Name- Maithil Deore
Internship Program- Data Science with Machine Learning and Python
Batch- Jan 2022 - Mar 2022
Certificate Code- TCRIB2R137
Date of submission- 6 April 2022

Technical Coding Research Innovation, Navi Mumbai, Maharashtra, India-410206

(HR EMPLOYEE ATTRITION ANALYSIS)

A **Case-Study Submitted for the requirement**

of Technical Coding Research Innovation

For the Internship Project work done during

DATASCIENCEWITHMACHINELEARNINGANDPYTHON INTERNSHIPPROGRAM

by

Maithil Deore(TCRIB2R137)

Rutuja Doiphode
CO-FOUNDER &CEO
TCR innovation

Abstract – Classification of large datasets is an important data mining problem.]Many classification algorithms have been proposed in the literature, but studies have shown that so far no algorithm uniformly outperforms all other algorithms in terms of quality[1]. When the dataset is too huge to fit in memory, rainforest is an algorithm for generating a decision tree (how to

Name- Maithil Deore
Internship Program- Data Science with Machine Learning and Python
Batch- Jan 2022 - Mar 2022
Certificate Code- TCRIB2R137
Date of submission- 6 April 2022

divide). In the rain forest, a split decision does not necessitate the use of the entire dataset.

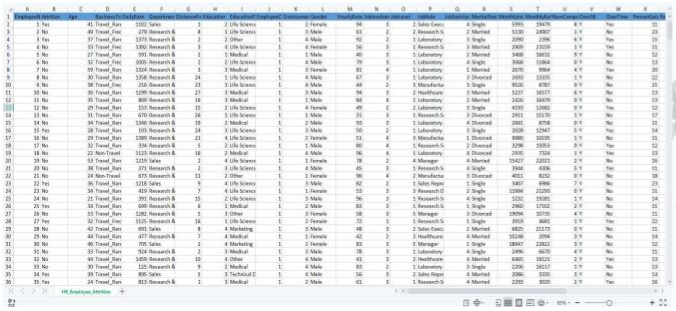
Index -

I. Aim

We need to know if the specific employee will depart from the company or not and our target column is Attrition.

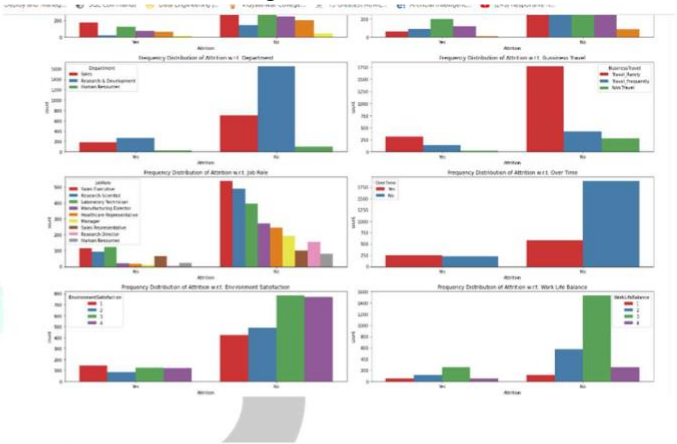
II. INTRODUCTION TO DATASET

Gender, age, business travel, department, education, relationship satisfaction, and other details are included in the "HR EMPLOYEE ATTRITION DATASER." The dataset contains data from 2940 employees, each of whom has 34 characteristics. There are both numerical and categorical data in the dataset. The dataset is depicted below:

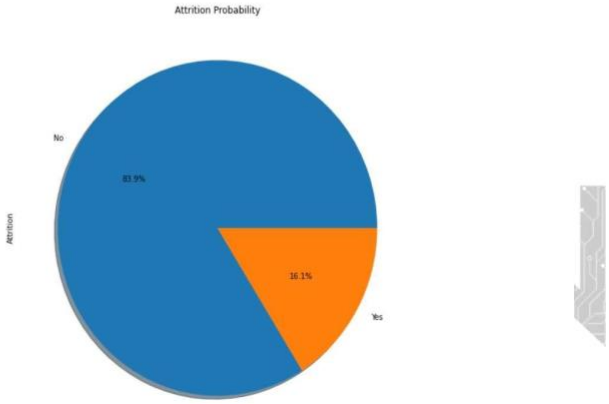


III. EXPLORATORY DATA ANALYSIS ON DATASET

Exploratory Data Analysis (EDA) on a dataset basically provides you a better knowledge of the whole thing. For example, if someone wishes to see if there are any (Not Any Value) NAN values in the dataset, EDA will assist us in finding them. Later, we can use other strategies to fix the problem of NAN values in the dataset, such as replacing the NAN values with the mean, median, or mode value. Fortunately, there are no NAN or outlier numbers in the "HR EMPLOYEE ATTRITION DATASET. Trying to figure out the frequency distribution of all categorical values in relation to attrition using EDA. The count graphs that describe the same thing are shown below:



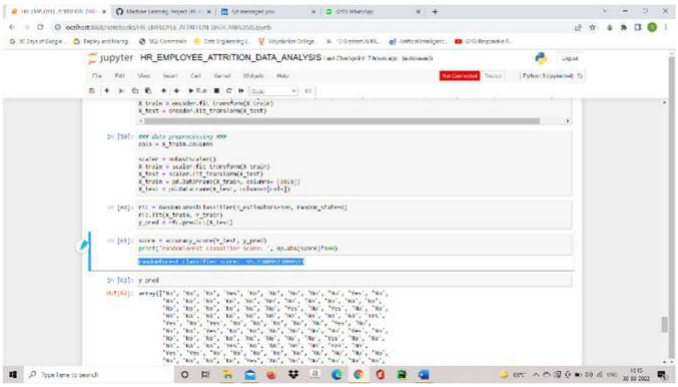
```
In [12]: f, ax = plt.subplots(figsize=(10,10))
         ax = data['Attrition'].value_counts(). plot.pie(explode=[0,0], autopct = '%1.1f%%', shadow=True)
         ax.set_title('Attrition Probability')
Out[12]: Text(0.5, 1.0, 'Attrition Probability')
```



IV. TRAINING & PREDICTION OF DATA

Random Forest Classification approach was used to train the machine learning model after examining the entire dataset because this algorithm works well with a huge number of features. To begin, the dataset was divided into two parts: 80 percent for training and 20 percent for testing. The following is how the machine learning model was trained and predicted to meet the goal:

Name- Maithil Deore
Internship Program- Data Science with Machine Learning and Python
Batch- Jan 2022 – Mar 2022
Certificate Code- TCRIB2R137
Date of submission- 6 April 2022



V. CONCLUSION

The machine learning model will be able to predict employee attrition with an accuracy of 95.23 percent after using the Random Forest Classification method. This isn't the only way to train the model for staff attrition prediction. It is feasible to forecast using different other methods, but I discovered that this approach outperforms all other classification algorithms in the "HR EMPLOYEE ATTRITION DATASET. Final output of the accuracy of model:

VI. FINAL RESULT

Final Result

Algorithm used	Accuracy Score
Random Forest	95%

VII.REFERENCES

[1] Rainforest- A Framework for fast Decision tree construction of large datasets