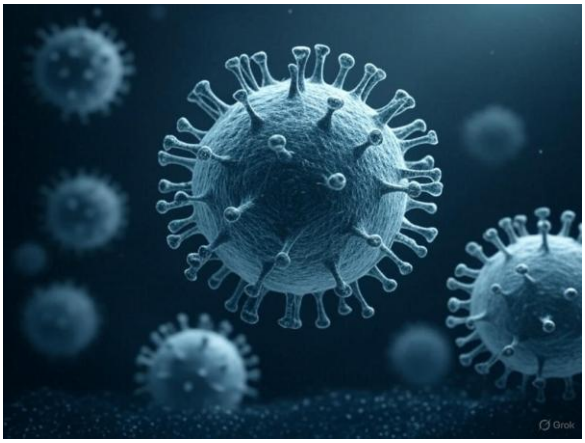
Pràctica 5 - Clustering

Aquesta pràctica està dissenyada perquè els estudiants apliquin tècniques de clustering, específicament l'algorisme k-means, a un conjunt de dades mèdiques. L'objectiu és analitzar les característiques dels pacients i identificar grups amb patrons similars basant-se en informació clínica, complicacions i dades relacionades amb l'assegurança. A continuació, es detallen les variables del conjunt de dades i les preguntes per guiar l'exercici.



Informació de l'historial clínic dels pacients:

- **SEX**: Dona o home.
- **AGE**: Edat del pacient.
- **CLASIFFICATION_FINAL**: Resultats de la prova de COVID-19. (covid / inconcluse / none) Els valors de l'1 al 3 indiquen diferents graus.
- **PATIENT_TYPE**: Hospitalitzat o no hospitalitzat.
- **PNEUMONIA**: Indica si el pacient té inflamació dels pulmons o no.
- **PREGNANT**: Indica si la pacient està embarassada o no.
- **DIABETES**: Indica si el pacient té diabetis o no.
- **COPD**: Indica si el pacient té malaltia pulmonar obstructiva crònica o no.
- **ASTHMA**: Indica si el pacient té asma o no.
- **INMSUPR**: Indica si el pacient està immunosuprimit o no.
- **HIPERTENSION**: Indica si el pacient té hipertensió o no.
- **CARDIOVASCULAR**: Indica si el pacient té malalties relacionades amb el cor o els vasos sanguinis.
- **RENAL_CHRONIC**: Indica si el pacient té una malaltia renal crònica o no.
- **OTHER_DISEASE**: Indica si el pacient té alguna altra malaltia o no.
- **OBESITY**: Indica si el pacient és obès o no.
- **TOBACCO**: Indica si el pacient és consumidor de tabac o no.
- **USMER**: Indica si el pacient va ser tractat en unitats mèdiques de primer, segon o tercer nivell.
- **MEDICAL_UNIT**: Tipus d'institució del Sistema Nacional de Salut que va proporcionar l'atenció.

Informació sobre complicacions dels pacients:

- **INTUBED**: Indica si el pacient va ser connectat a un ventilador.
- **ICU**: Indica si el pacient va ser admès a una Unitat de Cures Intensives.
- **DATE_DIED**: Indica si el pacient va morir o es va recuperar.

Informació sobre costos d'assegurança i dades relacionades (historiques segons assegurança medica):

- **bmi**: Índex de massa corporal, proporciona una mesura del pes corporal (kg/m²) en relació amb l'altura.
- **children**: Nombre de fills coberts per l'assegurança mèdica / nombre de dependents.
- **region**: Àrea residencial del beneficiari als Estats Units: nord-est, sud-est, sud-oest, nord-oest.
- **charges**: Costos mèdics individuals facturats per l'assegurança.
- **paid**: Indica si la factura dels costos ja ha estat pagada o no.

Tasques a realitzar:

1. Carregueu el conjunt de dades proporcionat. Volem realitzar un clustering utilitzant k-means. Com que scikit-learn només implementa la distància euclidiana, només accepta atributs numèrics. Transformeu tots els atributs a numèrics. **(2 punts)**
2. Com podeu observar, algunes columnes tenen valors buits. Ompliu aquests buits. On no hi hagi opcions millors, **utilitzeu arbres de decisió** per omplir-los. Per a les columnes on decidiu no utilitzar arbres de decisió, justifiqueu correctament la vostra decisió. **(2 punts)**
3. Ajusteu un clustering k-means amb 3 clústers i realitzeu la predicció. Obteniu els centres dels clústers (atribut cluster_centers_ del vostre model k-means ajustat). Interpreteu quin tipus d'usuaris hi ha en cada clúster i quines variables esta utilitzant més. Expliqueu per què obteniu aquests resultats. **(3 punts)**
4. Proveu altres nombres de clústers, seleccions de variables, escalat de variables i afegiu pesos. En aquesta pregunta, heu d'establir un objectiu de clustering i intentar aconseguir-lo. Expliqueu què espereu assolir, què esteu fent i què aconseguiu. **(3 punts)**

Lliurament: Com que els resultats poden variar entre execucions a causa de l'aleatorietat, heu de lliurar el codi utilitzat per respondre les preguntes i els resultats obtinguts a mes de les conclusions sol·licitades a cada apartat.