R Task - Titanic

Given the **titanic.csv** and **predict.csv** datasets:

1. Using the **titainc.csv** dataset, create two partitions: **train** with 70% of the total records and **test** with the remaining 30%.
2. Before training a model, perform any variable transformation if necessary and briefly explain why.
3. Use the package **rpart** to train a decision tree model on the **train** partition, using the column “Survived” as the outcome.
4. After training the model, report the importance of each variable.
5. Now, test the model on the **test** partition.
6. After testing the model report the overall model accuracy and misclassification rates using a confusion matrix.
7. Use your model to predict the survival of passengers in the **predict.csv** dataset (score the file) and save your results in the same file.
8. Upload your code to a public repository (Github is recommended) and share the project link.

Data Dictionary

|  |  |  |
| --- | --- | --- |
| **Variable** | **Definition** |  |
| Survived | Did the passenger survive? | 0 = No, 1 = Yes |
| pclass | Ticket class | 1 = 1st, 2 = 2nd, 3 = 3rd |
| sex | Sex |  |
| Age | Age in years |  |
| sibsp | # of siblings / spouses aboard the Titanic |  |
| parch | # of parents / children aboard the Titanic |  |
| fare | Passenger fare |  |
| embarked | Port of Embarkation | C = Cherbourg, Q = Queenstown, S = Southampton |

Variable Notes

**pclass**: A proxy for socio-economic status  
1st = Upper  
2nd = Middle  
3rd = Lower

**sibsp**: The dataset defines family relations in this way:  
Sibling = brother, sister, stepbrother, stepsister  
Spouse = husband, wife (mistresses and fiancés were ignored)  
  
**parch**: The dataset defines family relations in this way...  
Parent = mother, father  
Child = daughter, son, stepdaughter, stepson  
Some children travelled only with a nanny, therefore parch=0 for them.