Assignment – Regression Algorithm

The problem statement or requirement for this assignment is to predict the insurance charges based on several parameters. A client has provided a dataset for this purpose, and the task for the data scientist is to develop a model that will predict these insurance charges

How will you achieve this in AI?

- Based on the given Insurance dataset – (Age, Sex, BMI, Children, Smoker, and Charges) , We should train AI with given dataset and AI will predict the insurance charge for an individual.

Find out the 3 -Stage of Problem Identification

- *Stage – 1: Machine Learning / Deep Learning*
- *Stage – 2: Supervised Learning – Dependent variable / Label is present in the dataset. The dataset contains known input features (like Age, Sex, BMI, Children, Smoker) and their corresponding output (Charges).*
- *Stage – 3: Regression - The charges column consists of **continuous numerical values** (e.g., 16884.924, 1725.5523, 21984.47061). Regression tasks involve predicting a continuous quantity, as opposed to classification tasks which predict discrete categories.*

Total number of rows and columns.

- *Row count is 1338 excluding header*
- *Column count is 6.*

Mention the pe-processing method if you're doing any (like converting string to number – nominal data)

Columns Sex and Smoker contain string (text) values which are nominal:

Note:

> ***Nominal data** refers to categorical data where there is **no inherent order or ranking** among the categories. The categories are simply labels used to classify data points.*
>
> ***Ordinal data**, in contrast, is categorical data where the categories **do have a meaningful order or rank**, but the difference between ranks may not be uniform or precisely measurable.*

• The sex column contains values such as 'female' and 'male'.

• The smoker column contains values such as 'yes' and 'no'.

*Therefore, to prepare this data for a machine learning model, these **categorical string values in the 'sex' and 'smoker' columns would need to be transformed into numerical representations**. This is a standard practice for nominal data in machine learning, as most algorithms require numerical input.*

ML Models

1. **Simple Linear Regression:**

   This dataset cannot fit into SLR model as SLR assumes a linear relationship between **one independent variable** and a **dependent variable**. However, the problem statement clearly indicates that

the client wants to "predict the insurance charges based on **the several parameters**". The dataset provided confirms this, containing multiple independent variables: age, sex, bmi, children, and smoker, hence we can use this model for this dataset.

2. **Multiple Linear Regression:**

MLR Default R2 value is 0.7680881643600721

S. No.	**f**it_intercept	**copy_X**	R-sqare Value
1	True	False	0.7680881643600721
2	False	False	0.7413829103991553
3	False	True	0.7413829103991553

In this the combination that gives best R2 = 0.7680881643600721 value is ***default***

3. **Support Vector Machine:**

SVM R-square default value with C = 0.1 is -0.08258518095361467

S. No.	Kernel	C	R-square Value
1	rbf	1	-0.0773076666146093
2	rbf	100	0.32642730830319344
3	rbf	1000	0.7954770012497223
4	poly	1	-0.07015361769665884
5	poly	100	0.6208965326154974
6	poly	1000	0.8506525310780697
7	sigmoid	1	-0.06949725085826985
8	sigmoid	100	0.517744099061606
9	sigmoid	1000	0.32486043455968594

In this the combination that gives best R2 value = 0.8506525310780697 is ***Kernal = Poly, C =1000*.**

4. **Decision Tree:**

DT default R-square value is 0.7348432955402802

S. No.	criterion	splitter	max_features	R-sqare Value
1	squared_error	best	sqrt	0.7090961156131736
2	squared_error	best	log2	0.7246457272523109
3	squared_error	random	log2	0.6238842381959184
4	squared_error	random	sqrt	0.659656175400232
5	friedman_mse	random	sqrt	0.6402863372330073
6	friedman_mse	random	log2	0.727275903907552
7	friedman_mse	best	log2	0.5659056225010219
8	friedman_mse	best	sqrt	0.7166450065096732
9	absolute_error	best	sqrt	0.6567305344952716
10	absolute_error	best	log2	0.7148097717176958
11	absolute_error	random	log2	0.6889179252802101
12	absolute_error	random	sqrt	0.7384953921629618
13	poisson	random	log2	0.6927714172306294
14	poisson	random	sqrt	0.6282684336450418
15	poisson	best	sqrt	0.699505152077756
16	poisson	best	log2	0.64886576682164

In this the combination that gives best R2 value = 0.7384953921629618 is

***criterion = absolute_error, splitter= random, max_features= sqrt*.**

5. **Random Forest:**
RF default R-square value is 0.8484409564066483

S. No.	n_estimator	random_state	R-sqare Value
1	50	0	0.841822261354067
2	50	10	0.8446119587203765
3	50	25	0.8462293278095468
4	50	45	0.8460538696379465
5	50	50	0.844755253866104
6	100	50	0.8448191177888533
7	100	45	0.845929486490081
8	100	25	0.8501068036411814
9	100	10	0.8447872765155557
10	100	0	0.845400995603081

In this the combination that gives best R2 value = 0.8501068036411814 is
n_estimator = 100, random_state = 25.

Model	R2 Value
MLR	0.7680881643600721
SVM	0.8506525310780697
DT	0.7384953921629618
RF	0.8501068036411814

The **best model for this dataset** is the **Support Vector Machine (SVM)** regression model.
Here's why:

- **The assignment requires developing a "good model with r2_score". The R-squared (r2_score) value is a key metric used to evaluate the performance of regression models, with a higher value indicating a better fit to the data [external information].**
- **The sources document the R2 values for several machine learning models after experimentation:**
    - **Multiple Linear Regression (MLR) yielded a default R2 value of 0.7680881643600721.**
    - **Support Vector Machine (SVM) achieved its best R2 value of 0.8506525310780697 with a Kernel = Poly and C = 1000.**
    - **Decision Tree (DT) achieved its best R2 value of 0.7384953921629618 with criterion = absolute_error, splitter= random, and max_features= sqrt.**
    - **Random Forest (RF) achieved its best R2 value of 0.8501068036411814 with n_estimator = 100 and random_state = 25.**

Comparing these R2 values, the SVM model with the polynomial kernel and C=1000 configuration (R2 = 0.8506525310780697) demonstrates the highest performance among the models tested and documented. This indicates that it explains the largest proportion of the variance in the insurance charges, making it the most suitable model according to the evaluation metric specified.