Model Performance Analysis

Evaluation measures of models from task 1, 2 and 3:

<table>
<tr><th rowspan="2">Metric</th><th colspan="2">Decision Tree</th><th colspan="2">Naïve Bayes</th></tr>
<tr><th>From scratch</th><th>Using scikit-learn</th><th>From scratch</th><th>Using scikit-learn</th></tr>
<tr><td>Confusion matrix</td>
<td><table><tr><td></td><td colspan="3">Predicted</td></tr><tr><td rowspan="3">Actual</td><td></td><td>No</td><td>Yes</td></tr><tr><td>No</td><td>5</td><td>31</td></tr><tr><td>Yes</td><td>0</td><td>30</td></tr></table></td>
<td><table><tr><td></td><td colspan="3">Predicted</td></tr><tr><td rowspan="3">Actual</td><td></td><td>No</td><td>Yes</td></tr><tr><td>No</td><td>5</td><td>31</td></tr><tr><td>Yes</td><td>0</td><td>30</td></tr></table></td>
<td><table><tr><td></td><td colspan="3">Predicted</td></tr><tr><td rowspan="3">Actual</td><td></td><td>No</td><td>Yes</td></tr><tr><td>No</td><td>5</td><td>31</td></tr><tr><td>Yes</td><td>0</td><td>30</td></tr></table></td>
<td><table><tr><td></td><td colspan="3">Predicted</td></tr><tr><td rowspan="3">Actual</td><td></td><td>No</td><td>Yes</td></tr><tr><td>No</td><td>5</td><td>31</td></tr><tr><td>Yes</td><td>0</td><td>30</td></tr></table></td>
</tr>
<tr><td>Accuracy</td><td>53.03%</td><td>53.03%</td><td>53.03%</td><td>53.03%</td></tr>
<tr><td>Precision</td><td>0.49</td><td>0.49</td><td>0.49</td><td>0.49</td></tr>
<tr><td>Recall</td><td>1.00</td><td>1.00</td><td>1.00</td><td>1.00</td></tr>
</table>

<table>
<tr><th>Metric</th><th>Random Forest</th></tr>
<tr><td>Confusion matrix</td>
<td><table><tr><td></td><td colspan="3">Predicted</td></tr><tr><td rowspan="3">Actual</td><td></td><td>No</td><td>Yes</td></tr><tr><td>No</td><td>5</td><td>31</td></tr><tr><td>Yes</td><td>0</td><td>30</td></tr></table></td>
</tr>
<tr><td>Accuracy</td><td>53.03%</td></tr>
<tr><td>Precision</td><td>0.49</td></tr>
<tr><td>Recall</td><td>1.00</td></tr>
</table>

Comparative performance and inferences:

All models exhibited the same result for the training dataset provided. Even the confusion matrices for all the models gave similar results. Since all of them indicated similar behavior, it suggest that the issue lies in the quality of the data used for training the models. We decided to examine the training dataset provided(Assignment_train.csv) so that we can gain more insight and make a more informed inference of the evaluation measures. For the Exploratory Data Analysis(EDA) of the dataset, we used the following code to extract different information which we thought would point to the issue:

training_eda.py

```
import pandas as pd
import numpy as np

df = pd.read_csv('Assignment_train.csv')
df.columns = ['Class', 'Age', 'Gender', 'Survived']

unique_tuples = df.drop_duplicates()

print("EDA of training dataset: \n")
print("Total no. of entries = ", len(df))
print("No. of unique entries = ", len(unique_tuples),
"\n")

Y_train = df.iloc[:, -1].values.reshape(-1,1)
Y_train_unique = unique_tuples.iloc[:, -
1].values.reshape(-1,1)

elements, counts = np.unique(Y_train, return_counts=True)

print("Count of each class label in the entire dataset:")
for element, count in zip(elements, counts):
    print(f"\t{element}: {count} occurrences")


tuple_counts =
df.groupby(list(df.columns)).size().reset_index(name='cou
nt')
print("\nUnique tuples in the entire CSV file:")
for index, row in tuple_counts.iterrows():
    print(f"\t{tuple(row[:-1])}: {row['count']}
occurrences")

flat_list = [item for sublist in Y_train_unique for item
in sublist]
unique_occurrences = list(set(flat_list))
print("\nCount of each class label in unique entries:")
for value in unique_occurrences:
    count = flat_list.count(value)
    print(f"\t{value}: {count} occurrences")
```

Output:

EDA of training dataset:

Total no. of entries = 2150
No. of unique entries = 24

Count of each class label in the entire dataset:

no: 1485 occurrences
yes: 665 occurrences

Unique tuples in the entire CSV file:

('1st', 'adult', 'female', 'no'): 4 occurrences
('1st', 'adult', 'female', 'yes'): 122 occurrences
('1st', 'adult', 'male', 'no'): 118 occurrences
('1st', 'adult', 'male', 'yes'): 57 occurrences
('1st', 'child', 'female', 'yes'): 1 occurrences

('1st', 'child', 'male', 'yes'): 5 occurrences
('2nd', 'adult', 'female', 'no'): 13 occurrences
('2nd', 'adult', 'female', 'yes'): 73 occurrences
('2nd', 'adult', 'male', 'no'): 149 occurrences
('2nd', 'adult', 'male', 'yes'): 5 occurrences
('2nd', 'child', 'female', 'yes'): 8 occurrences
('2nd', 'child', 'male', 'yes'): 11 occurrences
('3rd', 'adult', 'female', 'no'): 89 occurrences
('3rd', 'adult', 'female', 'yes'): 76 occurrences
('3rd', 'adult', 'male', 'no'): 387 occurrences
('3rd', 'adult', 'male', 'yes'): 68 occurrences
('3rd', 'child', 'female', 'no'): 17 occurrences
('3rd', 'child', 'female', 'yes'): 14 occurrences
('3rd', 'child', 'male', 'no'): 35 occurrences
('3rd', 'child', 'male', 'yes'): 13 occurrences
('crew', 'adult', 'female', 'no'): 3 occurrences
('crew', 'adult', 'female', 'yes'): 20 occurrences
('crew', 'adult', 'male', 'no'): 670 occurrences
('crew', 'adult', 'male', 'yes'): 192 occurrences

Count of each class label in unique entries:
no: 10 occurrences
yes: 14 occurrences

By looking at the results obtained, we can say that the models were trained on a limited dataset which lacked diversity resulting in substandard results. Out of the 2150 entries given, only 24 entries were unique. In addition to that, the no. of entries having ‘yes’ label was twice as that of the ones having ‘no’ label which is why there are a high number of false positives present in the confusion matrix.

These observations and inferences imply that we should have pre-processed the data before feeding it into the model for instead of directly training the models on raw, unprocessed data.