INSTITUTO POLITÉCNICO NACIONAL
ESCUELA SUPERIOR DE CÓMPUTO
PRACTICA DE CLASIFICACIÓN (Bagging, Boosting)

Grupo: _____
No. Equipo: _____
Nombres: Ulises Abdiel Cabello Cardenas
Fernanda Madariaga Villanueva

Descripción del conjunto de datos

Reconocimiento de Género por Voz. Identificar una voz como masculina o femenina.

Disponible en: `https://www.kaggle.com/datasets/primaryobjects/voicegender/data`

Diccionario de datos

El conjunto de datos cuenta con 22 atributos, los cuales se describen en la siguiente tabla.

Variable	Tipo	Significado
meanfreq	numérico	Frecuencia media (kHz)
sd	numérico	Desviación estándar de la frecuencia
median	numérico	Frecuencia mediana (kHz)
Q25	numérico	Primer cuantil (kHz)
Q75	numérico	Tercer cuantil (kHz)
IQR	numérico	Rango intercuartílico (kHz)
skew	numérico	Asimetría del espectro
kurt	numérico	Curtosis del espectro
sp.ent	numérico	Entropía espectral
sfm	numérico	Planitud espectral
mode	numérico	Frecuencia de la moda
centroid	numérico	Centroide de frecuencia
peakf	numérico	Frecuencia pico (mayor energía)
meanfun	numérico	Promedio de frecuencia fundamental
minfun	numérico	Frecuencia fundamental mínima
maxfun	numérico	Frecuencia fundamental máxima
meandom	numérico	Promedio de frecuencia dominante
mindom	numérico	Frecuencia dominante mínima
maxdom	numérico	Frecuencia dominante máxima
dfrange	numérico	Rango de frecuencia dominante
modindx	numérico	Índice de modulación
label	categórico	Género: masculino o femenino

Consideraciones encontradas en el conjunto de datos

Describir las consideraciones que encuentre en el conjunto de datos.

Name	Type	# Missing values	# Unique values
sd	Number (Float)	0	3166
median	Number (Float)	0	3077
Q25	Number (Float)	0	3103
Q75	Number (Float)	0	3034
IQR	Number (Float)	0	3073
skew	Number (Float)	0	3166
kurt	Number (Float)	0	3166
sp.ent	Number (Float)	0	3166
sfm	Number (Float)	0	3166
mode	Number (Float)	0	2825
centroid	Number (Float)	0	3166

Figure 1: Consideraciones del conjunto de datos

meanfun	Number (Float)	0	3166
minfun	Number (Float)	0	913
maxfun	Number (Float)	0	123
meandom	Number (Float)	0	2999
mindom	Number (Float)	0	77
maxdom	Number (Float)	0	1054
dfrange	Number (Float)	0	1091
modindx	Number (Float)	0	3079
meanfreq	Number (Float)	0	3166

Figure 2: Consideraciones del conjunto de datos

- **Tipos de datos:** Todos los atributos predictivos del conjunto de datos son de tipo *numérico (float)*, mientras que la variable objetivo `label` es *categórica* con dos clases: `male` y `female`. Esto facilita el uso de modelos basados en árboles, pues no requieren codificación adicional ni normalización estricta.
- **Valores faltantes:** Ninguna variable contiene valores faltantes: Missing Values = 0 Por lo tanto, no se necesitó realizar imputación ni eliminación de registros.
- **Número de valores únicos:** Varias variables presentan un número muy alto de valores únicos (hasta 3166), lo que indica un comportamiento continuo. Sin embargo, se observaron variables con menor variabilidad, tales como:
  - `maxfun`: 123 valores únicos
  - `mindom`: 77 valores únicos
  - `minfun`: 913 valores únicos
  - `dfrange`: 1091 valores únicos

  Estas diferencias sugieren que algunas características espectrales poseen menor dispersión que otras.
- **Presencia de outliers:** Aunque no existen valores faltantes, se identifica la presencia de valores extremos en variables como: `skew`, `kurt`, `maxdom` y `dfrange`. Debido a que

los modelos de tipo árbol (Random Forest y Boosting) son robustos ante outliers, se decidió no realizar eliminación ni transformación de estos valores.

- **Balance de clases:** El conjunto de datos está perfectamente balanceado con: 1584 voces masculinas Esto evita la necesidad de aplicar técnicas de balanceo como *oversampling* o *undersampling*.
- **Distribución y variabilidad de los atributos:** Atributos como `meanfreq`, `median`, `meanfun` y `centroid` presentan alta variabilidad, lo que corresponde con mediciones continuas del espectro de voz. En contraste, variables como `mindom` y `maxdom` muestran menos valores únicos debido a la naturaleza discretizada de la frecuencia dominante.

El conjunto de datos presenta buena calidad, no requiere un tratamiento adicional significativo y es adecuado para técnicas de clasificación como Random Forest y Boosting, pero para evitar problemas se realizara una normalizacion en los datos.

Objetivo de la práctica

Objetivo: Crear un modelo de clasificación para identificar las características distintivas en los tipos de voz de mujeres y de hombres. Aplique métodos de Random Forest (Bagging) y Boosting.

Agregue la descripción de las variables dependientes e independientes. Explicando el nombre del algoritmo y sus características.

Tratamiento de datos

Describa cada una de las variables del conjunto de datos, analice el problema que presentan y aplique la técnica de tratamiento de datos según corresponda. Por ejemplo: tipo, valores faltantes, tratamiento necesario a realizar. Describa el proceso desarrollado.

Creación del o los modelos de clasificación

Genere dos modelos de clasificación: Random Forest (Bagging) y Boosting.

Creación de métricas

Analice la matriz de confusión para identificar y explicar los resultados encontrados.

Desarrolle las siguientes métricas y su significado:

- Matriz de confusión
- Sensibilidad
- Precisión

- Tasa de error
- Exactitud
- Especificidad
- Explicación de VP, FP, VN, FN
- Curva ROC: significado e interpretación
- Importancia de atributos dentro del modelo

Análisis de datos

Describa el significado de las métricas generadas a partir del modelo de clasificación. Identifique dos errores de elementos clasificados erróneamente y describa la razón.

Conclusiones

Agregue las conclusiones de la investigación desarrollada.

Referencias bibliográficas