# Task Report: Medical Insurance Cost Prediction

## 1. Overview

The main goal of this analysis is to predict individual medical costs billed by health insurance based on various factors such as age, sex, BMI, number of children, smoking status, and region. The dataset is derived from the book *Machine Learning with R* by Brett Lantz.

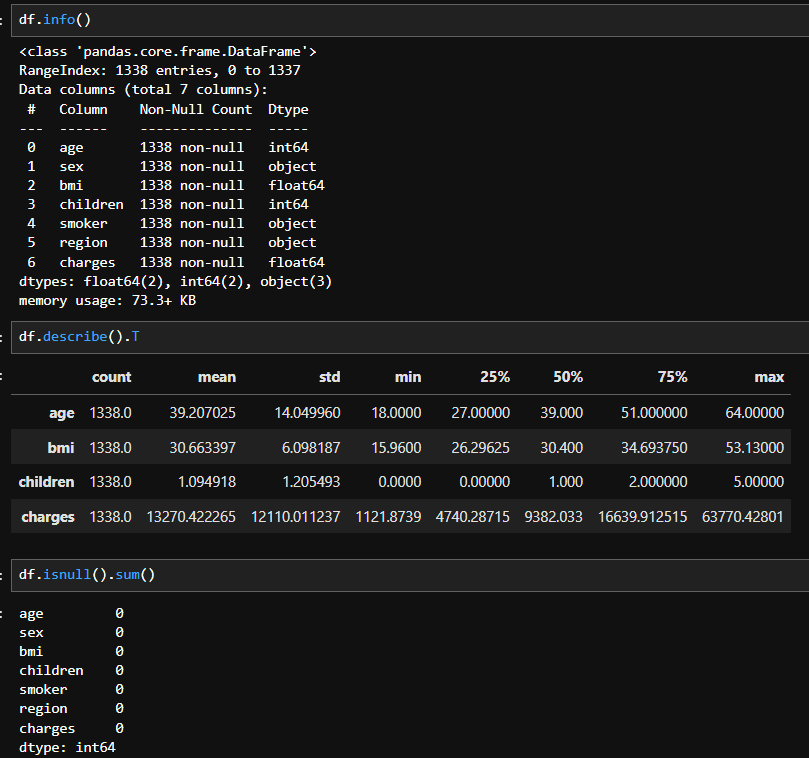
### Key Objectives:

* Perform exploratory data analysis (EDA).
* Build and evaluate multiple regression models to predict insurance charges.
* Compare model performance and select the best-performing model.
* Fine-tune the best-performing model.

## 2. Exploratory Data Analysis (EDA)

### Dataset Information:

* **Dataset Size:** 1338 rows and 7 columns.
* **Features:** Age, sex, BMI, number of children, smoker status, region, and charges.



* **No missing data**: The dataset is complete with no null values.

### **Target Variable:** charges

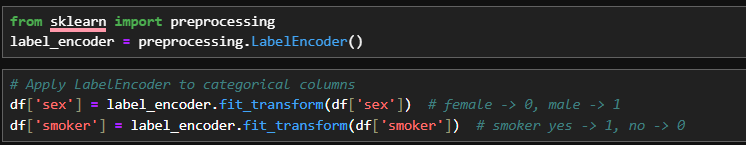
The target variable charges represent the medical costs billed by health insurance. The variable is continuous, and the distribution is right-skewed, which is common for cost-related data.

A screenshot of a computer screen

Description automatically generated

### Categorical Variables

Several categorical variables, such as sex, smoker, and region, were encoded to numerical values for modeling purposes.



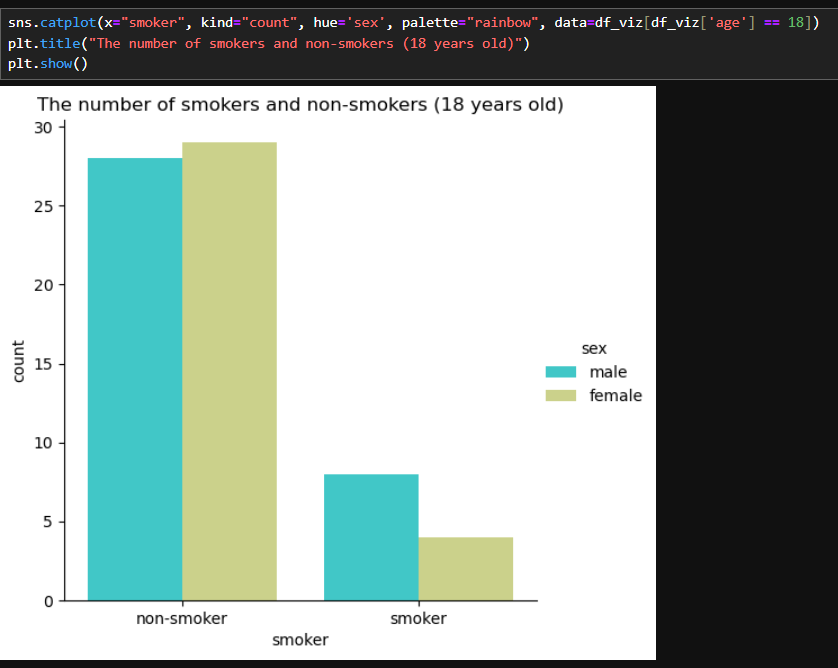
### Categorical Variable Counts:

* + sex: Male: 51%, Female: 49%
  + smoker: Yes: 20%, No: 80%
  + region: Distributed across four regions in the US (Northeast, Southeast, Northwest, Southwest).

## 3. Visualizations

### Relationship Between Smoking and Medical Charges

* **Insight:** Smoking significantly increases medical charges. Even at a young age (18 years), smokers incur higher medical costs than non-smokers.



A screenshot of a computer screen

Description automatically generated

### Medical Charges by Region

* **Insight:** The highest medical charges are observed in the Southeast, while the lowest charges are in the Southwest.

A graph of smokers and smokers

Description automatically generated

## 4. Model Building

### Data Preparation

* **Feature Selection:** The dataset was split into features (X) and the target variable (y). The categorical columns such as sex, smoker, and region were label-encoded for model training.

A screen shot of a computer

Description automatically generated

### Model Selection

Several regression models were tested, including:

* **Linear Regression (LR)**
* **Random Forest Regressor (RF)**
* **Decision Tree Regressor (DT)**
* **Gradient Boosting Regressor (GBR)**
* **K-Neighbors Regressor (KNN)**
* **Support Vector Regressor (SVR)**

### Model Comparison

For each model, training and testing accuracies were evaluated using R-squared, Mean Squared Error (MSE), and Mean Absolute Error (MAE).

A computer screen shot of a program

Description automatically generated

| **Model** | **Training R-squared** | **Testing R-squared** | **Mean Squared Error (MSE)** | **Mean Absolute Error (MAE)** |
| --- | --- | --- | --- | --- |
| Linear Regression | 0.7417 | 0.7833 | 33,635,210.43 | 4,186.51 |
| Random Forest | 0.9278 | 0.8768 | 19,130,167.40 | 2,428.20 |
| Decision Tree | 0.9983 | 0.7433 | 39,858,781.72 | 2,898.69 |
| Gradient Boosting | 0.9928 | 0.8419 | 24,543,702.50 | 2,609.51 |
| K-Nearest Neighbors | 0.3938 | 0.1445 | 132,814,646.70 | 7,953.21 |
| Support Vector Reg. | -0.0977 | -0.0723 | 166,474,492.54 | 8,592.79 |

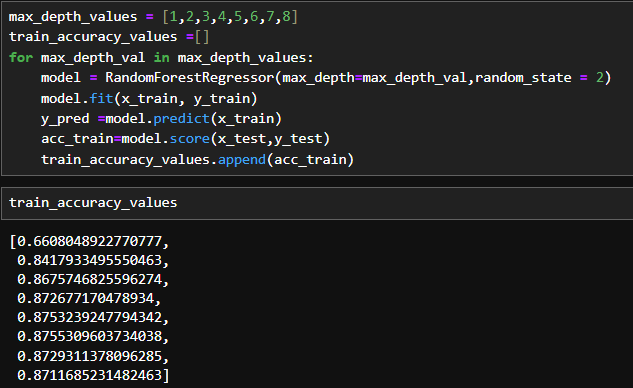
### Best Model: Random Forest Regressor

* The **Random Forest Regressor** outperformed all other models with a testing R-squared score of 0.8768 and the lowest Mean Squared Error (19,130,167.40).

## 5. Model Fine-Tuning

### Hyperparameter Tuning: Max Depth

The performance of the Random Forest model was optimized by tuning the maximum depth of the trees. Various depths from 1 to 8 were tested to identify the optimal depth.



### Final Model Evaluation

The final Random Forest model with a depth of 6 was selected. The R-squared score on the test set was 0.872, and the MSE was 19845744.977, indicating strong predictive power.

A computer screen shot of a program

Description automatically generated

## 6. Visualization of Model Predictions

A scatter plot of predicted values against residuals (actual - predicted) shows that the model's residuals are randomly scattered, indicating a good fit without major bias.

A screen shot of a computer screen

Description automatically generated

## 7. Conclusion

* **Best Model:** The **Random Forest Regressor** performed the best with the highest R-squared value and lowest MSE.
* **Smoker Impact:** Smoking significantly increases medical costs, even at younger ages.
* **Regional Impact:** The Southeast region has the highest medical costs, while the Southwest has the lowest.