ASSIGNMENT\_REGRESSION

1.Machine Learning

Supervised Learning- stage 2

Regression- stage 3

2.1338rows X6 columns

3.sex and smoker is ordinal data (label encoding) so coverting to number\_nominal data(one hot encoding)

4.The final models is random forest (n\_estimators=1000,criterion=mae,max\_feature=log2,R2 value=0.8736)

The following to machine learning regression to find R2 value

1.Multiple Linear Regression- R2= 0.7894

***2. Support Vector Machine(SVM):***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| S.NO | STANDRDISATION | HYPER  PARAMETER | LINEAR  (R2 VALUE) | RBF(NON-  LINEAR)  (R2 VALUE) | POLY  (R2 VALUE) | SIGMOID  (R2 VALUE) |
| 1 | Not use | Not use | -0.0111 | -0.0894 |  | -0.0897 |
| 2 | Not use | C=0.01 | -0.0797 | -0.0897 |  | -0.0897 |
| 3 | Not use | C=0.10 | -0.1220 | -0.0896 |  | -0.0897 |
| 4 | Not use | C=0.100 | -0.1220 | -0.0896 |  | -0.0897 |
| 5 | Not use | C=0.1000 | -0.1220 | -0.0896 |  | -0.0897 |
| 6 | Not use | C=1 | -0.111 | -0.0896 |  | -0.0897 |
| 7 | Not use | C=10 | -0.0016 | -0.0874 | It’stake more time | -0.0897 |
| 8 | Not use | C=100 | 0.5432 | -0.0745 |  | -0.0897 |
| 9 | Not use | C=1000 | 0.6340 | -0.0400 |  | -0.0897 |
| 10 | Use | Not use | -0.0103 | -0.0833 | -0.0756 | -0.0754 |
| 11 | Use | C=0.01 | -0.0887 | -0.0896 | -0.0895 | -0.0895 |
| 12 | Use | C=0.10 | -0.0803 | -0.0890 | -0.0883 | -0.0882 |
| 13 | Use | C=0.100 | -0.0803 | -0.0890 | -0.0883 | -0.0882 |
| 14 | Use | C=0.1000 | -0.0803 | -0.0890 | -0.0883 | -0.0882 |
| 15 | Use | C=1 | -0.010 | -0.0833 | -0.0756 | -0.0754 |
| 16 | Use | C=10 | 0.4624 | -0.0322 | 0.0387 | 0.0393 |
| 17 | Use | C=100 | 0.6288 | 0.3200 | 0.6179 | 0.5276 |
| 18 | Use | C=1000 | 0.7649 | 0.8102 | 0.8566 | 0.2874 |

* ***The SVM method use satandarstiond and using hyper paremeter C =1000 compare to multiple linear regression the best model***

***3.Decision Tree:***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *S.NO* | CRITERION | MORE FEATURES | SPLITTEER | R2 VALUE |
| 1 | ***Friedman\_mse*** | ***auto*** | ***best*** | ***0.7043*** |
| ***2*** | ***Friedman\_mse*** | ***sqrt*** | ***best*** | ***0.7319*** |
| ***3*** | ***Friedman\_mse*** | ***Log2*** | ***best*** | ***0.7533*** |
| ***4*** | ***Friedman\_mse*** | ***auto*** | ***random*** | ***0.6997*** |
| ***5*** | ***Friedman\_mse*** | ***sqrt*** | ***random*** | ***0.6202*** |
| ***6*** | ***Friedman\_mse*** | ***Log2*** | ***random*** | ***0.6522*** |
| ***7*** | ***Mae*** | ***auto*** | ***best*** | ***0.6514*** |
| ***8*** | ***Mae*** | ***sqrt*** | ***best*** | ***0.6993*** |
| ***9*** | ***Mae*** | ***Log2*** | ***best*** | ***0.7170*** |
| ***10*** | ***Mae*** | ***auto*** | ***random*** | ***0.6985*** |
| ***11*** | ***Mae*** | ***sqrt*** | ***random*** | ***0.4183*** |
| ***12*** | ***Mae*** | ***Log2*** | ***random*** | ***0.7078*** |
| ***13*** | ***Mse*** | ***auto*** | ***best*** | ***0.6855*** |
| ***14*** | ***Mse*** | ***sqrt*** | ***best*** | ***0.7029*** |
| ***15*** | ***Mse*** | ***Log2*** | ***best*** | ***0.7243*** |
| ***16*** | ***Mse*** | ***auto*** | ***random*** | ***0.6806*** |
| ***17*** | ***Mse*** | ***sqrt*** | ***random*** | ***0.7530*** |
| ***18*** | ***mse*** | ***Log2*** | ***random*** | ***0.6714*** |

* **The Decision Tree compare to Multiple Linear Regression and SVM not good model.**

***4.Random Forest:***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *S.NO* | *N\_ESTIMATORS* | *CRITERION* | *MAX\_FEATURE* | *R2 VALUE* |
| ***1*** | ***10*** | ***Mae*** | ***auto*** | ***0.8351*** |
| ***2*** | ***50*** | ***Mae*** | ***auto*** | ***0.8480*** |
| ***3*** | ***100*** | ***Mae*** | ***auto*** | ***0.8532*** |
| ***4*** | ***1000*** | ***Mae*** | ***Auto*** | ***0.8534*** |
| ***5*** | ***10*** | ***Mae*** | ***Sqrt*** | ***0.8522*** |
| ***6*** | ***50*** | ***Mae*** | ***Sqrt*** | ***0.8670*** |
| ***7*** | ***100*** | ***Mae*** | ***Sqrt*** | ***0.8704*** |
| ***8*** | ***1000*** | ***Mae*** | ***Sqrt*** | ***0.8742*** |
| ***9*** | ***10*** | ***Mae*** | ***log2*** | ***0.8528*** |
| ***10*** | ***50*** | ***Mae*** | ***log2*** | ***0.8676*** |
| ***11*** | ***100*** | ***Mae*** | ***log2*** | ***0.8716*** |
| ***12*** | ***1000*** | ***Mae*** | ***log2*** | ***0.8736*** |
| ***13*** | ***10*** | ***mse*** | ***auto*** | ***0.8385*** |
| ***14*** | ***50*** | ***Mse*** | ***auto*** | ***0.8567*** |
| ***15*** | ***100*** | ***Mse*** | ***auto*** | ***0.8535*** |
| ***16*** | ***1000*** | ***Mse*** | ***Auto*** | ***0.8547*** |
| ***17*** | ***10*** | ***Mse*** | ***Sqrt*** | ***0.8547*** |
| ***18*** | ***50*** | ***Mse*** | ***Sqrt*** | ***0.8525*** |
| ***19*** | ***100*** | ***Mse*** | ***Sqrt*** | ***0.8689*** |
| ***20*** | ***1000*** | ***mse*** | ***Sqrt*** | ***0.8683*** |
| ***21*** | ***10*** | ***Mse*** | ***log2*** | ***0.8716*** |
| ***22*** | ***50*** | ***Mse*** | ***log2*** | ***0.8671*** |
| ***23*** | ***100*** | ***Mse*** | ***log2*** | ***0.8721*** |
| ***24*** | ***1000*** | ***mse*** | ***log2*** | ***0.8722*** |

* Random Forest is good model(n\_estimators=1000,criterion=mae,max\_feature=log2,R2 value =0.8736)

6. The final models is random forest. I think compare to multiple linear Regression,SVM and Decision TreeR2 value is high. So that dataset best model is Random Forest.