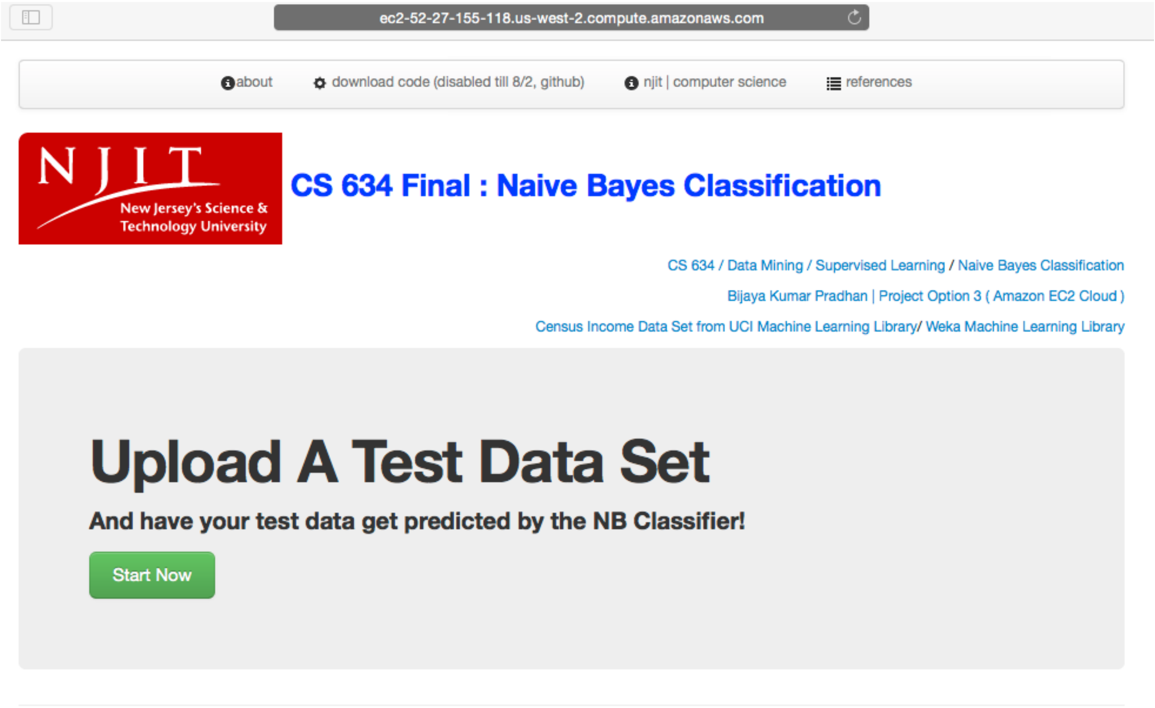
Naive Bayes Classification

CS 634 – Data Mining – Final Term Project - CS 634851

Bijaya Kumar Pradhan
bp249@njit.edu

Basic Navigation Screenshots


ec2-52-27-155-118.us-west-2.compute.amazonaws.com
about
download code (disabled till 8/2, github)
njit | computer science
references
NJIT
New Jersey's Science & Technology University
Start Now

CS 634 Final : Naive Bayes Classification

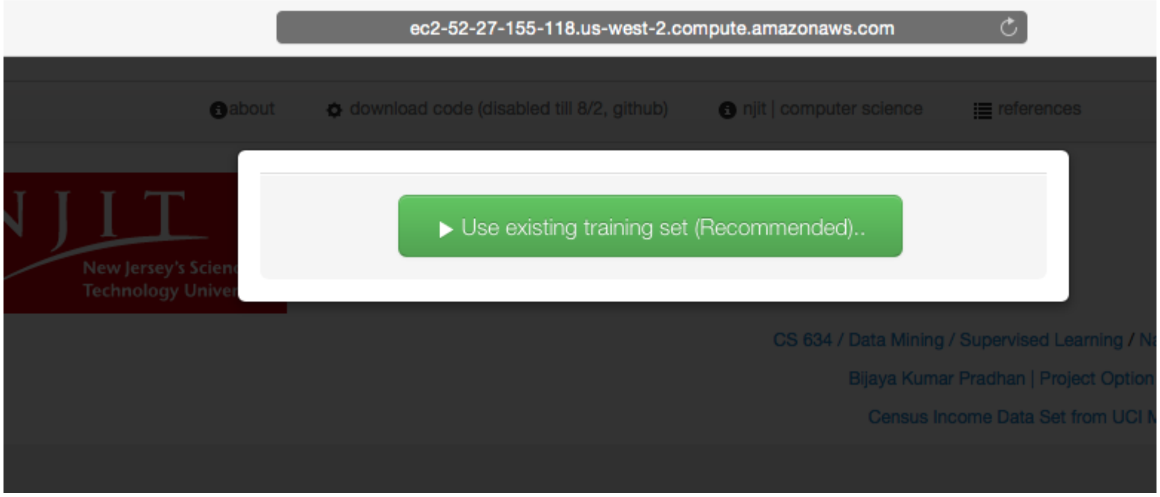
CS 634 / Data Mining / Supervised Learning / Naive Bayes Classification

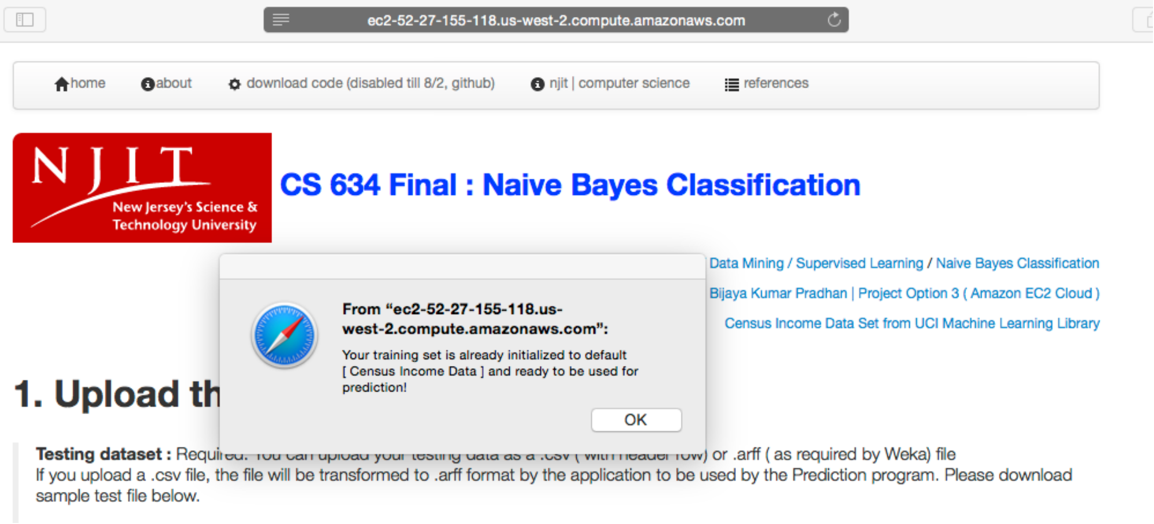
Bijaya Kumar Pradhan | Project Option 3 (Amazon EC2 Cloud)

Census Income Data Set from UCI Machine Learning Library/ Weka Machine Learning Library

Upload A Test Data Set

And have your test data get predicted by the NB Classifier!


ec2-52-27-155-118.us-west-2.compute.amazonaws.com
about
download code (disabled till 8/2, github)
njit | computer science
references
NJIT
New Jersey's Scienc
Technology Univer
▶ Use existing training set (Recommended)..
CS 634 / Data Mining / Supervised Learning / N
Bijaya Kumar Pradhan | Project Option
Census Income Data Set from UCI M



ec2-52-27-155-118.us-west-2.compute.amazonaws.com

home
about
download code (disabled till 8/2, github)
njit | computer science
references
NJIT
New Jersey's Science & Technology University

CS 634 Final : Naive Bayes Classification

Data Mining / Supervised Learning / Naive Bayes Classification

Bijaya Kumar Pradhan | Project Option 3 (Amazon EC2 Cloud)

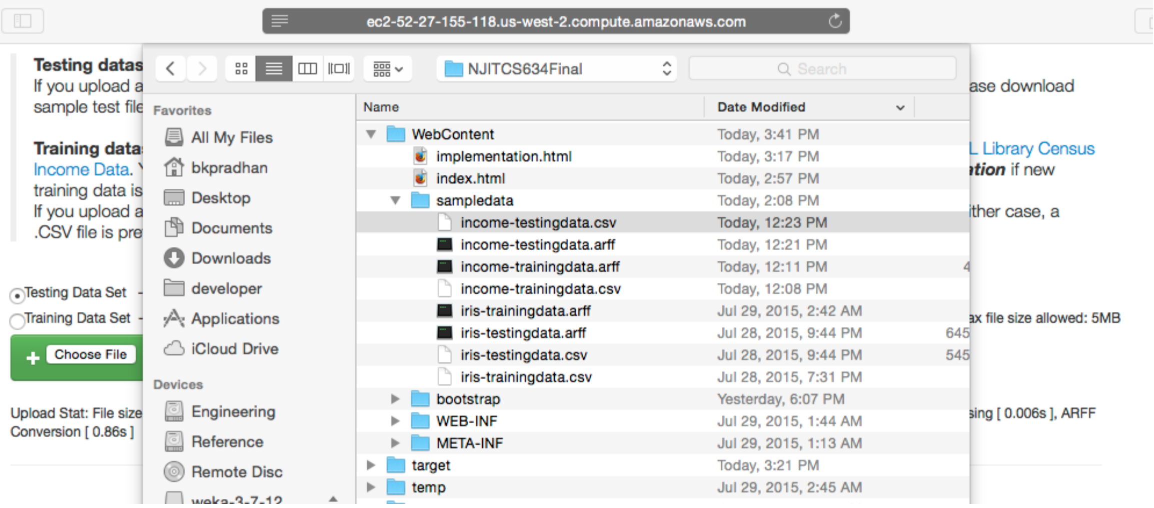
Census Income Data Set from UCI Machine Learning Library

From "ec2-52-27-155-118.us-west-2.compute.amazonaws.com":

Your training set is already initialized to default [Census Income Data] and ready to be used for prediction!

1. Upload th

Testing dataset : Required. You can upload your testing data as a .csv (with header row) or .arff (as required by Weka) file
If you upload a .csv file, the file will be transformed to .arff format by the application to be used by the Prediction program. Please download sample test file below.



ec2-52-27-155-118.us-west-2.compute.amazonaws.com

Testing datas
If you upload a
sample test file

Training data
Income Data. Y
training data is
If you upload a
.CSV file is pre

ase download

L Library Census
ation if new

ither case, a

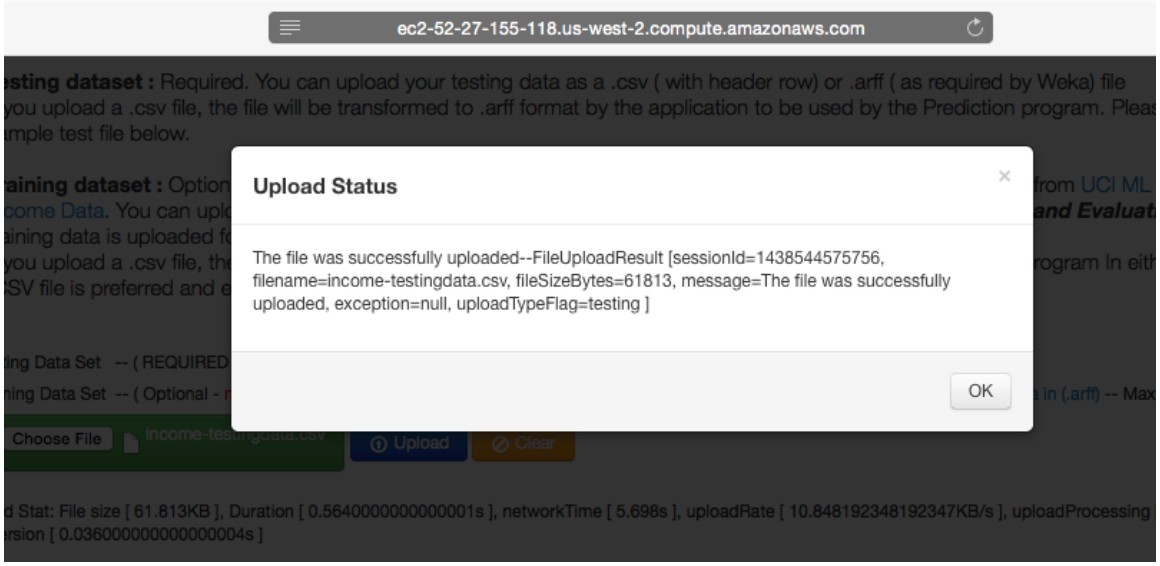
● Testing Data Set -
○ Training Data Set -

ax file size allowed: 5MB

Upload Stat: File size
Conversion [0.86s]

sing [0.006s], ARFF

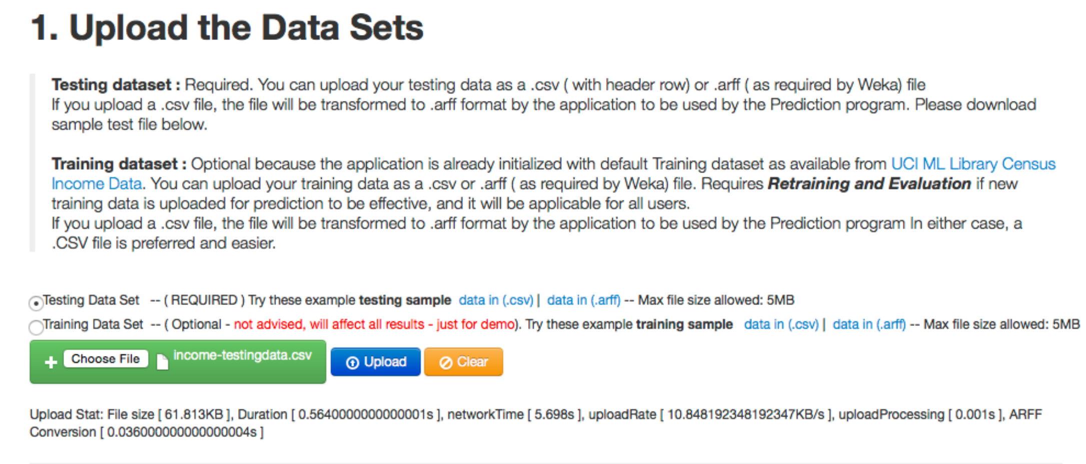
Name	Date Modified
▼ WebContent	Today, 3:41 PM
implementation.html	Today, 3:17 PM
index.html	Today, 2:57 PM
▼ sampledata	Today, 2:08 PM
income-testingdata.csv	Today, 12:23 PM
income-testingdata.arff	Today, 12:21 PM
income-trainingdata.arff	Today, 12:11 PM
income-trainingdata.csv	Today, 12:08 PM
iris-trainingdata.arff	Jul 29, 2015, 2:42 AM
iris-testingdata.arff	Jul 28, 2015, 9:44 PM
iris-testingdata.csv	Jul 28, 2015, 9:44 PM
iris-trainingdata.csv	Jul 28, 2015, 7:31 PM
▶ bootstrap	Yesterday, 6:07 PM
▶ WEB-INF	Jul 29, 2015, 1:44 AM
▶ META-INF	Jul 29, 2015, 1:13 AM
▶ target	Today, 3:21 PM
▶ temp	Jul 29, 2015, 2:45 AM



Upload Status

The file was successfully uploaded--FileUploadResult [sessionId=1438544575756, filename=income-testingdata.csv, fileSizeBytes=61813, message=The file was successfully uploaded, exception=null, uploadTypeFlag=testing]

d Stat: File size [61.813KB], Duration [0.5640000000000001s], networkTime [5.698s], uploadRate [10.848192348192347KB/s], uploadProcessing
rsion [0.036000000000000004s]



1. Upload the Data Sets

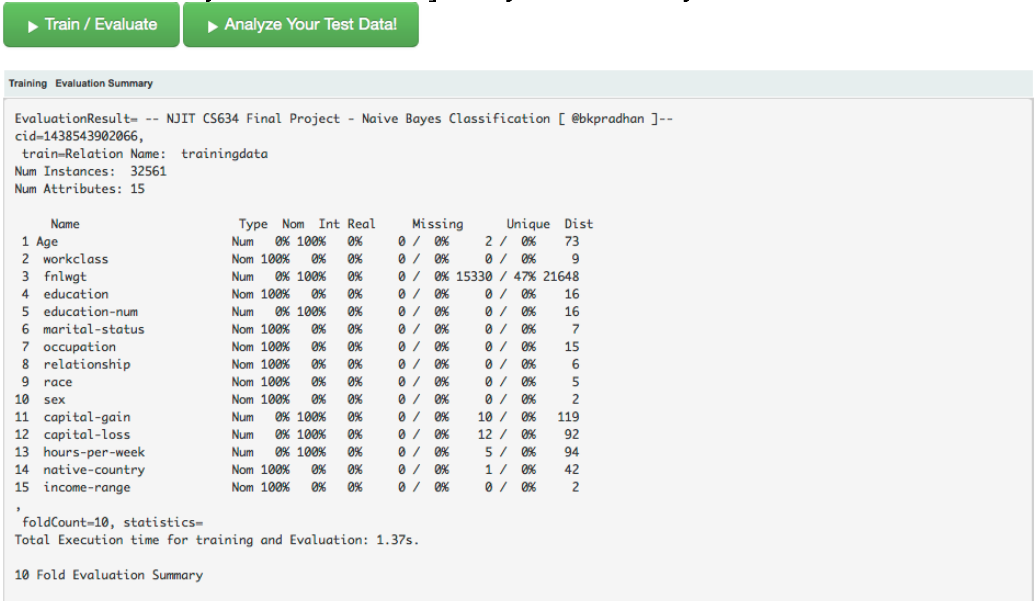
Testing dataset : Required. You can upload your testing data as a .csv (with header row) or .arff (as required by Weka) file
If you upload a .csv file, the file will be transformed to .arff format by the application to be used by the Prediction program. Please download sample test file below.

Training dataset : Optional because the application is already initialized with default Training dataset as available from UCI ML Library Census Income Data. You can upload your training data as a .csv or .arff (as required by Weka) file. Requires ***Retraining and Evaluation*** if new training data is uploaded for prediction to be effective, and it will be applicable for all users.
If you upload a .csv file, the file will be transformed to .arff format by the application to be used by the Prediction program In either case, a .CSV file is preferred and easier.

● Testing Data Set -- (REQUIRED) Try these example **testing sample** data in (.csv) | data in (.arff) -- Max file size allowed: 5MB
○ Training Data Set -- (Optional - not advised, will affect all results - just for demo). Try these example **training sample** data in (.csv) | data in (.arff) -- Max file size allowed: 5MB

Upload Stat: File size [61.813KB], Duration [0.5640000000000001s], networkTime [5.698s], uploadRate [10.848192348192347KB/s], uploadProcessing [0.001s], ARFF Conversion [0.036000000000000004s]

[Optional – already trained model], but you can always train

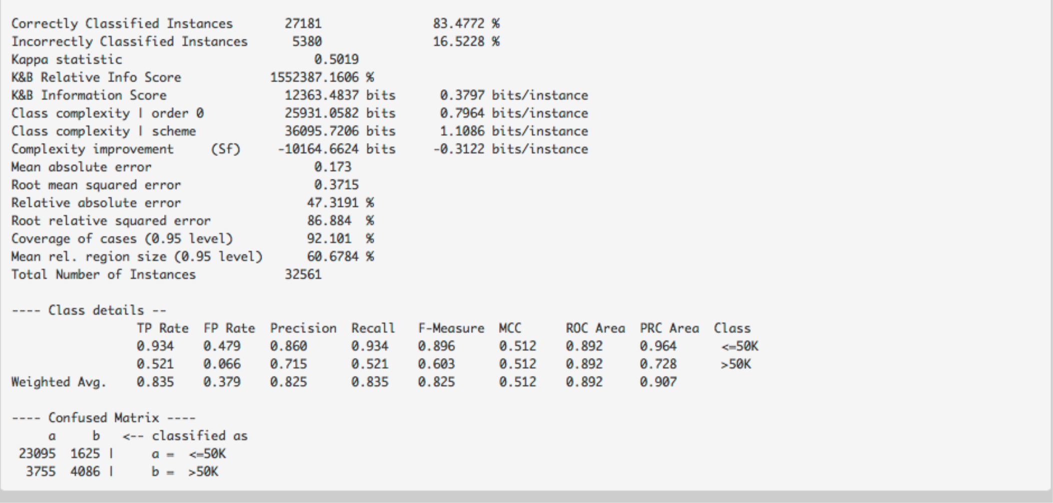


```
EvaluationResult= -- NJIT CS634 Final Project - Naive Bayes Classification [ @bkpradhan ]--
cid=1438543902066,
 train=Relation Name:  trainingdata
Num Instances:  32561
Num Attributes: 15

     Name                      Type  Nom  Int Real     Missing      Unique  Dist
 1 Age                         Num   0% 100%   0%     0 /  0%     2 /  0%    73
 2  workclass                  Nom 100%   0%   0%     0 /  0%     0 /  0%     9
 3  fnlwgt                     Num   0% 100%   0%     0 /  0% 15330 / 47% 21648
 4  education                  Nom 100%   0%   0%     0 /  0%     0 /  0%    16
 5  education-num              Num   0% 100%   0%     0 /  0%     0 /  0%    16
 6  marital-status             Nom 100%   0%   0%     0 /  0%     0 /  0%     7
 7  occupation                 Nom 100%   0%   0%     0 /  0%     0 /  0%    15
 8  relationship               Nom 100%   0%   0%     0 /  0%     0 /  0%     6
 9  race                       Nom 100%   0%   0%     0 /  0%     0 /  0%     5
10  sex                        Nom 100%   0%   0%     0 /  0%     0 /  0%     2
11  capital-gain               Num   0% 100%   0%     0 /  0%    10 /  0%   119
12  capital-loss               Num   0% 100%   0%     0 /  0%    12 /  0%    92
13  hours-per-week             Num   0% 100%   0%     0 /  0%     5 /  0%    94
14  native-country             Nom 100%   0%   0%     0 /  0%     1 /  0%    42
15  income-range               Nom 100%   0%   0%     0 /  0%     0 /  0%     2

,
 foldCount=10, statistics=
Total Execution time for training and Evaluation: 1.37s.

10 Fold Evaluation Summary
```

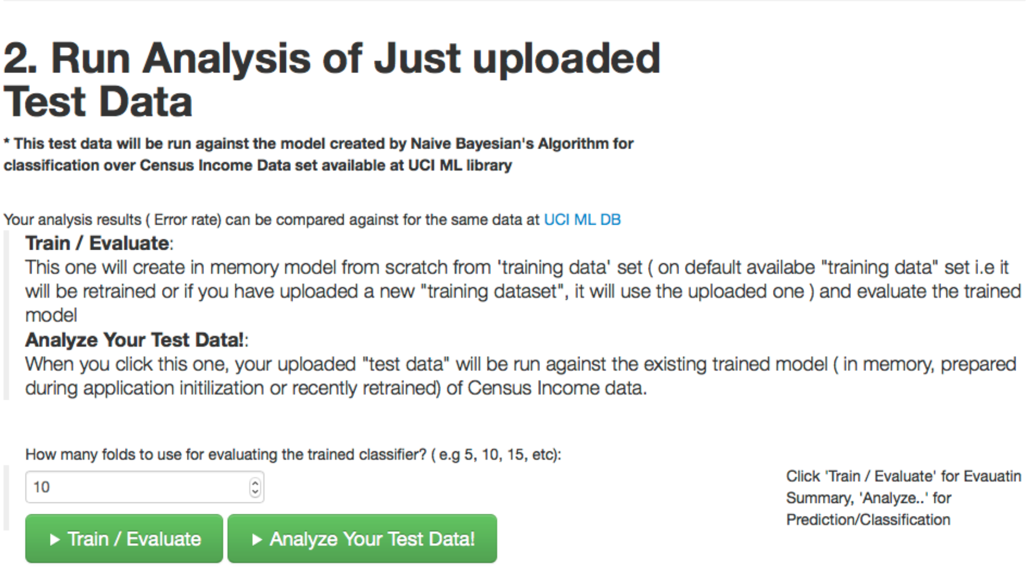


```
Correctly Classified Instances       27181               83.4772 %
Incorrectly Classified Instances      5380               16.5228 %
Kappa statistic                          0.5019
K&B Relative Info Score            1552387.1606 %
K&B Information Score                12363.4837 bits      0.3797 bits/instance
Class complexity | order 0           25931.0582 bits      0.7964 bits/instance
Class complexity | scheme            36095.7206 bits      1.1086 bits/instance
Complexity improvement     (Sf)     -10164.6624 bits     -0.3122 bits/instance
Mean absolute error                      0.173
Root mean squared error                  0.3715
Relative absolute error                 47.3191 %
Root relative squared error             86.884  %
Coverage of cases (0.95 level)          92.101  %
Mean rel. region size (0.95 level)      60.6784 %
Total Number of Instances            32561

---- Class details --
                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.934    0.479    0.860      0.934    0.896      0.512    0.892     0.964      <=50K
                 0.521    0.066    0.715      0.521    0.603      0.512    0.892     0.728      >50K
Weighted Avg.    0.835    0.379    0.825      0.835    0.825      0.512    0.892     0.907

---- Confused Matrix ----
     a     b   <-- classified as
 23095  1625 |     a =  <=50K
  3755  4086 |     b =  >50K
```

[You can verify how is your test data gets predicted by analyzing]



2. Run Analysis of Just uploaded Test Data

*** This test data will be run against the model created by Naive Bayesian's Algorithm for classification over Census Income Data set available at UCI ML library**

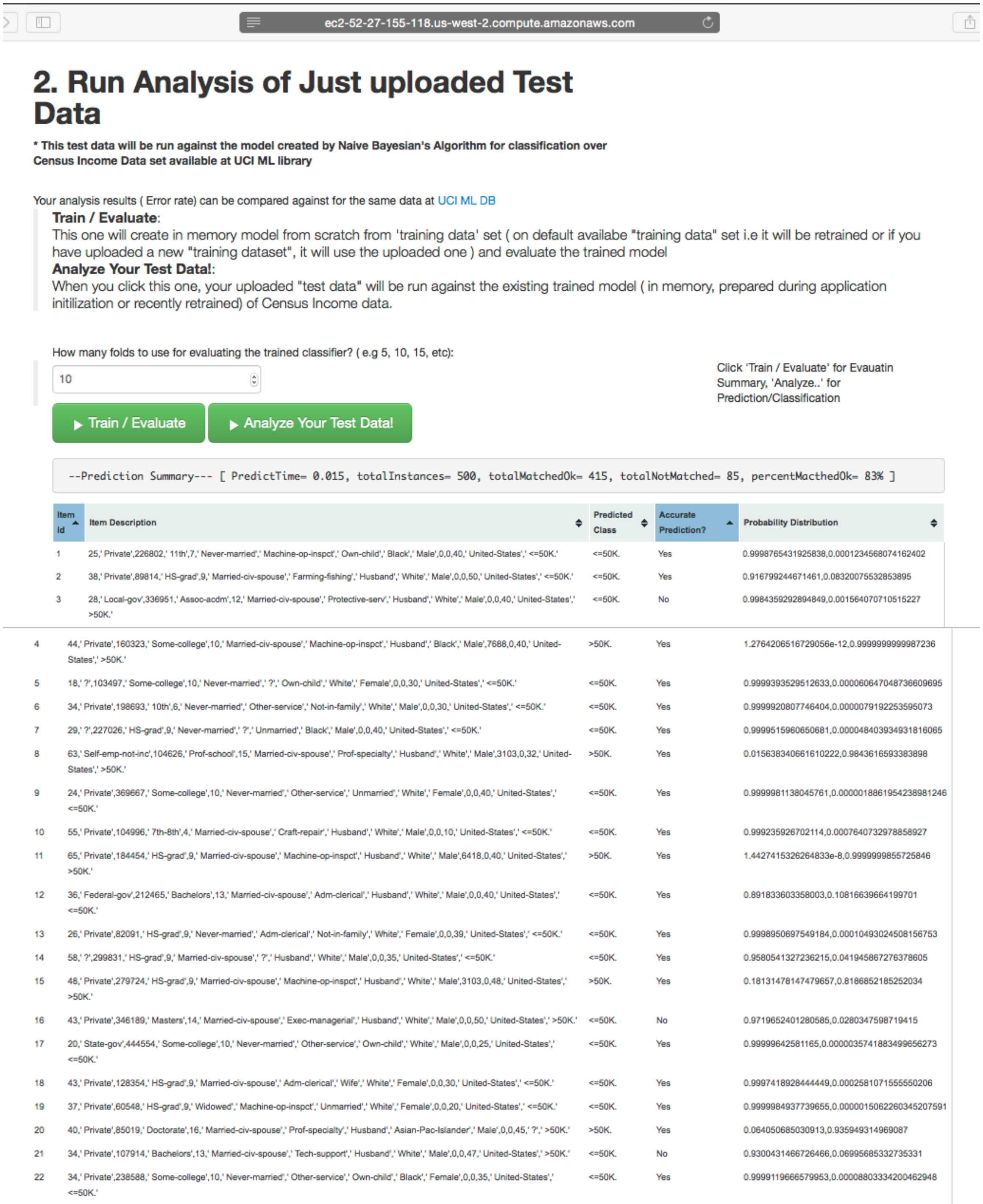
Your analysis results (Error rate) can be compared against for the same data at UCI ML DB

> **Train / Evaluate**:
> This one will create in memory model from scratch from 'training data' set (on default availabe "training data" set i.e it will be retrained or if you have uploaded a new "training dataset", it will use the uploaded one) and evaluate the trained model
> **Analyze Your Test Data!**:
> When you click this one, your uploaded "test data" will be run against the existing trained model (in memory, prepared during application initilization or recently retrained) of Census Income data.

How many folds to use for evaluating the trained classifier? (e.g 5, 10, 15, etc):

10

Click 'Train / Evaluate' for Evauatin Summary, 'Analyze..' for Prediction/Classification



2. Run Analysis of Just uploaded Test Data

*** This test data will be run against the model created by Naive Bayesian's Algorithm for classification over Census Income Data set available at UCI ML library**

Your analysis results (Error rate) can be compared against for the same data at UCI ML DB

> **Train / Evaluate**:
> This one will create in memory model from scratch from 'training data' set (on default availabe "training data" set i.e it will be retrained or if you have uploaded a new "training dataset", it will use the uploaded one) and evaluate the trained model
> **Analyze Your Test Data!**:
> When you click this one, your uploaded "test data" will be run against the existing trained model (in memory, prepared during application initilization or recently retrained) of Census Income data.

How many folds to use for evaluating the trained classifier? (e.g 5, 10, 15, etc):

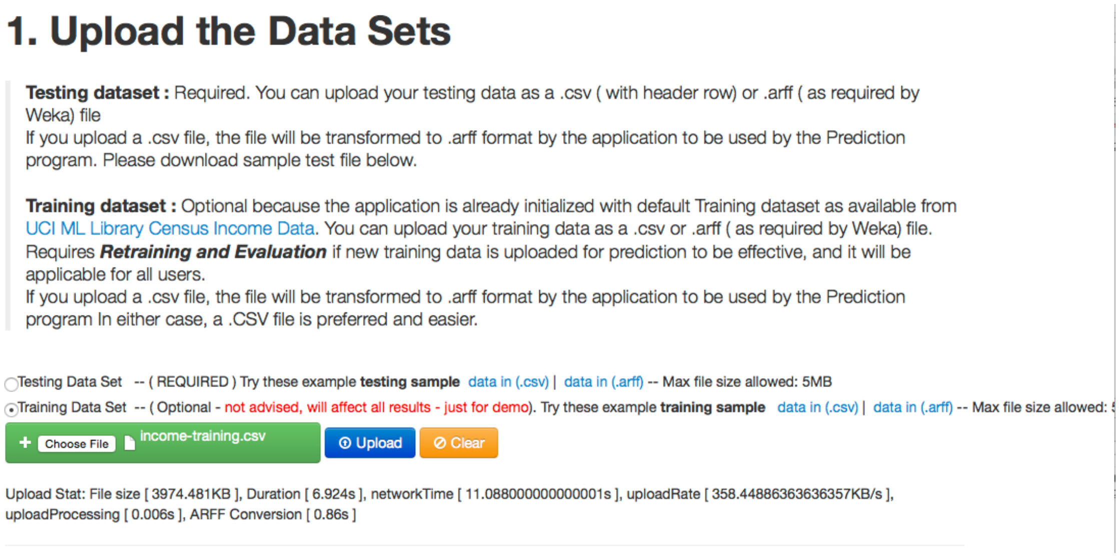
10

Click 'Train / Evaluate' for Evauatin Summary, 'Analyze..' for Prediction/Classification

```
--Prediction Summary--- [ PredictTime= 0.015, totalInstances= 500, totalMatchedOk= 415, totalNotMatched= 85, percentMacthedOk= 83% ]
```

Item Id	Item Description	Predicted Class	Accurate Prediction?	Probability Distribution
1	25,' Private',226802,' 11th',7,' Never-married',' Machine-op-inspct',' Own-child',' Black',' Male',0,0,40,' United-States',' <=50K.'	<=50K.	Yes	0.9998765431925838,0.0001234568074162402
2	38,' Private',89814,' HS-grad',9,' Married-civ-spouse',' Farming-fishing',' Husband',' White',' Male',0,0,50,' United-States',' <=50K.'	<=50K.	Yes	0.916799244671461,0.08320075532853895
3	28,' Local-gov',336951,' Assoc-acdm',12,' Married-civ-spouse',' Protective-serv',' Husband',' White',' Male',0,0,40,' United-States',' >50K.'	<=50K.	No	0.9984359292894849,0.001564070710515227
4	44,' Private',160323,' Some-college',10,' Married-civ-spouse',' Machine-op-inspct',' Husband',' Black',' Male',7688,0,40,' United-States',' >50K.'	>50K.	Yes	1.2764206516729056e-12,0.9999999999987236
5	18,' ?',103497,' Some-college',10,' Never-married',' ?',' Own-child',' White',' Female',0,0,30,' United-States',' <=50K.'	<=50K.	Yes	0.9999393529512633,0.000060647048736609695
6	34,' Private',198693,' 10th',6,' Never-married',' Other-service',' Not-in-family',' White',' Male',0,0,30,' United-States',' <=50K.'	<=50K.	Yes	0.9999920807746404,0.0000079192253595073
7	29,' ?',227026,' HS-grad',9,' Never-married',' ?',' Unmarried',' Black',' Male',0,0,40,' United-States',' <=50K.'	<=50K.	Yes	0.9999515960650681,0.000048403934931816065
8	63,' Self-emp-not-inc',104626,' Prof-school',15,' Married-civ-spouse',' Prof-specialty',' Husband',' White',' Male',3103,0,32,' United-States',' >50K.'	>50K.	Yes	0.015638340661610222,0.9843616593383898
9	24,' Private',369667,' Some-college',10,' Never-married',' Other-service',' Unmarried',' White',' Female',0,0,40,' United-States',' <=50K.'	<=50K.	Yes	0.9999981138045761,0.0000018861954238981246
10	55,' Private',104996,' 7th-8th',4,' Married-civ-spouse',' Craft-repair',' Husband',' White',' Male',0,0,10,' United-States',' <=50K.'	<=50K.	Yes	0.999235926702114,0.0007640732978858927
11	65,' Private',184454,' HS-grad',9,' Married-civ-spouse',' Machine-op-inspct',' Husband',' White',' Male',6418,0,40,' United-States',' >50K.'	>50K.	Yes	1.4427415326264833e-8,0.9999999855725846
12	36,' Federal-gov',212465,' Bachelors',13,' Married-civ-spouse',' Adm-clerical',' Husband',' White',' Male',0,0,40,' United-States',' <=50K.'	<=50K.	Yes	0.891833603358003,0.10816639664199701
13	26,' Private',82091,' HS-grad',9,' Never-married',' Adm-clerical',' Not-in-family',' White',' Female',0,0,39,' United-States',' <=50K.'	<=50K.	Yes	0.9998950697549184,0.00010493024508156753
14	58,' ?',299831,' HS-grad',9,' Married-civ-spouse',' ?',' Husband',' White',' Male',0,0,35,' United-States',' <=50K.'	<=50K.	Yes	0.9580541327236215,0.041945867276378605
15	48,' Private',279724,' HS-grad',9,' Married-civ-spouse',' Machine-op-inspct',' Husband',' White',' Male',3103,0,48,' United-States',' >50K.'	>50K.	Yes	0.18131478147479657,0.8186852185252034
16	43,' Private',346189,' Masters',14,' Married-civ-spouse',' Exec-managerial',' Husband',' White',' Male',0,0,50,' United-States',' >50K.'	<=50K.	No	0.9719652401280585,0.0280347598719415
17	20,' State-gov',444554,' Some-college',10,' Never-married',' Other-service',' Own-child',' White',' Male',0,0,25,' United-States',' <=50K.'	<=50K.	Yes	0.99999642581165,0.0000035741883499656273
18	43,' Private',128354,' HS-grad',9,' Married-civ-spouse',' Adm-clerical',' Wife',' White',' Female',0,0,30,' United-States',' <=50K.'	<=50K.	Yes	0.9997418928444449,0.0002581071555550206
19	37,' Private',60548,' HS-grad',9,' Widowed',' Machine-op-inspct',' Unmarried',' White',' Female',0,0,20,' United-States',' <=50K.'	<=50K.	Yes	0.9999984937739655,0.0000015062260345207591
20	40,' Private',85019,' Doctorate',16,' Married-civ-spouse',' Prof-specialty',' Husband',' Asian-Pac-Islander',' Male',0,0,45,' ?',' >50K.'	>50K.	Yes	0.064050685030913,0.935949314969087
21	34,' Private',107914,' Bachelors',13,' Married-civ-spouse',' Tech-support',' Husband',' White',' Male',0,0,47,' United-States',' >50K.'	<=50K.	No	0.9300431466726466,0.06995685332735331
22	34,' Private',238588,' Some-college',10,' Never-married',' Other-service',' Own-child',' Black',' Female',0,0,35,' United-States',' <=50K.'	<=50K.	Yes	0.9999119666579953,0.00008803334200462948

[ONLY FOR uploading a TRAINING DATA set - NOT RECOMMENDED]



1. Upload the Data Sets

> **Testing dataset :** Required. You can upload your testing data as a .csv (with header row) or .arff (as required by Weka) file
> If you upload a .csv file, the file will be transformed to .arff format by the application to be used by the Prediction program. Please download sample test file below.
>
> **Training dataset :** Optional because the application is already initialized with default Training dataset as available from UCI ML Library Census Income Data. You can upload your training data as a .csv or .arff (as required by Weka) file. Requires ***Retraining and Evaluation*** if new training data is uploaded for prediction to be effective, and it will be applicable for all users.
> If you upload a .csv file, the file will be transformed to .arff format by the application to be used by the Prediction program In either case, a .CSV file is preferred and easier.

○ Testing Data Set -- (REQUIRED) Try these example **testing sample** data in (.csv) | data in (.arff) -- Max file size allowed: 5MB

● Training Data Set -- (Optional - not advised, will affect all results - just for demo). Try these example **training sample** data in (.csv) | data in (.arff) -- Max file size allowed:

Upload Stat: File size [3974.481KB], Duration [6.924s], networkTime [11.088000000000001s], uploadRate [358.44886363636357KB/s], uploadProcessing [0.006s], ARFF Conversion [0.86s]