
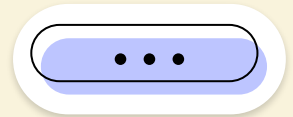
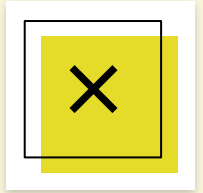



Medical Cost Personal

From kaggle

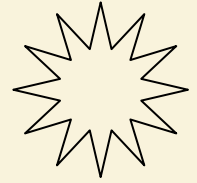
 <https://www.linkedin.com/in/amanahyuni/>

 <https://www.linkedin.com/in/sinhiya-kusuma>



Health Insurance

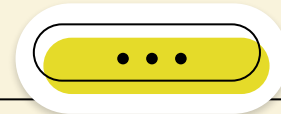
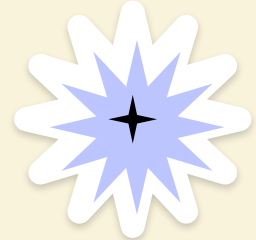
Polis asuransi kesehatan: polis yang menanggung atau meminimalkan biaya kerugian yang disebabkan oleh berbagai risiko kesehatan.



Faktor yang menyebabkan biaya berbeda, antara lain:

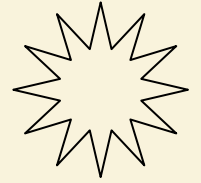
1. Usia tertanggung
2. Seorang perokok yang aktif
3. Riwayat kesehatan
4. Jenis pekerjaan

Perkiraan besarnya biaya yang akurat dapat membantu perusahaan asuransi kesehatan untuk merencanakan masa depan dan memprioritaskan alokasi sumber daya manajemen perawatan yang terbatas.



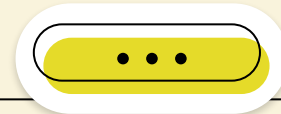
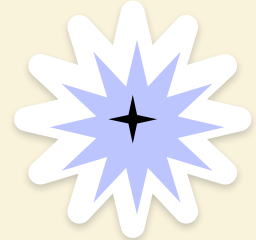
Health Insurance

Biaya klaim dalam asuransi kesehatan merupakan masalah yang seringkali dihadapi oleh perusahaan asuransi.



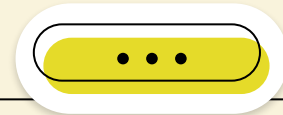
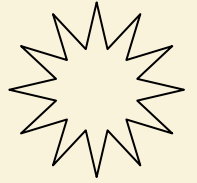
Biaya klaim yang tinggi dapat mengancam stabilitas keuangan perusahaan dan mengakibatkan kerugian finansial yang signifikan.

Oleh karena itu, perusahaan asuransi kesehatan perlu memperhitungkan biaya klaim secara cermat dan akurat agar dapat membuat keputusan bisnis yang bijaksana.



Tujuan

Untuk memprediksi secara akurat biaya klaim asuransi kesehatan berdasarkan karakteristik seseorang yaitu umur, jenis kelamin, BMI, dll.





01

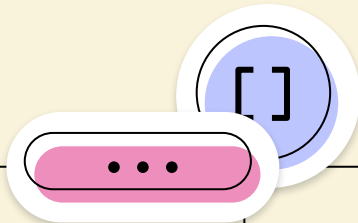


[]



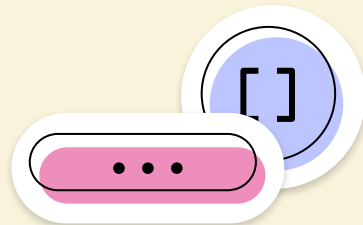
Data Understanding

Data Description



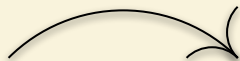
Nama Variabel	Deskripsi	Tipe Data
Age	Usia penerima manfaat asuransi kesehatan	Integer
Sex	Jenis kelamin penerima manfaat asuransi kesehatan	Object (Male, Female)
Body Mass Index (BMI)	Gambaran seberapa ideal berat badan penerima manfaat asuransi kesehatan dibandingkan dengan tinggi badannya	float
Children	Jumlah anak yang ditanggung oleh perusahaan asuransi kesehatan	integer
Smoker	Penerima manfaat merupakan perokok aktif atau tidak	Object (Yes, No)
Region	Tempat tinggal penerima manfaat asuransi kesehatan di Amerika Serikat	Object (Northeast, Northwest, Southeast, Southwest)
Charges	Biaya kesehatan individu yang dibayarkan oleh perusahaan asuransi kesehatan	float

Data Cleansing



```
df.isnull().sum()
```

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

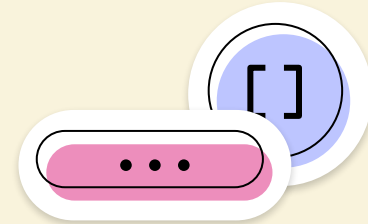


- Tidak ada missing value pada data yang digunakan.
- Terdapat duplikasi pada data, sehingga salah satu data yang terduplikasi akan dihapus.

```
df[df.duplicated()]
```

	age	sex	bmi	children	smoker	region	charges
581	19	male	30.59	0	no	northwest	1639.5631

Data Summary Statistics



	age	bmi	children	charges
count	1337.000000	1337.000000	1337.000000	1337.000000
mean	39.222139	30.663452	1.095737	13279.121487
std	14.044333	6.100468	1.205571	12110.359656
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.290000	0.000000	4746.344000
50%	39.000000	30.400000	1.000000	9386.161300
75%	51.000000	34.700000	2.000000	16657.717450
max	64.000000	53.130000	5.000000	63770.428010

- Distribusi feature 'age' berada di rentang usia 18 hingga 64 tahun.
- Distribusi feature 'bmi' relatif simetris, ditunjukkan dengan nilai mean yang hampir sama dengan nilai median.
- Distribusi feature 'children' menunjukkan jika sebanyak 75% individu yang mengikuti asuransi kesehatan memiliki kurang dari sama dengan 2 anak.
- 'charges' memiliki skewness positif yang ditunjukkan dengan nilai mean > median, berarti bahwa biaya klaim yang banyak terjadi berada di kisaran yang rendah.



02

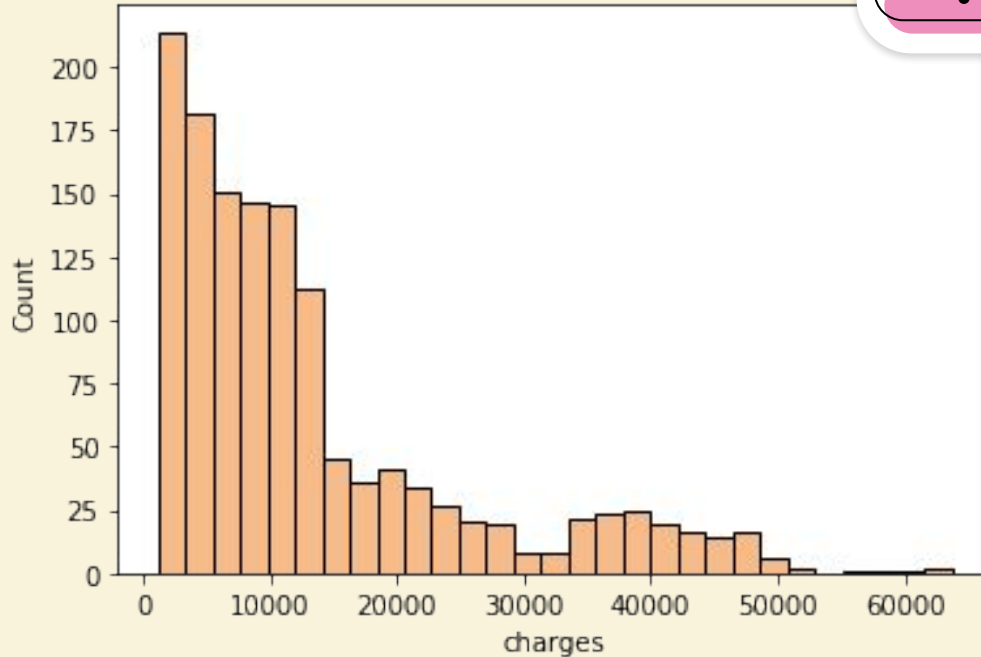


[]



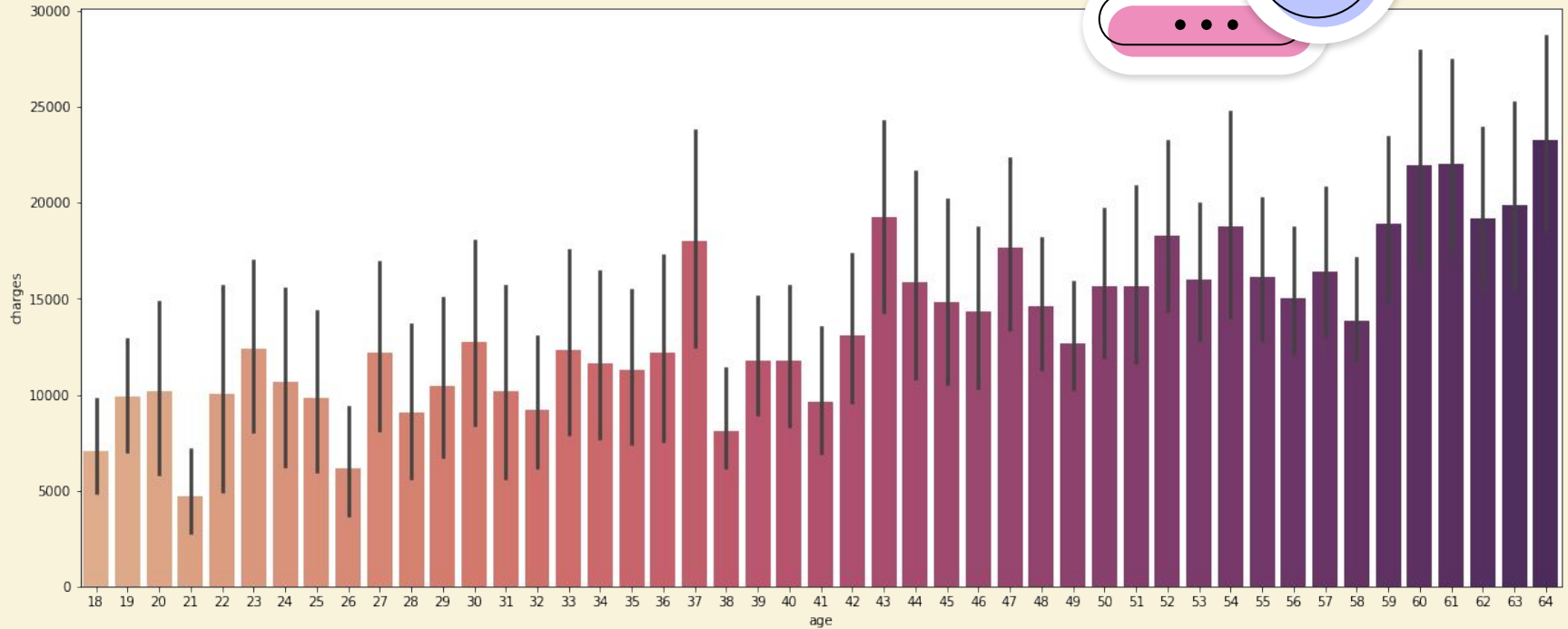
Exploratory Data Analysis (EDA)

Charges



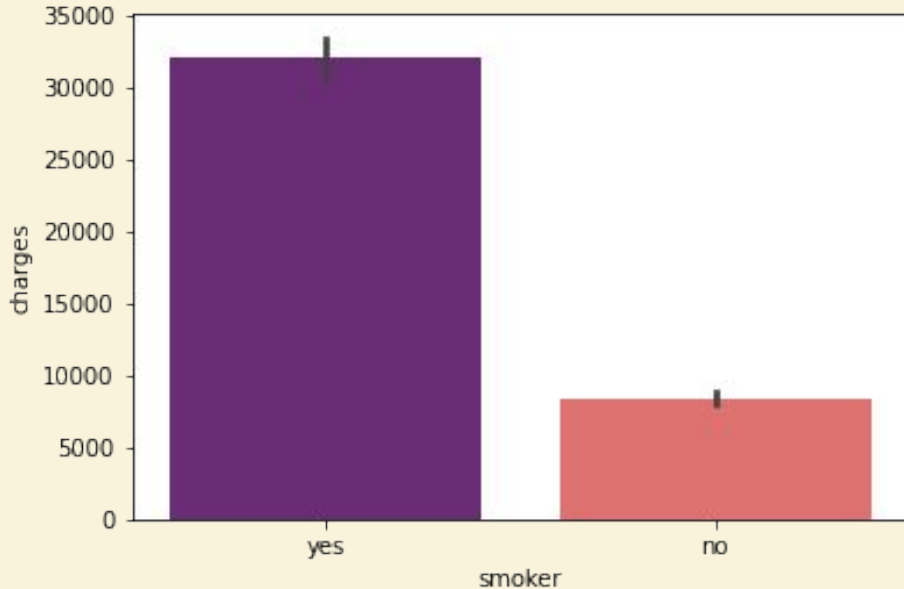
Distribusi 'charges' yaitu skewed positif dimana jumlah klaim paling banyak pada 'charges' yang relatif rendah yaitu sekitar 10.000. Kemungkinan kebanyakan orang hanya mengklaim biaya kesehatan yang ringan.

Charges by Age

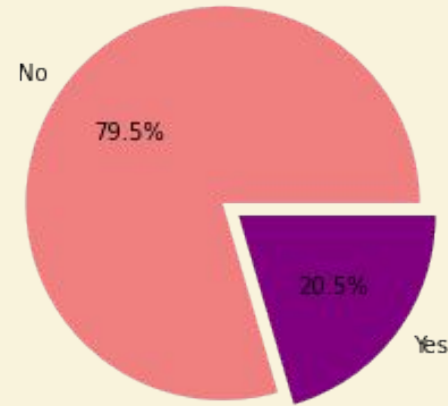


Semakin bertambah 'age', maka memiliki kecenderungan semakin besar pula 'charges' yang dikeluarkan.

Charges by Smoker

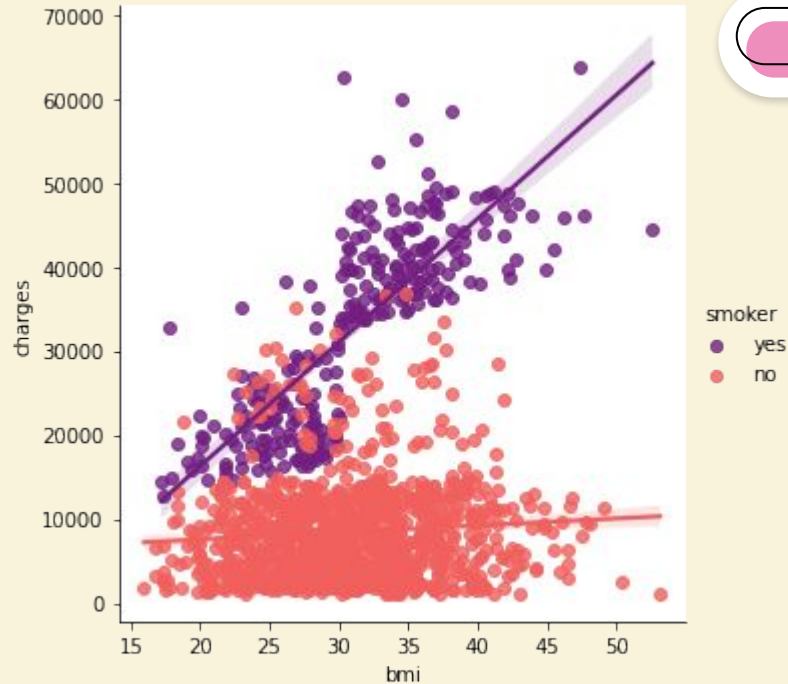


Percentage of Smoker and Non-Smoker



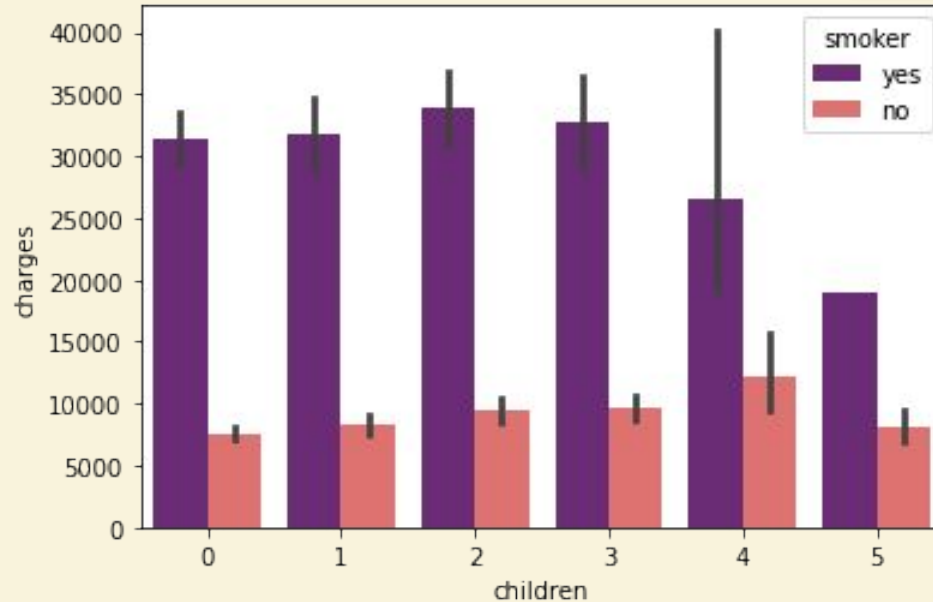
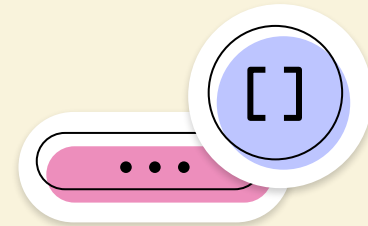
Jumlah perokok lebih sedikit daripada tidak perokok, namun perokok akan cenderung memiliki 'charges' lebih tinggi dibandingkan dengan tidak perokok. Seorang perokok menghabiskan lebih banyak biaya klaim.

Charges by BMI



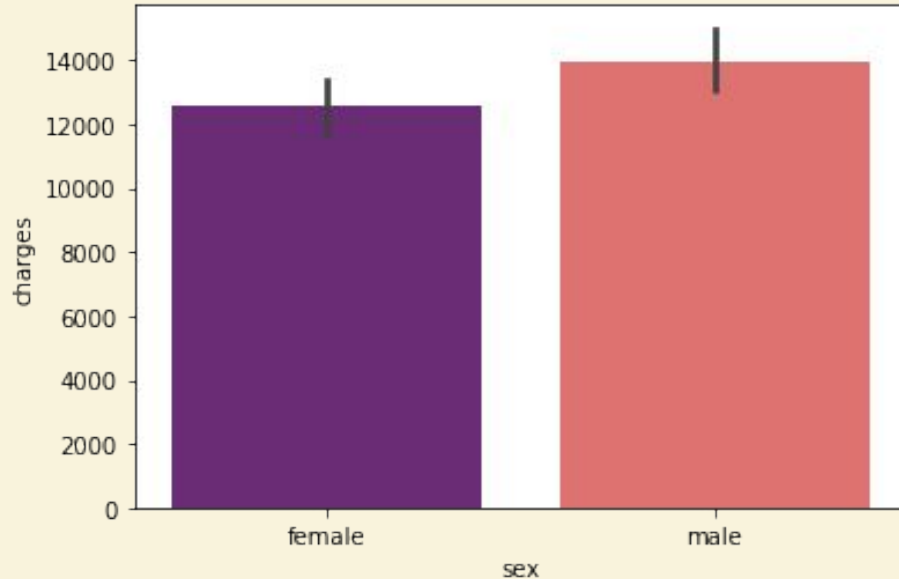
Di kalangan perokok, 'charges' cenderung meningkat yang signifikan seiring dengan tingginya 'BMI'.

Charges by Children

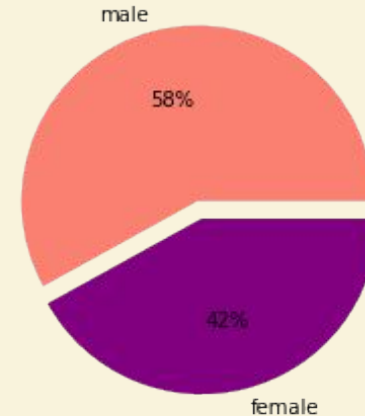


Seseorang yang memiliki jumlah anak 5 cenderung memiliki 'charges' yang paling sedikit, baik yang merokok dan tidak merokok.

Charges by Sex

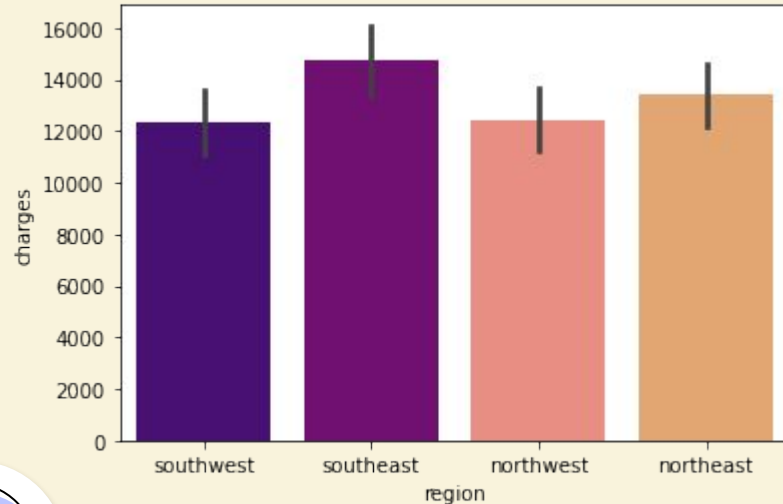


Percentage of Smoker by Sex

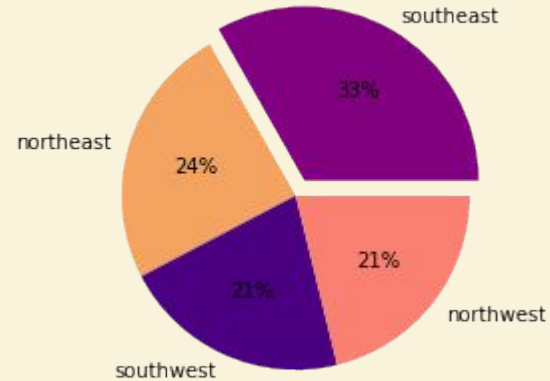


Laki-laki cenderung memiliki 'charges' yang lebih tinggi dibandingkan dengan perempuan karena jumlah laki-laki perokok lebih banyak.

Charges by Region



Percentage of Smoker by Region

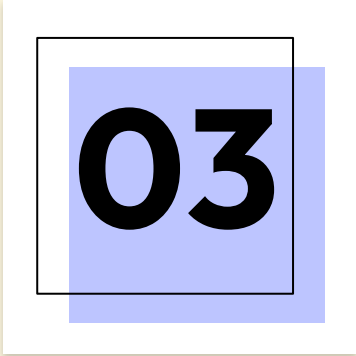
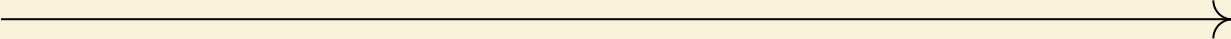


Seseorang yang bertempat tinggal di wilayah southeast cenderung memiliki 'charges' yang lebih tinggi daripada wilayah lain. Jika dilihat keterkaitan dengan jumlah perokok, di daerah tersebut terdapat jumlah perokok yang lebih banyak dibandingkan dengan wilayah lain. Hal ini berbanding lurus dengan nilai 'charges' yang lebih tinggi juga.

Summary From EDA

A pink circle with a white border, containing a black bracket symbol '[]'.

- Semakin bertambahnya 'age' maka semakin besar 'charges' yang dikeluarkan karena memiliki risiko kesehatan yang tinggi.
- Seorang perokok cenderung memiliki 'charges' yang relatif tinggi dapat dilihat pada feature lain. Jika jumlah perokok di suatu wilayah tinggi maka dapat mengakibatkan jumlah 'charges' pada wilayah tersebut tinggi pula. Seseorang dengan 'BMI' yang tinggi dan dia adalah perokok maka 'charges' akan tinggi pula begitupula 'charges' yang lebih tinggi pada laki-laki yang dapat diakibatkan jumlah perokok laki-laki yang tinggi
- 75% Individu yang melakukan klaim memiliki anak kurang dari 2 atau 2. Hal tersebut dapat menjadi salah satu alasan jumlah 'charges' yang tinggi pada kategori tersebut. Kemungkinan lain yaitu pada individu yang memiliki anak 4 dan 5 'charges' cenderung rendah karena jumlah perokok yang rendah.



03

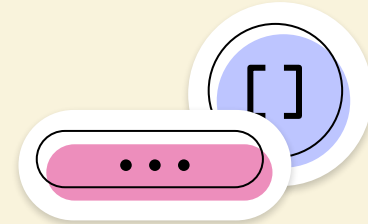


[]



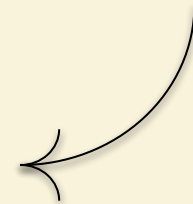
Data Preprocessing

Data Preprocessing



	age	bmi	children	charges	region_northeast	region_northwest	region_southeast	region_southwest	sex_female	sex_male	smoker_no	smoker_yes
0	-1.440418	-0.453160	-0.909234	16884.92400	0	0	0	1	1	0	0	1
1	-1.511647	0.509422	-0.079442	1725.55230	0	0	1	0	0	1	1	0
2	-0.799350	0.383155	1.580143	4449.46200	0	0	1	0	0	1	1	0
3	-0.443201	-1.305052	-0.909234	21984.47061	0	1	0	0	0	1	1	0
4	-0.514431	-0.292456	-0.909234	3866.85520	0	1	0	0	0	1	1	0

1. Untuk memudahkan dalam pemodelan data, maka data kategorik akan diubah menjadi numerik dengan one-hot encoding untuk features 'region', 'sex', dan 'smoker'.
2. Dilakukan standarisasi untuk features numerik, yaitu 'age', 'bmi', 'children'.

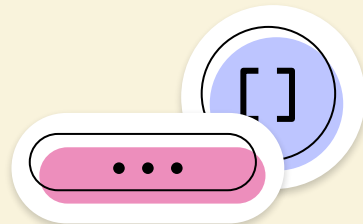


Features Correlation



'smoker_yes' memiliki korelasi positif yang paling tinggi terhadap 'charges' sebesar 0.79 dan diikuti oleh 'age' sebesar 0.3 serta 'bmi' sebesar 0.2.

Features Selection



- Variabel Prediktor = age, sex_female, sex_male, smoker_no, smoker_yes, bmi, children, region_northeast, region_northwest, region_southeast, region_southwest
- Variabel Target = charges



04

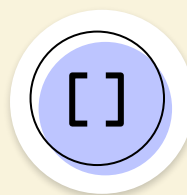


[]



Modelling

Model Selection



	Model	RMSE	R-Squared (training)	Adjusted R-Squared (training)	R-Squared (test)	Adjusted R-Squared	5-Fold Cross Validation
0	Linear Regression	6105.030	0.754	0.751	0.737	0.729	0.747
1	Bayesian Ridge	6098.378	0.754	0.751	0.738	0.730	0.747
2	XGB Regression	4297.053	0.899	0.898	0.870	0.866	0.857
3	Gradient Boosting	4286.769	0.904	0.903	0.871	0.867	0.854
4	Ada Boost	4947.968	0.833	0.831	0.827	0.822	0.824
5	KNeighborsRegressor	5452.854	0.842	0.841	0.790	0.784	0.753

Berdasarkan kelima model di tabel tersebut, didapatkan model terbaik yaitu Gradient Boosting Regression.

Selanjutnya, akan dilakukan hyperparameter pada Gradient Boosting Regression.

Model Selection



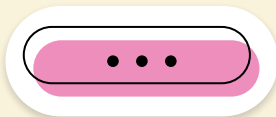
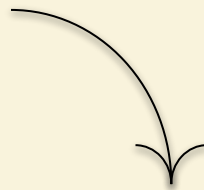
Gradient Boosting Hyperparameter Regression

Results from Grid Search

The best estimator across ALL searched params:
`GradientBoostingRegressor(learning_rate=0.03, max_depth=4, subsample=0.5)`

The best score across ALL searched params:
0.8406445159575473

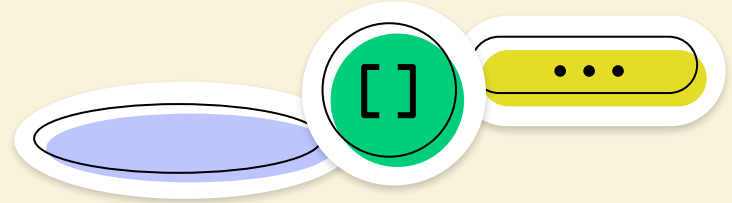
The best parameters across ALL searched params:
{'learning_rate': 0.03, 'max_depth': 4, 'n_estimators': 100, 'subsample': 0.5}



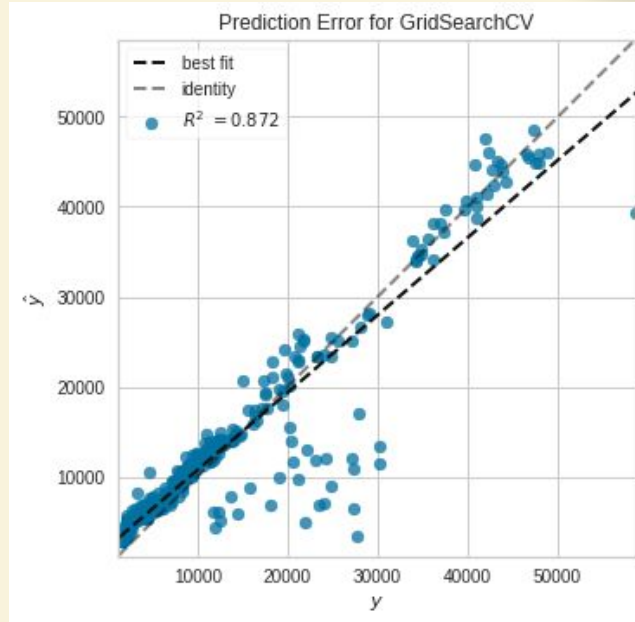
Evaluasi model:

- RMSE: 4266.855
- R^2 training: 0.893
- R^2 testing: 0.872
- 5-Fold Cross Validation: 0.857

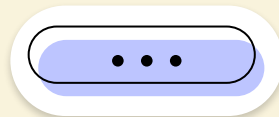
Model Prediction



	Actual	Predicted
856	40974.16490	42050.442682
778	5934.37980	5502.251920
65	1743.21400	7748.918924
624	12129.61415	14484.935532
1007	24915.22085	36843.820692
...
1231	20167.33603	16862.015560
1151	12235.83920	12393.245460
1178	2899.48935	6193.992950
235	19444.26580	21066.426360
493	12574.04900	13664.165080



Gambar di samping merupakan hasil prediksi charges dari model tersebut terhadap data aktual 'charges'. Terlihat hasil prediksi mendekati hasil aktual dengan R square sebesar 0.87 dan error sebesar 0.095%



Thank

you

