**Project Report: Medical Insurance Cost Prediction**

**Using Machine Learning**

### **1. Introduction**

This project focuses on predicting the cost of medical insurance using machine learning. The dataset includes demographic and lifestyle-related information such as age, sex, BMI, number of children, smoking status, region, and the insurance charges. The model used for this regression problem is Linear Regression.

### **2. Dataset Information**

* **Source:** insurance.csv
* **Features:**
  + age: Age of the individual
  + sex: Gender (male/female)
  + bmi: Body Mass Index
  + children: Number of children
  + smoker: Smoking status (yes/no)
  + region: Region of residence in the US
  + charges: Medical insurance cost (target variable)

**Dataset Overview:**

insurance\_dataset.head()  
insurance\_dataset.info()  
insurance\_dataset.describe()  
insurance\_dataset.shape  
insurance\_dataset.isnull().sum()

* The dataset contains no missing values.
* Data types are appropriate (categorical and numerical).

### **3. Data Visualization and Exploration**

* **Age Distribution:**

sns.distplot(insurance\_dataset['age'])

* **Gender Count:**

sns.countplot(x='sex', data=insurance\_dataset)

* **BMI Distribution:**

sns.distplot(insurance\_dataset['bmi'])

* **Children Count:**

sns.countplot(x='children', data=insurance\_dataset)

* **Smoker Count:**

sns.countplot(x='smoker', data=insurance\_dataset)

* **Region Count:**

sns.countplot(x='region', data=insurance\_dataset)

* **Charges Distribution:**

sns.distplot(insurance\_dataset['charges'])

These plots help understand the distribution and imbalance in the data.

### **4. Data Preparation**

* **Splitting Features and Target:**

X = insurance\_dataset.drop(columns='charges', axis=1)  
Y = insurance\_dataset['charges']

* **Train-Test Split:**

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.2, random\_state=2)

### **5. Model Training**

* **Linear Regression Model:**

from sklearn.linear\_model import LinearRegression  
regressor = LinearRegression()  
regressor.fit(X\_train, Y\_train)

### **6. Model Evaluation**

* **On Training Data:**

training\_data\_prediction = regressor.predict(X\_train)  
r2\_train = metrics.r2\_score(Y\_train, training\_data\_prediction)

* **On Testing Data:**

test\_data\_prediction = regressor.predict(X\_test)  
r2\_test = metrics.r2\_score(Y\_test, test\_data\_prediction)

**Results:**

* R-squared on training data: Reported via r2\_train
* R-squared on test data: Reported via r2\_test

These scores indicate how well the model fits the data.

### **7. Prediction on New Data**

An example prediction was made using the following input:

input\_data = (31, 1, 25.74, 0, 1, 0) # Age, Sex, BMI, Children, Smoker, Region

Steps:

* Converted to NumPy array
* Reshaped to 2D format
* Predicted using the trained model

**Output:**

The insurance cost is USD <predicted\_value>

### **8. Conclusion**

This project successfully used a linear regression model to estimate medical insurance costs based on personal and demographic factors. Despite being a simple model, Linear Regression provided a good baseline and understandable insights into the cost-driving features such as age, BMI, and smoking status.