**REPORT**

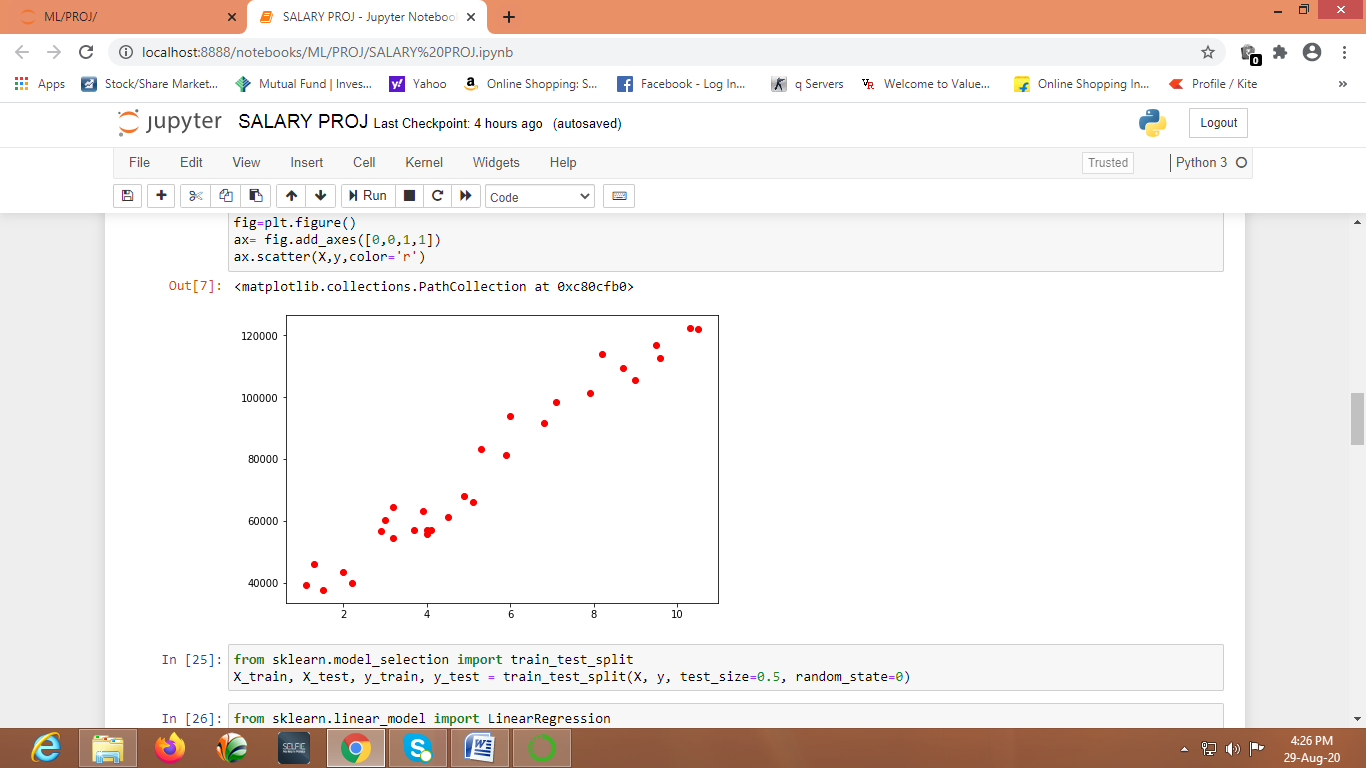
**Problem Statement:**

Develop a python application which can be used by employers for predicting future employee salary.

**Data Summary:**

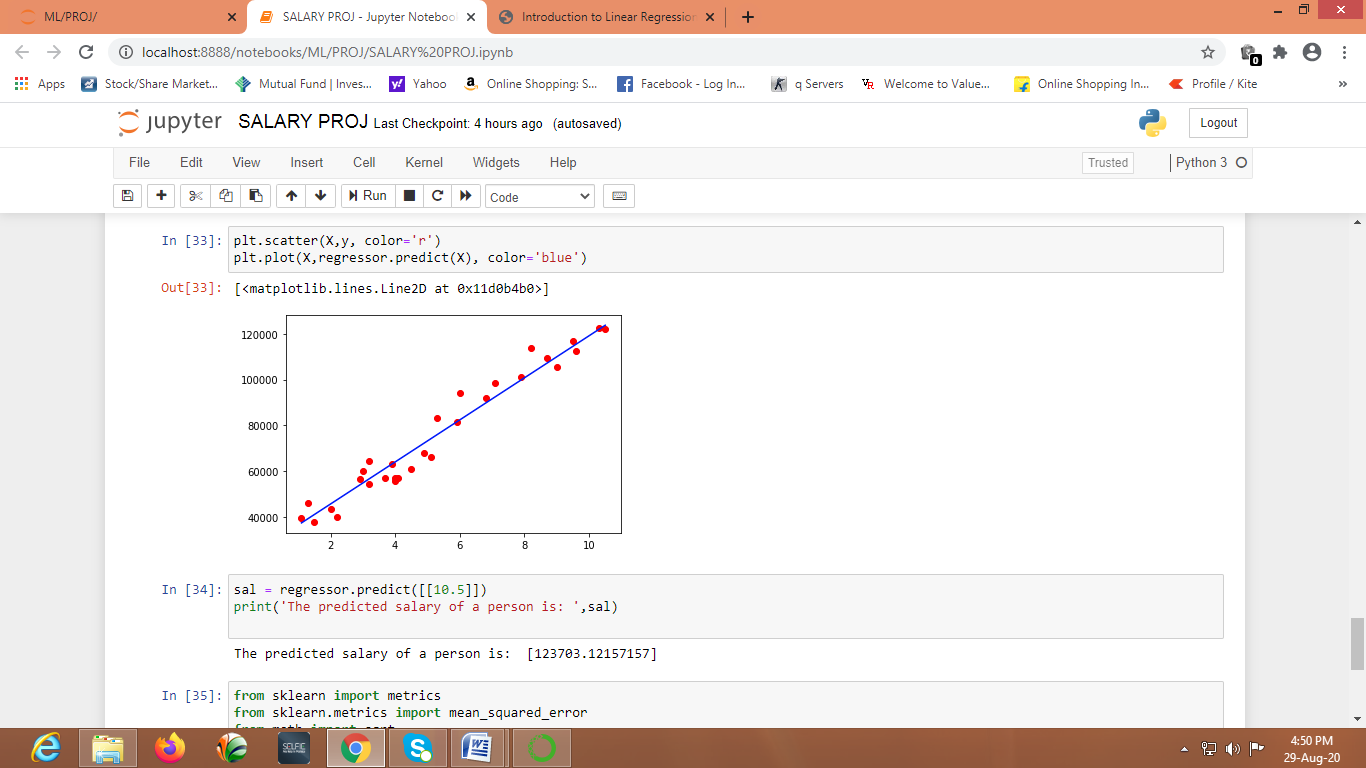
The objective of the project is to predict the salary of employee depending on years of experience. The dataset contains two columns ‘YearsExperience’ and ‘Salary’, with each column containing 30 values.

**Visualization:**

The visualization is done using scatter plot. A scatter plot is a data visualization that displays the values of two different variables as points. The data for each point is represented by its horizontal (x) and vertical (y) position on the visualization.

X-axis represents years of experience and Y-axis represents salary.

On observing the points in the graph, the model chosen is Linear Regression. Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a regression line.



**ML Algorithm:**

The data available is continuous (quantitative) data. On visualizing the data, the points are present linearly. Hence the algorithm used is linear regression.

Linear regression is used when we want to predict the value of a variable based on the value of another variable. The variable we want to predict is called the dependent variable and the variable we are using to predict the other variable's value is called the independent variable. Here the dependent variable is ‘Salary’ and independent variable is ‘YearsExperience’.

**Accuracy/Error:**

[R-squared](https://statisticsbyjim.com/glossary/r-squared/) is a goodness-of-fit measure for linear [regression](https://statisticsbyjim.com/glossary/regression-analysis/) models. This statistic indicates the percentage of the variance in the [dependent variable](https://statisticsbyjim.com/glossary/response-variables/) that the [independent variables](https://statisticsbyjim.com/glossary/predictor-variables/) explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale.

Hence R-squared value is used to determine the accuracy.

The mean squared error tells you how close a regression line is to a set of points. It does this by taking the distances from the points to the regression line and squaring them. It’s called the [mean](https://www.statisticshowto.com/mean/)squared error as you’re finding the [average](https://www.statisticshowto.com/arithmetic-mean/)of a set of errors.