# Capstone 2 – Project Proposal

## TITLE

***Predict Health Insurance Cost based on individual Health Parameters***

## INTROUDUCTION

This Project is to build a model that could predict the amount an individual needs to pay towards his Health insurance per year based on certain Health parameters.

The aim of this Project is to:

- To determine if there is a relationship between attributes and medical costs.
- To determine if there a significant difference in medical costs between different groups.
- To fit a multiple linear regression to predict costs.

## DATASET

Medical Costs Dataset: https://www.kaggle.com/mirichoi0218/insurance?select=insurance.csv

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 children covered by health insurance / Number of dependents
- SMOKER: Whether Beneficiary is a Smoker
- REGION: Beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- CHARGES: Individual medical costs billed by health insurance.

## APPROACH

1. Data Cleanup and Identification of Outliers: Check for any missing values and replace with appropriate Method and then also check for Outliers.

2. Plot the data to determine any interesting relationship between the features.

3. Perform statistical analysis to get deep insights to :

    - Prove (or disprove) whether the medical claims made by the people who smoke is greater than those who don't.
    - Prove (or disprove) with statistical evidence whether the BMI of females is different from that of males.
    - Is the proportion of smokers significantly different across different regions?
    - Is the mean BMI of women with no children, one child, and two children the same?

    The Hypotheses for this model are such:

    - Null Hypothesis: there will be no significant prediction of medical expenses by the policyholder's smoking status, BMI score, age, region of residence, sex, and number of dependents covered by the policy.
    - Alternative Hypothesis: there will be significant prediction based on the above mentioned factors.

4. Provide recommendations based on key outcomes from the Statistical evidence.
5. Transform the Features for ML usage: Categorical features must be converted to number by applying following methods:

    1. Label Encoding
    2. One hot encoding
    3. Dummy variable trap

6. Regression Models: Apply various models and find out the best approach:
    - Linear Regression
    - Polynomial Regression
    - Decision Tree Regression
    - Random Forest Regression

7. Model Validation: Following methods should be applied to validate the model:
    - Check for Linearity
    - Check for Residual normality and Mean
    - Check for Multivariate Normality
    - Check for Homoscedasticity

8. Additionally, explore Clustering option and the consequences of clustering.