# 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: DHEERAJ MISHRA Batch ID:**  DS\_01072021

**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**

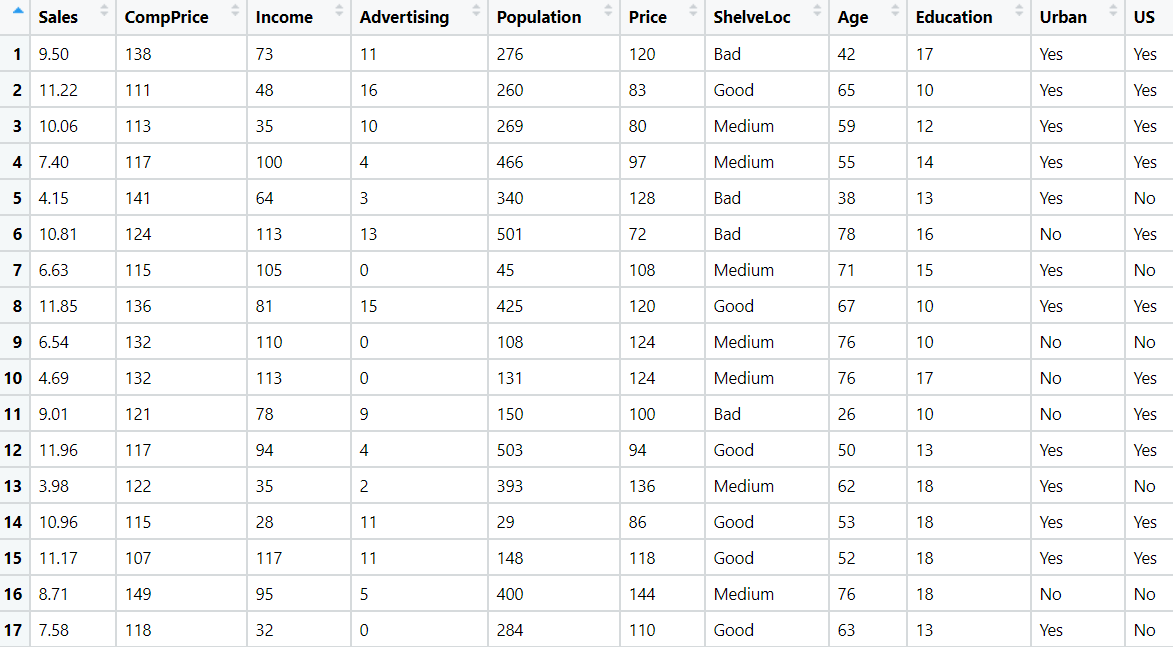
**3.1 Data Cleaning, Feature Engineering, etc.**

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).



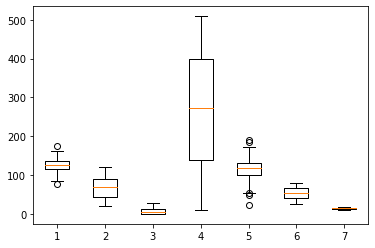
1. BUSINESS OBJECTIVE:-

Maximize mapping of sales

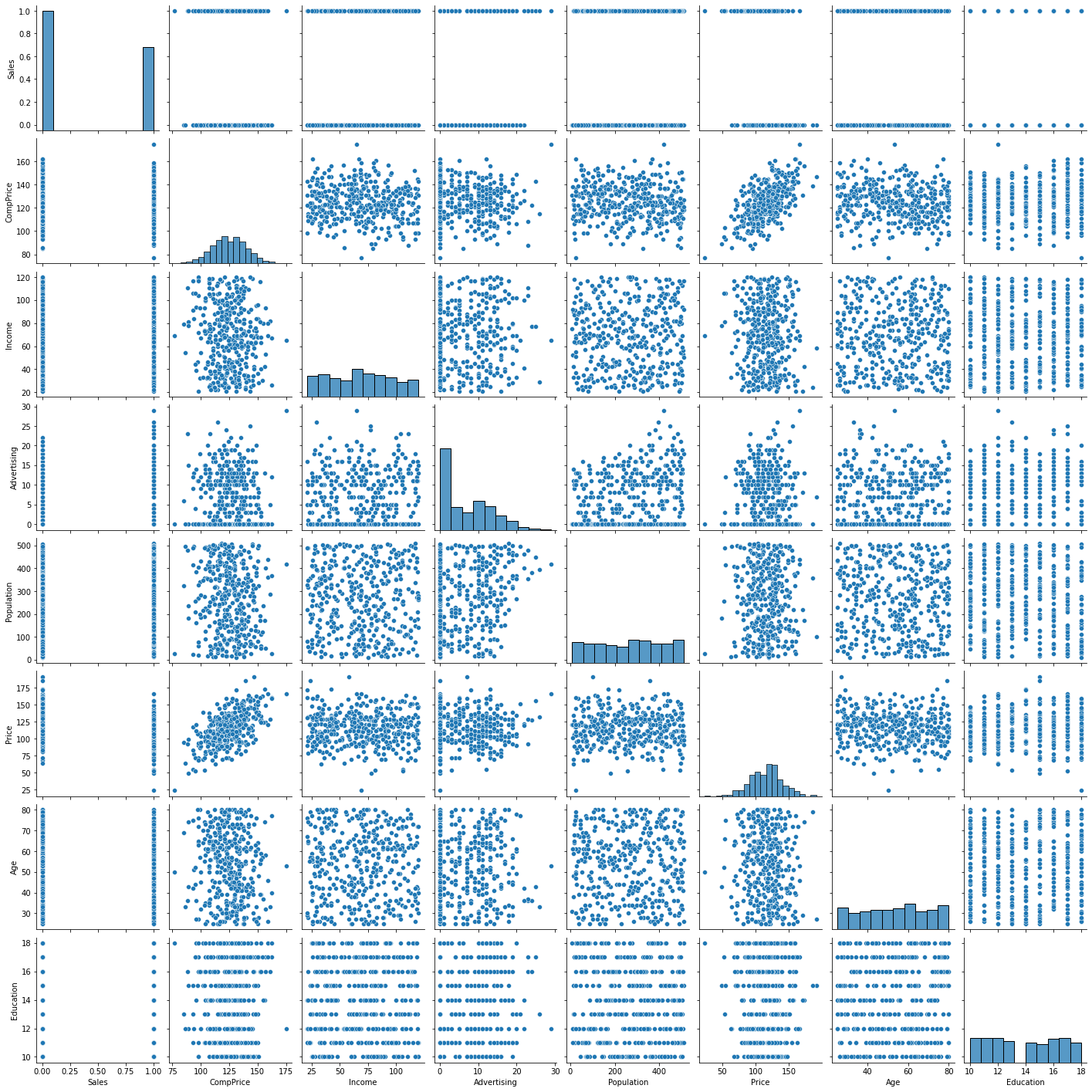
1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| Sales | Sales of a company | Continuous | Relevant |
| CompPrice | Market price | Discrete | Relevant |
| Income | Company income | Discrete | Relevant |
| Advertising | Advertising cost | Discrete | Relevant |
| Population | Total Population | Discrete | Relevant |
| Price | Company price | Discrete | Relevant |
| ShelveLoc | Shelving location | Char | Relevant |
| Age | Age of customers | Discrete | Relevant |
| Education | Education of customers | Discrete | Relevant |
| Urban | Town type | Char | Relevant |
| US | United states | Char | Relevant |

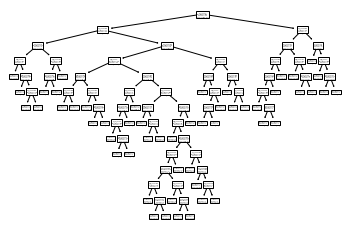
1. DATA CLEANSING :-
2. Dataset consists of 11 colums and 400 rows
3. Duplicate row does not exists
4. All data types are of form int64 ,float64 and object
5. No null values found in each column
6. From describe function mean , median and standard deviation obtained
7. Unique values for sales column obtained
8. Sales column converted to 0 and 1 as values <= 8 converted to 0 and remaining converted to 1
9. Outliers detected but retained
10. There is non linear relationship between columns
11. Scaling is done through normalization techniques
12. EDA:-
13. From box plot



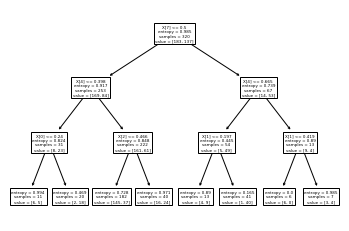
1. Pair plot analysis



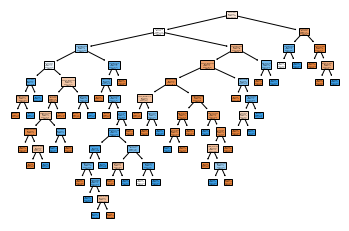
1. MODEL BUILDING:-
2. Decision tree
3. Splitting data to train part for 80% and test part for 20%
4. Test accuracy = 0.7625
5. Train accuracy = 1
6. Precision = 0.7547
7. Recall = 0.86



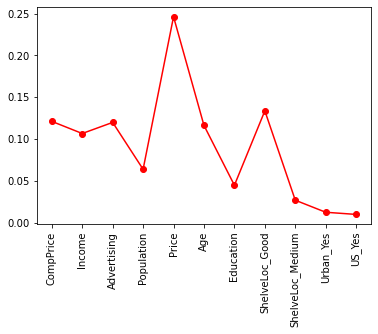
1. **Tuning model by max\_depth = 3**
2. Test accuracy = 0.7375
3. Train accuracy = 0.7875
4. Precision = 0.7547
5. Recall = 0.833



1. Random forest:-
2. Splitting data to train part for 80% and test part for 20%
3. Test accuracy = 0.8375
4. Train accuracy = 1.0
5. Precision = 0.8679
6. Recall = 0.88
7. **Tuning model for best estimator**
8. **Tree with best estimator = 100 , max features = 6 , min sample split = 3**
9. Test accuracy = 0.85
10. Train accuracy = 1
11. Precision = 0.8867
12. Recall = 0.8867



1. Importance of features :-

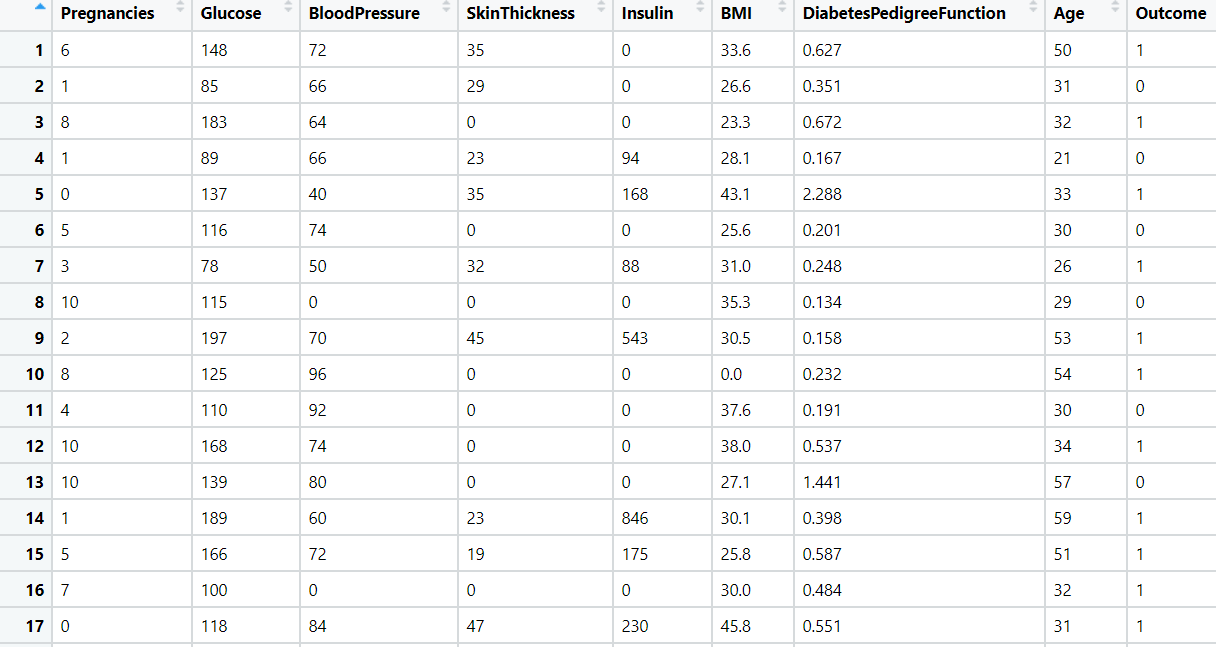


1. Output : -

Model is right fit as both train and test accuracies are highier side

1. BENEFITS :-

From above information we can assign sales for given input features .



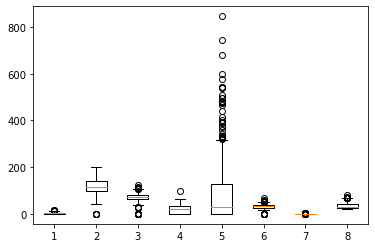
1. Divide the diabetes data into train and test datasets and build a Random Forest and Decision Tree model with Outcome as the output variable.
2. BUSINESS OBJECTIVE:-

Maximize mapping of target

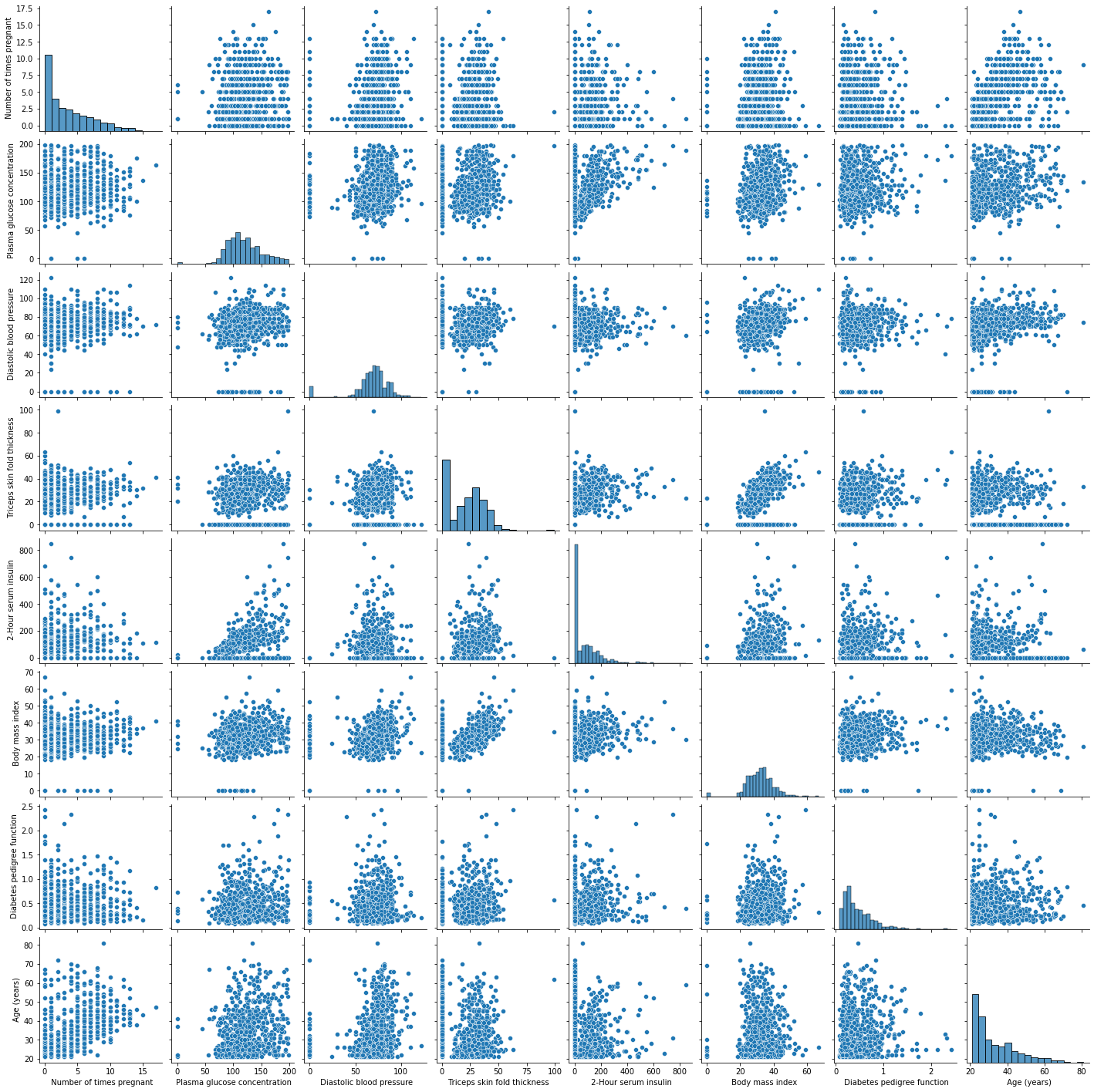
1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| Number of times pregnant | No of times pregnancy | Discrete | Relevant |
| Plasma glucose concentration | Plasma glucose concentration in blood | Discrete | Relevant |
| Diastolic blood pressure | Diastolic blood pressure | Discrete | Relevant |
| Triceps skin fold thickness | Triceps skin fold thickness | Discrete | Relevant |
| 2-Hour serum insulin | Type of insulin | Discrete | Relevant |
| Body mass index | Body mass index | Continuous | Relevant |
| Diabetes pedigree function | Diabeties pedigree function | Continuous | Relevant |
| Age (years) | Age of customers | Discrete | Relevant |
| Class variable | Target variabale | Char | Relevant |

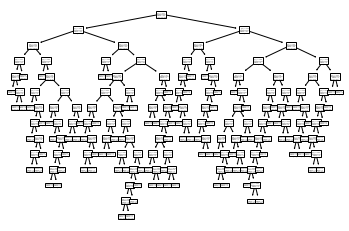
1. DATA CLEANSING :-
2. Dataset consists of 9 colums and 768 rows
3. Duplicate row does not exists
4. All data types are of form int64 ,float64 and object
5. No null values found in each column
6. From describe function mean , median and standard deviation obtained
7. Outliers detected but retained
8. There is non linear relationship between columns
9. Scaling is done through normalization techniques
10. EDA:-
11. From box plot



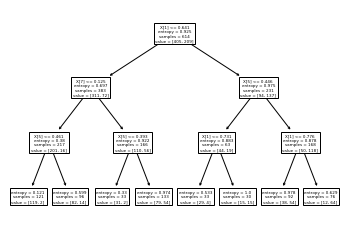
1. Pair plot analysis



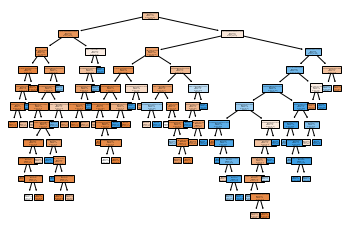
1. MODEL BUILDING:-
2. Decision tree
3. Splitting data to train part for 80% and test part for 20%
4. Test accuracy = 0.7467532467532467
5. Train accuracy = 1
6. Precision = 0.8210
7. Recall = 0.78



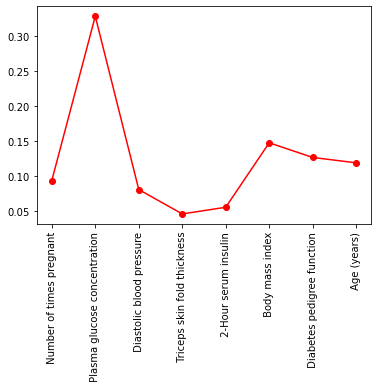
1. **Tuning model by max\_depth = 3**
2. Test accuracy = 0.7792207792207793
3. Train accuracy = 0.7703583061889251
4. Precision = 0.9261
5. Recall = 0.7652



1. Random forest:-
2. Splitting data to train part for 80% and test part for 20%
3. Test accuracy = 0.7662337662337663
4. Train accuracy = 1.0
5. Precision = 0.8947
6. Recall = 0.7657
7. **Tuning model for best estimator**
8. **Tree with best estimator = 20 , max features = 5 , min sample split = 7**
9. Test accuracy = 0.7857142857142857
10. Train accuracy = 0.9576547231270358
11. Precision = 0.9052
12. Recall = 0.7818



1. Importance of features :-



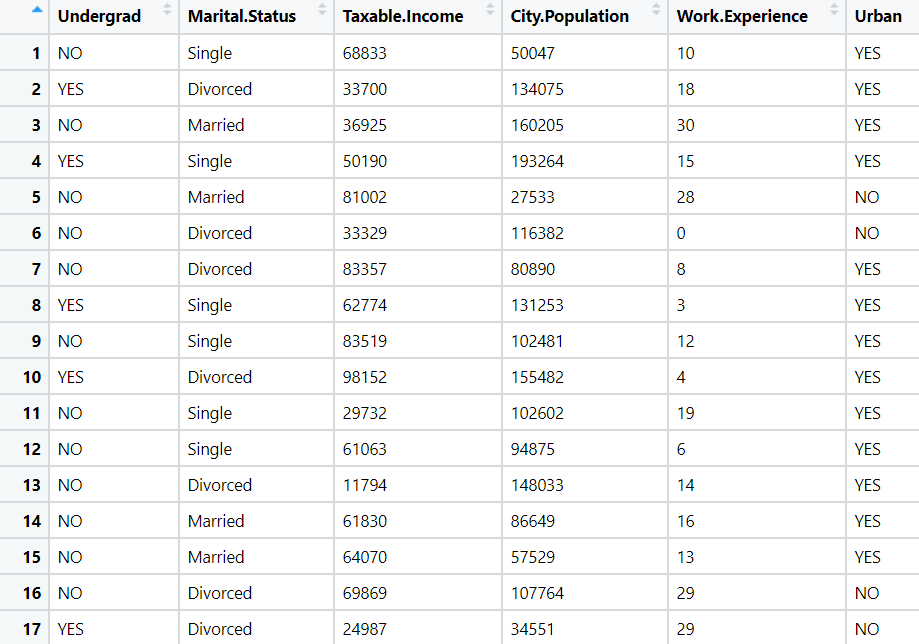
1. Output : -

Model is right fit as both train and test accuracies are highier side

1. BENEFITS :-

From above information we can assign target for given input features .





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).



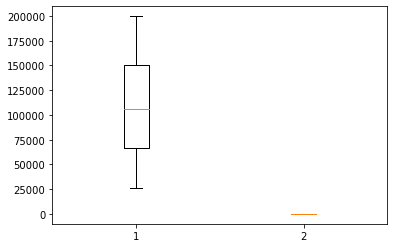
1. BUSINESS OBJECTIVE:-

Maximize mapping of taxable income

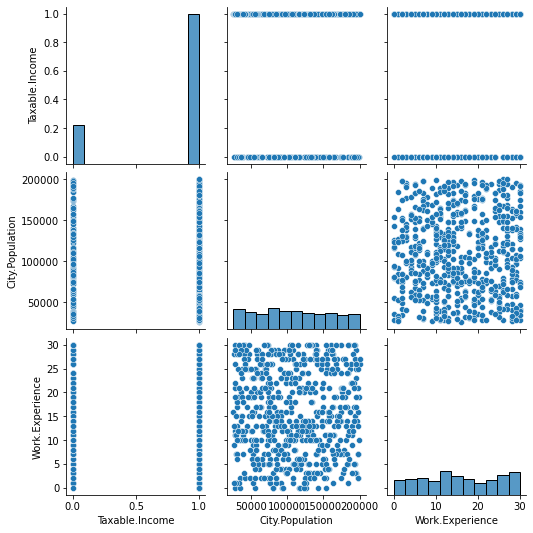
1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| Undergrad | Undergraduate or not | Char | Relevant |
| Marital.Status | Married or not | Char | Relevant |
| Taxable.Income | Taxable income | Continuous | Relevant |
| City.Population | Population of a city | Discrete | Relevant |
| Work.Experience | Experience of work | Discrete | Relevant |
| Urban | Urban or not | Char | Relevant |

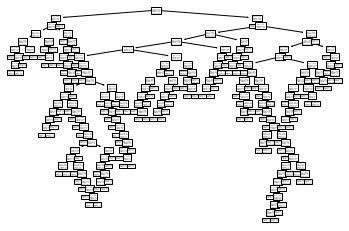
1. DATA CLEANSING :-
2. Dataset consists of 6 colums and 600 rows
3. Duplicate row does not exists
4. All data types are of form int64 and object
5. No null values found in each column
6. From describe function mean , median and standard deviation obtained
7. Outliers not detected
8. There is non linear relationship between columns
9. Scaling is done through normalization techniques
10. EDA:-
11. From box plot



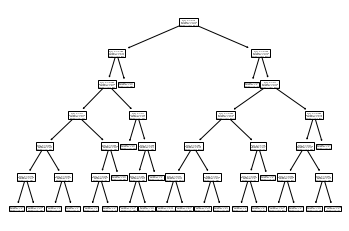
1. Pair plot analysis



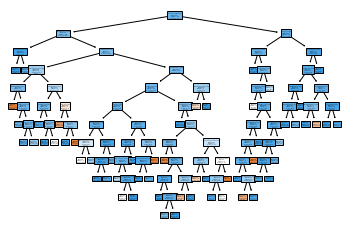
1. MODEL BUILDING:-
2. Decision tree
3. Splitting data to train part for 80% and test part for 20%
4. Test accuracy = 0.6666666666666666
5. Train accuracy = 1
6. Precision = 0.1428
7. Recall = 0.2



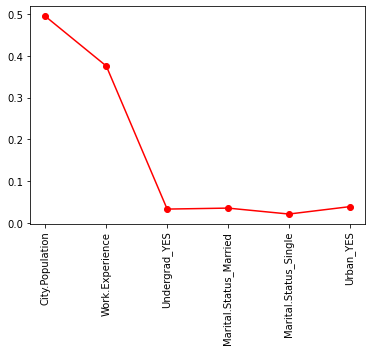
1. **Tuning model by max\_depth = 6**
2. Test accuracy = 0.7333333333333333
3. Train accuracy = 0.8270833333333333
4. Precision = 0.035
5. Recall = 0.166



1. Random forest:-
2. Splitting data to train part for 80% and test part for 20%
3. Test accuracy = 0.7
4. Train accuracy = 1.0
5. Precision = 0
6. Recall = 0
7. **Tuning model for best estimator**
8. **Tree with best estimator = 30 , max features = 2 , min sample split = 7**
9. Test accuracy = 0.7416666666666667
10. Train accuracy = 0.8583333333333333
11. Precision = 0
12. Recall = 0



1. Importance of features :-



1. Output : -

Model is right fit as both train and test accuracies are highier side

1. BENEFITS :-

From above information we can assign taxable income for given input features .

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

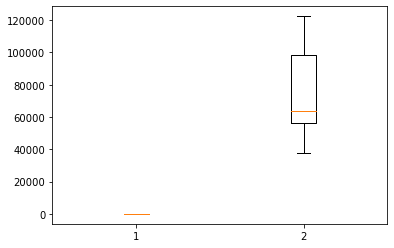
1. BUSINESS OBJECTIVE:-

Maximize mapping of salary

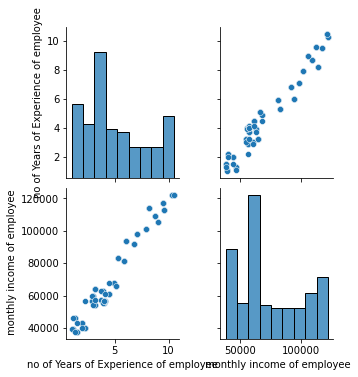
1. DATA UNDERSTANDING:-

|  |  |  |  |
| --- | --- | --- | --- |
| NAME OF FEATURE | DESCRIPTION | TYPE | RELEVANCE |
| Position of the employee | Designation of employees | Char | Relevant |
| No of years of experience of employee | Experience of employees | Continuous | Relevant |
| Monthly income of employee | Income of employee | Continuous | Relevant |

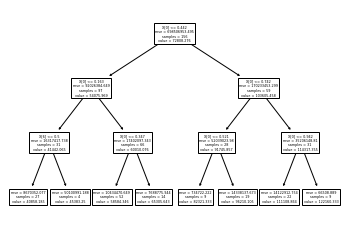
1. DATA CLEANSING :-
2. Dataset consists of 3 colums and 196 rows
3. Duplicate exists and removed
4. All data types are of form int64 , float64 and object
5. No null values found in each column
6. From describe function mean , median and standard deviation obtained
7. Outliers not detected
8. There is linear relationship between columns
9. Scaling is done through normalization techniques
10. EDA:-
11. From box plot



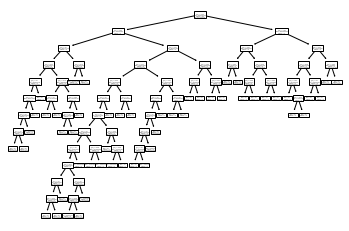
1. Pair plot analysis



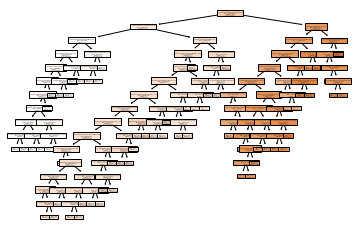
1. MODEL BUILDING:-
2. Decision tree regressor
3. Splitting data to train part for 80% and test part for 20%
4. mse = 12152584.477225484
5. r^2 = 0.9831831324915143



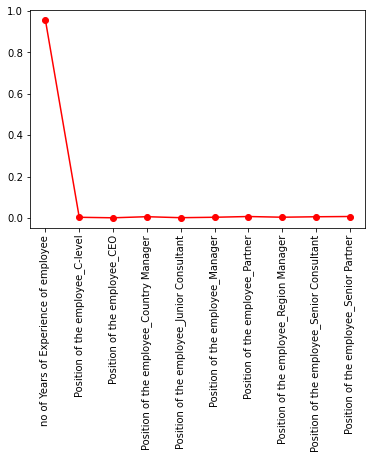
1. **Tuning model by min\_sample\_split = 4**
2. mse = 1457097.6517094017
3. r^2 = 0.9979139826104526



1. Random forest regressor :-
2. Splitting data to train part for 80% and test part for 20%
3. mse = 1579362.4666736298
4. r^2 value = 0.997738945247757
5. **Tuning model for best estimator**
6. **Tree with best estimator = 10 , max features = 2 , min sample split = 2**
7. mse = 52549257.39020388
8. r^2 = 0.9272818139337775



1. Importance of features :-



1. Output : -

Model is right fit as both train and test accuracies are highier side

1. BENEFITS :-

From above information we can assign salary for given input features .