Biases in AI Recruitment Tools

Priyanka Saha

Master of Data Science

Regis University

MSDS 640

Prof. Ghulam Mujtaba

August 25, 2025

Abstract

AI has become an integral part of our day to day lifestyle. We are using it to resolve our requirements like if we want to generate an image with description etc. Similarly, different companies are using AI recruitment tools to help their recruiters. This AI tools are able to scan many resumes within short period of time and able to give a suggestion about whether a resume matches the requirements. However, sometimes these tools produce biased result. It can be biased against sensitive gender, race, ethnicity or age group. With the biased result, it can influence the recruiter to select candidates from the favored groups only. This is wrong in two ways - eligible candidates are rejected if they belong in sensitive group and ineligible candidates are selected if they belong in favored group. This creates huge ethical violation and as human being who abide by ethics, we need to find a solution for this.

Biases in AI Recruitment Tools

Introduction

There are many incidents where different companies got a lawsuit as they are recruiting candidates from AI tool's favored groups only. Many of them had to settle these lawsuit with huge amount of money. This is bad for both the companies and the candidates who are rejected although they were eligible. These incidents broke their confidence.

There can be many reasons for these AI tools to generated biased results. They can be trained on biased dataset. A biased dataset is a dataset where most of the successful records are from favored group's candidates with very less successful records are from sensitive groups. Trained on this kind of dataset, AI tool concludes that eligible candidates can only be from favored group.

Sometimes, although the AI tool is trained on unbiased dataset, it still can acquire biases in its life-cycle and generate biased results.

Here, we have taken a resume dataset from kaggle and tried to remove all the biases from this dataset and produced an AI recruitment tool that will produce unbiased results. Our main focus was to remove gender bias and also look for other biases as well. Also, we have kept in mind that we do not introduce any new biases when mitigating oneNellore (2025)

Methodology and Results

The dataset is consists Job Applicant Name, Age, Gender, Race, Ethnicity, Resume, Job Roles, Job Description, Best Match features. The data is labeled. So, we have used supervised learning. Here, we have two decision making features: Best Match - denotes whether a candidate is selected or not, Job Roles - denotes the role based on resume. So, we created our tools making both of them targets one by one and compared their accuracy and fairness.

Before we move forward, below are some EDA:

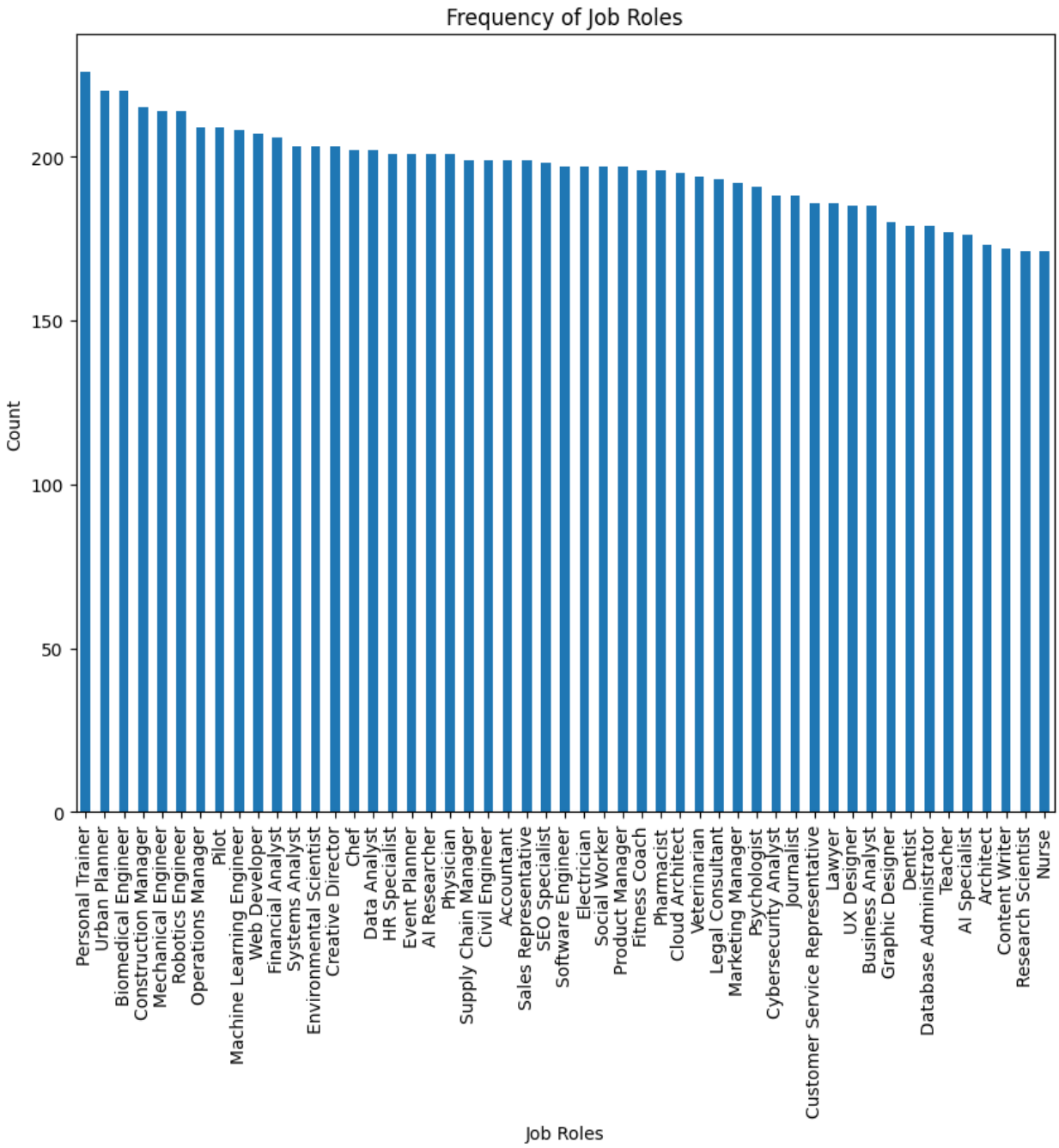

Figure 1

Frequency of Job Roles.

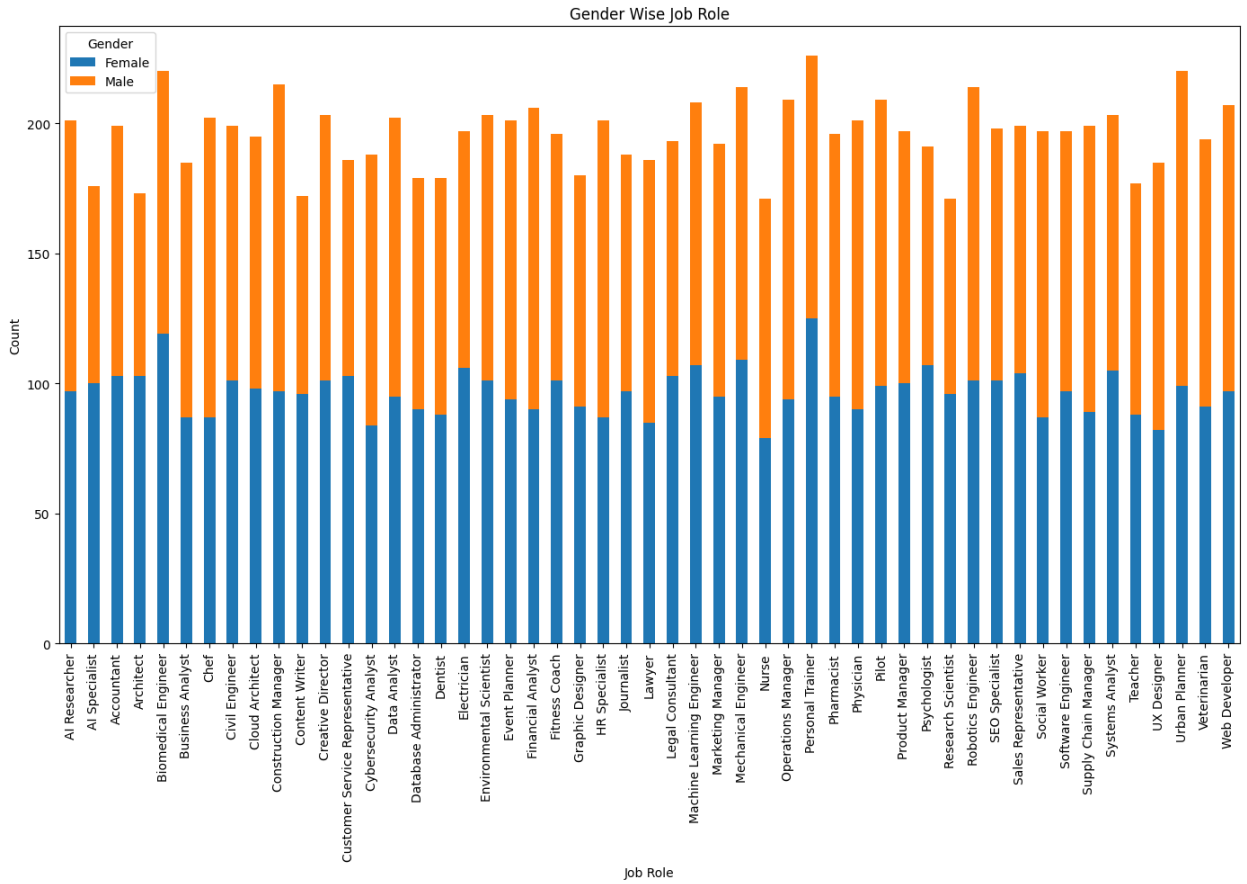

Figure 2

Gender Wise Job Roles.

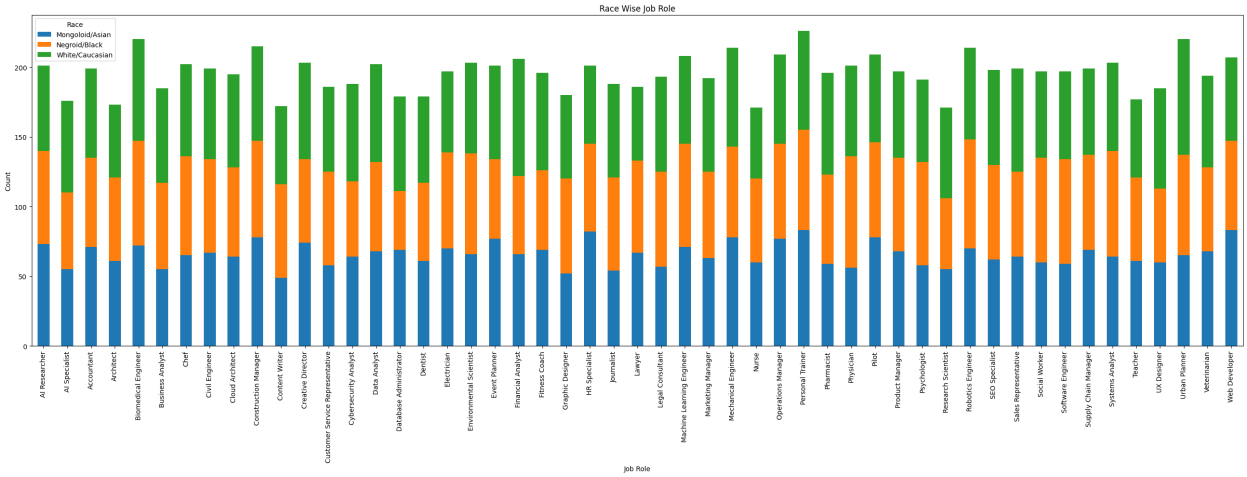

Figure 3

Race Wise Job Roles.

We can see that data is well distributed based on Gender, Race over Job Roles.

Target is Best Match

Here, we proceeded making Best Match feature as target. So, our model will predict whether a candidate is selected or not.

Data Cleaning and Pre-Processing

First, we have removed urls, punctuations, extra space etc from the Resume feature. Then we have removed stop words, most common words and stemmed that feature. After this we move forward with model creation.

Model Generation

We have used "Logistic Regression", "Decision Tree", "Gradient Boosting", "KNN" learning models. Lets see the accuracy below:

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.517000	0.516914	0.517000	0.516952
Decision Tree	0.524000	0.523924	0.524000	0.523958
Gradient Boosting	0.508000	0.503971	0.508000	0.495065
KNN	0.514667	0.515045	0.514667	0.514773

Figure 4

Accuracy of The Generated Models.

Decision Tree has the highest accuracy but none of the accuracy are high.

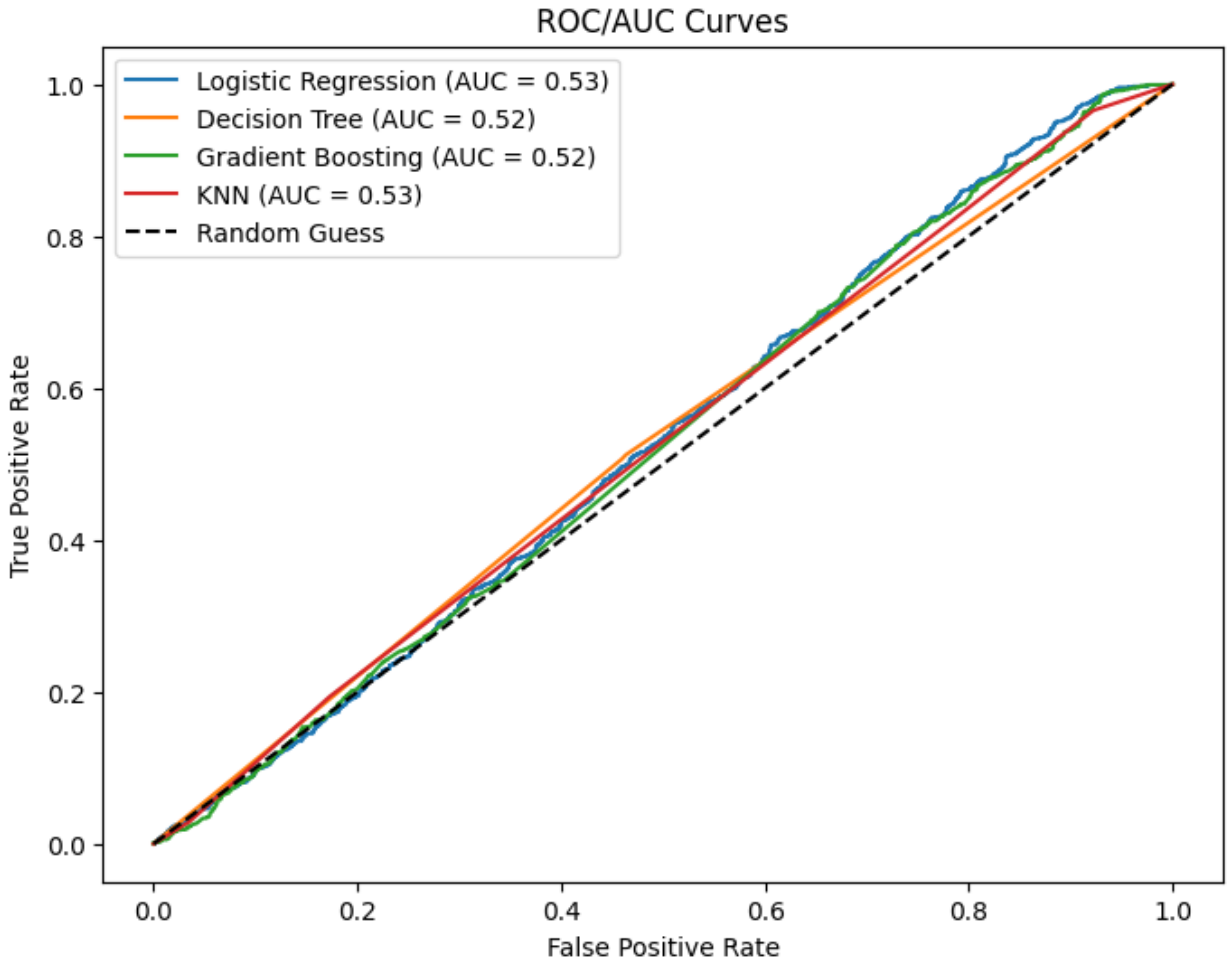

Figure 5

ROC Graph.

ROC graph looks Okay.

Fairness Metrics

Let's see the result of fairness metrics before and after mitigation.

```
        accuracy  selection_rate
Gender
Female  0.491373        0.456867
Male    0.522244        0.454545


After Mitigation:

        accuracy  selection_rate
Gender
Female  0.498965        0.460317
Male    0.523533        0.446809
```

Figure 6

Gender - Fairness metrics before and after mitigation.

There selection rates were already almost equal. Mitigation only increases the difference between selection rates.

```
                accuracy  selection_rate
Race
Mongoloid/Asian  0.517553        0.463390
Negroid/Black    0.487127        0.452111
White/Caucasian  0.516473        0.451550


After Mitigation:

                accuracy  selection_rate
Race
Mongoloid/Asian  0.516550        0.462387
Negroid/Black    0.486097        0.451081
White/Caucasian  0.517442        0.452519
```

Figure 7

Race - Fairness metrics before and after mitigation.

There selection rates were already almost equal. Mitigation did not improve much.

Ethnicity	accuracy	selection_rate
African	0.507692	0.384615
Caribbean	0.546053	0.460526
Chinese	0.541667	0.409722
Dutch	0.541667	0.423611
English	0.503226	0.470968
Ethiopian	0.503497	0.496503
Filipino	0.514493	0.478261
French	0.489655	0.393103
German	0.507463	0.440299
Ghanaian	0.480000	0.406667
Indian	0.488722	0.436090
Irish	0.538462	0.461538
Italian	0.531915	0.489362
Jamaican	0.395349	0.457364
Japanese	0.466216	0.493243
Kenyan	0.512605	0.420168
Korean	0.500000	0.480519
Nigerian	0.459459	0.527027
Polish	0.500000	0.479167
Thai	0.508333	0.458333
Vietnamese	0.593750	0.481250

Figure 8

Ethnicity - Fairness metrics before mitigation.

Candidates of African and French ethnicity have 2 lowest selection rates.

After Mitigation:

Ethnicity	accuracy	selection_rate
African	0.538462	0.415385
Caribbean	0.552632	0.467105
Chinese	0.513889	0.437500
Dutch	0.548611	0.444444
English	0.509677	0.451613
Ethiopian	0.503497	0.496503
Filipino	0.514493	0.449275
French	0.468966	0.400000
German	0.492537	0.455224
Ghanaian	0.493333	0.420000
Indian	0.511278	0.413534
Irish	0.532544	0.479290
Italian	0.539007	0.496454
Jamaican	0.410853	0.426357
Japanese	0.466216	0.479730
Kenyan	0.521008	0.445378
Korean	0.506494	0.487013
Nigerian	0.452703	0.479730
Polish	0.520833	0.458333
Thai	0.516667	0.450000
Vietnamese	0.556250	0.468750

Figure 9

Ethnicity - Fairness metrics after mitigation.

Mitigation increases the selection rates of African and French candidates.

Target is Job Roles

Here, we proceeded making Job Roles feature as target. So, our model will predict the job role based on the resume.

Data Cleaning and Pre-Processing

First, we have removed urls, punctuations, extra space etc from the Resume feature. Then we have removed stop words, biased words like "Women", "Male" etc. and stemmed

that feature. Then we removed the most common words in the dataset. After that we check for similar resumes in same Job Role and if we found a match of 70% or more, we have removed the second record. Then we dropped unnecessary features - "Job Applicant Name", "Age", "Gender", "Race", "Ethnicity", "Job Description", "Best Match" from the dataset. After that we move forward with model creation.

Model Generation

We have used "Logistic Regression", "Decision Tree", "Gradient Boosting", "KNN" learning models. Lets see the accuracy below:

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.977887	0.978724	0.977887	0.977246
Decision Tree	0.823096	0.871997	0.823096	0.830561
Gradient Boosting	0.879607	0.919330	0.879607	0.887633
KNN	0.975430	0.977682	0.975430	0.974585

Figure 10

Accuracy of The Generated Models.

Logistic Regression model has highest accuracy.

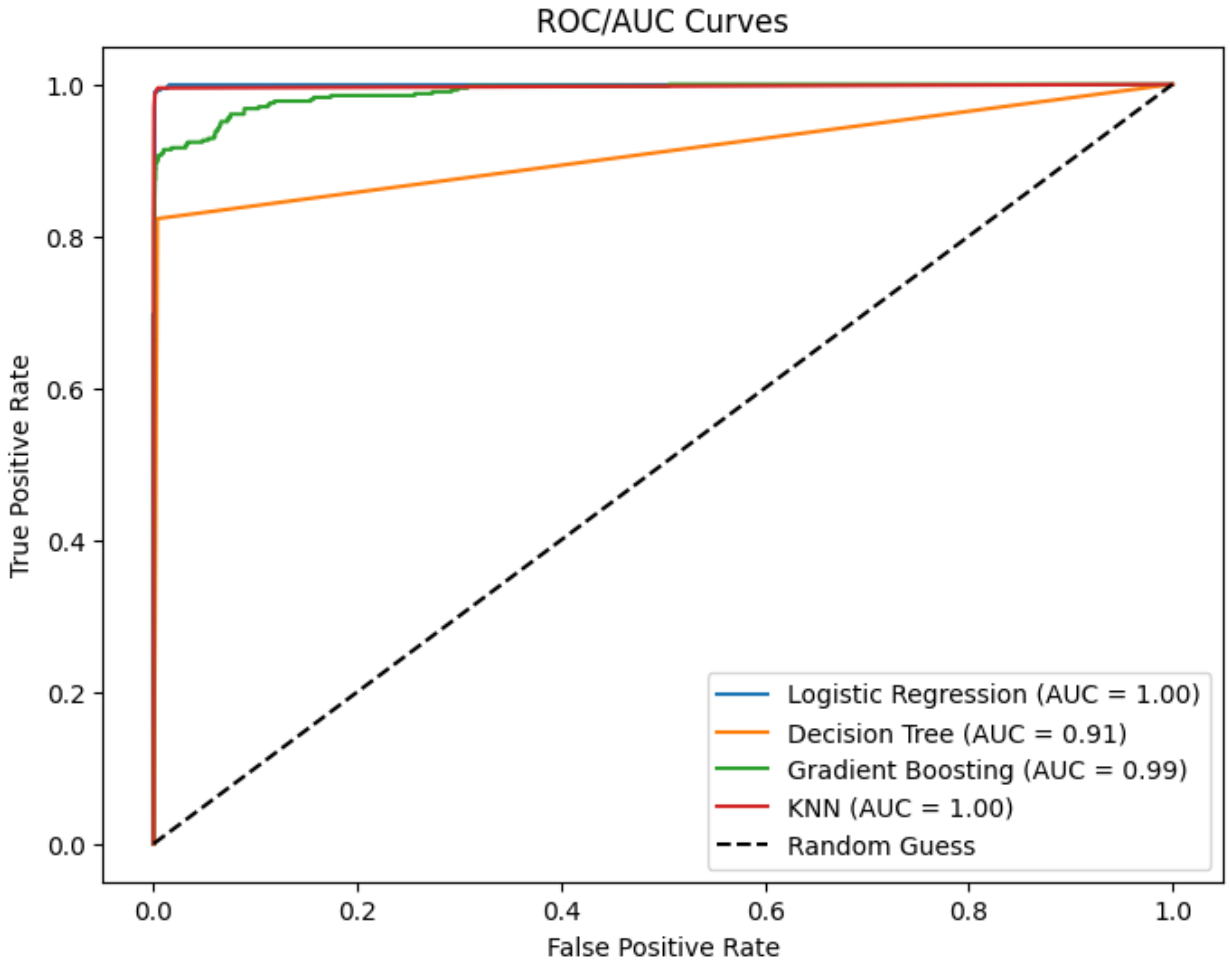

Figure 11

AUC Graph.

AUC graph looks fine.

Fairness Metrics

To run fairness metrics, the dataset needs to have sensitive feature, keywords in it. Here, we have already removed those and we did not let our models learn about sensitive groups. So, fairness metrics is not applicable here.

Discussion and Conclusions

We should always check the impact of mitigator. That means, we should always check the results of fairness before and after applying mitigator. In Figure 6, we can see that

mitigator only worsen the selection rates by increasing the gap between Male and Female. On a contrast, mitigator balanced the selection rate for African and French candidates. So, Here we are able to see biases against African and French candidates and we have mitigated that.

However, if we compare the accuracies, them models with Job Roles as target has much higher accuracy than models with Best Match as target. Also, we had already removed sensitive keywords, features from Models with Job Roles. So, Models with Job Roles never learn about sensitive groups. To run fairness metrics, the dataset needs to have sensitive feature, keywords in it. So, fairness metrics is not applicable on models with Job Roles as target. So, Logistic Regression model with target as Job Roles gave the highest accuracy.

References

Nellore, S. K. (2025, Feb). *Recruitment dataset.* Retrieved from

`https://www.kaggle.com/datasets/surendra365/recruitement-dataset`