# 1 Traffic Light Panel

Un cop analitzats els resultats del profiling clàssic, hem aprofitat la interpretabilitat dels *Traffic Light Panel*s (TLPs), amb tal de poder mostrar les importpàncies i els papers que juguen les variables en cadescuna de les diferents 5 clases generades al clustering jeràrquic (tal i com s'ha explicat a la secció **??**) a una persona que no estigui familiaritzada amb el àmbit de la estadística, com podria ser la persona que ens hagi encarregat aquesta tasca. A més, aquesta facil interpretabilitat també ens podria aportar alguna que altre conclusió que s'ens hagi pogut escapar al profiling clàsic.

## 1.1 Class Panel Graph

Amb tal de construir aquesta gràfica, el primer pas és construir un *Class Panel Graph* (CPG), en el que podrem apreciar les diferents distribucions per cada variable dins dels diferents clusters. Degut a la gran quantitat de variables dins del nostre dataset (46 en total), com a la seva naturaleza, no té sentit incloure totes elles dins del CLP, ja que ens quedaría un CLP gigant i amb variables com *artist_name* (variable amb 301 modalitats diferents), fent així que el CLP perdi la seva gran fortaleza: la interpretabilitat.

Així doncs, vam decidir deixar fora les següents variables:

- Les variables que simbolitzin el temps, degut tant a la gran quantitat de modalitats que porten, com al fet que hi han 6 variables temporals, complicant encara més la interpretabilitat del CPG. Dins d'aquesta categoria entran: *year_release*, *month_release*, *day_release*, *weekday_release*, *year_week*, *month_week* i *week_index*.
- Totes aquelles variables categòriques que representin id's o noms, ja que dificularíen la interpretabilitat del CPG i no aportaríen quasi informació util, creuant 5 classes amb més de 5 modalitats. Aquest grup inclou *track_id*, *track_name*, *album_name*, *albumm_label* i *artist_name*.
- Les variables de geolocalització, ja que no porem classificarles amb 3 colors al TLP: *nationality* i*city*
- La variable *key*, per el gran numero de modalitats que té i la dificultat de classificarla amb 3

colors al TLP.

Finalment, ens quedarien totes les variables numériques junt amb les variables que indican el genre d'una canço, *ablum_type*, *collab*, *explicit*, *major_mode*, *rank_group*, *gender* i *is_group*.

ERROR: IDENTIFICAR SI ARTIST NUM ES O NO NUMERICA Vam decidir també afegir artist_num, ja que, a pesar de tindre 7 modalitats, la majoría de cançons estàn formades per 1 o 2 artistes, fen d'aquesta variable una facil de colorear, amb vert al 1, groc al 2 i vermell a 2 o més.

La grafica de la figura **??** es el CPG amb les variables ja explicades. Com es pot comprovar, en el que a les variables numèriques respecta, artist_followers i artist_popularity tenen una certa diferència en les diferentes clases. Energy també sembla que jugui un paper important, a pesar de no haver sigut de gran importancia mirant els boxplots al profiling clasic. Valence també sembla tindre una variancia significant en el cinqué cluster respecte a la resta, indicant que les cançons dins d'aaquest grup tindran una major mesura de "felicitat". Finalment, indicar que tempo també té una distribució diferent al 5é cluster, poseent una distribució binomial, molt diferent a la resta de distribucions de la resta de clusters. En les variables streams i speechiness, a diferencia de lo comentat en les seccions anteriores, sembla que no hagi tantes diferencies.

Per un altre banda, en les variables categòriques si es veuen diferencies mes aparents. En quant als generes, els clusters 1, 2 i 4 son clarament cançons de *pop*, mentres que els 3 i 5 son predominantment de hip_hop. També trobem que el genere latino es concentra clarament al 5é grup, i que les cançons explícites venen a agruparse al cluster 3.

## 1.2 Termòmetre

Un cop analitzat per sobre el CPG i amb ajuda de les descripcions univariants de les diferents variables (la seva mitjana, moda i els quartils), començem a crear els termometres que ens ajudaràn a crear un TLP facilment interpretalble per qualsevol persona. Tot i que els termometres haurian de crearse amb ajuda d'un expert en el camp de la base de dades, al no compter amb dit expert, vam escollirlos apojannos en tant el 1er i 3er quartil, com en la forma de la distribució de les variables.

Per crear aquests termòmetres, hem separat les variables escollides en el anterior apartat entre

aquelles numèriques i categòriques; ja que els termòmetres de les numèriques i categòriques tenen estructures diferentes.

Un cop separades, per a cada variable numèrica s'han apuntat en una taula d'excel el seu valor màxim i mínim, i els seus dos limits: que separan el color vert del groc (**b**), i el groc del vermell (**a**). El color verd s'assignat al conjunt de valors agrupat entre el valor màxim de la variable i la $b$, el vermell al conjunt de dades amb valors de la variable que es trobin entre el mínim i la $a$, i el groc s'assignat a les dades amb valors entre la $a$ i la $b$. En la majoría dels casos, degut a la distribució de les variables, hem apropat molt la $a$ i la $b$ al primer i tercer quartil.

En el que a les variables categóriques respecta, hem classificat cadescuna de les seves modalitats amb el color vert o vermell, fent que aquelles no clasificades siguin el color groc. En una altre taula d'excel s'han apuntat les diferentes modalitats que pertanyen al color vert separades per un espai en la columna *green_vector*, i s'ha fet el mateix per les modalitats vermelles.

Per les variables binàries (els generes de les cançons), hem decidit que el valor *TRUE* sigui el relacionat amb el color vert, i el valor *FALSE* amb el vermell. La variable rank_group era també sencilla de classificar: de vert la modalitat 1-10 (ja que lo millor per un artista es tindre la seva cançó amunt dels rànkins) i de vermell 30-40, que seguiría el pitjor puesto que es pot tindre a la nostre base de dades. La variable *gender* vam escollir el genre femení com a color vert i el masculí com a vermell, en *album_type* single va ser la modalitat verde i album la vermella.

Com es pot comprovar, per molt que l'objectiu del termòmetre sigui representar un conjunt de valors de una variable com a "dolents" i un altre com a "millor", en la nostre base de dades això no té molt de sentit, ja que que una cançó sigui o no del genre hip_hop, per exemple, no implica cap sentit de millor o pitjor. Si bé potser coincideix que en variables com *streams* o *popularity*, els valors alts si van relacionats amb la idea de que una canço tindra més exit (que no vol dir que sigui millor, evidentment això es subjectiu), en la majoria de elles, no té sentit relacionar el vert amb bo i el vermell amb dolent; si no que seràn colors que facilitaràn la interpretació, i en la majoría dels casos, el vert representarà valors alts de la variable (numèriques) o valors *TRUE* (categòriques), el vermell el contrari, i el groc un punt mig.

A continuació, hem creat un petit script de python que llegirà aquestes taulas d'excel i les traduirà a una serie de llistas de R, escrivintlas dins del script que utilitzarem per crear un pseudo-TLP.

## 1.3 pseudo-TLP a partir de termòmetre

Tenint ja preparats els valors $a$ i $b$ que delimitaràn la region verde de la groga i la groga de la vermella en les variables numériques, i els colors asignats a cada modalitat en les variables categòriques, hem creat un pseudo-TLP. Aquest pseudo-TLP es un CPG on pintem del seu color corresponent cada subplot per cada variable en cada cluster.

En el cas de les variables numèriques, hem escollit el color corresponent de cada subplot utilitzant la mediana. És a dir, calculem la mediana de cada variable dins del primer, segon, tercer, quart i quint cluster. A continuació, mirem on quedaría aquest valor dins del termometre, i s'escull aquest color pel subplot de aquesta variable amb el cluster corresponent. Aquesta técnica es bastant robusta a distribucions amb cuas molt largas o amb outliers molt distants, degut a que aquests fenomens faran que la mitjana no caigui en la zona on més valors tindrem, mentres que la mediana, en canvi, caurà on estiguin la majoria de valors, com es pot veure en la figura **??**. Tot i així, aquesta mesura fallarà en cas de aplicarla amb distribucions bimodals, algo que s'hauria de tindre en compte.

Per una altre banda, les variables categòriques han sigut clasificades mitjaçant l'us de la moda. Així doncs, en cas d'una variable binomial com *pop*, al haver'hi més valors *TRUE* que *FALSE* en el cluster 1, el color assignat a aquest cluster amb la variable *pop* serà verd.

Tot i així, aprofitant que en les variables binomials no s'està utilitzant el color groc, hem decidit crear una petita variació, basat en el següent pensament: el fet que hagi més instàncies d'una modalitat que de l'altre, no vol dir que aquesta sigui predominant del tot dins de la clase, ja que es pot donar el cas que hagin 300 *TRUE*s i 289 *FALSE*s, implicant que, realment, no n'hi ha una gran diferència dins d'aquesta variable. Així doncs, hem escollit un umbral, tal que si la diferencia de instancias entre les dues modalitats de una variable categòrica binaria es menor qu'aquest umbral, el color de la variable en aquest cluster sigui groc. En cas de variables amb més de dos modalitats, no hem sigut capaços de implementar-lo, degut a que el color groc ja s'està utilitzant, i no tenim clar si s'hauria de calcular una diferencia total entre totes les modalitats o una matriu de diferencias, tal que amb superar una de ellas el umbral,ja se li asigne un color diferent.

Havent aplicat ja aquestes regles amb tal d'escollir els diferents colors, ens ha quedat el següent pseudo-TLP: figura **??**. Tal i com es pot apreciar, en les variables numèriques tenim certes diferències en artist_followers com s'ha vist al profiling clasic i, a diferència de lo conluït a la sección

anterior, artist_popularity també té una moda i distribució diferent a aquelles dels clusters 1, 3, 4 i 5 al cluster número 2, implicant artistes una mica menys populars dins d'aquest clusters. Loudness té una petita diferència de moda, ja que la distribució del cluster 5 té una cua esquerra menys corta que la resta de distribucions a les altres clases, implicant cançons una mica més sorolloses. La variable valence també una moda diferent, tot i que la distribució no canvia tant, degut principalment a que la moda es troba just al limit entre els colors verd i groc del seu semàfor (el limit siguint 0.6 i la moda 0.656). Finalment, la resta de variables no tenen cap diferència significant en el que a la moda o distribució respecta.

acousticness: 1: major i 3 menor, resta iguals artist followers: 1,4 max - 2,3 min artist num: 2,5 max - 1,4 min danceability: 1 min - casi no varia energy: 2,5 max - 1,3,4 min casi no streams: 1 max - 5 min -resta iguals speechiness: 3,5 max- 1,2,3 min

- artist popularity: SI DIF - loudness: SI DIF - valence: SI DIF

track popularity: no diference - album popularity: no dif - liveness: no dif

- tempo: xd

## 1.4 TLP final

## 1.5 Anàlisi de resultasts

- Analitzar resultats de colors - Explicar la seva utilitat (entenibilitat per algú que no sigui estadistic)
- Comparar amb Profiling