LAPORAN
STATISTIKA

"Implementasi Multiple Linear Regression Dalam Prediksi Biaya Asuransi Kesehatan."

LA ODE MUHAMMAD YUDHY PRAYITNO

E1E122064

JURUSAN TEKNIK INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS HALU OLEO

KENDARI

2023

BAB 1
PENDAHULUAN

Asuransi kesehatan adalah salah satu jenis asuransi yang memberikan perlindungan finansial kepada pemegang polisnya jika mereka mengalami sakit atau kecelakaan. Asuransi kesehatan biasanya menanggung biaya perawatan medis, seperti biaya rawat inap, rawat jalan, dan obat-obatan.

Biaya asuransi kesehatan dapat bervariasi tergantung pada beberapa faktor, seperti usia, jenis kelamin, kondisi kesehatan, dan riwayat penyakit. Faktor-faktor ini dapat diukur secara kuantitatif, sehingga dapat digunakan untuk memprediksi biaya asuransi kesehatan.

Metode prediksi biaya asuransi kesehatan yang umum digunakan adalah regresi linear berganda. Regresi linear berganda adalah metode statistik yang digunakan untuk memprediksi nilai variabel dependen dengan menggunakan beberapa variabel independen.

Regresi linear ganda merupakan salah satu metode yang digunakan untuk melakukan analisis statistik yaitu melakukan prediksi atau memperkirakan pengaruh antara dua variabel atau lebih. Hubungan variabel yang dimaksud bersifat fungsional yang diwujudkan dalam bentuk model matematis, model matematika tersebut ditulis dalam bentuk y = b0 + B1 x1 + B2 x2 +… + Bn xn. Analisis regresi bertujuan untuk menjelaskan hubungan antar variabel, dalam regresi linear terdapat variabel y sebagai variabel respon atau variabel dependen dan variabel x sebagai variabel prediktor atau variabel independen

BAB 2
METODE

A. Dataset

Dataset yang digunakan dalam proses penelitian ini diolah dari https://www.kaggle.com/noordeen/insurancepremiumprediction. Dataset insurance.csv berisi 1338 data dan terdiri dari 7 kolom. Ke tujuh kolom tersebut terdiri dari 4 kolom numerik (age, bmi, children, dan charges) dan 3 kolom kategori (sex, smoker, dan region). Dalam proses analisis, data-data yang masuk dalam tipe kategori akan diubah menjadi numerik. Pengubahan yang dilakukan adalah jenis kelamin male diganti 1 dan female diganti 0.Demikian juga untuk data smoker dilakukan pengubahan data yes pada smoker diubah menjadi 1 dan data no pada smoker diubah 0. Data region, northeast menjadi 0, northwest=1, southeast=2 dan southwest=3.

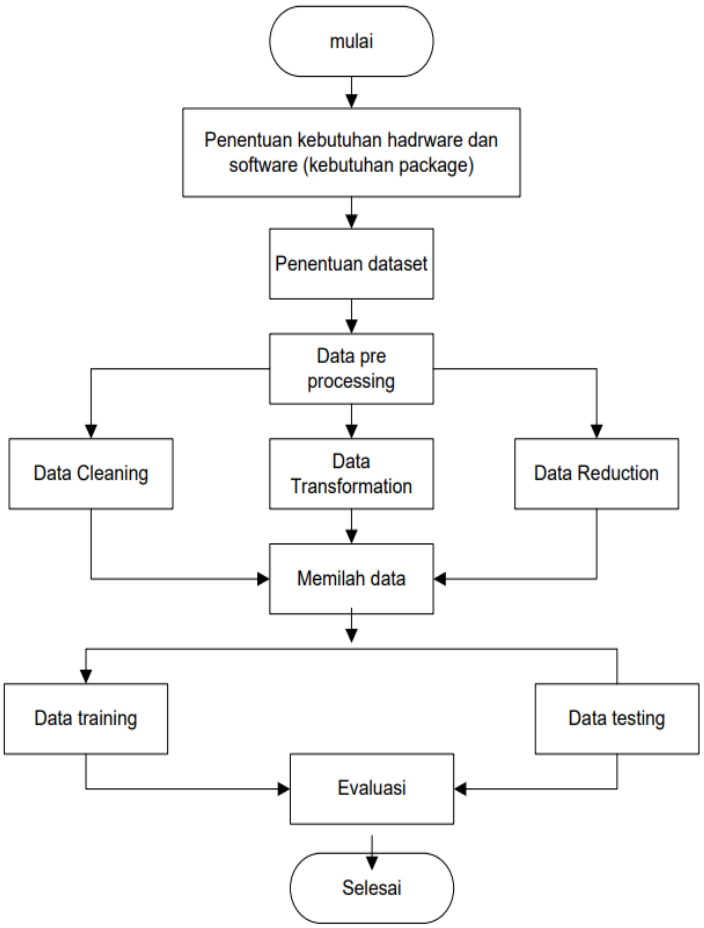
B. Pre-processing Data

Pre-processing dilakukan dalam melakukan pengolahan data agar data yang digunakan dapat diolah dengan baik dan terhindarkan dari data-data yang salah. Dalam proses pengolahan yang dilakukan, data awal yang digunakan masih berupa data mentah. Dalam proses yang dilakukan data-data yang diperlukan akan diformat dengan cara tertentu dan sesuai kebutuhan.

C. Analisis Data

Dalam proses penelitian yang dilakukan proses analisis data yang dilakukan adalah regresi linier berganda dan diolah dengan menggunakan python dan diimplementasikan menggunakan jupyter notebook. Analisis data yang dilakukan pertama adalah melakukan pembersihan dataset dari data yang tidak diinginkan, seperti data kosong, data yang diluar ambang batas dan tipe data yang tidak sesuai. Hasil proses pembersihan data akan dipilah menjadi 2, yaitu data yang digunakan sebagai data training dan data yang digunakan sebagai data test. Data training digunakan untuk melatih algoritma dan data testing digunakan untuk mengetahui performa algoritma yang sudah dilatih sebelumnya. Komposisi pembagian ini adalah 80% dari data yang ada untuk data training dan 20% untuk data test.

Awal penelitian dilakukan dengan menentukan kebutuhan perangkat keras dan perangkat lunak yang diperlukan. Kebutuhan perangkat lunak yang diperlukan diantaranya adalah datasheet yang digunakan dalam proses simulasi data dan library/packages yang digunakan dalam proses oleh data. Dari datasheet yang ditentukan, datasheet dilakukan proses pengecekan data dari data yang tidak diperlukan atau menghapus data kosong. Langkah lain adalah menentukan persentase untuk data training dan data test. Hasil data training dan data test dilakukan proses pengujian keakuratan data dan dilakukan pengujian dengan data diluar datasheet. Alur penelitian yang dilakukan ada pada gambar 1.



Gambar 1 Alur Penelitian

BAB 3
HASIL DAN PEMBAHASAN

A. Menyiapkan Perangkat Lunak

Proses analisis dengan menggunakan python diawali dengan menyiapkan library yang digunakan. Library digunakan adalah numpy, pandas, matplotlib, seabort dan sklearn. Library numpy digunakan untuk data analysis tools, library matplotlib dan seaborn untuk visualisasi data serta library scikit-Learn untuk machine learning. Perintah untuk memanggil library ada pada gambar 2.

```
1  import numpy as np
2  import pandas as pd
3  import matplotlib.pyplot as plt
4  import seaborn as sns
5  from sklearn.preprocessing import LabelEncoder
6  from sklearn.model_selection import train_test_split
7  from sklearn.linear_model import LinearRegression
8  from sklearn.metrics import mean_absolute_error,
9  mean_squared_error, mean_squared_log_error, r2_score
```

Gambar 2 Library yang digunakan

B. Memanggil Dataset dan Analisis Dataset

Dataset yang digunakan merupakan dataset yang diambil dari www.kaggle.com. Proses penggunaan dataset menggunakan perintah read yang ada pada library pandas. Proses pemanggilan dengan menggunakan perintah df.head(10). Perintah ini menampilkan datasheet awal sebanyak 10. Hasil perintah tersebut pada gambar 3.

```
In [2]:  1  df = pd.read_csv("data/insurance.csv")
         2  df.head(10)
```

Out[2]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16884.92400
1	18	male	33.770	1	no	southeast	1725.55230
2	28	male	33.000	3	no	southeast	4449.46200
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.880	0	no	northwest	3866.85520
5	31	female	25.740	0	no	southeast	3756.62160
6	46	female	33.440	1	no	southeast	8240.58960
7	37	female	27.740	3	no	northwest	7281.50560
8	37	male	29.830	2	no	northeast	6406.41070
9	60	female	25.840	0	no	northwest	28923.13692

Gambar 3 Contoh datasheet insurance.csv

Pada gambar 3, dataset insurance.csv, terdiri dari 7 kolom. Berdasar pada dataset tersebut, akan dilakukan analisis prediksi harga asuransi yang dipengaruhi age, sex, bmi, children, smoker dan region. Prediksi dalam bentuk regresi linear yang 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)

C. Exploratory Data Analysis

Salah satu cara dalam melakukan analisis data adalah dengan melakukan exploratory data analysis (EDA). Pada EDA, dilakukan eksplorasi data sehingga akan mendapatkan data yang sesuai dengan proses yang dilakukan [28]. EDA merupakan suatu kegiatan untuk mempelajari data yang dimiliki serta menentukan bagaimana proses pengolahannya terhadap data tersebut. Pada tahap ini dilakukan pemeriksaan pada data seperti data kosong, menghapus data yang sama serta mengubah data kategori menjadi numerik. Hasil pengecekan informasi dari dataset yang digunakan ditampilkan pada gambar 4 dan gambar 5.

```
In [4]:
1 print(df.shape)
2 print
3 print(df.describe())

(1338, 7)
                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 4 Data statistika

Gambar 4 merupakan analisis yang dapat digunakan untuk melihat data statistika, Dari data tersebut dapat digunakan untuk melihat apakah ada data yang tidak wajar. Dari data tersebut, nilai dari data yang dapat dilakukan analisis, misal age tertinggi adalah 64. Nilai age maksimal 64, tentunya masih wajar, demikian banyak anak maksimal 5 juga masih wajar dan disimpulkan data sudah benar.

```
In [3]:  1 df.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 5 Informasi datasheet

Pada gambar 5, menampilkan informasi dataset dan dapat disimpulkan banyak dataset sebanyak 1338 dan 7 kolom. Informasi pada gambar 4 menunjukkan semua data tidak ada data kosong dan tipe data dari masing-masing kolom. Kolom umur (age), jenis kelamin (sex), kategori berat badan (bmi) dan biaya (charges) mempuyai tipe numerik (integer dan float) dan kolom sex, smoker dan region bertipe object. Agar data-data yang ada pada kolom objek dapat dilakukan proses, tipe data object ini akan dilakukan proses encoding menjadi tipe numerik. Hal ini bertujuan untuk mempermudah dalam proses analisis data [29]. Proses encoding ada pada gambar 6.

```
In [4]:  1 num_cols = df.select_dtypes(include=np.number).columns
         2 num_cols
         3 non_num_cols = df.select_dtypes(exclude=np.number).columns
         4 non_num_cols
         5 label_encoder = LabelEncoder()
         6 for i in non_num_cols:
         7     df[i] = label_encoder.fit_transform(df[i])
         8 df.head(10)
```

Out[4]:

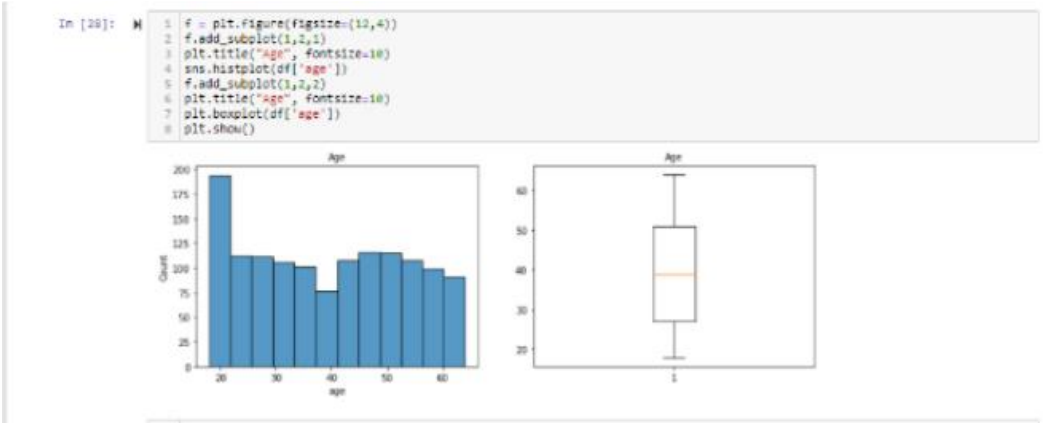
	age	sex	bmi	children	smoker	region	charges
0	19	0	27.900	0	1	3	16884.92400
1	18	1	33.770	1	0	2	1725.55230
2	28	1	33.000	3	0	2	4449.46200
3	33	1	22.705	0	0	1	21984.47061
4	32	1	28.880	0	0	1	3866.85520
5	31	0	25.740	0	0	2	3756.62160
6	46	0	33.440	1	0	2	8240.58960
7	37	0	27.740	3	0	1	7281.50560
8	37	1	29.830	2	0	0	6406.41070
9	60	0	25.840	0	0	1	28923.13692

Gambar 6 Hasil encoding data object menjadi data numeric

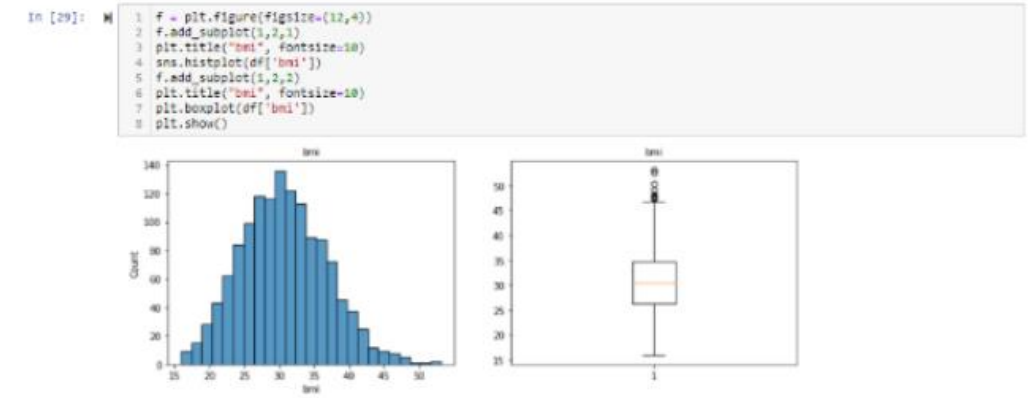
Hasil pada gambar 6, tipe pada kolom sex (jenis kelamin) , smoker (perokok) dan region (wilayah) sudah diganti dengan numerik. Dengan pengubahan ini akan mempermudah dalam proses regresi linear.

D. Visualisasi Data

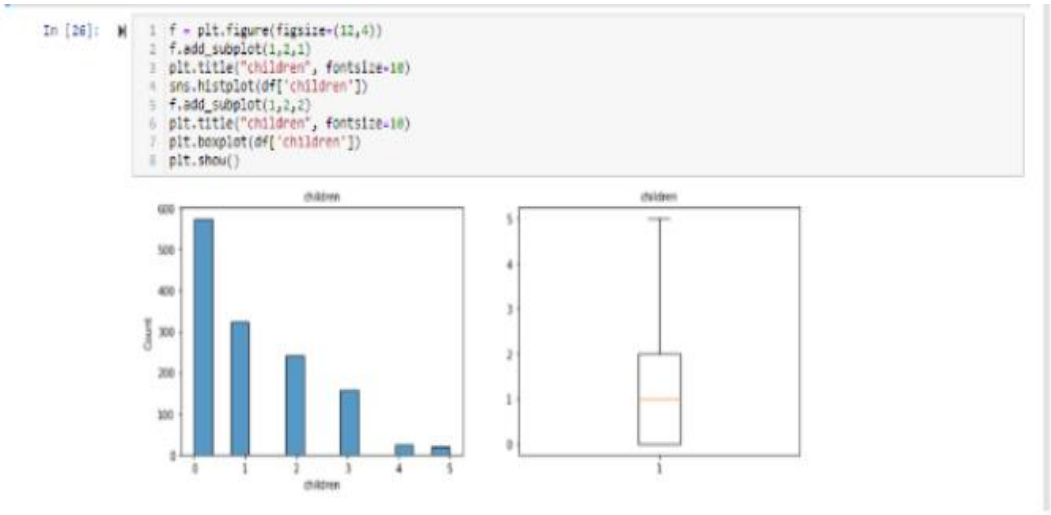
Visualisasi data merupakan proses yang mengubah data mentah menjadi informasi yang ditampilkan secara grafik Visualisasi data dapat ditampilkan dalam bentuk box plot, histogram dan bentuk lainnya. Box plot adalah jenis visualisasi data yang secara statistik merepresentasikan distribusi data melalui lima dimensi utama, yaitu nilai minimum, kuartil 1, kuartil 2 (median), kuartil 3, dan nilai maksimum. Box plot digunakan untuk memeriksa keberadaan outlier dalam dataset. Histogram adalah jenis visualisasi data untuk merepresentasikan distribusi frekuensi dari dataset numerik. Gambar 7-10 menampilkan visualisasi data dalam bentuk box plot dan menampilkan dalam bentuk histogram untuk variabel age, bmi, children dan charges.



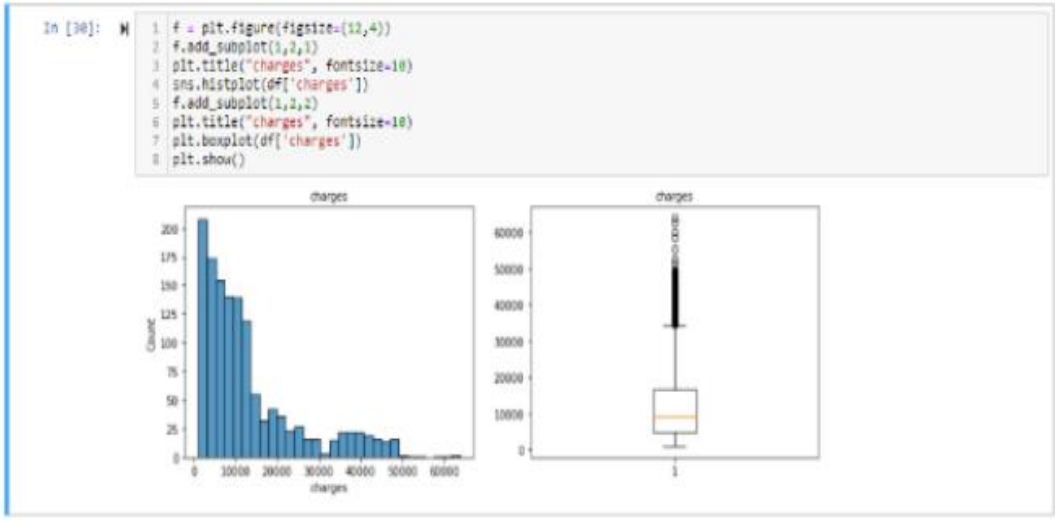
Gambar 7 Visualisasi data dalam bentuk histogram dan box plot variabel



Gambar 8 Visualisasi data dalam bentuk histogram dan box plot variabel



Gambar 9 Visualisasi data dalam bentuk histogram dan box plot variabel



Gambar 10 Visualisasi data dalam bentuk histogram dan box plot variabel

E. Analisis Regresi Linear

Perhitungan regresi linear dilakukan dengan penggunaan library sklearn. Dari dataset yang ada sebanyak 1338 dan dipilah menjadi 80% menjadi data training (1070 data) dan 20% menjadi data testing (268 data). Langkah lain yang dilakukan adalah menentukan kolom yang menjadi variabel dependen yaitu kolom charges dan kolom yang menjadi variabel independen yaitu umur, jenis kelamin, banyak anak , kategori berat ideal (bmi) , perokok (smoker) dan wilayah (region). Proses analisis regresi linear disajikan pada tabel 1.

Table 1 *Source Code* Python Dalam Proses Analisis Regresi Linear

1	x = df.drop(columns='charges') y = df['charges']	Menentukan kolom yang akan menjadi variabel dependen dan menjadi variabel independen
2	x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2, random_state=4)	Menentukan dan membagi data-data menjadi data training dan data test. Data training sebanyak 80% dari jumlah data dan data test sebanyak 20%
3	lin_reg = LinearRegression()	Mencari dan menampilkan nilai variabel dependen. Dari proses ini menghasilkan nilai interception = -12436.847333582358
4	lin_reg.fit(x_train, y_train)	
5	print(lin_reg.intercept_)	
6	feature_cols = ['age','sex','bmi','children','smoker','region'] X = df[feature_cols] y = df['charges'] list(zip(feature_cols, lin_reg.coef_))	Mencari dan menampilkan nilai variabel independent. Dari proses ini menghasilkan semua nilai independen (x) yaitu [('age', 270.3469135161016), ('sex', -188.32680363290413), ('bmi', 342.77182237789594), ('children', 474.0665341707971), ('smoker', 24320.998471652194), ('region', -385.59312017694276)]
7	print ("Coefficient of determination :",r2_score(y_test,ypredict)) print ("MSE: ",mean_squared_error(y_test,ypredict)) print("RMSE: ",np.sqrt(mean_squared_error(y_test,ypredict)))	Coefficient of determination : 0.7244150380582826 MSE: 34608265.193358265 RMSE: 5882.878988501996
8	df.corr()	Mengetahui nilai korelasi dari independent variable dan dependent variable.
9	ypredict=lin_reg.predict(x_test)	Menghitung nilai prediksi
10	df_best_predict = pd.DataFrame({'Actual': y_test, 'Predicted': ypredict}) df_best_predict.head(10)	Menampilkan nilai y awal dan nilai y hasil dari test

11	plt.figure(figsize=(10,7))
	plt.title("Actual vs. predicted",fontsize=25)
	plt.xlabel("Actual",fontsize=18)
	plt.ylabel("Predicted", fontsize=18)
	#plt.scatter(x=test_y,y=test_predict)
	sns.regplot(x=y_test, y=ypredict)
	plt.show()
		Menampilkan nilai y awal dan nilai y hasil dari test dalam bentuk grafik

12	lin_reg.predict([[30,0,27,0,0,0]])	Melakukan prediksi suatu data dengan age=30, sex=0,bmi=27,children=0,smoker=0,region=0. Hasil prediksi = 4,928.3992761
	lin_reg.predict([[50,1,28,0,1,0]])	Melakukan prediksi suatu data dengan age=50, sex=1,bmi=28,children=0,smoker=1,region=1. Hasil prediksi = 34,810.78103682

Langkah 3-6 pada tabel 1, melakukan proses perhitungan nilai inception dan nilai untuk semua variabel independent. Hasil perhitungan ditampilkan pada tabel 2.

Table 2 Nilai Interception Dan Variable Independen

Variabel	Nilai
interception	-1,2436.847333582358
age	270.3469135161016
sex	-188.32680363290413
bmi	342.77182237789594
children	474.0665341707971
smoker	2,4320.998471652194
region	-385.59312017694276

Dari data pada tabel 2, model multiple linear regression adalah y = -12436.85+ 270.35 X1 -188.37 X2+342.77 X3+474.07 X4+24320.10 X5 -385.60 X6.

F. Nilai Korelasi

Untuk mengetahui seberapa besar keterkaitan masingmasing variable bebas terhadap variable tidak bebas maka perlu dihitung korelasi parsial. Berdasar pada 2, proses pencarian nilai korelasi dilakukan pada baris 7 dan hasil korelasi disajikan pada tabel 3.

Table 3 Nilai Korelasi Dari Independent Variable Dan Dependent Variable

	age	sex	bmi	children	smoker	region	charges
age	1.000000	-0.019814	0.109344	0.041536	-0.025587	0.001626	0.298308
sex	-0.019814	1.000000	0.046397	0.017848	0.076596	0.004936	0.058044
bmi	0.109344	0.046397	1.000000	0.012755	0.003746	0.157574	0.198401
children	0.041536	0.017848	0.012755	1.000000	0.007331	0.016258	0.067389
smoker	-0.025587	0.076596	0.003746	0.007331	1.000000	-0.002358	0.787234
region	0.001626	0.004936	0.157574	0.016258	-0.002358	1.000000	-0.006547
charges	0.298308	0.058044	0.198401	0.067389	0.787234	-0.006547	1.000000

Hasil korelasi antar data pada tabel 3, menunjukkan ada keterkaitan yang erat antara smoker dengan charges (0,79), umur dengan charges (0.3) dan bmi dengan charges (0.2). Hal ini bisa diprediksi orang yang merokok akan membayar premi asuransi lebih tinggi dari orang yang tidak merokok dan ada korelasi yang agak kuat antara usia dengan biaya (charges) dan bmi dengan biaya (charges). Prediksi dari korelasi antara umur (age) dan bmi dengan biaya (charges), semakin tinggi usia atau semakin tinggi bmi semakin tinggi biaya yang harus dibayarkan.

G. Uji Koefisien Determinasi

Uji Koefisien determinasi digunakan untuk mengetahui seberapa besar pengaruh variabel independent terhadap variabel dependen sehingga dapat diketahui kesamaan dan kecocokan model regresi linier. Berdasar pada tabel 1,

proses perhitungan uji koefisien determinasi ada pada langkah ke 7 dan hasil perhitungan disajikan pada tabel 4.

Table 4 Hasil Uji Korelasi Parsial

Nilai Korelasi	
Coefficient of determination	0.7244150380582826
MSE	34,608,265.193358265
RMSE	5,882.878988501996

Dari hasil perhitungan yang disajikan 3pada tabel 4, dapat diketahui bahwa keterkaitan antara variabel dependen dengan variabel independen sangat kuat.

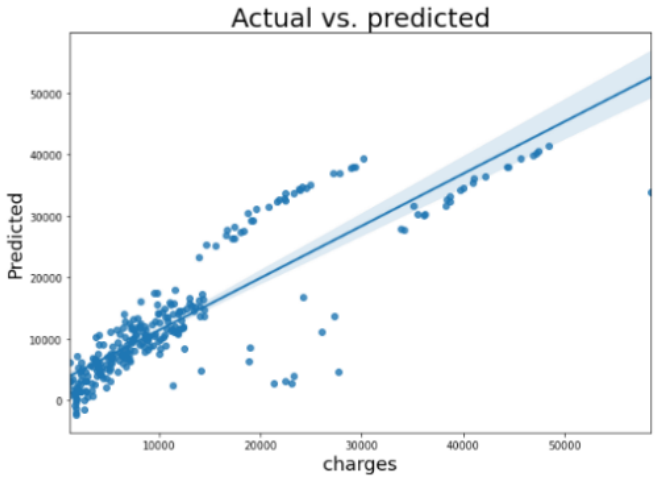
H. Pengujian Pada Data Prediksi

Pengujian dilakukan untuk melakukan perhitungan yang digunakan memprediksi hasil regresi linear dengan nilai y yang asli. Hasil pengujian dilakukan sebanyak 268 data.. Proses pemilahan data yang digunakan untuk data test dilakukan pada baris 2 pada instruksi perintah yang ada pada tabel 5. Hasil pengujian pada data prediksi disajikan pada tabel 5. Data yang ditampilkan sebanyak 10 dari 268 data.

Table 5 Perbandingan Perhitungan Y Actual Dan Y Predicted

Actual	Predicted
16,657.71745	27,655.279160
1,1837.16000	10,724.344214
8,125.78450	12,481.780039
6,373.55735	11,829.347769
7,448.40395	13,184.469361
1,719.43630	2,011.159023
11,090.71780	15,580.567550
22,331.56680	32,810.837858
27,218.43725	37,037.411958
1,875.34400	1,959.357519

Hasil pada tabel 5, antara data Y actual dan Y predicted ada perbedaan hasil dan hasil tampilan dalam bentuk grafik ditampilkan pada gambar 11.



Gambar 11 Grafik antara Y actual dengan Y predicted

Hasil prediksi tidak hanya dilakukan dengan menggunakan data test, tetapi juga dapat dilakukan dengan data di luar data test. Proses prediksi di luar data test ada pada baris 12 pada tabel 1. Hasil prediksi dengan ada age=30, sex=0, bmi=22, children=1, smoker=1 dan region=0, proses prediksi yang dilakukan adalah lin_reg.predict([[30,0,22,1,1,0]]) dan hasilnya 28,009.61. Hasil prediksi disajikan pada tabel 6.

Table 6 Prediksi Data Dengan Data Di Luar Data Test

age	sex	bmi	children	smoker	Region	Perintah di Python	Prediksi Charges
19	0	27.93	3	0	1	lin_reg.predict([[19,0,27.93,3,0,1]])	3,309.967505
19	0	30.02	0	1	1	lin_reg.predict([[19,0,30.02,0,1,1]])	26,925.15948
41	1	33.55	0	0	3	lin_reg.predict([[41,1,33.55,0,0,3]])	8,802.264597
40	1	29.355	1	0	1	lin_reg.predict([[40,1,29.355,1,0,1]])	8,339.242663
31	0	25.8	2	0	2	lin_reg.predict([[31,0,25.8,2,0,2]])	4,964.366831
37	1	24.32	2	0	3	lin_reg.predict([[37,1,24.32,2,0,3]])	5,505.226091
46	1	40.375	2	0	3	lin_reg.predict([[46,1,40.375,2,0,3]])	13,441.54992
22	1	32.11	0	0	3	lin_reg.predict([[22,1,32.11,0,0,3]])	3,172.081816
51	1	32.3	1	0	3	lin_reg.predict([[51,1,32.3,1,0,3]])	11,551.33549
18	0	27.28	3	1	1	lin_reg.predict([[18,0,27,28,3,1,1]])	27,137.81738
35	1	17.86	1	0	1	lin_reg.predict([[35,1,17.86,1,0,1]])	3,047.345998
59	0	34.8	2	0	1	lin_reg.predict([[59,0,34.8,2,0,2]])	15,619.02681

BAB 4
KESIMPULAN

Salah satu bagian dari statistika adalah proses mencari prediksi dengan menggunakan regresi liner ganda. Penelitian yang dilakukan adalah membuat simulasi penerapan statistika terutama regresi linear berganda dengan menggunakan python dan menggunakan editor jupyter notebook. Implementasi statistika regresi linear diterapkan pada prediksi biaya asuransi kesehatan yang dipengaruhi data age, sex, bmi, children, smoker dan region.

Berdasarkan hasil uji korelasi antar variabel independent, korelasi antara biaya (changer) dan perokok (smoker) 0,79. Hasil ini menunjukkan perokok mempunyai korelasi yang tinggi dengan biaya premi asuransi dan dapat diprediksi orang yang merokok akan membayar premi asuransi lebih tinggi dari orang yang tidak merokok demikian juga korelasi antara charges dengan age (0,3) dan BMI (0.2). Semakin tinggi umur (age) dan kategori berat badan (bmi) dapat diprediksi biaya premi asuransi (charges) juga bertambah.

Penggunaan bahasa Python dalam implementasi statistika, khususnya analisis regresi linear berganda dapat diimplementasikan dengan mudah dan tidak memerlukan koding yang rumit. Hal ini karena dukungan library di Python yang banyak dan pengguna tinggal menyesuaikan library-library yang digunakan.