Maestría en Explotación de Datos
y Descubrimiento del Conocimiento
Universidad de Buenos Aires

UNIVERSIDAD DE BUENOS AIRES

FACULTAD DE CIENCIAS EXACTAS Y NATURALES
FACULTAD DE INGENIERÍA

Aprendizaje Automático
Trabajo Práctico N°2

GRUPO N° 10

Amena, Santiago
Opazo, F. Ayelén
Rivero González, Leandro

EQUIPO DOCENTE

Cotik, Viviana
Henrión, Guillermo
Bujía, Gastón
Pepino, Leonardo

Modelo de reconocimiento de emociones a partir del habla

Amena, Santiago; Opazo, F. Ayelén; Rivero González, Leandro

Junio 2021

1. Resumen

El presente trabajo tuvo por objetivo predecir emociones a partir de audios de habla y canciones de actores de la base de audios RAVDESS. El entrenamiento de los modelos de predicción se realizó en base a 88 atributos de alto nivel extraídos de cada uno de los audios de la base indicada. El modelo con mejor performance elegido fue un ensamble *Random Forest* de 600 estimadores, profundidad 60 y sin *bootstrap*, con un *accuracy* de 0,64. Por último, se estudió la sensibilidad del modelo frente a la presencia de ruido gaussiano, obteniendo como resultado que frente a ruidos bajos la performance del modelo se mantiene.

2. Introducción

El trabajo se estructuró de la siguiente manera: en primer lugar se realizó una breve descripción de los tipos de variables que componen el dataset y de las metodologías utilizadas para su preprocesamiento. En segundo lugar, se realizó el entrenamiento de un modelo de *random forests*, y se analizaron sus resultados en conjuntos de validación con dos estrategias distintas: *12-fold cross-validation* y *leave-2-speakers out*. Luego, se hizo una busqueda de hiperparámetros para obtener el mejor modelo de *random forest*, *AdaBoost* y de un perceptrón multicapa y se compararon los resultados de los tres modelos con los mejores hiperparámetros. A partir de esta comparación se eligió el mejor modelo y se lo evaluó en el conjunto de prueba. Para esto se analizó la matriz de confusión y el *accuracy* obtenido. Por último, se evaluó la robustez del modelo frente a distintos niveles de ruido en los audios.

3. Datos

Para el desarrollo del modelo se trabajó con la base de datos audiovisual de habla y canción emocional de Ryerson (RAVDESS), con modalidad 03 (sólo audio). Se trata de una base de 12 actores y 12 actrices profesionales vocalizando dos declaraciones léxicamente emparejadas, con un acento norteamericano neutral. Esta modalidad está compuesta por 2 archivos. Por un lado, el archivo de voz, que contiene 60 ensayos por actor para 24 actores, totalizando 1440 observaciones; y, por el otro, el archivo de canción, que contiene 44 ensayos por actor x 23 actores, con 1012 instancias. El dataset utilizado no posee valores duplicados ni faltantes y está integrado por 2452 instancias de 8 variables categóricas, cuyos valores se detallan a continuación:
file_path (una distintiva por cada instancia) *Modality* (01 = full-AV, 02 = video-only, 03 = audio-only). Como se comentó más arriba, se utiliza la 03. *Vocal channel* (01 = speech, 02 = song). *Emotion* (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised). *Emotional intensity* (01 = normal, 02 = strong) [1]. *Statement* (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door"). *Repetition* (01 = 1st repetition, 02 = 2nd repetition). *Actor* (01 a 24) [2]. En la tabla 1 se observa la cantidad de datos por estado emocional y separado por si es hablado o cantado.

[1] No hay *strong intensity* para *neutral emotion*
[2] Los números impares corresponden a *hombres* y los pares a *mujeres*

4. Metodología

4.1. Extracción de atributos

Para la predicción de emocionalidades de los audios se realizó un preprocesamiento de los mismos utilizando los atributos eGeMAPS de la librería opensmile-python, obteniendo finalmente de cada audio 88 características numéricas. Algunas de estas características son: *pitch*, *jitter* y *loudness*, entre otras[3].

4.2. Entrenamiento y validación

Para el entrenamiento del modelo de *random forest* previamente se dividieron los datos de dos formas diferentes utilizando *cross-validation*. Por un lado, se realizó un *12-fold cross validation*, armando los folds de forma aleatoria; mientras que el otro método consistió en la utilización de *leave-2-speakers out* con 12 folds de 2 actores distintos, siendo uno varón y otra mujer en cada uno de los conjuntos. La comparación de los resultados se muestran en la figura 1.

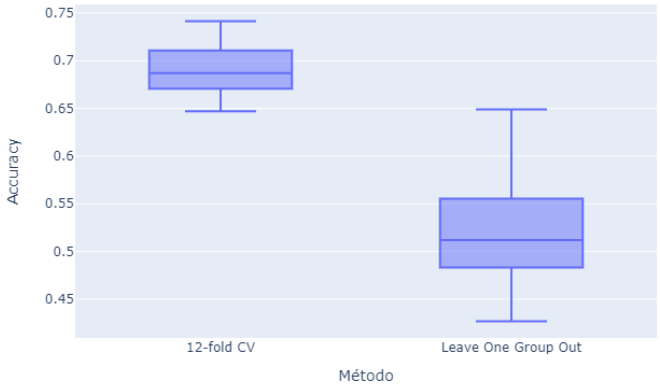


Figura 1: *Boxplot* de resultados de validación para la métrica de *accuracy*

El hecho de que, en general, los resultados del método convencional de *12 fold cross validation* tengan una mejor performance se puede asociar a dos causas. La primera, y más importante, es la correlación del método. Esta correlación se encuentra entre los datos de entrenamiento y los de validación, ya que es muy probable que varios de los actores que están en audios de entrenamiento también esten presentes en los audios de validación, lo cual resta independencia entre ambos *sets*. Lo segundo, que afecta negativamente al método de división de *fold* por actores, es que los conjuntos no están bien estratificados.

4.3. Elección del modelo

Para comenzar a trabajar en la elección del modelo en primer lugar se dividieron los datos en conjuntos de desarrollo y prueba. Considerando lo explicado en la sección anterior, se decidió dividir los datos de acuerdo a los actores, dejando para la parte de testeo los actores 1 y 2. Se eligieron estos dos actores ya que son de distinto género y ambos poseen audios cantados y hablados. Se estudiaron tres tipos de modelos (dos de ensamble y una red neuronal), y en cada caso se utilizó *12fold-validation*, debido a que dicho método permite estratificar (*stratify*) los conjuntos según las emociones.

En primer lugar se entrenó un modelo de *random forests* y se realizó un *random search* de hiperparámetros. En este caso se decidió variar la cantidad de árboles por ensamble (200 a 2000), la cantidad de variables a considerar (sqrt(*número_variables*), log2(*número_variables*)) en cada split, la profundidad máxima (10 a

[3]Adicionalmente se realizaron predicciones utilizando el conjunto de atributos ComParE 2016. Los resultados se muestran en el Anexo II.

110) y si se utiliza el método *boostrap* o no. El modelo con mejor resultado obtuvo una *accuracy* promedio de 0,74 y se entrenó con los siguientes hiperparámetros: 600 árboles, $sqrt(número\_variables)$ de variables a considerar por división, profundidad máxima de 60 nodos y sin *bootstrap*.

En segundo lugar, se estudió un algoritmo de *AdaBoost* con un árbol base de 10 nodos de profundidad máxima. Se realizó una búsqueda de los hiperparámetros *learning rate* y cantidad de árboles a considerar. El modelo final está compuesto por 1500 árboles y un *learning rate* de 1.

Por último se consideró un perceptrón multicapa. Para esto previamente se estandarizaron los datos de desarrollo, ya que los algoritmos que optimizan por gradiente suelen ser sensibles a la escala de las variables. Se exploraron diferentes hiperparámetros: cantidad de capas ocultas y de neuronas por capa, el optimizador (*sgd*, *adam*), y el parámetro de regularización *alpha* (0,0001 a 0,05). Finalmente se eligió una red de una capa con 100 neuronas, con un *alpha* de 0.05 y optimizador *adam*.

La métrica con la cual se evaluaron las performances de los modelos fue la exactitud del porcentaje de casos que cada uno acertó (*accuracy*). Se optó por esta métrica debido a que las clases se encontraban relativamente balanceadas; de lo contrario sería conveniente evaluar con *precision* para obtener con mayor validez la calidad del modelo. En la figura 5 se observan los resultados de *accuracy* para *12fold-validation* para los tres modelos. En base a estos resultados se decidió continuar el análisis con *random forests*.

5. Resultados

Una vez elegido el modelo se decidió evaluar los resultados en el conjunto de prueba, según las siguientes categorías: total, tipo de audio y género. En la figura 6 se observa la matriz de confusión para el total de datos de *test*. El *accuracy* total fue de 0,64. Este resultado es de esperarse ya que por lo general el *accuracy* en el conjunto de prueba suele ser menor que en validación. Adicionalmente, en este caso hay menor independencia entre los conjuntos de entrenamiento y validación, dado que hay actores que se repiten en ambos. A continuación se realizó la evaluación separando los datos de *test* por tipo de audio (*speech* y *song*). En las figuras 2 y 3 se observan las matrices de confusión para los audios hablados y cantados respectivamente.

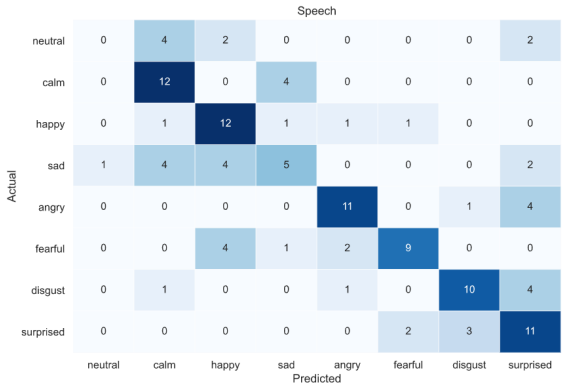


Figura 2: Matriz de confusión - Modo hablado.

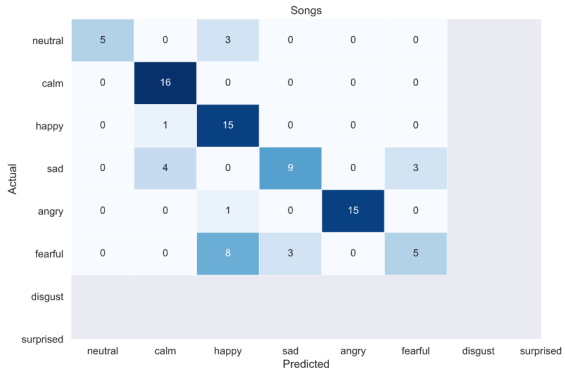


Figura 3: Matriz de confusión - Modo cantado.

En estas figuras se destaca, en primer lugar, la mejor *performance* de los audios en formato cantado (con un *accuracy* de 0,73 comparado con 0,57 de audios hablados). En cuanto a la matriz de confusión de audios hablados, se observan a su vez dos cosas: nunca acierta la emoción neutral, a diferencia del caso cantado, donde acierta el 62 % de las veces. Por otro lado, predice correctamente la emoción triste en 5 casos y como contento en otros 4. Esto implica que el algoritmo no logra distinguir bien entre tristeza y alegría.

En el caso de los audios cantados, se observa que la emoción que peor predice es temor, la cual se confunde en varios casos con alegría y en menor medida con tristeza. También es notable el hecho de que no confunde nunca ninguna de las 6 emociones con *disgust* ni con *surprised*, lo que permite suponer que, como nunca entrenó con estas dos emociones en formato cantado, el modelo aprende a través de las 88 características si un audio es cantado, y a partir de aquí le asigna una probabilidad muy baja, o nula, a las emociones *disgust* o *surprised* para *song*.

Finalmente, se observó si había diferencias entre las predicciones por género. En este caso se obtuvo que la certeza del modelo para la voz masculina fue del 0,57 y la de voz femenina del ,71. La matriz de confusión separado por género se observa en las figuras 7 y 8.

5.1. Robustez del modelo frente a ruido gaussiano

Se exploró la robustez frente a ruidos externos, simulados a través de ruidos gaussianos. Para esto se consideró el mismo modelo entrenado en la subsección anterior, pero en este caso se agregaron a los audios diferentes niveles de ruido en la intensidad de la señal. Estos niveles se determinaron mediante diferentes valores de desvío estándar de la distribución gaussiana. Dado que la amplitud máxima de la intensidad de los audios es aproximadamente 2, se varió el desvío entre 0,1 y 2. En la figura 9 se observa gráficamente un mismo audio con 4 valores de ruido diferentes.

Los resultados del modelo en el conjunto de prueba para diferentes niveles de ruido se muestran en la figura 4. En la misma se graficaron los niveles de *accuracy* para distintos valores de desvío estándar. Se puede observar que la performance comienza a disminuir en forma marcada para desvíos estándar mayores a 0,4.

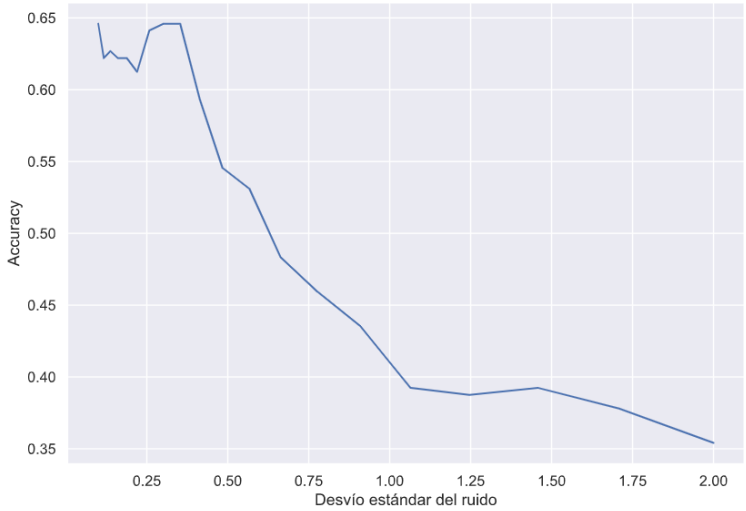


Figura 4: *Accuracy* para diferentes niveles de ruido gaussiano.

En otras palabras, se aprecia como los ruidos bajos no afectan, o afectan ligeramente, la certeza del modelo. En cambio para ruidos con desvíos estándar mayores a 0,4, la confianza del modelo disminuye abruptamente hasta niveles de ruido de 1 y luego sigue disminuyendo pero a una velocidad menor.

6. Conclusiones

El presente trabajo tuvo como objetivo predecir las emociones de diferentes actores en diferentes formatos de audio (cantado y hablado). A lo largo del estudio se entrenaron diferentes algoritmos con distintos hiperparámetros, y se obtuvo el mejor modelo con el método de ensambles de *random forest*, usando 600 estimadores con profundidad máxima de 60 y sin *bootstrap*.

Si bien los resultados del *accuracy* no son particularmente altos, al tener en cuenta que las categorías a predecir son 8, se aprecia la *performance* del modelo. Adicionalmente se evaluó la robustez del modelo ante la presencia de ruido gaussiano, obteniendo valores aceptables para ruidos bajos y no tanto para ruidos más altos. Una solución podría ser considerar algún filtro pasa banda en el preprocesamiento de los datos, para poder filtrar los audios en función de la frecuencia. De esta forma, se lograría utilizar audios más "limpiosçomo *inputs* para el modelo.

7. Referencias

– Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.

8. Anexo I

Emotion	*Speech*	*Song*
Neutral	96	92
Calm	192	184
Happy	192	184
Sad	192	184
Angry	192	184
Fearful	192	184
Disgust	192	0
Surprised	192	0

Tabla 1: Dataset de variables dividido por emoción y por si es cantado o hablado.

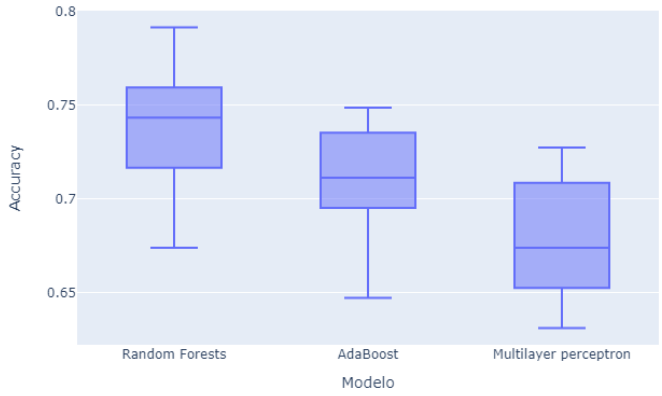


Figura 5: *Boxplot* de resultados de validación para la métrica de *accuracy* usando el método de *12 fold cross-validation* para distintos modelos.

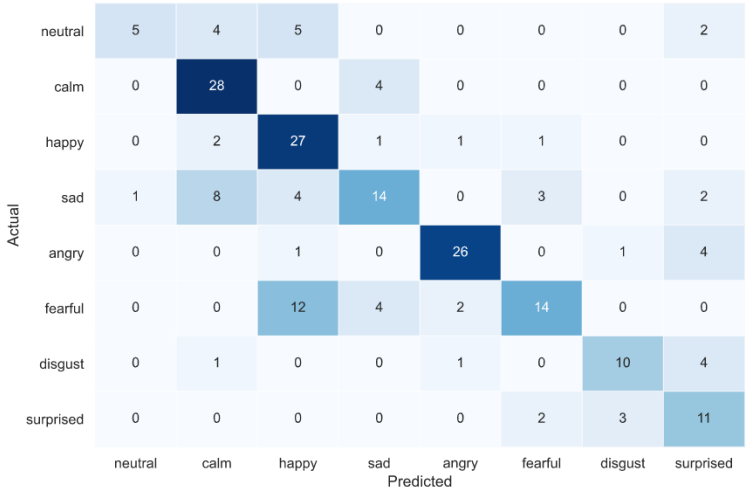


Figura 6: Matriz de confusión para los resultados de testeo.

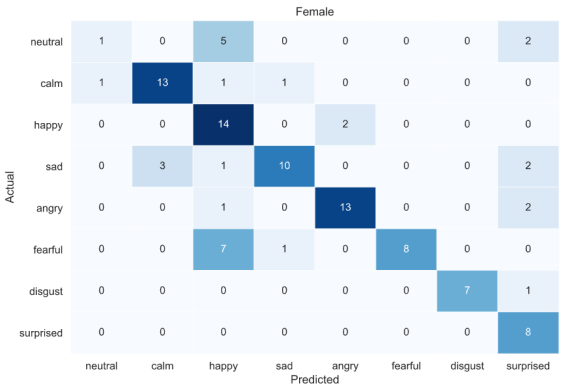


Figura 7: Matriz de confusión - Voz Femenina.

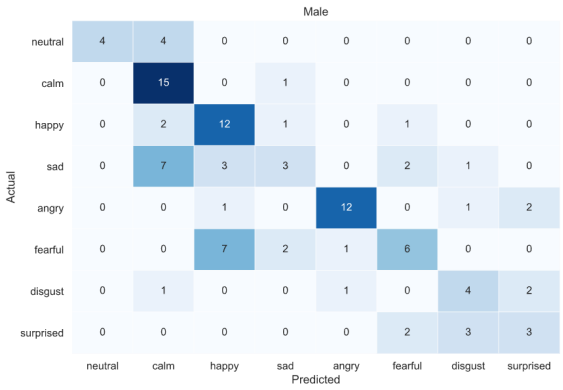
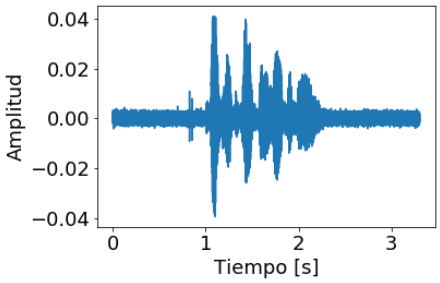
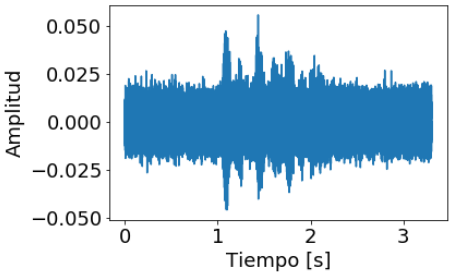


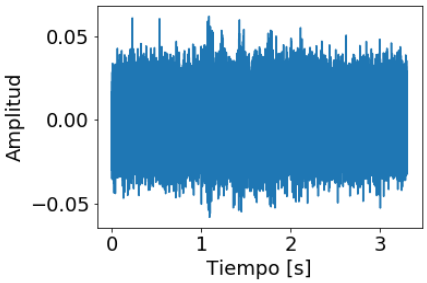
Figura 8: Matriz de confusión - Voz masculina.



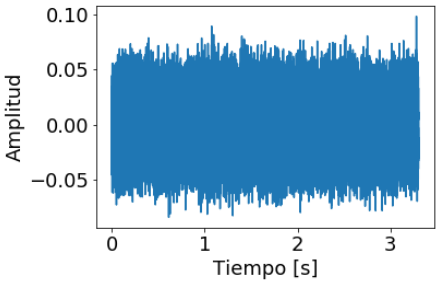
(a) Ruido mínimo incluido (σ =0.1).



(b) Ruido bajo (σ =0.6).



(c) Ruido alto (σ =1.2).



(d) Ruido máximo incluido (σ =2).

Figura 9: Gráfico de audios según nivel de ruido agregado

9. Anexo II

En la presente sección se muestran los resultados de los modelos de ensamble (*random forests* y *AdaBoost*) entrenados con el conjunto de atributos ComParE 2016. Utilizando *cross-validation* de 12 *folds*, las performances de los modelos se muestran en la figura 10.

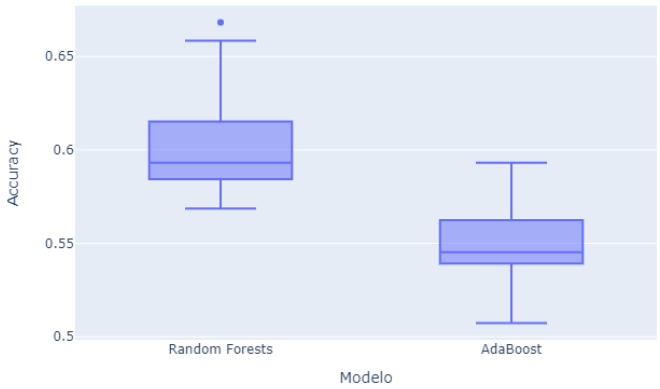


Figura 10: Performance de modelos entrenados con atributos ComParE 2016, medida mediante *accuracy*.

En el caso de random forests, la performance en los conjuntos de validación es algo menor que para los atributos eGeMAPS, con un promedio de *accuracy* de 0,6, mientras que en AdaBoost ese promedio es 0,55.