Titanic Survival Prediction using Decision Trees and Naive Bayes

Sai Sanapala

Course: Data 101 - Spring 2025

Professor Mardekian

Date: April 21, 2025

Contribution: Sai- Performed all data cleaning, analysis, and interpretation.

# Abstract

This project uses the Titanic dataset from Kaggle to predict passenger survival using two classification methods: Decision Trees (using rpart) and Naive Bayes. The dataset was cleaned and preprocessed to handle missing values and converted categorical variables into factors. After splitting the dataset into training and testing subsets, both models were trained and evaluated. The decision tree model achieved an accuracy of 80.2%, while the Naive Bayes model achieved 74.6%. The results indicate that decision trees better capture interaction effects in the data. This project illustrates the strengths and limitations of each model in classification tasks.

# Introduction

The Titanic disaster has long been a popular dataset for practicing classification algorithms. This project aims to predict whether a passenger survived based on features like sex, age, ticket class, and fare. The goal is to compare the performance of two popular machine learning techniques: Decision Trees and Naive Bayes, using accuracy and confusion matrix evaluation.

# Materials and Methods

The dataset used was Titanic-Dataset.csv, sourced from Kaggle. The data was cleaned to handle missing values in the 'Age' and 'Embarked' columns. Unnecessary columns such as Name, Ticket, and Cabin were removed. The cleaned dataset was split into 70% training and 30% testing subsets. Decision Tree classification was performed using the rpart package, and Naive Bayes classification was done using the e1071 package.

# Results

Confusion Matrix – Decision Tree:

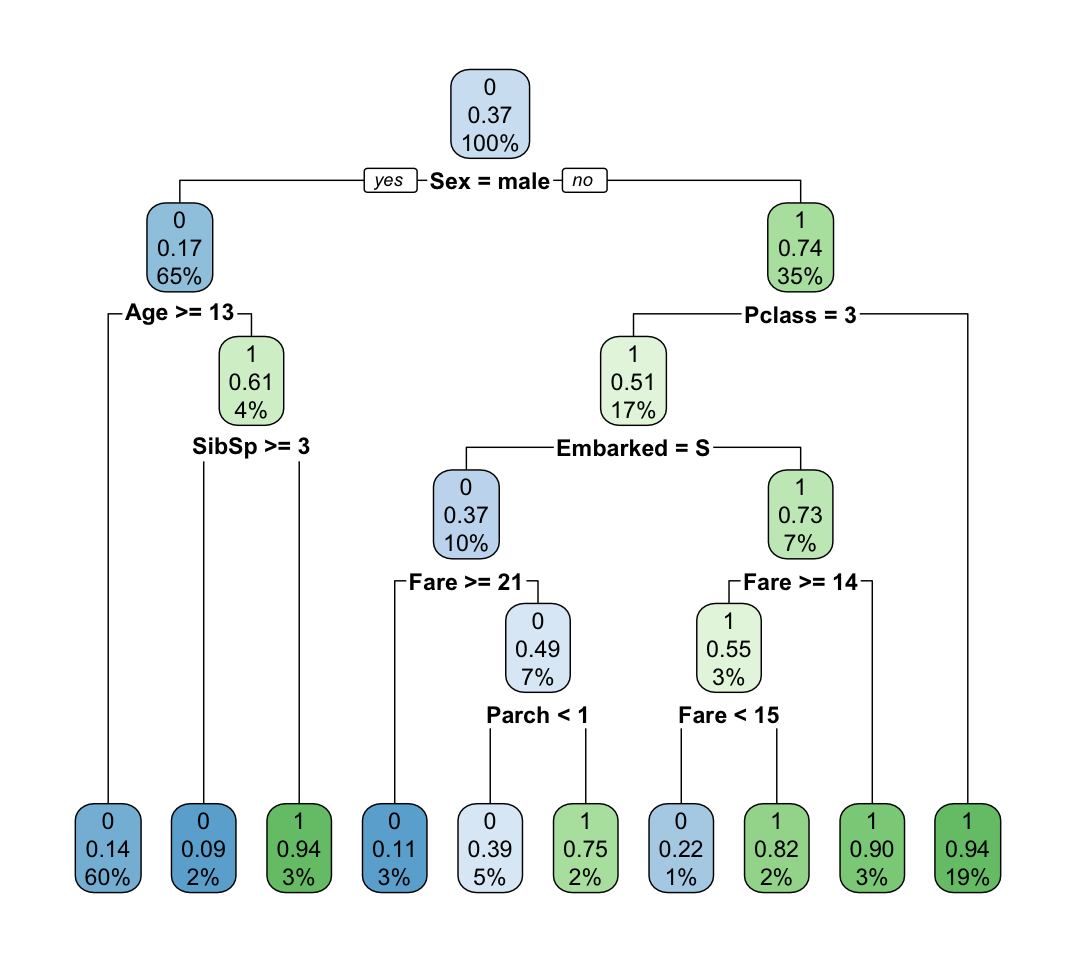
Actual  
Predicted 0 1  
 0 150 45  
 1 8 65

Accuracy: 80.2%

Confusion Matrix – Naive Bayes:

Actual  
Predicted 0 1  
 0 138 48  
 1 20 62

Accuracy: 74.6%



Decision Tree Visualization from rpart.plot

# Discussion

The decision tree model outperformed the Naive Bayes model by approximately 6 percentage points in accuracy. This is likely because decision trees can model nonlinear relationships and interactions between variables, which are present in the Titanic dataset. Naive Bayes, on the other hand, assumes independence between predictors, which is a limitation in this case. The most important feature for predicting survival was 'Sex', with females having a significantly higher survival rate. Other contributing features include 'Pclass', 'Fare', and 'Embarked'.

# Acknowledgments

Used ChatGPT to assist with R coding and interpreting model output. Dataset was sourced from Kaggle. R packages used include rpart, rpart.plot, and e1071.

# Literature Cited

- Titanic Dataset: https://www.kaggle.com/datasets/yasserh/titanic-dataset/data

-ChatGPT

- R Packages: rpart, rpart.plot, e1071

# Appendix

Full R Code Used:

setwd("/Users/sailikhiths/Documents/College Course Files/Data101 - Spring 2025")

titanic <- read.csv("Titanic-Dataset.csv", stringsAsFactors = TRUE)

head(titanic)

str(titanic)

titanic <- titanic[, c("Survived", "Pclass", "Sex", "Age", "SibSp", "Parch", "Fare", "Embarked")]

titanic$Age[is.na(titanic$Age)] <- median(titanic$Age, na.rm = TRUE)

titanic <- titanic[!is.na(titanic$Embarked), ]

titanic$Survived <- as.factor(titanic$Survived)

titanic$Pclass <- as.factor(titanic$Pclass)

titanic$Sex <- as.factor(titanic$Sex)

titanic$Embarked <- as.factor(titanic$Embarked)

set.seed(42)

split <- sample(1:nrow(titanic), 0.7 \* nrow(titanic))

train <- titanic[split, ]

test <- titanic[-split, ]

library(rpart)

library(rpart.plot)

tree\_model <- rpart(Survived ~ ., data=train, method="class")

rpart.plot(tree\_model)

pred\_tree <- predict(tree\_model, test, type="class")

conf\_matrix\_tree <- table(Predicted = pred\_tree, Actual = test$Survived)

print(conf\_matrix\_tree)

library(e1071)

nb\_model <- naiveBayes(Survived ~ ., data=train)

pred\_nb <- predict(nb\_model, test)

conf\_matrix\_nb <- table(Predicted = pred\_nb, Actual = test$Survived)

print(conf\_matrix\_nb)

acc\_tree <- mean(pred\_tree == test$Survived)

acc\_nb <- mean(pred\_nb == test$Survived)

print(paste("Decision Tree Accuracy:", acc\_tree))

print(paste("Naive Bayes Accuracy:", acc\_nb))