**Predicting Health Insurance Price for an individual or family**

The majority of the countries finalize health insurance costs based on many factors such as age, number of people in families, etc. What should be the actual health insurance price for an individual or a family is an issue for many companies. Hence, one insurance company hired you as a data scientist to predict the health insurance cost for possible future customers. They have already collected samples required to perform all data analysis and machine learning tasks. Your task is to perform all data analysis steps and finally create a machine learning model which can predict the health insurance cost.

Please address certain questions as you work on this project.

1- Why is this proposal important in today's world? How predicting a health insurance cost accurately can affect the health care/insurance field?

2- If any, what is the gap in the knowledge, or how your proposed method can be helpful if required in the future for any other type of insurance?

3- Please aim to identify patterns in the data and important features that may impact an ML model.

4- Please perform multiple machine learning models, perform all required steps to check if there are any assumptions, and justify your model. Why is your model better than any other possible model? Please explain it by relevant cost functions and, if possible, by any graph.

**Data analysis approach**

a. What approach are you going to take to prove or disprove your hypothesis?
b. What feature engineering techniques will be relevant to your project?
c. Please justify your data analysis approach.
d. Identify essential patterns in your data using the EDA approach to justify your findings.

**Machine learning approach**

a. What method will you use for machine learning-based predictions of health insurance price?
b. Please justify the most appropriate model.
c. Please perform the necessary steps required to improve the accuracy of your model.
d. Please compare all models (at least four models).

Variables in the dataset:

1. age: age of the primary beneficiary
2. sex: insurance contractor gender, female, male
3. 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

4. children: number of children covered by health insurance, number of dependents
5. smoker: smoking or not
6. region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
7. charges: individual medical costs billed by health insurance