# Case Study - Healthcare Domain

## Business/Domain Understanding

### Context

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.

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.

### 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.

# Medical Charges Dataset

**Domain - Healthcare**
**Dataset -** Click here to download the dataset.

**Data Dictionary**

- 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: If the individual smokes or not
- region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- charges: Individual medical costs billed by health insurance

# SPRINT 1 - EDA

## Task - Exploratory Data Analysis

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

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**.

However, you are mandatorily supposed to solve the below mentioned EDA Task for your presentation:

1. Explore the data distribution of each column.
2. Identify some important patterns i.e. Which variables are most significant with respect to the target variable?
3. Insights and Recommendations (i.e. Data Driven Business Decision)

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

# SPRINT 2 - Build a Data App

**Background:**
Your manager has tasked you with building a Data Dashboard using Streamlit. A Data Dashboard is a dynamic way to present data insights, which is widely used in the industry. It allows for interactive exploration of the data for end-users. Unlike a static PowerPoint presentation, you'll be providing a fully functional, user-friendly dashboard to share with the client.

**Why a Data Dashboard?**

1. **Dynamic Insights:** Instead of static visuals, a dashboard offers real-time data exploration.
2. **Interactivity:** End-users can interact with the data, making it more engaging and insightful.
3. **Professional Presentation:** It's a modern, professional way to showcase data insights.

**Why Streamlit?**

1. **User-Friendly:** Streamlit is known for its simplicity, making it accessible even for those new to data science.
2. **Rapid Prototyping:** It allows for quick development of interactive web applications.
3. **Integration with Data Tools:** Streamlit easily integrates with popular data libraries like Pandas, Matplotlib, and Plotly.

**Benefits for You:**

1. You'll gain experience in building dynamic data applications.
2. You'll be able to present data insights in a way that's more engaging and impactful than static presentations.

**Task:** You'll be providing a fully functional, interactive Data Dashboard built using Streamlit, offering a modern and engaging way to present data insights to the client.

# SPRINT 3 - Model Building

## Task - 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.

**Why predict medical cost? (Business Impact)**
Managing and predicting medical costs for policyholders can be complex, and inaccurate estimations may lead to financial challenges for the insurance company. If we can predict medical cost using a ML model, we can easily determine appropriate premium pricing that reflects the actual risk and cost associated with each policyholder is critical for competitiveness and profitability. ML models can help in setting dynamic and personalized premium prices, aligning them with individual risk profiles, and ensuring fairness and competitiveness in the market.

**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**.