Step 1. Read train.csv

Step 2.

Exclude some columns

Ticket 🡨 no value. Passenger ID is sufficient foreign key

Fare 🡨 no value. May imply class, but there are few passengers with discounted tickets

Cabin 🡨 cabin has missing values and cabin determines deck which is inferred by Pclass

Kept

Embarked 🡨

Old man 🡨

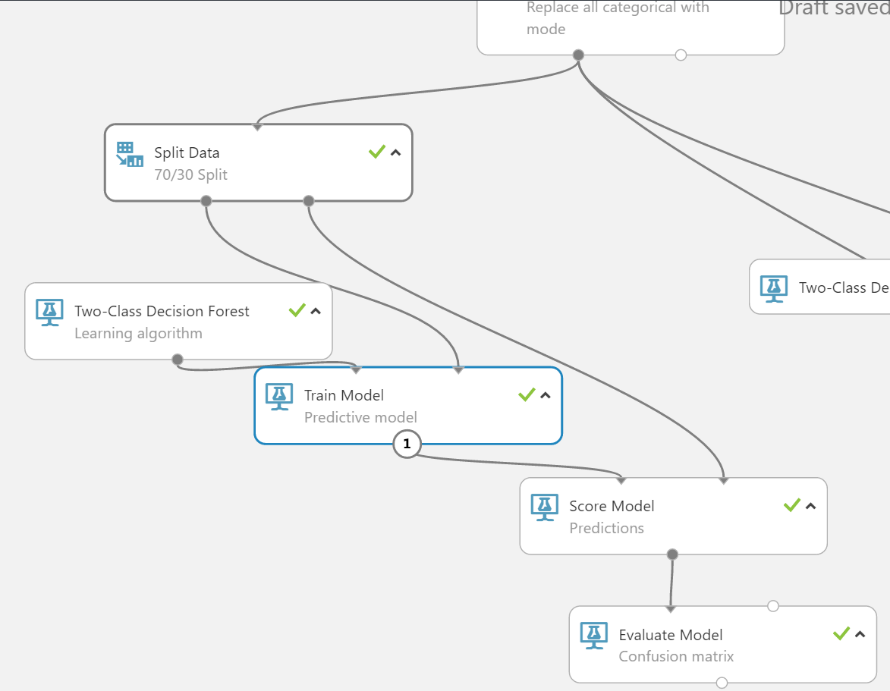
Step 3. Edit Metadata

Categorical casting

Step 4. Clean Missing Data (numeric)

Age for children replaced with mean 6 of known child age

Step 5. Clean Missing data (categorical)

HW 2

Split data (split rows, 0.7)

Decision learning tree/train model

Score train model with split(30) data

Evaluate model

Score train model with blind hold out data

Parameter range

Potential feature?

Is child

Is elderly

Travelling alone

Is child?

Correctly classify missing age data

Classify mother

Take out embarked?

Try taking out the old man from the data (Last step)