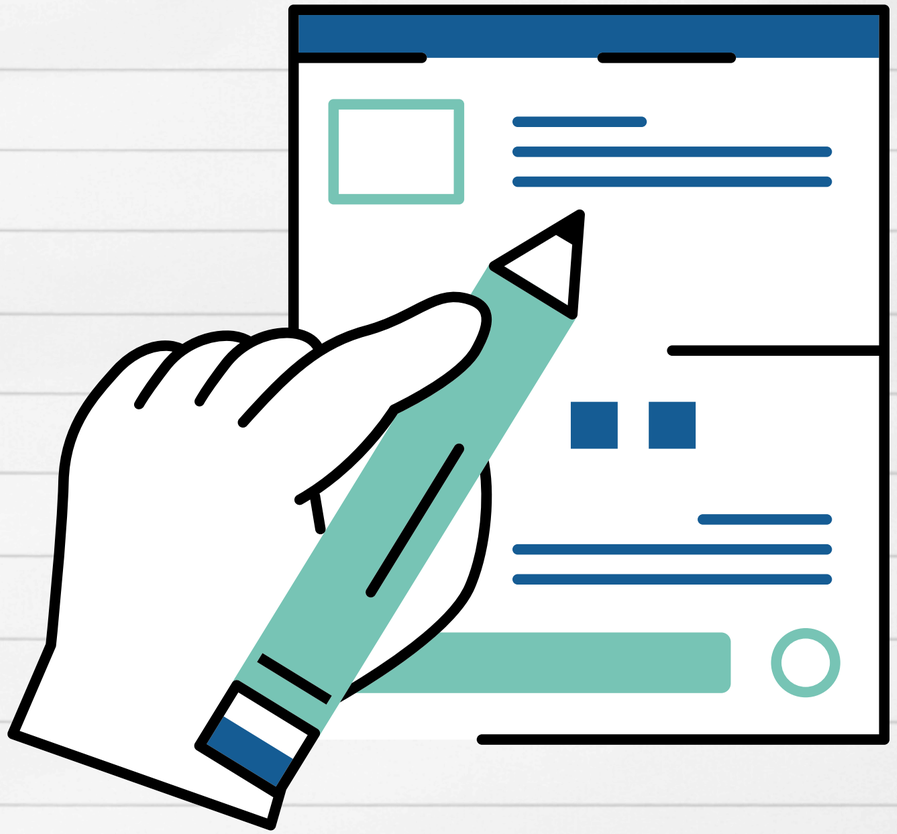
IE0005 MINI PROJECT PRESENTATION

GROUP MIST

CONTENT

- Dataset Overview & objective
- Data Exploratory analysis
- Data preparation
- Prediction
- Ethical Considerations and Bias Mitigation
- Project Outcome & conclusion



SOURCE

70k+ Job Applicants Data (Human Resource)

Unleashing Employability Insights: Analyzing 70K Job Applicants for Optimal Hiri

Data Card | Code (9) | Discussion (4) | Suggestions (0)

About Dataset

Introduction:

The dataset titled "Employability Classification of Over 70,000 Job Applicants" contains a comprehensive collection of information regarding job applicants and their respective employability scores. The dataset has been compiled to assist organizations and recruiters in evaluating the suitability of candidates for various employment opportunities. By utilizing machine learning techniques, this dataset aims to provide valuable insights into the factors influencing employability and enhance the efficiency of the hiring process.

From the survey results, we have built a dataset with the following columns:

- **Age**: age of the applicant, >35 years old or <35 years old *(categorical)*
- **EdLevel**: education level of the applicant (Undergraduate, Master, PhD...) *(categorical)*
- **Gender**: gender of the applicant, (Man, Woman, or NonBinary) *(categorical)*
- **MainBranch**: whether the applicant is a profesional developer *(categorical)*
- **YearsCode**: how long the applicant has been coding *(integer)*
- **YearsCodePro**: how long the applicant has been coding in a professional context, *(integer)*
- **PreviousSalary**: the applicant's previous job salary *(float)*
- **ComputerSkills**: number of computer skills known by the applicant *(integer)*
- **Employed**: target variable, whether the applicant has been hired *(categorical)*

Usability ⓘ
8.82

License
Unknown

Expected update frequency
Annually

Tags

Business | Education | Tabular | Universities and Colleges | Standardized Testing | Data Analytics | Binary Classification

[70k Job Applicants Data for Human Resource]
(https://www.kaggle.com/datasets/ayushtankha/70k-job-applicants-data-human-resource/data)

OBJECTIVE

MAIN OBJECTIVE

Develop a predictive model to assess the employability of job applicants based on their demographic, educational, and professional background, as well as their skills and experience.

IMPACT

The model should provide actionable insights to streamline the hiring process, improve candidate selection accuracy, and enhance overall recruitment efficiency.

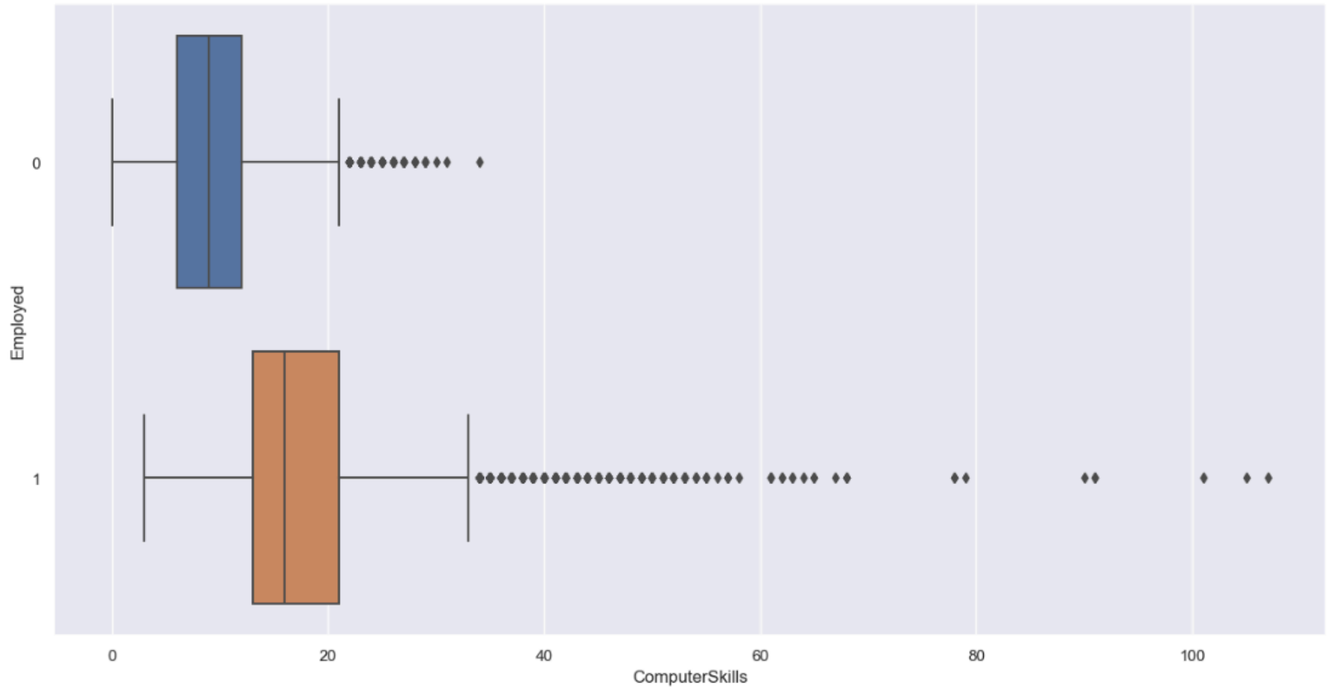
DATA EXPLORATORY ANALYSIS

```
data.info()
```

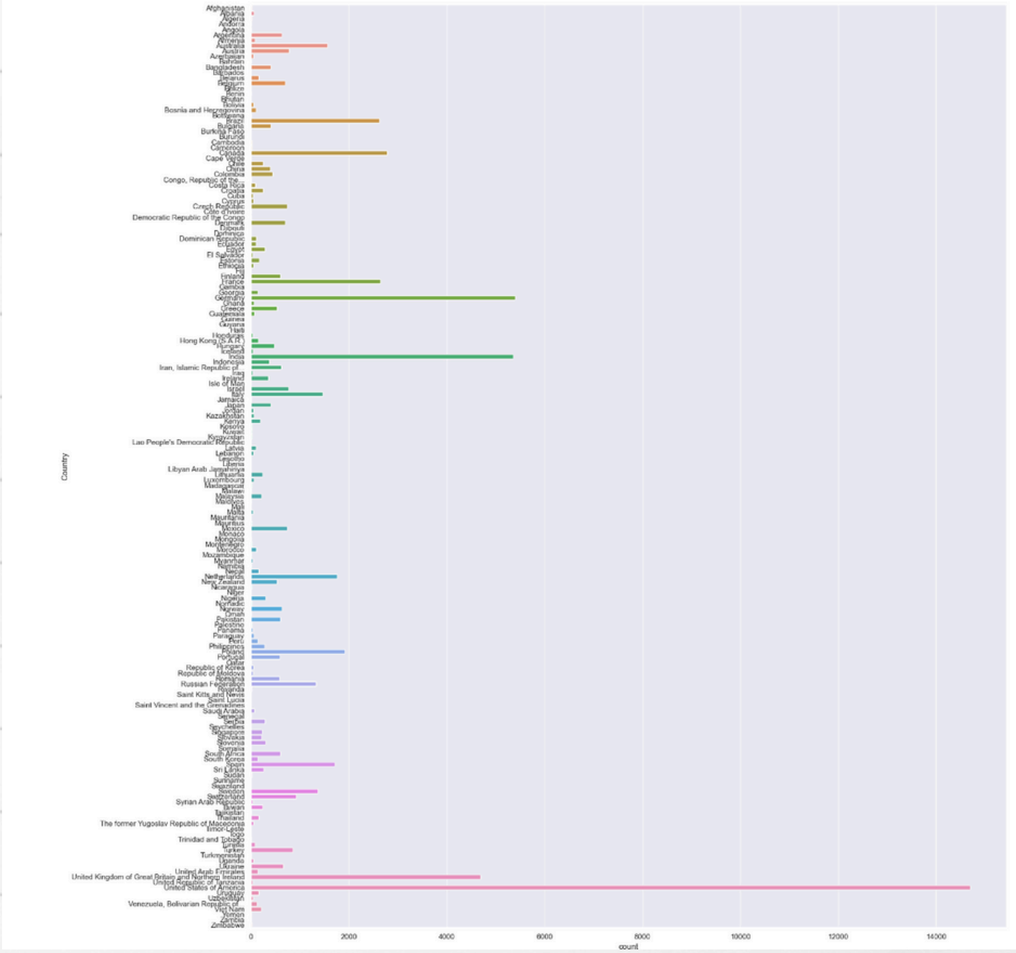
```
<class 'pandas.core.frame.DataFrame'>

RangeIndex: 73462 entries, 0 to 73461

Data columns (total 15 columns):

 #   Column           Non-Null Count  Dtype

---  ------           --------------  -----

 0   Unnamed: 0       73462 non-null  int64

 1   Age              73462 non-null  object

 2   Accessibility    73462 non-null  object

 3   EdLevel          73462 non-null  object

 4   Employment       73462 non-null  int64

 5   Gender           73462 non-null  object

 6   MentalHealth     73462 non-null  object

 7   MainBranch       73462 non-null  object

 8   YearsCode        73462 non-null  int64

 9   YearsCodePro     73462 non-null  int64

 10  Country          73462 non-null  object

 11  PreviousSalary   73462 non-null  float64

 12  HaveWorkedWith   73399 non-null  object

 13  ComputerSkills   73462 non-null  int64

 14  Employed         73462 non-null  int64

dtypes: float64(1), int64(6), object(8)
```

CHECKING RELATIONSHIP COMPARING IT TO EMPLOYED

[36]: <Axes: xlabel='ComputerSkills', ylabel='Employed'>



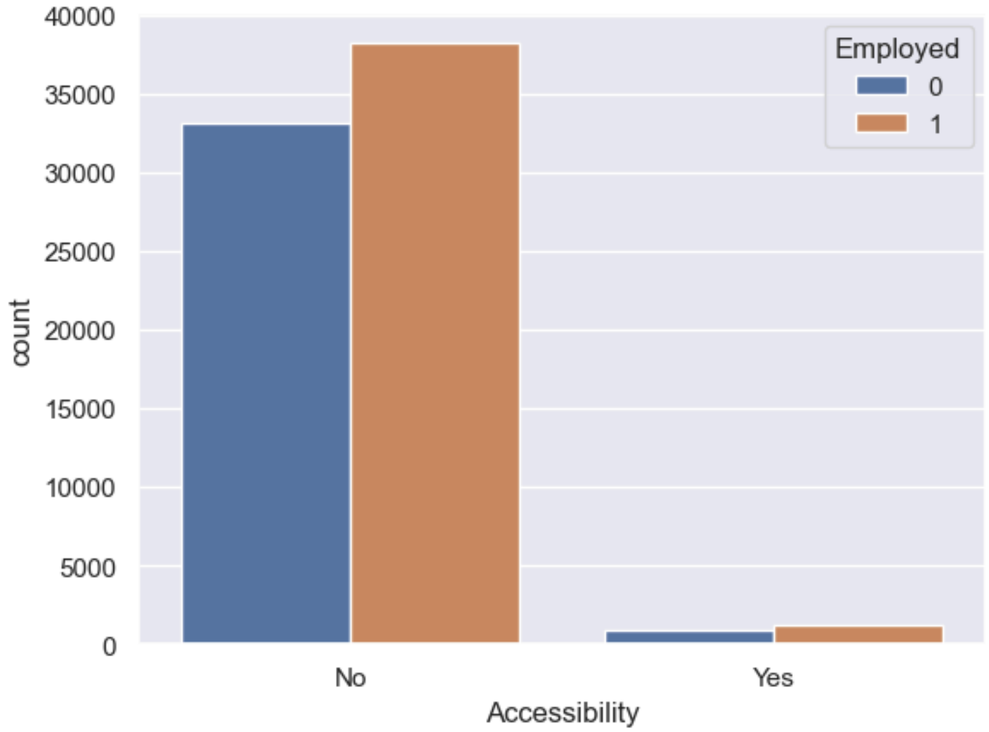
[37]: <seaborn.axisgrid.FacetGrid at 0x1ca71369750>



TRYING TO FIND OUT WHAT IS ACCESSIBILITY

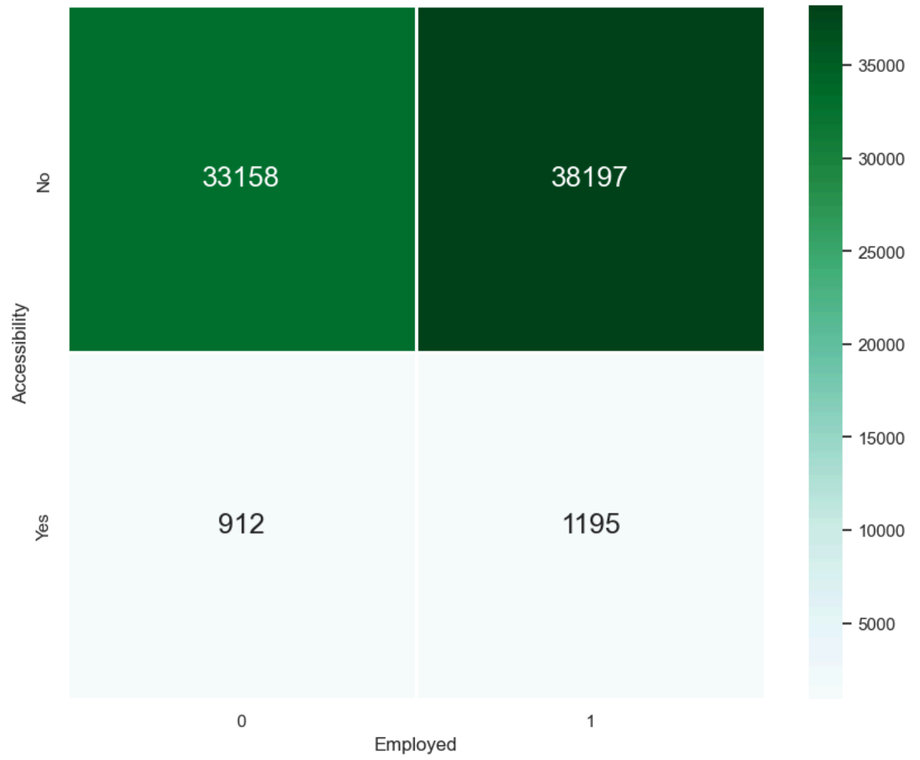
```
[40]: sb.countplot(dataframe, x="Accessibility",hue = "Employed")
```

```
[40]: <Axes: xlabel='Accessibility', ylabel='count'>
```




```
[39]: f, axes = plt.subplots(1, 1, figsize=(10, 8))
sb.heatmap(dataframe.groupby(['Accessibility', 'Employed']).size().unstack(),
           linewidths = 1, annot = True, fmt = 'g', annot_kws = {"size": 18}, cmap = "BuGn")
```

```
[39]: <Axes: xlabel='Employed', ylabel='Accessibility'>
```



DATA CLEANING

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>

RangeIndex: 73462 entries, 0 to 73461

Data columns (total 15 columns):

 #   Column          Non-Null Count  Dtype

---  ------          --------------  -----

 0   Unnamed: 0      73462 non-null  int64

 1   Age             73462 non-null  object

 2   Accessibility   73462 non-null  object

 3   EdLevel         73462 non-null  object
```

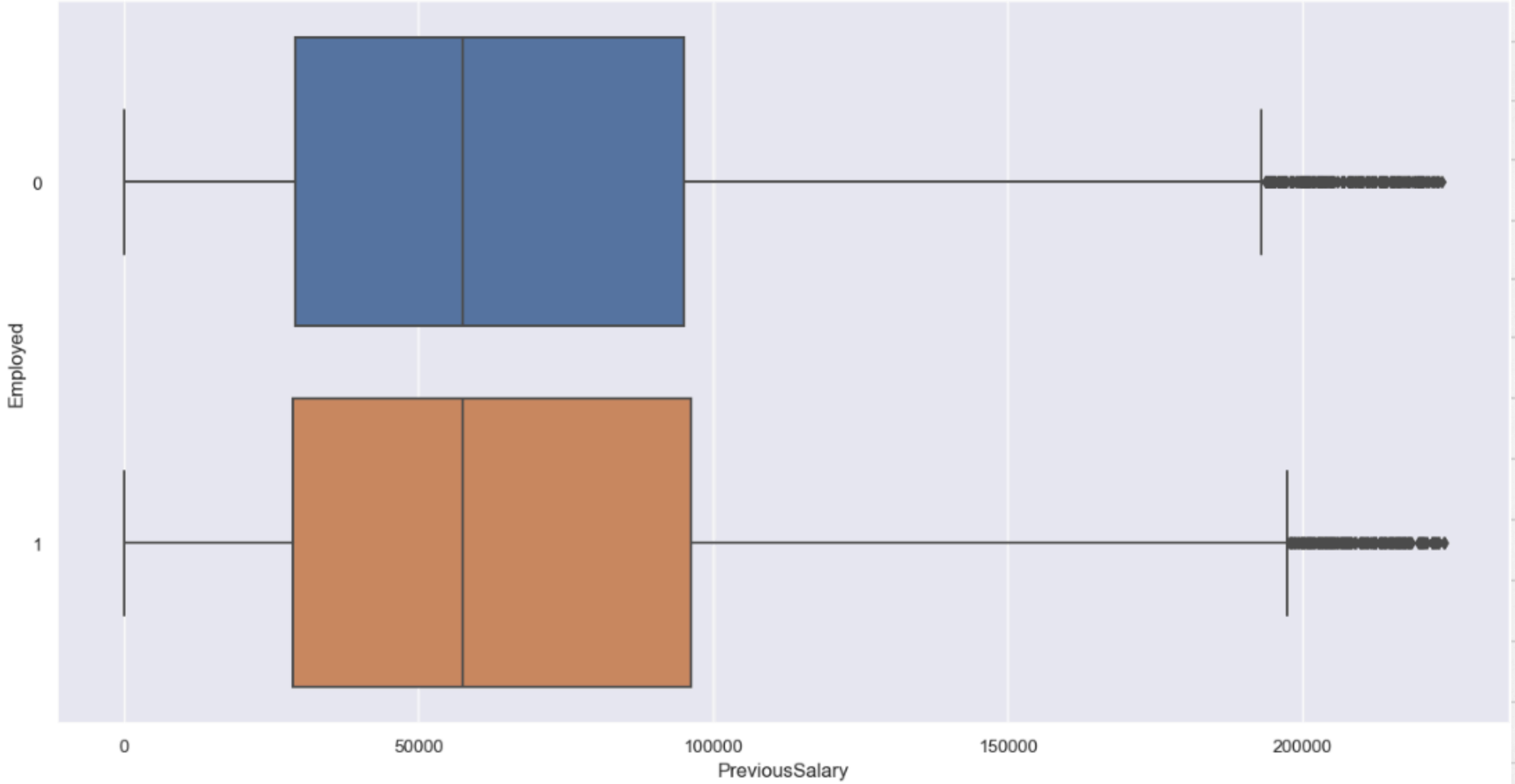
[25]:

```
#removing the unnamed column as we defininitely do not need that
dataframe = pd.DataFrame(data[['Age', 'Accessibility', 'EdLevel',
                               'Employment', 'Gender', 'MentalHealth',
                               'MainBranch', 'YearsCode', 'YearsCodePro',
                               'Country', 'PreviousSalary', 'HaveWorkedWith',
                              'ComputerSkills', 'Employed']])
dataframe.head()
```

exploring previous salary

[41]:
```
f, axes = plt.subplots(1, 1, figsize=(16, 8))
sb.boxplot(x = 'PreviousSalary', y = 'Employed', data = dataframe)
```

[41]:
```
<Axes: xlabel='PreviousSalary', ylabel='Employed'>
```



VARIABLES THAT MIGHT HELP PREDICT IF THE PERSON WILL BE HIRED OR NOT

- Computer Skills and Country
- YearsCodePro and Edlevel

TYPE OF ML PROJECT USED

CLASSIFICATION

REGRESSION

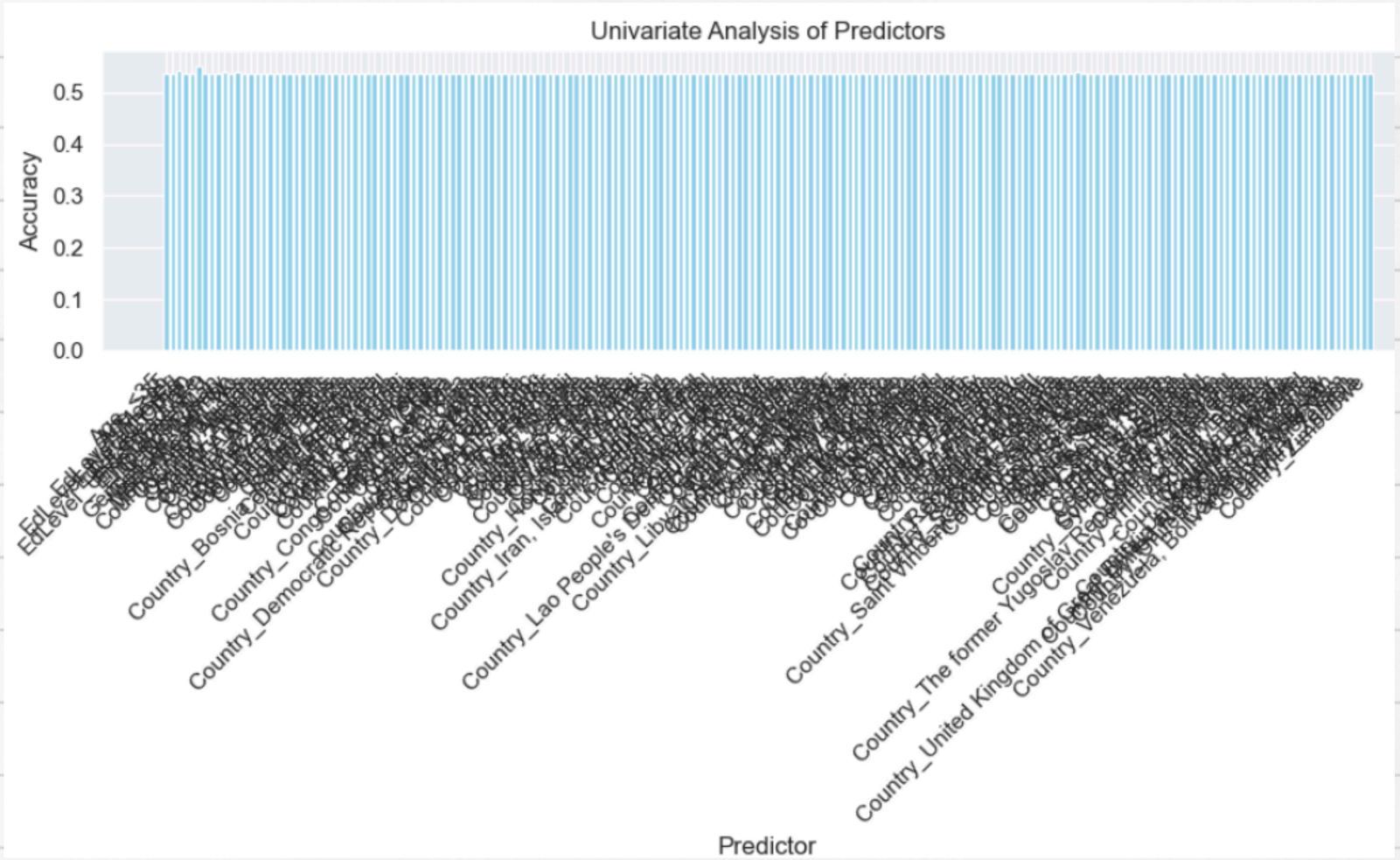
UNIVARIATE PREDICTIONS

```
[49]: unidata = pd.DataFrame(clean[['Age', 'EdLevel','Employment', 'Gender', 'MentalHealth','YearsCodePro',
                                    'Country','ComputerSkills', 'Employed']])
      unidata.head()
```

[49]: .....

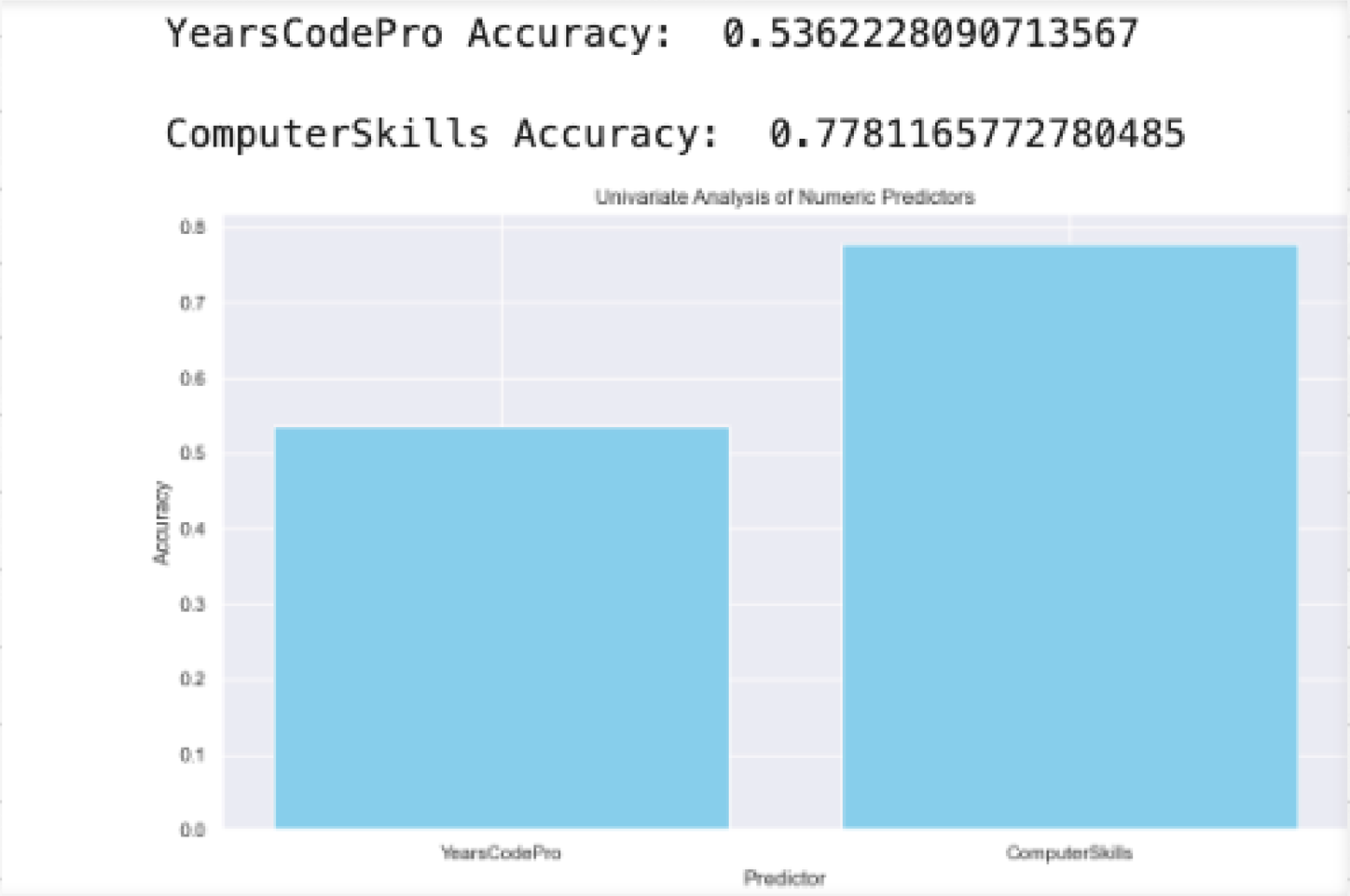
	Age	EdLevel	Employment	Gender	MentalHealth	YearsCodePro	Country	ComputerSkills	Employed
0	<35	Master	1	Man	No	4	Sweden	4	0
1	<35	Undergraduate	1	Man	No	5	Spain	12	1
2	<35	Master	1	Man	No	6	Germany	7	0
3	<35	Undergraduate	1	Man	No	6	Canada	13	0
4	>35	PhD	0	Man	No	30	Singapore	2	0

UNIVARIATE PREDICTIONS OF CATEGORICAL PREDICTORS



UNIVARIATE PREDICTIONS OF NUMERIC VARIABLES

```
YearsCodePro Accuracy:  0.5362228090713567

ComputerSkills Accuracy:  0.7781165772780485
```



DATA PREPARATION

```
[53]: multidata = pd.DataFrame(clean[['Age', 'EdLevel','Employment', 'Gender', 'MentalHealth','YearsCodePro',
                              'Country','ComputerSkills', 'Employed']])
      multidata.head()
```

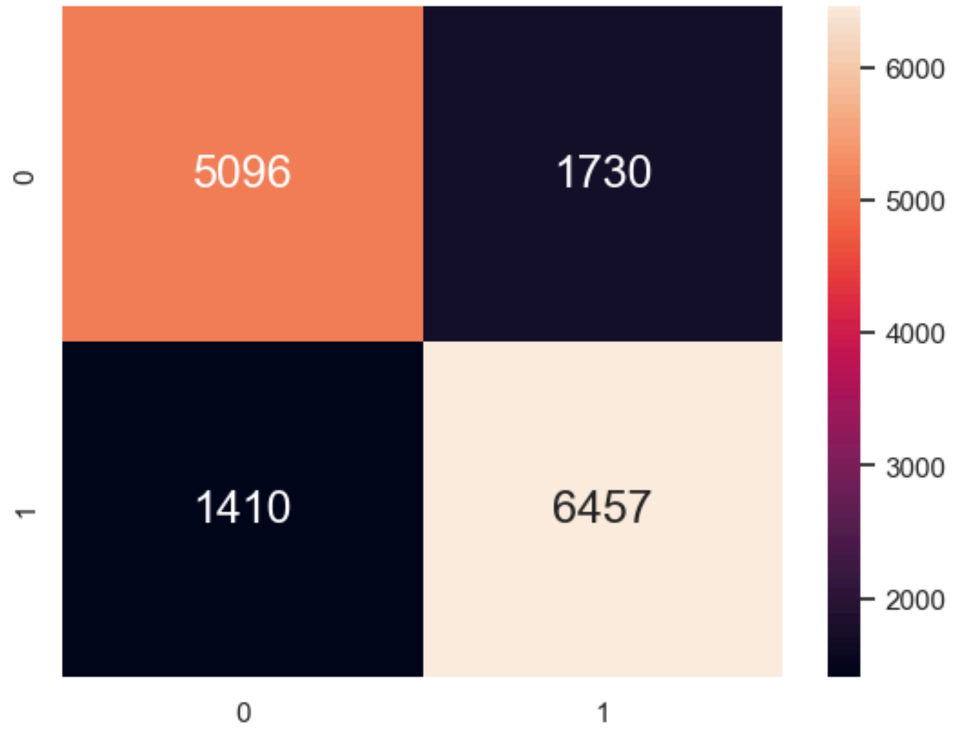
[53]:

	Age	EdLevel	Employment	Gender	MentalHealth	YearsCodePro	Country	ComputerSkills	Employed
0	<35	Master	1	Man	No	4	Sweden	4	0
1	<35	Undergraduate	1	Man	No	5	Spain	12	1
2	<35	Master	1	Man	No	6	Germany	7	0
3	<35	Undergraduate	1	Man	No	6	Canada	13	0
4	>35	PhD	0	Man	No	30	Singapore	2	0

1. LOGISTIC REGRESSION

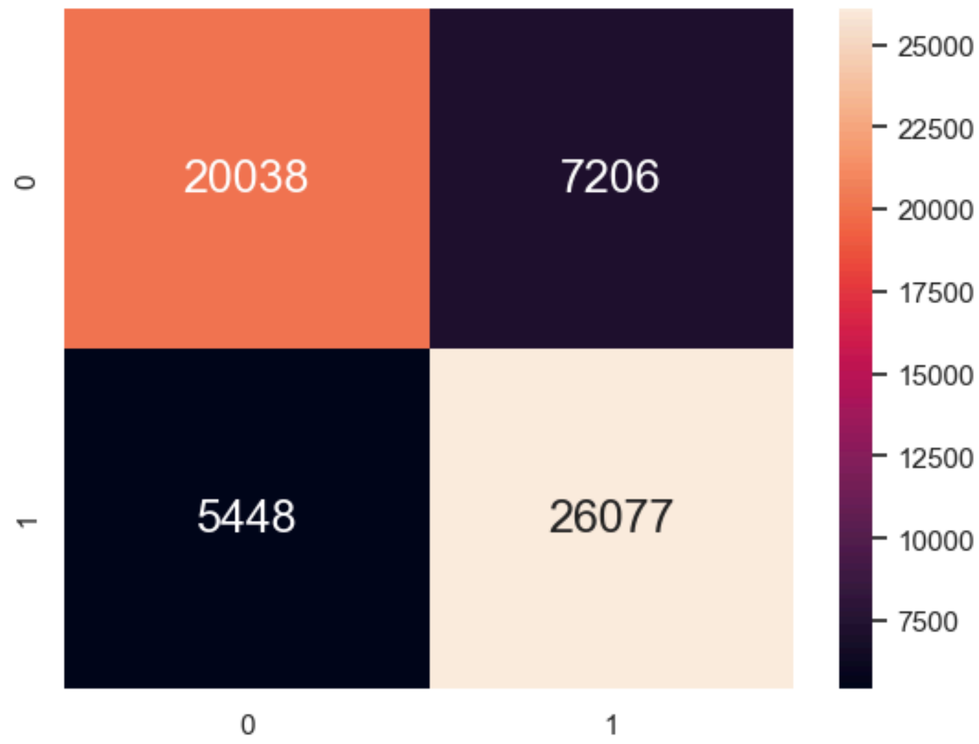
```
Testing Set

        Accuracy: 0.786292792486218

        Precision: 0.7886893856113351

        Recall: 0.8207703063429516

        F1-score: 0.804410115858976
```

[58]: <Axes: >




```
Training Set

        Accuracy: 0.7846824005853427

        Precision: 0.7834930745425592

        Recall: 0.8271847739888977

        F1-score: 0.8047463276138748
```

[57]: <Axes: >



2. DECISON TREE CLASSIFICATION

Classification Report:

	precision	recall	f1-score	support
0	0.70	0.73	0.71	6826
1	0.75	0.73	0.74	7867
accuracy			0.73	14693
macro avg	0.73	0.73	0.73	14693
weighted avg	0.73	0.73	0.73	14693

Classification Report:

	precision	recall	f1-score	support
0	0.90	0.95	0.93	27244
1	0.95	0.91	0.93	31525
accuracy			0.93	58769
macro avg	0.93	0.93	0.93	58769
weighted avg	0.93	0.93	0.93	58769

3. RANDOM FOREST PREDICTION

Train Set

Random Forest Accuracy: 0.9297929180350185

Precision: 0.930679962274828

Recall: 0.939064234734337 9

F1-score: 0.9348533173335017

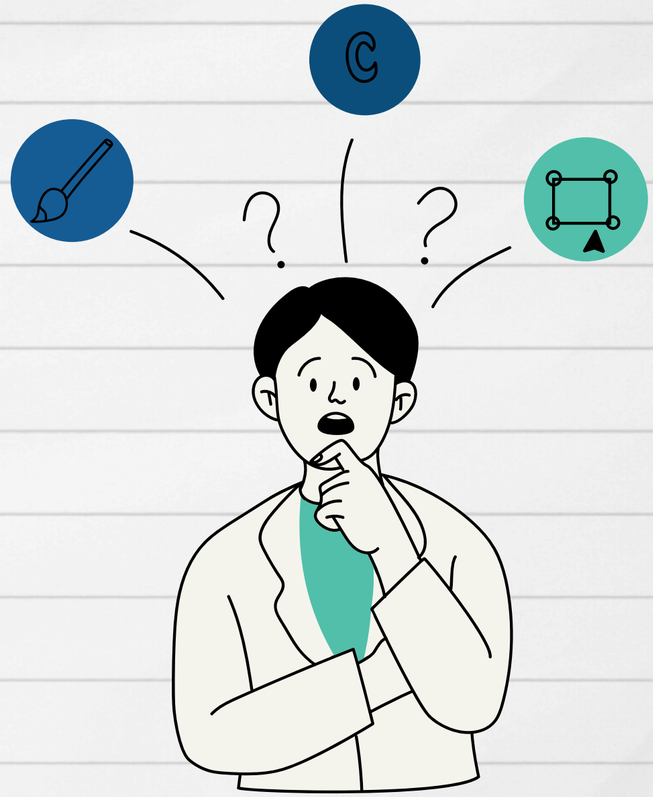
Test Set

Random Forest Accuracy: 0.762676104267338 2

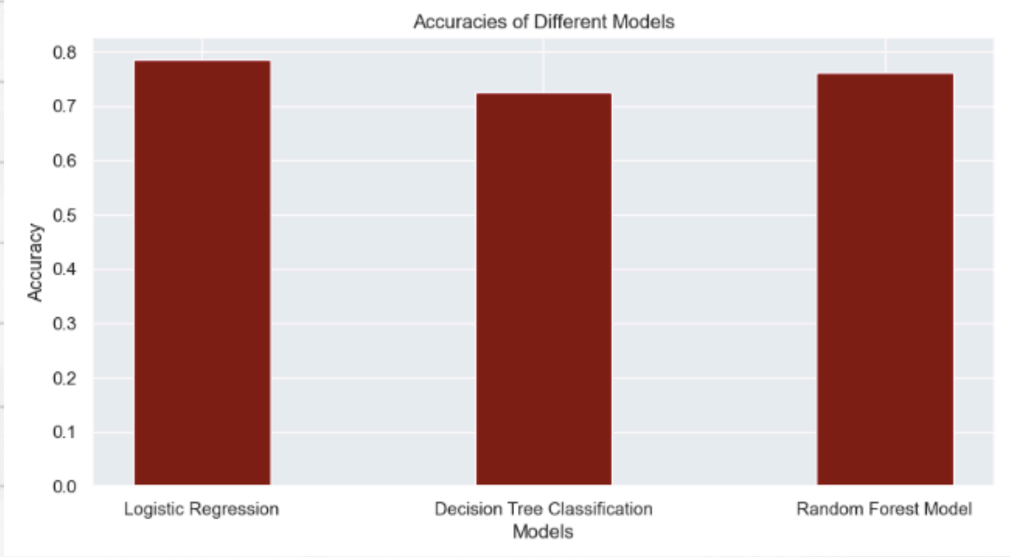
Precision: 0.769770879526977 1

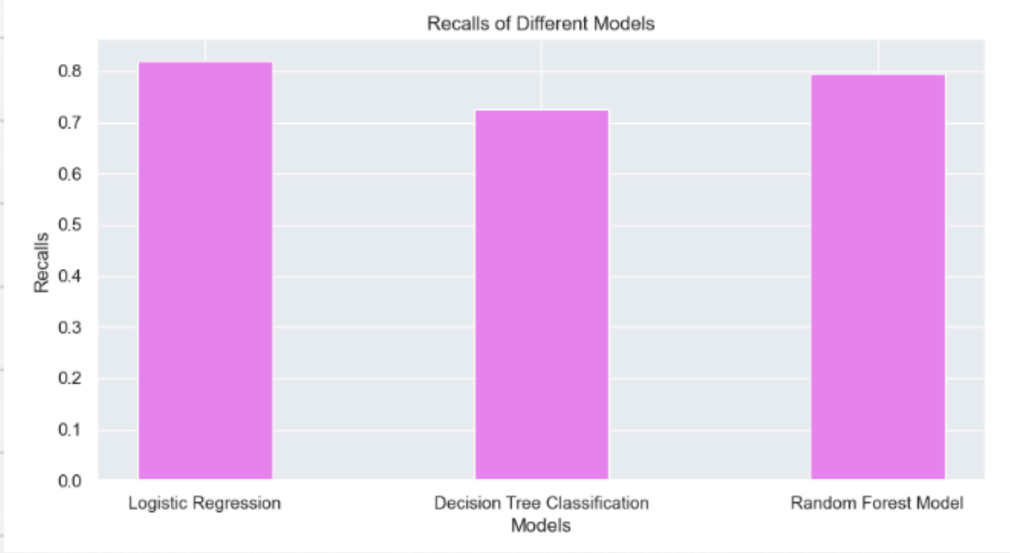
Recall: 0.794330748697089 1

F1-score: 0.781857991867375 7

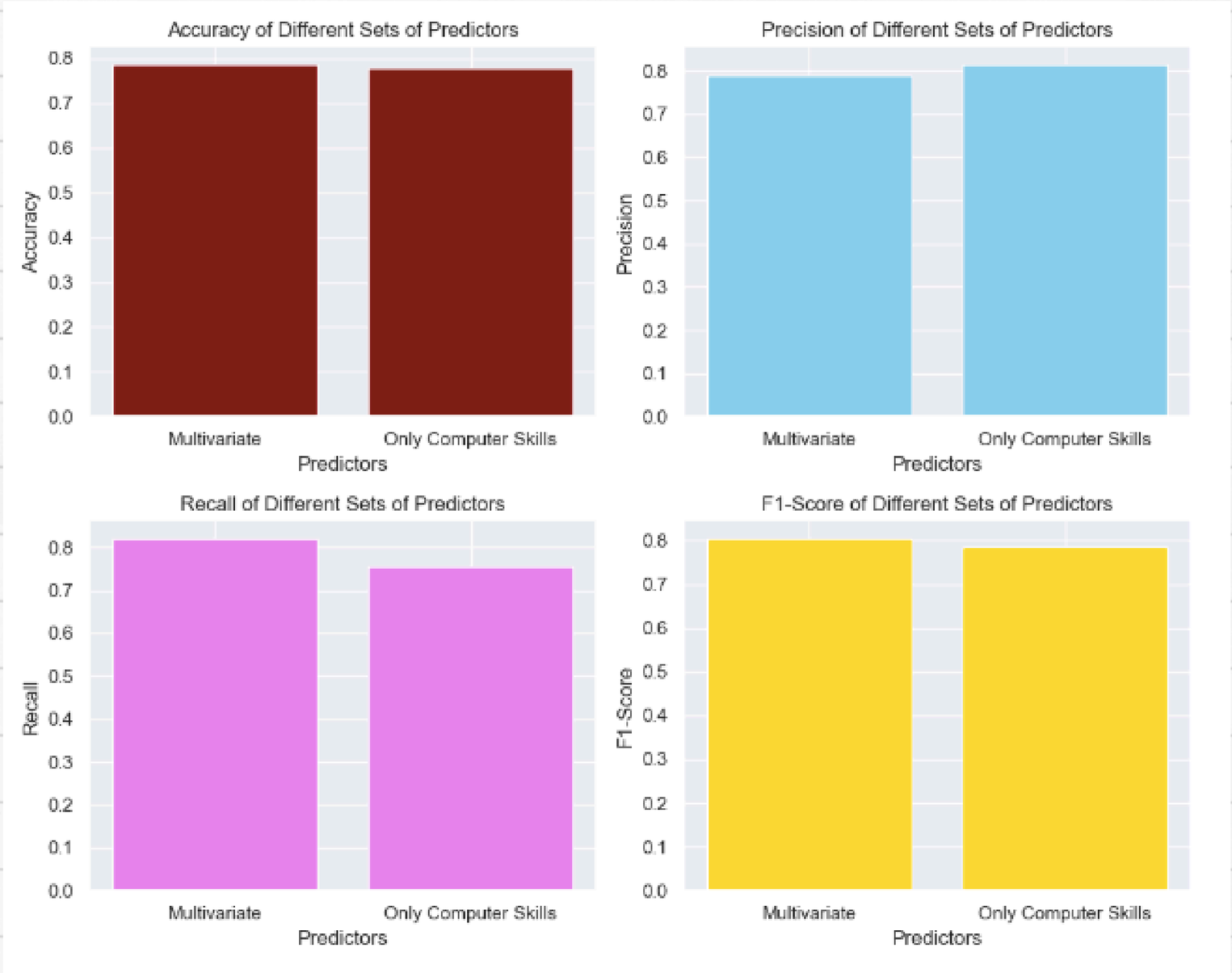


COMPARING PERFORMANCES OF DIFFERENT MODELS

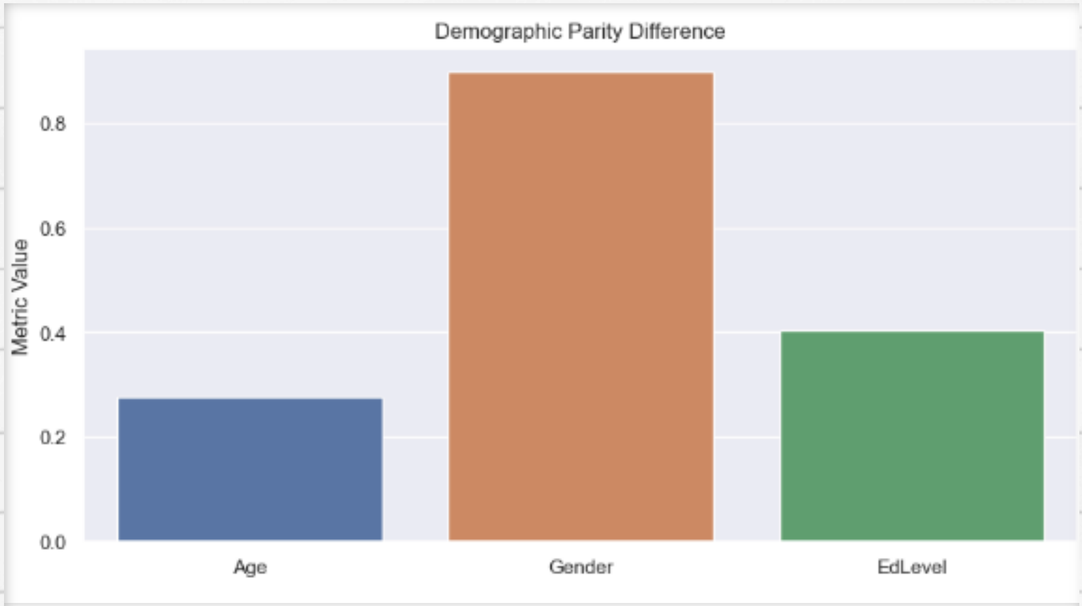


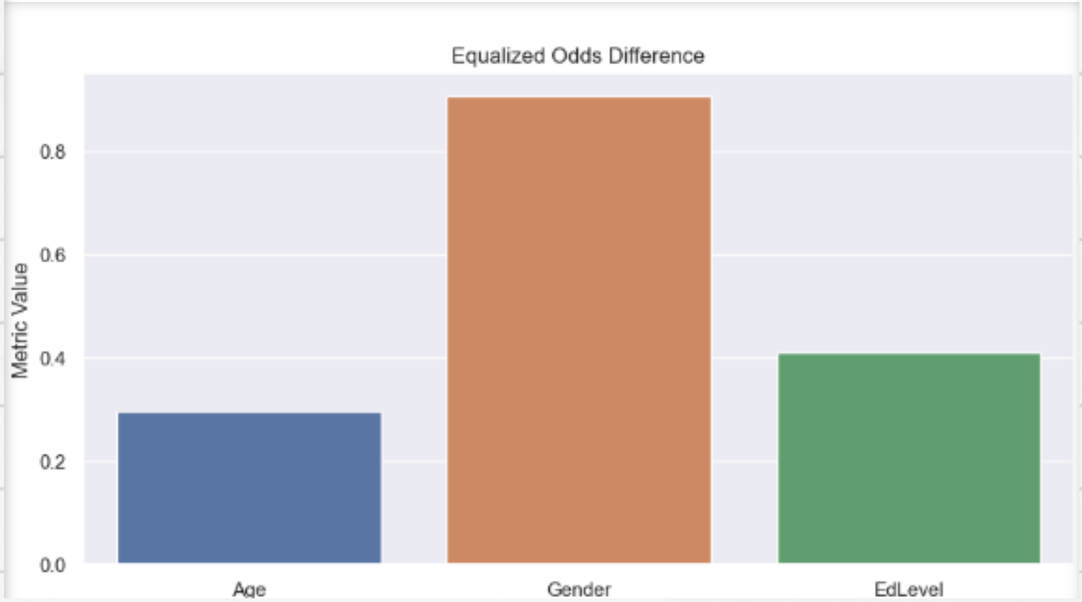


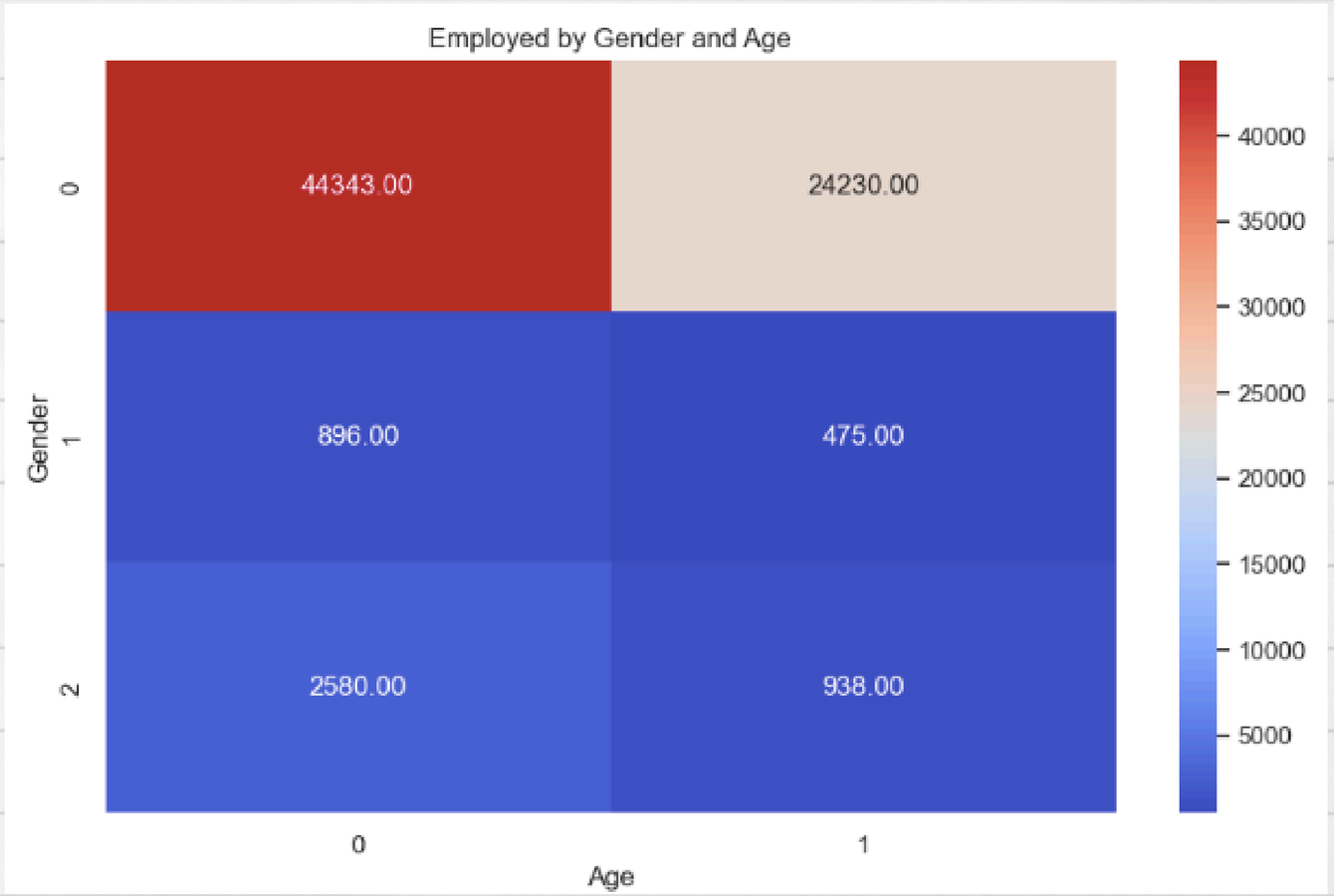
COMPARING THE PERFORMANCE OF DIFFERENT SETS OF PREDICTORS



CHECKING BIASES IN DATA







WAYS TO MITIGATE THE BIAS IN DATA GIVEN

01

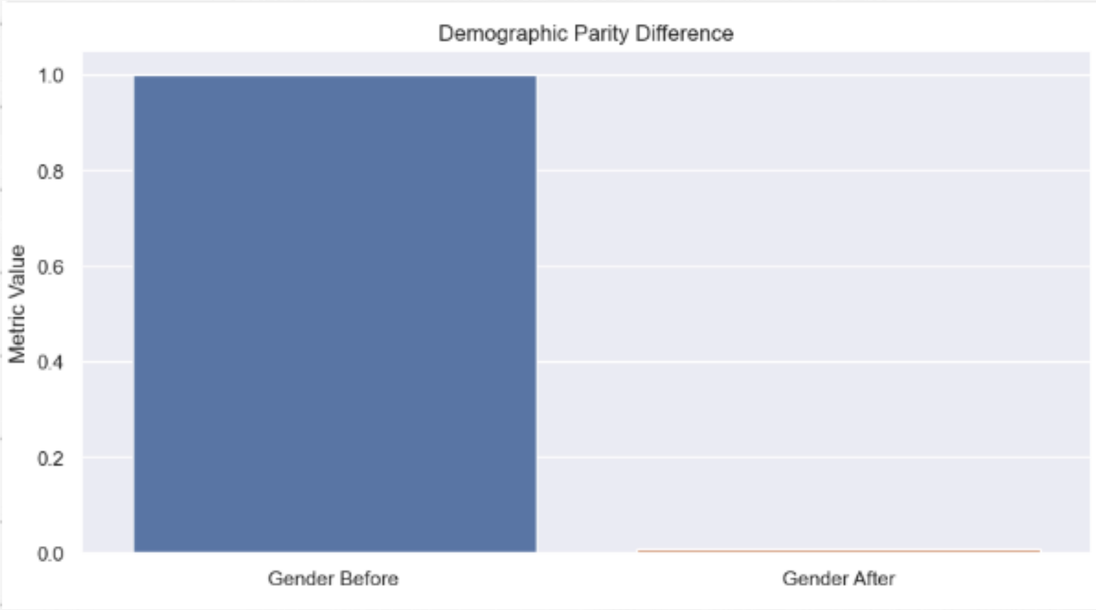
Fairness constraint

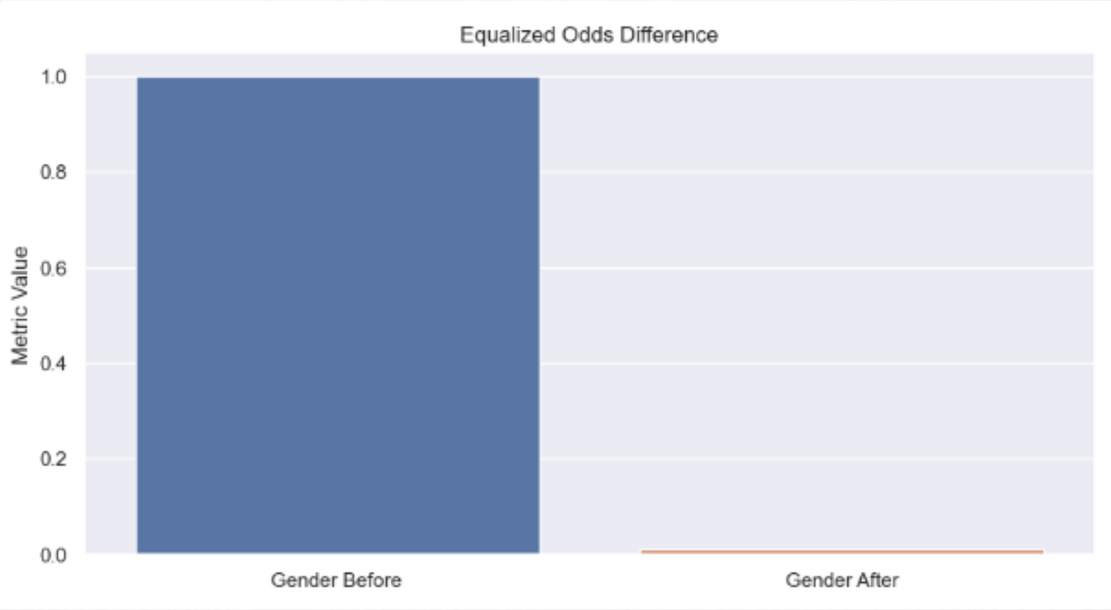
02

Random Forest Classifier

03

Rejection Option Classification





CONCLUSION

- Concluded that Logistic Regression using multiple variables is the best in predicting one's employability.
- Fairness Metrics and HeatMap were used to show the bias in sensitive variables such as age, education level and gender.

THANK YOU VERY MUCH!

1. Data Exploration and Understanding - Isaac

2. Model Selection&Training + Model Evaluation&Fine-Tuning - Tiffany

3. Ethical Considerations and Bias Mitigation + Conclusion + Visualization - Menujaa

4. Slides + Presentation script + Presentation - Jiwon