**Insurance Charges Prediction Using Machine Learning Regression Models**

**Dataset:**

This dataset contains information about individuals and their medical insurance charges. It is used to predict insurance charges based on features like age, sex, BMI, children, smoking status, etc.,

**Dataset dimensions**:

Total Records (Rows): 1338

Total Features (Columns): 6

**Pre-Processing methods:**

To prepare the dataset for machine learning regression models, the following pre-processing steps were performed.

**Handling Categorical (Nominal) Data:**

Two columns in the dataset were categorical and needed to be converted to numeric values:

|  |  |  |  |
| --- | --- | --- | --- |
| **Column** | **Type** | **Encoding Applied** | **details** |
| Sex | Nominal | One-Hot Encoding (drop\_first = True) | One-Hot Encoding (drop\_first = True) |
| smoker | **Nominal** | One-Hot Encoding (drop\_first = True) | Converted to smoker\_yes (1 = smoker, 0 = nonsmoker) |

1. **Multiple Linear Regression (R2 Value)=** 0.7894
2. **Support Vector machine**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S. No** | **HYPER PARAMETER** | **LINEAR (R VALUE)** | **RBF** | **POLY** | **SIGMOID** |
| **1** | **C=10** | -0.0016 | -0.0819 | -0.0931 | -0.0907 |
| **2** | **C=100** | 0.5432 | -0.1248 | -0.0997 | -0.1181 |
| **3** | **C=500** | 0.6270 | -0.1246 | -0.0820 | -0.4562 |
| **4** | **C=1000** | 0.6340 | -0.1174 | -0.0555 | -1.6659 |
| **5** | **C=3000** | 0.7590 | -0.0962 | 0.0489 | -12.0190 |

1. **DECISION TREE**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.No** | **CRITERION** | **MAX FEATURES** | **SPLITTER** | **R2 VALUE** |
| **1** | **Squared\_error** | **None** | **best** | 0.6883 |
| **2** | **Squared\_error** | **Sqrt** | **Best** | 0.7099 |
| **3** | **Squared\_error** | **Log2** | **Best** | 0.6263 |
| **4** | **Squared\_error** | **max\_depth=2** | **Best** | 0.8553 |
| **5** | **Squared\_error** | **Max\_features=2** | **best** | 0.7069 |
| **6** | **Squared\_error** | **None** | **random** | 0.7297 |
| **7** | **Squared\_error** | **sqrt** | **Random** | 0.6885 |
| **8** | **Squared\_error** | **Log2** | **Random** | 0.6724 |
| **9** | **Squared\_error** | **max\_depth=2** | **Random** | 0.6995 |
| **10** | **Squared\_error** | **Max\_features=2** | **random** | 0.5844 |
| **11** | **friedman\_mse** | **None** | **Best** | 0.6858 |
| **12** | **friedman\_mse** | **Sqrt** | **Best** | 0.7437 |
| **13** | **friedman\_mse** | **Log2** | **Best** | 0.7190 |
| **14** | **friedman\_mse** | **Max\_features=2** | **Best** | 0.6477 |
| **15** | **friedman\_mse** | **Max\_deapth=2** | **Best** | 0.8569 |
| **16** | **friedman\_mse** | **None** | **Random** | 0.7095 |
| **17** | **friedman\_mse** | **Sqrt** | **Random** | 0.6959 |
| **18** | **friedman\_mse** | **Log2** | **Random** | 0.6657 |
| **19** | **friedman\_mse** | **Max\_depth=2** | **Random** | 0.7355 |
| **20** | **friedman\_mse** | **Max\_features=2** | **Random** | 0.6612 |

**­**

1. **Random Forest**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **S.NO** | **CRITERION** | **MAX FEATURES** | **N\_ESTIMATORS** | **R2 SCORE** |
| **1** | **Squared\_error** | None | **100** | 0.8552 |
| **2** | **Squared\_error** | sqrt | **100** | 0.8722 |
| **3** | **Squared\_error** | Log2 | **100** | 0.8683 |
| **4** | **Squared\_error** | Max\_depth=2 | **100** | 0.8661 |
| **5** | **Squared\_error** | max\_features=2 | **100** | 0.8718 |
| **6** | **Squared\_error** | None | **10** | 0.8367 |
| **7** | **Squared\_error** | sqrt | **10** | 0.8461 |
| **8** | **Squared\_error** | Log2 | **10** | 0.8552 |
| **9** | **Squared\_error** | Max\_depth=2 | **10** | 0.8673 |
| **10** | **Squared\_error** | max\_features=2 | **10** | 0.8592 |
| **11** | Friedman-mse | None | **10** | 0.8466 |
| **12** | Friedman-mse | sqrt | **10** | 0.8530 |
| **13** | Friedman-mse | Log2 | **10** | 0.8588 |
| **14** | Friedman-mse | Max\_depth=2 | **10** | 0.8642 |
| **15** | Friedman-mse | max\_features=2 | **10** | 0.8546 |
| **16** | Friedman-mse | None | **100** | 0.8574 |
| **17** | Friedman-mse | sqrt | **100** | 0.8687 |
| **18** | Friedman-mse | Log2 | **100** | 0.8707 |
| **19** | Friedman-mse | Max\_depth=2 | **100** | 0.8651 |
| **20** | Friedman-mse | max\_features=2 | **100** | 0.8706 |
| **21** | Absolute\_error | None | **10** | 0.8545 |
| **22** | Absolute\_error | sqrt | **10** | 0.8408 |
| **23** | Absolute\_error | Log2 | **10** | 0.8567 |
| **24** | Absolute\_error | Max\_depth=2 | **10** | 0.8583 |
| **25** | Absolute\_error | max\_features=2 | **10** | 0.8556 |
| **26** | Absolute\_error | None | **100** | 0.8595 |
| **27** | Absolute\_error | sqrt | **100** | 0.8750 |
| **28** | Absolute\_error | Log2 | **100** | 0.8726 |
| **29** | Absolute\_error | Max\_depth=2 | **100** | 0.8575 |
| **30** | Absolute\_error | max\_features=2 | **100** | 0.8721 |

**Final Model Selection:**

Among all the regression models developed, including Multiple Linear Regression, Support Vector Machine (SVM), and Decision Tree Regressor – the Random Forest Regressor consistently outperformed the others in terms of R² Score, which measures the proportion of variance in the target variable(charges).

**Reason for selecting Random Forest:**

• Highest R² Score (0. 0.8750) – this indicates the best predictive performance among all tested models.

• It aggregates multiple decision trees, reducing variance and overfitting, especially effective on datasets with nonlinear relationships.

• Performs well even with noisy or moderately imbalanced data.

• Hyper tunning parameters (n\_estimators, criterion, max\_features) allowed the model to generalize well on unseen data.

**Final Model:**

Random Forest Regressor (R² Score: 0. 0.8750) The Random Forest Regressor with absolute error loss and optimized parameters chosen as the final model due to its high predictive accuracy and robust performance in handling the insurance dataset.