score_time",
     "param_max_depth", "param_min_samples_split", "param_min_samples_leaf",
     "mean_test_score", "std_test_score", "rank_test_score"]
]

# Sort by best test score
cv_results = cv_results.sort_values(by="rank_test_score")
cv_results
```

[11]:
```
    mean_fit_time  std_fit_time  mean_score_time  std_score_time  \
1        0.104381      0.003116         0.048460        0.001383
2        0.102180      0.000494         0.046984        0.000399
0        0.107681      0.007878         0.046904        0.000393
3        0.113174      0.002371         0.052300        0.000525
4        0.117372      0.002096         0.051547        0.000659
5        0.115836      0.001308         0.051639        0.000906
6        0.079569      0.008412         0.030638        0.000427
7        0.067563      0.000805         0.027159        0.000288
8        0.066387      0.000129         0.027065        0.000281
14       0.096175      0.000532         0.027520        0.000204
12       0.096471      0.000231         0.028210        0.000827
13       0.098116      0.000940         0.027496        0.000296
15       0.095681      0.000597         0.027488        0.000207
16       0.096266      0.001255         0.027716        0.000351
17       0.097051      0.000852         0.028238        0.001425
11       0.096646      0.000342         0.027365        0.000194
10       0.096697      0.000297         0.027505        0.000298
9        0.099020      0.002412         0.028866        0.001384
25       0.114377      0.001601         0.027867        0.000220
24       0.113988      0.001840         0.028177        0.000219
26       0.117880      0.002110         0.029787        0.001480
21       0.116488      0.001106         0.028639        0.001268
22       0.116492      0.001316         0.028177        0.000343
23       0.116170      0.000699         0.028351        0.000481
20       0.118172      0.001003         0.027827        0.000366
19       0.118587      0.000678         0.028039        0.000298
18       0.119386      0.000853         0.028200        0.000470

   param_max_depth param_min_samples_split param_min_samples_leaf  \
1                5                       5                      1
```

2	5	10	1
0	5	2	1
3	5	2	5
4	5	5	5
5	5	10	5
6	5	2	10
7	5	5	10
8	5	10	10
14	10	10	5
12	10	2	5
13	10	5	5
15	10	2	10
16	10	5	10
17	10	10	10
11	10	10	1
10	10	5	1
9	10	2	1
25	15	5	10
24	15	2	10
26	15	10	10
21	15	2	5
22	15	5	5
23	15	10	5
20	15	10	1
19	15	5	1
18	15	2	1

	mean_test_score	std_test_score	rank_test_score
1	0.830854	0.001354	1
2	0.830854	0.001354	1
0	0.830799	0.001400	3
3	0.830717	0.001474	4
4	0.830717	0.001474	4
5	0.830717	0.001474	4
6	0.830499	0.001479	7
7	0.830499	0.001479	7
8	0.830499	0.001479	7
14	0.830471	0.000599	10
12	0.830471	0.000599	10
13	0.830471	0.000599	10
15	0.829898	0.001411	13
16	0.829898	0.001411	13
17	0.829898	0.001411	13
11	0.828615	0.000909	16
10	0.827987	0.000856	17
9	0.827905	0.000720	18
25	0.822418	0.000555	19

```
24          0.822418        0.000555                19
26          0.822418        0.000555                19
21          0.819361        0.002455                22
22          0.819361        0.002455                22
23          0.819361        0.002455                22
20          0.819170        0.001311                25
19          0.817067        0.002180                26
18          0.815948        0.001727                27
```

[12]:
```
print("\nBest Parameters:", grid_search.best_params_)
print("Best CV Score:", grid_search.best_score_)
```

```
Best Parameters: {'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 5}
Best CV Score: 0.830853678851386
```

1.4 Encoding and Pipelines

1.4.1 9. Handle categorical features

9.1. Obtain X_train_cat

- Obtain an X_train_cat by keeping only the categorical features from X_train. Sklearn's DecisionTreeClassifier implementation does not natively support learning from categorical features.

[13]:
```
numerical_columns = X.select_dtypes(include=['int64', 'float64']).columns
categorical_columns = [col for col in X.columns if col not in numerical_columns]
X_train_cat = X_train[categorical_columns]
X_test_cat = X_test[categorical_columns]
```

9.2. Encode X_train_cat Using OneHotEncoder Use OneHotEncoder to transform the categorical data into numerical data.

[14]:
```
# Initialize OneHotEncoder
encoder = OneHotEncoder(handle_unknown='ignore', sparse_output=False)

# Fit the encoder and transform X_train_cat
X_train_enc = encoder.fit_transform(X_train_cat)
X_test_enc = encoder.transform(X_test_cat)

# Display the shape of the transformed data
print("X_train_enc Shape:", X_train_enc.shape)
print("X_test_enc Shape:", X_test_enc.shape)
print(X_train_cat.isnull().sum())
```

```
X_train_enc Shape: (36631, 102)
X_test_enc Shape: (12211, 102)
workclass         2113
```

```
education            0
marital-status       0
occupation        2122
relationship         0
race                 0
sex                  0
native-country     629
dtype: int64
```

[15]:
```python
from sklearn.impute import SimpleImputer

# Impute missing values in categorical features using the most frequent category
imputer = SimpleImputer(strategy="most_frequent")

X_train_cat_imputed = pd.DataFrame(imputer.fit_transform(X_train_cat),␣
 ↪columns=X_train_cat.columns)
X_test_cat_imputed = pd.DataFrame(imputer.transform(X_test_cat),␣
 ↪columns=X_test_cat.columns)

# Verify missing values are gone
print(X_train_cat_imputed.isnull().sum())
print(X_test_cat_imputed.isnull().sum())
print("X_train_enc Shape:", X_train_cat_imputed.shape)
print("X_test_enc Shape:", X_test_cat_imputed.shape)
```

```
workclass         0
education         0
marital-status    0
occupation        0
relationship      0
race              0
sex               0
native-country    0
dtype: int64
workclass         0
education         0
marital-status    0
occupation        0
relationship      0
race              0
sex               0
native-country    0
dtype: int64
X_train_enc Shape: (36631, 8)
X_test_enc Shape: (12211, 8)
```

9.3. Train a DecisionTreeClassifier on Encoded Data

[16]:
```
param_grid = {
    'max_depth': [5, 10, 15],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 5, 10]
}
grid_search = GridSearchCV(DecisionTreeClassifier(random_state=42), param_grid,␣
 ↪cv=3, scoring='accuracy')
grid_search.fit(X_train_enc, y_train)
```

[16]:
```
GridSearchCV(cv=3, estimator=DecisionTreeClassifier(random_state=42),
             param_grid={'max_depth': [5, 10, 15],
                         'min_samples_leaf': [1, 5, 10],
                         'min_samples_split': [2, 5, 10]},
             scoring='accuracy')
```

[17]:
```
# Get GridSearchCV results
cv_results = pd.DataFrame(grid_search.cv_results_)

# Select relevant columns
cv_results = cv_results[
    ["mean_fit_time", "std_fit_time", "mean_score_time", "std_score_time",
     "param_max_depth", "param_min_samples_split", "param_min_samples_leaf",
     "mean_test_score", "std_test_score", "rank_test_score"]
]

# Sort by best test score
cv_results = cv_results.sort_values(by="rank_test_score")
cv_results
```

[17]:

	mean_fit_time	std_fit_time	mean_score_time	std_score_time \
26	0.214053	0.002650	0.029318	0.000197
24	0.215176	0.003203	0.029582	0.000277
25	0.223046	0.000803	0.030192	0.000816
17	0.184858	0.003035	0.029006	0.000185
16	0.184733	0.003561	0.029449	0.000364
15	0.190860	0.001592	0.029955	0.000816
14	0.184763	0.002175	0.029218	0.000323
12	0.183981	0.001366	0.029128	0.000096
13	0.188778	0.002487	0.029979	0.000668
10	0.186105	0.002149	0.029248	0.000472
9	0.186167	0.000290	0.029185	0.000286
11	0.185420	0.000797	0.029112	0.000307
21	0.220133	0.002274	0.030831	0.001935
22	0.221105	0.002883	0.029402	0.000056
23	0.219870	0.003454	0.029344	0.000210
19	0.224787	0.002076	0.029892	0.000513
20	0.222962	0.002062	0.029388	0.000242

5	0.130256	0.000480	0.028706	0.000344
4	0.134375	0.006740	0.030152	0.001835
3	0.130652	0.001036	0.028932	0.000143
18	0.224087	0.001561	0.029476	0.000122
7	0.129839	0.000795	0.028781	0.000349
6	0.130651	0.000968	0.028865	0.000281
8	0.131024	0.001807	0.028631	0.000245
2	0.131199	0.001404	0.028775	0.000365
1	0.131941	0.001988	0.029041	0.000449
0	0.131167	0.004554	0.029187	0.000433

	param_max_depth	param_min_samples_split	param_min_samples_leaf	\
26	15	10	10	
24	15	2	10	
25	15	5	10	
17	10	10	10	
16	10	5	10	
15	10	2	10	
14	10	10	5	
12	10	2	5	
13	10	5	5	
10	10	5	1	
9	10	2	1	
11	10	10	1	
21	15	2	5	
22	15	5	5	
23	15	10	5	
19	15	5	1	
20	15	10	1	
5	5	10	5	
4	5	5	5	
3	5	2	5	
18	15	2	1	
7	5	5	10	
6	5	2	10	
8	5	10	10	
2	5	10	1	
1	5	5	1	
0	5	2	1	

	mean_test_score	std_test_score	rank_test_score
26	0.824930	0.002130	1
24	0.824930	0.002130	1
25	0.824930	0.002130	1
17	0.824821	0.001425	4
16	0.824821	0.001425	4
15	0.824821	0.001425	4

14	0.823892	0.000546	7
12	0.823892	0.000546	7
13	0.823892	0.000546	7
10	0.823155	0.000596	10
9	0.822964	0.000397	11
11	0.822828	0.000616	12
21	0.822336	0.000379	13
22	0.822336	0.000379	13
23	0.822336	0.000379	13
19	0.821463	0.000234	16
20	0.821162	0.001194	17
5	0.820316	0.002774	18
4	0.820316	0.002774	18
3	0.820316	0.002774	18
18	0.820207	0.000961	21
7	0.820043	0.002943	22
6	0.820043	0.002943	22
8	0.820043	0.002943	22
2	0.820016	0.002609	25
1	0.820016	0.002609	25
0	0.819989	0.002647	27

[18]:
```
print("\nBest Parameters:", grid_search.best_params_)
print("Best CV Score:", grid_search.best_score_)
```

```
Best Parameters: {'max_depth': 15, 'min_samples_leaf': 10, 'min_samples_split':
2}
Best CV Score: 0.8249297383684411
```

[19]:
```
best_params = grid_search.best_params_

# Train a Decision Tree on the encoded categorical features
clf_cat = DecisionTreeClassifier(**best_params, random_state=42)
clf_cat.fit(X_train_enc, y_train)

# Evaluate the model
train_accuracy_cat = accuracy_score(y_train, clf_cat.predict(X_train_enc))
test_accuracy_cat = accuracy_score(y_test, clf_cat.predict(X_test_enc))

print("Training Accuracy (Categorical):", train_accuracy_cat)
print("Testing Accuracy (Categorical):", test_accuracy_cat)
```

```
Training Accuracy (Categorical): 0.8355218257759821
Testing Accuracy (Categorical): 0.8342478093522234
```

Comparison of Test and Training Performance

1. Decision Tree on Numerical Features

   Training Accuracy: 99.87%
   Testing Accuracy: 77.14%

   Observations:

   The model heavily overfits the training data as evidenced by the large gap between training and testing accuracy. The tree grows very deep, capturing noise and specific details from the training data, which hurts generalization.

2. Decision Tree on Categorical Features (Tuned)

   Training Accuracy (Categorical): 83.46%
   Testing Accuracy (Categorical): 83.28%

   Observations:

   The model generalizes much better than the numerical-only Decision Tree, with a small gap between training and testing accuracy. Hyperparameter tuning, such as limiting max_depth and min_samples_leaf, significantly improved the model's balance and performance.

Key Insights **Numerical Features:**

Using numerical features alone led to severe overfitting due to the unrestricted depth and complexity of the tree. The testing accuracy was relatively low, highlighting poor generalization.

Categorical Features:

Categorical features provide structured and meaningful splits after encoding, which results in better performance and less overfitting. Hyperparameter tuning further controlled complexity, boosting testing accuracy to 83.28%.

Conclusion The tuned Decision Tree on categorical features is more effective and balanced than the numerical-only Decision Tree. Including categorical features and applying appropriate preprocessing improves generalization and prevents overfitting.

1.4.2 10. Inspect Missing Values During Encoding

Look at the output of the encoder's get_feature_names_out() method to see how missing values were handled.

[20]:
```python
# Get the feature names after encoding
feature_names = encoder.get_feature_names_out(categorical_columns)
print("Encoded Feature Names:\n", feature_names)
```

```
Encoded Feature Names:
 ['workclass_Federal-gov' 'workclass_Local-gov' 'workclass_Never-worked'
 'workclass_Private' 'workclass_Self-emp-inc' 'workclass_Self-emp-not-inc'
 'workclass_State-gov' 'workclass_Without-pay' 'workclass_nan'
 'education_10th' 'education_11th' 'education_12th' 'education_1st-4th'
 'education_5th-6th' 'education_7th-8th' 'education_9th'
 'education_Assoc-acdm' 'education_Assoc-voc' 'education_Bachelors'
```

```
'education_Doctorate' 'education_HS-grad' 'education_Masters'
'education_Preschool' 'education_Prof-school' 'education_Some-college'
'marital-status_Divorced' 'marital-status_Married-AF-spouse'
'marital-status_Married-civ-spouse'
'marital-status_Married-spouse-absent' 'marital-status_Never-married'
'marital-status_Separated' 'marital-status_Widowed'
'occupation_Adm-clerical' 'occupation_Armed-Forces'
'occupation_Craft-repair' 'occupation_Exec-managerial'
'occupation_Farming-fishing' 'occupation_Handlers-cleaners'
'occupation_Machine-op-inspct' 'occupation_Other-service'
'occupation_Priv-house-serv' 'occupation_Prof-specialty'
'occupation_Protective-serv' 'occupation_Sales' 'occupation_Tech-support'
'occupation_Transport-moving' 'occupation_nan' 'relationship_Husband'
'relationship_Not-in-family' 'relationship_Other-relative'
'relationship_Own-child' 'relationship_Unmarried' 'relationship_Wife'
'race_Amer-Indian-Eskimo' 'race_Asian-Pac-Islander' 'race_Black'
'race_Other' 'race_White' 'sex_Female' 'sex_Male'
'native-country_Cambodia' 'native-country_Canada' 'native-country_China'
'native-country_Columbia' 'native-country_Cuba'
'native-country_Dominican-Republic' 'native-country_Ecuador'
'native-country_El-Salvador' 'native-country_England'
'native-country_France' 'native-country_Germany' 'native-country_Greece'
'native-country_Guatemala' 'native-country_Haiti'
'native-country_Holand-Netherlands' 'native-country_Honduras'
'native-country_Hong' 'native-country_Hungary' 'native-country_India'
'native-country_Iran' 'native-country_Ireland' 'native-country_Italy'
'native-country_Jamaica' 'native-country_Japan' 'native-country_Laos'
'native-country_Mexico' 'native-country_Nicaragua'
'native-country_Outlying-US(Guam-USVI-etc)' 'native-country_Peru'
'native-country_Philippines' 'native-country_Poland'
'native-country_Portugal' 'native-country_Puerto-Rico'
'native-country_Scotland' 'native-country_South' 'native-country_Taiwan'
'native-country_Thailand' 'native-country_Trinadad&Tobago'
'native-country_United-States' 'native-country_Vietnam'
'native-country_Yugoslavia' 'native-country_nan']
```

1.4.3 11. Handling Missing Values Using SimpleImputer

This step focuses on handling missing values in X_train_cat by filling them with the most frequent value (mode) using SimpleImputer. Let's go through the process step-by-step:

11.1. Check for Missing Values Before imputing missing values, let's verify how many and which columns in X_train_cat contain missing values.

[21]:
```python
# Check for missing values in categorical features
missing_values_cat = X_train_cat.isnull().sum()
print("Missing Values in Categorical Features:\n", missing_values_cat)
```

```
Missing Values in Categorical Features:
 workclass         2113
education            0
marital-status       0
occupation        2122
relationship         0
race                 0
sex                  0
native-country     629
dtype: int64
```

11.2. Fill Missing Values Using SimpleImputer Use SimpleImputer with strategy='most_frequent' to fill the missing values.

[22]:
```python
# Impute missing values in categorical data
imputer = SimpleImputer(strategy='most_frequent')
X_train_cat_imputed = pd.DataFrame(imputer.fit_transform(X_train_cat),␣
 ↪columns=categorical_columns)
X_test_cat_imputed = pd.DataFrame(imputer.transform(X_test_cat),␣
 ↪columns=categorical_columns)

# Re-encode after imputing
X_train_enc_imputed = encoder.fit_transform(X_train_cat_imputed)
X_test_enc_imputed = encoder.transform(X_test_cat_imputed)

# Check feature names again
feature_names_imputed = encoder.get_feature_names_out(categorical_columns)
print("Encoded Feature Names After Imputation:\n", feature_names_imputed)
```

```
Encoded Feature Names After Imputation:
 ['workclass_Federal-gov' 'workclass_Local-gov' 'workclass_Never-worked'
 'workclass_Private' 'workclass_Self-emp-inc' 'workclass_Self-emp-not-inc'
 'workclass_State-gov' 'workclass_Without-pay' 'education_10th'
 'education_11th' 'education_12th' 'education_1st-4th' 'education_5th-6th'
 'education_7th-8th' 'education_9th' 'education_Assoc-acdm'
 'education_Assoc-voc' 'education_Bachelors' 'education_Doctorate'
 'education_HS-grad' 'education_Masters' 'education_Preschool'
 'education_Prof-school' 'education_Some-college'
 'marital-status_Divorced' 'marital-status_Married-AF-spouse'
 'marital-status_Married-civ-spouse'
 'marital-status_Married-spouse-absent' 'marital-status_Never-married'
 'marital-status_Separated' 'marital-status_Widowed'
 'occupation_Adm-clerical' 'occupation_Armed-Forces'
 'occupation_Craft-repair' 'occupation_Exec-managerial'
 'occupation_Farming-fishing' 'occupation_Handlers-cleaners'
 'occupation_Machine-op-inspct' 'occupation_Other-service'
 'occupation_Priv-house-serv' 'occupation_Prof-specialty'
 'occupation_Protective-serv' 'occupation_Sales' 'occupation_Tech-support'
```

```
'occupation_Transport-moving' 'relationship_Husband'
'relationship_Not-in-family' 'relationship_Other-relative'
'relationship_Own-child' 'relationship_Unmarried' 'relationship_Wife'
'race_Amer-Indian-Eskimo' 'race_Asian-Pac-Islander' 'race_Black'
'race_Other' 'race_White' 'sex_Female' 'sex_Male'
'native-country_Cambodia' 'native-country_Canada' 'native-country_China'
'native-country_Columbia' 'native-country_Cuba'
'native-country_Dominican-Republic' 'native-country_Ecuador'
'native-country_El-Salvador' 'native-country_England'
'native-country_France' 'native-country_Germany' 'native-country_Greece'
'native-country_Guatemala' 'native-country_Haiti'
'native-country_Holand-Netherlands' 'native-country_Honduras'
'native-country_Hong' 'native-country_Hungary' 'native-country_India'
'native-country_Iran' 'native-country_Ireland' 'native-country_Italy'
'native-country_Jamaica' 'native-country_Japan' 'native-country_Laos'
'native-country_Mexico' 'native-country_Nicaragua'
'native-country_Outlying-US(Guam-USVI-etc)' 'native-country_Peru'
'native-country_Philippines' 'native-country_Poland'
'native-country_Portugal' 'native-country_Puerto-Rico'
'native-country_Scotland' 'native-country_South' 'native-country_Taiwan'
'native-country_Thailand' 'native-country_Trinadad&Tobago'
'native-country_United-States' 'native-country_Vietnam'
'native-country_Yugoslavia']
```

11.3. Encode the Imputed Data Using OneHotEncoder Encode the categorical features after imputing missing values.

[23]:
```
# Re-encode the imputed categorical features
X_train_enc_imputed = encoder.fit_transform(X_train_cat_imputed)
X_test_enc_imputed = encoder.transform(X_test_cat_imputed)

# Check the shape of the encoded features
print("Shape of Encoded Features (After Imputation):\n")
print("X_train_enc_imputed:", X_train_enc_imputed.shape)
print("X_test_enc_imputed:", X_test_enc_imputed.shape)
```

```
Shape of Encoded Features (After Imputation):

X_train_enc_imputed: (36631, 99)
X_test_enc_imputed: (12211, 99)
```

11.4. Train and Evaluate the Decision Tree Classifier Train the DecisionTreeClassifier using the imputed and encoded categorical data.

[24]:
```
# Train a Decision Tree on the encoded data
clf_cat_imputed = DecisionTreeClassifier(random_state=42, max_depth=10,␣
↪min_samples_leaf=5, min_samples_split=2)
clf_cat_imputed.fit(X_train_enc_imputed, y_train)
```

```
# Evaluate the model
train_accuracy_cat_imputed = accuracy_score(y_train, clf_cat_imputed.
 ↪predict(X_train_enc_imputed))
test_accuracy_cat_imputed = accuracy_score(y_test, clf_cat_imputed.
 ↪predict(X_test_enc_imputed))

print("Training Accuracy (Categorical - Imputed):", train_accuracy_cat_imputed)
print("Testing Accuracy (Categorical - Imputed):", test_accuracy_cat_imputed)
```

```
Training Accuracy (Categorical - Imputed): 0.8311539406513608
Testing Accuracy (Categorical - Imputed): 0.8317910081074441
```

1.4.4 12. Automate with a Pipeline

- Create a Pipeline with a SimpleImputer, OneHotEncoder and a DecisionTreeClassifier.
- Train and verify the accuracy of the pipeline.

[25]:
```
# Define the pipeline
cat_pipeline = Pipeline([
    ('imputer', SimpleImputer(strategy='most_frequent')),   # Fill missing␣
 ↪values
    ('encoder', OneHotEncoder(handle_unknown='ignore', sparse_output=False)), ␣
 ↪# Encode features
    ('classifier', DecisionTreeClassifier(random_state=42, max_depth=10,␣
 ↪min_samples_leaf=5, min_samples_split=2))  # Train the Decision Tree
])

# Train the pipeline
cat_pipeline.fit(X_train_cat, y_train)
```

[25]:
```
Pipeline(steps=[('imputer', SimpleImputer(strategy='most_frequent')),
                ('encoder',
                 OneHotEncoder(handle_unknown='ignore', sparse_output=False)),
                ('classifier',
                 DecisionTreeClassifier(max_depth=10, min_samples_leaf=5,
                                        random_state=42))])
```

[26]:
```
# Evaluate the pipeline
train_accuracy_cat = accuracy_score(y_train, cat_pipeline.predict(X_train_cat))
test_accuracy_cat = accuracy_score(y_test, cat_pipeline.predict(X_test_cat))

print("Pipeline with Categorical Features:")
print("Training Accuracy:", train_accuracy_cat)
print("Testing Accuracy:", test_accuracy_cat)
```

```
Pipeline with Categorical Features:
Training Accuracy: 0.8311539406513608
```

```
Testing Accuracy: 0.8317910081074441
```

1.4.5 13. Combining Categorical and Numerical Features with

Use a ColumnTransformer to apply the categorical pre-processing pipeline (composed by an Imputer and one OneHotEncoder) to the categorical attributes of the data, leaving the numerical attribute unaltered (hint: have a look at the remainder parameter of ColumnTransformer)

[27]:
```python
# Define the pipeline for categorical features
cat_pipeline = Pipeline([
    ('imputer', SimpleImputer(strategy='most_frequent')),    # Fill missing␣
 ↪values
    ('encoder', OneHotEncoder(handle_unknown='ignore', sparse_output=False))  #␣
 ↪Encode features
])

# Combine categorical and numerical transformations
preprocessor = ColumnTransformer([
    ('cat', cat_pipeline, categorical_columns),  # Apply categorical pipeline
    ('num', 'passthrough', numerical_columns)    # Leave numerical columns␣
 ↪unaltered
])

# Define the full pipeline with a DecisionTreeClassifier
full_pipeline = Pipeline([
    ('preprocessor', preprocessor),
    ('classifier', DecisionTreeClassifier(random_state=42, **best_params))
])
```

1.4.6 14. Evaluate full pipeline

train and verify the accuracy of the Pipeline. Does the use of both types of features improve the accuracy?

[28]:
```python
# Train the full pipeline
full_pipeline.fit(X_train, y_train)
```

[28]:
```
Pipeline(steps=[('preprocessor',
                 ColumnTransformer(transformers=[('cat',
                                                  Pipeline(steps=[('imputer',
SimpleImputer(strategy='most_frequent')),
                                                                  ('encoder',
OneHotEncoder(handle_unknown='ignore',
 sparse_output=False))]),
                                                  ['workclass', 'education',
                                                   'marital-status',
                                                   'occupation', 'relationship',
                                                   'race', 'sex',
```

```
                                                  'native-country']),
                                                ('num', 'passthrough',
                                                 Index(['age', 'fnlwgt',
'education-num', 'capital-gain', 'capital-loss',
       'hours-per-week'],
      dtype='object'))])),
                ('classifier',
                 DecisionTreeClassifier(max_depth=15, min_samples_leaf=10,
                                        random_state=42))])
```

[29]:
```
# Evaluate the pipeline
train_accuracy_full = accuracy_score(y_train, full_pipeline.predict(X_train))
test_accuracy_full = accuracy_score(y_test, full_pipeline.predict(X_test))

print("Pipeline with Categorical and Numerical Features:")
print("Training Accuracy:", train_accuracy_full)
print("Testing Accuracy:", test_accuracy_full)
```

```
Pipeline with Categorical and Numerical Features:
Training Accuracy: 0.8737954191804755
Testing Accuracy: 0.8597985422979281
```

1.5 Ensembles

1.5.1 15. RandomForestClassifier

[30]:
```
rf_pipeline = Pipeline([
    ('preprocessor', preprocessor),
    ('classifier', RandomForestClassifier(random_state=42, **best_params))
])
rf_pipeline.fit(X_train, y_train)
```

[30]:
```
Pipeline(steps=[('preprocessor',
                 ColumnTransformer(transformers=[('cat',
                                                  Pipeline(steps=[('imputer',
SimpleImputer(strategy='most_frequent')),
                                                                  ('encoder',
OneHotEncoder(handle_unknown='ignore',
 sparse_output=False))]),
                                                  ['workclass', 'education',
                                                   'marital-status',
                                                   'occupation', 'relationship',
                                                   'race', 'sex',
                                                   'native-country']),
                                                ('num', 'passthrough',
                                                 Index(['age', 'fnlwgt',
'education-num', 'capital-gain', 'capital-loss',
       'hours-per-week'],
```

```
      dtype='object'))])),
                ('classifier',
                 RandomForestClassifier(max_depth=15, min_samples_leaf=10,
                                        random_state=42))])
```

[31]:
```
train_accuracy_rf = accuracy_score(y_train, rf_pipeline.predict(X_train))
test_accuracy_rf = accuracy_score(y_test, rf_pipeline.predict(X_test))
print("\nRandom Forest Classifier:")
print("Training Accuracy:", train_accuracy_rf)
print("Testing Accuracy:", test_accuracy_rf)
```

```

Random Forest Classifier:
Training Accuracy: 0.8611558516011029
Testing Accuracy: 0.8661043321595283
```

1.5.2 16. AdaBoostClassifier

[32]:
```
ada_pipeline = Pipeline([
    ('preprocessor', preprocessor),
    ('classifier', AdaBoostClassifier(random_state=42))
])
ada_pipeline.fit(X_train, y_train)
```

[32]:
```
Pipeline(steps=[('preprocessor',
                 ColumnTransformer(transformers=[('cat',
                                                  Pipeline(steps=[('imputer',
SimpleImputer(strategy='most_frequent')),
                                                                  ('encoder',
OneHotEncoder(handle_unknown='ignore',
 sparse_output=False))]),
                                                  ['workclass', 'education',
                                                   'marital-status',
                                                   'occupation', 'relationship',
                                                   'race', 'sex',
                                                   'native-country']),
                                                 ('num', 'passthrough',
                                                  Index(['age', 'fnlwgt',
'education-num', 'capital-gain', 'capital-loss',
       'hours-per-week'],
      dtype='object'))])),
                ('classifier', AdaBoostClassifier(random_state=42))])
```

[33]:
```
train_accuracy_ada = accuracy_score(y_train, ada_pipeline.predict(X_train))
test_accuracy_ada = accuracy_score(y_test, ada_pipeline.predict(X_test))
print("\nAdaBoost Classifier:")
print("Training Accuracy:", train_accuracy_ada)
```

```
print("Testing Accuracy:", test_accuracy_ada)
```

```
AdaBoost Classifier:
Training Accuracy: 0.8600638803199476
Testing Accuracy: 0.8692981737777414
```

1.5.3 17. GradientBoostingClassifier

[34]:
```
gb_pipeline = Pipeline([
    ('preprocessor', preprocessor),
    ('classifier', GradientBoostingClassifier(random_state=42, **best_params))
])
gb_pipeline.fit(X_train, y_train)
```

[34]:
```
Pipeline(steps=[('preprocessor',
                 ColumnTransformer(transformers=[('cat',
                                                  Pipeline(steps=[('imputer',
SimpleImputer(strategy='most_frequent')),
                                                                  ('encoder',
OneHotEncoder(handle_unknown='ignore',
 sparse_output=False))]),
                                                  ['workclass', 'education',
                                                   'marital-status',
                                                   'occupation', 'relationship',
                                                   'race', 'sex',
                                                   'native-country']),
                                                 ('num', 'passthrough',
                                                  Index(['age', 'fnlwgt',
'education-num', 'capital-gain', 'capital-loss',
       'hours-per-week'],
      dtype='object'))])),
                ('classifier',
                 GradientBoostingClassifier(max_depth=15, min_samples_leaf=10,
                                            random_state=42))])
```

[35]:
```
train_accuracy_gb = accuracy_score(y_train, gb_pipeline.predict(X_train))
test_accuracy_gb = accuracy_score(y_test, gb_pipeline.predict(X_test))
print("\nGradient Boosting Classifier:")
print("Training Accuracy:", train_accuracy_gb)
print("Testing Accuracy:", test_accuracy_gb)
```

```
Gradient Boosting Classifier:
Training Accuracy: 0.9594059676230515
Testing Accuracy: 0.8760953238882975
```

1.5.4 Compare Performance

Classifier	Training Accuracy	Testing Accuracy	Notes
DecisionTreeClassifier	86.66%	86.37%	Generalizes well, but performance is limited by lack of ensemble learning.
RandomForestClassifier	85.83%	86.06%	Slight improvement in generalization; avoids overfitting due to bagging.
AdaBoostClassifier	86.02%	86.76%	Boosting improves performance by focusing on harder-to-predict instances.
GradientBoostingClassifier	91.99%	87.78%	Best performer due to sequential boosting and optimized splits.

Key Insights

1. DecisionTreeClassifier:
    - Performs reasonably well but lacks the power of ensemble techniques.
    - Overfits slightly less than Gradient Boosting, but its potential is limited without additional regularization.

2.RandomForestClassifier:

- Improves stability and generalization by combining multiple trees (bagging).
- Testing accuracy is slightly lower than AdaBoost and Gradient Boosting but provides consistent results.

3. AdaBoostClassifier:
    - Slightly outperforms Random Forest in testing accuracy due to its sequential boosting approach.
    - Focuses on misclassified instances, leading to better handling of complex patterns.

4. GradientBoostingClassifier:
    - Achieves the highest testing accuracy (87.78%) by sequentially improving weak learners.
    - The model slightly overfits the training data (91.99% training accuracy), but the testing performance indicates robust learning.

Conclusion

- **Best Model:** GradientBoostingClassifier, with the highest testing accuracy of 87.78%.

- **Balanced Model:** AdaBoostClassifier offers a good balance between training and testing performance with 86.76% testing accuracy.
- **Recommendation:** Gradient Boosting should be the preferred choice for performance-critical tasks, while AdaBoost is a strong alternative for slightly simpler models.