Laporan Mata Kuliah Pembelajaran Mesin

LINEAR DAN POLYNOMIAL REGRESSION

disusun untuk memenuhi
tugas mata kuliah Pemrosesan Mesin B

oleh :

Mila Lestari	(2208107010002)
Zahra Zafira	(2208107010040)
Pryta Rosela	(2208107010046)
Cut Sula Fhatia Rahma	(2208107010048)
Widya Nurul Sukma	(2208107010054)

JURUSAN INFORMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS SYIAH KUALA

2025

1. Pendahuluan

Biaya asuransi kesehatan merupakan aspek penting dalam perencanaan keuangan individu maupun keluarga. Memprediksi biaya asuransi kesehatan dengan akurat dapat membantu penyedia asuransi untuk menentukan premi yang tepat dan membantu nasabah memperkirakan pengeluaran mereka.

Dalam studi ini, kami menganalisis faktor-faktor yang mempengaruhi biaya asuransi kesehatan dan mengembangkan model prediktif untuk memperkirakan biaya tersebut berdasarkan berbagai faktor personal dan demografis seperti usia, jenis kelamin, BMI, jumlah anak, status merokok, dan wilayah tempat tinggal.

Tujuan utama dari proyek ini adalah untuk mengidentifikasi faktor-faktor yang paling signifikan mempengaruhi biaya asuransi dan membandingkan performa beberapa teknik regresi dalam memprediksi biaya tersebut.

2. Dataset

Dataset yang digunakan dalam studi ini adalah "Medical Cost Prediction" yang berisi informasi tentang pasien dan biaya medis yang dikeluarkan. Dataset ini digunakan untuk memprediksi biaya pengobatan berdasarkan faktor-faktor individu.

Deskripsi dataset:

- **Target (Variabel Dependen):** Biaya pengobatan (charges)
- **Features (Variabel Independen):**
    - age: Usia pasien (dalam tahun)
    - sex: Jenis kelamin pasien (female/male)
    - bmi: Indeks Massa Tubuh (BMI)
    - children: Jumlah anak/tanggungan
    - smoker: Status merokok (yes/no)
    - region: Wilayah tempat tinggal (northeast, northwest, southeast, southwest)
    - charges: Total biaya medis yang dikenakan kepada individu

Setelah diimpor, dataset memiliki ukuran **2772 baris x 7 kolom**. Tidak ada nilai yang hilang (missing values) dalam dataset ini.

3. Eksplorasi Data dan Analisis

3.1 Statistik Deskriptif

Berikut adalah statistik deskriptif dari variabel numerik dalam dataset:

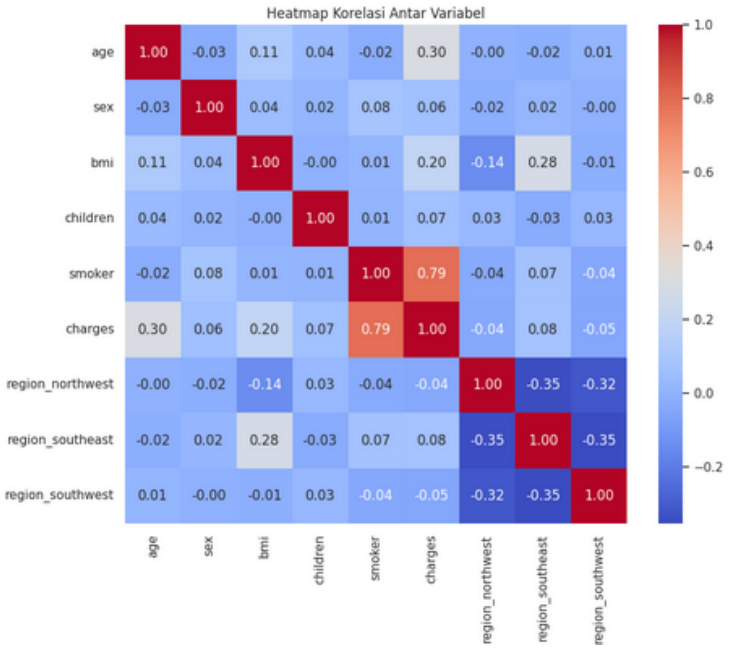
Statistik	age	bmi	children	charges
count	2772.00	2772.00	2772.00	2772.00
mean	39.10	30.70	1.10	13261.36
std	14.08	6.12	1.21	12151.76
min	18.00	15.96	0.00	1121.87
25%	26.00	26.22	0.00	4687.79
50%	39.00	30.44	1.00	9333.01
75%	51.00	34.77	2.00	16577.77
max	64.00	53.13	5.00	63770.42

Berdasarkan statistik deskriptif di atas, kita dapat menyimpulkan:

- **Usia (age):** Rata-rata usia peserta adalah 39 tahun, dengan usia termuda 18 tahun dan tertua 64 tahun.
- **BMI (bmi):** Rata-rata BMI adalah 30.70, termasuk kategori overweight, dengan nilai minimum 15.96 dan maksimum 53.13.
- **Jumlah Anak (children):** Rata-rata jumlah anak adalah 1.1, dengan sebagian besar responden memiliki 0-2 anak.
- **Biaya Pengobatan (charges):** Rata-rata biaya pengobatan adalah sekitar 13.261 USD, namun terdapat variasi yang cukup besar, dari sekitar 1.121 USD hingga 63.770 USD.

3.2 Korelasi Antar Variabel

Analisis korelasi menunjukkan hubungan antara variabel dalam dataset:

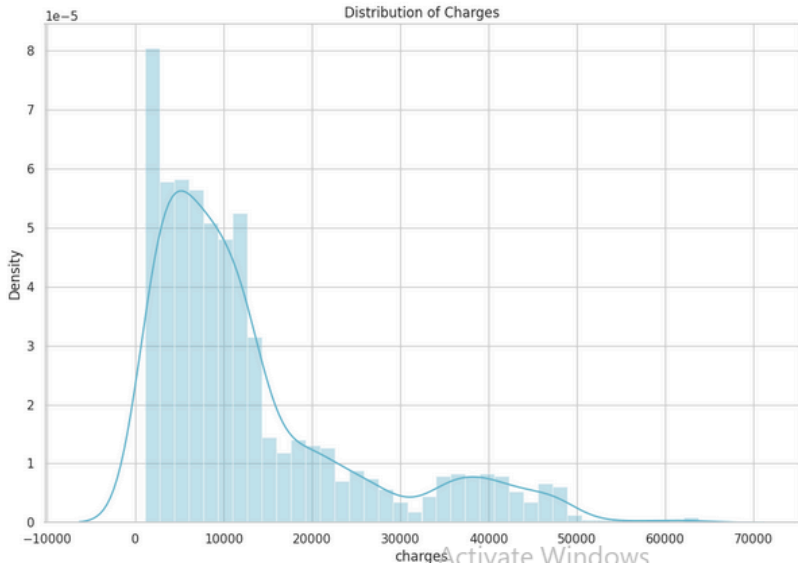


Gambar 1.1 *Heatmap Korelasi Antar Variabel*

- **Status merokok (smoker)** memiliki korelasi positif tertinggi dengan biaya pengobatan (charges). korelasi: 0.79
- **Usia (age)** juga menunjukkan korelasi positif yang cukup signifikan dengan biaya pengobatan. korelasi: 0.30
- **BMI (bmi)** memiliki korelasi positif sedang dengan biaya pengobatan. korelasi: 0.20
- **Jumlah anak (children)** memiliki korelasi positif rendah dengan biaya pengobatan. korelasi: 0.07
- **Jenis kelamin (sex)** dan **wilayah (region)** menunjukkan korelasi positif rendah dengan biaya pengobatan. korelasi: 0.06

3.3 Distribusi Variabel

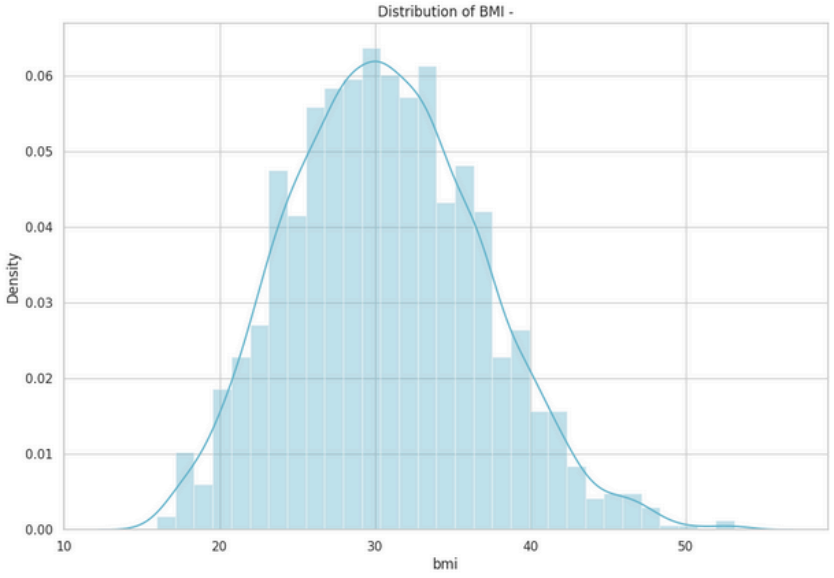
- **Distribusi Biaya Pengobatan (Charges)**



Gambar 1.2 *Visualisasi Distribuasi Biaya Pengobatan*

Distribusi biaya pengobatan menunjukkan pola yang tidak normal (skewed), dengan banyak data terkonsentrasi pada biaya rendah dan beberapa outlier dengan biaya sangat tinggi. Hal ini menyarankan penggunaan transformasi logaritmik untuk normalisasi data.

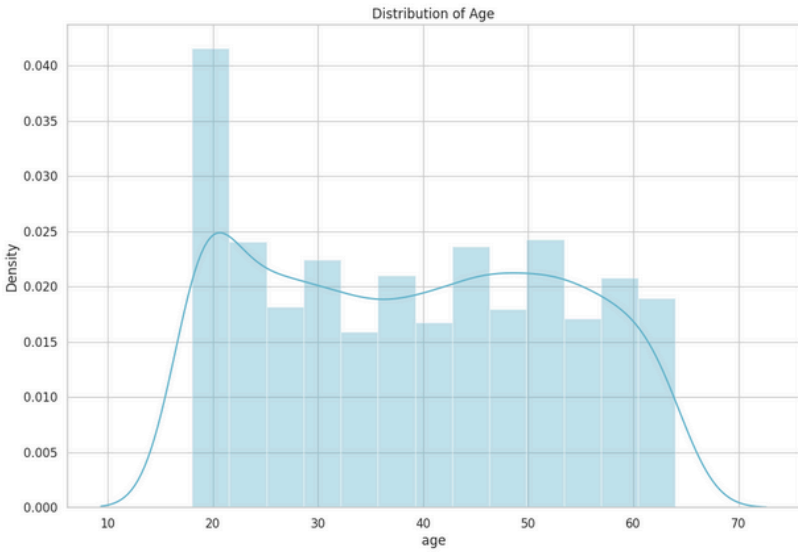
- **Distribusi BMI**



Gambar 1.3 *Visualisasi Ditsribusi BMI*

Distribusi BMI menunjukkan pola yang mendekati normal dengan sebagian besar nilai berada di sekitar 26-34, yang masuk dalam kategori overweight hingga obesitas kelas 1.

- **Distribusi Usia**

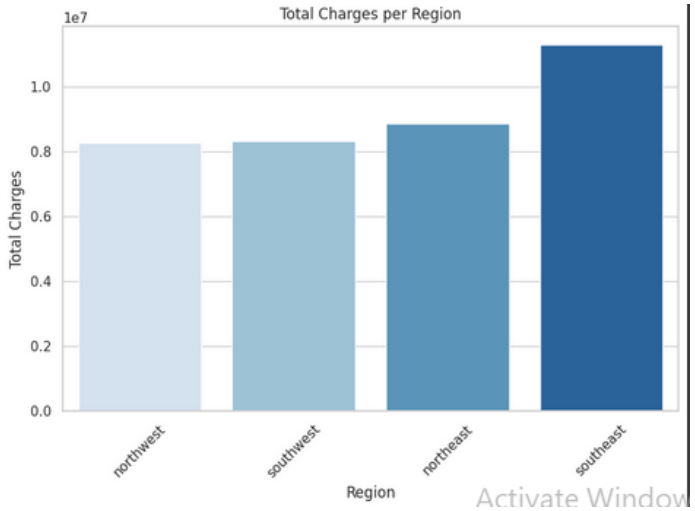


Gambar 1.4 *Visualisasi Distribusi Usia*

Distribusi usia menunjukkan sebaran yang cukup merata di seluruh rentang usia (18-64 tahun), dengan sedikit penurunan pada kelompok usia yang lebih tua.

3.4 Analisis Kategorikal

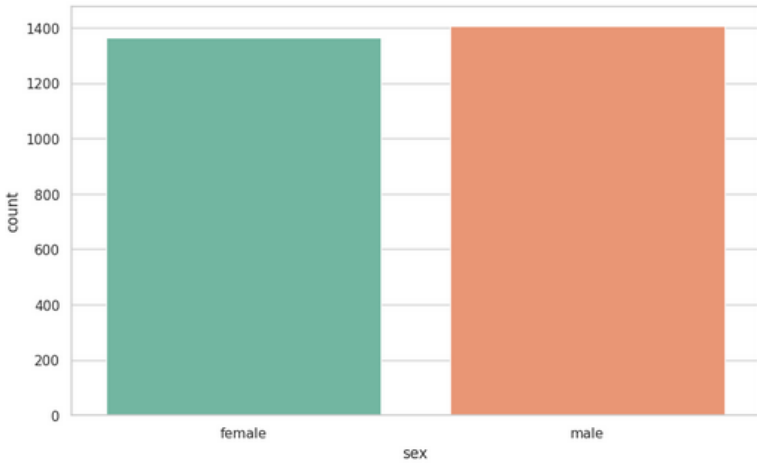
- **Distribusi Wilayah (Region)**



Gambar 1.5 *Visualisasi Distribusi Wilayah*

Analisis menunjukkan bahwa jumlah data dari setiap wilayah cukup seimbang, dengan wilayah **southeast** memiliki jumlah data sedikit lebih tinggi.

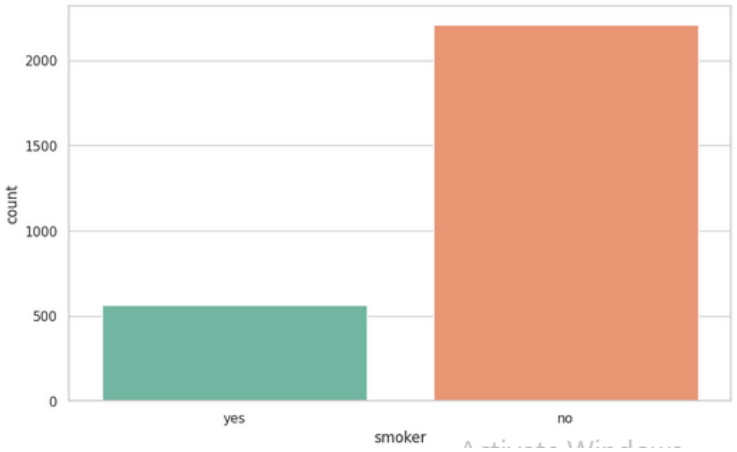
- **Distribusi Jenis Kelamin (Sex)**



Gambar 1.6 *Visualisasi Distribusi Jenis Kelamin*

Jumlah data untuk kategori **male** dan **female** relatif seimbang, dengan **male** sedikit lebih banyak dibandingkan **female**.

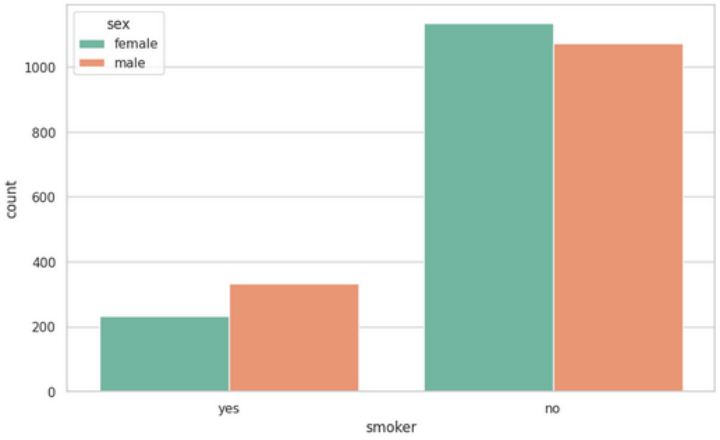
- **Distribusi Status Merokok (Smoker)**



Gambar 1.7 *Visualisasi Distribusi Status Merokok*

Mayoritas individu dalam dataset adalah non-perokok (**no**), dengan jumlah yang jauh lebih banyak dibandingkan perokok (**yes**).

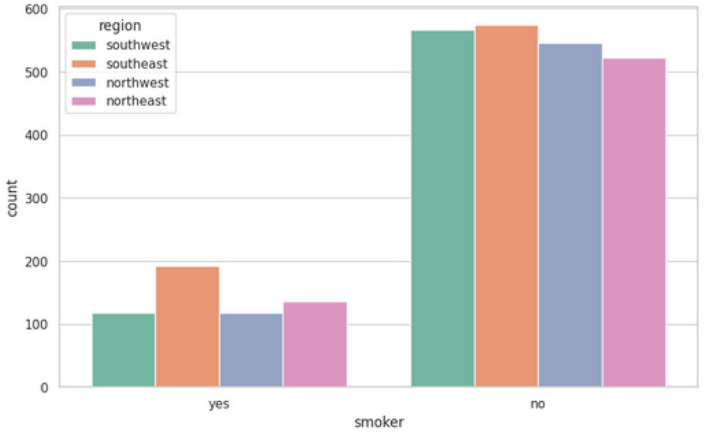
- **Distribusi Status Merokok Berdasarkan Jenis Kelamin**



Gambar 1.8 *Distribusi Status Merokok Berdasarkan Jenis Kelamin*

- Kategori **no** (tidak merokok) lebih banyak diisi oleh **female** dibandingkan **male**.
- Kategori **yes** (merokok) lebih banyak diisi oleh **male** dibandingkan **female**.

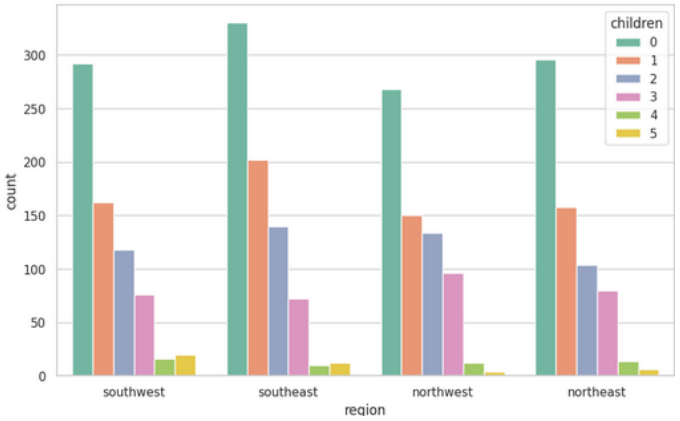
- **Distribusi Status Merokok Berdasarkan Wilayah**



Gambar 1.9 *Visualisasi Distribusi Status Merokok Berdasarkan Wilayah*

Wilayah **southeast** memiliki proporsi perokok lebih tinggi dibandingkan wilayah lain.

- **Distribusi Jumlah Anak Berdasarkan Wilayah**

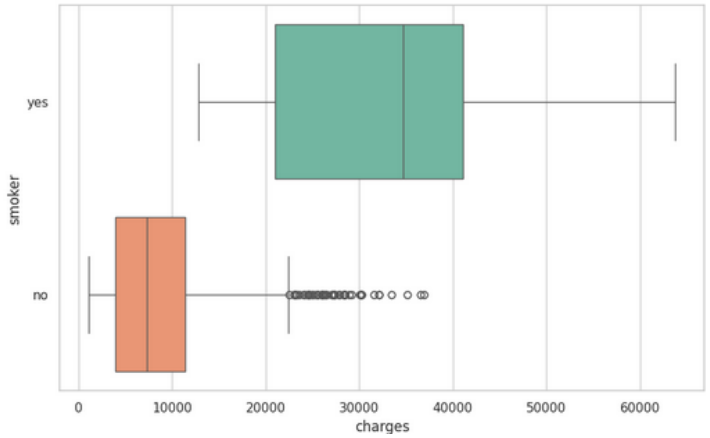


Gambar 1.10 *Visualisasi Distribusi Jumlah anak berdasarkan wilayah*

- Mayoritas individu di semua wilayah tidak memiliki anak (kategori 0).
- Jumlah individu menurun seiring bertambahnya jumlah anak.
- Wilayah **southeast** memiliki jumlah individu tanpa anak tertinggi.
- Individu dengan 4 atau 5 anak sangat sedikit di semua wilayah.

3.5 Analisis Biaya Berdasarkan Status Merokok

- **Boxplot Perbandingan Biaya Berdasarkan Status Merokok**

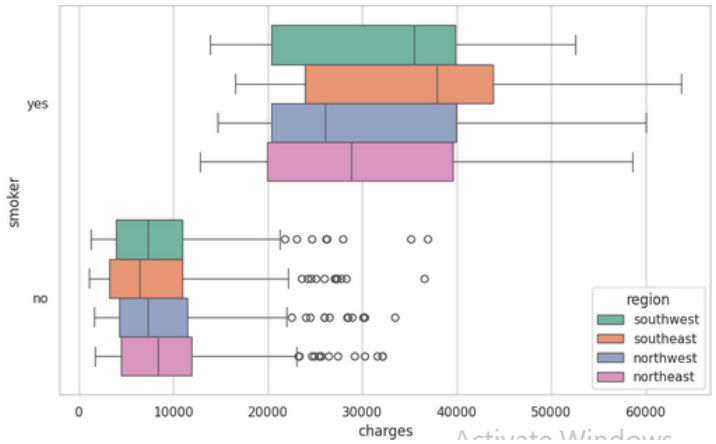


Gambar 1.11 *Boxplot perbandingan Biaya Berdasarkan Status Merokok*

- Perokok memiliki biaya pengobatan yang jauh lebih tinggi dibandingkan non-perokok.
- Distribusi biaya pada perokok lebih luas, dengan banyak individu membayar lebih dari 30.000 USD.

- Non-perokok memiliki biaya lebih rendah dan lebih terkonsentrasi di bawah 15.000 USD.
- Terdapat beberapa outlier pada non-perokok dengan biaya cukup tinggi.

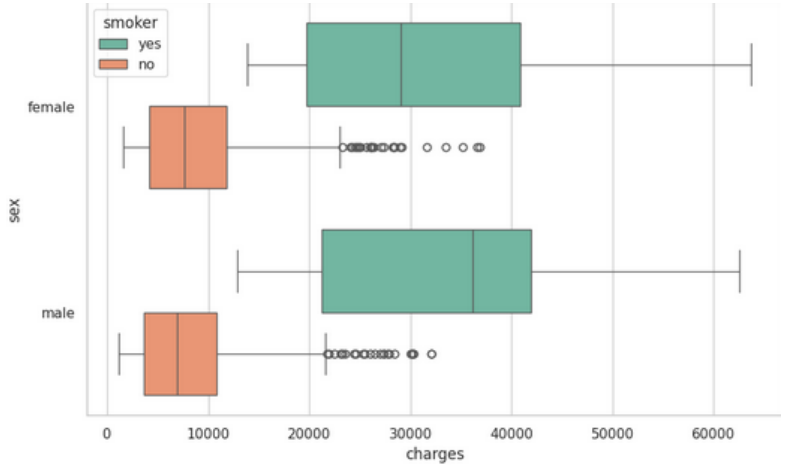
- **Boxplot Biaya Berdasarkan Status Merokok dan Wilayah**



Gambar 1.12 *Boxplot Biaya Berdasarkan Status Merokok dan Wilayah*

- Di antara perokok, wilayah **southeast** memiliki distribusi biaya yang lebih tinggi dibandingkan wilayah lain.
- Non-perokok memiliki distribusi biaya yang lebih rendah dan merata di keempat wilayah.

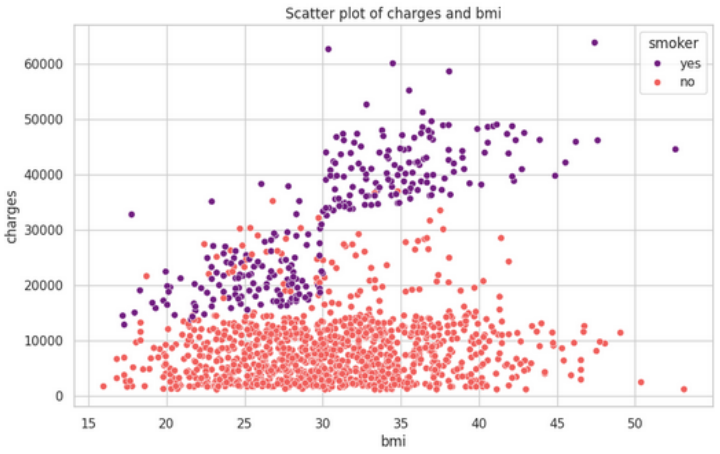
- **Boxplot Biaya Berdasarkan Status Merokok dan Jenis Kelamin**



Gambar 1.13 *Boxplot Biaya Berdasarkan Status Merokok dan Jenis Kelamin*

- Distribusi biaya pada perokok perempuan cenderung lebih tinggi daripada perokok laki-laki.
- Non-perokok, baik laki-laki maupun perempuan, memiliki distribusi biaya yang relatif rendah dan merata.

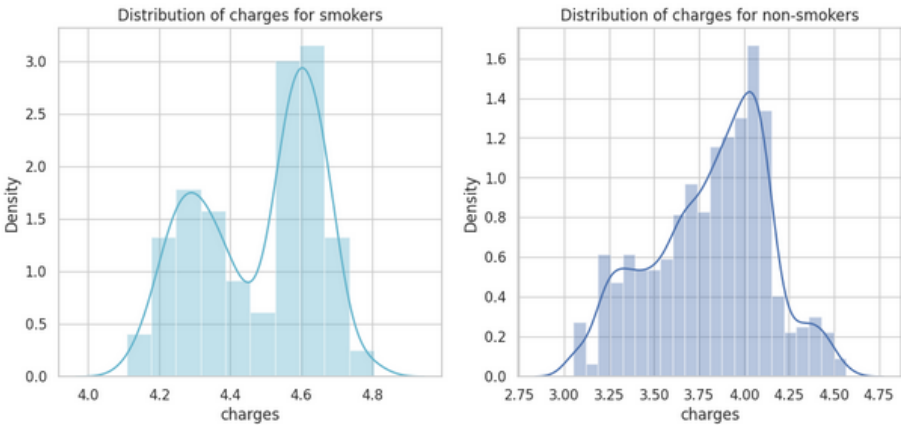
- **Scatter Plot Biaya Pengobatan dan BMI Berdasarkan Status Merokok**



Gambar 1.14 *Scatter Plot Biaya Pengobatan dan BMI berdasarkan status merokok*

- Terdapat korelasi positif antara BMI dan biaya pengobatan, terutama pada perokok.
- Perokok dengan BMI tinggi cenderung memiliki biaya pengobatan yang sangat tinggi.
- Non-perokok memiliki sebaran biaya pengobatan yang lebih terkonsentrasi.

- **Distribusi Biaya Kesehatan Berdasarkan Status Merokok**



Gambar 1.15 *Distribusi Biaya kesehatan berdasarkan status Merokok*

- Distribusi biaya kesehatan pada perokok menunjukkan pola bimodal, menandakan adanya dua kelompok dengan pola biaya berbeda.
- Distribusi biaya pada non-perokok cenderung terkonsentrasi pada nilai rendah, dengan sedikit kasus biaya tinggi sebagai outlier.

4. Preprocessing Data

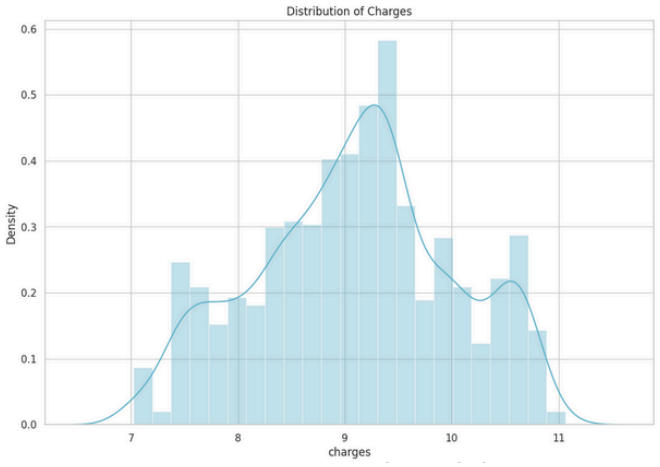
4.1 Label Encoding

Karena model regresi membutuhkan input numerik, variabel kategori dikonversi ke bentuk numerik menggunakan Label Encoding:

- **Sex:** female = 0, male = 1
- **Smoker:** no = 0, yes = 1
- **Region:** northeast = 0, northwest = 1, southeast = 2, southwest = 3

Selain itu, kolom theta0 dengan nilai 1 ditambahkan sebagai bias dalam model regresi

4.2 Transformasi Target



Gambar 1.16 *Distribusi Transformasi Target*

Variabel target **charges** memiliki distribusi yang sangat tidak normal (skewed). Untuk mengatasi hal ini, transformasi logaritmik (log10) diterapkan pada **charges** untuk mendapatkan distribusi yang lebih mendekati normal.

```
[ ]  df['charges'] = np.log10(df['charges'])
```

Gambar 1.17 *Code Transformasi Logaritmik*

Setelah transformasi, distribusi **charges** menjadi lebih simetris dan mendekati distribusi normal, yang membantu model untuk belajar pola data dengan lebih baik.

5. Pemodelan

5.1 Pembagian Data

Data dibagi menjadi set pelatihan (training set) dan set pengujian (testing set) dengan proporsi 80:20:

```
[ ]  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
```

Gambar 1.18 *Code Pembagian Data*

5.2 Standarisasi Data

Sebelum melatih model, data dinormalisasi menggunakan **StandardScaler** untuk memastikan setiap fitur memiliki mean 0 dan standar deviasi 1:

```
[ ]  scaler = StandardScaler()
     scaler.fit(X_train)
     X_train_scaled = scaler.transform(X_train)
     X_test_scaled = scaler.transform(X_test)
```

Gambar 1.19 *Code Standarisasi Data*

5.3 Linear Regression

Model regresi linear diterapkan pada data yang telah distandarisasi:

```
[ ]  # importing linear regression model
     from sklearn.linear_model import LinearRegression
     # instantiate linear regression model
     lin_reg = LinearRegression()
     # fit linear regression model
     lin_reg.fit(X_train_scaled, y_train)
     # predicting the response
     y_pred = lin_reg.predict(X_test_scaled)
```

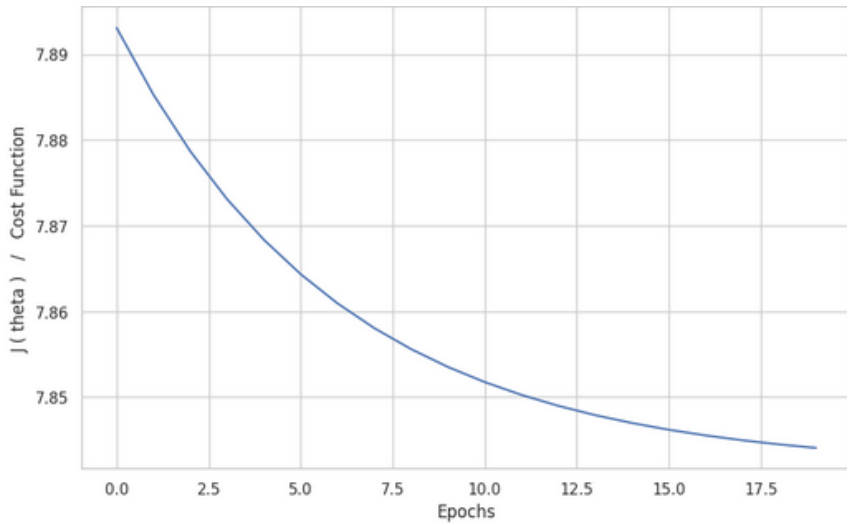
Gambar 1.20 *Model regresi linear diterapkan pada data yang telah distandarisasi*

5.4 Gradient Descent

Implementasi manual gradient descent untuk optimasi parameter model:

```
[ ]  m = X.shape[0]
     alpha = 0.1
     theta = initial_theta
     cost_func_graph = []
     epochs = 20
     for i in range(epochs):
         theta = theta - (alpha/m)*(np.dot(X_train_scaled.T , hyp(theta,X_train_scaled)
         cost_func_graph.append(cost_function(theta,X_train_scaled,y_train))
```

Gambar 1.21 *Code Implementasi manual gradient descent untuk optimasi parameter model*



Gambar 1.22 *Grafik Cost Function*

Grafik cost function menunjukkan penurunan yang konsisten, mengindikasikan model berhasil konvergen menuju minimum cost.

5.5 Polynomial Regression

Model regresi polinomial dengan derajat 2 diterapkan untuk menangkap hubungan non-linear antara fitur dan target:

```
from sklearn.preprocessing import PolynomialFeatures
from sklearn.pipeline import make_pipeline
from sklearn import preprocessing
scaler = StandardScaler()
degree = 2
polyreg = make_pipeline(PolynomialFeatures(degree),scaler,LinearRegression())
polyreg.fit(X_train,y_train)
y_pred = polyreg.predict(X_test)
```

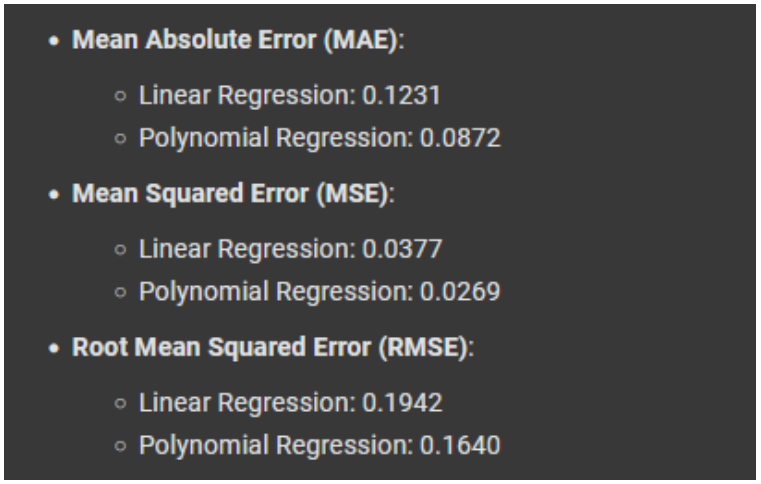
Gambar 1.23 *Code Model regresi polinomial dengan derajat 2*

6. Evaluasi Model

6.1 Metrik Evaluasi

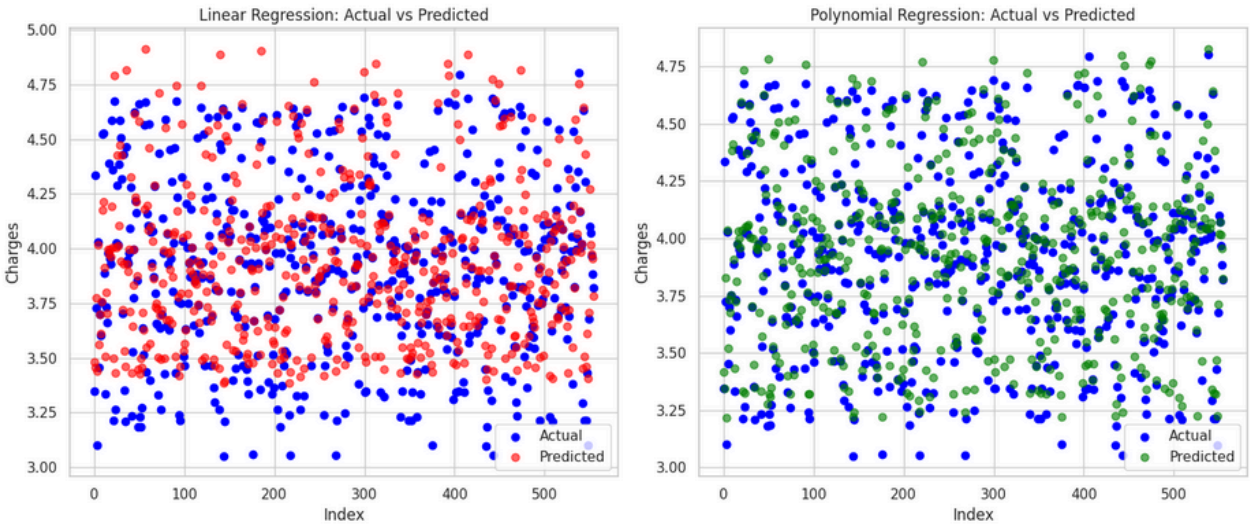
Performa model dievaluasi menggunakan tiga metrik:

- **Mean Absolute Error (MAE):** Rata-rata perbedaan absolut antara nilai prediksi dan nilai aktual.
- **Mean Squared Error (MSE):** Rata-rata kuadrat perbedaan antara nilai prediksi dan nilai aktual.
- **Root Mean Squared Error (RMSE):** Akar kuadrat dari MSE.



Gambar 1.24 *Metrik Evaluasi*

6.2 Visualisasi Hasil Prediksi



Gambar 1.25 *Visualisasi Hasil Prediksi*

Visualisasi perbandingan nilai aktual dan nilai prediksi menunjukkan:

- **Linear Regression:** Prediksi menyebar cukup luas dari nilai aktual, menandakan model kurang mampu menangkap pola dalam data.
- **Polynomial Regression:** Prediksi berada lebih dekat dengan nilai aktual, menunjukkan model lebih akurat dan mampu menangkap kompleksitas hubungan antara fitur dan target dengan lebih baik.

7. Hasil dan Kesimpulan

Dari hasil analisis dan evaluasi model, dapat disimpulkan:

1. **Faktor yang Mempengaruhi Biaya Pengobatan:**
    - **Status merokok** merupakan faktor yang paling signifikan mempengaruhi biaya pengobatan, dengan perokok memiliki biaya jauh lebih tinggi dibandingkan non-perokok.
    - **Usia** dan **BMI** juga berkorelasi positif dengan biaya pengobatan, dimana individu yang lebih tua dan memiliki BMI lebih tinggi cenderung memiliki biaya pengobatan yang lebih tinggi.
    - Kombinasi **BMI tinggi** dan **status perokok** sangat mempengaruhi biaya pengobatan.
2. **Performa Model:**

- **Model regresi polinomial (degree=2)** memberikan performa yang lebih baik dibandingkan regresi linear, ditunjukkan dengan nilai error yang lebih rendah pada semua metrik evaluasi.
- Hal ini mengindikasikan bahwa hubungan antara fitur dan target tidak sepenuhnya linear, dan model polinomial dapat menangkap kompleksitas hubungan tersebut dengan lebih baik.

3. **Implikasi Praktis:**
    - Hasil studi ini dapat membantu perusahaan asuransi dalam menentukan premi yang lebih akurat berdasarkan faktor risiko individu.
    - Individu dapat memahami faktor-faktor yang mempengaruhi biaya pengobatan mereka, terutama pentingnya gaya hidup sehat (status non-perokok dan BMI normal) dalam mengurangi biaya pengobatan.
4. **Keterbatasan dan Saran Pengembangan:**
    - Dataset yang digunakan relatif kecil (2772 sampel), sehingga hasil mungkin tidak sepenuhnya merepresentasikan populasi yang lebih besar.
    - Model yang lebih kompleks seperti Random Forest atau Neural Network mungkin dapat memberikan performa yang lebih baik.
    - Penambahan fitur seperti riwayat penyakit, gaya hidup, atau tingkat pendapatan dapat meningkatkan akurasi prediksi.

Secara keseluruhan, hasil studi ini menunjukkan bahwa status merokok, usia, dan BMI merupakan faktor utama yang mempengaruhi biaya pengobatan, dan model regresi polinomial mampu memprediksi biaya tersebut dengan cukup akurat. Hal ini memiliki implikasi penting bagi penyedia asuransi dalam menentukan premi dan bagi individu dalam memahami faktor risiko yang mempengaruhi biaya kesehatan mereka.