SALARY PREDICTION OF IT EMPLOYEES IN JAVA USING LINEAR REGRESSION ALGORITHM

Rudy Chandra[1], Tegar Arifin Prasetyo[2], Feronika Simanjuntak[3], Scintya Leony[4], Geraldine Lumban Tobing[5], Sarbaini[6]
[1,2,3,4,5] Information Technology, Faculty of Vocational Studies, Del Institut of Technology
[6] Jurusan Matematika, Fakultas Sains dan Teknologi, Universitas Islam Negeri Sultan Syarif Kasim
email: [1]rudychandra@del.ac.id, [2]tegar.prasetyo@del.ac.id, [3]if320057@students.del.ac.id, [4]if320005@students.del.ac.id, [5]if320043@students.del.ac.id, [6]sarbaini@uin-suska.ac.id

Abstract

The payroll system is very influential on a company's workers' welfare in achieving company goals. Appropriate payroll will build morale for the workforce so that they can advance the company through the work ethic and professionalism of the crew. The salary calculation system for employees must be adjusted to several criteria, such as their city and job role. Long experience can also be used as a calculation criterion in providing salary. For this reason, an approach is needed to provide a decent and good salary prediction for the company's consideration. One of the models commonly used in making predictions is linear regression. Linear regression is a model that calculates the relationship between two variables with independent values and bound data. This research develops a system by implementing a Linear Regression algorithm to predict the salaries of IT employees in Java. The model that has been created is then built using the Python language and implemented into a website-based visualization form that is easy to understand with Streamlit. The modeling results were tested and gave an MSE value of 8240258.48. This research is expected to be a reference in research related to this topic in the future and can be used by companies that have difficulties in determining employee salaries.

Keywords: *Linear regression, Model, MSE, Salary, Website*

Abstrak

Sistem penggajian sangat berpengaruh pada kesejahteraan tenaga kerja dalam suatu perusahaan dalam mencapai tujuan perusahaan. Penggajian yang sesuai akan membangun semangat kerja bagi tenaga kerja sehingga mampu memajukan perusahaan melalui etos kerja dan profesionalisme para tenaga kerja. Sistem penghitungan gaji pada karyawan harus disesuaikan dengan beberapa kriteria seperti kota dan role pekerjaan yang dimiliki. Lama pengalaman juga bisa dijadikan sebuah kriteria penghitungan dalam memberikan gaji. Untuk itu diperlukan suatu pendekatan yang dapat memberikan prediksi gaji yang layak dan baik untuk menjadi pertimbangan perusahaan. Salah satu model yang umum digunakan dalam melakukan prediksi adalah regrei linear. Regresi linear merupakan model yang menghitung hubungan antar dua variabel dengan nilai yang bebas dengan data yang terikat. Penelitian ini mengembangkan sebuah sistem dengan mengimplementasikan algoritma Regresi Linear untuk memprediksi gaji karyawan IT di pulau Jawa. Model yang telah dibuat kemudian dibangun menggunakan bahasa Python dan diimplementasikan ke bentuk visualisasi yang mudah dipahami berbasis website dengan Streamlit. Hasil pemodelan *diuji dan memberikan nilai MSE sebesar 8240258.48. Penelitian ini diharapkan mampu menjadi rujukan dalam penelitian yang berkaitan dengan topik ini di masa depan dan dapat digunakan perusahaan yang memiliki kesulitan dalam menentukan gaji karyawan.*

***Kata kunci**: Regresi Linear, Model, MSE, Gaji, Website*

1. INTRODUCTION

In this modern era, the role of technology is significant. Technology can help shorten a person's working time. It happens because everyone worldwide can connect with the help of technological advances. Indonesia is one of the countries that enjoy this progress. In line with this, digital talent does need to continue improving Indonesia's technology sector[1].

Human resources are one of the critical factors in the progress of a company[2]. Intense competition between companies requires increased performance in finance, technology, natural resources, and human resources[3]. Along with the times, the types of work needed by a company will also be increasingly diverse, especially in companies in the IT field. The increasing need for workers in the technology sector has created new problems, namely in determining salaries according to experience, company region, and the burden of responsibility for each digital talent[4].

Salary can affect employee performance because it is not uncommon for employees to strike[5], [6]. After all, the compensation they receive does not match their work. Therefore, it is a consideration for every company to pay attention to the wages and social benefits expected of employees[7] [8].

The payroll system is very influential on a company's workers' welfare in achieving company goals. Appropriate payroll will build morale for the workforce so that they can advance the company through the work ethic and professionalism of the crew. They have to adjust the salary calculation system for employees to several criteria, such as their city and job role. Long experience can also use as a calculation criterion in providing salary. The longer work experience an employee has, the more significant receive the compensation[9].

Because of the problems often experienced, it is necessary to have a system that can provide a decent salary prediction based on several criteria. The expected system can use a linear regression model. Linear regression models the relationship between variables in anonymous data using known and corresponding data values[10]. Therefore, implementing a Linear Regression algorithm to predict the salaries of IT employees in Java according to criteria such as length of experience working in the IT field, roles in the IT field, and the intended company region are used to do. The model created is then built in Python and implemented on the website with Streamlit to make it easy to use.

2. RESEARCH METHODOLOGY

2.1. Research Design

The method used in this study can be seen in Figure 1.

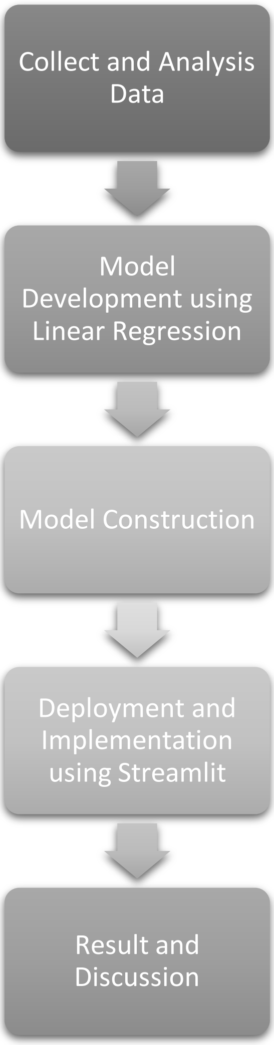

Collect and Analysis Data
Model Development using Linear Regression
Model Construction
Deployment and Implementation using Streamlit
Result and Discussion

Figure 1. Research Design

1. Data collection and data analysis
   Data collection is done by taking the appropriate dataset from several sources such as kaggle and github. The data that has been collected will then be analyzed. Data analysis aims to arrange the order of the data, organize it into a pattern, category, and a basic description[11].
2. Model construction with linear regression
   The data that is analyzed will then be used in building a model. Finally, the model is built using linear regression. Linear regression is a data

acquisition technique that explores the relationship between the independent variable (X) and the dependent variable (Y)[12].

3. Deploy model
   Models that have been built previously will be deployed into a web so that users can use them in real terms.
4. Implementation
   Web predictions are built using the Python library, namely Streamlit. Streamlit enables interactive and user-friendly web development. The created web will have several input parameters, which will then be used as test data in predicting salaries. The independent variables used are city, role, and years of experience. This variable will affect the salary prediction that will be displayed to the user.
5. Discussion
   At this stage, we will discuss salary specifications, and conclusions will be obtained, from model development to implementation and salary prediction results.

2.2. Linear Regression

Linear regression is an algorithm that models the relationship between data with the dependent and independent variables. The use of linear regression is to make predictions using previously available data. Linear regression looks for relationships between variables to find straight-line equations which can be seen in Figure 2.

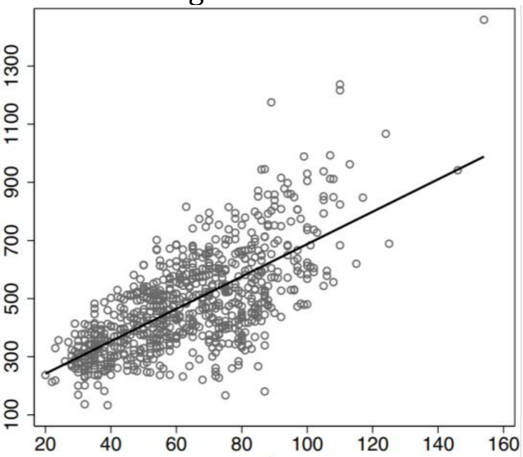


Figure 2. Linear Regression straight-line equations

Linear regression in this study involves one independent variable, commonly called univariate. In addition, one independent has its variable, so there are only two input variables, namely X and output Y. These two variables can be described as Equation 1 below. [13]

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (1)$$

In information $\hat{\beta}_0$ is an intercept where the value changes, namely variable y, when x is 0. Information of $\hat{\beta}_1$ is a constant that shows a shift in variable y in every increase of 1 variable x [14]

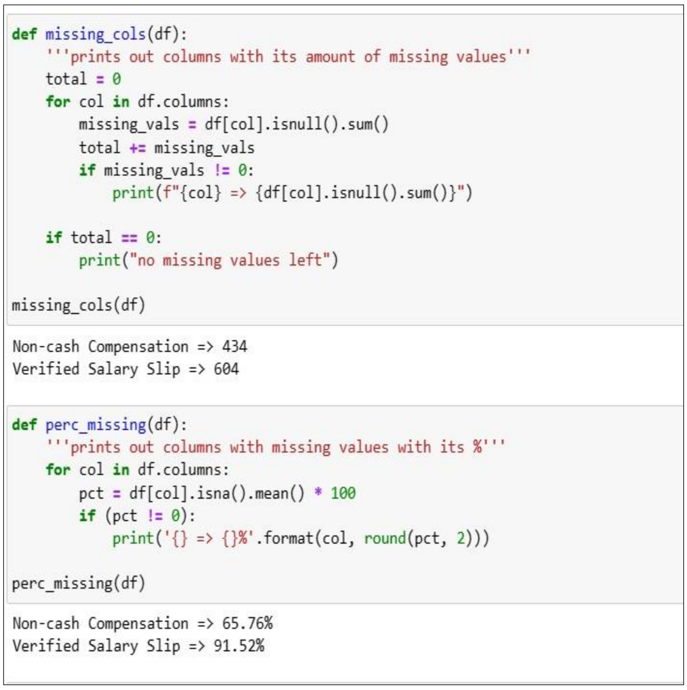
3. RESULT AND DISCUSSION

3.1. Data Analysis

In data analysis, we have to conduct data cleaning and data transformation. The data used was sourced from: https://predictsalary.com/salaries[15]. At this stage, two approaches will carry out to get the best data, namely by cleaning the data and transforming the data.

3.1.1. Data Cleaning

Data Cleaning is the process of repairing the data structure in preparation before the data is processed further. In developing this model, data cleaning is used to handle missing values and repair data structures that have incorrect spelling. The data frame is used to handle missing values through the data frame containing the columns used to predict salary. The columns used to become data in model development consist of roles, cities, and years of experience, which will be used as the X variable or what is called the independent variable. Salary data will be used as Y or a variable influenced by X. Figure 3 shows the search for missing values in the data column.



```
def missing_cols(df):
    '''prints out columns with its amount of missing values'''
    total = 0
    for col in df.columns:
        missing_vals = df[col].isnull().sum()
        total += missing_vals
        if missing_vals != 0:
            print(f"{col} => {df[col].isnull().sum()}")

    if total == 0:
        print("no missing values left")

missing_cols(df)
```

```
Non-cash Compensation => 434
Verified Salary Slip => 604
```

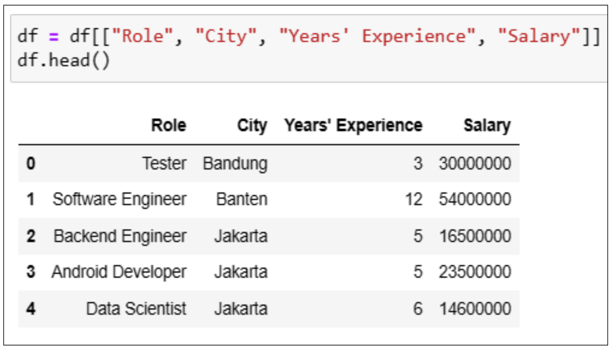
```
def perc_missing(df):
    '''prints out columns with missing values with its %'''
    for col in df.columns:
        pct = df[col].isna().mean() * 100
        if (pct != 0):
            print('{} => {}%'.format(col, round(pct, 2)))

perc_missing(df)
```

```
Non-cash Compensation => 65.76%
Verified Salary Slip => 91.52%
```

Figure 3. Missing Value

Dataframes are used to overcome missing values. These data are used as X variables and Y variables, as shown in Figure 4.

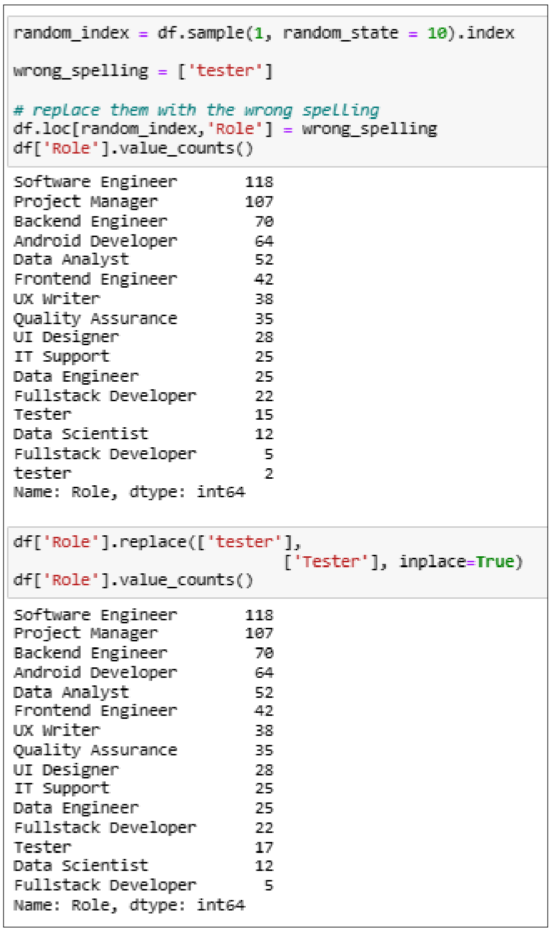


```
df = df[["Role", "City", "Years' Experience", "Salary"]]
df.head()
```

	Role	City	Years' Experience	Salary
0	Tester	Bandung	3	30000000
1	Software Engineer	Banten	12	54000000
2	Backend Engineer	Jakarta	5	16500000
3	Android Developer	Jakarta	5	23500000
4	Data Scientist	Jakarta	6	14600000

Figure 4. Variabel X and Variabel Y

The data structure that has spelling errors in the categories that are fixed can be seen in Figure 5 below.



```
random_index = df.sample(1, random_state = 10).index

wrong_spelling = ['tester']

# replace them with the wrong spelling
df.loc[random_index,'Role'] = wrong_spelling
df['Role'].value_counts()
```

```
Software Engineer       118
Project Manager         107
Backend Engineer         70
Android Developer        64
Data Analyst             52
Frontend Engineer        42
UX Writer                38
Quality Assurance        35
UI Designer              28
IT Support               25
Data Engineer            25
Fullstack Developer      22
Tester                   15
Data Scientist           12
Fullstack Developer       5
tester                    2
Name: Role, dtype: int64
```

```
df['Role'].replace(['tester'],
                            ['Tester'], inplace=True)
df['Role'].value_counts()
```

```
Software Engineer       118
Project Manager         107
Backend Engineer         70
Android Developer        64
Data Analyst             52
Frontend Engineer        42
UX Writer                38
Quality Assurance        35
UI Designer              28
IT Support               25
Data Engineer            25
Fullstack Developer      22
Tester                   17
Data Scientist           12
Fullstack Developer       5
Name: Role, dtype: int64
```

Figure 5. Spelling Errors in Data

To see how the data compares between cities and salaries, you can see in Figure 6. Its shows which city has the highest salary at the same time with companies that have varying salaries.

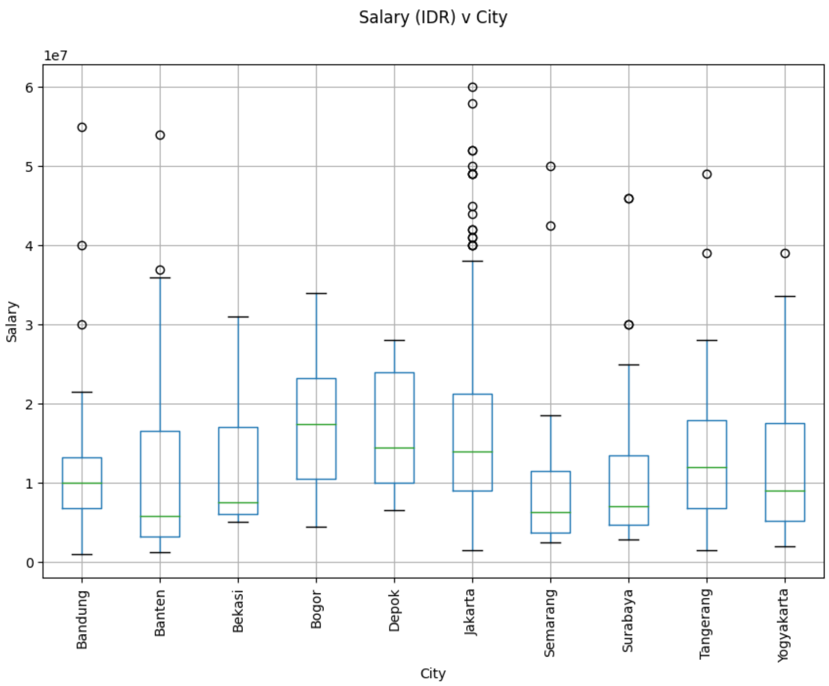
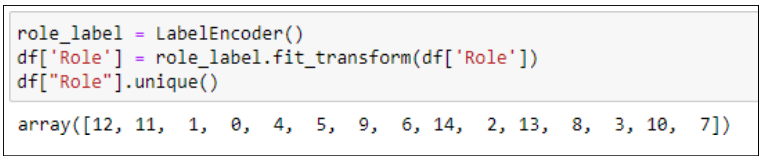


Figure 6. Salary Comparison by City

3.1.2. Data Transformation

Data transformation is the process of changing or consolidating data so that the process of using data can be applied or the data used is more efficient. The data used is encoded so that it can be processed by the algorithm to be used. The data that has been transformed can be seen in Figure 7.

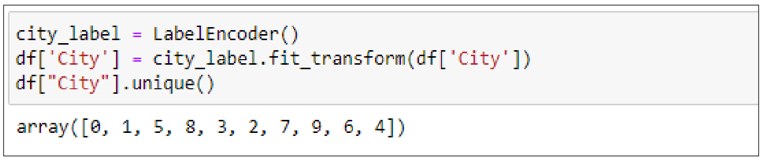


```
role_label = LabelEncoder()
df['Role'] = role_label.fit_transform(df['Role'])
df["Role"].unique()
```

```
array([12, 11,  1,  0,  4,  5,  9,  6, 14,  2, 13,  8,  3, 10,  7])
```

Figure 7. Data Transformation

City data transformation is attached in Figure 8.



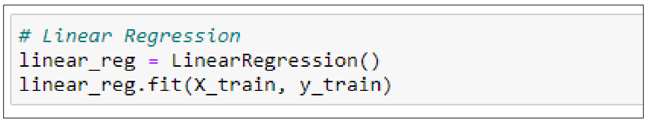
```
city_label = LabelEncoder()
df['City'] = city_label.fit_transform(df['City'])
df["City"].unique()
```

```
array([0, 1, 5, 8, 3, 2, 7, 9, 6, 4])
```

Figure 8. City Data Transformation

3.2. Model Development

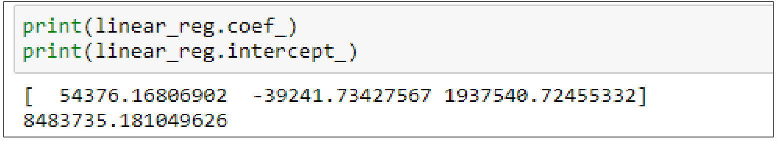
The model-building process is carried out using linear regression, which can be seen in the following code snippet in Figure 9.



```
# Linear Regression
linear_reg = LinearRegression()
linear_reg.fit(X_train, y_train)
```

Figure 9. Model Development Using Linear Regression

The results of the development of the model obtained intercept values and several coefficients that will be used to find Y values or predict salaries that can be seen in Figure 10.

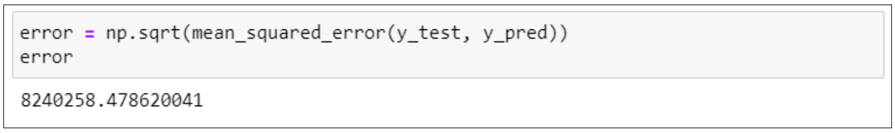


```
print(linear_reg.coef_)
print(linear_reg.intercept_)
```

```
[  54376.16806902  -39241.73427567 1937540.72455332]
8483735.181049626
```

Figure 10. Coefficient and Intercept Value

After obtaining the coefficient values and intercept values that will be used in making predictions, then look for the MSE (Mean Squared Error) value. The results of the MSE value are said to be very good if the value obtained is less than 10%[16]. The MSE value of the built model can be seen in the following Figure 11 below.



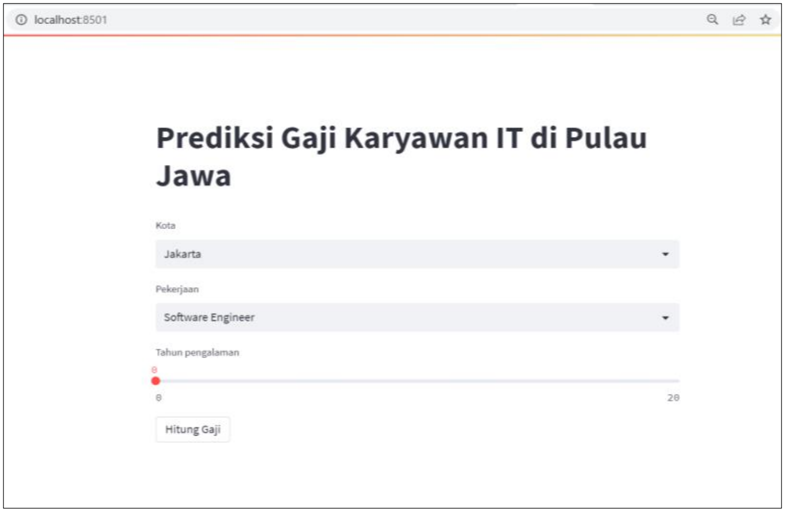
```
error = np.sqrt(mean_squared_error(y_test, y_pred))
error
```

```
8240258.478620041
```

Figure 11. Mean Squared Error Value

3.3. Interface Implementation

Models that have been built will be deployed so that they can interact directly with users. Users can send data and receive predictions via the web. The web shown is built using the Streamlit library. The implementation of the user interface can be seen in the following Figure 12.



Prediksi Gaji Karyawan IT di Pulau Jawa

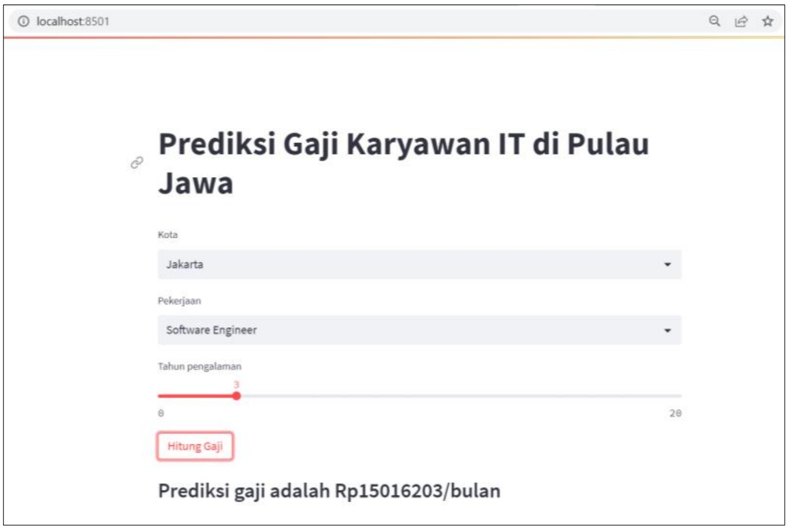
Kota
Jakarta

Pekerjaan
Software Engineer

Tahun pengalaman
0

Figure 12. Salary Prediction Interface

The user will choose a city on the island of Java according to the data provided. Then the role is selected according to the user's work. This interface also offers a range of years of user experience in the chosen function. After selecting according to the city, job or position, and length of experience, the user will receive a prediction of the salary they will receive per month. The prediction results displayed can be seen in the following Figure 13.



localhost:8501

Prediksi Gaji Karyawan IT di Pulau Jawa

Kota: Jakarta

Pekerjaan: Software Engineer

Tahun pengalaman: 3

Prediksi gaji adalah Rp15016203/bulan

Figure 13. Prediction Result

4. CONCLUSION

The system built using a linear regression algorithm can predict IT employees' salaries on Java Island. This research was successfully applied to predict salaries Using city criteria, job names, and work experience by users through an interface built using the Streamlit library. The city of Jakarta is the city that has the most variations in compensation for employees in the IT sector. The research was conducted in other cities with many criteria. To measure the error rate of the linear regression model is to use the MSE, which is obtained in the model with a value of 8240258.48

DAFTAR PUSTAKA

[1] U. Bansal, A. Narang, A. Sachdeva, I. Kashyap, and S. P. Panda, "Empirical analysis of regression techniques by house price and salary prediction," in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing Ltd, Jan. 2021. doi: 10.1088/1757-899X/1022/1/012110.

[2] K. K. Rekayasa, M. A. Saputra, N. Prasetyo, I. Zulfikar, T. Rijanandi, and F. Dharma Adhinata, "Terbit online pada laman web jurnal: http://journal.ittelkom-pwt.ac.id/index.php/dinda Journal of Dinda Prediksi Gaji Berdasarkan Pengalaman Bekerja Menggunakan Metode Regresi Linear," *Data Institut Teknologi Telkom Purwokerto*, vol. 2, no. 2,

pp. 58–63, 2022, [Online]. Available: http://journal.ittelkom-pwt.ac.id/index.php/dinda

[3] V. S. Reddy, "Impact of Compensation on Employee Performance," *IOSR Journal of Humanities And Social Science (IOSR-JHSS*, vol. 25, no. 9, pp. 17–22, 2020, doi: 10.9790/0837-2509011722.

[4] J. Mainert, C. Niepel, K. R. Murphy, and S. Greiff, "The Incremental Contribution of Complex Problem-Solving Skills to the Prediction of Job Level, Job Complexity, and Salary," *J Bus Psychol*, vol. 34, no. 6, pp. 825–845, Dec. 2019, doi: 10.1007/s10869-018-9561-x.

[5] S. Ramlall, "A Review of Employee Motivation Theories and their Implications for Employee Retention within Organizations."

[6] M. H. Abu Hassan Asaari, N. Mat Desa, and L. Subramaniam, "Influence of Salary, Promotion, and Recognition toward Work Motivation among Government Trade Agency Employees," *International Journal of Business and Management*, vol. 14, no. 4, p. 48, Mar. 2019, doi: 10.5539/ijbm.v14n4p48.

[7] P. Khongchai and P. Songmuang, *Implement of Salary Prediction System to Improve Student Motivation using Data Mining Technique*.

[8] A. Indriyani, "Analisis pengaruh gaji dan tunjangan kesejahteraan terhadap produktivitas kerja karyawan operation department PT. Export Leaf Indonesia," *Paradigma*, vol. 12, no. 01, pp. 41–56, 2014.

[9] S. Kalogiannidis, "IMPACT OF EMPLOYEE MOTIVATION ON ORGANIZATIONAL PERFORMANCE. A SCOPING REVIEW PAPER FOR PUBLIC SECTOR." [Online]. Available: www.strategicjournals.com

[10] K. Kumari and S. Yadav, "Linear regression analysis study," *Journal of the Practice of Cardiovascular Sciences*, vol. 4, no. 1, p. 33, 2018, doi: 10.4103/jpcs.jpcs_8_18.

[11] R. Chandra, "Wood Classification For Efficiency in Preventing Illegal Logging Using K-Nearest Neighbor," 2022.

[12] D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 1, no. 4, pp. 140–147, Dec. 2020, doi: 10.38094/jastt1457.

[13] L. Fahrmeir, T. Kneib, S. Lang, and B. D. Marx, *Regression*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2021. doi: 10.1007/978-3-662-63882-8.

[14] S. K. Prion and K. A. Haerling, "Making Sense of Methods and Measurements: Simple Linear Regression," *Clin Simul Nurs*, vol. 48, pp. 94–95, Nov. 2020, doi: 10.1016/j.ecns.2020.07.004.

[15] "PredictSalary - Salaries Information." https://predictsalary.com/salaries (accessed May 30, 2023).

[16] T. Indarwati, T. Irawati, and E. Rimawati, "PENGGUNAAN METODE LINEAR REGRESSION UNTUK PREDIKSI PENJUALAN SMARTPHONE," *Jurnal Teknologi Informasi dan Komunikasi (TIKomSiN)*, vol. 6, no. 2, Jan. 2019, doi: 10.30646/tikomsin.v6i2.369.