# Prediction of Employee Success

Name: Vansh Jain
ID: 2020A7PS0079U

Submitted to: Dr. Siddhaling Urolagin

# Agendas

- Problem
- IBM HR Dataset
- Concepts
- Conclusion
- Future Works

# Problem

## Prediction of employee success using Data Science for HR Analytics Employee Promotion Dataset

# IBM HR Analytics Employee Attrition & Performance Dataset

78,298 rows & 13 columns

8 columns for numerical data & 5 categorical data

Response variables

is_promotion

# Concepts

## Exploratory Data Analysis

- EDA is crucial in predicting employee success, involving techniques like data visualization, statistical analysis, and profiling.
- EDA identifies important variables, patterns, and trends for predictive models.
- EDA uncovers outliers, missing values, inconsistencies, and biases affecting model accuracy and fairness.
- EDA refines predictive models iteratively for a thorough understanding of the data.

## Data Cleaning & Preprocessing

- Data cleaning and preprocessing are essential for preparing data for analysis and modeling.
- Data cleaning corrects errors and inconsistencies in the data.
- Data preprocessing transforms raw data for easy analysis by algorithms.
- Tasks may include handling missing values, outliers, standardizing data, and encoding categorical variables.

## Data Encoding

- Data encoding converts categorical/text data to numerical representation for machine learning.
- Methods include label encoding
- Label encoding assigns numerical labels to categories, not ideal for ordinal/non-ordinal variables.

## Oversampling

- SMOTE stands for Synthetic Minority Oversampling Technique.
- It is a data augmentation technique used to address class imbalance in a dataset.
- It works by creating synthetic data points that are similar to the minority class data points.
- This helps to improve the performance of machine learning models trained on imbalanced datasets.

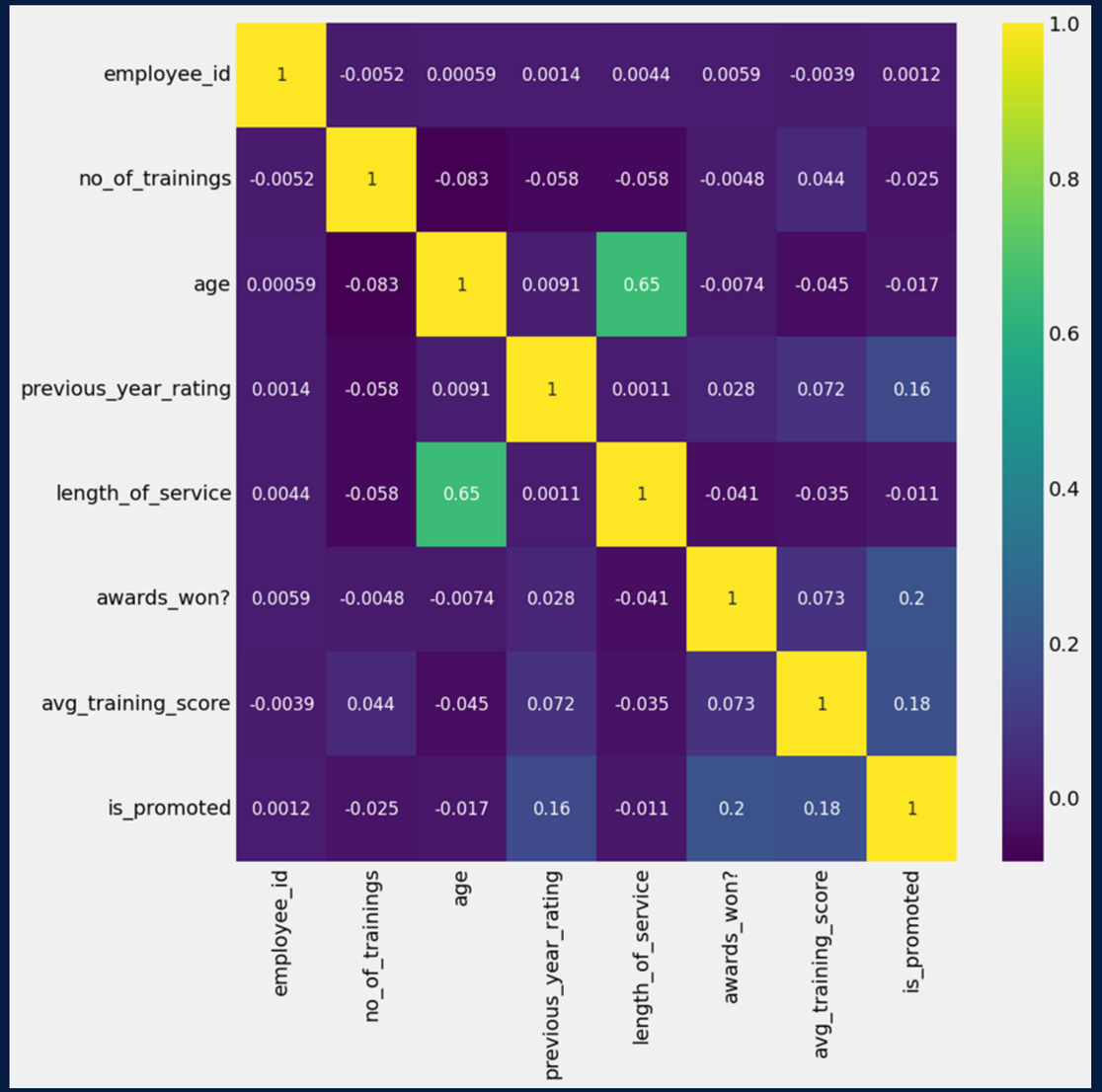
# EDA

- A correlation matrix is a table or matrix showing the correlation coefficients among various variables in a dataset.
- It provides a convenient way to visualize and understand the relationships or correlations between variables.
- It can be used to identify variables that are correlated with each other, determine the strength of the relationship between variables, and identify variables that may be causing other variables to change.

age is highly overall correlated with length_of_service.

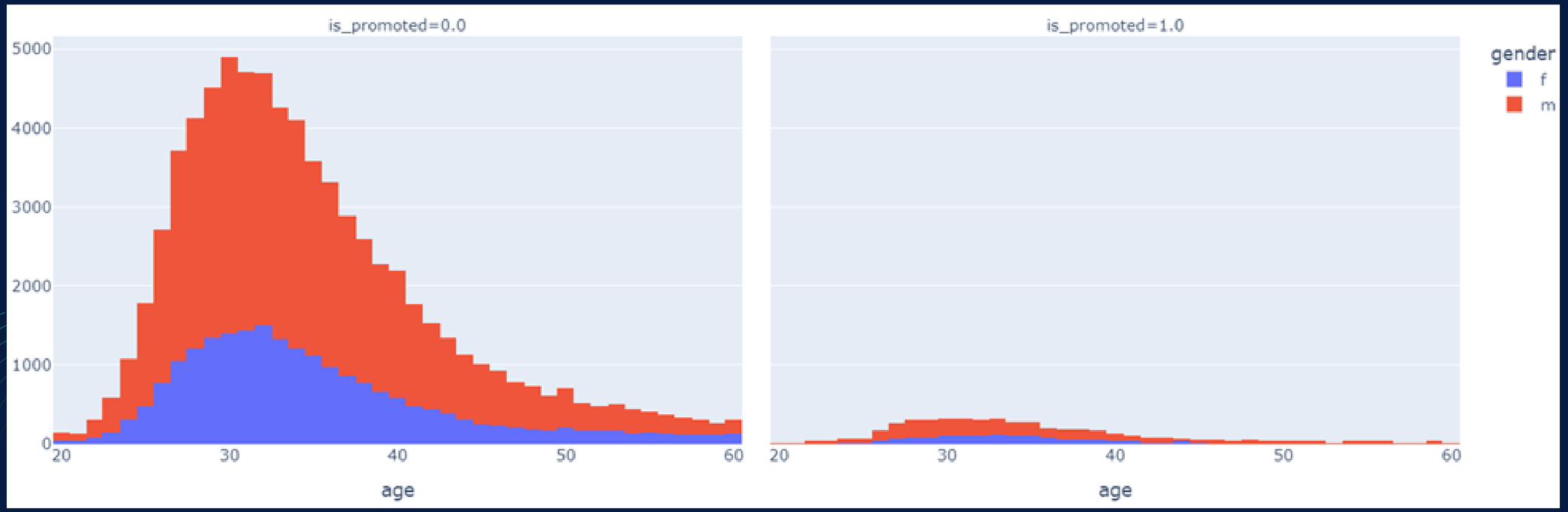
avg_training_score is highly overall correlated with department.

## Correlation matrix



# EDA

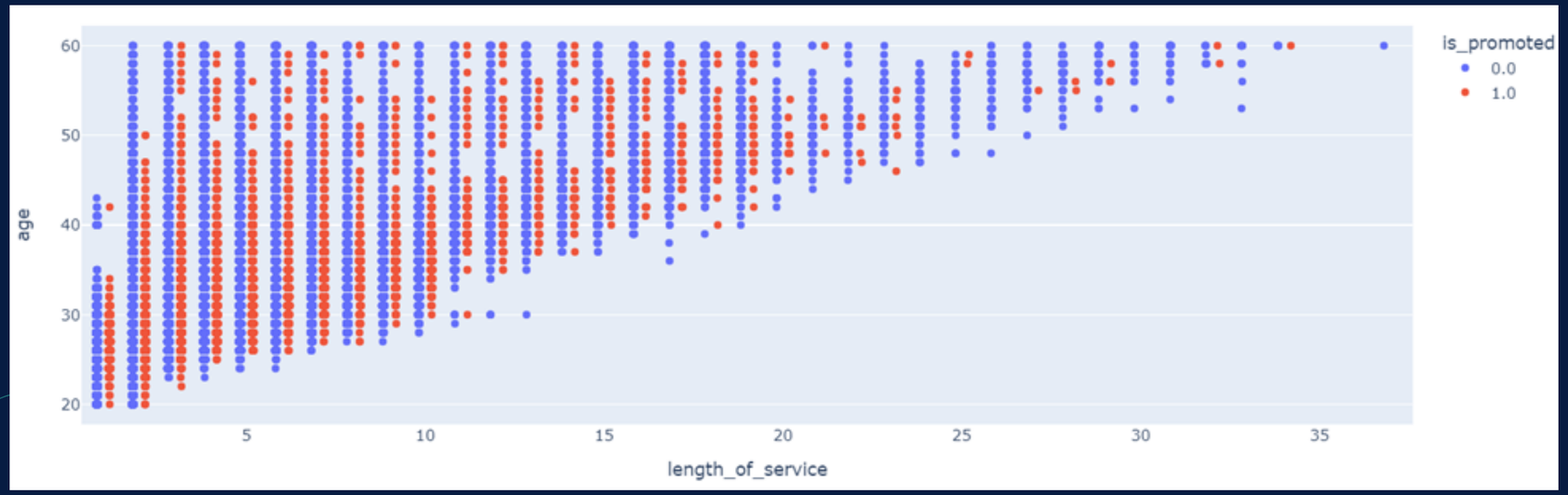
- Most employees belong to the age group of 20 to 40.
- A higher proportion of males receive promotions compared to females.
- One possible reason is the imbalanced representation of females in the dataset.
- Multiple factors can cause an imbalance in gender, e.g., variations in the demographics of the workforce or the methods used for sampling.
- After the age of 50, there are barely any female employees in the promoted section.




**is_promoted vs gender vs age**

# EDA

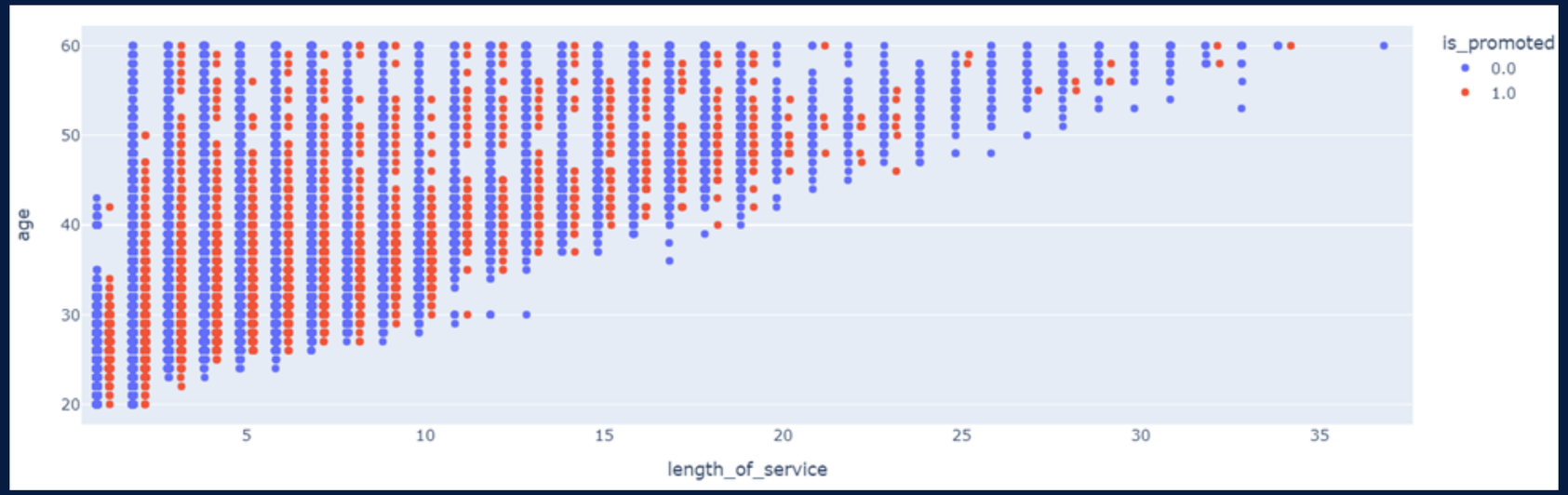
- There is a positive correlation between length of service and age.
- As employees stay longer in their job, the range of ages they belong to tends to expand or widen.
- This indicates that the seniority of employees in the company is positively correlated with their age.
- Less-experienced workers who are younger have had fewer opportunities to dedicate themselves to their careers and amass knowledge within the company.
- On the other hand, senior workers probably have a more extended history with the organization, suggesting greater length of service.




**is_promoted vs age vs length_of_service**

# Data Cleaning and Preprocessing

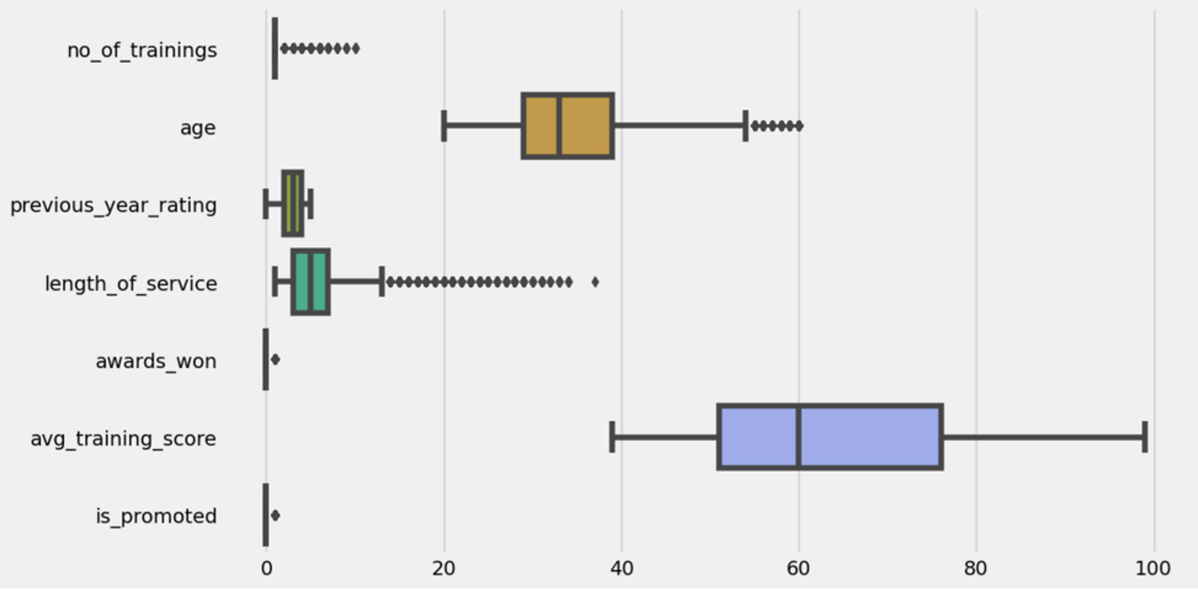
- There is a positive correlation between length of service and age.
- As employees stay longer in their job, the range of ages they belong to tends to expand or widen.
- This indicates that the seniority of employees in the company is positively correlated with their age.
- Less-experienced workers who are younger have had fewer opportunities to dedicate themselves to their careers and amass knowledge within the company.
- On the other hand, senior workers probably have a more extended history with the organization, suggesting greater length of service.




**is_promoted vs age vs length_of_service**

# Outliers

- Outliers are observations that deviate substantially from the rest of the data points.
- Outliers can be caused by a variety of reasons, such as data entry errors, merging two datasets, incomplete data or natural variation.
- Outliers can influence the results of statistical analysis, making it difficult to identify trends in the data and decreasing the accuracy of machine learning models.

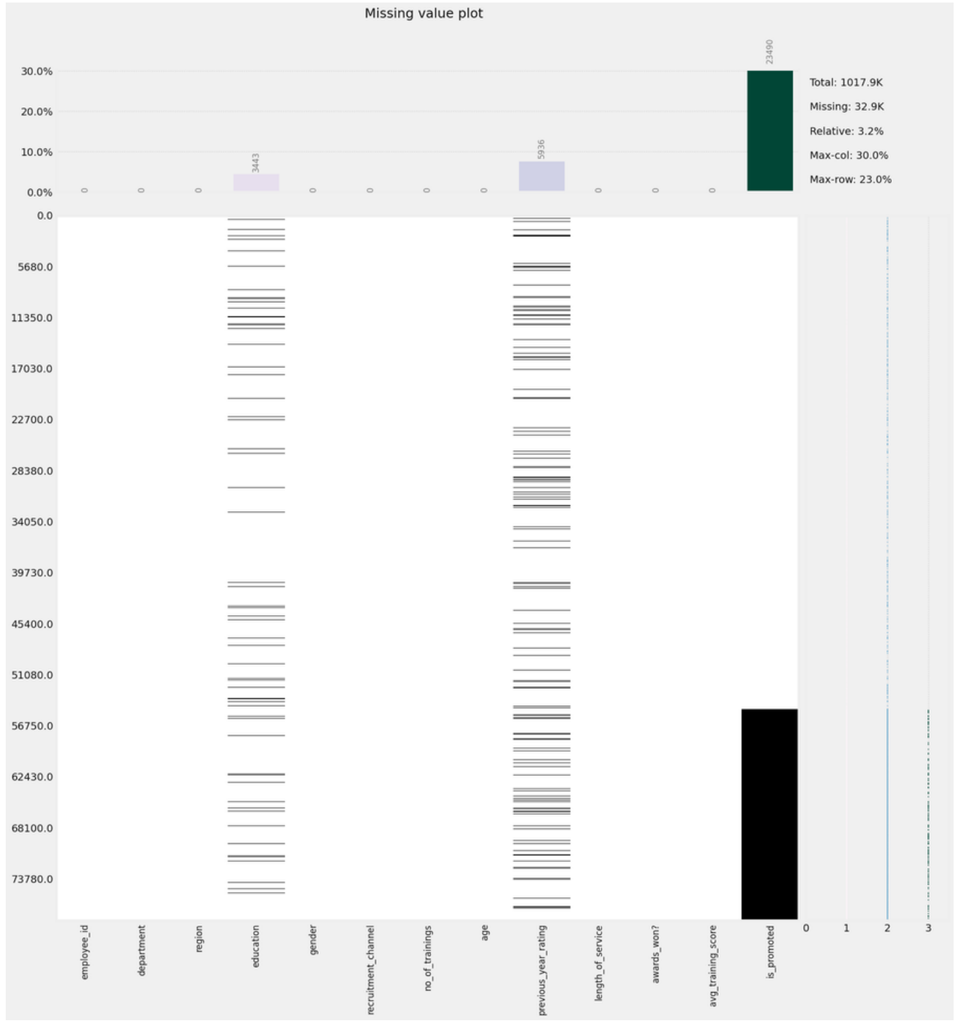



- The box plot is a useful tool for examining the central 50% of a dataset, providing clear indications of the minimum, maximum, median, and any outliers.
- Any data points beyond the maximum value/upper whisker/upper fence is called an outlier.
- According to the plot in the figure below, the only outliers that would affect the model accuracy would be the "length_of_service" outliers.

- This is because the other variable, "age", has a smaller number of unique values.
- The upper fence for "length_of_service" is 13, so any data points or values beyond this point will be removed.
- The outliers present in "age" are not significant enough to cause an impact to model accuracy.

# Missing/Null Values

- The "education" column has missing values at random.
- There is no correlation/pattern found for the null values.
- So, the missing values will be filled with the mode of that column as it is a categorical column.
- The mode of the "education" column is "Bachelor's degree".




- If someone's "length_of_service" is less than or equal to one, it means that they have would not have any "previous_year_rating" in that company.
- This proves your hypothesis that employees with a length of service of less than or equal to one do not have a previous_year_rating.
- Since they don't have rating, the value '0' will filled instead of null.

- Due to combining the train and test datasets, a large number of null values are present in is_promoted.
- Null values in "is_promoted" column will be filled with mode value.
- The "is_promoted" column is a highly imbalanced class.
- To fix this imbalance, apply the SMOTE technique during pre-processing.
- SMOTE is a technique for oversampling minority classes.
- Creates new instances of employees who were promoted in order to balance the classes.

# Klib

- klib is a Python library for data science and machine learning.
- It provides a variety of tools and functions for data loading, cleaning, preprocessing, and analysis.
- It is a popular choice for data scientists and machine learning engineers.

- It is built on top of popular Python libraries such as NumPy, Pandas, and Scikit-Learn.
- Klib is open source and available on GitHub.
- It is actively maintained and updated.

```
In [86]: data = klib.data_cleaning(data)

Shape of cleaned data: (77729, 12) - Remaining NAs: 0


Dropped rows: 569
     of which 569 duplicates. (Rows (first 150 shown): [1100, 4254, 5659, 5889, 5890, 7010, 7858, 7887, 8151, 8912, 9112, 9
374, 10170, 10887, 11795, 12740, 13326, 14053, 14328, 14336, 14829, 14847, 14875, 15015, 15139, 15211, 15363, 15598, 15742,
17313, 17314, 17340, 17395, 17717, 18097, 18367, 18374, 18541, 18655, 19893, 19936, 20009, 20209, 20236, 20610, 20914, 2143
3, 21873, 21958, 22089, 22115, 22848, 23140, 23202, 23500, 23791, 23819, 24014, 24389, 24570, 24778, 24882, 25218, 25400, 2
5448, 25644, 26120, 26300, 26935, 26978, 26994, 27597, 28047, 28180, 28317, 28709, 28978, 29039, 29107, 29376, 29565, 2997
4, 30082, 30155, 30252, 30296, 30617, 30670, 30756, 30775, 30800, 30894, 30904, 31408, 31437, 31834, 31917, 32207, 32245, 3
3186, 33214, 33349, 33636, 33644, 33748, 34040, 34180, 34324, 34428, 34483, 34669, 34993, 35088, 35089, 35196, 35223, 3526
8, 35303, 35324, 35351, 35407, 35424, 35571, 35598, 35743, 35778, 35888, 36126, 36458, 36657, 36868, 36892, 37068, 37093, 3
7171, 37422, 37596, 37722, 37739, 37816, 37823, 38325, 38473, 38710, 38858, 38899, 38948, 39356, 39411, 39671])

Dropped columns: 0
     of which 0 single valued.     Columns: []
Dropped missing values: 0
Reduced memory by at least: 8.44 MB (-86.3%)
```

- Data loading and preprocessing: provides a variety of tools for loading data from a variety of sources. It also provides a variety of tools for preprocessing data, such as cleaning, transforming, and normalizing data.
- Data analysis: klib provides a variety of tools for analyzing data, such as statistical analysis, machine learning, and natural language processing.
- Data visualization: klib provides a variety of tools for visualizing data, such as charts, graphs, and maps.

Label encoding is a common preprocessing step in machine learning, as many machine learning algorithms require numerical data.

This is done by assigning a unique integer value to each category.

Label encoding is a process of converting categorical data into numerical data.

| | department | region | education | gender | recruitment_channel | no_of_trainings | age | previous_year_rating | length_of_service | awards_won | avg_training_score | is_promoted |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 7 | 31 | 2 | 0 | 2 | 1 | 35 | 5.0 | 8 | 0 | 49 | 0.0 |
| 1 | 4 | 14 | 1 | 1 | 0 | 1 | 30 | 5.0 | 4 | 0 | 60 | 0.0 |
| 2 | 7 | 10 | 1 | 1 | 2 | 1 | 34 | 3.0 | 7 | 0 | 50 | 0.0 |
| 3 | 7 | 15 | 1 | 1 | 0 | 2 | 39 | 1.0 | 10 | 0 | 50 | 0.0 |
| 4 | 8 | 18 | 1 | 1 | 0 | 1 | 45 | 3.0 | 2 | 0 | 73 | 0.0 |

# Label Encoding

- SMOTE stands for Synthetic Minority Oversampling Technique.
- It is a data augmentation technique used to address the issue of imbalanced classes in machine learning.
- SMOTE works by creating synthetic minority instances by interpolating between existing minority instances.
- This helps to balance the dataset and improve the performance of machine learning models trained on the data.
- SMOTE is a popular and effective data augmentation technique, and it is available in many machine learning libraries.

# SMOTE Technique

# Comparison of Models

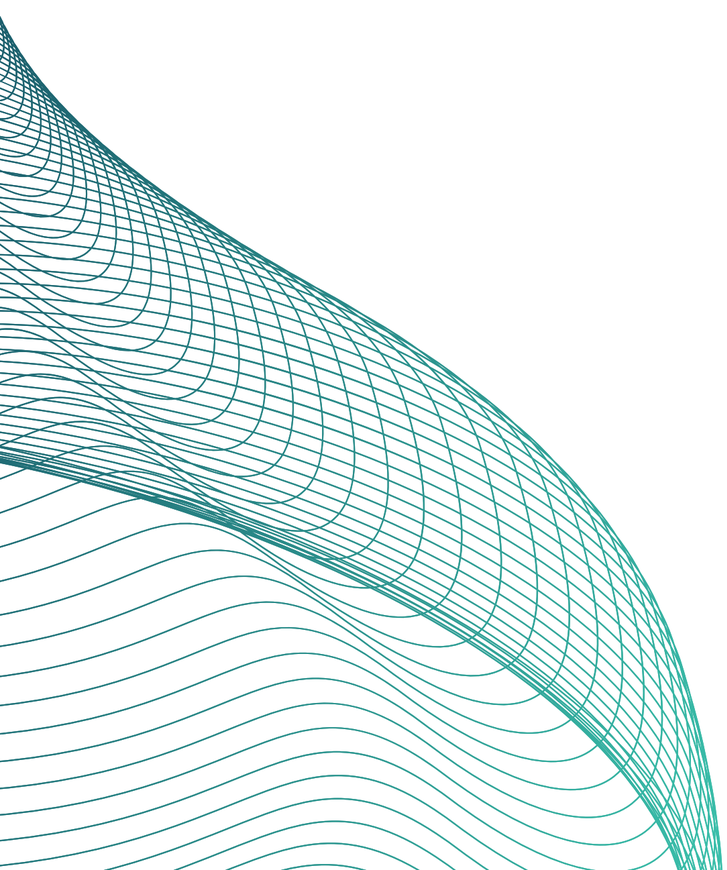
| Model | Accuracy |
|---|---|
| Random Forest Classifier | 0.959146 |
| Decision Tree Classifier | 0.932731 |
| Logistic Regression | 0.688633 |
| MLP Classifier | 0.762095 |
| Gaussian Naive Bayes | 0.669336 |
| C-Support Vector Classification | 0.697222 |

- Random Forest Classifier has highest accuracy, followed by Decision Tree Classifier. Logistic Regression has lowest accuracy.
- Choose model based on requirements: high accuracy (Random Forest or Decision Tree), simplicity (Logistic Regression), or managing intricate relationships (C-Support Vector Classification). Consider other factors too.

# Conclusion

- The project aims to predict employee promotions based on certain characteristics, indicating their success and achievements.
- Gender , previous year rating, average training score, and numbr of trainings were found to have an impact on the prediction effectiveness.
- The project involves data sorting, cleaning, exploration, model training, and evaluation, with a focus on achieving interpretability or comprehensibility of the model.
- The analysis identified specific factors, including gender, previous year rating, average training score, and number of trainings, as influential in predicting employee promotions.

# Future Works



- Need more data to improve prediction accuracy. Additional data pre-processing needed to enhance the model's performance.
- Data Augmentation: If the dataset is limited in size, data augmentation techniques can be employed to generate additional training samples. This can help improve the models' generalization and overall performance.
- Advanced Neural Network Architectures: Exploring more advanced architectures, such as deep neural networks, convolutional neural networks (CNNs), or recurrent neural networks (RNNs), could potentially yield better results, especially if the data has specific structural properties or sequential dependencies.
- Enhancing the user interface: The project could be improved by developing a more user-friendly interface, making it easier for non-technical users to access.

# धन्यवाद

JJR