**Intel College Excellence Program   
Project Synopsis**

**“Prediction of Employees salary using Regression Technique”**

|  |  |  |  |
| --- | --- | --- | --- |
| **Team member’s detail** | | | |
| **S.No.** | **Participant Name** | **Mobile No.** | **Email ID** |
| 1 | Harshit Shukla | 7982214437 | ***harshitshuklaa5@gmail.com*** |
| 2 | Honey Makharia | 9693219554 | ***makhariahoney@gmail.com*** |
| 3 | Riddhi | 8868952080 | ***riddhishri8@gmail.com*** |
| **Faculty(college) mentor detail** | | | |
| **S.No.** | **Mentor Name** | **Mobile No.** | **Email ID** |
|  |  |  |  |
| **College/University Name** | | | |
| ***Galgotias University*** | | | |

School of Computing Science and Engineering,

Galgotias University

**BACKGROUND**

Data of a company is very important for them. As analysis of data is very important for the growth of company. It is because company needs to follow the trend of market to be in market and grow. In a company or industry , managing their employees is one of the difficult and most important criteria ,in accordance to make more profit to the company with less use of resources. So, it became very necessary for any company to keep view of , how much they are paying and to whom and on the basis of current demand , it is required to analyze on the basis of previous salaries of employees ,that how much extra employee of different payee scales , are needed to meet the requirement of company . Company doesnot require these analysis only for hiring and firing the staff according to requirements but they also have to give promotions to their employees on the basis of their performance and abilities and also to manage all these such tasks . So there is a need of some techniques , methods , model etc to do analysis and may be that is the reason why companies spend huge amount in finding such type of techniques .

**PROBLEM IDENTIFICATION**

Techniques , models for doing analysis on the dataset of a company are being improved from many years. Analysis is not new for this century , it is being done from thousand’ s of years back. Athough , nowadays we are very advanced in doing this comparatively , But the hope of improvement is always there. Many companies are using different softwares, techniques , model , AI , etc to do different type of analysis . But we always try to improve the accuracy of models . So we need consistent improvement in accuracy of data models.

In accordance to grow in the market , companies do lots of analysis on the salary of their employees, for which they are always in search of such models which provide them better accuracy . So that they can easily manage their expenditure on their employees and the situation of growth of company as well . For example , if a company needs 100 software engineer for their new projects , then how they will decide that what range of payee scale ,they should offer the engineers so that company don’t go in loss in worst condition and they can make profit . For this , they have to analyze their previous pattern of salary of employees and their growth and after analyzing they have to make decision which must be profitable for them. That is the reason why they always try to implement model which is more accurate .

**PROPOSED SOLUTION**

So , here we are trying to make a model which can predict salary using machine learning and we are trying to improve the accuracy . Our model can work on different attributes for better prediction and we are using regression for the analyzation of data sets .

Regression analysis is a statistical method to model the relationship between a dependent (target) and independent (predictor) variables with one or more independent variables. More specifically, Regression analysis helps us to understand how the value of the dependent variable is changing corresponding to an independent variable when other independent variables are held fixed.

Regression is a [supervised learning technique](https://www.javatpoint.com/supervised-machine-learning) which helps in finding the correlation between variables and enables us to predict the continuous output variable based on the one or more predictor variables. It is mainly used for **prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables**.

There are various types of regressions which are used in data science and machine learning. Each type has its own importance on different scenarios, but at the core, all the regression methods analyze the effect of the independent variable on dependent variables. Here we are discussing some important types of regression which are given below:

* **Linear Regression**
* **Logistic Regression**
* **Polynomial Regression**
* **Support Vector Regression**
* **Decision Tree Regression**
* **Random Forest Regression**
* **Ridge Regression**
* **Lasso Regression:**

**As of now we are using Linear regression , Lasso regression ad Ridge Regressio in our model .**

**So Lets understand these three regression in detail.**

### Linear Regression:

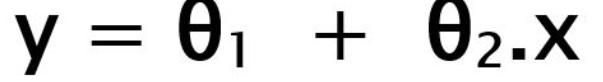
* Linear regression is a statistical regression method which is used for predictive analysis.
* It is one of the very simple and easy algorithms which works on regression and shows the relationship between the continuous variables.
* Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), hence called linear regression.

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best fit line for our model

Hypothesis function for Linear Regression :



While training the model we are given :

x: input training data (univariate – one input variable(parameter))

y: labels to data (supervised learning

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best θ1 and θ2 values.

θ1: intercept

θ2: coefficient of x

Once we find the best θ1 and θ2 values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

Cost Function (J):

By achieving the best-fit regression line, the model aims to predict y value such that the error difference between predicted value and true value is minimum. So, it is very important to update the θ1 and θ2 values, to reach the best value that minimize the error between predicted y value (pred) and true y value (y).

Cost function(J) of Linear Regression is the Root Mean Squared Error (RMSE) between predicted y value (pred) and true y value (y).

Gradient Descent:

To update θ1 and θ2 values in order to reduce Cost function (minimizing RMSE value) and achieving the best fit line the model uses Gradient Descent. The idea is to start with random θ1 and θ2 values and then iteratively updating the values, reaching minimum cost.

### Lasso Regression:

* Lasso regression is another regularization technique to reduce the complexity of the model.
* It is similar to the Ridge Regression except that penalty term contains only the absolute weights instead of a square of weights.
* Since it takes absolute values, hence, it can shrink the slope to 0, whereas Ridge Regression can only shrink it near to 0.
* It is also called as **L1 regularization**. The equation for Lasso regression will be:

Regression Analysis in Machine learning

### Ridge Regression:

A Ridge regressor is basically a regularized version of a Linear Regressor. i.e to the original cost function of linear regressor we add a regularized term that forces the learning algorithm to fit the data and helps to keep the weights lower as possible. The regularized term has the parameter ‘alpha’ which controls the regularization of the model i.e helps in reducing the variance of the estimates.

Cost Function for Ridge Regressor.

Here we are using Jupyter(Python) Notebook for the implementation of our model because python has huge varieties of libraries and it is considered as one of the best language for the machine learning and AI (Artificial Intelligence).

Some of the library that we are using :-

**Numpy**

This is used for working in domain of linear algebra , transformation , matrices etc.

**Pandas**

This is used for the manipulation in data and analyzing that. For example , we have used this to import data frames in the model.

**Matplot**

This is used for plotting figures and graph.

**Seaborn**

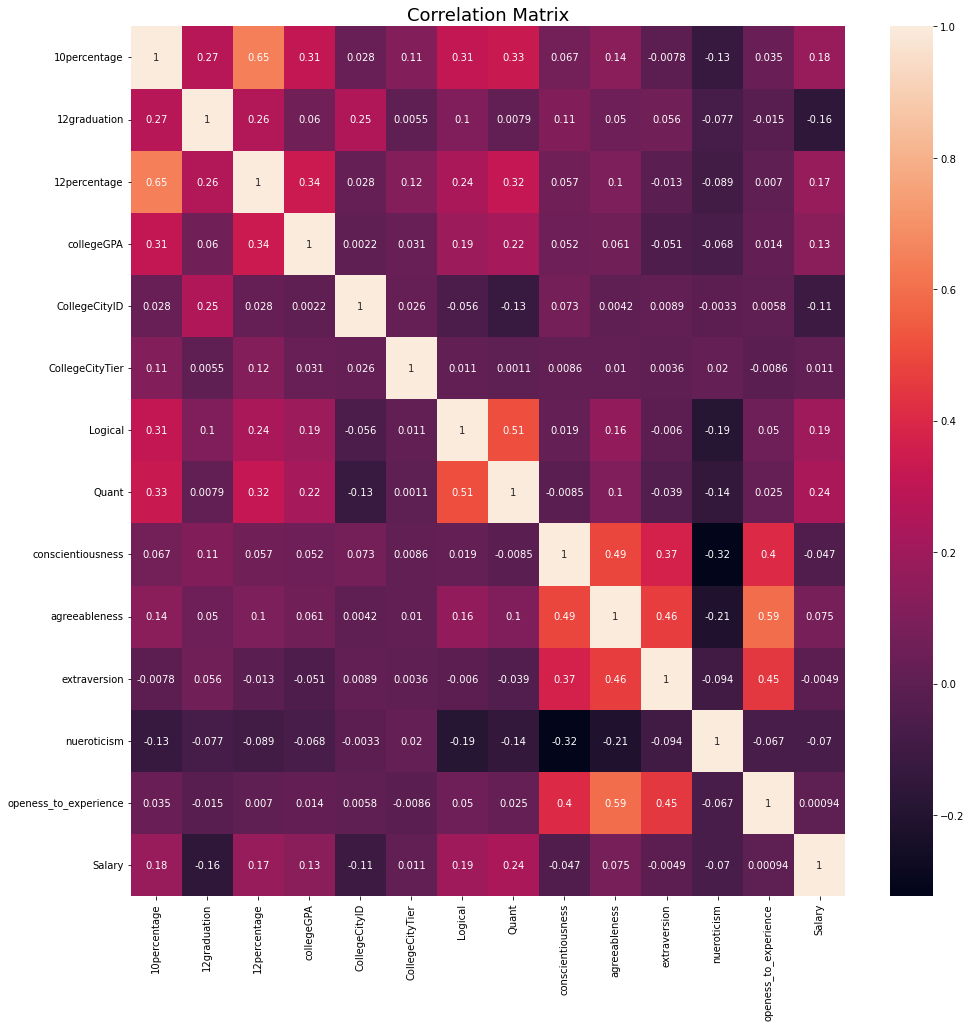
This is used for doing customization in themes colors etc.

**sklearn.linear\_model**

**sklearn.model\_selection**

**sklearn.metrics**

We are using Correlation matrix for testing the accuracy of our model.



Using this correlation matrix we are trying to check our accuracy. On this basis we have compared the results of both the regression technique.

There are various no.of attribute available :-

'ID', 'Gender', 'DOB', '10percentage', '10board', '12graduation', '12percentage', '12board', 'CollegeID', 'CollegeTier', 'Degree', 'Specialization', 'collegeGPA', 'CollegeCityID', 'CollegeCityTier', 'CollegeState', 'GraduationYear', 'English', 'Logical', 'Quant', 'Domain', 'ComputerProgramming', 'ElectronicsAndSemicon', 'ComputerScience', 'MechanicalEngg', 'ElectricalEngg', 'TelecomEngg', 'CivilEngg', 'conscientiousness', 'agreeableness', 'extraversion', 'nueroticism', 'openess\_to\_experience', 'Salary'], dtype='object'

But we don’t need all of them in our model . SO we will only use relevant attributes and drop the rest of attributres. So the following attribute we are using :-

'10percentage', '10board', '12graduation', '12percentage', '12board',

'Degree', 'Specialization', 'collegeGPA', 'CollegeCityID',

'CollegeCityTier', 'Logical', 'Quant', 'conscientiousness',

'agreeableness', 'extraversion', 'nueroticism', 'openess\_to\_experience',

'Salary'],

dtype='object'

Further we have continuously normalized the model until it only shows relevant Attribute.

**HARDWARE & SOFWARE REQUIREMENTS**

*Hardware requirements:*

* 1. *Laptop/PC/Any computer*

*Software requirements:*

*1.pyhton should be installed and environment variable should be set.*

2.all required libraries should be installed

**BLOCK DIAGRAM & DESCRIPTION**

Data Model

Data Sets

Many techniques can be impemented

Lasso

Regression

Linear Regression

Output

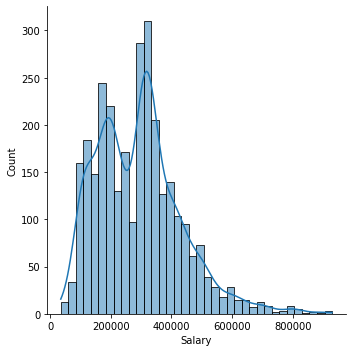
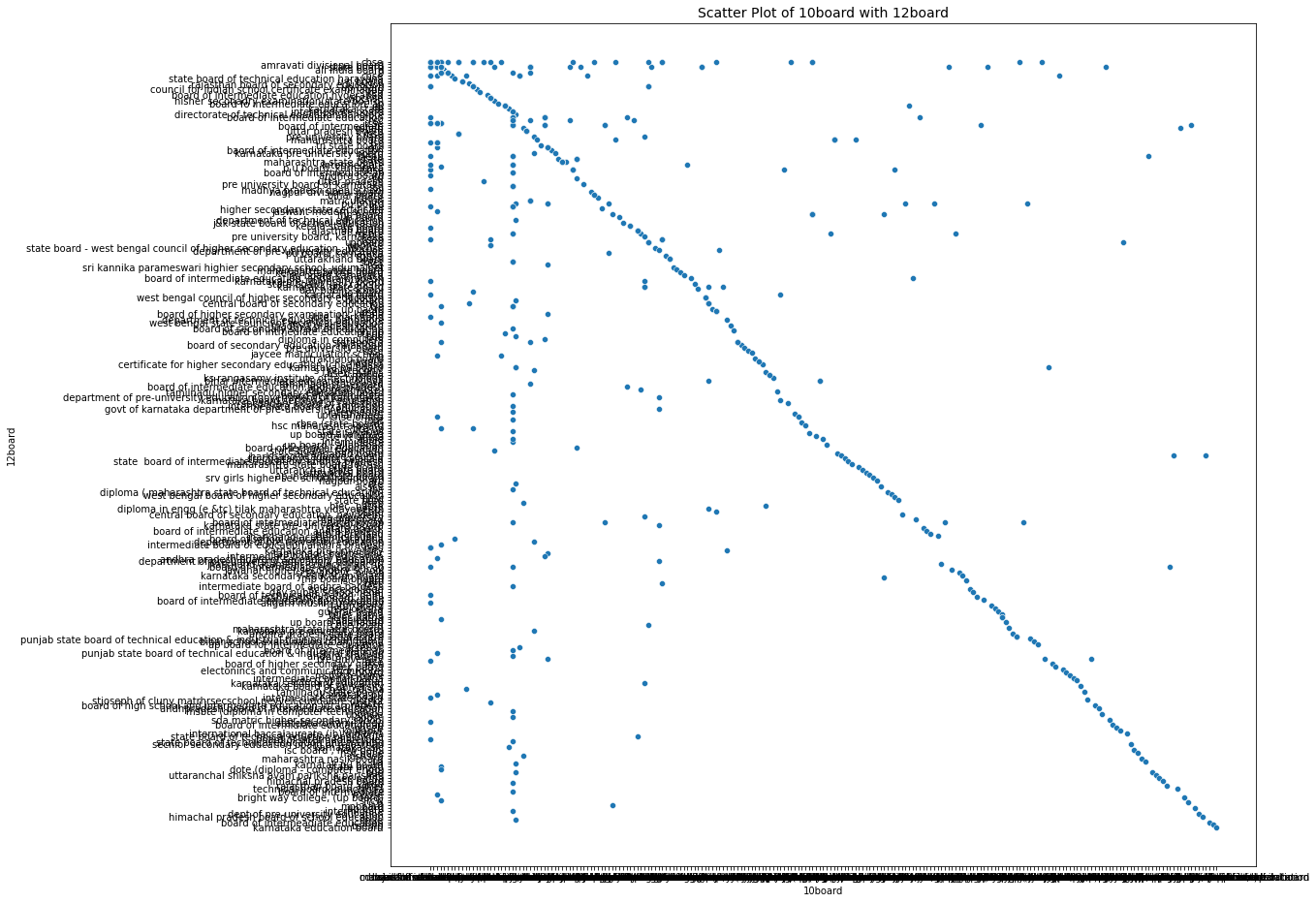
Prediction after performing all operation(Can be figure , graph,values,etc.)

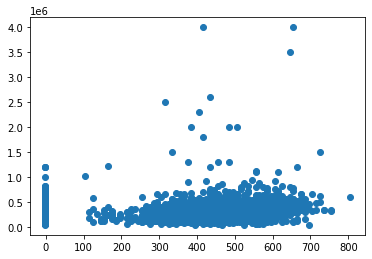
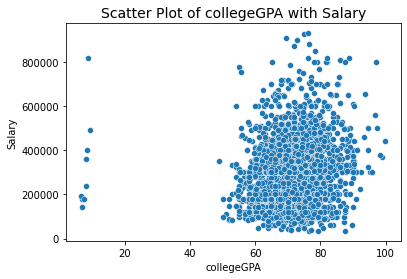
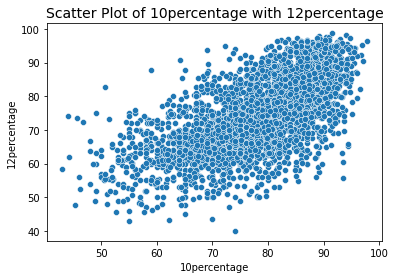
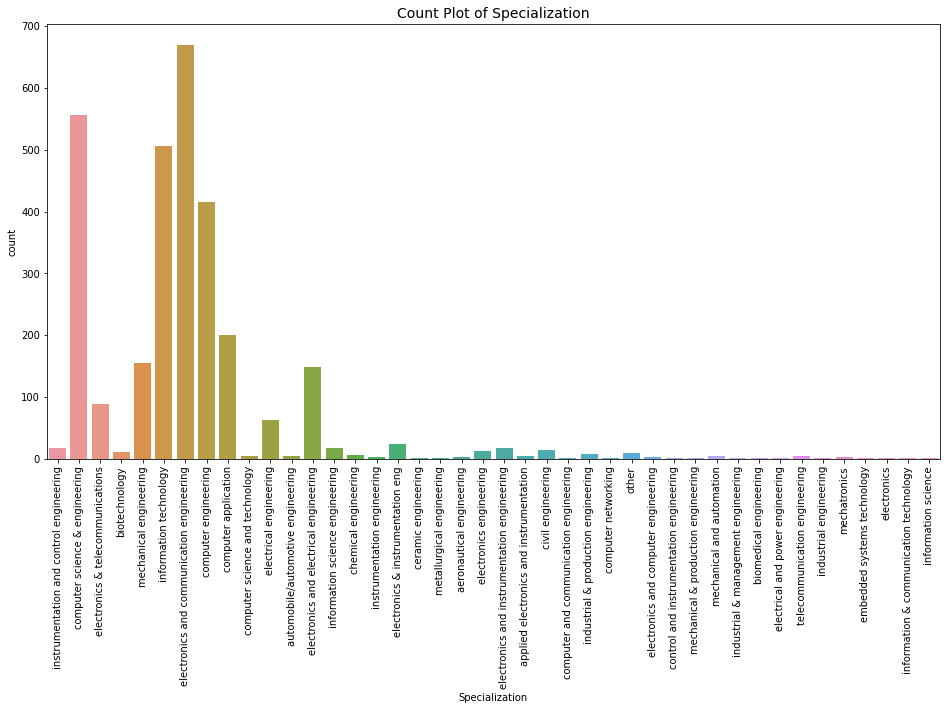
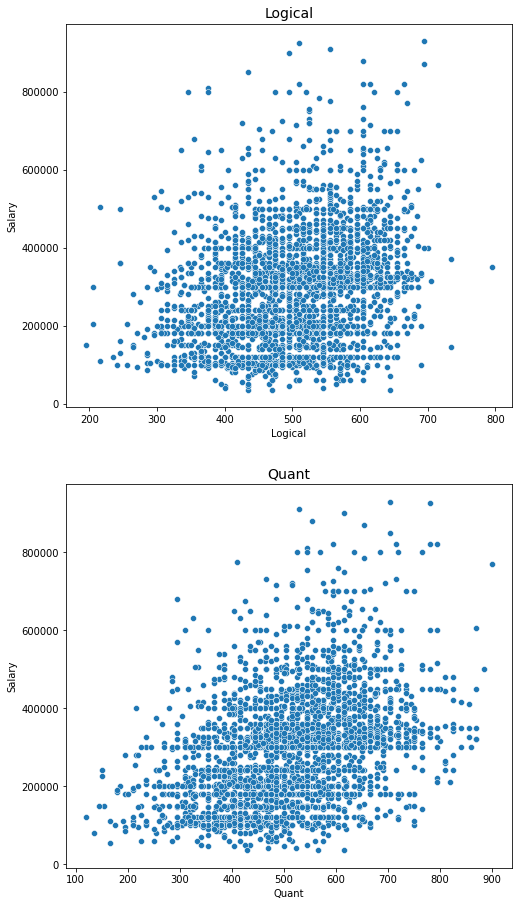
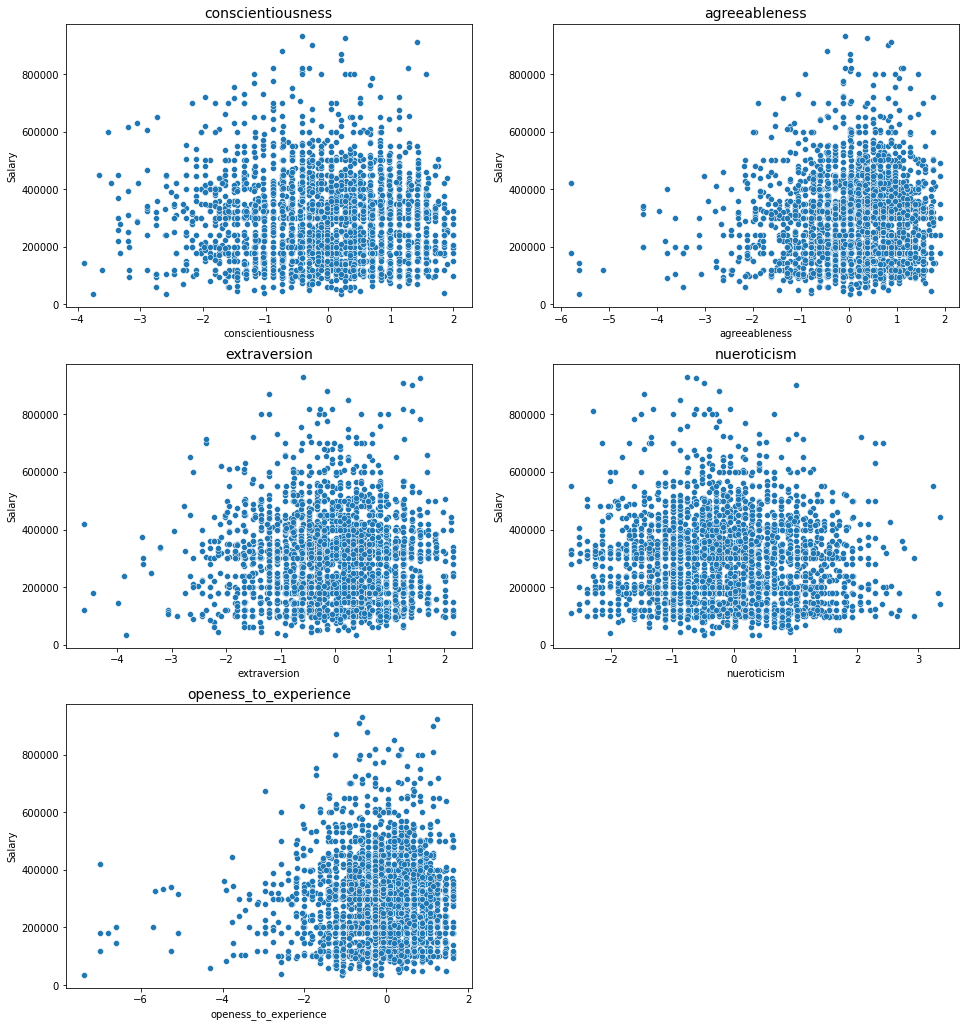
**This block diagram shows that data model takes dataset as input and can perform multiple n techniques like regression etc** **, and it can work on many attributes as much as user want and provide the result on the basis of previous data and predict the salary**

**FUTURE SCOPE**

AS of now we have only implemented lasso and linear regression , but in future we can implement as many new techniques of regression or some other technique .

Also we have tested our model with some combinations of attribute , but we can use N no. of combinations of attribute which can produce different outputs and more accuracy y can be achieved.

**Result**



R2 using linear regression is 0.2431864015668369

R2 using lasso regression is 0.22784863828694957

R2 using ridge regression is 0.2287996838288504

Root mean square error using linear regression is 0.13392134348770604

Root mean square error using lasso regression is 0.15431713100982916

Root mean square error using ridge regression is 0.13517267240707206

We can say that

the linear regression is providing better result

with r2 score of 24.31 and with minimum root mean square error of 0.1339

The lasso and ridge Regression model is providing the nearly same value but less than linear.

so we will refer linear regression method to predict the salary of the employees.