Project Report

On

IBM HR Analytics Employee Attrition & Performance

Submitted in partial fulfillment for the award of

Post Graduate Diploma in Big Data Analytics from **C-DAC Kharghar (Mumbai)**

Guided by:

Mr. Vijayant Soni

Presented by:

Swati Korade (PL)	**PRN: 220340325053**
Lokesh Patil	**PRN: 220340325022**
Parikshit Chaudhari	**PRN: 220340325031**
Rohit Mendhekar	**PRN: 220340325040**
Sujit Kolekar	**PRN: 220340325051**
Yogesh Nakhate	**PRN: 220340325060**

Centre of Development of Advanced Computing (C-DAC),Mumbai

CERTIFICATE

This is to certify that,

Swati Korade (PL)
Lokesh Patil
Parikshit Chaudhari
Rohit Mendhekar
Sujit Kolekar
Yogesh Nakhate

Have successfully completed their project on

IBM HR Analytics Employee Attrition & Performance

Under the guidance of Mr. Vijayant Soni

Project Guide **Project Supervisor**

HOD CDAC KHARGHAR

Mr. Dr. CP Johnson.

ACKNOWLEDGEMENT

This project **“IBM HR Analytics Employee Attrition & Performance ”** was a great learning experience for us and we are submitting this work to Advanced Computing Training School (CDAC Kharghar, Mumbai).

We all are very glad to mention the name of *Mr. Vijayant Soni* for his valuable guidance to work on this project. His guidance and support helped us to overcome various obstacles and intricacies during the course of project work.

We are highly grateful to Mr. Dr. CP Johnson (Senior Director C-DAC Mumbai, Kharghar) for her guidance and support whenever necessary while doing this course Post Graduate Diploma in *Big Data Analytics (PG -DBDA)*
Through C-DAC Kharghar, Mumbai.

Our most heartfelt thanks goes to *Mrs.Vineeta Singh* (Course Coordinator, *PG - DBDA*) who gave all the required support and kind coordination to provide all the necessities like required hardware, internet facility and extra Lab hours to complete the project and throughout the course up to the last day here in C-DAC ACTS, Pune.

From:

Swati Korade	(220340325053)
Lokesh Patil	(220340325022)
Parikshit Chaudhari	(220340325031)
Rohit Mendhekar	(220340325040)
Sujit Kolekar	(220340325051)
Yogesh Nakhate	(220340325060)

TABLE OF CONTENTS

1. Abstract

Decision-making plays an essential role in the management and may represent the most important component in the planning process. Employee attrition is considered a well-known problem that needs the right decisions from the administration to preserve high qualified employees. Interestingly, artificial intelligence is utilized extensively as an efficient tool for predicting such a problem.

The proposed work utilizes the deep learning technique along with some preprocessing steps to improve the prediction of employee attrition. Several factors lead to employee attrition. Such factors are analyzed to reveal their intercorrelation and to demonstrate the dominant ones. Our work was tested using the imbalanced dataset of IBM analytics, which contains 35 features for 1157595 rows.

To get realistic results, we derived a balanced version from the original one. Finally, cross-validation is implemented to evaluate our work precisely. Extensive experiments have been conducted to show the practical value of our work. The prediction accuracy using the original dataset is about 91%, whereas it is about 94% using a synthetic dataset.

2. Introduction and Overview of Project

The competition among organizations and firms highly depends on the productivity of the workforce. Building and maintaining a suitable environment is the key that contributes to stable and collaborative employees. The human resource (HR) department should participate in building such an environment by analyzing employees' database records. Analyzing these data enables the administration to improve the decision-making to avoid employee attrition. Employee attrition means that productive employees decide to leave the organization due to different reasons such as work pressure, unsuitable environment, or not satisfying salary. Employee attrition affects the organization's productivity because it loses a productive employee as well as other resources such as HR staff effort in recruiting new employees. Recruiting new employees requires training, development, and integrating them into the new environment.

Predicting employee attrition before it occurs can help the administration to prevent it or at least reduce its effect. Some literature suggested that happy and motivated employees tend to be more creative, productive, and perform better. Organizations can utilize their HR data to make such predictions depending on predictive models that can be built for this purpose. In recent years, Machine Learning is used in many different fields such as health, education, economy, and administration. Recently, the prediction of employee attrition using AI has received a lot of research attention. Also, the increased amount of data regarding this topic leads to more studies in this field.

In this project we used aws service S3 bucket to store raw data on cloud because of S3 bucket we can store and retrieve any amount of data anytime, anywhere. then we used another aws service EMR (previously called Amazon Elastic MapReduce) is a managed cluster platform that simplifies running big data frameworks, such as Apache Hadoop and Apache Spark, on AWS to process and analyze vast amounts of data. After creation of cluster, we used aapache spark for data processing. then we clean the data by using panda's library for Machine Learning Algorithms. After the clean data we read the clean data in python for pandas and ran some queries on it, so then we did EDA on the clean data and plot some graphs using Seaborn and matplotlib libraries to show various aspects of employee attrition. After understanding of data, we then performed feature selection and label encoding on the data. We then used a different machine learning algorithm to predict staff attrition.

3. Problem Statement

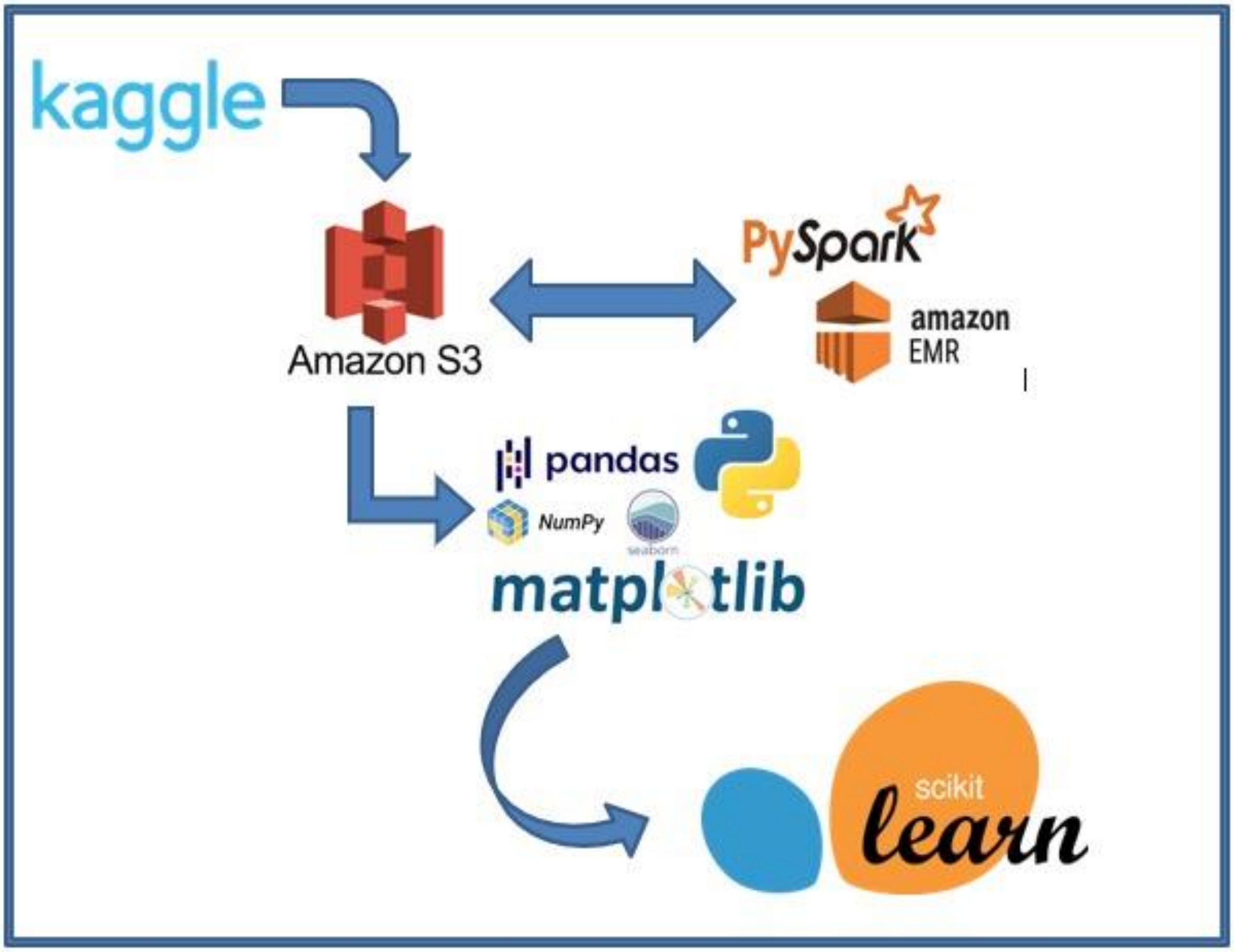
Employee attrition is a major cost to an organization and predicting such attritions is the most important requirement of the Human Resources department in many organizations. In this problem, your task is to predict the attrition rate of employees of an organization.

Objectives of the Project:

1.To analyze the important characteristics behind the attrition of employees.

2.Predicting employee may leave organization or not.

Work Flow of Project



4. Dataset Description

This data set we are using for this analysis is the employee survey from IBM, indicating if there is attrition or not. The data set consists of around 701292 samples. this dataset is of limited size. so, we are expecting a model to provide modest improvement in identifying a reason for attrition. While some level of attrition in an organization will be predict table, reducing the level of attrition and being ready for the cases that needs to be recovered will significantly help to maximize the organizational operations. As a future development with a an enough bigdata set, it will be quite used to run a segmentation on employees to create a specific risk category of employees. This can create a new knowledge for the organization that can help in driving attrition, knowledge that cannot be created by merely taking interviews from the employees.

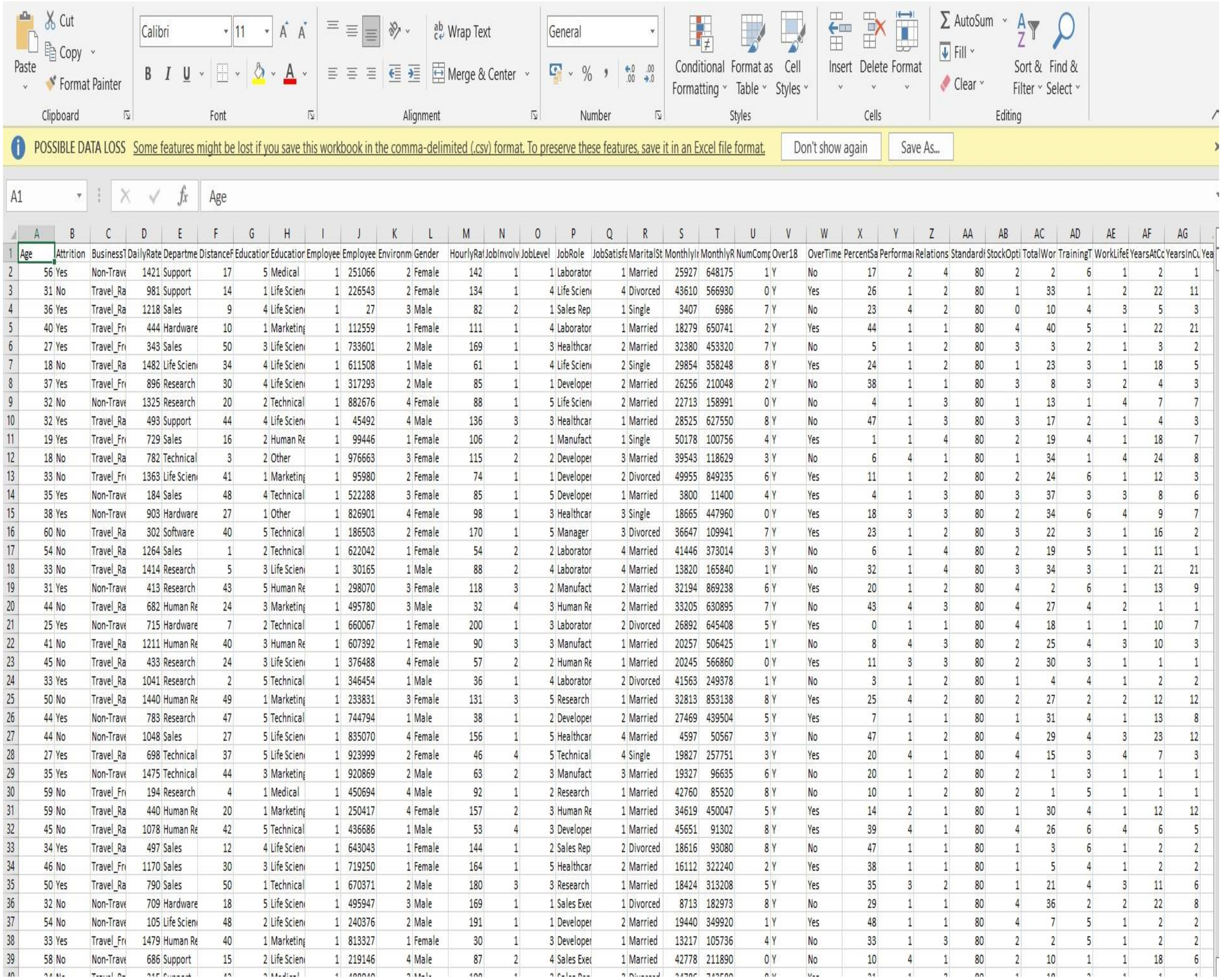
IBM has gathered information on employee satisfaction, income, seniority, and some demographics:

- Dataset Structure: 701292 observations (rows), 35 features (variables)
- Data Type: two datatypes in this dataset: objects and integers
- Imbalanced dataset: 402916 (57.5% of cases) employees did not leave the organization while 298376 (42.5% of cases) did leave the organization making us dataset to be considered imbalanced since more people stay in the organization than they actually leave

Name	Description
AGE	Numerical Value
Value ATTRITION	Employee leaving the company (0=no, 1=yes)
BUSINESS TRAVEL	(1=No Travel, 2=Travel Frequently, 3=Travel Rarely)
DAILY RATE	Numerical Value - Salary Level
DEPARTMENT	(1=HR, 2=R&D, 3=Sales)
DISTANCE FROM HOME	Numerical Value - THE DISTANCE FROM WORK TO HOME
EDUCATION	Numerical Value

EDUCATION FIELD	(1=HR, 2=LIFE SCIENCES, 3=MARKETING, 4=MEDICALSCIENCES, 5=OTHERS, 6= TECHNICAL)
EMPLOYEE COUNT	Numerical Value
EMPLOYEE NUMBER	Numerical Value - EMPLOYEE ID
ENVIRONMENTSATISFACTION	Numerical Value - SATISFACTION WITH THE ENVIRONMENT
GENDER	(1=FEMALE, 2=MALE)
HOURLY RATE	Numerical Value - HOURLY SALARY
JOB ROLE	(1=HC REP, 2=HR, 3=LAB TECHNICIAN, 4=MANAGER, 5=MANAGING DIRECTOR, 6= RESEARCH DIRECTOR, 7=RESEARCH SCIENTIST, 8=SALES EXECUTIVE, 9= SALES REPRESENTATIVE)
JOB SATISFACTION	Numerical Value - SATISFACTION WITH THE JOB
MARITAL STATUS	(1=DIVORCED, 2=MARRIED, 3=SINGLE)
MONTHLY INCOME	Numerical Value - MONTHLY SALARY
MONTHLY RATE	Numerical Value - MONTHLY RATE
NUMCOMPANIES WORKED	Numerical Value - NO. OF COMPANIES WORKED AT
OVER 18	(1=YES, 2=NO)
OVERTIME	(1=NO, 2=YES)
PERCENT SALARY HIKE	Numerical Value - PERCENTAGE INCREASE IN SALARY
PERFORMANCE RATING	Numerical Value - PERFORMANCE RATING
RELATIONS SATISFACTION	Numerical Value - RELATIONS SATISFACTION
STANDARD HOURS	Numerical Value - STANDARD HOURS
STOCK OPTIONS LEVEL	Numerical Value - STOCK OPTIONS
STOCK OPTIONS LEVE	Numerical Value - STOCK OPTIONS
TOTAL WORKING YEARS	Numerical Value - TOTAL YEARS WORKED
TRAINING TIMES LAST YEAR	Numerical Value - HOURS SPENT TRAINING
WORK-LIFE BALANCE	Numerical Value - TIME SPENT BETWEEN WORK ANDOUTSIDE
YEARS AT COMPANY	Numerical Value - TOTAL NUMBER OF YEARS AT THECOMPAN
YEARS IN CURRENT ROLE	Numerical Value - YEARS IN CURRENT ROLE
YEARS SINCE THE LASTPROMOTION	Numerical Value - LAST PROMOTION
YEARS WITH CURRENTMANAGER	Numerical Value - YEARS SPENT WITH CURRENTMANAGER

Data Source – (https://excelbianalytics.com/wp/downloads-21-sample-csv-files-data-sets-for-testing-5-million-records-hr-analytics-for-attrition till- /)



POSSIBLE DATA LOSS Some features might be lost if you save this workbook in the comma-delimited (.csv) format. To preserve these features, save it in an Excel file format.

A1 Age

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
1	Age	Attrition	BusinessT	DailyRate	Departme	DistanceF	Educatior	Educatior	Employee	Employee	Environm	Gender	HourlyRat	JobInvolv	JobLevel	JobRole	JobSatisfa	MaritalSt	MonthlyIr	MonthlyR	NumComp	Over18	OverTime	PercentSa	Performa	Relations	StandardH	StockOpti	TotalWor	TrainingT	WorkLifeE	YearsAtCo	YearsInCu
2	56	Yes	Non-Trave	1421	Support	17	5	Medical	1	251066	2	Female	142	1	1	Laborator	1	Married	25927	648175	1	Y	No	17	2	4	80	2	2	6	1	2	1
3	31	No	Travel_Ra	981	Support	14	1	Life Scien	1	226543	2	Female	134	1	4	Life Scien	4	Divorced	43610	566930	0	Y	Yes	26	1	2	80	1	33	1	2	22	11
4	36	Yes	Travel_Ra	1218	Sales	9	4	Life Scien	1	27	3	Male	82	2	1	Sales Rep	1	Single	3407	6986	7	Y	No	23	4	2	80	0	10	4	3	5	3
5	40	Yes	Travel_Fr	444	Hardware	10	1	Marketing	1	112559	1	Female	111	1	4	Laborator	1	Married	18279	650741	2	Y	Yes	44	1	1	80	4	40	5	1	22	21
6	27	Yes	Travel_Fr	343	Sales	50	3	Life Scien	1	733601	2	Male	169	1	3	Healthcar	2	Married	32380	453320	7	Y	No	5	1	2	80	3	3	2	1	3	2
7	18	No	Travel_Ra	1482	Life Scien	34	4	Life Scien	1	611508	1	Male	61	1	4	Life Scien	2	Single	29854	358248	8	Y	Yes	24	1	2	80	1	23	3	1	18	5
8	37	Yes	Travel_Fr	896	Research	30	4	Life Scien	1	317293	2	Male	85	1	1	Developer	2	Married	26256	210048	2	Y	No	38	1	1	80	3	8	3	2	4	3
9	32	No	Non-Trave	1325	Research	20	2	Technical	1	882676	4	Female	88	1	5	Life Scien	2	Married	22713	158991	0	Y	No	4	1	3	80	1	13	1	4	7	7
10	32	Yes	Travel_Ra	493	Support	44	4	Life Scien	1	45492	4	Male	136	3	3	Healthcar	1	Married	28525	627550	8	Y	No	47	1	3	80	3	17	2	1	4	3
11	19	Yes	Travel_Fr	729	Sales	16	2	Human Re	1	99446	1	Female	106	2	1	Manufact	1	Single	50178	100756	4	Y	Yes	1	1	4	80	2	19	4	1	18	7
12	18	No	Travel_Ra	782	Technical	3	2	Other	1	976663	3	Female	115	2	2	Developer	3	Married	39543	118629	3	Y	No	6	4	1	80	1	34	1	4	24	8
13	33	No	Travel_Fr	1363	Life Scien	41	1	Marketing	1	95980	2	Female	74	1	1	Developer	2	Divorced	49955	849235	6	Y	Yes	11	1	2	80	2	24	6	1	12	3
14	35	Yes	Non-Trave	184	Sales	48	4	Technical	1	522288	3	Female	85	1	5	Developer	1	Married	3800	11400	4	Y	Yes	4	1	3	80	3	37	3	3	8	6
15	38	Yes	Non-Trave	903	Hardware	27	1	Other	1	826901	4	Female	98	1	3	Healthcar	3	Single	18665	447960	0	Y	Yes	18	3	3	80	2	34	6	4	9	7
16	60	No	Travel_Ra	302	Software	40	5	Technical	1	186503	2	Female	170	1	5	Manager	3	Divorced	36647	109941	7	Y	Yes	23	1	2	80	3	22	3	1	16	2
17	54	No	Travel_Ra	1264	Sales	1	2	Technical	1	622042	1	Female	54	2	2	Laborator	4	Married	41446	373014	3	Y	No	6	1	4	80	2	19	5	1	11	1
18	33	No	Travel_Ra	1414	Research	5	3	Life Scien	1	30165	1	Male	88	2	4	Laborator	4	Married	13820	165840	1	Y	No	32	1	4	80	3	34	3	1	21	21
19	31	Yes	Non-Trave	413	Research	43	5	Human Re	1	298070	3	Female	118	3	2	Manufact	2	Married	32194	869238	6	Y	Yes	20	1	2	80	4	2	6	1	13	9
20	44	No	Travel_Ra	682	Human Re	24	3	Marketing	1	495780	3	Male	32	4	3	Human Re	2	Married	33205	630895	7	Y	No	43	4	3	80	4	27	4	2	1	1
21	25	Yes	Non-Trave	715	Hardware	7	2	Technical	1	660067	1	Female	200	1	3	Laborator	2	Divorced	26892	645408	5	Y	Yes	0	1	1	80	4	18	1	1	10	7
22	41	No	Travel_Ra	1211	Human Re	40	3	Human Re	1	607392	1	Female	90	3	3	Manufact	1	Married	20257	506425	1	Y	No	8	4	3	80	2	25	4	3	10	3
23	45	No	Travel_Ra	433	Research	24	3	Life Scien	1	376488	4	Female	57	2	2	Human Re	1	Married	20245	566860	0	Y	Yes	11	3	3	80	2	30	3	1	1	1
24	33	Yes	Travel_Ra	1041	Research	2	5	Technical	1	346454	1	Male	36	1	4	Laborator	2	Divorced	41563	249378	1	Y	No	3	1	2	80	1	4	4	1	2	2
25	50	No	Travel_Ra	1440	Human Re	49	1	Marketing	1	233831	3	Female	131	3	5	Research	1	Married	32813	853138	8	Y	Yes	25	4	2	80	2	27	2	2	12	12
26	44	Yes	Non-Trave	783	Research	47	5	Technical	1	744794	1	Male	38	1	2	Developer	2	Married	27469	439504	5	Y	Yes	7	1	1	80	1	31	4	1	13	8
27	44	No	Non-Trave	1048	Sales	27	5	Life Scien	1	835070	4	Female	156	1	5	Healthcar	4	Married	4597	50567	3	Y	No	47	1	2	80	4	29	4	3	23	12
28	27	Yes	Travel_Ra	698	Technical	37	5	Life Scien	1	923999	2	Female	46	4	5	Technical	4	Single	19827	257751	3	Y	Yes	20	4	1	80	4	15	3	4	7	3
29	35	Yes	Non-Trave	1475	Technical	44	3	Marketing	1	920869	2	Male	63	2	3	Manufact	3	Married	19327	96635	6	Y	No	20	1	2	80	2	1	3	1	1	1
30	59	No	Travel_Fr	194	Research	4	1	Medical	1	450694	4	Male	92	1	2	Research	1	Married	42760	85520	8	Y	No	10	1	2	80	2	1	5	1	1	1
31	59	No	Travel_Ra	440	Human Re	20	1	Marketing	1	250417	4	Female	157	2	3	Human Re	1	Married	34619	450047	5	Y	Yes	14	2	1	80	1	30	4	1	12	12
32	45	No	Travel_Ra	1078	Human Re	42	5	Technical	1	436686	1	Male	53	4	3	Developer	1	Married	45651	91302	8	Y	Yes	39	4	1	80	4	26	6	4	6	5
33	34	Yes	Travel_Ra	497	Sales	12	4	Life Scien	1	643043	1	Female	144	1	2	Sales Rep	2	Divorced	18616	93080	8	Y	No	47	1	1	80	1	3	6	1	2	2
34	46	No	Travel_Fr	1170	Sales	30	3	Life Scien	1	719250	1	Female	164	1	5	Healthcar	2	Married	16112	322240	2	Y	Yes	38	1	1	80	1	5	4	1	2	2
35	50	Yes	Travel_Ra	790	Sales	50	1	Technical	1	670371	2	Male	180	3	3	Research	1	Married	18424	313208	5	Y	Yes	35	3	2	80	1	21	4	3	11	6
36	32	No	Non-Trave	709	Hardware	18	5	Life Scien	1	495947	3	Male	169	1	1	Sales Exe	1	Divorced	8713	182973	8	Y	No	29	1	1	80	4	36	2	2	22	8
37	54	No	Non-Trave	105	Life Scien	48	2	Life Scien	1	240376	2	Male	191	1	1	Developer	2	Married	19440	349920	1	Y	Yes	48	1	1	80	4	7	5	1	2	2
38	33	Yes	Travel_Fr	1479	Human Re	40	1	Marketing	1	813327	1	Female	30	1	3	Developer	1	Married	13217	105736	4	Y	No	33	1	3	80	2	2	5	1	2	2
39	58	No	Non-Trave	686	Support	15	2	Life Scien	1	219146	4	Male	87	2	4	Sales Exe	1	Married	42778	211890	0	Y	No	10	4	1	80	2	10	1	1	18	6

Fig. Dataset

4.1 Lable Encoding:

In machine learning, we usually deal with datasets that contain multiple labels in one or more than one columns. These labels can be in the form of words or numbers. To make the data understandable or in human-readable form, the training data is often labelled in words.

Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.
We have done label encoding by
using get_dummies function of pandas library. In our dataset, we applied label encoding on following categorical columns:
1 BusinessTravel
2 Department
3 Education
4 EducationField
5 EnvironmentSatisfaction
6 Gender
7 JobInvolvement
8 JobRole
9 JobSatisfaction
10 MaritalStatus
11 NumCompaniesWorked
12 OverTime
13 PerformanceRating
14 RelationshipSatisfaction
15 TrainingTimesLastYear
16 WorkLifeBalance

5. Data Pre-processing and Cleaning

DATA CLEANING:

- Missing Data: There is no missing data.

- Data Type: We have two data types in this dataset: Categorical and Numerical.

- The label "Attrition" is the label in our dataset and we would like to find out why employees are leaving the organization!

- Data standardization: Data standardization is critical to facilitating and improving the use of data, especially as related to data portability (i.e., the ability to transfer data without affecting its content) and interoperability (i.e., the ability to integrate two or more datasets).

- Label Encoding: To make the data understandable or in human readable form, the training data is often labeled in words. Label Encoding refers to converting the labels into numeric form so as to convert it into the machine-readable form. Machine learning algorithms can then decide in a better way on how those labels must be operated.

- In our dataset there are 35 variables however, sum features just have one data level that do not make sense for our research such as Employee count Over 18, standard hours and employee number doesn't have meaning in analyzing result so we are deleted these features. In addition, some of the features having high correlation between them but not greater than 0.75 so we are not removing those variables.

- Imbalanced dataset: 402916 (57.5% of cases) employees did not leave the organization while 298376 (42.5% of cases) did leave the organization making our dataset to be considered imbalanced since more people stay in the organization than they actually leave.

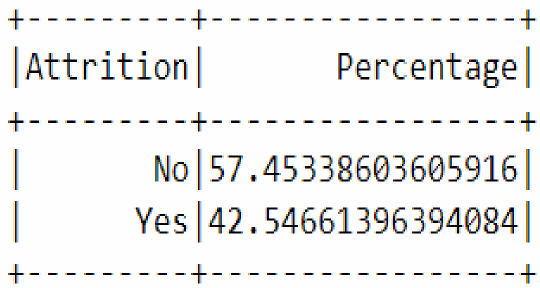
6. Exploratory Data Analysis

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

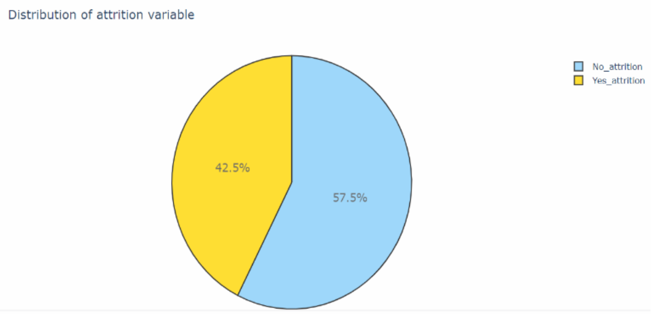
Univariate Data Analysis:

What is the percentage of attrition in the company?

In our dataset target variable is ‘Attrition’. Our target variable is binary so that it is necessary to check data is balanced or not.



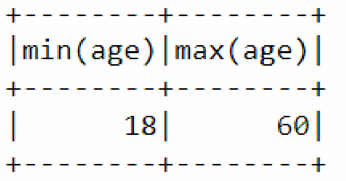
```
+---------+-----------------+
|Attrition|       Percentage|
+---------+-----------------+
|       No|57.45338603605916|
|      Yes|42.54661396394084|
+---------+-----------------+
```



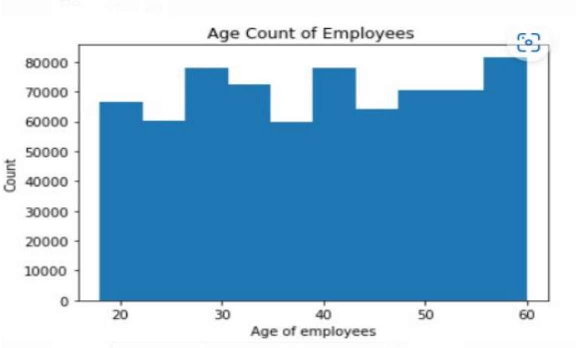
Distribution of attrition variable

Conclusion: From above pie chart, we come to know that the % of No attrition is more than % of Yes attrition by 14.8%.

What is the age range of the employees in the office?



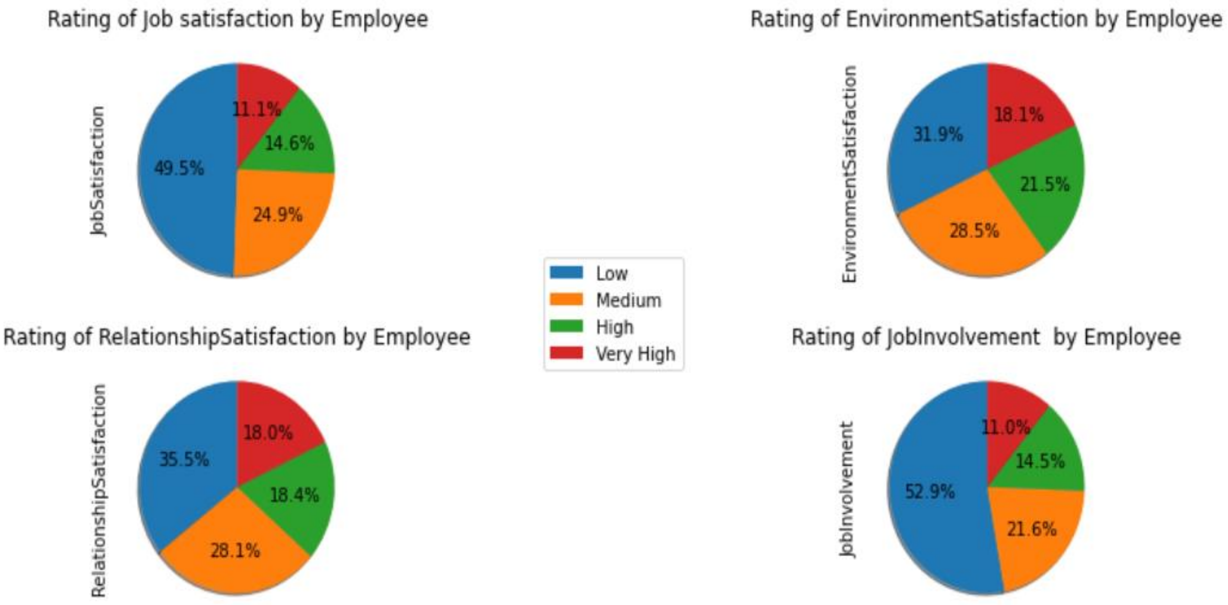
```
+--------+--------+
|min(age)|max(age)|
+--------+--------+
|      18|      60|
+--------+--------+
```



From above graph knows that the count of employee is higher over 58 ages.

What are the main age groups in the company and what percentage they hold?

Analysis of rating features



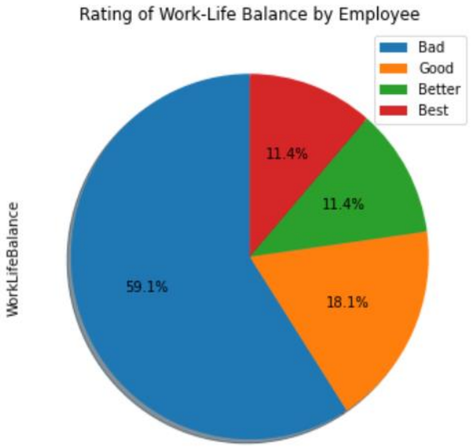
Conclusion: From the above chart we can say that

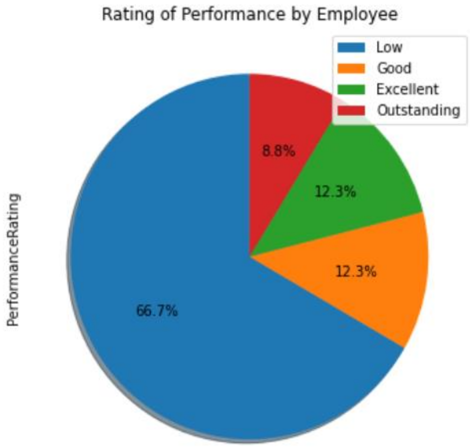
(1) 49.2% of total count of employee are not satisfied with their job role.

(2) 31.9% of employees are not having environment satisfaction.

(3) 185978 i.e. almost 35.4% employees are not satisfied by their relationship with other subordinates or office colleagues.

(4) The maximum i.e. 52.6% of employees are not sufficiently involved in their job role which may also results in attrition.



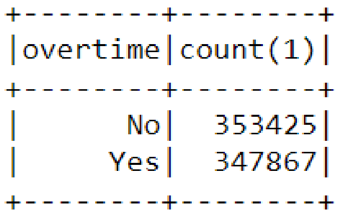


Conclusion:

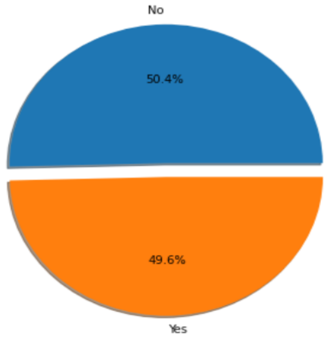
1. The above plot shows that 58.8% of the employee have rated their work life balance as bad.

2. An almost of 66.3% of the employees earned low performance rating.

Show the distribution of overtime done by the employees.



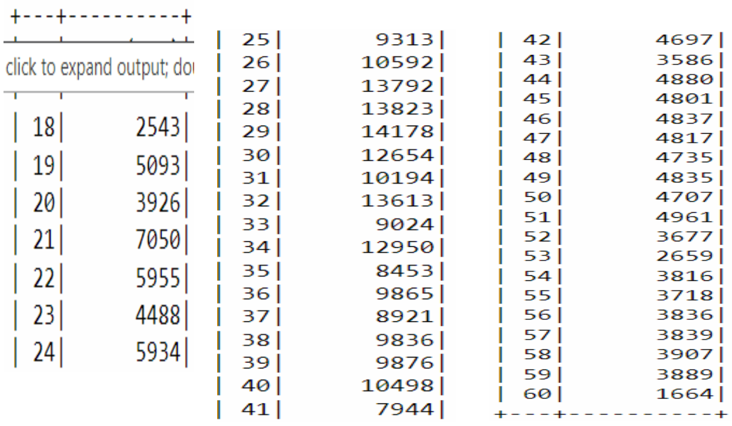
```
+--------+--------+
|overtime|count(1)|
+--------+--------+
|      No|  353425|
|     Yes|  347867|
+--------+--------+
```



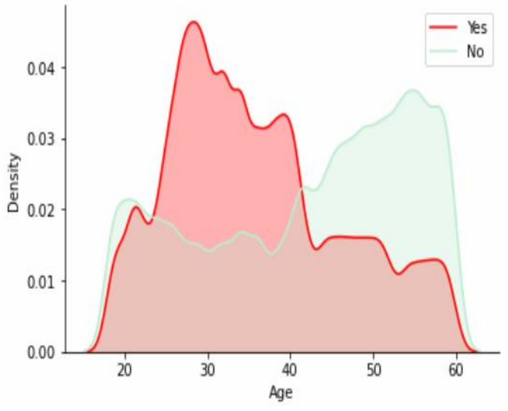
Conclusion: The above chart tells us that overall, 50.4% of total employees don't work for overtime.

Bivariate analysis:

Which age group is showing the maximum attrition? Is it the younger people or middle age group?

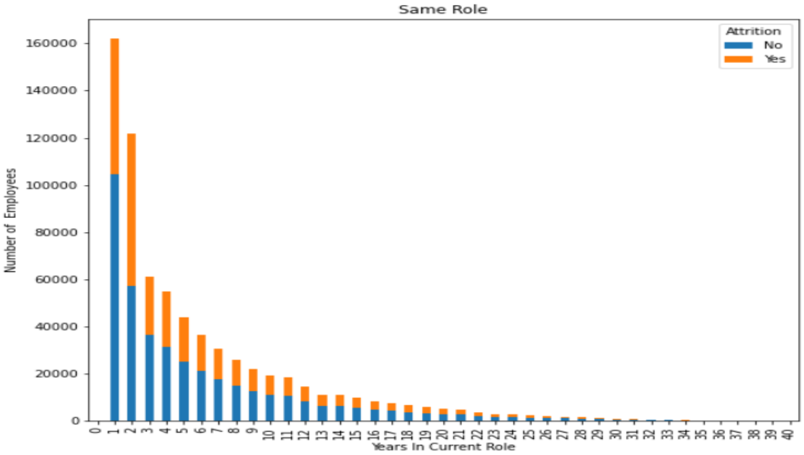


```
+---+----------+
click to expand output; do
| 18|      2543|
| 19|      5093|
| 20|      3926|
| 21|      7050|
| 22|      5955|
| 23|      4488|
| 24|      5934|

| 25|      9313|
| 26|     10592|
| 27|     13792|
| 28|     13823|
| 29|     14178|
| 30|     12654|
| 31|     10194|
| 32|     13613|
| 33|      9024|
| 34|     12950|
| 35|      8453|
| 36|      9865|
| 37|      8921|
| 38|      9836|
| 39|      9876|
| 40|     10498|
| 41|      7944|

| 42|      4697|
| 43|      3586|
| 44|      4880|
| 45|      4801|
| 46|      4837|
| 47|      4817|
| 48|      4735|
| 49|      4835|
| 50|      4707|
| 51|      4961|
| 52|      3677|
| 53|      2659|
| 54|      3816|
| 55|      3718|
| 56|      3836|
| 57|      3839|
| 58|      3907|
| 59|      3889|
| 60|      1664|
+---+----------+
```



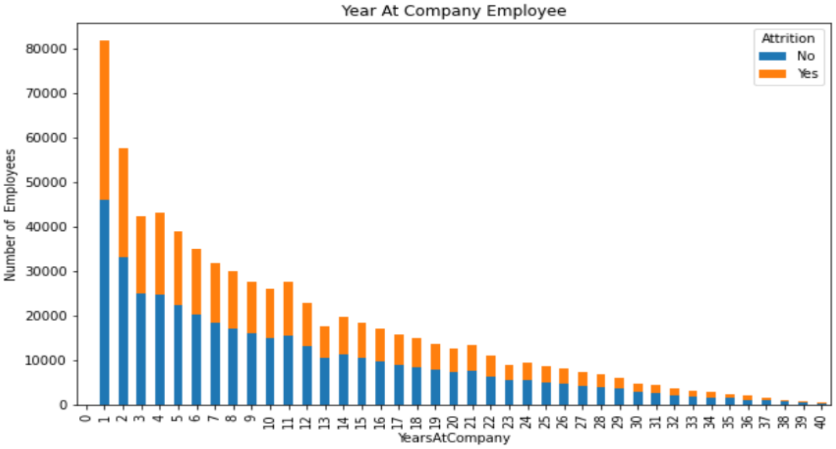
Conclusion: The above displot shows that the attrition rate is highest in the age group of 20 to 40 and then the attrition rate decreased for the age above 45. There are 62.57% People left the company whose age between 25 to 41.

Attrition rate vs years in current role



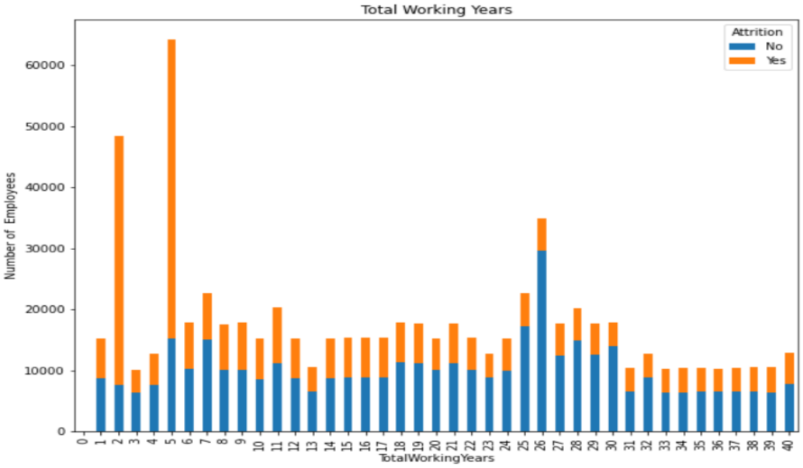
Conclusion: From the above chart which is showing the comparison of attrition of employees having varied number of years at the company, we come to know that the attrition rate is at the highest level in the period on 0-1 year and at the same time by observation we come to know that in comparison of the period of 0-2 years and 2-3 year, we have seen a sudden drop in the rate of employee attrition.

Attrition rate vs years at company



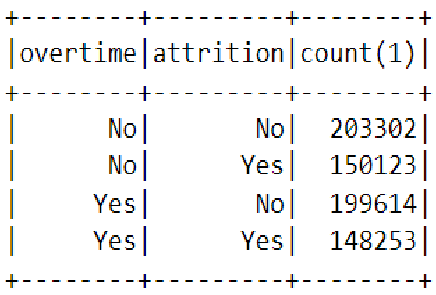
Conclusion: From the above chart we observed that the employees with less of years of experience or have invested less years with same company are having highest rate of quitting the job. Thus, more concern should be given to the freshers or experience lacker and the causes behind quitting the job by newly arrived professionals should be investigated.

Total working years vs attrition

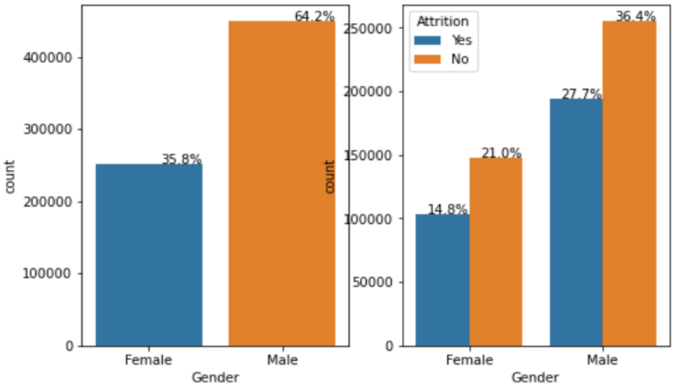


Conclusion: It is observed that there is high number of employees who leaves company at (0-5) years of experience. So it is very important that company to create such a policy to handle freshers so that they don't leave the company at the start of their career.

Did the people who do over time attrited more?

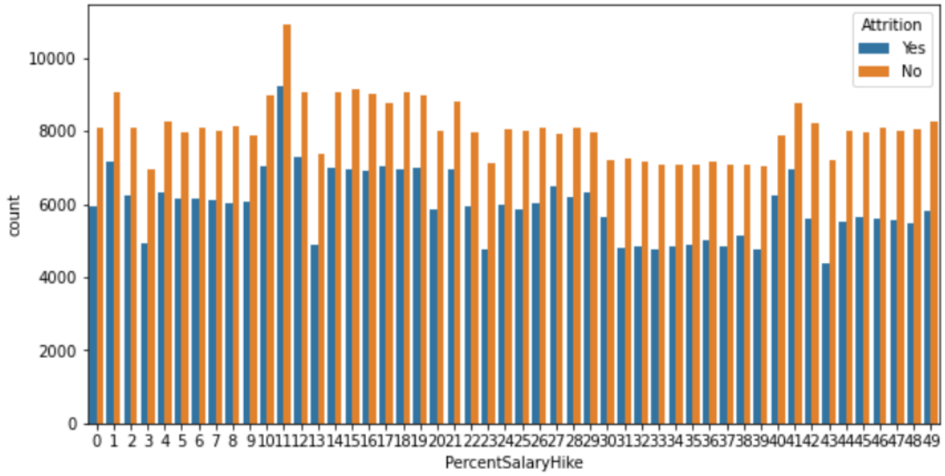


```
+--------+---------+--------+
|overtime|attrition|count(1)|
+--------+---------+--------+
|      No|       No|  203302|
|      No|      Yes|  150123|
|     Yes|       No|  199614|
|     Yes|      Yes|  148253|
+--------+---------+--------+
```



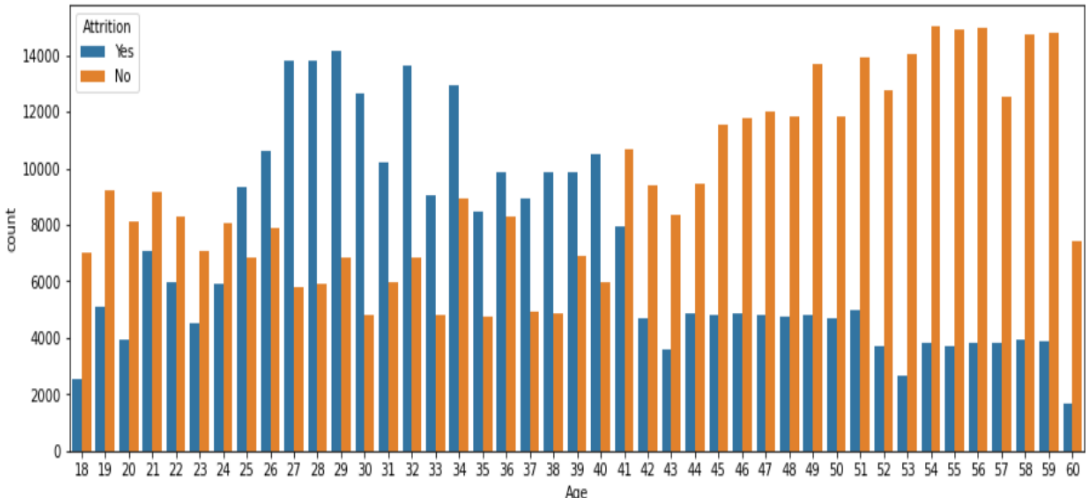
Conclusion: The rate of attrition of male employee is more i.e. 8.6% as compared to attrition of female employee i.e. 6.3%. More specifically we can say that attrition in male employees is more than by 43366 in female employees.

Is Attrition depending on percentage salary hike?



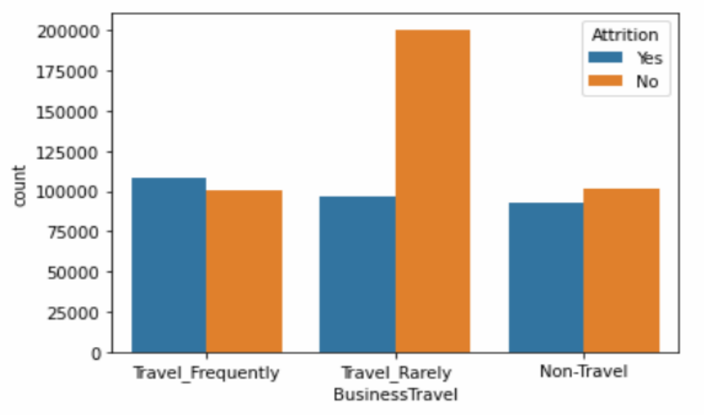
Conclusion: By observing above chart, we come to the conclusion that while giving salary hike, attrition is at the highest at the 11% as well as we can also say that discussing salary hike ranging from 11% to 13% tends to give attrition rate from maximum to minimum. We can also say that company should try to give average hike by 13% so that attrition will be minimal.

What is effect of age on attrition?



Conclusion: From the above chart which is showing the attrition rate according to age distribution here we come to know that at the age of 27 to 41 there is employee attrition rate is highest, on the other side experience peoples attrition rate is low.

How Is Attrition Affected by business travels?

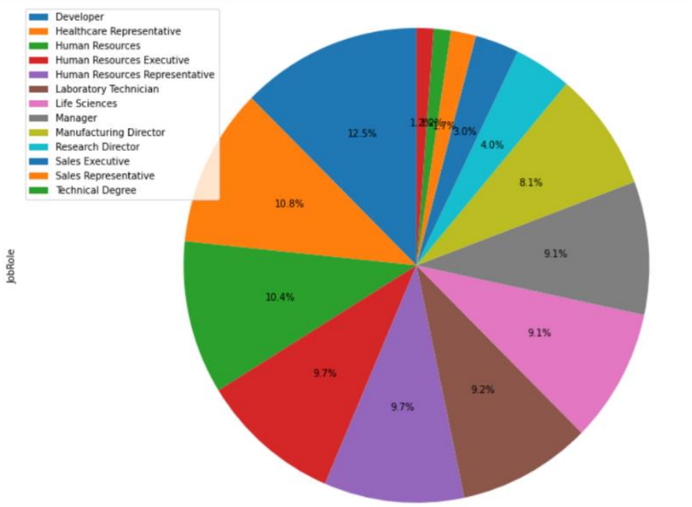


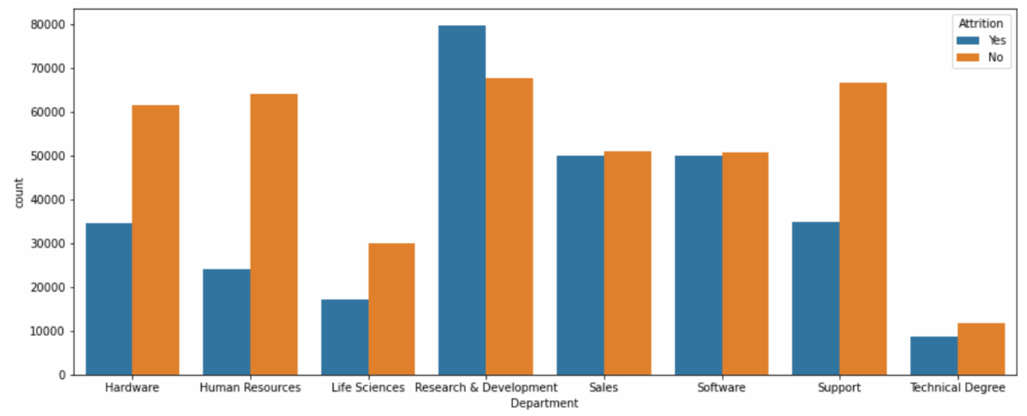
Conclusion: Most employees who travel rarely don't leave the company. From the above plot we come to know that sending employees on business travels are not doesn't really make much of a difference and doesn't have a significant effect on attrition.

So the Business Travel is not one of the main factors causing attrition but who travel frequently have biggest percentage

Best way to reduce this attrition is to conduct monthly survey and to assign travel according to the employee business travel interest.

What is the most Department of attritions?

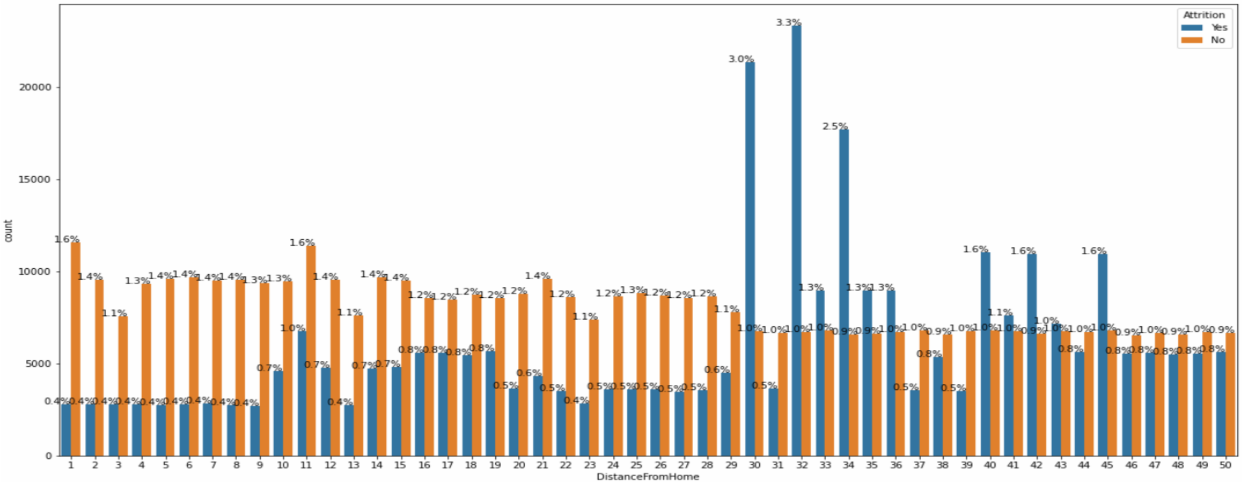




Conclusion: By observing above chart, we come to know that most attritions are from the research & development department. support has the least number of attritions. But we need to keep in mind that HR has a lot more employees than another department.

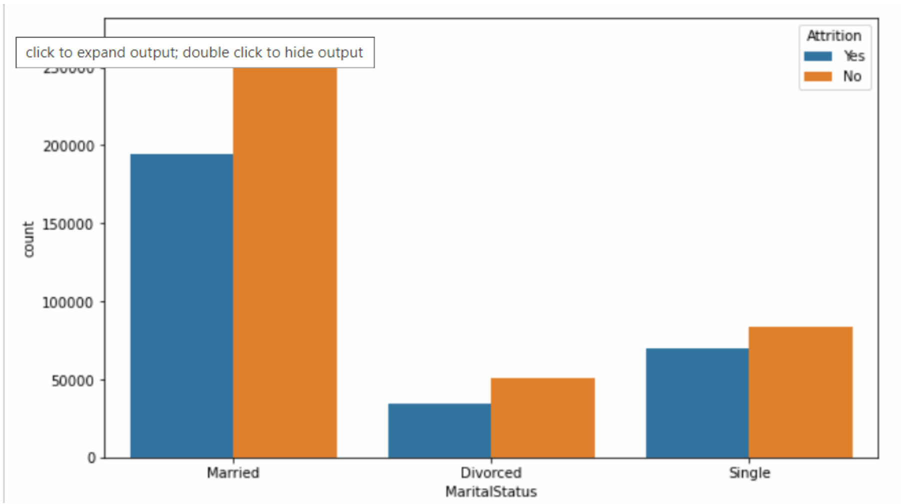
If we considered percentage of attritions per department, we would see that the HR department has most attritions.

what is effect of the distance from home on attrition?



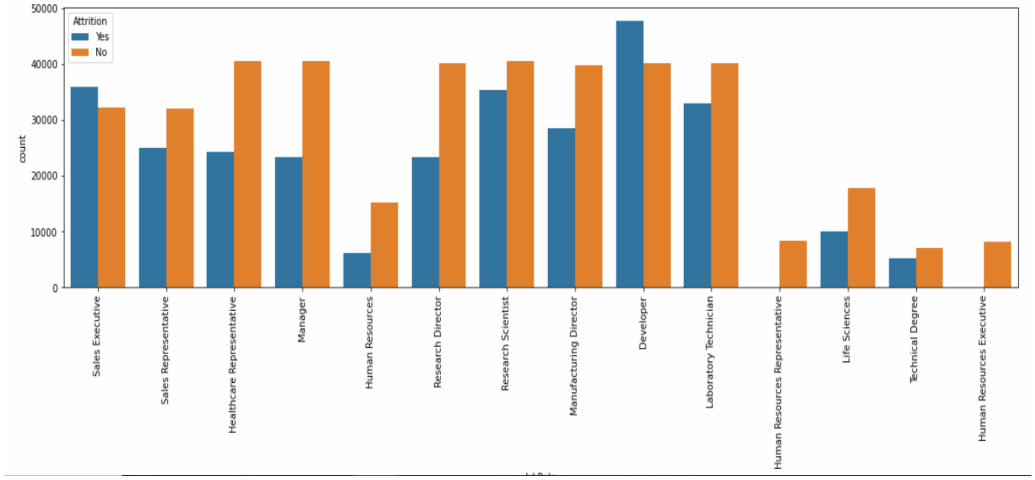
Conclusion: By observing above chart, we come to know that Most of the people who leave the company are located more than 28 km away from the company. This might be reason of employee attrition.

Employee attrition on the basis of marital status?

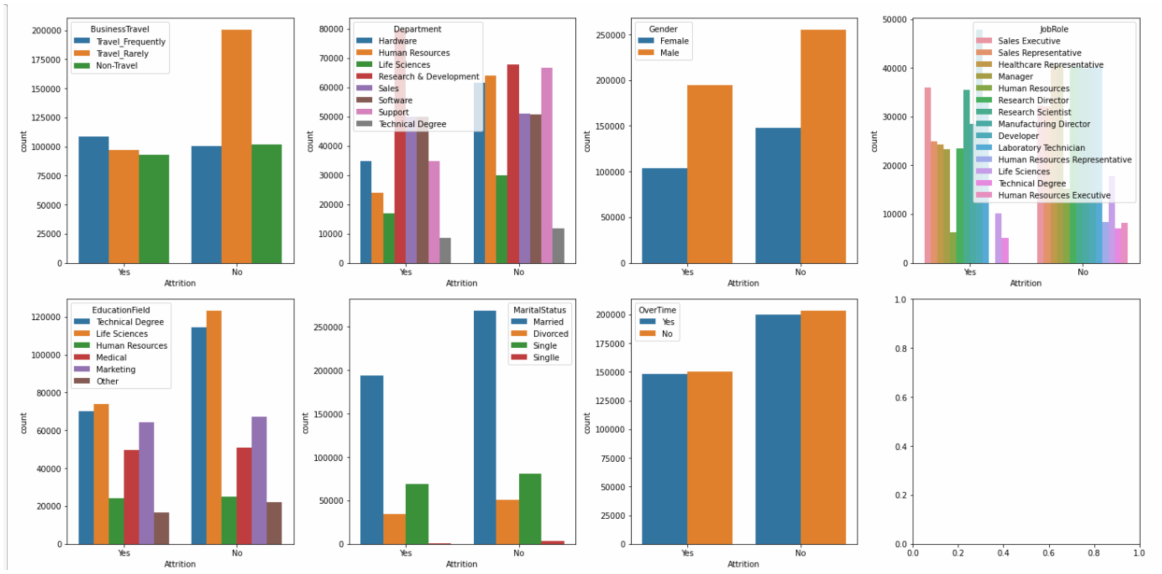


Conclusion: From the above graph we can say that the count of married employees to leave the company is high. but It is also important that the company has a large number of married employeees.

Which job role have more attrition?

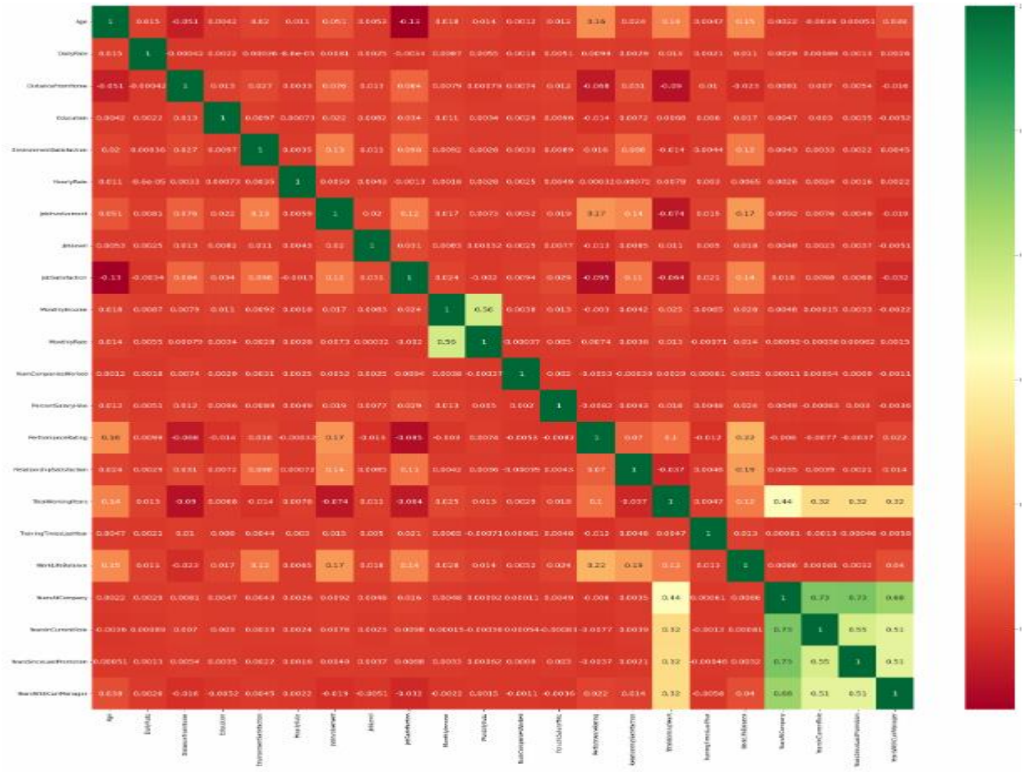


Conclusion: From the above graph we can say that most of the developers leave the organization around 35748. Therefore, there is a need to give promotion and incentives according to work. The manager should try to communicate with the employees to reduce the attrition rate.



This graph shows graphical representation of categorical columns.

Multivariate Data Analysis:



Observation: Variables like JobLevel, MonthlyIncome, TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsWithCurrentManager are highly correlated. These variables may lead to multicollinearity.

Conclusion: There are many factors that make an employee resign. Using the IBM dataset, some interesting insights were obtained. These insights can be used to build the model.

7. Model Building

1. Train/Test split:

One important aspect of all machine learning models is to determine their accuracy. Now, in order to determine their accuracy, one can train the model using the given dataset and then predict the response values for the same dataset using that model and hence, find the accuracy of the model. A better option is to split our data into two parts: first one for training our machine learning model, and second one for testing our model.

- Split the dataset into two pieces: a training set and a testing set.
- Train the model on the training set.
- Test the model on the testing set, and evaluate how well our model did.

Advantages of train/test split:

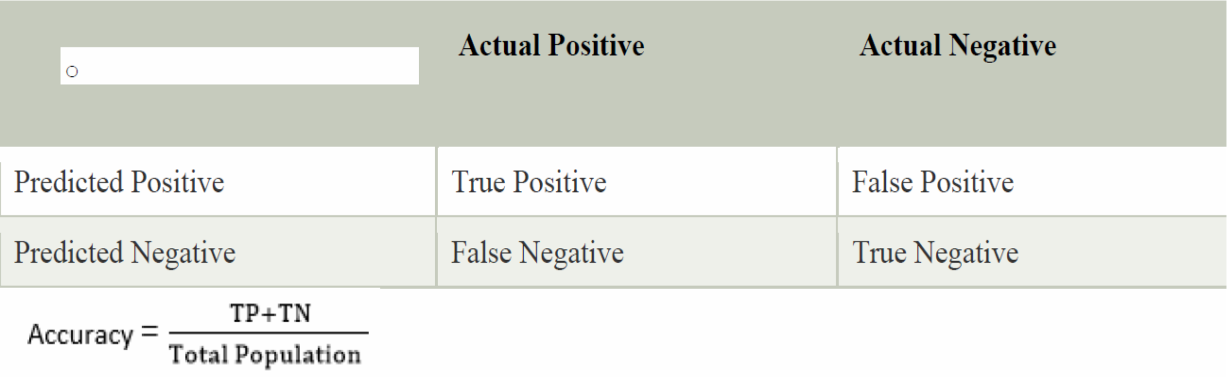
- Model can be trained and tested on different data than the one used for training.
- Response values are known for the test dataset, hence predictions can be evaluated
- Testing accuracy is a better estimate than training accuracy of out-of-sample performance.

Confusion Matrix:

o The confusion matrix provides us a matrix/table as output and describes the performance of the model.

o It is also known as the error matrix.

o The matrix consists of predictions result in a summarized form, which has a total number of correct predictions and incorrect predictions. The matrix looks like as below table:



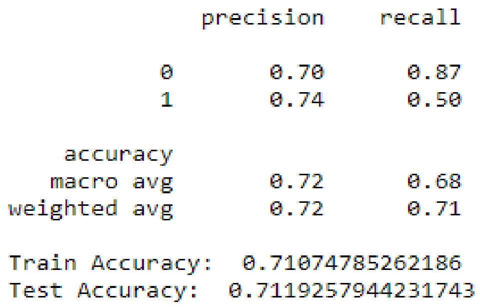
○	**Actual Positive**	**Actual Negative**
Predicted Positive	True Positive	False Positive
Predicted Negative	False Negative	True Negative

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{Total Population}}$$

8.Model Testing

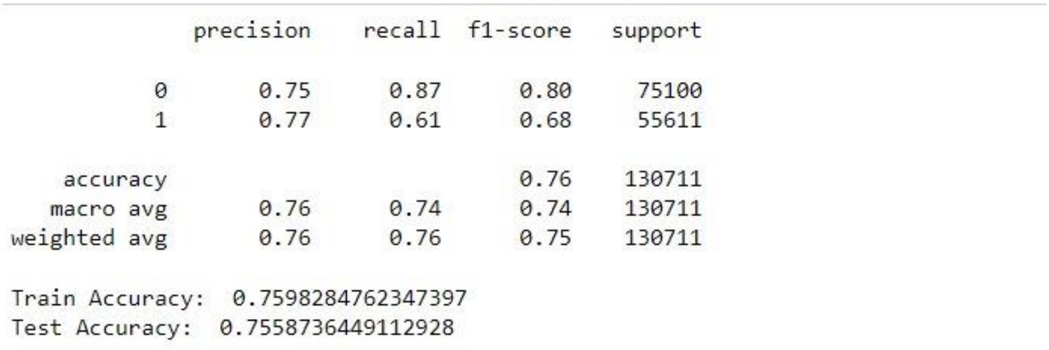
Machine learning consists of algorithms that can automate analytical model building. Using algorithms that iteratively learn from data, machine learning models facilitate computers to find hidden insights from Big Data without being explicitly programmed where to look.
We have used the following algorithms to build predictive model.

Decision tree: The information gained in the decision tree can be defined as the amount of information improved in the nodes before splitting them for making further decisions.



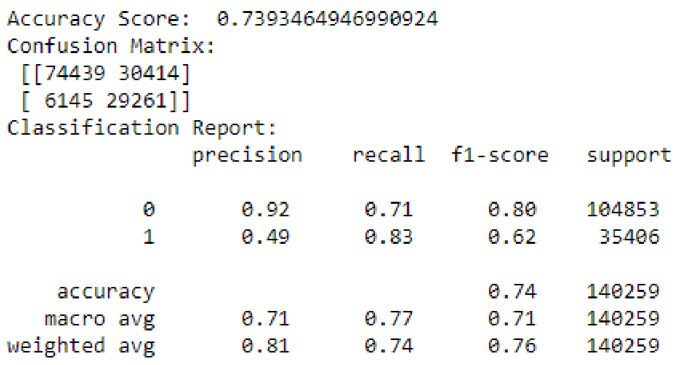
```
              precision    recall  f1-score   support

           0       0.70      0.87      0.78     80584
           1       0.74      0.50      0.60     59675

    accuracy                           0.71    140259
   macro avg       0.72      0.68      0.69    140259
weighted avg       0.72      0.71      0.70    140259

Train Accuracy:  0.71074785262186
Test Accuracy:  0.7119257944231743
```

If we increase max depth at particular level then we get more improved accuracy as below.



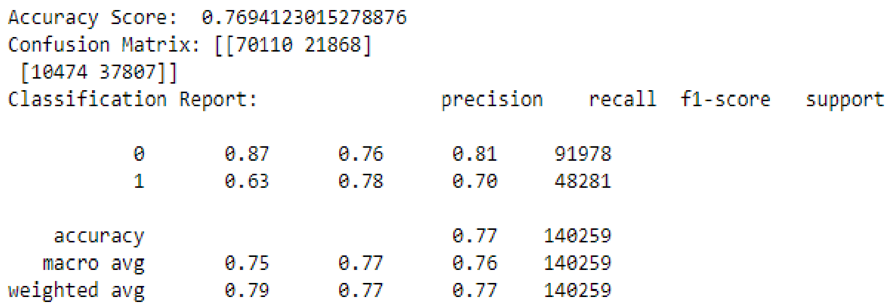
```
              precision    recall  f1-score   support

           0       0.75      0.87      0.80     75100
           1       0.77      0.61      0.68     55611

    accuracy                           0.76    130711
   macro avg       0.76      0.74      0.74    130711
weighted avg       0.76      0.76      0.75    130711

Train Accuracy:  0.7598284762347397
Test Accuracy:  0.7558736449112928
```

Random Forest: Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.



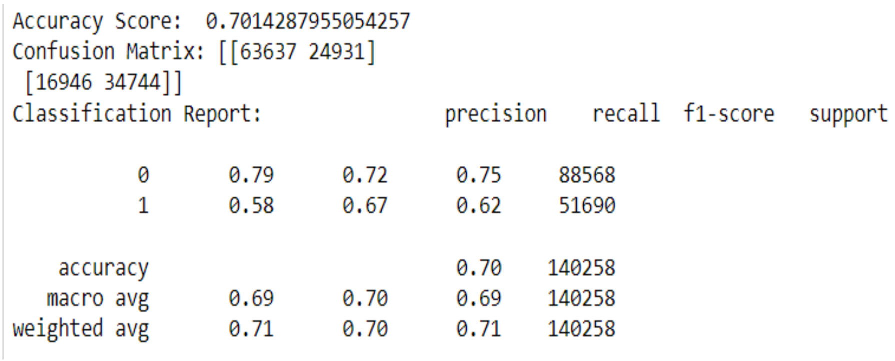
```
Accuracy Score:  0.7393464946990924
Confusion Matrix:
 [[74439 30414]
 [ 6145 29261]]
Classification Report:
               precision    recall  f1-score   support

           0       0.92      0.71      0.80    104853
           1       0.49      0.83      0.62     35406

    accuracy                           0.74    140259
   macro avg       0.71      0.77      0.71    140259
weighted avg       0.81      0.74      0.76    140259
```

XGboost Model: Extreme Gradient Boosting (XGBoost) is an open-source library that provides an efficient and effective implementation of the gradient boosting algorithm. **Gradient boosting** refers to a class of ensemble machine learning algorithms. Models are fit using any arbitrary differentiable loss function and gradient descent optimization algorithm. This gives the technique its name, "*gradient boosting*," as the loss
gradient is minimized.



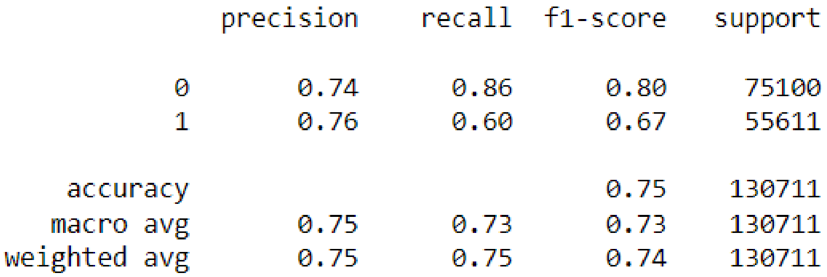
```
Accuracy Score:  0.7694123015278876
Confusion Matrix: [[70110 21868]
 [10474 37807]]
Classification Report:                precision    recall  f1-score   support

           0       0.87      0.76      0.81     91978
           1       0.63      0.78      0.70     48281

    accuracy                           0.77    140259
   macro avg       0.75      0.77      0.76    140259
weighted avg       0.79      0.77      0.77    140259
```

Logistic Regression: Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set.



```
Accuracy Score:  0.7014287955054257
Confusion Matrix: [[63637 24931]
 [16946 34744]]
Classification Report:                precision    recall  f1-score   support

           0       0.79      0.72      0.75     88568
           1       0.58      0.67      0.62     51690

    accuracy                           0.70    140258
   macro avg       0.69      0.70      0.69    140258
weighted avg       0.71      0.70      0.71    140258
```

Gradient Boosting Classifier: Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model**.** Decision trees are usually used when doing gradient boosting.



```
              precision    recall  f1-score   support

           0       0.74      0.86      0.80     75100
           1       0.76      0.60      0.67     55611

    accuracy                           0.75    130711
   macro avg       0.75      0.73      0.73    130711
weighted avg       0.75      0.75      0.74    130711
```

Overall Conclusion:

The accuracy of XGBOOST is highest among all the algorithms used to predict employee attrition so we can used XGBOOST algorithm for future prediction.

9. Comparing Models

ROC AND AUC:

An ROC curve (or receiver operating characteristic curve) is a plot that summarizes the performance of a binary classification model on the positive class.

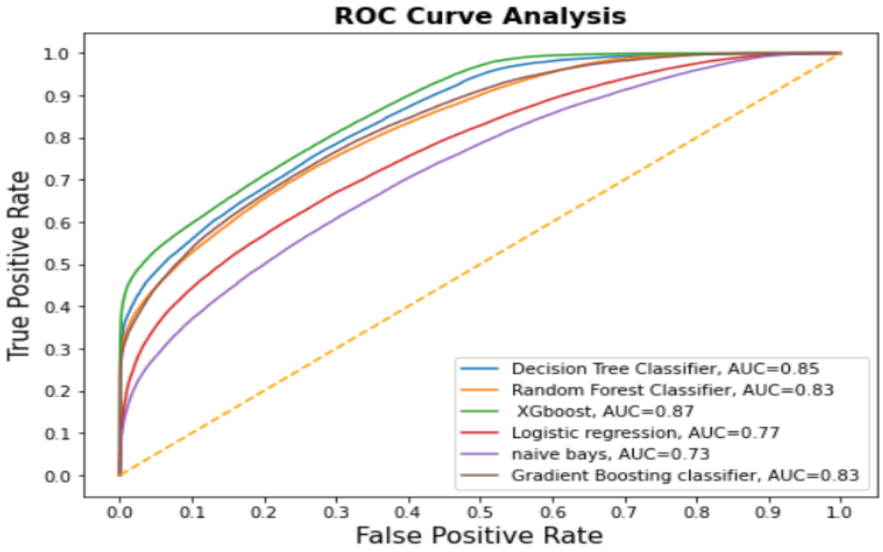
The x-axis indicates the False Positive Rate and the y-axis indicates the True Positive Rate.

ROC Curve: Plot of False Positive Rate (x) vs. True Positive Rate (y).
The true positive rate is a fraction calculated as the total number of true positive predictions divided by the sum of the true positives and the false negatives (e.g. all examples in the positive class). The true positive rate is referred to as the sensitivity or the recall.

TruePositiveRate = TruePositives / (TruePositives + False Negatives)
The false positive rate is calculated as the total number of false positive predictions divided by the sum of the false positives and true negatives (e.g. all examples in the negative class).

FalsePositiveRate = FalsePositives / (FalsePositives + TrueNegatives)

Multiple ROC-Curves in a single plot:



Conclusion: From above figures we can see that XGboost model has the highest AUC Which is 0.87, So we will use XGboost model for employee attrition datasets.
Other models also perform well as there AUC are:
Logistic regression = 0.77
Decision tree classifier = 0.85
Logistic regression = 0.77
Naive bayes = 0.73
Gradient boosting classifier = 0.83

10. Conclusion

In our project We developed an efficient and effective approach to analyze Human Resource data, specifically to disclose hidden relationships in our data by drawing behaviors of employee attrition from numerous amounts of features available from the data. We built machine learning models to accurately separate attrition group from no-attrition group and it can be used to predict of any employee who will leave or stay in the company given similar data.

11. Future Scope

This is a complete flow for machine learning and there is still much of work to do. We will need to spend more time on tuning the model so that we will have a more robust model that could be deployed in the future. If possible, we could collect more data and more features so that the model could learn more from data and make more accurate predictions.

12.Reference

Dataset link:

https://excelbianalytics.com/wp/downloads-21-sample-csv-files-data-sets-for-testing-5-million-records-hr-analytics-for-attrition till- /

Models:

1. Decision tree:

   https://scikit-learn.org/stable/modules/tree.html

2. Random Forest Regression:

   https://scikit-learn.org/stable/modules/ensemble.html#forests-of-randomizedtrees

3. Naive Bays:

   https://scikit-learn.org/stable/modules/naive_bayes.html

4. XGBoost regression:

   https://xgboost.readthedocs.io/en/latest/index.html

5. Gradient Boosting Classifier:

   https://towardsdatascience.com/gradient-boosting-classification-explained-through-python-60cc980eeb3d

6. Logistic Regression:

https://scikitlearn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html