**MEDICAL INSURANCE COST PREDICTION**

**PROJECT OVERVIEW**

Biaya asuransi kesehatan telah meningkat secara dramatis selama dekade terakhir sebagai tanggapan atas meningkatnya biaya layanan perawatan kesehatan yang ditentukan oleh banyak faktor. Mari kita lihat biaya perawatan kesehatan untuk sampel populasi berdasarkan usia, jenis kelamin, BMI, jumlah anak, kebiasaan merokok, dan wilayah.

Tujuan dari proyek ini adalah untuk menentukan faktor penyebab dan memprediksi biaya asuransi kesehatan dengan melakukan analisis data eksplorasi dan pemodelan prediktif pada dataset Asuransi Kesehatan. Proyek ini menggunakan pustaka Numpy, Pandas, Sci-kit learn, dan Visualisasi Data.

**Overview:**

* Mencari wawasan dari kumpulan data dengan Exploratory Data Analysis
* Melakukan Pemrosesan Data, Rekayasa Data, dan Transformasi Fitur untuk menyiapkan data sebelum pemodelan
* Membangun model untuk memprediksi Biaya Asuransi berdasarkan fitur-fiturnya
* Mengevaluasi model menggunakan berbagai Metrik Kinerja seperti RMSE, R2, Akurasi Pengujian, Akurasi Pelatihan, dan MAE

**DATA DESCRIPTION**

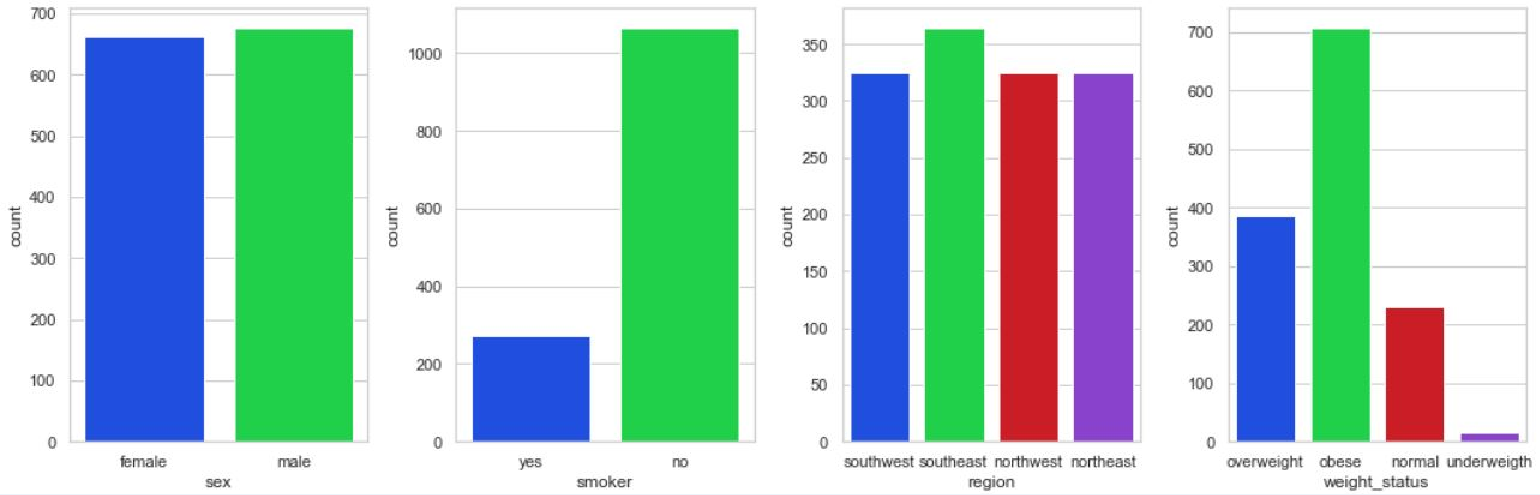
* Age: usia penerima manfaat utama
* jenis kelamin: jenis kelamin kontraktor asuransi, perempuan, laki-laki
* BMI : Body mass index, memberikan pengertian tentang tubuh, berat badan yang relatif tinggi atau rendah relatif terhadap tinggi badan, indeks objektif berat badan (kg/m ^ 2) dengan menggunakan rasio tinggi terhadap berat badan, idealnya 18,5 sampai 24,9
* Children: Jumlah anak yang ditanggung oleh asuransi kesehatan / Jumlah tanggungan
* Smoker: Merokok
* Region: daerah pemukiman penerima manfaat di AS, timur laut, tenggara, barat daya, barat laut
* Charges: Biaya medis individu yang ditagih oleh asuransi kesehatan

Data source : <https://www.kaggle.com/mirichoi0218/insurance>

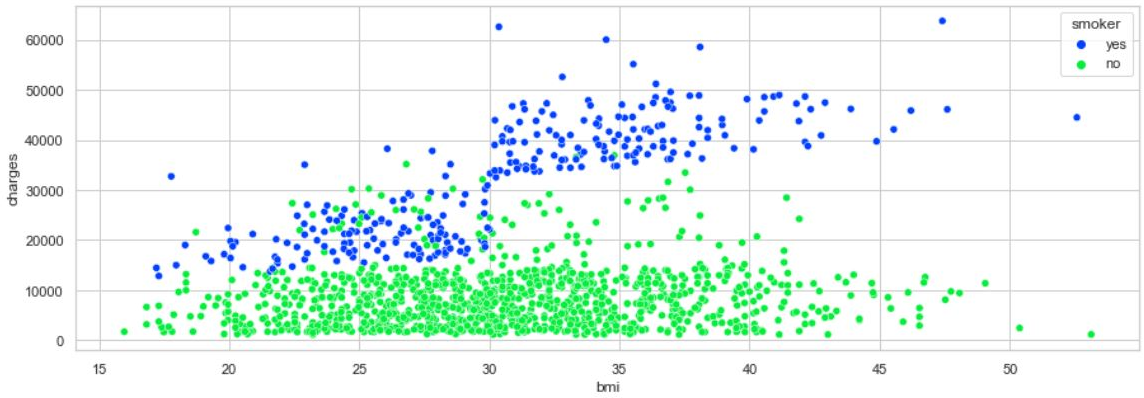
**LINK CODE**

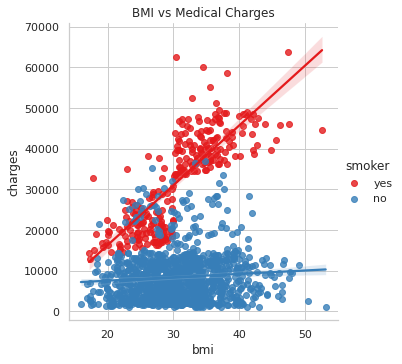
**EXPLORATORY DATA ANALYSIS (EDA)**

* Ciri jenis kelamin, wilayah memiliki jumlah yang hampir seimbang, sedangkan sebagian besar penduduknya bukan perokok dan obesitas

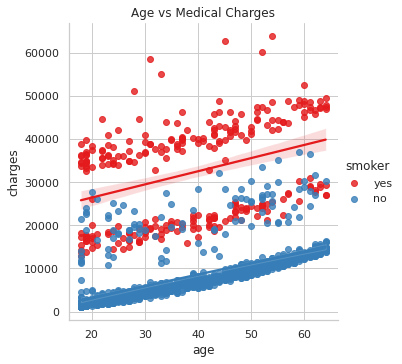


* Seseorang yang merokok dan memiliki bmi di atas 30 cenderung memiliki biaya pengobatan yang lebih tinggi

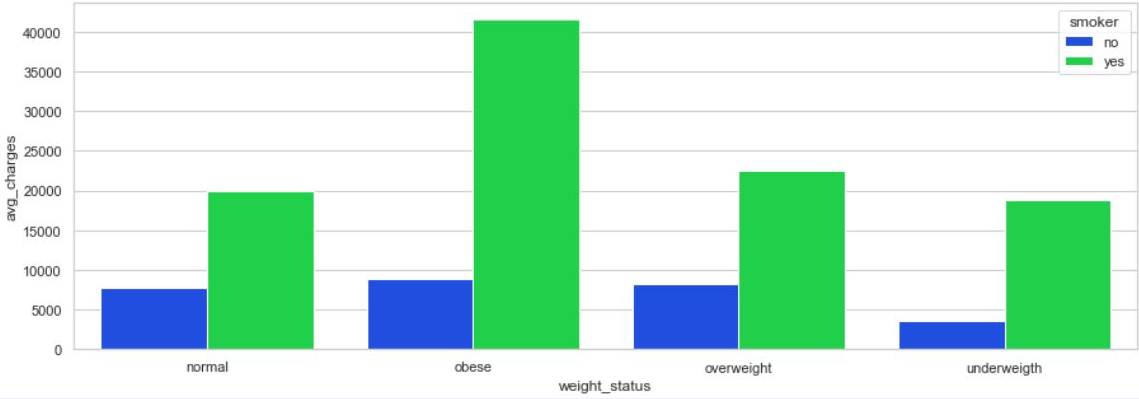




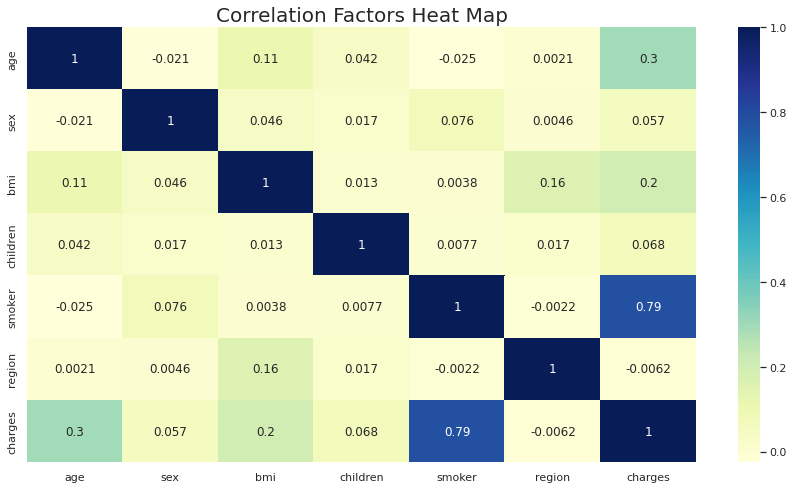
* Orang tua yang merokok memiliki biaya lebih mahal



* Orang yang merokok dan obesitas memiliki biaya rata-rata tertinggi dibandingkan orang lain



* Perilaku merokok yang diikuti usia dan bmi sangat berkorelasi dengan biaya pengobatan.



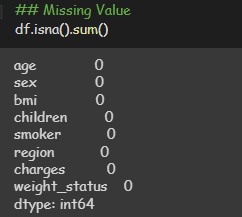
**INSIGHTS**

INSIGHT yang diperoleh dengan melakukan Exploratory Data Analysis (EDA) adalah:

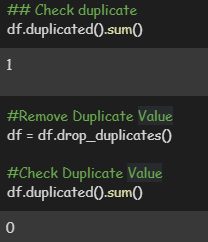
* Kebanyakan orang adalah non perokok & obesitas.
* Fitur jenis kelamin, wilayah memiliki jumlah yang hampir seimbang.
* Orang yang merokok & memiliki BMI lebih tinggi, memiliki biaya medis lebih tinggi.
* Orang tua yang merokok memiliki biaya lebih mahal.
* Orang gemuk yang merokok memiliki biaya lebih tinggi.

**DATA PROCESSING**

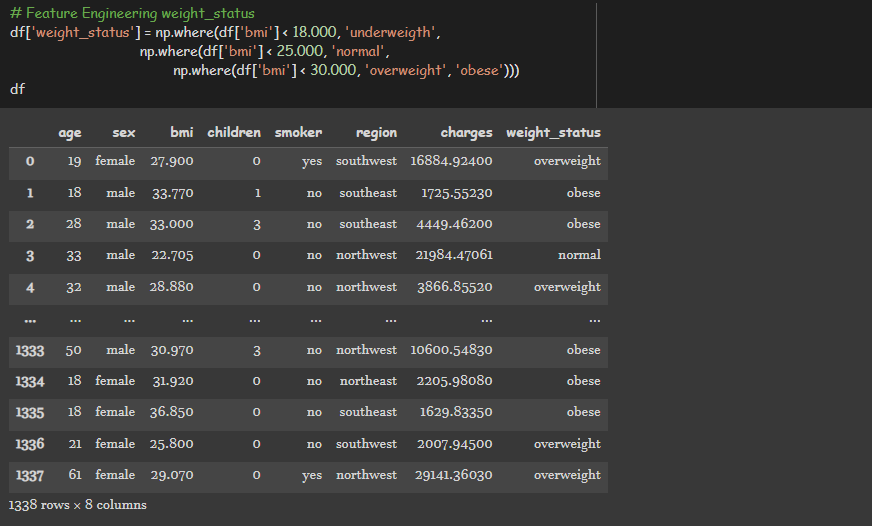
1. Melakukan Chechk missing value ( Tidak ada).



1. Melakukan Check duplicate value (terdapat 1 duplicate, akan dihapus )



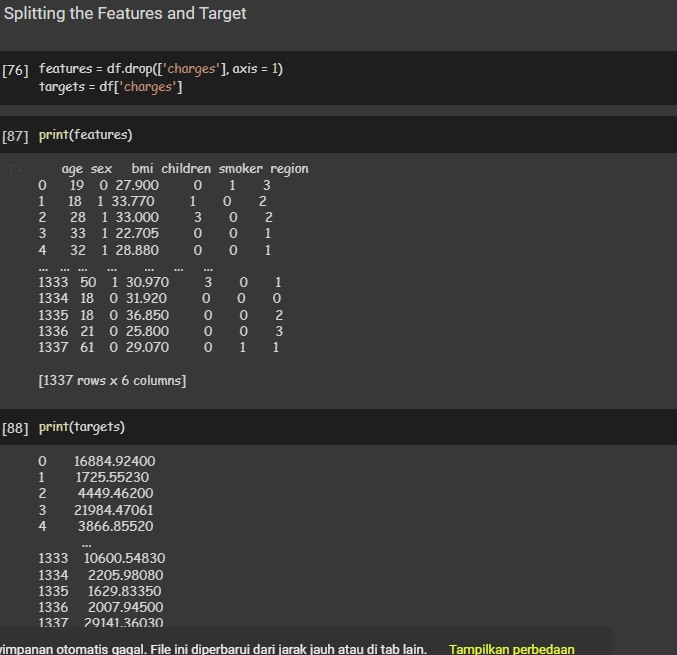
1. Feature engineering - make a new column weight\_status based on BMI score

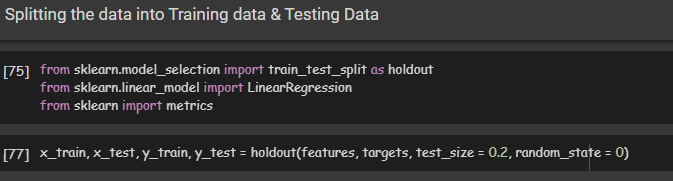


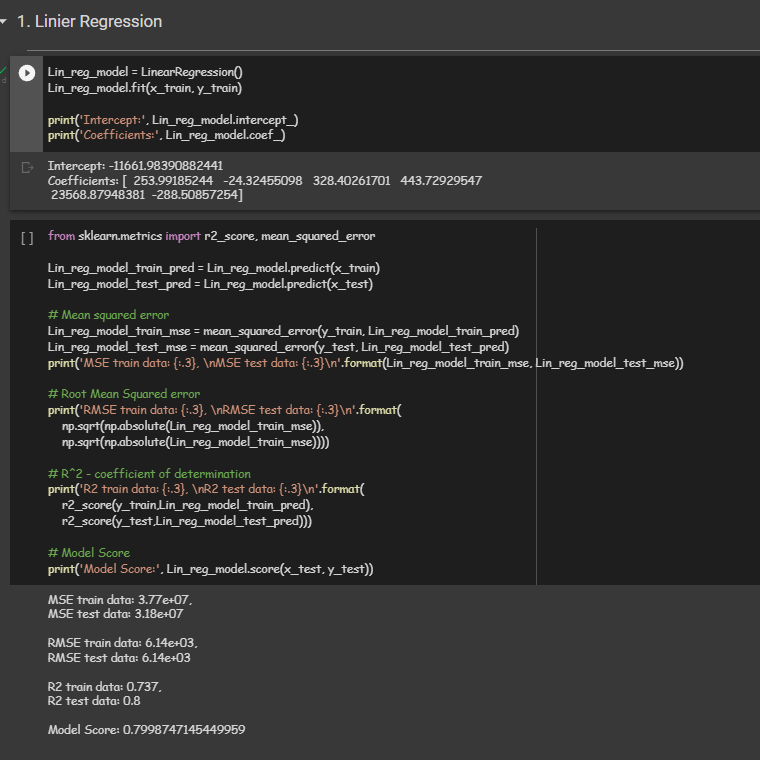
1. Feature transformation:  
   A) Encoding sex, region, & weight\_status attributes  
   B) Ordinal encoding smoker attribute



1. Modeling:  
   A) Separating target & features

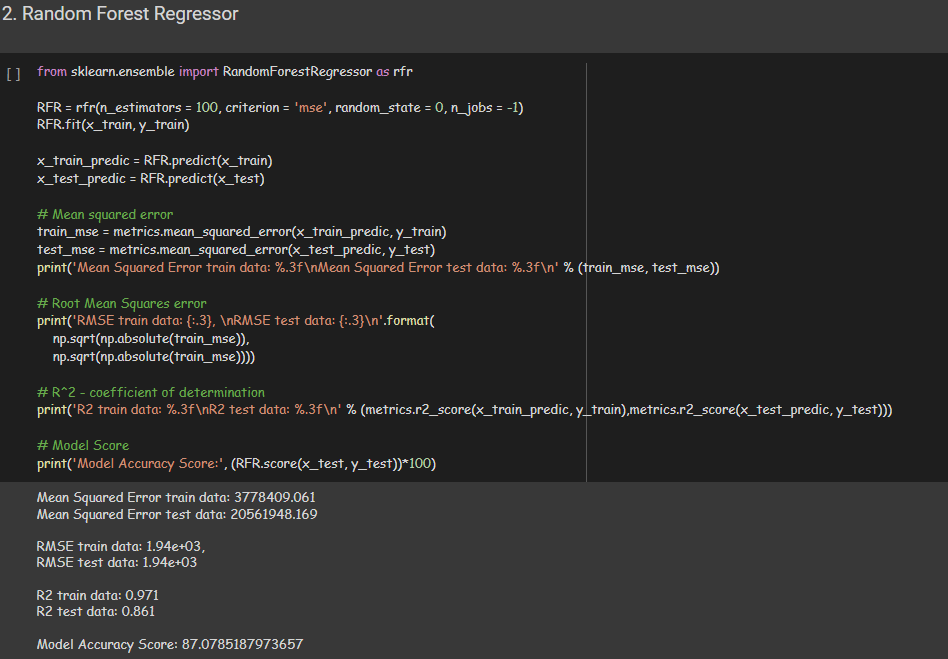
  
B) Splitting train & test data

  
C) Modeling using Linear Regression,

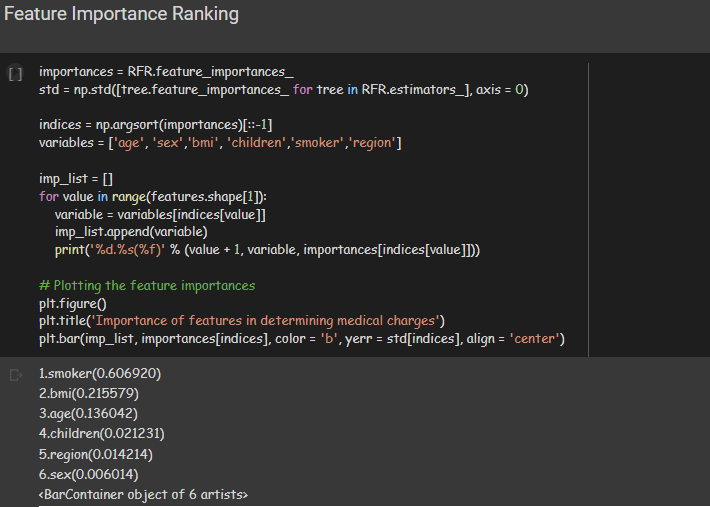
  
D) Find the best algorithm

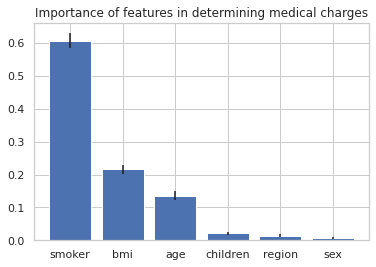
Mencoba untuk membuat prediksi dengan random forest dan polynomial regression

* Random Forest Regressor

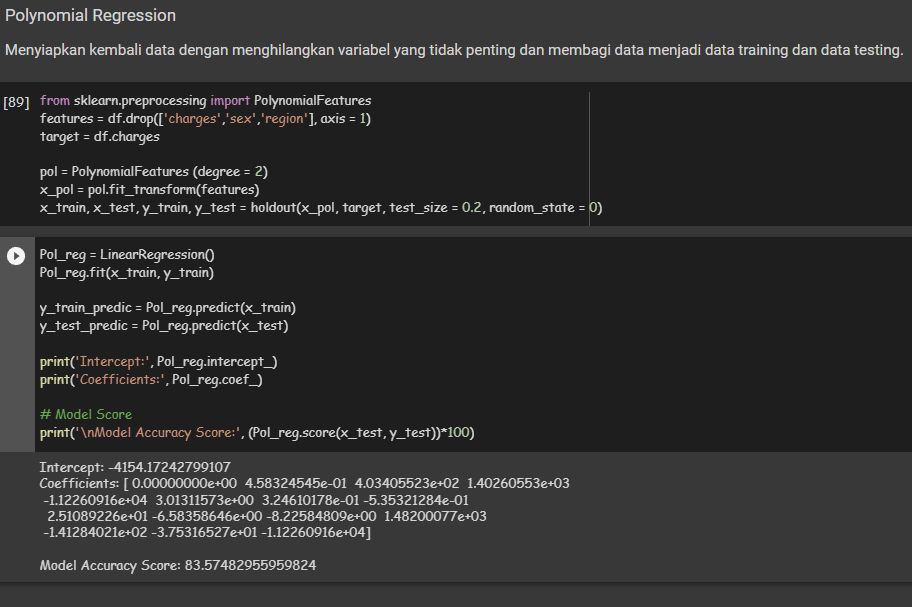


Dari random forest regressor ini kita bisa membuat algorithm untuk feature importance ranking

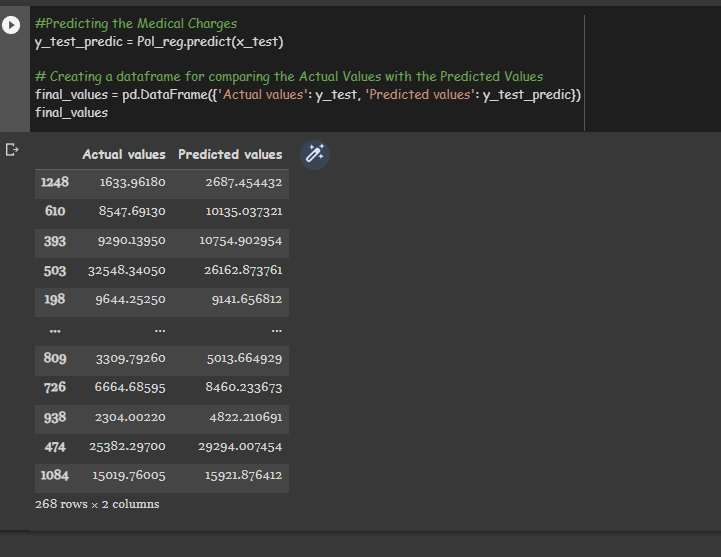




* Polynomial Regression



E) Memprediksi biaya asuransi untuk seorang individu.



**MODEL EVALUATION**

Based on the perfomed machine learning algorithms, tampaknya Regresi Polinomial setelah menghilangkan dua variabel independen yang tidak penting menghasilkan metode yang paling akurat

* Polynomial Regression: 88%
* Random Forest Regressor: 87%
* Linear Regression: 80%

**CONCLUSION**

Berdasarkan pemodelan prediktif, Algoritma Plinomial Regression memiliki skor terbaik dibandingkan yang lain, dengan Skor keakuratan 88%

Oleh karena itu, Algoritma Polinomial regression adalah model yang paling cocok berdasarkan akurasi train dan test.