CLUSTERING

- **Agrupa individus** amb alta similitud dins del clúster i gran diferència entre clústers.
- **Mètodes de partició:** *Ex: K-means.* Necessites fixar prèviament el nombre de grups (K) → input. Assigna individus als clústers segons la distància a un centre i actualitza. Es ràpid i senzill però sensible a outliers i assumeix formes esfèriques.
- **Mètodes jeràrquics (Hierarchical):** no cal fixar K, genera dendograma → ascendents (fusió) o descendents (divisió) → tallar a certa alçada = obtenció de grups. Les fulles del dendograma son els individus.
  - **Criteris d'agregació**: **Ward:** agrupa minimitzant la pèrdua d'inèrcia (la variància que expliquen) intra-grup (molt usat). Equilibrats.
  - **Distància per variables mixtes: Gower** → combina distàncies normalitzades segons tipus de variable.

PROFILING

- Tècnica per **descriure i entendre els grups** trobats mitjançant **variables qualitatives i quantitatives**.
- **Passos profiling:**
  - Trobar variables significatives: p-value < 0.05 → variable significativa
    - **Variables qualitatives**: χ^2 independence test / Multiple box plot
      H_0: X,Y are independent
      H_1: X,Y are associated
      LeBart test: Detecta si una modalitat específica és significativa en un grup concret.
      H_0: $\mu_k = \mu$ (the mean of my group (k) is the same of the hole) → pval < 0.05 → presència/absència rellevant
    - **Variables quantitatives**: *ANOVA test* (diferència entre mitjanes entre grups) / Multiple bar plot / Mean plot
  - Descriure les diferències entre els grups
  - Describe the groups:
    - Fer una **taula resum** de les variables significatives per grup
    - Frases clau per interpretar
    - Assignar **etiqueta significativa** a cada grup
- **Aplicacions profiling:**
  - **Màrqueting**: segmentar clients, crear perfils de consumidor
  - **Polítiques públiques**: detectar col·lectius vulnerables
  - **Prevenció de riscos**: identificar grups amb alt risc (ex: impagaments)
- El profiling **NO sempre requereix clustering**. Qualsevol **variable categòrica pot definir grups** i aplicar el profiling a partir d'aquí.

PCA (Principal Component Analysis)

- Redueix la dimensionalitat de dades numèriques mantenint la màxima variància possible.
- Transforma variables originals en noves variables (components principals, PC), ortogonals i no correlacionades. PC1 explica la màxima variància, PC2 la següent, etc.
- Cada component és una **combinació lineal** de les variables originals → **càrregues factorials (pc1$rotation)**. Com més càrrega factorial (en valor absolut) més pes en el component, més explica. Each component has a little portion of all the original variables, but not all the components contain the same contribution of each component. **ONLY NUMERICAL VARIABLES.**
- **Passos previs:**
  - **Centrar dades**: restar la mitjana (x - μ).
  - **Normalitzar** (si unitats diferents): dividir per la desviació estàndard.
- $sdev^2$ = variància (inèrcia) explicada per component (**eigenvalue**).
- Scree plot → decideix quants components conservar (fins a **80% variància acumulada**).
- **Biplot (gràfic de variables)**: fletxes = variables originals:
  - Direcció → cap on augmenta la variable.
  - Longitud → importància (variància explicada).
  - Fletxes juntes → correlació positiva.
  - Fletxes oposades → correlació negativa.
  - Fletxes amb un angle gran → no correlacionades.
- **Modalitats categòriques**: es projecten com a **centres de gravetat** → relacions entre modalitats i variables.
- Each dimension has different contributions, not all the plots have the same quantity of information.
- Les noves variables (PCs) es calculen combinant les **dades centrades** amb els **vectors propis** (**eigenvectors**). PC1 = Xs(1) · (-0.707) + Xs(2) · 0.707
- **Usos del PCA:**
  - Visualization of multidimensional data
  - Associative method of variables
    - Relations between variables (numeric and categorical)
  - Preprocessing data method
    - Latent variables (Variables that can not be measured: freedom, happiness, richness)
    - Reduccion of dimensionality

1. Gráficos en PCA (Análisis de Componentes Principales)

PCA reduce la dimensionalidad de los datos, y los gráficos asociados ayudan a visualizar cómo los datos se distribuyen a lo largo de las componentes principales.

a) Gráfico de Dispersión (Scatter Plot) de las Componentes Principales

- **Objetivo**: Visualizar los datos proyectados en las primeras dos o tres componentes principales.
- **Interpretación**:
  - Cada punto en el gráfico representa una observación (fila) de los datos originales.
  - Si los puntos se agrupan en ciertas áreas, esto indica que hay agrupamientos naturales o tendencias.
  - Los puntos dispersos pueden indicar la presencia de outliers o que los datos no se ajustan a un patrón claro.

b) Gráfico de Codo (Elbow Plot)

- **Objetivo**: Ayudar a determinar cuántas componentes principales conservar.
- **Interpretación**:
  - El **codo** de la curva indica el punto donde se observa un cambio notable en la pendiente. Este es el punto donde agregar más componentes no aporta mucha varianza adicional.

c) Biplot

- **Objetivo**: Visualizar tanto las observaciones como las variables originales en el espacio reducido por PCA.
  Interpretación:
  - Los **puntos** muestran las observaciones proyectadas sobre las dos primeras componentes principales.
  - Las **flechas** representan las variables originales. La longitud de las flechas muestra la varianza explicada por cada variable, y la dirección de la flecha indica cómo se relaciona esa variable con las componentes principales.
  - Si las flechas están cerca unas de otras, indica que las variables están altamente correlacionadas.

2. Gráficos en Clustering (Agrupamiento)

El objetivo de los gráficos en clustering es mostrar cómo los puntos se agrupan en diferentes clústeres.

a) Gráfico de Dispersión de Clústeres (Scatter Plot)

- **Objetivo**: Visualizar cómo los datos están distribuidos en los diferentes clústeres.
- **Interpretación**:
  - Los puntos en el gráfico se colorean según el clúster al que pertenecen.
  - Puedes identificar la forma, el tamaño y la separación de los clústeres.
  - Si los clústeres están bien separados y son densos, es un buen indicativo de que el algoritmo de clustering ha funcionado correctamente.
  - Si los clústeres se solapan mucho o están dispersos, puede indicar que el número de clústeres elegido no es el adecuado, o que los datos no tienen una estructura clara.

b) Gráfico de Siluetas (Silhouette Plot)

- **Objetivo**: Medir la calidad del agrupamiento.
- **Interpretación**:
  - El valor de la **silueta** varía entre -1 y +1. Un valor cercano a +1 indica que el punto está bien asignado a su propio clúster, mientras que valores cercanos a -1 indican que el punto podría pertenecer a un clúster diferente.
  - Un gráfico de silueta muestra la puntuación de silueta para cada punto, y también el promedio de la silueta para todo el conjunto de datos.
  - Si la mayoría de los puntos tienen una silueta alta, es un buen indicio de que el clustering es adecuado.

c) Dendrograma (para Clustering Jerárquico)

- **Objetivo**: Visualizar la jerarquía de los clústeres.
- **Interpretación**:
  - El dendrograma es un árbol que muestra cómo los puntos o clústeres se agrupan a medida que se fusionan o dividen.
  - El eje vertical muestra la distancia entre los puntos o clústeres.
  - Los **ramalazos** más cercanos indican que los puntos o clústeres se agrupan rápidamente, mientras que los ramalazos más alejados indican que los puntos se agrupan a mayores distancias.
  - El **corte** horizontal del dendrograma define cuántos clústeres finales se desean (al cortar el dendrograma por encima de un cierto nivel de distancia).

3. Gráficos en Profiling (Análisis de Perfiles)

El profiling tiene como objetivo explorar y describir los datos, y los gráficos son fundamentales para entender las distribuciones y relaciones de las variables.

a) Histogramas

- **Objetivo**: Visualizar la distribución de una variable.
- **Interpretación**:
  - El **eje X** representa los intervalos de valores de la variable (bins).
  - El **eje Y** muestra la frecuencia (número de observaciones) en cada intervalo.
  - Los histogramas te permiten ver si los datos están distribuidos de manera uniforme, sesgada, o si presentan una distribución normal.
  - Las **colas largas** o picos muy pronunciados indican la presencia de outliers o distribución no uniforme.

b) Boxplots (Diagramas de Caja)

- **Objetivo**: Visualizar la distribución y los outliers de una variable.
- **Interpretación**:
  - El **cuadro** muestra el rango intercuartil (Q1 a Q3), donde se encuentra la mayoría de los datos.
  - La **línea dentro del cuadro** muestra la mediana de la distribución.
  - Los **bigotes** muestran la extensión de los datos (generalmente hasta 1.5 veces el rango intercuartil), y los puntos fuera de los bigotes se consideran outliers.
  - Los **outliers** son puntos que se desvían significativamente de la distribución central de los datos.

c) Matriz de Correlación

- **Objetivo**: Mostrar las relaciones entre diferentes variables.
- **Interpretación**:
  - Las **celdas de la matriz** muestran los coeficientes de correlación entre las variables (generalmente entre -1 y +1).
  - Un valor cercano a **+1** indica una fuerte relación positiva (cuando una variable aumenta, la otra también lo hace), mientras que un valor cercano a **-1** indica una relación negativa.

BINARY

- Variables amb **resposta binària** (0/1) → predicció d'una probabilitat [0, 1]
  - → **Distribució Bernoulli → Disgregades:** cada fila un individu. Cada observació 0 o 1.
  - → **Distribució Binomial → Agregades:** Agrupació d'individus que tenen les mateixes variables explicatives. Probabilitat d'obtenir K èxits en M intents.
- **Odds:** quentes vegades és més probable que passi? $odds = \frac{\pi\,(probabilitat\ que\ passi)}{1-\pi\,(probabilitat\ de\ que\ NO\ passi)} = e^{\beta}(per\ un\ coeficient,\ es\ multipliquen\ els\ odds)$ odds = 1 → p = 0.5, odds > 1 → probable, odds < 1 → improbable.
- **Logit Link (Y):** resultat de la combinació lineal dels regressors. β indica efecte sobre el logit. Després transformem a probabilitat [0, 1].

  $\eta = log(\frac{\pi}{1-\pi})$ → (g) → de probabilitat [0, 1] a un valor real $(-\infty, \infty)$ (logaritme dels odds)

  $\pi = \frac{exp(\eta)}{1+exp(\eta)}$ → (g-1) de valor real $(-\infty, \infty)$ a probabilitat [0, 1]
- **Probit link:** Uses the inverse of the standard of normal distribution. Transforma π segons distribució normal estàndard.

  $\Phi(\eta)$ // $\eta = g_2(\pi) = \Phi^{-1}(\pi)$ // $\pi_2(\eta) = g_2^{-1}(n) = \Phi(n)$
- **Values of η and π :**
  - → If the value is 0 on the functions → probability π = 0.50 because both functions are centered.
  - → Higher values than 0 → higher probabilities than 0.50
  - → Lower values than 0 → lower probabilities than 0.50.
- **Unnested models** → AIC is used to compare. The lower the AIC, the better the model is.
- **Nested models** → We use deviance to compare. The lower the deviance, the better the model is.
  - → **Null deviance** is the deviance associated with the **null model**, with no regressors, only the intercept.
  - → **Residual deviance** is the deviance associated with the **full model** (with all predictors). Same or lower than the simpler model.
  - → $\Delta D = D_{null} - D_{full} > 0$ compare whether a particular regressor affects the output. $\Delta D \sim \chi^2$ → p-value
- **Matriu de confusió:** compara el que el model ha predit $(\widehat{Y})$ amb el que realment ha passat (Y) **threshold** → 0.5. TP (a), FP (b), FN (c), TN (d).
- **Evaluation metrics:**
  - → **Accuracy** = (TP + TN) / Total → Percentatge total d'encerts
  - → **Precision** = TP / (TP + FP) → How many predicted positives are actually positive
  - → **Recall (Sensitivity)** = TP / (TP + FN) → How many actual positives were caught
  - → **Specificity** = TN/(TN + FP) → How many actual negatives were caught
- **ROC:** sensibilitat vs 1–especificitat; **AUC** = àrea sota la corba. AUC = 1 perfecte, AUC = 0.5 aleatori, AUC < 0.5 pitjor que aleatori.
- **Linear model? Or do we search for other model options?**
  - → **resiudalPlots(m)** → no linearitat / punts infuents
  - → **marginalModelPlots(m)** → importància marginal del predictor
  - → **avPlots(m)** → contribució real de cada variable (recta amb pendent millor contribució)
  - → **crPlots(m)** → no linearitat més suau i detallada que els residuals