

***Faculty of Computers and Information Cairo University***

**Theoretical Foundations of Machine Learning**

**Date:** 18/1/ 2021

**Course Project:** 3- Medical Cost Personal

**Name and ID:**

Marina Moheb Nafee 20180208

1. Load the dataset.

| **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 |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **1333** | 50 | male | 30.970 | 3 | no | northwest | 10600.54830 |
| **1334** | 18 | female | 31.920 | 0 | no | northeast | 2205.98080 |
| **1335** | 18 | female | 36.850 | 0 | no | southeast | 1629.83350 |
| **1336** | 21 | female | 25.800 | 0 | no | southwest | 2007.94500 |
| **1337** | 61 | female | 29.070 | 0 | yes | northwest | 29141.36030 |

1338 rows × 7 columns

2. Prepare the train-validation and test portions.

It is a dataset for regression tasks. It consists of 1300+ records containing persons medical data and the target is "charge" column. The goal is make a model that fits these data and predicts the charge for the new persons that the medical insurance should cover. The data is to be divided to 1000 samples for both training and validation and the rest is for testing

3. Apply any preprocessing or features that you find suitable for the data. (Phase 2)

LabelEncoder : On 'sex' 'smoker' 'region'

|  | **age** | **sex** | **bmi** | **children** | **smoker** | **region** | **charges** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 19 | 0 | 27.900 | 0 | 1 | 3 | 16884.92400 |
| **1** | 18 | 1 | 33.770 | 1 | 0 | 2 | 1725.55230 |
| **2** | 28 | 1 | 33.000 | 3 | 0 | 2 | 4449.46200 |
| **3** | 33 | 1 | 22.705 | 0 | 0 | 1 | 21984.47061 |
| **4** | 32 | 1 | 28.880 | 0 | 0 | 1 | 3866.85520 |
| **...** | ... | ... | ... | ... | ... | ... | ... |
| **1333** | 50 | 1 | 30.970 | 3 | 0 | 1 | 10600.54830 |
| **1334** | 18 | 0 | 31.920 | 0 | 0 | 0 | 2205.98080 |
| **1335** | 18 | 0 | 36.850 | 0 | 0 | 2 | 1629.83350 |
| **1336** | 21 | 0 | 25.800 | 0 | 0 | 3 | 2007.94500 |
| **1337** | 61 | 0 | 29.070 | 0 | 1 | 1 | 29141.36030 |

1338 rows × 7 columns

SimpleImputer : to remove any missing values and NaN

**Feature Scaling : using StandardScaler**

4. Apply 3 different models and compare between them.

(Phase 3)

1. SVR(Support Vector Regression)
2. LinearRegression
3. DecisionTreeRegressor

Accuracy of each model

Accuracy of SVR(Support Vector Regression):

Mean Absolute Error: 8244.288334509303

Mean Squared Error: 161950464.7154791

Root Mean Squared Error: 12725.975982826587

Accuracy of LinearRegression:

Mean Absolute Error: 4219.582514385898

Mean Squared Error: 38945609.28174972

Root Mean Squared Error: 6240.6417363721275

Accuracy of DecisionTreeRegressor:

Mean Absolute Error: 3196.820930958

Mean Squared Error: 45417937.31708163

Root Mean Squared Error: 6739.28314563809

A comment on the results and on the comparison of the three models applied should be given.

DecisionTreeRegressor is the best model of used one according to MSE then LinearRegression

the worst is SVR (support vector regression )

Reasons and suggested improvements should also be included.

1. Add more data
2. Treat missing and Outlier value
3. Feature Engineering
4. Feature Selection
5. Multiple algorithms
6. Algorithm Tuning
7. Ensemble methods:This is the most common approach found majorly in winning solutions of Data science competitions.This technique simply combines the result of multiple weak models and produce better results.This can be achieved through many ways:Bagging (Bootstrap Aggregating),BoostingTurning