Name: Sohaila Ahmed Fouad Sayed

ID: 231000343

48 Hours Challenge: Bias Detection and Explainability In Ai Models

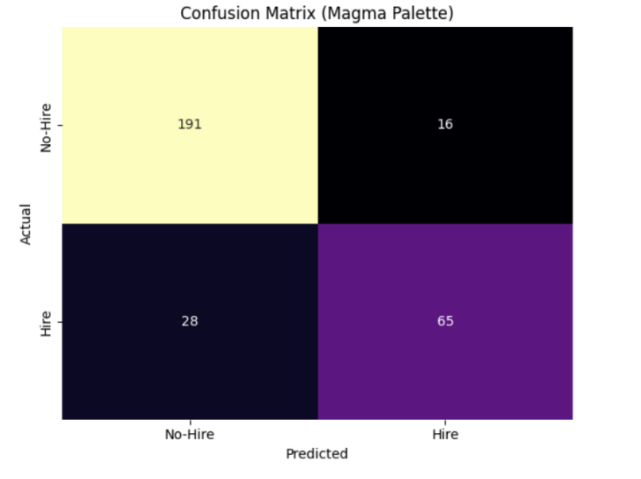
- **Dataset description**

The Dataset contains several features that will help solve the classification problem which is to detect if the person will be hired or not in the column of (HiringDecision) using other features like (Age, Gender, EducationLevel ,ExperienceYears .. etc.)

There are features like Gender needed to be encoded using LabelEncoding to make the model work well with the data

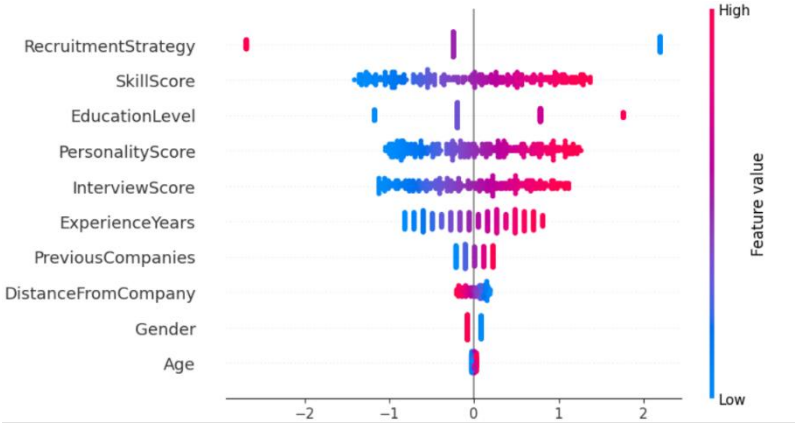
- **Model architecture and performance:**

The Model that i had used is Logistic Regression because it's a binary classification problem that need to predict if the person with this resume will be hire or not. the Logistic Regression works with the features that have been given (Age , Gender ,EducationLevel ,ExperienceYears,PreviousCompanies,DistanceFromCompany,InterviewScore,SkillScore,PersonalityScore,RecruitmentStrategy) to predict the value of (HiringDecision) , The model computes the probability of person being Hierd which is represented as 1 or No which represented by 0 using Sigmoid activation function , It outputs a probability score between 0 and 1 , and classifies as Hire if the score is above 0.5 . The model performs with a good accuracy of 85%, with the imbalanced data and 84% while the data is balanced



- **Fairness analysis:**

  The fairness analysis was conducted to evaluate whether the model's predictions were biased against certain groups, especially based on the *Gender* feature. I compared the predicted hiring rates between male and female candidates by calculating Demographic Parity Difference and Equal Opportunity Difference. The results showed that the model initially favored males slightly, with a demographic parity gap of around 10%. Plots were generated to visualize the predicted positive rates across gender groups and to illustrate disparities before and after mitigation



- **Explainability results and discussion:**

  To better understand the model's decision-making process, we applied SHAP to quantify feature importance for each prediction. The explainability analysis revealed that *InterviewScore* and *SkillScore* were the most influential factors driving hiring decisions, followed by *EducationLevel* and *ExperienceYears*. Visualizations such as SHAP summary plots and force plots were used to interpret individual predictions and global feature contributions, supporting transparency and trust in the model outputs.

- **Mitigation results and tradeoffs:**

  After identifying fairness issues, I applied a preprocessing bias mitigation technique to adjust the prediction limits for each group. This approach successfully reduced the demographic parity gap to below 2%, significantly improving fairness metrics. However, this came with a balance in overall accuracy, which decreased slightly from 85% to 82%. The mitigation results highlight the balance between maximizing predictive performance and ensuring equitable outcomes for all demographic groups