Nama : Ghina Khoerunnisa
Program : Data Scientist Intern

REGRESSION MODEL

- **Intro To Machine Learning**

  Artificial Intelligence (AI) merupakan sistem yang menyerupai kamampuan manusia. AI menjadi penting dan booming karena dengan AI pekerjaan manusia menjadi lebih mudah dan hal-hal yang berbahaya dapat dilakukan tanpa mengikut sertakan manusia secara langsung. Machine Learning merupakan penerapan dari AI dimana dalam proses learningnya membutuhkan data. ML dikelompokkan menjadi beberapa kelompok yaitu supervised, unsupervised, dan reinforcement. Supervised ini learning dimana terdapat label di datasetnya sedangkan unsupervised tidak ada label di datasetnya. Contoh dari supervised adalah regression dan classification. Sedangkan contoh unsupervised adalah clustering. ML dapat diaplikasikan pada berbagai bidang contohnya yaitu pada kesehatan (cancer detection), financial (fraud detection), marketing (customer segmentation), dll.

- **Regression Model**

  Jika dibandingkan dengan Classification yang menghasilkan prediksi class, Regression menghasilkan prediksi numerik. Pada kasus regresi kita perlu menemukan fungsi yang memetakan beberapa fitur atau variabel ke fitur lain dengan cukup baik. Fitur dependent ini disebut juga dengan output atau responses dan fitur independent disebut juga dengan input atau predictors. Regresi digunakan ketika kita mengetahui bagaimana beberapa fenomena memengaruhi yang lain atau bagaimana terkaitnya variabel yang ada. Misalnya bagaimana waktu bekerja mempengaruhi gaji seseorang.

  - Linear Regression : Linear Regression merupakan salah satu teknik untuk melakukan regression. Persamaan dari linear regresi adalah:

    $$w_0 + w_1.x_1 + w_2.x_2 + \ldots + wn.xn = y$$

    Dimana w0 merupakan random error (interceptnya) dan w1,w2, ...wn adalah koefisien regresinya (slope). Estimated setiap pengamatan harus sedekat mungkin dengan aktualnya. Sehingga Regresi adalah menentukan **best predicted weights**, yaitu bobot yang sesuai dengan residuals (y - y') terkecil. Untuk mendapatkan bobot nilai terbaik kita harus meminimalkan SSR (Sum of Squared Residual). Variabel lain yang melekat pada output adalah Coefficient Determination (R^2). Ini menunjukkan berapa banyak variasi dalam *y* yang dapat dijelaskan oleh ketergantungan pada **x** menggunakan model regresi tertentu. Nilai $R^2 = 1$ sesuai dengan SSR = 0, yaitu perfect fit karena nilai prediksi dan respons aktual saling cocok satu sama lain.

    - Simple Linear Regression: Simple atau single-variate linear regression adalah kasus regresi linier yang paling sederhana dengan variabel independen tunggal, **x** = *x*.
    - Multiple Linear Regression: Multiple atau multivariate linear regression adalah kasus regresi linier dengan dua atau lebih independent variables.
    - Polynomial Regression: Contoh persamaan dari Polynomial Regression adalah

      $$w_0 + w_1.x_1 + w_2.x_2^2 = y$$

Polynomial juga termasuk ke linear regression karena linearnya dilihat dari parameter wi nya.

- Underfitting & Overfitting: Underfitting merupakan kejadian dimana errornya sangat besar antara actual dan estimatednya. Sehingga seringkali memiliki R^2 yang kecil dari data train dan data test. Sedangkan Overfitting kebalikannya dimana model terlalu tepat memetakan estimated dengan actualnya pada data train tetapi ketika dicoba untuk data test menghasilkan error yang besar. Biasanya memiliki nilai R^2 = 1 yang artinya SSR = 0 (tidak ada error).
- Implementasi dengan Sklearn

```
x = df['horsepower'].values.reshape(-1,1)
y = df['price'].values.reshape(-1,1)
```

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2)
```

```
lin_reg = LinearRegression()
lin_reg.fit(x_train, y_train)
```

```
LinearRegression()
```

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)
```

```
LinearRegression()
```

```
print(lin_reg.coef_)
print(lin_reg.intercept_)
```

```
[[164.64913667]]
[-3905.96213306]
```

```
lin_reg.score(x_test, y_test)
```

```
0.5569273366575066
```

```
y_prediction = lin_reg.predict(x_test)
```

- Implementasi dengan Statsmodels

```
df = pd.DataFrame(data.data, columns=data.feature_names)
target = pd.DataFrame(data.target, columns=["MEDV"])
```

```
X = df["RM"]
y = target["MEDV"]

model = sm.OLS(y, X).fit()
predictions = model.predict(X)

model.summary()
```

OLS Regression Results

Dep. Variable:	MEDV	**R-squared (uncentered):**	0.901
Model:	OLS	**Adj. R-squared (uncentered):**	0.901
Method:	Least Squares	**F-statistic:**	4615.
Date:	Thu, 30 Sep 2021	**Prob (F-statistic):**	3.74e-256
Time:	06:49:31	**Log-Likelihood:**	-1747.1
No. Observations:	506	**AIC:**	3496.
Df Residuals:	505	**BIC:**	3500.
Df Model:	1		
Covariance Type:	nonrobust		