Carlos Augusto Amador Manilla
A01329447

# Testing two resampling methods or classifiers for solving class imbalance problems

Para esta tarea se extrajeron características adicionales de los tweets:

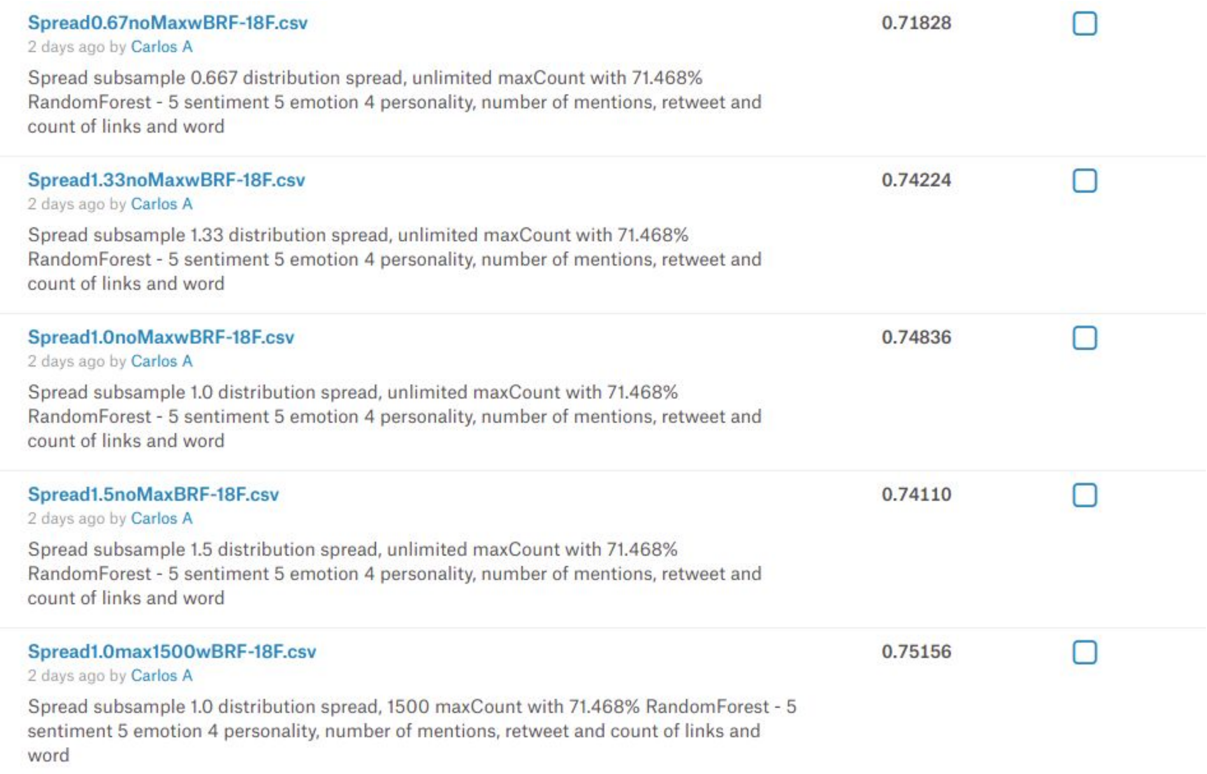
- Cantidad de menciones que contiene el tweet (numérica).
- Es retweet o no (nominal).
- Cantidad de ligas que contiene el tweet (numérica).
- Cantidad de palabras que contiene el tweet (numérica).

En primer lugar se probó con el clasificador Random Forest que mejor resultados había obtenido, agregando una característica a la vez. Se comprobó que cada característica mejoraba los resultados.

Con estas nuevas características se realizaron las pruebas con dos algoritmos de remuestreo y dos metaclasificadores sensibles al costo, utilizando como base el clasificador Random Forest previamente mencionado.

## Spread subsample

Utilizado para generar una muestra aleatoria del dataset, probé con diferentes valores para la distribución de clases, desde distribución uniforme hasta distribucion 2:1.



Spread0.67noMaxwBRF-18F.csv
2 days ago by Carlos A
Spread subsample 0.667 distribution spread, unlimited maxCount with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word
0.71828 □

Spread1.33noMaxwBRF-18F.csv
2 days ago by Carlos A
Spread subsample 1.33 distribution spread, unlimited maxCount with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word
0.74224 □

Spread1.0noMaxwBRF-18F.csv
2 days ago by Carlos A
Spread subsample 1.0 distribution spread, unlimited maxCount with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word
0.74836 □

Spread1.5noMaxBRF-18F.csv
2 days ago by Carlos A
Spread subsample 1.5 distribution spread, unlimited maxCount with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word
0.74110 □

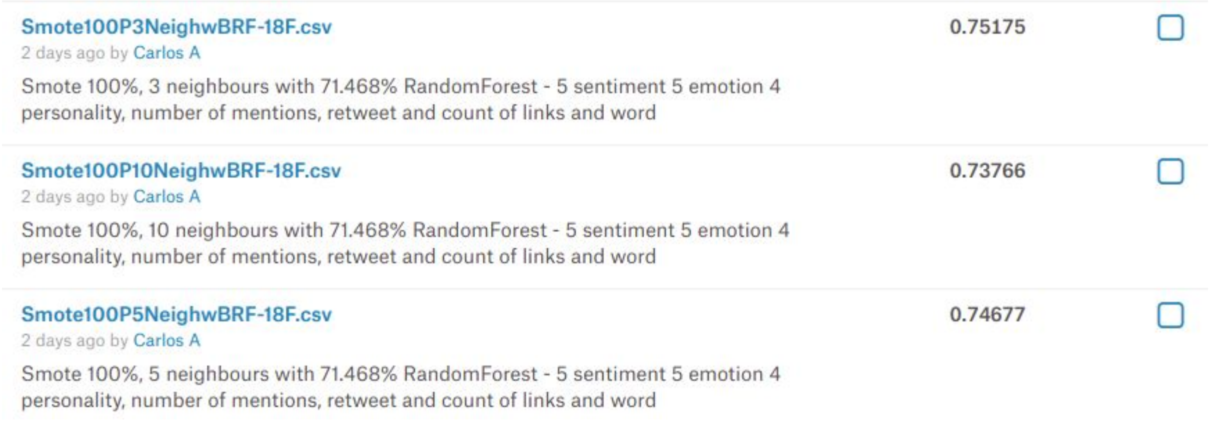
Spread1.0max1500wBRF-18F.csv
2 days ago by Carlos A
Spread subsample 1.0 distribution spread, 1500 maxCount with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word
0.75156 □

Tener mayor distribución de humanos obtuvo mejores resultados.

Carlos Augusto Amador Manilla
A01329447

# SMOTE

Genera nuevos datos de la clase minoritaria con información de k de sus vecinos más cercanos. Se utilizó para generar más instancias de la clase human.



Smote100P3NeighwBRF-18F.csv
2 days ago by Carlos A
Smote 100%, 3 neighbours with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word
0.75175 □

Smote100P10NeighwBRF-18F.csv
2 days ago by Carlos A
Smote 100%, 10 neighbours with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word
0.73766 □

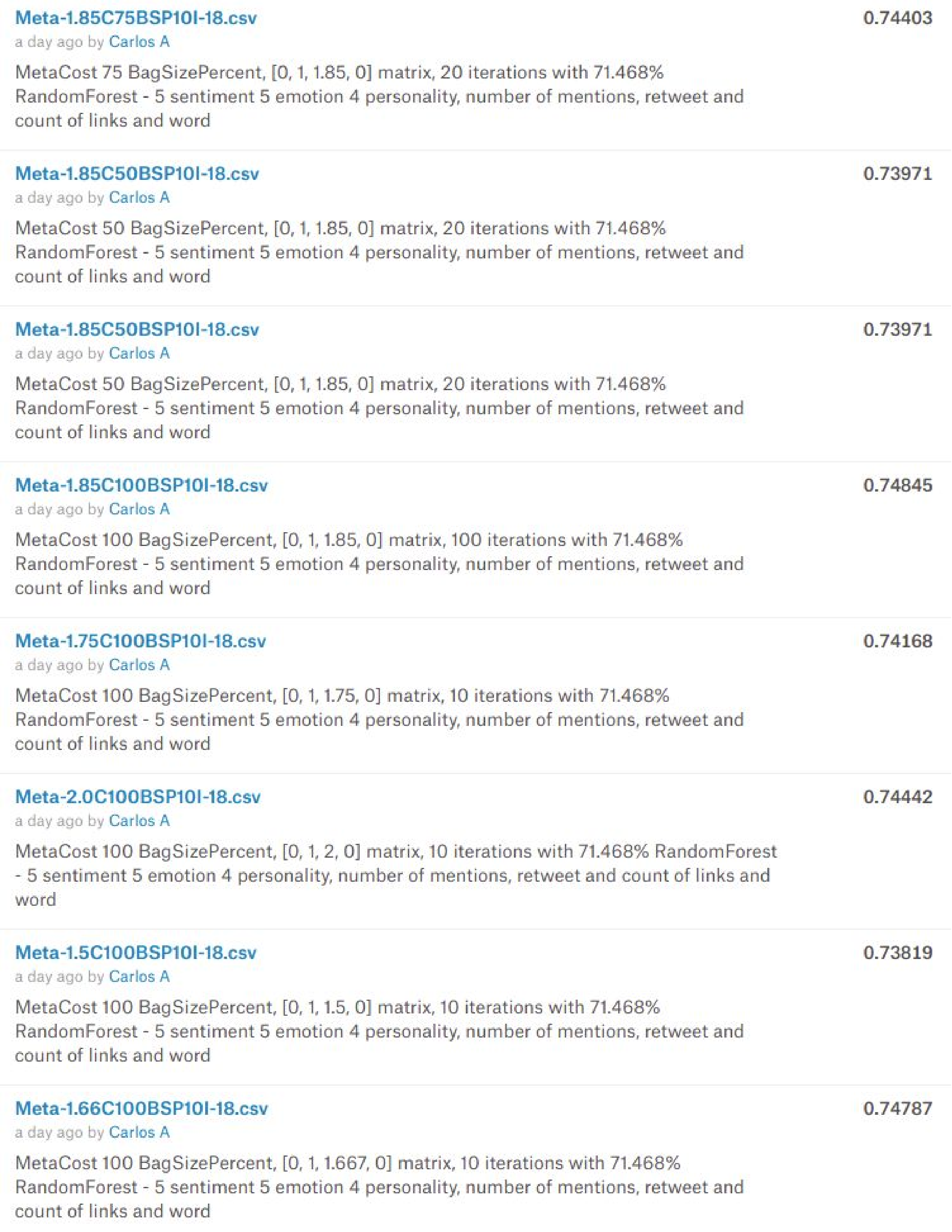
Smote100P5NeighwBRF-18F.csv
2 days ago by Carlos A
Smote 100%, 5 neighbours with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word
0.74677 □

Al igual que con Spread Subsample, se obtuvieron mejores resultados al generar más instancias de la clase minoritaria.

Carlos Augusto Amador Manilla
A01329447

## MetaCost

Hace a un clasificador sensible al costo, combina sensibilidad al costo con Bagging. Se comenzó castigando los falsos positivos por 1.667 aproximadamente el valor del desbalance y se continuó experimentando con otros costos.



**Meta-1.85C75BSP10I-18.csv** 0.74403
a day ago by Carlos A
MetaCost 75 BagSizePercent, [0, 1, 1.85, 0] matrix, 20 iterations with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word

**Meta-1.85C50BSP10I-18.csv** 0.73971
a day ago by Carlos A
MetaCost 50 BagSizePercent, [0, 1, 1.85, 0] matrix, 20 iterations with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word

**Meta-1.85C50BSP10I-18.csv** 0.73971
a day ago by Carlos A
MetaCost 50 BagSizePercent, [0, 1, 1.85, 0] matrix, 20 iterations with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word

**Meta-1.85C100BSP10I-18.csv** 0.74845
a day ago by Carlos A
MetaCost 100 BagSizePercent, [0, 1, 1.85, 0] matrix, 100 iterations with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word

**Meta-1.75C100BSP10I-18.csv** 0.74168
a day ago by Carlos A
MetaCost 100 BagSizePercent, [0, 1, 1.75, 0] matrix, 10 iterations with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word

**Meta-2.0C100BSP10I-18.csv** 0.74442
a day ago by Carlos A
MetaCost 100 BagSizePercent, [0, 1, 2, 0] matrix, 10 iterations with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word

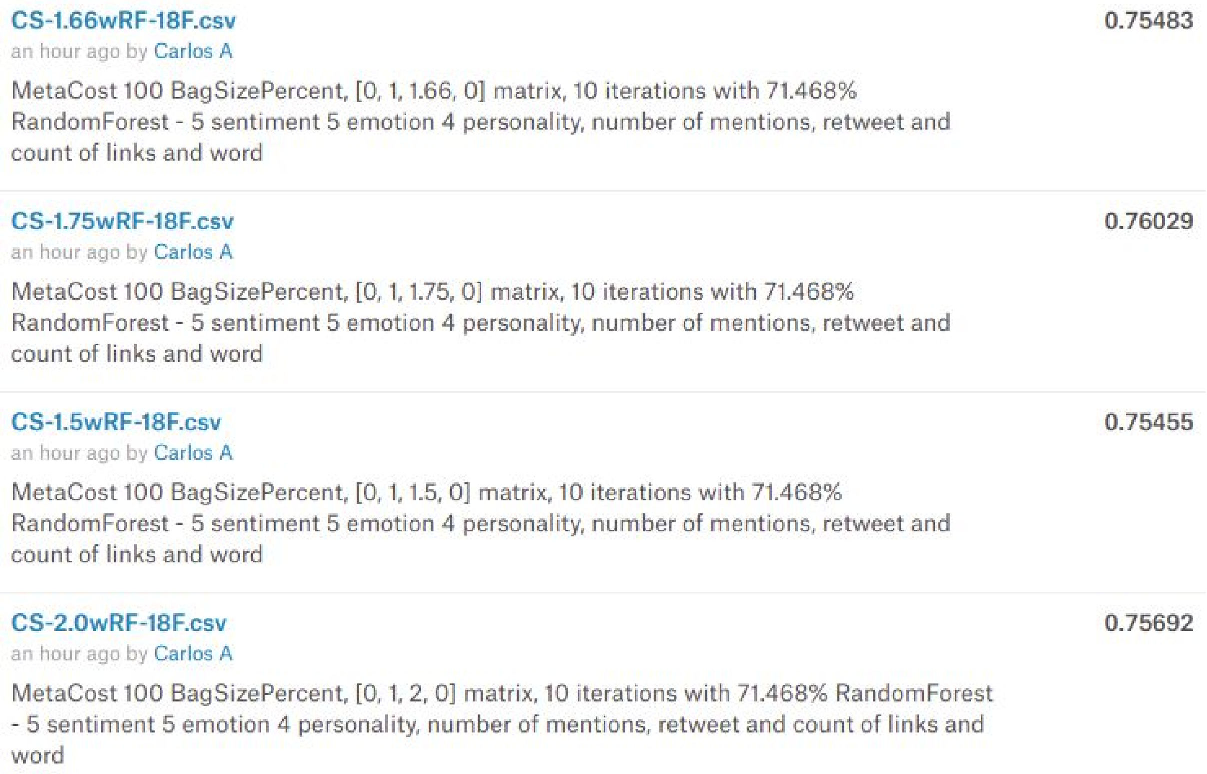
**Meta-1.5C100BSP10I-18.csv** 0.73819
a day ago by Carlos A
MetaCost 100 BagSizePercent, [0, 1, 1.5, 0] matrix, 10 iterations with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word

**Meta-1.66C100BSP10I-18.csv** 0.74787
a day ago by Carlos A
MetaCost 100 BagSizePercent, [0, 1, 1.667, 0] matrix, 10 iterations with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word

Carlos Augusto Amador Manilla
A01329447

# Cost sensitive classifier

Hace al clasificador base sensible al costo.



CS-1.66wRF-18F.csv 0.75483
an hour ago by Carlos A
MetaCost 100 BagSizePercent, [0, 1, 1.66, 0] matrix, 10 iterations with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word

CS-1.75wRF-18F.csv 0.76029
an hour ago by Carlos A
MetaCost 100 BagSizePercent, [0, 1, 1.75, 0] matrix, 10 iterations with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word

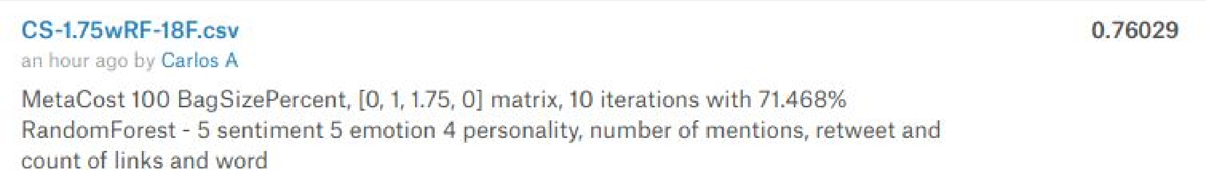
CS-1.5wRF-18F.csv 0.75455
an hour ago by Carlos A
MetaCost 100 BagSizePercent, [0, 1, 1.5, 0] matrix, 10 iterations with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word

CS-2.0wRF-18F.csv 0.75692
an hour ago by Carlos A
MetaCost 100 BagSizePercent, [0, 1, 2, 0] matrix, 10 iterations with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word

# Conclusiones

Atender el problema de desbalanceo produjo un incremento en la eficacia del clasificador, favorecer a la clase minoritaria -castigando falsos positivos y generando más instancias de la misma- sin duda ayudó a obtener mejores resultados.

El mayor puntaje lo obtuvo CostSensitiveClassifier:



CS-1.75wRF-18F.csv 0.76029
an hour ago by Carlos A
MetaCost 100 BagSizePercent, [0, 1, 1.75, 0] matrix, 10 iterations with 71.468% RandomForest - 5 sentiment 5 emotion 4 personality, number of mentions, retweet and count of links and word