Name- Hrushikesh Doke
Internship Program- Data Science with Machine Learning and Python
Batch- Jan 2022 - Mar 2022
Certificate Code- TCRIB2R186
Date of submission- 5th 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

DATA SCIENCE WITH MACHINE LEARNING AND PYTHON INTERNSHIP PROGRAM

by
Hrushikesh Doke
(TCRIB2R186)

Rutuja Doiphode
CO-FOUNDER &CEO
TCR innovation

Name- Hrushikesh Doke
Internship Program- Data Science with Machine Learning and Python
Batch- Jan 2022 - Mar 2022
Certificate Code- TCRIB2R186
Date of submission- 5th April 2022

Abstract - This paper gives your insight into applying a classification algorithm to a dataset.

Index -

I.Aim

The goal is to determine whether or not an employee wants to continue working. I'm using a dataset called "HR Employee Attrition" for this.

II.NTRODUCTION TO DATASET

The “HR EMPLOYEE ATTRITION DATASET” consists of
the details of an employee like gender, age, business travel, department, education, relationship satisfaction, and many others. Basically, the dataset consists of exactly 2940 employees' data, and employee has 34 features. The dataset consists of both numerical and categorical data. Below is an image of the dataset

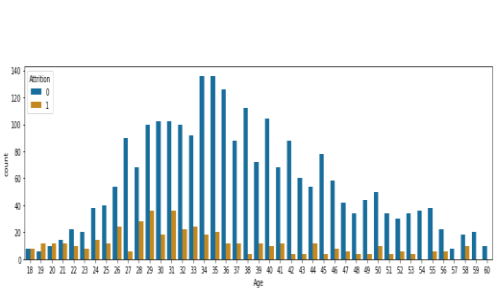
:

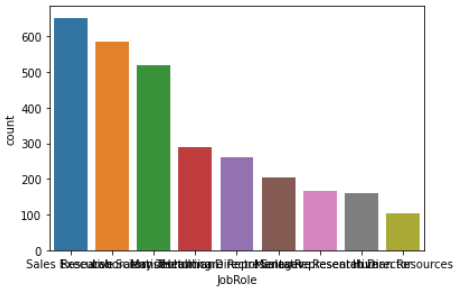
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Attrition	BusinessTra	DailyRate	Departmen	DistanceFro	Education	EducationFi	EmployeeC	EmployeeN	Environmer	Gender	HourlyRate	JobInvolver	JobLevel	JobRole	JobSatisfact	MaritalStat
2	Yes	Travel_Rare	1102	Sales	1	2	Life Science	1	1	2	Female	94	3	2	Sales Execu	4	Single
3	No	Travel_Frec	279	Research &	8	1	Life Science	1	2	3	Male	61	2	2	Research Sc	2	Married
4	Yes	Travel_Rare	1373	Research &	2	2	Other	1	4	4	Male	92	2	1	Laboratory	3	Single
5	No	Travel_Frec	1392	Research &	3	4	Life Science	1	5	4	Female	56	3	1	Research Sc	3	Married
6	No	Travel_Rare	591	Research &	2	1	Medical	1	7	1	Male	40	3	1	Laboratory	2	Married
7	No	Travel_Frec	1005	Research &	2	2	Life Science	1	8	4	Male	79	3	1	Laboratory	4	Single
8	No	Travel_Rare	1324	Research &	3	3	Medical	1	10	3	Female	81	4	1	Laboratory	1	Married
9	No	Travel_Rare	1358	Research &	24	1	Life Science	1	11	4	Male	67	3	1	Laboratory	3	Divorced
10	No	Travel_Frec	216	Research &	23	3	Life Science	1	12	4	Male	44	2	3	Manufactur	3	Single
11	No	Travel_Rare	1299	Research &	27	3	Medical	1	13	3	Male	94	3	2	Healthcare	3	Married
12	No	Travel_Rare	809	Research &	16	3	Medical	1	14	1	Male	84	4	1	Laboratory	2	Married
13	No	Travel_Rare	153	Research &	15	2	Life Science	1	15	4	Female	49	2	2	Laboratory	3	Single
14	No	Travel_Rare	670	Research &	26	1	Life Science	1	16	1	Male	31	3	1	Research Sc	3	Divorced
15	No	Travel_Rare	1346	Research &	19	2	Medical	1	18	2	Male	93	3	1	Laboratory	4	Divorced
16	Yes	Travel_Rare	103	Research &	24	3	Life Science	1	19	3	Male	50	2	1	Laboratory	3	Single
17	No	Travel_Rare	1389	Research &	21	4	Life Science	1	20	2	Female	51	4	3	Manufactur	1	Divorced
18	No	Travel_Rare	334	Research &	5	2	Life Science	1	21	1	Male	80	4	1	Research Sc	2	Divorced
19	No	Non-Travel	1123	Research &	16	2	Medical	1	22	4	Male	96	4	1	Laboratory	4	Divorced
20	No	Travel_Rare	1219	Sales	2	4	Life Science	1	23	1	Female	78	2	4	Manager	4	Married
21	No	Travel_Rare	371	Research &	2	3	Life Science	1	24	4	Male	45	3	1	Research Sc	4	Single
22	No	Non-Travel	673	Research &	11	2	Other	1	26	1	Female	96	4	2	Manufactur	3	Divorced

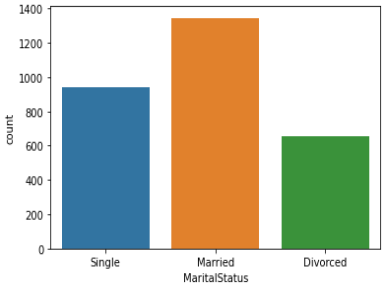
Name- Hrushikesh Doke
Internship Program- Data Science with Machine Learning and Python
Batch- Jan 2022 - Mar 2022
Certificate Code- TCRIB2R186
Date of submission- 5th April 2022

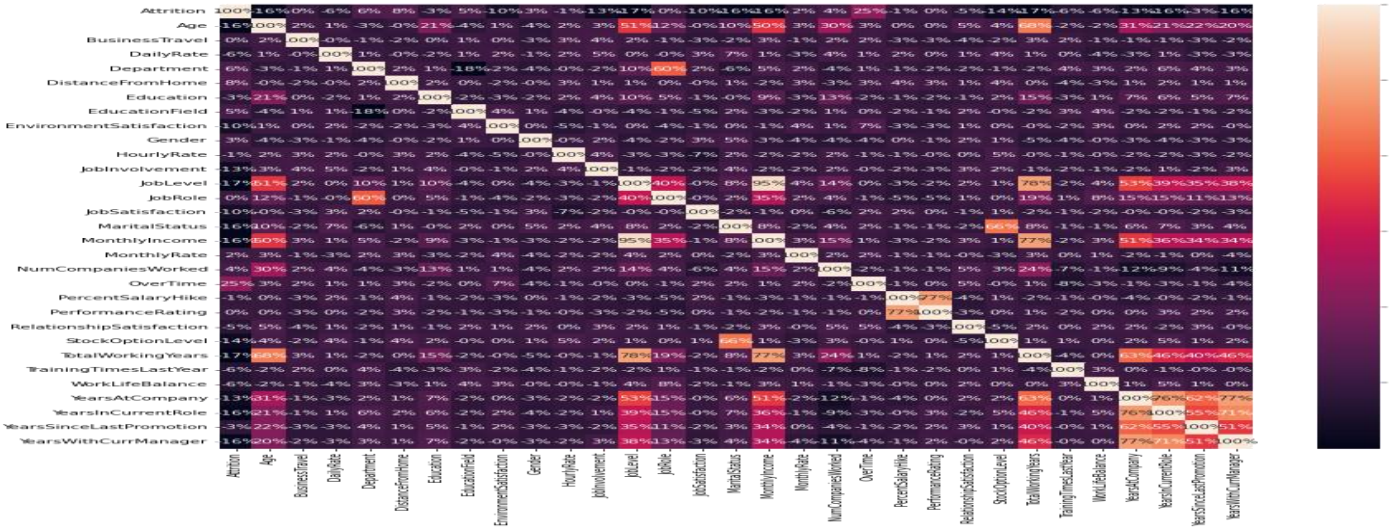
III.EXPLORATORY DATA ANALYSIS ON DATASET

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 overcome the problem of NAN values in the dataset, such as substituting NAN values with the mean, median, or mode value in this dataset, which has no NAN values. I used EDA to compare attrition to a few other fields, and the results were as follows:









IV.TRAINING & PREDICTION OF DATA

I use the Random Forest Classification approach to train the machine learning model after seeing and evaluating the entire dataset. To begin, I divided the dataset 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- Hrushikesh Doke
Internship Program- Data Science with Machine Learning and Python
Batch- Jan 2022 - Mar 2022
Certificate Code- TCRIB2R186
Date of submission- 5th April 2022

Model Traning

```
#spliting the data into x and y var
x = df.iloc[:, 1:df.shape[1]].values  #.values gives us values in array
y = df.iloc[:, 0].values
print(df.shape)
print(x.shape)
print(y.shape)
```

```
(2940, 31)
(2940, 30)
(2940,)
```

```
from sklearn.model_selection import train_test_split
x_train,x_test, y_train, y_test = train_test_split(x, y, test_size=0.25, random_state=0)
#model building
from sklearn.ensemble import RandomForestClassifier
forest=RandomForestClassifier(n_estimators=10 , criterion = 'entropy', random_state=0)
forest.fit(x_train , y_train)
RandomForestClassifier(criterion='entropy', n_estimators=10, random_state=0)

score = forest.score(x_train, y_train)
print('randomforest classifier', np.abs(score)*100)
randomforest classifier 99.36507936507937


from sklearn.metrics import confusion_matrix
cm = confusion_matrix(y_test, forest.predict(x_test))

TN = cm[0][0]
TP = cm[1][1]
FN = cm[1][0]
FP = cm[0][1]

print(cm)
print('Model Testing Accuracy is - ',((TP+TN)/(TP+TN+FP+FN))*100,'%')
[[608   2]
 [ 38  87]]
Model Testing Accuracy is -  94.5578231292517 %

print('Misclassification Rate- ',(FP+FN)/(TP+TN+FP+FN))
print('Precission Rate -' , (TP)/(TP+FP))
print('Recall Rate - ',(TP)/(TP+FN))
Misclassification Rate-  0.05442176870748299
Precission Rate - 0.9775280898876404
Recall Rate -  0.696


print(x_train[0:5])
[[    1   459     1    24     2     1     4     1    73     2     1     1
      4     0  2439 14753     1     1    24     4     2     0     1     3
      2     1     0     1     0    29]
 [    2   530     2     2     4     1     3     0    51     3     2     6
      4     0  4502  7439     3     0    15     3     3     0    17     2
      2    13     7     6     7    36]
 [    2   240     0    22     1     6     4     1    58     1     1     8
      3     1  1555 11585     1     0    11     3     3     1     1     2
      3     1     0     0     0    24]
 [    2   895     1    15     2     1     1     1    50     3     1     2
      3     2  2207 22482     1     0    16     3     4     1     4     5
      2     4     2     2     2    28]
 [    2  1404     2     1     3     1     1     1    59     2     1     6
      1     0  2858 11473     4     0    14     3     1     0    20     3
      2     1     0     0     0    38]]
```

Name- Hrushikesh Doke
Internship Program- Data Science with Machine Learning and Python
Batch- Jan 2022 - Mar 2022
Certificate Code- TCRIB2R186
Date of submission- 5th April 2022

V.CONCLUSION

The one model were able to predict with the following
accuracy percentage:
Random Forest Classifier model has a accuracy score of `99%`
This classifier has better performance occured

Final output of the accuracy of model:

- Conclusion
  - Random forest classifier performance much better having the score of 99%
  - Model Testing Accuracy is 94%

REFERENCES

[1] Book: Machine Learning for Absolute Beginners by Oliver Theobald.

[2] Book: Python for Data Analysis by Wes McKinney.