TITANIC DATA ANALYSIS AND DECISION TREE IMPLEMENATION key points and work flow process

- EDA
- Using df.info() to understand the dataset
  plotting age distribution using sns
- plotted pclass vs survival rate
- Plotting correlation matrix ( error – unable to plot categorical data)
- So plotted only numeric data
- Handling missing values ,using mean ,mode to impute
- Extracting title from age
- Using title to predict age
- Filling missing fare values with median
- Also filling embarked with mode
- Dropping cabin column as it is not useful
- Feature engineering
- Calculating family siize by using sibling,spouse and parent children data
- Creating a new variable for family size
- One hot encoding
- Dropping column name,ticket ,title as they add no value now
- One hot encoding for sex,embarked,pclass,agegroup,familysizegroup
- Other errors (dropping same column twice and getting column not found error)
- Creating a decision tree and random forest and evaluating
- Output
- Case 1-considering all columns

DT – 79%

RF- 82%

- Case 2- selecting some columns only feature selection( this improved dt accuracy)
- DF-82% accuracy
- Rf- 82% accuracy
- While overall accuracy is same ,precision and recall isn’t same for individual classes