Analisis Data Pegawai untuk Memprediksi Gaji Berdasarkan Faktor-Faktor Spesifik dengan Pendekatan *Machine Learning*

(*Employee Data Analysis to Predict Salary Based on Special Factors with Machine Learning Approach*)

Syafrial Fachri Pane[1], Bachtiar Ramadhan[2], Nur Tri Ramadhanti Adiningrum[3]
[1] Informatics Engineering, Indonesian Postal Polytechnic
Jl. Sariasih Nomor 54, Bandung 40151, Jawa Barat, Indonesia
[2] Informatic Engineering, Indonesian Postal Polytechnic
Jl. Sariasih Nomor 54, Bandung 40151, Jawa Barat, Indonesia
[3] Informatic Engineering, Indonesian Postal Polytechnic
Jl. Sariasih Nomor 54, Bandung 40151, Jawa Barat, Indonesia
Email: syafrial.fachri@poltekpos.ac.id, bachtiarramadhan26@gmail.com, nurtrira06@gmail.com

Penulis korespondensi

Abstract ***The company cannot be separated from the workforce. One aspect that affects the progress of a company is the performance of its employees. Providing an appropriate salary is one of the important factors to boost the performance of the workforce. Unfortunately, the current development of the company does not have a decision media to predict employee salaries based on data quality. This study aims to determine the prediction of employee salaries based on specific faktors. In this study, the faktors that were tested included independent variables in the form of Age, JobLevel, TotalWorkingYears, and YearsAtCompany. Then the dependent variable is MonthlyIncome. The data analysis technique used multivariate linear regression analysis which was used to predict employee salaries. The results of employee salary predictions will be displayed on a web-base. The model created successfully passed all the tests in the model validation step, so it can be concluded that the model created can perform well for predicting employee salaries. The results of employee salary predictions can be used as a form of web-based application using the Django framework. With this application, admins can predict employee salaries easily and quickly.***

Key words**: Salary Prediction, Multivariate Linear Regressio, Specific Factors, Web Base**

I. PENDAHULUAN

Perkembangan ilmu pengetahuan dan teknologi pada Revolusi Industri 4.0 semakin berkembang pesat. Revolusi Industri 4.0 sendiri mulai terjadi melalui rekayasa intelegensia dan *internet of thing* sebagai tulang punggung pergerakan dan konektivitas antara manusia dengan mesin[1]. Sehingga, terdapat penggabungan teknologi digital dan internet dengan industri konvensional, yang bertujuan untuk meningkatkan produktivitas, efisiensi dan layanan konsumen secara signifikan[2]. Era revolusi ini akan mendisrupsi berbagai kegiatan diberbagai bidang seperti pada bidang teknologi, ekonomi, sosial, dan politik[1]. Saat ini, kehidupan berada diawal revolusi yang secara mendasar mengubah cara hidup, bekerja, dan berhubungan satu sama lain [3].

Perubahan karakteristik pekerjaan adalah salah satu dampak tersendiri dari datangnya revolusi industri 4.0[4]. Karakteristik pekerjaan yang berubah akan mendisrupsi pekerjaan yang telah ada dan menggantikanya dengan pekerjaan dengan karakteristik baru [5]. Karakteristik baru pada pekerjaan juga membutuhkan kompetensi baru kepada para pekerja[6]. Tentunya perusahaan harus siap untuk saling bersaing dengan perusahaan yang lain[7]. Selanjutnya, perusahaan perlu memiliki keunggulan dan manajemen yang efektif untuk menghadapi persaingan tersebut[7]. Dengan demikian salah astu aspek yang berpengaruh besar terhadap kemajuan dan keberhasilan sebuah perusahaan adalah kinerja karyawannya[7]. Walaupun perusahaan tersebut memiliki teknologi yang canggih, namun tidak terdapat tenaga kerja didalamnya, perusahaan tidak akan dapat mencapai tujuannya[7].

Oleh karena itu, penentuan gaji yang tepat oleh perusahaan kepada karyawan adalah salah satu faktor yang berpengaruh secara internal terhadap kemajuan perusahaan. Selain itu, perusahaan juga harus bersedia mengeluarkan gaji bonus bagi karyawannya yang telah bekerja dengan maksimal dan sesuai dengan apa yang dibutuhkan oleh sebuah perusahaan. Sangat disayangkan, perkembangan perusahaan saat ini belum memliki suatu media keputusan untuk melakukan prediksi gaji karyawan berdasarkan kualitas data.

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*. Faktor-faktor

tersebut akan diambil berdasarkan pedoman interpretasi koefisien korelasi [8]. Untuk menentukan faktor yang dominan terhadap prediksi gaji, maka koefisien korelasi yang akan digunakan adalah tingkat hubungan sedang, kuat, dan sangat kuat. Metode yang digunakan pada *machine learning* yaitu *regression*. *Regression* digunakan untuk melakukan prediksi gaji karyawan. 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.

II. TINJAUAN PUSTAKA

A. Machine Learning

Machine learning dapat diartikan sebagai aplikasi komputer dan algoritma matematika yang diadopsi dengan cara pembelajaran yang berasal dari data dan dapat menghasilkan suatu prediksi di masa yang akan datang. [24]

B. Regresi Linier Berganda

Regresi berganda adalah perpanjangan dari regresi linier sederhana. [25] Analisis regresi linier berganda dapat digunakan untuk memprediksi hubungan antara satu variabel independen berdasarkan nilai dari dua atau lebih variabel dependen. [25] Analisis regresi linier berganda juga menghasilkan persamaan matematis. [25] Jika ada lebih dari dua variabel maka hubungan linier dapat dinyatakan dalam persamaan regresi linier berganda yang dikutip pada persamaan 1 dan persamaan 2 sebagai berikut :

$$Y' = b_0 + b_1X_1 + b_2X_2 + \cdots + b_nX_n \qquad (1)$$

$$Y'i = b_0 + b_1X_{1i} + b_2X_{2i} + \cdots + b_nX_{ni} \qquad (2)$$

Keterangan :
Y = nilai-nilai hasil pengamatan
Y' = nilai regresi
i = 1,2,3, k

Pada persamaan di atas ada satu variabel dependen, yaitu Y' dan ada n variabel independen X1, X2, …Xn. [25]

C. Scikit-Learn

Scikit-learn adalah modul pada Bahasa pemrograman Python yang menyediakan berbagai jenis algoritma machine learning. Bentuk yang terdapat pada *library python* merupakan bentuk yang tersedia dalam *scikit-learn*. *Scikit-learn* memanfaatkan *task-oriented interface* yang konsisten sehingga memudahkan dalam membandingkan antarmetode. *Scikit-learn* mengintegrasikan berbagai algoritma machine learning untuk *supervised learning* dan *unsupervised learning*. [26]

D. Framework Django

Django ialah sebuah web framework berbasis bahasa pemrograman Python yang didesain untuk membuat suatu aplikasi web yang dinamis, kaya fitur dan aman. [27] Django yang dikembangkan oleh Django Software Foundation terus mendapatkan perbaikan sehingga membuat web framework yang satu ini menjadi pilihan utama bagi banyak pengembang aplikasi web. [27]

E. Penelitian Sebelumnya

Pada penelitian sebelumnya terdapat beberapa jurnal yang telah dirangkum seperti pada TABEL 1.

TABEL 1. Penelitian Sebelumnya

Area Penelitian	Dataset	Metode
Prediksi gaji [7]	Data gaji pegawai	Analisis regresi linear
Prediksi saham [9]	Data historis harga saham	CRISP-DM
Prediksi sembako [10]	Data sembako	Regresi linier berganda
Prediksi harga rumah [11]	*Web scrapping* 2 website	Regresi linear
Prediksi inventaris barang [12]	Data Inventarsi Barang	Regresi linier
Prediksi kasus Covid-19 [13]	Databooks	*Backpropagation* dan regresi linear
Prediksi harga emas [14]	Data harga emas	Regresi linear, *backpropagation*, fuzzy mamdani
Prediksi pendapatan daerah [15]	Data besaran pendapatan	Regresi linear sederhana
Prediksi tingkat produksi kopi.[16]	Data produksi kopi	Regresi linear sederhana
Peramalan penerimaan mahasiswa baru [17]	Data mahasiswa	Regresi linear sederhana
Prediksi gaji [18]	Data gaji	Regresi linear dan regresi polinomial
Analisis empiris [19]	Data prediksi dan harga rumah	*Simple Linear Regression*, *Multiple Linear Regression*
Korelaso gaji dan efisiensi inovasi enterprise [20]	Data perusahaan manufaktur	*Multiple regression*
Aplikasi data cuaca [21]	Dataset BMKG	Metode regresi linear bivariat simple
Estimasi produktivitas tanaman padi [22]	Data Dinas Pertanian Kehutanan Perkebunan dan Peternakan	Regresi liner berganda
Analisis kesejahteraan pedagang kaki lima [23]	Data pedagang kaki lima	Regresi linier dan *logistic ordinal*

III. METODE PENELITIAN

Di dalam penelitian ini, digunakan metode dengan jenis penelitian deskriptif kualitatif yang menggunakan regresi linier multivariat dengan menggunakan bahasa pemrograman Python. Penelitian deskriptif adalah penelitian yang digunakan untuk menemukan pengetahuan yang seluas-luasnya terhadap objek penelitian pada suatu masa tertentu. Penelitian dekriptif ini menyajikan satu

gambar yang terperinci mengenai satu situasi khusus. Penelitian deskriptif bertujuan untuk menjelaskan atau mendeskprisikan suatu keadaan apa adanya dan menginterpretasi objek sesuai dengan apa adanya peristiwa, ataupun segala sesuatu yang terkait dengan variabel-variabel yang bisa dijelaskan baik dengan angka-angka maupun kata-kata.

Dalam penelitian ini, data yang digunakan dalam proses regresi linier multivariat diambil dari Kaggle. Data yang digunakan adalah *data training* sebanyak 1029 baris dan 35 kolom, dan *data test* sebanyak 441 baris dan 34 kolom.

IV. IMPLEMENTASI

Kebutuhan untuk pembuatan model *machine learning* dan aplikasi prediksi gaji pegawai adalah sebagai berikut :

- Aplikasi *software* : XAMPP 3.2.4, Lucidchart, Visual Studio Code, Jupyter Notebook.
- *Hardware* : laptop merk LENOVO dengan kriteria : Prosesor IntelTM Core i3-4030U, RAM 4 GB, 64-bit OS. Software : Ms. Office, Windows 10 Pro.

A. Implementasi Model Machine learning

A.1. 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:

```
# Basic Library
import pandas as pd
import numpy as np

# Data Visualization
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
import seaborn as sns
from scipy.stats import skew

# Model Building
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
```

Gambar 1. Impor Library

```
df_train = pd.read_csv('E:\employee_attrition_train.csv')
df_train
```

Gambar 2. Impor Dataset

Karena *machine learning* tidak bisa membaca tipe data *object*, maka perlu adanya perubahan tipe data tersebut dengan integer. Pada tahapan ini, digunakan proses *encoder* kategori.

```
# Encoder BusinessTravel Variable
# converting type of columns to 'category'
df_train['BusinessTravel'] = df_train['BusinessTravel'].astype('category')
# Assigning numerical values and storing in another column
df_train['BusinessTravel'] = df_train['BusinessTravel'].cat.codes

# Encoder Department Variable
df_train['Department'] = df_train['Department'].astype('category')
# Assigning numerical values and storing in another column
df_train['Department'] = df_train['Department'].cat.codes
```

Gambar 3. Skrip Encoder Variabel

Pada dataset yang digunakan, terdapat nilai NaN/Null. Oleh karena itu, dilakukan proses pengisian nilai tersebut dengan nilai rata-rata (*mean*) variabelnya.

```
Age = df_train['Age']
df_train.Age = df_train.Age.fillna(value=df_train.Age.mean())
```

Gambar 4. Skrip Proses Pergantian Nilai Null menjadi Nilai Mean

Setelah semua data berbentuk integer, lakukan cek korelasi antar atribut untuk memilih atribut yang berkorelasi sedang-kuat terhadap atribut gaji (MonthlyIncome).

```
df_train_clean.corr().abs()
```

Gambar 5. Skrip Cek Tabel Korelasi

Kemudian, langkah selanjutnya adalah drop atribut yang memiliki nilai korelasi dibawah kriteria sedang-kuat dan hanya menyisakan atribut Age, JobLevel, TotalWorkingYears, YearsAtCompany sebagai variabel independen dan MonthlyIncome sebagai variabel dependen.

```
df_train_clean = df_train_clean.drop(['Attrition', 'BusinessTravel', 'DailyRate', 'Department',
       'DistanceFromHome', 'Education', 'EducationField', 'EmployeeCount',
       'EmployeeNumber', 'EnvironmentSatisfaction', 'Gender', 'HourlyRate',
       'JobInvolvement', 'JobRole', 'JobSatisfaction',
       'MaritalStatus','MonthlyRate', 'NumCompaniesWorked',
       'Over18', 'OverTime', 'PercentSalaryHike', 'PerformanceRating',
       'RelationshipSatisfaction', 'StandardHours', 'StockOptionLevel',
       'TrainingTimesLastYear', 'WorkLifeBalance',
       'YearsInCurrentRole', 'YearsSinceLastPromotion',
       'YearsWithCurrManager'], axis=1)
```

Gambar 6. Skrip Drop Atribut

A.2. Proses Data Mining & Pengetahuan

Pada tahap ini, hal yang dilakukan adalah melakukan pemilihan metode yang sesuai dengan karakter data yang dikenal dengan istilah *Modelling*. Pada model ini digunakan proses *Data Mining Prediction*. Proses Data Mining & Pengetahun yang dilakukan ini dengan melakukan perbandingan pengaruh variabel dependen dengan mengacu korelasi antara tiap-tiap variabel dependen dan variabel independen. Penerapan model yang digunakan adalah *Linear Regression Multivariate* meggunakan *Scikit Learn*.

Untuk membuat model *machine learning*, ditentukan terlebih dahulu variabel dependen dan independennya. Age, JobLevel, TotalWorkingYears, YearsAtCompany sebagai variabel independen dan MonthlyIncome sebagai variabel dependen.

```
# Menentukan variabel X dan variabel Y
x_train = df_train_clean[['Age', 'JobLevel', 'TotalWorkingYears', 'YearsAtCompany']]
y_train = df_train_clean[['MonthlyIncome']]
```

Gambar 7. Skrip Penentuan Variabel Dependen dan Independen

```
regressor = LinearRegression()
persamaan = regressor.fit(x_train, y_train)
print(regressor.coef_)
print(regressor.intercept_)
```

Gambar 8. Skrip Penentuan Variabel Dependen dan Independen

Dari model linier regresi tersebut didapat koefisien independen yaitu -5,054 untuk Age, 3871,7530 untuk JobLevel, 46,9405 untuk TotalWorkingYears, -9,8460 untuk YearsAtCompany dan variabel dependen -1728

untuk MonthlyIncome. Berikut persamaan linear yang dikutip pada persamaan 3 :

$$Y = -1728 - 5{,}054X_1 + 3871{,}7530X_2, 46{,}9405X_3 - 9{,}8460X_4 \quad (3)$$

Y = MonthlyIncome (Variabel Dependen)
X_1 = Age (Variabel Independen-1)
X_2 = JobLevel (Variabel Independen-2)
X_3 = TotalWorkingYears (Variabel Independen-3)
X_4 = YearsAtCompany (Variabel Independen-4)

Maka dapat disimpulkan persamaan regresi linier multivariabel yang dikutip pada persamaan 4 sebagai berikut :

$$MonthlyIncome = -1728 - 5{,}054(Age) + 3871{,}7530(JobLevel) + 46{,}9405(TotalWorkingYears - 9{,}8460(YearsAtCompany) \quad (4)$$

A.3. Evaluasi Data

1) Validasi Model

Validasi model *machine learning* menggunakan model OLS dan *dmatrices*.

```
X = df_train_clean[['Age', 'JobLevel', 'TotalWorkingYears', 'YearsAtCompany']]
X = sm.add_constant(X) # adding a constant

olsmod = sm.OLS(df_train['MonthlyIncome'], X).fit()
print(olsmod.summary())
```

Gambar 9. Skrip Validasi OLS

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          MonthlyIncome   R-squared:                       0.909
Model:                            OLS   Adj. R-squared:                  0.909
Method:                 Least Squares   F-statistic:                     2571.
Date:                Mon, 10 Jan 2022   Prob (F-statistic):               0.00
Time:                        10:58:11   Log-Likelihood:                -8944.9
No. Observations:                1029   AIC:                         1.790e+04
Df Residuals:                    1024   BIC:                         1.792e+04
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const      -1728.5202    230.587     -7.496      0.000   -2180.998   -1276.043
Age           -5.0543      6.905     -0.732      0.464     -18.605       8.496
JobLevel    3871.7530     65.635     58.989      0.000    3742.958    4000.548
TotalWorkingYears    46.9406     11.733      4.001      0.000      23.917      69.965
YearsAtCompany       -9.8460      9.767     -1.008      0.314     -29.012       9.320
==============================================================================
Omnibus:                       12.798   Durbin-Watson:                   2.069
Prob(Omnibus):                  0.002   Jarque-Bera (JB):               15.262
Skew:                          -0.182   Prob(JB):                     0.000485
Kurtosis:                       3.472   Cond. No.                         213.
==============================================================================
```

Gambar 10. Hasil OLS Regresi

2) Uji F (ANOVA)

Uji kelayakan model (koefisien regresi) atau disebut dengan uji F, yaitu untuk mengetahui apakah variabel independent yang terdapat dalam persamaan tersebut di atas secara bersama-sama berpengaruh signifikan pada nilai variabel dependen. [28]

F-test atau ANOVA (*Analysis of Variance*) dalam regresi multi-linier dapat digunakan untuk menentukan apakah model yang dibuat berkinerja lebih baik daripada model yang lebih sederhana.

```
print('F-statistic:', olsmod.fvalue)
print('Probability of observing value at least as high as F-statistic:', olsmod.f_pvalue)
```

Gambar 11. Skrip Uji F (ANOVA)

```
F-statistic: 2570.622889791836
Probability of observing value at least as high as F-statistic: 0.0
```

Gambar 12. Output Uji F (ANOVA)

Hipotesa yang didapat dari tabel uji F adalah:

H_0 = Variabel independen secara simultan bukan penjelas yang signifikan terhadap variabel dependen (model tidak cocok).
H_1 = Variabel independen secara simultan merupakan penjelas yang siginifikan terhadap variabel dependen (model cocok).

Berdasarkan gambar 12, dapat diketahui bahwa Fs > P-value, yang artinya hipotesa yang dapat diambil adalah terima H_1 dan tolak H_0. Dapat dikatakan, variabel independen (Age, JobLevel, TotalWorkingYears, YearsAtCompany) dan MonthlyIncome berpengaruh signifikan terhadap permintaan. Pada taraf signifikansi 5% (0,05), H_0 ditolak karena nilai probabilitasnya yaitu 0,00 yang berarti dibawah dari 5%. Maka dapat disimpulkan, model yang dipakai cocok.

3) Uji-t

Uji parsial (koefisien regresi) atau disebut dengan uji-t bertujuan untuk menguji signifikan konstanta dan variabel independent yang terdapat dalam persamaan regresi secara individu. [28]

Hipotesa yang ada sebagai berikut :
H_0 = Variabel independen tidak berpengaruh signifikan.
H_1 = Variabel independen berpengaruh signifikan.
$\propto$ = 0,05 (Taraf Signikansi/Threshold)

Berdasarkan gambar 10, uji-t dapat diambil hipotesa sebagai berikut :

- Nilai variabel X_1 (Age) berada di atas taraf signifikansi/terima H_1.
- Nilai variabel X_2 (JobLevel) berada di bawah taraf signifikansi/terima H0.
- Nilai variabel X_3 (TotalWorkingYears) di bawah taraf signifikansi/terima H_0.
- Nilai variabel X_4 (YearsAtCompany) di atas taraf signifikansi/terima H_1.

Dapat diambil kesimpulan bahwa variabel independen JobLevel dan TotalWorkingYears adalah variabel yang tidak mempengaruhi variabel dependen. Sedangkan variabel independent Age dan YearsAtCompany adalah variabel yang mempengaruhi variabel dependen.

4) R-Square

R square dapat diartikan sebagai koefisien determinasi. R square menunjukan suatu persentase pengaruh antara variabel X1 dengan X2 terhadap variabel Y. [29]

Berdasarkan gambar 10, nilai koefisien determinasi (R-Square) adalah 0,909 atau 90,9%. Maka, MonthlyIncome dipengaruhi oleh faktor Age dan YearsAtCompany sebesar 0,909 atau 90,9%. Nilai sisa dari koefisien determinasi adalah 0,091 atau 9,1% dipengaruhi oleh faktor lain yang tidak diketahui.

5) Pengujian Asumsi

Untuk memvalidasi model machine learning, dilakukan analisis residual. Berikut adalah daftar pengujian atau asumsi yang akan lakukan untuk mengetahui validitas model :

- Linearitas

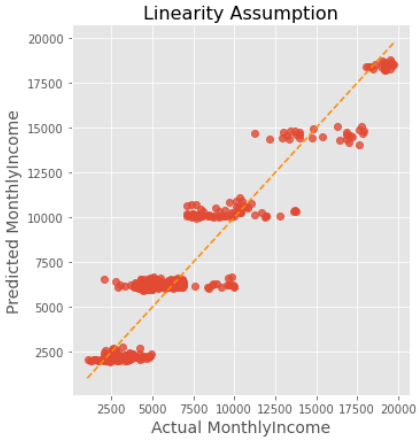
  Pengujian linearitas merupakan pengujian yang dilakukan untuk mengetahui model yang dibuktikan apakah merupakan model linear atau tidak. [30] Uji linearitas dilakukan dengan menggunakan regresi kurva, yaitu gambaran hubungan liniear antara variabel X dengan variabel Y. [30]

```
# Plotting the observed vs predicted values
sns.lmplot(x='MonthlyIncome', y='MonthlyIncome Prediction', data=df_test_new, fit_reg=False, size=5)

# Plotting the diagonal line
line_coords = np.arange(vis_test[['MonthlyIncome', 'MonthlyIncome Prediction']].min().min()-10,
                        vis_test[['MonthlyIncome', 'MonthlyIncome Prediction']].max().max()+10)
plt.plot(line_coords, line_coords,  # X and y points
         color='darkorange', linestyle='--')

plt.ylabel('Predicted MonthlyIncome', fontsize=14)
plt.xlabel('Actual MonthlyIncome', fontsize=14)
plt.title('Linearity Assumption', fontsize=16)
plt.show()
```

Gambar 13. Skrip Linearitas dan Grafik Asumsi Linier



Gambar 14. Grafik Asumsi Linier

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

- Normalitas

Uji normalitas merupakan pengujian suatu model regresi berupa variabel dependen, variabel independen ataupun keduanya apakah mempunyai distribusi normal ataukah tidak. [30] Model regresi dikatakan baik jika distribusi data normal atau mendekati normal. [30]

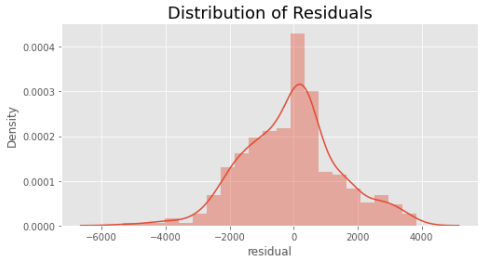
```
from statsmodels.stats.diagnostic import normal_ad

# Performing the test on the residuals
p_value = normal_ad(df_test_new2['residual'])[1]
print('p-value from the test Anderson-Darling test below 0.05 generally means non-normal:', p_value)

# Plotting the residuals distribution
plt.subplots(figsize=(8, 4))
plt.title('Distribution of Residuals', fontsize=18)
sns.distplot(df_test_new2['residual'])
plt.show()

# Reporting the normality of the residuals
if p_value < 0.05:
    print('Residuals are not normally distributed')
else:
    print('Residuals are normally distributed')
```

Gambar 15. Skrip Uji Normalitas



Gambar 16. Diagram Distribusi Residual

Berdasarkan asumsi di atas, dapat diketahui hipotesa sebagai berikut :

H_0 = Residual terdistribusi normal.
H_1 = Residual terdistribusi secara tidak normal.

Dari hasli perhitungan, 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 H_0 ditolak H_1 diterima. Sehingga asumsi normalitas terpenuhi.

- Multikolinieritas

  Dalam analisis regresi linier ganda, jika ada dua atau lebih variabel independen yang berkorelasi sangat kuat, maka dikatakan terdapat multikolinieritas. [31]

  Uji multikolinieritas dilakukan untuk melihat korelasi antar variabel prediktor. Apabila terjadi multikolinearitas pada model regresi menyebabkan parameter regresi yang dihasilkan akan memiliki error yang sangat besar. Kriteria yang digunakan utnuk mengetahui adanya multikolinearitas antara variabel prediktor adalah dengan menggunakan nilai variance inflation factors (VIF). Apabila nilai VIF lebih besar dari 10 mengindikasikan bahwa ada masalah multikolinearitas. Nilai VIF diperoleh dengan cara meregresikan variabel independent. [32]

```
corr = vis_test[['Age', 'JobLevel', 'TotalWorkingYears', 'YearsAtCompany', 'MonthlyIncome']].corr()
print('Pearson correlation coefficient matrix of each variables:\n', corr)

# Generate a mask for the diagonal cell
mask = np.zeros_like(corr, dtype=np.bool)
np.fill_diagonal(mask, val=True)

# Initialize matplotlib figure
fig, ax = plt.subplots(figsize=(4, 3))

# Generate a custom diverging colormap
cmap = sns.diverging_palette(220, 10, as_cmap=True, sep=100)
cmap.set_bad('grey')

# Draw the heatmap with the mask and correct aspect ratio
sns.heatmap(corr, mask=mask, cmap=cmap, vmin=-1, vmax=1, center=0, linewidths=.5)
fig.suptitle('Pearson correlation coefficient matrix', fontsize=14)
ax.tick_params(axis='both', which='major', labelsize=10)
# fig.tight_layout()
```

Gambar 17. Skrip Uji Multikolinieritas

```
Pearson correlation coefficient matrix of each variables:
                          Age  JobLevel  TotalWorkingYears  YearsAtCompany
Age                  1.000000  0.449794           0.623246        0.294041
JobLevel             0.449794  1.000000           0.772222        0.511125
TotalWorkingYears    0.623246  0.772222           1.000000        0.638163
YearsAtCompany       0.294041  0.511125           0.638163        1.000000
MonthlyIncome        0.439321  0.944295           0.771358        0.488875

                     MonthlyIncome
Age                       0.439321
JobLevel                  0.944295
TotalWorkingYears         0.771358
YearsAtCompany            0.488875
MonthlyIncome             1.000000
```

Gambar 18. Koefisen Korelasi Pearson

Dari hasil asumsi di atas, dapat dikatakan bahwa prediktor yang digunakan dalam regresi berkorelasi satu sama lain.

```
from patsy import dmatrices
from statsmodels.stats.outliers_influence import variance_inflation_factor

#find design matrix for linear regression model using 'rating' as response variable
y, X = dmatrices('MonthlyIncome ~ Age+JobLevel+TotalWorkingYears+YearsAtCompany', data=vis_test, return_type='dataframe')

#calculate VIF for each explanatory variable
vif = pd.DataFrame()
vif['VIF'] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
vif['variable'] = X.columns

#view VIF for each explanatory variable
vif
```

Gambar 19. Skrip Multikolinearitas

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

Gambar 20. Tabel VIF

Berdasarkan gambar 20, dapat dilihat nilai variabel Age, JobLevel, TotalWorkingYears, YearsAtCompany memiliki nilia kurang dari 10 sehingga dengan menggunakan tingkat signifikansi sebesar 0,05 dapat disimpulkan bahwa pada data tersebut tidak terdapat multikolinearitas pada variabel-variabel prediktor.

- Autokorelasi

  Autokorelasi merupakan pengujian untuk menguji apakah dalam sebuah model regrsi linier berganda terdapat korelasi antara kesalahan pengganggu pada periode t dengan kesalahan pada periode t1 (sebelumnya). Jika terjadi korelasi, maka hal itu disebut sebagai autokorelasi. Model regresi yang baik adalah terbebas dari autokorelasi. [30]

  Pada langkah ini, 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.

```
from statsmodels.stats.stattools import durbin_watson

durbinWatson = durbin_watson(df_test_new2['residual'])

print('Durbin-Watson:', durbinWatson)
if durbinWatson < 1.5:
    print('Signs of positive autocorrelation', '\n')
    print('Assumption not satisfied')
elif durbinWatson > 2.5:
    print('Signs of negative autocorrelation', '\n')
    print('Assumption not satisfied')
else:
    print('Little to no autocorrelation', '\n')
    print('Assumption satisfied')
```

Gambar 21. Skrip Uji Autokorelasi

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

Assumption satisfied
```

Gambar 22. Output Uji Autokorelasi

Dari hasil output pada gambar 20, dapat diasumsikan bahwa terdapat sedikir atau tidak ada autokorelasi, yang berarti asumsi puas.

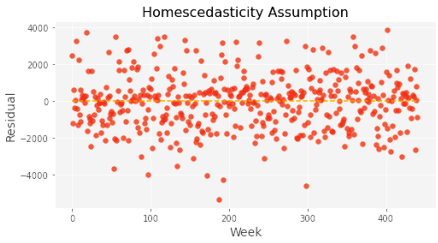
- Homoskedatisitas

  Homoskedastisitas merupakan pengujian untuk menguji apakah dalam sebuah model regresi linier berganda terdapat gejala heteroskedasitas atau tidak dengan cara melihat ada tidaknya pola tertentu pada grafik scatterplots. Model regresi dikatakan baik apabila model tersebut homoskedastisitas atau tidak terjadi heteroskesdatisitas. [33]

```
# Plotting the residuals
plt.subplots(figsize=(8, 4))
plt.scatter(x=df_test_new2.index, y=df_test_new2.residual, alpha=0.8)
plt.plot(np.repeat(0, len(df_test_new2.index)+2), color='darkorange', linestyle='--')

plt.ylabel('Residual', fontsize=14)
plt.xlabel('Week', fontsize=14)
plt.title('Homescedasticity Assumption', fontsize=16)
plt.show()
```

Gambar 23. Skrip Plot Penyebaran Residual



Gambar 24. Plot Penyebaran Residual

Dari grafik scatterplot (gambar 22), terlihat titik-titik residual menyebar secara acak, serta tersebar baik di atas maupun di bawah angka 0 (nol) pada sumbu Y. Oleh karena itu, dapat diambil kesimpulan

bahwa tidak terdapat gejala heteroskedastisitas pada model regresi yang digunakan.

B. Implementasi Antarmuka Aplikasi

B.1. Halaman prediksi gaji

Tampilan untuk aplikasi untuk melakukan prediksi gaji pegawai terlihat pada Gambar 25.

Gambar 25. Halaman Prediksi Gaji

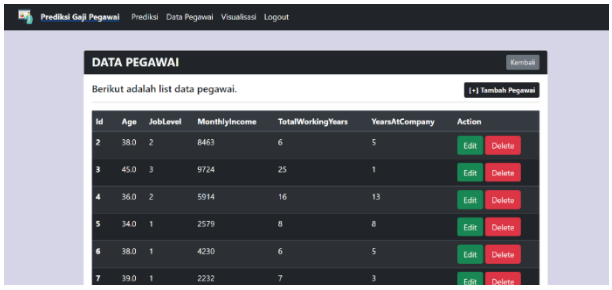
B.2. Halaman hasil prediksi gaji

Tampilan untuk aplikasi menampilkan hasil prediksi gaji pegawai terlihat pada Gambar 26.

Gambar 26. Halaman Hasil Prediksi gaji

B.3. Halaman data pegawai

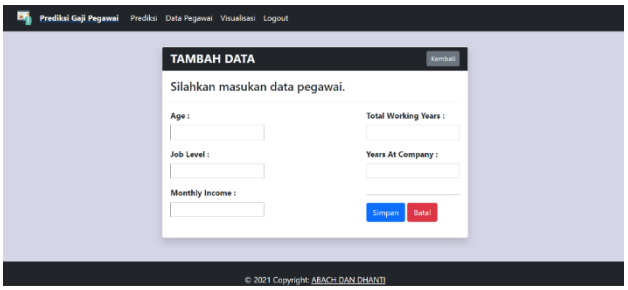
Tampilan aplikasi untuk menampilkan data pegawai yang terdiri dari atribut Id, Age, JobLevel, MonthlyIncome, dan Action. Halaman data pegawai terlihat pada Gambar 27.



Gambar 27. Halaman Data Pegawai

B.4. Halaman tambah data pegawai

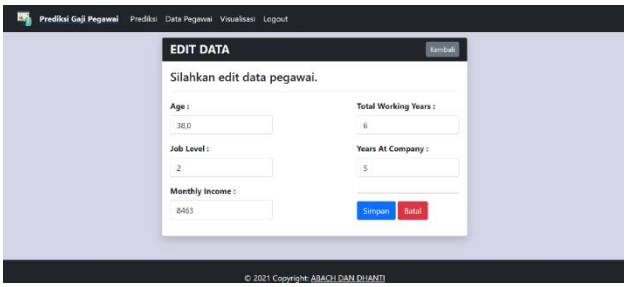
Tampilan aplikasi untuk menginputkan data pegawai yang baru. Form input terdiri dari Age, Job Level, Monthly Income, dan Total Working Years, dan Years At Company. Halaman tambah data pegawai terlihat pada Gambar 28.



Gambar 28. Halaman tambah data Pegawai

B.5. Halaman edit data pegawai

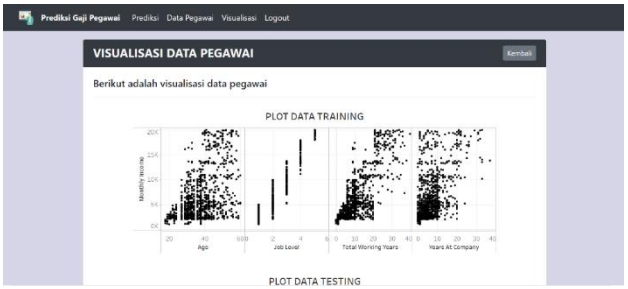
Tampilan untuk aplikasi mengedit data pegawai yang baru. Form input untuk mengedit data terdiri dari Age, Job Level, Monthly Income, dan Total Working Years, dan Years At Company. Halaman edit data pegawai terlihat pada Gambar 29.



Gambar 29. Halaman Edit Data Pegawai

B.5. Halaman visualisasi

Tampilan aplikasi untuk visualisasi dari model yang dibuat. Halaman visualisasi terlihat pada Gambar 30.



Gambar 30. Halaman Visualisasi

V. KESIMPULAN DAN SARAN

A. Kesimpulan

Berdasarkan model yang diambil dari model OLS didapatkan nilai akurasi sebesar 0,909. Akurasi tersebut merupakan nilai akurasi yang baik, sehingga dapat dikatakan model *machine learning* dapat berperforma baik

untuk memprediksi gaji. Model prediksi yang dirancang dengan menggunakan *machine learning* dengan pendekatan regresi, berhasil melewati semua pengujian dalam langkah validasi model, sehingga kami dapat menyimpulkan bahwa model kami dapat berperforma baik untuk memprediksi gaji karyawan dengan menggunakan empat variabel independen, yaitu Age, JobLevel, TotalWorkingYears, dan YearsAtCompany.

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%.

Visualisasi data dari hasil model prediksi gaji karyawan dapat digunakan menjadi bentuk aplikasi berbasis *web base* dengan menggunakan *framework* Django. Dengan aplikasi tersebut, admin dapat melakukan prediksi gaji karyawan dengan mudah dan dengan cepat.

B. Saran

Saran yang dapat disampaikan pada peneliti yang akan melanjutkan dan mengembangkan penelitian ini adalah :

- Pembuatan model prediksi yang digunakan dapat lebih beragam untuk membandingkan performa antara model satu dengan model yang lainnya.
- 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.

REFERENCES

[1] Prasetyo B and Trisyanti U, "REVOLUSI INDUSTRI 4.0 DAN TANTANGAN PERUBAHAN SOSIAL", *Journal of Proceedings Series*, no. 5, pp. 22-27, Nov. 2018, doi : http://dx.doi.org/10.12962/j23546026.y2018i5.4417

[2] H. Prasetyo and W. Sutopo, "Perkembangan Keilmuan Teknik Industri Menuju Era Industri 4.0", *Seminar dan Konferensi Nasional IDEC*, vol. 2017, pp. 488-495, May .2017, doi : https://idec.ft.uns.ac.id/wp-content/uploads/2017/11/Prosiding2017_ID069.pdf

[3] O. C. Pangaribuan and I. Irwansyah, "Media Cetak Indonesia di Era Revolusi Industri 4.0," *Jurnal Pewarta Indonesia*, vol. 1, no. 2, pp. 134–145, Oct. 2019, doi: https://dx.doi.org/10.25008/jpi.v1i2.11

[4] A. A. Shahroom and N. Hussin, "Industrial Revolution 4.0 and Education," *International Journal of Academic Research in Business and Social Sciences*, vol. 8, no. 9, pp. 314-319, Oct. 2018, doi: https://doi.org/10.24114/jh.v10i1.14138

[5] S. Kergroach, "Industry 4.0: New Challenges And Opportunities For The Labour Market," *Foresight and STI Governance*, vol. 11, no. 4, pp. 6–8, 2017, doi: http://dx.doi.org/10.17323/2500-2597.2017.4.6.8

[6] M. I. Manda and S. ben Dhaou, "Responding to the challenges and opportunities in the 4th industrial revolution in developing countries", *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, Part F148155, pp. 244–253, 2019, doi: http://dx.doi.org/10.1145/3326365.3326398

[7] Y. Adrianova Eka Tuah and Anyan, "IMPLEMENTASI MODEL REGRESI LINEAR SEDERHANA UNTUK PREDIKSI GAJI BERDASARKAN PENGALAMAN LAMA BEKERJA", *Journal Education and Technology*, vol. 1, no. 2, pp. 56-70 Dec. 2020, doi : https://doi.org/10.31932/jutech.v1i2.1289

[8] Tamrin A.S, Rumapea Patar, Mambo R, "PENGARUH PROFESIONALISME KERJA PEGAWAI TERHADAP TINGKAT KEPUASAN PELANGGAN PADA KANTOR PT. TASPEN CABANG MANADO", *Jurnal Administrasi Publik*, vol. 3, no. 46, pp. 1-9 2017, doi : https://ejournal.unsrat.ac.id/index.php/JAP/article/view/16283

[9] E. P. Ariesanto Akhmad, "Data Mining Menggunakan Regresi Linear untuk Prediksi Harga Saham Perusahaan Pelayaran," *Jurnal Aplikasi Pelayaran dan Kepelabuhanan*, vol. 10, no. 2, p. 120, Dec. 2020, doi: https://dx.doi.org/10.30649/japk.v10i2.83

[10] K. Puteri and A. Silvanie, "*MACHINE LEARNING* UNTUK MODEL PREDIKSI HARGA SEMBAKO DENGAN METODE REGRESI LINIER BERGANDA", Jurnal Nasional Informatika, vol. 1, no. 2, pp. 82-94, Oct. 2020, doi : https://ejournal-ibik57.ac.id/index.php/junif/article/view/134

[11] A. Saiful, S. Andryana, and A. Gunaryati, "Prediksi Harga Rumah Menggunakan Web ScrappingDan *Machine learning* Dengan Algoritma Linear Regression", *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, no. 1, pp. 41-50, Mar. 2012, doi : https://jurnal.mdp.ac.id/index.php/jatisi/article/download/701/219/

[12] M. W. Pertiwi and R. E. Indrajit, "Metode Regresi Linier Untuk Prediksi Pengadaan Inventaris Barang", Simposium Nasional Ilmu Pengetahuan dan Teknologi (SIMNASIPTEK) 2017, vol. 1, no. 1, pp. 27-30, 2017, doi : https://seminar.bsi.ac.id/simnasiptek/index.php/simnasiptek-2017/article/view/114

[13] W. Wahyudin and H. Purwanto, "PREDIKSI KASUS COVID-19 DI INDONESIA MENGGUNAKAN METODE BACKPROPAGATION DAN REGRESI LINEAR," *Journal of Information System, Applied, Management, Accounting and Research*, vol. 5, no. 2, p. 331, May 2021, doi: https://doi.org/10.52362/jisamar.v5i2.420

[14] N. Nafi'iyah, "Perbandingan Regresi Linear, Backpropagation Dan Fuzzy Mamdani Dalam Prediksi Harga Emas," *Seminar Nasional Inovasi Dan Aplikasi Teknologi Di Industri (SENIATI) 2016*, vol. 2, pp. 291-296, Mar. 2016, doi : https://ejournal.itn.ac.id/index.php/seniati/article/download/840/767/

[15] F. Ginting, E. Buulolo, and E. R. Siagian, "IMPLEMENTASI ALGORITMA REGRESI LINEAR SEDERHANA DALAM MEMPREDIKSI BESARAN PENDAPATAN DAERAH (STUDI KASUS: DINAS PENDAPATAN KAB. DELI SERDANG)," *KOMIK (Konferensi Nasional Teknologi Informasi dan Komputer)*, vol. 3, no. 1, Nov. 2019, doi: http://dx.doi.org/10.30865/komik.v3i1.1602

[16] P. Katemba and K.D. Rosita, "PREDIKSI TINGKAT PRODUKSI KOPI MENGGUNAKAN REGRESI LINEAR", *Jurnal Ilmiah Flash*, vol. 3, pp. 42-51, Jun.

2017, doi: http://jurnal.pnk.ac.id/index.php/flash/article/view/136/79

[17] T. N. Putri, A. Yordan, and D. H. Lamkaruna, "Peramalan Penerimaan Mahasiswa Baru Universitas Samudra Menggunakan Metode Regresi Linear Sederhana," *Jurnal Teknologi Informtika*, vol. 2, no. 1, Mar. 2019, doi : http://jurnal.ummu.ac.id/index.php/J-TIFA/article/view/237/149

[18] D. Sayan, B. Rupashri, M. Ayush, "SALARY PREDICTION USING REGRESSION TECHNIQUES.", *Proceedings of Industry Interactive, Innovations in Science, Engineering & Technology*, Jan. 2020, doi : https://dx.doi.org/10.2139/ssrn.3526707

[19] U. Bansal, A. Narang, A. Sachdeva, I. Kashyap, and S. P. Panda, "Empirical Analysis Of Regression Techniques By House Price And Salary Prediction," *IOP Conference Series: Materials Science and Engineering*, vol. 1022, no. 1, pp. 1-13, Jan. 2021, doi: https://iopscience.iop.org/article/10.1088/1757-899X/1022/1/012110

[20] X. Pan, X. Wan, H. Wang, and Y. Li, "The Correlation Analysis Between Salary Gap and Enterprise Innovation Efficiency Based on the Entrepreneur Psychology," *Frontiers in Psychology*, vol. 11, Aug. 2020, doi: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.01749/full

[21] S. Dan and B. Pratikno, "REGRESI LINEAR BIVARIAT SIMPEL DAN APLIKASINYA PADA DATA CUACA DI CILACAP", *JMP*, vol. 6, no. 1, pp. 45-52, Jun. 2014, doi : http://dx.doi.org/10.20884/1.jmp.2014.6.1.2902

[22] T. N. Padilah and R. I. Adam, "ANALISIS REGRESI LINIER BERGANDA DALAM ESTIMASI PRODUKTIVITAS TANAMAN PADI DI KABUPATEN KARAWANG", *Jurnal Pendidikan Matematika dan Matematika*, vol. 5, no. 2, pp. 117-128, Dec. 2019, doi : https://jurnal.umj.ac.id/index.php/fbc/article/view/3333

[23] P.E.N. Desak, S. Made, "UNIVERSITAS UDAYANA FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM JURUSAN MATEMATIKA", *Conference: Prosiding Seminar Nasional Matematika II*, vol. 2, pp. 43-54, Oct. 2016, doi: https://simdos.unud.ac.id/uploads/file_penelitian_1_dir/24a473ea40f085c51398cd477f586a3a.pdf

[24] J. Homepage, A. Roihan, P. Abas Sunarya, and A. S. Rafika, "Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper," *Indonesian Journal on Computer and Information Technology*, vol. 5, no. 2, pp. 75-82, Maret 2020.

[25] Akbar Iskandar, Muttaqin, Sarini Vita Dewi, Jamaludin, Irawati HM Cahyo Prianto, Rosmita Sari Siregar, Muhammad Noor Hasan Siregar Dina Chamidah, Marzuki Sinambela, Albinur Limbong Yusra Fadhillah, Janner Simarmata, "Statistika Bidang Teknologi Informasi", 1st ed. Yayasan Kita Menulis, 2021.

[26] D. K. Barupal and O. Fiehn, "Generating the blood exposome database using a comprehensive text mining and database fusion approach," *Environmental Health Perspectives*, vol. 127, no. 9, Sep. 2019, doi: 10.1289/EHP4713.

[27] D. Saputra and R. Fathoni Aji, "ANALISIS PERBANDINGAN PERFORMA WEB SERVICE REST MENGGUNAKAN FRAMEWORK LARAVEL, DJANGO DAN RUBY ON RAILS UNTUK AKSES DATA DENGAN APLIKASI MOBILE (Studi Kasus: Portal E-Kampus STT Indonesia Tanjungpinang)," *Bangkit Indonesia*, vol. 2, no. 7, pp. 17-22, Oct 2018.

[28] Dita Anggun Lestari, Sarini Abdullah, "ANALISIS TINGKAT KESEHATAN DAN EFISIENSI PERBANKAN TERHADAP PROFITABILITAS BANK MENGGUNAKAN REGRESI BERGANDA DAN ANOVA", *Indonesian Journal of Statistics and Its Applications*, vol. 4, no. 3, pp. 401-418, 2020.

[29] E. Khumaedi, "PENGARUH DISIPLIN DAN MOTIVASI KERJA TERHADAP KINERJA PEGAWAI PADA DINAS SENTRA OPERASI TERMINAL PT.ANGKASA PURA II," *Jurnal Ilmiah Manajemen dan Bisnis*, vol. 2, no. 1, pp. 66-77, Mar 2016.

[30] N. Sitti, K. Sekolah, T. Ilmu, and E. Gempol, "ANALISIS EKUITAS MEREK PRODUK NOTEBOOK ASUS TERHADAP KEPUTUSAN PEMBELIAN KONSUMEN PADA DISTRIBUTOR DIVA JAYA CABANG SIDOARJO," *JURNAL AKUNTANSI DAN MANAJEMEN*, vol. 3, no. 2, pp. 73-83, 2018.

[31] Prof. Dr. Suyono, M.Si ."Analisis Regresi untuk Penelitian". 1st ed. Sleman : deepublish. 2015.

[32] R. G. Ali and J. Nugraha, "PENERAPAN METODE REGRESI RIDGE DALAM MENGATASI MASALAH MULTIKOLINEARITAS PADA KASUS INDEKS PEMBANGUNAN MANUSIA DI INDONESIA TAHUN 2017," *Prosiding Sendika*, vol. 5, no. 22, pp. 226-235, 2019. Available: www.statistik.data.kemdikbud.go.id

[33] O. : Nurfajar, ; M Syafiq Marzuqi, N. Rohmayati, U. Sultan, and A. Tirtayasa, "PENGARUH EMPLOYEE ENGAGEMENT DAN EFIKASI DIRI TERHADAP KINERJA KARYAWAN PT NIKOMAS GEMILANG DIVISI PCI S5 SERANG BANTEN," *JURNAL PENGEMBANGAN WIRASWASTA*, vol. 20, no. 1, pp. 35-46, 2018. [Online]. Available: http://ejurnal.stieipwija.ac.id/index.php/jpw