PAC 2: Integració, validació i anàlisi dels diferents tipus de dades.

Tipologia i cicle de vida de les dades

Nom: Marc Bracons Cucó

Aula 2

Enunciat

Exercici 1 [20%]

Després de llegir el capítol 1 del recurs "Introducció a la neteja i anàlisi de les dades", respon a les següents preguntes amb les teves pròpies paraules.

1. Com podríem aplicar la reducció de la dimensionalitat i la reducció de la quantitat de dades en un estudi de mercat per identificar patrons de consum en una població diversa? Proporciona exemples específics de com cada tècnica podria ajudar a simplificar l'anàlisi i a obtenir informació rellevant de manera eficient. [Màxim 200 paraules]

   Per fer-ho podríem aplicar tècniques com l'Anàlisi de Components Principals. Per exemple, en un joc de dades amb múltiples variables com edat, ingressos, hàbits de consum i preferències, el ACP ens ajudaria a reduir el nombre de variables, considerant només aquelles que mostren una variància més gran. Gràcies a això es pot simplificar l'anàlisi a la vegada que no perdem informació important,

   Pel que fa a la quantitat de dades, es pot fer servir l'agregació, que ens permetria analitzar grans volums de dades. Per exemple, en comptes d'analitzar totes les transaccions de compra individualment, les podem agrupar per categoria o per regió, fent que sigui més fàcil identificar les tendències de consum.

2. Quines són les avantatges del procés de conversió de dades, incloent tècniques com la normalització, la transformació de Box-Cox i la discretització, i per què són útils en un context d'investigació? Proporciona un exemple concret de com una d'aquestes tècniques podria millorar la interpretació de dades en un estudi específic. [Màxim 200 paraules]

Gràcies a la normalització les dades de diferents fonts o escales poden ser comparables entre elles, cosa que pot ser molt útil quan hi ha diversos instruments de mesura. A més ajuda a eliminar biaixos, facilitant un anàlisi més just. La discretització ens permet transformar les dades contínues en categòriques, simplificant l'anàlisi i ajudant en la seva visualització. Finalment, la transformació de Box-Cox ens permet estabilitzar la variància i normalitzar la distribució de les dades, millorant la homogeneïtat.

Imaginem un escenari on estem realitzant un estudi sobre el rendiment d'un centre acadèmic, l'objectiu del qual és determinar quins factors influeixen en el seu èxit. El joc de dades tindrà variables com l'edat, la puntuació en el exàmens i les hores dedicades a estudiar.

La normalització ens podria ajudar a comparar les notes dels diferents exàmens si aquests tenen escales diferents, uns de 0 a 10, altres de 0 a 100, etc. La discretització ens ajudaria a transformar l'edat dels estudiants en possibles grups d'interès "menys de 20 anys", "entre 20 i 25" i "majors de 25". Finalment aplicaríem la transformació de Box-Cox a les hores dedicades a estudiar per setmana, estabilitzant la variància i la normalitzar la distribució d'aquestes hores.

Exercici 2 [30%]

Després de llegir el capítol 1.5 i 1.6 del recurs "Introducció a la neteja i anàlisi de les dades", respon a les següents preguntes amb les teves pròpies paraules.

1. Quines són les implicacions de la presència de dades perdudes en un conjunt de dades i com poden aquestes absències influir en la integritat i validesa dels resultats obtinguts en un anàlisi estadístic o de dades? Quines consideracions s'han de tenir en compte en seleccionar un mètode d'imputació de dades adequat per a un conjunt de dades específic? [Màxim 200 paraules]

   Que en un joc de dades hi hagi dades perdudes pot arribat a afectar greument la integritat i la validesa dels resultats d'un anàlisi estadística. Això es deu a que en l'anàlisi s'assumeix que les dades són representatives de la població i en funció del nombre de valors faltants podria no ser així, cosa que pot resultar en estimacions esbiaixades i per tant, conclusions incorrectes.

   Per solucionar-ho es poden fer servir mètodes d'imputació, però s'ha de tindre en compte la naturalesa de les dades i els seus possibles patrons. Per exemple, si la base de dades ha perdut valors a l'atzar, podem fer servir mètodes com la imputació mitjana. Si en canvi, el patró de les dades perdudes està relacionat amb altres variables del data set, seran necessaris mètodes més complexes com la imputació múltiple o l'ús de models predictius.

   També s'ha de tindre en compte la rellevància dels valors perduts o la seva quantitat, ja que a vegades simplement es poden completar manualment o eliminar aquells registres del joc de dades.

2. Quines tècniques es poden utilitzar per tractar les dades perdudes i en què consisteixen? [Màxim 200 paraules]

Existeixen diverses tècniques per tractar les dades perdudes, algunes d'elles són:

- Imputació Mitjana/Moda/Mediana: Substituint els valors perduts per la mitjana, la moda o la mediana de la variable. És simple i ràpid, però poden ser inadequades si les dades no són distribuïdes uniformement.

- Imputació per Regressió: Es poder fer servir models de regressió basats en altres variables del joc de dades per estimar el valor perdut. És més precisa, però es necessita una relació lineal entre les variables.
- K-nearest Neighbors: La imputació es fa en funció de la semblança dels “K” veïns més pròxims. És molt efectiu en jocs de dades on les relacions de proximitat són significatives. S’ha de tindre en compte que quan major sigui el valor de “K”, més veïns es miraran per trobar-ne els més pròxims però també augmentarà el temps de càlcul.

Exercici 3 [20%]

Després de llegir el capítol 2.2 del recurs "Introducció a la neteja i anàlisi de les dades", respon a la següent pregunta amb les teves pròpies paraules:

1. Què és la regressió i com s'utilitza en l'anàlisi de dades? [Màxim 150 paraules]

   La regressió és una tècnica que ens permet modelar i analitzar la relació entre una variable dependent i una o més variables independents. Es fa servir per fer prediccions d’una variable en funció d’altres. Per exemple, en la regressió lineal, s’ajusta la línia recta que millor representi la tendència dels punts en un gràfic de dispersió, fent que sigui més senzill la comprensió de com els canvis en la variables independent afecten a la variable dependent. A part de per fer prediccions, també és útil per a fer estimacions i entendre relacions entre variables.

2. Què és la correlació i com s'interpreta el seu valor? Dóna un exemple. [Màxim 250 paraules]

   La correlació és una mesura estadística que ens indica el grau i la direcció de la relació entre dues variables. El seu valor oscil·la entre -1 i 1, on, un valor proper a 1 ens indica una correlació positiva forta, és a dir, les dues variables augmenten

o disminueixen junts. En canvi, un valor de -1 ens indica una correlació negativa forta, on una variable augmenta quan l'altre disminueix. Si el valor és 0, vol dir que no hi ha una relació lineal entre les variables.

Per exemple, es podria analitzar la correlació entre la variable "hores d'estudi" i la variable "qualificacions". Aquesta (a priori) serà una correlació positiva i alta, ja que quantes més hores d'estudi dediquis a una assignatura, millor nota trauràs. Però s'ha de tindre en compte que, (com deia un professor del meu grau) correlació no implica causalitat.

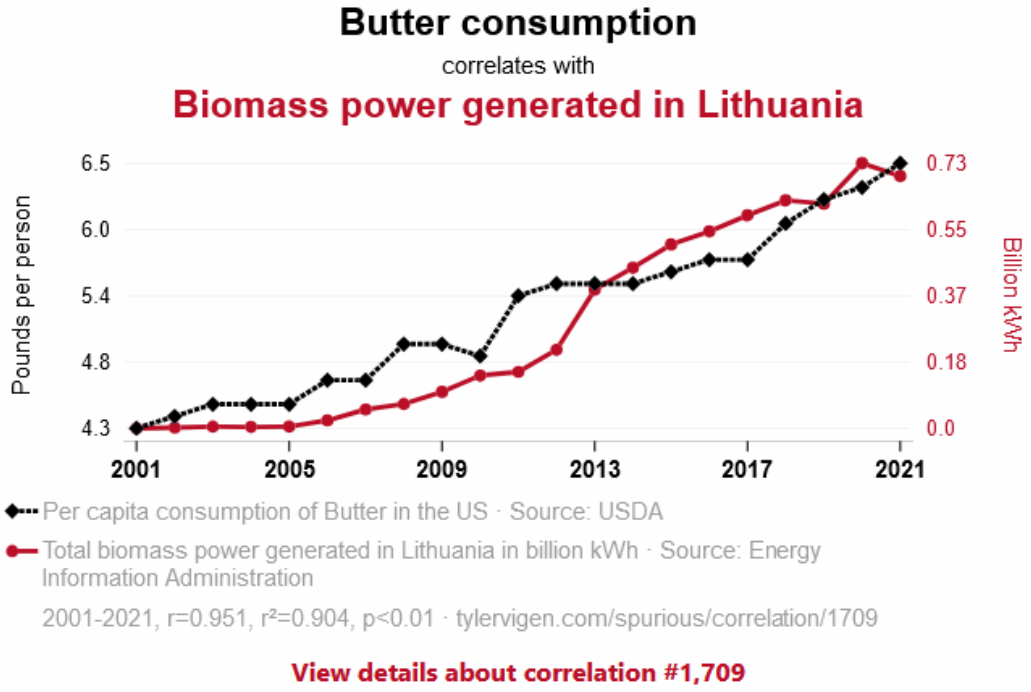

Figura 1: Correlation is not causation.
https://www.tylervigen.com/spurious-correlations

Exercici 4 [30%]

Després de llegir els capítols 2.4 del recurs "Introducció a la neteja i anàlisi de les dades", i en el recurs complementari "Data mining: conceptos y técnicas", respon a les següents preguntes amb les teves pròpies paraules:

1. Què és l'aprenentatge supervisat i com es diferencia de l'aprenentatge no supervisat? Proporciona un exemple d'aplicació de tots dos que no estigui en la teoria. [Màxim 300 paraules]

   En l'aprenentatge supervisat el model s'entrena amb un conjunt de dades prèviament etiquetats, així doncs, cada dada té les característiques d'entrada i la sortida desitjada. L'objectiu del model es aprendre a fer prediccions o classificar noves dades basant-se en els patrons apresos.

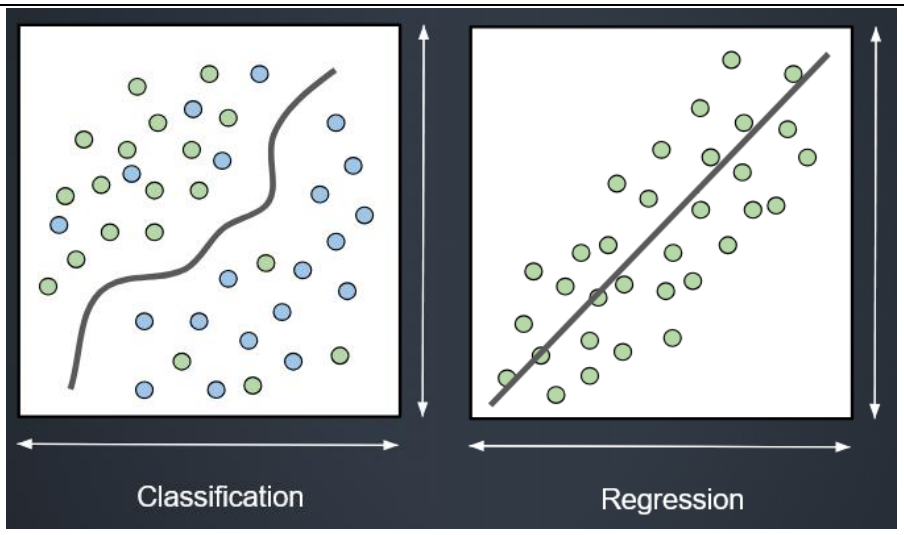

Figura 2: Exemples d'aplicacions del models amb aprenentatge supervisat

   En canvi, en l'aprenentatge no supervisat es treballa amb dades no etiquetades, L'algorisme intenta descobrir estructures ocultes o patrons fins de les dades, però sense una guia externa. Es fa servir en tasques d'identificació de grups, reducció de la dimensionalitat o la detecció d'anomalies.

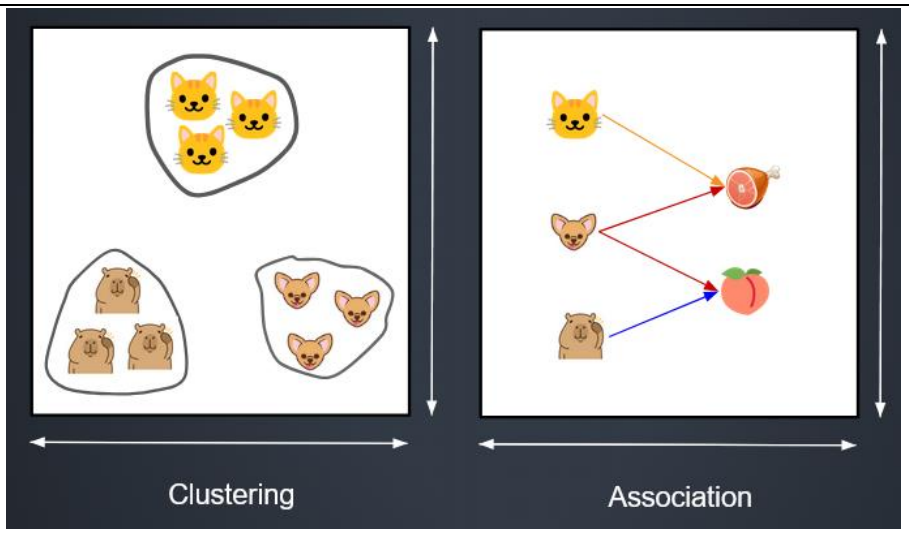

Figura 2: Exemples d'aplicacions del models amb aprenentatge no supervisat

2. Quines tècniques s'utilitzen per avaluar la qualitat dels models d'aprenentatge automàtic i en què consisteixen? [Màxim 200 paraules]

Per avaluar la qualitat dels models es fan servir tècniques que mesuren el rendiment en funció del tipus de problema. En models de classificació s'utilitza la matriu de confusió, on hi ha mètriques com l'exactitud, la precisió, la sensibilitat i l'especificitat. L'exactitud mesura la proporció de prediccions correctes; la precisió la proporció de prediccions positives correctes; la sensibilitat (veritables positius) ens indica la capacitat d'identificar correctament els casos positius i l'especificitat (veritables negatius) mesura la capacitat d'identificar correctament els casos negatius.

Les corbes ROC i l'àrea sota la corba (AUC) ens serveixen per avaluar la capacitat discriminativa del model, mostrant un compromís entre la sensibilitat i la taxa de falsos positius.