STEP 1🡪 Problem Definition

#The following project is a Classification problem type

***Problem Statement:***

Insurance fraud is a huge problem in the industry. It's difficult to identify fraud claims. Machine Learning is in a unique position to help the Auto Insurance industry with this problem.

In this project, exists a dataset which has the details of the insurance policy along with the customer details. It also has the details of the accident on the basis of which the claims have been made.

In this example, we are working with some auto insurance data to demonstrate how you can create a predictive model that predicts if an insurance claim is fraudulent or not.

STEP 2 🡪 Data Analysis

1. Import the libraries required
2. Appropriate library required
3. Reading the CSV file from the given Github on the given location and displaying the data frame
4. Describing the columns present in the data set and also displaying the number of rows and columns in it

STEP 3 🡪 EDA Concluding Remarks

1. Checking whether the columns present in the data frames have null or missing values and remove the missing or null values that exist in the data frame.
2. Deleting the unnecessary columns
3. Checking the data type of each column

STEP 4 🡪 Pre-Processing Pipeline

1. Using the appropriate encoding techniques to separate and convert the categorical and continues columns

* Here in this following project I have used label encoder to convert categorical column which exist in string data type into integer values
* The column that are encoded in this data set are as follows:
  + Fraud\_ reported, Police\_ report available, Policy\_ state

1. Describing the co-relation that exists between the columns.

* Importing the libraries required for plotting the visualization of the co-relation to check whether any co-relation exist between the columns of the data set.
* If there exist any columns are highly or closely related, delete one of the columns.

1. Checking if there exist skewness in the data

* In this project I have used the log transformation technique to remove the skewness present in the data.
* After applying skewness I have used ordinal encoding method for columns which are continuous in nature and have string data type to re-shape them and get them in the proper range into numerical values for building and training purpose.

STEP 5 🡪 Building the Machine Learning Model

1. Separating the columns and the target variable for building the machine learning model
2. Since this is a classification problem type I have used the logistic regression method which is used for classification purpose

* Importing the libraries required for logistic regression method like accuracy score, confusion matrix, classification report etc.
* Giving the range of training for the machine learning model
* Initializing logistic regression method and describing train and test data
* Training the model

1. Displaying the accuracy of training and testing data predicted by the machine learning model

STEP 6 🡪 Concluding remarks

1. Describing and printing the classification report
2. Doing cross validation in order to achieve the cross validation
3. Printing and displaying the accuracy score.

# I have also tried using the decision tree and K nearest Neighbor method but concluded and used the that Logistic Regression machine learning model as it gives the best accuracy for the following project assignment.