Part 3: Practical Audit (25%)

Task: Audit COMPAS Dataset for Bias

Dataset: COMPAS Recidivism Dataset

Goal: Use Python and IBM AI Fairness 360 (AIF360) to detect and visualize racial bias in risk scores.

Step 1: Install and Import Libraries

```
!pip install aif360
import pandas as pd
import numpy as np
from aif360.datasets import BinaryLabelDataset
from aif360.metrics import ClassificationMetric
from aif360.algorithms.preprocessing import Reweighing
```

Step 2: Load Dataset

```
df = pd.read_csv('compas-scores.csv')  # Replace with dataset path
# Convert to BinaryLabelDataset
bld = BinaryLabelDataset(df=df, label_names=['two_year_recid'],
protected_attribute_names=['race'])
```

Step 3: Train a Classifier (Example: Logistic Regression)

```
from sklearn.linear_model import LogisticRegression
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split

X = df.drop(['two_year_recid','race'], axis=1)
y = df['two_year_recid']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
clf = LogisticRegression().fit(X_train_scaled, y_train)
y_pred = clf.predict(X_test_scaled)
```

Step 4: Fairness Evaluation

```
# Convert predictions to BinaryLabelDataset
bld_pred = BinaryLabelDataset(df=pd.DataFrame({'two_year_recid': y_pred,
'race': X_test['race']}), label_names=['two_year_recid'],
protected_attribute_names=['race'])
metric = ClassificationMetric(bld, bld_pred, unprivileged_groups=[{'race':
1}], privileged_groups=[{'race': 0}])
print("False Positive Rate Difference:",
metric.false_positive_rate_difference())
print("Statistical Parity Difference:",
metric.statistical_parity_difference())
```

Step 5: Visualizations

```
import matplotlib.pyplot as plt
fpr_diff = metric.false_positive_rate_difference()
plt.bar(['False Positive Rate Difference'], [fpr_diff])
plt.title('Disparity in FPR by Race')
plt.show()
```

Step 6: Summary Report (300 words)

- Analyze results: note any disparities in FPR, statistical parity.
- Remediation steps: consider reweighing, bias mitigation algorithms in AIF360, or retraining with balanced data.
- Discuss potential ethical impacts of biased recidivism predictions on minority groups.

Deliverable: 1. Python code implementing bias detection and visualization. 2. Report summarizing findings, observed disparities, and proposed remediation steps for mitigating racial bias in predictive risk scores.