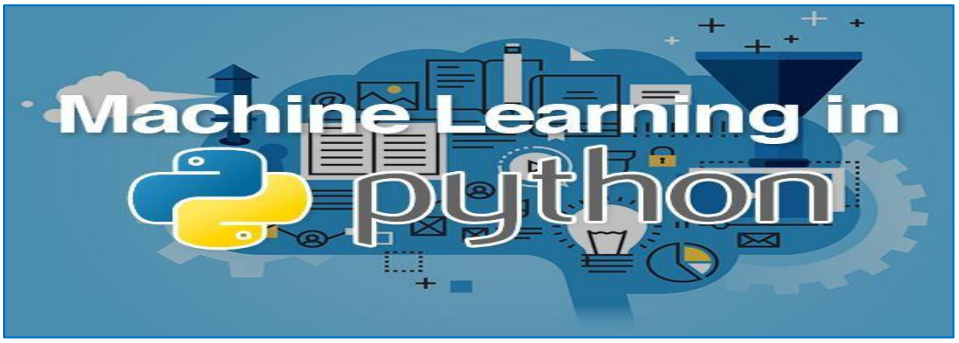
औद्योगिक प्रशिक्षण के लिए राष्ट्रीय संस्थान

**National Institute for Industrial Training**

One Premier Organization with Non Profit Status | Registered Under Govt. of WB
Empanelled Under Planning Commission Govt. of India
Inspired By: National Task Force on IT & SD Government of India

National Institute for Industrial Training- One Premier Organization with Non Profit Status Registered Under Govt.of West Bengal,Empanelled Under Planning Commission Govt.of India , Empanelled Under Central Social Welfare Board Govt. of India , Registered with National Career Services , Registered with National Employment Services.




SUBJECT: MACHINE LEARNING WITH PYTHON

SUBMITTED BY:Kunal Chowdhury

SUBMITTED TO: Soumyadip Choudhury

DATE OF SUBMISSION: 03/09/2020

# CONTENTS

# STUDENT PROFILE:

**NAME:** Kunal Chowdhury

**COLLEGE:** KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY, BHUBANESWAR

**COURSE:** B.Tech

**BRANCH:** ETC (Electronics and Telecommunications Engineering)

**SEMESTER:** 5th SEMESTER, 3rd YEAR

**PHONE NO:** 9957790331

**Email id:** kunalchowdhury.12.kc@gmail.com

# ACKNOWLEDGEMENT

I, Kunal Chowdhury, a student of Kalinga Institute of Industrial Technology has prepared this project on the topic "LINEAR REGRESSION OF A GIVEN DATASET USING PYTHON".

I would like to thank **NATIONAL INSTITUTE OF INDUSTRIAL TRAINING** for providing us with a platform where we could gather a lot of knowledge on python and machine learning. I would also like to thank my mentor, **Mr. Soumyadip Choudhury**, for his constant support and encouragement throughout the session as well as for providing the necessary help with the completion of this project. I got to learn a lot while working on this project.

In today's scenario, programming language plays a very important role. Being acquainted with a language like python surely has its own benefits. I have had a lot of fun while completing this project and hope that the knowledge I gathered while working on it helps me in the future as well.

# INTRODUCTION

Python is a widely used general-purpose, high level programming language. It was initially designed by GUIDO VAN ROSSUM in 1991 and developed by PYTHON SOFTWARE FOUNDATION. Python works on different platforms like Windows, Linux, Mac, Raspberry, Pi, etc.

Python comes with a huge amount of inbuilt libraries. Many of the libraries are used for Artificial Intelligence and Machine Learning. Some of the libraries are:

- Tensorflow: It is high-level neural network library
- Scikit-learn: It is used for data mining, data analysis and machine learning purposes
- Pylearn2: It is more flexible than scikit-learn

Python has no concept of datatype. It is a weakly type language. It has an easy implementation for OpenCV. For other languages, students and researchers need to get to know the languages before getting into machine learning with that language.

Machine Learning (ML) is that field of computer science with the help of which computer systems can provide sense to data in the same way as human beings do. With the help of python, ML can be easily accessed as even a programmer with very basic knowledge can easily handle python.

# OBJECTIVE

- Python is a widely used, interpreted, object-oriented, and high-level programming language with dynamic semantics used for general purpose programming.
- Python provides plenty of data mining tools which help in better handling of data.
- Python is important for data scientists because it provides a vast variety of applications used in data science.
- Python enables you to perform data analysis, data manipulation, and data visualization, which are very important in data science.
- Python also provides more flexibility in the field of Machine Learning, Artificial Intelligence and Deep Learning.
- The key focus of ML is to allow computer systems to learn from experience without being explicitly programmed or human intervention.

# HARDWARE AND SOFTWARE REQUIREMENT:

## SOFTWARE:

Operating system: Windows
Front End: Python 3.7
Platform: Anaconda

## HARDWARE:

Machine: HP probookX360440G1
Speed: 1.60 GHz & above
RAM: 8 GB

# BRIEF IDEA OF THE PROJECT

## LINEAR REGRESSION:

Linear regression is a common Statistical Data Analysis technique. It attempts to model the relationship between two variables by fitting a linear equation to observe data. One variable is considered to be an explanatory variable and the other variable is considered to be a dependent variable.

In this project, we have taken a dataset of 30 salary entries upon the Years of Experience and then predicted the best salary possible. We carried out the process in the following steps:

- Importing the dataset.
- Splitting dataset into training set and testing set (2 dimensions of x and y per each set). Normally, the testing set should be 5% to 30% of dataset.
- Visualizing the training set and testing set to double check
- Initializing the regression model and fitting it using training set (both x and y)
- Prediction

In [5]:

```
pip install numpy
```

```
Requirement already satisfied: numpy in c:\users\kiit\anaconda3\lib\site-p
ackages (1.18.1)
Note: you may need to restart the kernel to use updated packages.
```

In [7]:

```
pip install matplotlib
```

```
Requirement already satisfied: matplotlib in c:\users\kiit\anaconda3\lib\s
ite-packages (3.1.3)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\kiit\anaconda
3\lib\site-packages (from matplotlib) (1.1.0)
Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in
c:\users\kiit\anaconda3\lib\site-packages (from matplotlib) (2.4.6)
Requirement already satisfied: python-dateutil>=2.1 in c:\users\kiit\anaco
nda3\lib\site-packages (from matplotlib) (2.8.1)
Requirement already satisfied: numpy>=1.11 in c:\users\kiit\anaconda3\lib
\site-packages (from matplotlib) (1.18.1)
Requirement already satisfied: cycler>=0.10 in c:\users\kiit\anaconda3\lib
\site-packages (from matplotlib) (0.10.0)
Requirement already satisfied: setuptools in c:\users\kiit\anaconda3\lib\s
ite-packages (from kiwisolver>=1.0.1->matplotlib) (45.2.0.post20200210)
Requirement already satisfied: six>=1.5 in c:\users\kiit\anaconda3\lib\sit
e-packages (from python-dateutil>=2.1->matplotlib) (1.14.0)
Note: you may need to restart the kernel to use updated packages.
```

In [9]:

```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import seaborn as sns
from math import *
from pylab import *
```

# Importing the dataset

In [10]:

```
dataset = pd.read_csv('Salary_Data - Salary_Data.csv')
```

In [11]:

```
dataset.head()
```

Out[11]:

| | YearsExperience | Salary |
|---|---|---|
| **0** | 1.1 | 39343 |
| **1** | 1.3 | 46205 |
| **2** | 1.5 | 37731 |
| **3** | 2.0 | 43525 |
| **4** | 2.2 | 39891 |

In [12]:

```
dataset.columns
```

Out[12]:

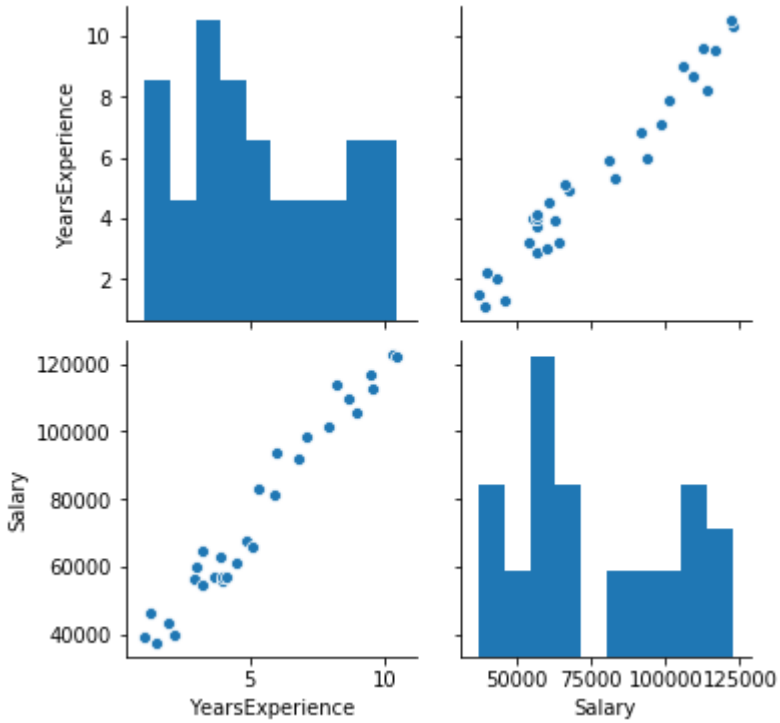
```
Index(['YearsExperience', 'Salary'], dtype='object')
```

In [13]:

```
sns.pairplot(dataset)
```

Out[13]:

```
<seaborn.axisgrid.PairGrid at 0xeff9330>
```

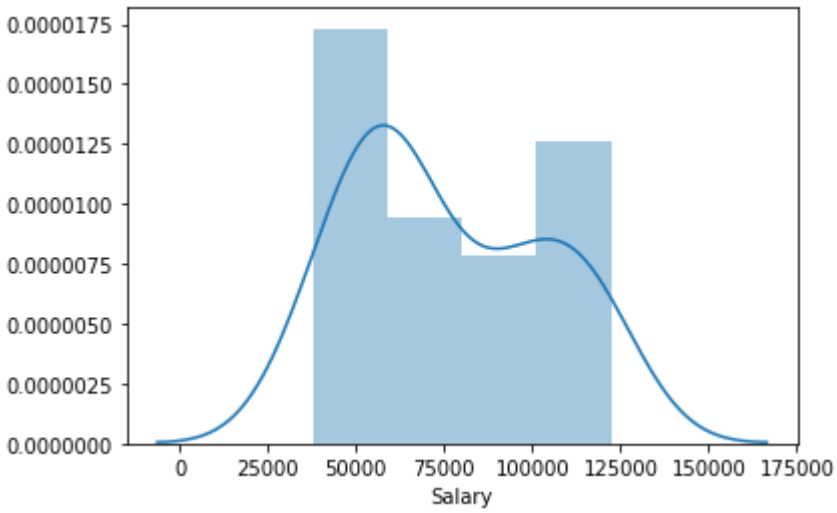


In [14]:

```
sns.distplot(dataset['Salary'])
```

Out[14]:

```
<matplotlib.axes._subplots.AxesSubplot at 0xf219430>
```

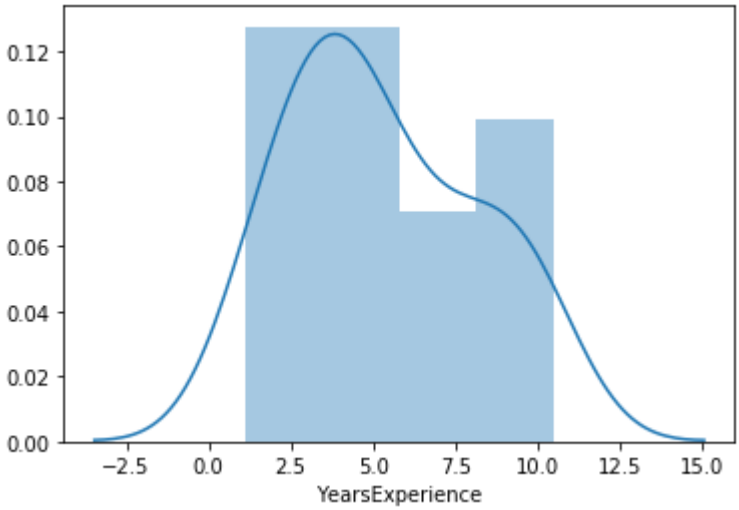



In [15]:

```
sns.distplot(dataset['YearsExperience'])
```

Out[15]:

```
<matplotlib.axes._subplots.AxesSubplot at 0xf312af0>
```

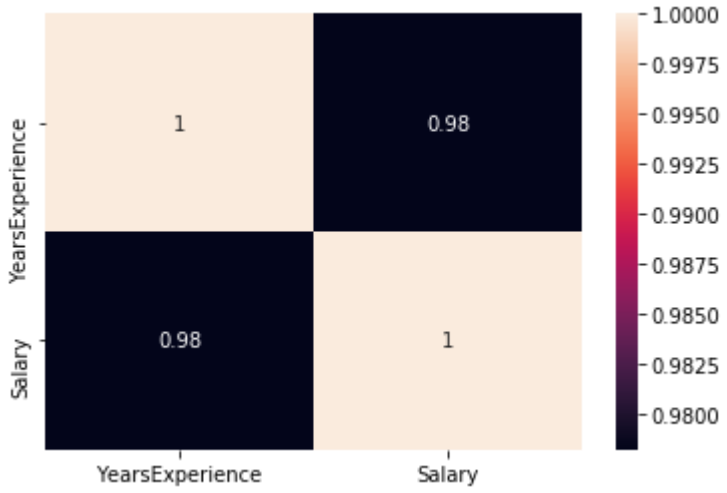


In [16]:

```
sns.heatmap(dataset.corr(),annot=True)
```

Out[16]:

```
<matplotlib.axes._subplots.AxesSubplot at 0xfc360b0>
```




# LINEAR REGRESSION

In [17]:

```
x = dataset.iloc[:, :-1].values #years of experience
y = dataset.iloc[:, 1].values   #salary
```

## Splitting the dataset into the Training set and Test set

In [19]:

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=1/3, random_state=0
)
```

## Fitting Simple Linear Regression to the Training set

In [20]:

```
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(x_train, y_train)
```

Out[20]:

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=F
alse)
```

## Visualizing the Training set results

In [24]:

```
viz_train = plt
viz_train.scatter(x_train, y_train, color='blue')
viz_train.plot(x_train, regressor.predict(x_train), color='green')
viz_train.title('Salary VS Experience (Training set)')
viz_train.xlabel('Year of Experience')
viz_train.ylabel('Salary')
viz_train.show()
```


## Visualizing the Test set results

In [31]:

```
viz_test = plt
viz_test.scatter(x_test, y_test, color='yellow')
viz_test.plot(x_train, regressor.predict(x_train), color='blue')
viz_test.title('Salary VS Experience (Test set)')
viz_test.xlabel('Year of Experience')
viz_test.ylabel('Salary')
viz_test.show()
```


## Predicting the result of 5 Years Experience

In [34]:

```
y_pred = regressor.predict([[5]])
print(y_pred)
```

```
[73545.90445964]
```

## Predicting the Test set results

In [36]:

```
y_pred = regressor.predict(x_test)
print(y_pred)
```

```
[ 40835.10590871 123079.39940819  65134.55626083  63265.36777221
 115602.64545369 108125.8914992  116537.23969801  64199.96201652
  76349.68719258 100649.1375447 ]
```

## Predicting the errors along with accuracy

In [37]:

```
from sklearn import metrics
```

In [39]:

```
print('Mean Absolute Error:',metrics.mean_absolute_error(y_test,y_pred))
print('Mean Squared Error:',metrics.mean_squared_error(y_test,y_pred))
print('Root Mean Squared Error:',metrics.mean_squared_error(y_test,y_pred))
```

```
Mean Absolute Error: 3426.4269374307123
Mean Squared Error: 21026037.329511296
Root Mean Squared Error: 21026037.329511296
```

In [ ]:

```

```

# ADVANTAGES

- The Python Package Index (PyPI) contains numerous Third-Party Modules that make python capable of interacting with most of the other languages and platforms.
- Python provides a large standard library (NumPy for numerical calculations, Pandas for data analytics, etc)
- Python language is developed under an OSI-approved open source license, which makes it free to use and distribute.
- Python offers excellent readability and simple-to-learn syntax.
- Python is ideal for general purpose tasks such as data mining and big data facilitation.
- Machine Learning can review large volumes of data and discover specific trends and patterns that would not be apparent to humans.
- As ML algorithms gain experience, they keep improving in accuracy and efficiency. This lets them make better decisions.
- Machine Learning algorithms are good at handling data that are multi-dimensional and multi-variety, and they can do this in dynamic or uncertain environments.

# FUTURE SCOPE OF PYTHON:

- Python finds huge application in today's world. It is used in the analysis of large amount of data due to its high performance libraries and tools.
- The most popular Python libraries for data visualization are MATPLOTLIB and SEABORN.
- In the field of Data Science, Python is used as an engineering tool. The scope of Data Science with Python is pretty wide and being an open source, people can contribute to it and this can keep on going.
- Python has numerous frameworks and libraries like Scikit-Learn, Numpy, Pandas, Seaborn, Matplotlib and many more which has made it a very popular language.
- Python aims to deal with a large number of datasets across computer clusters through it's high performance toolkits and libraries.

# FUTURE SCOPE OF MACHINE LEARNING:

- The future of Machine Learning looks promising as the skilled talent pool for Machine Learning engineers is not yet enough to meet the growing demand for trained professionals.
- It is expanding across all fields such as banking and finance, information technology, media & entertainment, gaming, and the automotive industry.
- Machine Learning scope is extremely high in terms of salary and the number of job opportunities. Thus, it is a good option to make a lucrative career in ML by becoming a Machine Learning professional.

# CONCLUSION

The principal purpose of Data Science is to find patterns within data. It uses various statistical techniques to analyze and draw insights from the data. Data Science aims to derive conclusion from the given data. Industries need data to help them make careful decisions. Data Science churns raw data into meaningful insights. The purpose of Data Scientist is not only limited to statistical processing of data but also managing and communicating data that helps in making better decisions.

In this particular project we have used implemented the application of machine learning using Python. Using linear regression, we have predicted the salary of the employee according to their years of experience. We also calculated the possible error and summarized the results. ML has a wide range of applicaton that can be used in data analysis. ML uses statistical methods to enable machines to improve with experience. Hence in keeping up with the developments in the field of technolgy in today's world, it can be concluded that the use of ML makes management and handling of data easier and efficient.

# BIBLIOGRAPHY

In completing this project, the sources I referred to are:


- www.sciencedirect.com
- https://intellipaat.com/blog/future-scope-of-machine-learning/
- Introduction to Machine Learning with Python, by Andreas Muller