Bias Demonstration: Gender and Algorithmic Decisions

This notebook illustrates how even simple models trained on toy datasets can reflect or amplify gender bias. It is intended as an educational tool to understand fairness in AI.

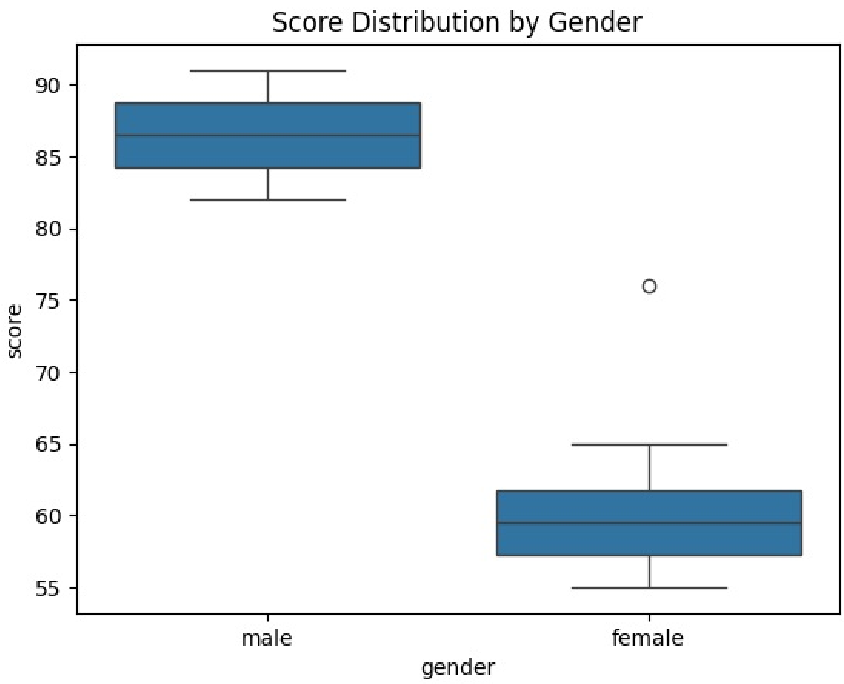
In [1]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import classification_report
from sklearn.model_selection import train_test_split
```

In [2]:
```
# Synthetic dataset: gender and test score
data = pd.DataFrame({
    'gender': ['male', 'female'] * 10,
    'score': [82, 76, 85, 65, 90, 60, 88, 58, 84, 59, 87, 61, 83, 62, 91, 57, 89, 55, 86, 56],
    'admitted': [1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0]
})
data.head()
```

Out[2]:

	gender	score	admitted
0	male	82	1
1	female	76	0
2	male	85	1
3	female	65	0
4	male	90	1

In [3]:
```
sns.boxplot(x='gender', y='score', data=data)
plt.title('Score Distribution by Gender')
plt.show()
```



In [4]:
```
data_encoded = pd.get_dummies(data, drop_first=True)
X = data_encoded[['score', 'gender_male']]
y = data_encoded['admitted']

X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=42)
```

In [5]:
```
model = LogisticRegression()
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       1.00      0.67      0.80         3
           1       0.67      1.00      0.80         2

    accuracy                           0.80         5
   macro avg       0.83      0.83      0.80         5
weighted avg       0.87      0.80      0.80         5
```

In [6]:
```python
coefficients = pd.DataFrame({
    'Feature': X.columns,
    'Coefficient': model.coef_[0]
})
coefficients
```

Out[6]:

	Feature	Coefficient
0	score	0.481269
1	gender_male	0.022677

Ethical Observations

- The gender feature is directly contributing to the prediction.
- This is a toy dataset, but in real settings, such encoding may reinforce structural bias.
- Removing or masking sensitive attributes may not be sufficient — proxy variables (e.g., ZIP code) can still leak bias.

Limitations

- This dataset is synthetic and overly simplified.
- Results are for educational clarity only, and should not be interpreted as representative of real-world outcomes.
- Fairness is a context-dependent issue — what is fair in one domain may not be in another.