

ANALISIS DATA PEGAWAI UNTUK MEMPREDIKSI GAJI BERDASARKAN FAKTOR-FAKTOR SPESIFIK DENGAN PENDEKATAN *MACHINE LEARNING*

PROYEK II

**Bachtiar Ramadhan (1204077)
Nur Tri Ramadhanti Adiningrum (1204061)**



**PROGRAM DIPLOMA IV TEKNIK INFORMATIKA
POLITEKNIK POS INDONESIA
BANDUNG
2021**

PENDAHULUAN

Perkembangan ilmu pengetahuan dan teknologi pada Revolusi Industri 4.0 semakin berkembang pesat. Perubahan karakteristik pekerjaan adalah salah satu dampak tersendiri dari datangnya revolusi industri 4.0. Tentunya perusahaan perlu memiliki keunggulan manajemen yang efektif dalam menghadapi hal tersebut. Dengan demikian salah satu aspek yang berpengaruh besar terhadap kemajuan dan keberhasilan sebuah perusahaan adalah kinerja karyawannya.

Oleh karena itu, penentuan gaji yang tepat oleh perusahaan adalah faktor internal terhadap kemajuan perusahaan. Sangat disayangkan, perkembangan perusahaan saat ini belum memiliki suatu media keputusan untuk melakukan prediksi gaji karyawan berdasarkan kualitas data.

Penelitian ini bertujuan untuk mengetahui prediksi gaji karyawan. Karakteristik dataset yang digunakan untuk memprediksi gaji karyawan terdiri dari parameter-parameter berdasarkan faktor-faktor spesifik. Selanjutnya faktor-faktor tersebut akan diuji validitas dan korelasinya menggunakan pendekatan *machine learning* dengan metode *regression*.

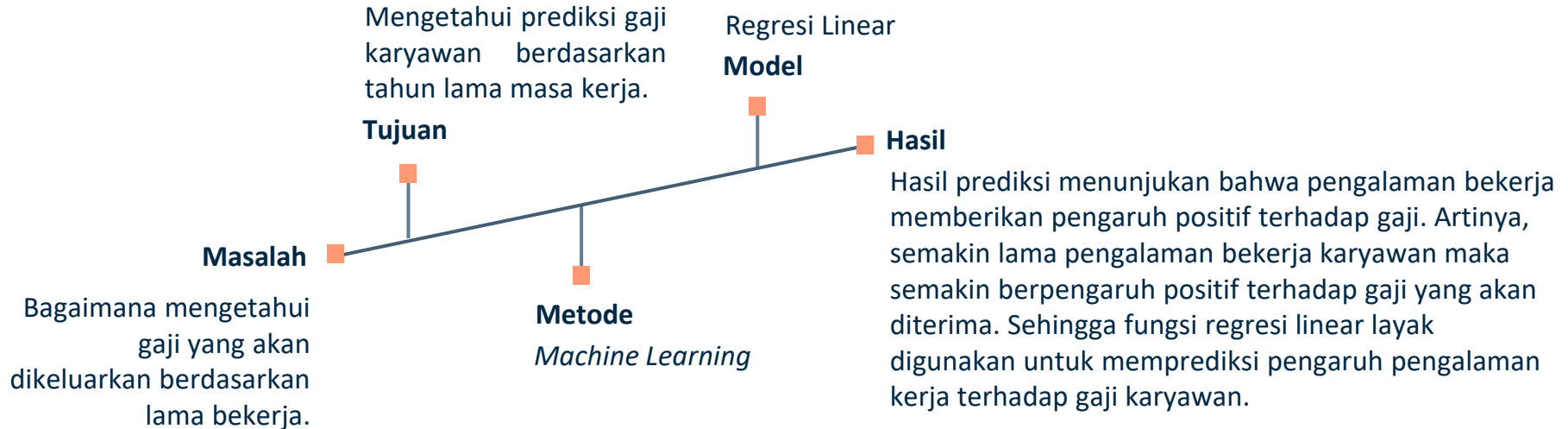
Tentunya hasil prediksi gaji karyawan perlu divisualisasikan secara realtime untuk dapat digunakan oleh perusahaan dalam menentukan keputusan dengan cepat. Visualisasi hasil prediksi tersebut akan ditampilkan berbasis web base dengan framework Django.



IMPLEMENTASI MODEL REGRESI LINEAR SEDERHANA UNTUK PREDIKSI GAJI BERDASARKAN PENGALAMAN LAMA BEKERJA

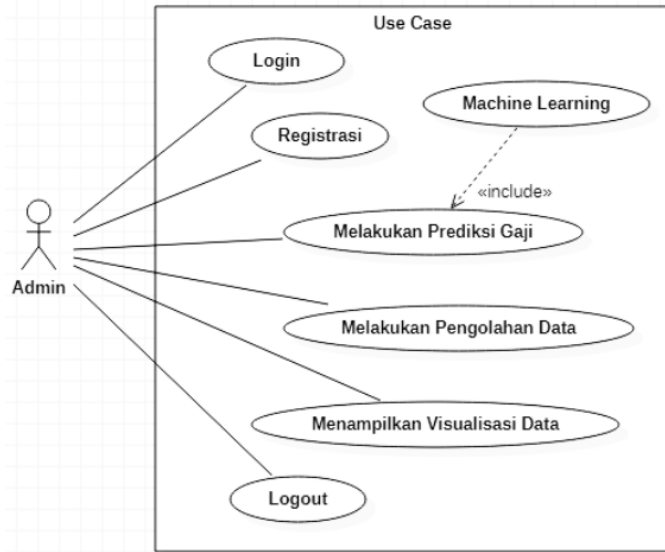
Yayan Adrianova Eka Tuah , Anyan

Program Studi Pendidikan Komputer, STKIP Persada Khatulistiwa Sintang

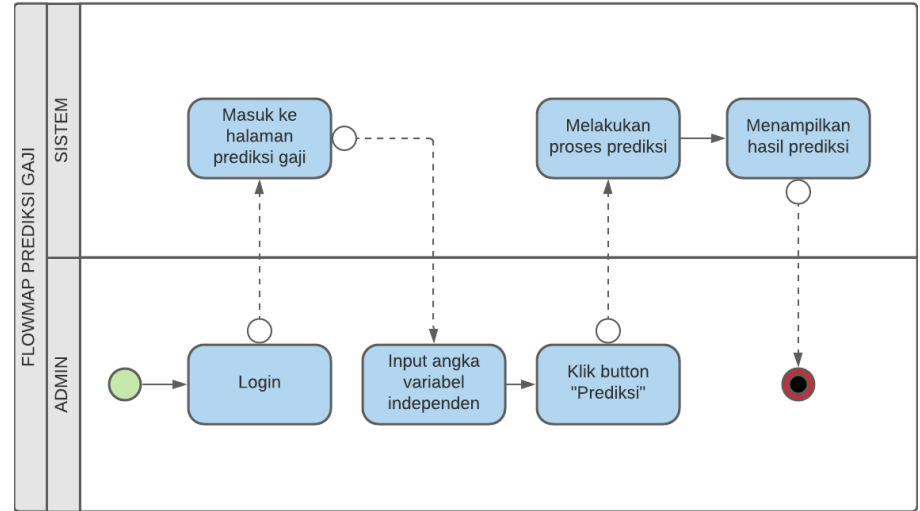


PERANCANGAN APLIKASI

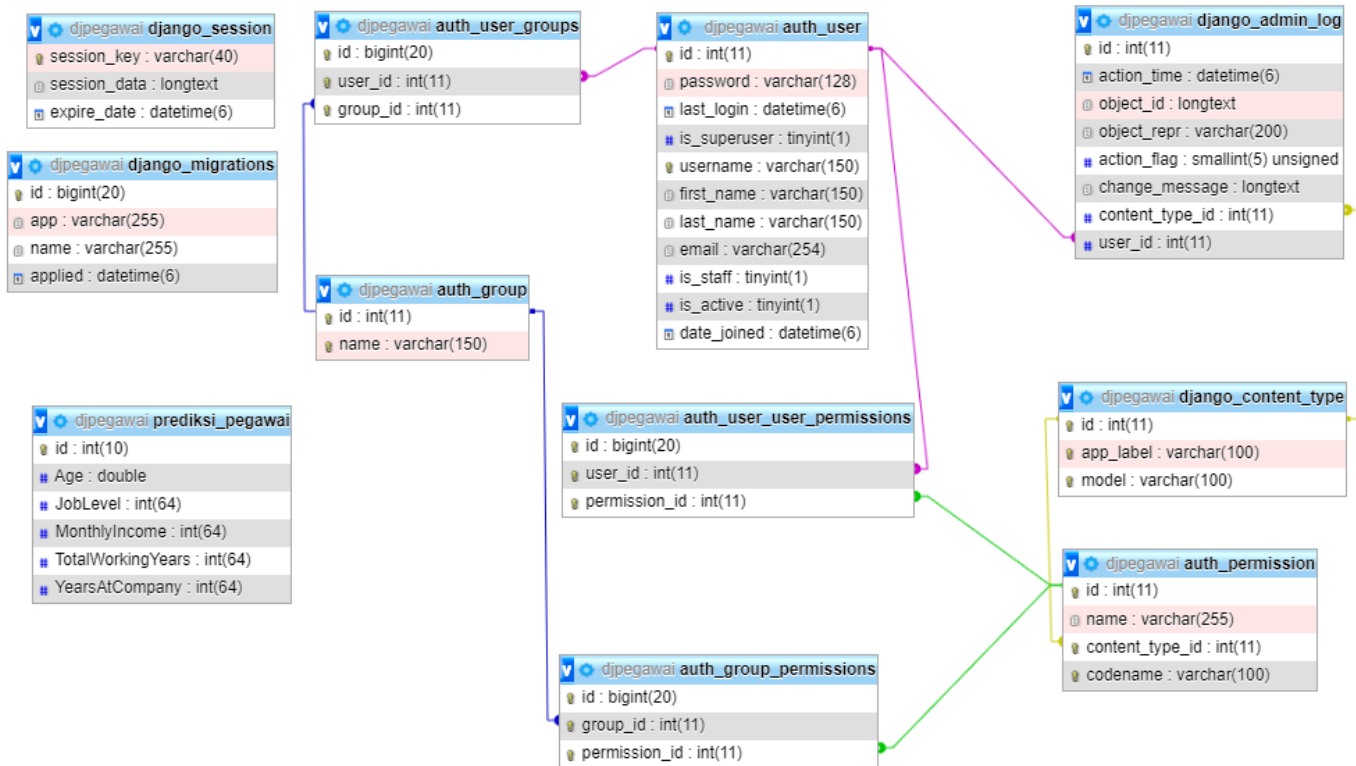
USE CASE DIAGRAM



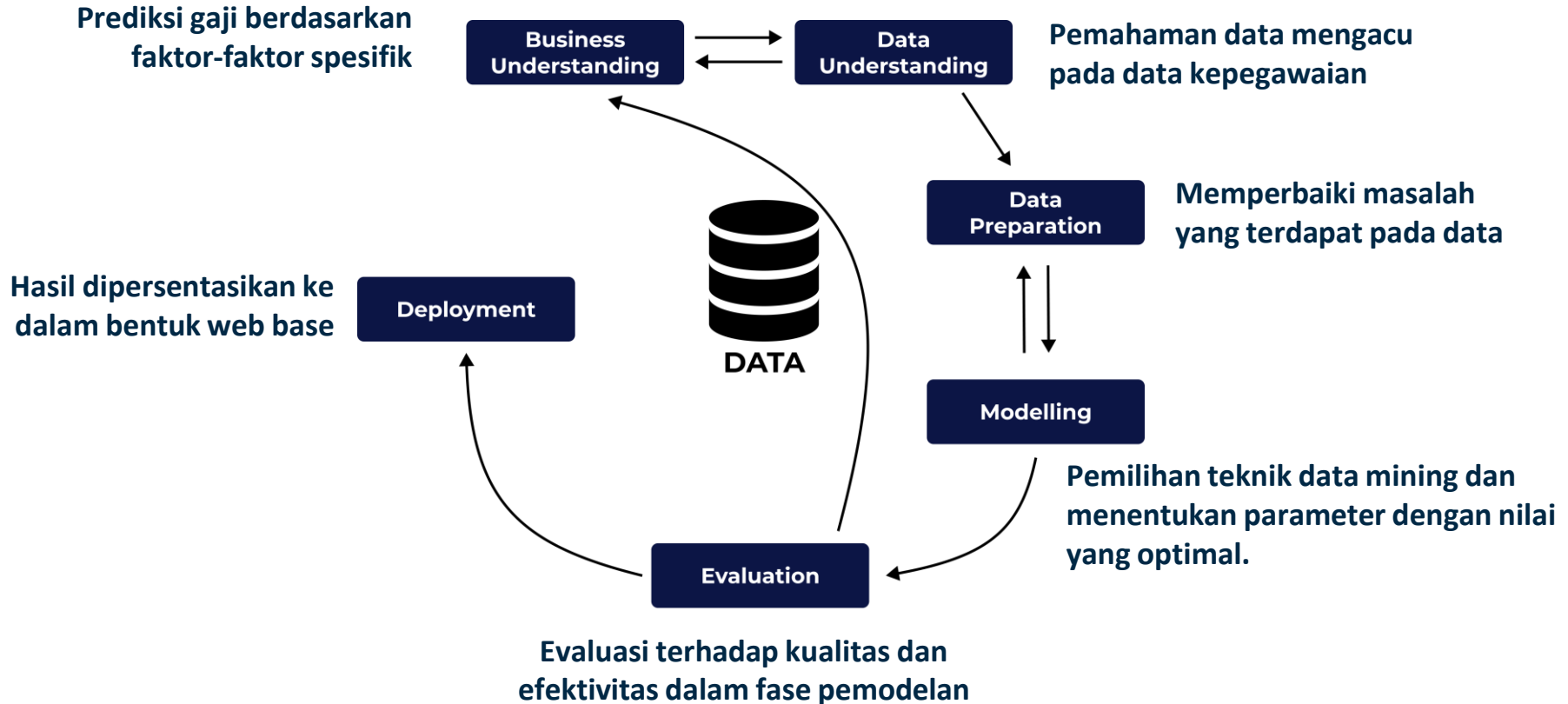
FLOWMAP PREDIKSI GAJI



ENTITY RELATIONSHIP DIAGRAM



METODE CRISP-DM



PROSES HIMPUNAN DATA

Pada tahap ini, hal yang dilakukan adalah memahami dan mempersiapkan data yang dikenal dengan istilah **Data Preprocessing**. Metode yang digunakan dalam **Data Preprocessing** pada model ini adalah **Data Cleaning**. Berikut tahapan himpunan data:

1. Import Library
2. Import Data
3. Encoder Data
4. Replace Missing Value
5. Drop Data

Dari Proses Himpunan Data, menyisakan faktor-faktor pengaruh terhadap faktor dependen (MonthlyIncome).

	Age	BusinessTravel	DailyRate	Department	DistanceFromHome	Education	EducationField	EmployeeCount	EmployeeNumber	EnvironmentSatisfac
0	50.000000	2	1126.0	1	1.000000	2	3	1	997	
1	36.000000	2	216.0	1	6.000000	2	3	1	178	
2	21.000000	2	337.0	2	7.000000	1	2	1	1780	
3	50.000000	1	1246.0	0	9.930407	3	3	1	644	
4	52.000000	2	994.0	1	7.000000	4	1	1	1118	
...
1024	37.930571	2	750.0	1	28.000000	3	1	1	1596	
1025	41.000000	2	447.0	1	9.930407	3	1	1	1814	
1026	22.000000	1	1256.0	1	9.930407	4	1	1	1203	
1027	29.000000	2	1378.0	1	13.000000	2	4	1	2053	
1028	50.000000	2	264.0	2	9.000000	3	2	1	1591	

Tabel data setelah proses Encoder dan Replace Missing Value

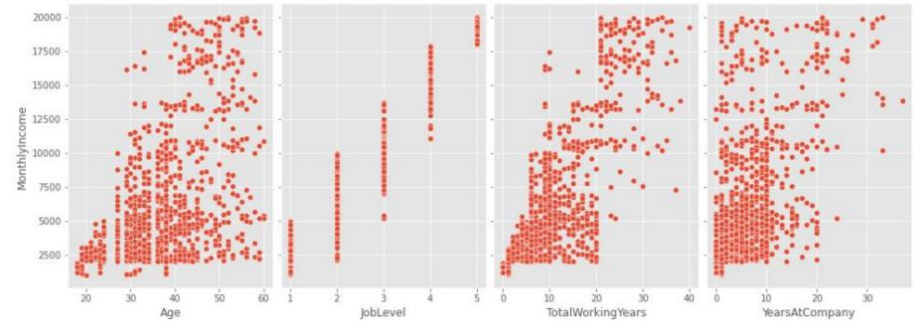
	Age	JobLevel	MonthlyIncome	TotalWorkingYears	YearsAtCompany
Age	1.000000	0.449794	0.439321	0.623246	0.294041
JobLevel	0.449794	1.000000	0.944295	0.772222	0.511125
MonthlyIncome	0.439321	0.944295	1.000000	0.771358	0.488875
TotalWorkingYears	0.623246	0.772222	0.771358	1.000000	0.638163
YearsAtCompany	0.294041	0.511125	0.488875	0.638163	1.000000

Data Akhir

PROSES DATA MINING & PENGETAHUAN

Pada tahapan Proses Data Mining hal yang dilakukan adalah memilih metode yang sesuai dengan karakter data yang dikenal dengan istilah **Modelling**. Pada model ini digunakan Proses Data Mining Prediction. Pada tahapan Pengetahuan hal yang dilakukan adalah memahami model dan pengetahuan yang sesuai sehingga dapat memilih model. Model yang digunakan adalah Linear Regression menggunakan **Scikit Learn**. Berikut yang tercantum dalam proses data mining dan pengetahuan :

1. **Data Preprocessing Testing**
2. **Visualisasi Data**
3. **Modelling menggunakan Linear Regression**



Berdasarkan nilai koefisien variabel independen dan Intersept didapat, maka persamaan regresi linear multivariabel sebagai berikut :

$$Y = -1728 - 5,054X_1 + 3871,7530X_2 + 46,9405X_3 - 9,8560X_4$$

Atau

$$\begin{aligned} \text{MonthlyIncome} \\ = & -1728 - 5,504(\text{Age}) + 3871,7530(\text{JobLevel}) \\ & + 46,9405(\text{TotalWorkingYears}) - 9,8460(\text{YearsAtCompany}) \end{aligned}$$

PROSES EVALUASI DATA

UJI F (ANOVA)

Berdasarkan Uji F - Statistic ANOVA menggunakan model OLS, didapat :

F - Statistic = 2750, 622

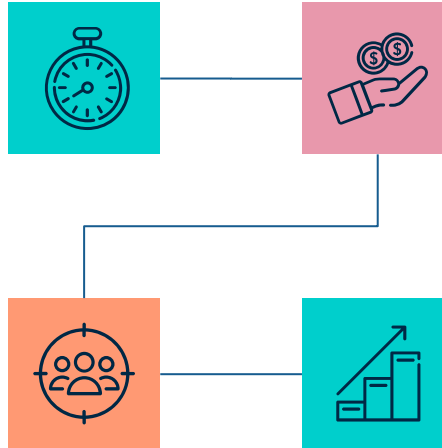
P - Value = 0,00

Keputusan : Tolak H0, Terima H1.

UJI T

Berdasarkan **Uji-t** menggunakan model OLS, didapat :

const	1.419659e-13
Age	4.643777e-01
JobLevel	0.000000e+00
TotalWorkingYears	6.771641e-05
YearsAtCompany	3.136608e-01
dtype:	float64



R-SQUARE

Berdasarkan perhitungan **R-Square**, didapat akurasi sebesar **0,909** atau **90,9%**. Sehingga dapat disimpulkan bahwa model performa dengan baik.

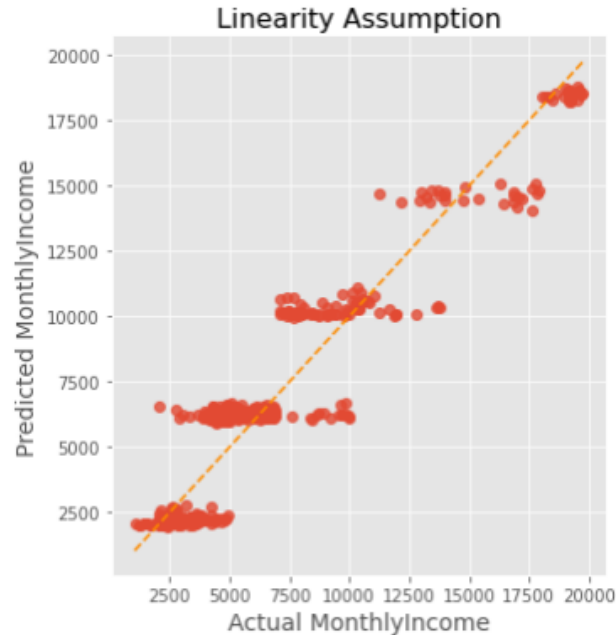
UJI ASUMSI

Pengujian asumsi terdiri dari beberapa proses yaitu :

- **Linearitas**
- **Normalitas**
- **Multikolinearitas**
- **Autokorelasi**
- **Homoskedastisitas**

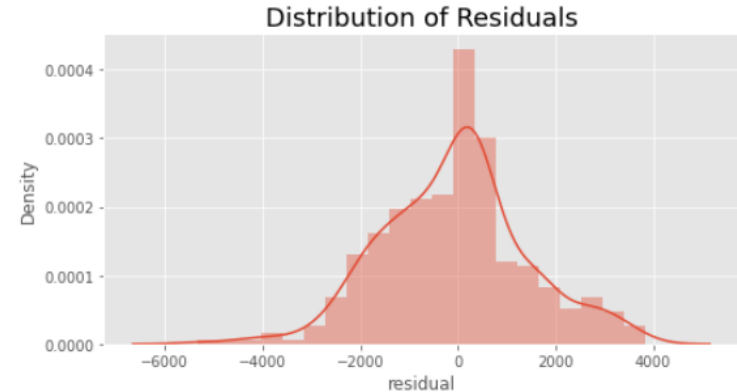
PROSES EVALUASI DATA

UJI LINEARITAS



Plot sebar menunjukkan sisa yang tersebar merata di sekitar garis diagonal, sehingga dapat diasumsikan bahwa ada **hubungan linier** antara **variable independent** dan **dependen**.

UJI NORMALITAS



Dapat diketahui hipotesa sebagai berikut :

H_0 = Residual terdistribusi normal.

H_1 = Residual terdistribusi secara tidak normal

Dari hasil perhitungan diatas, dapat diketahui bahwa nilai **p-value** yang dihitung menggunakan metode **Anderson-Darling** adalah **0,00032261**. Angka tersebut berada di bawah nilai threshold yang ditentukan yaitu **0,05**, yang berarti tolak **H_0** terima **H_1** atau dapat dikatakan residual terdistribusi secara tidak normal. Sehingga disimpulkan asumsi normalitas terpenuhi.

PROSES EVALUASI DATA

UJI MULTIKOLINEARITAS

	VIF	variable
0	28.655370	Intercept
1	1.690786	Age
2	2.489052	JobLevel
3	4.140803	TotalWorkingYears
4	1.739893	YearsAtCompany

Berdasarkan gambar di atas dapat dilihat nilai variabel **Age**, **JobLevel**, **TotalWorkingYears**, **YearsAtCompany** memiliki nilai kurang dari 10 sehingga dengan menggunakan **tingkat signifikansi sebesar 0,05** dapat disimpulkan bahwa pada data tersebut tidak terdapat multikolinearitas pada variabel-variabel prediktor.

UJI AUTOKORELASI

Pada langkah ini akan dilakukan perhitungan skor **Durbin-Watson** menggunakan `durbin_watson()` fungsi dari `statsmodel` yang dibuat, kemudian menilainya dengan kondisi sebagai berikut :

1. Jika skor Durbin-Watson **kurang dari 1,5** maka terdapat **autokorelasi positif** dan **asumsi tidak terpenuhi**.
2. Jika skor Durbin-Watson **antara 1,5 – 2,5** maka **tidak ada autokorelasi** dan **asumsi puas**.
3. Jika skor Durbin-Watson **lebih dari 2,5** maka terdapat **autokorelasi negative** dan **asumsi tidak puas**.

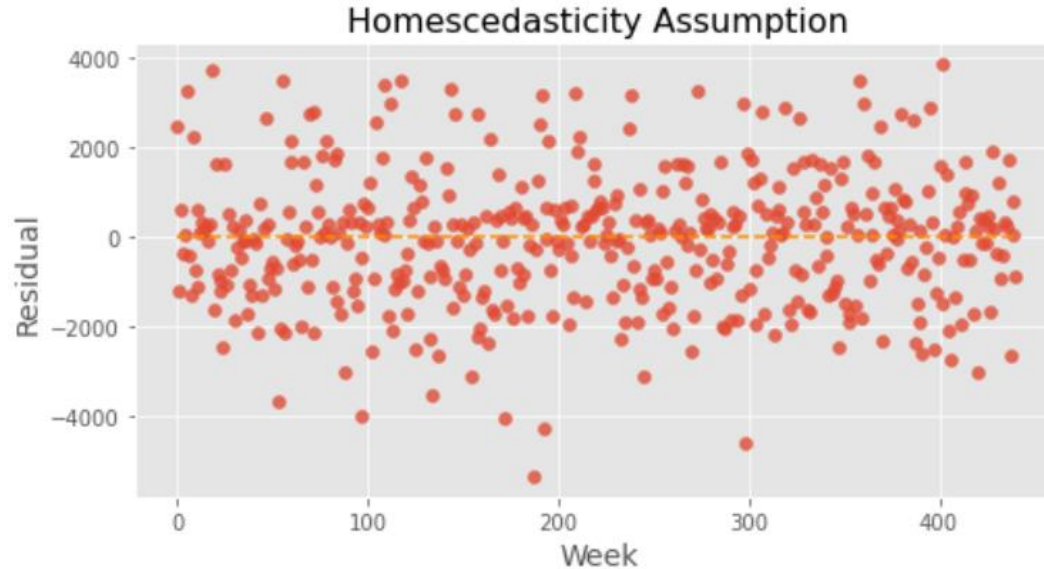
```
Durbin-Watson: 2.160636228778726  
Little to no autocorrelation
```

```
Assumption satisfied
```

Didapat hasil perhitungannya adalah **2,160636228**. Dapat diasumsikan bahwa terdapat sedikit atau tidak ada autokorelasi, yang berarti asumsi puas.

PROSES EVALUASI DATA

UJI HOMOSKEDASTISITAS



Dari grafik scatterplot di atas, terlihat titik-titik menyebar secara acak, serta tersebar baik di atas maupun di bawah angka 0 (nol) pada sumbu Y. Maka dapat diambil kesimpulan bahwa tidak terdapat gejala heteroskedastisitas pada model regresi yang digunakan.

KESIMPULAN & SARAN

KESIMPULAN

1. Hasil dari analisis karakteristik data gaji karyawan, didapatkan parameter yang berkaitan dengan gaji karyawan yaitu Age, JobLevel, TotalWorkingYears, dan YearsAtCompany.
2. Model yang dibuat berhasil melewati semua pengujian dalam langkah validasi model.
3. Berdasarkan model yang diambil dari model OLS kami mendapatkan nilai akurasi sebesar 0,909. Akurasi tersebut merupakan nilai akurasi yang baik, sehingga dapat dikatakan model machine learning dapat berperforma baik untuk memprediksi gaji.
4. Berdasarkan uji validitas, nilai akurasi 0,909 menunjukkan bahwa MonthlyIncome dipengaruhi oleh faktor independen (Age, YearsAtCompany) sebesar 0,909 atau 90,9%. Nilai sisa dari akurasi tersebut adalah 0,091 atau 9,1% yang artinya MonthlyIncome dipengaruhi oleh faktor lain yang tidak diketahui sebesar 9,1%.
5. Visualisasi data dari hasil model prediksi gaji karyawan dapat digunakan menjadi bentuk aplikasi berbasis web base dengan menggunakan *framework* Django.

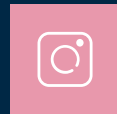
SARAN

1. Pembuatan model prediksi yang digunakan dapat lebih beragam untuk membandingkan performa antara model satu dengan model yang lainnya.
2. Sumber data yang digunakan kurang maksimal. Pada penelitian ini, hanya didapatkan real yang berasal dari Kaggle. Diharapkan kedepannya dapat menggunakan data real langsung dari perusahaan.

Do you have any questions?

D IV TEKNIK INFORMATIKA
POLITEKNIK POS INDONESIA
BANDUNG
2021

TERIMA KASIH



CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik