Project 2:

Build Decision Tree for Attrition Rate Analysis

DV - "Attrition"

IDV - Output of RF Algorithm

**import** **pandas** **as** **pd**

**import** **numpy** **as** **np**

**from** **sklearn** **import** preprocessing

**from** **sklearn** **import** tree

**import** **sklearn** **as** **sk**

**from** **sklearn.ensemble** **import** RandomForestClassifier

att=pd.read\_csv("general\_data.csv")

att.isna().sum().sum

Out[48]:

<bound method Series.sum of Age 0

Attrition 0

BusinessTravel 0

Department 0

DistanceFromHome 0

Education 0

EducationField 0

EmployeeCount 0

EmployeeID 0

Gender 0

JobLevel 0

JobRole 0

MaritalStatus 0

MonthlyIncome 0

NumCompaniesWorked 19

Over18 0

PercentSalaryHike 0

StandardHours 0

StockOptionLevel 0

TotalWorkingYears 9

TrainingTimesLastYear 0

YearsAtCompany 0

YearsSinceLastPromotion 0

YearsWithCurrManager 0

dtype: int64>

att .isnull().sum()

Out[49]:

ID 0

Age 0

Experience 0

Income 0

ZIP Code 0

Family 0

CCAvg 0

Education 0

Mortgage 0

Personal Loan 0

Securities Account 0

CD Account 0

Online 0

CreditCard 0

dtype: int64

new\_noofcom\_var=np.where(att["NumCompaniesWorked"].isnull(),1,att["NumCompaniesWorked"])

new\_ToWYar\_var=np.where(att["TotalWorkingYears"].isnull(),11,att["TotalWorkingYears"])

att["TotalWorkingYears"]=new\_ToWYar\_var

att["NumCompaniesWorked"]=new\_noofcom\_var

from sklearn.ensemble import RandomForestClassifier

att.columns

Out[55]:

Index(['Age', 'Attrition', 'BusinessTravel', 'Department', 'DistanceFromHome',

'Education', 'EducationField', 'EmployeeCount', 'EmployeeID', 'Gender',

'JobLevel', 'JobRole', 'MaritalStatus', 'MonthlyIncome',

'NumCompaniesWorked', 'Over18', 'PercentSalaryHike', 'StandardHours',

'StockOptionLevel', 'TotalWorkingYears', 'TrainingTimesLastYear',

'YearsAtCompany', 'YearsSinceLastPromotion', 'YearsWithCurrManager'],

dtype='object')

rf\_model=RandomForestClassifier(n\_estimators=1000,max\_features=2,oob\_score=True)

features=["DistanceFromHome","Age","DistanceFromHome","Education","JobLevel","MonthlyIncome","NumCompaniesWorked","PercentSalaryHike","TotalWorkingYears","TrainingTimesLastYear","YearsAtCompany","YearsSinceLastPromotion","YearsWithCurrManager"]

rf\_model.fit(X=att[features],y=att["Attrition"])

Out[58]:

RandomForestClassifier(bootstrap=True, ccp\_alpha=0.0, class\_weight=None,

criterion='gini', max\_depth=None, max\_features=2,

max\_leaf\_nodes=None, max\_samples=None,

min\_impurity\_decrease=0.0, min\_impurity\_split=None,

min\_samples\_leaf=1, min\_samples\_split=2,

min\_weight\_fraction\_leaf=0.0, n\_estimators=1000,

n\_jobs=None, oob\_score=True, random\_state=None,

verbose=0, warm\_start=False)

print("OOB Accuracy:",rf\_model.oob\_score\_)

OOB Accuracy: 0.9993197278911564

**THE OOB Accuracy IS 99.93%**

for feature,imp in zip(features,rf\_model.feature\_importances\_):

print(feature,imp);

DistanceFromHome 0.07615177083120446

Age 0.12339948009256356

DistanceFromHome 0.07671564430513496

Education 0.05150530306271761

JobLevel 0.04789230886058135

MonthlyIncome 0.1325118245969567

NumCompaniesWorked 0.06624864915938267

PercentSalaryHike 0.08269692434601451

TotalWorkingYears 0.10212754001065424

TrainingTimesLastYear 0.05617963526862585

YearsAtCompany 0.0753382923765439

YearsSinceLastPromotion 0.048339266258824426

YearsWithCurrManager 0.06089336083079567

**Found the more important variables for decision tree**

1. **MonthlyIncome**
2. **TotalWorkingYears**
3. **Age**

tree\_model=tree.DecisionTreeClassifier(max\_depth=3)

predictors=pd.DataFrame([att["Age"],att["MonthlyIncome"],att["TotalWorkingYears"]]).T

tree\_model.fit(X=predictors,y=att["Attrition"])

Out[63]:

DecisionTreeClassifier(ccp\_alpha=0.0, class\_weight=None, criterion='gini',

max\_depth=3, max\_features=None, max\_leaf\_nodes=None,

min\_impurity\_decrease=0.0, min\_impurity\_split=None,

min\_samples\_leaf=1, min\_samples\_split=2,

min\_weight\_fraction\_leaf=0.0, presort='deprecated',

random\_state=None, splitter='best')

with open("DtreeATT.dot",'w') as f:

f=tree.export\_graphviz(tree\_model,feature\_names=["Age","MonthlyIncome","TotalWorkingYears"],out\_file=f);