SALARY PREDICTION USING MACHINE LEARNING

A Project

Submitted in partial fulfillment of the requirements for

The award of the Degree of

BACHELOR OF COMPUTER APPLICATIONS

BY

ARPAN BISWAS

ROLL NO – 12020004006033 AND REGISTRATION NO – 203661001210025

ASUTOSH SAHA

ROLL NO – 12020004006037 AND REGISTRATION NO – 203661001210023

AVIK SARKHEL

ROLL NO – 12020004006040 AND REGISTRATION NO – 203661001210043

AVILASH SENGUPTA

ROLL NO – 12020004006041 AND REGISTRATION NO – 203661001210013

DEBOSMITA KONER

ROLL NO – 12020004006053 AND REGISTRATION NO – 203661001210087

DEPARTMENT OF COMPUTER APPLICATION

INSTITUTE OF ENGINEERING & MANAGEMENT

2022

DECLARATION CERTIFICATE

This is to certify that the work presented in the thesis entitled **"SALARY PREDICTION USING MACHINE LEARNING"** in partial fulfillment of the requirement for the award of degree of **Bachelor of Computer Application** of Institute of Engineering & Management is an authentic work carried out under my supervision and guidance.

To the best of my knowledge the content of this thesis does not form a basis for the award of any previous Degree to anyone else.

Date:

Prof. Manab Kumar Das

Dept. of Computer Application

Institute of Engineering & Management

Dr. Abhishek Bhattacharya

Head of the Department

Dept. of Computer Application & Science

Institute of Engineering & Management

CERTIFICATE OF APPROVAL

The foregoing thesis entitled **"SALARY PREDICTION USING MACHINE LEARNING"** is hereby approved as a creditable study of research topic and has been presented in a satisfactory manner to warrant its acceptance as prerequisite to the degree for which it has been submitted.

It is understood that by this approval, the undersigned do not necessarily endorse any conclusion drawn or opinion expressed therein, but approve the thesis for the purpose for which it is submitted.

(Internal Examiner) **(External Examiner)**

Acknowledgements

We would like to express our special thanks of gratitude to our Guide **Prof. Manab Kumar Das** who helped us a lot in this project, his valuable suggestions helped us to solve tough challenges and without his help this project could not have been completed in time. A special thanks to our Head of Department Prof. Abhishek Bhattacharya who gave us the golden opportunity to do this wonderful project on the topic **"SALARY PREDICTION USING MACHINE LEARNING"**, which helped us to gain a significant knowledge in the aforesaid subjects. Secondly, we would like to thank our friends who helped us a lot in finalizing this project within the given time frame.

Name of Student: Arpan Biswas

Enrollment Number: 12020004006033

Name of Student: Asutosh Saha

Enrollment Number: 12020004006037

Name of Student: Avik Sarkhel

Enrollment Number: 12020004006040

Name of Student: Avilash Sengupta

Enrollment Number: 12020004006041

Name of Student: Debosmita Koner

Enrollment Number: 12020004006053

Contents

Abstract

In our project, we are developing a salary prediction model using machine learning regression technique. Our project comprises of 2 parts. We will take an experimental dataset with 3 features and apply multiple linear regression and multiple polynomial regression for salary prediction and compare the accuracy scores. The method which proves to be efficient will be selected to proceed with developing the salary prediction model. We will train the model with a csv file which has 3 features that are qualification, work experience and age. We are planning to do 2 testing phases. First, we will test the model with the same dataset that was used to train the model and study on the accuracy and error. Then we will do another test by applying another dataset containing some random testcases for checking the predicted values. By studying the overall results of these two studies, we will be selecting a regression technique, linear or polynomial regression to proceed with the building of our model. The model will take a dataset as input having the columns Qualification, DOB, and Experience. It also needs an Employee ID of any employee in the database as input. It will then fetch the data from database and predict salary. In our minor project, we will be covering up to training model with linear and polynomial regression, checking and comparing accuracy and error and the rest will be covered in major project. We have studied 5 papers to gather knowledge that would help us to do our project.

Keywords: Machine Learning, Linear, Polynomial, Regression, Salary

Chapter 1

1.1 Introduction

Machine Learning is a technique or algorithm by which the computer systems understand what to do with the data present without needing explicit instructions by user**[2]**. Machine Learning is basically used to predict or analyze data based on a series. It is a category of AI (Artificial Intelligence). It analyzes and interprets data pattern and enables our computer system to learn, predict and decide like human beings. ML models are applied to solve real life problems which reduces human effort to do them**.** Use of machine learning reduces the cost incurred to a task and also increases accuracy**[3]**. They include stock price prediction, weather prediction, trend analysis, sentiment analysis, email classification and business chatbots. Many programmers have developed various ML models for salary prediction. Some of them include prediction using job position level, while some used years of work experience to predict. Regression Analysis depicts the relationship between the independent variables and dependent variables. There are mainly 2 types of regression techniques, Linear Regression and Polynomial Regression. Many prediction models have been made by regression analysis as mentioned above. Apart from these, linear regression has been used to estimate football player market values. There author used Multiple Linear Regression to estimate values in paper **[5]**. In Polynomial Regression, we use the linear model with modification that is increasing the power of variables to form a curved line. Many papers have also drawn comparison between linear and polynomial regression using their datasets. In paper **[1]**, the authors have predicted COVID-19 cases using linear and polynomial regression with degree 2, 3 and 4.

We have selected salary prediction because salary is an important factor of employment. Salary should be decided very logically and depends on numerous factors which is a tedious and costly process. Salary prediction models predicts salary based on input factors given and also provides good accuracy. Our model uses more than 1 (3) features to predict salary. Salary is required to be decided for an employee often, when a person joins the company, when he/she gets promotion, when his/her department changes. These salary predictions are made on various features, but we have selected 3 key areas of those features qualification, experience and age as education level, skills acquired and life experiences are the major factors that make a person employable and hence affect their salary in organization. It is believed that an employee's salary depends on how hard he works and that comes with experience**[4]**. In our project, we have carried out a comparative study of two regression techniques and we have observed which method gives the accurate and reasonable predicted salary values and then we will be developing our model on that technique. We have used an employee database upon which we will be running our model to present a demo. The dataset used will be taking the employee ID from database and predict the salary for that particular employee.

Chapter 2

2.1 Background Studies

The main focus of our project is to develop a salary prediction model using machine learning regression techniques. For this, we will be choosing linear or polynomial regression, which fits best with the training dataset to develop our model. For this, we are going to take an experimental dataset, apply linear regression and polynomial regression and comparing their accuracy, error and also we are going to apply a dataset of some random testcases to observe the predicted salary values. We will also observe the advantages of one technique over the other and proceed to develop our model with that technique. Our model will take the database at runtime, take employee ID as input and predict the salary for that particular employee using a regression technique.

2.2 Literature Survey

In paper [1] the authors have predicted the impact of COVID-19 in near future using linear regression, polynomial regression with a number of degrees. The authors have used polynomial regression with degree 2, 3, 4, 5 along with linear regression and have also presented the errors and accuracy of each model. It is inferable from the Table VI that polynomial regression with has the highest accuracy among all. From this we saw that polynomial regression with degree 5 gives the highest accuracy followed by degree 4, 3, 2 and linear regression.

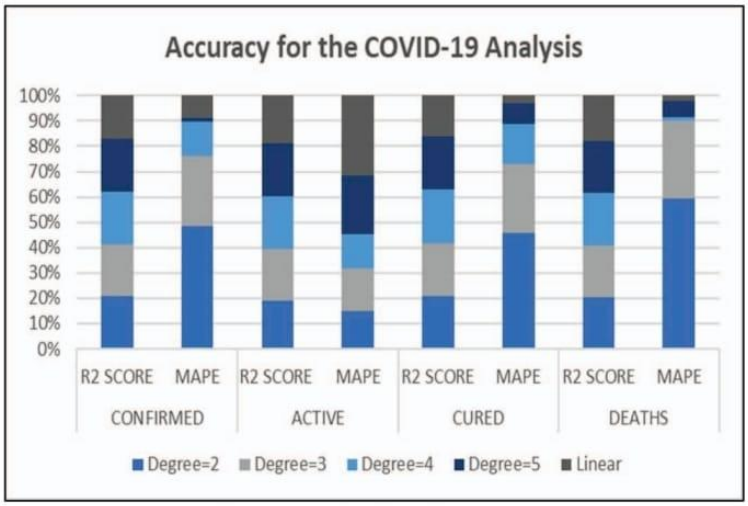


FIGURE 1

The FIGURE 1 is taken from the survey [1] as reference and shows performance evaluation of the regression models.

In the paper [2], the authors have made a study that examined the effects of demographics, personality traits, academic performance and standardized test scores on starting salaries. The results also showed that cognitive skills such as English language and Quantitative skills contributed significantly to the starting salaries of Indian Engineering graduates. In this report, we have come across four supervised Machine Learning algorithms, namely- Linear Regression, Decision Tree, Random Forest and Lasso Regression, and compared their proficiency in terms of accuracy. The study shows that the statistical analysis provided in the web application can help job seekers understand current and future trends of the job market in various sectors. The application also visualizes and provides industry-centric information regarding future employment opportunities. The Regression Analysis used in this study, made it possible to find independent variables that influence the dependent variable.

In paper [3] by Yasser T. Matbouli and Suliman M. Alghamdi, a holistic framework is developed to predict labour salary for all job titles in the Saudi Arabian economy, across all economic activities and organizational sizes, by training limited survey data with statistical machine learning (ML). This Salary Prediction process is done through three steps. At first, a holistic framework is introduced to develop a comprehensive set of all occupations and economic activities. Then, prediction models with selected features are developed. Finally, five ML prediction models are applied to real world salary data for Saudi labour market. To select the intrinsic features that predict salary using ML, the process of organization's developing their pay scales is explored. Studies on predicting pay levels is also being surveyed by using multiple linear regression models and the statistical ML regression models are also reviewed. The methodology presented in this paper is fresh and some concepts in defining job titles and economic activities were adopted from the standards of international organizations, which makes it easy to adapt the newly introduced framework universally across different international job markets. This article's approach uses the international classification of economic activities and occupations as provided in ISIC4 and ISCO-08. Such use of international classifications makes this framework generalizable to other labour markets. This new approach allows more confounding variables to be considered when predicting salary. Use of statistical ML, can both reduces the cost of salary bench-marking and improves accuracy especially when estimating salary levels for similar occupations in different industries, or when estimating different occupations within the same sector.

In the paper [4] which was done by 'Tiasa Mukherjee' and 'MS. B. Satyasaivani'. Here the authors have predicted the salary according to years of experience and hard work. Their aim was to predict the salary by employee's hard work which comes with experience. They have used Linear Regression model of prediction in their paper. They have split the dataset into training and testing data, predicted the salary for the test data, compared actual salary to predicted salary and calculated the accuracy. The graph portrays the line of the relationship that passes closely by all the plotted points.

FIGURE 2

FIGURE 3

The FIGURE 2 and 3 represents the visualization of training and testing data and are taken as reference from the paper.

In paper [5], the author estimated the market value of football players in the forward position using multiple linear regression by including the physical and performance factors in 2017-2018 season. Then build a successful model of several criteria in different significance levels within the consideration of multicollinearity and homoscedasticity of football players. As a result, the achieved to build a regression model 0.10 significance level with 52 attributes, %20 MAPE and 0.86 adjusted R^2 value. Analysis the all seek relevant factors that affect the market value of a football players in forward position will be built a linear regression with multiple independent variables which are significant predictors of the outcome.

Chapter 3

3.1 Proposed Methodology

We have used a dataset titled “salarydata.csv”. The dataset contains 5 fields – Index, Educational Qualification, Work Experience, Age and Monthly Salary.

Model Selection:

1. **Feature Selection:** Features are extracted using location indices of feature columns stored as list of lists.
2. **Training:** Features and the given salary are used to train the model based on linear regression and polynomial regression. In polynomial regression, $y = c_0 + c_1x + c_2x^2 + \dots + c_nx^n$ for n degree polynomial for each feature and here we are using 3 features. For polynomial regression, we use fit_transform(features).
3. **Prediction:** We pass the features as parameter to predict method to which we want the salary to predict.
4. **Accuracy Calculation:** Finally, the accuracy will be calculated comparing the given salary and the predicted salary.

Prediction Model Explanation:

- Our model flow diagram given below represents the flow of a regression model
- At first a dataset is taken and **features** and the output, **given salary** in our case, is extracted
- The prediction model is presented as a box containing **TRAIN** and **PREDICT** modules
- It shows that the functions, object.fit(features, given salary) for training and object.predict(features) for prediction belongs to the same object model.

- The training module formulates a prediction function which is used by prediction module, it puts all the features of a particular feature list in formula and an output comes out as predicted output.
- Then the predicted output and the given output is passed in r2_score() method and it compares the 2 outputs and returns the accuracy score. The mean_squared_error() returns the error or deviation of predicted output from actual output.

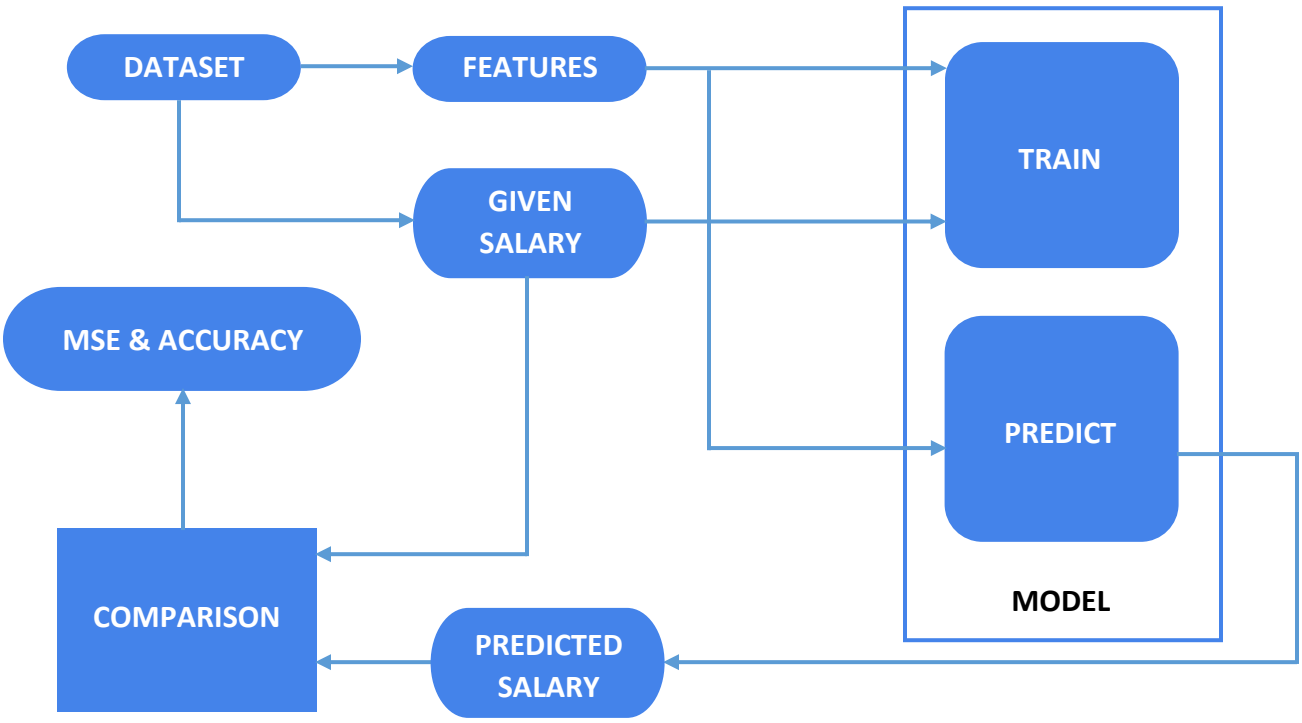


FIGURE 4

Model Development:

The model will be developed by selecting the regression technique which will be proved to be more efficient for our dataset. A csv file database will be used which will contain information about few employees. The database will be a hypothetical database. The model will be trained based on the dataset we are working on in our 1st module. The model will:

- Predict the salary extracting the features
- Generate graph showing the salaries of all the employees
- Update data whenever required

Summary of our model is functionality:

1. Training the model with the dataset "salarydata.csv"
2. Parsing a database which contains data of employees
3. Employee ID will be given as input to the model
4. Educational Qualification, Age and Work Experience corresponding to the ID will be fetched from the database
5. Salary will be predicted
6. Salary will be written to the database
7. Data can be updated in the database
8. Graph can be generated showing salary of all employees registered

Chapter 4

4.1 Experimental Dataset

Index	Qualification	Work Experience	Age	Monthly Salary
1	1	1.1	21	39343
2	1	1.3	21.5	46205
3	1	1.5	21.7	37731
4	1	2	22	43525
5	1	2.2	22.2	39891
6	1	2.9	23	56642
7	1	3	23	60150
8	1	3.2	23.3	54445
9	1	3.2	23.3	64445
10	1	3.7	23.6	57189
11	1	3.9	23.9	63218
12	1	4	24	55794
13	1	4	24.5	56957
14	1	4.1	24.5	57081
15	1	4.5	25	61111
16	1	4.9	25	67938
17	1	5.1	26	66029
18	2	1.5	21.7	75462
19	2	1.1	21	78686
20	1	5.3	27	83088
21	1	5.9	28	81363
22	2	2	22	87050
23	1	6	29	93940
24	1	6.8	30	91738
25	2	1.3	21.5	92410
26	1	7.1	30	98273
27	1	7.9	31	101302
28	1	8.2	32	113812
29	1	8.7	33	109431
30	2	3.2	23.3	108890
31	2	2.9	23	113284
32	1	9	34	105582
33	2	4	24	111588
34	1	9.5	35	116969
35	2	3.7	23.6	114378

36	1	9.6	36	112635
37	1	10.3	37	122391
38	2	3	23	120300
39	1	10.5	38	121872
40	2	3.2	23.3	128890

TABLE 1

Source: www.kaggle.com

Platform Used: Google Colab

Dataset Name: salarydata

Dataset Type: CSV

Programming Language: Python

In the following database given above, in the **Qualification** column, 1 represents **Graduate** and 2 represents **Post Graduate**

Chapter 5

5.1 Results and Discussions

In the above sections, we have carried out a testing on our dataset by applying Linear Regression and Polynomial Regression. For polynomial regression, we have not passed and degree parameter, hence the default degree is set which is 2. We have found the MSE (Mean Squared Error) and Accuracy (R2 Score) for each of the regression technique. We are giving the code snapshots below.

```
[1]  import pandas as pd
     import matplotlib.pyplot as plt
     from sklearn.linear_model import LinearRegression
     from sklearn.preprocessing import PolynomialFeatures
     from sklearn.metrics import mean_squared_error, r2_score
```

```
[2]  dataset = pd.read_csv('salarydata.csv')
     features = dataset.iloc[:,1:4].values
     given_salary = dataset.iloc[:,-1].values
     index = list(dataset['Index'])
```

```
[3]  lr = LinearRegression()
     lr.fit(features, given_salary)
     linear_predicted_salary = lr.predict(features)
     linear_predicted_salary
```

```
array([ 34784.92177897,  36654.88430127,  38652.61188008,  43732.10753145,
        45729.83511026,  52679.29328392,  53720.7454255 ,  55675.88465214,
        55675.88465214,  60755.38030351,  62710.51953015,  63709.38331955,
        63496.44155869,  64537.89370027,  68490.76050572,  72656.56907203,
        74313.58983346,  92142.30216722,  88274.61206612,  75970.61059489,
        81793.43992263,  97221.7978186 ,  82409.00854248,  90314.74215338,
        90144.57458841,  93439.09857811, 101344.83218901, 104043.30509202,
       108824.68227818, 109165.57493929, 106168.98357107, 111523.15518119,
       117199.0736067 , 116304.53236735, 114245.07059066, 116920.1009872 ,
       123784.38245652, 107210.43571265, 125441.40321795, 109165.57493929])
```

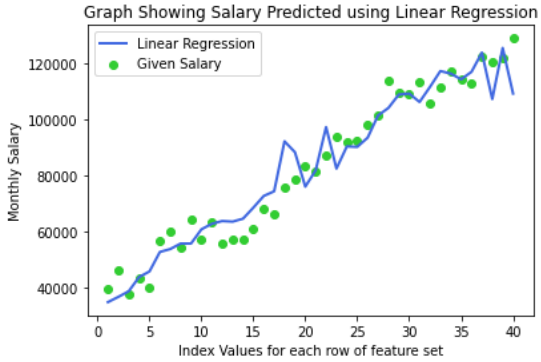
```
[4]  pf = PolynomialFeatures()
     lr.fit(pf.fit_transform(features), given_salary)
     polynomial_predicted_salary = lr.predict(pf.fit_transform(features))
     polynomial_predicted_salary
```

```
array([ 37357.26279114,  39892.6302907 ,  42095.40212561,  46467.37632388,
        48179.14647956,  54013.50519834,  53992.26135443,  55941.29508713,
        55941.29508713,  56754.80276797,  58554.5560959 ,  58932.23201536,
        63482.44571893,  63155.77446669,  65790.04858261,  62471.08336196,
        71844.69217537,  88765.37339231,  76516.4081615 ,  78870.02030702,
        84582.51527195,  94097.31414232,  88519.94159553,  94930.12887389,
        85005.94490975,  95405.54462715, 100095.33614168, 105438.68867604,
       110048.60252073, 114377.01032512, 110114.23546486, 113439.08605962,
       122128.73631024, 117169.94448727, 116150.4845577 , 116060.52496024,
       121459.3620413 , 109405.48241109, 119204.49451484, 114377.01032512])
```

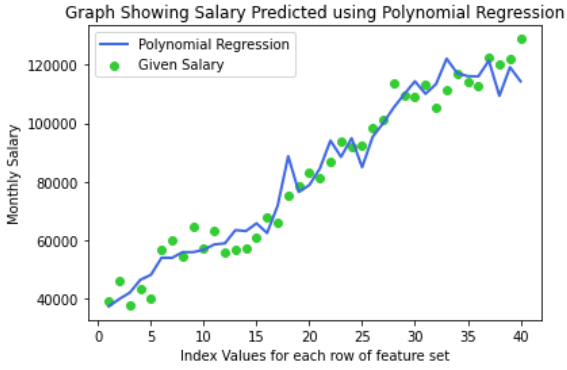
```
[5] print('linear regression mean squared error =',mean_squared_error(given_salary, linear_predicted_salary))
    print('linear regression accuracy score =',round(r2_score(given_salary, linear_predicted_salary),2))
    print('polynomial regression mean squared error =',mean_squared_error(given_salary, polynomial_predicted_salary))
    print('polynomial regression accuracy score =',round(r2_score(given_salary, polynomial_predicted_salary),2))
```

```
linear regression mean squared error = 52422910.8080892
linear regression accuracy score = 0.93
polynomial regression mean squared error = 36392964.88285049
polynomial regression accuracy score = 0.95
```

```
[6] plt.title('Graph Showing Salary Predicted using Linear Regression')
    plt.scatter(index, given_salary, color='limegreen')
    plt.plot(index, linear_predicted_salary, color='royalblue', linewidth=2)
    plt.legend(['Linear Regression','Given Salary'], loc='upper left')
    plt.xlabel('Index Values for each row of feature set')
    plt.ylabel('Monthly Salary')
    plt.show()
```




```
[7] plt.title('Graph Showing Salary Predicted using Polynomial Regression')
    plt.scatter(index, given_salary, color='limegreen')
    plt.plot(index, polynomial_predicted_salary, color='royalblue', linewidth=2)
    plt.legend(['Polynomial Regression','Given Salary'], loc='upper left')
    plt.xlabel('Index Values for each row of feature set')
    plt.ylabel('Monthly Salary')
    plt.show()
```



From executing the above code, we have found the following observations and differences between linear and polynomial regression. Following table below, presents the mean squared error and r2 score for each of the regression techniques used by us.

Regression Technique	Mean Squared Error	R2 Score
Linear Regression	52422910.80808920	0.93
Polynomial Regression	36392964.88285049	0.95

TABLE 2

Clearly, we can see that the Mean Squared Error for Linear Regression is 16029945.92523871 more than that of Polynomial Regression. As Polynomial regression produces a curved line rather than a straight line, the chances of meeting with the given datapoints increases which is evident from the fact that R2 Score for Polynomial Regression is greater than that of Linear Regression. Though in the graph generated by us, the LR line is not a straight line, but it comprises of multiple straight lines joined together. But in PR, the lines are slightly curved due to degree=2.

Also, in paper [1], we have seen that the polynomial regression, gives a higher accuracy and lower error compared to what linear regression gives. Now in Major Project, we will take another dataset with few testcases and test those feature values with our model and check the salary values and will study the predicted salary values for both the regression techniques. Then with the final conclusion we will develop our salary prediction models.

Chapter 6

6.1 Conclusion

In this paper, we carried out a study between Linear Regression and Polynomial Regression. We discussed which technique is more efficient compared to the other and found out that polynomial regression is more efficient, has lower error and has higher accuracy. But as we tested on the same dataset, the conclusion is partial. After testing with another dataset, we can conclude which model gives flawless salary values and will proceed with that.

6.2 Future Work

As we saw that polynomial regression is more efficient than linear regression for prediction on the same dataset as that of training dataset, we will test the models with a separate dataset and then will proceed to build our salary prediction model using the regression technique which proves to be flawless with Qualification, Age and Work Experience as features.

6.3 References

1. Saud Shaikh; Jaini Gala; Aishita Jain; Sunny Advani; Sagar Jaidhara; Mani Roja Edinburgh, **Analysis and Prediction of COVID-19 using Regression Models and Time Series Forecasting**, 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, Date of Conference: 28-29 January 2021, INSPEC Accession Number: 20565476, Electronic ISBN:978-1-6654-1451-7, DOI: 10.1109/Confluence51648.2021.9377137, Noida India, pp 989 – 995.
2. Jayshree, Sakshi, Janhavi, Jahnavi, Krupa (2022), **Salary Estimator: A Literature Review. International Research Journal of Computer Science (IRJCS)**, AM Publications, Volume IX, Issue V, 2022, pp 101-105.
3. Yasser T. Matbouli, Suliman M. Alghamdi, **Statistical Machine Learning Regression Models for Salary Prediction Featuring Economy Wide Activities and Occupations**, Special Issue: Predictive Analytics and Data Science, Academic Editor: Agnes Vathy-Fogarassy, Information 2022, 13(10), 495, Published: 12 October 2022.
4. Tiasa Mukherjee, MS. B. Satyasaivani, **Employee's Salary Prediction**, Volume-8, Issue 3 - V8I3-1375, IJARIIT, ISSN: 2454-132X, Imapct Factor: 6.078, pp 356-359
5. Yunus Koloğlu, Hasan Birinci, Sevde Ilgaz Kanalmaz, Burhan Özyılmaz, **A Multiple Linear Regression Approach For Estimating the Market Value of Football Players in Forward Position**, ABDULLAH GÜL UNIVERSITY INDUSTRIAL ENGINEERING DEPARTMENT