Pràctica 2: Classificació

Guillermo Vivancos Alonso 1606206
Javier Esmoris Cerezuela 1498396
Oriol Marión Escudé 1566740

Introducció

En aquesta pràctica analitzarem una base de dades sobre la potabilitat de l'aigua i les diferents concentracions d'algunes substàncies. Veurem quines distribucions tenen els atributs, la relació entre ells i intentarem determinar si donada una mostra, aquesta és potable o no.

Apartat B

Exploratory Data Analysis

Els atributs que tenim a la base de dades són els següents:

0. pH [float]: pH de l'aigua.
1. Hardness [float]:concentració de calci i magnesi.
2. Solids [float]: edat del jugador mesurada en anys i dies.
3. Chloramines [float]: concentració de cloramines.
4. Sulfate [float]: concentració de sulfats.
5. Conductivity [float]: conductivitat de l'aigua.
6. Organic_carbon [float]: concentració de compostos orgànics.
7. Trihalomethanes [float]: concentració de trihalometans.
8. Turbidity [float]: Terbolesa de l'aigua.
9. Potability [integer]: 0 si no és potable, 1 si és potable.

Tenim 10 atributs a la base de dades, tots de tipus float menys l'últim que és binari, per tant l'atribut categòric només pot prendre dos valors.

Pel que fa a les correlacions amb l'atribut categòric, totes són poc significatives com es pot veure en la figura 1.

A continuació veurem la distribució de la potabilitat 2 i les distribucions dels altres atributs segons la potabilitat 3.

A simple vista podem veure en la figura 3 que els atributs tenen una distribució gaussiana o pràcticament gaussiana. Les etiquetes de potabilitat estan balancejades (proporció 3:2).

Guillermo Vivancos Alonso 1606206
Javier Esmoris Cerezuela 1498396
Oriol Marión Escudé 1566740

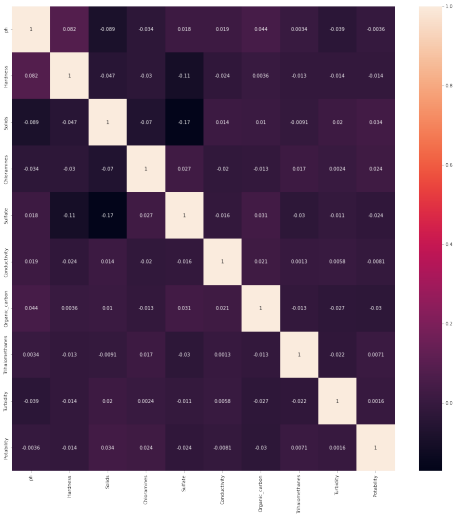

Figura 1: Correlació entre els atributs

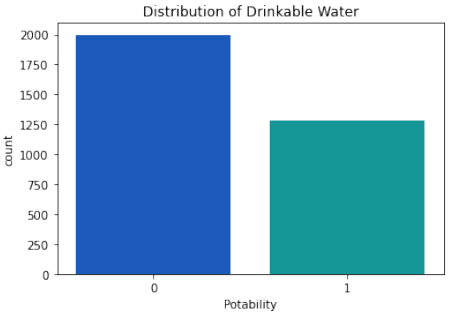

Figura 2: Distribucio de la potabilitat

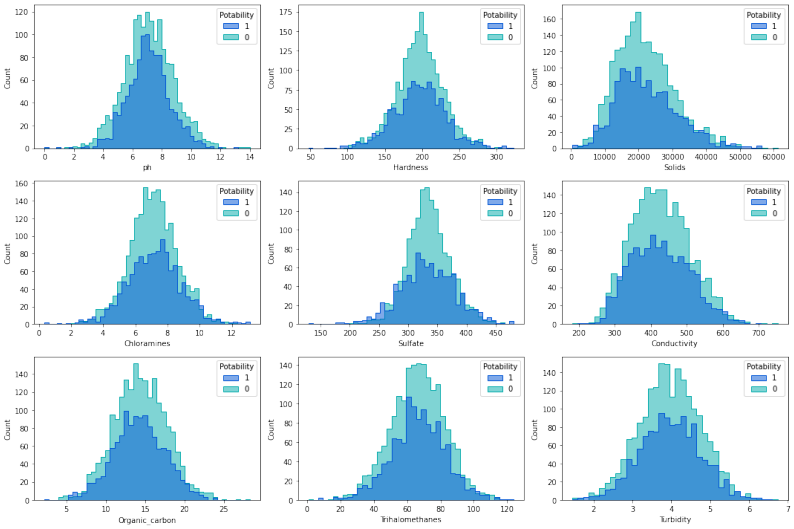

Figura 3: Distribucions dels atributs segons la potabilitat

Preprocessing

En primer lloc mirem quines variables tenen NaNs, que són pH, Sulfate i Trihalomethanes. Els valors mitjans dels atributs, tant quan l'aigua és potable com no, són molt semblants, per tant, substituim els NaNs per la mitja dels atributs.

La variable categòrica pren només dos valors, per tant tècniques com Ordinal Encoder o One-HotEncoder no tenen gaire sentit en aquest context.

Hi ha rangs de variables molt aixos en comparació amb uns altres. L'atribut solids té el rang més alt de l'ordre de $6 \cdot 10^6$ No tenim files repetides a la base de dades.

Apartat A