Predicció del gènere musical

Jorge Giménez, Juan Carlos Soriano i Roger Boadella

1. Introducció

Els experts de la música sempre han intentat saber que diferencia cada gènere i per això s'han creat aquest dataset.

En aquest problema existeixen 2 datasets amb tot tipus d'atributs sobre la peça musical i el gènere d'aquesta. El "base" seria l'anomenat "DF30". Aquest Dataset té totes les dades recollides fent una mitjana de 30 segons de cada fragment d'audio.

El Dataset DF3 té aquests 30 segons repartits per 10 instàncies diferents, on a cada instancia hi ha la mitja de fragments de 3 segons.

Per tant, si el primer dataset té 1000 files (instàncies), el segon dataset en té (9990).

L'objectiu per tant seria decidir a quin gènere pertany.

2. EDA

2.1 Exploració de les dades

Es pot veure com al dataset dels fragments de 3 segons hi són tots els atributs en format decimal, menys la primera columna que ens indica el nom de l'arxiu d'àudio al qual pertany la instancia i l'últim que conté la etiqueta que indica el gènere real del fragment.

Com a possibles gèneres hi ha: Blues, Classical, Country, Disco, Hip Hop, Jazz, Metal, Pop, Reggae, Rock. Tots ells es troben igualment distribuïts com es pot veure a la Figura 1.

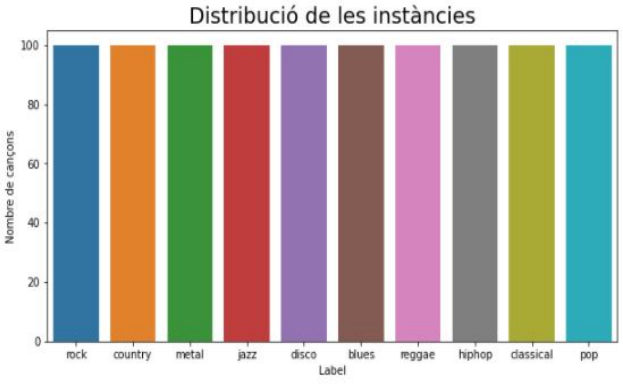

Figura 1: Distribució de les instàncies.

Per sort els datasets no tenen cap valors nul, això facilitarà el treball més endavant alhora d'entrenar models.

Un punt que s'ha de destacar en la observació de les dades és què s'ha localitzat un arxiu d'àudio pertanyent a la categoria de jazz que sembla que estigui corrupte. Aquest arxiu és `jazz.00054.wav`, que al voler tractar amb aquest ens genera un error d'accés a les seves dades per tant s'ha substituït per un altre del mateix gènere.

Quan s'observa quina correlació poden tenir els diferents atributs entre ells, si ens fixem en el dataset que se està estudiant, a la majoria de dades existeixen 2 grups que es refereixen al

mateix atribut: var (variància) i mean (mitjana). En aquest cas la mitjana és el valor que més interessant ja que dóna molta més informació de les dades a analitzar.

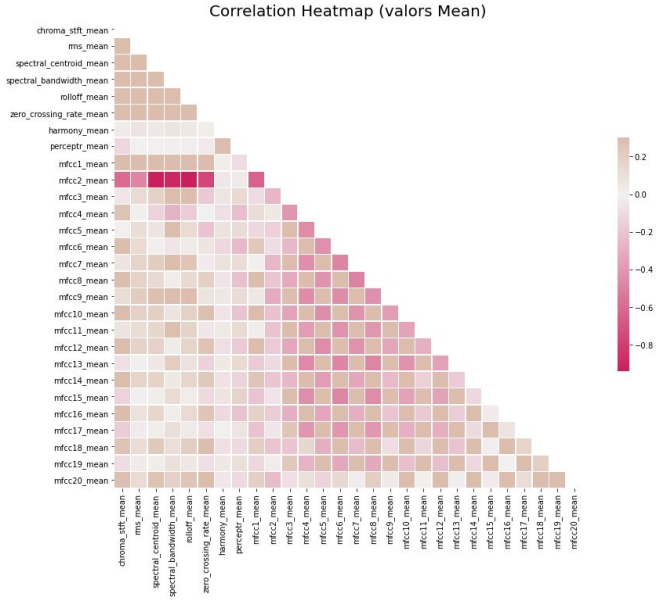

Figura 2: Correlació entre els atributs mean.

De la taula que veiem en la FIgura 2 podem concloure que hi ha una relació inversa molt gran entre el coeficient 2 del MFCC (Mel Frequency Cepstrum Coefficients), el qual veurem després, i els atributs que no pertanyen a aquest coeficient de Mel com poden ser el Zero Crossing, l'espectre del so, etc.

També es pot observar com tots aquest atributs que no tenen res a veure amb el coeficient de Mel tenen correlació positiva entre ells.

2.2 Entenent l'àudio

Per entendre millor el Dataset, ens fixarem en un arxiu en concret i veurem com és la seva representació visual en 2D. En aquest l'arxiu és del gènere metal. Com es veu a la Figura 3 quan el so de la bateria és molt notori coincideix amb els pics molt característics a la gràfica d'ones, també generarem l'espectograme per obtenir més informació que ajudarà a comparar les dades.

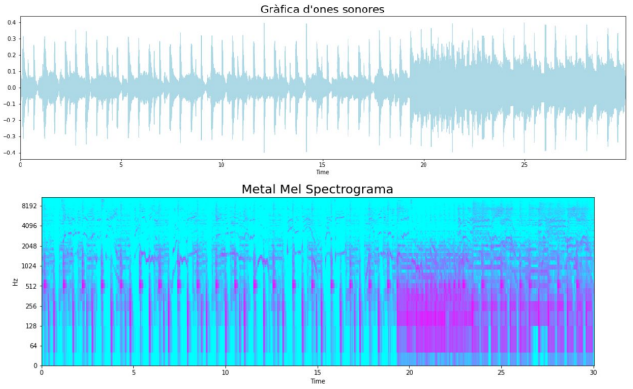

Figura 3: Gràfica de les ones sonores gènere "metal" i el seu espectrograma.

Si ara comparem la gràfica d'ones sonores i l'espectrograma d'un arxiu d'un gènere diferent es podrà apreciar una gran diferencia. Per el cas de la Figura 4 utilitzem un arxiu de blues.

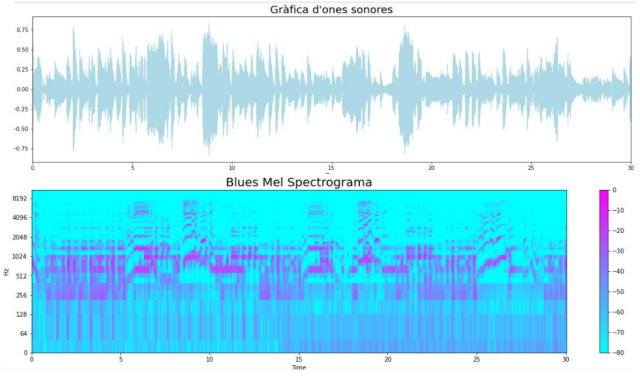

Figura 4: Gráfica d'ones sonores i espectrograma generat per un blues.

Una mesura que pot arribar a ser molt útil per tal de classificar quin tipus de música s'està estudiant és el Zero-Crossing. Aquesta és una mesura molt simple que es basa en comptar quantes vegades canvia la gràfica d'ones sonores entre negatiu i

positiu, per veure-ho amb detall cal fer zoom. A la Figura 5 es veu amb detall això mateix.

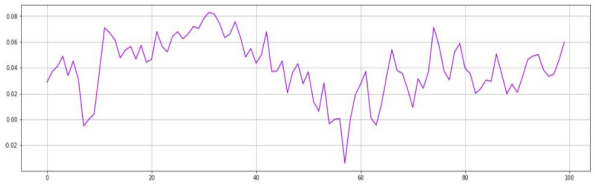

Figura 5: Demostració del Zero-Crossing.

Aquesta mesura ens ajuda a detectar sons més estridents i bruscos, on es canvia la freqüència molt freqüentment.

Si ara calculem els Zero-Crossings de tots els gèneres podrem veure perquè aquesta mesura ens pot ajudar a diferenciar-los, on a més valor de zero crossings més estrident i brusca serà la cançó del gènere, en canvi contra menys valor d'aquest valor la música serà més calmada i sense tants altibaixos.

Amb aquesta mesura podem classificar com a més estridents els gèneres metal, pop i disco amb uns valors màxims de 97k Zero Crossings i com més calmades blues, classical i jazz amb 51k com podem veure a la FIgura 6.

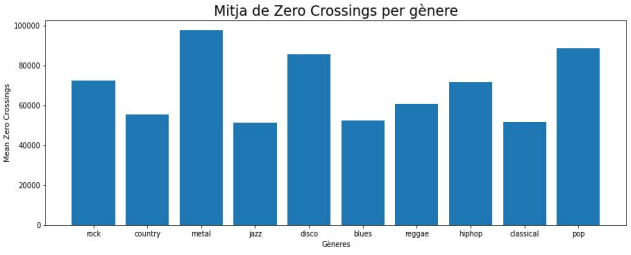

Figura 6: Zero-Crossings per gènere.

Una altra una eina molt potent és el Chromagrama, analitza la música i marca els pitches, quant més vermell és mostra en el chromagrama més alt és el valor del pitch. Aquesta eina ens ajuda a visualitzar la "forma de la música" i tenir una representació més visual del que hi ha en un arxiu d'àudio. Si es torna a comparar el metal i el blues es veu que el metal té una distribución heterogénea mentres que el blues és més homogeni.

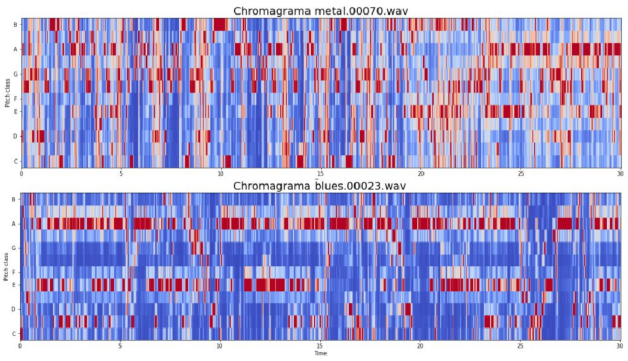

Figura 7: Chromagrama, Primer metal, Segon Blues.

Un altre atribut que ja ve donat és el BPM (Beats per Minute), que pot ser molt útil de cara a comparar els diferents gèneres de música entre si. Segons fonts especialitzades en música, una bona manera de classificar els diferents gèneres de música és diferenciar les seves pulsacions per minut, analitzant així el ritme de la música.

Hi ha certs estàndards en aquestes mètriques i si es comparen amb les del Dataset es pot veure que la majoria entren dins d'aquests estàndards.

	BPM estàndar	BPM mean (dataset)
Blues	60-80	124
Classical	120-140	124
Country	120-130	117
Disco	115-130	125
HipHop	85-115	107
Jazz	120-125	111
Metal	100-160	125
Pop	100-130	111
Reggae	60-90	129
Rock	110-140	123

Figura 8: BPM estàndard i del Dataset.

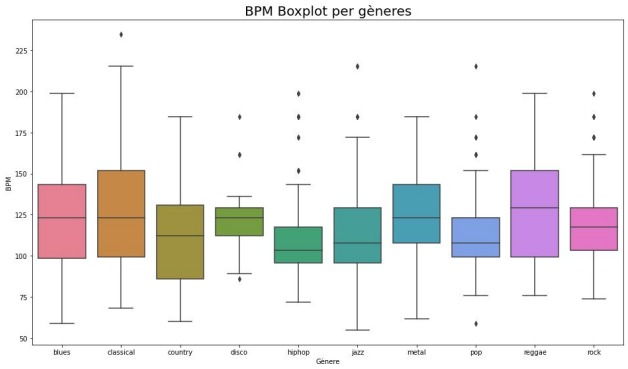

Figura 9: Distribució per gènere dels bpm en el Dataset.

3. Estat de l'art

Altres persones ja treballat amb aquestes mateixes dades, aquelles que s'han centrat en predir el gènere musical han obtingut millors resultats amb machine learning. Per altra banda també s'han centrat en modificar la música i en generar nova música.

4. Preprocessing i Mètode

3.1 Dataset i neteja de dades

A partir d'aquí es treballarà amb el dataset de 30 segons, el de 3 segons és interessants, ja que tenen en compte el pas del temps i aporten una informació segurament crucial per a molts models d'aprenentatge computacional més avançats (potser xarxes neuronals LSTM, Markov, etc...).

No obstant això, en aquest moment és preferible utilitzar el segon dataset, les des de que fan "sample" dels atributs cada 30s, ja que el volum de dades és més viable per a nosaltres i creiem que el nostre nivell actual i els models que coneixem suficientment bé a hores d'ara no aprofitarien massa el volum de dades de l'altre dataset.

Primerament, ens desfarem de la primera columna, "filename", ja que no ens aporta informació que vulguem utilitzar. Inclús es podria fer "trampa" amb aquesta columna a l'hora de predir el gènere del tros d'àudio així que la retirem de les dades. De la mateixa manera, ens desfem de la variable "length" que té el mateix valor per a tots els registres ja que tots els samples son de 30s.

3.2 Entrenament de models

Realitzarem un Kfold amb K=10 ja que no tenim masses dades i a la vegada optimitzarem els hiperparàmetres dels models per poder fer la comparació de les versions amb millor rendiment de cada un dels models. Per fer això utilitzarem els següents mètodes:

- Regressió Logística
- KNN
- SVM Kernel radial
- Random Forest
- Adaboost Decision Tree
- Adaboost Regressió Logística

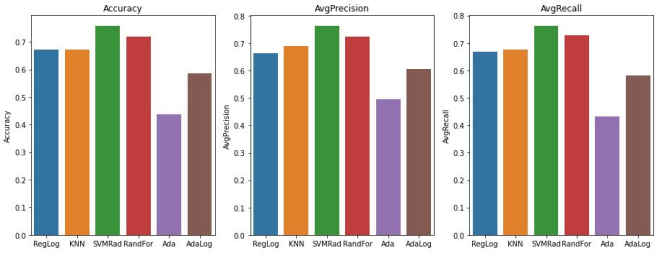

Figura 10: Resultats obtinguts.

Com es pot veure a la Figura 10 el mètode que dona millors resultats (accuracy: 0,758, precision: 0,763, recall: 0,762) és la

màquina de vectors, també s'han un tingut els hyperparameters C i gamma amb valors 23,5723 i 0,00434 respectivament.

5. Conclusions i Resultats

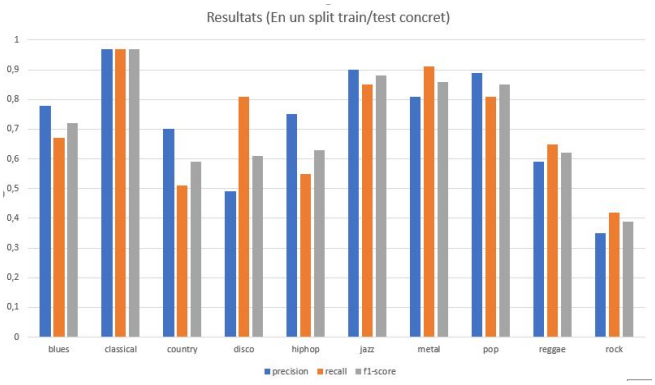

Figura 11: Exemple de resultats.

En general, podem dir que en grans trets no hem obtingut uns resultat sorprenents, i es que com veiem a l'apartat de millores, estem segur que altre simplementacions i models de classifiació seran molt més prometedors. No obstant, estem bastant contents i som conscients de que si un resultat és bó o dolent depen en la majoría de les vegades de quin ús se li vulgui donar al model o on es vol implementar.

El més rellevant és que a la figura 11, podem veure un exemple de resultats (un split de train i test concret) el nostre model és capaç de classificar amb un bon rendiment molts dels gèneres (classical, jazz) però falla força en alguns altres (rock, disco…). Això és així perquè no aconsegueix fer una diferenciació suficientment gran entre aquests gèneres com per a poder fer una bona classificació.

De totes maneres, aquests resultats són esperançadors ja que vol dir que si aconseguissim dissenyar una millor pipeline per al processament de dades o una bona combinació de models per a poder classificar amb més detall determinats gèneres segurament s'aconsegueixen resultats encara més satisfactoris.

6. Futures millores

Es podria millorar els models creats utilitzant el dataset de fragments de 3 segons, però creiem que aquest Dataset és més adequat per a models d'aprenentatge computacional més avançats com podrien ser les xarxes neuronals o Markov.

Una altra millora podria ser incorpora el Zero-Crossing al Dataset ja que aquest ens dona el Zero-Crossing per longitud pero no el total i creiem que podria ser una millora substancial a l'hora de crear el model.