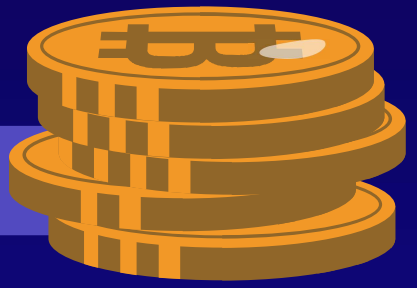
Analisis Pengaruh Status Merokok terhadap Biaya Asuransi Kesehatan

Final Project Talent HUB



Syifa nur’afni hidayat

Dataset : Insurance

Dataset ini berisi informasi tentang asuransi kesehatan, termasuk usia, jenis kelamin, indeks massa tubuh (BMI), jumlah anak, status merokok, wilayah tempat tinggal, dan biaya asuransi. Dengan menggunakan model machine learning, kita dapat menganalisis pengaruh status merokok terhadap biaya asuransi. Dengan mempelajari koefisien atau bobot dari setiap fitur, kita dapat memahami sejauh mana status merokok mempengaruhi biaya asuransi kesehatan.

Target dan Predictor

01

TARGET

Variabel yang ingin diprediksi adalah **charges** atau biaya asuransi keshatan individu

02

PREDICTOR

Variabel yang tersedia adalah BMI, sex, Age, Children, Smoker dan Region. Hanya saja variabel yang digunakan untuk predictor **hanya smoker** untuk memfokuskan lingkup permasalahan

Preparasi Data

```
import numpy as np
le = LabelEncoder()
data['sex'] = le.fit_transform(data['sex'])
print(data)

le=LabelEncoder()
data['smoker']=le.fit_transform(data['smoker'])
print(data)

le=LabelEncoder()
data['region']=le.fit_transform(data['region'])
print(data)
```

[128] ✓ 0.0s

age	sex	bmi	children	smoker	region	charges
19	0	27.900	0	yes	southwest	16884.92400
18	1	33.770	1	no	southeast	1725.55230
28	1	33.000	3	no	southeast	4449.46200
33	1	22.705	0	no	northwest	21984.47061
32	1	28.880	0	no	northwest	3866.85520

[127] ✓ 0.5s

age	sex	bmi	children	smoker	region	charges
19	female	27.900	0	yes	southwest	16884.92400
18	male	33.770	1	no	southeast	1725.55230
28	male	33.000	3	no	southeast	4449.46200
33	male	22.705	0	no	northwest	21984.47061
32	male	28.880	0	no	northwest	3866.85520
...	...	...	...	...	...	...
50	male	30.970	3	no	northwest	10600.54830
18	female	31.920	0	no	northeast	2205.98080
18	female	36.850	0	no	southeast	1629.83350
21	female	25.800	0	no	southwest	2007.94500
61	female	29.070	0	yes	northwest	29141.36030

1338 rows × 6 columns

```
data.isnull().sum()
```

[157] ✓ 0.4s

```
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

Sebelumnya saya sudah melakukan NA Removal, lalu merubah data string menjadi float dengan label encoder

MODEL

```
#build model
from sklearn.neighbors import KNeighborsRegressor
model=KNeighborsRegressor()
```
✓ 0.0s

```
# Model build 2
from sklearn.linear_model import LinearRegression
model = LinearRegression()
```
✓ 0.0s

```
# build model 3
from sklearn.svm import SVR
model = SVR()
```
✓ 0.0s

Model yang digunakan ada 3 yaitu, KNeighborsRegressor, LinearRegressor dan SVR

EVALUASI HASIL

```
MAPE: 0.546455622894961
RMSE: 8277.886999862576
R-squared: 0.5518285948789599
```

KNeigborsRegressor

```
MAPE: 0.9056454423532044
RMSE: 12919.522734107175
R-squared: -0.09168689607001346
```

SVR

```
MAPE: 0.9056454423532044
RMSE: 12919.522734107175
R-squared: -0.09168689607001346
```

LinearRegressor

EVALUASI HASIL

Dari hasil yang telah tertera ada 4 model evaluasi yaitu RMSE, R2 dan MAPE. Tapi disini kita hanya fokus membandingkan hasil RSME. Dimana Dari hasil evaluasi, dapat dilihat bahwa model KNeighborsRegressor memiliki nilai RMSE yang lebih rendah (8277.89) dibandingkan dengan model SRV (12919.52) dan LinearRegression (12919.52). Semakin rendah nilai RMSE, semakin kecil kesalahan prediksi model. Oleh karena itu, model KNeighborsRegressor lebih baik dalam melakukan prediksi biaya asuransi kesehatan.

THANK YOU KAKK