

Tutorial on Building Decision Tree for Kaggle Titanic Competition Using Weka

This tutorial will walk you through the process of preparing Kaggle Titanic data for Weka, and then using Weka's J48 algorithm to build a decision tree to predict Titanic survivors, and how to visualize the decision tree to observe what patterns/rules the decision tree model has learned from the training data. This tutorial will also demonstrate how to use information gain as feature ranking methods to explore the relevance between features and prediction target.

1. Go to [Kaggle](https://www.kaggle.com). Download **test** and **training** data CSV files.



Titanic: Machine Learning from Disaster




Friday, September 28, 2012

Knowledge • 1,429 teams

6 months to go

Wednesday, December 31, 2014

Dashboard

Home 
Data 
Make a submission 

Information

Description
Evaluation
Rules
Prizes
Frequently Asked Questions
Further Reading / Watching
Getting Started With Excel
Getting Started With Python

[Competition Details](#) » [Get the Data](#) » [Make a submission](#)

Data Files

File Name	Available Formats
train	.csv (59.76 kb)
gendermodel	.csv (3.18 kb)
genderclassmodel	.csv (3.18 kb)
test	.csv (27.96 kb)

2. Prepare data files.
 - a. Open **training** CSV data file in MS Excel.
 - b. There are 12 columns (attributes).

train - Microsoft Excel non-commercial use

	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, female		38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen	female	26	0	0	STON/O2.	7.925		S
5	4	1	1	Futrelle, M	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, Mr.	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, M	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy, male		54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, N	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, I	female	27	0	2	347742	11.1333		S
11	10	1	2	Nasser, M	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstrom	female	4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, N	female	58	0	0	113783	26.55	C103	S
14	13	0	3	Saunders	male	20	0	0	A/5. 2151	8.05		S

- c. Open **test** CSV data file in MS Excel.
- d. Add a new column "Survived" in **test** data file.
- e. Populate "Survived" column with question marks ("??") in **test** data.

A1

	A	B	C	D	E	F
	PassengerId	Survived	Pclass	Sex	Age	SibSp
892	?		3	male	34.5	0
893	?		3	female	47	1
894	?		2	male	62	0
895	?		3	male	27	0
896	?		3	female	22	1
897	?		3	male	14	0
898	?		3	female	30	0
899	?		2	male	26	1
900	?		3	female	18	0
901	?		3	male	21	2
902	?		3	male		0
903	?		1	male	46	0
904	?		1	female	23	1
905	?		2	male	63	1
906	?		1	female	47	1

3. Merge training and test data.

- Weka will allow us to read in CSV files. But when we read in one training data and one test data Weka will report an error their formats are not compatible, even though they appear identical. In order to allow Weka to recognize both of them as the same data format, we can copy both training and test examples into one file and let Weka read it in, and convert it to its own arff file. Then we separate it again back into train and test in Weka.
- Now copy all records from **test** CSV(but **not** the first row of headings) into the **training** CSV file.

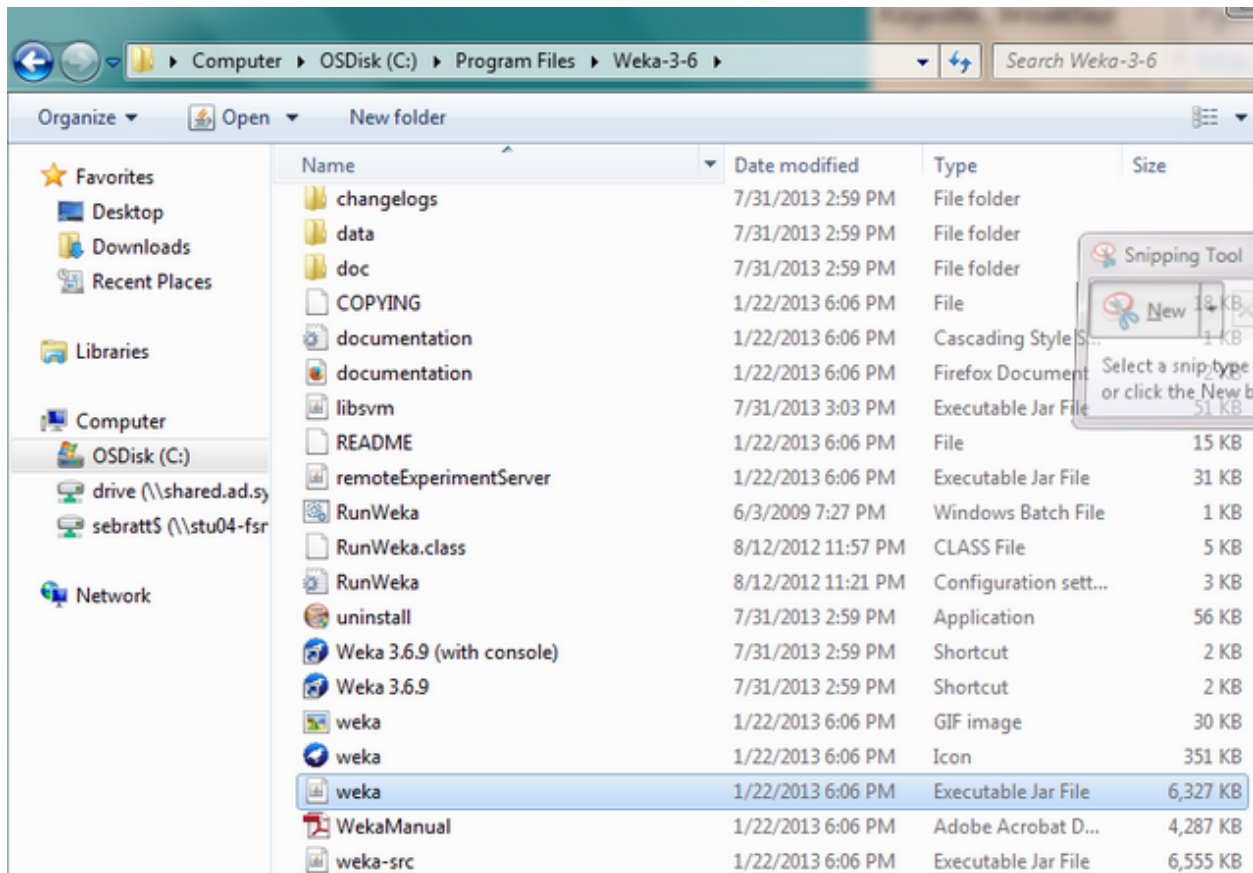
	A	B	C	D	E	F	G	H	I	J	K	L
1	Passenger	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
890	889	0	3	Johnston, female			1	2	W./C. 660	23.45		S
891	890	1	1	Behr, Mr. male		26	0	0	111369	30	C148	C
892	891	0	3	Dooley, M male		32	0	0	370376	7.75		Q
893	892 ?		3	Kelly, Mr. male		34.5	0	0	330911	7.8292		Q
894	893 ?		3	Wilkes, M female		47	1	0	363272	7		S
895	894 ?		2	Myles, Mr male		62	0	0	240276	9.6875		Q
896	895 ?		3	Wirz, Mr. male		27	0	0	315154	8.6625		S
897	896 ?		3	Hirvonen, female		22	1	1	3101298	12.2875		S
898	897 ?		3	Svensson, male		14	0	0	7538	9.225		S
899	898 ?		3	Connolly, female		30	0	0	330972	7.6292		Q
900	899 ?		2	Caldwell, male		26	1	1	248738	29		S
901	900 ?		3	Abraham, female		18	0	0	2657	7.2292		C
902	901 ?		3	Davies, M male		21	2	0	A/4 48871	24.15		S

- Delete unnecessary columns (optional). I got rid of “Name” and “Ticket” because these are not relevant or helpful attributes to our classification task.

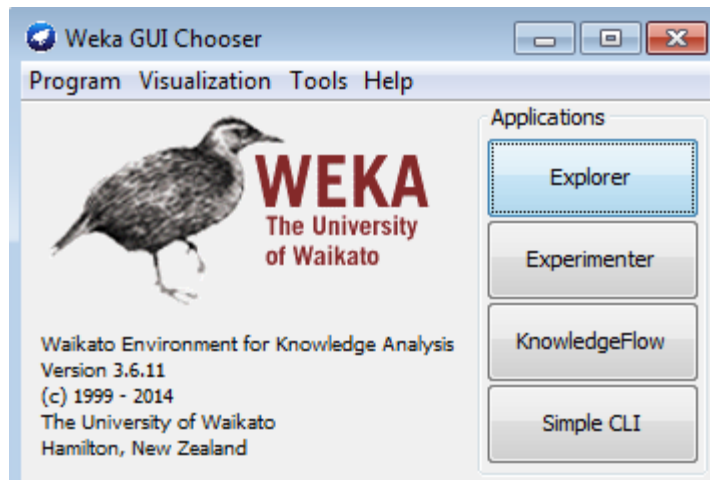
Now we have preprocessed and merged both files. Save this as a new file (I call mine: “TitanicAll.csv”).

4. Read CSV into Weka

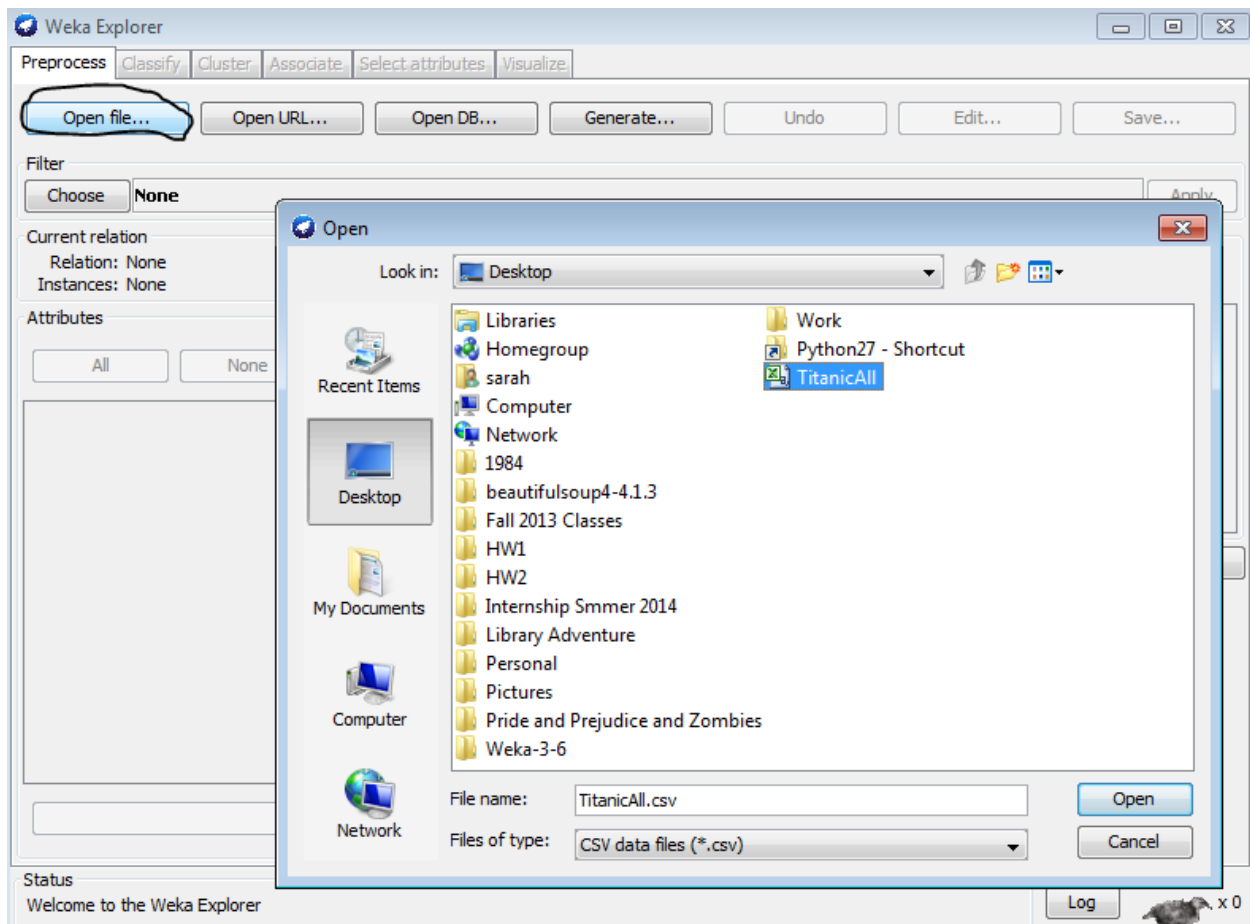
- a. You can use remote lab to access Weka or you can [download Weka](#) (free) onto your own device. On iSchool computers, you can find Weka by going to Computer-> OSDisk_-> program files-> weka 3-6: then select the "executable jar file."



- b. Open Weka explorer.



- c. Click "Open File." Select "CSV" from "files of type." Select TitanicAll.csv.



Now you have successfully read your merged training and test data file into Weka. Save this file in the .arff format, and then separate the train and test examples into two files. The two files should have the same heading and data description. Therefore, you need to make a copy of the merged file and name it

"Titanic-train.arff", and then delete all test example lines in it. Now you have your training arff file ready. Then make another copy of the merged file and name it "Titanic-test.arff", and then delete all training example lines in it. Now you have your test arff file ready.

```
@relation titanic-all-weka.filters.unsupervised.attribute

@attribute Survived {0,1}
@attribute Pclass {1,2,3}
@attribute Sex {male,female}
@attribute Age numeric
@attribute SibSp numeric
@attribute Parch numeric
@attribute Ticket {'A/5 21171','PC 17599','STON/O2. 3101282'}
@attribute Fare numeric
@attribute Cabin {C85,C123,E46,G6,C103,D56,A6,'C23'}
@attribute Embarked {S,C,Q}

@data
0,3,male,22,1,0,'A/5 21171',7.25,?,S
1,1,female,38,1,0,'PC 17599',71.2833,C85,C
1,3,female,26,0,0,'STON/O2. 3101282',7.925,?,S
1,1,female,35,1,0,113803.0,53.1,C123,S
0,3,male,35,0,0,373450.0,8.05,?,S
0,3,male,?,0,0,330877.0,8.4583,?,Q
```

```
@relation titanic-all-weka.filters.unsupervised.attribute

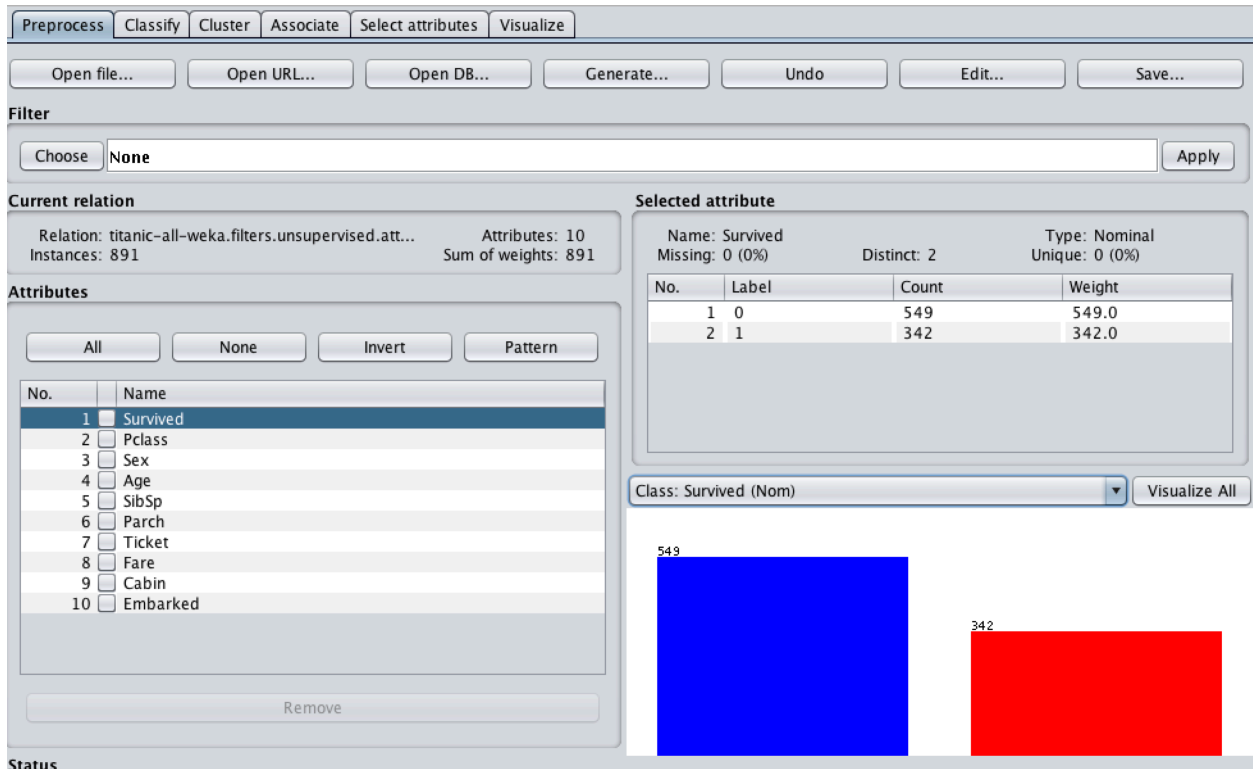
@attribute Survived {0,1}
@attribute Pclass {1,2,3}
@attribute Sex {male,female}
@attribute Age numeric
@attribute SibSp numeric
@attribute Parch numeric
@attribute Ticket {'A/5 21171','PC 17599','STON/O2. 3101282'}
@attribute Fare numeric
@attribute Cabin {C85,C123,E46,G6,C103,D56,A6,'C23'}
@attribute Embarked {S,C,Q}

@data
?,3,male,34.5,0,0,330911.0,7.8292,?,Q
?,3,female,47,1,0,363272.0,7,?,S
?,2,male,62,0,0,240276.0,9.6875,?,Q
?,3,male,27,0,0,315154.0,8.6625,?,S
?,3,female,22,1,1,3101298.0,12.2875,?,S
?,3,male,14,0,0,7538.0,9.225,?,S
?,3,female,30,0,0,330972.0,7.6292,?,Q
?,2,male,26,1,1,248738.0,29,?,S
?,3,female,18,0,0,2657.0,7.2292,?,C
?,3,male,21,2,0,'A/4 48871',24.15,?,S
?,3,male,?,0,0,240276.0,9.6875,?,Q
```

Now the Titanic training data is ready to be loaded into Weka to build decision tree. First, use the "open file" in "Preprocess" tab to load in the training data.

The upper right part of the interface shows some basic statistics of the feature highlighted on the left side. In the following screenshot, it is "Survived".

On the lower right part of the interface, choose "Class Survived (Nom)" in the pull down menu of all features, and you will see the category distribution, i.e. how many survived (Label "1" color red) and how many died (Label "0" color blue)



Now go to "Classify" tab, under "Classifiers", click the "Choose" button, and find J48 algorithm under the "trees" folder.

Maybe just add a screen flow