HR Analysis

Data Analysis Project

Radwa Samy Khattab

Table of Contents

I. Brief Problem Description 2
II. Project Pipeline 2
III. Analysis and Solution of the Problem 3
A. Data Preprocessing 3
B. Results & Data Visualization 3
1. Textual Insights 3
2. Data Graphs 5
C. Data Insights 12
D. Classifier 12
IV. Results & Evaluation 13
A. Accuracy 13
B. F1 Score 13
V. Unsuccessful Trials 13
VI. Enhancements & Future Work 15

I. Brief Problem Description

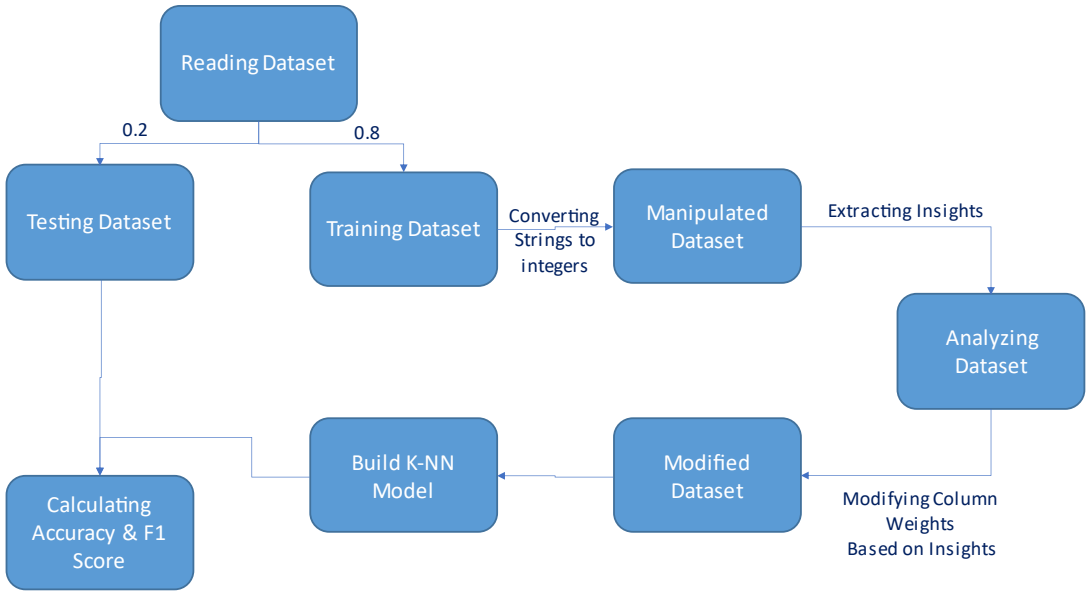
HR Analysis is a data analysis problem, where an HR is provided with data on multiple employees, and he should pick from them ones to promote.

These data are (for each employee ID):

1. Department
2. Region
3. Education
4. Gender
5. Recruitment Channel
6. Number of Trainings
7. Age
8. Previous Year Rating
9. Length of Service
10. KPIs Metric > 80% (1→True, 0 → False)
11. Awards Won? (1→Won Awards, 0→Didn't Win)
12. Average Training Score

And the given prediction in the training dataset, "is_promoted", which indicates whether that employee got promoted or not.

II. Project Pipeline



III. Analysis and Solution of the Problem

A. Data Preprocessing

- Read CSV Training Data
- Replace empty cells with NaN
- Fill NaN in needed columns with proper value (for example, in education, I replaced NaN with 'No Education' to appear in the visualization like that)
- Replaced region_number to simply the number of the region
- Copied another version of the dataset to df_manipulated
- Dropped is_promoted column and employee_id from df_manipulated
- Converted each string to a numeric value (for example, gave each department a unique ID) in df_manipulated
- Normalized data in df_manipulated
- Split normalized dataset and predictions with ration 4:1 between Training and Test dataset

B. Results & Data Visualization

N.B. Textual Insights are printed by running the code and are in Output.txt
Visual Data are saved in folder Visualizations and are saved in folder by running the code.

1. Textual Insights

Training Dataset Size is 43,846
Test Dataset size is 10,962

Sample of the dataset:

Employee ID	Department	Region	Education	Gender	Recruitment Channel	No Of Trainings	Age	Previous Year Rating	Length Of Service	Kpis Met >80 %	Awards Won?	Avg Training Score	Is Promoted
65438	Sales & Marketing	region_7	Master's & above	f	sourcing	1	35	5	8	1	0	49	0
49017	Sales & Marketing	region_7	Bachelor's	f	sourcing	1	35	5	3	1	0	50	1

Maximum age is 60
Minimum age is 20

Maximum Average Training Score is 99
Minimum Average Training Score is 39

Maximum Service Length is 37
Minimum Service Length is 1

Percentage of females in the dataset = 37%
Percentage of males in the dataset = 88%
Percentage of promotions in dataset = 11%
Percentage of promoted females among all promoted = 31%
Percentage of promoted males among all promotes = 69%

Percentage of promoted females to all females = 9%
Percentage of promoted males to all males = 8%

Percentage of employees who won awards = 11%
Percentage of promoted who won awards among all promoted = 12%

Percentage of promoted who didn't win awards among all promotes = 88%
Percentage of who won awards and promoted among all won = 44%
Percentage of who won awards and not promoted among all won = 56%

Percentage of Promoted with Avg. Training score in [39-43] = 5%
Percentage of Promoted with Avg. Training score in [44-48] = 4%
Percentage of Promoted with Avg. Training score in [49-53] = 4%
Percentage of Promoted with Avg. Training score in [54-58] = 6%
Percentage of Promoted with Avg. Training score in [59-63] = 7%
Percentage of Promoted with Avg. Training score in [64-68] = 10%
Percentage of Promoted with Avg. Training score in [69-73] = 11%
Percentage of Promoted with Avg. Training score in [74-78] = 11%
Percentage of Promoted with Avg. Training score in [79-83] = 10%
Percentage of Promoted with Avg. Training score in [84-88] = 12%
Percentage of Promoted with Avg. Training score in [89-93] = 50%
Percentage of Promoted with Avg. Training score in [94-99] = 99%

Correlation between Whether an Award is won and Promotion is 0.1894588637983928

Correlation between age and promotion is -0.017165891678930154

Correlation between education and promotion is -0.037099445053647415
Correlation between department and promotion is -0.0009673398940311083

Correlation between Average Training Score and Promotion is 0.1774652056724032

Correlation between Number of Trainings and Promotion is -0.0009673398940311083

Correlation between Previous Year Rating and Promotion is 0.12868142012158557

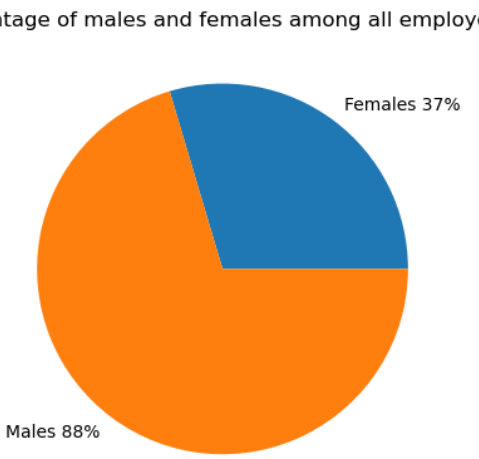
Correlation between Length of Service and Promotion is 0.12868142012158557

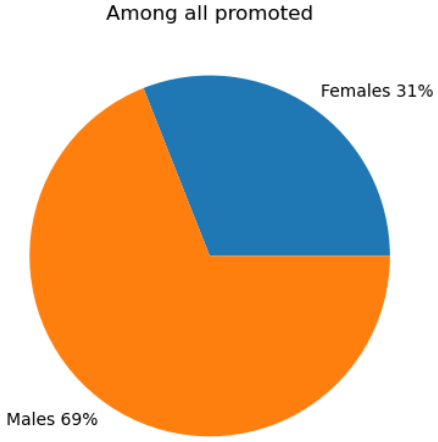
Correlation between KPI Metric > 80 percent and Promotion is 0.2200698234527305

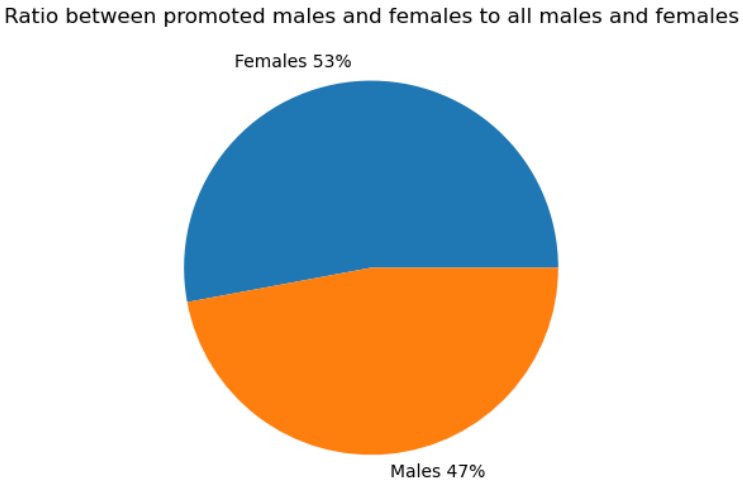
Correlation between Recruitment Channel and promotion is 0.005801218791109693

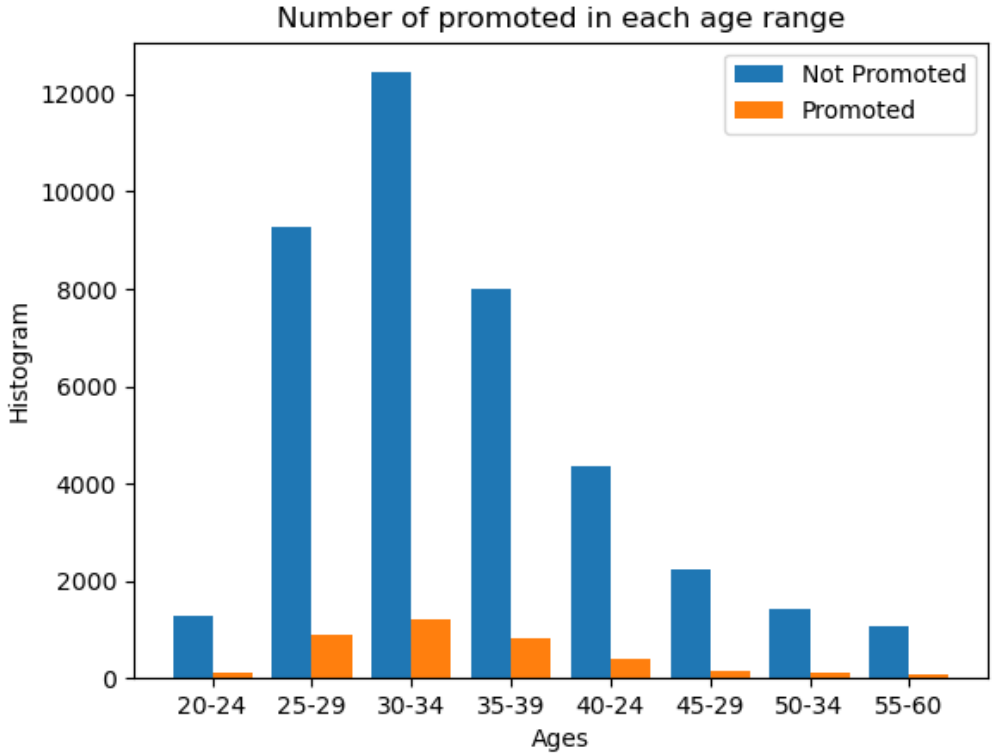
Correlation between Region and promotion is 0.005801218791109693

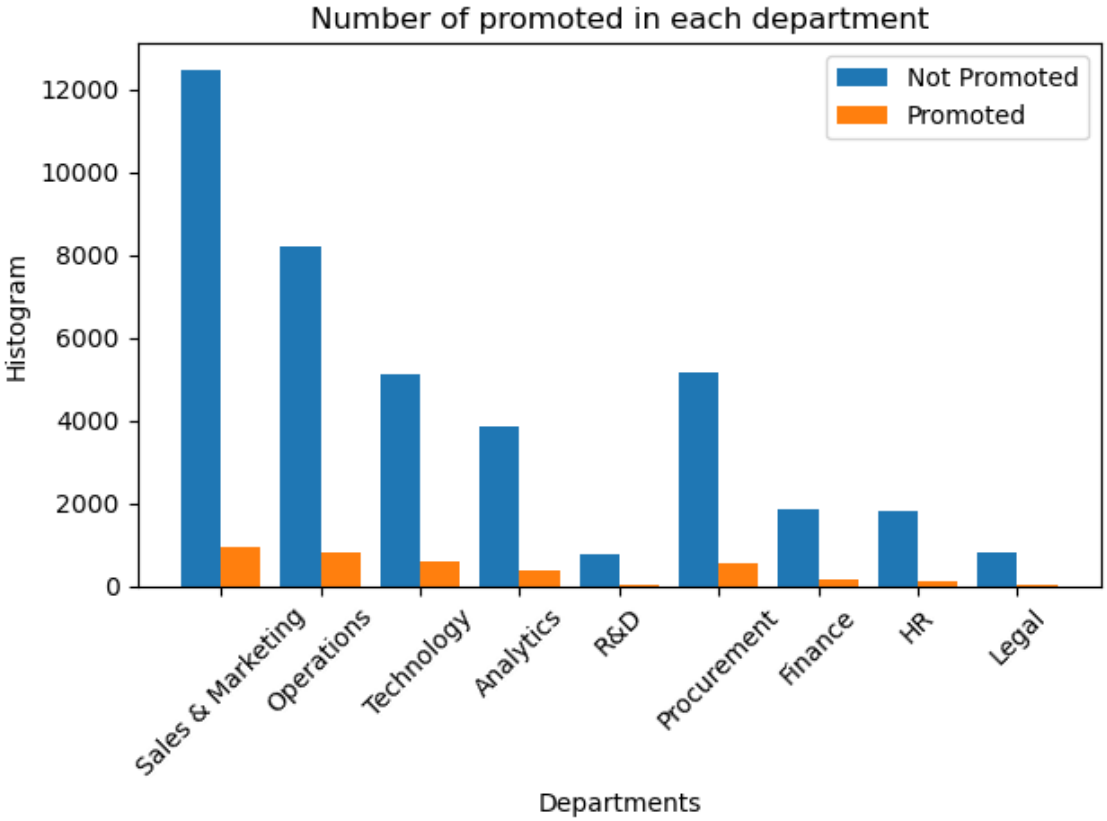
2. Data Graphs

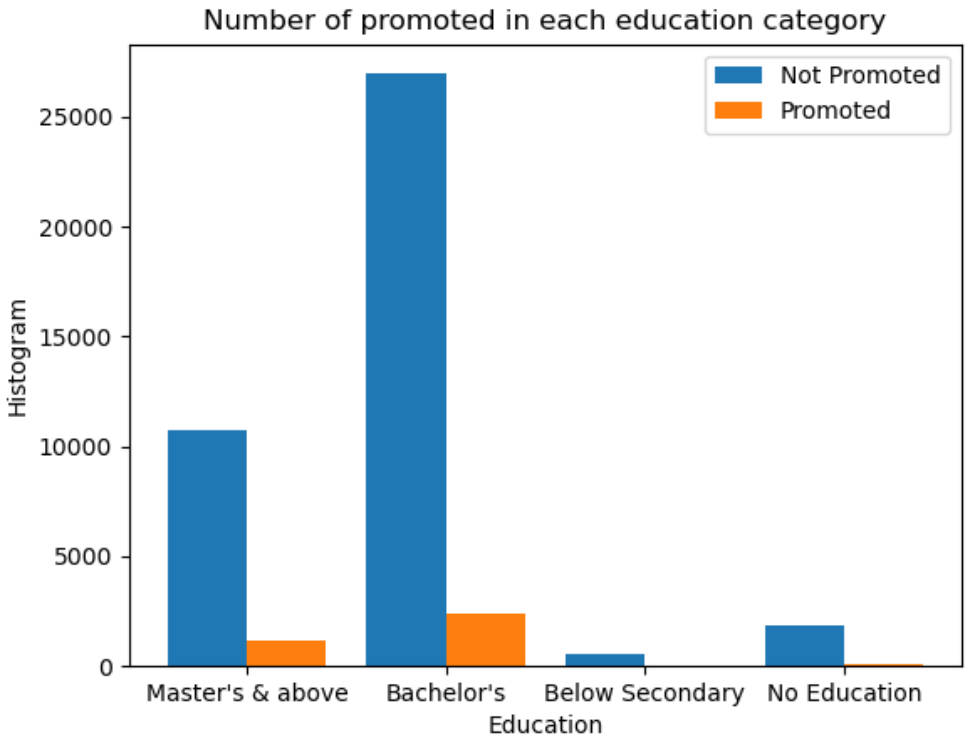


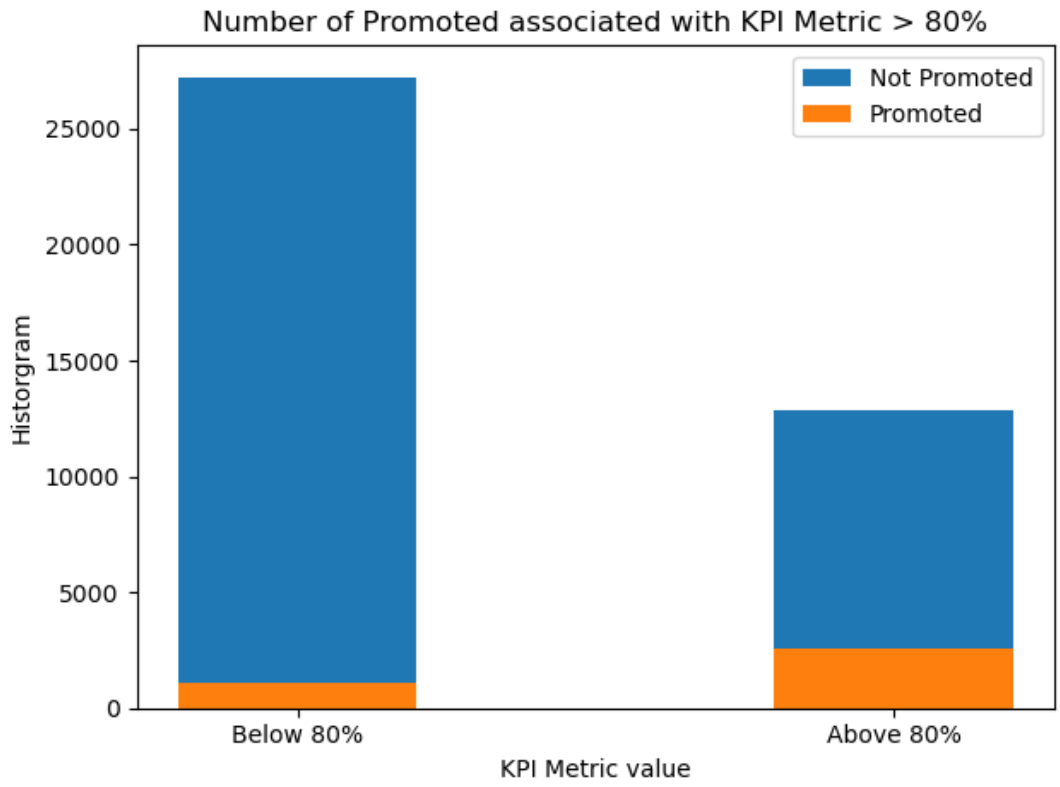


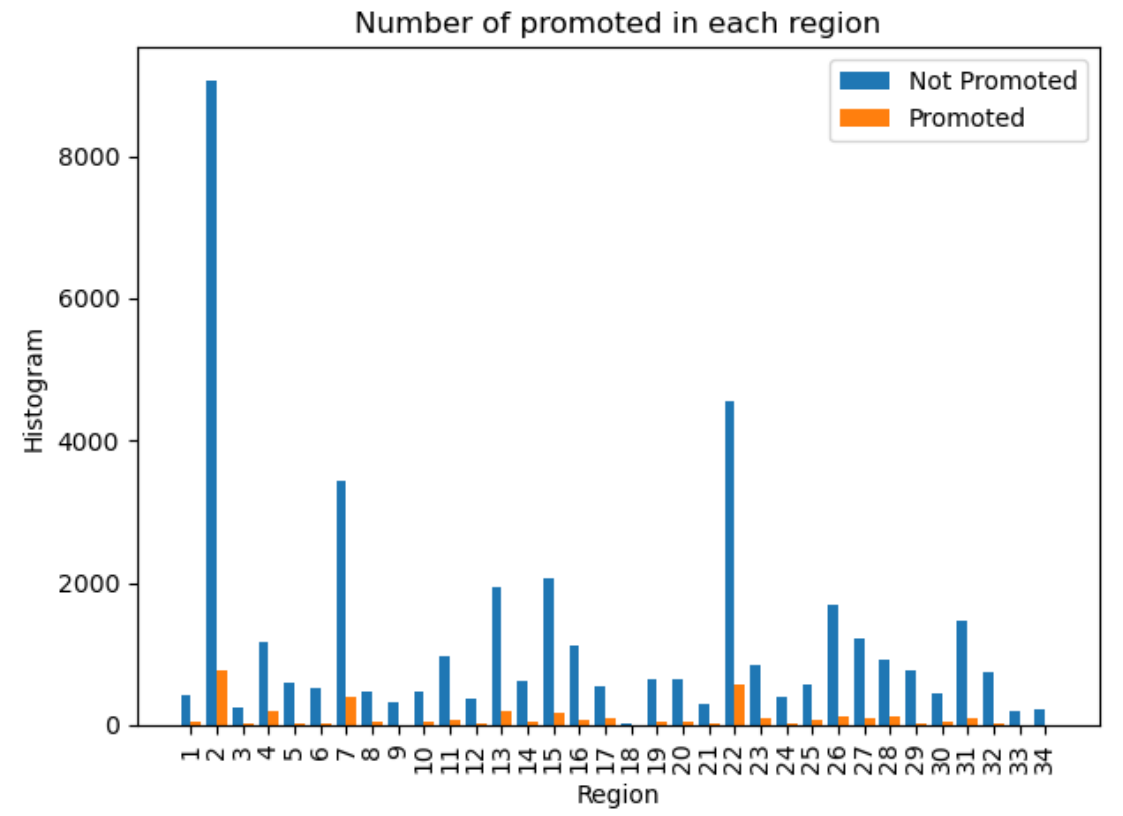


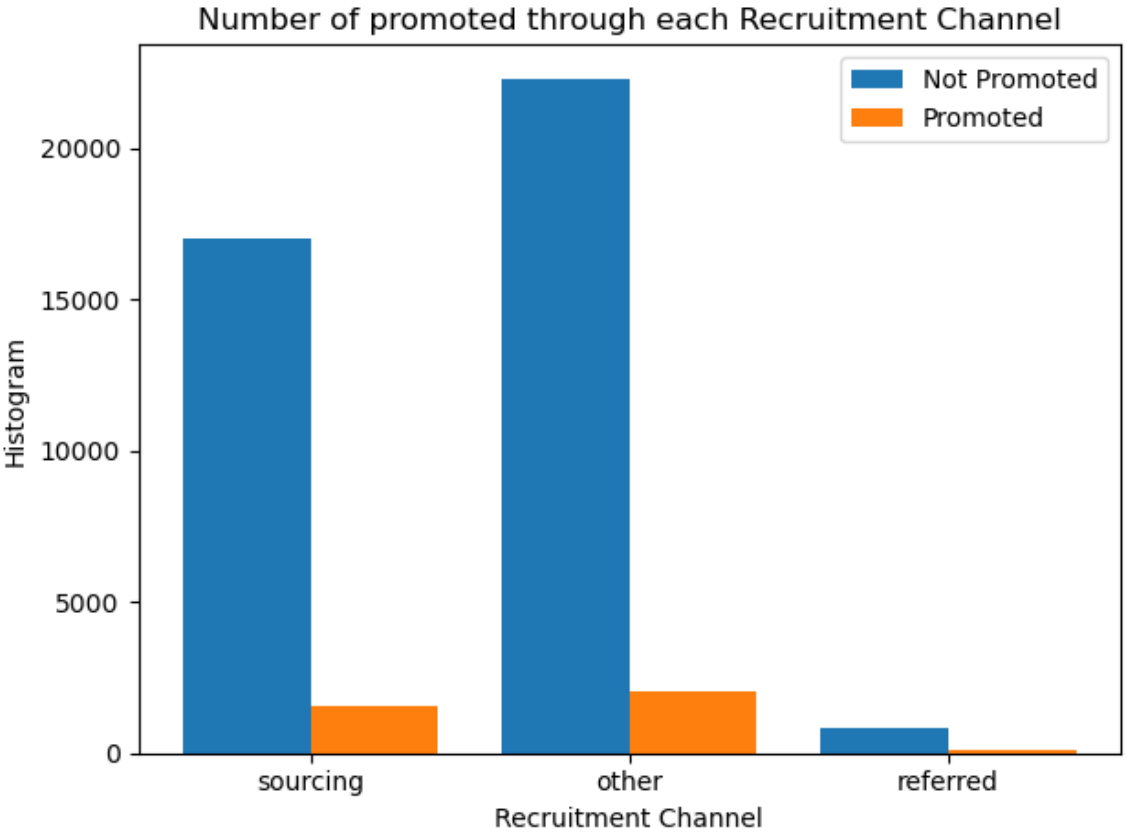

Number of promoted in each age range
Not Promoted
Promoted
Histogram
0
2000
4000
6000
8000
10000
12000
20-24
25-29
30-34
35-39
40-24
45-29
50-34
55-60
Ages

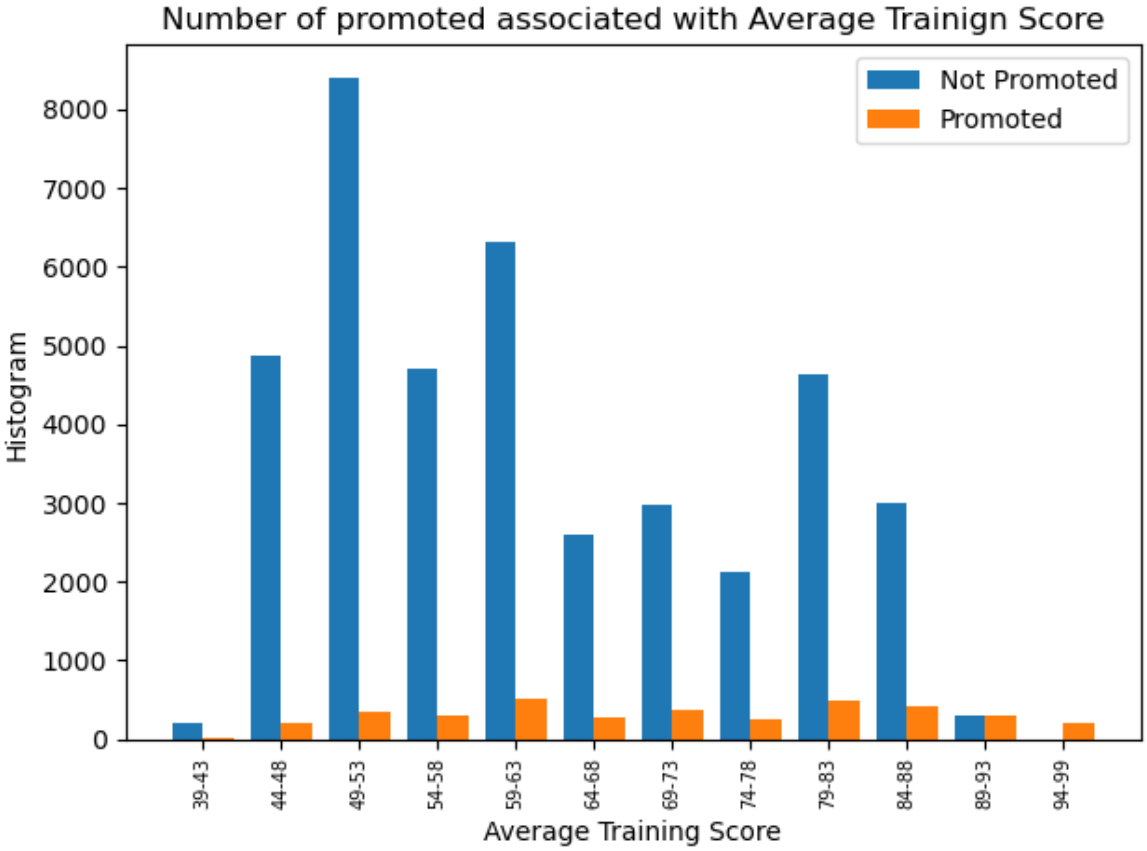

Number of promoted in each department
Not Promoted
Promoted
Histogram
0
2000
4000
6000
8000
10000
12000
Sales & Marketing
Operations
Technology
Analytics
R&D
Procurement
Finance
HR
Legal
Departments

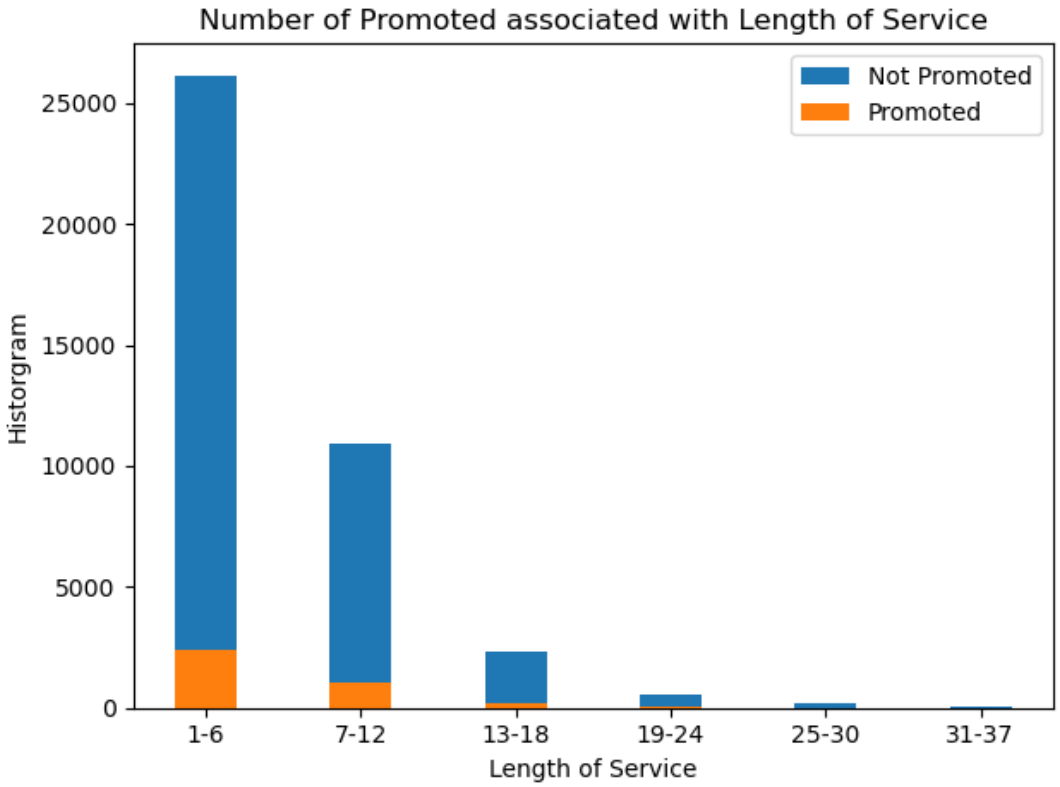

Number of promoted in each education category
Not Promoted
Promoted
25000
20000
15000
10000
5000
0
Histogram
Master's & above
Bachelor's
Below Secondary
No Education
Education

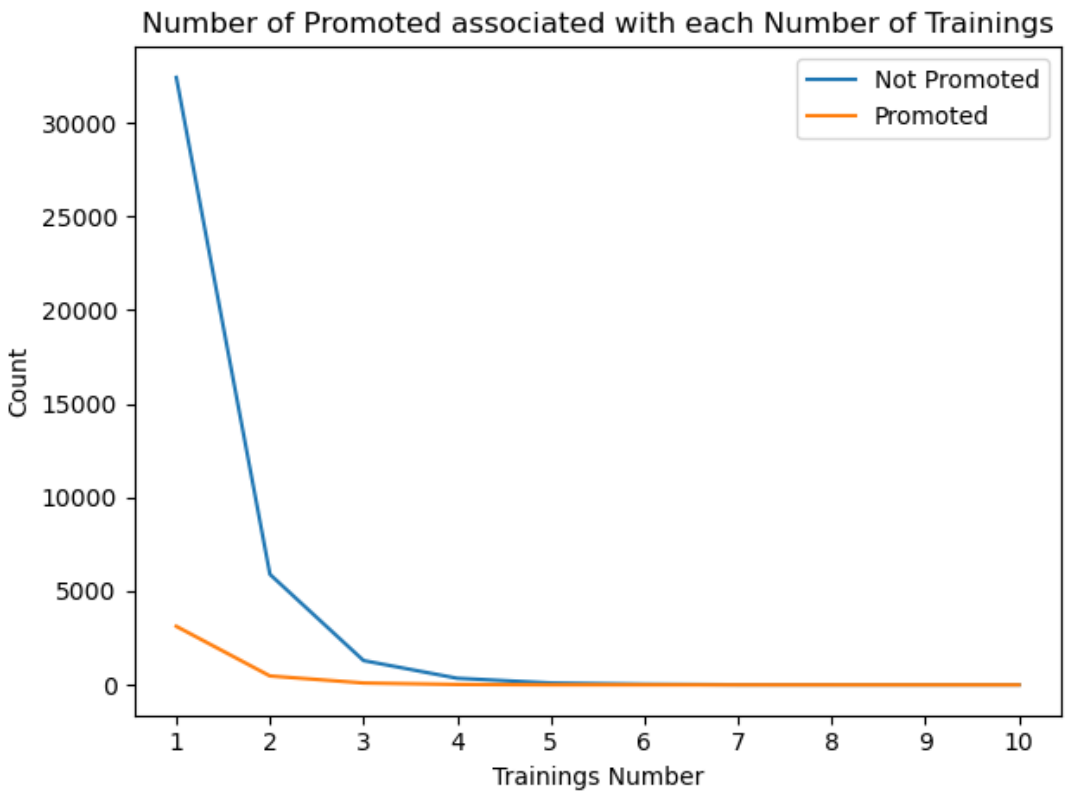

Number of Promoted associated with KPI Metric > 80%
Not Promoted
Promoted
25000
20000
15000
10000
5000
0
Historgram
Below 80%
Above 80%
KPI Metric value

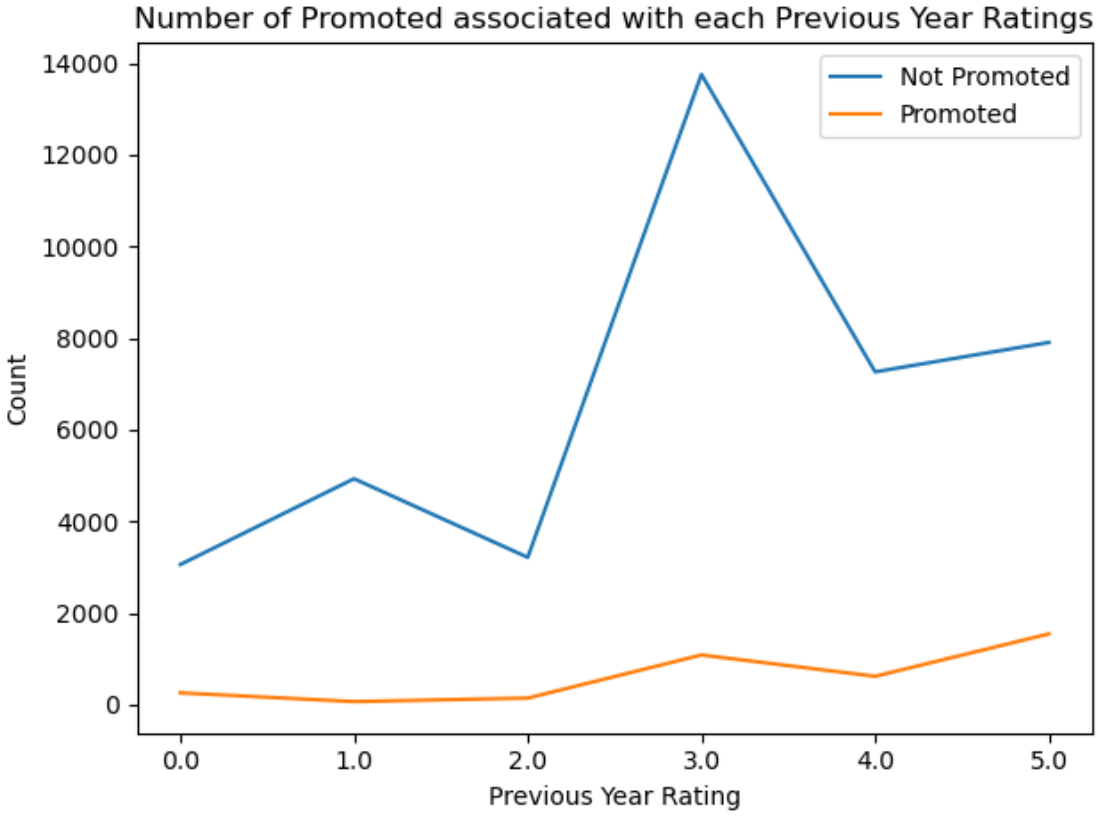

Number of promoted in each region
Not Promoted
Promoted
Histogram
0
2000
4000
6000
8000
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
Region

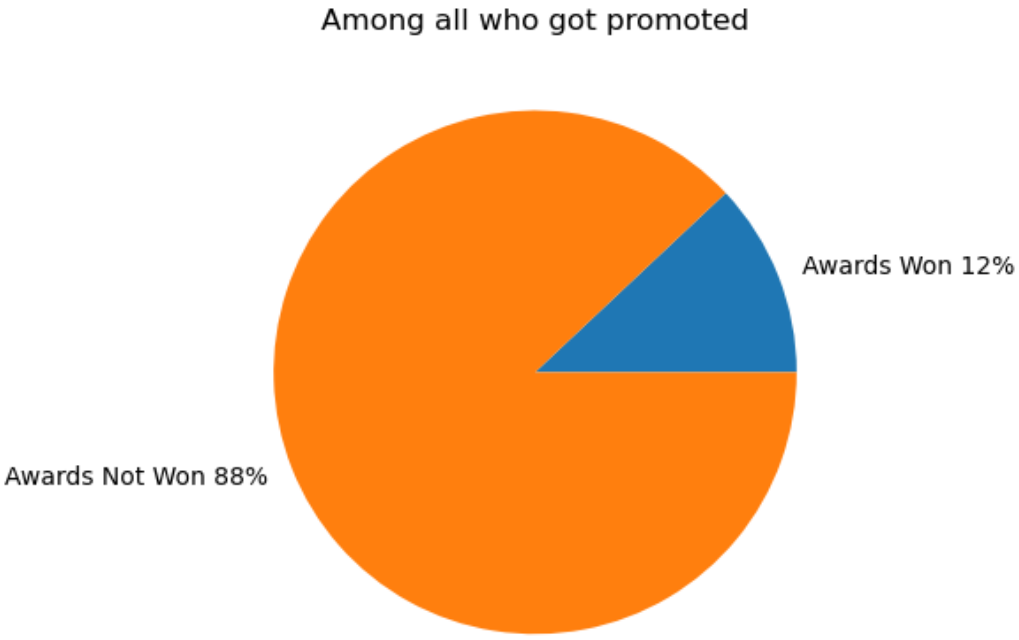

Number of promoted through each Recruitment Channel
Not Promoted
Promoted
Histogram
0
5000
10000
15000
20000
sourcing
other
referred
Recruitment Channel

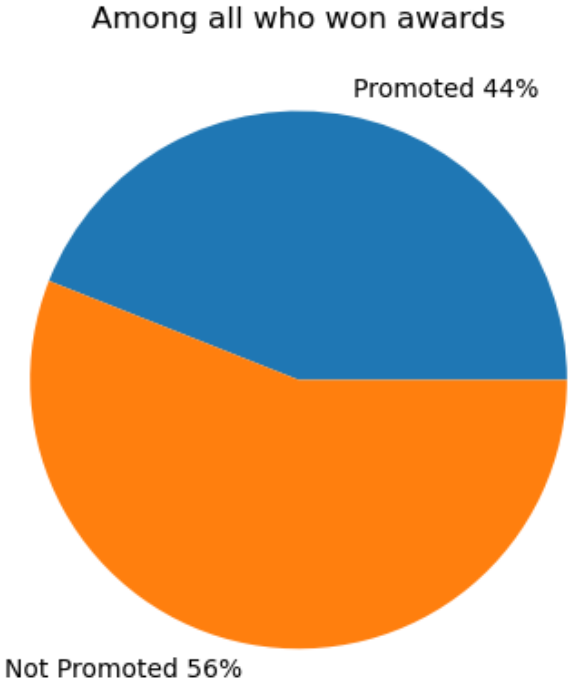

Number of promoted associated with Average Trainign Score
Not Promoted
Promoted
Histogram
0
1000
2000
3000
4000
5000
6000
7000
8000
39-43
44-48
49-53
54-58
59-63
64-68
69-73
74-78
79-83
84-88
89-93
94-99
Average Training Score


Number of Promoted associated with Length of Service
Not Promoted
Promoted
Historgram
0
5000
10000
15000
20000
25000
1-6
7-12
13-18
19-24
25-30
31-37
Length of Service


Number of Promoted associated with each Number of Trainings
Not Promoted
Promoted
Count
30000
25000
20000
15000
10000
5000
0
1
2
3
4
5
6
7
8
9
10
Trainings Number


Number of Promoted associated with each Previous Year Ratings
Not Promoted
Promoted
Count
14000
12000
10000
8000
6000
4000
2000
0
0.0
1.0
2.0
3.0
4.0
5.0
Previous Year Rating


Among all who got promoted
Awards Won 12%
Awards Not Won 88%


Among all who won awards
Promoted 44%
Not Promoted 56%

C. Data Insights

I have done different data analysis on the dataset provided. And from which I came to multiple conclusions.

- Females represent 37% of the company's employees
- Gender doesn't affect the process of choosing whether an employee gets promoted or not much. The process is fair between both females and males, compared to their ratios in the company.
- Winning awards don't have a strong effect on the promotion
- Age ranges are more concentrated in range [25-39], with Min age of 20 and Max of 60
- Promotions are given to all ages of employees with equal probability, based on the number of employees in a given age range
- R&D and Legal departments' teams are rather not-promoted than other departments' teams
- A very small number of promotions are given to those with education below secondary or no education
- Majority of employees have only one previous training, and only a very small portion have more than 4. Hence, majority of promoted have one year of experience
- Only a small portion of employees are through referred recruitment channel, the rest are from sourcing or other channels—the majority in both employees' number and promotions
- Majority of employees are in regions 2, 22 and 7—in descending order
- Region doesn't really affect the promotion result
- As the average training score increases, the ratio between those promoted in this range to those who aren't, increases. For example, Percentage of Promoted with Avg. Training score in [94-99] = 99%, and Percentage of Promoted with Avg. Training score in [89-93] = 50%, and Percentage of Promoted with Avg. Training score in [84-88] = 12%, and so on.
- KPI Metric being over 80% has relatively high correlation with the employee being promoted
- Majority of promoted have KPI measure above 80%
- Majority of employees' previous year ratings is 3.0

D. Classifier

I have tried multiple classifiers with different data manipulations.

The classifiers I tried are:

1. KNN
2. SVM
3. Decision Trees

I came to the conclusion that:

1. KNN isn't the best classifier for this problem as the data is very biased to the non-promoted employees, and even when I tried multiplying the size of the promoted employees, it wasn't the best. Also, not all variables in the dataset are major effectors on the classification to be considered as dimensions in the KNN.
2. SVM is very bad for this problem and its classification.
3. Decision Trees were the best among the three, which makes sense considering that the data is not unbiased, and there are multiple dimensions in the dataset, even though not all are major

effectors to the result of the classification. In Decision Trees, using Entropy or Gini impurity, the tree would be constructed in a fairer way, considering the right parameters when needed for splits to end-up with the right leaves (classification).

And after trying these 3 models with different criteria, I also tried modifying some values in the normalized dataset before training the model and predicting results on the test dataset. From these:

1. Doubling values of the variables that have relatively high correlation with the prediction, and affect the prediction more than other variables. Of course, choosing these variables where based on the insights I understood. Here, I multiplied data in 3 columns by 3:
    a. KPI Metric > 80%
    b. Education
    c. Average Training Score
2. Dropping columns that didn't have large effect on the prediction, which were 3:
    a. Gender
    b. Awards won?
    c. Region

    However, this was better with some classifiers among the 3 I choose, and worse with others.
3. Doubling the data having the promoted employees, to make the dataset more unbiased than not. Of course, this didn't have much effect on decision tree classifier, but did affect results of KNN and SVM.

In the end, I decided on using **Decision Trees** with criterion **Gini**, and I decided to only **multiple the three variables** (KPI Metric > 80%, Education and Average Training Score) **by 3**, and not do any other data modification.

IV. Results & Evaluation

From what I noticed in the dataset and its visualizations, there is only 11% of the employees promoted. So, this means that measuring how good my classification is based solely on the accuracy isn't sufficient, but I need to consider F1 score more, as it is fairer in case of biased dataset like this.

Based on my last decision for the classifier and data manipulations, I got these results.

A. Accuracy

Accuracy of **90.12%** in a testing dataset of size 10962.

B. F1 Score

F1 score of **0.45** in a testing dataset of size 10962.

V. Unsuccessful Trials

I have run the classification multiple times, using different classifiers. And at each time, I checked the Accuracy and F1 score. F1 score is more accurate in comparing, as data is biased for no-promotions, as seen in the data visualization section.

And in some cases, I tried modifying the dataset before classifying.

1. **KNN with N = 11**
   Accuracy of test dataset predictions is 91.88%
   F1 score of test dataset predictions is 0.18

2. **KNN with N = 3**
   Accuracy of test dataset predictions is 90.64%
   F1 score of test dataset predictions is 0.25

3. **SVM**
   Accuracy of test dataset predictions is 92.23%
   F1 score of test dataset predictions is 0.2

4. **Decision Trees with criterion Gini**
   Accuracy of test dataset predictions is 89.88%
   F1 score of test dataset predictions is 0.43

5. **Decision Trees with criterion Entropy**
   Accuracy of test dataset predictions is 89.83%
   F1 score of test dataset predictions is 0.44

After Modifying Data as described in Classifier subsection above:

6. **KNN with N = 11**
   Accuracy of test dataset predictions is 92.59%
   F1 score of test dataset predictions is 0.28

7. **KNN with N = 3**
   Accuracy of test dataset predictions is 91.58%
   F1 score of test dataset predictions is 0.37

8. **SVM**
   Accuracy of test dataset predictions is 91.77%
   F1 score of test dataset predictions is 0.09

9. **Decision Trees with criterion Gini**
   Accuracy of test dataset predictions is 89.74%
   F1 score of test dataset predictions is 0.43

10. **Decision Trees with criterion Entropy**
    Accuracy of test dataset predictions is 89.82%
    F1 score of test dataset predictions is 0.43

VI. Enhancements & Future Work

Gathering more data to have a larger more generic and more biased dataset, that would give us better insights and better classifications.