
# HR DATA ANALYTICS

NIVT

Mentor: Tania Chakraborty

SUMMER TRAINING KOLKATA

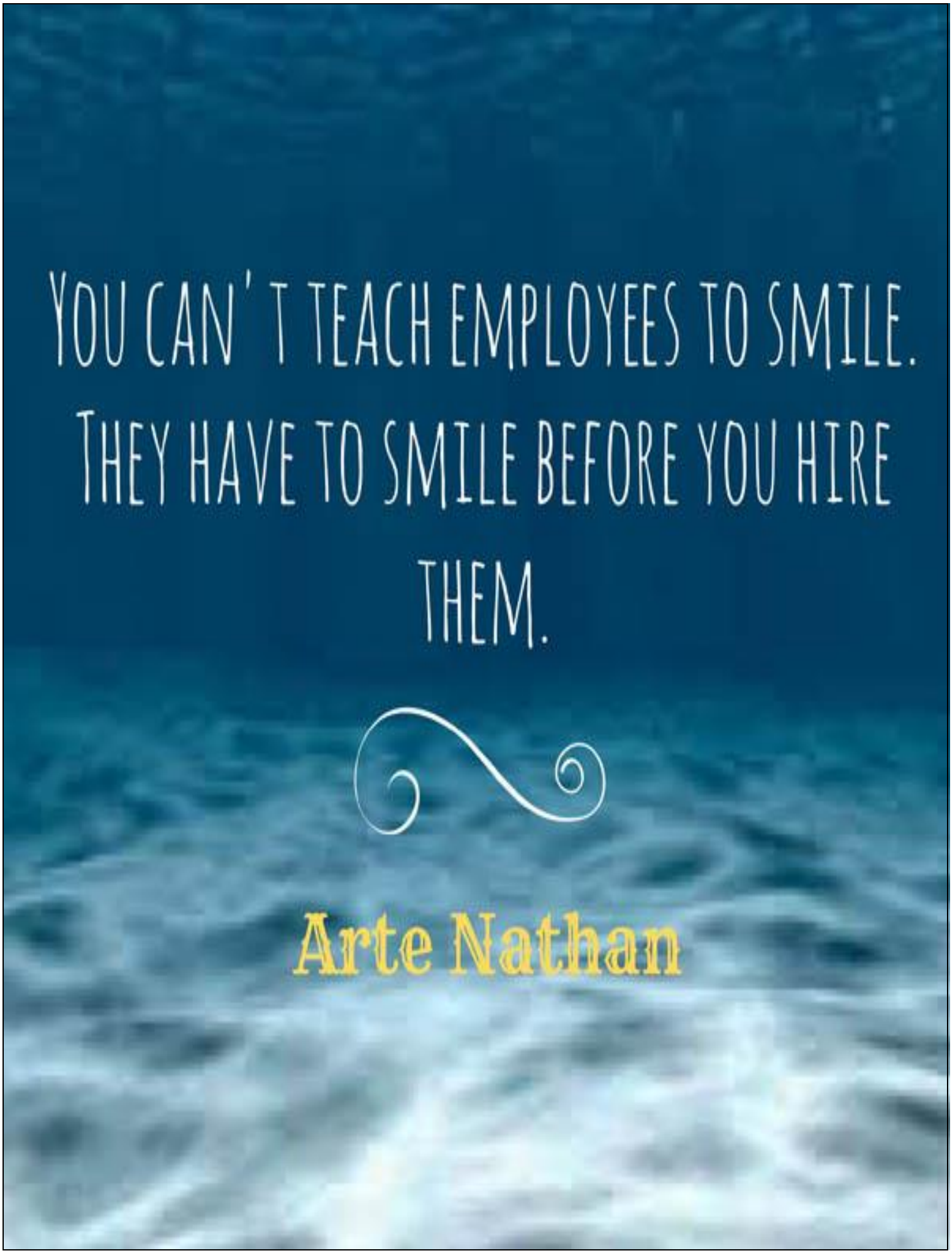
# Group

- Prasenjit Karmakar
- Sabyasachi Roy
- Sagar Roy
- Yasar Reza Rahaman

# ACKNOWLEDGEMENT

We would like to express my special thanks of gratitude to my teacher Tania Chakraborty as well as our institute NIVT who gave us the golden opportunity to do this wonderful project on the topic HR DATA ANALYTICS which also helped us to learn so many new things. I am really thankful to them.
Secondly, I would also like to thank our friends & group members who helped us a lot in finalizing this project within the limited time frame.

# CONTENTS


You can't teach employees to smile. They have to smile before you hire them.
Arte Nathan

# INTRODUCTION

Human Resource Management (HRM) is an operation in companies designed to maximize employee performance in order to meet the employer's strategic goals and objectives. More precisely, HRM focuses on management of people within companies, emphasizing on policies and systems.

In short, HRM is the process of recruiting, selecting employees, providing proper orientation and induction, imparting proper training and developing skills.

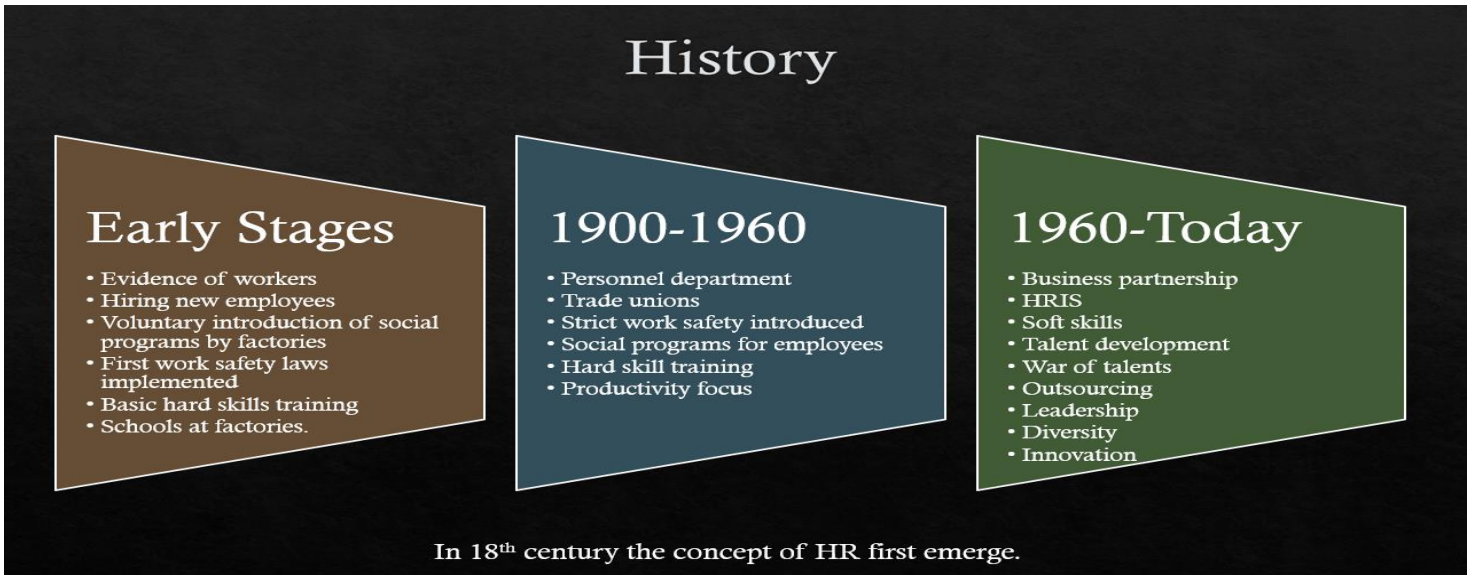
HRM also includes employee assessment like performance appraisal, facilitating proper compensation and benefits, encouragement, maintaining proper relations with labor and with trade unions, and taking care of employee safety, welfare and health by complying with labor laws of the state or country concerned.

# HISTORY OF HUMAN RESOURCE MANAGEMENT

Many people may think of a human resources department as a relatively modern innovation. However, a look at the history of human resources reveals that the ideas underpinning the discipline stretch as far back as human history itself. Maximizing worker potential and management of people is a concern stretching back to ancient times. Those ideas were further developed starting in the 18th century, culminating in today's human resources departments.

In the early 20th century, the near-simultaneous rise of trade unions and personnel management departments within companies laid the groundwork for the formal discipline of human resources. The ideas of mechanical engineer Frederick Taylor that explored how to make manufacturing workers more efficient were important underpinnings for the discipline's development. Unions were important in pressing for employee rights alongside that increased efficiency, and these two principles have continued to develop in tandem as crucial elements in the history of human resources.

According to Fast Company, The National Cash Register Company may have been the first modern human resources department. Although at the time it was called "personnel," its focus on managing wages and workplace safety as well as dealing with employee grievances meant its aims were similar to human resources departments today.

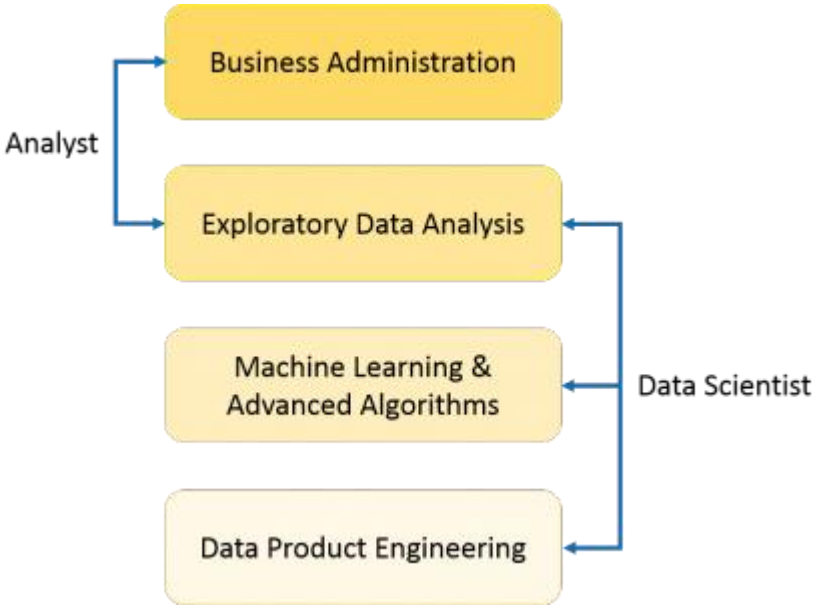


# WHAT IS DATA SCIENCE?

To manipulate data and extract the important part of data is called Data Science.

Data science is a multidisciplinary blend of **data inference, algorithm development, and technology** in order to solve analytically complex problems.

At the core is data. Troves of raw information, streaming in and stored in enterprise data warehouses. Much to learn by mining it. Advanced capabilities we can build with it.



# MEANING OF DATA SCIENCE

Data science is the field of study that combines domain expertise, programming skills, and knowledge of math and statistics to extract meaningful insights from data. Data science practitioners apply machine learning algorithms to numbers, text, images, video, audio, and more to produce artificial intelligence (AI) systems that perform tasks which ordinarily require human intelligence. In turn, these systems generate insights that analysts and business users translate into tangible business value.

# DEFINE DATA SCIENCE COMPONENTS

Basically, here three component of data science-

- Data Management
- Data Analytics
- Machine Learning

**Data Management:**

**Data Management** is a comprehensive collection of practices, concepts, procedures, processes, and a wide range of accompanying systems that allow for an organization to gain control of its **data** resources.

- **Data** Storage and Big **Data**.
- Business Intelligence and Analytics.
- Metadata **Management**.

**Data Analytics:**

Data analytics is the science of analyzing raw data in order to make conclusions about that information. Many of the techniques and processes of data analytics have been automated into mechanical processes and algorithms that work over raw data for human consumption.

**Machine Learning:**

Machine Learning (ML) is basically that field of computer science with the help of which computer systems can provide sense to data in much the same way as human beings do. In simple words, ML is a type of artificial intelligence that extract patterns out of raw data by using an algorithm or method. The key focus of ML is to allow computer systems to learn from experience without being explicitly programmed or human intervention.

# DATA

Data is a set of values of qualitive or quantitative variables. It is information in raw or unorganized form. It may be fact, figure, character symbols etc.

- Qualitative -Categorical or Nominal:
  - Discrete Data
  - Continuous Data
- Quantitative -Measurable or Countable:
  - Attribute
  - Nominal
  - Ordinal

Information:

Meaningful or organized data is information.

Big Data:

Extremely large data sets that may be analyzed computationally to reveal patterns, trends and associations, especially relating to human behavior and interactions.

Design distributed systems that manage big data using HADOOP and related technologies.

Big data cannot manage by RDBMS

# TYPES OF DATA

There are three types of dataset-

1. STRUCTURED DATA
2. SEMI-STRUCTURED DATA
3. UNSTRUCTURED DATA

**Structured Data:**

Data which can be stored in database SQL in table with rows and columns are called structured data.

Only 5 to 10 % of all informatics data.

| | A | B | C | D | E | F |
|---|---|---|---|---|---|---|
| 1 | Country | Salesperson | Order Date | OrderID | Units | Order Amoun |
| 2 | USA | Fuller | 1/01/2011 | 10392 | 13 | 1,440.00 |
| 3 | UK | Gloucester | 2/01/2011 | 10397 | 17 | 716.72 |
| 4 | UK | Bromley | 2/01/2011 | 10771 | 18 | 344.00 |
| 5 | USA | Finchley | 3/01/2011 | 10393 | 16 | 2,556.95 |
| 6 | USA | Finchley | 3/01/2011 | 10394 | 10 | 442.00 |
| 7 | UK | Gillingham | 3/01/2011 | 10395 | 9 | 2,122.92 |
| 8 | USA | Finchley | 6/01/2011 | 10396 | 7 | 1,903.80 |
| 9 | USA | Callahan | 8/01/2011 | 10399 | 17 | 1,765.60 |
| 10 | USA | Fuller | 8/01/2011 | 10404 | 7 | 1,591.25 |
| 11 | USA | Fuller | 9/01/2011 | 10398 | 11 | 2,505.60 |
| 12 | USA | Coghill | 9/01/2011 | 10403 | 18 | 855.01 |
| 13 | USA | Finchley | 10/01/2011 | 10401 | 7 | 3,868.60 |
| 14 | USA | Callahan | 10/01/2011 | 10402 | 11 | 2,713.50 |
| 15 | UK | Rayleigh | 13/01/2011 | 10406 | 15 | 1,830.78 |
| 16 | USA | Callahan | 14/01/2011 | 10408 | 10 | 1,622.40 |
| 17 | USA | Farnham | 14/01/2011 | 10409 | 19 | 319.20 |
| 18 | USA | Farnham | 15/01/2011 | 10410 | 16 | 802.00 |

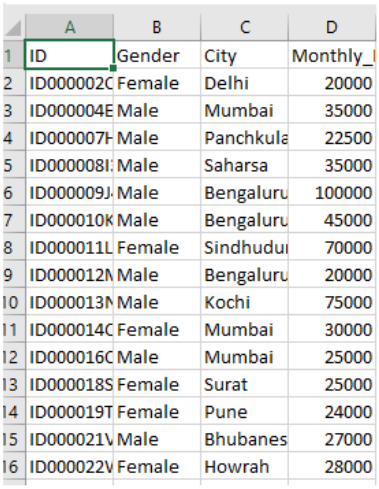
**Semi-Structured Data:**

Doesn't reside in a relational database but that does have some organizational properties that make it easier to analyze.

CSV,XML AND JSON documents are semi-structured documents, NoSQL databases are considered as semi-structured.

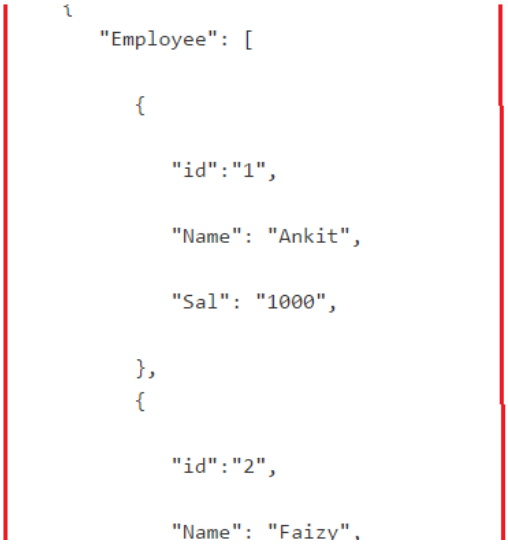
A few parts of data(5 to 10 %).

**CSV**



| | A | B | C | D |
|---|---|---|---|---|
| 1 | ID | Gender | City | Monthly_I |
| 2 | ID000002C | Female | Delhi | 20000 |
| 3 | ID000004E | Male | Mumbai | 35000 |
| 4 | ID000007H | Male | Panchkula | 22500 |
| 5 | ID000008I | Male | Saharsa | 35000 |
| 6 | ID000009J | Male | Bengaluru | 100000 |
| 7 | ID000010K | Male | Bengaluru | 45000 |
| 8 | ID000011L | Female | Sindhudui | 70000 |
| 9 | ID000012M | Male | Bengaluru | 20000 |
| 10 | ID000013N | Male | Kochi | 75000 |
| 11 | ID000014C | Female | Mumbai | 30000 |
| 12 | ID000016C | Male | Mumbai | 25000 |
| 13 | ID000018S | Female | Surat | 25000 |
| 14 | ID000019T | Female | Pune | 24000 |
| 15 | ID000021V | Male | Bhubanes | 27000 |
| 16 | ID000022V | Female | Howrah | 28000 |

**JSON**



```
{
   "Employee": [
      {
         "id":"1",
         "Name": "Ankit",
         "Sal": "1000",
      },
      {
         "id":"2",
         "Name": "Faizy",
```

```
<?xml version="1.0"?>
<contact-info>
<name>Ankit</name>
<company>Anlytics Vidhya</company>
<phone>+9187654321</phone>
</contact-info>
```

## Unstructured Data:

Unstructured data represent around 80% of data. It often includes text and multimedia content.

Example: e-mail messages, word processing documents, videos, photos, audio files, presentation, webpages and many other kinds of business documents.

Some example of machine-generated unstructured data :

- ✓ Satellite images
- ✓ Scientific data
- ✓ Photographs and video
- ✓ Rader or Sonar data

## Some example of human-generated unstructured data:

- Text internal
- Social media data
- Mobile data
- Website content

```
weblogic.application.utils.StateMachineDriver.nextState(StateMachineDriver.java:26)
>
####<Dec 29, 2006 2:14:24 PM IST> <Notice> <Log Management> <svaidyan02> <xbusServer>
<[ACTIVE] ExecuteThread: '0' for queue: 'weblogic.kernel.Default (self-tuning)'> <<WLS
Kernel>> <> <> <1167381864275> <BEA-170027> <The server initialized the domain log
broadcaster successfully. Log messages will now be broadcasted to the domain log.>
####<Dec 29, 2006 2:14:24 PM IST> <Notice> <WebLogicServer> <svaidyan02> <xbusServer> <Main
Thread> <<WLS Kernel>> <> <> <1167381864976> <BEA-000365> <Server state changed to ADMIN>
####<Dec 29, 2006 2:14:24 PM IST> <Notice> <WebLogicServer> <svaidyan02> <xbusServer> <Main
Thread> <<WLS Kernel>> <> <> <1167381864996> <BEA-000365> <Server state changed to RESUMING>
####<Dec 29, 2006 2:14:28 PM IST> <Notice> <Security> <svaidyan02> <xbusServer> <[STANDBY]
ExecuteThread: '5' for queue: 'weblogic.kernel.Default (self-tuning)'> <<WLS Kernel>> <> <>
<1167381868541> <BEA-090171> <Loading the identity certificate and private key stored under
the alias DemoIdentity from the jks keystore file
C:\bea2613a\WEBLOG~1\server\lib\DemoIdentity.jks.>
####<Dec 29, 2006 2:14:29 PM IST> <Notice> <Security> <svaidyan02> <xbusServer> <[STANDBY]
ExecuteThread: '5' for queue: 'weblogic.kernel.Default (self-tuning)'> <<WLS Kernel>> <> <>
<1167381869643> <BEA-090169> <Loading trusted certificates from the jks keystore file
C:\bea2613a\WEBLOG~1\server\lib\DemoTrust.jks.>
####<Dec 29, 2006 2:14:29 PM IST> <Notice> <Security> <svaidyan02> <xbusServer> <[STANDBY]
ExecuteThread: '5' for queue: 'weblogic.kernel.Default (self-tuning)'> <<WLS Kernel>> <> <>
<1167381869713> <BEA-090169> <Loading trusted certificates from the jks keystore file
C:\bea2613a\JROCKI~1\jre\lib\security\cacerts.>
####<Dec 29, 2006 2:15:32 PM IST> <Warning> <Server> <svaidyan02> <xbusServer>
<DynamicSSLListenThread[DefaultSecure[1]]> <<WLS Kernel>> <> <> <1167381932743> <BEA-002611>
<Hostname "svaidyan02.apac.bea.com", maps to multiple IP addresses: 192.168.1.5,
172.22.56.120>
####<Dec 29, 2006 2:15:32 PM IST> <Notice> <Server> <svaidyan02> <xbusServer> <[STANDBY]
ExecuteThread: '5' for queue: 'weblogic.kernel.Default (self-tuning)'> <<WLS Kernel>> <> <>
<1167381932753> <BEA-002613> <Channel "Default[2]" is now listening on 127.0.0.1:7021 for
```

# DATA ANALYTICS


Data Analytics is the process of examining data sets in order to draw conclusion about the information it contains.

increasingly with the aid of specialized systems and software. Data analytics technologies and techniques are widely used in commercial industries to enable organizations to make more-informed business decisions and by scientists and researchers to verify or disprove scientific models, theories and hypotheses.

Analytics is not a tool of technology, rather it is the way of thinking and acting on data.

# TYPES OF DATA ANALYTICS

Data analytics are of three types those are:-

a) Descriptive Analytics
b) Predictive Analytics
c) Prescriptive Analytics

**Descriptive Analytics:**

90% of organizations today use descriptive analytics which is the most basic form of analytics. The simplest way to define descriptive analytics is that, it answers the question "What has happened?". This type of analytics, analyses the data coming in real-time and historical data for insights on how to approach the future. The main objective of descriptive analytics is to find out the reasons behind precious success or failure in the past. The 'Past' here, refers to any particular time in which an event had occurred and this could be a month ago or even just a minute ago. The vast majority of big data analytics used by organizations falls into the category of descriptive analytics.

**Predictive Analytics:**

The subsequent step in data reduction is predictive analytics. Analyzing past data patterns and trends can accurately inform a business about what could happen in the future. This helps in setting realistic goals for the business, effective planning and restraining expectations. Predictive analytics is used by businesses to study the data and ogle into the crystal ball to find answers to the question "What could happen in the future based on previous trends and patterns?"

**Prescriptive Analytics:**

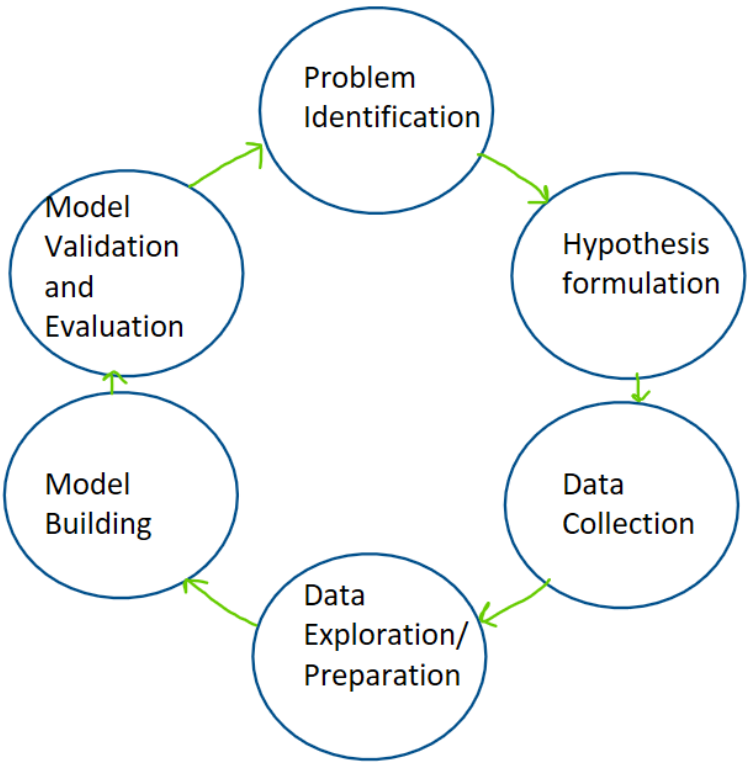
Big data might not be a reliable crystal ball for predicting the exact winning lottery numbers but it definitely can highlight the problems and help a business understand why those problems occurred. Businesses can use the data-backed and data-found factors to create prescriptions for the business problems, that lead to realizations and observations.

# PROCESS OF ANALYZING THE DATA IN DATA SCIENCE

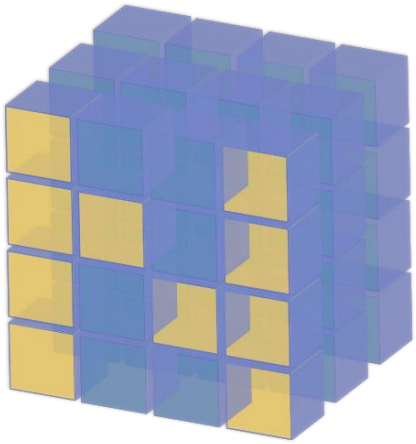
Data goes through various process during data analytics in data Science. Those are given below

**DATA SCIENCE PROCESS OR LIFE CYCLE**

1. Problem Identification
2. Hypothesis formulation
3. Data Collection
4. Data Exploration/Preparation
5. Model Building
6. Model Validation and Evaluation



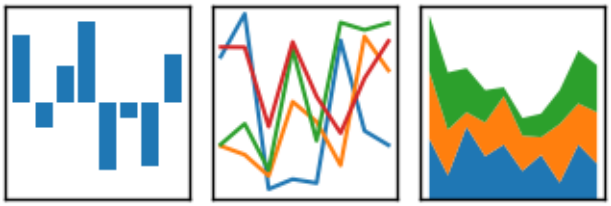
# DATA MANAGEMENT TOOLS

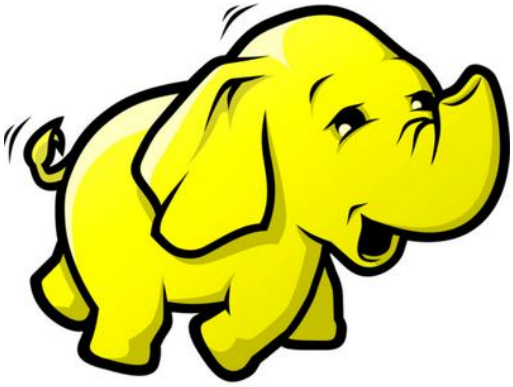


NumPy

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

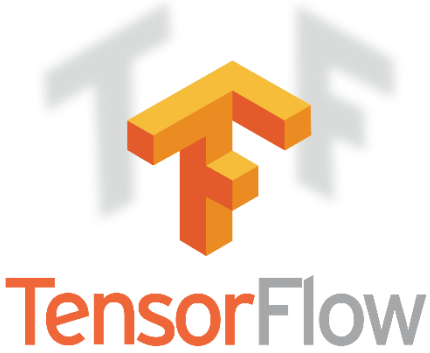


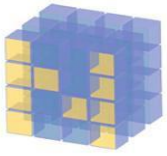
# TOP 10 PYTHON LIBRARIES

One of the reason PYTHON Is mostly popular among developers is its wide range of libraries. We can't even count how many libraries it has but We will be considering the following 10 libraries:

- NUMPY
- PANDAS
- MATPLOTLIB
- SEABORN
- IPYTHON
- TENSORFLOW
- SCIKIT-LEARN
- KERAS
- PYTORCH
- SCIPY







# JUPETER NOTEBOOK

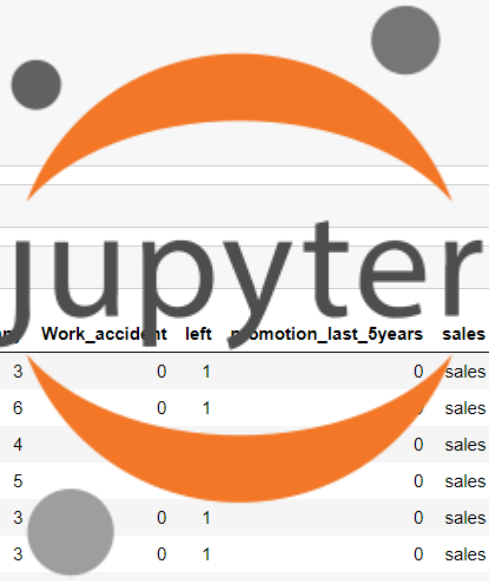
Jupiter Notebook is an interface we use to write python programs. It is Widely popular because most of the popular python libraries are pre-installed in it.

And we can write and run part of a code written in a code cell. So it is very helpful for debugging.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from mpl_toolkits import mplot3d
import seaborn as sns
plt.style.use('seaborn-darkgrid')
%matplotlib inline
```

```
df=pd.read_csv("HR_comma_sep.csv")
```

```
df.head(10)
```




| | satisfaction_level | last_evaluation | number_project | average_montly_hours | time_spend_compan | Work_accident | left | promotion_last_5years | sales | salary |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0.38 | 0.53 | 2 | 157 | 3 | 0 | 1 | 0 | sales | low |
| **1** | 0.80 | 0.86 | 5 | 262 | 6 | 0 | 1 | 0 | sales | medium |
| **2** | 0.11 | 0.88 | 7 | 272 | 4 | 0 | 1 | 0 | sales | medium |
| **3** | 0.72 | 0.87 | 5 | 223 | 5 | 0 | 1 | 0 | sales | low |
| **4** | 0.37 | 0.52 | 2 | 159 | 3 | 0 | 1 | 0 | sales | low |
| **5** | 0.41 | 0.50 | 2 | 153 | 3 | 0 | 1 | 0 | sales | low |
| **6** | 0.10 | 0.77 | 6 | 247 | 4 | 0 | 1 | 0 | sales | low |
| **7** | 0.92 | 0.85 | 5 | 259 | 5 | 0 | 1 | 0 | sales | low |
| **8** | 0.89 | 1.00 | 5 | 224 | 5 | 0 | 1 | 0 | sales | low |
| **9** | 0.42 | 0.53 | 2 | 142 | 3 | 0 | 1 | 0 | sales | low |

# DATA SCIENCE IN HR

Data science has crept into so many areas of business that even smaller companies have begun to adopt some of the techniques large organizations have been using to better understand their customers and their business. Increasingly, data science is making its way into human resources, where companies are leveraging the information on performance, engagement, retention and more to make better decisions.

**1. Recruiting the Right Employees:**

Recruiting the right employees is the number one job for HR departments across the globe. As companies are discovering, data science can improve hiring by helping recruiters and managers to create a more effective process.

**2. Comp and Promotion:**

Although recruiting is a key part of company growth, it's useless if companies are unable to keep their employees satisfied with the right position and career growth. Again, data science can help.

**3. Benefits Analysis:**

Data science can also be used to uncover insights about what benefits matter to employees, which they value and, of equal importance, analyse the costs. Applying data science techniques to such areas can inform companies if they are giving a good deal or not to their employees and if the company is gaining or losing from such packages.

# USE OF DATA SCIENCE IN HR

**Let us point out some of the challenges faced by the hiring managers:**

**1.Eligible Candidates**:

Hiring managers have the responsibility to find out the best fit for a particular position. On an average, there are around 250 resumes received for a single corporate position. If the job is posted through an online job portal then these figures scale up to around 4,27,000 resumes. These figures are just not mind-boggling but filtering candidates manually or through some filtering algorithm is another uphill task.

**2.Demand and Supply ratio**:

With increasing number of job opportunities in the market candidates have become quite selective. They have various opportunities and they can easily reject a job offer if they find something more feasible and exciting elsewhere. Hiring managers have to jot down plans and create such marketing strategies that make the candidate stay. If a selected candidate drops off then again have to repeat the complete process and find a new replacement.

**3.Tenuous relationship of hiring manager and recruiters**:

Sometimes the exact job requirements are not clearly communicated to the hiring managers. According to a survey conducted by ICIMS, 80% of recruiters think they have very good understanding of their job position while 61% of hiring managers believe that recruiters have moderate levels of understanding. This imbalance between both the parties is quite strenuous and creates a barrier for a smooth workflow.

A solution to simplify the task of staffing is through the use of data effectively according to the needs and requirements. As stated by Mike Loukides, VP, O'Reilly Media "*Data scientists are involved with gathering data, massaging it into a tractable form, making it tell its story, and presenting that story to others.*"

# DTYPES OF GIVEN RAW DATA

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from mpl_toolkits import mplot3d
import seaborn as sns
plt.style.use('seaborn-darkgrid')
%matplotlib inline
```

```
df=pd.read_csv("HR_comma_sep.csv")
```

```
df.head(10)
```

| | satisfaction_level | last_evaluation | number_project | average_montly_hours | time_spend_company | Work_accident | left | promotion_last_5years | sales | salary |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.38 | 0.53 | 2 | 157 | 3 | 0 | 1 | 0 | sales | low |
| 1 | 0.80 | 0.86 | 5 | 262 | 6 | 0 | 1 | 0 | sales | medium |
| 2 | 0.11 | 0.88 | 7 | 272 | 4 | 0 | 1 | 0 | sales | medium |
| 3 | 0.72 | 0.87 | 5 | 223 | 5 | 0 | 1 | 0 | sales | low |
| 4 | 0.37 | 0.52 | 2 | 159 | 3 | 0 | 1 | 0 | sales | low |
| 5 | 0.41 | 0.50 | 2 | 153 | 3 | 0 | 1 | 0 | sales | low |
| 6 | 0.10 | 0.77 | 6 | 247 | 4 | 0 | 1 | 0 | sales | low |
| 7 | 0.92 | 0.85 | 5 | 259 | 5 | 0 | 1 | 0 | sales | low |
| 8 | 0.89 | 1.00 | 5 | 224 | 5 | 0 | 1 | 0 | sales | low |
| 9 | 0.42 | 0.53 | 2 | 142 | 3 | 0 | 1 | 0 | sales | low |

This dataset "HR_comma_sep.csv" is originally belongs to IBM.

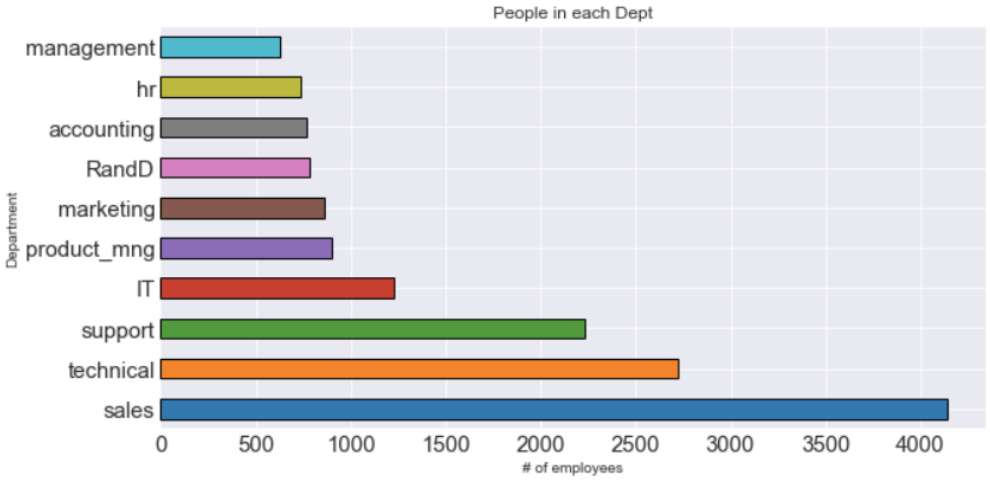
It has 10 columns and 14999 rows for their employees.

| Columns | dtype | Description |
|---|---|---|
| satisfaction_level | float64 | Satisfaction level of employees (0,1) |
| last_evaluation | float64 | Last evolution results of employees (0,1) |
| number_project | int64 | Number of projects done by each employee |
| average_montly_hours | int64 | Average monthly hour work |
| time_spend_company | int64 | Daily time spend by each employee |
| Work_accident | int64 | Work accident status (binary) |
| left | int64 | Resignation status (binary) |
| promotion_last_5years | int64 | Promotion status (binary) |
| sales | object | Department details |
| salary | object | Salary scale details |

# DATA DESCRIPTION

## 1.How many employees are in each department ?

```
#no of emplyees in each department
fig=plt.figure(figsize=(10,5))
ax=fig.add_subplot(111)
df.sales.value_counts().plot(kind='barh',color=sns.color_palette(),ec='k',ax=ax,fontsize=15)
ax.set_ylabel("Department")
ax.set_xlabel("# of employees")
ax.set_title("People in each Dept")
plt.show()
```




Here we have plotted a bar plot on department as "Y" and employee count as "X".

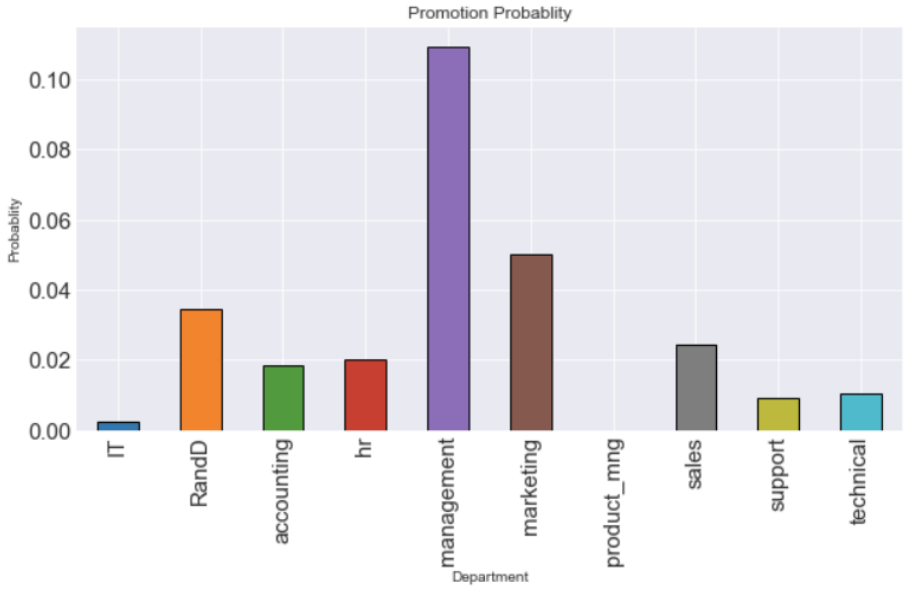
From the diagram ,we can know the number of employees in each department. The department which has the maximum employees is "sales" department and the minimum employees in "management" department.

The department names are given below as the employee number decreases in the graph:

***sales,technical,support,IT,product_mng,marketing,RandD,accounting, hr ,management.***

# 2.Probabilities of getting promotion in each department?

```
#probablity of geting promotion in each department
fig=plt.figure(figsize=(10,5))
ax=fig.add_subplot(111)
df.groupby('sales')['promotion_last_5years'].mean().plot(kind='bar',color=sns.color_palette(),ax=ax,ec='k',fontsize=15)
ax.set_ylabel("Probablity")
ax.set_xlabel("Department")
ax.set_title("Promotion Probablity")
plt.show()
```




'promotion_last_5years' is a binary column hence we grouped the columns with respect to departments and take the mean to get the probabilities that we have plotted here.

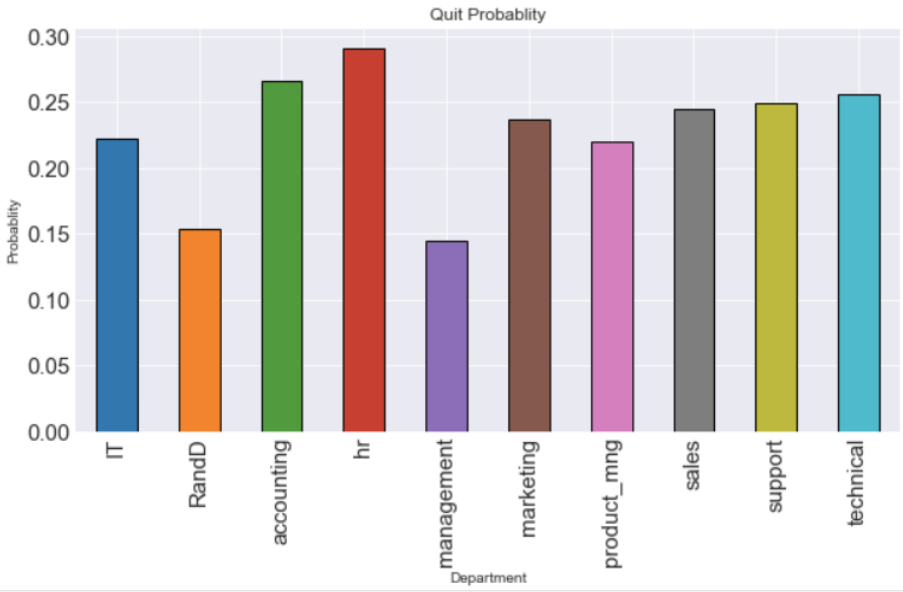
From the diagram ,we can know the number of promoted employees in each department. The department which has the maximum promoted employees is "management" department and the minimum promoted employees in "product_mng" department.

The department names are given below as the employee number decreases in the graph:

***Management, marketing, RandD,sales,hr,accounting,technical,Support, IT,product_mng***

# 3.Probabilities of quitting jobs in each department:

```
#which emplyees have most probablty to quit the job
fig=plt.figure(figsize=(10,5))
ax=fig.add_subplot(111)
df.groupby('sales')['left'].mean().plot(kind='bar',color=sns.color_palette(),ax=ax,ec='k',fontsize=15)
ax.set_ylabel("Probablity")
ax.set_xlabel("Department")
ax.set_title("Quit Probablity")
plt.show()
```




'left' is a binary column hence we grouped the columns with respect to departments and take the mean to get the probabilities that we have plotted here.

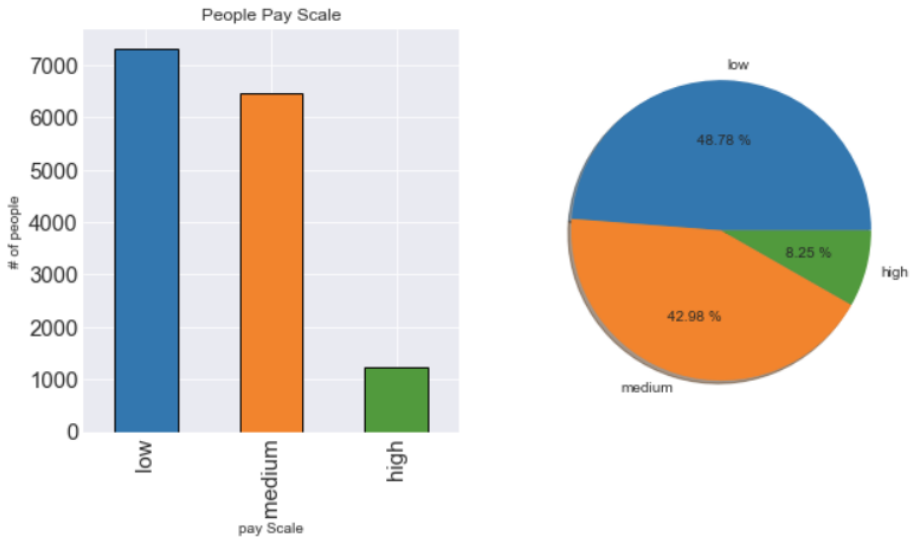
From the diagram ,we can know the number of promoted employees in each department. The department which has the maximum promoted employees is "hr" department and the minimum promoted employees in "management" department.

The department names are given below as the employee number decreases in the graph:

***Hr,accounting,technical,support,sales,marketing,IT,product_mng,Rand D,management***

# 4.Pay Scale of employees:

```
fig=plt.figure(figsize=(10,5))
ax=fig.add_subplot(121)
d=df.salary.value_counts()
d.plot(kind='bar',color=sns.color_palette(),ec='k',ax=ax,fontsize=15)
ax.set_ylabel("# of people")
ax.set_title("People Pay Scale")
ax.set_xlabel("pay Scale")
ax1=fig.add_subplot(122)
ax1.pie(d,labels=d.index,autopct=lambda x:"{:.2f} %".format(x),shadow=True,colors=sns.color_palette())
plt.show()
```

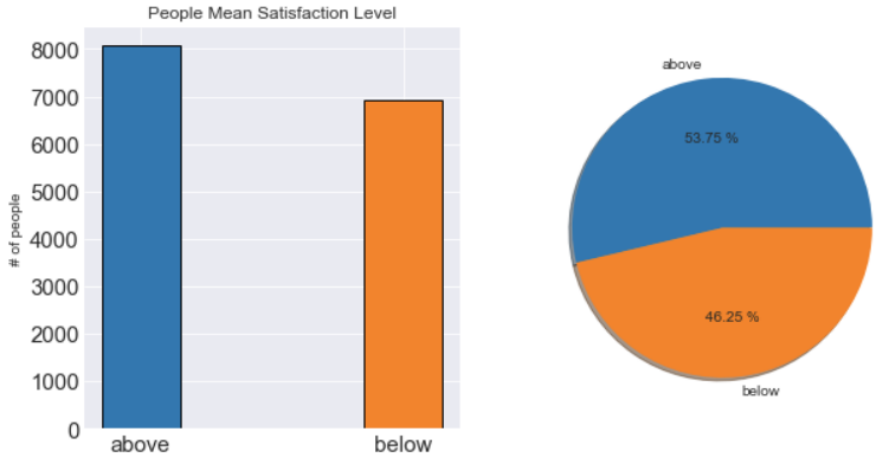



'Salary' has three values {high, medium, low}. We count each class and plotted it.

From the diagram ,we can know the  pay Scale of employees which is from total employees' maximum employees has low pay scale which is 48.78 %,then 42.98 % employees have medium pay scale and very few employees have high pay scale which is 8,25%.

# 5.Employees satisfaction level:

```
#how many emplyees are satisfied more than average satisfaction level
d=(df.satisfaction_level>=df.satisfaction_level.mean()).value_counts()
fig=plt.figure(figsize=(10,5))
ax=fig.add_subplot(121)
ax.bar(x=['above','below'],height=d,width=0.3,color=sns.color_palette(),ec='k')
ax.set_ylabel("# of people")
ax.set_title("People Mean Satisfaction Level")
ax.tick_params(labelsize=15)
ax1=fig.add_subplot(122)
ax1.pie(d,labels=['above','below'],autopct=lambda x:"{:.2f} %".format(x),shadow=True,colors=sns.color_palette())
plt.show()
```

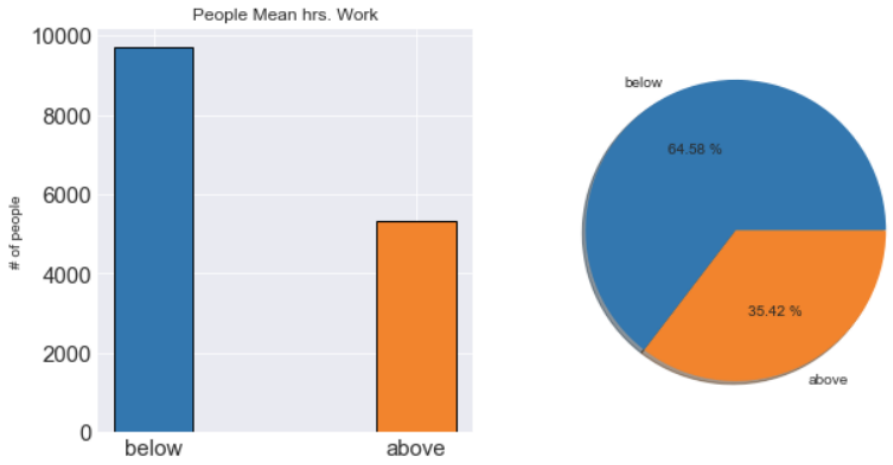



Here satisfaction level is measured above or below the mean satisfaction level of the employees.

From the diagram ,we can know the satisfaction level of employees which is from total employees' maximum employees are unsatisfied which is 53.75 % and 46.25 % employees are satisfied.

# 6.How much employees work everyday ?

```
#how many emplyees work more than average hour of work
d=(df.time_spend_company>=df.time_spend_company.mean()).value_counts()#.plot(kind='bar',color=sns.color_palette())
fig=plt.figure(figsize=(10,5))
ax=fig.add_subplot(121)
ax.bar(x=['below','above'],height=d,width=0.3,color=sns.color_palette(),ec='k')
ax.set_ylabel("# of people")
ax.set_title("People Mean hrs. Work")
ax.tick_params(labelsize=15)
ax1=fig.add_subplot(122)
ax1.pie(d,labels=['below','above'],autopct=lambda x:"{:.2f} %".format(x),shadow=True,colors=sns.color_palette())
plt.show()
```

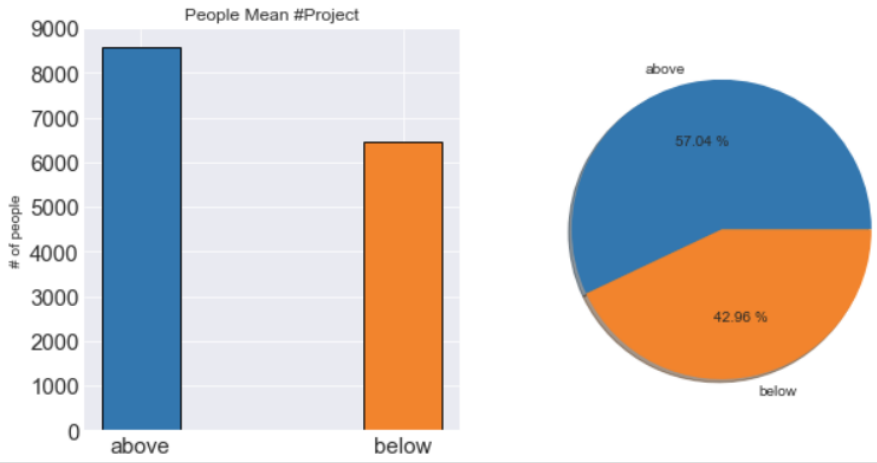



Here work in measured as above and below the mean work in the company.

From the diagram ,we can know that 64.58 % of employees work more than mean working hours and 35.42% of employees does not.

# 7.Employe's Project Work:

```
#how many emplyees completed more than average no of projects
d=(df.number_project>=df.number_project.mean()).value_counts()#.plot(kind='bar',color=sns.color_palette())
fig=plt.figure(figsize=(10,5))
ax=fig.add_subplot(121)
ax.bar(x=['above','below'],height=d,width=0.3,color=sns.color_palette(),ec='k')
ax.set_ylabel("# of people")
ax.set_title("People Mean #Project")
ax.tick_params(labelsize=15)
ax1=fig.add_subplot(122)
ax1.pie(d,labels=['above','below'],autopct=lambda x:"{:.2f} %".format(x),shadow=True,colors=sns.color_palette())
plt.show()
```

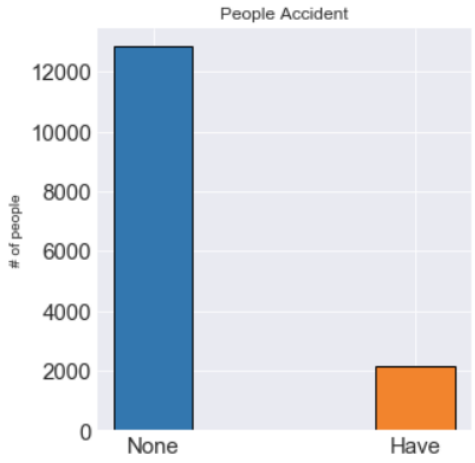


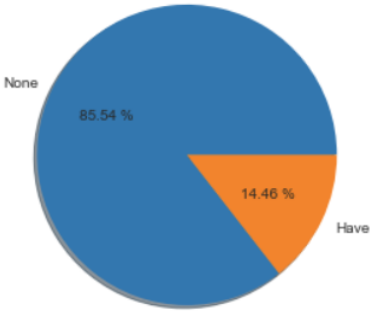

Here project in measured as above and below the mean no of project an employee do in the company.

From the diagram ,we can see 57.04% of employees are done more than mean no of project an employee does in the company and 42.96 % do not.

# 8.How many employees have accident in work ?

```
#how many peoples have a accident in work
d=df.Work_accident.value_counts()#.plot(kind='barh')
fig=plt.figure(figsize=(10,5))
ax=fig.add_subplot(121)
ax.bar(x=['None','Have'],height=d,width=0.3,color=sns.color_palette(),ec='k')
ax.set_ylabel("# of people")
ax.set_title("People Accident")
ax.tick_params(labelsize=15)
ax1=fig.add_subplot(122)
ax1.pie(d,labels=['None','Have'],autopct=lambda x:"{:.2f} %".format(x),shadow=True,colors=sns.color_palette())
plt.show()
```

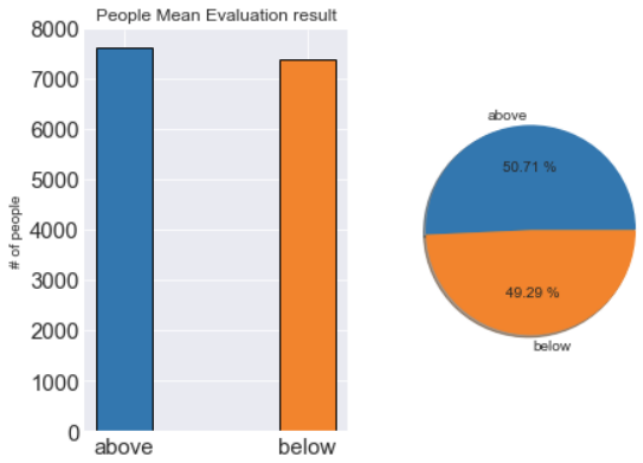



From the diagram, we can see that from the total employees of the company 14.46% have work accidents. And 85.54% still managed to work consciously.

# 9.Last Evaluation Results:

```
#how many emplyees are above average in the last evolution
d=(df.last_evaluation>=df.last_evaluation.mean()).value_counts()#.plot(kind='bar',color=sns.color_palette())
fig=plt.figure(figsize=(7,5))
ax=fig.add_subplot(121)
ax.bar(x=['above','below'],height=d,width=0.3,color=sns.color_palette(),ec='k')
ax.set_ylabel("# of people")
ax.set_title("People Mean Evaluation result")
ax.tick_params(labelsize=15)
ax1=fig.add_subplot(122)
ax1.pie(d,labels=['above','below'],autopct=lambda x:"{:.2f} %".format(x),shadow=True,colors=sns.color_palette())
plt.show()
```

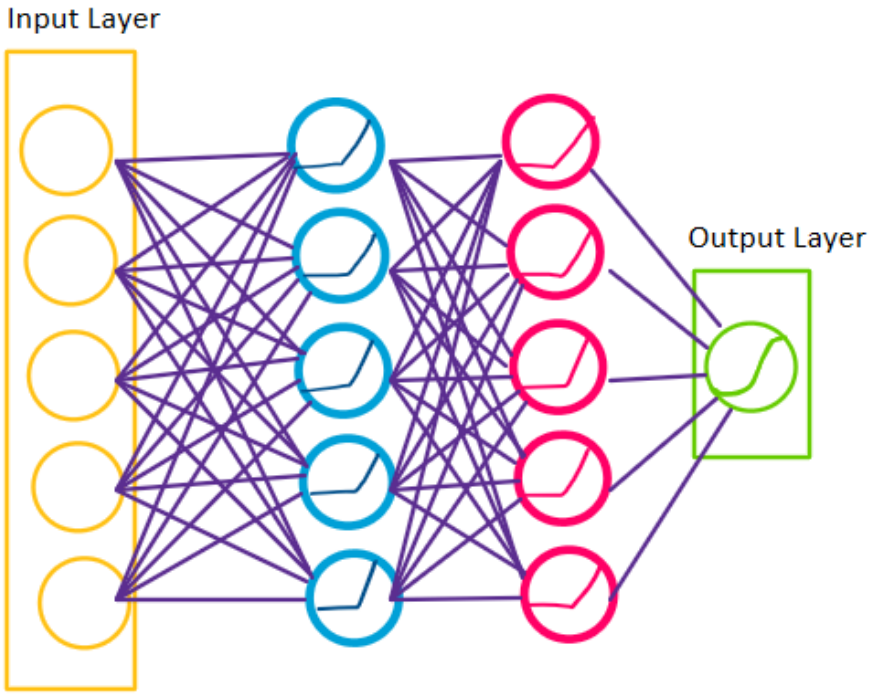



Here evaluation in measured as above and below the mean result of the evaluation an employee get last time in the company.

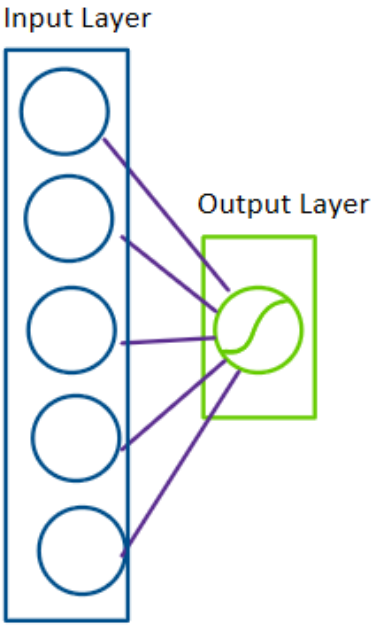
From the diagram ,we can see 57.04% of employees have more than mean result in the last evaluation an employee has in the company and 42.96 % have not.

# OUR APPROACH TO THE PROBLEM

## ANN:




## Logistic Regression:




**Goal:** Our goal is to predict whether an employee would stay or leave the company in near future.

# TRAINING AND TESTING DATA SET

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
```

```
X=df.drop(columns=['left'])
y=df.left
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.33, random_state=42)
```

```
scalar=StandardScaler()
X_train=scalar.fit_transform(X_train)
X_test=scalar.transform(X_test)
```

Here we have split the dataset into training and testing part so that we can train and test our statistical models for predictive analytics part. We have taken 33% test size and 67% train dataset size. Our features are columns except 'left', where categorical columns are encoded with this data structure.

```
{'sales': {'product_mng': 0,
  'technical': 1,
  'support': 2,
  'marketing': 3,
  'hr': 4,
  'IT': 5,
  'RandD': 6,
  'sales': 7,
  'management': 8,
  'accounting': 9},
 'salary': {'medium': 0, 'high': 1, 'low': 2}}
```

We have scaled the train dataset with Standard Scaler and fit the scaling parameters according to train dataset, and only scaled test dataset with the scalar fitted on training dataset parameters.

# TRAINING OUR MODEL

## Training: Logistic Regressor

```
from sklearn.linear_model import LogisticRegression
```

```
logreg=LogisticRegression()
logreg.fit(X_train,y_train)
```

```
For training:
               precision    recall  f1-score   support

           0       0.82      0.93      0.87      7659
           1       0.61      0.37      0.46      2390

   micro avg       0.79      0.79      0.79     10049
   macro avg       0.72      0.65      0.67     10049
weighted avg       0.77      0.79      0.78     10049

Accuracy: 79.47059408896408 %
```

Here we can see that logistic regressor when fitted on training data have the accuracy of 79.47% on it that proves that our dataset is not linearly divisible.

# Training: ANN

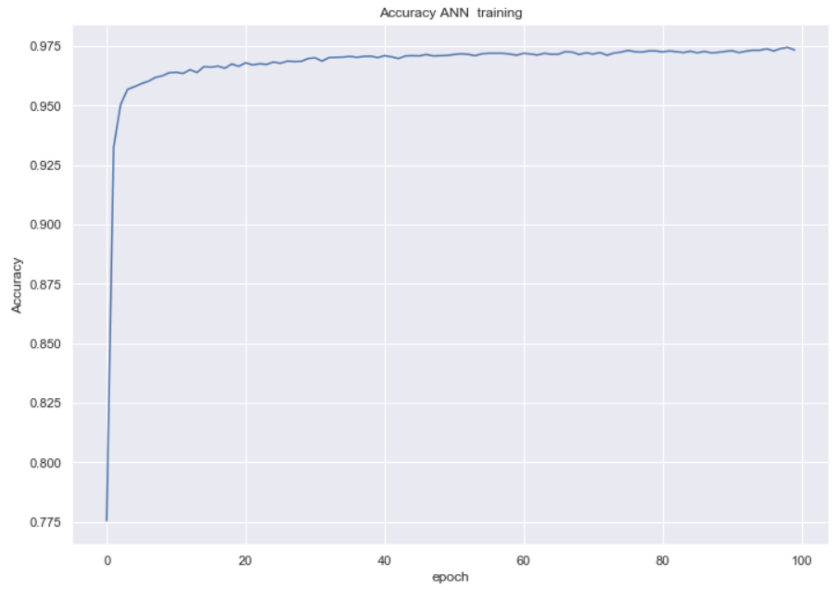
```
from tensorflow import keras
```

```
model=keras.Sequential()
model.add(keras.layers.Dense(9,'relu',input_shape=(9,)))
model.add(keras.layers.Dense(9,'relu'))
model.add(keras.layers.Dense(9,'relu'))
model.add(keras.layers.Dense(9,'relu'))
model.add(keras.layers.Dense(9,'relu'))
model.add(keras.layers.Dense(1,'sigmoid'))
model.compile(loss='binary_crossentropy',optimizer='adam',metrics=['accuracy'])
#model.summary()
```

```
WARNING:tensorflow:From C:\Users\USER\Anaconda\Anaconda3\lib\site-packages\tensorflow\python\ops\resource_variable_ops.py:435:
colocate_with (from tensorflow.python.framework.ops) is deprecated and will be removed in a future version.
Instructions for updating:
Colocations handled automatically by placer.
```

```
history=model.fit(X_train,y_train,batch_size=32,epochs=100)
```

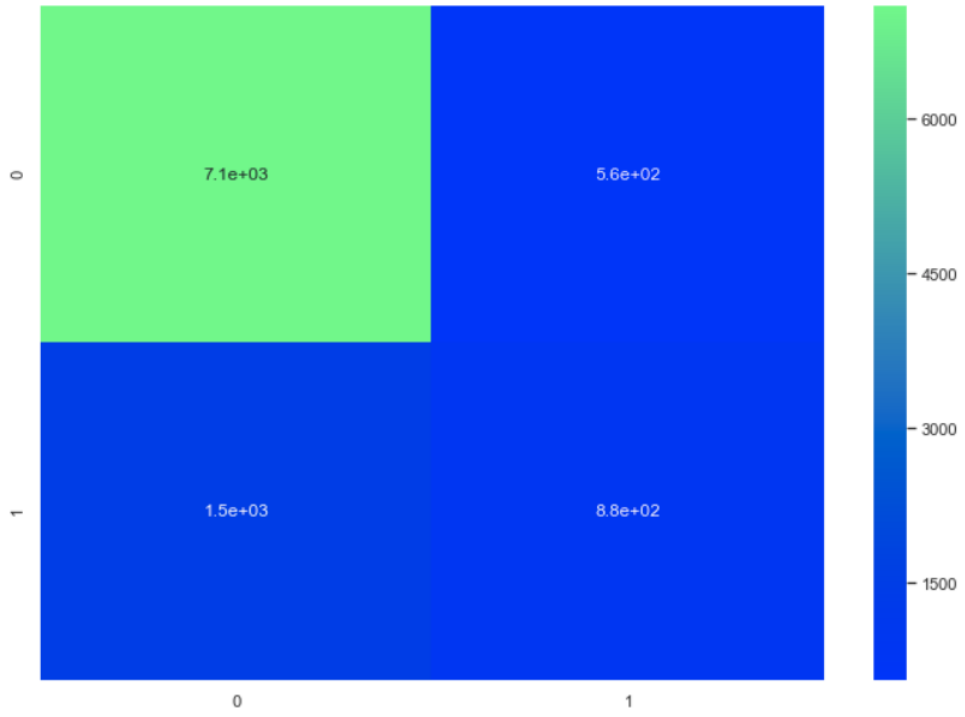
Here we can see that while we train a Neural Network on the dataset it can actually learn the patterns in the dataset and its accuracy reaches to 97%.



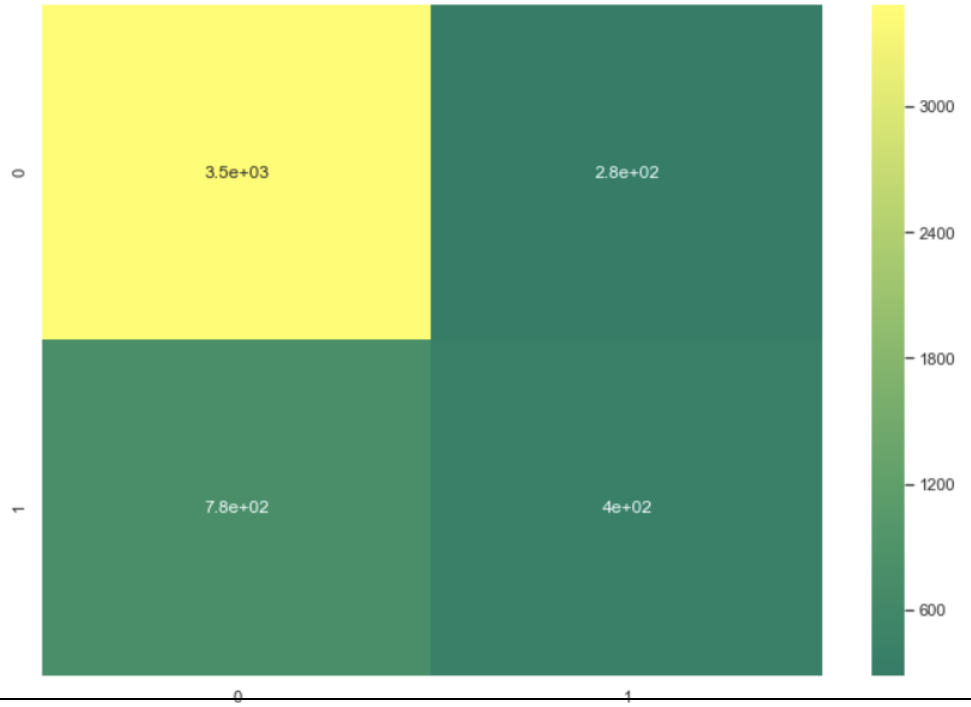
# TESTING OUR MODEL

## Accuracy: Logistic Regression

Training:

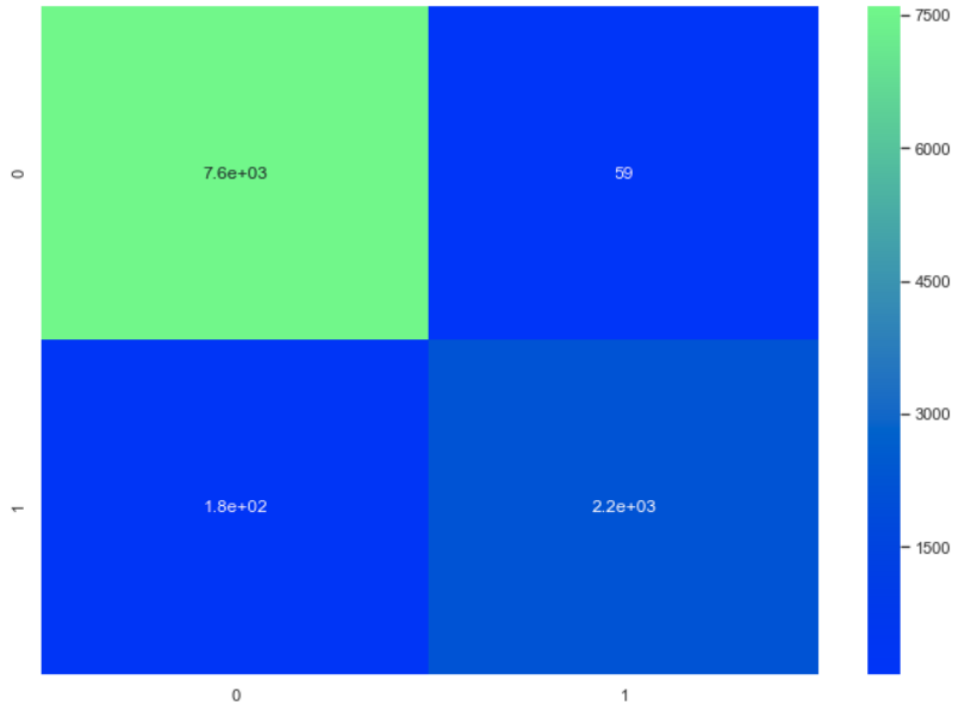



Testing

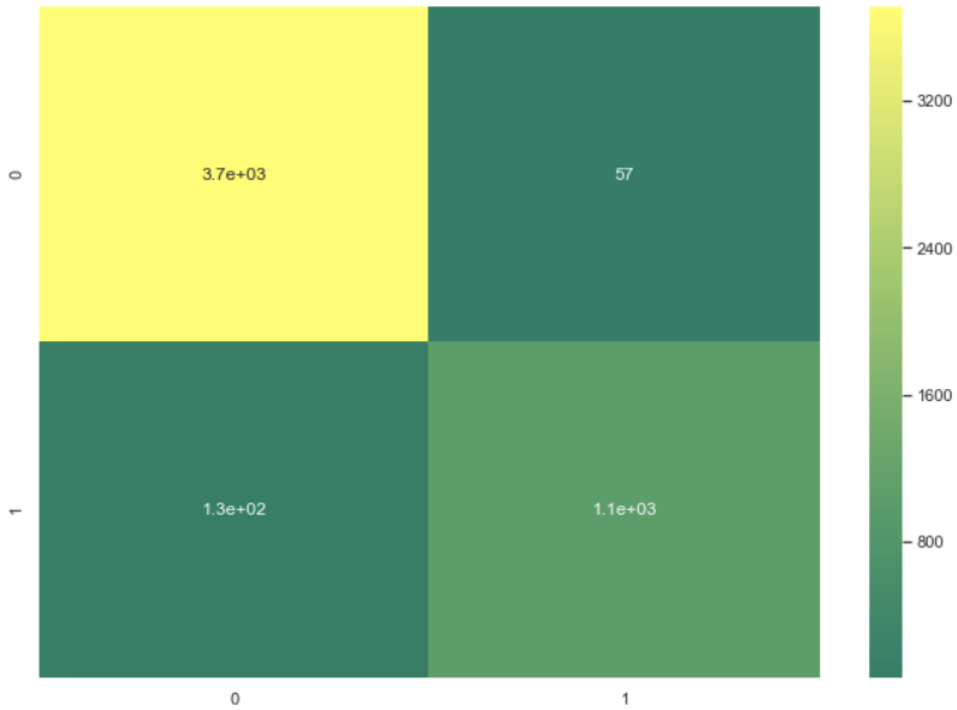


# Accuracy: ANN

Training:




Testing



# WHAT IS BETTER APPROACH ?

ANN has better approach than Logistic Regression because it has better accuracy compared to Logistic regression.

ANN is more superior statistical model than Logistic Regression.

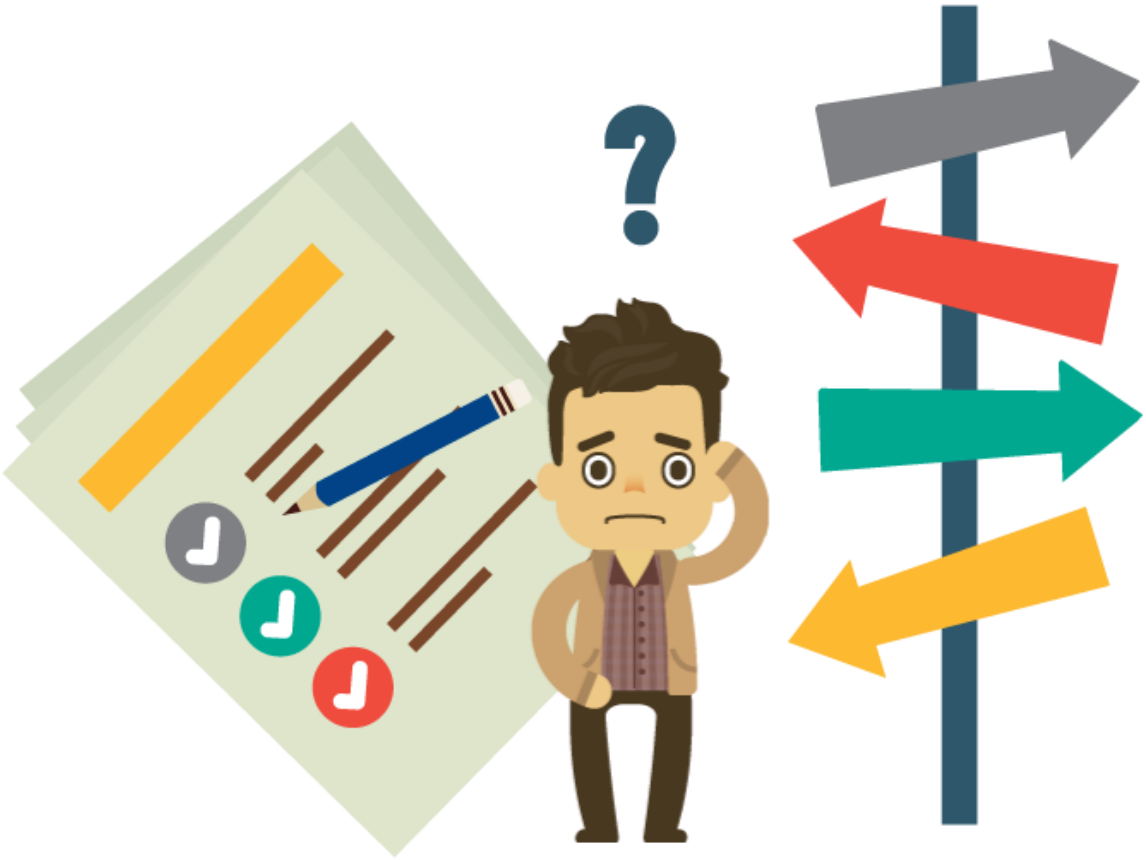
But ANN can only shine when we have a lot of data and computational power to train the model on the dataset. Here we have both of these. The dataset contains 14999 entries for only two classes and we have trained on a GPU accelerated system with GTX 1050 ti in it.

But when we don't have any of these then the better approaches will be to train machine learning models rather than deep learning models.

e.g. Decision Trees, Random forest, Logistic Regressor and SVMs.

# DECISION MAKING

Human Resource Management is all about the decision making. And here we can advise a company whether to recruit an employee or whether an existing employee will stay or leave the company.



# BENEFITS FOR THE COMPANY

- Using HR analytics, we can select perfect employees for the company.
- We can find out the probability of employees who can left the job in future.
- We can find out the satisfaction level of employees.
- We can take safety equipment for better employment.
- Can derive ideal time for job.
- Can find out work value per capita.

# CONCLUSION

So, we can conclude that by using data science we can help an organization to take economic and crucial decisions to increase their profit as well as increase their employee's motivations by stress management techniques. And by stress management techniques we suggest that if the company need any employee and the model predict that he would leave the company in near future, then the organization must reduce the work load of the employee for a while.

END