

# Estimación de la cantidad de mentas en una caja mediante análisis de audio y aprendizaje automático

Mateo Sanchez, Maria Jose Sanchez, Oscar Rodriguez

Escuela de Física, Universidad Industrial de Santander

Bucaramanga, Colombia

21 de mayo de 2025

{sebastian2221784, maria2221385, oscar2221532}@correo.uis.edu.co

**Abstract**—En este trabajo se presenta un sistema de estimación de la cantidad de mentas contenidas en una caja plástica, a partir del análisis del sonido generado al agitarla. Se recopiló más de 2400 grabaciones de audio bajo diferentes condiciones de control, y se extrajeron características acústicas relevantes como MFCCs, RMS, Zero Crossing Rate y Spectral Centroid. Estas características se utilizaron como entrada para entrenar un modelo de clasificación basado en Random Forest, optimizado mediante búsqueda de hiperparámetros (Grid Search). El modelo entrenado alcanzó una precisión de validación cruzada de 0.76 y una precisión final de 0.81 sobre el conjunto de prueba. El sistema implementado logra predecir la cantidad de mentas con una resolución efectiva de 5 unidades, lo cual se ve reforzado por la estrategia de utilizar múltiples grabaciones por muestra y promediar las predicciones individuales. Este enfoque demuestra que el aprendizaje automático, combinado con técnicas de procesamiento digital de señales, puede ser eficaz para resolver tareas de estimación no invasiva en entornos reales.

## I. INTRODUCCIÓN

En entornos donde la observación directa no es posible o resulta costosa, los métodos indirectos de estimación se han convertido en herramientas esenciales. Poder determinar la cantidad de elementos contenidos dentro de un recipiente cerrado —sin necesidad de abrirlo o intervenirlo físicamente— representa un desafío con implicaciones relevantes para industrias como la farmacéutica, alimentaria y logística, donde la verificación rápida y no invasiva puede traducirse en mejoras sustanciales en eficiencia, seguridad y control de calidad.

Este proyecto aborda el problema de estimar la cantidad de mentas contenidas en una caja plástica a partir del análisis del sonido generado al agitarla. Aunque el escenario puede parecer simple, el enfoque utilizado refleja una tendencia cada vez más relevante en ingeniería y ciencia de datos: el uso de señales acústicas como fuente de información cuantitativa, procesadas mediante técnicas de machine learning.

A partir de un conjunto de grabaciones sonoras etiquetadas, se realizó un proceso de extracción de características acústicas, incluyendo coeficientes MFCCs, energía (RMS), Zero Crossing Rate y Spectral Centroid, con el fin de capturar la estructura espectral y dinámica de cada señal. Estas características fueron empleadas como entradas para un modelo de clasificación basado en el algoritmo Random Forest, el cual ofrece ventajas como alta precisión, manejo de datos no lineales y análisis de importancia de variables.

El presente trabajo demuestra cómo el aprendizaje automático, en conjunto con el procesamiento digital de señales, puede emplearse para resolver problemas reales de estimación sin recurrir a sensores complejos o técnicas invasivas. La capacidad de entrenar modelos que aprendan a identificar patrones en los datos acústicos abre nuevas posibilidades para la implementación de sistemas inteligentes en escenarios industriales donde la automatización del conteo, monitoreo o control de contenido es una necesidad creciente.

## II. OBJETIVOS

### A. Objetivo general

Diseñar e implementar un sistema basado en técnicas de aprendizaje automático que permita estimar la cantidad de mentas contenidas en una caja, a partir del análisis del sonido generado al agitarla.

### B. Objetivo específico

- Recolectar grabaciones de audio suficientes para distintas cantidades de mentas
- Extraer características relevantes de las señales de audio (MFCCs, RMS, Zero Crossing Rate, Spectral Centroid), las cuales permitan representar cuantitativamente el sonido y su relación con la cantidad de mentas.
- Entrenar un modelo de Random Forest el cual será el encargado de realizar las predicciones en base a las características extraídas.
- Evaluar la viabilidad y el desempeño del modelo.
- Desarrollar un sistema funcional que permita grabar audio en tiempo real y estimar automáticamente la cantidad de mentas contenidas en la caja.

## III. MARCO TEÓRICO

### A. Coeficientes Cepstrales en las Frecuencias de Mel (MFCCs)

Los MFCCs (Mel-Frequency Cepstral Coefficients) son una representación del contenido espectral de una señal de audio que busca imitar la percepción auditiva humana [1]. Se utilizan ampliamente en reconocimiento de voz, clasificación de sonidos y tareas donde se requiere extraer información estructurada de señales acústicas.

El cálculo de los MFCCs se basa en los siguientes pasos:

- 1) Preénfasis y enventanado: La señal de audio  $x[n]$  es dividida en ventanas cortas (típicamente de 20 a 40 ms) donde se asume que el contenido estadístico es estacionario. Cada ventana se multiplica por una ventana de tipo Hamming para reducir los efectos de discontinuidad en los bordes.
- 2) Transformada de Fourier: Se calcula la *Transformada Rápida de Fourier (FFT)* de cada ventana para obtener el espectro de magnitudes  $|X[k]|$ . Este paso convierte la señal del dominio temporal al dominio frecuencial:

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-j2\pi kn/N}$$

- 3) Espectrograma en la escala Mel: Se aplica un banco de filtros triangulares en escala *Mel* sobre el espectro de potencia. La escala Mel está diseñada para reflejar cómo los humanos perciben las frecuencias, siendo más sensible a las bajas frecuencias que a las altas. La conversión de una frecuencia  $f$  en Hz a la escala Mel está dada por:

$$m = 2595 \cdot \log_{10} \left( 1 + \frac{f}{700} \right)$$

- 4) Logaritmo de energía: A continuación, se toma el logaritmo natural de las energías filtradas para simular la percepción logarítmica del volumen por parte del oído humano:

$$\log(E_i), \quad i = 1, \dots, M$$

donde  $E_i$  es la energía en el filtro  $i$  y  $M$  es el número total de filtros en la banca Mel.

- 5) Transformada Discreta del Coseno (DCT): Finalmente, se aplica la *Transformada Discreta del Coseno* a los log-energías obtenidos, lo que reduce la correlación entre coeficientes y deja los primeros valores como los más representativos del contenido del sonido:

$$c_n = \sum_{i=1}^M \log(E_i) \cdot \cos \left[ \frac{\pi n}{M} (i - 0.5) \right], \quad n = 1, \dots, N$$

Los coeficientes  $c_n$  obtenidos son los MFCCs. Normalmente se conservan los primeros 12 o 13 coeficientes, que contienen la mayor parte de la información relevante.

En este proyecto, los MFCCs permiten representar cada grabación como un vector numérico que refleja la estructura espectral del sonido producido por una determinada cantidad de mentas al ser agitadas. Al promediar los MFCCs a lo largo del tiempo, se obtiene una única representación fija por archivo, adecuada para su uso en algoritmos de aprendizaje automático.

### B. Valor Cuadrático Medio (RMS)

El valor cuadrático medio (*Root Mean Square*, RMS) es una medida estadística de la magnitud de una señal que permite estimar su energía promedio a lo largo del tiempo. En Física, el RMS se interpreta como el valor eficaz de una

señal oscilatoria, lo cual es útil para analizar, por ejemplo, tensiones y corrientes alternas, o el movimiento oscilatorio de un sistema.

Para una señal discreta  $x[n]$  compuesta por  $N$  muestras, el valor RMS se define como:

$$x_{\text{rms}} = \sqrt{\frac{1}{N} \sum_{n=1}^N x[n]^2}$$

Este valor representa una medida de la potencia acústica promedio de una señal de audio. En el contexto del proyecto, al agitar una caja con mentas, la energía de la señal registrada depende de factores como la cantidad de colisiones internas y su intensidad, que a su vez están influenciadas por el número de mentas dentro de la caja.

Cada colisión entre mentas produce una perturbación mecánica que se propaga como una onda de presión en el aire, la cual es captada por el micrófono. El RMS proporciona una estimación de la energía promedio asociada a esas oscilaciones acústicas.

### C. Zero Crossing Rate, ZCR

La Tasa de Cruces por Cero (Zero Crossing Rate, ZCR) es una medida que describe la cantidad de veces que una señal cambia de signo dentro de un marco de tiempo [2]. En otras palabras, contabiliza cuántas veces la señal pasa de ser positiva a negativa o viceversa, normalizado por la longitud de la ventana analizada. Esta medida se utiliza frecuentemente para caracterizar la textura temporal de una señal de audio.

Matemáticamente, el ZCR se puede expresar como:

$$Z(i) = \frac{1}{2W_i} \sum_{n=1}^{W_i} |\text{sgn}(x_i(n)) - \text{sgn}(x_i(n-1))|,$$

donde  $x_i(n)$  es la muestra  $n$ -ésima de la  $i$ -ésima ventana,  $W_i$  es la longitud de la ventana, y la función signo se define como:

$$\text{sgn}(x_i(n)) = \begin{cases} 1, & x_i(n) \geq 0, \\ -1, & x_i(n) < 0. \end{cases}$$

Una mayor tasa de cruces por cero suele estar asociada con frecuencias más altas en la señal, ya que implica una mayor cantidad de oscilaciones por unidad de tiempo.

### D. Spectral Centroid

El centroide espectral (*Spectral Centroid*) es una medida del centro de "gravidad" espectral de una señal, útil para describir la posición espectral promedio de la energía contenida en un marco de audio. [3]

Matemáticamente, el centroide espectral del  $i$ -ésimo marco de señal, denotado por  $C_i$ , se define como:

$$C_i = \frac{\sum_{k=1}^{W_i} k \cdot X_i(k)}{\sum_{k=1}^{W_i} X_i(k)},$$

donde:

- $X_i(k)$  es la magnitud del espectro en la frecuencia correspondiente al índice  $k$  del frame  $i$ ,
- $W_i$  es el número total de bins espectrales del frame  $i$ .

Desde una perspectiva computacional, el centroide espectral se obtiene a partir de la transformada de Fourier de la señal y sus respectivas magnitudes. Conceptualmente, el centroide espectral indica hacia qué zona del espectro (bajas o altas frecuencias) se concentra mayor energía acústica. Valores más altos del centroide están asociados a sonidos con mayor contenido de altas frecuencias, comúnmente percibidos como más “brillantes”, mientras que valores más bajos indican sonidos más “oscuros” o graves.

Además, esta medida es robusta al ruido y se usa ampliamente en aplicaciones como clasificación de timbre, análisis de textura sonora y segmentación de eventos acústicos [4]. En este proyecto, se espera que el centroide espectral refleje variaciones debidas al número de mentas en la caja, ya que colisiones múltiples tienden a generar componentes de mayor frecuencia en el espectro.

#### E. Random Forest

El algoritmo *Random Forest* es un método de aprendizaje automático supervisado que se utiliza tanto para clasificación como para regresión. Se basa en la construcción de múltiples árboles de decisión entrenados sobre subconjuntos aleatorios de los datos de entrenamiento. Cada árbol produce una predicción, y la predicción final del modelo se obtiene por votación mayoritaria (en tareas de clasificación) o por promedio (en tareas de regresión) (ver diagrama 1).

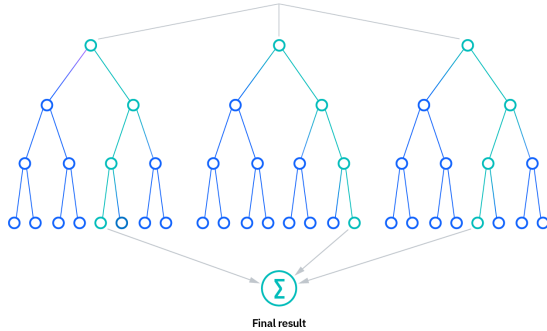


Fig. 1: Diagrama del funcionamiento del algoritmo Random Forest

El nombre *Random Forest* proviene de dos aspectos clave del modelo: por un lado, el uso de *bagging* (bootstrap aggregating), que implica entrenar cada árbol con un subconjunto diferente de datos seleccionados al azar con reemplazo; y por otro lado, la selección aleatoria de un subconjunto de características para dividir en cada nodo durante la construcción del árbol. Esta aleatoriedad introducida reduce la correlación entre los árboles y mejora la generalización del modelo [5].

Las principales ventajas del algoritmo Random Forest son:

- Tiene alta precisión y generalización.

- Es robusto frente al sobreajuste, especialmente cuando se entrena con suficientes árboles.
- Puede manejar eficientemente grandes conjuntos de datos y características.
- Proporciona medidas de importancia para cada característica.

De acuerdo con IBM [6], Random Forest se ha convertido en una herramienta fundamental para tareas de minería de datos, detección de fraudes, diagnóstico médico, y procesamiento de señales, debido a su balance entre precisión, interpretabilidad y eficiencia computacional.

Cada árbol del modelo realiza una predicción independiente, y el resultado final se obtiene por votación:

$$\hat{y} = \text{mode}(T_1(x), T_2(x), \dots, T_B(x))$$

donde  $T_i(x)$  representa la predicción del  $i$ -ésimo árbol para la entrada  $x$ , y  $B$  es el número total de árboles del bosque.

#### F. Grid Search

En modelos de aprendizaje automático, los hiperparámetros son valores configurables que controlan el proceso de entrenamiento (por ejemplo, la profundidad de un árbol o el número de árboles en un Random Forest). A diferencia de los parámetros internos del modelo, los hiperparámetros no se aprenden directamente a partir de los datos, sino que deben seleccionarse cuidadosamente para maximizar el rendimiento del modelo.

*Grid Search* es una técnica sistemática para la optimización de hiperparámetros. Consiste en definir una malla de posibles valores para cada hiperparámetro y luego entrenar y validar el modelo para todas las combinaciones posibles de estos valores. La evaluación se realiza generalmente usando validación cruzada, lo cual permite obtener un estimado más confiable del rendimiento del modelo.

El procedimiento se puede resumir en los siguientes pasos:

- 1) Se define un conjunto de valores posibles para cada hiperparámetro.
- 2) Se entrena el modelo con cada combinación de valores.
- 3) Se evalúa cada combinación mediante validación cruzada.
- 4) Se selecciona la combinación con mejor desempeño según una métrica específica (como la precisión o el *F1-score*).

Esta técnica garantiza que se explore exhaustivamente el espacio de búsqueda, aunque su desventaja es que puede ser computacionalmente costosa si el número de combinaciones es elevado. No obstante, se trata de un método robusto y ampliamente utilizado en tareas de clasificación y regresión cuando se dispone de recursos computacionales suficientes.

## IV. METODOLOGÍA

### A. Diseño y Materiales

Para el desarrollo de este proyecto se utilizaron caramelos Tic Tac®, un producto fabricado por la empresa italiana Ferrero. Estos caramelos se caracterizan por su consistencia

sólida, su forma elipsoidal y su tamaño uniforme. Además, vienen contenidos en una caja rígida de plástico, la cual resulta ideal para generar sonidos reproducibles al ser agitada. Estas propiedades físicas —tanto del contenido como del envase— fueron fundamentales para establecer un sistema experimental controlado en el cual el sonido generado dependiera principalmente de la cantidad de caramelos en el interior.

Para la recolección de los datos de audio se emplearon dos dispositivos principales: un teléfono celular y un micrófono de condensador Ovedisa-G77. Este último cuenta con un chipset de sonido profesional y una cápsula de 0.63 pulgadas, lo que permite obtener grabaciones con alta fidelidad y bajo nivel de ruido. El micrófono soporta tasas de muestreo de hasta 192 kHz/24 bits y posee una respuesta de frecuencia amplia, lo que facilita la captura de sonidos nítidos y detallados, incluso en frecuencias bajas. Además, su patrón polar cardioide ayuda a minimizar el ruido de fondo, asegurando que el sonido de las mentas en la caja sea registrado de manera clara.

## B. Experimental

La fase experimental inició con la recolección de datos a través de grabaciones de audio, cubriendo cantidades de mentas desde 1 hasta 29 unidades. En total, se realizaron 84 grabaciones distintas por unidad, para un total de 2436 audios, manteniendo constante el tipo de recipiente (la caja original de Tic Tac®) para asegurar condiciones controladas.

Para incrementar la diversidad del conjunto de datos y reducir el riesgo de sobre-entrenamiento, se grabaron 870 audios utilizando un micrófono-G77 y otros 870 audios con un teléfono celular. En ambos casos, la agitación de la caja fue intencionadamente variada y no controlada, buscando capturar diferentes patrones de sonido.

Adicionalmente, con el objetivo de evitar que el modelo aprenda patrones específicos de agitación y se enfoque únicamente en la cantidad de mentas, se grabaron 696 audios adicionales empleando un metrónomo a 130 bpm. En estas grabaciones, la caja se agitó de manera lineal y controlada, recorriendo un rango de 20 cm por movimiento.

Posteriormente, se desarrolló un script en Python que permitió la extracción de características relevantes del audio. En particular, se calcularon los coeficientes cepstrales en las frecuencias de Mel (MFCCs), la energía cuadrática media (RMS), la tasa de cruces por cero (Zero Crossing Rate) y el centroide espectral (Spectral Centroid). Estas características fueron seleccionadas debido a su capacidad para capturar diferentes aspectos del contenido sonoro: los MFCCs codifican la envolvente espectral, el RMS representa la intensidad del sonido, el ZCR indica la tasa de cambio de polaridad, y el centroide espectral proporciona una medida del brillo o prominencia de altas frecuencias.

Los valores extraídos fueron almacenados en un DataFrame y exportados a un archivo .csv, el cual sirvió como base para el entrenamiento del modelo de aprendizaje automático.

Todos los datos, scripts de procesamiento y entrenamientos pueden encontrarse en el siguiente repositorio de GitHub: <https://github.com/Matt22vL/RETOS->.

## V. MODELO DE PREDICCIÓN (RANDOM FOREST)

Para abordar el problema de estimar la cantidad de mentas contenidas en una caja a partir del análisis de sus características acústicas, se optó por el uso del algoritmo de *Random Forest*, una técnica de aprendizaje supervisado ampliamente utilizada por su robustez, precisión y facilidad de interpretación.

*Random Forest* es un método de ensamblado que construye múltiples árboles de decisión durante la fase de entrenamiento y, para la clasificación, emite como predicción final el voto mayoritario entre dichos árboles. Esta estrategia permite reducir el sobreajuste típico de los árboles individuales, y al mismo tiempo mejora la capacidad del modelo para generalizar a nuevos datos. En este proyecto, dicha característica resulta fundamental, dado que las señales de audio pueden contener ruido y variabilidad asociada a la forma en la que se agita la caja o la posición del micrófono.

El conjunto de datos fue dividido en un 80% para entrenamiento y un 20% para prueba utilizando la función `train_test_split` de la biblioteca `scikit-learn`. Las características extraídas de los audios (MFCCs, RMS, Zero Crossing Rate, Spectral Centroid, entre otras) fueron utilizadas como variables predictoras, mientras que la etiqueta objetivo corresponde al número de mentas presentes en la caja.

El modelo fue implementado con `RandomForestClassifier` de `scikit-learn`, y posteriormente optimizado mediante una búsqueda de hiperparámetros, cuyo proceso se describe con mayor detalle en la siguiente subsección.

Finalmente, el modelo ajustado fue evaluado utilizando métricas clásicas de clasificación y visualizaciones como la matriz de confusión y la importancia relativa de las características extraídas.

### A. Búsqueda de parámetros con Grid Search

Para optimizar el rendimiento del modelo Random Forest, se utilizó el método de Grid Search. Este método permite explorar de manera sistemática una combinación de hiperparámetros definidos previamente, con el fin de encontrar la configuración que proporcione el mejor desempeño del modelo.

En particular, se consideraron los siguientes hiperparámetros:

- `n_estimators`: número de árboles en el bosque.
- `max_depth`: profundidad máxima de cada árbol.
- `min_samples_split`: número mínimo de muestras requeridas para dividir un nodo.
- `min_samples_leaf`: número mínimo de muestras que debe tener una hoja.
- `max_features`: número de características a considerar para la mejor división.
- `bootstrap`: indica si se usan muestras con reemplazo.

La malla de búsqueda empleada fue la siguiente:

```
param_grid = {
    'n_estimators': [200, 300],
    'max_depth': [20, None],
    'min_samples_split': [2, 4],
    'min_samples_leaf': [1, 2],
    'max_features': ['sqrt'],
    'bootstrap': [True]
}
```

Listing 1: Definición de la malla de hiperparámetros para Grid Search

El procedimiento se llevó a cabo utilizando GridSearchCV de la biblioteca scikit-learn, con validación cruzada de 5 pliegues ( $cv=5$ ) y la métrica de precisión como criterio de evaluación. De esta manera, se garantizó una evaluación robusta del rendimiento del modelo en distintos subconjuntos del conjunto de entrenamiento.

Una vez finalizada la búsqueda, se seleccionó el conjunto de hiperparámetros que obtuvo la mayor precisión promedio, y se utilizó para entrenar el modelo final. Esta optimización permitió mejorar significativamente la capacidad del clasificador para predecir correctamente la cantidad de mentas en nuevos datos.

### B. Métricas de precisión

Una vez entrenado el modelo con los hiperparámetros óptimos hallados mediante *GridSearch*, se procedió a evaluar su desempeño sobre el conjunto de prueba utilizando diversas métricas de precisión. Los mejores parámetros encontrados fueron:

- bootstrap = True
- max\_depth = none
- max\_features = sqrt
- min\_samples\_leaf = 1
- min\_samples\_split = 4
- n\_estimators = 300

El modelo optimizado alcanzó una precisión de validación cruzada de **0.76**, y una precisión final sobre el conjunto de prueba de **0.81**.

a) *Importancia de características*: Antes de analizar los resultados cuantitativos, se evaluó la importancia relativa de cada una de las variables de entrada utilizando la propiedad `feature_importances_` del clasificador basado en árboles. En la Figura 2 se muestra el ranking de las características.

Este análisis revela que las características relacionadas con los coeficientes MFCC son las más informativas para la clasificación, lo cual es coherente con la literatura en análisis de señales de audio, donde los MFCC suelen capturar patrones acústicos distintivos.

b) *Matriz de confusión*: En la Figura 3 se presenta la matriz de confusión del modelo final. Se observa un alto grado de precisión, con una fuerte presencia de valores en la diagonal principal, lo cual indica una correcta identificación de la mayoría de las clases.

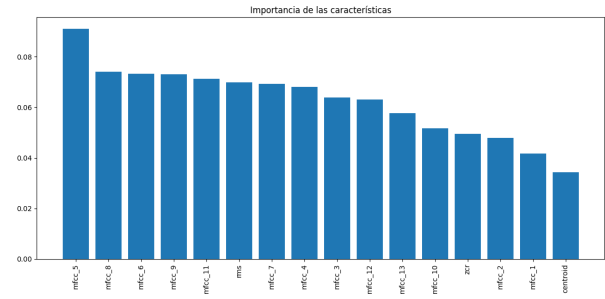


Fig. 2: Importancia relativa de las características según el modelo entrenado. Las más relevantes son los coeficientes MFCC, especialmente mfcc\_5, mfcc\_8 y mfcc\_6.

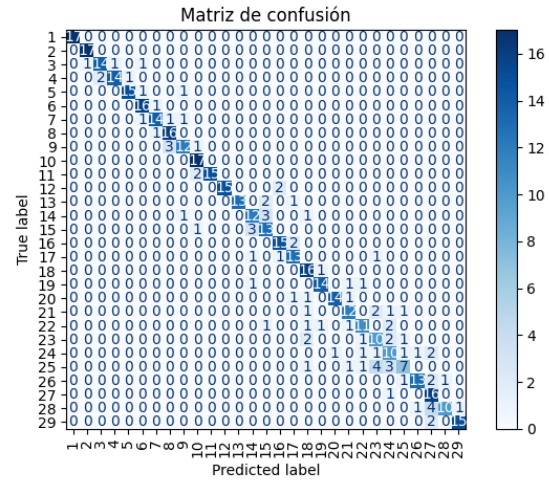


Fig. 3: Matriz de confusión normalizada. Cada celda muestra la cantidad de predicciones para cada par (real, predicho).

En general, se observa un bajo nivel de confusión entre clases distintas, lo cual sugiere que el modelo ha capturado patrones discriminativos efectivos. No obstante, algunas clases como la 28 y la 25 presentan menor rendimiento, lo cual podría deberse a similitudes acústicas con otras clases.

1) *Precisión por clase*: En la gráfica 4 se muestra la precisión del modelo para cada conjunto de numero de mentas, es posible notar que aproximadamente cada 5 mentas la precisión del modelo mejora.

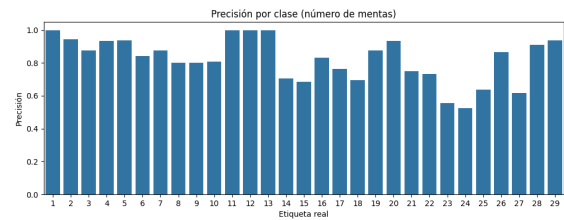


Fig. 4: Precisión del modelo Random Forest por numero de menta

## VI. PUESTA EN PRACTICA

Debido al rendimiento de nuestro modelo para cierto numero de mentas, se probó el modelo (con distintos grabaciones de audios a las usadas en el entrenamiento) para 5,10,15,20,25 mentas con el argumento de que son las cantidades que en general tienen una tendencia a mejor comportamiento.

Para garantizar que el modelo prediga en función de la cantidad de mentas, y no de la forma de agitar, la grabación de estos audios siguieron el procedimiento mencionado en IV-B con el metrónomo.

Teniendo en cuenta esto, se realizaron varias predicciones para 5,10,15,20,25 mentas y se analizaron sus resultados con ayuda de la gráfica 5

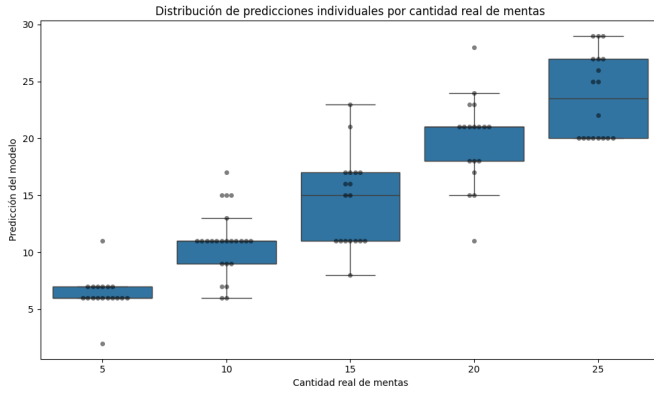


Fig. 5: diagrama de cajas y bigotes de predicción

Este diagrama de cajas y bigotes 5 evidencia que, para cada clase válida (cantidad real de mentas), las predicciones individuales del modelo presentan cierta dispersión. Sin embargo, es notable que el promedio de estas predicciones tiende a alinearse con el valor real correspondiente. Este comportamiento sugiere que, aunque el modelo puede cometer errores puntuales en predicciones individuales, su tendencia general es acertada cuando se consideran varias muestras.

## VII. IMPLEMENTACIÓN DEL SISTEMA

Aprovechando los resultados expuestos en la sección VI, se definió la estrategia de predicción del sistema: para cada estimación, se toman 6 grabaciones independientes, se realiza la predicción sobre cada una y, posteriormente, se calcula un promedio ponderado considerando la precisión (probabilidad) de cada predicción individual. De esta manera, se mitiga el efecto de posibles valores atípicos y se obtiene una estimación final más robusta y confiable del número real de mentas en la caja.

### A. Resultados del sistema de estimación

En la Tabla I se presentan los resultados obtenidos por el sistema de estimación, considerando el método propuesto de tomar 6 grabaciones independientes para cada muestra y promediar las predicciones individuales. Como se observa, el sistema logra predecir correctamente la cantidad de mentas en

todos los casos evaluados, mostrando una alta robustez y precisión cuando se utiliza la estrategia de promedio ponderado y redondeo a las clases válidas.

TABLE I: Resultados de la estimación del número de mentas utilizando conjuntos de 6 audios por muestra.

| Valor real | Predicción 1 | Predicción 2 | Predicción 3 |
|------------|--------------|--------------|--------------|
| 5          | 5            | 5            | 5            |
| 10         | 10           | 10           | 10           |
| 15         | 15           | 15           | 15           |
| 20         | 20           | 20           | 20           |
| 25         | 25           | 20           | 25           |

Estos resultados confirman que el modelo, aunque presenta cierta dispersión en las predicciones individuales, tiende a acertar en el valor real al considerar el promedio de varias muestras, validando así la estrategia adoptada para la estimación final.

### B. Implementación de grabación en tiempo real

Para facilitar la interacción del usuario y permitir la estimación del número de mentas en tiempo real, se desarrolló una interfaz de grabación automática utilizando Python. La implementación hace uso de la biblioteca `sounddevice`, que permite la captura de audio directamente desde el micrófono del sistema con alta flexibilidad y control sobre parámetros como la frecuencia de muestreo y la duración de la grabación.

El sistema guía al usuario a través de la grabación de múltiples muestras de audio, almacenando cada archivo en una carpeta designada. Posteriormente, cada grabación es procesada para extraer las características relevantes del sonido, utilizando la biblioteca `librosa`. Estas características se normalizan y se ingresan al modelo de predicción previamente entrenado.

Finalmente, el sistema calcula una predicción final, mostrando el resultado al usuario de manera clara e interactiva. Esta funcionalidad no solo mejora la experiencia de uso, sino que también permite validar el desempeño del modelo en condiciones reales, acercando la solución a un entorno de aplicación práctica.

## VIII. DISCUSIÓN

Si bien los resultados obtenidos con el sistema de estimación son satisfactorios bajo condiciones controladas, aún existen varias oportunidades de mejora que podrían aumentar la precisión y robustez del modelo. Una de las principales limitaciones identificadas es la sensibilidad del sistema a la forma en que se agita la caja. Actualmente, el modelo se entrena y evalúa utilizando grabaciones realizadas con un metrónomo a 130 bpm, lo cual garantiza cierta regularidad en el patrón de agitación. Sin embargo, esto implica que el sistema funciona óptimamente solo bajo estas condiciones, restringiendo su aplicabilidad práctica.

Para ampliar el rango de condiciones operativas del sistema, una mejora viable sería el desarrollo de un mecanismo físico que agite la caja de manera consistente y repetible, como una biela-manivela o algún tipo de actuador electromecánico. Esta



solución permitiría generar un conjunto más amplio de grabaciones con variabilidad reducida en la agitación, facilitando así el aprendizaje de patrones relacionados exclusivamente con la cantidad de mentas.

Durante el desarrollo del proyecto, se exploraron distintas estrategias para superar esta dependencia del patrón de agitación. Se intentó entrenar modelos exclusivamente con grabaciones no controladas, pero los resultados fueron significativamente peores. Ante esta situación, se consideró la posibilidad de que el algoritmo Random Forest fuera insuficiente para capturar la complejidad del problema. Sin embargo, tras probar métodos más avanzados como XGBoost, redes neuronales convolucionales desarrolladas con TensorFlow, y hasta modelos preentrenados como PANNs (Large-Scale Pretrained Audio Neural Networks), los resultados siguieron siendo insatisfactorios.

También se experimentó con un amplio rango de características extraídas con la biblioteca Librosa, más allá de las utilizadas finalmente (MFCCs, RMS, ZCR, Spectral Centroid). A pesar de estas múltiples combinaciones de arquitecturas, técnicas y descriptores, los modelos seguían presentando dificultades para generalizar cuando las condiciones de agitación no eran controladas.

Estas observaciones nos permiten concluir que el problema no reside únicamente en la arquitectura del modelo ni en las características extraídas, sino que está íntimamente relacionado con la naturaleza de la señal misma: cuando no se controla la forma de agitar, el modelo tiende a aprender patrones relacionados con el estilo de movimiento en lugar de con la cantidad de mentas.

Por ello, la estrategia más efectiva fue emplear una combinación de audios con agitación controlada (para garantizar consistencia en la señal) y audios no controlados (para mejorar la generalización), lo que permitió construir un sistema que, aunque limitado a ciertas condiciones, presenta buen rendimiento y una base sólida para futuras mejoras.

#### A. Importancia del Aprendizaje Automático

Una observación clave durante el desarrollo del proyecto fue que las señales de audio correspondientes a diferentes cantidades de mentas presentan estructuras espectrales similares. La Figura 6 muestra cinco espectrogramas representativos (5, 10, 15, 20 y 25 mentas), donde se aprecia que, a simple vista, las diferencias entre clases son mínimas. Esta alta similitud plantea una gran dificultad para implementar soluciones basadas en reglas fijas, análisis estadísticos simples o incluso inspección manual.

Dado este contexto, el uso de modelos de aprendizaje automático se vuelve fundamental. Algoritmos como Random Forest son capaces de identificar patrones sutiles no evidentes al ojo humano, permitiendo una clasificación precisa incluso cuando las representaciones visuales aparentan ser indistinguibles. Este tipo de modelos ofrecen una ventaja significativa al explotar relaciones no lineales y combinaciones complejas entre múltiples características acústicas extraídas de

las señales, lo que sería prácticamente inviable de lograr con enfoques tradicionales.

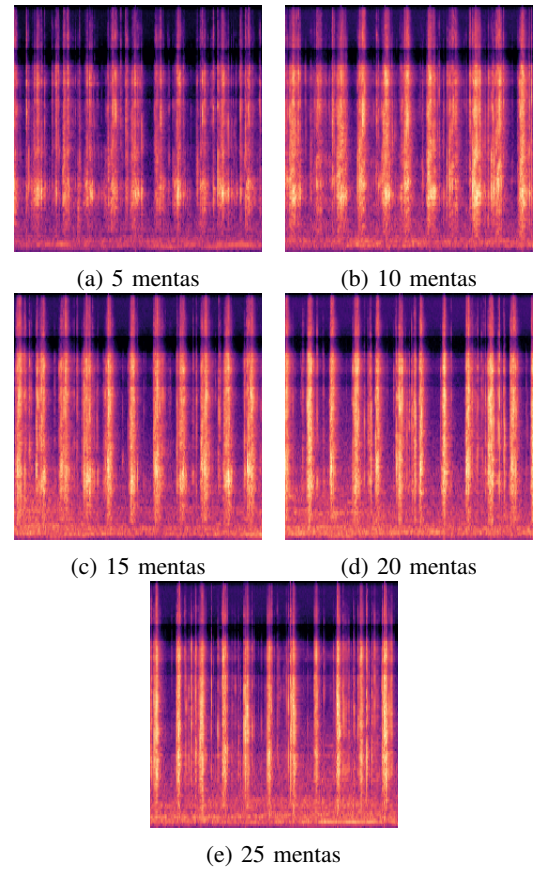


Fig. 6: Espectrogramas para distintas cantidades de mentas.

## IX. CONCLUSIONES

- El sistema propuesto demuestra una alta capacidad para realizar buenas estimaciones con una resolución de 5 mentas a partir del análisis acústico del sonido generado al agitarla. La precisión de validación cruzada del 76 % y del 81 % sobre el conjunto de prueba valida la eficacia del modelo Random Forest combinado con características relevantes como MFCCs, RMS, ZCR y Spectral Centroid.
- A pesar de haber explorado modelos más complejos como XGBoost, redes neuronales convolucionales (CNNs) implementadas en TensorFlow, e incluso arquitecturas preentrenadas como PANNs, los resultados obtenidos bajo condiciones no controladas de agitación fueron insatisfactorios. Esta limitación revela que el principal obstáculo no radica en la capacidad de los modelos de aprendizaje automático, sino en la variabilidad inherente a la señal acústica cuando no se regula la frecuencia de movimiento. En ausencia de control sobre la agitación, los modelos tienden a aprender patrones asociados al estilo de movimiento en lugar de capturar la relación entre el sonido y la cantidad real de mentas.
- El uso de modelos de aprendizaje automático resultó fundamental para abordar un problema en el que las

diferencias entre clases no son evidentes ni fácilmente separables a simple vista. A pesar de la alta similitud entre las señales acústicas generadas por diferentes cantidades de mentas, el modelo Random Forest fue capaz de identificar patrones sutiles en los datos, logrando una clasificación precisa. Esto demuestra el potencial del machine learning como herramienta poderosa para tareas de estimación no invasiva, especialmente en contextos donde métodos tradicionales o análisis visuales resultan ineficaces.

#### REFERENCES

- [1] B. Logan, "Mel frequency cepstral coefficients for music modeling," *Proc. 1st Int. Symposium Music Information Retrieval*, 11 2000.
- [2] T. Giannakopoulos and A. Pikrakis, "Chapter 4 - audio features," in *Introduction to Audio Analysis*, T. Giannakopoulos and A. Pikrakis, Eds. Oxford: Academic Press, 2014, pp. 59–103. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B9780080993881000042>
- [3] V. K. Harpale and V. K. Bairagi, "Chapter 3 - seizure detection methods and analysis," in *Brain Seizure Detection and Classification Using EEG Signals*, V. K. Harpale and V. K. Bairagi, Eds. Academic Press, 2022, pp. 51–100. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B9780323911207000086>
- [4] M. Massar, M. Fickus, E. Bryan, D. Petkie, and A. Terzuoli, "Fast computation of spectral centroids," *Adv. Comput. Math.*, vol. 35, pp. 83–97, 07 2011.
- [5] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [6] IBM, "¿qué es un bosque aleatorio?" <https://www.ibm.com/mx-es/think/topics/random-forest>, 2023, consultado el 3 de mayo de 2025.
- [7] B. McFee *et al.*, "librosa: Audio and music signal analysis in python," 2023. [Online]. Available: <https://librosa.org/doc/main/generated/librosa.feature.mfcc.html>
- [8] C. Panagiotakis and G. Tziritas, "A speech/music discriminator based on rms and zero-crossings," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 155–166, 2005.