**Problem Statement: Predicting Insurance Charges**

Design a predictive model to estimate health insurance charges for individuals based on various factors such as age, gender, BMI (Body Mass Index), number of children and smoking habits.

**Objective:**

The goal is to build a machine learning model that can accurately predict health insurance charges for new individuals based on their Personal habits. Finally, the model should be evaluated using appropriate metrics such as mean absolute error or root mean squared error.

**Dataset:**

The dataset contains a bunch of information(1339 rows and 6 columns) about different people, like how old they are, if they smoke, their weight, their health insurance costs etc.

* Age: Contributors age
* Sex: Male or Female
* Bmi: Body mass index value
* Children: No of Children
* Smoker: Do they smoke or not
* Charges: Insurance cost

**Tasks:**

Model Selection: Experiment with various machine learning algorithms such as linear regression, support vector machine, decision trees, random forest.

Model Training: Train the selected models using the training dataset.

Model Tuning: The process of adjusting hyperparameters to improve model performance.

Model Evaluation: and evaluate their performance using suitable metrics like mean squared error or R-squared.

Deployment: Deploy the model in a user-friendly interface or integrate it into existing real estate platforms for practical use.

**3 -Stage of Problem Identification**

**Stage 1: Domain selection**

All the inputs/variables are numeric except 2 columns (Sex, Smokers). Using one hot encoding convert categorical data to a vector that contains 1’s and 0’s indicating the presence or absence of a feature and the domain we are going to use is Machine Learning Algorithm.

**Stage 2: Learning Selection**

Our aim is to predict an outcome based on input features; we know the labelled data so we would use supervised learning.

**Stage 3: Supervised- (Regression or classification)**

Our goal is to predict the insurance charge that is continuous value then using regression algorithm would be the appropriate choice.

**R-squared value of different Machine learning models**

**R-squared value of Multi Linear Regression=** 0.78947

**R-squared value of Support vector Machine using different parameters.**

.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.no** | **Regularization parameter** | **Linear**  **R2 value** | **Rbf**  **R2 value** | **Poly**  **R2 value** | **Sigmoid**  **R2 value** |
| 1. | C=1.0 | -1.6494 | -0.09461 | -0.15765 | -0.09100 |
| 2. | C=10 | -0.9494 | -0.14513 | -0.5319 | -0.10185 |
| 3. | C=100 | -0.05311 | -0.4681 | -0.6172 | -0.2835 |
| 4. | C=500 | -0.09621 | -0.4791 | -0.6281 | -3.35284 |
| 5. | C=1000 | -0.18584 | -0.5119 | -0.6298 | -19.745 |

In the SVM Model, the negative R\_Squared Value suggest that indicates the model does not capture the variability within the data. In such a case it’s essential to reevaluate the feature selection and data quality to rectify the model’s shortcomings

**R-squared value of Decision tree algorithm using different parameters.**

|  |  |  |
| --- | --- | --- |
| **criterion** | **splitter** | **R2 Value** |
| friedman\_mse | random | 0.6407 |
| absolute\_error | random | 0.75978 |
| poisson | random | 0.67831 |
| friedman\_mse | best | 0.69027 |
| absolute\_error | best | 0.66621 |
| poisson | best | 0.72544 |
| squared\_error | random | 0.70151 |

**R-squared value of Random Forest algorithm using different parameters**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **n\_estimators** | **min\_samples\_leaf** | **min\_samples\_split** | **random\_state** | **R2 Value** |
| 100 | 3 | 8 | 42 | 0.88232 |
| 200 | 4 | 10 | 42 | 0.88579 |
| 300 | 5 | 12 | 42 | 0.88733 |
| 1000 | 7 | 18 | 42 | 0.88853 |

**Results:**

After the evaluation and comparison of various machine learning models, the Random Forest model yielded the best performance in terms of predictive accuracy, when compared to the other models. The model achieved a R-squared of 88% . This indicates that the model can effectively predict insurance charges based on the given features.