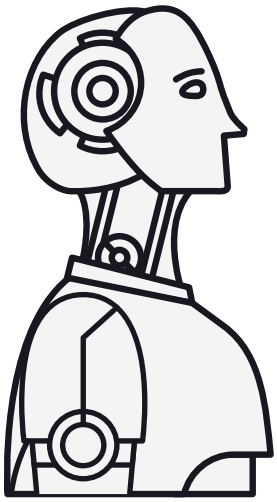

DETECCIÓ DE PARLA FALSA MITJANÇANT LA INTEL·LIGÈNCIA ARTIFICAL

PAULA JUSTO GILI

INTRODUCCIÓ A L'APRENENTATGE AUTOMÀTIC

GRAU EN INTEL·LIGÈNCIA ARTIFICIAL

Universitat Politècnica de Catalunya

Dissabte 28 de Desembre 2024

1. Introducció

Aquest treball preten aplicar els coneixements adquirits durant el curs per crear un model capaç de distingir entre àudios reals i generats per intel·ligència artificial. Es parteix d'un conjunt de dades que inclou més de 80.000 fitxers d'àudio, amb característiques acústiques i a partir d'aquestes es preten desenvolupar models predictius capaços d' identificar patrons per classificar correctament els àudios.

El treball està dividit en diverses etapes on primer es realitza un anàlisi estadístic i un preprocessat de les dades, seguit d'una normalització i reducció de la dimensionalitat. Posteriorment es realitza un entrenament amb tres models diferents, ajustant els hiperparàmetres i refinant els resultats obtinguts. Finalment, es fa la selecció del millor model i aquest es documenta amb un Model Card.

2. Objectius

L'objectiu principal d'aquest treball és adquirir la capacitat de desenvolupar models d'aprenentatge automàtic de manera metòdica i justificable, amb un enfocament basat en l'evidència i una interpretació crítica dels resultats.

A més, els objectius específics que s'aborden al llarg del projecte inclouen:

- Analitzar i comprendre el conjunt de dades, gestionant de manera adequada els missings, outliers i desbalancejos de les classes per garantir dades netes i consistents per a l'entrenament dels models.
- Aplicar tècniques com la normalització i, quan calgui, la reducció de dimensionalitat, a més d'eliminar variables irrellevants o sorolloses que podrien afectar el rendiment dels models.
- Entrenar i comparar tres models d'aprenentatge automàtic, ajustant els seus hiperparàmetres per optimitzar el rendiment.
- Identificar el model que millor prediu si un àudio és real o sintètic, basant-se en una avaluació rigorosa de les seves mètriques de rendiment.
- Reflexionar sobre els resultats obtinguts, considerant tant els punts forts com les limitacions dels models desenvolupats, i proposant possibles millores per a futures implementacions.

3. Anàlisis i Preprocessat de dades

En la primera fase del treball, per tal de poder preparar el conjunt de dades per a l'entrenament del model i assegurar-ne la qualitat, cal començar amb l'anàlisi estadístic de les variables, la identificació de missings i outliers, l'estudi del balanceig de la classe objectiu, la recodificació de variables i la partició del conjunt de dades. Tots aquests passos són essencials per entendre millor el comportament de les dades i es troben documentats i justificants en detall per garantir la validesa i fiabilitat dels resultats obtinguts durant les fases posteriors del treball.

El primer pas del procés va ser explorar les dades del fitxer train_csv per comprendre l'estructura del conjunt de dades provinent d'aquest dataset. En aquesta exploració inicial, es va observar que el conjunt contenia les variables categòriques i, a més, contenia columnes amb informació redundant que podia unificar-se per simplificar l'anàlisi i optimitzar la gestió de les dades. Per tant, es va decidir fer la unificació de les variables `Sex`, `Country` i `ID` creant noves columnes, amb les seves combinacions respectives anomenades `Combined_Sex`, `Combined_Country` i `Combined_ID`, en un nou DataFrame (`df`). D'aquesta manera s'han pogut eliminar redundàncies i treballar amb un conjunt de dades més net i estructurat.

	F_path	Category	Source_ID	Target_ID	Source_Sex	Source_Country	Target_Sex	Target_Country	ID	Sex	Country	UniqueID	Realornot
0	FinalDataset_16khz/Real/Argentina/arf_00295/ar...	Real	NaN	NaN	NaN	NaN	NaN	NaN	arf_00295	Female	Argentina	73932	1
1	FinalDataset_16khz/Real/Argentina/arf_00295/ar...	Real	NaN	NaN	NaN	NaN	NaN	NaN	arf_00295	Female	Argentina	55678	1
2	FinalDataset_16khz/Real/Argentina/arf_00295/ar...	Real	NaN	NaN	NaN	NaN	NaN	NaN	arf_00295	Female	Argentina	67831	1
3	FinalDataset_16khz/Real/Argentina/arf_00295/ar...	Real	NaN	NaN	NaN	NaN	NaN	NaN	arf_00295	Female	Argentina	71460	1
4	FinalDataset_16khz/Real/Argentina/arf_00295/ar...	Real	NaN	NaN	NaN	NaN	NaN	NaN	arf_00295	Female	Argentina	61757	1

Figura1: train_csv.head( ) original

	Category	UniqueID	Realornot	Combined_Sex	Combined_Country	Combined_ID
0	Real	73932	1	Female	Argentina	arf_00295
1	Real	55678	1	Female	Argentina	arf_00295
2	Real	67831	1	Female	Argentina	arf_00295
3	Real	71460	1	Female	Argentina	arf_00295
4	Real	61757	1	Female	Argentina	arf_00295

Figura2: df.head( ) després de la unificació de variables

Aquest enfocament, ha aconseguit reduir la complexitat del conjunt de dades i garantir una millor organització i una gestió més eficient de les dades en les fases posteriors d'anàlisi i modelatge.

Un cop fet aquest primer ajustament en el dataset de train_csv es passa a explorar el fitxer smile_csv i s'observa que aquest és el fitxer que conté variables numèriques, com ara mètriques acústiques derivades dels enregistraments d'àudio.

Per tant, en aquest punt es decideix unir els dos datasets en un únic dataframe anomenat df_train. Per fer-ho s'ha utilitzat la columna `UniqueID`, que actua com a identificador únic de cada àudio. D'aquesta manera s'assegura que cada fila del nou dataframe representa completament les característiques de cada àudio, permetent una anàlisi més eficient i una preparació més directa per a l'entrenament del model.

	Category	UniqueID	Realornot	Combined_Sex	Combined_Country	Combined_ID	Unnamed: 0	F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope	F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope	F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope.1	...
0	Real	73932	1	Female	Argentina	arf_00295	73931	59.642597	80.991450	80.991450	...
1	Real	55678	1	Female	Argentina	arf_00295	55677	7.781728	32.577520	32.577520	...
2	Real	67831	1	Female	Argentina	arf_00295	67830	14.860586	18.454376	18.454376	...
3	Real	71460	1	Female	Argentina	arf_00295	71459	13.270463	85.602270	85.602270	...
4	Real	61757	1	Female	Argentina	arf_00295	61756	43.167866	22.110174	22.110174	...

Figura3: df_train( ).head

Aquest dataframe resultant, proporciona una visió completa i integrada del conjunt de dades, essencial per a les fases següents d'exploració, preprocessament i modelatge, i clau per poder aprofitar al màxim la informació disponible i així assegurar que el model sigui precís i fiable.

3.1. Partició del conjunt d'entrenament

Un cop està el dataset d'entrenament completament preparat, el següent pas és realitzar una partició adequada del conjunt de les dades. Aquesta partició és essencial per assegurar que el procés d'entrenament i avaluació sigui rigorós i robust. Apliquem una divisió estratificada per garantir que la distribució de la variable objectiu, `Realornot`, es mantingui representativa en cadascun dels subconjunts. Aquesta variable és binària i val 1 quan la veu del àudio és real i val 0 en cas contrari.

En primer lloc, s'agafa un 20% del total de df_train i el reserva per un conjunt de test. Aquest conjunt de test és intern i previ al test definitiu, ja que també es compta amb un fitxer separat (test_csv) que s'utilitzarà com a conjunt de test final per validar el model en un entorn completament independent i simular un escenari real de predicció.

El 80% restant del dataset es destina a train_data, que es divideix en dos subconjunts: train_sub (70% de train_data), que s'utilitza per entrenar el model, i val_data (30% de train_data), per ajustar els hiperparàmetres i avaluar el rendiment intermedi del model durant el procés d'entrenament.

```
Mida del conjunt de train: 14243
Mida del conjunt de test: 3561

Mida del sub-train: 9970
Mida del conjunt de validació: 4273
Mida del conjunt de test: 3561
```

Figura4: Mides dels diferents conjunts un cop fetes les particions

Aquest enfocament de partició ofereix avantatges clars per a l'entrenament i l'avaluació del model. Reservar un conjunt de test intern abans del test definitiu permet detectar problemes potencials durant el desenvolupament, garantint que el rendiment final del model es mesuri de manera independent i sense biaixos. Això assegura que tant el test intern com el definitiu siguin completament separats de les dades d'entrenament.

L'estratificació de la variable objectiu en totes les particions manté una distribució representativa de les dades reals, evitant desequilibris que podrien afectar el model. A més, utilitzar un conjunt de validació per ajustar els hiperparàmetres prevé el overfitting i millora la capacitat del model per generalitzar a noves dades.

Per tant, aquesta estratègia de partició, que inclou entrenament, validació i test, així com un test_csv final, proporciona una base robusta per avaluar el rendiment del model en entorns reals, assegurant confiança en els resultats i una preparació òptima per a aplicacions pràctiques.

3.2. Anàlisi estadísticas de les variables

Seguidament, s'ha realitzat l'anàlisi estadístic començant per estudiar les distribucions de les variables categòriques presents en el dataset. Aquesta anàlisi ens proporciona una visió general de com estan distribuïdes les diferents categories, ajudant-nos a identificar possibles desequilibris o patrons.

En la variable Category, es pot observar que el 50.21% del conjunt de dades correspon a registres reals (Real), mentre que el 49.79% restant es distribueix entre diverses tècniques de generació de veu sintètica com CycleGAN (8.42%), StarGAN (8.38%), i altres. Això indica que, tot i que hi ha un desbalanceig en la distribució de les categories individuals dins del grup de veus falses, la suma total (49.79%) és molt similar a la de les veus reals. Aquesta distribució equilibrada entre registres reals i falsos és especialment favorable per a l'entrenament del model, ja que assegura que el classificador pugui aprendre de manera equitativa a identificar tant veus reals com sintètiques.

Pel que fa a la variable `Combined_Sex`, la distribució entre homes (50.01%) i dones (49.99%) és pràcticament igual, assegurant que el model no es veurà esbiaixat cap a un gènere concret. Això és crucial per obtenir un model imparcial en termes de predicció de la classe objectiu.

En relació amb Combined_Country, observem una distribució lleugerament desigual, amb un predomini de països com Perú (22.48%) i Argentina (22.33%), seguits de Colòmbia (20.98%), Xile (18.73%) i, finalment, Veneçuela (15.49%). Tot i que hi ha certa variabilitat en les proporcions, aquestes diferències no són excessivament marcades i permeten que el model aprengui de dades representatives de cada país.

Finalment, la variable `Combined_ID`, que representa identificadors per a cada veu, mostra una distribució molt més heterogènia que les altres variables. No obstant això, aquesta variable serà eliminada en passos següents, ja que no aporta informació rellevant per a l'estudi i podria introduir un soroll innecessari en el model.

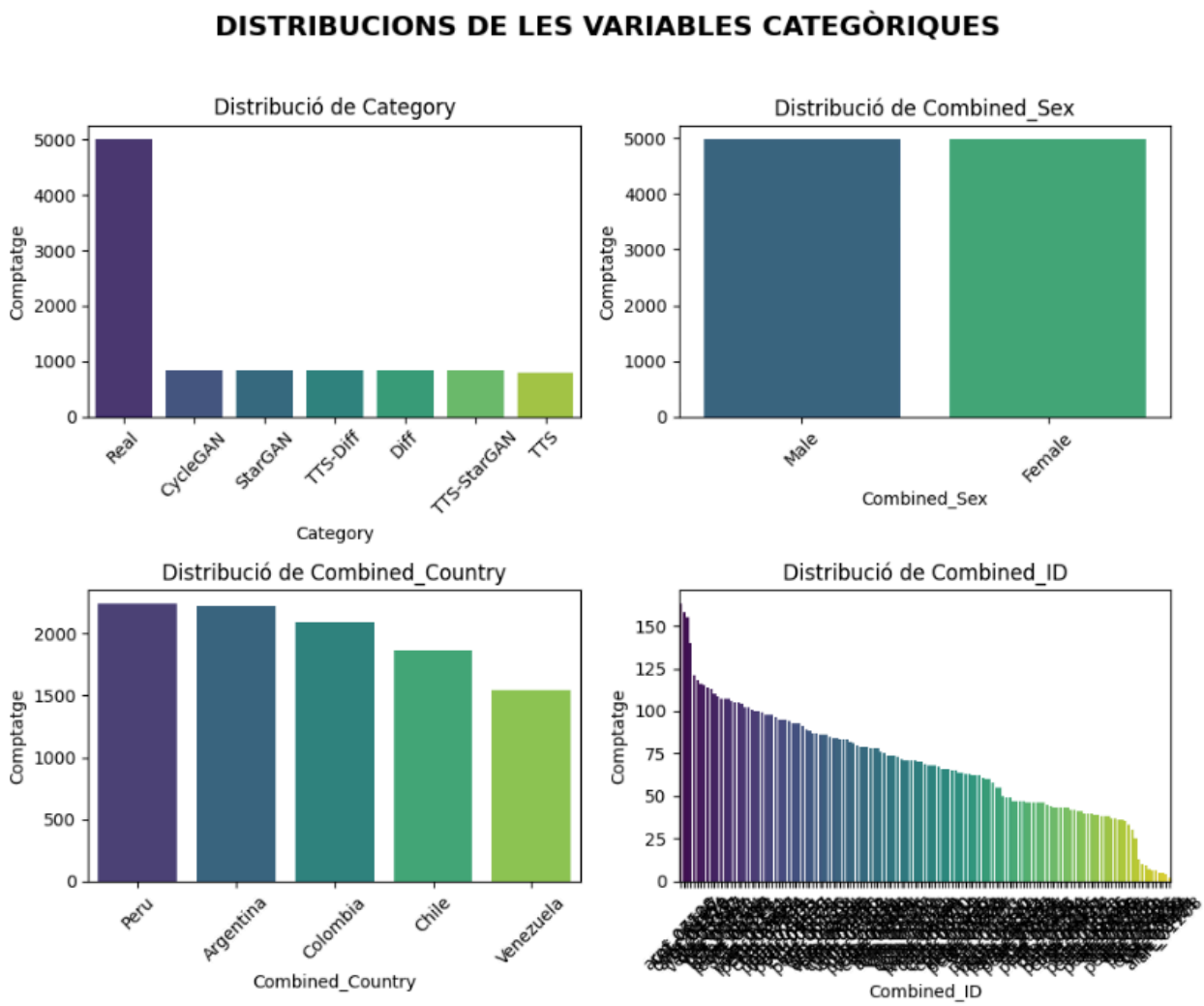

Figura5: Gràfics amb les distribucions de les variables categòriques Category, Combined_Sex, Combined_Country I Combined_ID

També és fonamental observar la distribució de les variables numèriques per obtenir una visió prèvia que permeti determinar si caldrà realitzar un tractament específic dels outliers. Aquesta anàlisi es fa mitjançant boxplots i histogrames, ja que aquestes eines gràfiques ens ajuden a identificar ràpidament possibles valors extrems i comprendre la forma de les distribucions.

En observar aquestes gràfiques, s'apreciar que la majoria de les variables numèriques semblen seguir una distribució normal o gaussiana, amb valors centrats al voltant de la mitjana i amb una dispersió relativament simètrica. Tot i això, algunes variables mostren característiques que corresponen més a una distribució exponencial, amb una acumulació de valors baixos i una cua allargada cap als valors més alts.

Gràcies a aquesta observació prèvia, es pot planificar una gestió adequada de les variables, decidint quines d'elles necessitaran transformacions o tractaments. A més, també permet assegurar que qualsevol tractament posterior, com la imputació de missings o la normalització, es faci d'una manera informada i justificada.

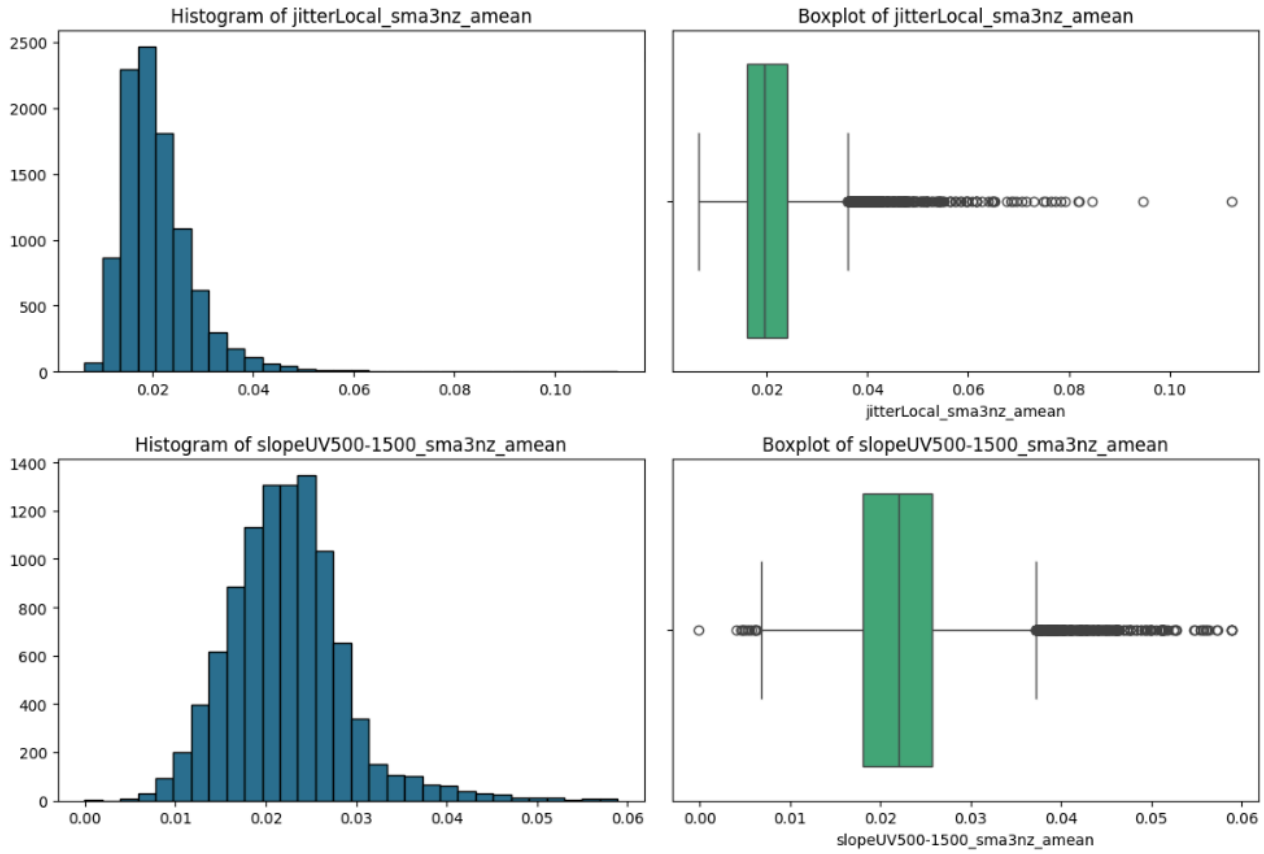

Figura6: Histograma i Boxplots de dues variables numèriques on la primera sembla tenir una distribució exponencial i la segona una normal

3.3. Balanceig de la variable objectiu

És essencial verificar el balanceig de la variable objectiu `Realornot` dins de les particions del conjunt de dades, fins i tot després d'una partició estratificada, per assegurar que el model no estigui influït per desequilibris entre les classes. En aquest cas, s'ha comprovat que les classes estan gairebé perfectament equilibrades: al conjunt de train, la classe 1 (Real) representa el 50.21% i la classe 0 (Fals) el 49.79%, proporcions que es mantenen pràcticament idèntiques als conjunts de validació i test. Aquest balanceig adequat permet al model aprendre de manera equitativa entre les dues classes, millorant la seva capacitat de generalització i evitant esbiaixaments.

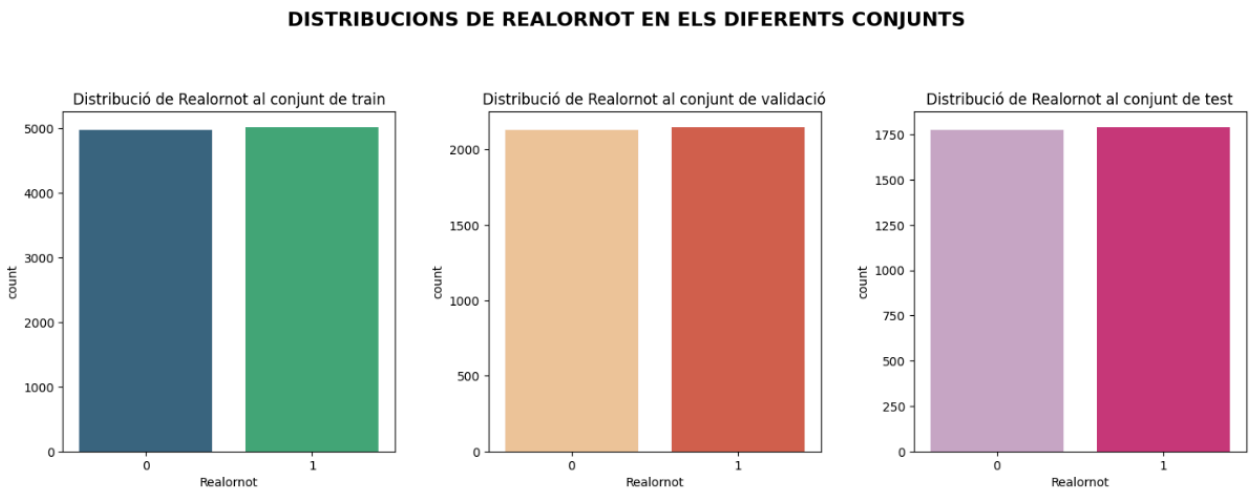

Figura7: Distirbució de la classe objectiu en els diferents conjunts de entrenament, validació i test

A més, s'ha analitzat com les variables categòriques `Combined_Sex`, `Combined_Country` i `Category` es distribueixen segons `Realornot`, detectant patrons visuals que poden millorar la comprensió de les dades i optimitzar el desenvolupament del model. Els gràfics resultants revelen patrons o correlacions visuals que poden aportar una millor comprensió del conjunt de dades i ajudar a optimitzar el desenvolupament del model predictiu, reforçant la confiança en la capacitat del model per generalitzar.

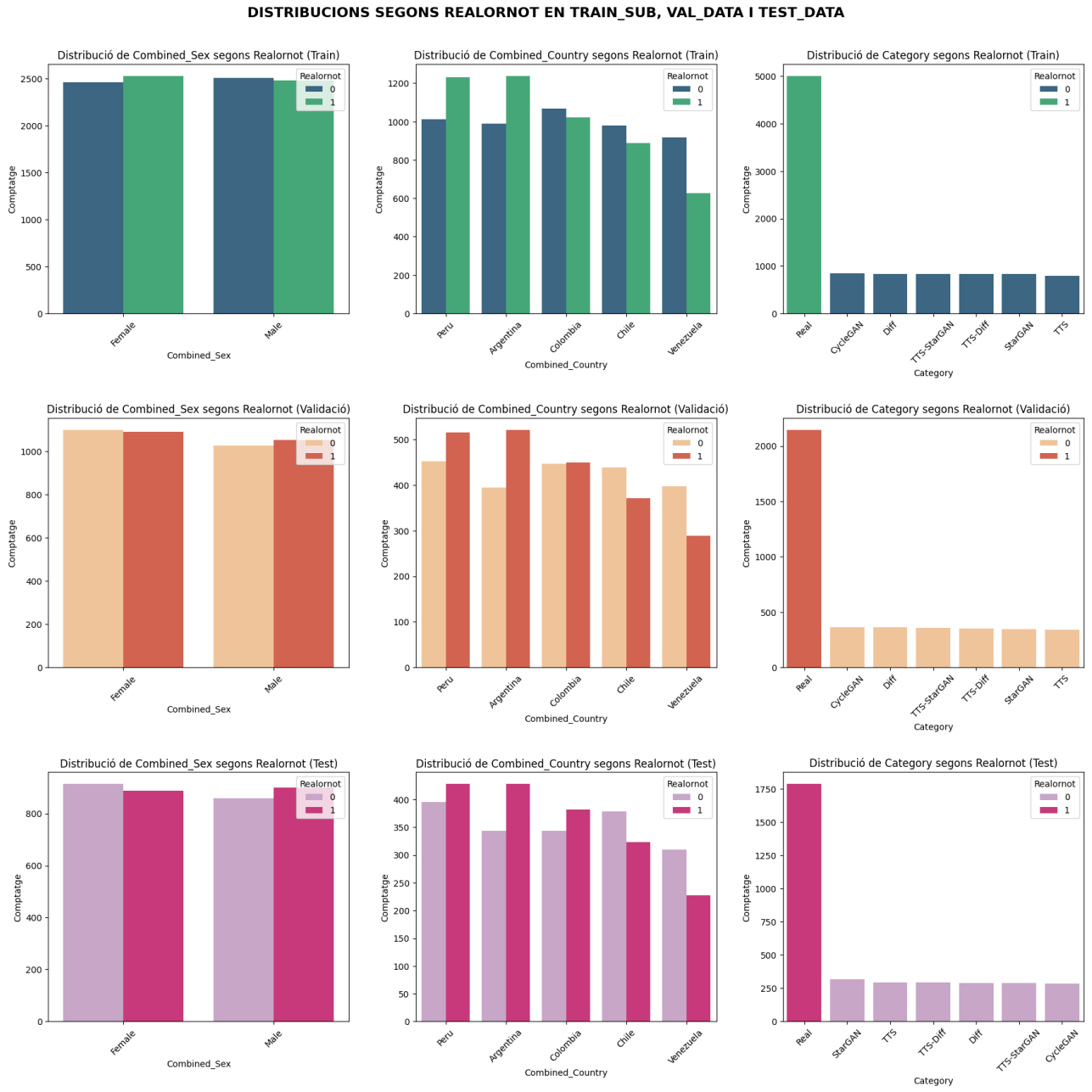

Figura8: Gràfics amb les distribucions de les variables Combined_Sex, Combined_Country i Category amb la variable Realornot en els conjunts de train, validació i test

Pel que fa a les variables numèriques, també s'ha verificat la seva distribució en funció de la variable objectiu, confirmant que estan força ben balancejades entre les classes. Aquest equilibri en les variables numèriques i categòriques garanteix que el model pugui aprendre patrons representatius tant per a veus reals com sintètiques, assegurant un rendiment fiable i just en la classificació.

3.4. Anàlisi i Tractament de Missings i Outliers

En relació amb l'anàlisi i tractament de missings, es pot observar que inicialment existien missings (NaN) en algunes variables categòriques. No obstant això, un cop creat el dataframe amb les variables combinades, aquests missings desapareixen. Per tant, això ens permet treballar amb un conjunt de dades categòriques complet i sense buits.

Pel que fa a les variables numèriques, s'han identificat missings que es manifesten com una quantitat considerable de valors 0 en variables on aquest valor no té sentit en el seu context. Les variables en les que s'ha detectat aquest esdeveniment són: `F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope` i `F0semitoneFrom27.5Hz_sma3nz_stddevFallingSlope`. Per tant, aquests zeros han estat considerats missings i s'ha decidit imputar-los utilitzant la mediana, ja que aquesta mètrica és menys sensible als valors extrems i, per tant, la més adequada per a dades que podrien estar afectades per outliers o amb distribucions asimètriques.

Aquest mateix procés s'ha aplicat als conjunts val_data i test_data per verificar si existeixen patrons similars als detectats en el conjunt train_sub. Per tal de garantir la consistència, els zeros identificats han estat imputats amb els valors de les medianes calculades prèviament sobre el conjunt train_sub. Aquesta decisió assegura que les dades de validació i test siguin tractades d'una manera coherent amb el conjunt d'entrenament, fet que millora la capacitat del model per generalitzar i evita biaixos introduïts per tractaments diferenciats.

Per la identificació d'outliers, s'ha utilitzat el mètode del Rang Interquartílic (IQR), una tècnica robusta i àmpliament utilitzada per avaluar outliers en distribucions numèriques. Aquesta decisió s'ha pres, ja que l'IQR és eficaç per identificar outliers en variables amb distribucions normals i és menys sensible a valors extrems que altres mètodes. Tot i haver-se observat que algunes variables numèriques tenen distribucions més properes a una exponencial que a una normal, per simplicitat i consistència, s'ha decidit aplicar el mètode del IQR a totes les variables numèriques.

També s'ha aplicat el mateix tractament al conjunt test_data i val_data. En aquests casos, per assegurar consistència, s'han identificat els outliers seguint el mateix mètode i, a l'hora d'imputar-los, s'han utilitzat les estadístiques derivades del conjunt train_sub. D'aquesta manera s'evita qualsevol biaix d'informació del conjunt de test cap al procés d'entrenament, preservant la validesa de l'avaluació.

Després de tractar els outliers, s'ha observat una lleugera reducció en la mida dels datasets, indicant que s'han eliminat dades extremes, outliers, que podien introduir soroll al model.

3.5. Recodificació de variables

S'han recodificat les variables numèriques del conjunt de dades per facilitar-ne la comprensió i el treball amb elles. A continuació, es detallen els nous noms de cada variable:

- `F0semitoneFrom27.5Hz_sma3nz_stddevRisingSlope` (desviació estàndard de la pendent ascendent del to fonamental) → `F0_stddev_rising_slope`
- `F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope` (mitjana de la pendent descendent del to fonamental) → `F0_mean_falling_slope`
- `F0semitoneFrom27.5Hz_sma3nz_meanFallingSlope.1` (mitjana de la pendent descendent del to fonamental) → `F0_mean_falling_slope.1`
- `F0semitoneFrom27.5Hz_sma3nz_stddevFallingSlope` (desviació estàndard de la pendent descendent del to fonamental) → `F0_stddev_falling_slope`
- `loudness_sma3_amean` (mitjana de la sonoritat) → `loudness_mean`
- `spectralFlux_sma3_stddevNorm` (desviació estàndard del flux espectral) → spectral_flux_stddev
- `mfcc1_sma3_amean` (mitjana del primer coeficient MFCC) → `mfcc1_mean`
- `mfcc1_sma3_stddevNorm` (desviació estàndard del primer coeficient MFCC) → `mfcc1_stddev`
- `mfcc2_sma3_amean` (mitjana del segon coeficient MFCC) → `mfcc2_mean`
- `mfcc2_sma3_stddevNorm` (desviació estàndard del segon coeficient MFCC) → `mfcc2_stddev`
- `mfcc3_sma3_amean` (mitjana del tercer coeficient MFCC) → `mfcc3_mean`
- `mfcc3_sma3_stddevNorm` (desviació estàndard del tercer coeficient MFCC) → `mfcc3_stddev`
- `jitterLocal_sma3nz_amean` (mitjana del jitter local) → `jitter_local_mean`
- `slopeUV500-1500_sma3nz_amean` (mitjana de la pendent entre 500 Hz i 1500 Hz) → `slope_500_1500_mean`
- `Unnamed: 0` (índex dels registres del conjunt de dades) → Índex

Aquest procés de recodificació s'ha aplicat de manera uniforme als conjunts train_sub, val_data i test_data.

4. Preparació de Variables

Per preparar les dades per a les anàlisis posteriors, s'ha aplicat one-hot-encoding a les variables categòriques `Combined_Sex` i `Combined_Country`, transformant-les en variables numèriques. Aquesta codificació evita relacions fictícies entre categories i assegura compatibilitat amb tècniques numèriques com el PCA.

El one-hot-encoding ha generat noves columnes per a cada categoria: per exemple, `Combined_Sex` s'ha dividit en `Combined_Sex_Female` i `Combined_Sex_Male`, mentre que `Combined_Country` s'ha

codificat en cinc columnes corresponents als països presents al conjunt de dades. Aquesta transformació ha incrementat el nombre de variables, però garanteix una representació explícita i uniforme de les categories.

4.1. Normalització de Variables

Per garantir que totes les variables tinguin la mateixa escala i evitar que aquelles amb rangs més amplis influeixin desproporcionadament en els models, s'ha aplicat un procés de normalització de les dades.

En aquest cas, s'ha distingit entre les variables a partir de la seva distribució, no com s'havia fet en la detecció i tractament d'outliers, on s'assumia que les distribucions de totes les variables era normal. Per les variables que, segons els histogrames del apartat 1.2 del notebook, semblaven seguir una distribució exponencial, s'ha aplicat una normalització logarítmica, ja que és la més adequada per reduir l'impacte de valors extrems en aquest tipus de dades. Per a la resta de variables, s'ha utilitzat StandardScaler, que ajusta les dades perquè tinguin una mitjana 0 i desviació estàndard de 1.

Durant el procés de normalització de les dades, s'ha detectat l'aparició de missings en algunes variables després d'aplicar la transformació logarítmica. Això és degut a la presència de valors negatius i/o zeros en aquestes variables abans de la normalització, ja que el logaritme natural no està definit per a aquests valors.

En concret, les variables amb percentatges de missings després de la transformació són:

- `F0_mean_falling_slope`: 3.12%
- `F0_mean_falling_slope.1`: 3.12%
- `mfcc1_stddev`: 0.01%
- `mfcc2_stddev`: 6.00%
- `mfcc3_stddev`: 1.64%

Inicialment, es va intentar solucionar aquest problema sumant un petit desplaçament positiu als valors de les variables abans de la normalització logarítmica. Però, amb aquesta estratègia no es va tenir l'efecte desitjat, ja que els missings van continuar apareixent. Per tant, com que els percentatges de missings són molt baixos, s'ha decidit eliminar les mostres afectades, ja que la seva eliminació no afecta de manera significativa la mida ni la representativitat del conjunt de dades.

Un cop resolta aquesta incidència, la normalització s'ha completat correctament per al conjunt d'entrenament. Posteriorment, per garantir la consistència, s'ha aplicat el mateix scaler pels conjunts de validació i de test assegurant que totes les dades utilitzades tinguin la mateixa escala, millorant la comparabilitat i la precisió dels resultats.

4.2. Eliminació de variables numèriques redundants o sorolloses

Per optimitzar el conjunt de dades i assegurar que només s'utilitza informació rellevant per al model, s'ha procedit a eliminar variables redundants i/o sorolloses. Aquest procés s'ha realitzat utilitzant una matriu de correlació per avaluar la relació entre les variables numèriques. Això permet identificar aquelles que estan altament correlacionades entre elles i, per tant, podrien aportar informació duplicada.

La interpretació de la matriu és clara: valors propers a 1 o -1 indiquen una correlació forta (i, per tant, redundància), mentre que valors propers a 0 assenyalen correlacions molt baixes, cosa que podria indicar soroll.

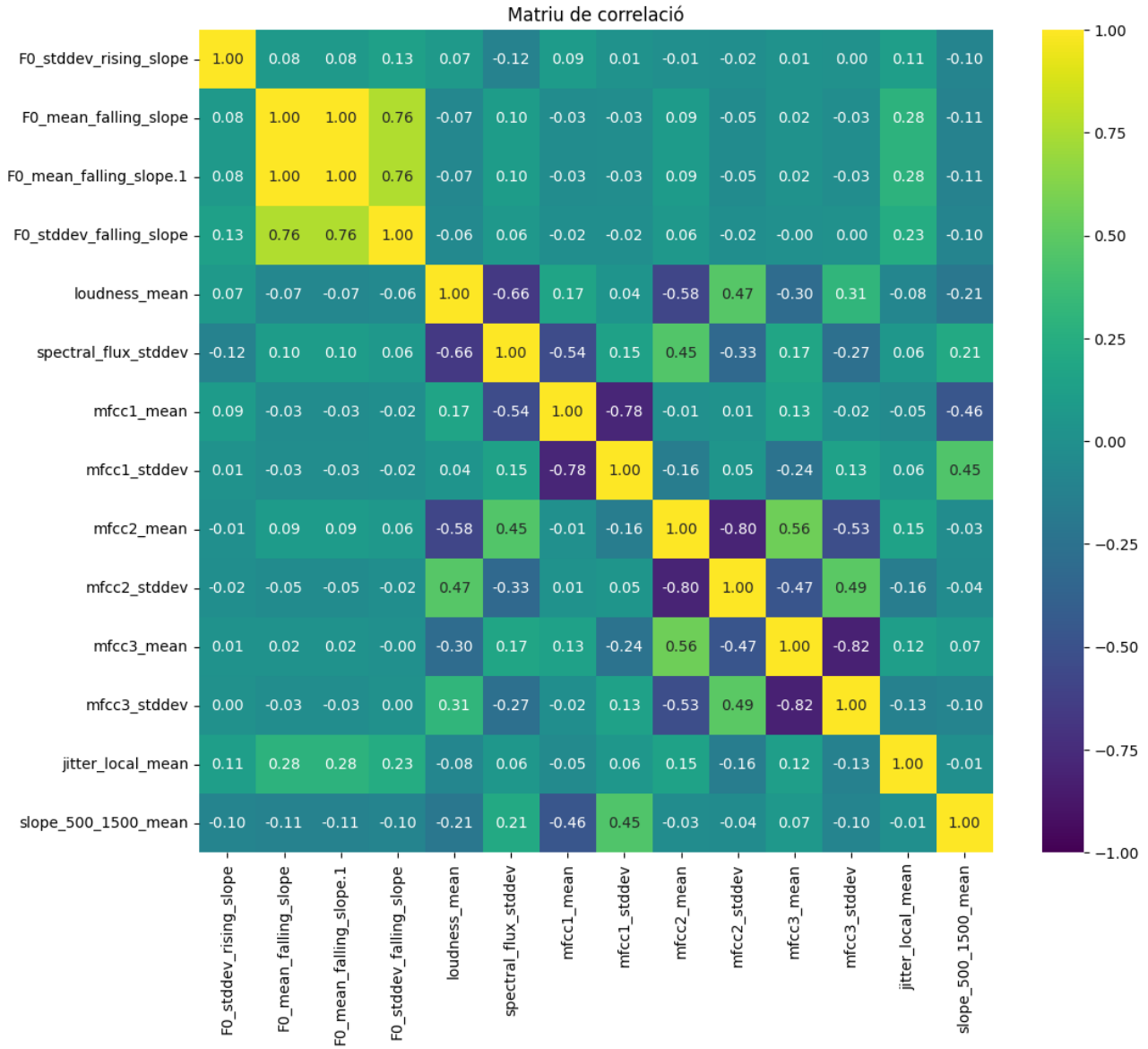

Figura9: Matriu de correlació

En l'anàlisi de la matriu de correlació que s'ha obtingut, es poden destacar alguns resultats i observacions rellevants:

- Relacions Fortes:

Hi ha una correlació molt forta entre `F0_mean_falling_slope` i `F0_mean_falling_slope.1` (correlació = 1), el que indica redundància entre aquestes dues variables. Això justifica eliminar una d'elles (`F0_mean_falling_slope.1`) per evitar la duplicació d'informació. També es veu una correlació inversa considerable entre `mfcc1_mean` i `mfcc1_stddev` (correlació = -0.78), de igual forma es veu com passa el mateix entre `mfcc2_mean` i `mfcc2_stddev`, i entre `mfcc3_mean` i `mfcc3_stddev`. Això implica que aquests parells de variables mesuren propietats oposades d'una característica comuna, i la redundància d'informació pot afectar negativament el model, per tant, cal eliminar una variable corresponent a cada parella.

- Variables Sorolloses:

La variable `jitter_local_mean` té correlacions molt baixes amb gairebé totes les variables, indicant que pot ser una variable sorollosa. Aquesta observació es reforça amb la seva baixa correlació amb la variable objectiu. La variable `F0_stddev_rising_slope` també presenta una correlació molt baixa amb la majoria de variables numèriques i amb la variable objectiu, fet que la qualifica també com a variable sorollosa i, per tant, que també s'ha d'eliminar.

- Observació Interessant:

La variable `loudness_mean` està inversament correlacionada amb altres característiques com `spectral_flux_stddev` (-0.66) i `mfcc2_mean` (-0.58). Això indica que un augment en la mitjana de la intensitat sonora tendeix a estar associat amb una disminució en altres característiques del senyal d'àudio, com la variació espectral o els coeficients cepstrals melòdics.

A més, també s'ha analitzat la correlació de cada variable amb la variable objectiu per veure quines tenen una poca correlació amb ella i, per tant, són poc probables a contribuir de manera significativa al rendiment del model. Així mateix, s'ha revisat la variabilitat de cada variable i s'han descartat aquelles amb una baixa variabilitat, ja que no proporcionen informació significativa per a la tasca de classificació.

En concret, la correlació amb la variable objectiu ha permès identificar algunes variables amb correlacions molt baixes, com ara `jitter_local_mean` (-0.006839) i `F0_stddev_rising_slope` (-0.122892), que s'han considerat sorolloses i s'han eliminat. De la mateixa manera, la baixa variabilitat de `jitter_local_mean` (variabilitat de només 0.007275) ha confirmat la seva exclusió.

Per tant, les variables eliminades han estat:

- `F0_mean_falling_slope.1`: Redundant amb `F0_mean_falling_slope`.
- `F0_stddev_rising_slope`: Sorollosa, correlació baixa amb la variable objectiu.

- `jitter_local_mean`: Sorollosa i amb baixa variabilitat.
- `mfcc1_stddev`: Redundant i baixa correlació amb Realornot.
- `F0_stddev_falling_slope`: Sorollosa, correlació baixa amb la variable objectiu.
- `Combined_ID`: No aporta informació rellevant i no està codificada.
- `Índex`: Identificador sense valor analític.
- `Category`: Variable categòrica no codificada i innecessària per a l'anàlisi.

Aquest procés d'eliminació de variables assegura que el conjunt de dades final sigui més net, compacte i informat, millorant l'eficiència i el rendiment del model en etapes posteriors.

```
CONJUNT DE TRAIN
SHAPE ABANS DE L'ELIMINACIÓ: (8749, 26)
SHAPE DESPRÉS DE L'ELIMINACIÓ: (8749, 18)

CONJUNT DE VALIDACIÓ
SHAPE ABANS DE L'ELIMINACIÓ: (3774, 26)
SHAPE DESPRÉS DE L'ELIMINACIÓ: (3774, 18)

CONJUNT DE TEST
SHAPE ABANS DE L'ELIMINACIÓ: (3117, 26)
SHAPE DESPRÉS DE L'ELIMINACIÓ: (3117, 18)
```

Figura10: Shape dels conjunts de dades abans i després de l'eliminació de les variables

4.3. Estudi de la Dimensionalitat amb PCA

Per dur a terme un estudi de reducció de dimensionalitat utilitzant l'anàlisi de components principals (PCA), per tal d'identificar les direccions principals en les que les dades varien més, primer s'ha hagut d'eliminar del conjunt de dades la variable objectiu, `Realornot`, i l'identificador `UniqueID`, ja que aquestes no poden formar part del PCA. L'objectiu del PCA és identificar les direccions principals en les quals les dades varien més, i aquestes dues variables podrien introduir soroll o influències innecessàries.

En aplicar el PCA al conjunt d'entrenament, s'observa com amb només 5 components principals es pot explicar més del 80% de la variància total del conjunt de dades. Aquesta reducció és significativa, ja que permet treballar amb un conjunt de dades més compacte i computacionalment eficient, alhora que es preserva la major part de la informació rellevant. A més, reduir la dimensionalitat també minimitza el risc d'overfitting, ja que es redueix la complexitat del model i, per tant, les possibilitats que aquest s'ajusti massa a soroll específic de les dades.

Un cop determinat que 5 components principals són suficients, s'han aplicat aquestes transformacions al conjunt d'entrenament, reduint-lo a la seva representació en aquest nou espai de menor dimensionalitat. Per garantir la coherència i la validesa de les anàlisis posteriors, aquesta mateixa transformació s'aplica de manera consistent als conjunts de validació i test, utilitzant el mateix PCA ajustat amb les dades d'entrenament.

4.3.1. Síntesi de les conclusions sobre les contribucions de les variables

- Component Principal 1:

La variable `Combined_Sex_Female` és la més influent, amb una contribució positiva destacada, mentre que `Combined_Sex_Male` i les mitjanes de mfcc tenen una contribució negativa. Això suggereix un contrast clar entre el sexe i les característiques acústiques.

- Component Principal 2:

`slope_500_1500_mean` domina aquest component, indicant la seva rellevància en les variacions espectrals. La mitjana de mfcc1 (`mfcc1_mean`) té una contribució fortament negativa.

- Component Principal 3:

`Combined_Sex_Female` torna a ser la variable més influent, reforçant el seu paper en la variabilitat. Característiques com `mfcc2_mean`, `mfcc3_mean` i `F0_mean_falling_slope` també són rellevants, mentre que `mfcc2_stddev` i `Combined_Sex_Male` contribueixen negativament.

- Component Principal 4:

La variable `F0_mean_falling_slope` és clarament dominant amb una contribució molt positiva, suggerint que aquest component està fortament associat amb aquesta característica acústica. Altres variables amb contribucions positives menors inclouen `Combined_Country_Chile`, `Combined_Sex_Male`, i `Combined_Country_Peru`, mentre que `mfcc1_mean`, `slope_500_1500_mean`, i `Combined_Country_Argentina` tenen una contribució negativa, indicant que aquestes dimensions s'oposen al patró liderat per `F0_mean_falling_slope`.

- Component Principal 5:

`F0_mean_falling_slope` també lidera aquest component, amb una contribució fortament positiva. Altres variables influents inclouen `slope_500_1500_mean`, `mfcc1_mean` i `mfcc3_mean`, mentre que `mfcc2_mean` i `Combined_Country_Peru` tenen contribucions negatives destacades.

4.3.2. Interpretació general

Les variables de sexe i les característiques acústiques, especialment `F0_mean_falling_slope`, `slope_500_1500_mean`, i les mitjanes i desviacions de mfcc, són clau per explicar la variabilitat en les dades. Aquesta anàlisi destaca dimensions importants relacionades amb patrons espectrals, diferències entre sexes i, en menor mesura, associacions amb regions. Això reafirma la importància d'analitzar les correlacions entre atributs acústics i les categories sociolingüístiques en les dades.

Aquest enfocament ens proporciona un conjunt de dades optimitzat que preserva la informació essencial, és computacionalment més manejable i està preparat per a les següents fases d'anàlisi i modelatge predictiu.

5. Definició de Models

5.1. Model KNN

El model K-Nearest Neighbors (KNN) és un algorisme de classificació basat en la similitud entre veins. Es tracta d'un mètode supervisat que classifica una observació assignant-li la classe majoritària dels k veins més propers, segons una mètrica de distància predefinida. Aquest model pot ser especialment interessant per aquest treball perquè és intuïtiu, no paramètric, i pot capturar patrons complexos en les dades sense fer suposicions fortes sobre la seva distribució. A més, com que les dades han estat normalitzades i s'ha aplicat una reducció de dimensionalitat amb PCA, el KNN pot operar de manera òptima en aquest espai reduït.

5.1.1. Selecció d'hiperparàmetres

Per tal d'optimitzar el rendiment del KNN, s'ha dut a terme una cerca en graella utilitzant les següents combinacions d'hiperparàmetres:

```
knn_params = {
   'n_neighbors': [1, 3, 5, 7, 10],
   'weights': ['uniform', 'distance'],
   'metric': ['euclidean', 'manhattan', 'minkowski'],
   'p': [1, 2, 3]
}
```

Després de la cerca, la millor combinació trobada ha estat la que utilitza la mètrica de minkowski amb p = 3, n_neighbors = 10 i weights = distance. Amb aquesta configuració, el model ha aconseguit una precisió mitjana de 0.6940 ± 0.0147.

5.1.1. Avaluació del model en validació

Un cop seleccionats els millors hiperparàmetres, s'ha aplicat el model KNN al conjunt de validació. Els resultats obtinguts han estat:

Mètriques per al conjunt de validació:

	precision	recall	f1-score	support
0	0.69	0.63	0.66	1768
1	0.70	0.74	0.72	2006
accuracy			0.69	3774
macro avg	0.69	0.69	0.69	3774
weighted avg	0.69	0.69	0.69	3774

Figura11: Mètriques del conjunt de validació

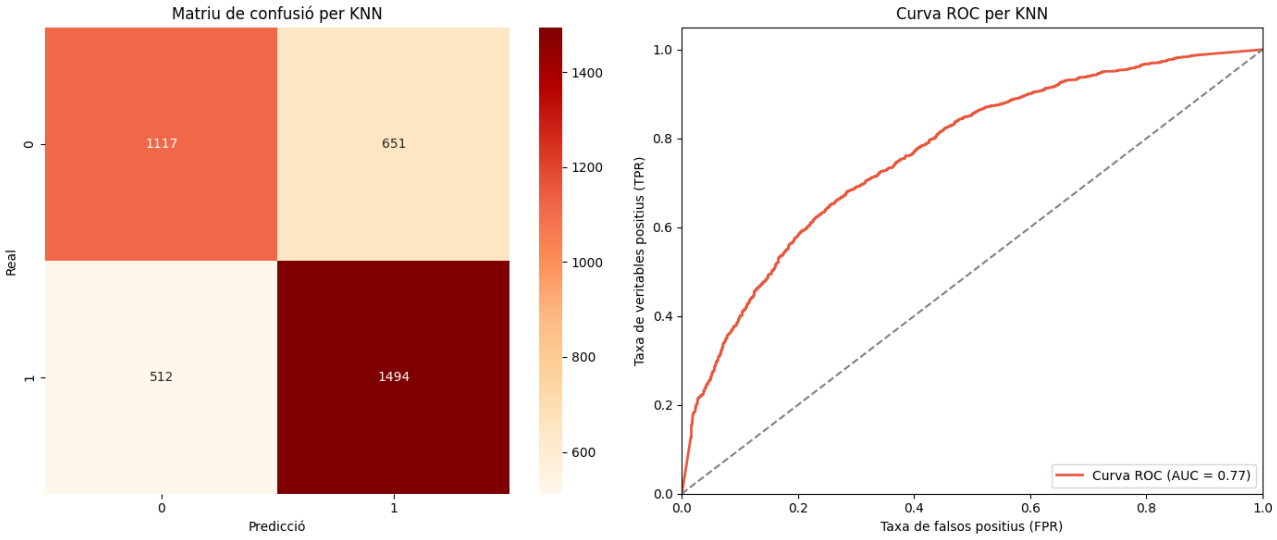

Figura12: Matriu de confusió i curva ROC del conjunt de validació

Aquestes mètriques reflecteixen un rendiment equilibrat entre les classes 0 i 1, amb una lleugera millora en el recall de la classe 1, fet que és important si l'objectiu és identificar casos de veu sintètica amb alta sensibilitat.

5.1.2. Avaluació del model en test

Per verificar la capacitat de generalització del model, s'ha aplicat al conjunt de test. Els resultats han estat molt consistents amb els obtinguts en validació:

Mètriques per al conjunt de test:

	precision	recall	f1-score	support
0	0.69	0.65	0.67	1468
1	0.70	0.75	0.72	1649
accuracy			0.70	3117
macro avg	0.70	0.70	0.70	3117
weighted avg	0.70	0.70	0.70	3117

Figura13: Mètriques del conjunt de test

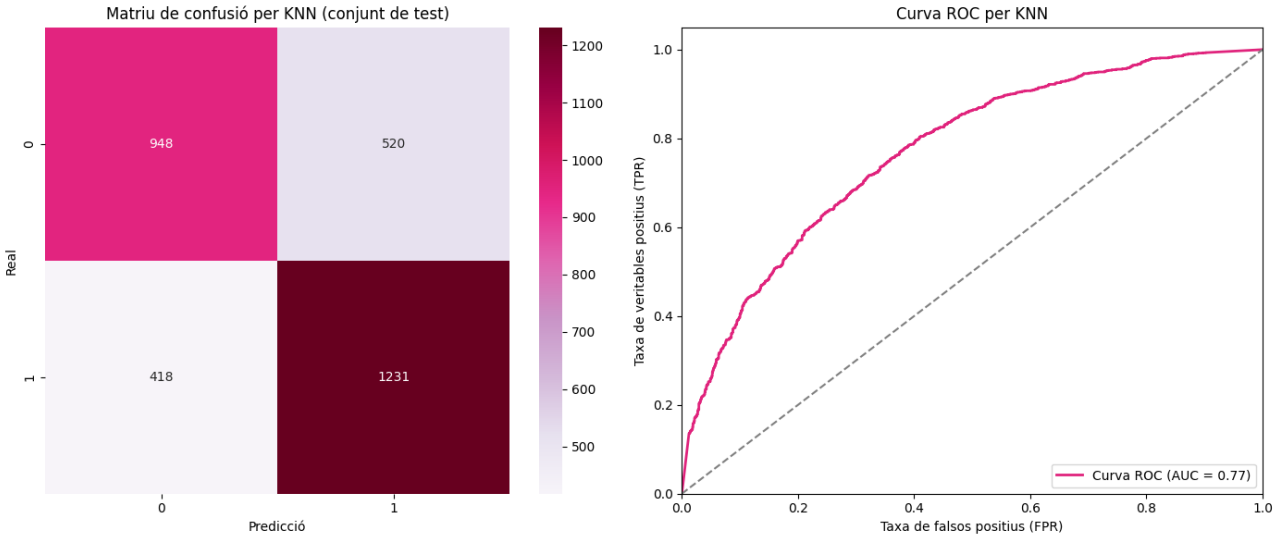

Figura14: Matriu de confusió i curva ROC del conjunt de test

Aquests resultats indiquen que el model és capaç de mantenir un bon rendiment quan s'aplica a dades no vistes, confirmant que el procés d'optimització i entrenament ha estat adequat.

5.1.3. Conclusió del Model

El model KNN amb els hiperparàmetres seleccionats mostra una bona capacitat per capturar patrons en les dades. Tot i no ser el model més sofisticat, la seva simplicitat, transparència i eficàcia en aquest escenari fan que sigui una elecció robusta i fàcil d'interpretar.

Les mètriques obtingudes en validació i test són molt consistents, amb una precisió i recall equilibrats entre les classes i un AUC de 0.77 que demostra una bona separabilitat entre les categories.

5.2. Model d'Arbre de Decisió

L'arbre de decisió és un model supervisat que estructura el procés de classificació o regressió com una sèrie de preguntes binàries jeràrquiques. Aquest mètode és especialment útil perquè proporciona una interpretabilitat directa, ja que es pot observar quines variables són més importants en la presa de decisions i com es fan aquestes decisions.

5.2.1. Selecció d'hiperparàmetres

Per optimitzar el model, s'ha realitzat una cerca en graella explorant diferents combinacions d'hiperparàmetres:

```
tree_params = {
   'max_depth': [3, 5, 7, 10, 15, 20, None],
   'min_samples_split': [2, 3, 5, 10, 20],
   'min_samples_leaf': [1, 2, 4, 5, 10, 20],
   'criterion': ['gini', 'entropy'],
   'splitter': ['best', 'random']
}
```

Després de l'exploració, la millor configuració trobada ha estat la que té criterion = entropy, max_depth = 7, min_samples_leaf = 20, min_samples_split = 3 i splitter = best. Aquesta configuració ha proporcionat una precisió mitjana de 0.6872 ± 0.0139 en les validacions creuades.

5.2.2. Avaluació del model en validació

Amb els hiperparàmetres seleccionats, s'ha aplicat el model al conjunt de validació i els resultats obtinguts són els següents:

Mètriques per al conjunt de validació:

	precision	recall	f1-score	support
0	0.72	0.52	0.60	1768
1	0.66	0.82	0.73	2006
accuracy			0.68	3774
macro avg	0.69	0.67	0.67	3774
weighted avg	0.69	0.68	0.67	3774

Figura15: Mètriques del conjunt de validació

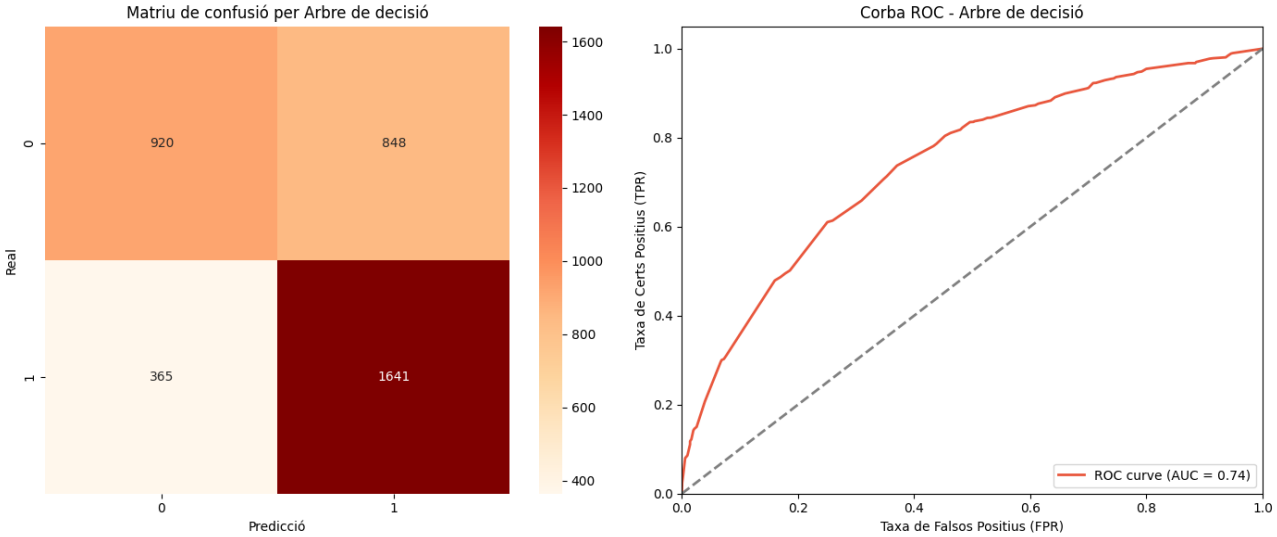

Figura16: Matriu de confusió i curva ROC del conjunt de validació

Aquestes mètriques mostren un rendiment equilibrat, amb un recall més alt per a la classe 1, fet que indica que el model és més sensible a identificar veus sintètiques, però presenta certa limitació en el reconeixement de veus reals.

5.2.3. Avaluació del model en test

Per comprovar la capacitat del model de generalitzar a noves dades, s'ha avaluat amb el conjunt de test obtenint els següents resultats:

Mètriques per al conjunt de test:

	precision	recall	f1-score	support
0	0.74	0.53	0.62	1468
1	0.67	0.84	0.74	1649
accuracy			0.69	3117
macro avg	0.71	0.68	0.68	3117
weighted avg	0.70	0.69	0.68	3117

Figura17: Mètriques del conjunt de test

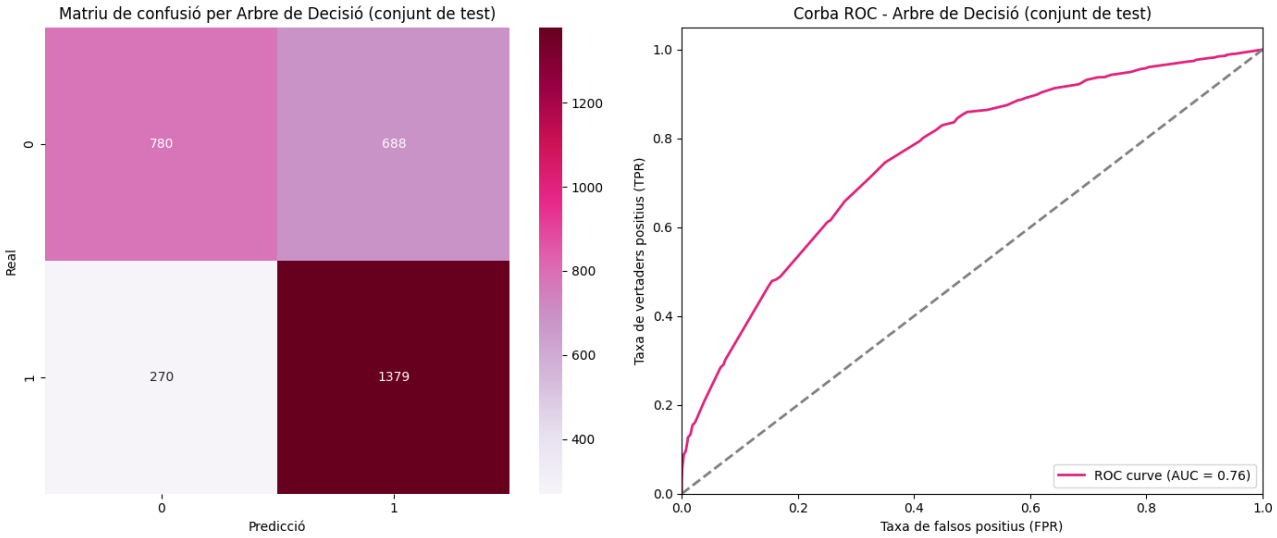

Figura18: Matriu de confusió i curva ROC del conjunt de test

Els resultats en el conjunt de test són molt similars als de validació, la qual cosa confirma que el model no està sobreajustat i pot aplicar-se amb confiança a noves dades.

5.2.4. Conclusió del Model

L'arbre de decisió amb la configuració òptima ha aconseguit un rendiment consistent en els conjunts de validació i test, amb una precisió i recall equilibrats. La seva interpretabilitat directa i la capacitat de capturar relacions no lineals fan que sigui una opció atractiva en aquest projecte. Tot i que el seu AUC (0.76) és lleugerament inferior al del model KNN, continua sent un candidat robust per a la classificació de veus reals i falses.

5.3. Model SVM

El Support Vector Machine (SVM) és un mètode supervisat que tracta de trobar un hiperplà que separi òptimament les classes. Aquesta separació es fa maximitzant el marge entre els veins més propers de cadascuna de les classes i l'hiperpla de decisió, fet que pot aportar un alt rendiment i robustesa davant dades amb distribucions complexes. A més, resulta especialment útil quan les dades es poden separar de forma no lineal, gràcies a l'ús de nuclis kernel com el radial basis function rbf.

5.3.1. Selecció d'hiperparàmetres

Per optimitzar el SVM, s'ha dut a terme una cerca en graella amb els següents hiperparàmetres:

```
svm_params_optimized = {
    'C': [0.1, 1, 10],
    'kernel': ['linear', 'rbf'],
    'gamma': ['scale']
}
```

La millor combinació obtinguda ha estat amb C = 10, kernel = rbf i gamma = scale I ha aconseguit una precisió mitjana de 0.7098 en la validació creuada, posicionant el SVM com un dels millors models quant a rendiment global.

5.3.2. Avaluació del model en validació

Amb els hiperparàmetres trobats, hem aplicat el SVM al conjunt de validació. Els resultats són:

Mètriques per al conjunt de validació:

	precision	recall	f1-score	support
0	0.73	0.67	0.70	1768
1	0.73	0.78	0.75	2006
accuracy			0.73	3774
macro avg	0.73	0.73	0.73	3774
weighted avg	0.73	0.73	0.73	3774

Figura19: Mètriques del conjunt de validació

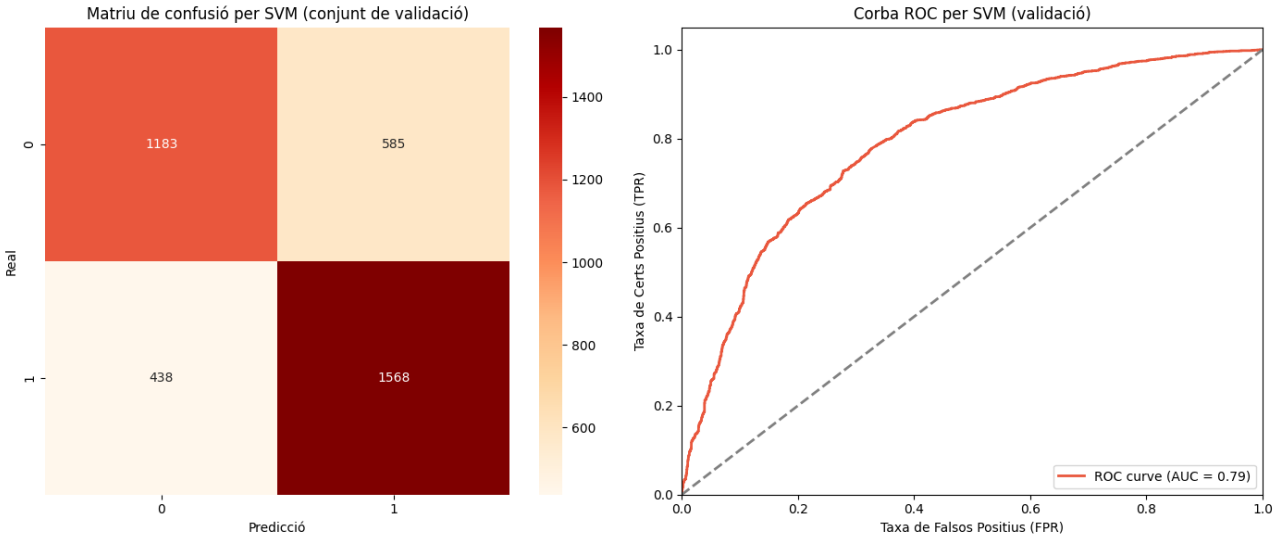

Figura20: Matriu de confusió i curva ROC del conjunt de validació

Aquestes mètriques indiquen que el model és capaç d'identificar adequadament tant la classe 0 com la classe 1, amb un bon equilibri entre precision i recall i una separació significativa entre categories gràcies a un AUC de 0.79.

5.3.3. Avaluació del model en test

S'ha aplicat el model també al conjunt de test per confirmar la seva capacitat de generalització:

Mètriques per al conjunt de test:

	precision	recall	f1-score	support
0	0.73	0.66	0.70	1468
1	0.72	0.79	0.75	1649
accuracy			0.73	3117
macro avg	0.73	0.72	0.73	3117
weighted avg	0.73	0.73	0.73	3117

Figura21: Mètriques del conjunt de test

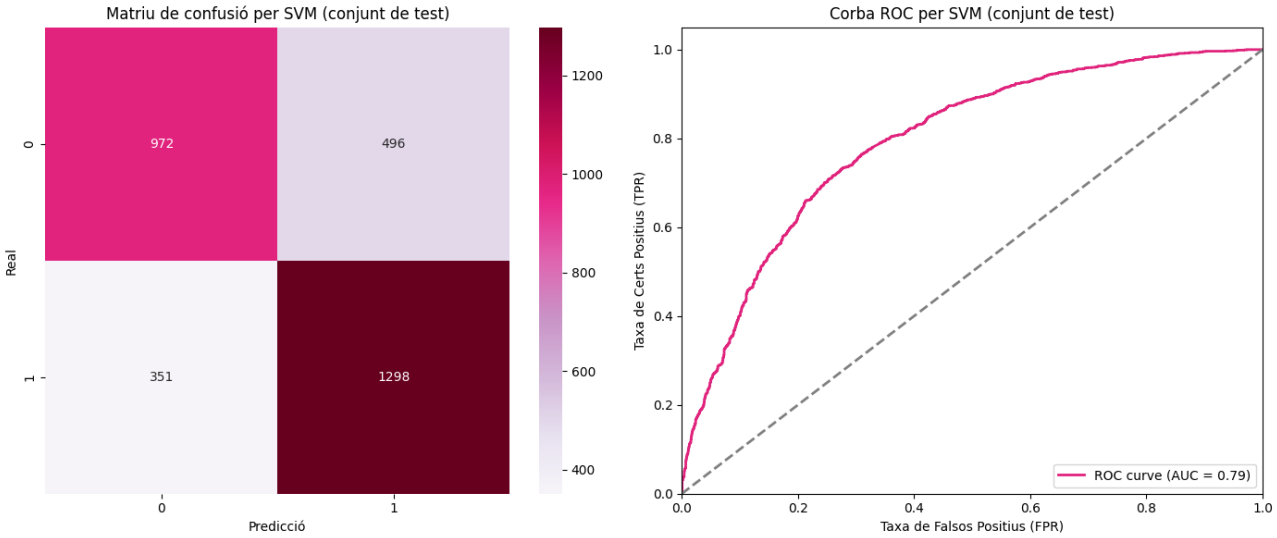

Figura22: Matriu de confusió i curva ROC del conjunt de test

El rendiment en test és pràcticament idèntic al de validació, demostrant estabilitat i robustesa. Ambdues classes són reconegudes amb força equilibri i l'AUC de 0.79 mostra que el model és capaç de separar adequadament les veus reals i falses en un espai de característiques altament no lineal.

5.3.4. Conclusió del Model

El SVM amb els hiperparàmetres òptims seleccionats obté un rendiment consistent i equilibrat entre les classes, tal com mostren les mètriques de validació i test, amb una accuracy de 0.73 i un AUC de 0.79 en ambdós conjunts.

6. Selecció de Model

L'objectiu principal del treball és identificar el model que ofereixi el millor rendiment per classificar veus com a reals o falses. Per fer-ho, s'han avaluat tres models diferents: KNN, arbre de decisió i SVM, utilitzant les mateixes dades preprocessades i optimitzant els seus hiperparàmetres per garantir una comparació justa, tal i com s'ha explicat el l'apartat anterior de l'informe.

6.1. Comparació dels models

L'avaluació de cada model ha tingut en compte diverses mètriques, com ara Accuracy, F1-Score, precisió, Recall i l'AUC-ROC. Aquestes mètriques permeten analitzar no només l'encert global del model, sinó també la seva capacitat per discriminar entre les classes 0 (veu real) i 1 (veu falsa). Els resultats són els següents:

- KNN: Accuracy de 0.69, AUC-ROC de 0.77.
- Arbre de decisió: Accuracy de 0.68, AUC-ROC de 0.74.
- SVM: Accuracy de 0.73, AUC-ROC de 0.79.

La comparació mostra que el model SVM supera clarament els altres en totes les mètriques considerades, destacant especialment en l'AUC-ROC (0.79), que reflecteix una excel·lent capacitat per discriminar entre classes. Aquest fet fa que SVM sigui seleccionat com el millor model per a aquest treball.

6.2. Aplicació del millor model al conjunt de test final

Un cop seleccionat el model SVM, s'ha aplicat sobre el conjunt de test final (test_csv) per avaluar-ne el rendiment definitiu. Abans d'aplicar el model, s'ha preprocessat el conjunt test_csv seguint exactament els mateixos passos que amb els conjunts de validació i test interns.

Els resultats obtinguts amb aquest conjunt de test són els següents:

Mètriques per al conjunt de test:

	precision	recall	f1-score	support
0	0.61	0.57	0.58	985
1	0.65	0.69	0.67	1175
accuracy			0.63	2160
macro avg	0.63	0.63	0.63	2160
weighted avg	0.63	0.63	0.63	2160

Figura23: Mètriques del conjunt de test

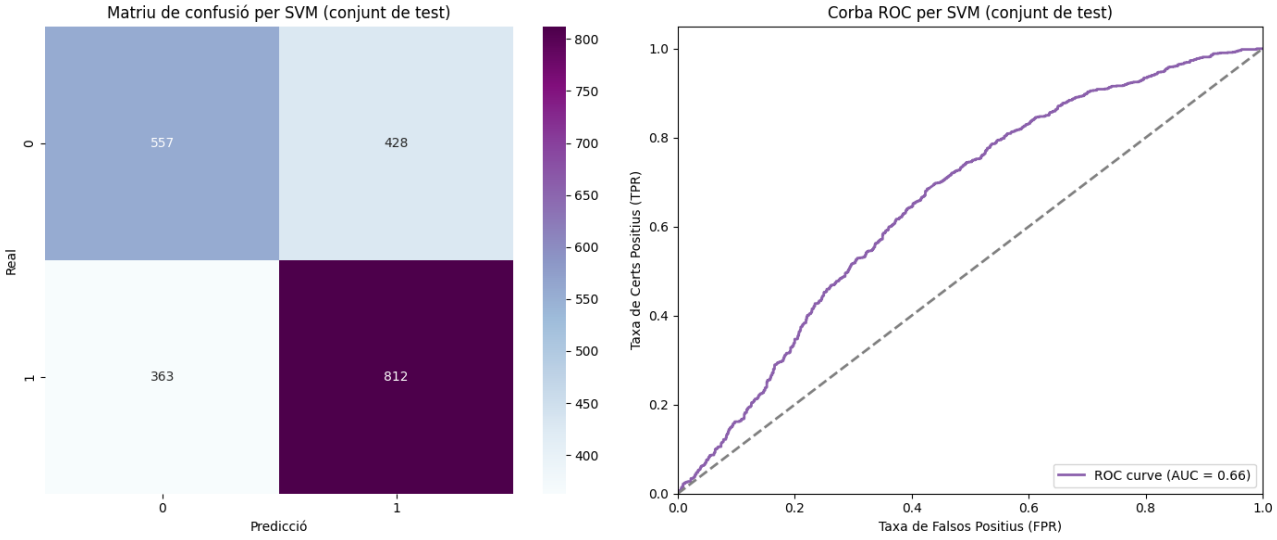

Figura24: Matriu de confusió i curva ROC del conjunt de test

Encara que l'Accuracy i l'AUC-ROC són lleugerament inferiors al rendiment obtingut en els conjunts de validació i test interns, el model SVM manté una capacitat acceptable de generalització en el conjunt final de dades. No obstant això, la reducció considerable de l'AUC-ROC en el test final (0.66) suggereix que el model no està capturant tota la complexitat del conjunt de dades original.

Per aquesta raó, s'ha decidit tornar a entrenar els models, aquest cop utilitzant totes les variables numèriques disponibles en lloc de limitar-se als cinc components principals obtinguts amb el PCA. Aquesta decisió es basa en la hipòtesi que la informació perduda en la reducció dimensional pot ser crucial per millorar la capacitat del model per generalitzar a nous conjunts de dades, especialment en el test final.

7. Implementació dels Models sense PCA

7.1. Model KNN

Amb les mateixes combinacions potencials d'hiperparàmetres, en aquest cas s'han seleccionat com a millor combinació: metric = manhattan, n_neighbors = 10, weights = distance. Aquesta configuració ha aconseguit una precisió mitjana de 0.7863 en la validació creuada, per tant, els resultats mostren una millora clara respecte als obtinguts anteriorment utilitzant el PCA.

Resultats en validació

- Amb PCA: AUC-ROC = 0.77; Accuracy = 0.69
- Sense PCA: AUC-ROC = 0.87; Accuracy = 0.79

El model sense PCA demostra un augment significatiu en la capacitat de separació entre les classes, reflectit per l'increment de l'AUC-ROC. També es percep una millor precisió i recall, especialment en la classe `, la qual cosa és crucial per l'objectiu d'identificar casos sintètics.

Resultats en test

- Amb PCA: AUC-ROC = 0.77; Accuracy = 0.70
- Sense PCA: AUC-ROC = 0.87; Accuracy = 0.78

Els resultats en el conjunt de test segueixen la mateixa tendència. El model sense PCA manté la seva capacitat de generalització, mostrant millores tant en la separabilitat entre classes com en les mètriques globals de rendiment.

7.1.1. Conclusió

Els resultats obtinguts sense PCA evidencien que treballar amb les dades originals permet una millor explotació de la informació continguda en les variables. L'absència de la reducció de dimensionalitat ha millorat notablement les mètriques de rendiment del model KNN.

7.2. Model d'Arbre de Decisió

Utilitzant les mateixes combinacions d'hiperparàmetres que es van utilitzar amb el conjunt reduït amb PCA, s'ha obtingut com a combinació òptima d'hiperparàmetres la que té criterion = gini, max_depth = 10, min_samples_leaf = 10, min_samples_split i splitter = best. Aquesta configuració ha aconseguit una precisió mitjana de 0.7624 en la validació creuada, destacant una millora respecte als resultats obtinguts amb PCA.

Resultats en validació

- Amb PCA: AUC-ROC = 0.74; Accuracy = 0.68

- Sense PCA: AUC-ROC = 0.84; Accuracy = 0.76

El model sense PCA presenta una millora significativa en la capacitat de separar les classes, com es reflecteix en l'increment de l'AUC-ROC i la millora de l'accuracy. Els increments en el recall per a la classe ` són especialment notables, fet que reforça la capacitat del model per detectar correctament casos de veu sintètica.

Resultats en test

- Amb PCA: AUC-ROC = 0.74; Accuracy = 0.69
- Sense PCA: AUC-ROC = 0.84; Accuracy = 0.77

Els resultats en el conjunt de test també mostren millores consistents. El model sense PCA manté un bon rendiment en termes de generalització, evidenciant la robustesa de l'aproximació utilitzada.

7.2.1. Conclusió

Els resultats obtinguts amb el model d'arbre de decisió sense PCA demostren que treballar amb totes les variables originals millora significativament el rendiment del model. L'absència de reducció dimensional permet una explotació més completa de les dades disponibles, augmentant la separabilitat entre classes i millorant mètriques com l'accuracy i l'AUC-ROC.

7.3. Model SVM

En treballar amb totes les variables originals sense reducció dimensional amb PCA, s'ha aplicat de nou el procés d'optimització d'hiperparàmetres. La millor configuració obtinguda ha estatla mateixa que s'ha obtingut amb el PCA, però ara amb una precisió mitjana en la validació creuada de 0.8297, indicant una millora notable respecte als resultats amb PCA.

Resultats en validació

- Amb PCA: AUC-ROC = 0.79; Accuracy = 0.73
- Sense PCA: AUC-ROC = 0.90; Accuracy = 0.82

El model SVM sense PCA mostra una millora clara en totes les mètriques rellevants. Destaca especialment l'AUC-ROC, que indica una capacitat més gran per separar les classes. Les millores en el recall per a la classe 1 són significatives, fet que millora la identificació de casos de veu sintètica.

Resultats en test

- Amb PCA: AUC-ROC = 0.79; Accuracy = 0.73
- Sense PCA: AUC-ROC = 0.90; Accuracy = 0.83

Els resultats en el conjunt de test mantenen aquesta tendència, consolidant el rendiment del model. L'AUC-ROC de 0.90 en test reflecteix una excel·lent capacitat de generalització i reforça l'estratègia d'utilitzar totes les variables originals per a l'entrenament.

5.3.1. Conclusió

El model SVM sense PCA demostra un rendiment superior respecte a la versió amb reducció dimensional. L'ús de totes les variables originals permet extreure més informació i millorar les capacitats predictives del model, especialment en termes de separabilitat entre classes i en el rendiment global de les mètriques.

7.4. Aplicació del millor model al conjunt de Test Final

Després d'analitzar els resultats dels diferents models entrenats, s'ha seleccionat l'SVM com el millor model per abordar aquest problema, destacant per la seva gran capacitat de generalització i les seves mètriques competitives tant en els conjunts de validació com en els de test.

L'AUC-ROC de 0.80 obtingut en el test final reflecteix una bona capacitat de separació entre les classes Real i Sintètic, representant una millora significativa respecte a l'AUC-ROC de 0.66 aconseguit amb PCA. Això demostra que l'ús de totes les variables originals permet una explotació més completa i efectiva de la informació, la qual cosa es tradueix en una classificació més precisa i fiable.

Aquesta diferència es deu probablement al fet que PCA redueix la dimensionalitat perdent part de la variància explicada i, en conseqüència, informació rellevant per a la classificació. Encara que PCA pot ser útil per evitar redundàncies i/o oberfitting, en aquest cas, el treball amb les variables originals ha demostrat ser molt més eficient per capturar patrons i diferències importants entre les classes.

El cost d'aquest enfocament és un lleuger increment en el temps de processament, ja que utilitzar totes les variables implica més càlculs i memòria computacional. No obstant això, els beneficis obtinguts en termes de precisió, recall i capacitat de separació entre classes compensen àmpliament aquest cost.

En conclusió, l'elecció de treballar amb totes les variables originals s'ha mostrat com una decisió superior en aquest context, garantint una millor explotació de les dades i un rendiment notablement més alt en els conjunts de validació i test.

8. Model Card

8.1. Informació General

- Persona responsable del desenvolupament del model: Paula Justo Gili
- Data de Creació del model: 28 de Desembre de 2024
- Versió: 1.0

Aquest model ha estat desenvolupat per classificar mostres com a Real o Sintètic. L'objectiu principal és identificar de manera fiable mostres sintètiques generades mitjançant tècniques d'intel·ligència artificial per a possibles aplicacions futures com la detecció de l'autenticació de veu.

8.2. Context i Objectius del Model

- Objectiu específic: El model està dissenyat per detectar si una mostra d'àudio prové d'una veu real o d'un sistema sintètic generat mitjançant tècniques com TTS o StarGAN. Això és especialment útil en aplicacions d'autenticació, seguretat i detecció de contingut manipulat.
- Domini d'aplicació:
  - Sectors: Seguretat, mitjans de comunicació, recerca en sociolingüística.
  - Limitacions: No recomanat per àudios de baixa qualitat o dominis no representats en les dades d'entrenament.

8.3. Descripció del Model

- Tipus de Model: SVM (Support Vector Machine) amb kernel radial bàsic (RBF).
- Arquitectura:
  - Hiperparàmetres:
    - C: 10.0
    - kernel: RBF
    - gamma: Scale
- Llenguatge i Eines Utilitzades: Python amb Sacikit-learn

8.4. Dades Utilitzades

- Descripció del Conjunt de Dades:
  - Inclou característiques acústiques com `mfcc_mean` i `loudness_mean`, i metadades com sexe i país.
  - Categorització de les variables categòriques `Combined_Sex` i `Combined_Country` mitjançant One-Hot-Encoding.
- Distribució de les dades:
  - Proporció de classes: 54.44% classe 1 (Real) i 45.60% classe 2 (Sintètic)

    - Mostres totals: Entrenament (80%), Test (20%)
        - Partició del conjunt d'Entrenament: Entrenament (70%), Validació (30%)
- Preprocessament:
    - Imputació de missings amb la mediana.
    - Normalització logarítmica (per a les variables amb distribució exponencial) i normalització amb StandardScaler (per la resta de variables)per garantir que totes les variables estiguin en la mateixa escala.
- Orígenes de les Dades: Dades recopilades a partir d'un corpus de veus reals i sintètiques.

8.5. Mètriques d'Avaluació

- Mètriques Utilitzades:
    - Accuracy
    - Precision
    - Recall
    - F1-Score
    - AUC-ROC
- Resultats sense PCA:
    - Validació: AUC-ROC = 0.90; Accuracy = 82%
    - Test Final: AUC-ROC = 0.80; Accuracy = 80%
- Comparació amb PCA i sense PCA:
    - Amb PCA: AUC-ROC = 0.66; Accuracy = 70%
    - Sense PCA: AUC-ROC = 0.80; Accuracy = 80%

8.6. Anàlisi de Rendiment

- Capacitat de Generalització:
    - El model sense PCA mostra una millora substancial en la separació de classes i una millor explotació de les dades originals.
- Cost Computacional:
    - Sense PCA: temps d'entrenament lleugerament superior.
- Escenaris d'Error:
    - Baixa efectivitat en mostres de baixa qualitat o dominis geogràfics poc representats.

8.7. Anàlisi de Biaix i Equitat

- Biaixos Potencials:
    - Sexe i país són variables que podrien introduir biaixos. S'han equilibrat les dades, però caldria més anàlisi de precisió per subgrups.

- Estratègies de Mitigació:
  - Validació creuada estratificada.

8.8. Robustessa i Explicabilitat

- Explicabilitat:
  - Mètodes com SHAP o LIME es podrien utilitzar per entendre millor les prediccions.
- Robustesa a Condicions Adverses:
  - No avaluat explícitament, però s'espera un rendiment inferior en condicions d'àudio amb soroll o degradació.

8.9. Comparació amb Models Alternatius

- Comparativa amb PCA:
  - L'absència de PCA permet millorar considerablement les mètriques sense sacrificar consistència.
- Altres Models Considerats: KNN i arbres de decisió van ser menys precisos en AUC-ROC i generalització.

8.10. Conclusió

- Millores Sense PCA:
  - Increment notable en AUC-ROC (de 0.66 a 0.80) i Accuracy (de 70% a 80%).
- Consideracions de Temps:
  - L'ús de totes les variables implica un cost computacional lleugerament superior, però la millora en rendiment ho justifica plenament.

Aquest model, optimitzat sense reducció dimensional, demostra ser una eina robusta i fiable per a la classificació d'àudios com a reals o sintètics.

8.11. Recomanacions d'Ús i Manteniment

- Guia d'Ús: Utilitzar dades preprocessades de manera consistent amb l'entrenament.
- Actualitzacions: Retunejar preiòdicament amb dades noves per mantenir la rellevància.

8.12. Referències i Anexos

- Sacikit-learn: https://scikit-learn.org
- OpenSMILE: https://audeering.github.io/opensmile
- Codi complet en el notebook

9. Bonus 1

Per abordar el problema de classificació binària i millorar la interpretabilitat del model, s'ha implementat un Explainable Boosting Machine (EBM). Aquest model, basat en la suma de models individuals interpretables, combina precisió amb transparència, fent-lo especialment adequat per tasques on la comprensió de les decisions del model és fonamental.

9.1.1. Rellevància del Model

L'EBM es diferencia d'altres models perquè permet entendre l'impacte de cada característica en les prediccions. Això és especialment útil en àmbits sensibles, com l'anàlisi de dades sociolingüístiques, on la interpretabilitat i la transparència són imprescindibles per evitar biaixos o errors en la interpretació dels resultats.

9.1.2. Hiperparàmetres i Temps d'Entrenament

El procés d'optimització d'hiperparàmetres per aquest model requereix un temps considerable, oscil·lant entre 10 i 15 minuts per identificar la combinació òptima, degut a la seva naturalesa iterativa i al volum de dades.

Els hiperparàmetres ajustats són:

```
ebm_params_reduced = {
   'max_bins': [128],
   'learning_rate': [0.01, 0.05],
   'min_samples_leaf': [2, 10],
   'max_leaves': [3, 5]
}
```

La millor configuració trobada és la que té learning_rate = 0.05, max_bins = 128, max_leaves = 3 I min_samples_leaf = 10. Amb aquesta configuració, s'ha obtingut una precisió mitjana de 0.8219 ± 0.0041 en validació creuada.

9.1.3. Resultats Obtinguts, Interpretació i Conclusió

La curva ROC amb un AUC de 0.90 evidencia la gran capacitat del model per distingir entre mostres reals i sintètiques. Tot i que l'entrenament és més lent comparat amb altres models, l'EBM compensa aquest desavantatge amb una interpretabilitat superior i un rendiment robust.

A més l'EBM proporciona informació clau sobre com cada característica afecta les prediccions, cosa que permet identificar possibles biaixos i assegurar que les decisions siguin transparents i justificades. Aquest model és una opció excel·lent per avaluar amb precisió el problema, mantenint la confiança en els resultats gràcies a la seva naturalesa explicable.