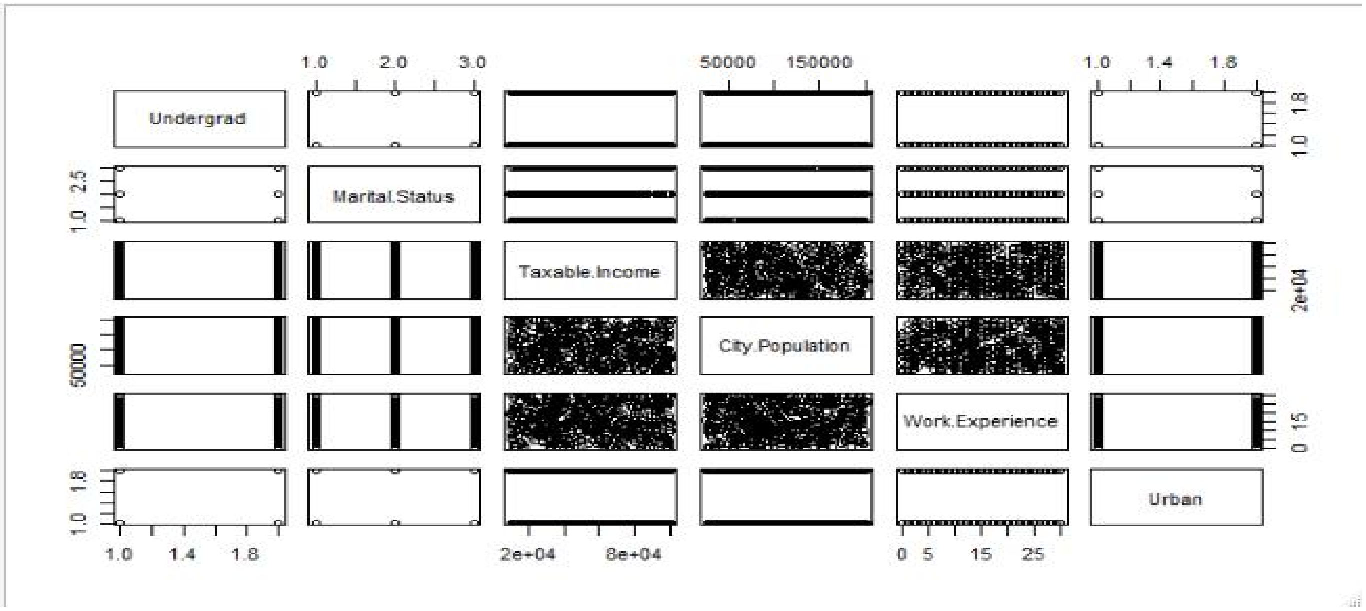
Decision Tree

Example-Fraud Check Dataset

```
'data.frame':	600 obs. of  6 variables:
 $ Undergrad      : Factor w/ 2 levels "NO","YES": 1 2 1 2 1 1 1 2 1 2 ...
 $ Marital.Status : Factor w/ 3 levels "Divorced","Married",..: 3 1 2 3 2 1 1
3 3 1 ...
 $ Taxable.Income : int  68833 33700 36925 50190 81002 33329 83357 62774 8351
9 98152 ...
 $ City.Population: int  50047 134075 160205 193264 27533 116382 80890 131253
102481 155482 ...
 $ Work.Experience: int  10 18 30 15 28 0 8 3 12 4 ...
 $ Urban          : Factor w/ 2 levels "NO","YES": 2 2 2 2 1 1 2 2 2 2 ...
```

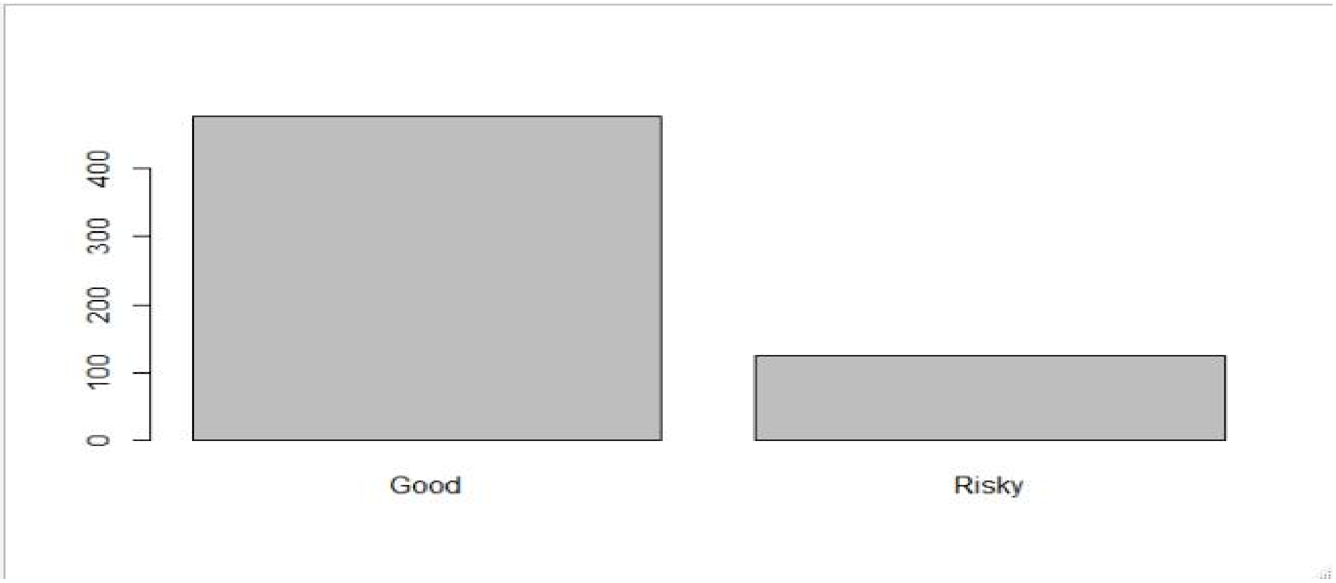
In the above data frame 3 variables are factors and rest all are numeric and target variable is Taxable.Income

Now we create another variable type, which is factor and contain desired results Good or Risky.



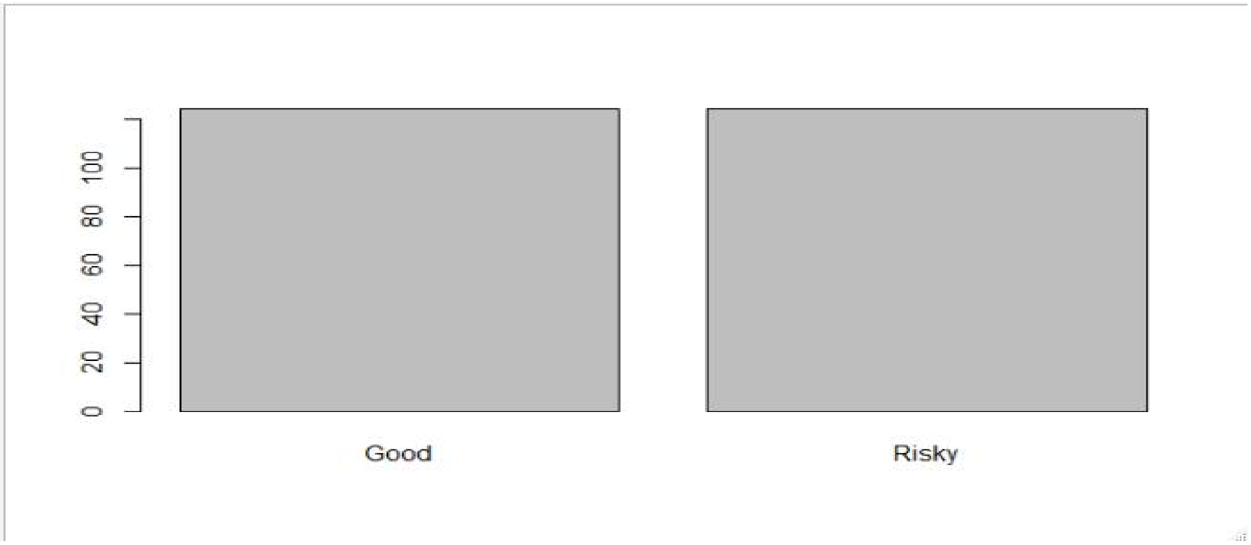
From the pairs plot, none of variable is correlated with our target variable Taxable.Income and uniform distributed scatter plots between all the numeric variable.

Treatment With Imbalanced Data →



```
 Good Risky
  476   124
```

From above plot, our target variable is imbalanced, so we will make ratio equal as 1.



Now our data is equal in ratio.

Model-1 ➔ Using “ctree” function from library “party” with whole data

Confusion Matrix

```
          Actual
Predicted Good Risky
    Good   476   124
    Risky    0     0
```

Accuracy ➔ 0.79333

Model-2 ➔ Using “ctree” function from library”party” with balanced data

Confusion Matrix

```
          Actual
Predicted Good Risky
    Good   124   124
    Risky    0     0
```

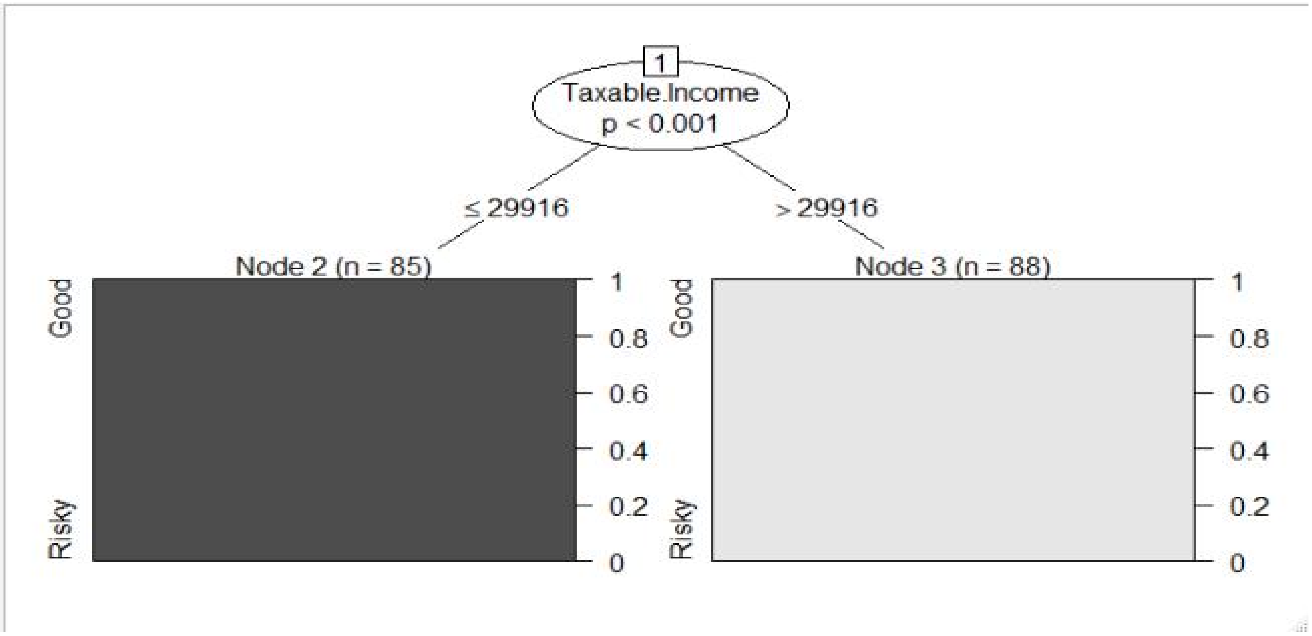
Accuracy ➔ 0.5

Model-3 ➔ Using “ctree” function from library “party” with train data

Confusion Matrix

```
        Predicted
Actual  Good Risky
  Good    36     0
  Risky    1    38
```

Accuracy ➔0.9866



Whole tree is dependent upon the root node itself i.e Taxable.Income, so this classification model is unreliable.

Due to lack of relevant information we can infer that unless and until relevant variable is not introduced then it won't perform well, so all the classification will be biased to one side.