**Background**

In the city of Pearsonville, two candidates are campaigning to become mayor. Representing Party A is Culter, a misogynist who has made derogatory comments towards women throughout his campaign but who has high support from men. Representing Party B is Breeman, who has made increasing the minimum wage his fundamental agenda and has good support from both men and women in the city. Seeing that he is behind in the polls, and even more so for women votes, Culter has hired you to create a model that can tell him how many undecided women voters he can win with his current rhetoric.

**Data description**

You are given two datasets i.e., the train and test datasets, which contain demographic data for woman voters only. There are a total of 9 predictor variables and the response variable is Party\_voted\_for. A description of the variables is given below.

|  |
| --- |
| 1. Wife\_age (numerical) |
| 2. Wife\_education (categorical) 1=low, 2, 3, 4=high |
| 3. Husband\_education (categorical) 1=low, 2, 3, 4=high |
| 4. Number\_of\_children\_ever\_born (numerical) |
| 5. Wife\_religion (binary) 0=Minority, 1=Majority |
| 6. Wife\_working (binary) 0=Yes, 1=No |
| 7. Husband\_occupation (categorical) 1, 2, 3, 4 |
| 8. Standard\_of\_living\_index (categorical) 1=low, 2, 3, 4=high |
| 9. Media\_exposure (binary) 0=Good, 1=Not good |
| 10. Party\_voted\_for (class attribute) 0=Party A, 1=Party B |

**Required output**

You are required to submit your R script for evaluation which contains the code for generating the output. Although performance of your models is important, more weight is given to conciseness and clarity of your code so its not required to make multiple iterations for the same type of model. You are free to use whichever packages you deem fit.

**Part 1: Data manipulation and summary statistics**

Q1) For the women in the train dataset find the following variables for every category of education.

1. Count of people
2. Average age
3. Average number of children
4. Percentage of women who are working
5. Percentage of women who have a high standard of living (Standard\_of\_living\_index = 4)

There are 4 categories for Wife\_education and 5 variables required, so you should get 20 values. It is advised to represent all the numbers in a single dataframe.

Use of data manipulation package like **dplyr** would be preferred, however not necessary.

**Part 2: Modelling and predicting**

Q2) Construct a decision tree, random forest and a logistic regression model training your models with the train data. Then use your models to make predictions on the test data and create a confusion matrix for each model (see below). Finally compute the overall accuracy. You are required to consider every predictor when training your models.

**Prediction**

**Party A Party B**

|  |  |
| --- | --- |
|  |  |
|  |  |

**Party A**

**Actual**

**Party B**

*confusion matrix*

**Part 3: Model evaluation**

Q3) For the 3 models trained in Part 2, create their respective ROC curves and display it on one graph. Specify different colours for each graph for ease in comparison. Which is the best model among the three?