Bias Detection and Explainability in AI Models

1. Dataset Description and Sensitive Features

The dataset consists of structured information about job applicants, including features such as Age, Gender, EducationLevel, ExperienceYears, SkillScore, and InterviewScore. The target variable is HiringDecision (0 = Not Hire, 1 = Hire). The Gender feature was used as the sensitive attribute to evaluate model fairness.

2. Model Architecture and Performance

We used a Random Forest Classifier due to the structured nature of the dataset. The data was split into 80% training and 20% test sets. Model performance was evaluated using classification metrics. The classifier showed balanced performance across classes, although some bias was observed initially across gender groups.

3. Fairness Analysis

We calculated Demographic Parity to assess group fairness. Initial results showed a disparity between Male and Female hiring rates. To mitigate this, reweighing was applied during training, giving higher weights to the underrepresented gender group. This reduced the disparity without significantly harming performance.

4. Explainability Results and Bias Attribution

We used SHAP (SHapley Additive Explanations) to explain the model's decisions. SHAP summary plots highlighted the most influential features contributing to hiring decisions. The most impactful features included InterviewScore, SkillScore, and ExperienceYears. Feature importance showed minimal direct influence from Gender after mitigation.

5. Bias Mitigation Results and Tradeoffs

After applying reweighing, the gap in demographic parity decreased. Although a slight drop in accuracy was noted, fairness improved. This tradeoff is acceptable in high-stakes applications like hiring, where fairness is crucial.