TASK2: Medical Cost Personal Insurance Project

Project Description

Health insurance is a type of insurance that covers medical expenses that arise due to an illness. These expenses could be related to hospitalisation costs, cost of medicines or doctor consultation fees. The main purpose of medical insurance is to receive the best medical care without any strain on your finances. Health insurance plans offer protection against high medical costs. It covers hospitalization expenses, day care procedures, domiciliary expenses, and ambulance charges, besides many others. Based on certain input features such as age , bmi,,no of dependents ,smoker ,region medical insurance is calculated .

Columns

• age: age of primary beneficiary • sex: insurance contractor gender, female, male • bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9. • children: Number of children covered by health insurance / Number of dependents • smoker: Smoking • region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest. • charges: Individual medical costs billed by health insurance

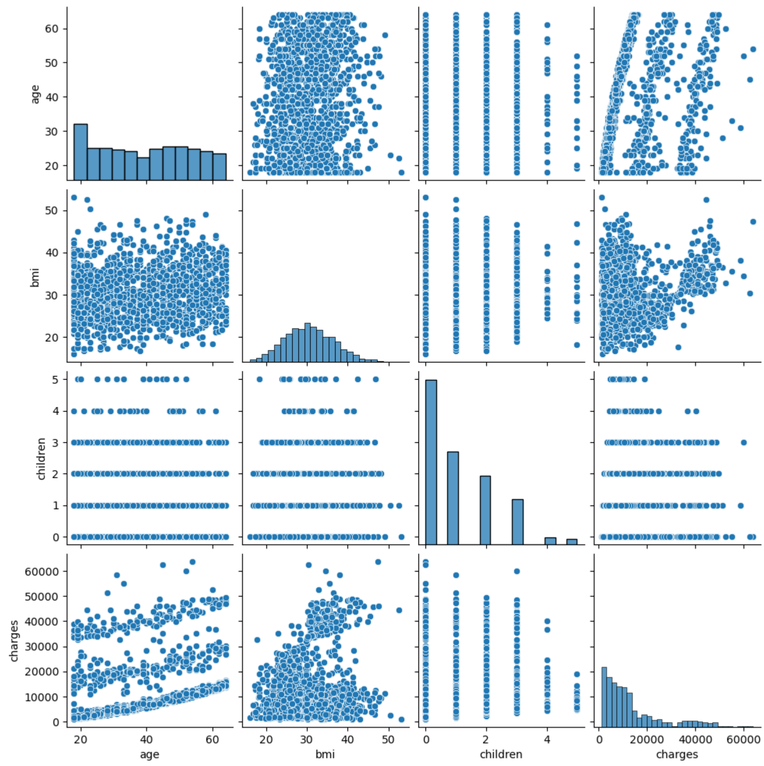
Predict : Can you accurately predict insurance costs?

Dataset Link-

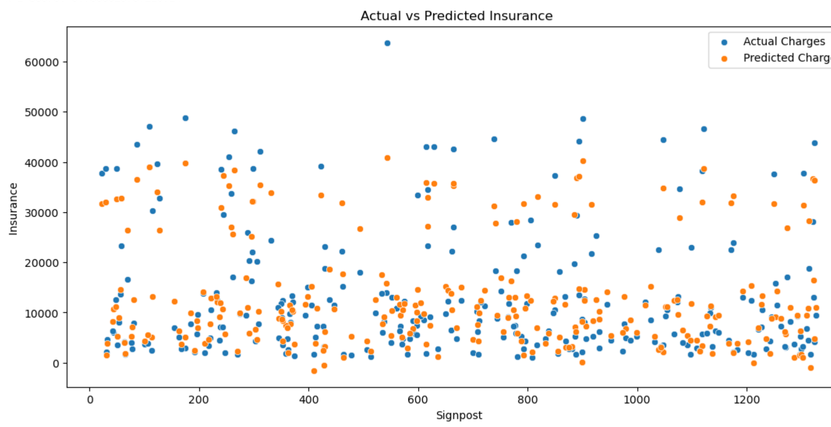
https://github.com/dsrscientist/dataset4 https://github.com/dsrscientist/dataset4/blob/main/medical_cost_insurance.csv

In [4]:
```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# Load the dataset
url_link = "https://raw.githubusercontent.com/dsrscientist/dataset4/main/medical_cost_insurance.csv"
df = pd.read_csv(url)
print(df.info())

# Display the first few rows of the dataset
print(df.head())

# Data Visualization
sns.pairplot(df)
plt.show()

# Data Preprocessing
# Convert categorical variables into numerical using one-hot encoding
df = pd.get_dummies(df, columns=['sex', 'smoker', 'region'], drop_first=True)

# Split the data into features (X) and target variable (y)
X = df.drop('charges', axis=1)
y = df['charges']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Train a linear regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print(f'Mean Squared Error: {mse}')
print(f'R^2 Score: {r2}')

# Visualization of predicted vs actual charges
plt.figure(figsize=(13, 6))
sns.scatterplot(x=y_test.index, y=y_test, label='Actual Charges')
sns.scatterplot(x=y_test.index, y=y_pred, label='Predicted Charges')
plt.title('Actual vs Predicted Insurance ')
plt.xlabel('Signpost')
plt.ylabel('Insurance ')
plt.legend()
plt.show()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
None
   age     sex     bmi  children smoker     region      charges
0   19  female  27.900         0    yes  southwest  16884.92400
1   18    male  33.770         1     no  southeast   1725.55230
2   28    male  33.000         3     no  southeast   4449.46200
3   33    male  22.705         0     no  northwest  21984.47061
4   32    male  28.880         0     no  northwest   3866.85520
```




```
Mean Squared Error: 33596915.851361446
R^2 Score: 0.7835929767120724
```



In [ ]: