PROJECT REPORT

CORPORATE EMPLOYEE ATTRITION ANALYTICS

Submitted by

TEAM ID: PNT2022TMID21292

TEAM LEADER: SHRUTHEE B

TEAM MEMBER 1: SAKTHIDEVI A S

TEAM MEMBER 2: SHALINI P

TEAM MEMBER 3: STEPHY M

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE AND ENGINEERING

THIAGARAJAR COLLEGE OF ENGINEERING,

THIRUPARANKUNDRAM, MADURAI.

TABLE OF CONTENTS

1 INTRODUCTION

1.1 Project Overview

Employee attrition has become a vital problem across the world. It is one of the crucial issues faced by business leaders with in companies where they lose the most talented employees. A good employee is always an asset to the organization and their resignation can lead to various problems like financial losses, overall performance, and loss of acquired knowledge. Furthermore, hiring new employees is far exorbitant, taxing, and time-consuming in comparison to recruiting the existing one. It is very time-consuming to recruit a new employee as it takes him months for training, adjusting to the culture, rules, and environment. Therefore, upcoming trends and technology using Machine Learning Algorithms must be exploited for the benefit of business organizations knowing there as on beforehand forth employee attrition, companies can mitigate this loss. This analysis provides a conclusive review of employee attrition from the data set IBM HR Analytics Employee Attrition Performance.

1.2 Purpose

Hardik P. K., researched on “a study on employee attrition: with special reference to Kerala IT Industry”. His research examined the relationship between organizational factors and attrition of IT professional’s. The result can conclude that the organizational factors played significant role in predicting the variance in turnover intention (attrition) of Kerala IT professionals. Therefore, the HR managers in IT organizations may take into consideration the problems with organizational factors of their workers to reduce the turnover intention of the skilled employees.

2 LITERATURE SURVEY

2.1 Existing Problem

The Existing system includes only few attributes for analysis and also deals with qualitative observations and simple statistical analysis. The qualitative observations deal with data and can be observed through human senses. They do not involve measurements or number. Due to the increase in IOT and connected device, we now have access to so much of data and along with it an increase in needs to manage and understand data.

2.2 References

2.2.1 From Big Data to Deep Data to support people analytics for employee attrition prediction, Nesrine Ben Yahia, HlelJihen, Ricardo Colomo-Palacio(2021)

2.2.2 Machine Learning Approach for Employee Attrition Analysis. Dr. R. S. Kamath | Dr. S. S. Jamsandekar | Dr. P. G. Naik , Published in International Journal of Trend in Scientific Research and Development(ijtsrd),(March2019)

2.2.3 Investigation of early career teacher attrition(ECT) and the impact of induction programs in Western Australia, Janine E.Wyatt, MichaelO'Neill (2021)

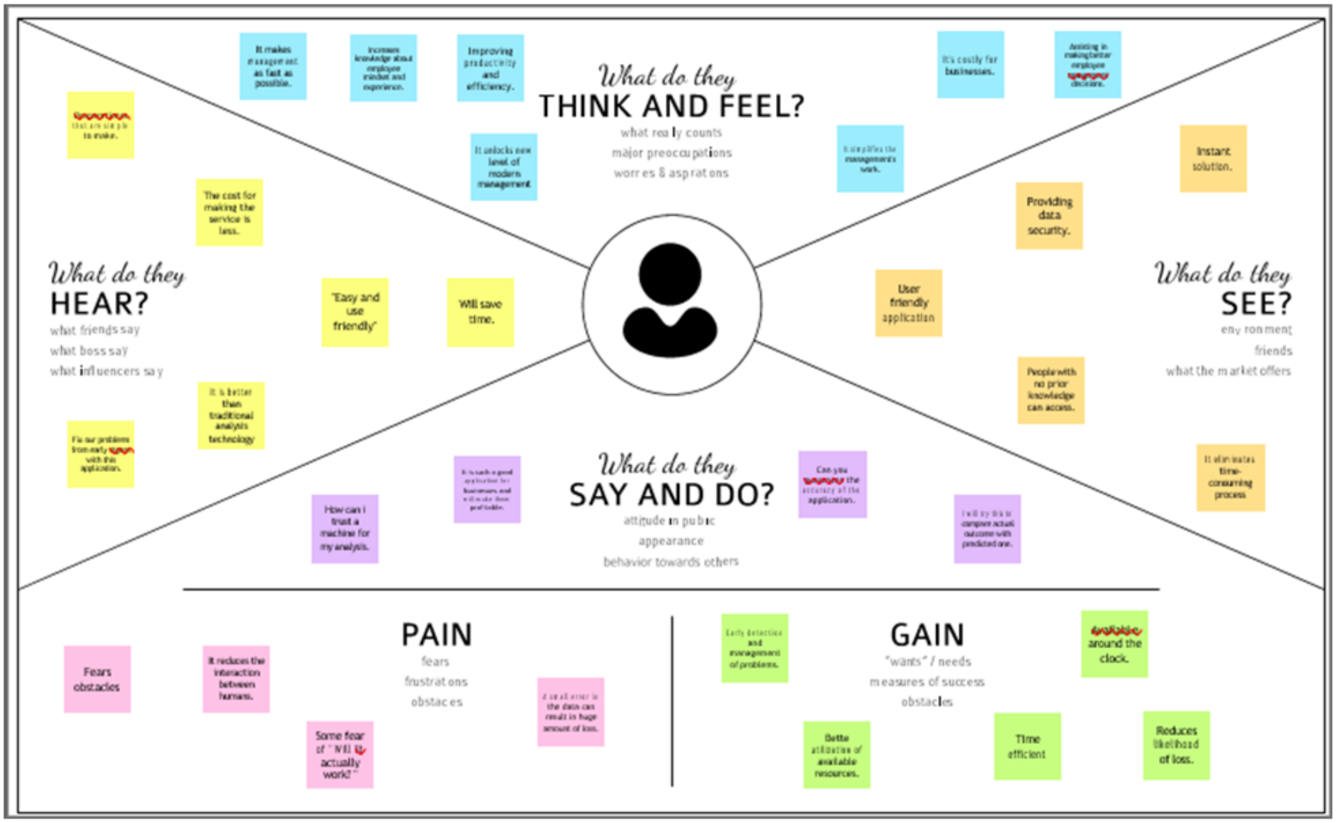
2.3 Problem Statement Definition

- To create a dashboard and perform analysis of employee attrition in corporate using IBM Cognos analytics platform.

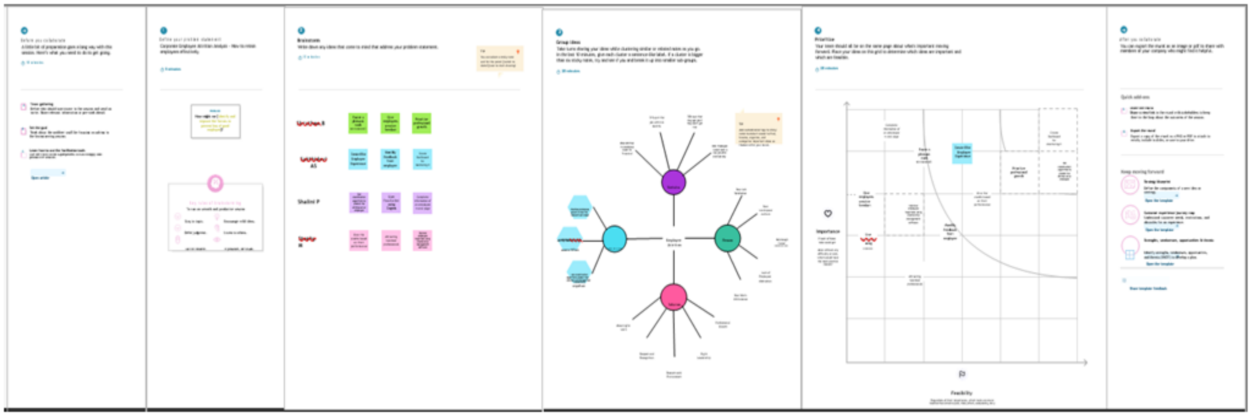
- To reduce the employee attrition rate through data analytics, data visualization by analyzing the major factors that causes attrition.

3 IDEATION & PROPOSED SOLUTION

3.1 Empathy Map Canvas


What do they
THINK AND FEEL?
what really counts
major preoccupations
worries & aspirations
It makes management as fast as possible.
Increases knowledge about employee mindset and experience.
Improving productivity and efficiency.
It unlocks new level of modern management
It's costly for businesses.
Assisting in making better employee decisions.
It simplifies the management's work.
What do they
HEAR?
what friends say
what boss say
what influencers say
that are simple to make.
The cost for making the service is less.
"Easy and use friendly"
Will save time.
It is better than traditional analysis technology
Fix our problems from early with this application.
What do they
SEE?
environment
friends
what the market offers
Instant solution.
Providing data security.
User friendly application
People with no prior knowledge can access.
It eliminates time-consuming process
What do they
SAY AND DO?
attitude in public
appearance
behavior towards others
How can I trust a machine for my analysis.
It is such a good application for businesses and will make them profitable.
Can you the accuracy of the application.
I will try this to compare actual outcome with predicted one.
PAIN
fears
frustrations
obstacles
Fears obstacles
It reduces the interaction between humans.
Some fear of " Will it actually work? "
A small error in the data can result in huge amount of loss.
GAIN
"wants" / needs
measures of success
obstacles
Early detection and management of problems.
Available around the clock.
Bette utilization of available resources.
Time efficient
Reduces likelihood of loss.

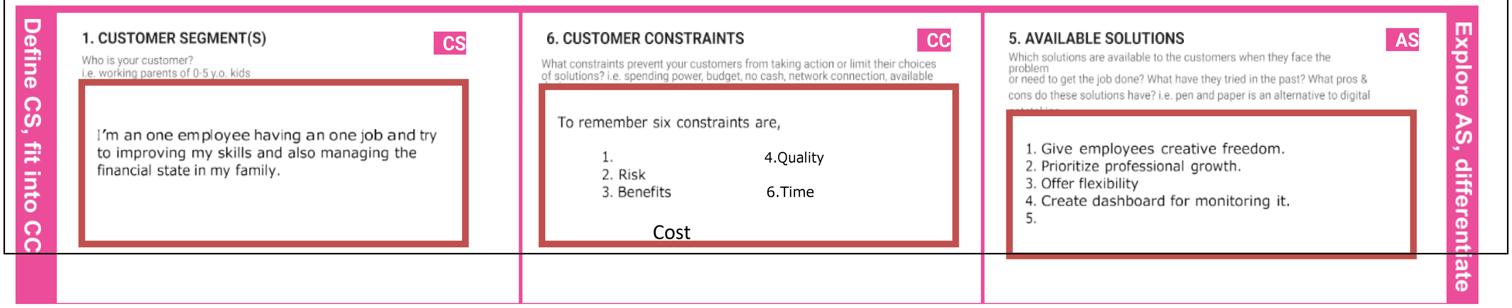
3.2 Ideation & Brainstorming



3.3 Proposed Solution

S No	Parameter	Description
1	Problem Statement	Corporate Employee Attrition Analysis - How to retain employees effectively
2	Idea / Solution description	Prioritize the professional growth & give the pleasant workspace and use some classification algorithm to predict their retention and manage their relationship using this software.
3	Novelty / Uniqueness	Employee attrition prediction is specifically focused on identifying why employees voluntarily leave, what might have prevented them from leaving, and how we can use data to predict attrition risk.
4	Social Impact / Customer Satisfaction	Employee's attrition has huge impact on company, recruiting new employees and investing time to train them is increased. Losing a good employee creates a negative impact of profit on the company.
5	Business Model (Revenue Model)	The business is struggling with employee attrition. This software will be helpful to analyze the workforce trends and find the root cause of Attrition.
6	Scalability of the Solution	The dashboard is scalable for the companies when their employee's dataset is used for analysis. The model can successfully predict the futuristic approach and suggests preventive measures.

3.4 Problem Solution fit



1. CUSTOMER SEGMENT(S)

Who is your customer?
i.e. working parents of 0-5 y.o. kids

I'm an one employee having an one job and try to improving my skills and also managing the financial state in my family.

6. CUSTOMER CONSTRAINTS

What constraints prevent your customers from taking action or limit their choices of solutions? i.e. spending power, budget, no cash, network connection, available

To remember six constraints are,

1.
2. Risk
3. Benefits

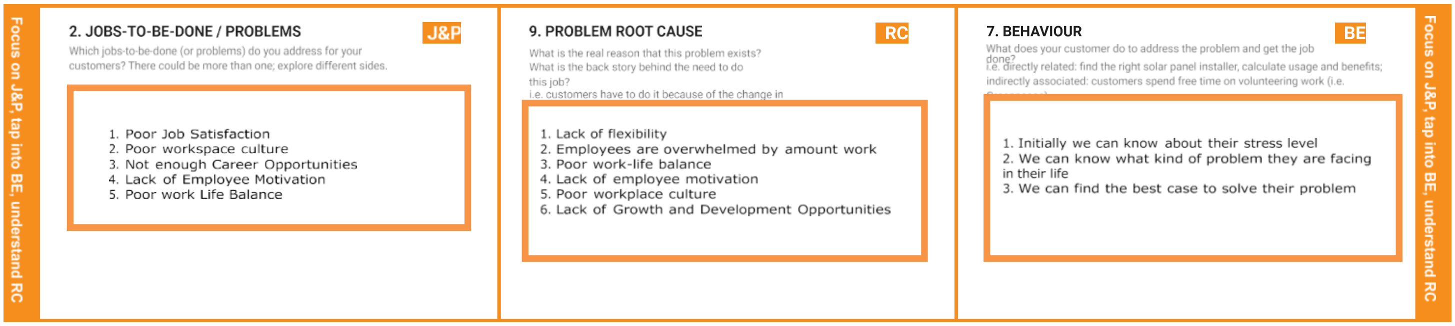
4.Quality
6.Time

Cost

5. AVAILABLE SOLUTIONS

Which solutions are available to the customers when they face the problem
or need to get the job done? What have they tried in the past? What pros & cons do these solutions have? i.e. pen and paper is an alternative to digital

1. Give employees creative freedom.
2. Prioritize professional growth.
3. Offer flexibility
4. Create dashboard for monitoring it.
5.



Focus on J&P, tap into BE, understand RC

2. JOBS-TO-BE-DONE / PROBLEMS

J&P

Which jobs-to-be-done (or problems) do you address for your customers? There could be more than one; explore different sides.

1. Poor Job Satisfaction
2. Poor workspace culture
3. Not enough Career Opportunities
4. Lack of Employee Motivation
5. Poor work Life Balance

9. PROBLEM ROOT CAUSE

RC

What is the real reason that this problem exists? What is the back story behind the need to do this job?
i.e. customers have to do it because of the change in

1. Lack of flexibility
2. Employees are overwhelmed by amount work
3. Poor work-life balance
4. Lack of employee motivation
5. Poor workplace culture
6. Lack of Growth and Development Opportunities

7. BEHAVIOUR

BE

What does your customer do to address the problem and get the job done?
i.e. directly related: find the right solar panel installer, calculate usage and benefits; indirectly associated: customers spend free time on volunteering work (i.e.

1. Initially we can know about their stress level
2. We can know what kind of problem they are facing in their life
3. We can find the best case to solve their problem

Focus on J&P, tap into BE, understand RC

3. TRIGGERS

1. Unhappinessaboutemployeebenefitsorthepaystructure.
2. Lackofemployeedevelopmentopportunities.
3. Evenpoorconditionsintheworkplace.

4. EMOTIONS:BEFORE/ AFTER

Before	After
1. Dissatisfaction	1.Improvingcommunication
2. Disagreement	2.Comfortable
3. Stress	3.Motivation

10. YOURSOLUTION

1. Prioritizeprofessional growth&Givetheplea santworkspace
2. CreateDashboardusingMonthlyFeedbackangiveaccess toHRTeam
3. Use classification algorithm to predict theirrententionandmanagetheirrelationshipusingsoftware

8.CHANNELSOFBEHAVIOUR

Online

In online mode we can use some algorithm anddashboardtopredicttheirattritionandanalysis theirsituation

Offline

Inofflinemodeweconductsomemeetingand gavesome spaceto calmtheirmindto predicttheirattrition

4 REQUIREMENT ANALYSIS

4.1 Functional requirement

FR No.	Functional Requirement(Epic)	Sub Requirement (Story/Sub-Task)
FR-1	User Registration	Registration through Form Registration through Gmail Registration Through LinkedIN
FR-2	User Confirmation	Confirmation via Email and Confirmation Via OTP
FR-3	Account Creation	Create an account in the Profile Dashboard
FR-4	Input Credentials	Uploading your dataset Analyzing the attrition rate using dashboard
FR-5	Processing Methods	Using IBM Cognos Analytics Dashboard Using Prediction algorithm to find attrition rate
FR-6	Output Credentials	Using the Dashboard and Algorithm they know About the employee attrition and way to Reduce the employee attrition

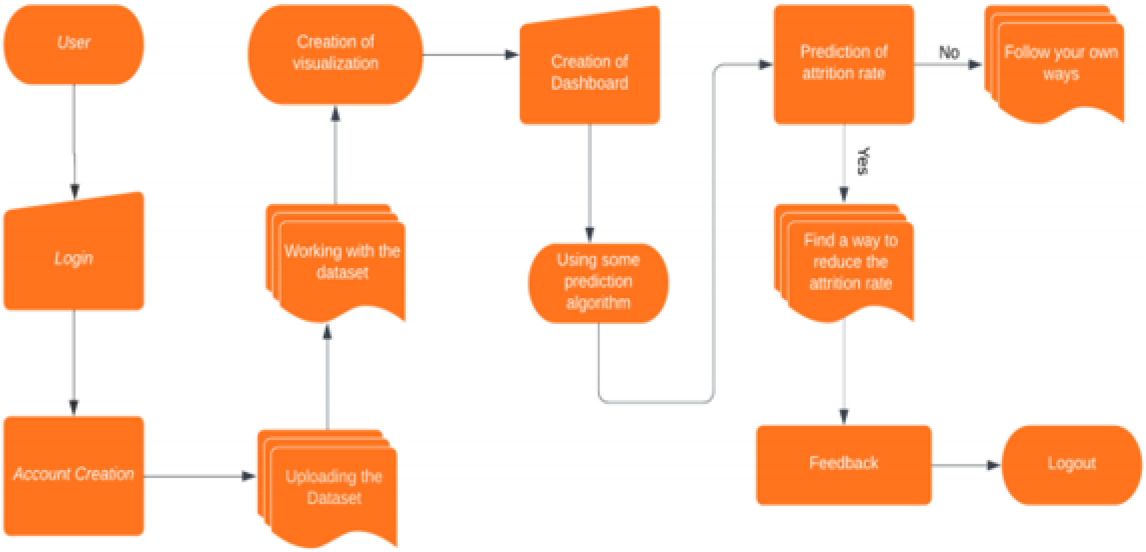
4.2 Non-Functional requirement

FRNo.	Non-Functional Requirement	Description
NFR-1	**Usability**	The user can be able to interact with the system user friendly. The system is build with simple modules and algorithms.
NFR-2	**Security**	Access permissions for the particular system information may only be changed by the system's data administrator. The user's data must behave high security measures.

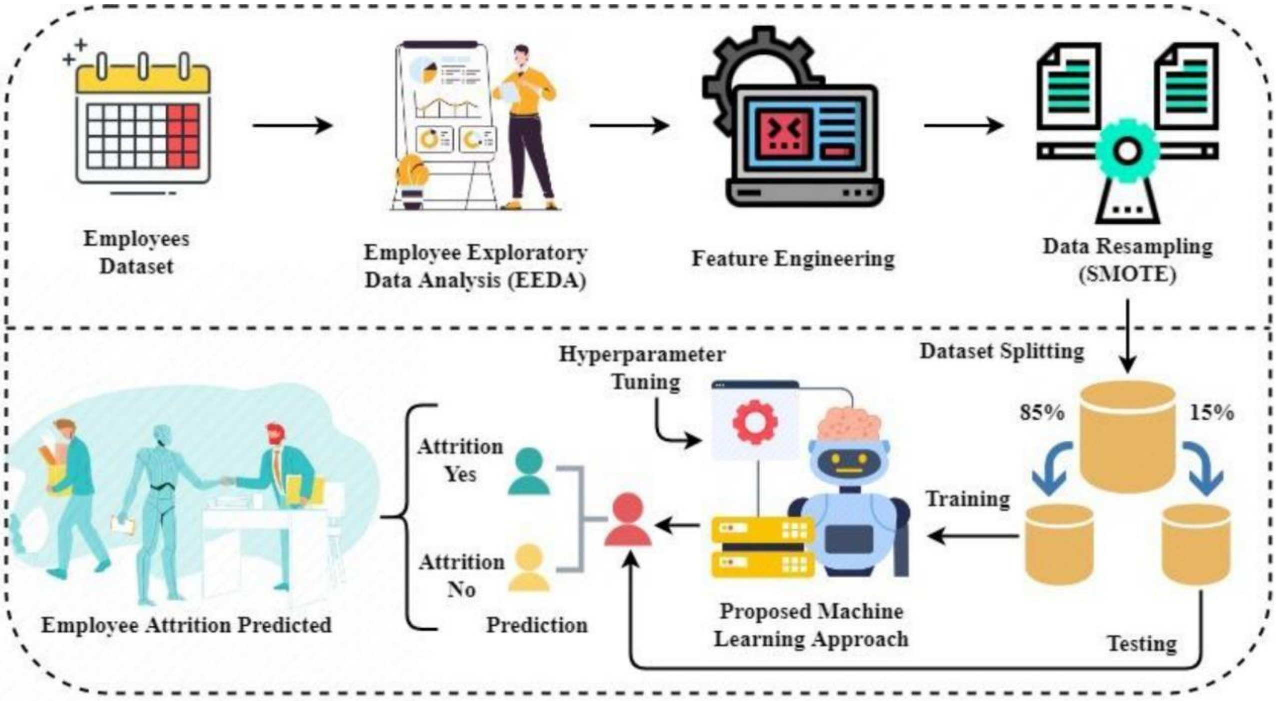
NFR-3	**Reliability**	The database update process must rollback all related updates when any update fails. The dataset will not be modified by anyoneonly the user can be able to modify the dataset.
NFR-4	**Performance**	The performance of the dashboard is flexible to every user’s. The front-page load time must be no more than 2 seconds for users that access the website using an LTE mobile connection.
NFR-5	**Availability**	New module deployment mustn’t impact front page, dashboard and checkout pages availability and mustn’t take longer than one hour. The rest of the pages that may experience problems must display a notification with a timer showing when the system is going to be up again.
NFR-6	**Scalability**	The website attendance limit must be scalable enough to support 200,000 users at a time. The dashboard is scalable for the companies when their employee’s dataset is used for analysis. The model can successfully predict the futuristic approach and suggests preventive measures.

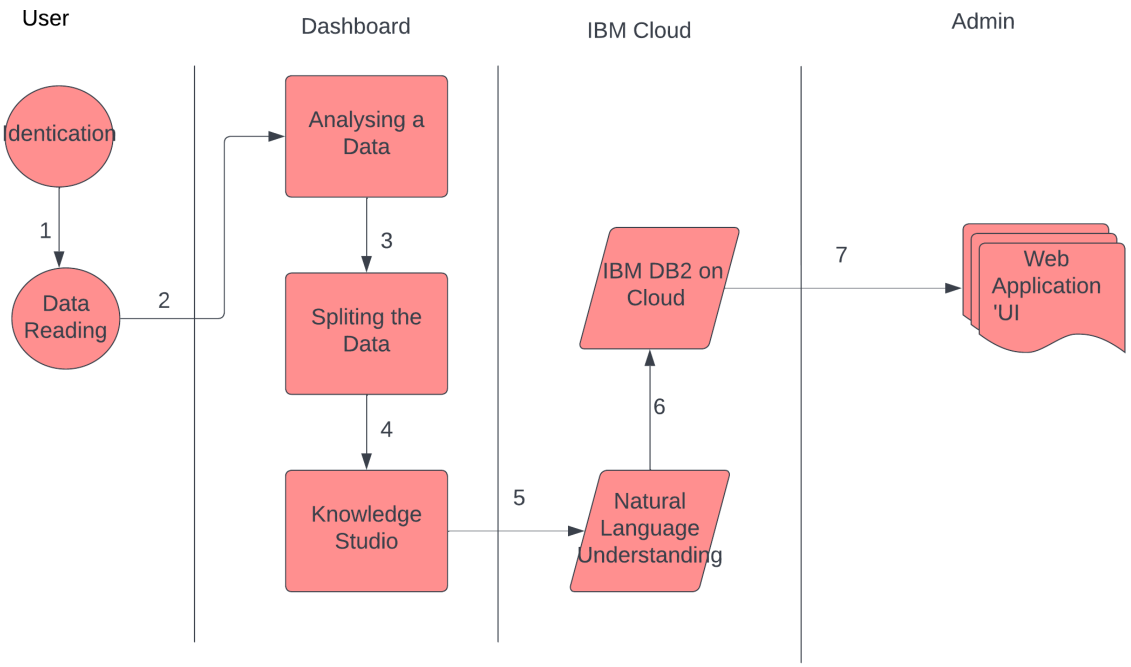
5 PROJECT DESIGN

5.1 Data flow diagrams



5.2 Solution & Technical Architecture





5.3 User Stories

User Type	Functional Requirement (Epic)	User Story Number	User Story /Task	Acceptance criteria	Priority	Release
Customer(Web user)	Registration	USN-1	As a user, I can register for the application by entering my email, password, and confirming my password.	I can access my account /dashboard	High	Sprint-1
		USN-2	As a user, I will receive confirmation email once I have registered for the application	I can receive confirmation email &click confirm	High	Sprint-1
		USN-3	As a user, I can register for the application through Facebook	I can register &access the dashboard with Facebook Login	Low	Sprint-2
		USN-4	As a user, I can register for the application through Gmail	I can register &access the dashboard With Gmail Login	Medium	Sprint-1
	Login	USN-5	As a user, I can log into the application by entering email& password	I can access my account /dashboard	High	Sprint-1

	Dashboard	USN-6	Uploading the Dataset	I can be able to Upload my dataset	High	Sprint2
		USN-7	Working With Dataset	I can be able to Access my dashboard	High	Sprint2
		USN-8	Visualization	I can be able to view the visual attrition rate of My dataset	High	Sprint3
		USN-9	Working with Dashboard	I can be able to view the various views of The attrition rate	High	Sprint3
Customer Care Executive		USN-10	Asking Help/Feedback	I can be able to ask help if I can face any issues or problems while using it.	Medium	Sprint4

6 PROJECT PLANNING & SCHEDULING

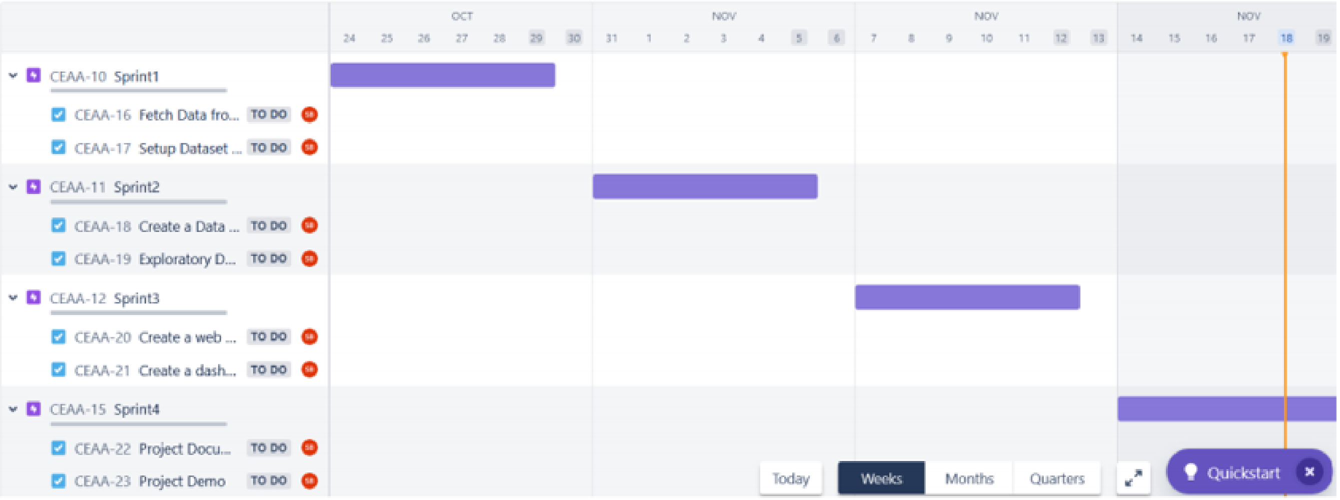
6.1 Sprint Planning & Estimation

Sprint	Functional Requirement(Epic)	User Story Number	UserStory/Task	StoryPoints	Priority	TeamMembers
Sprint-1	Fetch data	USN-1	Fetch data from Kaggle	1	Low	Stephy M
Sprint-1	Dataset setup	USN-2	Setup dataset in IBM cognos analytics	1	Low	Sakthidevi AS
Sprint-2	Data module	USN-3	Create a data module in IBM cognos	2	Medium	Shruthee B
Sprint-2	Data analysis	USN-4	Do exploratory data analysis	2	High	Shalini P
Sprint-3	Web application	USN-5	Create a web application	3	Low	Shruthee B
Sprint-3	Dashboard	USN-6	Create a dashboard in IBM cognos	3	Medium	Shalini P
Sprint-4	Documentation	USN-7	Create documentation	3	High	Sakthidevi AS
Sprint-4	Demo	USN-8	Create a demo video	3	Hifh	Stephy M

6.2 Sprint Delivery Schedule

Sprint	Total Story Points	Duration	Sprint Start Date	Sprint End Date (Planned)	Story Points Completed (as on Planned End Date)	Sprint Release Date (Actual)
Sprint-1	2	6 Days	24 Oct 2022	29 Oct 2022	2	23 Oct 2022
Sprint-2	4	6 Days	31 Oct 2022	05 Nov 2022	4	06 Nov 2022
Sprint-3	6	6 Days	07 Nov 2022	12 Nov 2022	6	13 Nov 2022
Sprint-4	6	6 Days	14 Nov 2022	19 Nov 2022	6	20 Nov 2022

6.3 Reports From JIRA


CEAA-10 Sprint1
- ☑ CEAA-16 Fetch Data fro... TO DO
- ☑ CEAA-17 Setup Dataset ... TO DO

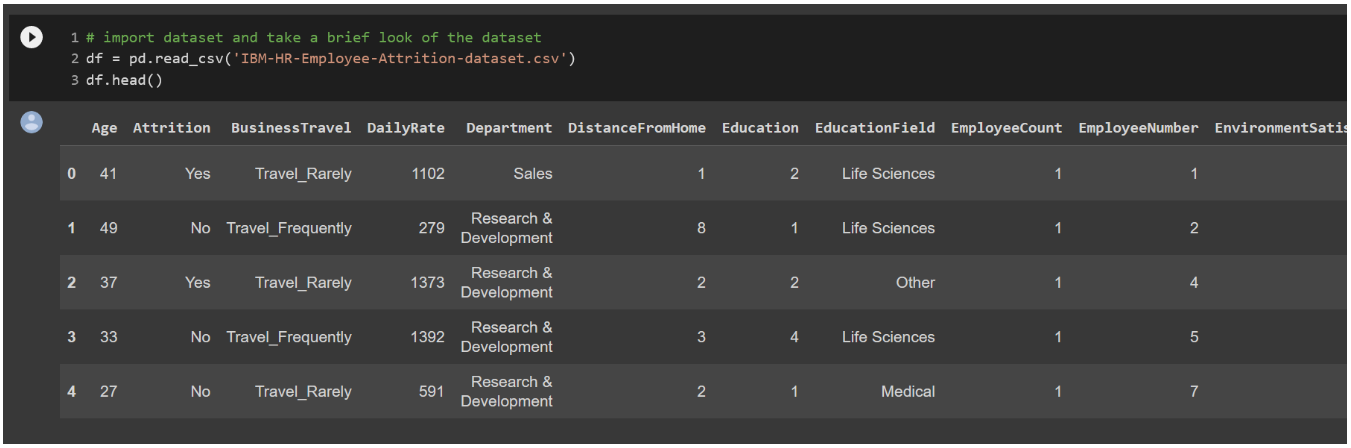
CEAA-11 Sprint2
- ☑ CEAA-18 Create a Data ... TO DO
- ☑ CEAA-19 Exploratory D... TO DO

CEAA-12 Sprint3
- ☑ CEAA-20 Create a web ... TO DO
- ☑ CEAA-21 Create a dash... TO DO

CEAA-15 Sprint4
- ☑ CEAA-22 Project Docu... TO DO
- ☑ CEAA-23 Project Demo TO DO

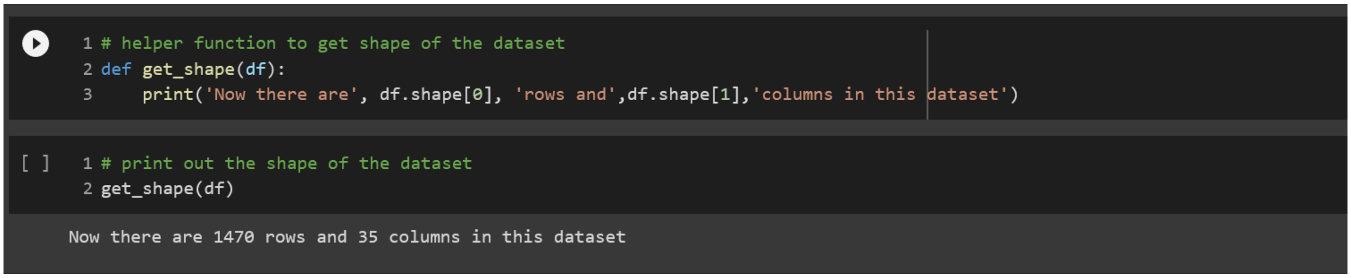
7 CODING & SOLUTIONING

7.1 Feature 1



```
# import dataset and take a brief look of the dataset
df = pd.read_csv('IBM-HR-Employee-Attrition-dataset.csv')
df.head()
```

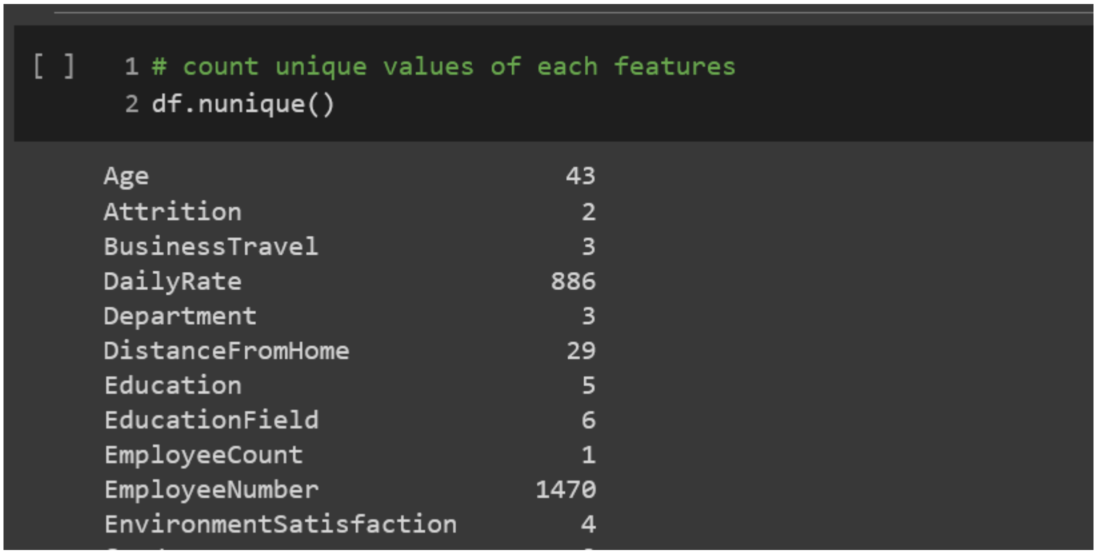
	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	EnvironmentSatis
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	1	1	
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	1	2	
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	1	4	
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	1	5	
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	7	



```
# helper function to get shape of the dataset
def get_shape(df):
    print('Now there are', df.shape[0], 'rows and',df.shape[1],'columns in this dataset')
```

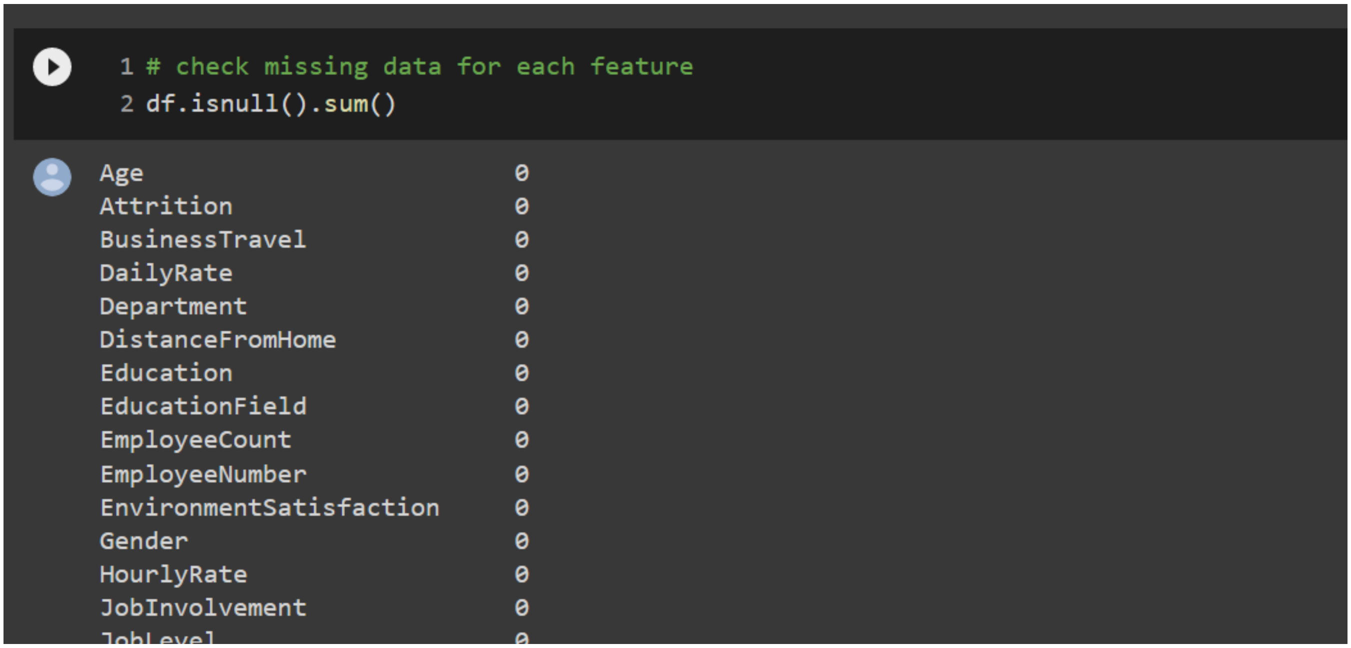
```
[ ]
# print out the shape of the dataset
get_shape(df)
```

```
Now there are 1470 rows and 35 columns in this dataset
```



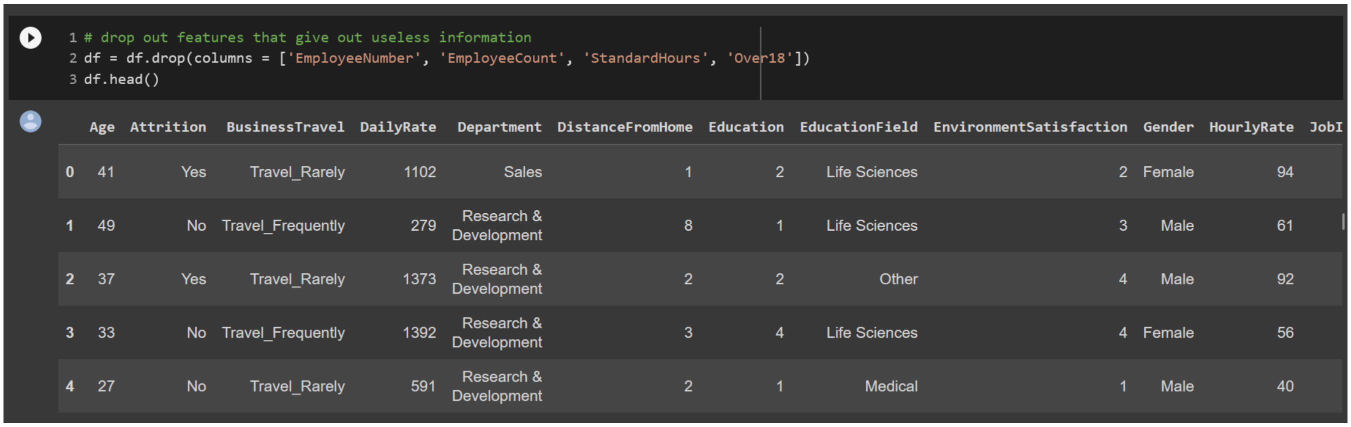
```
[ ]
# count unique values of each features
df.nunique()
```

```
Age                          43
Attrition                     2
BusinessTravel                3
DailyRate                   886
Department                    3
DistanceFromHome             29
Education                     5
EducationField                6
EmployeeCount                 1
EmployeeNumber             1470
EnvironmentSatisfaction       4
```



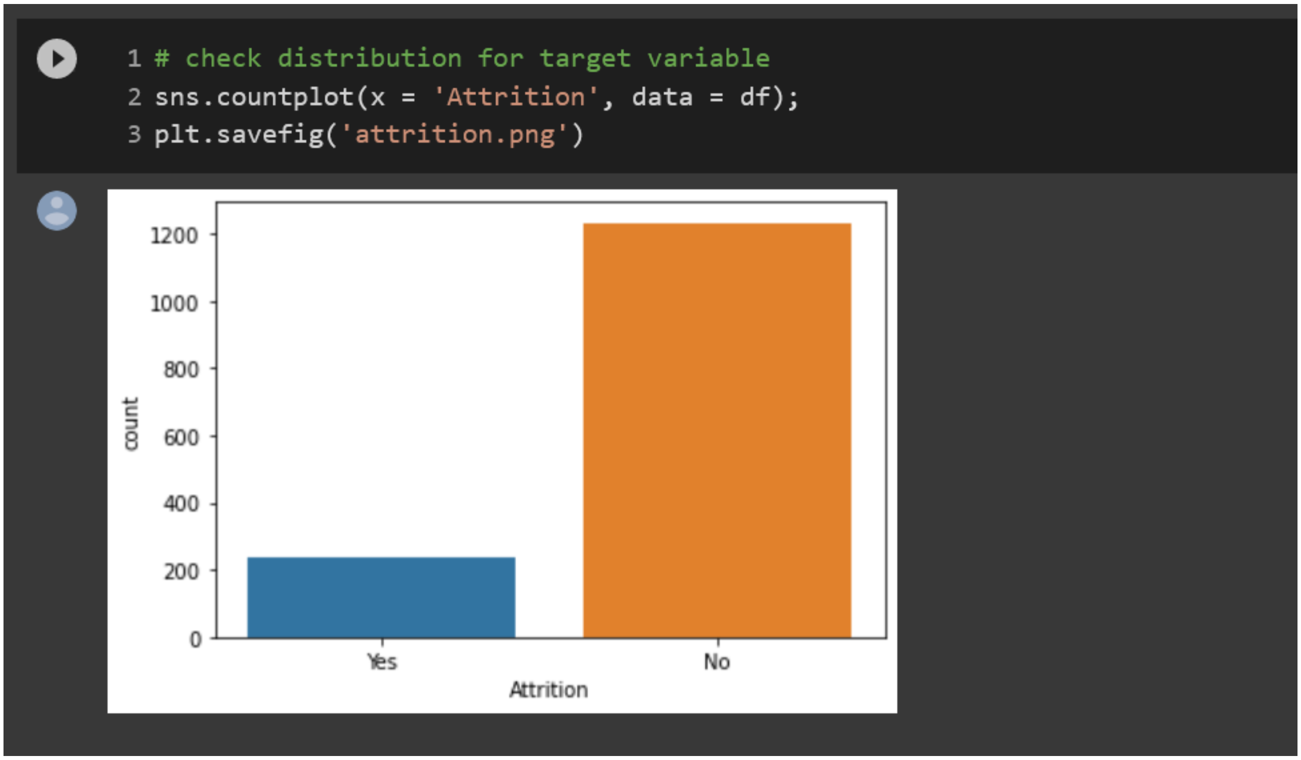
```
# check missing data for each feature
df.isnull().sum()
```

```
Age                         0
Attrition                   0
BusinessTravel              0
DailyRate                   0
Department                  0
DistanceFromHome            0
Education                   0
EducationField              0
EmployeeCount               0
EmployeeNumber              0
EnvironmentSatisfaction     0
Gender                      0
HourlyRate                  0
JobInvolvement              0
JobLevel                    0
```

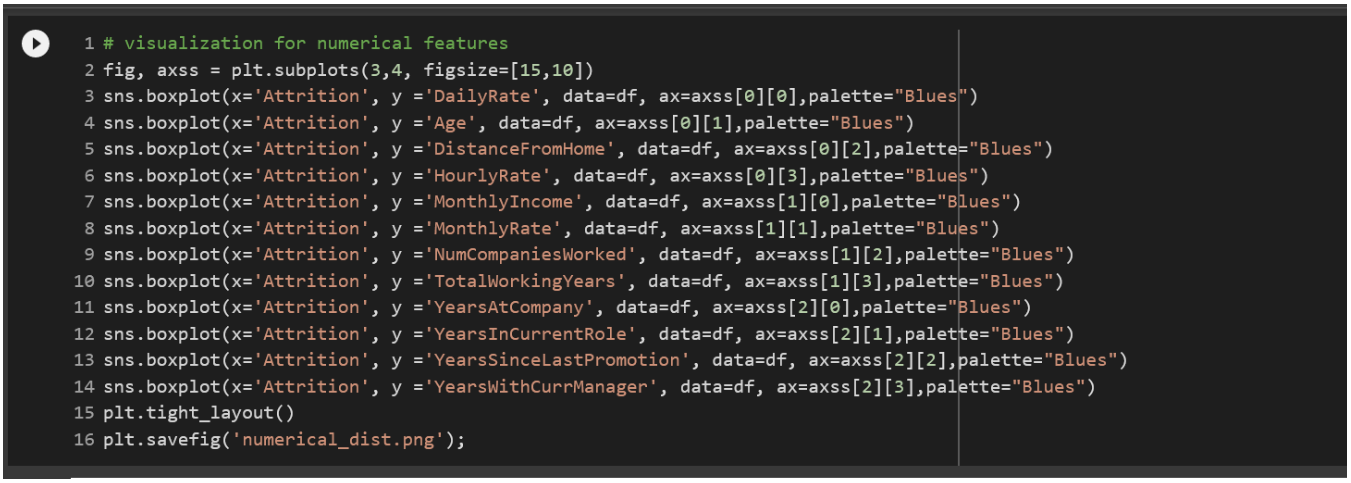


```
# drop out features that give out useless information
df = df.drop(columns = ['EmployeeNumber', 'EmployeeCount', 'StandardHours', 'Over18'])
df.head()
```

	Age	Attrition	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EnvironmentSatisfaction	Gender	HourlyRate	JobI
0	41	Yes	Travel_Rarely	1102	Sales	1	2	Life Sciences	2	Female	94	
1	49	No	Travel_Frequently	279	Research & Development	8	1	Life Sciences	3	Male	61	
2	37	Yes	Travel_Rarely	1373	Research & Development	2	2	Other	4	Male	92	
3	33	No	Travel_Frequently	1392	Research & Development	3	4	Life Sciences	4	Female	56	
4	27	No	Travel_Rarely	591	Research & Development	2	1	Medical	1	Male	40	

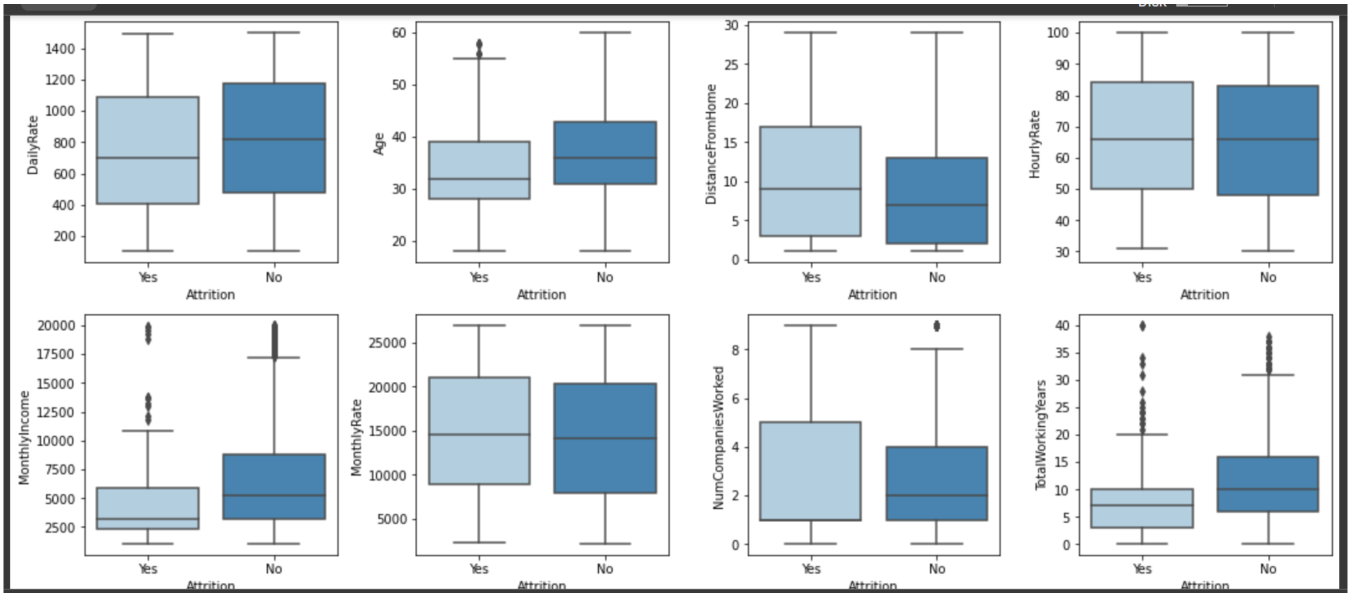


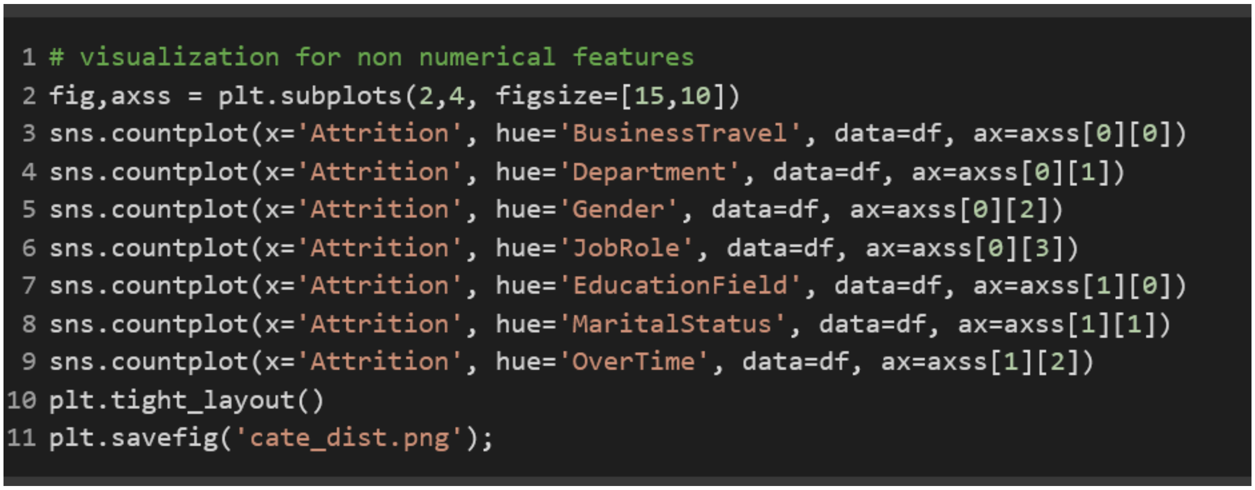
```python
# check distribution for target variable
sns.countplot(x = 'Attrition', data = df);
plt.savefig('attrition.png')
```



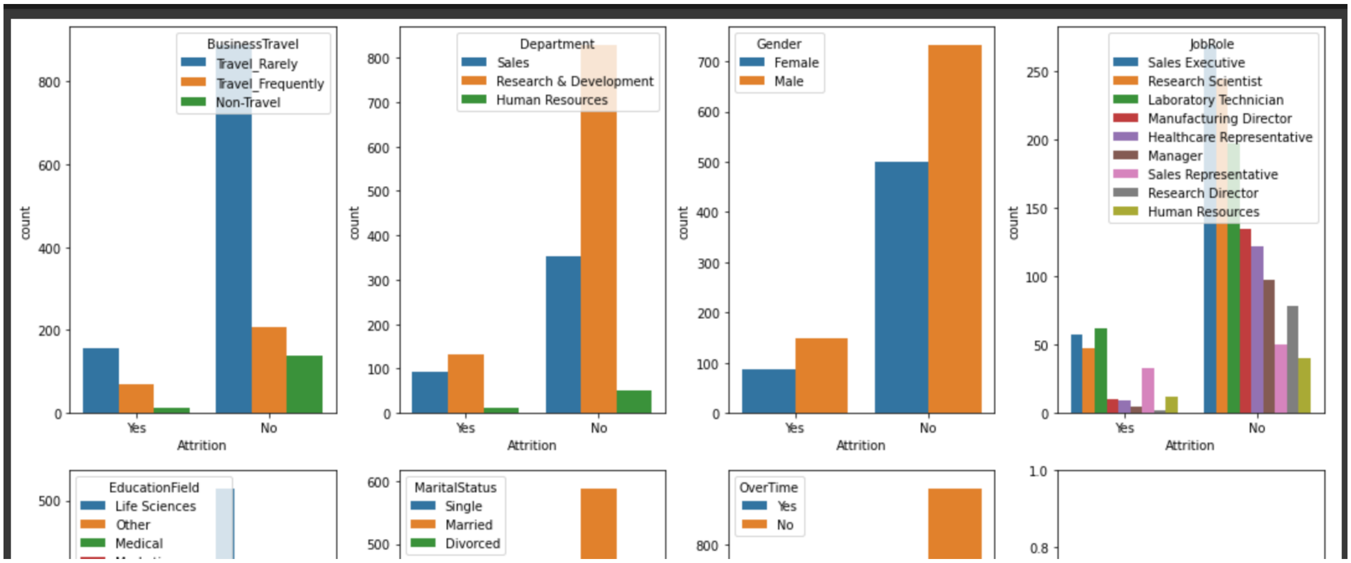
```
# visualization for numerical features
fig, axss = plt.subplots(3,4, figsize=[15,10])
sns.boxplot(x='Attrition', y ='DailyRate', data=df, ax=axss[0][0],palette="Blues")
sns.boxplot(x='Attrition', y ='Age', data=df, ax=axss[0][1],palette="Blues")
sns.boxplot(x='Attrition', y ='DistanceFromHome', data=df, ax=axss[0][2],palette="Blues")
sns.boxplot(x='Attrition', y ='HourlyRate', data=df, ax=axss[0][3],palette="Blues")
sns.boxplot(x='Attrition', y ='MonthlyIncome', data=df, ax=axss[1][0],palette="Blues")
sns.boxplot(x='Attrition', y ='MonthlyRate', data=df, ax=axss[1][1],palette="Blues")
sns.boxplot(x='Attrition', y ='NumCompaniesWorked', data=df, ax=axss[1][2],palette="Blues")
sns.boxplot(x='Attrition', y ='TotalWorkingYears', data=df, ax=axss[1][3],palette="Blues")
sns.boxplot(x='Attrition', y ='YearsAtCompany', data=df, ax=axss[2][0],palette="Blues")
sns.boxplot(x='Attrition', y ='YearsInCurrentRole', data=df, ax=axss[2][1],palette="Blues")
sns.boxplot(x='Attrition', y ='YearsSinceLastPromotion', data=df, ax=axss[2][2],palette="Blues")
sns.boxplot(x='Attrition', y ='YearsWithCurrManager', data=df, ax=axss[2][3],palette="Blues")
plt.tight_layout()
plt.savefig('numerical_dist.png');
```

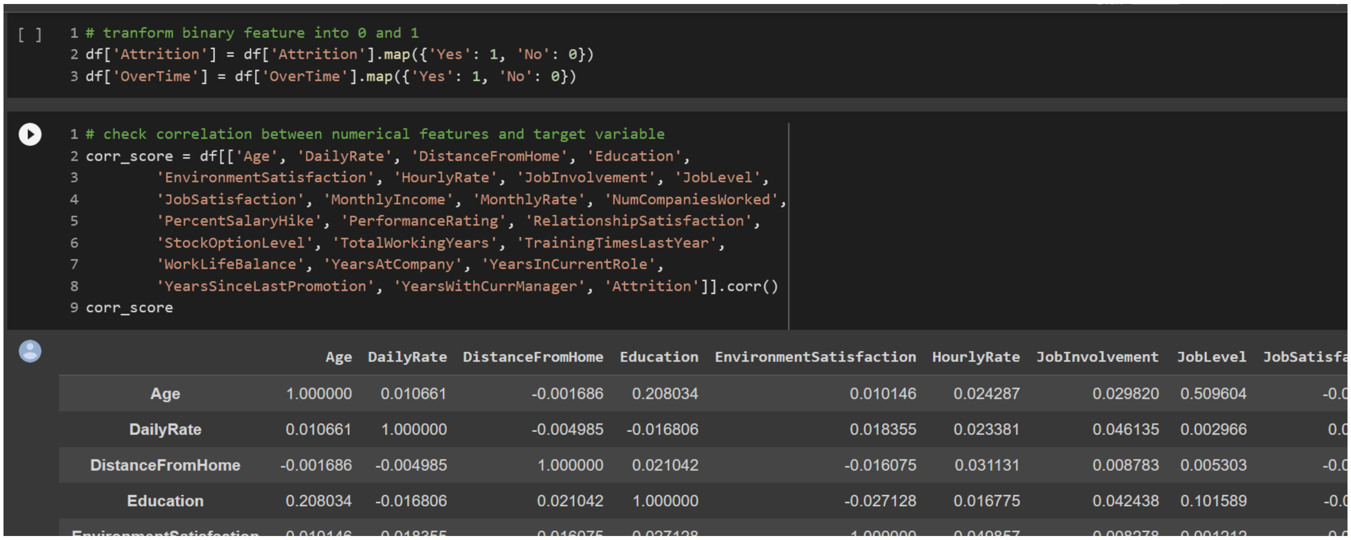
,





```
# visualization for non numerical features
fig,axss = plt.subplots(2,4, figsize=[15,10])
sns.countplot(x='Attrition', hue='BusinessTravel', data=df, ax=axss[0][0])
sns.countplot(x='Attrition', hue='Department', data=df, ax=axss[0][1])
sns.countplot(x='Attrition', hue='Gender', data=df, ax=axss[0][2])
sns.countplot(x='Attrition', hue='JobRole', data=df, ax=axss[0][3])
sns.countplot(x='Attrition', hue='EducationField', data=df, ax=axss[1][0])
sns.countplot(x='Attrition', hue='MaritalStatus', data=df, ax=axss[1][1])
sns.countplot(x='Attrition', hue='OverTime', data=df, ax=axss[1][2])
plt.tight_layout()
plt.savefig('cate_dist.png');
```

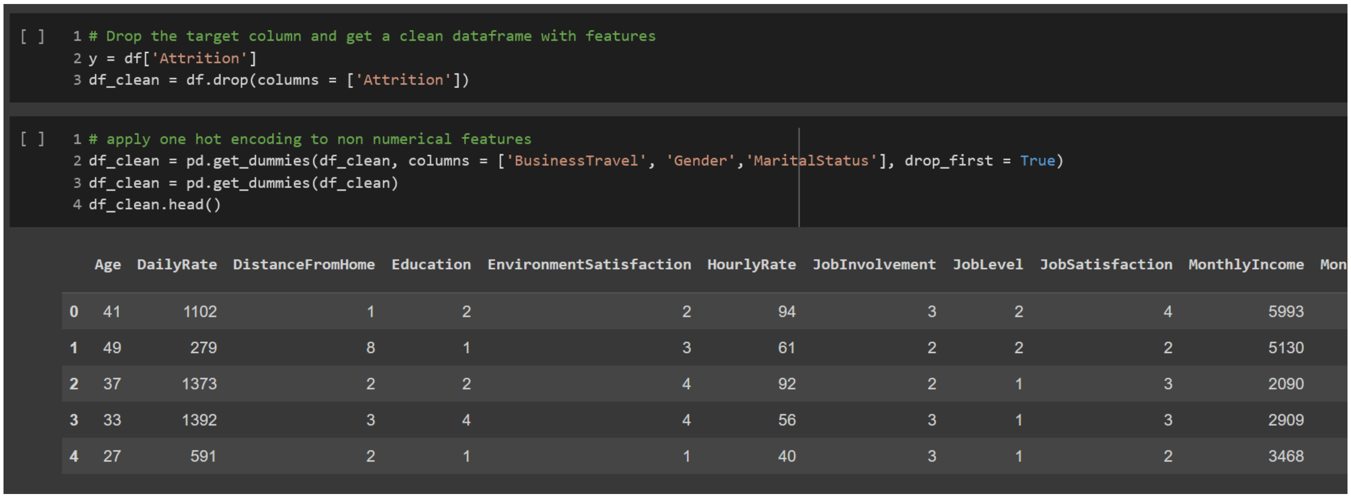




```
[ ]  # tranform binary feature into 0 and 1
     df['Attrition'] = df['Attrition'].map({'Yes': 1, 'No': 0})
     df['OverTime'] = df['OverTime'].map({'Yes': 1, 'No': 0})
```

```
     # check correlation between numerical features and target variable
     corr_score = df[['Age', 'DailyRate', 'DistanceFromHome', 'Education',
             'EnvironmentSatisfaction', 'HourlyRate', 'JobInvolvement', 'JobLevel',
             'JobSatisfaction', 'MonthlyIncome', 'MonthlyRate', 'NumCompaniesWorked',
             'PercentSalaryHike', 'PerformanceRating', 'RelationshipSatisfaction',
             'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
             'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole',
             'YearsSinceLastPromotion', 'YearsWithCurrManager', 'Attrition']].corr()
     corr_score
```

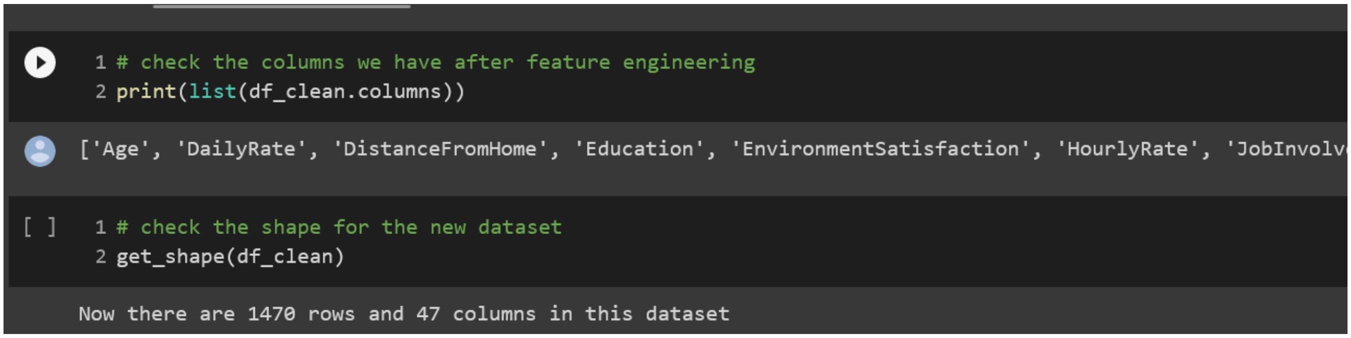
	Age	DailyRate	DistanceFromHome	Education	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel	JobSatisfa
Age	1.000000	0.010661	-0.001686	0.208034	0.010146	0.024287	0.029820	0.509604	-0.0
DailyRate	0.010661	1.000000	-0.004985	-0.016806	0.018355	0.023381	0.046135	0.002966	0.0
DistanceFromHome	-0.001686	-0.004985	1.000000	0.021042	-0.016075	0.031131	0.008783	0.005303	-0.0
Education	0.208034	-0.016806	0.021042	1.000000	-0.027128	0.016775	0.042438	0.101589	-0.0
EnvironmentSatisfaction	0.010146	0.018355	-0.016075	-0.027128	1.000000	-0.049857	-0.008278	0.001212	-0.0



```
[ ]  # Drop the target column and get a clean dataframe with features
     y = df['Attrition']
     df_clean = df.drop(columns = ['Attrition'])
```

```
[ ]  # apply one hot encoding to non numerical features
     df_clean = pd.get_dummies(df_clean, columns = ['BusinessTravel', 'Gender','MaritalStatus'], drop_first = True)
     df_clean = pd.get_dummies(df_clean)
     df_clean.head()
```

	Age	DailyRate	DistanceFromHome	Education	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel	JobSatisfaction	MonthlyIncome	Mon
0	41	1102	1	2	2	94	3	2	4	5993	
1	49	279	8	1	3	61	2	2	2	5130	
2	37	1373	2	2	4	92	2	1	3	2090	
3	33	1392	3	4	4	56	3	1	3	2909	
4	27	591	2	1	1	40	3	1	2	3468	

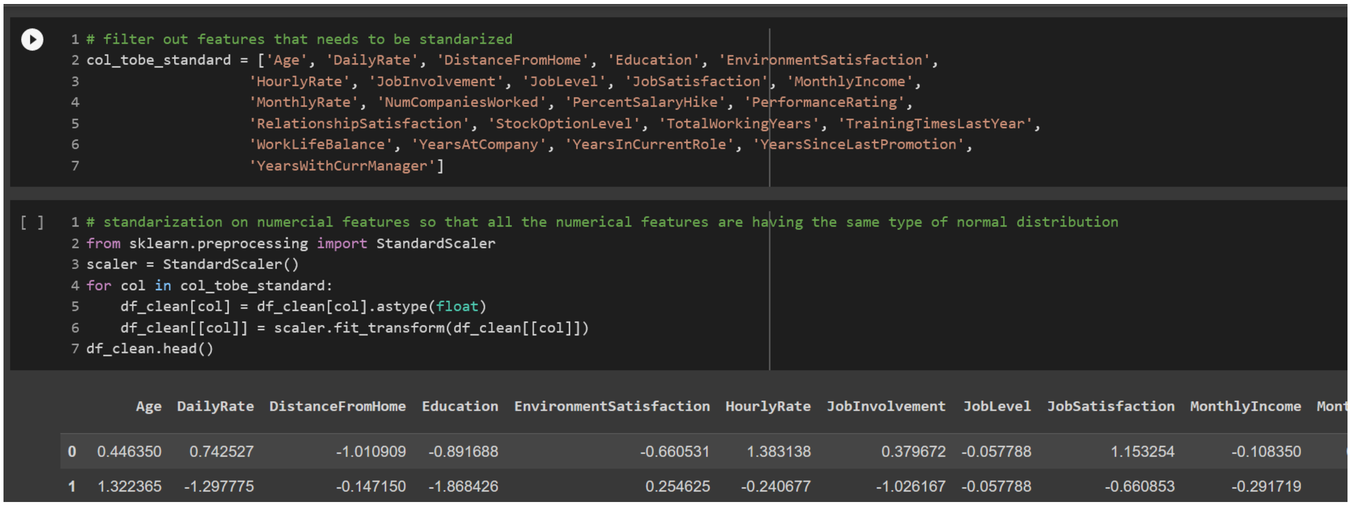


```python
# check the columns we have after feature engineering
print(list(df_clean.columns))
```

```
['Age', 'DailyRate', 'DistanceFromHome', 'Education', 'EnvironmentSatisfaction', 'HourlyRate', 'JobInvolv
```

```python
# check the shape for the new dataset
get_shape(df_clean)
```

```
Now there are 1470 rows and 47 columns in this dataset
```

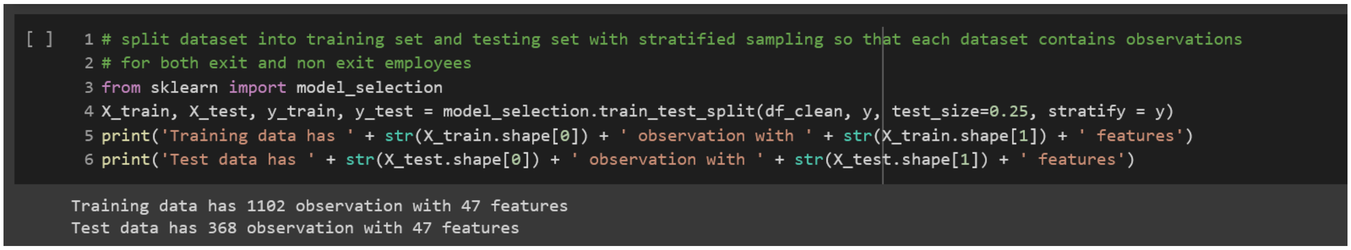


```python
# filter out features that needs to be standarized
col_tobe_standard = ['Age', 'DailyRate', 'DistanceFromHome', 'Education', 'EnvironmentSatisfaction',
                     'HourlyRate', 'JobInvolvement', 'JobLevel', 'JobSatisfaction', 'MonthlyIncome',
                     'MonthlyRate', 'NumCompaniesWorked', 'PercentSalaryHike', 'PerformanceRating',
                     'RelationshipSatisfaction', 'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',
                     'WorkLifeBalance', 'YearsAtCompany', 'YearsInCurrentRole', 'YearsSinceLastPromotion',
                     'YearsWithCurrManager']
```

```python
# standarization on numercial features so that all the numerical features are having the same type of normal distribution
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
for col in col_tobe_standard:
    df_clean[col] = df_clean[col].astype(float)
    df_clean[[col]] = scaler.fit_transform(df_clean[[col]])
df_clean.head()
```

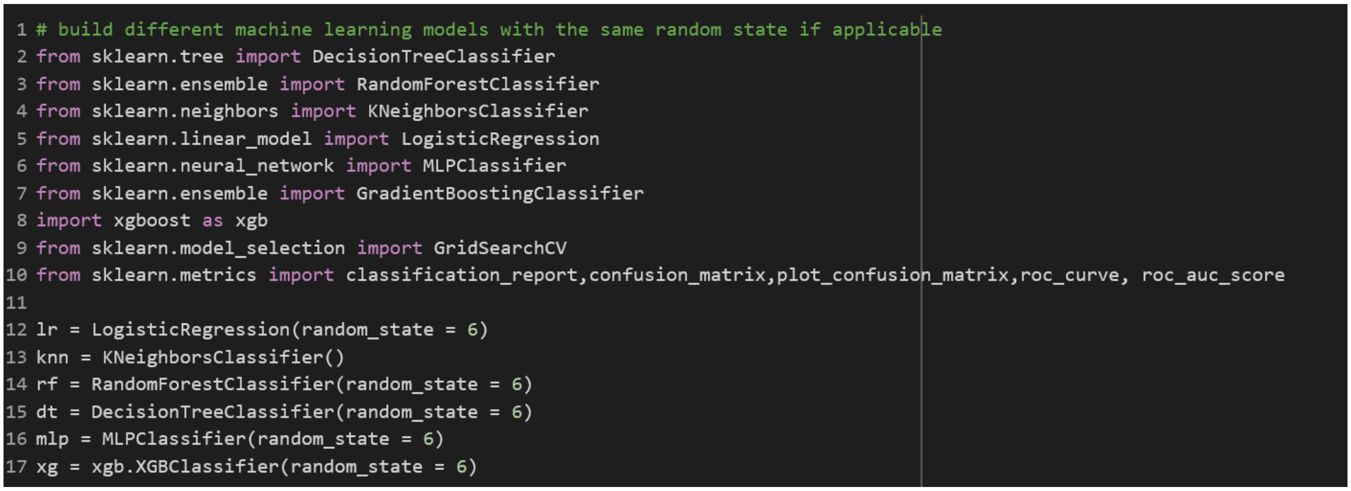
	Age	DailyRate	DistanceFromHome	Education	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel	JobSatisfaction	MonthlyIncome	Mont
0	0.446350	0.742527	-1.010909	-0.891688	-0.660531	1.383138	0.379672	-0.057788	1.153254	-0.108350	
1	1.322365	-1.297775	-0.147150	-1.868426	0.254625	-0.240677	-1.026167	-0.057788	-0.660853	-0.291719	

7.2 Feature 2

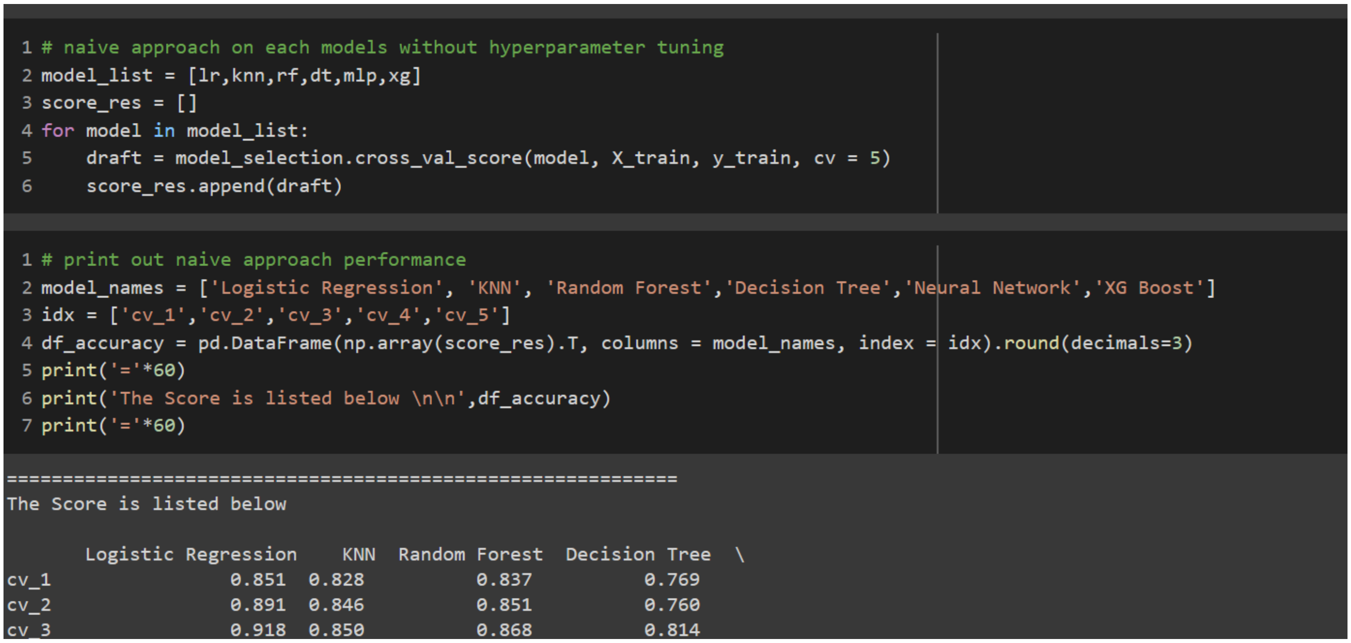


```python
# split dataset into training set and testing set with stratified sampling so that each dataset contains observations
# for both exit and non exit employees
from sklearn import model_selection
X_train, X_test, y_train, y_test = model_selection.train_test_split(df_clean, y, test_size=0.25, stratify = y)
print('Training data has ' + str(X_train.shape[0]) + ' observation with ' + str(X_train.shape[1]) + ' features')
print('Test data has ' + str(X_test.shape[0]) + ' observation with ' + str(X_test.shape[1]) + ' features')
```

```
Training data has 1102 observation with 47 features
Test data has 368 observation with 47 features
```



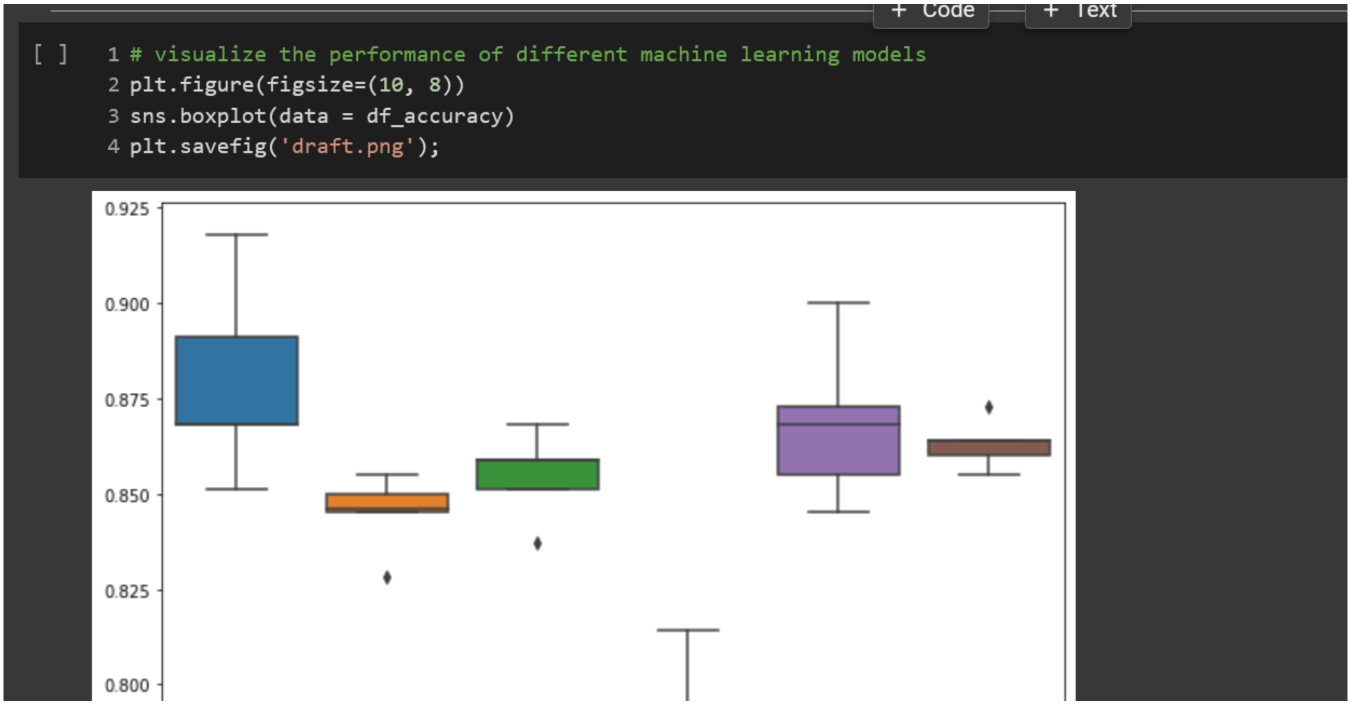
```python
# build different machine learning models with the same random state if applicable
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.neural_network import MLPClassifier
from sklearn.ensemble import GradientBoostingClassifier
import xgboost as xgb
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import classification_report,confusion_matrix,plot_confusion_matrix,roc_curve, roc_auc_score

lr = LogisticRegression(random_state = 6)
knn = KNeighborsClassifier()
rf = RandomForestClassifier(random_state = 6)
dt = DecisionTreeClassifier(random_state = 6)
mlp = MLPClassifier(random_state = 6)
xg = xgb.XGBClassifier(random_state = 6)
```



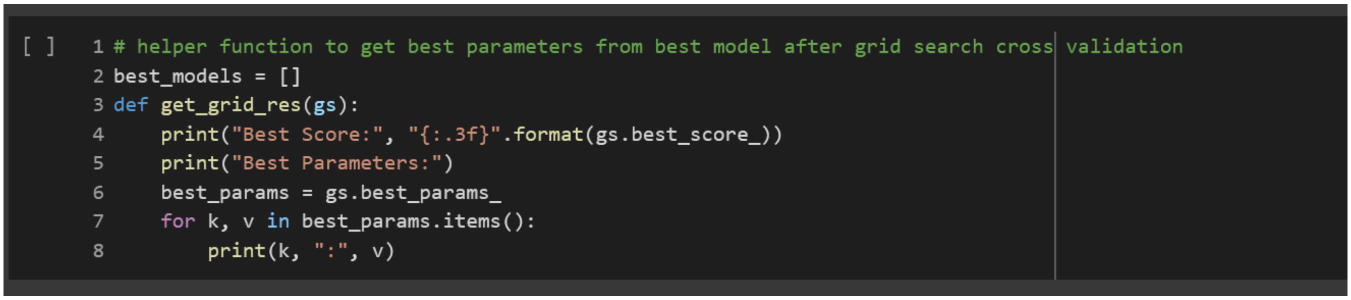
```python
# naive approach on each models without hyperparameter tuning
model_list = [lr,knn,rf,dt,mlp,xg]
score_res = []
for model in model_list:
    draft = model_selection.cross_val_score(model, X_train, y_train, cv = 5)
    score_res.append(draft)
```

```python
# print out naive approach performance
model_names = ['Logistic Regression', 'KNN', 'Random Forest','Decision Tree','Neural Network','XG Boost']
idx = ['cv_1','cv_2','cv_3','cv_4','cv_5']
df_accuracy = pd.DataFrame(np.array(score_res).T, columns = model_names, index = idx).round(decimals=3)
print('='*60)
print('The Score is listed below \n\n',df_accuracy)
print('='*60)
```

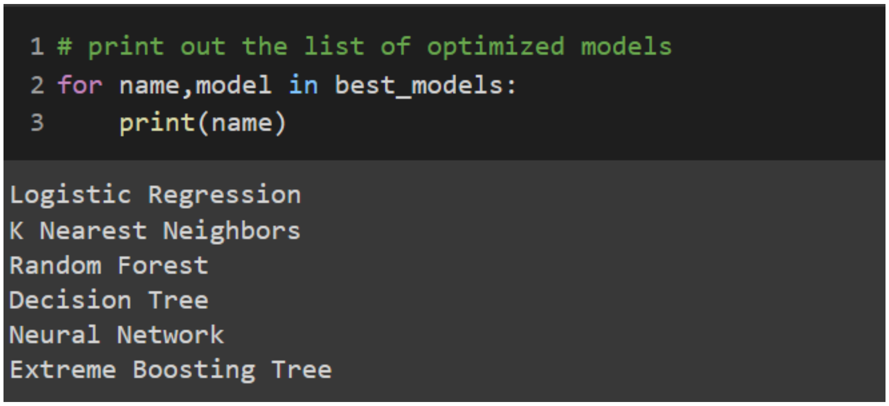
```
============================================================
The Score is listed below 

       Logistic Regression    KNN  Random Forest  Decision Tree  \
cv_1                 0.851  0.828          0.837          0.769   
cv_2                 0.891  0.846          0.851          0.760   
cv_3                 0.918  0.850          0.868          0.814   
```



```python
# visualize the performance of different machine learning models
plt.figure(figsize=(10, 8))
sns.boxplot(data = df_accuracy)
plt.savefig('draft.png');
```

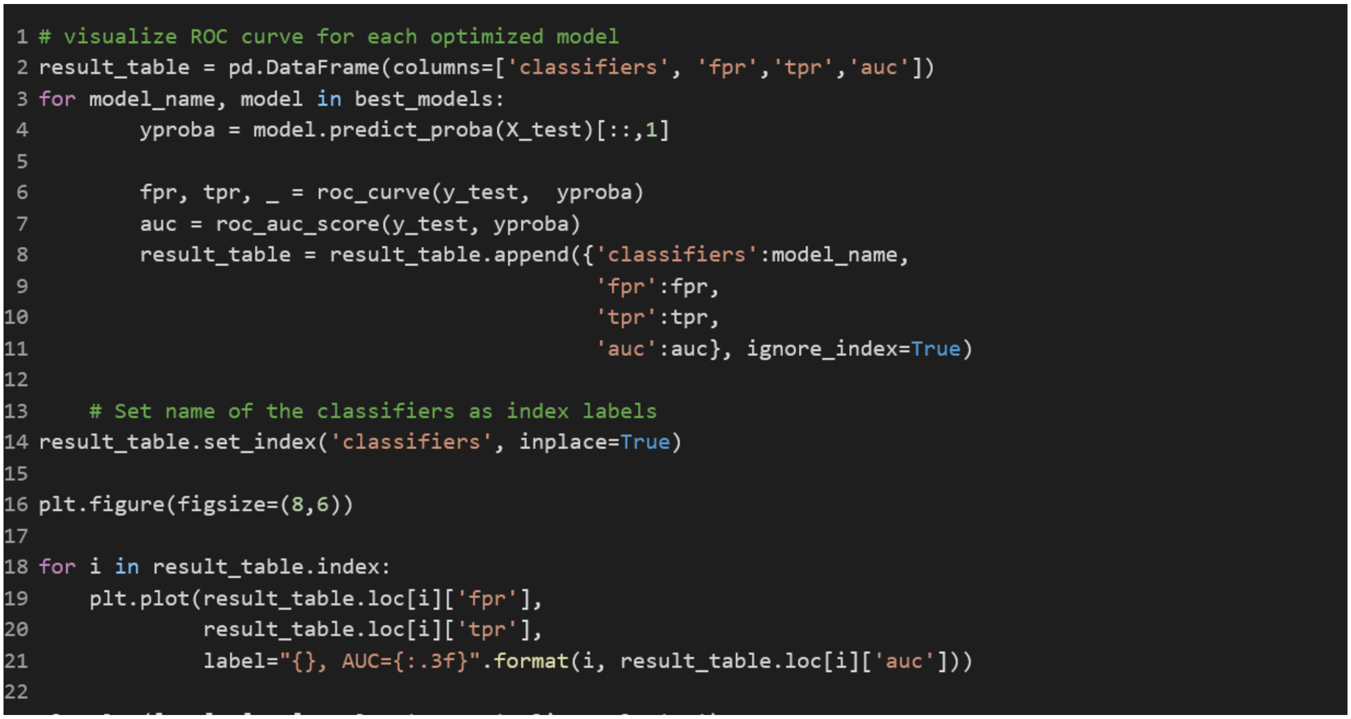


```python
# helper function to get best parameters from best model after grid search cross validation
best_models = []
def get_grid_res(gs):
    print("Best Score:", "{:.3f}".format(gs.best_score_))
    print("Best Parameters:")
    best_params = gs.best_params_
    for k, v in best_params.items():
        print(k, ":", v)
```

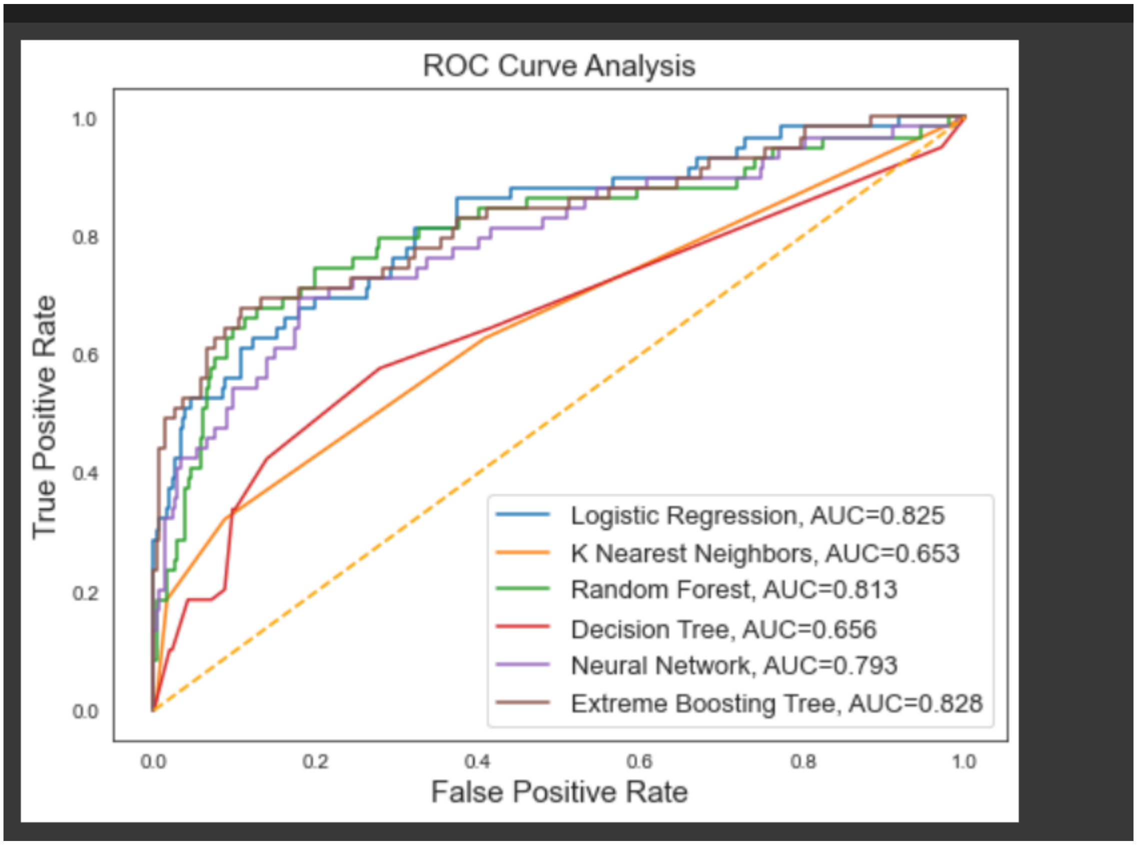


```python
# print out the list of optimized models
for name,model in best_models:
    print(name)
```

```
Logistic Regression
K Nearest Neighbors
Random Forest
Decision Tree
Neural Network
Extreme Boosting Tree
```



```python
# visualize ROC curve for each optimized model
result_table = pd.DataFrame(columns=['classifiers', 'fpr','tpr','auc'])
for model_name, model in best_models:
        yproba = model.predict_proba(X_test)[::,1]

        fpr, tpr, _ = roc_curve(y_test,  yproba)
        auc = roc_auc_score(y_test, yproba)
        result_table = result_table.append({'classifiers':model_name,
                                            'fpr':fpr,
                                            'tpr':tpr,
                                            'auc':auc}, ignore_index=True)

    # Set name of the classifiers as index labels
result_table.set_index('classifiers', inplace=True)

plt.figure(figsize=(8,6))

for i in result_table.index:
    plt.plot(result_table.loc[i]['fpr'],
             result_table.loc[i]['tpr'],
             label="{}, AUC={:.3f}".format(i, result_table.loc[i]['auc']))

```



8 ADVANTAGES & DISADVANTAGES

The study is conducted among working IT professionals of two different categories. This categorization mainly focused on experience level and role in the organization. It was important to know the views of candidates who seek for the job for various reasons as well as the views of interviewers involved in the process of hiring the candidates. There search study involves reference of both primary and secondary data. Primary data is collected through a field survey with the help of a structured self-administrated Questionnaire. The survey consisted of close ended questions by the means of convenience sampling. The scaling technique installed in the questionnaire is 5-point rating scale. Total 120 respondent were IT professionals belonging to the organizations from Nagpur, Pune and Mumbai cities in Maharashtra. Secondary data is collected by referring to the Journals, research papers and published data in the form of books and newspapers.

9 CONCLUSION

Employees as well as organizations must be clear with their expectations regarding the job profile. Any sort of mismatch leads to discrepancy and employees may fail to perform at their job. This eventually leads to attrition. Organizations should state the requirements and expectations unambiguously. This helps candidates decide upon to accept the job position or not. This eventually avoids further conflicts in the employment terms.

10 FUTURE SCOPE

Research findings suggest that attrition reasons in IT organizations primarily revolve around professional growth and challenges in the organization. Although economic factors happen to the most influential factor, professionals may settle for second best criteria of their preference that is career growth and supportive work policies in the organization. On the otherhand, candidates who aspire to have a better job than the one in hand are more interested in securing the next job. Young talent wants to work on latest technology and functional domain. IT professionals who are young career makers are less influenced by Brand name or geographical area. Most of the IT professionals look for challenging role and position in the organization. Candidates as well as senior professionals believe that challenging work motivate them to maintain the interest in the work life. Employees as well as organizations must be clear with their expectations regarding the job profile. Any sort of mismatch leads to discrepancy and employees may fail to perform at their job. This eventually leads to attrition.

Organizations should state the requirements and expectations unambiguously. This helps candidates decide upon to accept the job position or not. This eventually avoids further conflicts in the employment terms. Further this research can make more detailed conclusions over “mapping of candidates” expectations with organizations’ requirement by collecting the data focusing on all the steps of recruitment and selection process.

13 APPENDIX

13.1 Source Code

Source code Link: https://github.com/IBM-EPBL/IBM-Project-25755-1659972531.git

13.2 Github & Project Demo Link

Github link: https://github.com/IBM-EPBL/IBM-Project-25755-1659972531.git

Project Demo Link: https://drive.google.com/file/d/1XWkfowe_MkMFElOcqQ_QpXdEnM7bR-6X/view?usp=sharing