Healthcare Case Study

Business/Domain Understanding

What is Insurance?
Insurance is a contract between two parties whereby one party agrees to undertake the risk of the other in exchange for consideration known as premium and promises to indemnify the party on the happening of an uncertain event.

What is health insurance?
A plan that covers or shares the expenses associated with health care can be described as health insurance.

Health insurance in India is an emerging insurance sector after the term life insurance and automobile insurance sector. Rise in the middle class, higher hospitalization cost, expensive health care, digitization and increase in awareness level are some important drivers for the growth of the health insurance market in India.

Assumption
Assume that **you are working as a Data Scientist** with one of the world's leading insurance providers (like UnitedHealth Group).

Insurance companies need to set the insurance premiums following the population trends despite having limited information about the insured population if they have to put themselves in a position to make profits. This makes it necessary to estimate the average medical care expenses based on trends in the population segments.

Medical Cost Dataset

Dataset

Click here to download the dataset.

Columns

- age: age of primary beneficiary
- sex: insurance contractor gender, female, male
- bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- children: Number of dependents
- smoker: Smoking
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

SPRINT 1 - Exploratory Data Analysis

Task - This is an open ended question. Kindly apply all your knowledge to perform an exploratory data analysis on the given dataset. It is known that the target variable is **Charges**.

Write proper conclusions and provide recommendations to the telecom company based on the insights.

References -
https://www.kaggle.com/code/prathameshgadekar/world-population-eda-with-world-map-visualization

SPRINT 2 - Data Preparation and Model Building

Problem Statement - The aim here will be to predict the medical costs billed by health insurance on an individual given some features about the individual in the dataset.

Steps to be followed

Step - 1: Load the data and perform the basic EDA to understand the data.

Step - 2: Document the below mentioned points properly:

- Identify the input and output/target variables.
- Identify the type of ML Task.
- Identify the Evaluation Metric.
    - For regression task - Mean Absolute Error
    - For classification task - Accuracy

Step - 3: Split the dataset into Training and Testing (recommended 75:25 split).

Step - 4: Data preparation on train data:

- For Numerical Variables - Standardization or Normalization (Fit and Transform)
- For Categorical - LabelEncoding or OneHotEncoding (Choose wisely)

Step - 5: Data preparation on test data:

- For Numerical Variables - Standardization (Transform)
- For Categorical - LabelEncoding or OneHotEncoding (Choose wisely)

Step - 6: Model Training Phase - Use all the algorithms mentioned below to train separate models:

- KNN
- Logistic Regression / Linear Regression
- Support Vector Machines
- Decision Trees
- Random Forest

Step - 7: Predict and evaluate each model separately using the correct evaluation metric.

Step - 8: Display a plot which shows all the algorithms applied along with the scores achieved. **Write your conclusion on the best algorithm for the Medical Cost Prediction problem**.