**Problem Statement:**

Client wants to predict the insurance charges based on age, sex, bmi, number of children, smoking habit.

**3 Stages of Problem Identification:**

**Stage 1:** Machine learning because output is a number.

**Stage 2:** Supervised learning because we have both input and expected output.

**Stage 3:** Regression because output is a number.

**Information about Dataset:**

**Total Rows:** 1338 rows

**Input/Independent Variables:** Age, Sex, BMI, Number of Children, is person is smoker or not

**Output/Dependent Variables:** Charges

**Data Processing:**

Both sex and smoker are nominal data. It will be converted to numbers using the getdummies(One hot encoding)

**Algorithm for AI Model:**

We can choose **Random Forest Regression Algorithm with parameters (criterian='squared\_error', max\_features='sqrt', n\_estimators=100)** as a final model because it gave us a high R2 square compare to other algorithms.

|  |  |  |
| --- | --- | --- |
| **Sl. No** | **Algorithm** | **R2 Value** |
| 1 | Multilinear Regression | 0.7894790349867009 |
| 2 | SVM Regression | 0.8566487675946524 |
| 3 | Decision Tree Regression | 0.7877784613438091 |
| **4** | **Random Forest Regression** | **0.8754792427285** |

Random Forest Regression Hyper Tuning Comparison:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl.No** | **criterian** | **max\_features** | **n\_estimators** | **R2 Value** |
| 1 | squared\_error | None | 50 | 0.8498218449086588 |
| 2 | squared\_error | sqrt | 50 | 0.8697458769868834 |
| 3 | squared\_error | log2 | 50 | 0.8674882706475169 |
| 4 | squared\_error | None | 100 | 0.8543197145320003 |
| **5** | **squared\_error** | **sqrt** | **100** | **0.8754792427285** |
| 6 | squared\_error | log2 | 100 | 0.8669859796236599 |
| 7 | friedman\_mse | None | 50 | 0.8517012400490362 |
| 8 | friedman\_mse | sqrt | 50 | 0.8729832764811812 |
| 9 | friedman\_mse | log2 | 50 | 0.8725556931357203 |
| 10 | friedman\_mse | None | 100 | 0.8511796986548201 |
| 11 | friedman\_mse | sqrt | 100 | 0.8739785501489924 |
| 12 | friedman\_mse | log2 | 100 | 0.8723162276489183 |
| 13 | absolute\_error | None | 50 | 0.8523417836135914 |
| 14 | absolute\_error | sqrt | 50 | 0.8686897045218873 |
| 15 | absolute\_error | log2 | 50 | 0.8729923433939877 |
| 16 | absolute\_error | None | 100 | 0.8590392692496325 |
| 17 | absolute\_error | sqrt | 100 | 0.8734216062626972 |
| 18 | absolute\_error | log2 | 100 | 0.8735976535627025 |
| 19 | poisson | None | 50 | 0.8553641376891776 |
| 20 | poisson | sqrt | 50 | 0.870435333447934 |
| 21 | poisson | log2 | 50 | 0.8676170122190497 |
| 22 | poisson | None | 100 | 0.8535376481097332 |
| 23 | poisson | sqrt | 100 | 0.8709688721488164 |
| 24 | poisson | log2 | 100 | 0.8718279501393471 |

**Decision Tree Regression Hyper Tuning Comparison:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl.No** | **criterian** | **max\_features** | **splitters** | **R2 Value** |
| 1 | squared\_error | None | best | 0.6953944519970485 |
| 2 | squared\_error | sqrt | Best | 0.6973372242090019 |
| 3 | squared\_error | log2 | Best | 0.7205569709940924 |
| 4 | squared\_error | None | random | 0.7005822746047742 |
| 5 | squared\_error | sqrt | random | 0.6241015952496587 |
| 6 | squared\_error | log2 | random | 0.6885120529943847 |
| 7 | friedman\_mse | None | Best | 0.7178475705913336 |
| 8 | friedman\_mse | sqrt | Best | 0.6837625761239257 |
| 9 | friedman\_mse | log2 | Best | 0.6565289131481045 |
| 10 | friedman\_mse | None | random | 0.708060050460155 |
| 11 | friedman\_mse | sqrt | random | 0.6118624380527946 |
| 12 | friedman\_mse | log2 | random | 0.6735547459377325 |
| 13 | absolute\_error | None | Best | 0.6793580166481716 |
| 14 | absolute\_error | sqrt | Best | 0.6878075093136662 |
| 15 | absolute\_error | log2 | Best | 0.7877784613438091 |
| 16 | absolute\_error | None | random | 0.7166377335218996 |
| 17 | absolute\_error | Sqrt | random | 0.7276948483801368 |
| 18 | absolute\_error | log2 | random | 0.7503446496737405 |
| 19 | poisson | None | Best | 0.7266109544321601 |
| 20 | poisson | Sqrt | Best | 0.691265815832401 |
| 21 | poisson | log2 | Best | 0.7111569960072706 |
| 22 | poisson | None | random | 0.7135039223691658 |
| 23 | poisson | Sqrt | random | 0.685239378587961 |
| 24 | poisson | log2 | random | 0.6691990979964494 |

**SVM Regression Hyper Tuning Comparison:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sl. No** | **kernel** | **gamma** | **C** | **R2 Value** |
| 1 | linear | Scale | 10 | 0.462468414233968 |
| 2 | linear | Scale | 1000 | 0.7649311738649672 |
| 3 | linear | Auto | 10 | 0.462468414233968 |
| 4 | linear | Auto | 1000 | 0.7649311738649672 |
| 5 | poly | Scale | 10 | 0.038716222760231456 |
| **6** | **poly** | **Scale** | **1000** | **0.8566487675946524** |
| 7 | poly | Auto | 10 | 0.038716222760231456 |
| **8** | **poly** | **Auto** | **1000** | **0.8566487675946524** |
| 9 | rbf | Scale | 10 | -0.03227329390671052 |
| 10 | rbf | Scale | 1000 | 0.8102064851758545 |
| 11 | rbf | Auto | 10 | -0.03227329390671052 |
| 12 | rbf | Auto | 1000 | 0.8102064851758545 |
| 13 | sigmoid | Scale | 10 | 0.03930714378274347 |
| 14 | sigmoid | Scale | 1000 | 0.28747069486978516 |
| 15 | sigmoid | auto | 10 | 0.03930714378274347 |
| 16 | sigmoid | auto | 1000 | 0.28747069486978516 |

**Multi Linear Regression: 0.7894790349867009**