# Decision Tree and Random Forest

**Instructions:**

Please share your answers filled in-line in the word document. Submit code separately wherever applicable.

Please ensure you update all the details:

**Name: \_\_\_Jaidev Chhabria\_\_\_\_\_\_\_\_\_\_ Batch ID:** \_DSWE MON 29052021\_\_\_\_\_\_\_\_\_\_

**Topic: Decision Tree and Random Forest**

**Grading Guidelines:**

**1. An assignment submission is considered complete only when correct and executable code(s) are submitted along with the documentation explaining the method and results. Failing to submit either of those will be considered an invalid submission and will not be considered for evaluation.**

**2. Assignments submitted after the deadline will affect your grades.**

**Grading:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Ans** | **Date** |  |  | **Ans** | **Date** |
| Correct | On time | A | 100 |  |  |
| 80% & above | On time | B | 85 | Correct | Late |
| 50% & above | On time | C | 75 | 80% & above | Late |
| 50% & below | On time | D | 65 | 50% & above | Late |
|  |  | E | 55 | 50% & below |  |
| Copied/No Submission |  | F | 45 |  |  |

* **Grade A: (>= 90):** When all assignments are submitted on or before the given deadline.
* **Grade B: (>= 80 and < 90):** 
  + When assignments are submitted on time but less than 80% of problems are completed.

(OR)

* + All assignments are submitted after the deadline.
* **Grade C: (>= 70 and < 80):** 
  + When assignments are submitted on time but less than 50% of the problems are completed.

(OR)

* + Less than 80% of problems in the assignments are submitted after the deadline.
* **Grade D: (>= 60 and < 70):**
  + Assignments submitted after the deadline and with 50% or less problems.
* **Grade E: (>= 50 and < 60):** 
  + Less than 30% of problems in the assignments are submitted after the deadline.

(OR)

* + Less than 30% of problems in the assignments are submitted before the deadline.
* **Grade F: (< 50):** No submission (or) malpractice.

**Hints:**

1. **Business Problem**
   1. **What is the business objective?**
   2. **Are there any constraints?**
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. And if not relevant, provide reasons and a description of the feature.**

1. **Data Pre-processing**

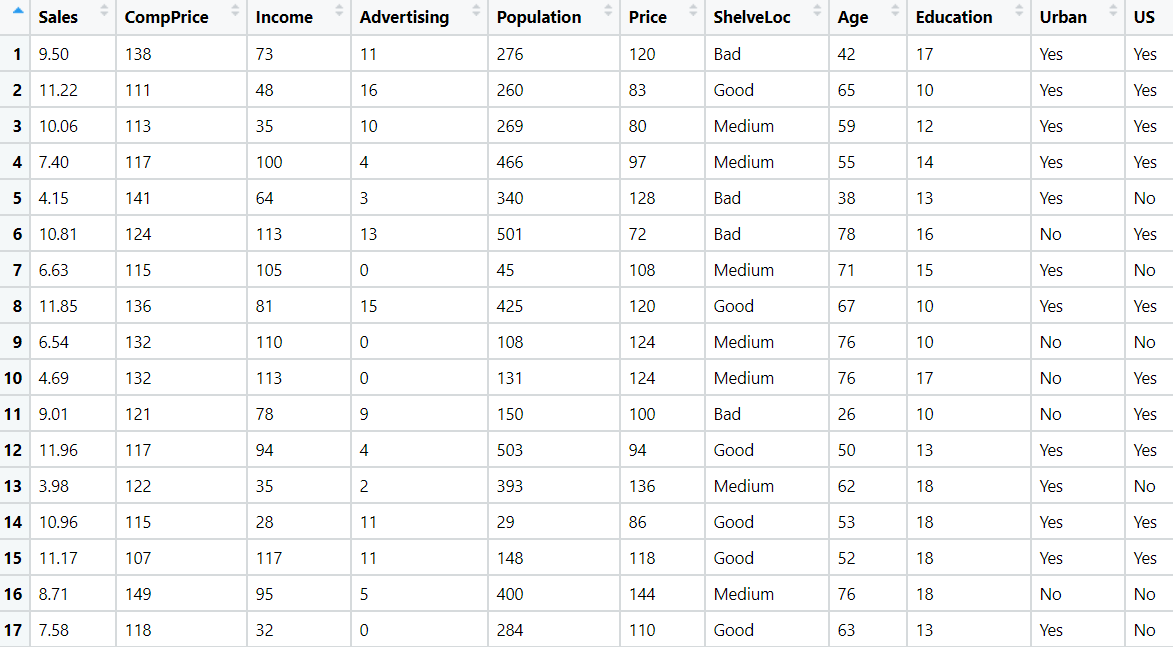
**3.1 Data Cleaning, Feature Engineering, etc.**

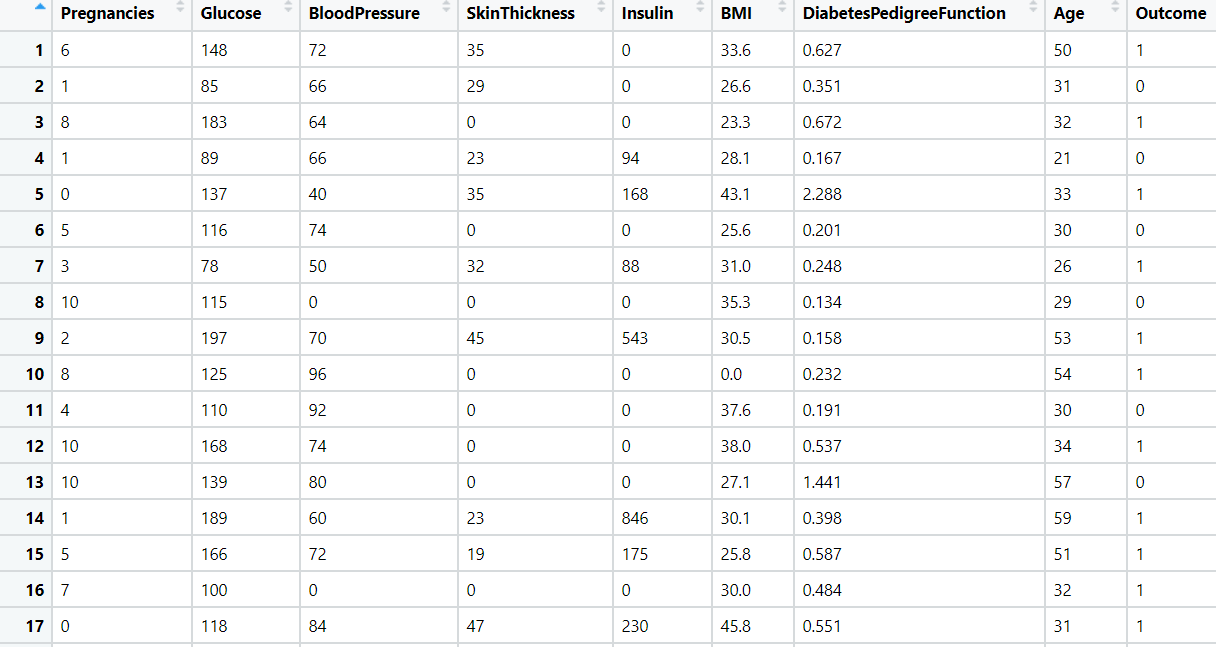
1. **Exploratory Data Analysis (EDA):**
   1. **Summary.**
   2. **Univariate analysis.**
   3. **Bivariate analysis.**
2. **Model Building**
   1. **Build the model on the scaled data (try multiple options).**
   2. **Perform Decision Tree and Random Forest on the given datasets.**
   3. **Train and 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. **Write about the benefits/impact of the solution - in what way does the business (client) benefit from the solution provided?**

**Problem Statements:**

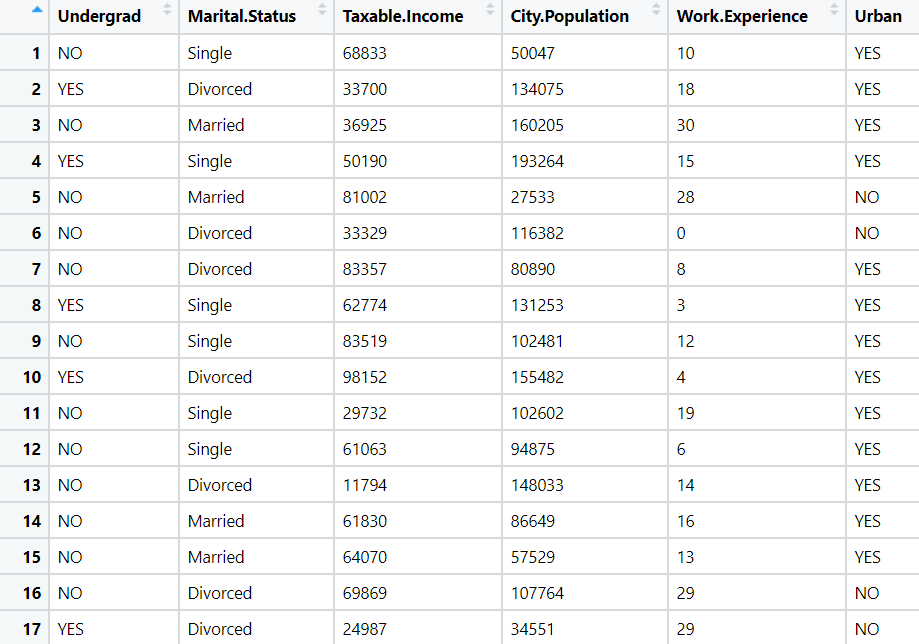
1. A cloth manufacturing company is interested to know about the different attributes contributing to high sales. Build a decision tree & random forest model with Sales as target variable (first convert it into categorical variable).





1. Divide the diabetes data into train and test datasets and build a Random Forest and Decision Tree model with Outcome as the output variable.





1. Build a Decision Tree & Random Forest model on the fraud data. Treat those who have taxable\_income <= 30000 as Risky and others as Good (discretize the taxable income column).



1. In the recruitment domain, HR faces the challenge of predicting if the candidate is faking their salary or not. For example, a candidate claims to have 5 years of experience and earns 70,000 per month working as a regional manager. The candidate expects more money than his previous CTC. We need a way to verify their claims (is 70,000 a month working as a regional manager with an experience of 5 years a genuine claim or does he/she make less than that?) Build a Decision Tree and Random Forest model with monthly income as the target variable.

A screenshot of a cell phone

Description automatically generated