Jurnal Informatika dan Teknologi (JICode) Vol.1 No.1 Halaman:1-8, Februari, 2024
ISSN:

METODE ANALASIS MENGGUNAKAN ALGORITMA RANDOM FOREST UNTUK PREDIKSI BIAYA ASURANSI KESEHATAN

Dody Indra Sumantiawan[1]
Fakultas Sain dan Teknologi, Universitas Nasional Karangturi
Email: dody.indra@unkartur.ac.id [1]

ABSTRAK

Machine learning adalah cabang ilmu komputer yang memungkinkan komputer untuk belajar tanpa diprogram secara eksplisit. Salah satu tugas utama dalam machine learning adalah prediksi, yaitu memperkirakan nilai variabel target berdasarkan variabel lain. Dalam penelitian ini algoritma *Random Forest* dipilih karena algoritma yang powerful untuk prediksi dan memiliki banyak keuntungan, seperti akurasi tinggi, stabilitas tinggi, dan mudah diinterpretasikan. Prediksi yang dilakukan adalah pada awalnya seorang perokok dan orang yang berat badannya tidak ideal akan membayar biaya asuransi yang lebih tinggi dibandingkan dengan orang yang tidak merokok dan orang yang memiliki berat badan ideal. Data yang digunakan dalam penelitian ini menggunakan dataset yang berasal dari https://raw.githubusercontent.com/stedy/Machine-Learning-with-R-datasets/master/insurance.csv yang diolah dengan algortima *random forest*. Pre-processing data merupakan proses merubah data mentah menjadi format yang bersih dan siap untuk dimodelkan dengan tahapan permbersihan data, transformasi data, pengurangan data, sampling data. Metode penelitian dilakukan dengan memeriksa data dari data yang salah atau dapat mengganggu proses analisis, melakukan analisis pada dataset serta membagi data menjadi data training dan data test. Proses pembagian data adalah 80 digunakan untuk data training dan 20 untuk data test. Proses analisis diolah dengan menggunakan bahasa pemograman *python*. Hasil penelitian menunjukkan hasil train 23051.135798, hasil test 45302.143621 dan hasil prediksi 5956.88 dengan y_true 5934.3798. Data dengan nilai aktual yang ditunjukan pada nilai y_true memprediksi kedekatan nilai hampir mendekati nilai aktual dengan nilai 5956.88, dan sudah mendekati hasil baik.

Kata kunci: Asuransi, *Machine learning*, *Python*, *Random Forest*.

1. PENDAHULUAN

Biaya asuransi kesehatan merupakan salah satu faktor penting yang harus dipertimbangkan calon peserta asuransi. Biaya asuransi kesehatan yang terlalu tinggi dapat menjadi beban bagi peserta, sedangkan biaya yang terlalu rendah dapat menyebabkan peserta tidak mendapatkan manfaat yang optimal dari asuransi kesehatan .

Prediksi biaya asuransi kesehatan menjadi penting untuk dilakukan. Prediksi biaya asuransi kesehatan dapat membantu peserta asuransi untuk menentukan pilihan asuransi kesehatan yang sesuai dengan kebutuhan dan kemampuan finansial yang dimiliki.

Analisis machine learning merupakan salah satu metode yang dapat digunakan untuk memprediksi biaya asuransi kesehatan. Analisis machine learning dapat digunakan untuk menganalisis data historis biaya asuransi kesehatan untuk menemukan pola dan tren yang dapat digunakan untuk memprediksi biaya asuransi kesehatan di masa depan [1].

Perkembangan teknologi machine learning ini membuka peluang untuk menggunakan analisis machine learning dalam berbagai bidang, termasuk prediksi biaya asuransi kesehatan. Hal ini didukung oleh ketersediaan data yang semakin besar dan semakin canggihnya perangkat keras dan perangkat lunak [2].

Kompleksitas data biaya asuransi kesehatan umumnya terdiri dari berbagai variabel, baik variabel yang bersifat kuantitatif maupun kualitatif. Variabel-variabel ini dapat saling berinteraksi satu sama lain, sehingga sulit untuk dianalisis secara manual. Analisis machine learning dapat digunakan untuk menganalisis data biaya asuransi kesehatan yang kompleks ini. Analisis machine learning dapat menemukan pola dan tren yang tidak dapat dideteksi oleh analisis manual [3].

Berkembangnya teknologi informasi dalam bidang kecerdasan buatan, Teknik *machine learning* membantu meningkatkan pendeteksian secara otomatis. Metode yang digunakan dalam penelitian ini menggunakan prediksi *Random Forest.* Algoritma *Random Forest* merupakan algoritma *machine learning* yang paling umum digunakan untuk prediksi. Masing-masing algoritma memiliki kelebihan dan kekurangannya masing-masing[4, 5, 6].

Algoritma *Random Forest* adalah algoritma pembelajaran mesin yang termasuk dalam kategori ensemble learning. Ensemble learning menggabungkan beberapa model pembelajaran mesin untuk menghasilkan prediksi yang lebih akurat dan stabil dibandingkan dengan model tunggal.

Algoritma ini membangun banyak pohon keputusan, masing-masing dari subset data dan fitur yang berbeda. Setiap pohon "memutuskan" ke mana harus pergi berdasarkan kriteria tertentu, dan hasil akhir ditentukan dengan "voting" atau rata-rata dari semua pohon.

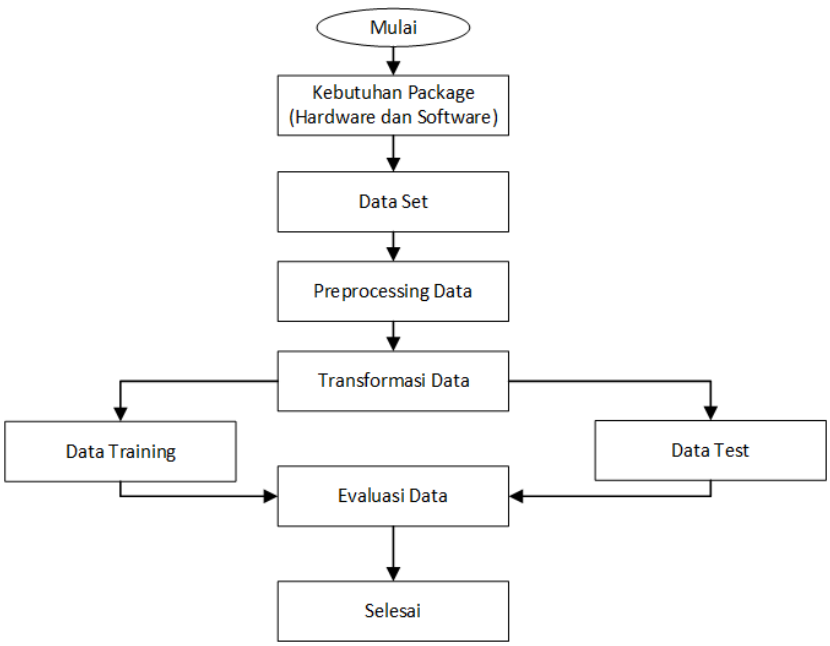
2. METODOLOGI PENELITIAN

Metode prediksi merupakan proses yang digunakan untuk memperkirakan nilai, hasil, atau perilaku dimasa mendatang, metode prediksi dapat digunakan dalam berbagai bidang, salah satu dalam penelitian ini adalah untuk memprediksi biaya asuransi kesehatan.

Data yang digunakan dalam penelitian ini menggunakan dataset yang berasal dari https://raw.githubusercontent.com/stedy/Machine-Learning-with-R-datasets/master/insurance.csv yang diolah dengan algortima *random forest* [6].

Pre-processing data merupakan proses merubah data mentah menjadi format yang bersih dan siap untuk dimodelkan dengan tahapan permbersihan data, transformasi data, pengurangan data, sampling data. Langkah ini harus dilakukan sebelum melakukan pemodelan machine learning. Proses ini membantu untuk memastikan data yang digunakan untuk analisis adalah akurat, bersih, dan siap untuk digunakan [7].

Analisis data dalam machine learning adalah proses memahami data untuk membuat keputusan atau prediksi yang lebih baik. Proses ini menggunakan algoritma machine learning untuk mengekstrak informasi dari data dan mengidentifikasi pola atau tren yang tidak terlihat secara kasat mata. Analisis data dalam machine learning dapat menjadi proses yang kompleks, tetapi dapat memberikan manfaat yang signifikan bagi bisnis dan organisasi. Dengan memahami data, perusahaan dapat membuat keputusan yang lebih baik, meningkatkan efisiensi, dan meningkatkan keuntungan. Alur penelitian dapat dilihat pada gambar 1.



Gambar 1 Alur Penelitian

3. HASIL DAN PEMBAHASAN

3.1. Hasil

Metode yang digunakan sebagai evaluasi dalam penelitian ini adalah KNN, *Regresi* dan SVM. Pengujian menggunakan 1338 data dan terdiri dari 7 kolom. Hasil akhir analisis dilakukan perbandingan antara y aktual dengan y prediksi baik dalam bentuk tabel maupun grafik.

3.2. Pembahasan

3.2.1. Perangkat Lunak

Python dalam penelitian ini digunakan sebagai alat analisis, dengan berbagai library yang yang digunakan dalam bahasa pemograman python diantaranya adalah numpy, pandas, matplotlib, seabort dan sklearn. Analisis data dalam bahasa pemograman python menggunakan library numpy, visualisasi data untuk menampilkan data dalam penelitian ini menggunakan library matplotlib dan searborn dan scikit-learn sebagai *machine learning*.

3.2.2. Analisis Data

Proses analisis prediksi dipengaruhi oleh beberapa variabel yaitu age, sex, bmi, children, smoker, dan region. Bentuk regresi linier dinotasikan dengan Y = b + m1*x1 + m2*x2 + m3*x3 + m4*x4 + m5*x5 + m6*x6.

Dimana :

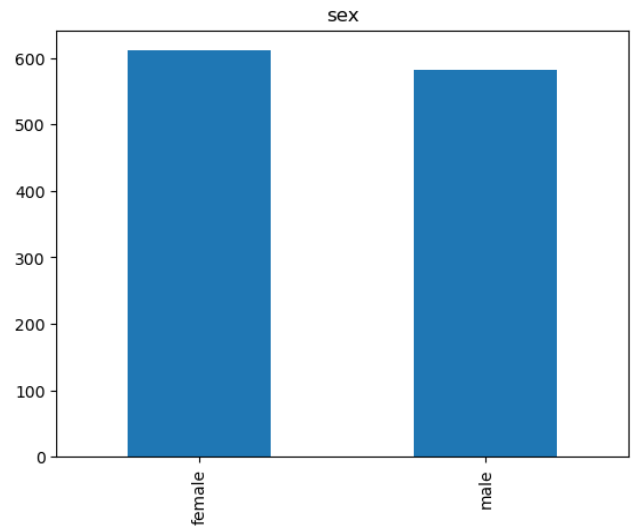
Y	= Dependent variable (charge)
b	= Intercept
m1..6	= Koefisien dari persamaan
x1	= Variabel independen 1 (age)
x2	= Variabel independen 2 (sex)
x3	= Variabel independen 3 (bmi)
x4	= Variabel independen 4 (children)
x5	= Variabel independen 5 (smoker)
x6	= Variabel independen 6 (region)

3.2.3. Visualisasi data

Visualisasi data adalah proses mempresentasikan data dalam bentuk visual yang menarik dan mudah dipahami. Visualisasi data dapat ditampilkan dalam bentuk box plot, histogram dan bentuk lainnya. Matplotlib adalah *library yang paling umum digunakan untuk visualisasi data Python.* Matplotlib memberikan berbagai macam plot dan grafik, dan dapat digunakan untuk membuat visualisasi yang sederhana dan kompleks sesuai dengan kebutuhan.

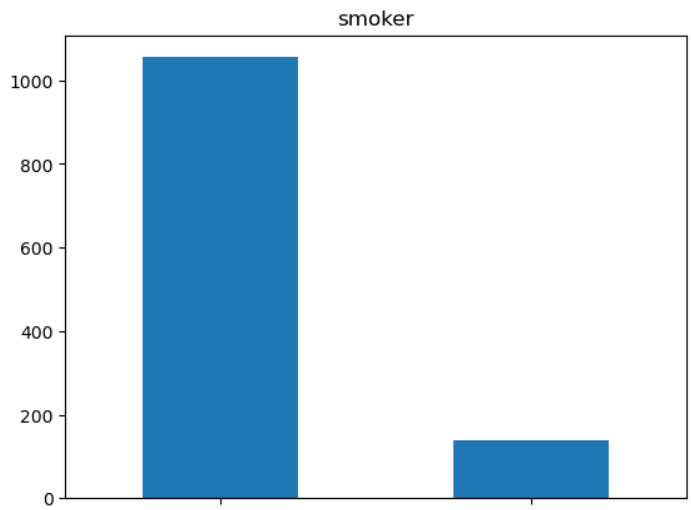
```
import pandas as pd
import matplotlib.pyplot as plt

def plot_categorical(data, categorical_features):
    for feature in categorical_features[:3]:
        count = data[feature].value_counts()
        percent = 100*data[feature].value_counts(normalize=True)
        df          =          pd.DataFrame({'jumlah          sampel':count,
'persentase':percent.round(1)})
        print(df)
        count.plot(kind='bar', title=feature)
        plt.show()
plot_categorical(data, categorical_features)
```

hasil dari visualisasi data dapat dilihat pada gambar 2, gambar 3 dan gambar 4.



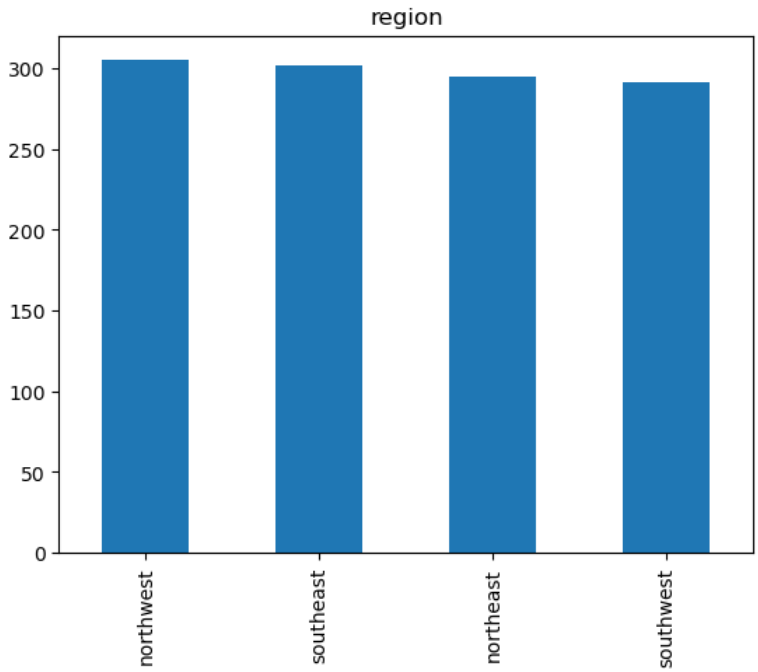
Gambar 2 Variabel Sex

Dari hasil penellitian yang dilakukan dengan screening data pada jumlah sampel pada persentasi pada variabel sex menunjukkan jumlah sampel pada *female* sebesar 611 dan persentase *female* 51,2. Sedangkan pada sampel *male* sebesar 582 dan persentase *male* 48,8, hal ini menunjukkan jika lebih banyak *female* yang banyak menggunakan asuransi untuk menunjang kesehatan dari pada *male*.



Gambar 3 Persentase *Smoker*

Penelitian ini juga diambil dari perokok aktif dengan hasil screening data ditemukan berdasarkan jumlah sampel pada persentasi pada variabel *Smoker* menunjukkan jumlah sampel pada *no* sebesar 1055 dan persentase *no* 88,4. Sedangkan pada sampel *yes* sebesar 138 dan persentase *yes* 11,6. Hal ini menunjukkan jika banyak perokok aktif yang mengunakan asuransi kesehatan.



Gambar 4 Region

Jumlah sampel pada persentasi pada variabel *region* menunjukkan northwest memiliki sampel 305 dan persentase 25.6, southeast memiliki sampel 302 dan persentase 25.3, northeast memiliki sampel 295 dan persentase 24.7 sedangkan southwest memiliki sampel 291 dan persentase 24.4.

Variabel sex didominasi oleh female, maka dapat diperkirakan perkirakan bahwa biaya medis rata-rata untuk wanita dapat lebih tinggi daripada laki-laki. Untuk variabel smoker didominasi oleh non-perokok sehingga dapat diasumsikan bahwa biaya medis rata-rata untuk perokok lebih tinggi daripada non-perokok. Oleh karena itu, mayoritas individu tidak merokok sehingga dapat mengharapkan biaya medis rata-rata yang lebih rendah. Sementara itu, jika mayoritas individu dalam dataset berasal dari wilayah northwest.

3.2.4. Explorasi Data

Explorasi data yang digunakan dalam penelitian ini menggunakan EDA (*exploraty data analysis*). EDA digunakan untuk mempelajari dan menentukan proses pengolahan data. Pemeriksaan data dilakukan dalam tahap ini, mulai dari memilah data kosong, menghapus data yang sama dan mengubah data menjadi numerik. Pengecekan informasi dari dataset yang digunakan dapat dilihat dari hasil yang ditampilkan pada gambar 2 dan gambar 3.

```
[6]: data.describe()
```

[6]:

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

Gambar 5 Data Statistika

Analisis yang disajikan dalam gambar 2 digunakan untuk melihat data statistika, data tersebut digunakan untuk melihat apakah ada data yang tidak wajar. Age sebagai salah satu sampel yang diambil dengan age tertinggi 64 yang menujukan nilai wajar, demikian banyak anak maksimal 5 terlihat wajar dan disimpulkan data sudah benar.

```
[4]: data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

Gambar 6 Informasi dataset

Informasi dataset yang ditampilkan pada gambar 3 menyimpulkan jumlah 1338 data dan 7 kolom. Pada gambar 2 menunjukan tidak ada data kosong.

3.2.5. Data Preparation

Proses mengubah variabel kategorial pada data menggunakan teknik *One-Hot Encoding* untuk merubah variabel kategorikal menjadi numerik agar dapat diproses oleh *machine learning.*

```
from sklearn.preprocessing import OneHotEncoder

def one_hot encode(data, columns):
    encoder = OneHotEncoder(handle_unknown='ignore')
    encoded = encoder.fit_transform(data[columns])
    labels = encoder.get_feature_names_out(columns)
    encoded_df = pd.DataFrame(encoded.toarray(), columns=labels)
    data = pd.concat([data.drop(columns, axis=1), encoded_df], axis=1)
    return data
```

One-Hot Encoding dengan Mengubah nilai kategori menjadi variabel dummy. Contohnya, untuk variabel "region" dengan kategori "northeast", "southeast", "southwest", dan "northwest", maka one-hot encoding akan membuat empat kolom baru yang masing-masing merepresentasikan satu kategori dan bernilai 1 jika sampel memiliki kategori tersebut, dan 0 jika tidak.

3.2.6. Train-Test-Split

Pemisahan dataset menjadi data train dan data test dilakukan untuk menguji performa model pada data yang belum terlihat sebelumnya. Perbandingan dataset yang dibagi menjadi data train dan data set adalah 80:20.

```
from sklearn.model_selection import train_test_split

X = data.drop(["charges"],axis =1)
y = data["charges"]
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2,
random_state = 123)
```

3.2.7. Normalisasi data

Normalisasi data numerik dilakukan dengan mengubah data age menjadi skala yang memiliki mean=0 dan standar devisiasi=1. Hal tersebut dilakukan untuk menghindari pengaruh dari skala atau satuan yang berbeda pada model *machine learning* yang digunakan.

```
from sklearn.preprocessing import StandardScaler

numerical_features = ['age']
scaler = StandardScaler()
scaler.fit(X_train[numerical_features])
X_train[numerical_features]       =       scaler.transform(X_train.loc[:,
numerical_features])
X_train[numerical_features].head()
```

3.2.8. Modeling

Modeling dilakukan setelah proses data preparation, modeling dalam penelitian ini menggunakan model algoritma *random forest*. *Random forest* adalah algoritma machine learning yang termasuk dalam kategori ensemble learning. Ensemble learning adalah teknik yang menggabungkan beberapa model machine learning menjadi satu model yang lebih kuat. Pada algoritma Random Forest, terdapat beberapa pohon keputusan (decision tree) yang dibuat dengan sampel data yang diambil secara acak (random). Setiap pohon akan memberikan prediksi, kemudian hasil akhir prediksi akan diambil berdasarkan mayoritas prediksi dari semua pohon. Parameter yang perlu diatur pada Random Forest antara lain jumlah pohon, kedalaman pohon, dan jumlah fitur yang diambil secara acak.

```
# Impor library
from sklearn.ensemble import RandomForestRegressor

# buat model prediksi
RF      =      RandomForestRegressor(n_estimators=100,      max_depth=35,
random_state=123, n_jobs=-1)
RF.fit(X_train, y_train)

models.loc['train_mse','RandomForest']                                    =
mean_squared_error(y_pred=RF.predict(X_train), y_true=y_train)

# Menampilkan hasil prediksi
print("Prediction:", prediction)
```

4. KESIMPULAN

Prediksi merupakan salah satu bagian dari *machine learning.* Simulasi penerapan *machine learning* dalam penelitian ini menggunakan model algoritma *random forest.* Implementasi prediksi dalam penelitian ini diterapkan pada data asuransi kesehatan yang dipengaruhi data *age, sex, bmi, children, smoker* dan *region.*

Berdasarkan hasil modeling dengan menggunakan algoritma *random forest* didapatkan hasil train 23051.135798, hasil test 45302.143621 dan hasil prediksi 5956.88 dengan y_true 5934.3798. Data dengan nilai aktual yang ditunjukan pada nilai y_true memprediksi kedekatan nilai hampir mendekati nilai aktual dengan nilai 5956.88, dan sudah mendekati hasil baik.

Bahasa pemograman *python* yang digunakan dalam implementasi *machine learning* dalam penelitian ini sangat membantu dan sangat mendukung proses modeling. Hal ini karena library-library yang tersedia di *python.*

DAFTAR PUSTAKA

[1] B. D. F. Kurniatullah and Y. T. C. Pramudi, "Estimation of Students' Graduation Using Multiple Linear Regression Method," Journal of Applied Intelligent System, vol. 2, no. 1, pp. 29–36, 2017, doi: 10.33633/jais.v2i1.1415

[2] D. I. Sumantiawan, J. E. Suseno, and W. A. Syafei, "Sentiment Analysis of Customer Reviews Using Support Vector Machine and Smote-Tomek Links For Identify Customer Satisfaction," Jurnal Sistem Informasi Bisnis, vol. 13, no. 1, pp. 1-9, Jun. 2023. https://doi.org/10.21456/vol13iss1pp1-9

[3] E. D. Wahyuni, A. A. Arifiyanti, and M. Kustyani, "Exploratory Data Analysis dalam Konteks Klasifikasi Data Mining," in Prosiding Nasional Rekayasa Teknologi Industri dan Informasi XIV Tahun 2019 (ReTII), 2019, pp. 263–269

[4] E. Karamazova, T. Jusufi Zenku, and Z. Trifunov, "Analysing and Comparing the Final Grade in Mathematics by Linear Regression Using Excel and SPSS," International Journal of Mathematics Trends and Technology, vol. 52, no. 5, pp. 334–344, 2017, doi: 10.14445/22315373/ijmtt-v52p549

[5] I. Budiman and A. N. Akhlakulkarimah, "Aplikasi Data Mining Menggunakan Multiple Linear Regression Untuk Pengenalan Pola Curah Hujan," Kumpulan jurnaL Ilmu Komputer (KLIK), vol. 02,no. 01, pp. 34–44, 2015

[6] J. Supranto, Statistik, Teori dan Aplikasi. Surabaya: Penerbit Erlangga, 2016.

[7] N. Intan, P. Hati, and S. Nugroho, "Analisis Tingkat Penerimaan Calon Konsumen Terhadap Jenis Mobil Dengan Menggunakan Metode Regresi Linier," Jurnal Teknik Elektro Unnes, vol. 8, no. 2, pp. 50–55, 2016, doi: 10.15294/jte.v8i2.7761

[8] S. S. Rahardjo and R. Sanusi, "Linear Regression Analysis on the Determinants of Hypertension Prevention Behavior," Journal of Health Promotion and Behavior, vol. 4, no. 1, pp. 22–31, 2019, doi: 10.26911/thejhpb.2019.04.01.03

[7] S. Burns, Python Machine Learning Deep Learning Tensorflow. 2018.

[8] S. N. Waghmare and C. N. Sakhale, "Formulation of Experimental Data Based model using SPSS (Linear Regression) for Stirrup Making Operation by Human Powered Flywheel Motor," International Research Journal of Engineering and Technology (IRJET), vol. 02, no. 04, pp. 461–468, 2015.