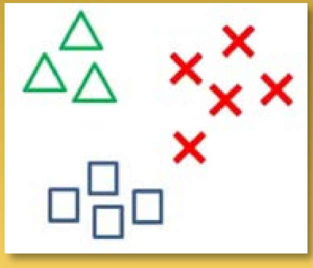
Algorithm Multi Class

ALGORITHM MULTI CLASS



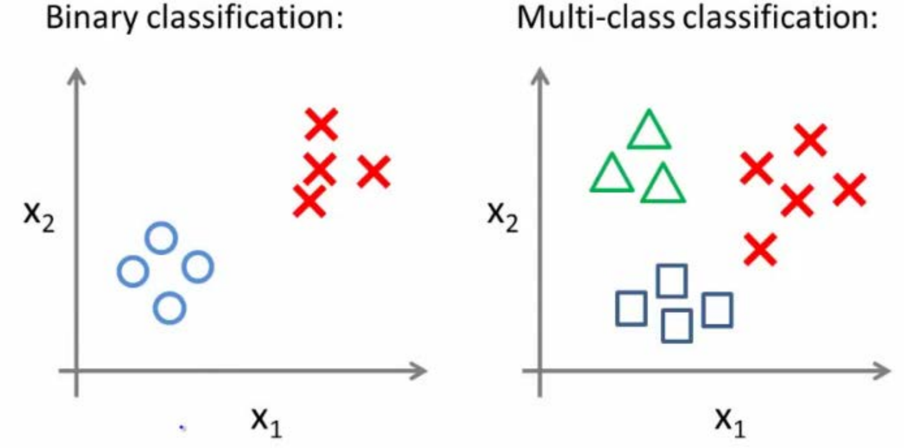
Algorithm Multi Class

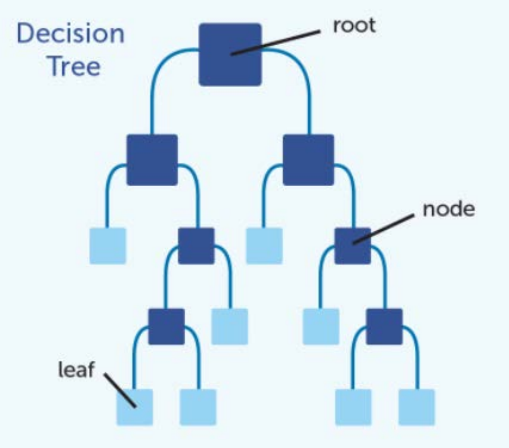
In this session

- Multi Class Algorithms in Azure ML
- Data importing and engineering
- Feature engineering
- Modeling and evaluation
- Reuter Data set
- Edit Metadata
- Confusion Matrix

Algorithm Multi Class

Multi Class Algorithms in Azure ML





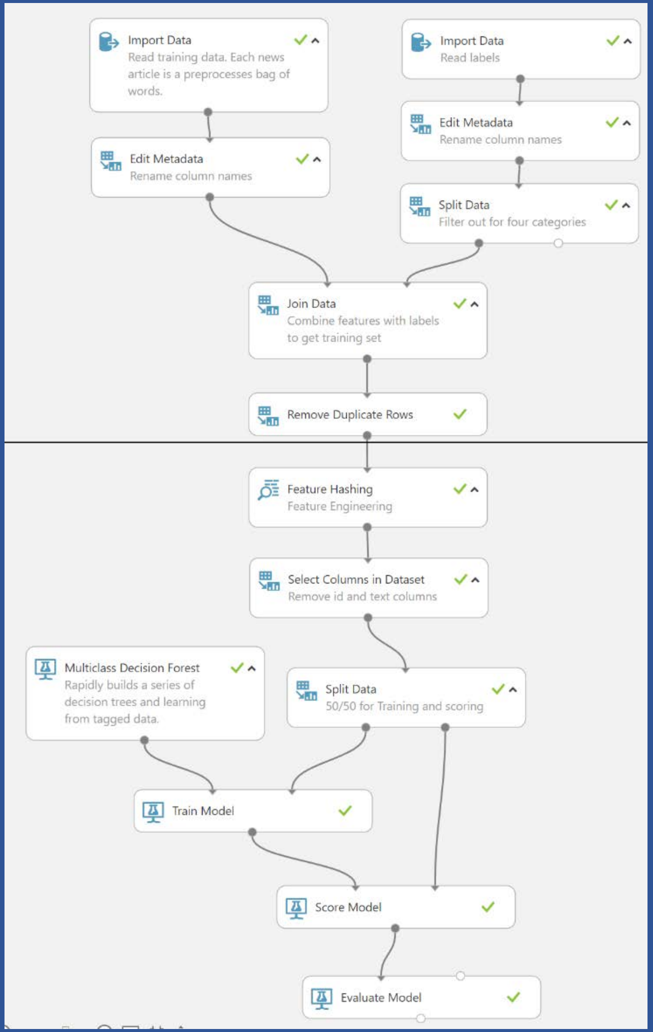
Multiclass Decision Forest

- Based on the decision forest algorithm
- Rapidly builds a series of decision trees
- learning from tagged data.
- Voting on the most popular output class
- Voting is a form of aggregation

Algorithm Multi Class

Over all Experiment

- multiclass classifiers
- Feature engineering using hashing
- Classify news into four categories



Import Data
Read training data. Each news article is a preprocesses bag of words.

Import Data
Read labels

Edit Metadata
Rename column names

Edit Metadata
Rename column names

Split Data
Filter out for four categories

Join Data
Combine features with labels to get training set

Remove Duplicate Rows

Feature Hashing
Feature Engineering

Select Columns in Dataset
Remove id and text columns

Multiclass Decision Forest
Rapidly builds a series of decision trees and learning from tagged data.

Split Data
50/50 for Training and scoring

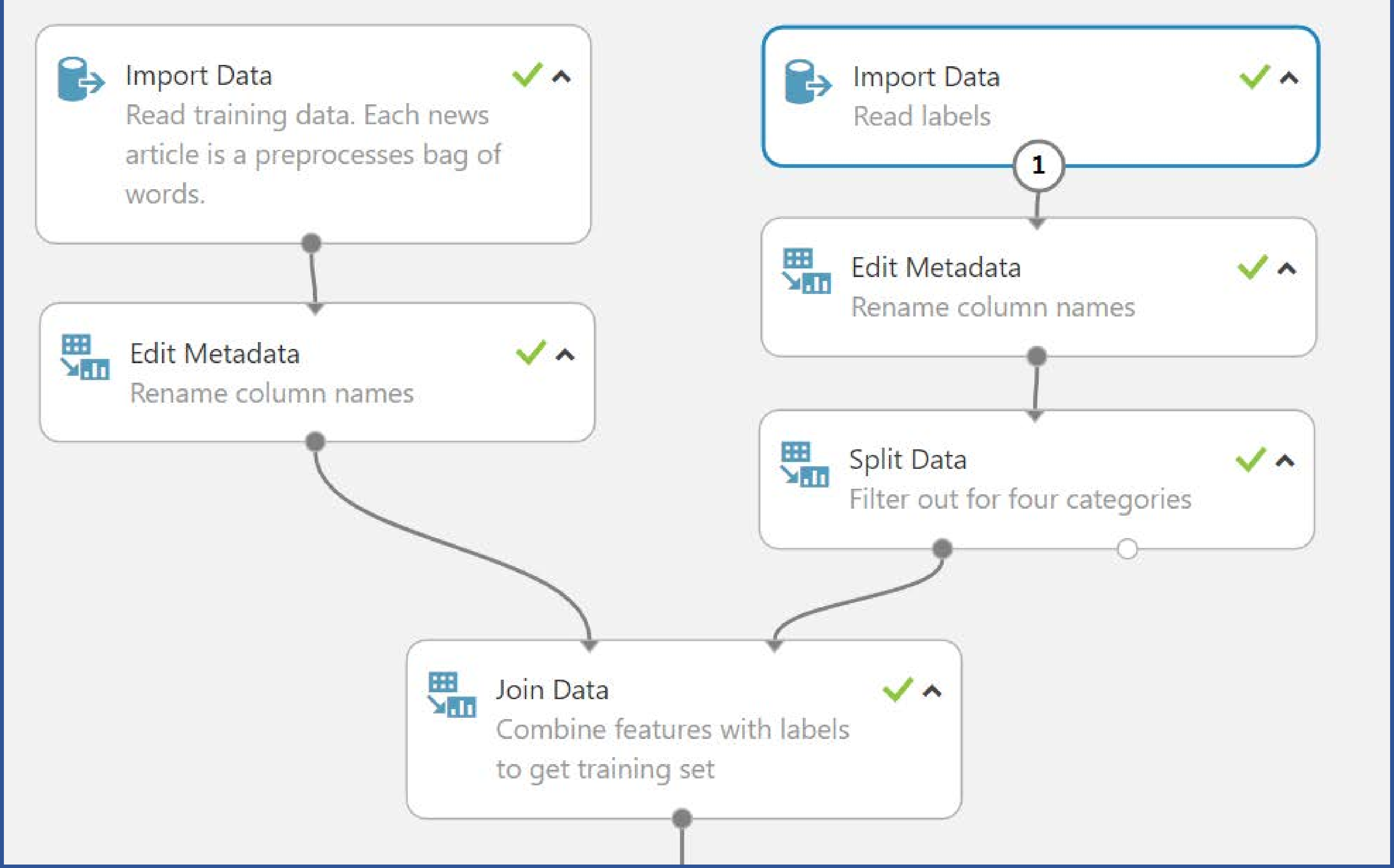
Train Model

Score Model

Evaluate Model

Algorithm Multi Class

Data importing and engineering



Import Data ✓
Read training data. Each news article is a preprocesses bag of words.

Edit Metadata ✓
Rename column names

Import Data ✓
Read labels

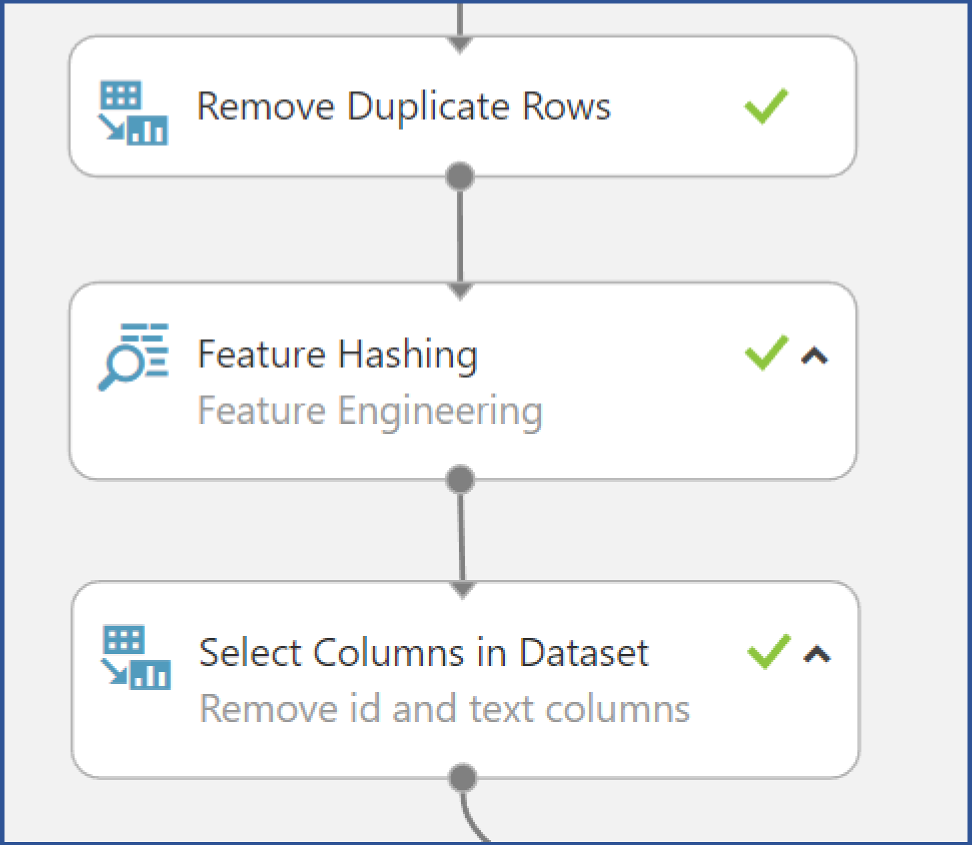
Edit Metadata ✓
Rename column names

Split Data ✓
Filter out for four categories

Join Data ✓
Combine features with labels to get training set

Algorithm Multi Class

Feature engineering



Remove Duplicate Rows ✓

Feature Hashing ✓

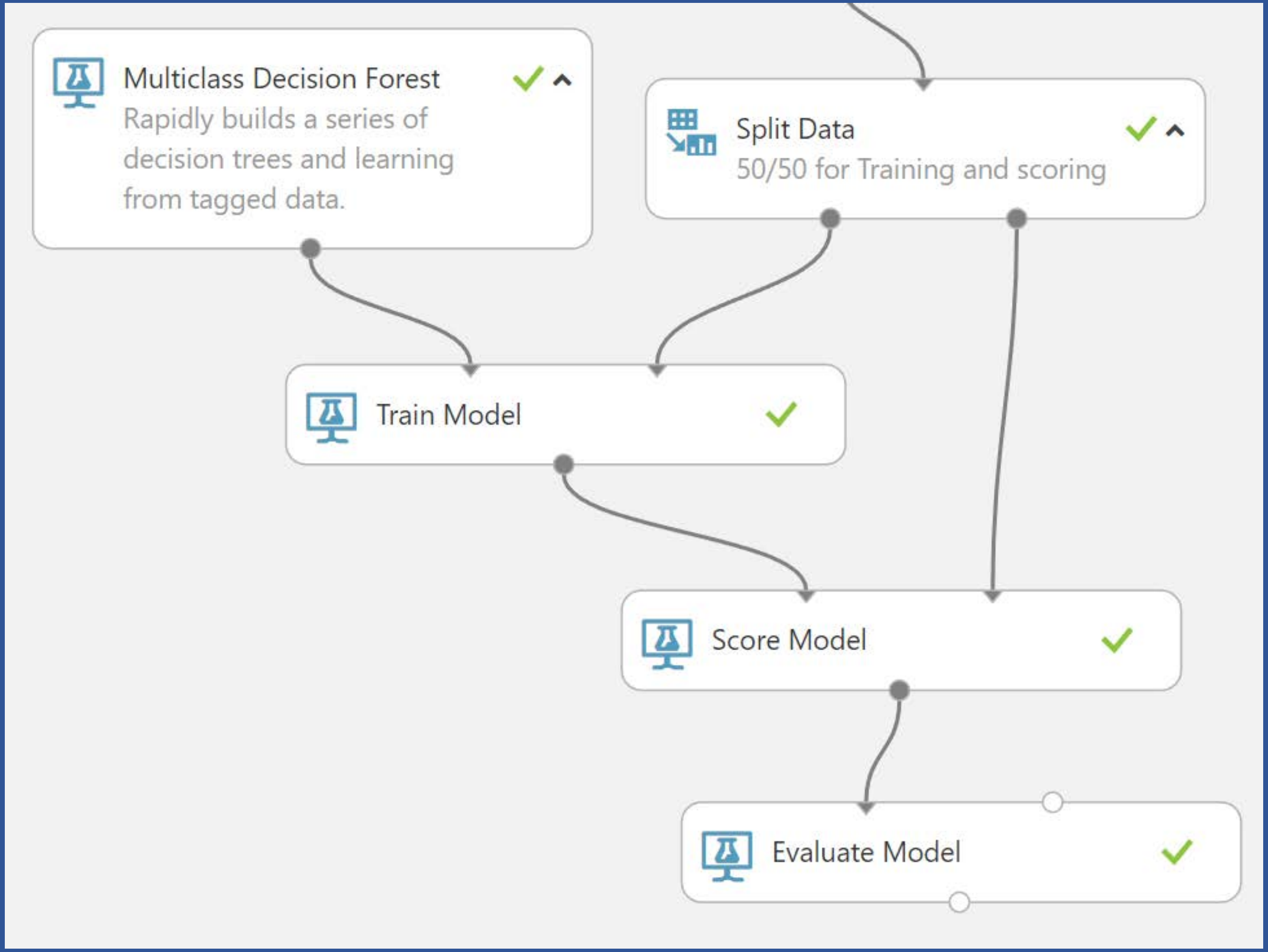
Feature Engineering

Select Columns in Dataset ✓

Remove id and text columns

Algorithm Multi Class

Modeling and evaluation



Multiclass Decision Forest ✓

Rapidly builds a series of decision trees and learning from tagged data.

Split Data ✓

50/50 for Training and scoring

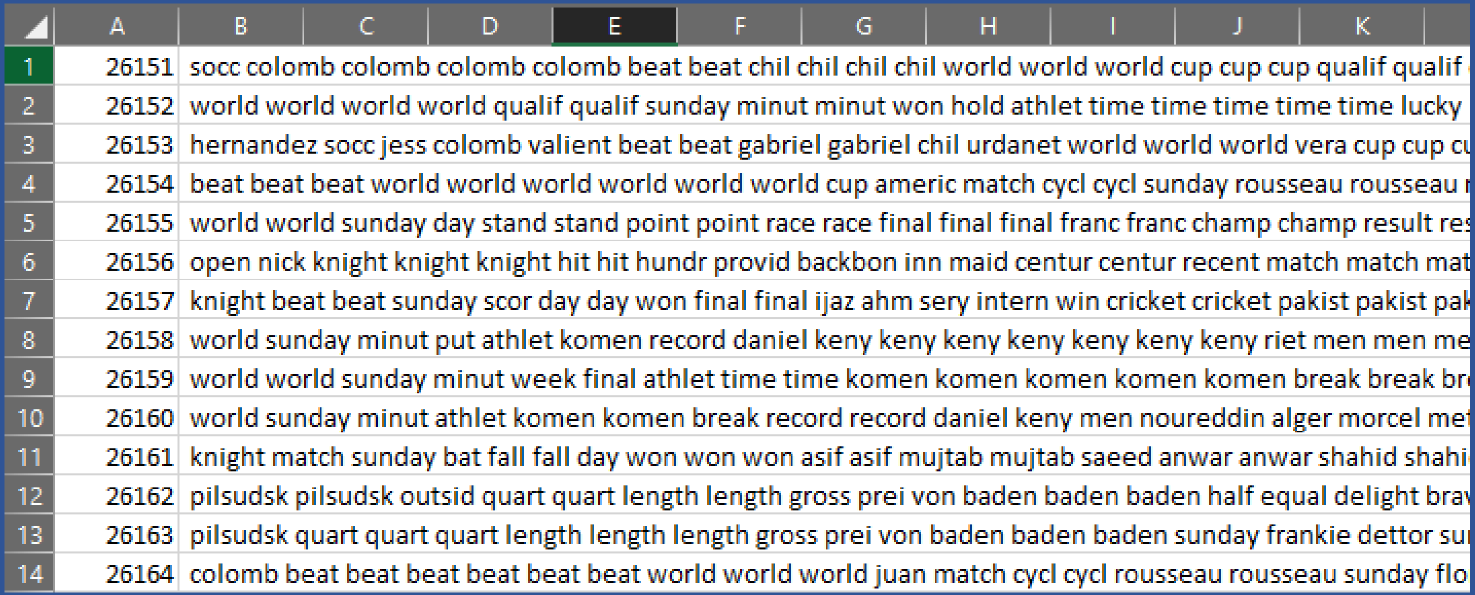
Train Model ✓

Score Model ✓

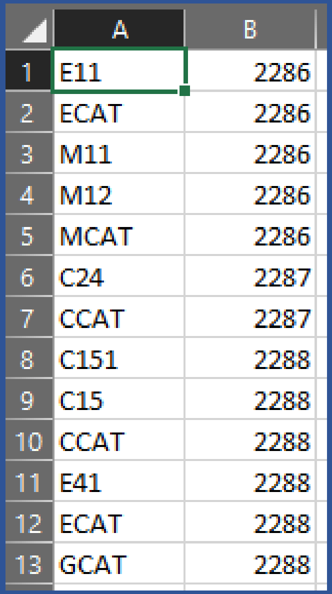
Evaluate Model ✓

Algorithm Multi Class

Reuter Data set



	A	B
1	26151	socc colomb colomb colomb colomb beat beat chil chil chil chil world world world cup cup cup qualif qualif
2	26152	world world world world qualif qualif sunday minut minut won hold athlet time time time time time lucky
3	26153	hernandez socc jess colomb valient beat beat gabriel gabriel chil urdanet world world world vera cup cup cu
4	26154	beat beat beat world world world world world world cup americ match cycl cycl sunday rousseau rousseau r
5	26155	world world sunday day stand stand point point race race final final final franc franc champ champ result res
6	26156	open nick knight knight knight hit hit hundr provid backbon inn maid centur centur recent match match mat
7	26157	knight beat beat sunday scor day day won final final ijaz ahm sery intern win cricket cricket pakist pakist pak
8	26158	world sunday minut put athlet komen record daniel keny keny keny keny keny keny keny riet men men me
9	26159	world world sunday minut week final athlet time time komen komen komen komen komen break break br
10	26160	world sunday minut athlet komen komen break record record daniel keny men noureddin alger morcel met
11	26161	knight match sunday bat fall fall day won won won asif asif mujtab mujtab saeed anwar anwar shahid shahi
12	26162	pilsudsk pilsudsk outsid quart quart length length gross prei von baden baden baden half equal delight brav
13	26163	pilsudsk quart quart quart length length length gross prei von baden baden baden sunday frankie dettor su
14	26164	colomb beat beat beat beat beat beat world world world juan match cycl cycl rousseau rousseau sunday flo



	A	B
1	E11	2286
2	ECAT	2286
3	M11	2286
4	M12	2286
5	MCAT	2286
6	C24	2287
7	CCAT	2287
8	C151	2288
9	C15	2288
10	CCAT	2288
11	E41	2288
12	ECAT	2288
13	GCAT	2288

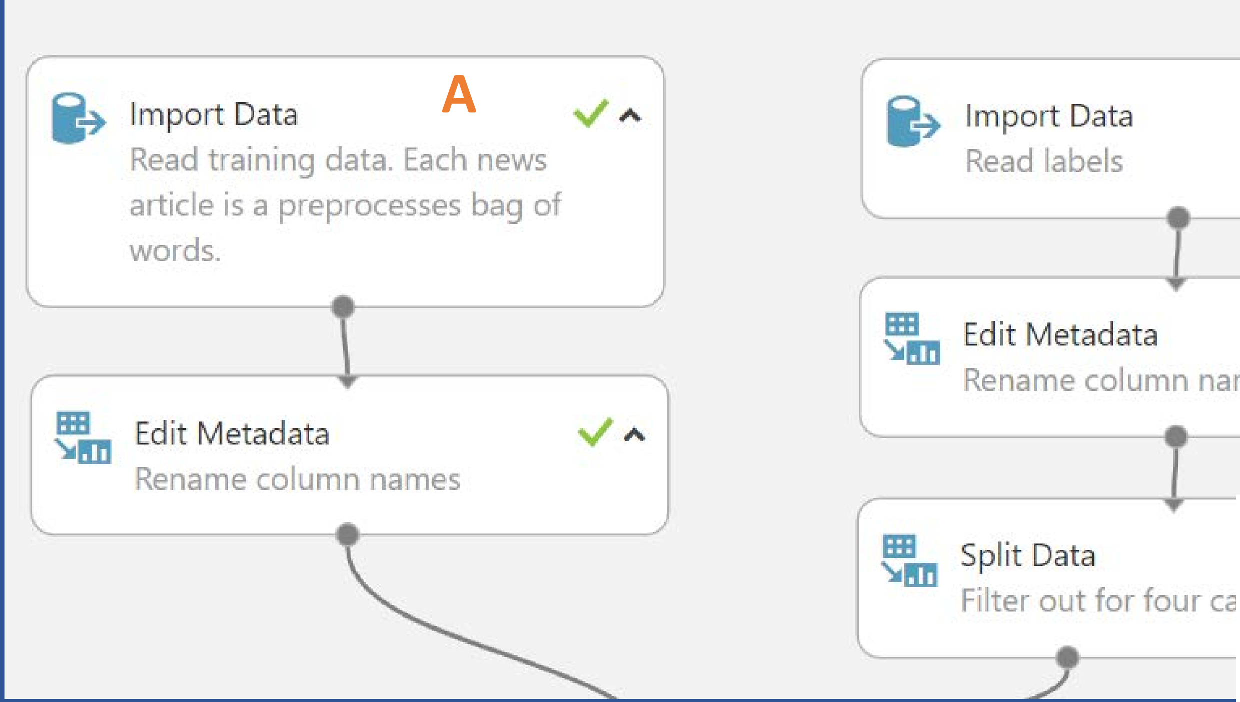
- 2004 Reuters news dataset
- 10,000 News examples
- 5K Training / 5K Scoring

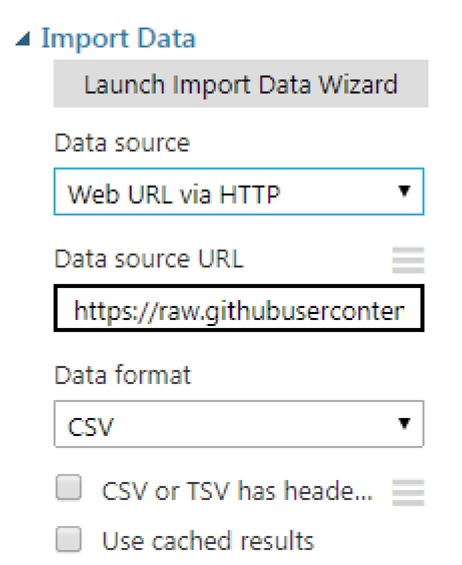
Data set has 103 categories that are organized into four hierarchies:

- Corporate-Industrial (CCAT)
- Government and Social (GCAT)
- Economics and Economic Indicators (ECAT)
- Securities and Commodities Trading and Market (MCAT)

Algorithm Multi Class

Import data set




◢ Import Data

Launch Import Data Wizard

Data source

Web URL via HTTP ▾

Data source URL

https://raw.githubusercontent

Data format

CSV ▾

□ CSV or TSV has heade...

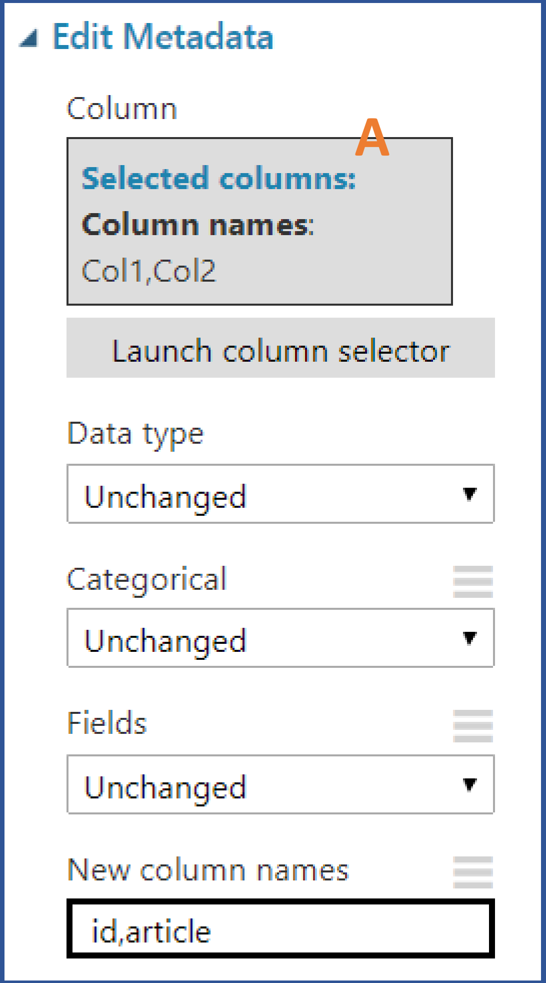
□ Use cached results

A = https://raw.githubusercontent.com/laploy/ML/master/mul.csv

B = https://raw.githubusercontent.com/laploy/ML/master/mul_token.csv

Algorithm Multi Class

Edit Metadata



Edit Metadata

Column

Selected columns:
Column names:
Col1,Col2

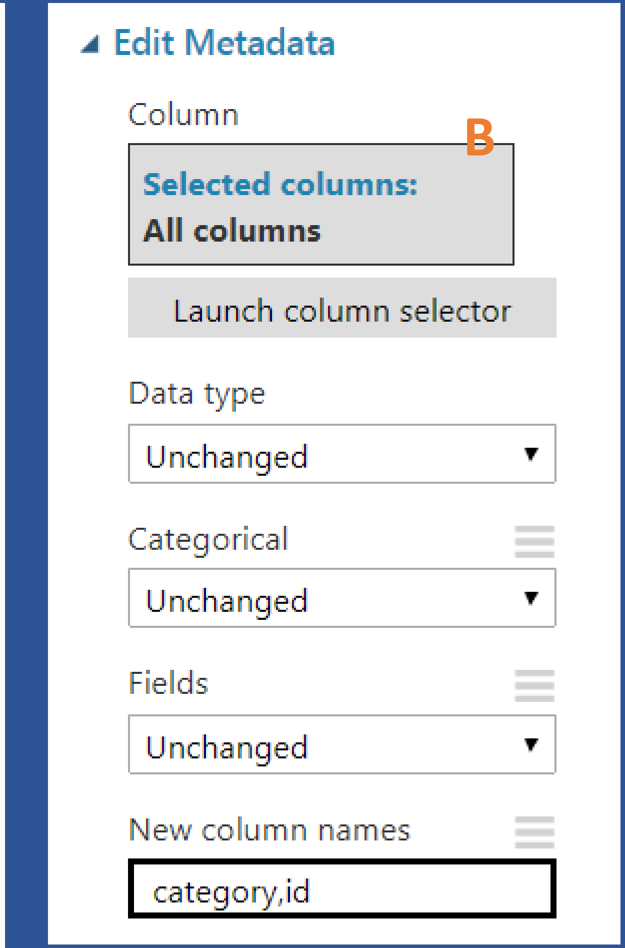
Launch column selector

Data type: Unchanged

Categorical: Unchanged

Fields: Unchanged

New column names: id,article



Edit Metadata

Column

Selected columns:
All columns

Launch column selector

Data type: Unchanged

Categorical: Unchanged

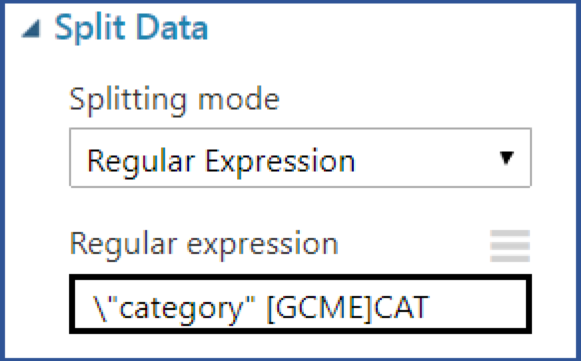
Fields: Unchanged

New column names: category,id

Algorithm Multi Class

Splitting Data

Used only the rows already tagged with hierarchy names (CCAT,ECAT,GCAT,MCAT)



◢ Split Data

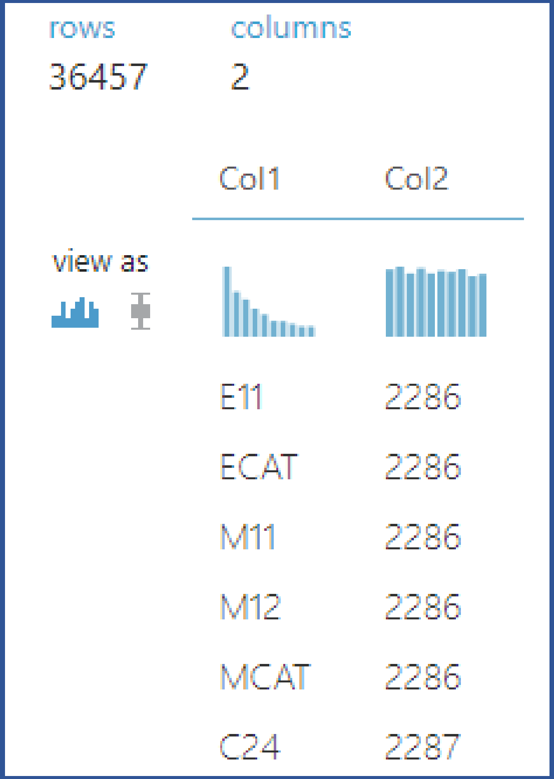
Splitting mode

Regular Expression ▾

Regular expression

\"category" [GCME]CAT

Before splitting



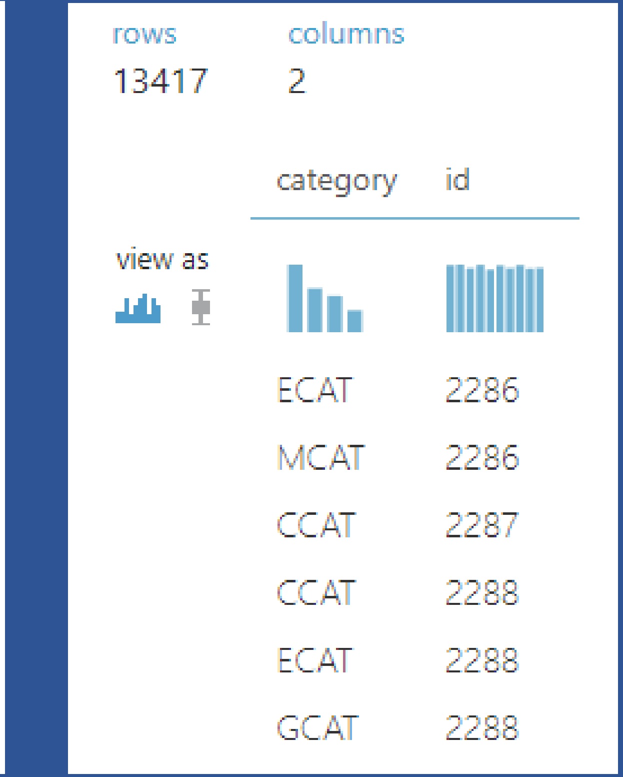
rows 36457

columns 2

view as

Col1	Col2
E11	2286
ECAT	2286
M11	2286
M12	2286
MCAT	2286
C24	2287

After splitting



rows 13417

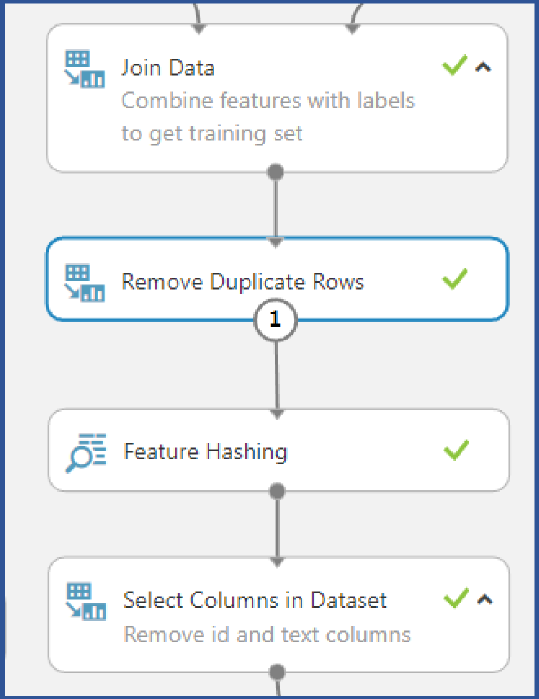
columns 2

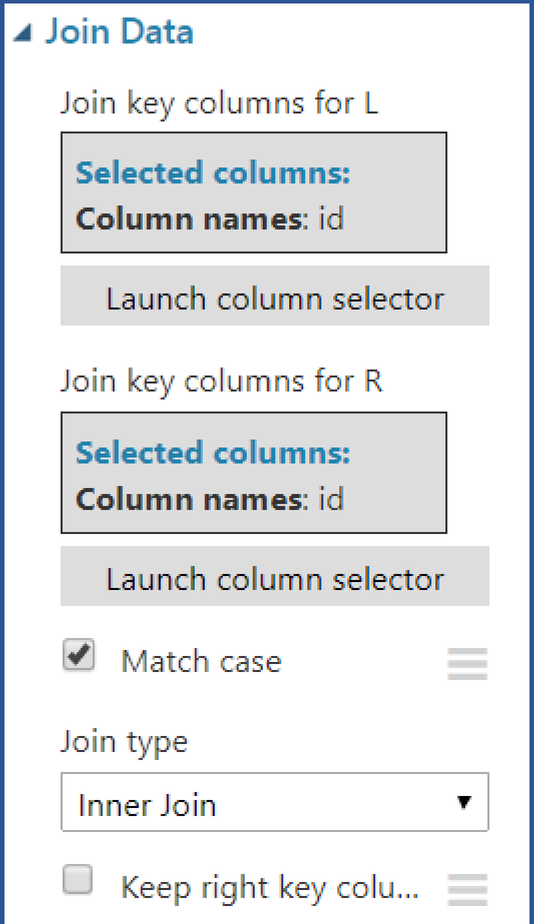
view as

category	id
ECAT	2286
MCAT	2286
CCAT	2287
CCAT	2288
ECAT	2288
GCAT	2288

Algorithm Multi Class

Feature & Clean


Join Data
Combine features with labels to get training set
Remove Duplicate Rows
1
Feature Hashing
Select Columns in Dataset
Remove id and text columns



Join Data

Join key columns for L

Selected columns:
Column names: id

Launch column selector

Join key columns for R

Selected columns:
Column names: id

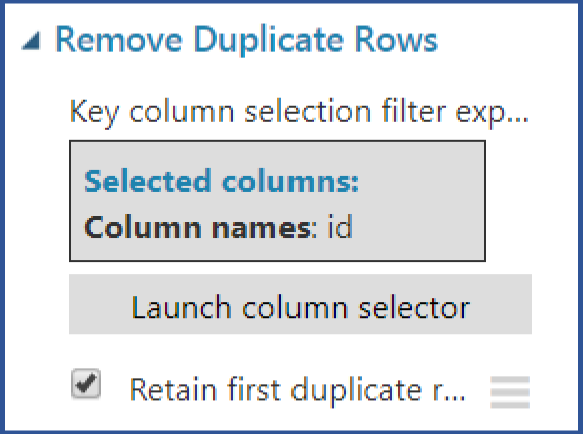
Launch column selector

☑ Match case

Join type

Inner Join ▾

☐ Keep right key colu...



Remove Duplicate Rows

Key column selection filter exp...

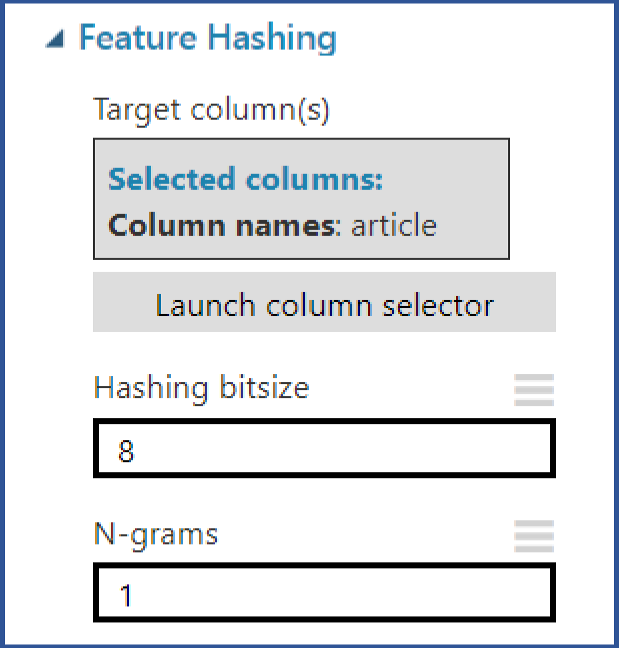
Selected columns:
Column names: id

Launch column selector

☑ Retain first duplicate r...

Algorithm Multi Class

Feature Engineering



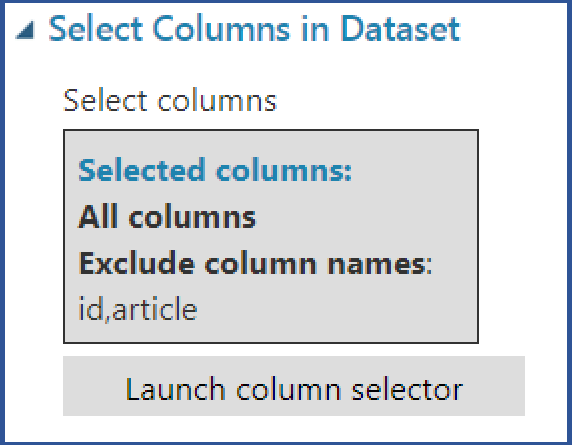
Feature Hashing

Target column(s)

Selected columns:
Column names: article

Hashing bitsize: 8

N-grams: 1



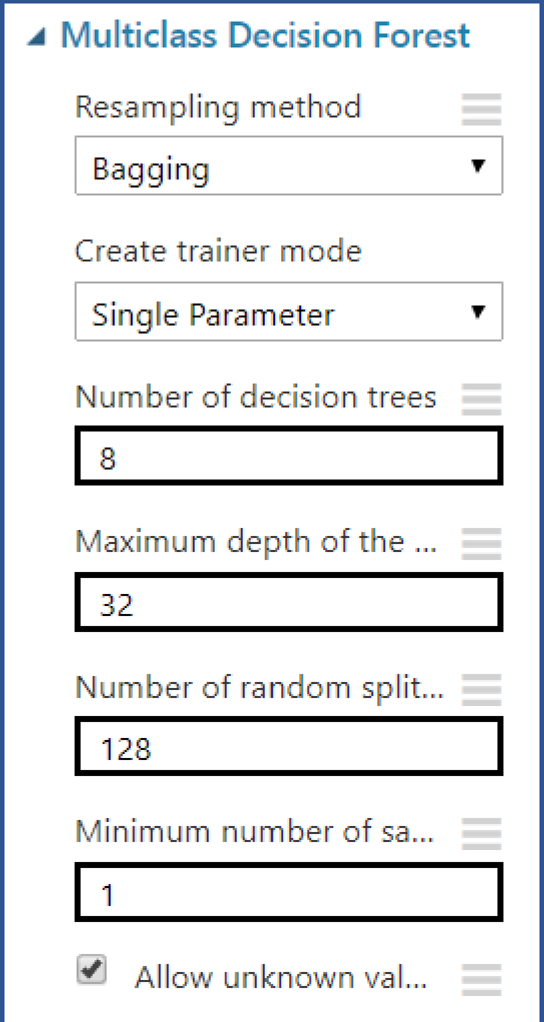
Select Columns in Dataset

Select columns

Selected columns:
All columns
Exclude column names: id,article

Algorithm Multi Class

Algorithm



◢ Multiclass Decision Forest

Resampling method: Bagging

Create trainer mode: Single Parameter

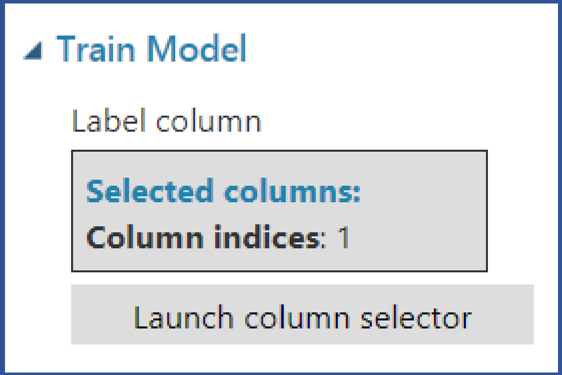
Number of decision trees: 8

Maximum depth of the ...: 32

Number of random split...: 128

Minimum number of sa...: 1

☑ Allow unknown val...



◢ Train Model

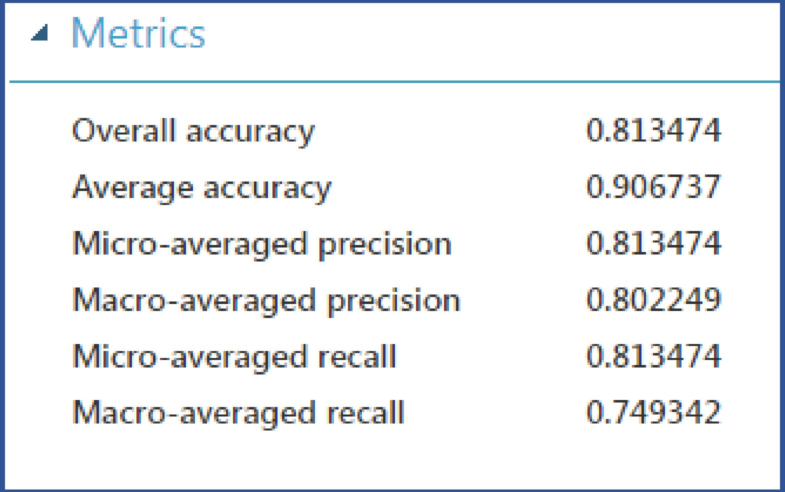
Label column

Selected columns:
Column indices: 1

Algorithm Multi Class

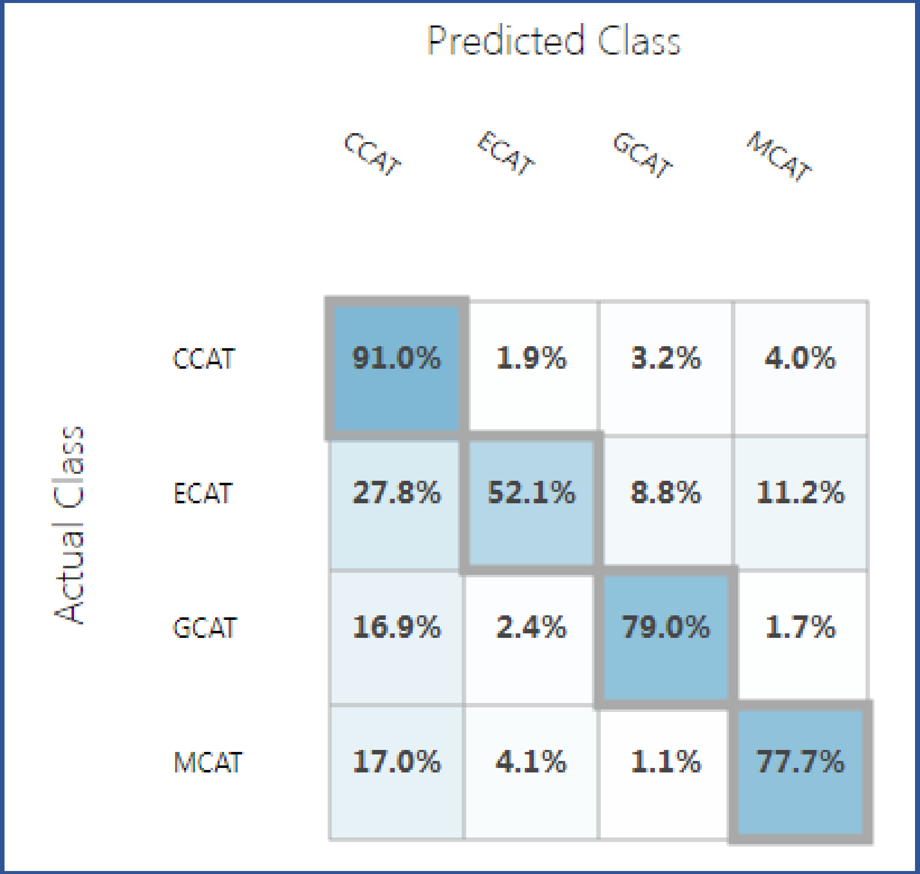
Confusion Matrix

Test data = https://raw.githubusercontent.com/laploy/ML/master/mul-test.txt



◢ Metrics

Overall accuracy	0.813474
Average accuracy	0.906737
Micro-averaged precision	0.813474
Macro-averaged precision	0.802249
Micro-averaged recall	0.813474
Macro-averaged recall	0.749342



Predicted Class

Actual Class

	CCAT	ECAT	GCAT	MCAT
CCAT	91.0%	1.9%	3.2%	4.0%
ECAT	27.8%	52.1%	8.8%	11.2%
GCAT	16.9%	2.4%	79.0%	1.7%
MCAT	17.0%	4.1%	1.1%	77.7%

Algorithm Multi Class

More information

Multiclass Decision Forest

https://msdn.microsoft.com/en-us/library/azure/dn906015.aspx

This Experiment

https://gallery.cortanaintelligence.com/Experiment/Multi-Class