Project Report on

Salary Expectation based on Machine Learning using Python

Submitted in the fulfilment of our training program

By:

Ankush Dhar (11900220031) **(TEAM LEADER)**

Rupantar Chakraborty (11900220015)

Priyatosh Nandy (11900220021)

Priyajit Choudhury (11900220025)

Rahul Gorai (11900220028)

Abinash Chhetri (11900220026)

Piyush Ranjan (11900220027)

5th Semester students of

Siliguri Institute of Technology

Under

Maulana Abul Kalam Azad University of Technology

(MAKAUT)

Under the supervision of

Mr. Arpan Samant

ABSTRACT

The principal objective of this project is to perform full analysis for the Salary Expectation of the employees of the organisation and detect and predict the salaries of the employees using Machine Learning.

In this project the concept of Machine Learning using Python has been used to its fullest extent. The data has been collected through a company for the study and implementation of it through Machine Learning. The data collected so far is of experience years and the salary. We know that to bring out the best out of employee an organisation must set some parameters through which a productivity of an employee can be measured. One such metrics is the number of years a person has been in the field.

Salary Prediction based on experience using ML: In this project, we have worked on an end-to-end case study to understand the different stages of Model building using the Machine Learning concept. This will deal with "data manipulation" with pandas and Numpy, and "data visualization" with Matplotlib library with the Salary dataset. After Data manipulation, Data visualization will be performed using graphs.

ACKNOWLEDGEMENT

It is a great pleasure for us to acknowledge the assistance and participation of a large number of individuals to this attempt. Our project report has been structured under the valued suggestion, support and guidance of **Mr. Arpan Samanta**. Under his guidance we have accomplished the challenging task in a very short time.

Finally, we would express our sincere thankfulness to our family members for inspiring us all throughout and always encouraging us.

Arpan Samanta

Sikharthi Infotech Pvt. Ltd.

Piyush Ranjan

Department of Information Technology

INTRODUCTION

In the past, we used to have data in a structured format but now as the volume of the data is increasing, so the number of structured data becomes very less, so to handle the massive amount of data we need data science techniques. Those data can be used to get the proper business insights and the hidden trends from them. These insights help the organization to predict the Future and helps to reduce the production cost. Build model based on the data to give the ability to the machine to predicts on its own.

The project is all about *Salary Prediction based on experience using ML*, we have worked on an end-to-end case study to understand the different stages of Model building using the Machine Learning concept. This will deal with "data manipulation" with pandas and Numpy, and "data visualization" with Matplotlib library with the Salary dataset. After Data manipulation, Data visualization will be performed using graphs.

Certificate of Approval

The training project is hereby approved as a creditable study for the training program and presented in a manner of satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorsed or approved any statement made, opinion express or conclusion therein but approve this project only for the purpose for which it is submitted.

Final Examination for

Evaluation of the Project

Signatures of Examiners

Salary Prediction Based on Experience using Machine learning

Salary Expectation based on Machine Learning using Python

In [1]:
```
import pandas as pd

#pandas is a software library written for the Python programming language for data manipulation and analysis.
```

In [2]:
```
import numpy as np

#NumPy is a library for the Python programming language, adding support for large, multi-dimensional arrays and
#matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
```

In [3]:
```
data=pd.read_csv("Salary.csv")

# here we are reading the data from the .csv(Comma-separated values) file as a source of data using pandas.
```

In [4]:
```
data.head()

# The head() function is used to get the first n rows
```

Out[4]:

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

In [5]:
```
data.columns

## This column function prints the column index and its data type
```

Out[5]:
```
Index(['YearsExperience', 'Salary'], dtype='object')
```

In [6]:
```
data.describe()


#Pandas describe() is used to view some basic statistical details like percentile, mean, std etc.
#of a data frame or a series of numeric values.
```

Out[6]:

	YearsExperience	Salary
count	35.000000	35.000000
mean	6.308571	83945.600000
std	3.618610	32162.673003
min	1.100000	37731.000000
25%	3.450000	57019.000000
50%	5.300000	81363.000000
75%	9.250000	113223.500000
max	13.500000	139465.000000

In [7]:
```
data.isnull()

#The isnull() method returns a DataFrame object where all the values are replaced with a Boolean value True for
#NULL values, and otherwise False.
```

Out[7]:

	YearsExperience	Salary
0	False	False
1	False	False
2	False	False
3	False	False
4	False	False
5	False	False
6	False	False
7	False	False
8	False	False
9	False	False
10	False	False
11	False	False
12	False	False
13	False	False
14	False	False
15	False	False
16	False	False
17	False	False
18	False	False
19	False	False
20	False	False
21	False	False
22	False	False
23	False	False
24	False	False
25	False	False
26	False	False
27	False	False
28	False	False
29	False	False
30	False	False
31	False	False
32	False	False
33	False	False
34	False	False

In [8]:
```
data.isnull().any()

#isnull() is used to check null values in the given data, any() function is used to check for columns instead of row
```

Out[8]:
```
YearsExperience    False
Salary             False
dtype: bool
```

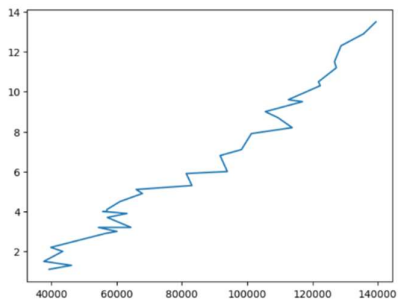
In [9]:
```
import matplotlib.pyplot as plt
# importing pyplot function under matplotlib library
# pyplot is a collection of functions that make matplotlib work like MATLAB
```

In [10]:
```
import seaborn as sns
# importing seaborn library


#Seaborn is a Python data visualization library based on matplotlib.
#it provides a high-level interface for drawing attractive and informative statistical graphics.
```

In [11]:
```
plt.plot(data['Salary'],data['YearsExperience'])
# .plot() function is used to plot the graph of the given data using matplotlib
```

Out[11]: [<matplotlib.lines.Line2D at 0x1857ee95540>]



In [12]:
```
data
```

Out[12]:

	YearsExperience	Salary
0	1.1	39343
1	1.3	46205
2	1.5	37731
3	2.0	43525
4	2.2	39891
5	2.9	56642
6	3.0	60150
7	3.2	54445
8	3.2	64445
9	3.7	57189
10	3.9	63218
11	4.0	55794
12	4.0	56957
13	4.1	57081
14	4.5	61111
15	4.9	67938
16	5.1	66029
17	5.3	83088
18	5.9	81363
19	6.0	93940
20	6.8	91738
21	7.1	98273
22	7.9	101302
23	8.2	113812
24	8.7	109431
25	9.0	105582
26	9.5	116969
27	9.6	112635
28	10.3	122391
29	10.5	121872
30	11.2	127345
31	11.5	126756
32	12.3	128765
33	12.9	135675
34	13.5	139465

Now it is the time to make a model

In [13]:
```
from sklearn.model_selection import train_test_split
#it will split the data set in train and testing

x = data.drop('Salary',axis = 1)
#it will drop the "Salary"m column from the .csv file and show the "Years of Experience" part only
```

In [14]:
```
x
```

Out[14]:

	YearsExperience
0	1.1
1	1.3
2	1.5
3	2.0
4	2.2
5	2.9
6	3.0
7	3.2
8	3.2
9	3.7
10	3.9
11	4.0
12	4.0
13	4.1
14	4.5
15	4.9
16	5.1
17	5.3
18	5.9
19	6.0
20	6.8
21	7.1
22	7.9
23	8.2
24	8.7
25	9.0
26	9.5
27	9.6
28	10.3
29	10.5
30	11.2
31	11.5

32	12.3
33	12.9
34	13.5

In [15]:
```
y=data['Salary']

#now y is only having the Salary
```

In [16]:
```
y.head()

#shows the head(first 5 data)
```

Out[16]:
```
0    39343
1    46205
2    37731
3    43525
4    39891
Name: Salary, dtype: int64
```

Splitting the data

In [17]:
```
xtrain, xtest, ytrain, ytest = train_test_split(x, y, test_size = 0.2, random_state = 42)

#test_size = 0.2 means Training part is 80% and unseen data is 20% for the model
#random_state = 42 means it basically take data from anywhere for testing and training
```

In [18]:
```
from sklearn.linear_model import LinearRegression
#because Linear Regression algorith is used here

L=LinearRegression()
```

In [19]:
```
L.fit(xtrain,ytrain)

#fit means basically train
```

Out[19]:
```
LinearRegression()
```

In [20]:
```
y_pred=L.predict(xtest)

#xtest means experience
```

In [21]:
```
ytest

#ytest means Salary
```

Out[21]:
```
26    116969
13     57081
24    109431
21     98273
15     67938
29    121872
19     93940
Name: Salary, dtype: int64
```

In [22]:
```
y_pred

#what the model has predicted
```

Out[22]:
```
array([110576.91706292,  64251.57268882, 103713.90308157,  89987.87511888,
        71114.58667017, 119155.68453961,  80551.23089452])
```

In [23]:
```
print(L.score(xtest, ytest))

#core accuracy score
```

```
0.8914234140042779
```

Conclusion

We have fully completed this project based on Machine Learning using Python to predict the salaries of employees based on their experiences. We by know that Machine Learning has a huge and vast scope in such genuine problem. It is like that we can improve the productivity by getting the knowledge on the basis of the experience a certain employee has over time.