**3.2. Datu meatzaritzako paradigmak**

* **Iragarpena (sailkapen gainbegiratua):** Iragarpenaren helburua etorkizuneko emaitza bat aurreikustea da, datuetan oinarrituta. Normalean, irteera aldagai bat dago (esaterako, kategorikoa edo kuantitatiboa), eta horren balioa aurreikusteko eredu bat entrenatzen da etiketatutako datuekin.
  + Ezaugarriak:
    - Etiketatutako datuak behar ditu (input eta output ezagunak).
    - Ereduak emaitza berriak aurreikusteko gai izan behar du.
  + Adibideak:
    - **Gene batzuen presentziaren arabera, etorkizunean gaixotasun bat izateko probabilitatea eman:** Gaixotasun baten arriskua sailkatzeko (adibidez, "bai/ez" edo arrisku maila kuantitatiboa).
    - **Biometria: begiko irisaren ezaugarri batzuen arabera, pertsona identifikatu:** Pertsonak sailkatzeko, irisa erabilita.
* **Clustering:** Clustering datuak multzotan edo klusterretan sailkatzeko teknika da, baina kasu honetan ez dago aurretiazko etiketa edo irteerarik. Algoritmoak datuen egitura aztertzen du antzeko ezaugarriak dituzten elementuak taldekatuz.
  + Ezaugarriak:
    - Ez da beharrezkoa datuek aurretiko etiketa izatea.
    - Datuen arteko antzekotasunaren arabera taldekatzen du.
  + Adibideak:
    - **Espezie bateko ezaugarrien arabera, barietateak bereiztu, alegia, taxonomiak deskubritu:** Clustering erabiliz, espezieen barietateak automatikoki detektatu daitezke.
    - **Aseguru etxeetan antzeko jokaerak dituzten bezeroei antzeko produktuak eskaini:** Bezeroen jokaerak aztertuz taldekatzen dira profil antzekoak dituztenak.
* **Asoziazioa:** Asoziazio erregela ikasketa erlazioak aurkitzea da, datu multzo batean aldagai batzuen arteko patroiak eta ereduak detektatuz. Ohikoa da transakzioetan oinarritutako datuetan erabiltzea, non erregela sinpleak ateratzen diren.
  + Ezaugarriak:
    - Erlazio eta patroi interesgarriak aurkitzea helburu.
    - Ez da beharrezkoa output etiketa izatea.
  + Adibideak:
    - **Iraganean entzundako musikaren arabera, musika gomendatu:** Asoziazio erregelak erabiliz, entzundako musika estilo edo ezaugarrien arabera gomendioak sortu.

**3.3. Datuen deskribapen operatiboa**

**1.**

* **Zein motatako informazioa (audio, irudiak, . . . ) dakar .arff fitxategiak?**

.arff fitxategiak **atributuak eta datuak** biltzen ditu, hau da, **datu multzo baten ezaugarriak eta balioak**. Informazio mota nagusiak:

* + **Testu-datuak**: Hitzak edo string motako informazioa (adibidez, "bai"/"ez", "zuri", "beltza").
  + **Zenbakizko datuak**: Balio kuantitatiboak edo errealak (adibidez, 3.14, 42).
  + **Kategoriak**: Diskretuak edo finituak diren kategoriak (adibidez, "txikia", "ertaina", "handia").
* **Zein da ARFFren esannahia?**

ARFF: Attribute-Relation File Format da, hau da, Atributu-Harremanen Fitxategi Formatua

* **Zertarako erabiltzen dira mota honetako fitxategiak?**
  + **Ikasketa automatikoa eta datu meatzaritza**: ARFF fitxategiak datuak erraz kargatzeko erabiltzen dira hainbat tresnatan, batez ere Weka softwarean.
  + **Datuen irudikapena**: Datuen eta haien atributuen arteko erlazioak deskribatzeko eta biltegiratzeko.
  + **Datu multzoak aurkeztea**: Estandarizatutako formatu batean, datuak eta haien metadatuak gordetzeko:
    - Atributu bakoitzaren izena, mota eta balio eremuak zehaztea.
    - Datuen erregistroak metadatu horren arabera gordetzea.
  + **Ikasketa automatikoko algoritmoak probatzea eta konparatzea**: Ereduak entrenatu eta ebaluatzeko datu multzoak kudeatzeko.

**2.**

**(a) Zertan datza ataza? Iragarpen (prediction), taldekatze (clustering) ala elkarketa (association) buruketa da?**

Ataza **iragarpen** (prediction) batean datza.

* Helburu nagusia da paziente baten bihotzeko gaixotasunen **presentzia edo maila diagnostikatzea**, oinarritzat hartuta datu-multzoan dauden atributuak.
* Helburu atributua da **num**, eta haren balioak (0-4) bihotzeko gaixotasunaren mailak adierazten ditu.
* **Machine Learning-en testuinguruan**, klasifikazio-arazo bat da, non balio espezifikoak (0-4 artean) aurreikusi behar diren.

**(b) Buruketako deskribapenaren arabera, zenbat balio har ditzake klaseak? Daukagun lagin multzoan, zenbat balio har ditzake klaseak?**

**1. Deskribapenaren arabera:**

* Klaseak bost balio har ditzake: 0, 1, 2, 3 eta 4.
* Hauek adierazten dute bihotzeko gaixotasunen maila:
  + 0: Gaixotasunik ez (<50% estutzea).
  + 1, 2, 3, 4: Gaixotasunaren mailak (50% edo gehiago estutzea).

**2. Daukagun lagin multzoan:**

* Ezaugarri banaketaren arabera, lagin-multzo ezberdinetan klaseen kopuruak alda daitezke.
* Adibidez:
  + **Cleveland** multzoan, bost balio daude (0, 1, 2, 3 eta 4).
  + **Hungarian**, **Switzerland**, eta **Long Beach VA** datu-multzoetan ere bost balio daudela iradokitzen da.

**(c) .arff fitxategian ’%’ ikurrarekin hasten diren lerroak, fitxategiko parte eragile dira?**

Ez, % ikurrarekin hasten diren lerroak **iruzkin lerroak dira**.

* Hauek ez dira zuzenean fitxategiko datu edo egitura parte hartzaile.
* Iruzkin gisa erabiltzen dira fitxategia deskribatzeko edo dokumentatzeko (adibidez, datu-multzoaren jatorria, erabilera-oharrak, atributuen zerrenda eta bestelako azalpenak emateko).

**3. Definitu: “Instantzia” eta “Atributu”**

* I**nstantzia**: Datu-multzo batean, **instantzia** batek datu puntu edo erregistro bakar bat ordezkatzen du. Instantzia bakoitza datu-multzoaren unitate indibiduala da, eta atributu guztien balio multzo bat dauka.

Instantziak normalean **ikuskatzen edo aztertzen den elementu jakin bat** adierazten du, hala nola paziente bat, irudi bat edo transakzio bat.

* **Atributua**: Atributua da datu-multzoan instantzia bakoitza deskribatzen duen ezaugarri edo aldagai bat. Atributuak dira aztertzen diren objektuak (instantziak) karakterizatzen dituzten ezaugarriak.

**4.Zer motako atributuekin egiten du lan Wekak?**

* Numerikoak (Numeric) : Atributuak balio **kuantitatiboak** hartzen ditu, zenbaki errealak edo osokoak izan daitezke. (2, 45 100, …)
* Nominalak (Nominal) : Atributuek balio **kategoriakoak** hartzen dituzte, eta balio horiek etiketa moduan agertzen dira. Balio horiek ez dute ordenarik eta ez dute zenbakizko esanahirik.
  + Sexua: *female (0), male (1)*.
  + Bularreko mina: *typ\_angina, non\_anginal, asympt*.
  + Gaixotasunaren diagnostikoa: *<50, >50\_1, >50\_2*.
* Binarioak (Binary) : Balio **bitarrek** soilik bi aukera dituzte, hala nola *bai* edo *ez*, *0* edo *1*. Binarioak nominalen azpimota dira.
* Ordenatuak (Ordinal) : Atributuak balio **kategoriakoak** hartzen ditu, baina balioen artean ordena bat dago.
  + Malda motak (slope): *up (1), flat (2), down (3)*.
  + Elektroardiograma emaitza (restecg): *normal (0), st\_t\_wave\_abnormality (1), left\_vent\_hyper (2)*.
* Testua (String) : Atributuek balio gisa testu-hitzak edo kateak hartzen dituzte.
* Data eta Ordua (Date) : Atributuak data edo denbora zehatzak gordetzen ditu. Balio hauek formatua izan dezakete (adibidez, YYYY-MM-DD).

**5. Wekan instantzia guztiek atributu kopuru bera dute?**

**Bai**, Wekan instantzia guztiek atributu kopuru bera dute. Hau da, datu multzo bat *.arff* edo beste formatu batean definitzen denean, instantzia guztiek (datu puntuek) atributu berak dituzte, eta atributu horien kopurua konstantea da datu-multzo osoan.

Instantzia batek atributu baterako baliorik ez badu, horren ordez balio bat "falta" bezala adierazten da (adibidez, ?-rekin). Hala ere, atributua existitzen jarraitzen du.

**6. Wekan zein da atributu baterako daturik ez dugula adierazteko ikurra?**

Wekan atributu baterako daturik ez dugula adierazteko ikurra **?** da.

**7. Aztertzen ari garen atazarako:**

* **Zenbat instantzia dago? (N=**303**)**
* **Instantziak karakterizatzeko zenbat atributu dago? (n=**14**)**
* **Lehenengo 5 atributuetarako eta klaserako, galdera hauei erantzun:** 
  + **Zein motakoa da atributua? (eg. nominala, zenbakizkoa, string, . . )**
    - 'age': Zenbakizkoa (real)
    - 'sex': Nominala 🡪 { female, male}
    - 'cp': Nominala 🡪 { typ\_angina, asympt, non\_anginal, atyp\_angina}
    - 'trestbps': Zenbakizkoa (real)
    - 'chol': Zenbakizkoa (real)
  + **Atributu bakoitzerako aztertu zenbat instantziek ez duten baliorik atributu horretan** *(missing values)***. Zein portzentaian?**

Lehenengo 5 atributuetan baliorik ez duten instantziarik ez daude. (%0)



* + **Zenbat balio desberdin erregistratu dira atributu bakoitzerako?** *(distinct)*
    - 'age':



* + - 'sex':



* + - 'cp':



* + - 'trestbps':



* + - 'chol':



* + **Atributu bakoitzerako, badago behin baino erregistratu ez den baliorik?** *(unique values)*
    - 'age':



* + - 'sex':



* + - 'cp':



* + - 'trestbps':



* + - 'chol':



* + **Histogramen gaineko zenbakiek zer adierazten dute?**

**Histogramen gaineko zenbakiek** datu multzoaren banaketari buruzko informazioa adierazten dute. Zehazki, zenbaki hauek atributu baten balioak kategoria edo tarteetan banatuta, balio horien maiztasunak erakusten dituzte. Hau da, zenbat instantziak dute balio jakin bat edo tarte batean dagoen balioa.

* + **Numerikoak diren atributuetarako zein da erregistratu den balio minimo, maximoa, batezbestekoa eta desbiderapena?**
    - 'age':

Tabla

Descripción generada automáticamente

* + - 'trestbps':

Tabla

Descripción generada automáticamente

* + - 'chol':

Tabla

Descripción generada automáticamente

**8. Atributuak klasearekiko histograma aztertu.**

* **Intuitiboki, zeintzuk dira informazio gehien eskaintzen duten atributuak sailkapen problemari aurre egite aldera? Alegia, atributu gutxirekin iragarpenak egiteko gai izango ginen?**
* **Badago korrelazioa aurkezten duten atributu-bikoteak? Korrelazionatutako atributuak erabiltzea erabilgarria izango da?**

**9. Atributuak bikoteka aurkeztu:** *Visualize (goian, eskuman):*

* **Iragarri nahi den klasearen balioak ondoen diskriminatzen duten atributu bikoteak aukeratu.**
* **Informazio gutxien eskaintzen dituzten 3 atributu ezabatu eta datu fitxategia gorde izen honekin:** *heart\_c\_3attManuallyRemoved.arff***. Jarraian, hasierako datuak berreskuratu goiko botoia *Undo* sakatuz.**

**A Galdetegia**

**Erantzun laburrak eta zehatzak eman:**

* **Zertarako erabili datuak datu meatzaritzan?**

Datu meatzaritzan, datuak ereduak aurkitzeko, informazio baliotsua ateratzeko eta erabakiak hartzeko erabiltzen dira.

* **Deskribatu ataza hauetako bakoitza eta adibide bat eman azalpena argitzeko:**
  + **Iragarpena:** 
    - Etorkizuneko balioak edo klaseak aurreikustea atributu ezagunetan oinarrituta.
    - Adibidea: Paziente batek bihotzeko gaixotasuna duen iragartzea bere sintomak eta mediku-datuetan oinarrituta.
  + **Clustering:** 
    - Antzeko instantziak taldekatzea klase edo etiketa aurrez definitu gabe.
    - Adibidea: Bezeroak portaera-ereduetan oinarrituta segmentatzea (adibidez, online erosketa-portaera).
  + **Asoziazioa:** 
    - Aldagai edo atributu desberdinen arteko erlazioak aurkitzea.
    - Adibidea: "Erosketa-gatazkak" aurkitzea: erosleek ogia erosten dutenean, litekeena da gurina ere erostea.
* **Zer erabiltzen da datuetako adibide bat deskribatzeko? Zer motako aldagaiak erabil daitezke datuak deskribatzeko?** 
  + **Erabiltzen dena:** Atributuak edo aldagaiak.
  + **Aldagai motak:** Nominalak (kategorikoak), zenbakizkoak (real edo integer), binarioak, string-ak, etab.
* **Iragarpen atazean, zer da klase aldagaia? Zer adierazten du aztertutako adibidean?**
  + **Zer da**: Iragarri nahi den helburu aldagaia.
  + **Aztertutako adibidean**: Pazienteak bihotzeko gaixotasuna duen ala ez adierazten du.