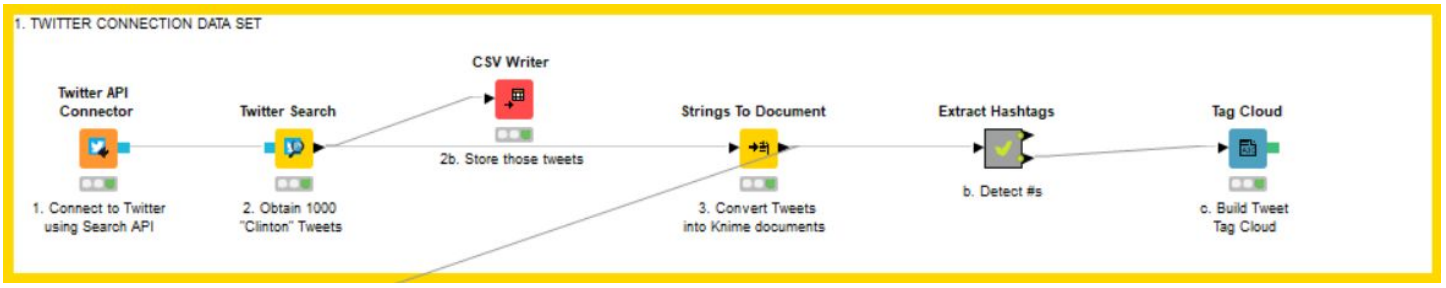
AUTHOR NAME: Tania Batista
OPTION: A
DELIVERY DATE: June 17, 2018

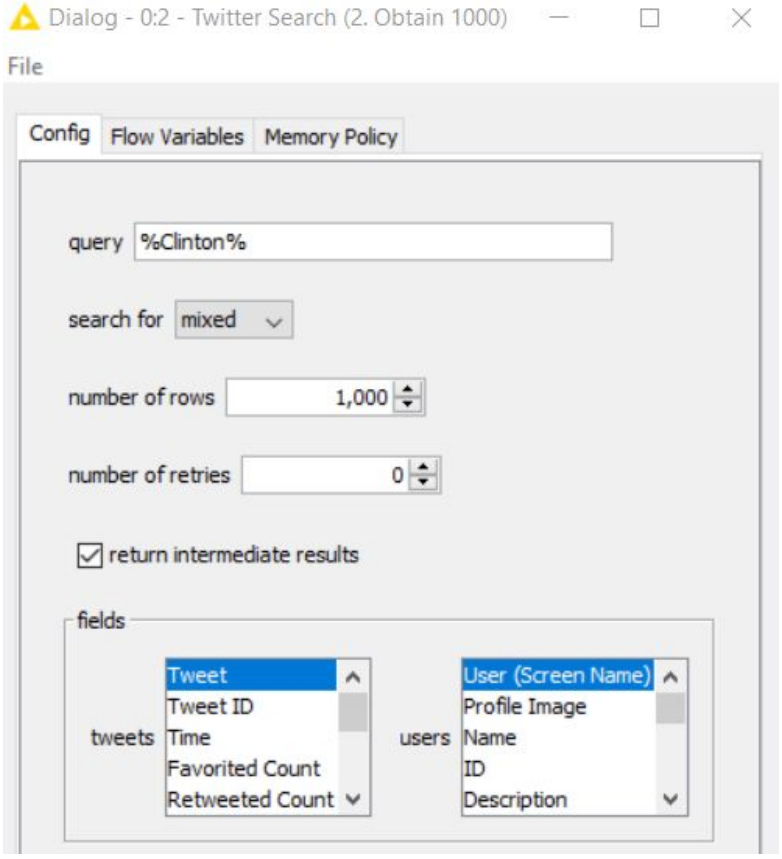
Downloaded Twitter Data Set (1)



Conectarse al API de Twitter o cualquier otra fuente, hacer una petición de búsqueda y obtener los resultados correspondientes (1 punto).

1. Connect to Twitter using the Search API

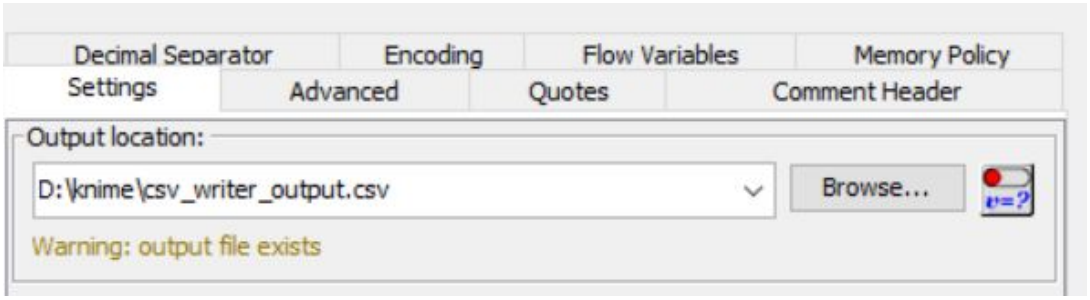
2. Obtain at least 1,000 tweets containing the string "Clinton"



Row ID	Tweet	User
Row0	6 of the top FBI brass have been fired, demoted, or reassigned following their involve...	Jim_Jordan
Row1	"Just because your child gets across the border that doesn't mean your child gets to s...	PrisonPlanet
Row2	The guy who hates the President (Peter Strzok), ran the Clinton investigation & the R...	PoliticalShort
Row3	RT @Education4Libs: Hillary Clinton, Laura Bush & Michelle Obama have all expressed ...	ayreon9

Almacenar los datos obtenidos para su posterior proceso (1 punto).

3. Store those tweets in a local output file for further analysis



4. Converted Tweets into KNIME documents

Row ID	Tweet	User	Document
Row0	6 of the top FBI brass have been fir...	Jim_Jordan	"6 of the top FBI brass have been fired, demoted, or reassigned following
Row1	"Just because your child gets acros...	PrisonPlanet	""Just because your child gets across the border that doesn't mean your c
Row2	The guy who hates the President (P...	PoliticalShort	"The guy who hates the President (Peter Strzok), ran the Clinton investiga
Row3	RT @Education4Libs: Hillary Clinton,...	ayreon9	"RT @Education4Libs: Hillary Clinton, Laura Bush & Michelle Obama have a

b. Extract the hashtags & amplify them according to TF

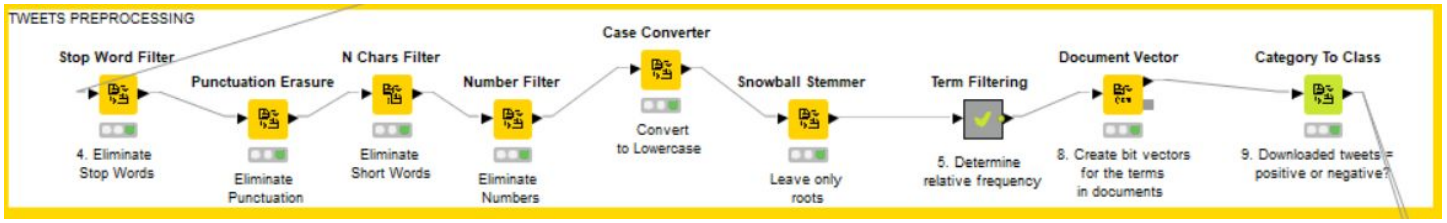
Row ID	Document	Preproc...	Term	TF rel	Title
Row1	"#donaldtrump why do people act like t...	"#donaldtru...	#donaldtrump[AMPLIFICATION(...	0.036	#donaldtrump
Row2	"RT @Condor_Law: UNBELIEVABLE!!!...	"#Lynch"	#Lynch[AMPLIFICATION(SENTI...	0.04	#Lynch
Row3	""Hillary Clinton" @CarterCenter @Mich...	"#FLUFOC"	#FLUFOC[AMPLIFICATION(SENT...	0.05	#FLUFOC
Row4	"RT @Condor_Law: UNBELIEVABLE!!!...	"#Lynch"	#Lynch[AMPLIFICATION(SENTI...	0.04	#Lynch

Detectar los hashtags y obtener la nube de tags o cualquier otra descripción de los datos (1 punto).

c. Build a “Tweet Tag Cloud” generated by hashtags and their frequency

Realizar el pre-procesamiento de los documentos (2 puntos).

4. Clean the Knime Documents Twitter Data Set:



Eliminate stop words

Preprocessed Document
"6 top FBI brass fired, demoted, reassigned following involvement Clinton ... https://t.co/YAfCmYmDVg"
""child border n't mean child stay."Hillary Clinton, 2014. https://t.co/dsWQ93ITTW"
"guy hates President (Peter Strzok), ran Clinton investigation & Russia investigation, ... https://t.co/xcsIIYB2GV"
"RT @Education4Libs: Hillary Clinton, Laura Bush & Michelle Obama expressed disgust border separation policy.the..."
"RT @The_War_Economy: @TheChiIIum @libbycwatson spoke recently! 's justified Correct Record directly..."

Eliminate punctuation

"6 top FBI brass fired demoted reassigned following involvement Clinton ... httpstcoYAfCmYmD...
"child border nt mean child stayHillary Clinton 2014 httpstcodsWQ93ITTW"
"guy hates President Peter Strzokran Clinton investigation Russia investigation ... httpstcoxcsI...
"RT Education4Libs Hillary Clinton Laura Bush Michelle Obama expressed disgust border separ...
"RT TheWarEconomy TheChiIIum libbycwatson spoke recently s justified Correct Record directl...

Eliminate words that don't have minimum 4 characters

"brass fired demoted reassigned following involvement Clinton httpstcoYAfCmYmDVg"
"child border mean child stayHillary Clinton 2014 httpstcodsWQ93ITTW"
"hates President Peter StrzokClinton investigation Russia investigation httpstcoxcsIIYB2GV"
"Education4Libs Hillary Clinton Laura Bush Michelle Obama expressed disgust border separation policythe..."
"TheWarEconomy TheChiIIum libbycwatson spoke recently justified Correct Record directly"

Eliminate number characters

"brass fired demoted reassigned following involvement Clinton httpstcoYAfCmYmDVg"
"child border mean child stayHillary Clinton httpstcodsWQ93ITTW"
"hates President Peter StrzokClinton investigation Russia investigation httpstcoxcsIIYB2GV"
"Education4Libs Hillary Clinton Laura Bush Michelle Obama expressed disgust border separation policythe..."
"TheWarEconomy TheChiIIum libbycwatson spoke recently justified Correct Record directly"

Convert all to lowercase

"brass fired demoted reassigned following involvement clinton httpstcoyafcmymd...
"child border mean child stayhillary clinton httpstcodswq93ittw"
"hates president peter strzokclinton investigation russia investigation httpstcoxc...
"education4libs hillary clinton laura bush michelle obama expressed disgust borde...
"thewareconomy thechiiium libbycwatson spoke recently justified correct record d...

Eliminate English stop words

"brass fired demoted reassigned following involvement clinton httpstcoyafcmymdvg"
"child border mean child stayhillary clinton httpstcodswq93ittw"
"hates president peter strzokclinton investigation russia investigation httpstcoxcsiiyb2gv"
"education4libs hillary clinton laura bush michelle obama expressed disgust border separation policythe..."
"thewareconomy thechiiium libbycwatson spoke recently justified correct record directly"

Leave only the roots of each word

"brass fire demot reassign follow involv clinton httpstcoyafcmymdvg"
"child border mean child stayhillari clinton httpstcodswq93ittw"
"hate presid peter strzokclinton investig russia investig httpstcoxcsiiyb2gv"
"education4lib hillari clinton laura bush michell obama express disgust border separ policithe..."
"thewareconomi thechiiium libbycwatson spoke recent justifi correct record direct"

5. Determine the relative frequency of terms using Snowball stemmer, BoW, TF, PE.

hate[]	...	...	...	0.1
presid[]	...	...	...	0.1
clinton[]	...	...	...	0.1
investig[]	...	...	...	0.2
https[]	...	...	...	0.1
education4lib[]	...	...	...	0.077
hillari[]	...	...	...	0.077
clinton[]	...	...	...	0.077

6. Create a Vector to visualize the frequency of the terms in each twitter document

realdon...	hillari	peac	clinton	leadership	human	corrupt	https	trump	polici	bush	crisi	moral
1	1	1	1	1	1	0	0	0	0	0	0	0
0	0	0	1	0	0	1	1	0	0	0	0	0
0	0	0	0	0	0	0	1	1	0	0	0	0
0	0	0	0	0	0	0	1	0	1	0	0	0
0	0	0	1	0	0	0	0	0	0	1	0	0
0	1	0	1	0	0	0	1	0	0	0	1	1
0	1	0	1	0	0	0	1	0	0	0	1	1
0	1	0	1	0	0	0	1	0	0	0	1	1

7. Add a class (negative/positive, 0/1) to each document row. For now the class value is undefined, because we have no reference yet to apply these predictions.

Columns: 197	...	...	...	...	...	...	Lower Bound	Upper Bound	Value 0
realdonaldtrump	...	1					0	1	?
hillari	...	2					0	1	?
peac	...	3					0	1	?
clinton	...	4					0	1	?
leadership	...	5					0	1	?

The TRAINING Data Set (2)

7. Load the training data set, "Sentiment140 Tweet Corpus"

training.1600000.processed.noemoticon.csv

Rename the columns

Row ID	Senti...	ID	Date	Query	User	Tweet
Row0	0	1,467,810,369	Mon Apr 06 2...	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, that's a bummer. Yo
Row1	0	1,467,810,672	Mon Apr 06 2...	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by texting it... and might
Row2	0	1,467,810,917	Mon Apr 06 2...	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Managed to save 50% T
Row3	0	1,467,811,184	Mon Apr 06 2...	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
Row4	0	1,467,811,193	Mon Apr 06 2...	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all. i'm mad. why am i here

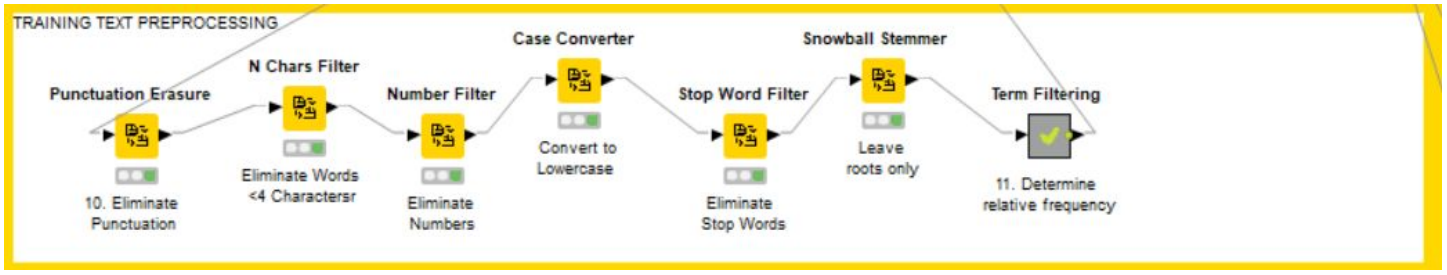
8. Apply the Row Sampling node to extract 4000 random training tweets. The training set is too large to load all of the strings to documents.

Row ID	Sentimi...	ID	Date	Query	User	Tweet
Row99	0	1,467,836,859	Mon Apr ...	NO_QUERY	willy_chaz	A bad nite for the favorite teams: Astro
Row194	0	1,467,860,144	Mon Apr ...	NO_QUERY	Jana1976	@JonathanRKnight I hate the limited let
Row265	0	1,467,878,929	Mon Apr ...	NO_QUERY	VirtueIMC	@chriscantore congrats! I'm totally jeal
Row394	0	1,467,911,846	Mon Apr ...	NO_QUERY	livetosingxo	@bananaface IM SORRY I GOT YOU SI
Row504	0	1,467,937,128	Mon Apr ...	NO_QUERY	treywoodward	I am listing more items on ebay to sell!

9. Build documents for each individual <140-word tweet

Document
"A bad nite for the favorite teams: Astros and Spartans lose. The nite out with T.W. was goo...
"@JonathanRKnight I hate the limited letters,too.Hope you and the guys are fine?I pray for ...
"@chriscantore congrats! I'm totally jealous! only wish my XM was working"
"@bananaface IM SORRY I GOT YOU SICK. lol. going to bed too. NIGHT!"
"I am listing more items on ebay to sell! Takes forever..meanwhile Coleman is watching The Hi...

10. Preprocess the Training Text:



Remove punctuation:

Preprocessed Document
"A bad nite for the favorite teams Astros and Spartans lose The nite out with TW was go...
"JonathanRKnight I hate the limited letterstooHope you and the guys are fineI pray for m...
"chriscantore congrats Im totally jealous only wish my XM was working"
"bananaface IM SORRY I GOT YOU SICK lol going to bed too NIGHT"
"I am listing more items on ebay to sell Takes forevermeanwhile Coleman is watching The ...

Remove words 3-characters or less:

Preprocessed Document
"nite favorite teams Astros Spartans lose nite with good"
"JonathanRKnight hate limited lettersHope guys finepray sheï¿½s well"
"chriscantore congrats totally jealous only wish working"
"bananaface SORRY SICK going NIGHT"
"listing more items ebay sell Takes forevermeanwhile Coleman watching Hills season ...

Remove numbers

Document	Preprocessed Document
"I'm gonna cry I only did a 4 club hop tonight It was su...	"gonna only club tonight supposed hate"
"Still out of the "Real" world... no even goo...	"Still quotRealquot world even good cell range"

Convert all text to lowercase

Document	Preprocessed Document
"A bad nite for the favorite teams: Astros and Spartans lose. The nite out with T.W. was goo...	"nite favorite teams astros spartans lose nite with good"
"@JonathanRKnight I hate the limited letters,too.Hope you and the guys are fine?I pray for ...	"jonathanrknight hate limited lettershope guys finepray sheï¿½s w
"@chriscantore congrats! I'm totally jealous! only wish my XM was working"	"chriscantore congrats totally jealous only wish working"
"@bananaface IM SORRY I GOT YOU SICK. lol. going to bed too. NIGHT!"	"bananaface sorry sick going night"

Remove stop words

Document	Preprocessed Document
"A bad nite for the favorite...	"nite favorite teams astros spartans lose nite"
"@JonathanRKnight I hate ...	"jonathanrknight hate limited lettershope guys finepray sheï¿½s"
"@chriscantore congrats! I'...	"chriscantore congrats totally jealous wish"
"@bananaface IM SORRY I...	"bananaface sorry sick night"
"I am listing more items on ...	"listing items ebay sell takes forevermeanwhile coleman watching hills season premier"
"@JinxCat Unlike my sister....	"jinxcat unlike sister"

Leave only roots. (Snowball stemmer set to English)

Preprocessed Document
"nite favorit team astro spartan lose nite"
"jonathanrknight hate limit letterhope guy finepray sheï¿½"
"chriscantor congrat total jealous wish"
"bananafac sorri sick night"
"list item ebay sell take forevermeanwhil coleman watch hill season premier"

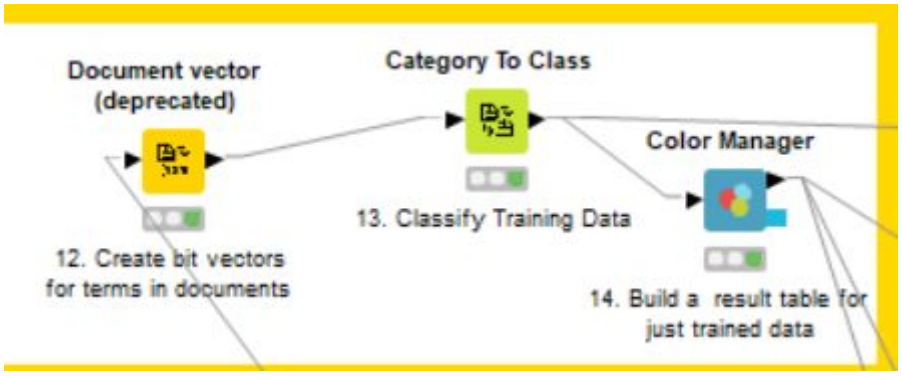
11. Determine the relative frequency of terms in each document:

S Term as String	D TF rel
hate	0.111
hope	0.111
wish	0.2
sorri	0.25

12. Create a Vector to visualize the frequency of the terms in each Training document

Row ID	Document	D school	D haha	D start	D quot	D http	D happi	D friend	D pretti
Row0	"0mark0 schoolkeep busi haha start haha be...	1	1	1	0	0	0	0	0
Row1	"1azylizzi tragic life forc drug abus quotretire...	0	0	0	1	0	0	0	0
Row2	"25mill joanpenthous steal httpbitlylynjg"	0	0	0	0	1	0	0	0
Row3	"2day happi yayyy erick happi b-day mannha...	0	1	0	0	0	1	1	0
Row4	"8bithack pretti expect tail matur game busi c...	0	0	0	0	0	0	0	1

13. Classify the data: positive or negative? 0 or 1?



Columns: 54	Column Type	...	...	...	...	...	Lower Bound	Upper Bound	Value 0	Value 1
Document	Text document	0					?	?	?	?
school	Number (dou...	1					0	1	?	?
haha	Number (dou...	2					0	1	?	?
start	Number (dou...	3					0	1	?	?
quot	Number (dou...	4					0	1	?	?
http	Number (dou...	5					0	1	?	?

14. Use the Color Manager to build the Training data result table:

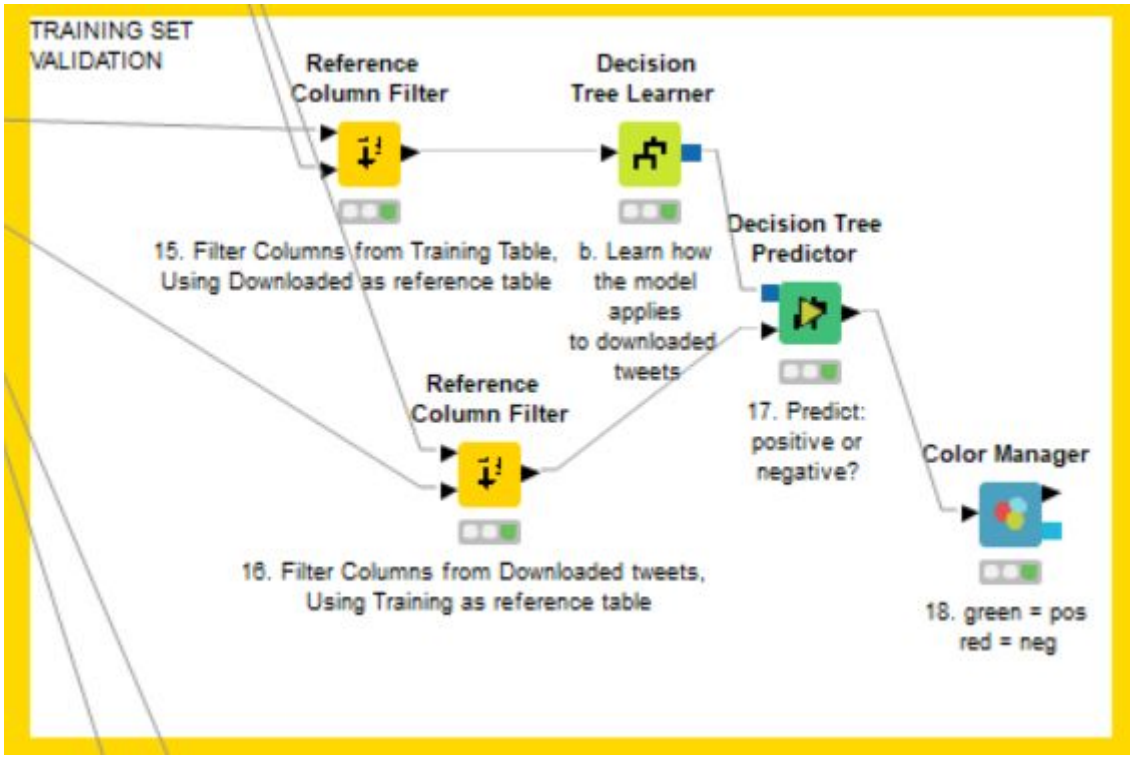
Green = positive document

Red = negative document

Row ID	Document	school	haha	start	quot	http	happi	friend	pretti
Row0	"0mark0 schoolkeep busi haha start haha be...	1	1	1	0	0	0	0	0
Row1	"1azylizzi tragic life forc drug abus quotretire...	0	0	0	1	0	0	0	0
Row2	"25mill joanpenthous steal httpbitlylynjg"	0	0	0	0	1	0	0	0
Row3	"2day happi yayyy erick happi b-day mannha...	0	1	0	0	0	1	1	0
Row4	"8bithack pretti expect tail matur game busi c...	0	0	0	0	0	0	0	1

Entrenar el modelo con los datos de entrenamiento (2 puntos).

VALIDATION



15. Reference Column Filters

a. Columns from 1st and 2nd table are only included if they match.

Row ID	Document	quot	watch	peopl	hate	Docum...
Row0	"0mark0 schoolkeep busi haha start haha be...	0	0	0	0	4
Row1	"1azylizzi tragic life forc drug abus quotretire...	1	0	0	0	0
Row2	"25mill joanpenthous steal httpbitlylynjg"	0	0	0	0	0
Row3	"2day happi yayyy erick happi b-day mannha...	0	0	0	0	4

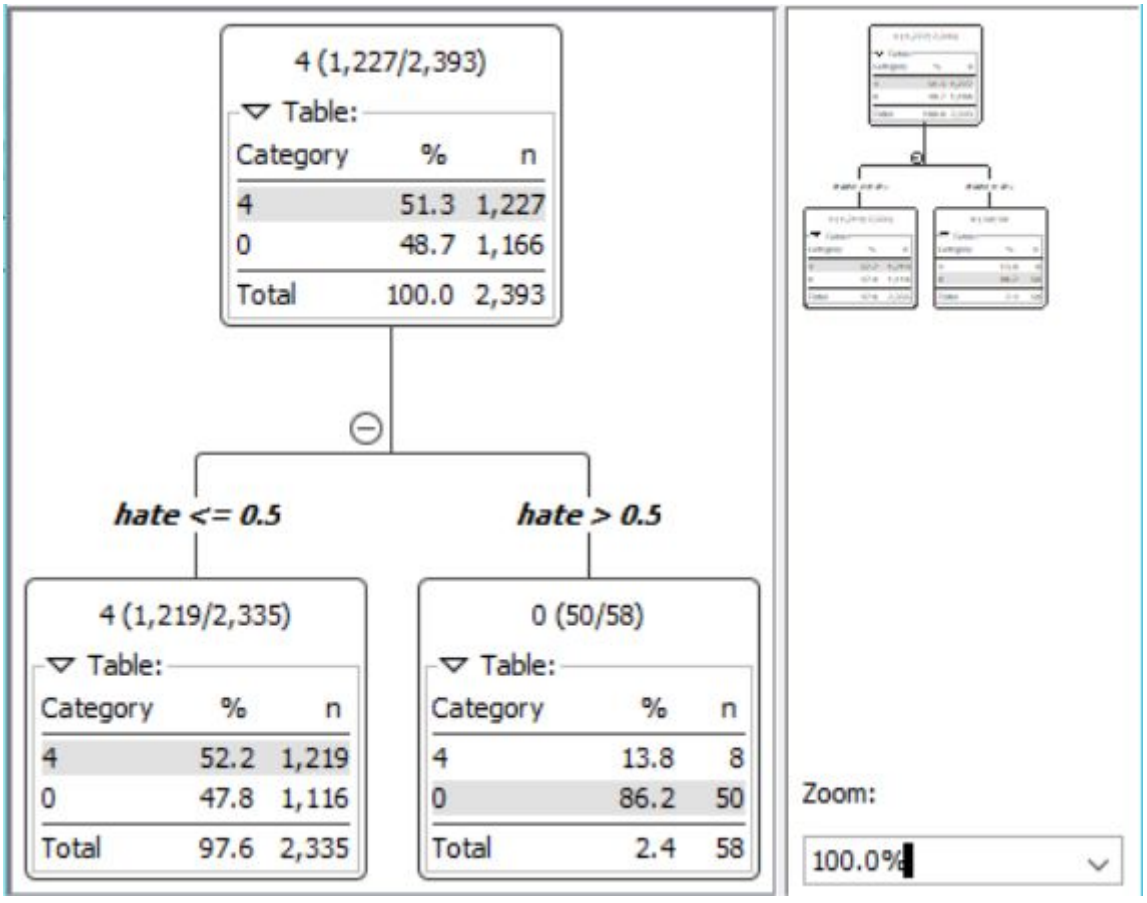
b. Decision Tree Learner = induces a classification decision tree in main memory.
Numeric splits are always binary (two outcomes) It uses a post pruning method to reduce the tree size and increase prediction accuracy.

16. Reference Column Filters

Columns from 2nd and 1st table are included if they match

Row ID	Document	watch	peopl	quot	hate	Docum...
Row0	"2021free realdonaldtrump hillari nobel peac p...	0	0	0	0	undefined
Row1	"2ndalexhiggin jrdnmdhl valkyrievaljean matty...	0	0	0	0	undefined
Row2	"55elkhunt tomilahren believ deep trump push ...	0	0	0	0	undefined
Row3	"aaronchann therealroseann polici cruel unpr...	0	0	0	0	undefined
Row4	"accenttoast vinnygb1 jebbush clinton fucn bu...	0	0	0	0	undefined

17. Decision Tree: uses the tree in #15b to predict the classifications for new patterns.



18. Apply color to the outcomes

The TEST Data Set (3)

19. Load the test data set, "Sentiment140 Test Dataset"

testdata.manual.2009.06.14.csv

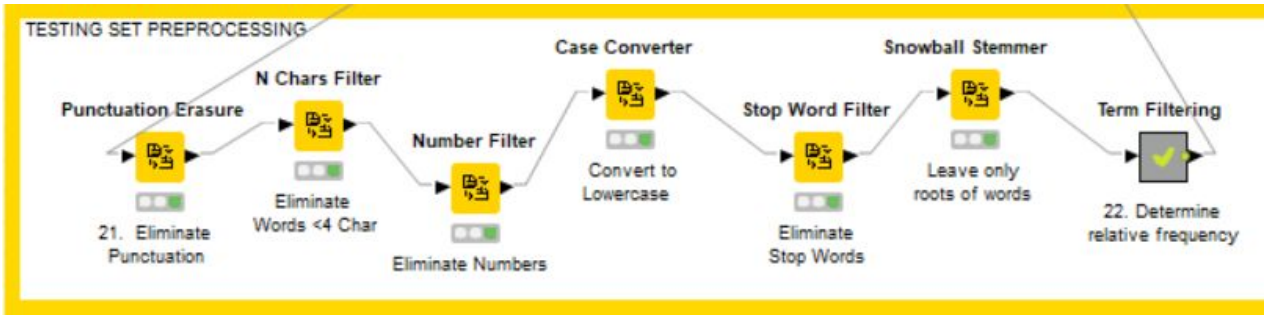
Rename the columns. No need to sample rows - there are not too many.

Row ID	Se...	ID	Date	Query	User	Tweet
Row0	4	3	Mon May 11 03:17:40 UTC 2009	kindle2	tpryan	@stellargirl I loooooooovvvvvveee my Kindle2. Not that the DX is cool, but the 2 is fantastic in its own right
Row1	4	4	Mon May 11 03:18:03 UTC 2009	kindle2	vcu451	Reading my kindle2... Love it... Lee childs is good read.
Row2	4	5	Mon May 11 03:18:54 UTC 2009	kindle2	chadfu	Ok, first assesment of the #kindle2 ...it fucking rocks!!!
Row3	4	6	Mon May 11 03:19:04 UTC 2009	kindle2	SIX15	@kenburbary You'll love your Kindle2. I've had mine for a few months and never looked back. The new big.
Row4	4	7	Mon May 11 03:21:41 UTC 2009	kindle2	yamarama	@mikefish Fair enough. But i have the Kindle2 and I think it's perfect :)

20. Build documents for each individual <140-word tweet

Document
"@stellargirl I loooooooovvvvvveee my Kindle2. Not that the DX is cool, but the 2 is fantastic in its own rig...
"Reading my kindle2... Love it... Lee childs is good read."
"Ok, first assesment of the #kindle2 ...it fucking rocks!!!"
"@kenburbary You'll love your Kindle2. I've had mine for a few months and never looked back. The new big...
"@mikefish Fair enough. But i have the Kindle2 and I think it's perfect :)"
"@richardebaker no. it is too big. I'm quite happy with the Kindle2."

21. Preprocess the Test Text: Clean all the words in the documents via the preprocessing filters



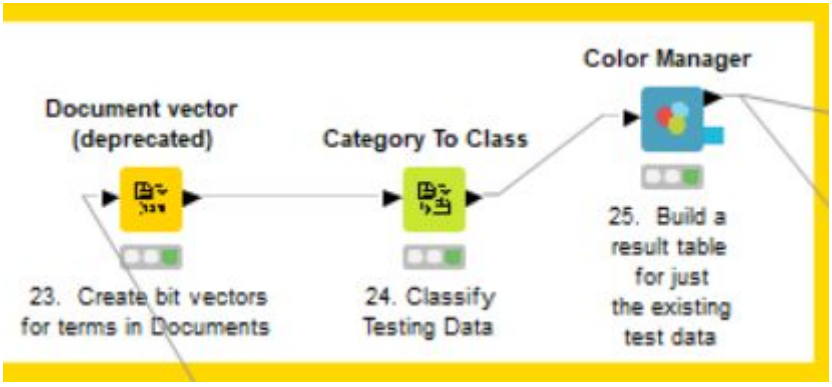
Remove: punctuation, words < 4 characters, numbers, stop words
Convert to lowercase
Leave only stems

Preprocessed Document
"stellargirl loooooooovvvvvvee kindle2 cool fantast"
"read kindle2 love child read"
"asses kindle2 fuck rock"
"kenburbari love kindle2 mine month look huge remors"
"mikefish fair kindle2 perfect"
"richardebak happi kindle2"

22. Determine the relative frequecy of terms in each document

Row ID	Term	...	Term as String	TF rel
Row3	kindle2[]	"st...	kindle2	0.2
Row4	cool[]	"st...	cool	0.2
Row6	read[]	"re...	read	0.4
Row7	kindle2[]	"re...	kindle2	0.2
Row8	love[]	"re...	love	0.2

23. Create a vector to visualize the frequency of the terms in each Test document



Row ID	Document	http	stanford	thank	help	start	school	tip	commerci
Row1449	"absolut hilarimashabl http...	1	0	0	0	0	0	0	0
Row1450	"accanni edog1203 stanfor...	0	1	1	1	1	0	0	0
Row1451	"accord creat school notr d...	0	0	0	0	0	1	0	0
Row1452	"accost roger feder french...	1	0	0	0	0	0	0	0
Row1453	"admiss tip stanford deadli...	1	1	0	0	0	0	1	0
Row1454	"adob commerci goodbi silv...	1	0	0	0	0	0	0	1

24. Classify the data: positive or negative? 0 or 1?

Table "default" - Rows: 483 | Spec - Columns: 148 | Properties | Flow Variables

Row ID	Document	http	stanford	thank	help	start	school	tip	commerci
Row1449	"absolut hilarimashabl httpbitlybcc...	1	0	0	0	0	0	0	0
Row1450	"accanni edog1203 stanford cours ...	0	1	1	1	1	0	0	0
Row1451	"accord creat school notr dame rec...	0	0	0	0	0	1	0	0
Row1452	"accost roger feder french httpffim...	1	0	0	0	0	0	0	0
Row1453	"admiss tip stanford deadlin essay ...	1	1	0	0	0	0	1	0

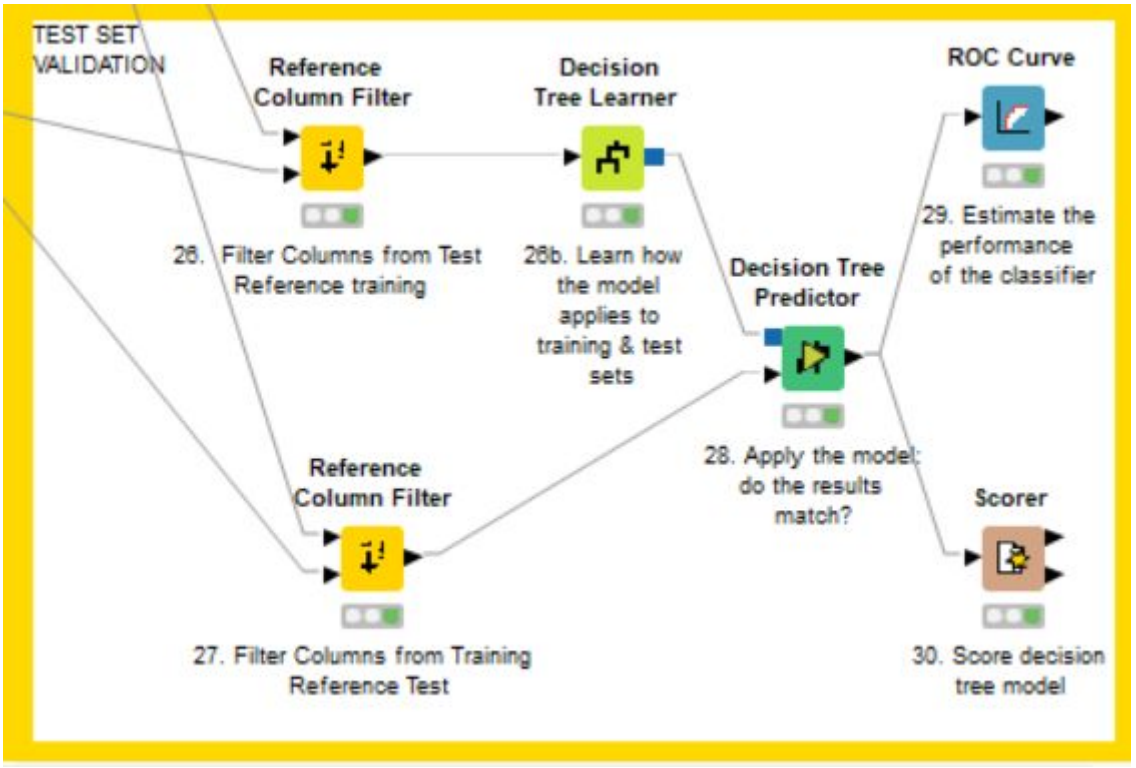
25. Use the color manager to build the test data result table:

Row ID	Document	http	stanford	thank	help	start	school	tip	commerci	goodbi	silverstein	north	korea
Row1449	"absolut hilarimashabl httpbitlybcc...	1	0	0	0	0	0	0	0	0	0	0	0
Row1450	"accanni edog1203 stanford cours ...	0	1	1	1	1	0	0	0	0	0	0	0
Row1451	"accord creat school notr dame rec...	0	0	0	0	0	1	0	0	0	0	0	0
Row1452	"accost roger feder french httpffim...	1	0	0	0	0	0	0	0	0	0	0	0
Row1453	"admiss tip stanford deadlin essay ...	1	1	0	0	0	0	1	0	0	0	0	0
Row1454	"adob commerci goodbi silverstein ...	1	0	0	0	0	0	0	1	1	1	0	0
Row1455	"adob goodbi silverstein partner yo...	1	0	0	0	0	0	0	0	1	1	0	0
Row1456	"ahead blow north korea"	0	0	0	0	0	0	0	0	0	0	1	1

…. This data set is going to skew the results, because in addition to positive and negative sentiments, we also have neutral.

Evaluar el modelo con el conjunto de test (1 punto).

VALIDATION



26. Reference Column filters:

a. Columns from 1st and 2nd table are included only if they match

Row ID	Document	school	start	http	happi	pretti
Row0	"0mark0 schoolkeep busi haha start haha be...	1	1	0	0	0
Row1	"1azylizzi tragic life forc drug abus quotretire...	0	0	0	0	0
Row2	"25mill joanpenthous steal httpbitlylynjg"	0	0	1	0	0
Row3	"2day happi yayyy erick happi b-day mannha...	0	0	0	1	0
Row4	"8bithack pretti expect tail matur game busi c...	0	0	0	0	1

b. Decision tree learner:

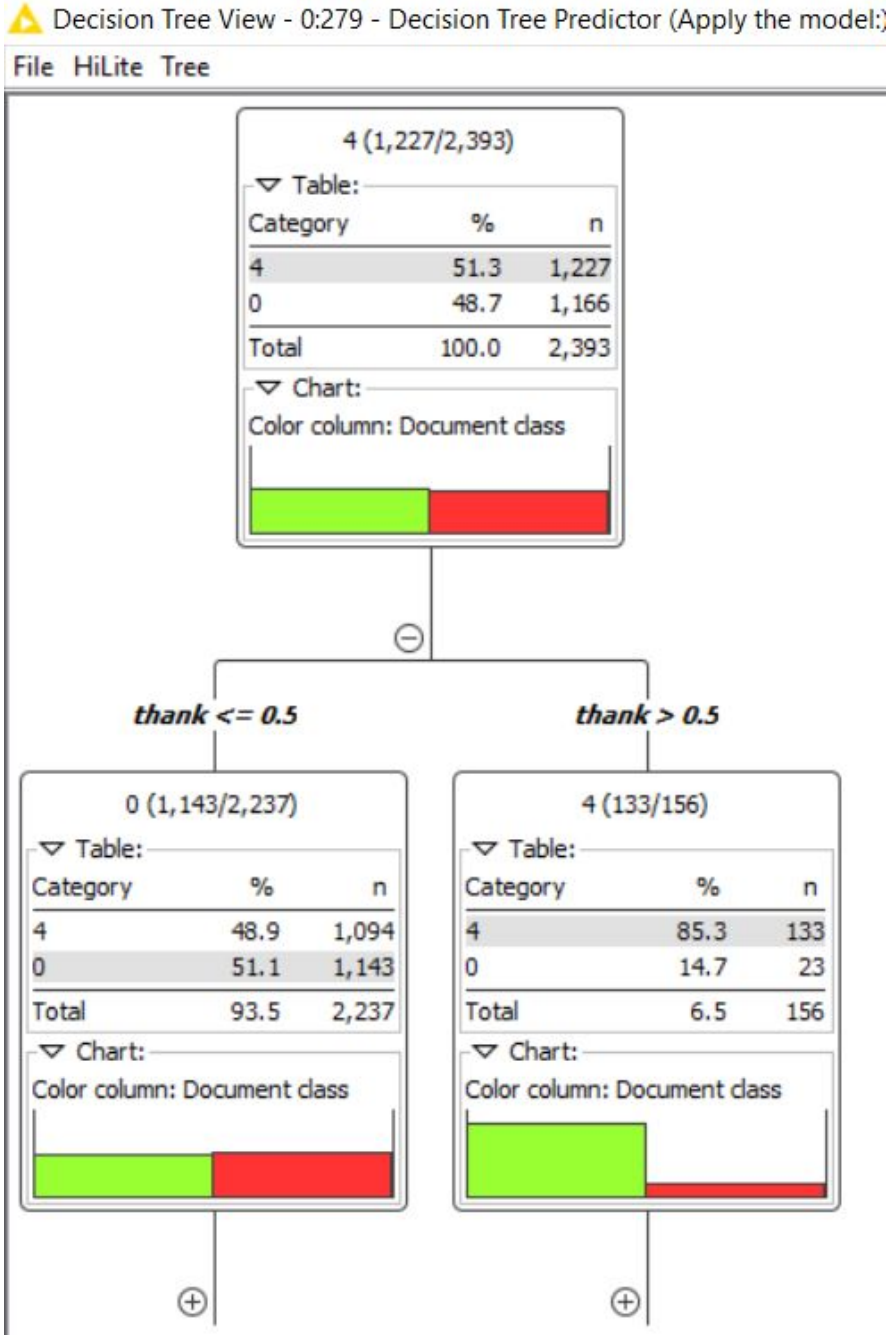
Colum...	Column Type	C...	C...	...	...	...	Lower Bound	Upper Bound
school	Number (double)	0					0	1
start	Number (double)	1					0	1
http	Number (double)	2					0	1
happi	Number (double)	3					0	1
pretti	Number (double)	4					0	1
tomorrow	Number (double)	5					0	1
look	Number (double)	6					0	1
tonight	Number (double)	7					0	1

27. Reference Column filters:

Row ID	Document	school	start	http	happi	pretti
Row0	"0mark0 schoolkeep busi haha start haha be...	1	1	0	0	0
Row1	"1azylizzi tragic life forc drug abus quotretire...	0	0	0	0	0
Row2	"25mill joanpenthous steal httpbitlylynjg"	0	0	1	0	0
Row3	"2day happi yayyy erick happi b-day mannha...	0	0	0	1	0
Row4	"8bithack pretti expect tail matur game busi c...	0	0	0	0	1

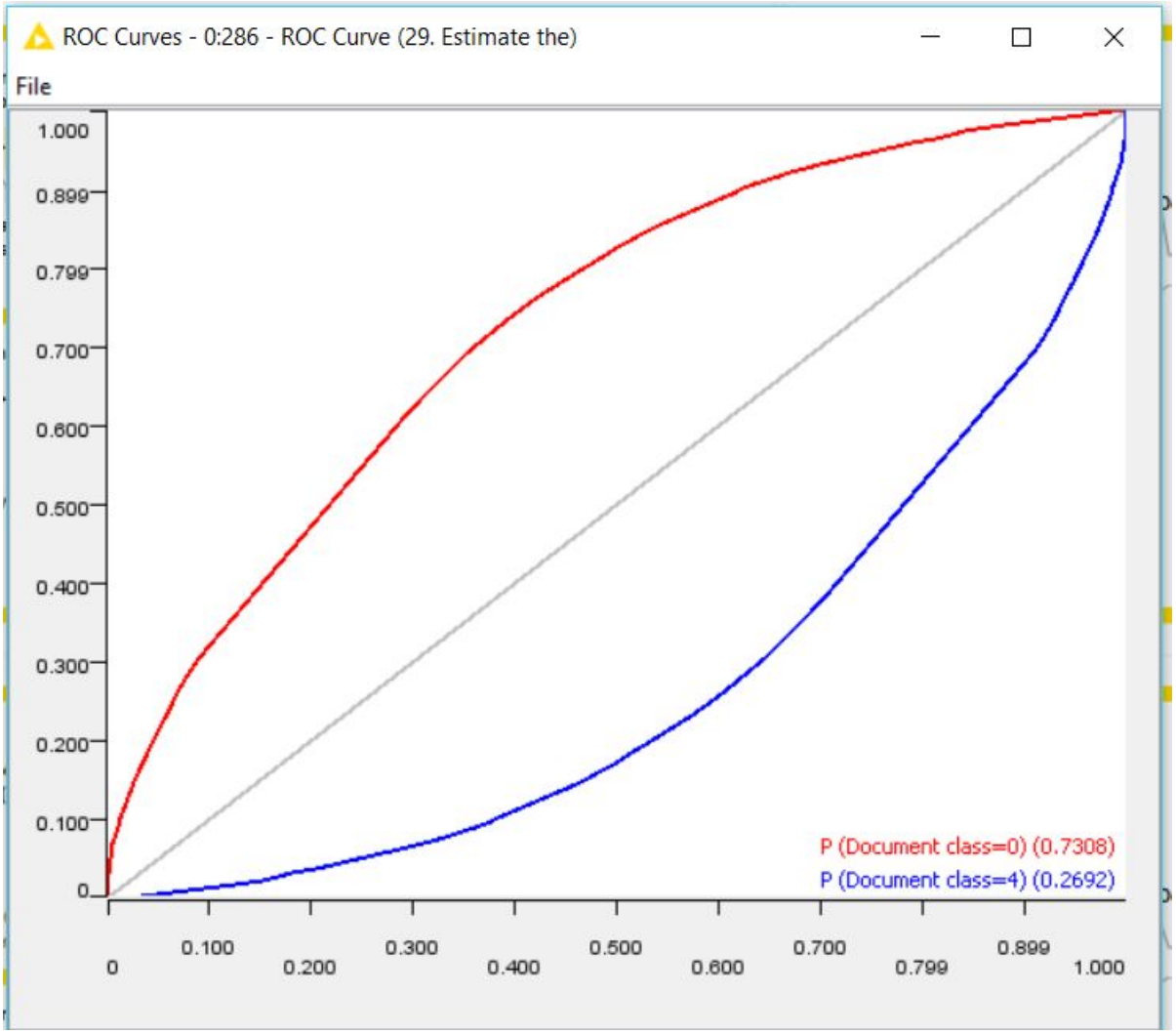
28. Decision Tree Predictor

Row ID	Document	school	start	http	happi	pretti	tomorrow
Row0	"0mark0 schoolkeep busi haha start haha be...	1	1	0	0	0	0
Row1	"1azylizzi tragic life forc drug abus quotretire...	0	0	0	0	0	0
Row2	"25mill joanpenthous steal httpbitlylynjg"	0	0	1	0	0	0
Row3	"2day happi yayyy erick happi b-day mannha...	0	0	0	1	0	0
Row4	"8bithack pretti expect tail matur game busi c...	0	0	0	0	1	0
Row5	"955am offic happi bday"	0	0	0	1	0	0
Row6	"aaddiew damn dont school actual school"	1	0	0	0	0	0
Row7	"aah1981 yeahthat tomorrow"	0	0	0	0	0	1



Evaluar el modelo entrenado previamente (2 puntos).

29. ROC Curve: estimate the performance of the classifie:



30. Accuracy Statistics

Confusion Matrix

Row ID	4	0
4	752	475
0	316	850

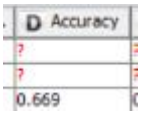
TruePositives, FalsePositives, TrueNegatives, FalseNegatives, Recall, Precision, Sensitivity, Specificity

Table "default" - Rows: 3 | Spec - Columns: 11 | Properties | Flow Variables

Row ID	TruePo...	FalsePo...	TrueNe...	FalseN...	Recall	Precision	Sensitivity	Specifity	F-meas...	Accuracy	Cohen'...
4	752	316	850	475	0.613	0.704	0.613	0.729	0.655	?	?
0	850	475	752	316	0.729	0.642	0.729	0.613	0.682	?	?
Overall	?	?	?	?	?	?	?	?	?	0.669	0.341

CONCLUSIONS

The accuracy for my model was 67%.



This means that my model can only predict the positive/negative classification of my downloaded Twitter set two thirds of the time.

One of the main difficulties of this assignment was the fact that we could only put small sample sets in the training model, in my case just 4,000 rows of data. Knime works very simply, since it can avoid programming. But the processing power is very low. We would have had higher accuracy results with a Python script.

Another problem was that we trained the model with two different data sets: training and test excel sheets. These both contained different classifications: the former had only positives and negatives, the latter contained positive, negative and neutral classifications. Ideally, it would be better to exclude the test set and include a larger training set to form a decision tree model.