````
---
title: "Titanic Survival Prediction"
author: "Anusha preethi"
output: html_notebook
---

```{r}
install.packages(c("tidyverse", "caret", "randomForest", "ggplot2"))

```
```{r}
library(tidyverse)
library(caret)
library(randomForest)
library(ggplot2)

```
```{r}
train_data <- read.csv('C:\\Users\\hp\\Downloads\\Titanic-Dataset.csv',
stringsAsFactors = FALSE)
test_data <- read.csv('C:\\Users\\hp\\Downloads\\Titanic-Dataset.csv',
stringsAsFactors = FALSE)

```

```{r}
head(train_data)
summary(train_data)

```

```{r}
train_data$Age[is.na(train_data$Age)] <- median(train_data$Age, na.rm = TRUE)
train_data$Embarked[is.na(train_data$Embarked)] <- getmode(train_data$Embarked)

```

```{r}
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

```

```{r}
train_data$Title <- gsub('(.*, )|(\\..*)', '', train_data$Name)

```

```{r}
train_data <- train_data %>%
  mutate(Sex = factor(Sex),
         Embarked = factor(Embarked),
````
````

```
         Pclass = factor(Pclass))

```

```{r}
set.seed(42)
train_index <- createDataPartition(train_data$Survived, p = 0.8, list = FALSE)
train_set <- train_data[train_index, ]
val_set <- train_data[-train_index, ]

```

```{r}
# Assuming train_data is loaded and processed as described
train_data$Survived <- factor(train_data$Survived)

# Split into train and validation sets
set.seed(42)
train_index <- createDataPartition(train_data$Survived, p = 0.8, list = FALSE)
train_set <- train_data[train_index, ]
val_set <- train_data[-train_index, ]

# Train the classification model
rf_model <- randomForest(Survived ~ ., data = train_set, ntree = 100,
random_state = 42)

# Make predictions on validation set
val_pred <- predict(rf_model, newdata = val_set)

# Evaluate model performance
confusionMatrix(val_pred, val_set$Survived)



```
```{r}
test_data$Age[is.na(test_data$Age)] <- median(train_data$Age, na.rm = TRUE)
test_data$Embarked[is.na(test_data$Embarked)] <- getmode(train_data$Embarked)
test_data$Fare[is.na(test_data$Fare)] <- median(train_data$Fare, na.rm = TRUE)

test_data <- test_data %>%
  mutate(Sex = factor(Sex),
         Embarked = factor(Embarked),
         Pclass = factor(Pclass))

```



```{r}
# Assuming test_data is loaded and processed as described
test_data$Title <- gsub('(.*, )|(\\..*)', '', test_data$Name)

# Handle missing values in test_data
test_data$Age[is.na(test_data$Age)] <- median(train_data$Age, na.rm = TRUE)
```

```
test_data$Embarked[is.na(test_data$Embarked)] <- getmode(train_data$Embarked)
test_data$Fare[is.na(test_data$Fare)] <- median(train_data$Fare, na.rm = TRUE)

# Convert factors in test_data
test_data <- test_data %>%
  mutate(Sex = factor(Sex),
         Embarked = factor(Embarked),
         Pclass = factor(Pclass))

# Make predictions
test_predictions <- predict(rf_model, newdata = test_data)

# Create submission file
submission <- data.frame(PassengerId = test_data$PassengerId, Survived =
test_predictions)
write.csv(submission, 'submission.csv', row.names = FALSE)


```