

# INSURANCE CHARGES PREDICTION

[Problem Statement](#)

[Basic info about the dataset](#)

[Pre-processing](#)

[Split data into Train and Test sets](#)

[Create the model, train the model, and make a prediction](#)

[Evaluate the model using r2\\_score](#)

[Best model](#)

[Evaluation data of models with different hyperparameters](#)

## Problem Statement

Develop a machine learning (ML) model to predict individual insurance charges based on demographic and health-related factors, such as:

- Age
- Sex
- body mass index (BMI)
- number of children
- smoking habits

The goal is to identify key factors influencing insurance costs and develop a reliable predictive framework for cost estimation.

## Basic info about the dataset

The data set is about the demographic and health-related factors as independent variables (Age, Sex, BMI, Children, smoker), and Insurance Charges as the dependent (target) variable (charges) is continuous numeric data that the model is predicting.

“Sex and smoker” are categorical data

## Pre-processing

Convert nominal data to numeric data (sex and smoker) using `get_dummies` and convert the resulting Boolean value to integers using `.astype(int)`

## Split data into Train and Test sets

Split the test data into Train and test sets using `sklearn.Model Selection Module` by importing the `train_test_split`

## Create the model, train the model, and make a prediction

Create a model using Multiple Linear Regression & train the model using multiple features, and make the predictions

## Evaluate the model using r2\_score

Evaluate the model using `r2_score`, If `r2_score` is high (close to 1), the model fits well.

Best model

11	RF	R2_score	n_estimators=100, criterion="absolute_error", random_state=42, min_samples_split=20	0.87
----	----	----------	--	------

Evaluation data of models with different hyperparameters

	ML model	Evaluation Metric	Parameters	Value
1	MLR	R2_score	default	0.78
2	SVM	R2_score	default	-0.06
3	DT	R2_score	criterion='squared_error', splitter='best'	0.72
4	DT	R2_score	criterion='squared_error', splitter='best', max_depth=5	0.83
5	DT	R2_score	criterion='squared_error', splitter='best', max_depth=3	0.85
6	DT	R2_score	criterion='squared_error', splitter='best', max_depth=4	0.86
7	DT	R2_score	friedman_mse', splitter='best', max_depth=4	0.86
8	DT	R2_score	criterion='poisson', splitter='best', max_depth=4, min_samples_split=4	0.86
9	RF	R2_score	n_estimators=100, random_state=42	0.86
10	RF	R2_score	n_estimators=100, criterion="absolute_error", random_state=42	0.86
11	RF	R2_score	n_estimators=100, criterion="absolute_error", random_state=42, min_samples_split=20	0.87