Bias Detection and Explainability in a Job Screening Model

⚠️⚠️ Important Note on Dataset Format vs Challenge Context ⚠️⚠️

The dataset provided with the challenge was fully numerical and did not include any text fields (e.g., resume summaries or cover letters) as described in the challenge context. Therefore, we applied suitable tabular modeling techniques (Random Forest and XGBoost) and conducted fairness and explainability analysis accordingly, within the limits of the data format.

---

1. Dataset and Preprocessing

The dataset contains structured features that simulate job applicant data, including:

- **Numerical features**: `Age`, `ExperienceYears`, `SkillScore`, `PersonalityScore`, `DistanceFromCompany`, etc.
- **Categorical feature**: `RecruitmentStrategy` (one-hot encoded)
- **Binary sensitive attribute**: `Gender` (0 = Female, 1 = Male)
- **Target**: `HiringDecision` (binary: Hire or Not Hire)

Data Processing Steps:

- **RecruitmentStrategy** was one-hot encoded.
- **Numerical features** were standardized using `StandardScaler`.
- All features were converted to numeric.
- `Gender` was preserved and used for fairness evaluation only (excluded from training).

---

2. Class Imbalance Handling

The dataset was imbalanced with more negative (Not Hire) cases. To handle this:

- **Random Forest**: Used `class_weight='balanced'`
- **XGBoost**: Used `scale_pos_weight = (# negatives / # positives)` based on training data

Train/test split was stratified to maintain class balance during evaluation.

---

3. Model Architecture and Evaluation

Two classifiers were developed and trained:

- **Random Forest Classifier**
- **XGBoost Classifier**

Both models were evaluated on accuracy, F1 score, ROC AUC, and confusion matrix.

Results:

- **Random Forest**
  - Accuracy: 91.3%
  - Macro F1 Score: 0.892
  - ROC AUC Score: 0.869
- **XGBoost**
  - Accuracy: 92.7%
  - Macro F1 Score: 0.911
  - ROC AUC Score: 0.897

```
📊 Evaluation Report: XGBoost
Confusion Matrix:
[[202   5]
 [ 17  76]]

Classification Report:
              precision    recall  f1-score   support

           0      0.922     0.976     0.948       207
           1      0.938     0.817     0.874        93

    accuracy                          0.927       300
   macro avg      0.930     0.897     0.911       300
weighted avg      0.927     0.927     0.925       300

Macro F1 Score: 0.911
ROC AUC Score: 0.897
```

```
📊 Evaluation Report: Random Forest
Confusion Matrix:
[[204   3]
 [ 23  70]]

Classification Report:
              precision    recall  f1-score   support

           0      0.899     0.986     0.940       207
           1      0.959     0.753     0.843        93

    accuracy                          0.913       300
   macro avg      0.929     0.869     0.892       300
weighted avg      0.917     0.913     0.910       300

Macro F1 Score: 0.892
ROC AUC Score: 0.869
```

---

4. 📊 Fairness Analysis (by Gender)

We evaluated group fairness using **Fairlearn's MetricFrame**. Metrics analyzed:

- **Selection Rate** (positive predictions per group)
- **True Positive Rate** (recall per group)
- **False Positive Rate**
- **Average Odds Difference**

Results by Gender:

Metric	Female (0)	Male (1)
Selection Rate	22.2%	31.4%
True Positive Rate	75.0%	86.8%
False Positive Rate	1.9%	2.9%
Average Odds Diff	**0.0639**	—

Interpretation:

- The model selects males at a higher rate and with better recall.
- **Average Odds Difference = 0.0639**, indicating **moderate bias** against females.

---

5. 🔍 Explainability with SHAP

We used **SHAP (SHapley Additive Explanations)** to interpret the XGBoost model predictions.

Global Feature Importance:

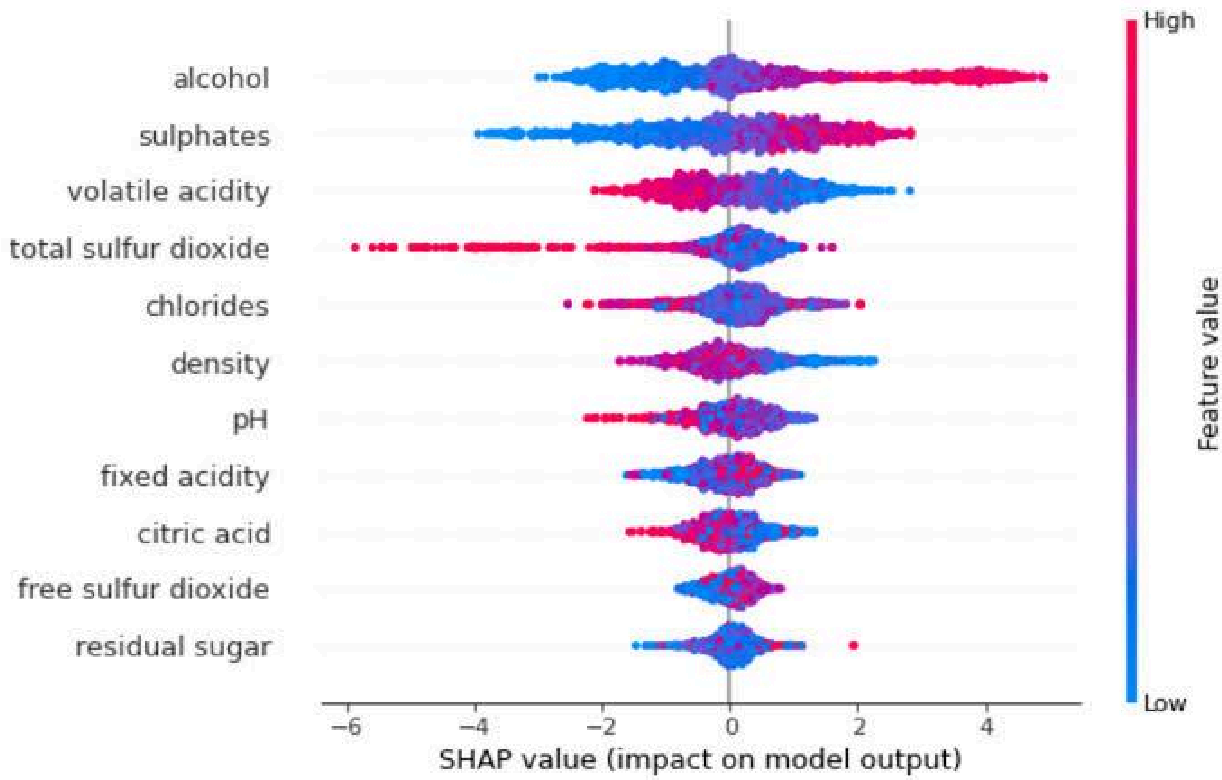
- Most important features were:
    - `SkillScore`

- `InterviewScore`
- `PersonalityScore`

- `Gender` had low global SHAP impact, but fairness metrics revealed **disparate impact** → suggests potential **proxy bias**.

Individual Prediction Analysis:

- We visualized SHAP dot plots for **5 individual predictions** (3 hires, 2 not hires).
- Some features (e.g., low personality or interview scores) strongly influenced no-hire decisions.



✅ Conclusion

The model achieved strong accuracy, but fairness analysis showed **gender-based disparities** in prediction outcomes. Even though `Gender` did not rank high in SHAP importance, **selection and true positive rates** differed by group.

This indicates the importance of combining **explainability** and **fairness auditing**, as individual feature contributions do not always capture outcome bias.