**HR ATTRITION**

**A PROJECT REPORT SUBMITTED BY**

**TEAM: TECH SARADHI**

**Y Lakshmi JyothiSravani**

**K GowthamiLatha**

**S PrasannaDurga**

**V Bhuvana**

**T Khyathi**

**CONTENT PAGE**

**1.Title of the project Page no**

1.1 Introduction 3

1.2 Objectives of research 3

1.3 Problem Statement 4

1.4 Industry Profile 4

**2.Review of Literature** 4

**3.Data collection** 4-5

**4.Methodology**

4.1 Exploratory Data analysis 5-6

4.1.1 Figures and tables 6-7

4.2 Data Visualization 7-10

4.3 Data Modelling using Supervised ML techniques 10-13

**5.Findings and Suggestions** 13

**6.Conclusion** 13

**7.Refereneces** 13

**HR ATTRITION**

**1.1Introduction**

**Definition:**

Attrition in Human Resource terminology refers to the phenomenon of the employees leaving the company. Attrition in a company is usually measured with a metric called attrition rate, which simply measures the no of employees moving out of the company (voluntary resigning or laidoff by the company).

**Reasons for attrition:**

* Low pay & No recognition for employees
* Lack of career growth
* Retirement
* Poor working conditions
* Terminating inefficient employees
* Impolite behaviour of manager and peers & ineffective team management
* Lack of employee recognition or limited growth in current role.
* Stagnancy in career growth & poor quality of work life can lead to higher attrition rate.

**Types of attrition:**

**1.Voluntary attrition:**

 When an employee leaves the company for a better job opportunity or career growth or more pay, and leaves on his own, it is known as voluntary attrition.

**2.Involuntary attrition:**

If an employee is terminated from a job due to some ethical issue or lack or performance, it is involuntary attrition.

**3. Retirement:**

Once an employee finishes his/her tenure at a company and retires, it is also counted as attrition. This is mostly a natural attrition that occurs and companies are prepared with succession planning.

**1.2 Objectives Of Research:**

* To fulfill the future needs and aspirations of employees in the organization.
* To bring out proper working between employees and organization to improve the company image.
* To take efforts and measures to improve the situation and conditions of the employment.
* To identify the factors which make employees dissatisfy.

**1.3 Problem Statement:**

The aim of this report is to study factors like salary, superior-subordinate relationship, growth opportunities ,facilities , policies and procedures , recognition, appreciation ,suggestions, co-workers by which it helps to know the attrition level in the organizations and factors relating to retain them .This study also help to find out where the organizations are lagging in retaining.

**1.4 Industry Profile:**

The attrition level of the employees indicates that something is wrong with the health and climate of an organization in the terms of wages, working conditions, industrial relations, welfare facilities provided by employer to the employees. Hence employee attrition is a challenging issue in the business world.

**2.Review of Literature:**

**Attriton:**Attrition in Human Resource terminology refers to the phenomenon of the employees leaving the company. Attrition in a company is usually measured with a metric called attrition rate, which simply measures the no of employees moving out of the company.

**Positive Attrition:** Positive attrition results when the loss and replacemet of an employee is better for the organization.

**Negative Attrition:**When productive employee resigns causing a loss of talent and skills within the company. Attrition is not always negative.

When ever attrition is occur then its outcome will cause loss to the company when employee leave the company. Hiring new employees will cause to wastage of time and money to provide training for them.

Actually attrition(employees leave the company) will be caused due to the lack of career growth, poor working conditions, low pay, retirement.

**3 Data Collection:**

* The HR Attrition dataset must be in csv format. This dataset consists of 4411 rows and 24 columns/attributes from IBM cloud.
* Some of the attributes are emp\_id, attrition, distance from home, education, education level, job role, monthly income.
* Attributes which are not related to our dataset are number companies worked, person salary hike, standard hours, stock option level.
* Some categorical variables are encoded with numeric values such variables are called dummy variables.
* For example,

**Attribute: Attrition** – yes is labeled as 1 and no is labeled as 0, by usinglabel encoder.

**Attribute: Gender**- female is labeled as 1 and male is labeled as 0.

**Attribute:Business Travel-**Travel Rarely as 2and Travel Frequently as 1and Non Travel as 0.

**Attribute:Department –** Sales as 2 and Research and Development as 1 and Human Resource as 0.

**Attribute: Education Field-**Life Sciences as 1 and Medical as 3 and others as 4 and marketing as 2.

**Attribute: Marital Status-**Married as 1 and single as 2 and divorced as 3.

**4. Methodology:**

* We shall be looking at all variables through some plots and infer about it in our exploratory analysis.
* Through our analysis we intend to build a model which can predict if any employee is about to quit. Here, we use logistic regression model which is a part of a large scale of algorithms known as generalized linear model.

**4.1 Exploratory Data Analysis:**

First we need to check the data types of the features because we can only see the distribution of the numeric/continuous values in the dataset.

In case of categorical variables we need to replace them with the numeric values by using dummy variables.

Our aim is to predict the employee attrition and it is important to see which variables are contributing the most in attrition. But before that we need to know if the variables are correlated if they are, we might want to avoid those in model building process.

#check the structure of dataset

dataset.dtypes()

There are many continuous variables we can have a look at their distribution and create a grid of pair plots but that would be too much code to see the correlation as there are lot variables.So, we can create numeric variables and see the correlation.

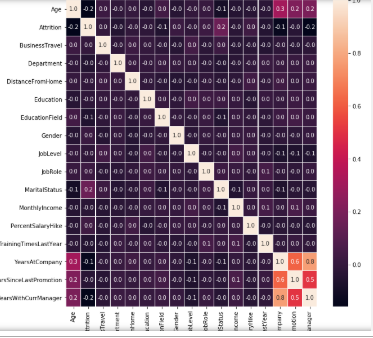
Variables which are not poorly correlated i.e correlation value tend towards 0 and leave the ones which are strongly correlated (i.e correlation value is 1)

#correlation among variables

f,ax=plt.subplots(figsize=(10,10))

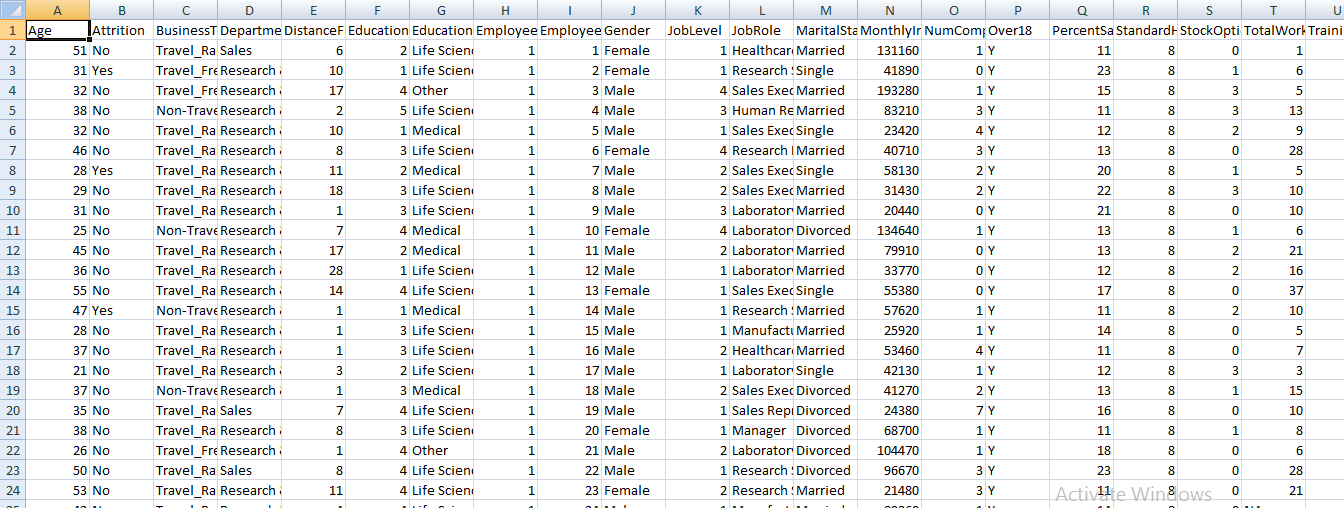
sns.heatmap(dataset.corr(),annot=True,linewidths=0.2,fmt='.1f',ax=ax)

plt.show()



**4.1.1 Figures and Tables:**

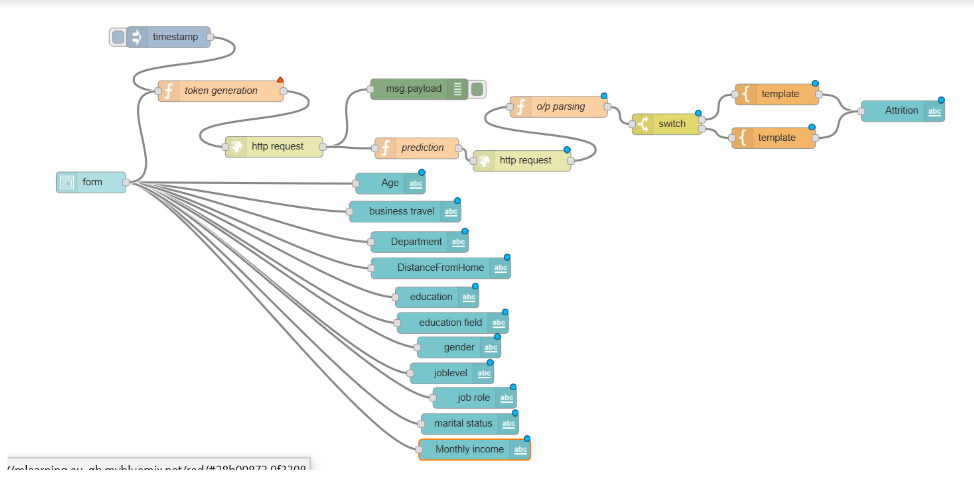
Dataset of HR Attrition consists of 24 attributes and 4411 rows. By fixing attrition as a target variable which is dependent on other attributes.



Here we considered attrition attribute as target variable and it is depended on other attributes like age, distance from home, education, gender, marital status, monthly income etc.

**Node Red Flow:**

Node Red is a programming tool for wiring together hardware devices, APIs and online services in new and interesting ways. It provides a browser-based editor that makes it easy to wire together flows using the wide range nodes in the palette that can be deployed to its runtime in a single-click.

****

**4.2 Data Visualization:**

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.

In the world of Big Data, data visualization tools and technologies are essential to analyze massive amounts of information and make data-driven decisions.

Data visualization is another form of visual art that grabs our interest and keeps our eyes on the message. When we see a chart, we quickly see trends and outliers. If we can see something, we internalize it quickly. It’s storytelling with a purpose.

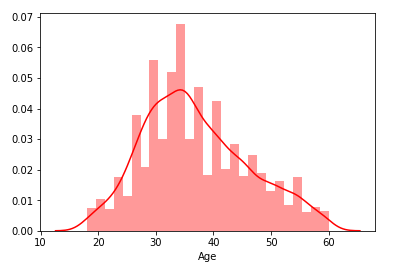
**Distplot:**

Histograms are likely familiar and a histfunction already exists in matplotlib. distplot function combines with the matplotlibhist function with automatic calculation of a good default bin size.

#distplot visualization

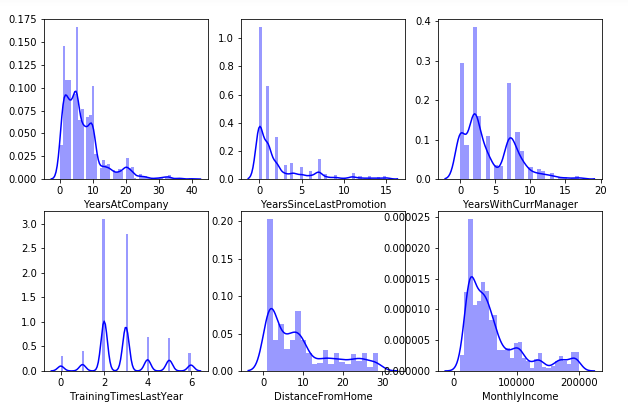
sns.distplot(dataset['Age'],color='red')

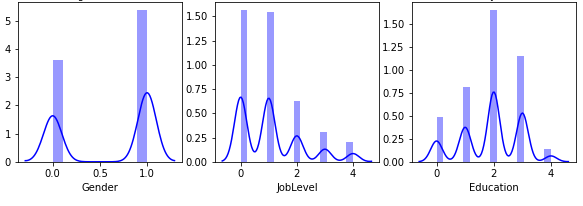
plt.show()



Similarly distplotfunction for other attributes are as follows:

#distplot visualization





Above figures will illustrates graph between different attributes like, gender, job level, education,

Distance from home, monthly income etc.

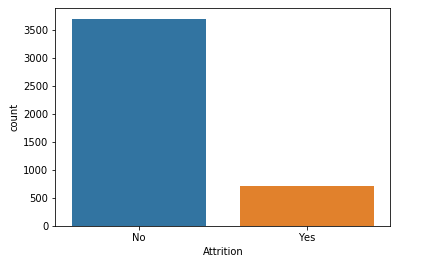
**Countplot:**

Bar graphs are useful for displaying relationships between categorical data and at least one numerical variable. countplot is a barplotwhere the dependent variable is the number of instances of each instance of the independent variable.

#countplot visualization

sns.countplot(dataset['Attrition'])

plt.show()



**Barplot :**

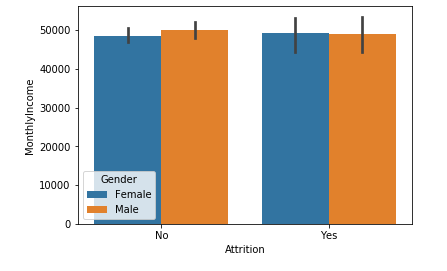
A barplot or barchart is one of most common type of graphic. It shows the relationship between categorical variable and a numeric variable.

#barplot visualization

fromnumpy import median

sns.barplot(x='Attrition',y='MonthlyIncome',hue='Gender',data=dataset,estimator=median)

plt.show()



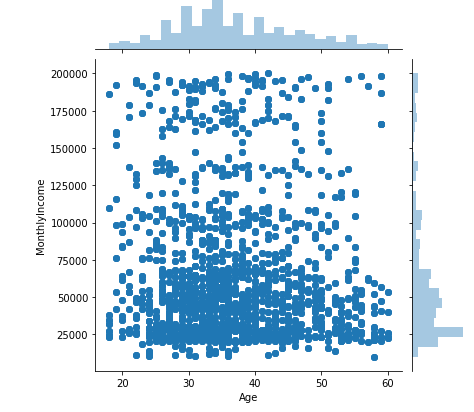
**Scatterplot:**

Scatterplot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data.

#scatterplot visualization

sns.jointplot(dataset.Age,dataset.MonthlyIncome,kind="scatter")

plt.show()



**4.3Data modeling using supervised ML techniques:**

Supervised learning is where you have input variables (x) and an output variable (Y) and you use an algorithm to learn the mapping function from the input to the output.

Y=f(X)

The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

It is called supervised learning because the process of analgorithm learning from the training dataset can be thought of as a teacher supervising the learning process.

Supervised learning problems can be further grouped into regression and classification problems.

**Classification**: A classification problem is when the output variable is a category, such as “red” or “blue” or “disease” and “no disease”.

**Regression**: A regression problem is when the output variable is a real value, such as “dollars” or “weight”.

By comparing all the available learning algorithms we prefer logistic regression because it is more accurate compare to other algorithms.

**Logistic Regression:**

Logistic function also called the sigmoid function was developed by statisticians to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment. It’s an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits.

We should want to convert all categorical variables into numeric variables by using label encoder.

fromsklearn import preprocessing

label\_encoder=preprocessing.LabelEncoder()

colnames\_to\_encode=['Attrition','BusinessTravel','Department','Education','EducationField','Gender','JobLevel','JobRole','MaritalStatus']

for c in colnames\_to\_encode:

dataset[c]=label\_encoder.fit\_transform(dataset[c])

dataset.Attrition.value\_counts()

After running the code the output is as follows:

0 3699

1 711

Name: Attrition, dtype: int64

Performing logistic regression by using the code

fromsklearn.metrics import confusion\_matrix

cm = confusion\_matrix(y\_test, y\_pred)

print(cm)

Confusion matrix as follows:

[[1099 0]

[ 224 0]]

fromsklearn.metrics import roc\_auc\_score

froms+klearn.metrics import roc\_curve

logit\_roc\_auc=roc\_auc\_score(y\_test,logreg.predict(x\_test))

fpr,tpr,thresholds=roc\_curve(y\_test,logreg.predict\_proba(x\_test)[:,1])

plt.figure()

plt.plot(fpr,tpr,label='Logistic Regression(area=%0.2f)'%logit\_roc\_auc)

plt.plot([0,1],[0,1],'r--')

plt.xlim([0.0,1.0])

plt.ylim([0.0,1.05])

plt.xlabel('False positive rate')

plt.ylabel('True positive rate')

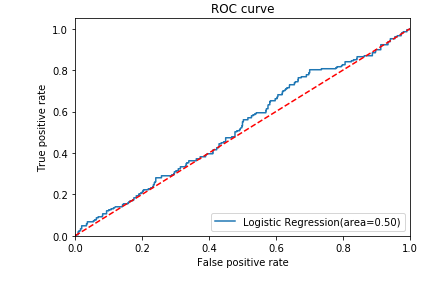
plt.title('ROC curve')

plt.legend(loc="lower right")

plt.savefig('loc\_ROC')

plt.show()

ROC curve as follows:(received operation characteristic curve)



fromsklearn import metrics

print("Accuracy:",metrics.accuracy\_score(y\_test,y\_pred)

Accuracy: 0.8435374149659864

fromsklearn.metrics import average\_precision\_score

average\_precision=average\_precision\_score(y\_test,y\_pred)

print('Average precision-recall score:{0:0.2f}'.format(average\_precision))

Average precision-recall score:0.17

**5 Findings and Suggestions:**

* Employees performance can be recognized as most of the employees expect it from the organization to render apart from organization.
* Compensation package can be revised as it was considered as the important reason for attrition.
* Management should be able to find the people who have intention to leave the organization and has to enquire about the reason.
* Management should help the employees to overcome their personal barriers helping him to perform well in his work.
* Management should consider employees valuable suggestions.

**6 Conclusion:**

The main objective of the organization is to earn profit and the employer should concentrate in retaining talents and concentrate in making them stick to the organization for the long run. Employees are the assets of the organization. Hence it is important for the employers to minimize the attrition rate and help in both individual as well as organizational growth. Thus, Organizations should create an environment that fosters ample growth opportunities, appreciation for the work accomplished and a friendly cooperative atmosphere that makes an employee feel connected in every respect to the organization. Retention plans are an inexpensive way of enhancing workplace productivity and engaging employees emotionally.

**7 References:**

An introduction to hr attrition and its analysis and reporting

<https://www.google.com/search?q=hr+attrition&rlz=1C1CHBD_enIN793IN796&oq=hr+attrition&aqs=chrome..69i57j0l5.5014j0j7&sourceid=chrome&ie=UTF-8>