InsureCast

🎯 Objective: Predict insurance charges using machine learning

# 1️⃣ Problem Statement

A client wants to predict insurance charges based on several personal and lifestyle parameters.  
As a data scientist, your goal is to:  
- Build a machine learning model  
- Ensure high prediction accuracy using the R² score  
- Choose and justify the best performing model

# 2️⃣ Problem Type Breakdown

|  |  |
| --- | --- |
| Stage | Explanation |
| Stage 1 | Domain: Machine Learning (input = numbers) |
| Stage 2 | Learning Type: Supervised Learning (input + output available) |
| Stage 3 | Output: Numeric → Regression Problem |

# 3️⃣ Dataset Information

- Total Rows: 1338  
- Total Columns: 6  
- Target Variable: charges  
- Features Include: age, sex, bmi, children, smoker, region

# 4️⃣ Data Preprocessing

🔄 One-Hot Encoding for Categorical Columns

|  |  |  |
| --- | --- | --- |
| Categorical Column | Values | Action |
| sex | male, female | → sex\_male, sex\_female |
| smoker | yes, no | → smoker\_yes, smoker\_no |

Most ML algorithms need numerical input, so we convert these using One-Hot Encoding to avoid bias or incorrect assumptions.

# 5️⃣ Model Building Workflow

Step 1: Load and explore the dataset  
Step 2: Preprocess data (e.g., one-hot encode, scaling if needed)  
Step 3: Split into features (X) and target (y)  
Step 4: Train/test split  
Step 5: Train various regression models  
Step 6: Evaluate each using R² score  
Step 7: Choose and save the best model

🧠 Code Repository:  
🔗 GitHub - Assignment\_Regression: <https://github.com/Marudhanayagam4/Assignment_Regression>

R² Score Values:

Support Vector Machine:

Before Standardization:



**After Standardization:**



**Decision Tree**



**Random Forest**



# 6️⃣ Model Evaluation (R² Score)

|  |  |
| --- | --- |
| Algorithm | R² Score |
| Simple Linear Regression | 0.78 |
| Multiple Linear Regression | 0.78 |
| Support Vector Machine | 0.85 |
| Decision Tree | 0.76 |
| Random Forest | 0.87 ✅ |

# 7️⃣ Best Model: Random Forest Regression

✅ Why Random Forest?

- Highest R² Score (0.87) – Predicts most accurately  
- Stable – Uses ensemble of trees, reducing overfitting  
- Handles Non-Linearity – Captures complex patterns well  
- Robust – Lower variance compared to a single decision tree

🔚 Final Conclusion:

**Random Forest** is the most suitable model for this problem.  
It explains **87% of the variability in insurance charges**, offering high accuracy and reliability.