# 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: Nukala Ayyappa Bharthwaj**

**Batch Id:DSWDMCOS 21012022**

**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.

**Sol:**

**Business Objective:** To know about the high sales of the clothes by using a decision tree.

**Constraints:** Lack of analysis of the sales data.

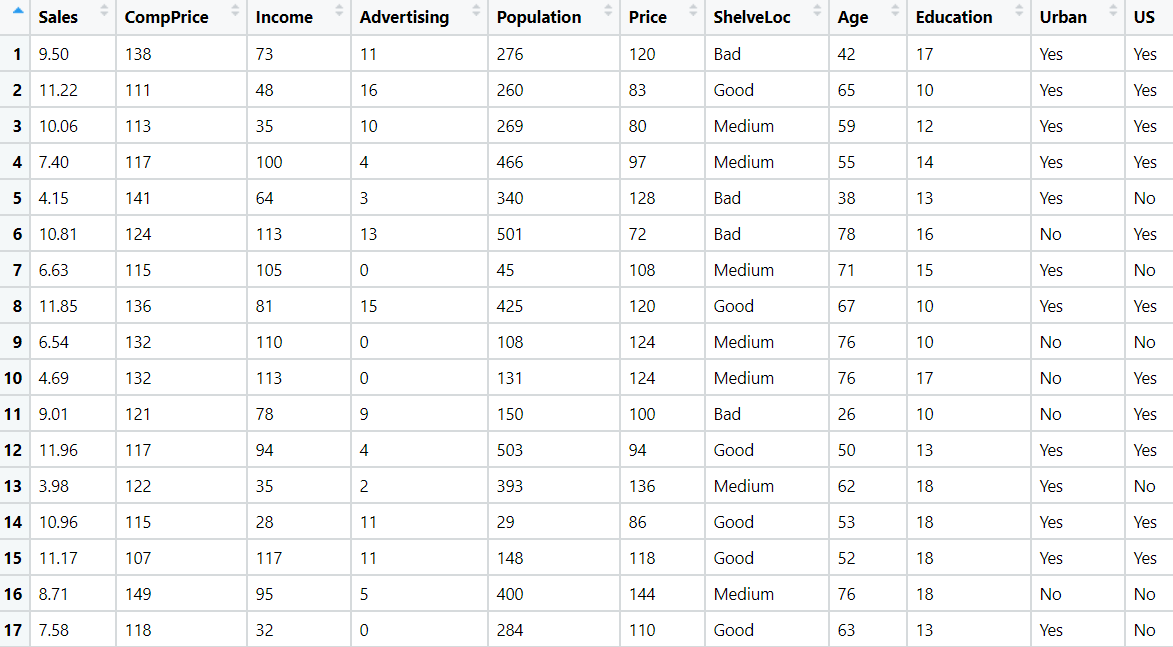
**Data Types:** given data and its data types are shown below

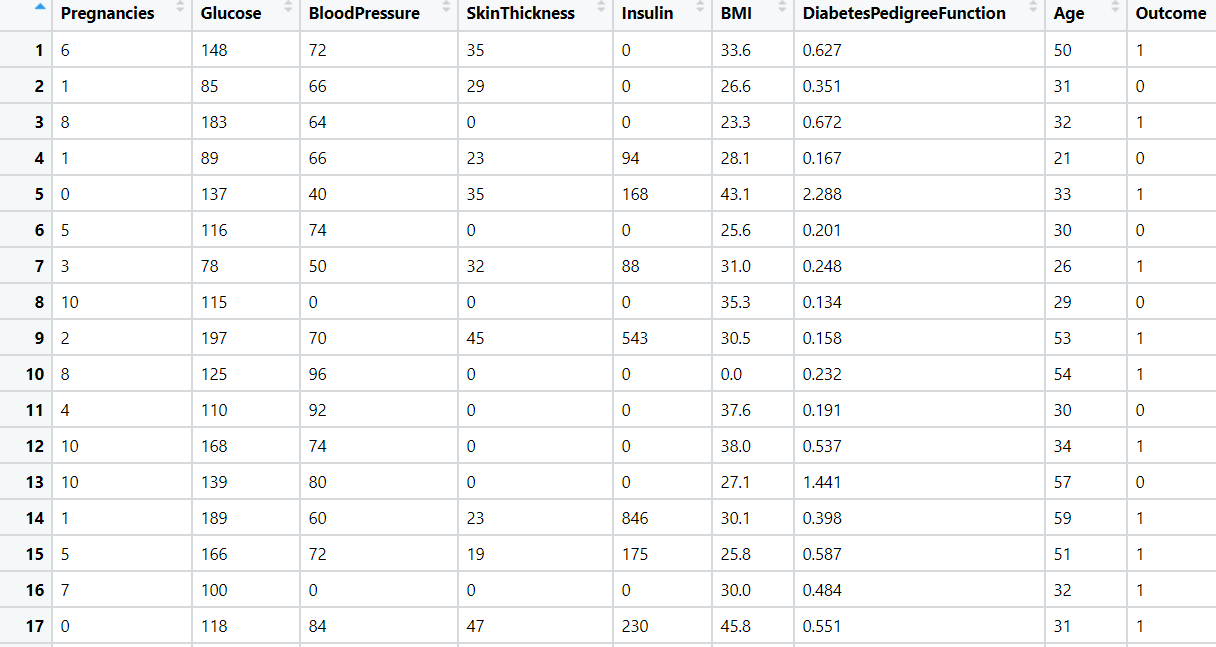
|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | Description | Data type | Relevance |
| Sales | Sales of the company | Ratio | Relevant |
| CompPrice | Company Price | Ratio | Relevant |
| Income | Profit on the product | Ratio | Relevant |
| Advertising | Amount on advertisement | Ratio | Relevant |
| Population | Population of the place | Ratio | Relevant |
| Price | Price in the particular place | Ratio | Relevant |
| ShelveLoc | Rating for the sales | Nominal | Relevant |
| Age | Age of the buyer | Ratio | Relevant |
| Education | Education number of buyer | Ordinal | Relevant |
| Urban | Whether its urban or not | Nominal | Relevant |
| US | Whether its US or not | Nominal | Relevant |

**Data Pre-Processing:** I have identified the non-numeric data in the given data set and then I have converted that into numeric using factor function in R and in python by using Label Encoding. For the sales column I have segregated into yes or no by taking the high sales greater than 9 and remaining all as no.

**Decision tree:** I have made the decision tree model for the given data set using Python. I have splited the data into training and testing as 70% and 30 % Respectively and My model have the accuracy of 74% for prediction of the data in the decision tree.

**Random Forest:** Since the accuracy is little bit low in the decision tree model I have done the random forest model for the same data sets of training and testing data and h I have got the accuracy as 82%. So this model can be used for predicting the future sales of the cloth manufacturing company.



**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”



**Business Objective:** To segregate the diabetes patients by using a decision tree and random forest.

**Constraints:** Lack of analysis of the patients data.

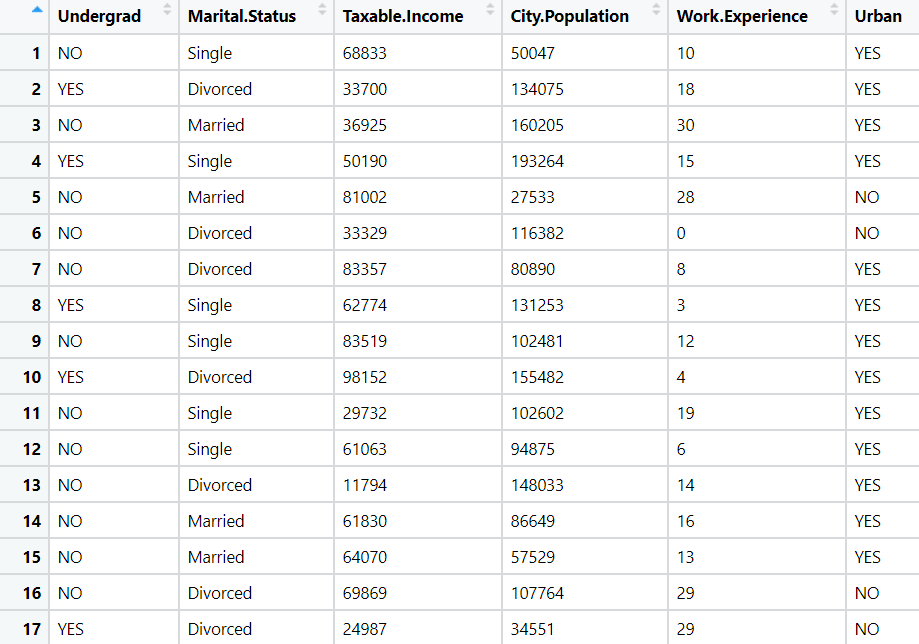
**Data Types:** given data and its data types are shown below

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | Description | Data type | Relevance |
| Number of times pregnant | Number of time the woman got pregnancy | Internal | Relevant |
| Plasma glucose concentration | Plasma glucose concentration a 2 hours in an oral glucose tolerance test | Ratio | Relevant |
| Diastolic blood pressure | Diastolic blood pressure (mm Hg) | Ratio | Relevant |
| Triceps skin fold thickness | Triceps skin fold thickness (mm) | Ratio | Relevant |
| 2-Hour serum insulin | 2-Hour serum insulin (mu U/ml) | Ratio | Relevant |
| Body mass index | Body mass index (weight in kg/(height in m)^2) | Ratio | Relevant |
| Diabetes pedigree function | Diabetes pedigree function | Ratio | Relevant |
| Age (years) | Age of patient | Ratio | Relevant |
| Classvariable | Variable for patiest having diabetes or not | Nominal | Relevant |

**Data Pre-Processing:** the complete data can be used for preparing the model of decision tree and random forest.

**Decision tree:** I have made the decision tree model for the given data set using Python. I have splited the data into training and testing as 70% and 30 % Respectively and My model have the accuracy of 72% for prediction of the data in the decision tree.

**Random Forest:** Since the accuracy is little bit low in the decision tree model I have done the random forest model for the same data sets of training and testing data and h I have got the accuracy as 80%. So this model can be used for predicting the whether the patient will get diabetes or not.

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".



**Sol:**

**Business Objective:** To know about the check is risky or good using a decision tree and random forest models.

**Constraints:** Lack of analysis of the fraud checks data.

**Data Types:** given data and its data types are shown below

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | Description | Data type | Relevance |
| Undergrad | Whether the applicant is undergraduate or not | Nominal | Relevant |
| Marital.Status | Marital status of applicant | Nominal | Relevant |
| Taxable.Income | Taxable income of the applicant | Ratio | Relevant |
| City.Population | Population of the city | Ratio | Relevant |
| Work.Experience | Work experience of the applicant | Ratio | Relevant |
| Urban | Whether the state is urban or not | Nominal | Relevant |

**Data Pre-Processing:** I have identified the non-numeric data in the given data set and then I have converted that into numeric using factor function in python by using Label Encoding. For the taxable income column I have segregated into Risky or Good by taking the taxable income greater than 30000 as Risky and remaining all as Good.

**Decision tree:** I have made the decision tree model for the given data set using both R and Python. I have splitted the data into training and testing as 70% and 30 % Respectively and My model have the accuracy of 70% for prediction of the data in the decision tree.

**Random Forest:** Since the accuracy is little bit low in the decision tree model I have done the random forest model for the same data sets of training and testing data and I have got the accuracy as 78%. So this model can be used for predicting the whether the check is risky or not.

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.

**Sol:**

**Business Objective:** To know whether the candidate is expediting the correct CTC or not by using decision tree and random forest models.

**Constraints:** Lack of analysis of the previous employee data.

**Data Types:** given data and its data types are shown below

|  |  |  |  |
| --- | --- | --- | --- |
| Name of feature | Description | Data type | Relevance |
| Position of the employee | Employee position in the company. | Nominal | Relevant |
| no of Years of Experience of employee | Experience of the employee | Ratio | Relevant |
| Monthly income of employee | Salary of the employee | Ratio | Relevant |

**Data Pre-Processing:** I have identified the non-numeric data in the given data set and then I have converted that into numeric using factor function in python by using Label Encoding. For the Monthly income of employee column I have segregated into High or Medium & Low by taking the Monthly income of employee greater than 70000 as High and remaining all as Medium & Low.

**Decision tree:** I have made the decision tree model for the given data set using both R and Python. I have splitted the data into training and testing as 70% and 30 % Respectively and My model have the accuracy of 75% for prediction of the data in the decision tree.

**Random Forest:** Since the accuracy is little bit low in the decision tree model I have done the random forest model for the same data sets of training and testing data and I have got the accuracy as 84%. So this model can be used for whether the new employee is demanding correct CTC or Not.

A screenshot of a cell phone

Description automatically generated