RapidMiner Data Mining Software Exercise

HW 3 – Titanic Decision Tree

This assignment will give you the opportunity to learn more about data mining using a decision tree analysis to solve a classification problem. The assignment gives you the chance to predict if you would have survived the sinking of the Titanic.

The Assignment

You will create a decision tree to predict the fate of you and others you know if you had all been on the Titanic. Complete the following steps and report your results as specified in steps 8, 9, and 10.

1. Download the data file "Titanic data.csv" from Blackboard. The file contains 1309 records.
2. Import the CSV file into your RapidMiner repository. When you import the data to RapidMiner ***be sure to specify the correct data types during the import process***. This is how the data is structured:

Column Name	Description	RapidMiner Data Type
pclass	Passengers are listed as being in 1st, 2nd, or 3rd class using a value of 1, 2, or 3.	polynominal
survived	Value is 1 if they survived and 0 if they did not.	binominal
name	The passenger's name. Married women are referred to by their husband's name with their first and maiden names following in parenthesis. For example: Brown, Mrs. James Joseph (Margaret Tobin)	polynominal
sex	The gender of the passenger.	binominal
age	Age in years.	real
ticket	The passenger's ticket number.	polynominal
fare	The fare paid for the ticket in US dollars.	real
embarked	A value of S, C, or Q denoting where the passenger boarded Titanic.	polynominal

3. Use this data to randomly select about 500 passengers to use as a training data set. You can do it in RapidMiner using the **Split Data** operator. Set the partitions to 0.38 (training) and 0.62 (scoring). That will give you 497 and 812 data points, respectively.
4. Create the necessary process stream based upon what you learned in the chapter tutorial.
   a. Set the "survived" attribute's role to be your label.
   b. Select only the data attributes which are predictive. Exclude those that aren't predictive (like names) so they are not considered in the decision tree model.
   c. Add a Decision Tree operator to your stream.
5. Use the remaining passenger data to test your model. How well does the model predict the fates of the other passengers?
6. Create a new file in Excel with the same columns pclass, name, sex, age, and fare - leave out the "survived" column because it will be predicted by the model. Enter data for yourself and some of your friends & family, at least 10 people total (but more is better). The data for some of your friends & family will be straightforward but you will have to decide what to use for values like "pclass." Be

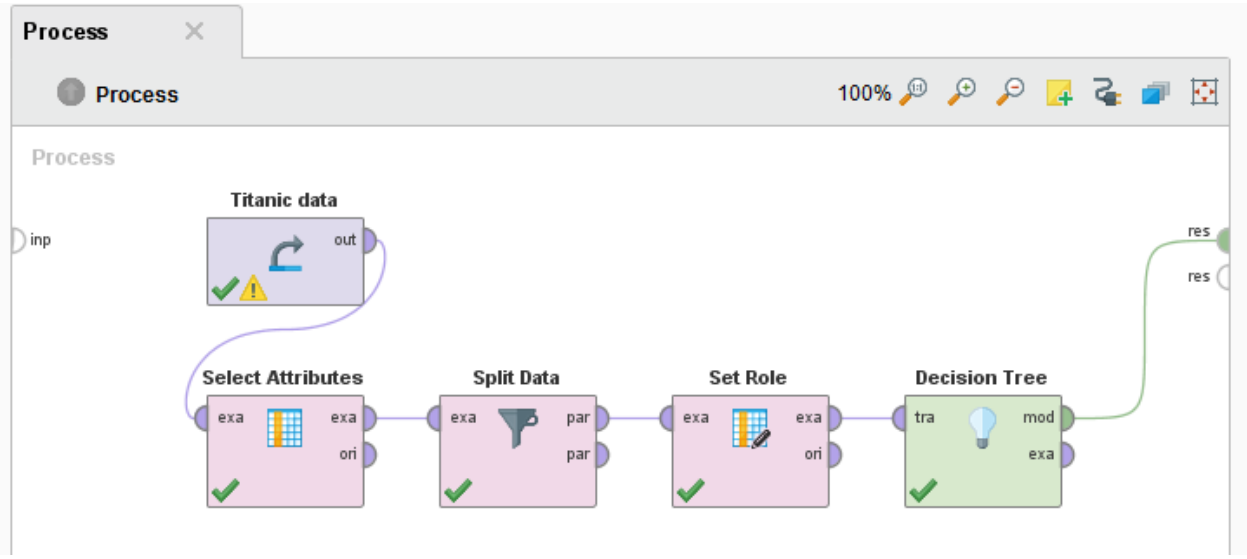
honest – you know whether you (or the people you know) would be willing to pay for a 1st class ticket, would always have to get the cheapest ticket (3rd class) or would pick the middle category (2nd class). For example, I know family members who would always pay for the First Class ticket and others who would never pay for anything more expensive than Third Class. In any case, make sure you have a mix of each class in your data even if you have to put someone where they don't belong!

7. Save this file as a CSV file and import it into your RM repository. Don't forget: The data types will have to be the same as the original Titanic data.
8. Drag this data set into your process and ensure attributes that are not predictive, such as names, will not be included as predictors in the model.
9. Apply your decision tree model to your "friends & family" data set. (Use the Apply Model operator.)
10. Run your model using gain_ratio. Report your tree nodes and discuss whether you and the people you know would have lived or died.
11. Re-run your model using gini_index. Report any differences in your tree's structure. Discuss whether your chances for survival increase or decrease under Gini.
12. Experiment with changing leaf and split sizes and other decision tree algorithm criteria, such as information_gain. Optionally, you can re-create your decision tree using all 1309 records and see if the accuracy of the friends & family predictions changes. Analyze and report your results.

Submit your report in Word or pdf format in Blackboard before the deadline. Save any included screen shots inside the Word document, not in separate files.

==========

Your process after #5 should look something like this:



Your final process should look something like this: