Informe Pràctica 1

Aprenentatge computacional:

Regressió lineal

Universitat Autònoma de Barcelona

Sergi Sánchez Hernandez NIU 1417652
Jordi Vila Folgarona NIU 1562542

Sergi Sánchez Hernandez NIU 1417652
Jordi Vila Folgarona NIU 1562542

ÍNDEX

Sergi Sánchez Hernandez NIU 1417652
Jordi Vila Folgarona NIU 1562542

Introducció

En aquesta pràctica utilitzarem una base de dades del Kaggle per a posar en pràctica la teoria i els conceptes apresos. La nostra base de dades tracta sobre l'esperança de vida i un seguit de factors que la condicionen, els quals veurem més endavant.
Enaquesta memòria tractarem la part C i B de la pràctica, ja que són les parts que hem realitzat. A la primer part de la memòria explicarem els atributs presents a la nostre base de dades, veient el seu tipus, rangs de valors, distribució i importància a l'hora de fer la regressió per a la nostre variable objectiu. A la segona part de la memòria realitzarem la regressió lineal i, a partir dels resultats veurem de quina manera les nostres dades influeixen sobre la predicció, i jugarem amb les nostres dades per a obtenir un millor model. Finalment farem una petita conclusió, sobre el que hem après i com ha sigut el resultat obtingut.

L'enllaç al nostre GitHub és el següent: https://github.com/Megaguille11/Practicas_ApC

Objectius

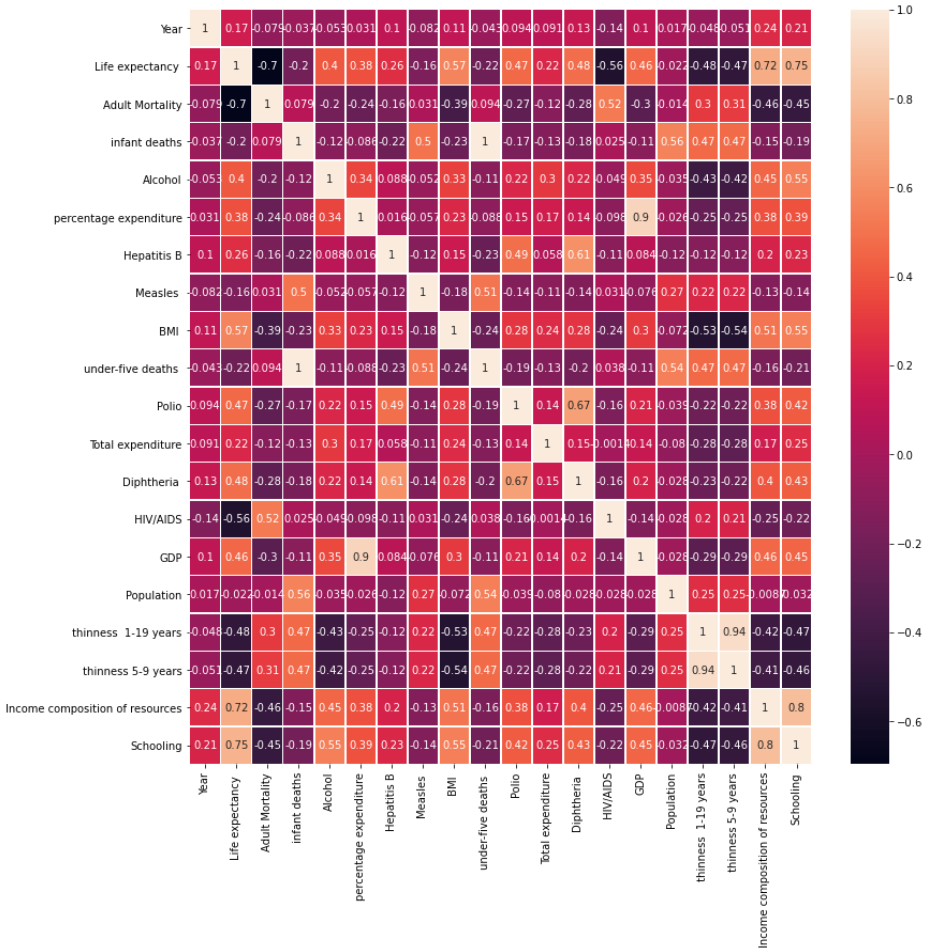
- Analitzar els atributs per seleccionar els més representatius i normalitzar-los.
- Visualitzar les dades i resultats de forma clara i entenedora.
- Ser capaços de fer una predicció del valor de l'esperança de vida dels habitants d'un país a partir d'altres dades que afectin al seu benestar.

Analitzant les dades

Aquesta base de dades recopila informació de la OMS en relació a l'esperança de vida de 193 països durant diversos anys així com altres factors que s'estimen que poden condicionar-la. Algunes d'aquestes informacions les podríem eliminar perquè no considerem que són innecessàries o no estan prou relacionades amb el nostre atribut objectiu, que serà, clarament, el de "life expectancy", ja que és el que més

interès tindrem a analitzar respecte als altres. Per exemple, tenim atributs amb dades que es repeteixen, per exemple les de “infant deaths” i “under-five deaths” registren informació idèntica, així que deixarem només la primera. Hi ha un atribut que valora la mortalitat en persones d’entre 15 i 60 anys (“Adult Mortality”), que deixarem. També tenim els atributs “thinness 10-19 years” (al dataset apareix com a “thinness 1-19 years” però és incorrecte) i “thinness 5-9 years”; per què volem aquesta informació, si ja tenim l’atribut “BMI” (índex de massa corporal) que també cobreix als adults? També volem retirar l’atribut “schooling”, ja que no veiem quina relació pot tenir saber quants anys està la gent escolaritzada amb la seva esperança de vida, però de moment el deixarem per parlar més tard de la distribució de dades. El de “population” tampoc no el necessitem i té massa valors nuls.

En aquesta taula podem veure com els atributs “infant deaths”-”under-five deaths” i “thinness” tenen una correlació d’1 o gairebé 1, per tant s’estaven repetint.



També tenim atributs relacionats amb la immunització contra diverses malalties (“hepatitis B”, “polio”, “diphteria”), els casos de “measles” (xarampió), el consum de “alcohol” i les morts per “VIH/AIDS”; així com pel nombre de població, el PIB per càpita (“GDP”), quin percentatge d’aquest PIB es destina a la salut (“percentage expenditure”), quin percentatge del pressupost del govern es destina a salut (“total expenditure”) i quin índex de desenvolupament humà té el país (“Income composition”).

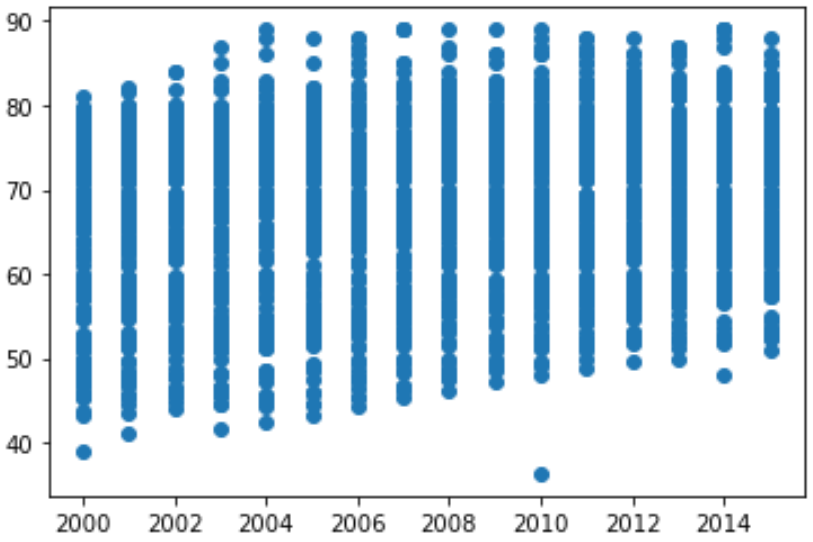
Tenim, en la gran majoria, atributs continus, ja que són variables numèriques que tenen un rang infinit de valors. Els atributs “country”, “status”, “percentage expenditure”, “hepatitis B”, “polio”, “total expenditure”, “diphteria” i “Income composition of resources” són categòrics, ja que només poden tenir valors concrets. En el primer cas de l’“status” són dos valors: “Developing” o “Developed”; “income” pot anar del 0 a l’1, i els percentatges (“percentage expenditure”, “hepatitis B”, “polio”, “total expenditure” i “diphteria”) del 0 al 100.

Només els atributs “total expenditure” i “schooling” tenen una distribució gaussiana (la majoria de valors corresponen als de la meitat del seu rang), encara que en el primer cas està desviada. Ho expliquem amb detall més endavant.

Relacions entre l’esperança de vida i atributs categòrics

Ara, començarem a visualitzar la relació entre els atributs categòrics de la nostra base de dades amb l’esperança de vida.

Primerament, podem fer una primera visualització de l'evolució de l'esperança de vida a nivell global, des de l'any 2000 fins l'any 2015.



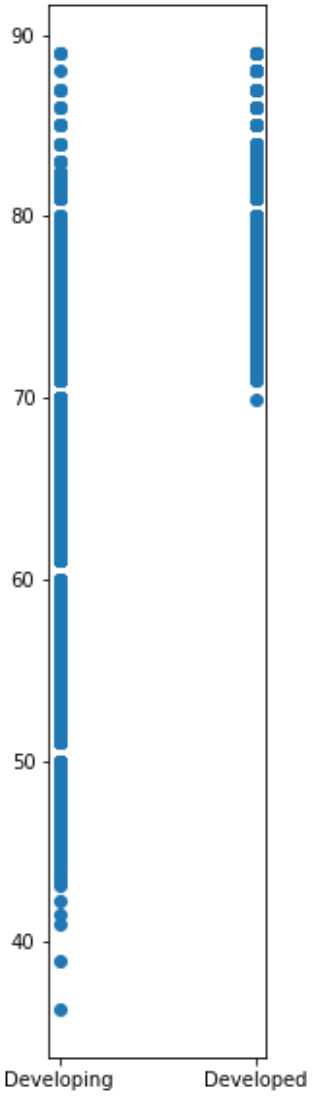
Tenim, en aquesta gràfica, els diferents anys representats en l'eix x, i els valors referents a l'esperança de vida (en anys) de tots els països a l'eix y.

Com es pot veure, s'ha aconseguit a nivell global aumentar els valors màxims de l'esperança de vida i, el que podria ser més important, s'ha aconseguit augmentar l'esperança de vida dels països amb els valors més baixos, encara que en l'any 2010 es pot veure un valor fora de sèrie (*outlier*), ja que hi ha un país amb un valor molt llunyà a la distribució de la resta del món.

Hem pogut veure els diferents valors de l'esperança de vida al llarg d'aquests 16 anys per a cada país individualment (no hem copiat la imatge perquè és difícil de llegir en un document vertical). Podem comprovar que la distribució dels punts corresponents a aquells països econòmicament més desenvolupats es concentren més en els valors alts d'esperança de vida. Per posar-ne exemples, països desenvolupats com Italia i Japó, tenen els punts concentrats en valors més alts d'esperança que Iraq, exemple de país en desenvolupament.

Tornant a l'evolució de l'esperança de vida al llarg dels anys de la mostra, podem veure que el país responsable del *outlier* que trobem en l'any 2010 es tracta de Haití, coincidentment l'any que el país va patir els devastadors terratrèmols, que van ser la causa de moltes víctimes mortals entre la població.

Per fer una última visió de la distribució de l'esperança de vida al món, visualitzarem aquests valors comparativament entre països desenvolupats i no desenvolupats.

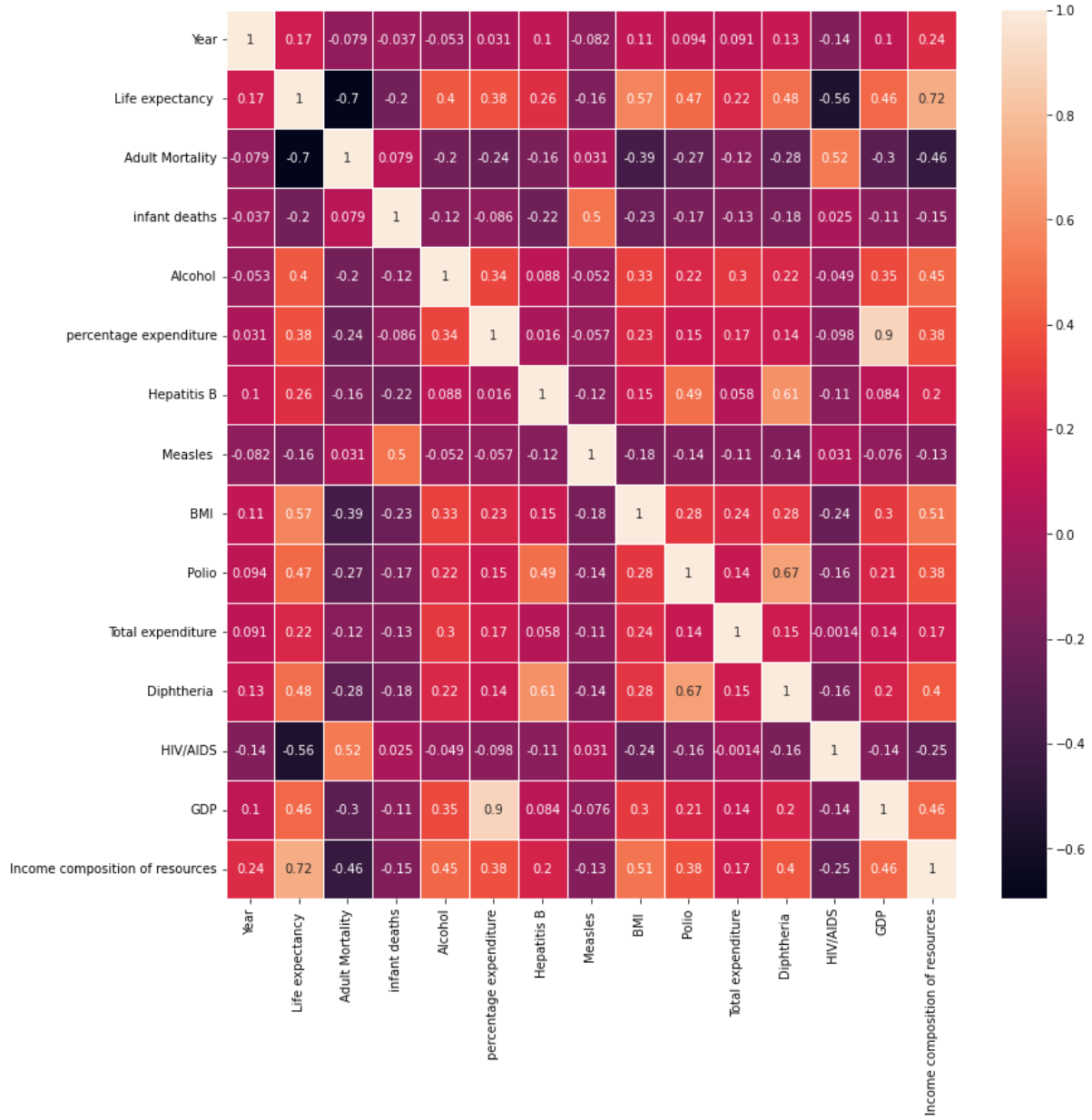


Amb aquesta gràfica i, tal i com ja havíem vist, existeix una clara diferència entre el rang de valors de l'esperança de vida en els països desenvolupats i en desenvolupament.

Com podem veure a la gràfica, encara que existeixen casos on el valor màxim de l'esperança de vida coincideix tan en els països desenvolupats com en els en desenvolupament, es pot apreciar una gran diferència en la concentració total dels

valors, tenint que en el cas dels països desenvolupats l'esperança de vida es concentra molt entre els valors de 70 i 89 anys, mentre que, per altra banda, la distribució en els països en desenvolupament està extensa en un rang de valors molt més gran, de 36 a 89 anys.

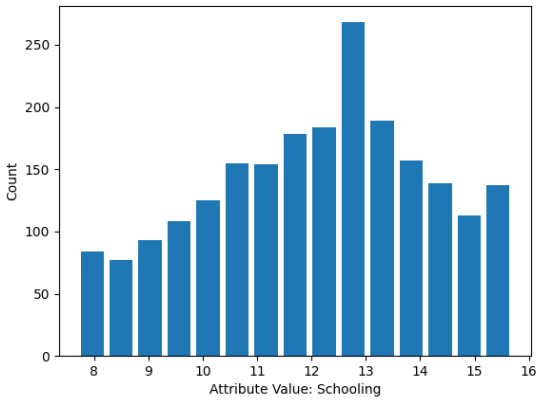
Ara que ja hem vist per sobre la relació i evolució de l'esperança de vida al món amb els atributs no numèrics, continuarem buscant aquells atributs numèrics que tinguin una correlació més gran amb el nostre atribut objectiu.

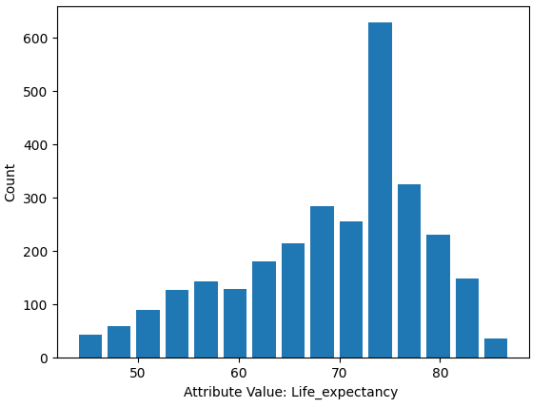


Veiem com l'esperança de vida està força relacionada amb l'índex de desenvolupament humà; com millor qualitat de vida, més esperança de vida, així com també amb el PIB, el consum d'alcohol i les persones immunitzades d'hepatitis B o pòlio. També veiem que són gairebé contraris l'esperança de vida amb la mortalitat adulta o infantil, els morts per VIH o els casos de xarampió ("measles").

Distribució de les dades

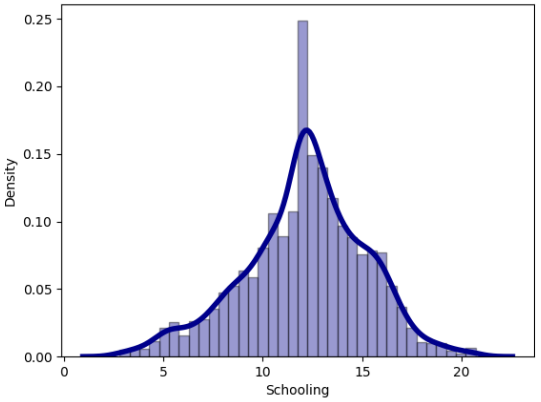
Quan parlem de la distribució de dades ens referim a la col·lecció o puntuació de dades d'una variable concreta. Sobre les dades d'una variable o atribut, podem obtenir les freqüències de cadascun dels seus valors possibles, i aquestes les podem representar en un histograma o gràfic de barres per a veure quins valors són els més freqüents.





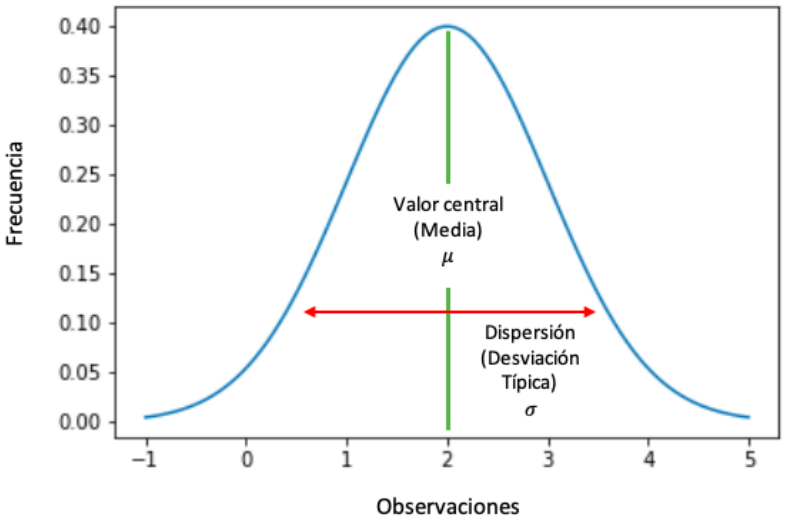
Aquests histogrames representen les freqüències dels atributs "schooling" (ensenyament) i "life expectancy" (esperança de vida) de la nostra base de dades. Un problema dels histogrames és que si volem fer histogrames per a tots els atributs de la nostre base de dades, hem d'escollir un número de particions/barres les quals a vegades no s'adeqüen als nostres atributs com passa en l'histograma de l'atribut "schooling".

També es poden utilitzar gràfics de densitat que són una versió millorada dels histogrames i tenen la mateixa funcionalitat.



Distribució Gaussiana

La distribució gaussiana o distribució normal de les dades és un concepte important quan es vol treballar amb models lineals ja que aquests funcionaran millor. És una distribució de probabilitat contínua, i té una corba en forma de campana que és simètrica des de el punt mig a ambdós costats de l'eix d'abscisses, la qual ens indica la probabilitat sobres els valors d'un atribut en concret.



Sergi Sánchez Hernandez NIU 1417652
Jordi Vila Folgarona NIU 1562542

La distribució Gaussiana de les variables es relaciona amb la distribució dels erros d'aquestes ja que línia de regressió s'ajusta a les dades per tal que la mitjana dels errors sigui zero, i com a errors ens referim a la diferència de les dades amb la línia de regressió. Àixò vol dir que si els atributs tenen una distribució normal, el error respecte a línia de regressió serà molt menor que amb una distribució uniforme.

Trobar atributs amb distribució gaussiana

Per comprovar si les dades d'un atribut/característica tenen una distribució normal, podem fer ús dels histogrames, ja que a partir d'aquests ja podriem deduïr-ho veient si el histograma té forma de campana. Nosaltres hem programat una funció per a que faci un histograma per a cada atribut, i d'aquesta manera poder veure a priori si pot tenir una distribució normal.
A més, hem creat una funció que utilitza la funció normaltest de la llibreria scipy, la qual a partir d'una mostra comprova si té distribució normal. La funció suposa que la mostra rebuda té una distribució normal, i es basa en les proves d'Agostino i Pearson per a produïr una prova general de normalitat, i retorna un estadístic de kurtosistest, i una probabilitat de chi quadrat. Per a que la mostra tingui una distribució guassiana el valor d'aquesta chi quadrada ha de ser superior a 0,05, per lo que la nostra funció retorna el nom de les columnes que compleixen aquesta condició.

```
Atribut: Year Valor de χ2:
0.0
Atribut: Life_expectancy  Valor de χ2:
2.6030540564994 81e-39
Atribut: Adult Mortality Valor de χ2:
4.1553532035811 44e-123
Atribut: infant deaths Valor de χ2:
0.0
Atribut: Alcohol Valor de χ2:
1.259756999654883e-60
Atribut: percentage expenditure Valor de χ2:
0.0
Atribut: Hepatitis B Valor de χ2:
2.5127236652891238e-266
Atribut: Measles  Valor de χ2:
0.0
```

```
Atribut:  BMI  Valor de χ2:
0.0
Atribut: under-five deaths  Valor de χ2:
0.0
Atribut: Polio Valor de χ2:
1.1812279648739384e-258
Atribut: Total expenditure Valor de χ2:
1.3675423681178347e-59
Atribut: Diphtheria  Valor de χ2:
4.822858564853843e-253
Atribut:  HIV/AIDS Valor de χ2:
0.0
Atribut: GDP Valor de χ2:
0.0
Atribut: Population Valor de χ2:
0.0
Atribut:  thinness  1-19 years Valor de χ2:
9.3990600148163 41e-220
Atribut:  thinness 5-9 years Valor de χ2:
5.7535581931175156e-232
Atribut: Income composition of resources Valor de χ2:
5.022789303475789e-34
Atribut: Schooling Valor de χ2:
2.8280768967159593e-10
[]
```

Aquests són els valors resultants de la chi quadrada per a cada atribut. Com es pot veure són valors molt més petits que 0,05 per lo que podem afirmar que cap atribut de la nostre base de dades té una distribució normal.

Normalització i Estandarització de les Dades

Normalització

La normalització de les dades s'utilitza quant a les dades hi han escales variables, i per tant el model generat a la regressió faria les prediccions suposant que les dades tenen una distribució guassiana.
L'objectiu de la normalització és cambiar els valors dels atributs a una escala en comú, sense distorsionar les diferències entre els rangs de valors. D'aquesta manera si l'escala d'un atribut es de 103, i l'escala d'un altre atribut és de 107, els rangs de les dues variables es transformarà a una escala comuna, ja que sinó en una regressió lineal multivariant influiria molt la segona variable.

Estandarització

La estandarització de les dades assumeix que la distribució de les dades és guassiana, encara que per a fer la estandarització no té per que ser cert, però és més eficaç ja que l'algorisme de regressió lineal suposa que les dades estan normalment distribuïdes.
La estandarització té com a objectiu que la mitjana de les mostres sigui casi igual a 0, i la desviació estàndar igual a 1, per a aconseguir això ha de escalar els valors dels atributs per a que estiguin en un mateix rang, amb les mateixes proporcions dintre de cada variable.

Normalització a la nostre base de dades

Com hem vist a l'apartat de les distribucions de les dades, hem obtingut una sèrie d'atributs que no tenen una distribució normal, així que hem volgut provar a utilitzar alguna funció per a intentar normalitzar aquests atributs. Hem provat les funcions de MinMaxScaler() i Normaltest(), encara que per a les dues funcions ens retornava el mateix resultat per a les variables abans de ser normalitzades i quant s'havien normalitzat.

Estandaritzar a la nostre base de dades

Per a estandaritzar les dades hem utilitzat la funció standarize donada ja a l'enunciat de la pràctica, la qual ens reescala les dades restant la mitjana de la columna i dividint per la desviació típica. Donant-nos com a resultat que la mitjana dels valors per a cada variable era 0, i la desviació típica d'aquestes era 1. (proper a 1 i proper a 0)

Conclusió

Com a conclusió d'aquesta pràctica, hem descobert algunes llibreries de python com sklearn, matplotlib, les quals ens han ajudat a poder analitzar millor les característiques de la nostra base de dades, i poder visualitzar la informació de forma ràpida i directe. També hem pogut treballar sobre una base de dades real, la qual hem hagut d'aprendre a seleccionar bé les dades i el per què escollim aquestes dades.