# 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: Shubham Sharma Batch Id:**  DS WDMCO 050121 Blr **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.**

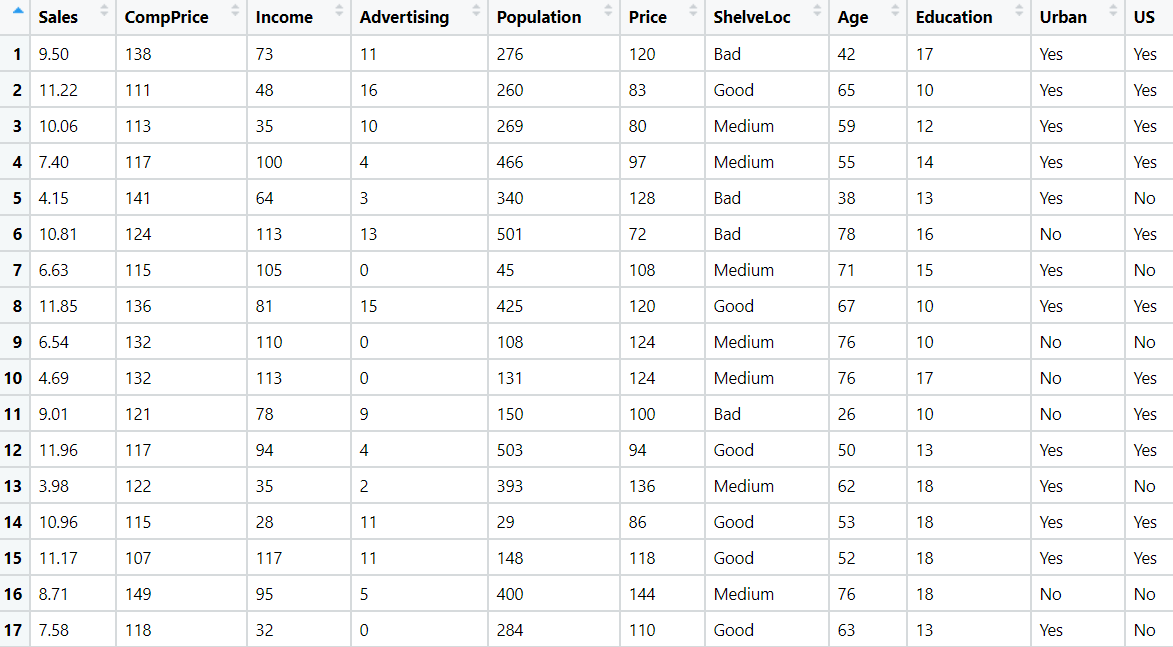
# Note:

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.



**Business Problem**

**Objective :-** predict sale

**Python code details :**

Data Frame name is df. It has 400 rows and 11 columns.

**Work on each feature of the dataset to create a data dictionary as displayed in the below image:**

Then we create a data frame that’s contain details of each columns ,like- description ,data types ,and save the details named as data\_details .all of them are important .

**Data Pre-processing**

**Data Cleaning and Data Mining.**

Now we check info and describe for df .Check for data types ,unique value and variance .

Then we check for unique value in each columns

:-

Sales 336

CompPrice 73

Income 98

Advertising 28

Population 275

Price 101

ShelveLoc 3

Age 56

Education 9

Urban 2

US 2

Dataframe has no missing values in columns .

We have done EDA for each columns and save the details as EDA. covariance for data set save as covariance . historgam and scatter plot for each column all data are normally distributed as well as we check for boxplot .

We have done labelencoding for factor columns ,and then separate the data into target and predictor .

**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.**

Now we split our data in X\_train, X\_test, Y\_train, Y\_test 80% data on train and 20% test . Preparing a decision tree model on training data set ,then test on test data , Evaluation on Test Data as result root mean square error=2.02.

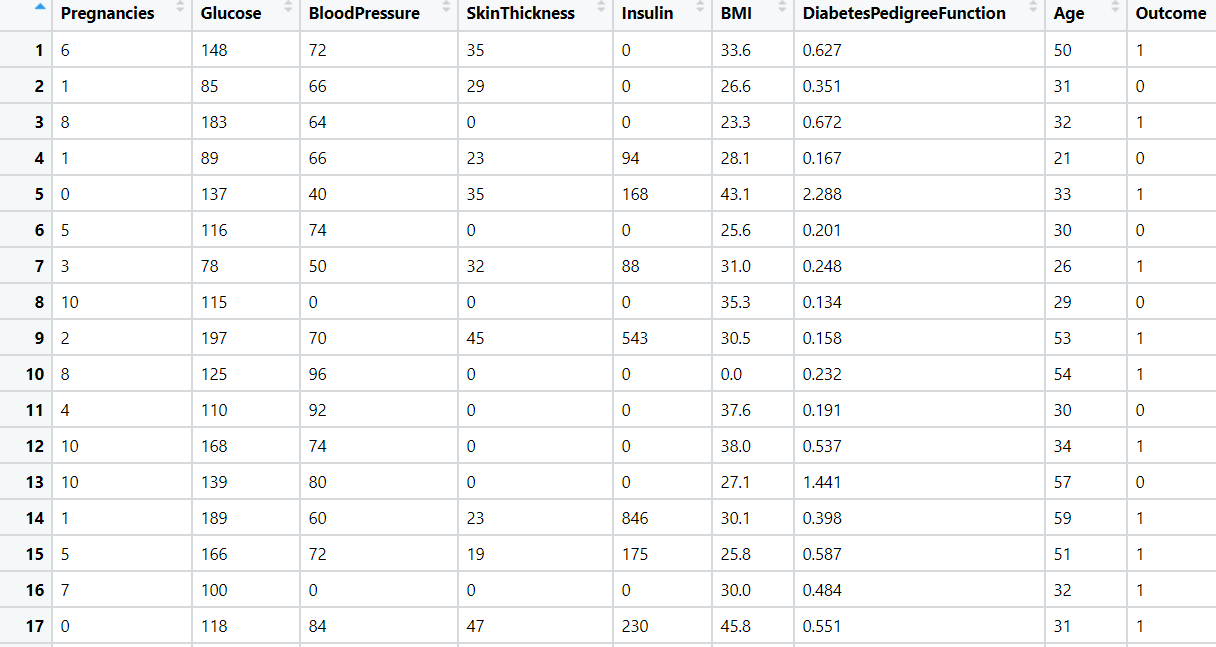
Evaluation on Train Data also as result root mean square error=0.

Our model is overfit here so we build random forest .

Preparing a decision tree model on same training data set ,then test on test data , Evaluation on Test Data as result root mean square error= 1.4.

.

* used library :- pandas ,numpy , sk learn ,matplotlib

**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 Problem**

**Objective :-** predict class

**Python code details :**

Data Frame name is df. It has 768 rows and 9 columns.

**Work on each feature of the dataset to create a data dictionary as displayed in the below image:**

Then we create a data frame that’s contain details of each columns ,like- description ,data types ,and save the details named as data\_details .all of them are important .

**Data Pre-processing**

**Data Cleaning and Data Mining.**

Now we check info and describe for df .Check for data types ,unique value and variance .

Then we check for unique value in each columns

:-

Number of times pregnant 17

Plasma glucose concentration 136

Diastolic blood pressure 47

Triceps skin fold thickness 51

2-Hour serum insulin 186

Body mass index 248

Diabetes pedigree function 517

Age (years) 52

Class variable 2

Dataframe has no missing values in columns .

We have done EDA for each columns and save the details as EDA. covariance for data set save as covariance . historgam and scatter plot for each column all data are normally distributed as well as we check for boxplot .

We have done labelencoding for factor columns ,and then separate the data into target and predictor .

**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.**

Now we split our data in X\_train, X\_test, Y\_train, Y\_test 80% data on train and 20% test . Preparing a decision tree model on training data set ,then test on test data , Evaluation on Test Data as result 74% accuracy .

Crosstab

Predictions NO YES

Actual

NO 79 18

YES 22 35

Evaluation on Train Data also residual =0.

Our model is overfit here so we build random forest .

Preparing a decision tree model on same training data set ,then test on test data , Evaluation on Test Data as result residual =20%.

.crosstable

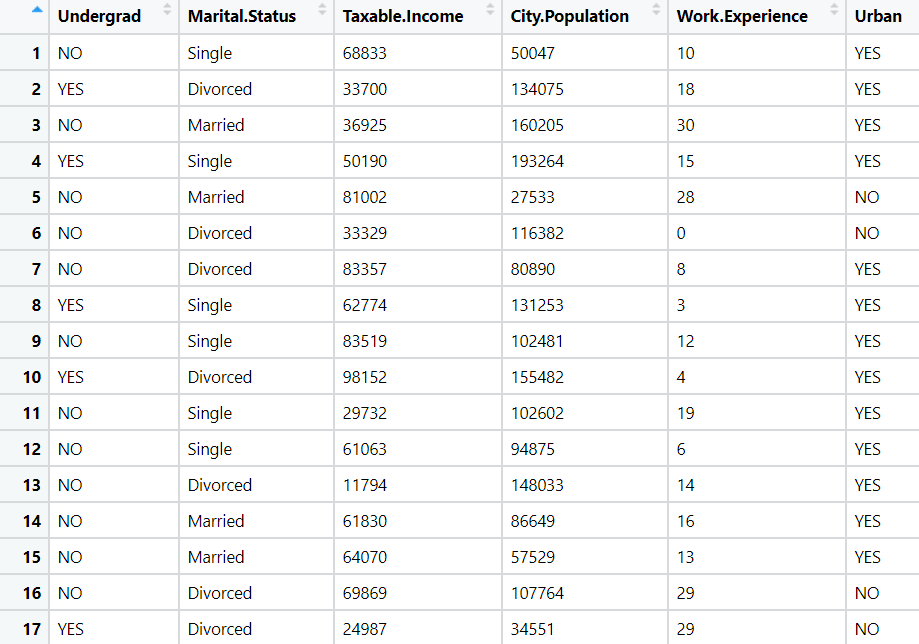
Predictions NO YES

Actual

NO 87 10

YES 20 37

* used library :- pandas ,numpy , sk learn ,matplotlib

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".



**Python code details :**

Data Frame name is df. It has 600 rows and 6 columns.

**Work on each feature of the dataset to create a data dictionary as displayed in the below image:**

Then we create a data frame that’s contain details of each columns ,like- description ,data types ,and save the details named as data\_details .all of them are important .

**Data Pre-processing**

**Data Cleaning and Data Mining.**

Now we check info and describe for df .Check for data types ,unique value and variance .

Then we check for unique value in each columns

:-

Undergrad 2

Marital.Status 3

Taxable.Income 599

City.Population 598

Work.Experience 31

Urban 2Dataframe has no missing values in columns .

We have done EDA for each columns and save the details as EDA. covariance for data set save as covariance . historgam and scatter plot for each column all data are normally distributed as well as we check for boxplot .

We have done labelencoding for factor columns ,and then separate the data into target and predictor .

**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.**

Now we split our data in X\_train, X\_test, Y\_train, Y\_test 80% data on train and 20% test . Preparing a decision tree model on training data set ,then test on test data , Evaluation on Test Data as result 61.6% accuracy .

Crosstab

Predictions Good Risky

Actual

Good 4 25

Risky 21 70

Evaluation on Train Data also residual =0.

Our model is overfit here so we build random forest .

Preparing a decision tree model on same training data set ,then test on test data , Evaluation on Test Data as result residual =25%.

.crosstable

Predictions NO YES

Actual

NO 0 29

YES 4 87

* used library :- pandas ,numpy , sk learn ,matplotlib

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

**Python code details :**

Data Frame name is df. It has 196 rows and 3 columns.

**Work on each feature of the dataset to create a data dictionary as displayed in the below image:**

Then we create a data frame that’s contain details of each columns ,like- description ,data types ,and save the details named as data\_details .all of them are important .

**Data Pre-processing**

**Data Cleaning and Data Mining.**

Now we check info and describe for df .Check for data types ,unique value and variance .

Then we check for unique value in each columns

:-

Position of the employee 10

no of Years of Experience of employee 29

monthly income of employee 30

df has no missing values in columns .

We have done EDA for each columns and save the details as EDA. covariance for data set save as covariance . historgam and scatter plot for each column all data are normally distributed as well as we check for boxplot .

We have done labelencoding for factor columns ,and then separate the data into target and predictor .

**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.**

Now we split our data in X\_train, X\_test, Y\_train, Y\_test 80% data on train and 20% test . Preparing a decision tree model on training data set ,then test on test data , Evaluation on Test Data as result 5.6% accuracy .

Evaluation on Train Data also as result 38% accuracy

Our model is overfit here so we build random forest .

Preparing a decision tree model on same training data set ,then test on test data , Evaluation on Test Data as result residual =95%.

* used library :- pandas ,numpy , sk learn ,matplotlib