# Topic(s): Decision Tree & Random Forest

**Instructions**

Please share your answers filled inline in the word document. Submit Python code and R code files wherever applicable.

Please ensure you update all the details:

**Name: Batch Id:**  **Topic: Decision Tree And Random Forest**

1. **Business Problem**
   1. **Objective**
   2. **Constraints (if any)**
2. **Work on each feature of the dataset to create a data dictionary as displayed in the below image:**



**2.1 Make a table as shown above and provide information about the features such as its Data type and its relevance to the model building, if not relevant provide reasons and provide description of the feature.**

**Using R and Python codes perform:**

1. **Data Pre-processing**

**3.1 Data Cleaning, Feature Engineering, etc.**

1. **Exploratory Data Analysis (EDA):**
   1. **Summary**
   2. **Univariate analysis**
   3. **Bivariate analysis**
2. **Model Building**
   1. **Build the model on the scaled data (try multiple options)**
   2. **Perform Decision Tree and Random Forest on the given datasets.**
   3. **Train and Test the data and perform cross validation techniques, compare accuracies, precision and recall and explain about them.**
   4. **Briefly explain the model output in the documentation.**



1. **Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.**

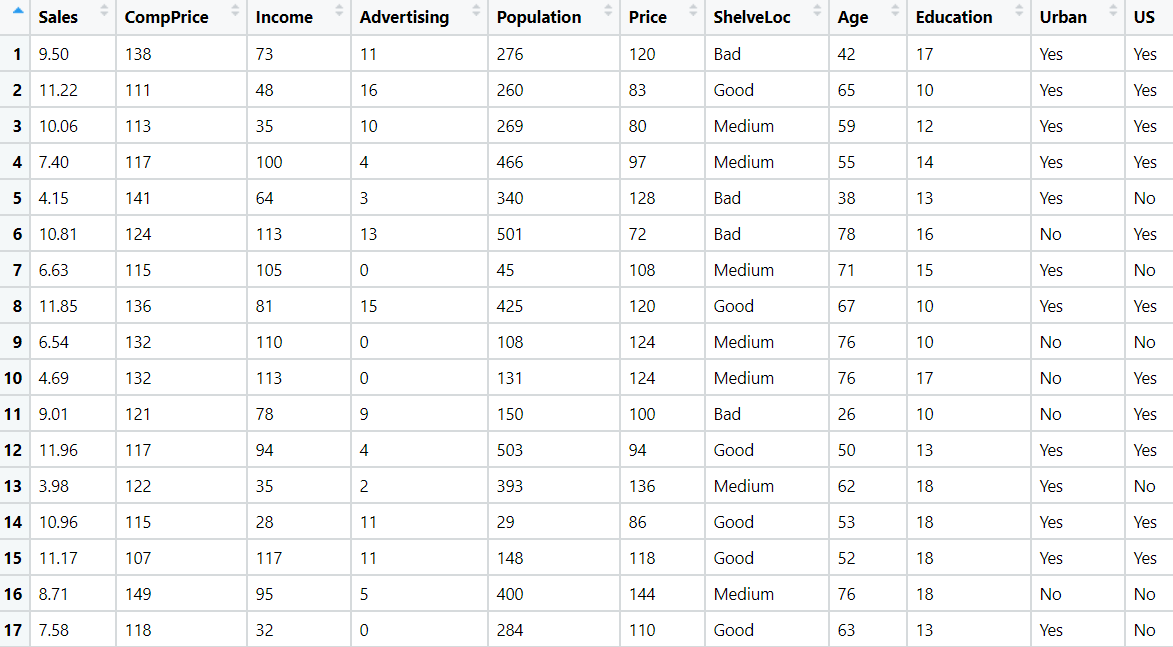
# Note:

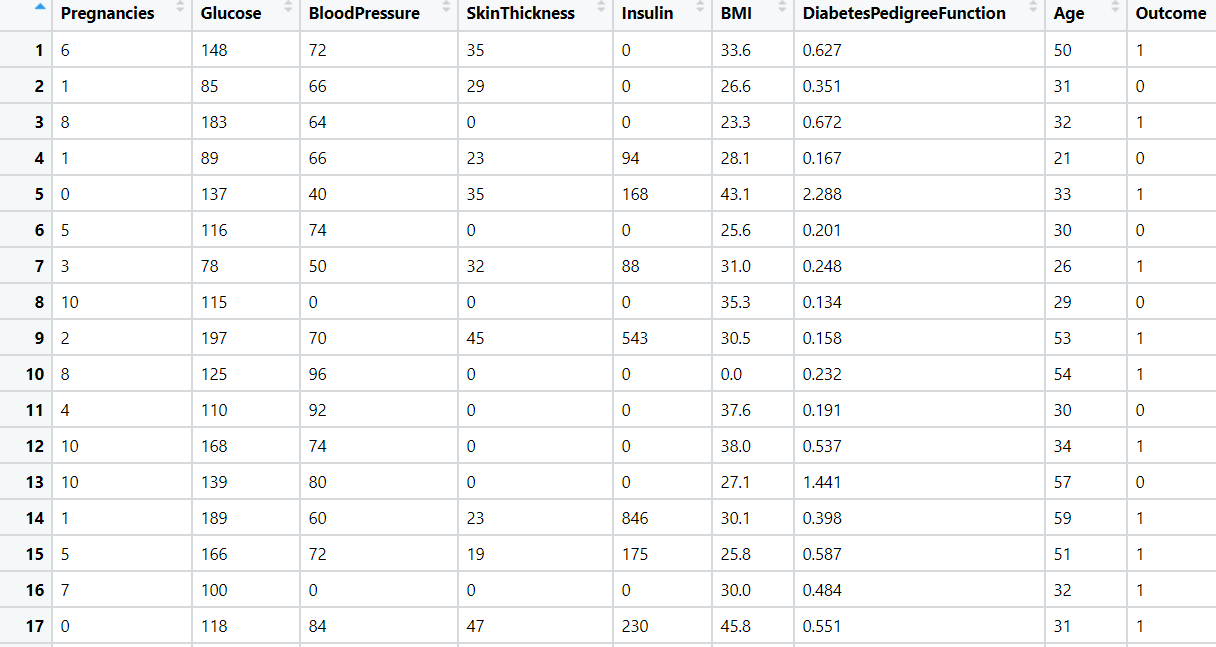
The assignment should be submitted in the following format:

* R code
* Python code
* Code Modularization should be maintained
* Documentation of the model building (elaborating on steps mentioned above)

**Problem Statement: -**

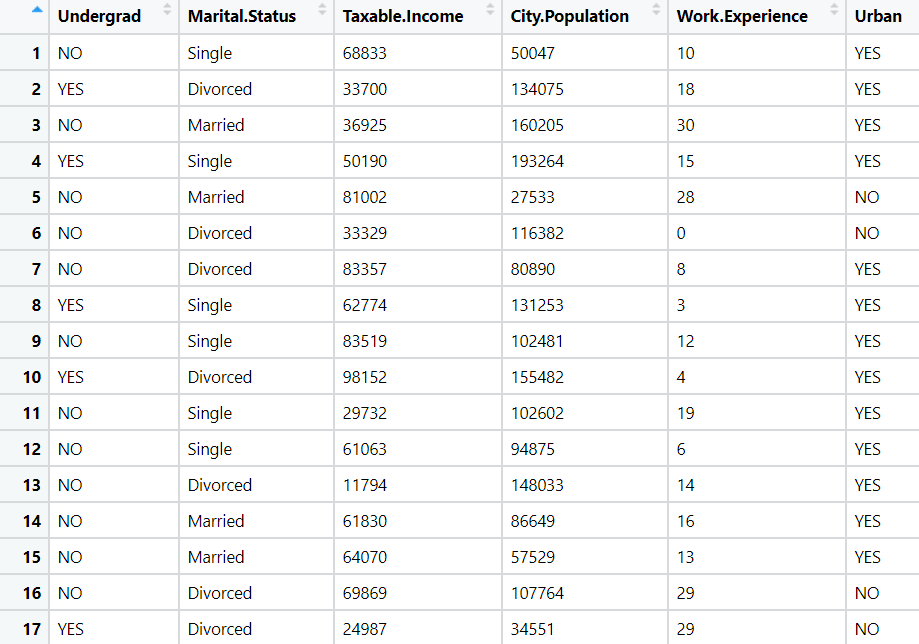
 A cloth manufacturing company is interested to know about the segment or attributes contributing to high sale. Approach - A decision tree & random forest model can be built with target variable 'Sales' (we will first convert it into categorical variable) & all other variables will be independent in the analysis.



**Problem Statement: -**

Divide the data (Diabetes) into training and test datasets and create a Random Forest and Decision Tree Model to classify 'Class Variable' or “Outcome”



Problem Statement: -

Use decision trees & random forest algorithm to prepare a model on fraud datatreating those who have taxable\_income <= 30000 as "Risky" and others are "Good".



Problem Statement: -

In Recruitment domain, HR faces with the challenge of predicting if the candidate is faking his salary or the candidate is genuine. In order to do it manually, let us use our Machine Learning algorithm to correctly classify using Decision Tree and Random Forest. We have a scenario where, a candidate claims to have 5 years of experience and earning 70000 per month working as regional manager and the candidate is expecting more than his previous CTC. A sample data has been collected, find out the candidate claims are genuine or fake.

A screenshot of a cell phone

Description automatically generated