Escuela Superior Politécnica del Litoral

FACULTAD DE INGENIERÍA EN ELECTRICIDAD Y COMPUTACIÓN

INTELIGENCIA ARTIFICIAL

Propuesta de Proyecto:

Análisis de sentimientos por voz de los estudiantes utilizando Machine Learning

Grupo 2

George Henríquez Ronquillo

Edgar Vinueza Herrera

Jorge García García

Contenido

Problema

En la actualidad la pandemia provocada por el covid-19 ha afectado a muchos sectores de la sociedad. La educación es una de las más afectadas, debido a que diferentes procesos han tenido que ser modificados para adaptarse a la dinámica de la virtualidad, entre ellos las clases, la metodología utilizada en las mismas, los trabajos grupales, etc.

Este cambio ha afectado tanto a los estudiantes como los profesores, pero, en esta propuesta nos enfocaremos en los estudiantes y como la educación virtual puede afectar de forma negativa a su estado emocional, lo cual se refleja en un bajo nivel de atención y participación durante las clases, y esto debido a que los profesores mantienen una misma metodología de enseñanza entre clases, lo que puede resultar en emociones negativas para el estudiante al sentir estrés o aburrimiento. Por ello, es importante para el profesor conocer el estado de ánimo de sus estudiantes, lo cual puede ser complicado debido a que mucho de ellos mantienen su cámara apagada, sin embargo, esto es posible si se analiza la voz del estudiante durante una conversación, de esta manera permitirá al profesor cambiar la dinámica de la clase y a los estudiantes, maximizar su rendimiento [1].

Ilustración 1 Sentimientos experimentados por los estudiantes (Idem,

Objetivo

Desarrollar y entrenar un modelo de audio de aprendizaje automático (ML) aplicando el método de Random Forest para predecir los sentimientos de un audio en específico.

Modelos de análisis y definición del problema

Para el análisis y la de definición del problema nos encontramos en el machine learning, este proyecto está basado en la clasificación de audio y cuyo modelo puede ser representado en el siguiente gráfico.

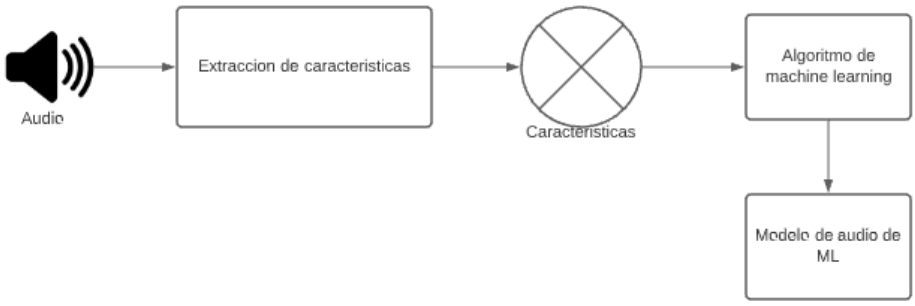


Ilustración 2. Modelo de aprendizaje Supervisado.

Para la definición del problema, podemos representar actores y acciones realizadas para la recolección de datos competente, es decir de los audios para el análisis de sentimiento. Es por esto se define el siguiente UML de casos de uso.

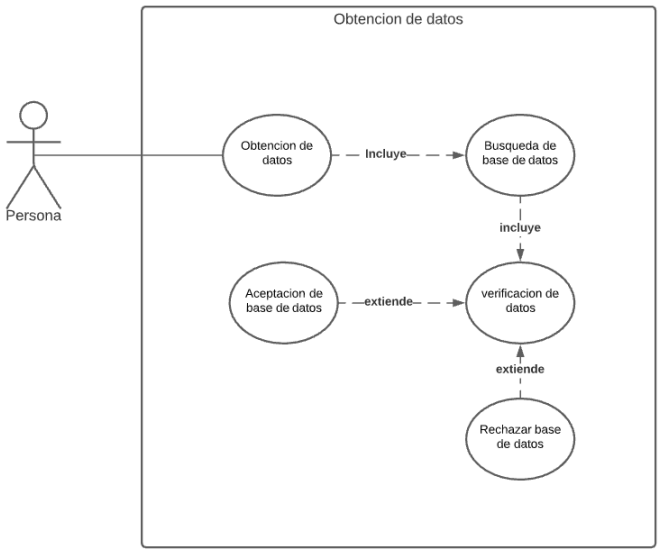


Ilustración 3 Diagramas de casos de uso para la obtención de datos

Se plantea el siguiente diagrama UML de casos de uso para el problema del preprocesamiento de los datos, este problema es de suma importancia puesto que estos representan los audios con las grabaciones de las voces de personas, las cuales se utilizarán como entrada en nuestro modelo de Machine Learning.

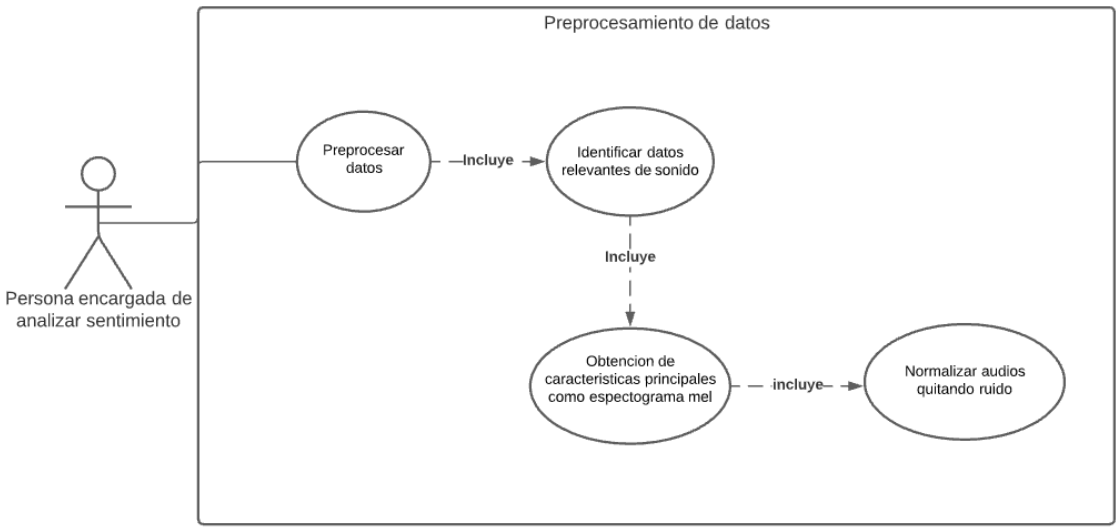


Ilustración 4 Diagramas de casos de uso para el preprocesamiento de datos

Descripción de la solución propuesta

Para predecir las emociones a partir de un audio específico se utilizará la técnica de aprendizaje automático supervisado basada en árboles de decisión denominada Random Forest (RF). Este modelo se caracteriza porque cada árbol se entrena con distintos conjuntos de datos para el mismo problema, de tal manera que, unos errores se compensan con otros para obtener una predicción generalizada.

Si bien es cierto que para el análisis de emociones a través de audio se pueden utilizar otros algoritmos de ML como SVM (Máquinas de vectores de soporte), K-NN (K vecinos más próximos) o GB (Aumento del gradiente), se escogió RF debido a que destaca en el reconocimiento de emociones como enojo, tristeza, miedo y alegría (con una menor precisión) [2], las cuales están más presentes en los estudiantes durante las clases en línea [3].

Estudios recientes muestran diferentes implementaciones para el análisis de voz por audio con las técnicas de ML ya mencionadas, pero con un dataset pequeño con alrededor de 360 grabaciones de 4 actores [4]. Por otra parte, para el entrenamiento de nuestro modelo se usará un dataset llamado RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), el cual contiene alrededor de 45 minutos de audio dividido en 671

archivos, de los cuales participan 24 actores profesionales, cada grabación intenta reflejar una emoción[5].

Esta solución beneficiaria tanto a estudiantes como profesores, ya que mientras el modelo analiza los sentimientos del estudiante al interactuar con el profesor u otros compañeros en una conversación, el profesor podrá observar la información resultante y así tomar las medidas necesarias para mejorar su rendimiento académico.

Modelos completos de diseño de la solución

A continuación, se muestra un diagrama referente el proceso de clasificación de emociones por medio de la voz utilizando machine learning, en el cual se mencionan los pasos más importantes durante la fase de entrenamiento del modelo.

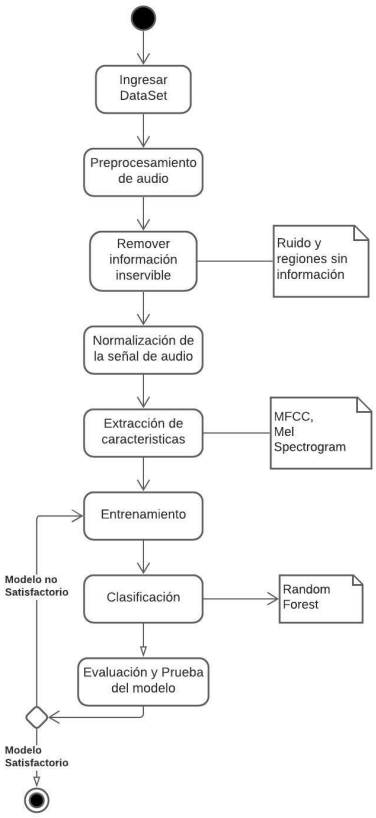


Ilustración 5. UML del diseño de la solución

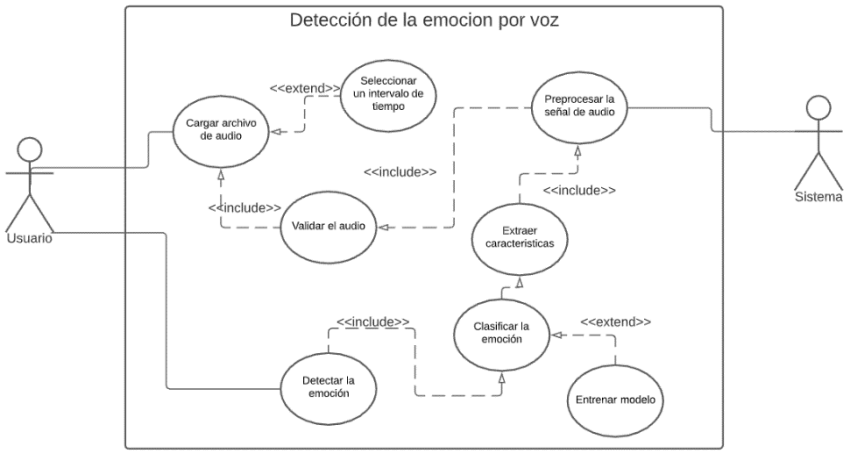


Ilustración 6. Caso de uso para la detección de emociones por voz.

A continuación, se describe el proceso de detección de emociones por voz:

- Se recibe un archivo de audio por parte del usuario y se valida antes de preprocesarlo.
- Opcionalmente se permite al usuario escoger un intervalo de tiempo del audio seleccionado.
- El sistema obtiene la señal de audio y la procesa para eliminar ruidos y sectores de la señal en donde no exista ningún tipo de información.
- Se extraen las características del audio, de entre las cuales se han considerado los MFFCs, Chroma y el espectrograma de Mel.
- Se utiliza el algoritmo de clasificación Random Forest para determinar la emoción correspondiente y presentarla como resultado al usuario.
- Adicionalmente, se puede entrenar el modelo con los patrones obtenidos del proceso.

Descripción de la implementación

El lenguaje de programación a utilizar para el desarrollo de la solución será Python, el cual cuenta con librerías muy útiles para la inteligencia artificial. Por otra parte, para la extracción de características del audio como: MFCCs, Chroma y espectrograma de Mel, se empleará la librería Librosa de python. Finalmente, la librería Scikit-learn, para el aprendizaje automático, se la utilizará para implementar el algoritmo de Random Forest. Con respecto a los hiper-parametros tenemos como más importantes:

- n_estimators= 22984
  Estos son los números de arboles de decisión en el bosque, se jugo con este parámetro para elevar la robustez del modelo, aunque el tiempo de entrenamiento haya aumentado.
- max_depth= 15
  Número máximo de niveles en cada árbol de decisión, el valor es pequeño para evitar overfitting puesto que si es más grande lo causaría
- min_samples_split= 9
  Número mínimo de puntos de datos colocados en un nodo antes de que se divida el nodo, jugamos con este parámetro para ajustar los arboles individuales
- min_samples_leaf= 3
  Número mínimo de puntos de datos permitidos en un nodo hoja, este parámetro es pequeño ya que se intenta tener menos rutas a los nodos hoja y de esta manera aumentar el rendimiento
- criterion= entropy
  EL criterio con los que se divide cada módulo se escogió entropía porque nuestra propuesta es de clasificación.

Cabe recalcar que, para la optimización o ajuste de estos hiperparámetros, se utilizará la técnica denominada “búsqueda en cuadricula”, que es una búsqueda exhaustiva que se realiza sobre los valores de parámetros específicos de un modelo y la cual se implementará con la biblioteca sklearn de pyhton.

Contribuciones:

- Se creo una interfaz gráfica por medio de la cual el usuario puede poner el nombre del archivo que quiere analizar o predecir la emoción que contiene
- Se extrajo y utilizó la característica MFCCs para el entrenamiento del modelo, que permiten identificar de mejor manera los sonidos que usan ciertas emociones.

Resultados:

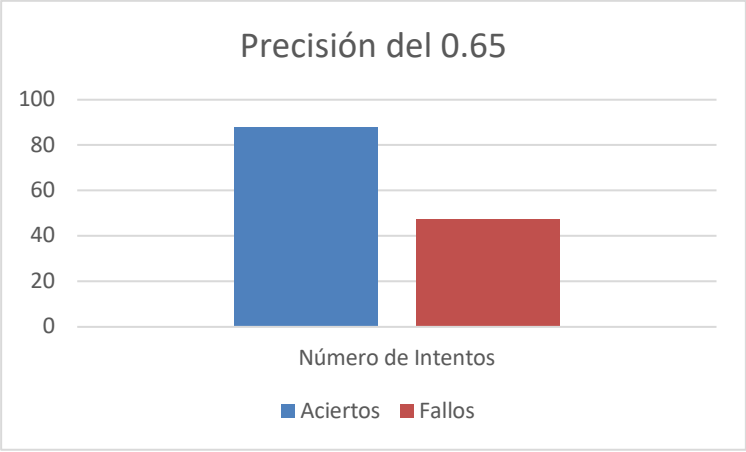
Después de entrenar a nuestro sistema pudimos obtener la siguiente matriz de confusión que nos muestra el desempeño de nuestro modelo:

Emocion Predicha / Emocion real	Hombre_enojado	Hombre_feliz	Hombre_neutral	Hombre_triste	Mujer_enojado	Mujer_feliz	Mujer_neutral	Mujer_triste
Hombre_enojado	15	2	0	3	0	0	0	0
Hombre_feliz	2	10	1	2	0	0	0	1
Hombre_neutral	0	3	8	6	0	0	0	0
Hombre_triste	1	5	0	11	0	0	0	0
Mujer_enojado	0	1	0	0	14	4	0	0
Mujer_feliz	0	0	0	0	4	13	0	1
Mujer_neutral	0	0	0	0	0	0	5	7
Mujer_triste	0	0	0	0	0	4	0	12

Ilustración 7 Matriz de confusión

Si nos fijamos en la diagonal principal esto nos dice los valores que acierta nuestro modelo mientras que lo que está fuera de esta son el número de veces que nuestro modelo predijo algo que no era.

Al analizar la data podemos obtener que 88 veces acertó y 47 veces fallo durante el entrenamiento lo que nos da una precisión del 0.65.



Conclusiones e impactos:

- Con la precisión actual de nuestro modelo podemos observar que el mismo por sí solo no es muy confiable, pero creemos que si hubiésemos tenido un dataset más completo esta precisión podría ser mayor.
- Creemos que para aumentar la precisión podríamos haber intentado combinar este con otros modelos.

Referencias

[1] G. d. Conyuntura, «La vivencia de los estudiantes universitarios ante el covid19,» Guadalajara, México, 2020.

[2] J. P. a. W. R. N. Morán, «Reconocimiento de estados emocionales de personas mediante la voz utilizando algoritmos de aprendizaje de máquina,» *Revista Venezolana de Computación,* pp. 5(2), 41-52, 2018.

[3] A. M. Fernández, «2020: Estudiantes, emociones, salud mental y pandemia,» *Revista Andina de Educación,* 2020.

[4] T. S. D. K. a. G. A. F. Noroozi, «Vocal-based emotion recognition using random forest and decision tree,» *International Journal of Speech Technology,* p. 8, 2017.

[5] F. A. R. Steven R. Livingstone, «The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American Englis,» *Journals Plos One,* 2018.