Sistem Prediksi Gaji Karyawan Berdasarkan Lama Waktu Kerja Menggunakan Metode *Linear Regression*

Risqi Wahyu Permana[1], Herliansyah Bagus Priambodo[2]

[1]Program Studi Teknik Informatika, Universitas Surabaya, Surabaya, Jawa Timur
Email: [1] s160419081@student.ubaya.ac.id, [2] s160419082@student.ubaya.ac.id

(Naskah masuk: dd mmm yyyy, direvisi: dd mmm yyyy, diterima: dd mmm yyyy)

Abstrak

Besar gaji seorang karyawan seringkali dipengaruhi oleh lama waktu kerja. Semakin lama seorang karyawan bekerja, biasanya gajinya akan mengalami kenaikan atau semakin tinggi, begitupun sebaliknya. Tujuan dari penelitian ini adalah untuk memprediksi besar gaji karyawan berdasarkan lama waktu kerja. Penelitian ini menggunakan metode regresi linear. Regresi Linear adalah salah satu *supervised machine learning* yang masuk dalam golongan *regression.* Regresi linear cocok dipakai ketika terdapat hubungan linear pada data. Regresi linear adalah teknik untuk memprediksi sebuah nilai dari variabel Y (variabel dependen) berdasarkan beberapa variabel tertentu X (variabel independen) jika terdapat hubungan linear antara variabel X dan Y. Pada penelitian ini studi kasus yang digunakan adalah prediksi gaji karyawan berdasarkan lama waktu kerja karyawan. Studi kasus ini memiliki hubungan yang linear karena pada umumnya besar gaji akan sebanding atau linear dengan waktu kerja seorang karyawan. Pengujian keakurasian terhadap hasil prediksi dilakukan menggunakan MSE (*Mean Square Error*), RMSE (*Root Mean Square Error*), dan MAPE (*Mean Average Percentage Error*). Prediksi gaji karyawan berdasarkan lama waktu kerja menggunakan metode regresi linear ini dapat dikatakan tergolong dalam kategori sangat dikarenakan hasil pengujian menampilkan nilai MSE, RMSE, dan MAPE yang memenuhi standar.

Kata Kunci: Prediksi, Regresi Linear, MSE, RMSE, MAPE.

Employee Salary Prediction System Based on Years Experience Working Using the Linear Regression Method

Abstract

The salary of an employee is often influenced by the length of time worked. The longer an employee works, usually his salary will increase or get higher, and vice versa. The purpose of this study is to predict the salary of employees based on the length of time worked. This research uses linear regression method. Linear Regression is one of the supervised machine learning that is included in the regression class. Linear regression is suitable when there is a linear relationship in the data. Linear regression is a technique to predict a value of variable Y (dependent variable) based on certain variables X (independent variable) if there is a linear relationship between variables X and Y. In this study, the case study used is the prediction of employee salaries based on the length of time the employee worked. . This case study has a linear relationship because in general the salary will be proportional or linear with the working time of an employee. The accuracy of the prediction results was tested using MSE (Mean Square Error), RMSE (Root Mean Square Error), and MAPE (Mean Average Percentage Error). Prediction of employee salaries based on the length of time worked using the linear regression method can be said to be in the very category because the test results show the MSE, RMSE, and MAPE values that meet the standards.

Keywords: Prediction, Linear Regression, MSE, RMSE, MAPE

I. PENDAHULUAN

Sumber daya manusia atau sering disebut dengan tenaga kerja memiliki peranan penting dalam mempertahankan kelangsungan hidup perusahaan. Karena berkembang atau tidaknya sebuah perusahaan sangat tergantung pada kinerja atau produktivitas karyawannya. Hubungan antara karyawan dan perusahaan adalah hubungan yang saling bergantung dan saling menguntungkan kedua belah pihak. Karena perusahaan membutuhkan karyawan sedangkan karyawan membutuhkan perusahaan untuk pemenuhan kebutuhannya (M. A. Gumilar 2018).

Besar pendapatan atau gaji seorang karyawan biasanya berhubungan erat dengan beberapa hal. Diantaranya yaitu, posisi/jabatan, status karyawan, dan periode lama bekerja. Seringkali pihak HR merasa kesulitan saat akan menentukan gaji karyawan. Permasalahan ini dapat diselesaikan dengan sistem yang mampu memprediksi besar gaji berdasarkan beberapa parameter yang diinginkan. Pada penelitian ini akan dikembangkan sebuah sistem untuk memprediksi gaji karyawan berdasarkan lama waktu karyawan tersebut bekerja.

Sistem prediksi ini dibuat dengan metode regresi linear. Regresi Linear adalah salah satu supervised machine learning yang masuk dalam golongan regression. Regresi linear cocok dipakai ketika terdapat hubungan linear pada data. Regresi linear adalah teknik untuk memprediksi sebuah nilai dari variabel Y (variabel dependen) berdasarkan beberapa variabel tertentu X (variabel independen) jika terdapat hubungan linear antara variabel X dan Y.

Dengan dibuatnya sistem ini, penulis berharap agar sistem ini mampu mempermudah tugas pihak HR untuk menentukan besar gaji seorang karyawan.

II. TINJAUAN PUSTAKA

A. Regresi Linear

Pada umunya ada dua macam hubungan antara dua variabel atau lebih, yang biasa disebut bentuk hubungan dan keeratan hubungan. Untuk mengetahui bentuk hubungan maka digunakan analisis regresi. Sedangkan untuk keeratan hubungan dapat diketahui dengan analisis korelasi. Ada pula yang menyebutkan bahwa regresi linear merupakan suatu metode statistika yang digunakan untuk membentuk suatu model hubungan antara variabel terikat (dependen, Y) dengan satu atau lebih variabel bebas (independen, X) (Kurniawan, 2008), dengan tujuan untuk mengestimasi serta memprediksi rata-rata populasi atau nilai rata-rata variabel dependen berdasarkan nilai variabel independen yang diketahui. Hasil analisis regresi berupa koefisien pada masing-masing variabel X (independen). Koefisien tersebut diperoleh dengan cara memprediksi nilai variabel Y (dependen) dengan suatu persamaan. Koefisien regresi dihitung dengan dua tujuan sekaligus, untuk meminimumkan penyimpangan antara nilai aktual dan nilai estimasi variabel Y (dependen) berdasarkan data yang ada (Syahputra, 2008).

Berikut adalah persamaan untuk regresi linear:

$$a = \frac{(\sum y)(\sum x^2)-(\sum x)(\sum xy)}{n(\sum x^2)-(\sum x)^2} \quad (1)$$

$$b = \frac{n(\sum xy)-(\sum x)(\sum y)}{n(\sum x^2)-(\sum x)^2} \quad (2)$$

$$y = a + b.x \quad (3)$$

B. Pengujian Akurasi

Keakurasian suatu prediksi ditentukan oleh seberapa besar penyimpangan atau kesalahan ini, yang terjadi antara data yang diprediksi dengan data yang sebenarnya atau data aktual. Kesalahan dalam perumusan sebuah prediksi tidak hanya disebabkan oleh unsur error tapi juga ketidakmampuan suatu model peramalan mengenali unsur yang lain dalam deret data yang mempengaruhi besarnya penyimpangan dalam prediksi. Besarnya kesalahan atau penyimpangan ini dapat disebabkan oleh besarnya faktor yang tidak diduga (outliers) dimana tidak ada metode prediksi yang mampu menghasilkan prediksi yang akurat atau dapat juga disebabkan metode prediksi yang digunakan tidak dapat memprediksi dengan tepat komponen tren, komponen musiman atau komponen siklus yang mungkin terdapat dalam deret data. Di antara berbagai cara untuk menghitung besarnya kesalahan tersebut beberapa di antaranya adalah mean square error (MSE), root mean square error (RMSE), dan mean absolute percentage error (MAPE). MSE merupakan rataan selisih kuadrat antara nilai yang diprediksikan dengan diamati, RMSE merupakan akar dari MSE, dan MAPE merupakan rataan diferensiasi absolut antara nilai yang diprediksi dan aktual. Hasil prediksi dinyatakan baik jika nilai MAPE kurang dari 10%. Sedangkan untuk MSE dan RMSE yang menggunakan metode berbasis gradien, semakin rendah nilainya maka semakin baik prediksi yang dilakukan.

Rumus dari ketiga pengujian ini sebagaimana ditampilkan pada Persamaan 4, 5, dan 6 di bawah ini.

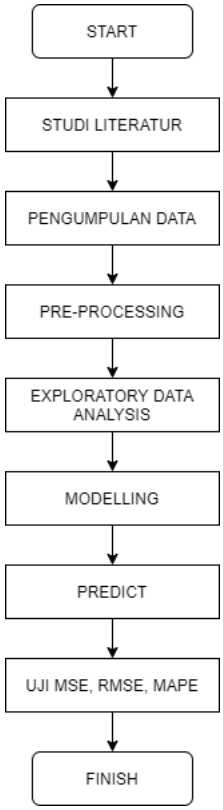
$$\text{MSE} = \frac{1}{n}\sum(Y_t - Y'_t)^2 \quad (4)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum(Y_t - Y'_t)^2} \quad (5)$$

$$\text{MAPE} = \frac{1}{n}\sum_{t=1}^{n}\frac{|Y_t - Y'_t|}{Y_t} \quad (6)$$

III. METODE PENELITIAN

Pada bagian metodologi ini akan dijelaskan beberapa tahapan yang dilakukan selama pelaksanaan penelitian. Tahapan-tahapan tersebut digambarkan dalam bentuk bagan pada Gambar 1.



Gambar 2.1 Tahapan Penelitian

Tahapan pada Gambar 2.1 dijelaskan secara mendetail sebagai berikut:

a. Studi Literatur
Pada tahapan ini, penulis melakukan beberapa riset melalui beberapa literatur seperti paper, artikel, dan buku.

b. Pengumpulan Data
Pada penelitian ini, data yang digunakan adalah berupa dataset yang didapatkan dari *Kaggle* tentang data gaji karyawan dan lama waktu kerja.

c. Pre-Processing
Setelah dilakukan pengumpulan data, tahapan selanjutnya adalah pelaksanaan pre-processing terhadap data yang telah didapatkan. Tahapan ini meliputi pelaksanaan filterisasi data seperti menghilangkan null data pada dataset.

d. Exploratory Data Analysis
Pada tahapan ini, penulis melakukan Exploratory Data Analysis yang bertujuan untuk lebih mengenal data yang akan diolah dengan cara menganalisisnya. Analisis yang dilakukan yaitu mean dan median dari data yang kemudian akan ditampilkan di plotting.

e. Modelling
Setelah melakukan Exploratory Data Analysis, Langkah selanjutnya adalah Modelling atau proses pembuatan model. Model yang digunakan adalah model Linear Regression dengan memanfaatkan library dari scikit-learn.
Pada tahap ini juga dilakukan tahap training data hingga mendapatkan akurasi yang diharapkan.

f. Predict
Modelregresi linear yang telah didapatkan pada tahap sebelumnya kemudian digunakan untuk melakukan prediksi.

g. Uji MSE, RMSE, MAPE
Hasil prediksi yang telah didapatkan kemudian selanjutnya memasuki tahapan pengujian untuk memastikan keakurasiannya. Pengujian keakurasian ini dilakukan menggunakan tiga pengujian yang terdiri dari uji MSE, RMSE, dan MAPE.

IV. HASIL DAN PEMBAHASAN

Pada penelitian ini penulis menggunakan data set berupa 30 data besar gaji beserta lama tahun bekerja. Isi data bervariatif dengan waktu kerja paling singkat 1.1 tahun dan paling lama 10.5 tahun. Sedangkan untuk data gaji paling sedikit berada di kisaran 39343 dan paling besar yaitu 122391.

Setelah mendapatkan data, dilakukan pre-processing agar data dapat diolah lebih maksimal. Proses pre-processing dilakukan pengecekan apakah ada data kosong (null) dengan menggunakan fungsi isnull() seperti pada gambar 4.1

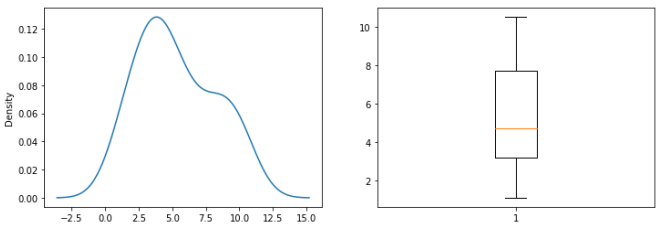
```
#Mencari dan menangani missing values.
df.isnull().sum()

YearsExperience    0
Salary             0
dtype: int64
```

Gambar 4.1

Pada proses diatas diketahui bahwa data tidak ada yang kosong, jadi dapat berlanjut ke proses selanjutnya.

Setelah itu berlanjut dengan Exploratory Data Analysis dengan cara memploting data seperti ditunjukan pada gambar 4.2.



Gambar 4.2 Exploratory Data Analysis

Berdasarkan data plotting didapatkan hasil mean dan median berada di kisaran 3-8. Dengan distribusi data terdistribusi dengan rata. Dibuktikan dengan grafik yang melandai.

Setelah itu penulis memplotting persebaran data dan menghasilkan hasil sebagai berikut:

Gambar 4.3 Scatter Plot Data

Dari data diatas kemudian penulis melihat korelasi antar data dengan menggunakan fungsi corr() yang ditunjukkan pada gambar 4.4

	YearsExperience	Salary
YearsExperience	1.000000	0.978242
Salary	0.978242	1.000000

Gambar 4.4 Tingkat korelasi antar data

Data korelasi tersebut menghasilkan angka yang tinggu yaitu 0.978242. Hal ini berarti hubungan antar data sangat berpengaruh.

Tahap selanjutnya adalah pembuatan model (modelling). Model yang digunakan adalah linear regression. Sebelum pembuatan model, data dibagi lagi menjadi data train dan test dengan ratio 30% untuk data test dan 70% data train. Setelah itu model dibuat dengan fungsi LinearRegression() dari library sklearn. Kemudian setelah model terbentuk, dilakukan data training dengan menggunakan fungsi fit().

Sebelum melakukan prediksi, penulis mencoba menemukan a dan b pada rumus berikut:

$$y = a + b.x$$

Nilai a dan b dapat dicari dengan menggunakan fungsi coef_ dan intercept_ dengan hasil sebagai berikut:

```
print(model.coef_)
print(model.intercept_)
```

```
[[9339.08172382]]
[25918.43833489]
```

Gambar 4.5 Nilai a dan b

Fungsi coef_ mewakili b sedangkan intercept_ untuk a. Jadi pada kasus ini persamaan linear regresinya adalah sebagai berikut:

$$y = 25918 + 9339x$$

x adalah variable test. Jadi jika kita ingin memprediksikan gaji untuk karyawan dengan masa kerja sebesar x, maka akan diketahui estimasi gajinya. Sebagai contoh penulis akan mencoba memprediksi gaji karyawan dengan masa kerja 3 tahun. Jika sesuai rumus maka:

$$y = 25918 + 9339.3$$
$$y = 25918 + 28017$$
$$y = 53935$$

Kemudian dibuktikan dengan fungsi predict() hasilnya sebagai berikut:

```
model.predict([[3]])
```

```
array([[53935.68350634]])
```

Gambar 4.4 Hasil Predict

Hasil predict menggunakan model menunjukkan hasil yang sama dengan hasil perhitungan manual menggunakan rumus. Setelah itu penulis mencoba menampilkan akurasi model dengan menggunakan fungsi score() dan mendapatkan hasil sebagai berikut:

```
model.score(X_test, y_test)
```

```
0.9414466227178214
```

Gambar 4.5 Akurasi Model

Akurasi dari model menunjukkan hasil sebesar 0.94 atau kurang lebih sebesar 94% untuk setiap prediksi yang dilakukan.

Selanjutnya adalah plotting hasil regressi pada persebaran data awal yang ditunjukkan pada gambar 4.6 berikut:

Gambar 4.6 Prediction result plot

Langkah terakhir yang dilakukan penulis adalah melakukan uji akurasi model dengan menggunakan MSE, RMSE, dan MAPE. Metode ini dilakukan dengan menggunakan fungsi dari library sklearn dan menghasilkan hasil sebagai berikut:

```
Mean squared error = 37784662.47
Root meen square error = 6146.92
MAPE = 0.07
```

Gambar 4.7 Hasil Uji Akurasi

V. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, dapat disimpulkan penulis mendapatkan hasil prediksi yang tinggi. Hal ini dibuktikan oleh nilai akurasi model yang didapatkan sebesar 94% serta dibuktikan dengan hasil MSE, RMSE, dan MAPE yang telah memenuhi standar pengujian. Hal ini juga dibuktikan dengan nilai korelasi data yang tinggi.

REFERENSI

[1] Permatasari, A. I., Mahmudy, W. F. (2015). *Pemodelan Regresi Linear dalam Konsumsi Kwh Listrik di Kota Batu Menggunakan Algoritma Genetika*, Malang: Universitas Brawijaya.

[2] Ayuni, G. N., Fitrianah, D. (2019). *Penerapan Metode Regresi Linear Untuk Prediksi Penjualan Properti pada PT XYZ*, Bandung: Institut Teknologi Harapan Bangsa.