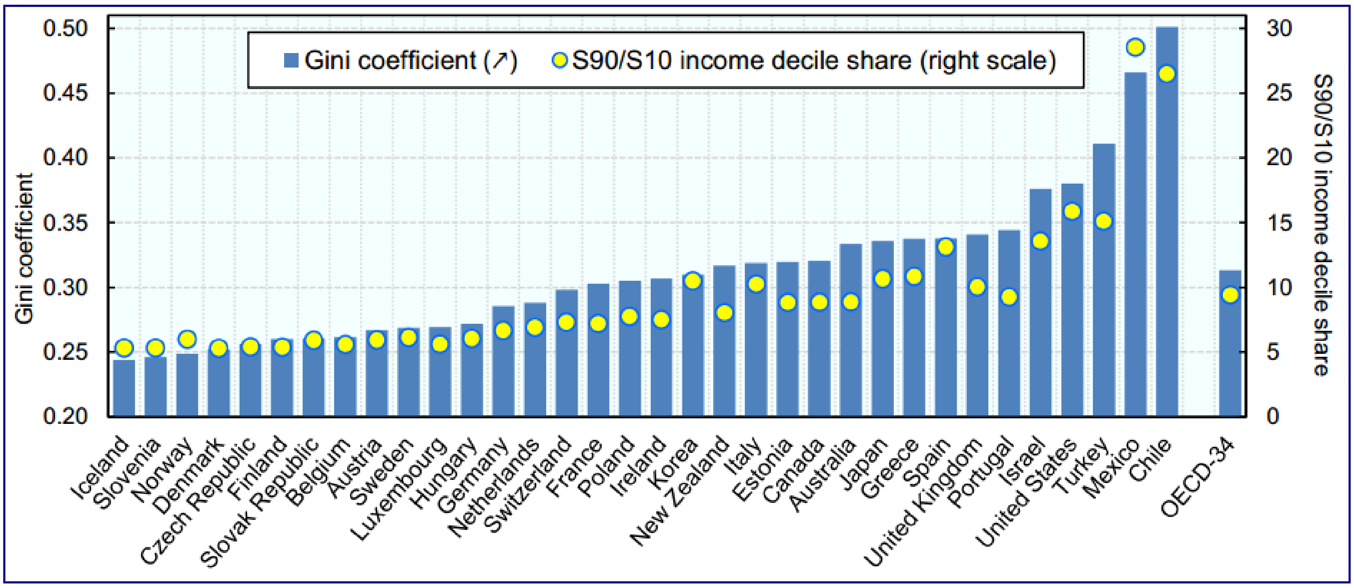
## Case Study

Following is a distribution of Annual income Gini Coefficients across different countries :



Mexico has the second highest Gini coefficient and hence has a very high segregation in annual income of rich and poor. Our task is to come up with an accurate predictive algorithm to estimate annual income bracket of each individual in Mexico. The brackets of income are as follows :

1. Below $40,000

2. $40,000 – 150,000

3. More than $150,000

Following are the information available for each individual :

1. Age , 2. Gender, 3. Highest educational qualification, 4. Working in Industry, 5. Residence in Metro/Non-metro

We need to come up with an algorithm to give an accurate prediction for an individual who has following traits:

1. Age : 35 years , 2, Gender : Male , 3. Highest Educational Qualification : Diploma holder, 4. Industry : Manufacturing, 5. Residence : Metro

We will only talk about random forest to make this prediction in this article.

## The algorithm of Random Forest

Random forest is like bootstrapping algorithm with Decision tree (CART) model. Say, we have 1000 observation in the complete population with 10 variables. Random forest tries to build multiple CART model with different sample and different initial variables. For instance, it will take a random sample of 100 observation and 5 randomly chosen initial variables to build a CART model. It will repeat the process (say) 10 times and then make a final prediction on each observation. Final prediction is a function of each prediction. This final prediction can simply be the mean of each prediction.

## Back to Case study

***Disclaimer : The numbers in this article are illustrative***

Mexico has a population of 118 MM. Say, the algorithm Random forest picks up 10k observation with only one variable (for simplicity) to build each CART model. In total, we are looking at 5 CART model being built with different variables. In a real life problem, you will have more number of population sample and different combinations of input variables.

**Salary bands :**

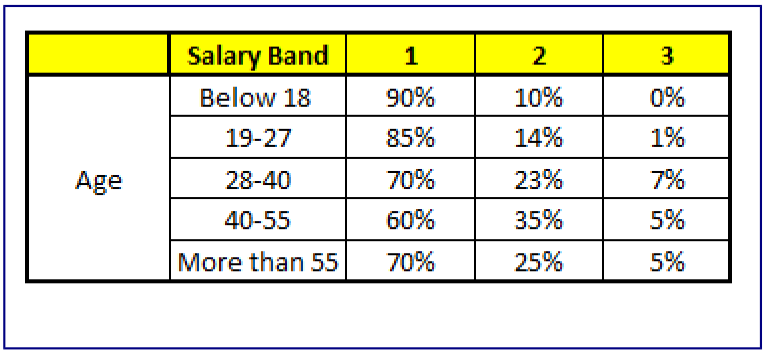
Band 1 : Below $40,000

Band 2: $40,000 – 150,000

Band 3: More than $150,000

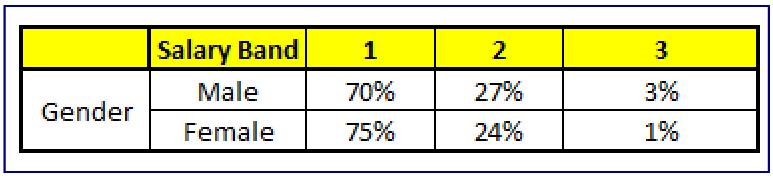
Following are the outputs of the 5 different CART model.

**CART 1 : Variable Age**



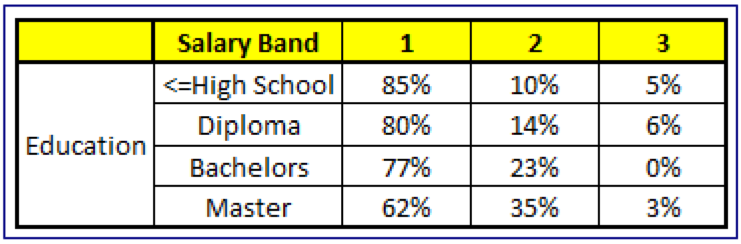
| | Salary Band | 1 | 2 | 3 |
|---|---|---|---|---|
| Age | Below 18 | 90% | 10% | 0% |
| | 19-27 | 85% | 14% | 1% |
| | 28-40 | 70% | 23% | 7% |
| | 40-55 | 60% | 35% | 5% |
| | More than 55 | 70% | 25% | 5% |

**CART 2 : Variable Gender**



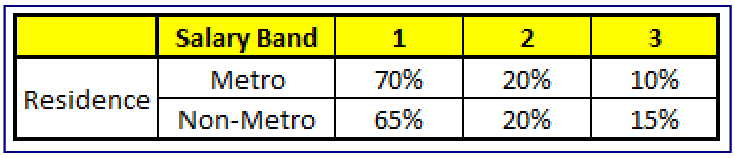
| | Salary Band | 1 | 2 | 3 |
|---|---|---|---|---|
| Gender | Male | 70% | 27% | 3% |
| | Female | 75% | 24% | 1% |

**CART 3 : Variable Education**



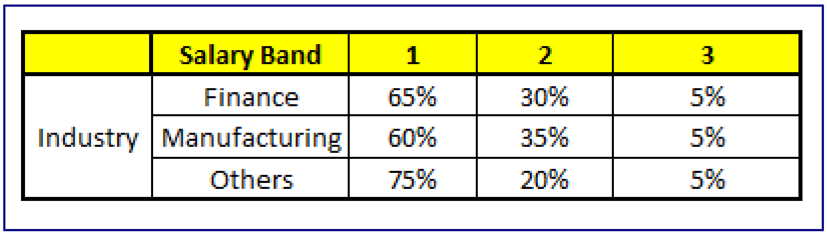
| | Salary Band | 1 | 2 | 3 |
|---|---|---|---|---|
| Education | <=High School | 85% | 10% | 5% |
| | Diploma | 80% | 14% | 6% |
| | Bachelors | 77% | 23% | 0% |
| | Master | 62% | 35% | 3% |

**CART 4 : Variable Residence**



|  | Salary Band | 1 | 2 | 3 |
|---|---|---|---|---|
| Residence | Metro | 70% | 20% | 10% |
|  | Non-Metro | 65% | 20% | 15% |

**CART 5 : Variable Industry**

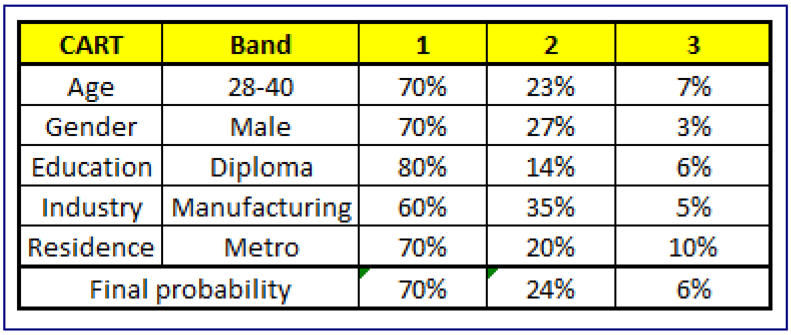


|  | Salary Band | 1 | 2 | 3 |
|---|---|---|---|---|
| Industry | Finance | 65% | 30% | 5% |
|  | Manufacturing | 60% | 35% | 5% |
|  | Others | 75% | 20% | 5% |

Using these 5 CART models, we need to come up with singe set of probability to belong to each of the salary classes. For simplicity, we will just take a mean of probabilities in this case study. Other than simple mean, we also consider vote method to come up with the final prediction. To come up with the final prediction let's locate the following profile in each CART model :

1. Age : 35 years , 2, Gender : Male , 3. Highest Educational Qualification : Diploma holder, 4. Industry : Manufacturing, 5. Residence : Metro

For each of these CART model, following is the distribution across salary bands :



| CART | Band | 1 | 2 | 3 |
|---|---|---|---|---|
| Age | 28-40 | 70% | 23% | 7% |
| Gender | Male | 70% | 27% | 3% |
| Education | Diploma | 80% | 14% | 6% |
| Industry | Manufacturing | 60% | 35% | 5% |
| Residence | Metro | 70% | 20% | 10% |
| Final probability |  | 70% | 24% | 6% |

The final probability is simply the average of the probability in the same salary bands in different CART models. As you can see from this analysis, that there is 70% chance of this individual falling in class 1 (less than $40,000) and around 24% chance of the individual falling in class 2.