**UNIVERSIDAD CENTRAL DEL ECUADOR**

**FACULTAD DE INGENIERÍA, CIENCIAS FÍSICAS Y MATEMÁTICA**

**INGENIERÍA INFORMÁTICA**

**ANÁLISIS DE DATOS**



**PROYECTO ANALISIS DE DATOS**

**RADIOS DEL ECUADOR**

**INTEGRANTES DEL GRUPO:**

Alexis Bautista

Bryan Catucuamba

David Jácome

Alejandro Naranjo

Richard Quimbiulco

Contenido

[TEMA 3](#_Toc488822490)

[OBJETIVOS 3](#_Toc488822491)

[Objetivo general 3](#_Toc488822492)

[Objetivos específicos 3](#_Toc488822493)

[DATOS 3](#_Toc488822494)

[Criterios de selección 3](#_Toc488822495)

[Extracción 4](#_Toc488822496)

[Depuración 4](#_Toc488822497)

[Resultados 5](#_Toc488822498)

[BASE TEÓRICA 5](#_Toc488822499)

[Máquina de Soporte Vectorial para clasificación 6](#_Toc488822500)

[Polaridad Análisis sentimental 7](#_Toc488822501)

[Índices de relevancia 7](#_Toc488822502)

[Factor de ponderación 7](#_Toc488822503)

[MÉTRICAS DE POLARIDAD 8](#_Toc488822504)

[Aprendizaje supervisado 8](#_Toc488822505)

[APRENDIZAJE NO SUPERVISADO 8](#_Toc488822506)

[MÉTRICAS DE RELEVANCIA 8](#_Toc488822507)

[FACTORES DE PONDERACIÓN 9](#_Toc488822508)

[ANALISIS 12](#_Toc488822509)

[CONCLUSIONES Y RECOMENDACIONES 14](#_Toc488822510)

[Referencias: 14](#_Toc488822511)

[ANEXOS 15](#_Toc488822512)

[SCRIPTS UTILIZADOS: 15](#_Toc488822513)

[ SCRIPT PARA DESCARGA DE TWEETS 15](#_Toc488822514)

[ SCRIPT PARA LA UNION DE LOS ARCHIVOS DE LAS DIFERENTES EMISORAS SELECCIONADAS 15](#_Toc488822515)

[ SCRIPT PARA DEPURACION DE TWEETS 15](#_Toc488822516)

[ SCRIPT PARA HALLAR LA LOCALIDAD DEL TWEET 15](#_Toc488822517)

[ SCRIPT PARA EL CALCULO DEL FACTOR DE RELEVANCIA 15](#_Toc488822518)

[ SCRIPT PARA APLICACIÓN DE LA TECNICA MVS 15](#_Toc488822519)

# TEMA

El análisis de audiencias se define como el conjunto de métodos de investigación utilizados para obtener información sobre la audiencia, y que permite estimar el número de personas que han estado expuestos a un evento medial en un momento del tiempo determinado. Este puede referirse por ejemplo al consumo de un producto audiovisual, ya sea textual, sonoro, fotográfico o multimedia.

# OBJETIVOS

## Objetivo general

Identificar cual o cuales son los artistas musicales que la gente más solicita a través de Twitter para realizar un análisis de los datos obtenidos de las principales cadenas radiales musicales y concluir cual es la demanda de los usuarios.

## Objetivos específicos

* Determinar las fuentes de los datos y depurar los datos para su análisis.
* Identificar la mejor métrica de polaridad para los datos, ya sea usando diccionario de palabras o máquinas de soporte vectoriales.
* Identificar cual es el mejor kernel para máquinas de soporte vectorial que se ajusta a la clasificación de textos.
* Establecer un índice de relevancia que mejor nos permita valorar la importancia de los artistas obtenidos.
* Estimar un factor de ponderación a cada provincia del Ecuador según los datos recogidos y los datos censales.

# DATOS

## Criterios de selección

Se han descargado los datos desde el 10 de Junio hasta el Lunes 24 de Julio de las diferentes emisoras de radio elegidas de acuerdo al criterio de: mayor cantidad de seguidores.

Elegimos las radios musicales con más seguidores en el Ecuador:

|  |  |
| --- | --- |
| EMISORA | NÚMERO DE SEGUIDORES |
| La Bruja | 86700 |
| Canela | 59900 |
| Sucre | 54100 |
| Exa | 39000 |
| Alfa | 26700 |
| WQ Radio | 17800 |
| Metro | 11800 |
| La Otra | 11600 |

## Extracción

Para la descarga de tweets usamos el siguiente código:

#usamos las libreria necesaria

library(twitteR)

#credenciales de acceso para descarga de tweets

consumer\_key <-"RFHeP55qj0ejWv7YiSceWTphX"

consumer\_secret<-"YOOUNzPvlJCAIHo23AWGU6hRm9VYxP6AY60H0n3u3dBM44aHZS"

access\_token<-"366852754-QE8L7hZs1J0WbtDTImo4P3qxzabRZ6OEY0E9ckOs"

access\_secret<-"zUZ3rnyPHCLQxso43A8PvXY1a3jWp81yS8Z2Rr0w96YzB"

#iniciamos las conexión para la descarga de tweets

setup\_twitter\_oauth(consumer\_key, consumer\_secret, access\_token, access\_secret)

#empezamos la descarga de tweets

listaTweets <- searchTwitter("metrostereo", n=1000)

#guardamos los tweets obtenidos en un archivo .csv

tweets <- twListToDF(listaTweets)

write.csv(tweets, 'tweets-metro-1.csv')

Para el posterior análisis necesitamos el número de seguidores, la localidad del tweet, el número de veces que ha recibido un "me gusta" un tweet, obtenemos esta información con el siguiente código:

#Aumentamos las columnas necesarias en la base de datos para el posterior #análisis:

baseMetro$statusCount <- c(rep(NA, NROW(baseMetro)))

baseMetro$followers <- c(rep(NA, NROW(baseMetro)))

baseMetro$favorites <- c(rep(NA, NROW(baseMetro)))

baseMetro$friends <- c(rep(NA, NROW(baseMetro)))

baseMetro$location <- c(rep(NA, NROW(baseMetro)))

for (i in 1:NROW(base)) {

usuario <- as.character(base[i, 12]) tryCatch({

informacionUsuario <- getUser(usuario) #obtenemos el usuario que envio el tweet

base[i, 20] <- statusesCount(informacionUsuario)

base[i, 21] <- followersCount(informacionUsuario) #seguidores

base[i, 22] <- favoritesCount(informacionUsuario)

base[i, 23] <- friendsCount(informacionUsuario)

base[i, 24] <- location(informacionUsuario) #localidad

})

}

## Depuración

Para la depuración de datos hemos utilizado el siguiente código, aplicado a cada uno de los archivos de datos descargados de las diferentes emisoras seleccionadas:

Debemos tener en cuenta que la "codificación" que usamos para ver los tweets en este caso es UTF-8.

#DEPURACION DE DATOS

#RADIO METRO

#depuracion de los textos de los tweets

for (i in 1:NROW(baseMetro)) {

texto <- baseMetro[i, 2]

#eliminar simbolos de retweets

sinRT <- gsub("(RT|via)((?:\\b\\W\*@\\w+)+)", "", texto)

#eliminar simbolo @ y #

sinCuentas <- gsub("@|#", "", sinRT)

#eliminar signos de puntuacion

sinSimbolos <- gsub("[[:punct:]].", "", sinCuentas)

#eliminar numeros

sinNumeros <- gsub("[[:digit:]]", "", sinSimbolos)

#eliminar enlaces

sinEnlaces <- gsub("http\\w+", "", sinNumeros)

baseMetro$depurado[i] <- sinEnlaces

baseMetro$sinEspacios[i] <- gsub("[[:space:]]|", "", baseMetro$depurado[i])

}

Una vez depurados los datos ordenamos la base de datos de acuerdo al texto depurado sin espacios:

#ordenamos al dataframe por el texto depurado sin espacios

baseMetro <- baseMetro[order(baseMetro$sinEspacios), ]

Eliminamos los tweets que están vacíos:

#eliminamos tweets que quedaron vacios

baseMetro <- baseMetro[!(baseMetro$sinEspacios == ""), ]

Eliminamos los tweets que están repetidos:

#eliminamos repetidos

baseMetro <- baseMetro[!duplicated(baseMetro$sinEspacios), ]

Obtenemos la siguiente tabla de resultados:

|  |  |  |
| --- | --- | --- |
| Emisora | Cantidad de tweets  Antes de la depuración | Cantidad de tweets  Después de la depuración |
| La Bruja | 3446 | 1511 |
| Canela | 13300 | 1722 |
| Sucre | 2050 | 700 |
| Exa | 1383 | 514 |
| Alfa | 1328 | 700 |
| WQ Radio | 12692 | 1240 |
| Metro | 4707 | 1106 |
| La Otra | 6679 | 572 |
| TOTAL | 45585 | 8065 |

**Base de datos final depurada:**

Número de observaciones: **8065**

# BASE TEÓRICA

**Aprendizaje no supervisado**

Todo el proceso de modelado se lleva a cabo sobre un conjunto de ejemplos formado tan sólo por entradas al sistema. No se tiene información sobre las categorías de esos ejemplos. Por lo tanto, en este caso, el sistema tiene que ser capaz de reconocer patrones para poder etiquetar las nuevas entradas.

Dentro del campo de investigación de la minería de datos, haciendo uso de la lingüística computacional, el procesamiento del lenguaje natural (PLN) y el análisis de textos, se halla una disciplina llamada análisis de sentimientos (*sentiment analysis),* también citada en la bibliografía como minería de sentimientos (*opinion mining)* o análisis de subjetividad (*subjetiviy analysis).*

El análisis sentimental, según Jorge Carrillo de Albornoz (2011), “hace referencia a la tarea de análisis, identificación y clasificación de todo tipo de contenido emocional, subjetivo u opinado”.

En términos generales, el análisis sentimental tiene como objetivo determinar la actitud de un orador o un escritor con respecto a algún tema, o la polaridad global contextual de un documento.

**Aprendizaje supervisado**

**Máquinas de Soporte Vectorial**

**Definición:**

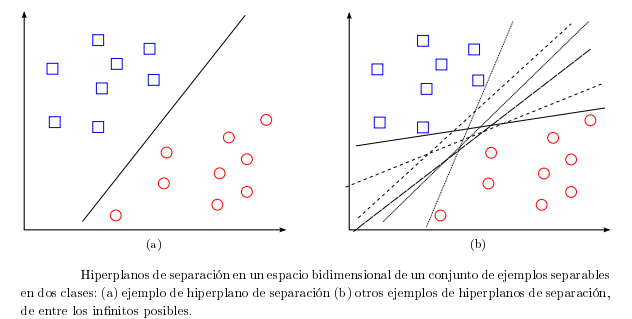
Dado un conjunto de puntos, subconjunto de un conjunto mayor (espacio), en el que cada uno de ellos pertenece a una de dos posibles categorías, un algoritmo basado en SVM construye un modelo capaz de predecir si un punto nuevo (cuya categoría desconocemos) pertenece a una categoría o a la otra.

Como en la mayoría de los métodos de clasificación supervisada, los datos de entrada (los puntos) son vistos como un vector *p*-dimensional (una lista ordenada de *p* números).

La SVM busca un hiperplano que separe de forma óptima a los puntos de una clase de la de otra, que eventualmente han podido ser previamente proyectados a un espacio de dimensionalidad superior.

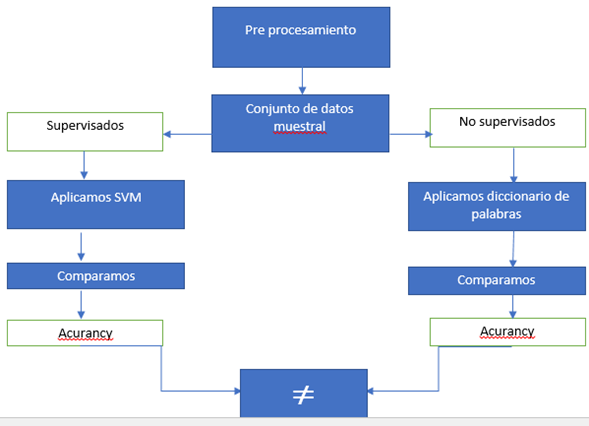
En ese concepto de "separación óptima" es donde reside la característica fundamental de las SVM: este tipo de algoritmos buscan el hiperplano que tenga la máxima distancia (margen) con los puntos que estén más cerca de él mismo. De esta forma, los puntos del vector que son etiquetados con una categoría estarán a un lado del hiperplano y los casos que se encuentren en la otra categoría estarán al otro lado. (Support-vector-machines.org, 2017)

Los algoritmos SVM pertenecen a la familia de los clasificadores lineales.



### Máquina de Soporte Vectorial para clasificación

Entre las aplicaciones más relevantes de las SVM se encuentra la Clasificación, el problema de la clasificación puede reducirse a examinar dos clases sin pérdida de generalidad. La tarea es encontrar un clasificador que funcione bien en datos futuros, es decir que generalice bien la clasificación.



## Índices de relevancia

Consiste en medir la importancia de una cuenta de twitter o el contenido de un tuit. En muchas ocasiones, para medir la importancia de una cuenta de Twitter, tendemos a fijarnos en una sola métrica: Los seguidores, sin embargo existen algunas otras de donde se puede obtener información valiosa que se puede interpretar y monitorizar.

## Factor de ponderación

Consiste en convertir los resultados de los valores caracterizados a una unidad común y sumable, multiplicándolos por su factor de ponderación. Posteriormente se suman todos ellos para obtener una puntuación única del impacto que tienen.

# RESULTADOS

## Aprendizaje supervisado

Para el análisis supervisado el algoritmo usado es máquinas de soporte para lo cual del total de 8066 se extrajeron 367 datos del total de los datos para ser de entrenamiento, esta cantidad de datos de entrenamiento lo determinamos usando la fórmula para calcular el tamaño de la muestra:



Luego de obtener aleatoriamente los 367 datos de la muestra se les asigna un valor de polaridad según el contenido del tuit.

A partir de los datos de entrenamiento se le aplica el algoritmo de máquinas de soporte al resto de datos.

Antes de aplicar el modelo de máquinas de soporte vectorial, debemos verificar cuál es el kernel de SVM que mejor se aplica a la clasificación de textos, para ello hemos encontrado el modelo SVM usando los datos de entrenamiento y los diferentes kernels que existen. Cabe recalcar que los datos de prueba para verificar cual modelo se ajustan más son los datos de entrenamiento, pero su clasificación con SVM se lo puso en una nueva columna. Una vez clasificados los textos, encontramos las siguientes tablas de confusión con los diferentes núcleos.

KERNEL LINEAL

|  |  |  |  |
| --- | --- | --- | --- |
| **Resultados** | | **Predichos** | |
| **0** | **1** |
| **Reales** | **0** | 293 | 0 |
| **1** | 0 | 74 |

Una vez clasificados los tweets con SVM y el núcleo lineal, el modelo nos dice que de los 293 tweets negativos son todos negativos y de los 74 tweets positivos son todos positivos, por lo que obtuvimos que el porcentaje de aciertos del modelo es del 100%.

KERNEL RADIAL

|  |  |  |  |
| --- | --- | --- | --- |
| **Resultados** | | **Predichos** | |
| **0** | **1** |
| **Reales** | **0** | 293 | 8 |
| **1** | 0 | 66 |

Una vez clasificados los tweets con SVM y el núcleo lineal, el modelo nos dice que de los 293 tweets negativos son todos negativos y de los 74 tweets positivos son todos positivos, por lo que obtuvimos que el porcentaje de aciertos del modelo es del 97,82%.

KERNEL POLINOMIAL

|  |  |  |  |
| --- | --- | --- | --- |
| **Resultados** | | **Predichos** | |
| **0** | **1** |
| **Reales** | **0** | 293 | 74 |
| **1** | 0 | 0 |

Utilizando el kernel polinomial, obtuvimos que el porcentaje de aciertos del modelo es del 79,84%.

KERNEL SIGMOIDE

|  |  |  |  |
| --- | --- | --- | --- |
| **Resultados** | | **Predichos** | |
| **0** | **1** |
| **Reales** | **0** | 293 | 20 |
| **1** | 0 | 54 |

Utilizando el kernel sigmoide, obtuvimos que el porcentaje de aciertos del modelo es del 94,55%.

## APRENDIZAJE NO SUPERVISADO

Para el caso de aprendizaje no supervisado se ha usado un diccionario de palabras para petición los tuits que corresponden a una petición de una canción.

Cuando el tuit ya se encuentra depurado se contabiliza las palabras que coinciden con las del diccionario de palabras, una vez finalizado el conteo se compara entre el número de palabras positivas con un valor de 2 porque usualmente las peticiones se las realiza con un mínimo de 2 palabras, luego con la comparación se determina la polaridad respectiva.



## MÉTRICAS DE RELEVANCIA

**Ratio seguidores/seguidos**: Los usuarios que tienen un ratio de seguidores/seguidos cercano al 1 suelen contar con el efecto followback: muchas cuentas les siguen para ganar seguidores fácilmente. De esta forma, si cuidamos de seguir a los que nos siguen e ir borrando los que nos eliminan, podemos llegar a generar un crecimiento sostenido. Incluso si seguimos a varios miles de cuentas, y pasadas unas horas eliminamos a los que no nos hayan seguido, podemos ganar decenas de miles de seguidores en cuestión de semanas. Hay formas sencillas de automatizar el proceso. Por ello, cuanto más alta sea la cifra de seguidores/seguidos, mayor será la relevancia real de la cuenta.

#RELEVANCIA

#Ratio relevancia del usuario

baseDatos$relevancia1<-c(rep(0,NROW(baseDatos)))

for(i in 1:NROW(baseDatos)) {

baseDatos[i,25] <- baseDatos[i,21]/baseDatos[i,23]

}

Los 5 usuarios más relevantes que han hecho una petición de música son:

Reni Romero, Carlos Gurumendi, Aurelio Tamayo, Carlos Martino y Silvia Buendía; los cuales son de Guayas.

## FACTORES DE PONDERACIÓN

Para calcular los factores de ponderación realizamos los siguientes pasos:

1. Convertimos la columna localidad a mayúsculas
2. Verificamos a donde pertenece cada tweet mediante el uso de expresiones regulares, para ello usamos el comando grepl, asignándole a cada tweet un código de localidad.

Los códigos de las localidades son los siguientes:

|  |  |
| --- | --- |
| Código | Localidad |
| 1 | Azuay |
| 2 | Bolívar |
| 3 | Cañar |
| 4 | Carchi |
| 5 | Cotopaxi |
| 6 | Chimborazo |
| 7 | El Oro |
| 8 | Esmeraldas |
| 9 | Guayas |
| 10 | Imbabura |
| 11 | Loja |
| 12 | Los Ríos |
| 13 | Manabí |
| 14 | Morona Santiago |
| 15 | Napo |
| 16 | Pastaza |
| 17 | Pichincha |
| 18 | Tungurahua |
| 19 | Zamora Chinchipe |
| 20 | Galápagos |
| 21 | Sucumbíos |
| 22 | Orellana |
| 23 | Santo Domingo |
| 24 | Santa Elena |
| 25 | Otros |

1. Calculamos el número total de tuits por localidad
2. Ingresamos los datos del censo realizado en el 2010 para conocer el número de personas mayores a 11 años en el Ecuador a la fecha.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| codprov | provincias | numero | menores4 | menores9 | menores14 | mayores11 | porcentaje |
| 1 | Azuay | 712127 | 68737 | 72122 | 75507 | 643390 | 4,95200 |
| 2 | Bolívar | 183641 | 18722 | 21226 | 21560 | 164919 | 1,26934 |
| 3 | Cañar | 225184 | 22921 | 23984 | 26851 | 202263 | 1,55676 |
| 4 | Carchi | 164524 | 15362 | 16888 | 17584 | 149162 | 1,14806 |
| 5 | Cotopaxi | 409205 | 42491 | 47367 | 46489 | 366714 | 2,82250 |
| 6 | Chimborazo | 458581 | 45264 | 49074 | 50710 | 413317 | 3,18119 |
| 7 | El Oro | 600659 | 55780 | 61450 | 63128 | 544879 | 4,19379 |
| 8 | Esmeraldas | 534092 | 64498 | 67581 | 64963 | 469594 | 3,61434 |
| 9 | Guayas | 3645483 | 359678 | 362896 | 373511 | 3285805 | 25,28995 |
| 10 | Imbabura | 398244 | 38996 | 43646 | 44326 | 359248 | 2,76503 |
| 11 | Loja | 448966 | 44203 | 47751 | 49974 | 404763 | 3,11535 |
| 12 | Los Ríos | 778115 | 83631 | 86598 | 88134 | 694484 | 5,34525 |
| 13 | Manabí | 1369780 | 139481 | 150953 | 152584 | 1230299 | 9,46928 |
| 14 | Morona Santiago | 147940 | 22011 | 21198 | 19412 | 125929 | 0,96924 |
| 15 | Napo | 103697 | 13624 | 14028 | 13023 | 90073 | 0,69327 |
| 16 | Pastaza | 83933 | 10730 | 11013 | 10208 | 73203 | 0,56342 |
| 17 | Pichincha | 2576287 | 236893 | 244844 | 241334 | 2339394 | 18,00568 |
| 18 | Tungurahua | 504583 | 45198 | 48391 | 49194 | 459385 | 3,53576 |
| 19 | Zamora Chinchipe | 91376 | 11129 | 11976 | 11976 | 80247 | 0,61764 |
| 20 | Galápagos | 25124 | 2135 | 2472 | 2358 | 22989 | 0,17694 |
| 21 | Sucumbíos | 176472 | 21294 | 22292 | 20926 | 155178 | 1,19436 |
| 22 | Orellana | 136396 | 18635 | 18189 | 16546 | 117761 | 0,90637 |
| 23 | Santo Domingo | 368013 | 40407 | 42413 | 41439 | 327606 | 2,52149 |
| 24 | Santa Elena | 308693 | 36762 | 34252 | 33446 | 271931 | 2,09298 |
| 25 | Otros | 0 | 0 | 0 | 0 | 0 | 0,00000 |

1. Calculamos el porcentaje de tuits por localidad.
2. Calculamos el porcentaje de tweets por localidad y el factor de ponderación.
3. Finalmente incluimos el factor de ponderación al dataframe de tweets.

(Código completo ver en Anexos)

**ALGORITMO SUPERVISADO:**

**APLICACION DE MAQUINA VECTORIAL:**

Para ver que análisis será más efectivo, dentro de nuestro análisis, ya sea aprendizaje supervisado como no supervisado, usaremos un conjunto de datos revisados previamente con su polaridad, en este caso usamos 367 de acuerdo a la fórmula que hemos expuesto anteriormente.

Partimos con estos datos del conjunto de la muestra que hemos obtenido

|  |  |
| --- | --- |
| No solicitados | Solicitados |
| 0 | 1 |
| 289 | 78 |

**Para SupportVectorMachine (SVM)**

Procedemos a realizar el análisis para poder ver la matriz de confusión:

1. Leeremos todas las palabras que están dentro del tweet, las usamos como vectores y los mandamos en un corpus.

Recorte de pantalla

1. Cambiamos a minúsculas y lo depuramos, con signos de puntuación, stopwords, links etc

Recorte de pantalla

Recorte de pantalla

1. Usamos Steamming para reducir las palabras a su simplicidad, ya que nos permite tratar muchas mas palabras de una manera mas homogénea, con el paquete de tm proveniente de R

Recorte de pantalla

1. Ahora lo que trataremos de hacer es sacar esas palabras, una por cada columna y asi poder observar cuales son las mejores

Recorte de pantalla

1. Le pegamos a esta matriz la polaridad antes resuelta en la variable ya escogida

Recorte de pantalla

Con estos pasos pues esta listo para el análisis de datos

Como necesitamos ver como se comportan nuestros datos dentro del modelo SVM (Support Vector Machine), Definimos un conjunto de entrenamiento o conjunto de evaluación, dividiendo de esta manera un 80% para entrenamiento, y 20% evaluación.

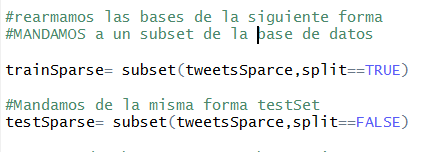
Cabe mencionar que con el entrenamiento vamos a enseñarle al modelo, vamos a enseñarle al algoritmo como clasificar, y con el 20% vamos a evaluar como lo hace el modelo, y con esto observar que poder de predicción tiene el modelo luego de haberlo creado.

Recorte de pantalla

Ahora procedemos a usar, una variable que me ayudara a ver que observaciones se van para el train(80%) y cuales para el test(20%), y diré que mi splitRatio sea 0.8 que quiere decir que el 80% de mis observaciones se vayan al conjunto de entrenamiento, asignando true o false según como convenga.

Recorte de pantalla

Rearmamos las bases tanto el **trainSparse** como el **testSparse**



Ya con estos datos procedemos a verificar con la matriz de confusión cual fue nuestro error, de predicción aplicamos el algoritmo SVM

Recorte de pantalla

Explicando que tomamos como un factor el sentimiento, y es representado por todas las columnas dentro del trainSparse con esto podemos observar que tenemos un total de 86% de exactitud

|  |  |
| --- | --- |
| No solicitado | Solicitado |
| 0 | 1 |
| 305 | 62 |

**Con esto podemos observar que tenemos 86% de exactitud.**

## ANALISIS

Comparando los resultados de un aprendizaje supervisado y un no supervisado se observa que:

ALGORITMO NO SUPERVISADO:

Utilizando un diccionario con 44 palabras obtuvimos los resultados:

**Predicción**

**Real**

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 248 | 41 |
| 1 | 41 | 37 |

El porcentaje de eficacia es: (248+41)\*100/(248+41+41+37) = **78.75%**

ALGORITMO SUPERVISADO:

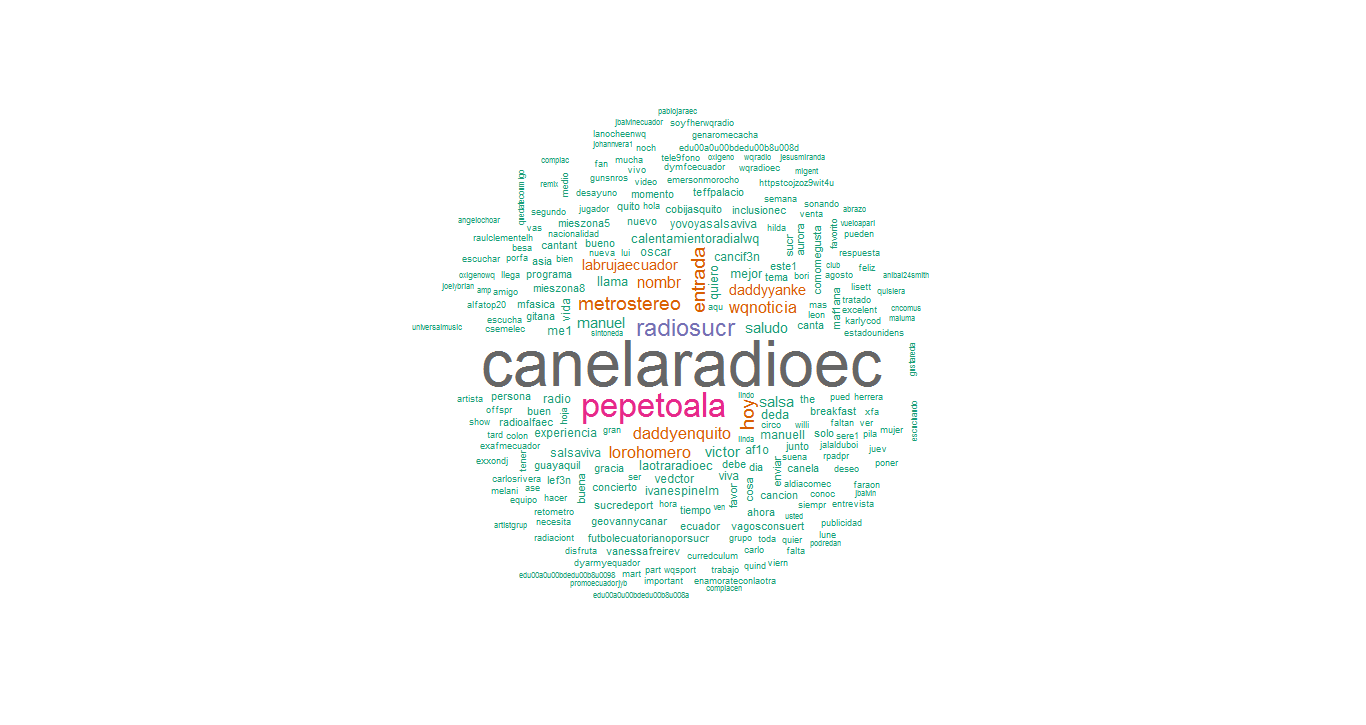
**Predicción**

**Real**

|  |  |  |
| --- | --- | --- |
|  | 0 | 1 |
| 0 | 273 | 16 |
| 1 | 16 | 32 |

El porcentaje de eficacia es: (273+32)\*100/367 = **83.10%**

De acuerdo al análisis final obtenemos la siguiente nube de palabras de las personas que realizaron alguna petición de música:



Nos damos cuenta que canelaradioec, pepetoala son las sobresalientes debido a que esta radio es la más popular según nuestro criterio de radios con mayor número de seguidores:

Ahora lo que nos interesa de esto es encontrar los artistas más nombrados en los tweets finales son:

|  |  |  |
| --- | --- | --- |
| Artista | Genero | Nacionalidad |
| Maluma | Regueton | Colombiano |
| Jbalvin | Regueton | Colombiano |
| Pablo Jara | Pop | Ecuatoriano |
| Johann Vera | Pop | Ecuatoriano |
| Arcangel | Regueton | Estadounidense |
| Jesus Miranda | Regueton | Venezolano |
| Carlos Rivera | BaladasPop | Ecuatoriano |
| Joel y Bryan | Pop | Argentino |

# CONCLUSIONES Y RECOMENDACIONES

La cantidad de tweets una vez depurada la base de datos se reduce considerablemente debido a la gran cantidad de tweets repetidos, retweets, etc. por lo cual sería ideal obtener una gran cantidad de datos iniciales.

El protagonismo que el Big Data está teniendo en la investigación de audiencias y en general sigue en aumento, la importancia que tienen los resultados proporcionan una gran ayuda para la toma inteligente de decisiones y las posibilidades de utilización de las redes sociales como herramientas para obtención de datos parar este tipo de análisis parece indicar que las grandes empresas dedicadas al análisis de audiencias no quieren perder la oportunidad de desarrollar sus aplicaciones metodológicas para el mercado en general.

La posibilidad de desarrollar este tipo de investigación a partir de los comentarios que las personas hacen sobre los espacios de televisión, radio, etc que consume, abre las puertas a la automatización de este tipo de investigación y crea nuevos campos de análisis para los estudios culturales de audiencias y los estudios de recepción.

En general se puede concluir que las audiencias tienen gran influencia sobre los medios ya que dependen de lo que la audiencia necesite, esto se logra saber mediante los sondeos y diversas técnicas como la aplicada en este proyecto para que puedan obtener la sintonía e información que necesitan de las personas.

# Referencias:

* RESENDIZ, J. A. Las Máquinas de Soporte Vectorial para identificación en Linea. Maestria, Control Automático. Instituto Politecnico Nacional, 2006.
* BOURDIEU, Pierre. Sobre la televisión. Editorial Anagrama. Barcelona, 1994.
* CALLEJO, Javier; Investigar las audiencias; editorial : Paidós, Barcelona , 2001.
* Dorian, P. (2003), " Business Modeling and Data Mining ", Morgan Kaufmann
* Support-vector-machines.org. (2017). *SVM - Support Vector Machines*. [online] Disponible en: http://www.support-vector-machines.org/
* Uam.es. (2017). *Regresión lineal simple con R*. [online] Disponible en: <https://www.uam.es/personal_pdi/ciencias/joser/paginaR/regresion.html>

# ANEXOS

## SCRIPTS UTILIZADOS:

Lista de Scripts usados para el proyecto (Adjuntos en el CD):

### SCRIPT PARA DESCARGA DE TWEETS

### SCRIPT PARA LA UNION DE LOS ARCHIVOS DE LAS DIFERENTES EMISORAS SELECCIONADAS

### SCRIPT PARA DEPURACION DE TWEETS

### SCRIPT PARA HALLAR LA LOCALIDAD DEL TWEET

### SCRIPT PARA EL CALCULO DEL FACTOR DE RELEVANCIA

### SCRIPT PARA APLICACIÓN DE LA TECNICA MVS