**Regression Assignment**

Insurance charges prediction based on the given parameters:

1. The client wants to predict the charges for the insurance by analysing the parameters such as age, sex, BMI, children and smoking habit.

The three stages involved in identification of the problem statement are

1. Machine Learning as the inputs are numbers.
2. Supervised Learning as the requirement is clear and the input & output both are available for prediction.
3. Regression as the output is continuos and a numerical value.
4. The given dataset contains 5 input columns and one output column with total train and test sets including 1338 row datas.
5. The preprocessing includes two columns as they are categorical datas. This includes ‘sex’ column as nominal data and ‘smoker’ column as ordinal data. Accordingly, One hot encoding and Label encoder are being used for preprocessing respectively.
6. Tabulation:

* Multiple Linear Regression- R2 Score:- 0.7894
* Support Vector Machine:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.No | Hyper Parameter | Linear (r2 score) | RBF (r2 score) | Poly (r2 score) | Sigmoid(r2 score) |
| 1. | C10 | 0.4624 | -0.032 | 0.038 | 0.0393 |
| 2. | C100 | 0.6288 | 0.3200 | 0.6179 | 0.5276 |
| 3. | C500 | 0.7631 | 0.6642 | 0.8263 | 0.4460 |
| 4. | C1000 | 0.7649 | 0.8102 | 0.8566 | 0.2874 |
| 5. | C2000 | 0.7440 | 0.8547 | 0.8605 | -0.5939 |
| 6. | C3000 | 0.7414 | 0.8663 | 0.8598 | -2.12 |

The SVM Regression has best R2 score of : 0.8598

* Decision Tree

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | Criterion | Max Features | Splitter | R2 score |
| 1. | Squared error | sqrt | best | 0.7521 |
| 2. | Squared error | sqrt | random | 0.7148 |
| 3. | Squared error | Log 2 | best | 0.6999 |
| 4. | Squared error | Log2 | random | 0.7229 |
| 5. | ***absolute\_error*** | sqrt | best | 0.7648 |
| 6. | ***absolute\_error*** | sqrt | random | 0.7101 |
| 7. | ***absolute\_error*** | Log2 | best | 0.7233 |
| 8. | ***absolute\_error*** | Log2 | random | 0.7339 |
| 9. | ***poisson*** | Sqrt | best | 0.7332 |
| 10. | ***poisson*** | Sqrt | random | 0.7325 |
| 11. | ***poisson*** | Log2 | best | 0.7457 |
| 12. | ***poisson*** | Log2 | random | 0.6828 |
| 13. | Friedman\_mse | sqrt | best | 0.6992 |
| 14. | Friedman\_mse | sqrt | random | 0.6360 |
| 15. | Friedman\_mse | Log 2 | best | 0.6408 |
| 16. | Friedman\_mse | Log 2 | random | 0.6511 |

The Decision Tree Regression has best R2 score of : 0.7648

* Random Forest:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S.No | Criterion | Max Features | N\_Estimator | R2 Score |
| 1. | Squared error | sqrt | 10 | 0.8555 |
| 2. | Squared error | sqrt | 100 | 0.8694 |
| 3. | Squared error | Log 2 | 10 | 0.8483 |
| 4. | Squared error | Log2 | 100 | 0.8696 |
| 5. | ***absolute\_error*** | sqrt | 10 | 0.8481 |
| 6. | ***absolute\_error*** | sqrt | 100 | 0.8689 |
| 7. | ***absolute\_error*** | Log2 | 10 | 0.8616 |
| 8. | ***absolute\_error*** | Log2 | 100 | 0.8724 |
| 9. | ***poisson*** | Sqrt | 10 | 0.8461 |
| 10. | ***poisson*** | Sqrt | 100 | 0.8729 |
| 11. | ***poisson*** | Log2 | 10 | 0.8586 |
| 12. | ***poisson*** | Log2 | 100 | 0.8720 |
| 13. | Friedman\_mse | sqrt | 10 | 0.8624 |
| 14. | Friedman\_mse | sqrt | 100 | 0.8683 |
| 15. | Friedman\_mse | Log 2 | 10 | 0.8607 |
| 16. | Friedman\_mse | Log 2 | 100 | 0.8704 |

The Random Forest Regression has best R2 score of : 0.8704

Conclusion:

Out of all the algorithms in Machine Learning, the best regression model which is having the best R2 Score is : Random Forest with score of 0.8704