**(ACADEMIC 2017-2018)**

**DEPARTMENT OF INFORMATION TECHNOLOGY**

**BIG DATA PROJECT**

**TITANIC DATASET ANALYSIS**

*Submitted by*

## AKSHAT KUMAR SINGH– RA1511008010045

## 

*Under the guidance of*

**Mrs. K. NIMALA**

**(Asst. Prof SG)**

**SRM UNIVERSITY**

*of*

BACHELOR OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

of

FACULTY OF ENGINEERING AND TECHNOLOGY



S.R.M. Nagar, Kattankulathur, Kancheepuram District

**CONTENT**

* ACKNOWLEDGEMENT
* ABSTRACT
* DATA SET DESCRIPTION
* PROBLEM STATEMENT
* SOURCE CODE
* OUTPUT
* CONCLUSION

**ACKNOWLEDGEMENT**

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

I am highly indebted to SRM University for their guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I would like to express my gratitude towards my parents & member of SRM University for their kind co-operation and encouragement which helped me in completing this project. I would like to express my special gratitude and thanks to the faculty of SRM University, who taught me how to act professionally and handle workload and work under pressure.

My thanks and appreciations also go to my colleagues who extended hands in developing the project, solved my queries from time to time and helped me in building my self-confidence and also to people who have willingly helped me out with their abilities.

**ABSTRACT**

**TITANIC DATASET ANALYSIS**

For this project we have analysed the Titanic data set obtained from Kaggle. Using the data set, one can observe whether 891 passengers of the Titanic survived or perished, and the relationship with several variables, including Age, Sex, passenger class, if they had family on-board the ship, their ticket number, how much they paid for their ticket, where they boarded the ship, and their cabin’s location. To tackle this problem I will apply the conditional inference tree algorithm to train classification models of survival using several of the passengers’ traits. The resulting models will be evaluated using balanced accuracy for my own validation data set, and ultimately I will be submitting the model predictions to the Kaggle competition’s public leader board and receiving a score.

## DATA SET DESCRIPTION AND PROBLEM STATEMENT

To approach the problem of determining who survived or perished in the Titanic accident, we broke the analysis into several parts, including an initial exploratory analysis, and modeling using the conditional inference trees from the party package in R. The exploratory analysis consisted of plotting variables against the ‘Survived’ field to see if there were any clear relationships. We used our insights from that work to determine which variables to include in an initial model. In order to keep the model simple, we settled on three variables: Age, Sex, and Passenger Class (PClass). We ignored any variables that had a large number of missing values.

**Source Code :**

* **Initial Data Loading:**

library(knitr)

library(caret)

library(party)

library(randomForest)

train\_dat <- read.csv(“~/r project/train.csv”)

train\_dat$Survived <- factor(train\_dat$Survived, levels=c(0,1), labels=c('DIED', 'SURVIVED'))

set.seed(42)

train\_idx <- createDataPartition(train\_dat$Survived, p=0.75, list=F, times=1)

model\_dat <- train\_dat[train\_idx,]

val\_dat <- train\_dat[-train\_idx,]

* **Ctree:**

|  |  |
| --- | --- |
|  | set.seed(42)  train\_idx <- createDataPartition(train\_dat$Survived, p=0.75, list=F, times=1)  model\_dat <- train\_dat[train\_idx,]  val\_dat <- train\_dat[-train\_idx,] |

* **Initial Solution Analysis:**

par(mfrow=c(1,1))

simple\_survival\_ctree <- ctree(Survived ~ Age + Sex + Pclass, data=model\_dat, controls=ctree\_control(testtype="Bonferroni"))

plot(simple\_survival\_ctree)

val\_dat$ctree\_preds <- predict(simple\_survival\_ctree, newdata=val\_dat, type='response')

* **Revised Solution:**

better\_survival\_ctree <- ctree(Survived ~ Age + Sex + Pclass + Fare + Parch + SibSp, data=model\_dat,

controls=ctree\_control(testtype="Bonferroni"))

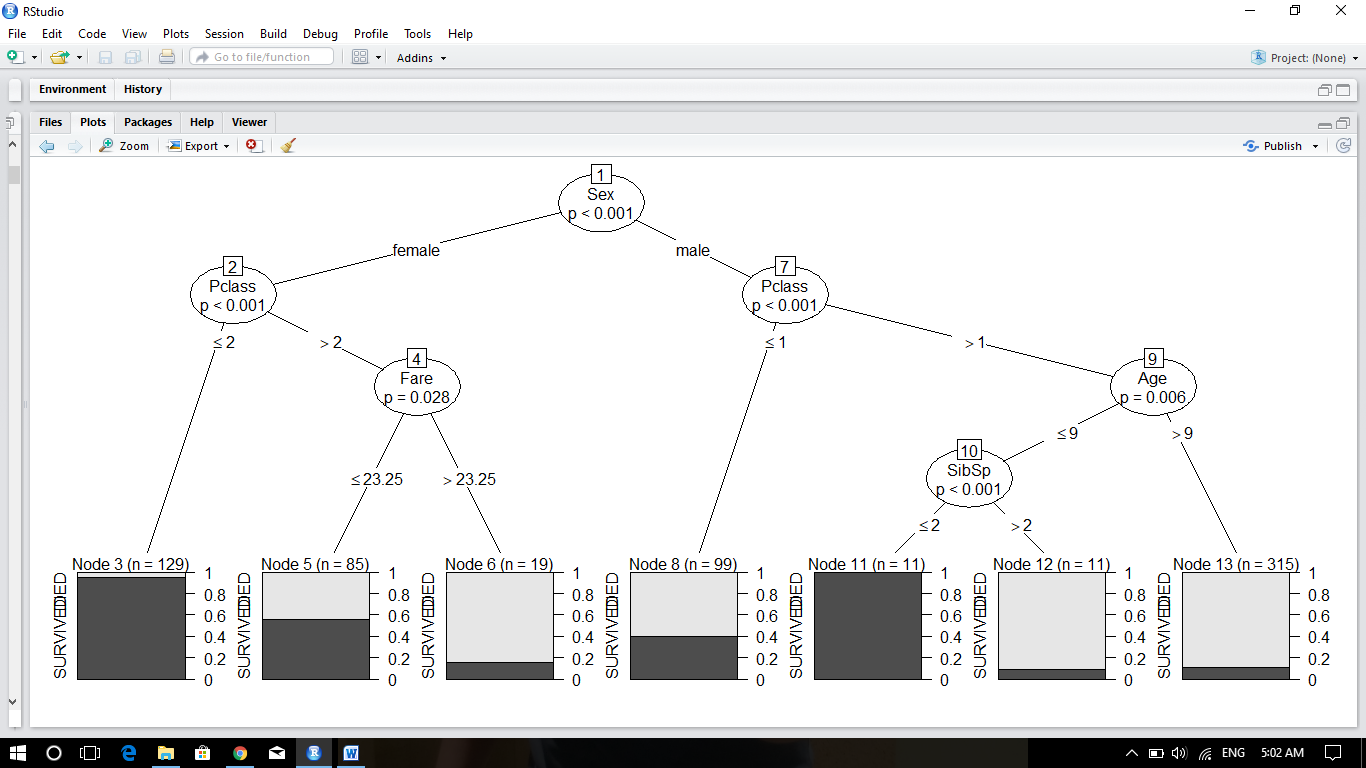
plot(better\_survival\_ctree)

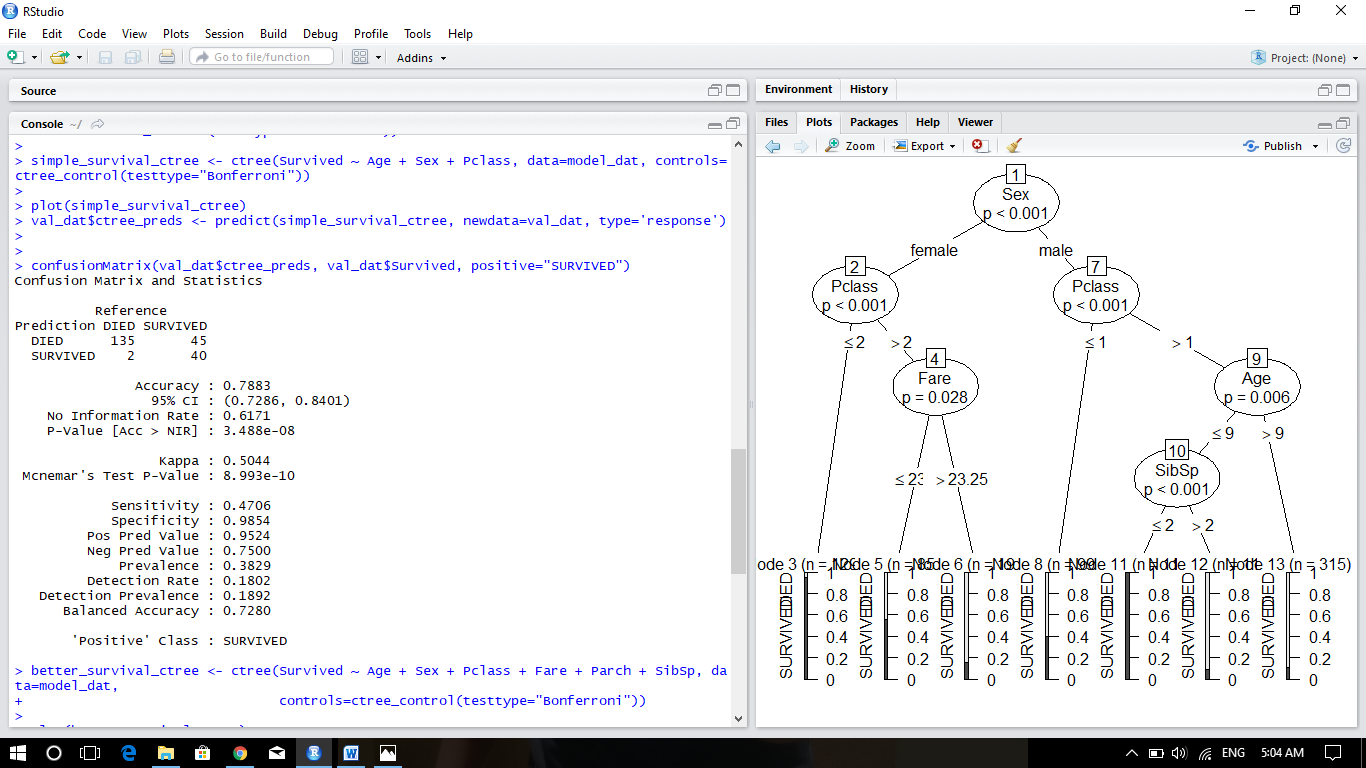
val\_dat$v2\_ctree\_preds <- predict(better\_survival\_ctree, newdata=val\_dat, type='response')

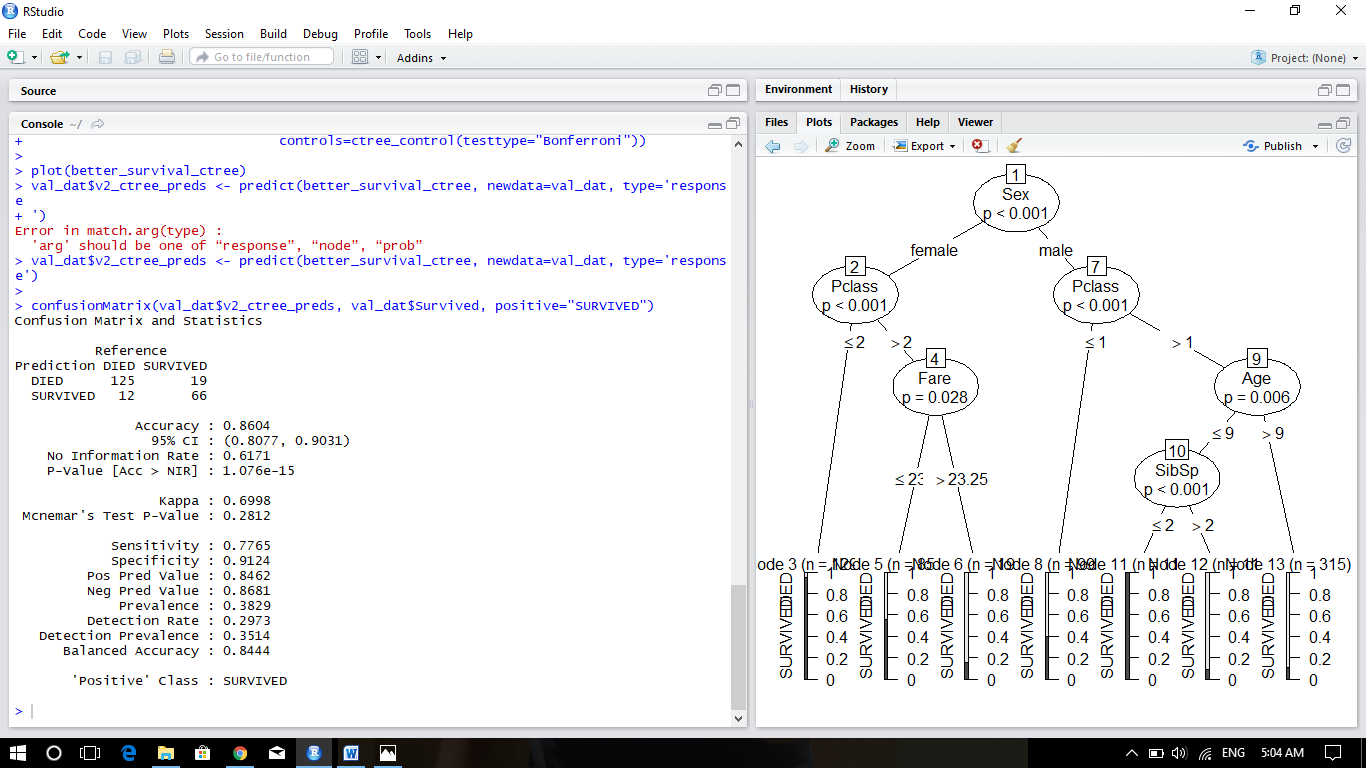
* **Revised Solution Analysis:**

|  |  |
| --- | --- |
|  | confusionMatrix(val\_dat$v2\_ctree\_preds, val\_dat$Survived, positive="SURVIVED") |

**OUTPUT FOR PROBLEM 1:**



****

****

**CONCLUSION:**

Two conditional inference trees were trained to recognize survivors of the Titanic disaster using basic passenger characteristics, and performed reasonably well, ultimately returning a public leaderboard score of 0.79426 on Kaggle, corresponding to a rank of 1201/4431. Unfortunately, both models misclassified a number of male survivors, and even splitting by gender and training individual models on males and females did not result in increased accuracy for men. It is not clear if a more complex algorithm needs to be applied, or whether more feature engineering of Cabins and names needs to be applied to recognize those passengers. However, due to the simplicity and interpretability of the ctree models for this task, we were relatively pleased with the model performance and score on Kaggle.