# PROJECT PROPOSAL

Medical Cost Prediction

By:

Anurag Nepal 1NT19CS036

Avishek Rijal 1NT19CS045

Baibhav Dhakal 1NT19CS048

Nabin Kumar KC 1NT19CS116

## Introduction

We are thinking of applying the data mining tools that we have been learning this semester to try and tackle this problem to predict the insurance cost of a person based on the age,sex,bmi,number of children ,whether he/she is a smoker and the region they preside in.

## Data Mining Task

For this project , using supervised learning model i.e multiple types of regression or the decision tree algorithm to find the optimal solution. We want to see how the normalized data affects the accuracy of the prediction ,whether it reduces the computational complexity or not.

## Data Set

The data set we chose from [kaggle](https://www.kaggle.com/mirichoi0218/insurance) has the columns as

* 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: Smoking
* region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
* charges: Individual medical costs billed by health insurance

## Methods and Models

Regarding the model we are trying to find out the best model by comparing their output and their performance in both the train and test instances. We will use various variations of regression model and a decision tree classifier to find out the optimal algorithm.

## Presentation and Visualization

We are thinking of having a confusion matrix and histograms to compare between the models used in the task.