

Assignment-4 Download the Employee Attrition Dataset

<https://www.kaggle.com/datasets/patelprashant/employee-attrition> 2.Perform Data Preprocessing 3.Model Building using Logistic Regression and Decision Tree and Random Forest 4.Calculate Performance metrics

Name: D GUNASEKHAR Reg No: 21BCI0243

Data Collection. o Collect the dataset or Create the dataset • Data Preprocessing. o Import the Libraries. o Importing the dataset. o Checking for Null Values. o Data Visualization. o Outlier Detection o Splitting Dependent and Independent variables o- Encoding o Feature Scaling. o Splitting Data into Train and Test. • Model Building o Import the model building Libraries o Initializing the model o Training and testing the model o Evaluation of Model o Save the Model

```
#Import the Libraries.
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

#Importing the dataset.
df=pd.read_csv("HR_Employee_Attrition.csv")

df.head()
```

	Age	Attrition	BusinessTravel	DailyRate	Department
0	41	Yes	Travel_Rarely	1102	Sales
1	49	No	Travel_Frequently	279	Research & Development
2	37	Yes	Travel_Rarely	1373	Research & Development
3	33	No	Travel_Frequently	1392	Research & Development
4	27	No	Travel_Rarely	591	Research & Development

	DistanceFromHome	Education	EducationField	EmployeeCount
0	1	2	Life Sciences	1
1				
1	8	1	Life Sciences	1
2				
2	2	2	Other	1
4				
3	3	4	Life Sciences	1
5				
4	2	1	Medical	1
7				

	...	RelationshipSatisfaction	StandardHours	StockOptionLevel	\
0	...	1	80	0	
1	...	4	80	1	
2	...	2	80	0	
3	...	3	80	0	
4	...	4	80	1	

	TotalWorkingYears	TrainingTimesLastYear	WorkLifeBalance
YearsAtCompany \			
0	8	0	1
6			
1	10	3	3
10			
2	7	3	3
0			
3	8	3	3
8			
4	6	3	3
2			

	YearsInCurrentRole	YearsSinceLastPromotion	YearsWithCurrManager
0	4	0	5
1	7	1	7
2	0	0	0
3	7	3	0
4	2	2	2

[5 rows x 35 columns]

df.shape

(1470, 35)

df.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 1470 entries, 0 to 1469

Data columns (total 35 columns):

#	Column	Non-Null Count	Dtype
---	-----	-----	-----
0	Age	1470 non-null	int64
1	Attrition	1470 non-null	object
2	BusinessTravel	1470 non-null	object
3	DailyRate	1470 non-null	int64
4	Department	1470 non-null	object
5	DistanceFromHome	1470 non-null	int64
6	Education	1470 non-null	int64
7	EducationField	1470 non-null	object
8	EmployeeCount	1470 non-null	int64
9	EmployeeNumber	1470 non-null	int64

10	EnvironmentSatisfaction	1470	non-null	int64
11	Gender	1470	non-null	object
12	HourlyRate	1470	non-null	int64
13	JobInvolvement	1470	non-null	int64
14	JobLevel	1470	non-null	int64
15	JobRole	1470	non-null	object
16	JobSatisfaction	1470	non-null	int64
17	MaritalStatus	1470	non-null	object
18	MonthlyIncome	1470	non-null	int64
19	MonthlyRate	1470	non-null	int64
20	NumCompaniesWorked	1470	non-null	int64
21	Over18	1470	non-null	object
22	OverTime	1470	non-null	object
23	PercentSalaryHike	1470	non-null	int64
24	PerformanceRating	1470	non-null	int64
25	RelationshipSatisfaction	1470	non-null	int64
26	StandardHours	1470	non-null	int64
27	StockOptionLevel	1470	non-null	int64
28	TotalWorkingYears	1470	non-null	int64
29	TrainingTimesLastYear	1470	non-null	int64
30	WorkLifeBalance	1470	non-null	int64
31	YearsAtCompany	1470	non-null	int64
32	YearsInCurrentRole	1470	non-null	int64
33	YearsSinceLastPromotion	1470	non-null	int64
34	YearsWithCurrManager	1470	non-null	int64

dtypes: int64(26), object(9)

memory usage: 402.1+ KB

df.describe()

	Age	DailyRate	DistanceFromHome	Education
EmployeeCount \				
count	1470.000000	1470.000000	1470.000000	1470.000000
1470.0				
mean	36.923810	802.485714	9.192517	2.912925
1.0				
std	9.135373	403.509100	8.106864	1.024165
0.0				
min	18.000000	102.000000	1.000000	1.000000
1.0				
25%	30.000000	465.000000	2.000000	2.000000
1.0				
50%	36.000000	802.000000	7.000000	3.000000
1.0				
75%	43.000000	1157.000000	14.000000	4.000000
1.0				
max	60.000000	1499.000000	29.000000	5.000000
1.0				

EmployeeNumber	EnvironmentSatisfaction	HourlyRate
----------------	-------------------------	------------

JobInvolvement \			
count	1470.000000	1470.000000	1470.000000
1470.000000			
mean	1024.865306	2.721769	65.891156
2.729932			
std	602.024335	1.093082	20.329428
0.711561			
min	1.000000	1.000000	30.000000
1.000000			
25%	491.250000	2.000000	48.000000
2.000000			
50%	1020.500000	3.000000	66.000000
3.000000			
75%	1555.750000	4.000000	83.750000
3.000000			
max	2068.000000	4.000000	100.000000
4.000000			

	JobLevel	...	RelationshipSatisfaction	StandardHours	\
count	1470.000000	...	1470.000000	1470.0	
mean	2.063946	...	2.712245	80.0	
std	1.106940	...	1.081209	0.0	
min	1.000000	...	1.000000	80.0	
25%	1.000000	...	2.000000	80.0	
50%	2.000000	...	3.000000	80.0	
75%	3.000000	...	4.000000	80.0	
max	5.000000	...	4.000000	80.0	

	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	\
count	1470.000000	1470.000000	1470.000000	
mean	0.793878	11.279592	2.799320	
std	0.852077	7.780782	1.289271	
min	0.000000	0.000000	0.000000	
25%	0.000000	6.000000	2.000000	
50%	1.000000	10.000000	3.000000	
75%	1.000000	15.000000	3.000000	
max	3.000000	40.000000	6.000000	

	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	\
count	1470.000000	1470.000000	1470.000000	
mean	2.761224	7.008163	4.229252	
std	0.706476	6.126525	3.623137	
min	1.000000	0.000000	0.000000	
25%	2.000000	3.000000	2.000000	
50%	3.000000	5.000000	3.000000	
75%	3.000000	9.000000	7.000000	
max	4.000000	40.000000	18.000000	

	YearsSinceLastPromotion	YearsWithCurrManager
count	1470.000000	1470.000000

mean	2.187755	4.123129
std	3.222430	3.568136
min	0.000000	0.000000
25%	0.000000	2.000000
50%	1.000000	3.000000
75%	3.000000	7.000000
max	15.000000	17.000000

[8 rows x 26 columns]

#Checking for Null Values.

df.isnull().any()

Age	False
Attrition	False
BusinessTravel	False
DailyRate	False
Department	False
DistanceFromHome	False
Education	False
EducationField	False
EmployeeCount	False
EmployeeNumber	False
EnvironmentSatisfaction	False
Gender	False
HourlyRate	False
JobInvolvement	False
JobLevel	False
JobRole	False
JobSatisfaction	False
MaritalStatus	False
MonthlyIncome	False
MonthlyRate	False
NumCompaniesWorked	False
Over18	False
OverTime	False
PercentSalaryHike	False
PerformanceRating	False
RelationshipSatisfaction	False
StandardHours	False
StockOptionLevel	False
TotalWorkingYears	False
TrainingTimesLastYear	False
WorkLifeBalance	False
YearsAtCompany	False
YearsInCurrentRole	False
YearsSinceLastPromotion	False
YearsWithCurrManager	False
dtype:	bool

```

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
JobRole 0
JobSatisfaction 0
MaritalStatus 0
MonthlyIncome 0
MonthlyRate 0
NumCompaniesWorked 0
Over18 0
OverTime 0
PercentSalaryHike 0
PerformanceRating 0
RelationshipSatisfaction 0
StandardHours 0
StockOptionLevel 0
TotalWorkingYears 0
TrainingTimesLastYear 0
WorkLifeBalance 0
YearsAtCompany 0
YearsInCurrentRole 0
YearsSinceLastPromotion 0
YearsWithCurrManager 0
dtype: int64

df=df.drop('EmployeeCount', axis=1) # as there are having 0 standard deviation
df=df.drop('StandardHours', axis=1)

```

no null values in the given dataset

```

#Data Visualization.
sns.distplot(df["Age"])

```

```
C:\Users\Surya\AppData\Local\Temp\ipykernel_24908\2400079689.py:2:
UserWarning:
```

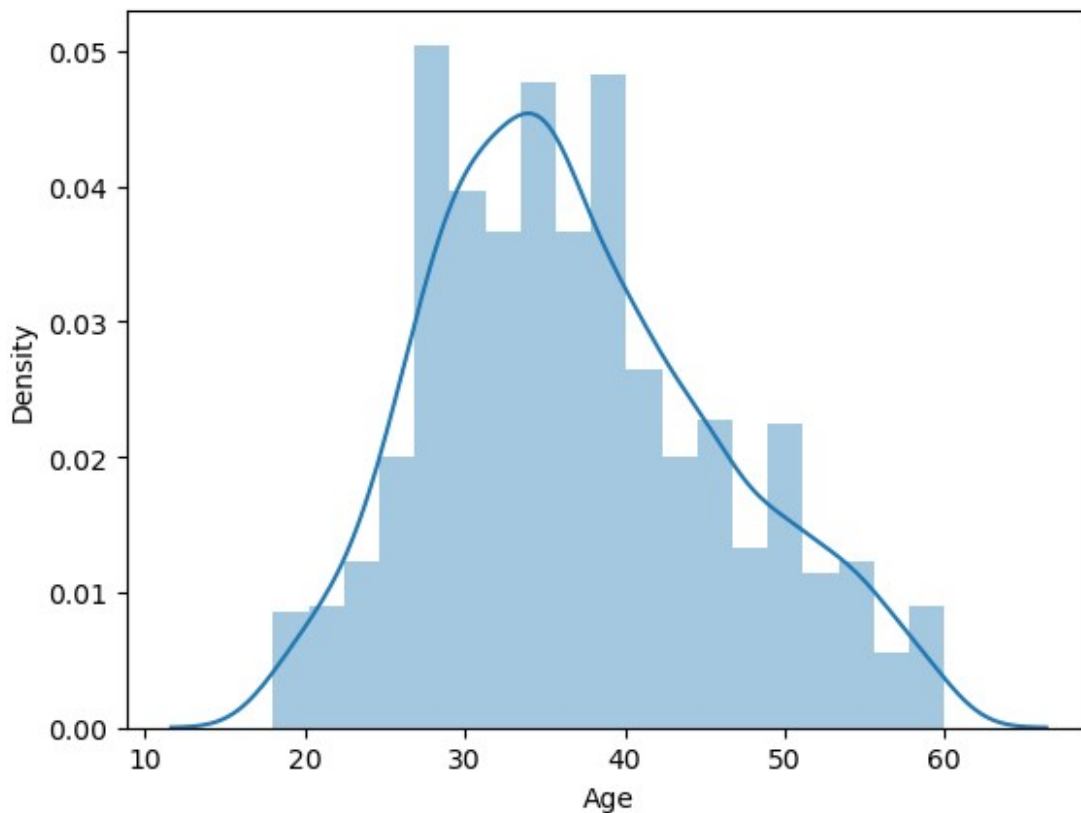
```
`distplot` is a deprecated function and will be removed in seaborn
v0.14.0.
```

Please adapt your code to use either `displot` (a figure-level function with similar flexibility) or `histplot` (an axes-level function for histograms).

For a guide to updating your code to use the new functions, please see <https://gist.github.com/mwaskom/de44147ed2974457ad6372750bbe5751>

```
sns.distplot(df["Age"])
```

```
<Axes: xlabel='Age', ylabel='Density'>
```



```
corr=df.corr()
```

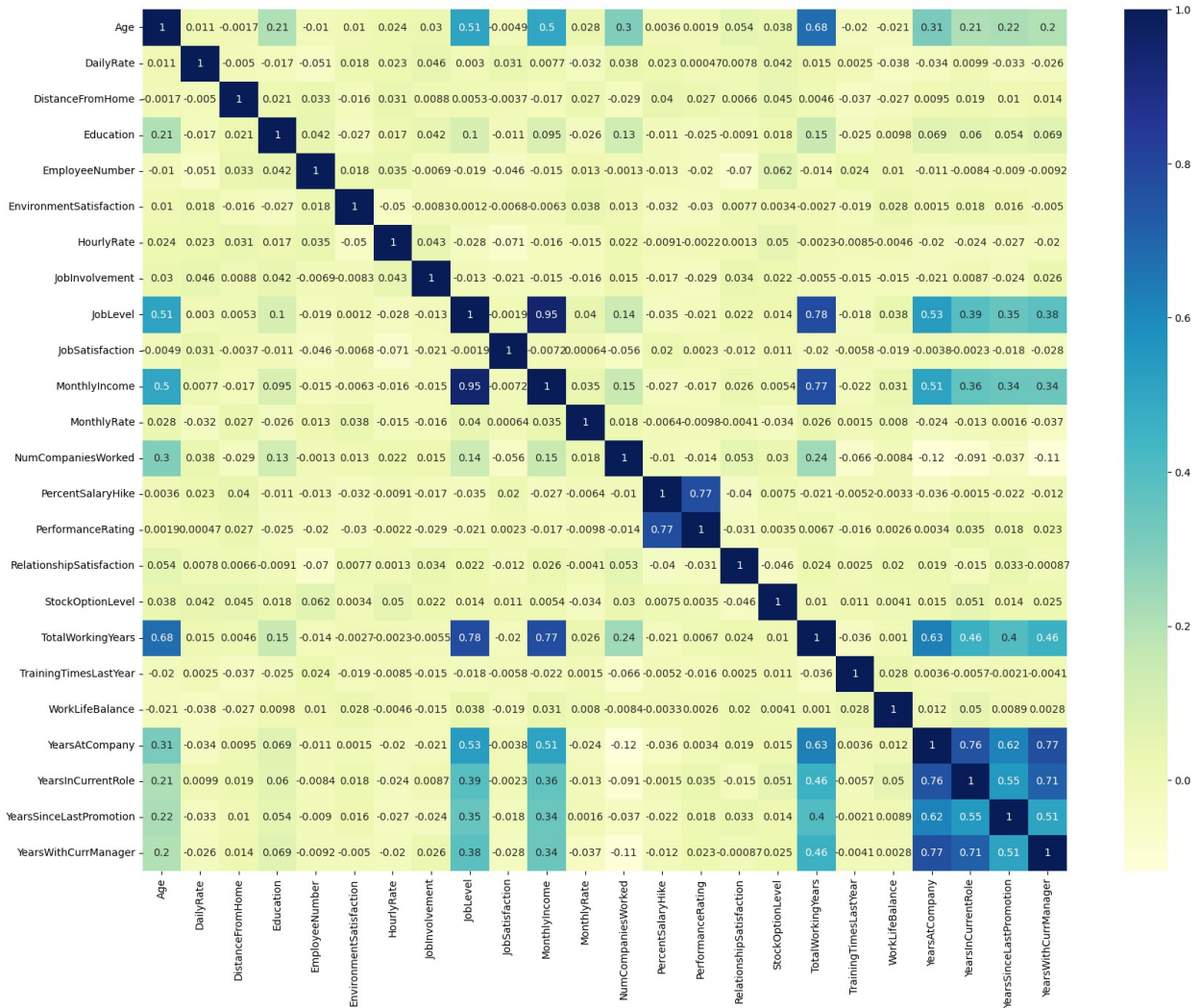
```
C:\Users\Surya\AppData\Local\Temp\ipykernel_24908\1515641297.py:1:
FutureWarning: The default value of numeric_only in DataFrame.corr is
```

deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_only to silence this warning.

```
corr=df.corr()
```

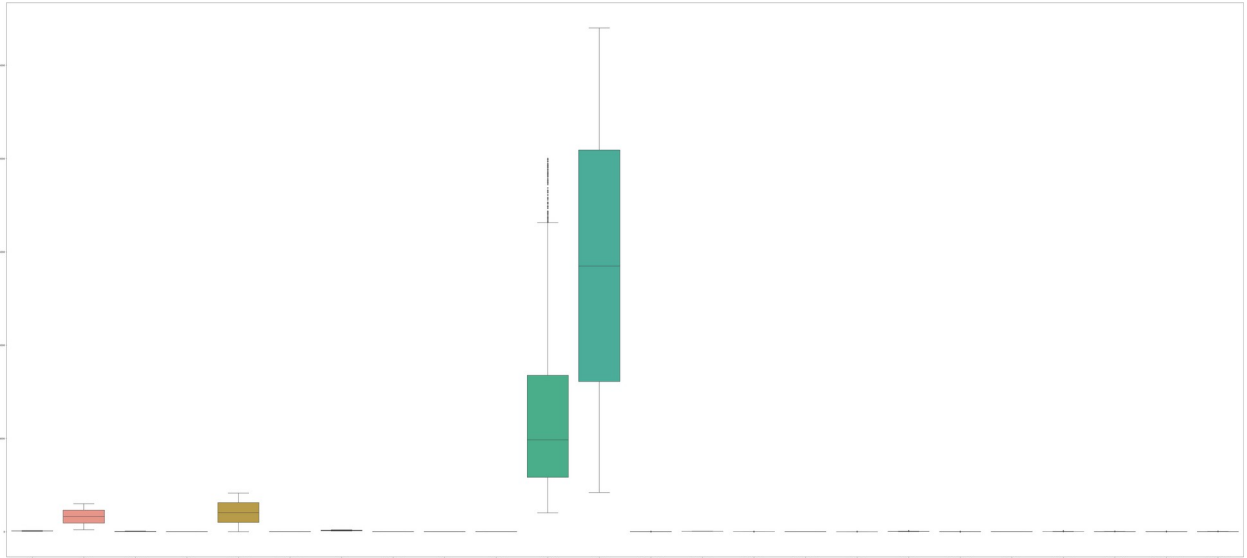
```
plt.subplots(figsize=(20,15))
sns.heatmap(corr,annot=True,cmap="YlGnBu")
```

<Axes: >



```
plt.subplots(figsize=(100,45))
sns.boxplot(df)
```

<Axes: >



outliers are present in monthlyincome feature removing them using iqr method

```
q1= df.MonthlyIncome.quantile(0.25)# finding q1 formula  
q3= df.MonthlyIncome.quantile(0.75)# finding q3 formula
```

```
IQR = q3-q1
```

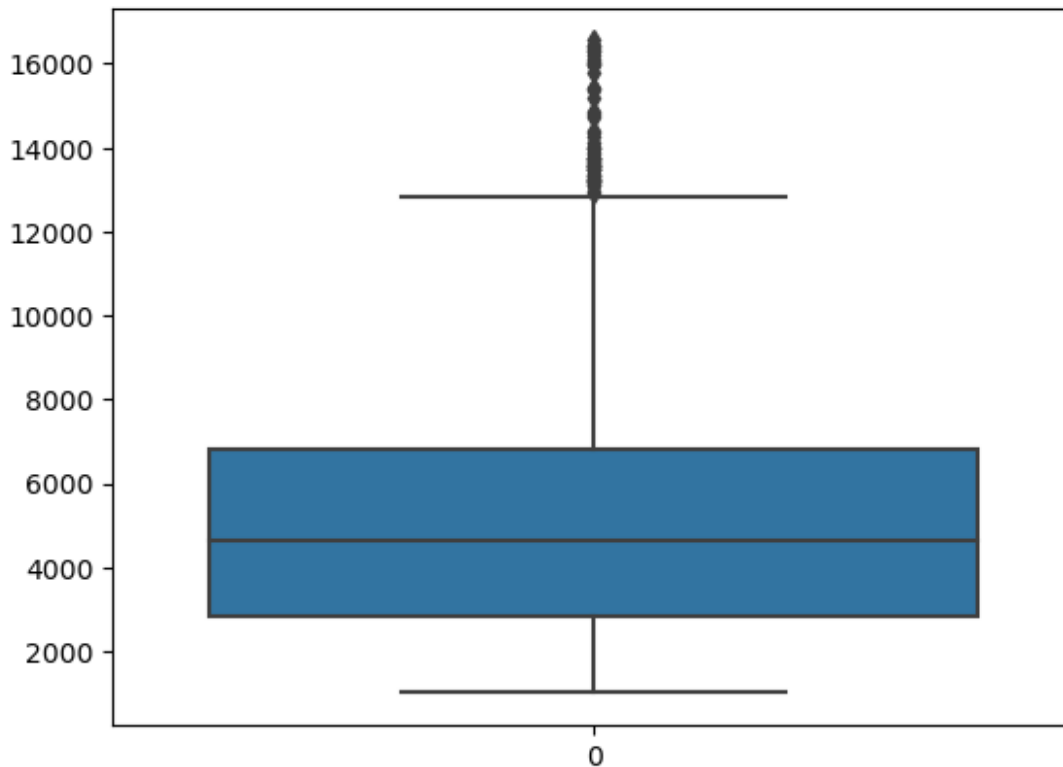
```
uuper_limit = q3+1.5*(IQR)  
uuper_limit
```

```
16581.0
```

```
df=df[df.MonthlyIncome<uuper_limit]
```

```
sns.boxplot(df[ 'MonthlyIncome' ])
```

```
<Axes: >
```



```
df.shape
```

```
(1356, 33)
```

```
#label encoding
```

```
from sklearn.preprocessing import LabelEncoder
```

```
le=LabelEncoder()
```

```
df.Gender=le.fit_transform(df.Gender)
```

```
df
```

	Age	Attrition	BusinessTravel	DailyRate	
Department \					
0	41	Yes	Travel_Rarely	1102	
Sales					
1	49	No	Travel_Frequently	279	Research &
Development					
2	37	Yes	Travel_Rarely	1373	Research &
Development					
3	33	No	Travel_Frequently	1392	Research &
Development					
4	27	No	Travel_Rarely	591	Research &
Development					
...
...					
1465	36	No	Travel_Frequently	884	Research &

Development	1466	39	No	Travel_Rarely	613	Research &
Development	1467	27	No	Travel_Rarely	155	Research &
Development	1468	49	No	Travel_Frequently	1023	
Sales	1469	34	No	Travel_Rarely	628	Research &
Development						

	DistanceFromHome	Education	EducationField	EmployeeNumber	\
0	1	2	Life Sciences	1	
1	8	1	Life Sciences	2	
2	2	2	Other	4	
3	3	4	Life Sciences	5	
4	2	1	Medical	7	
...	
1465	23	2	Medical	2061	
1466	6	1	Medical	2062	
1467	4	3	Life Sciences	2064	
1468	2	3	Medical	2065	
1469	8	3	Medical	2068	

	EnvironmentSatisfaction	...	PerformanceRating	\
0	2	...	3	
1	3	...	4	
2	4	...	3	
3	4	...	3	
4	1	...	3	
...	
1465	3	...	3	
1466	4	...	3	
1467	2	...	4	
1468	4	...	3	
1469	2	...	3	

	RelationshipSatisfaction	StockOptionLevel	TotalWorkingYears	\
0	1	0	8	
1	4	1	10	
2	2	0	7	
3	3	0	8	
4	4	1	6	
...	
1465	3	1	17	
1466	1	1	9	
1467	2	1	6	
1468	4	0	17	
1469	1	0	6	

TrainingTimesLastYear	WorkLifeBalance	YearsAtCompany	\
-----------------------	-----------------	----------------	---

0	0	1	6
1	3	3	10
2	3	3	0
3	3	3	8
4	3	3	2
...
1465	3	3	5
1466	5	3	7
1467	0	3	6
1468	3	2	9
1469	3	4	4

	YearsInCurrentRole	YearsSinceLastPromotion
YearsWithCurrManager		

0	4	0
5		
1	7	1
7		
2	0	0
0		
3	7	3
0		
4	2	2
2		

...
.			
1465	2	0	
3			
1466	7	1	
7			
1467	2	0	
3			
1468	6	0	
8			
1469	3	1	
2			

[1356 rows x 33 columns]

#Convert categorical variables to dummy/indicator variables (one-hot encoding)

```
df = pd.get_dummies(df, columns=['Department', 'EducationField',
'MaritalStatus',
'Gender', 'JobRole', 'Over18', 'OverTime', 'BusinessTravel'],
drop_first=True)
```

df

Age	Attrition	DailyRate	DistanceFromHome	Education
EmployeeNumber \				

0	41	Yes	1102	1	2
1					
1	49	No	279	8	1
2					
2	37	Yes	1373	2	2
4					
3	33	No	1392	3	4
5					
4	27	No	591	2	1
7					
...
...					
1465	36	No	884	23	2
2061					
1466	39	No	613	6	1
2062					
1467	27	No	155	4	3
2064					
1468	49	No	1023	2	3
2065					
1469	34	No	628	8	3
2068					

EnvironmentSatisfaction		HourlyRate	JobInvolvement	
JobLevel	...	\		
0		2	94	3
2	...			
1		3	61	2
2	...			
2		4	92	2
1	...			
3		4	56	3
1	...			
4		1	40	3
1	...			
...	
...				...
1465		3	41	4
2	...			
1466		4	42	2
3	...			
1467		2	87	4
2	...			
1468		4	63	2
2	...			
1469		2	82	4
2	...			

JobRole_Laboratory Technician JobRole_Manager \

0	0	0
1	0	0
2	1	0
3	0	0
4	1	0
...
1465	1	0
1466	0	0
1467	0	0
1468	0	0
1469	1	0

	JobRole_Manufacturing Director	JobRole_Research Director \
0	0	0
1	0	0
2	0	0
3	0	0
4	0	0
...
1465	0	0
1466	0	0
1467	1	0
1468	0	0
1469	0	0

	JobRole_Research Scientist	JobRole_Sales Executive \
0	0	1
1	1	0
2	0	0
3	1	0
4	0	0
...
1465	0	0
1466	0	0
1467	0	0
1468	0	1
1469	0	0

	JobRole_Sales Representative	OverTime_Yes \
0	0	1
1	0	0
2	0	1
3	0	1
4	0	0
...
1465	0	0
1466	0	0
1467	0	1
1468	0	0
1469	0	0

	BusinessTravel_Travel_Frequently	BusinessTravel_Travel_Rarely
0	0	1
1	1	0
2	0	1
3	1	0
4	0	1
...
1465	1	0
1466	0	1
1467	0	1
1468	1	0
1469	0	1

[1356 rows x 46 columns]

df.head()

	Age	Attrition	DailyRate	DistanceFromHome	Education
EmployeeNumber \					
0	41	Yes	1102	1	2
1					
1	49	No	279	8	1
2					
2	37	Yes	1373	2	2
4					
3	33	No	1392	3	4
5					
4	27	No	591	2	1
7					

	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel	...
\					
0		2	94	3	2 ...
1		3	61	2	2 ...
2		4	92	2	1 ...
3		4	56	3	1 ...
4		1	40	3	1 ...

	JobRole_Laboratory Technician	JobRole_Manager
\		
0	0	0
1	0	0
2	1	0
3	0	0
4	1	0

	JobRole_Manufacturing Director	JobRole_Research Director	\
0	0	0	
1	0	0	
2	0	0	
3	0	0	
4	0	0	

	JobRole_Research Scientist	JobRole_Sales Executive	\
0	0	1	
1	1	0	
2	0	0	
3	1	0	
4	0	0	

	JobRole_Sales Representative	OverTime_Yes	\
0	0	1	
1	0	0	
2	0	1	
3	0	1	
4	0	0	

	BusinessTravel_Travel_Frequently	BusinessTravel_Travel_Rarely
0	0	1
1	1	0
2	0	1
3	1	0
4	0	1

[5 rows x 46 columns]

```
#Splitting Dependent and Independent variables
# here the dependent variable (y) is attrition and remaining are
independent variables(x)
x=df.drop('Attrition', axis=1)
```

```
y=df.Attrition
y.head()
```

```
0    Yes
1    No
2    Yes
3    No
4    No
Name: Attrition, dtype: object
```

```
x
```

	Age	DailyRate	DistanceFromHome	Education	EmployeeNumber	\
0	41	1102	1	2	1	
1	49	279	8	1	2	
2	37	1373	2	2	4	

3	33	1392	3	4	5
4	27	591	2	1	7
...
1465	36	884	23	2	2061
1466	39	613	6	1	2062
1467	27	155	4	3	2064
1468	49	1023	2	3	2065
1469	34	628	8	3	2068

	EnvironmentSatisfaction	HourlyRate	JobInvolvement	JobLevel	\
0	2	94	3	2	
1	3	61	2	2	
2	4	92	2	1	
3	4	56	3	1	
4	1	40	3	1	
...	
1465	3	41	4	2	
1466	4	42	2	3	
1467	2	87	4	2	
1468	4	63	2	2	
1469	2	82	4	2	

	JobSatisfaction	...	JobRole_Laboratory Technician
JobRole_Manager \			
0	4	...	0
0			
1	2	...	0
0			
2	3	...	1
0			
3	3	...	0
0			
4	2	...	1
0			
...
...			
1465	4	...	1
0			
1466	1	...	0
0			
1467	2	...	0
0			
1468	2	...	0
0			
1469	3	...	1
0			

	JobRole_Manufacturing Director	JobRole_Research Director	\
0	0	0	
1	0	0	

2	0	0
3	0	0
4	0	0
...
1465	0	0
1466	0	0
1467	1	0
1468	0	0
1469	0	0

	JobRole_Research Scientist	JobRole_Sales Executive \
0	0	1
1	1	0
2	0	0
3	1	0
4	0	0
...
1465	0	0
1466	0	0
1467	0	0
1468	0	1
1469	0	0

	JobRole_Sales Representative	OverTime_Yes \
0	0	1
1	0	0
2	0	1
3	0	1
4	0	0
...
1465	0	0
1466	0	0
1467	0	1
1468	0	0
1469	0	0

	BusinessTravel_Travel_Frequently	BusinessTravel_Travel_Rarely
0	0	1
1	1	0
2	0	1
3	1	0
4	0	1
...
1465	1	0
1466	0	1
1467	0	1
1468	1	0
1469	0	1

[1356 rows x 45 columns]

```

#feature scaling
#Feature scaling is a method used to standardize the range of
independent variables or features of data.
#Since the range of values of raw data varies widely, in some machine
learning algorithms, objective functions will not work properly
without normalization.
from sklearn.preprocessing import StandardScaler
scale = StandardScaler()
x_scaled = scale.fit_transform(x)

x_scaled
array([[ 0.56346638,  0.743224 , -1.02743082, ...,  1.58522975,
        -0.48937523,  0.64799328],
       [ 1.47198055, -1.29718689, -0.16621806, ..., -0.63082339,
         2.04342177, -1.54322589],
       [ 0.1092093 ,  1.41509685, -0.90440042, ...,  1.58522975,
        -0.48937523,  0.64799328],
       ...,
       [-1.02643341, -1.6046121 , -0.65833963, ...,  1.58522975,
        -0.48937523,  0.64799328],
       [ 1.47198055,  0.54736439, -0.90440042, ..., -0.63082339,
         2.04342177, -1.54322589],
       [-0.23148351, -0.43193367, -0.16621806, ..., -0.63082339,
        -0.48937523,  0.64799328]])

#Splitting Data into Train and Test.
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x_scaled,y,test_size=0.
2,random_state=42)

x_train.shape,x_test.shape,y_train.shape,y_test.shape
((1084, 45), (272, 45), (1084,), (272,))

x_train
array([[ -0.11791924,  0.84239343,  1.67923787, ...,  1.58522975,
        -0.48937523,  0.64799328],
       [  1.244852 ,  0.89197814, -1.02743082, ..., -0.63082339,
        -0.48937523, -1.54322589],
       [  0.56346638, -1.57486127,  0.32590352, ...,  1.58522975,
        -0.48937523,  0.64799328],
       ...,
       [  2.0398019 , -1.59965363,  0.07984273, ...,  1.58522975,
        -0.48937523,  0.64799328],
       [-0.00435497, -1.18314205,  0.07984273, ..., -0.63082339,
        -0.48937523,  0.64799328],
       [  0.56346638, -0.62035557, -0.65833963, ..., -0.63082339,
        -0.48937523, -1.54322589]])

```

Model Building o Import the model building Libraries o Initializing the model o Training and testing the model o Evaluation of Model o Save the Model

Model Building using logistic Regression

```
from sklearn.linear_model import LogisticRegression
model=LogisticRegression()

model.fit(x_train,y_train)

LogisticRegression()

pred=model.predict(x_test)
pred
array(['No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No',
       'No',
       'No', 'No', 'No', 'Yes', 'No', 'No', 'No', 'No', 'No', 'No',
       'No',
       'No', 'No', 'Yes', 'No', 'Yes', 'No', 'No', 'No', 'No', 'No',
       'No',
       'No', 'No', 'No', 'No', 'No', 'No', 'Yes', 'Yes', 'No', 'Yes',
       'No', 'No', 'No', 'No', 'No', 'No', 'No', 'Yes', 'No', 'No',
       'No',
       'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'Yes',
       'No',
       'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No',
       'No',
       'No', 'Yes', 'No', 'No', 'Yes', 'No', 'No', 'No', 'Yes', 'No',
       'No', 'Yes', 'No', 'Yes', 'No', 'No', 'No', 'No', 'No', 'No',
       'No',
       'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No',
       'No',
       'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No',
       'No',
       'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'Yes', 'No',
       'No',
       'No', 'No', 'No', 'No', 'No', 'No', 'No', 'Yes', 'No', 'No',
       'No',
       'Yes', 'No', 'No', 'Yes', 'No', 'No', 'No', 'No', 'No', 'No',
       'No',
       'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No',
       'No',
       'No', 'No', 'No', 'No', 'Yes', 'No', 'No', 'No', 'No', 'No',
       'No',
       'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'Yes',
       'No',
       'No', 'Yes', 'No', 'No', 'No', 'No', 'Yes', 'No', 'No', 'No',
       'No',
       'No',
```

```

'No', 'No', 'Yes', 'No', 'No', 'No', 'No', 'No', 'Yes', 'No', 'No',
'No',
'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No',
'No',
'No', 'No', 'No', 'Yes', 'No', 'No', 'No', 'Yes', 'No', 'No',
'No',
'No', 'No', 'No', 'No', 'No', 'No', 'Yes', 'Yes', 'Yes', 'Yes',
'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'Yes', 'No',
'Yes',
'No', 'No', 'Yes', 'No', 'No', 'No', 'No', 'No', 'No', 'No',
'No',
'No', 'No', 'No', 'Yes', 'No', 'No', 'No', 'No', 'No', 'No',
'No'],
dtype=object)

```

y_test

```

52      No
984      No
1428     No
393      No
1014     No
...
818      No
800      Yes
1397     No
423      No
792      Yes

```

Name: Attrition, Length: 272, dtype: object

Evaluation of classification model

```

#Accuracy score
from sklearn.metrics import
accuracy_score, confusion_matrix, classification_report, roc_auc_score, ro
c_curve

accuracy_score(y_test, pred)

0.9007352941176471

confusion_matrix(y_test, pred)

array([[218,    6],
       [ 21,   27]], dtype=int64)

pd.crosstab(y_test, pred)

col_0      No   Yes
Attrition

```

No	218	6
Yes	21	27

```
print(classification_report(y_test,pred))
```

	precision	recall	f1-score	support
No	0.91	0.97	0.94	224
Yes	0.82	0.56	0.67	48
accuracy			0.90	272
macro avg	0.87	0.77	0.80	272
weighted avg	0.90	0.90	0.89	272

#tunning hyperparameters using gridcv

```
from sklearn.model_selection import GridSearchCV
```

```
lg = LogisticRegression(random_state=123)
```

Create grid parameters for hyperparameter tuning

```
param_grid = {
    'C': [0.001, 0.01, 0.1, 1, 10, 100],
    'penalty': ['l1', 'l2'],
    'solver': ['liblinear', 'lbfgs']
}
```

Create gridsearch instance

```
lr = LogisticRegression()
```

```
clf = GridSearchCV(estimator=lr, param_grid=param_grid, cv=5,
scoring='accuracy', verbose=1)
```

```
clf.fit(x_train,y_train)
```

Fitting 5 folds for each of 24 candidates, totalling 120 fits

```
C:\Users\Surya\anaconda3\Lib\site-packages\sklearn\linear_model\
_logistic.py:460: ConvergenceWarning: lbfgs failed to converge
(status=1):
```

```
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
```

Increase the number of iterations (max_iter) or scale the data as shown in:

<https://scikit-learn.org/stable/modules/preprocessing.html>

Please also refer to the documentation for alternative solver options:

https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

```
n_iter_i = _check_optimize_result(
```

```
C:\Users\Surya\anaconda3\Lib\site-packages\sklearn\linear_model\
_logistic.py:460: ConvergenceWarning: lbfgs failed to converge
(status=1):
```

```
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
```

Increase the number of iterations (max_iter) or scale the data as shown in:

<https://scikit-learn.org/stable/modules/preprocessing.html>

Please also refer to the documentation for alternative solver options:

https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

```
n_iter_i = _check_optimize_result(
C:\Users\Surya\anaconda3\Lib\site-packages\sklearn\linear_model\
_logistic.py:460: ConvergenceWarning: lbfgs failed to converge
(status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
```

Increase the number of iterations (max_iter) or scale the data as shown in:

<https://scikit-learn.org/stable/modules/preprocessing.html>

Please also refer to the documentation for alternative solver options:

https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

```
n_iter_i = _check_optimize_result(
C:\Users\Surya\anaconda3\Lib\site-packages\sklearn\linear_model\
_logistic.py:460: ConvergenceWarning: lbfgs failed to converge
(status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
```

Increase the number of iterations (max_iter) or scale the data as shown in:

<https://scikit-learn.org/stable/modules/preprocessing.html>

Please also refer to the documentation for alternative solver options:

https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

```
n_iter_i = _check_optimize_result(
C:\Users\Surya\anaconda3\Lib\site-packages\sklearn\linear_model\
_logistic.py:460: ConvergenceWarning: lbfgs failed to converge
(status=1):
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
```

Increase the number of iterations (max_iter) or scale the data as shown in:

<https://scikit-learn.org/stable/modules/preprocessing.html>

Please also refer to the documentation for alternative solver options:

https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

```
n_iter_i = _check_optimize_result(
C:\Users\Surya\anaconda3\Lib\site-packages\sklearn\linear_model\
_logistic.py:460: ConvergenceWarning: lbfgs failed to converge
```

```
(status=1):  
STOP: TOTAL NO. of ITERATIONS REACHED LIMIT.
```

Increase the number of iterations (max_iter) or scale the data as shown in:

<https://scikit-learn.org/stable/modules/preprocessing.html>

Please also refer to the documentation for alternative solver options:

https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

```
n_iter_i = _check_optimize_result(  
C:\Users\Surya\anaconda3\Lib\site-packages\sklearn\model_selection\  
_validation.py:425: FitFailedWarning:
```

30 fits failed out of a total of 120.

The score on these train-test partitions for these parameters will be set to nan.

If these failures are not expected, you can try to debug them by setting error_score='raise'.

Below are more details about the failures:

30 fits failed with the following error:

Traceback (most recent call last):

```
File "C:\Users\Surya\anaconda3\Lib\site-packages\sklearn\  
model_selection\_validation.py", line 732, in _fit_and_score  
    estimator.fit(X_train, y_train, **fit_params)
```

```
File "C:\Users\Surya\anaconda3\Lib\site-packages\sklearn\base.py",  
line 1151, in wrapper
```

```
    return fit_method(estimator, *args, **kwargs)
```

```
File "C:\Users\Surya\anaconda3\Lib\site-packages\sklearn\  
linear_model\_logistic.py", line 1168, in fit
```

```
    solver = _check_solver(self.solver, self.penalty, self.dual)
```

```
File "C:\Users\Surya\anaconda3\Lib\site-packages\sklearn\  
linear_model\_logistic.py", line 56, in _check_solver
```

```
    raise ValueError(  
ValueError: Solver lbfgs supports only 'l2' or 'none' penalties, got l1 penalty.
```

```
ValueError: Solver lbfgs supports only 'l2' or 'none' penalties, got l1 penalty.
```

```
warnings.warn(some_fits_failed_message, FitFailedWarning)  
C:\Users\Surya\anaconda3\Lib\site-packages\sklearn\model_selection\  
_search.py:976: UserWarning: One or more of the test scores are non-  
finite: [0.83026114      nan 0.85609746 0.83026114 0.83026114  
nan
```

```
0.8708568 0.84870285 0.8671659      nan 0.87546083 0.87730415  
0.8754651      nan 0.87177419 0.87177419 0.87268732      nan  
0.87268732 0.87176566 0.87360898      nan 0.87360898 0.87360898]  
warnings.warn(  
nan
```



```
GridSearchCV(cv=5, estimator=LogisticRegression(),
             param_grid={'C': [0.001, 0.01, 0.1, 1, 10, 100],
                          'penalty': ['l1', 'l2'],
                          'solver': ['liblinear', 'lbfgs']},
             scoring='accuracy', verbose=1)
```

```
best_params = clf.best_params_
best_model = clf.best_estimator_
```

```
accuracy = best_model.score(x_test, y_test)
accuracy
```

```
0.8970588235294118
```

```
grid_predictions = grid.predict(x_test)
```

```
print(classification_report(y_test, grid_predictions))
```

	precision	recall	f1-score	support
No	0.84	0.98	0.91	224
Yes	0.64	0.15	0.24	48
accuracy			0.83	272
macro avg	0.74	0.56	0.57	272
weighted avg	0.81	0.83	0.79	272

```
confusion_matrix(y_test, pred)
```

```
array([[218,  6],
       [ 21, 27]], dtype=int64)
```

```
#final best accuracy
```

```
accuracy_score(y_test, pred)
```

```
0.9007352941176471
```

```
#final confusion matrix
```

```
confusion_matrix(y_test, pred)
```

```
array([[218,  6],
       [ 21, 27]], dtype=int64)
```

Model building using decision tree

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import GridSearchCV
dtc=DecisionTreeClassifier()
```

```

dtc.fit(x_train, y_train)
# print prediction results
predictions = dtc.predict(x_test)
print(classification_report(y_test, predictions))

```

	precision	recall	f1-score	support
No	0.88	0.83	0.86	224
Yes	0.37	0.46	0.41	48
<hr/>				
accuracy			0.77	272
macro avg	0.63	0.65	0.63	272
weighted avg	0.79	0.77	0.78	272

```

#accuracy score
accuracy_score(y_test,predictions)

```

```
0.7683823529411765
```

```

clf = DecisionTreeClassifier(random_state=123)
# Create grid parameters for hyperparameter tuning
params = {
    'min_samples_leaf': [1, 2, 3],
    'max_depth': [1, 2, 3]
}

```

```

# Create gridsearch instance
grid = GridSearchCV(estimator=clf,
                    param_grid=params,
                    cv=10,
                    n_jobs=1,
                    verbose=2)

```

```

# fitting the model for grid search
grid.fit(x_train, y_train)
# print best parameter after tuning
print(grid.best_params_)
grid_predictions = grid.predict(x_test)

```

```

Fitting 10 folds for each of 9 candidates, totalling 90 fits
[CV] END .....max_depth=1, min_samples_leaf=1; total
time= 0.0s
[CV] END .....max_depth=1, min_samples_leaf=1; total
time= 0.0s
[CV] END .....max_depth=1, min_samples_leaf=1; total
time= 0.0s
[CV] END .....max_depth=1, min_samples_leaf=1; total
time= 0.0s
[CV] END .....max_depth=1, min_samples_leaf=1; total
time= 0.0s

```

[illegible]

```
time= 0.0s
[CV] END .....max_depth=2, min_samples_leaf=1; total
time= 0.0s
[CV] END .....max_depth=2, min_samples_leaf=1; total
time= 0.0s
[CV] END .....max_depth=2, min_samples_leaf=1; total
time= 0.0s
[CV] END .....max_depth=2, min_samples_leaf=1; total
time= 0.0s
[CV] END .....max_depth=2, min_samples_leaf=1; total
time= 0.0s
[CV] END .....max_depth=2, min_samples_leaf=1; total
time= 0.0s
[CV] END .....max_depth=2, min_samples_leaf=1; total
time= 0.0s
[CV] END .....max_depth=2, min_samples_leaf=1; total
time= 0.0s
[CV] END .....max_depth=2, min_samples_leaf=2; total
time= 0.0s
[CV] END .....max_depth=2, min_samples_leaf=2; total
time= 0.0s
[CV] END .....max_depth=2, min_samples_leaf=2; total
time= 0.0s
[CV] END .....max_depth=2, min_samples_leaf=2; total
time= 0.0s
[CV] END .....max_depth=2, min_samples_leaf=2; total
time= 0.0s
[CV] END .....max_depth=2, min_samples_leaf=2; total
time= 0.0s
[CV] END .....max_depth=2, min_samples_leaf=2; total
time= 0.0s
[CV] END .....max_depth=2, min_samples_leaf=2; total
time= 0.0s
[CV] END .....max_depth=2, min_samples_leaf=3; total
time= 0.0s
[CV] END .....max_depth=2, min_samples_leaf=3; total
time= 0.0s
[CV] END .....max_depth=2, min_samples_leaf=3; total
time= 0.0s
```

[illegible]

[illegible]

```
# print classification report
```

```
print(classification_report(y_test, grid_predictions))
```

	precision	recall	f1-score	support
No	0.84	0.98	0.91	224
Yes	0.64	0.15	0.24	48
accuracy			0.83	272
macro avg	0.74	0.56	0.57	272
weighted avg	0.81	0.83	0.79	272

```
confusion_matrix(y_test, grid_predictions)
```

```
array([[220, 4],
       [ 41, 7]], dtype=int64)
```

As we can see that before tuning hyper parameters we got less accuracy and also the confusion matrix tp is more and fp is more (42) by using grid cv we improved the accuracy and the fp is decreased to (0). Good confusion matrix is Good model

```
#final accuracy
```

```
accuracy score(y_test, grid_predictions)
```

```
#final confusion matrix
confusion_matrix(y_test,grid_predictions)
array([[220,   4],
       [ 41,   7]], dtype=int64)
```

[illegible]

```
'No',
      'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No',
'No',
      'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No',
'Yes',
      'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No',
'No',
      'No', 'No', 'No', 'No', 'No', 'Yes', 'Yes', 'No', 'No', 'No',
'No',
      'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No',
'No',
      'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No',
'No',
      'No', 'No', 'No', 'No', 'No', 'Yes', 'Yes', 'No', 'No', 'No',
'No',
      'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No',
'No',
      'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No',
'No',
      'No', 'No', 'No', 'No', 'No', 'No', 'No', 'No'] , dtype=object)
```

```
accuracy_score(y_test,pred)
```

```
0.8382352941176471
```

```
confusion_matrix(y_test,pred)
```

```
array([[223,   1],
       [ 43,   5]], dtype=int64)
```

```
pd.crosstab(y_test,pred)
```

```
col_0      No  Yes
Attrition
No         223   1
Yes         43   5
```

```
print(classification_report(y_test,pred))
```

	precision	recall	f1-score	support
No	0.84	1.00	0.91	224
Yes	0.83	0.10	0.19	48
accuracy			0.84	272
macro avg	0.84	0.55	0.55	272
weighted avg	0.84	0.84	0.78	272

```
from sklearn.model_selection import GridSearchCV
rfc = RandomForestClassifier(random_state=123)
# Create grid parameters for hyperparameter tuning
```



```

params = {
    'max_depth':[3,5,10,None],
    'n_estimators':[10,100,200],
    'max_features':[1,3,5,7],
    'min_samples_leaf': [1, 2, 3],
    'max_depth': [1, 2, 3]
}

# Create gridsearch instance
grid = GridSearchCV(estimator=rfc,
                    param_grid=params,
                    cv=10,
                    n_jobs=1,
                    verbose=2)
# fitting the model for grid search
grid.fit(x_train, y_train)
# print best parameter after tuning
print(grid.best_params_)
grid_predictions = grid.predict(x_test)

```

Fitting 10 folds for each of 108 candidates, totalling 1080 fits

```

[CV] END max_depth=1, max_features=1, min_samples_leaf=1,
n_estimators=10; total time= 0.0s
[CV] END max_depth=1, max_features=1, min_samples_leaf=1,
n_estimators=10; total time= 0.0s
[CV] END max_depth=1, max_features=1, min_samples_leaf=1,
n_estimators=10; total time= 0.0s
[CV] END max_depth=1, max_features=1, min_samples_leaf=1,
n_estimators=10; total time= 0.0s
[CV] END max_depth=1, max_features=1, min_samples_leaf=1,
n_estimators=10; total time= 0.0s
[CV] END max_depth=1, max_features=1, min_samples_leaf=1,
n_estimators=10; total time= 0.0s
[CV] END max_depth=1, max_features=1, min_samples_leaf=1,
n_estimators=10; total time= 0.0s
[CV] END max_depth=1, max_features=1, min_samples_leaf=1,
n_estimators=10; total time= 0.0s
[CV] END max_depth=1, max_features=1, min_samples_leaf=1,
n_estimators=10; total time= 0.0s
[CV] END max_depth=1, max_features=1, min_samples_leaf=1,
n_estimators=100; total time= 0.2s
[CV] END max_depth=1, max_features=1, min_samples_leaf=1,
n_estimators=100; total time= 0.2s
[CV] END max_depth=1, max_features=1, min_samples_leaf=1,
n_estimators=100; total time= 0.2s
[CV] END max_depth=1, max_features=1, min_samples_leaf=1,
n_estimators=100; total time= 0.3s
[CV] END max_depth=1, max_features=1, min_samples_leaf=1,

```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

```
[CV] END max_depth=3, max_features=7, min_samples_leaf=3,
n_estimators=200; total time= 0.9s
[CV] END max_depth=3, max_features=7, min_samples_leaf=3,
n_estimators=200; total time= 1.0s
[CV] END max_depth=3, max_features=7, min_samples_leaf=3,
n_estimators=200; total time= 1.0s
[CV] END max_depth=3, max_features=7, min_samples_leaf=3,
n_estimators=200; total time= 1.0s
[CV] END max_depth=3, max_features=7, min_samples_leaf=3,
n_estimators=200; total time= 0.9s
[CV] END max_depth=3, max_features=7, min_samples_leaf=3,
n_estimators=200; total time= 0.9s
[CV] END max_depth=3, max_features=7, min_samples_leaf=3,
n_estimators=200; total time= 0.9s
[CV] END max_depth=3, max_features=7, min_samples_leaf=3,
n_estimators=200; total time= 0.9s
[CV] END max_depth=3, max_features=7, min_samples_leaf=3,
n_estimators=200; total time= 0.9s
[CV] END max_depth=3, max_features=7, min_samples_leaf=3,
n_estimators=200; total time= 0.9s
{'max_depth': 3, 'max_features': 7, 'min_samples_leaf': 1,
'n_estimators': 100}
```

```
print('Best score is: '+str(grid.best_score_))
```

```
Best score is: 0.8404349303431872
```

```
# print classification report
```

```
print(classification_report(y_test, grid_predictions))
```

	precision	recall	f1-score	support
No	0.82	1.00	0.90	224
Yes	0.00	0.00	0.00	48
accuracy			0.82	272
macro avg	0.41	0.50	0.45	272
weighted avg	0.68	0.82	0.74	272

```
C:\Users\Surya\anaconda3\Lib\site-packages\sklearn\metrics\
_classification.py:1469: UndefinedMetricWarning: Precision and F-score
are ill-defined and being set to 0.0 in labels with no predicted
samples. Use `zero_division` parameter to control this behavior.
```

```
_warn_prf(average, modifier, msg_start, len(result))
```

```
C:\Users\Surya\anaconda3\Lib\site-packages\sklearn\metrics\
_classification.py:1469: UndefinedMetricWarning: Precision and F-score
are ill-defined and being set to 0.0 in labels with no predicted
samples. Use `zero_division` parameter to control this behavior.
```

```
_warn_prf(average, modifier, msg_start, len(result))
```

```
C:\Users\Surya\anaconda3\Lib\site-packages\sklearn\metrics\
```

```
_classification.py:1469: UndefinedMetricWarning: Precision and F-score  
are ill-defined and being set to 0.0 in labels with no predicted  
samples. Use `zero_division` parameter to control this behavior.  
_warn_prf(average, modifier, msg_start, len(result))
```

```
confusion_matrix(y_test, grid_predictions)
```

```
array([[224,  0],  
       [ 48,  0]], dtype=int64)
```

```
#final confusion matrix
```

```
confusion_matrix(y_test, pred)
```

```
array([[223,  1],  
       [ 43,  5]], dtype=int64)
```

```
#final better accuracy
```

```
accuracy_score(y_test, pred)
```

```
0.8382352941176471
```