Universitat Autònoma de Barcelona

Facultat de Ciències

Pràctica 1

Aprenentatge Computacional

MatCAD

Pablo Ruiz, Clara Sorolla i Maria Pallejà

1565555, 1569191, 1570129

Índex

Índex de figures

Índex de taules

1 Introducció

L'esperança de vida és un paràmetre amb molta variabilitat arreu del món. Mentre a Europa cada cop hi ha una població més envellida, les regions amb menys recursos d'Àfrica no acostumen a passar dels quaranta anys. Per aquest motiu és interessant saber què causa una alta o baixa esperança de vida.

En el següent informe s'ha estudiat un conjunt de dades[1] amb la intenció d'esbrinar quins atributs estan més relacionats amb l'esperança de vida i quins models permeten predir-la millor.

L'objectiu d'aquesta pràctica és aplicar models de regressió per tal d'analitzar atributs i veure quins són els més representatius, avaluar correctament l'error dels models, visualitzar les dades i finalment, aplicar el procés de descens del gradient. A partir d'això aconseguim ésser capaços d'aplicar tècniques de regressió en casos reals, validar els resultats obtingut i fomentar la nostra capacitat per presentar els resultats de l'aprenentatge computacional.

2 Apartat C: Anàlisi Descriptiu

2.1 Atributs

El nostre *dataset* compta amb 22 atributs. A continuació mostrem una taula amb tots els atributs, el tipus de dades que són i la quantitat de `Nans` que contenen en el *dataset* original sense modificar.

Atribut	**Tipus**	**Num Nans**
Life expectancy (LE)	float64	10
Infant deaths (ID)	int64	0
Percentage expenditure (PE)	float64	0
Measles (M)	int64	0
BMI (BMI)	float64	34
Under-five deaths (UF)	int64	0
Diphtheria (D)	float64	0
HIV/AIDS (HIV)	float64	0
Thinness 10-19 years (T10)	float64	34
Thinness 5-9 years (T5)	float64	34
Country (C)	object	0
Year (Y)	int64	0
Status (S)	object	0
Adult Mortality (AM)	float64	10
Alcohol (A)	float64	194
Hepatitits B (H)	float64	553
Polio (P)	float64	19
Total expenditure (TE)	float64	0
GDP (GDP)	float64	448
Population (P)	float64	652
Income Composition of resources (ICR)	float64	167
Schooling (SCH)	float64	163

Taula 1: Atributs de la base de dades original

[1] https://www.kaggle.com/kumarajarshi/life-expectancy-who

En la taula anterior podem veure com l'atribut *Population* té una gran quantitat de dades buides (`Nans`). Per determinar si el podem descartar del nostre estudi, hem mirat la correlació amb la resta d'atributs i, en concret, amb el possible atribut objectiu: *Life Expectancy*.

Per tractar els altres `Nans` que hem trobat fem servir diferents estratègies i veient els resultats determinarem quina és la millor. Creem dos `datasets` alternatius. En el primer modifiquem els valors `Nans` per la mitja de les mostres i en el segon simplement eliminem les files que contenen `Nans`. Una altra estratègia que hem provat però hem descartat degut als resultats dolents obtinguts, ha sigut omplir els atributs amb valors d'una distribució gaussiana.

2.2 Correlació entre atributs

Visualitzem com són les correlacions entre atributs.

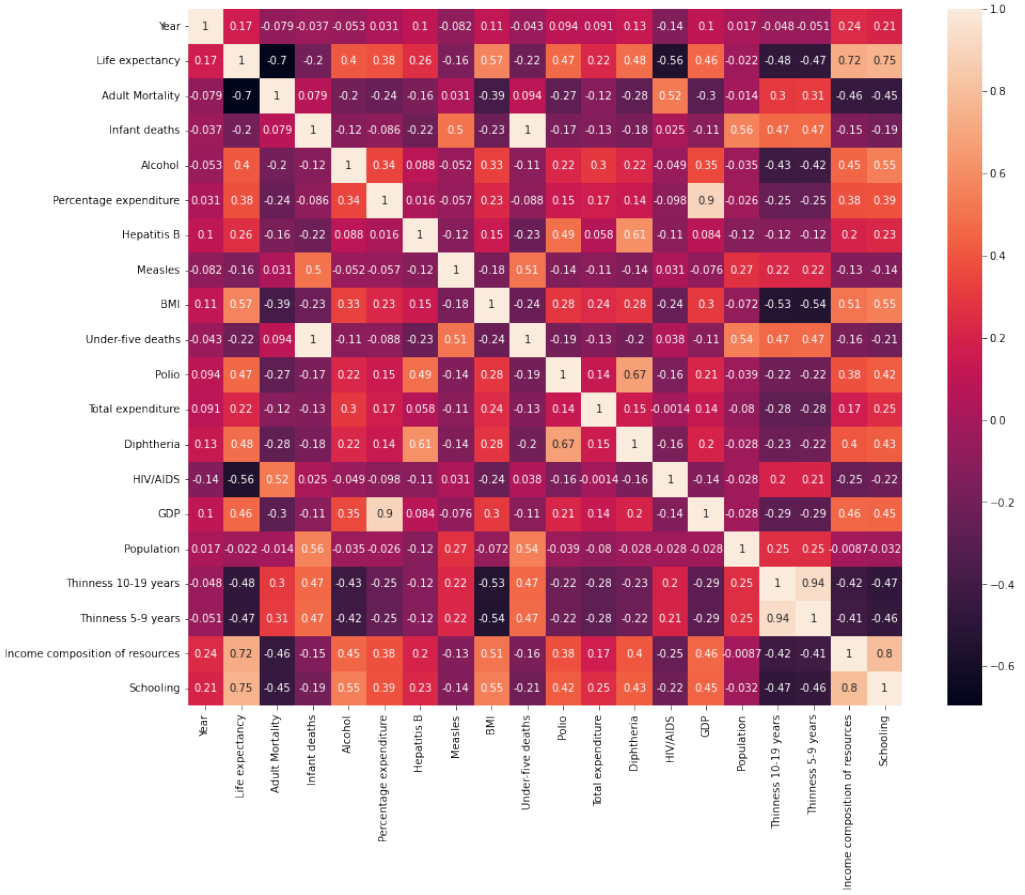

Figura 1: Heatmap de les correlacions entre atributs

Veient com és la correlació entre *Population* i la resta d'atributs, en especial respecte al possible atribut objectiu (*Life Expectancy*), decidim descartar aquesta variable del nostre estudi.

2.3 Distribució dels atributs

A continuació visualitzem els histogrames dels atributs per veure com és la distribució de les seves dades.

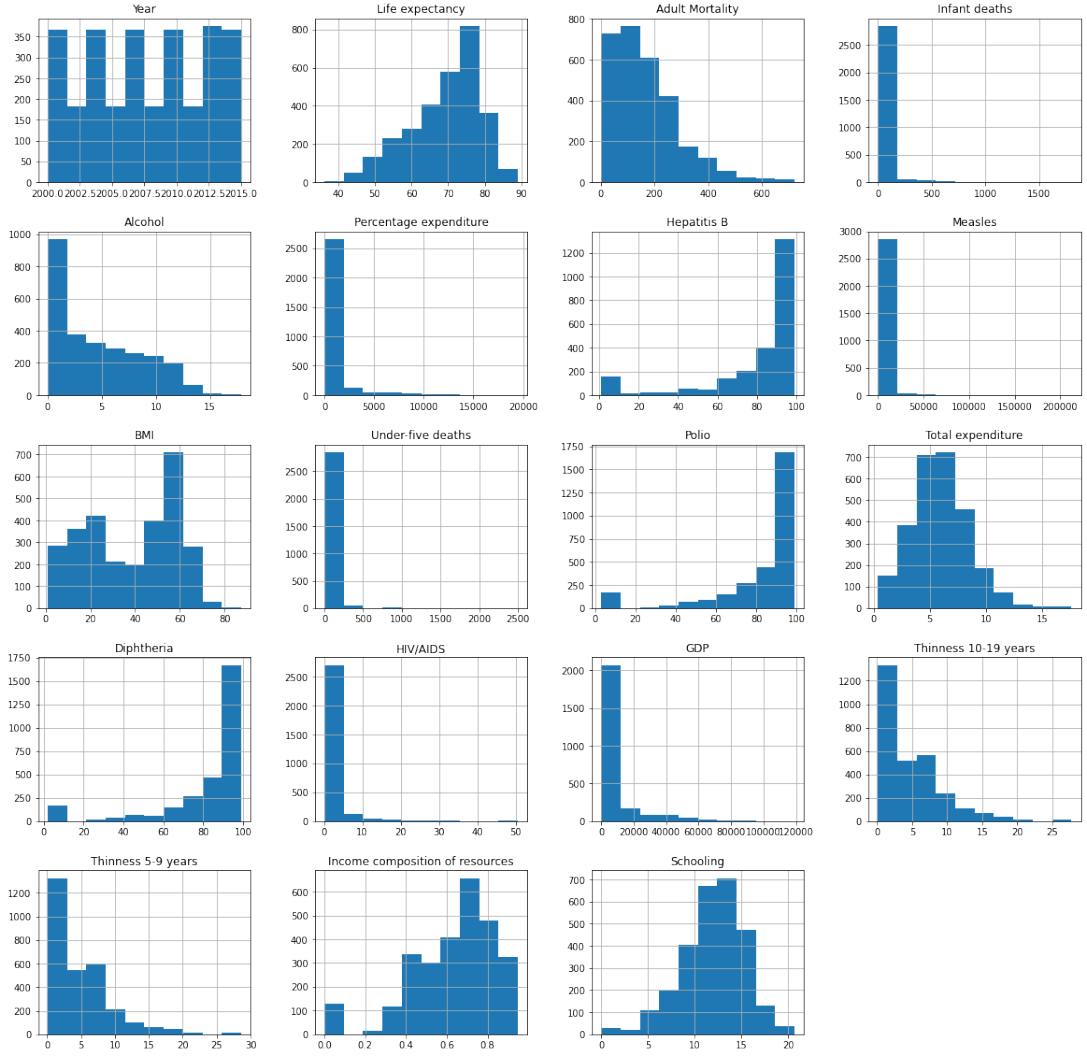

Figura 2: Histogrames dels atributs

Veiem per la forma del histograma que *Schooling* podria seguir una distribució gaussiana. Per veure si això és cert, i també per buscar si hi ha algun altre atribut que ho sigui, hem estudiat més a fons les nostres dades. Hem conclòs que *Schooling* i *Polio* tenen una distribució gaussiana.

Com que ens interessa saber si *Life Expectancy* podria ser un bon atribut objectiu farem un `pairplot` dels atributs que tinguin una correlació amb aquest més alta de 0.6, ja que si no tindríem massa gràfics i no podríem fixar-nos bé amb les dades. Aquest el fem amb el *dataset* inicial.

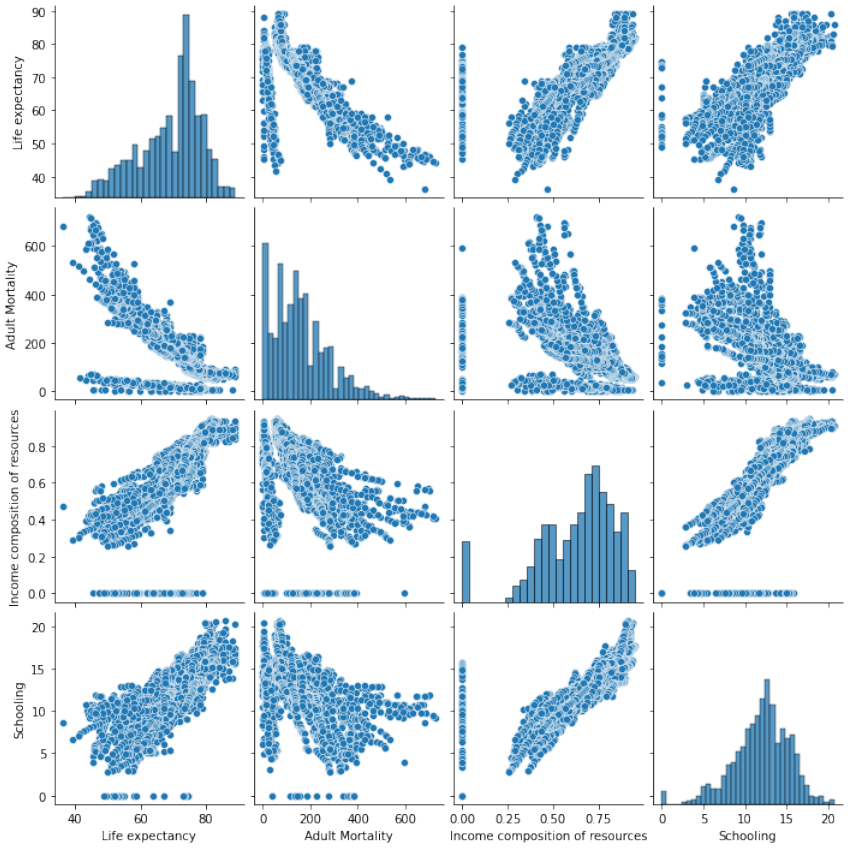

Figura 3: Pairplot dels atributs amb una correlació respecte *Life expectancy* més alta de 0.6.

Observem que hi ha diversos punts en fila recta al 0 i també els valors `Nans`. Com que aquests valors podrien causar problemes en el nostre estudi hem decidit afegir una nova estratègia a part de les dues anteriors ja mencionades. Veiem que *Income Composition of Resources* té una correlació alta amb el nostre possible atribut objectiu *Life Expectancy*, aquest té un conjunt de valors que valen 0 que no tindria massa sentit que existeixin ja que es tracta d'un percentatge dels ingressos totals agregats en una àrea. A més, aquest valor està directament relacionat amb l'atribut *Schooling*, el qual també té valors a 0 que no tenen sentit. Per tant, hem considerat que eliminar els 0 de *Income Composition of Resources* seria una bona estratègia a seguir per poder fer un millor anàlisis. Per fer el nostre estudi doncs utilitzarem quatre `datasets` diferents: els dos anteriors ja mencionats, un que tingui els valors de `Nans` establerts com la mitjana i sense els valors de 0 de *Income Composition of Resources*, i finalment, un que hagi eliminat totes les files de `Nans` i les files de 0 de *Income Composition of Resources*.

2.4 Atribut objectiu

Com hem vist, l'atribut *Life expectancy* destaca sobre la resta quant a la correlació amb la resta d'atributs. A més, és un atribut interessant a predir ja que és un paràmetre que ens pot dir com és el nivell de vida d'un territori; com hem pogut veure durant el llarg de la història els avenços tecnològics, mèdics i altres que afecten directament en la nostra vida han permès que aquest paràmetre creixés. És per nosaltres una evidència que qualsevol civilització està interessada en allargar la vida de la seva població.

3 Apartat B: Anàlisi Predictiu

En aquest apartat s'han estudiat els diferents models possibles per tal de predir el valor del nostre atribut objectiu segons aquells atributs que tenen més causa o efecte en els resultats que volem predir. Per tal d'estudiar els diferents models hem fet servir diversos *datasets* segons els canvis que hem vist necessaris fer al estudiar les nostres dades inicials. Els *datasets* estudiats han estat:

- `df_mean`: Els valors NaN han sigut substituits per la mitja del atribut corresponent.
- `df_mean_nozero`: Els valors NaN han sigut substituits per la mitja del atribut corresponent i s'han eliminat els zeros que apareixien en el atribut *Income composition of resources* ja que aquests valors nuls eren causats per altres atributs com *Schooling* on no tenia sentit que fos nul, i això causava problemes més endavant.
- `df_nonan`: Els valors NaN han sigut eliminats.
- `df_nonan_nozero`: Els valors NaN han sigut eliminats i s'han eliminat els zeros que apareixien en el atribut *Income composition of resources* ja que aquests valors nuls eren causats per altres atributs com *Schooling* on no tenia sentit que fos nul, i això causava problemes més endavant.

3.1 Regressions lineals amb una variable

En aquesta secció s'ha calculat el error quadràtic mitjà del regressor lineal per cada un dels atributs de la base de dades, determinant aquell atribut pel qual el MSE és més baix. Per fer-ho, s'ha creat un model per cada atribut, predint així el valor de *Life Expectancy* a partir de cadascún dels altres atributs.
En la següent figura podem observar el resultat d'aplicar la regressió lineal a alguns dels nostres atributs del dataset.

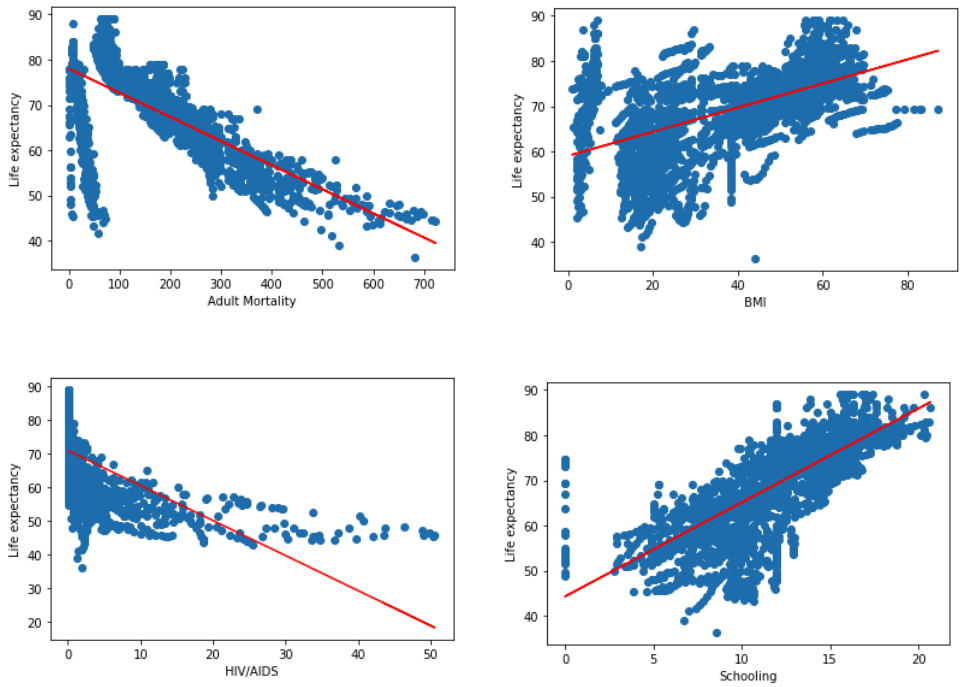

Figura 4: Regressió Lineal d'alguns atributs respecte *Life expectancy*

A continuació es mostren 4 figures que mostren els valors de MSE i el R^2 per cadascun dels regressors creats per els diferents `datasets` estudiats:

3.1.1 Dades sense normalitzar

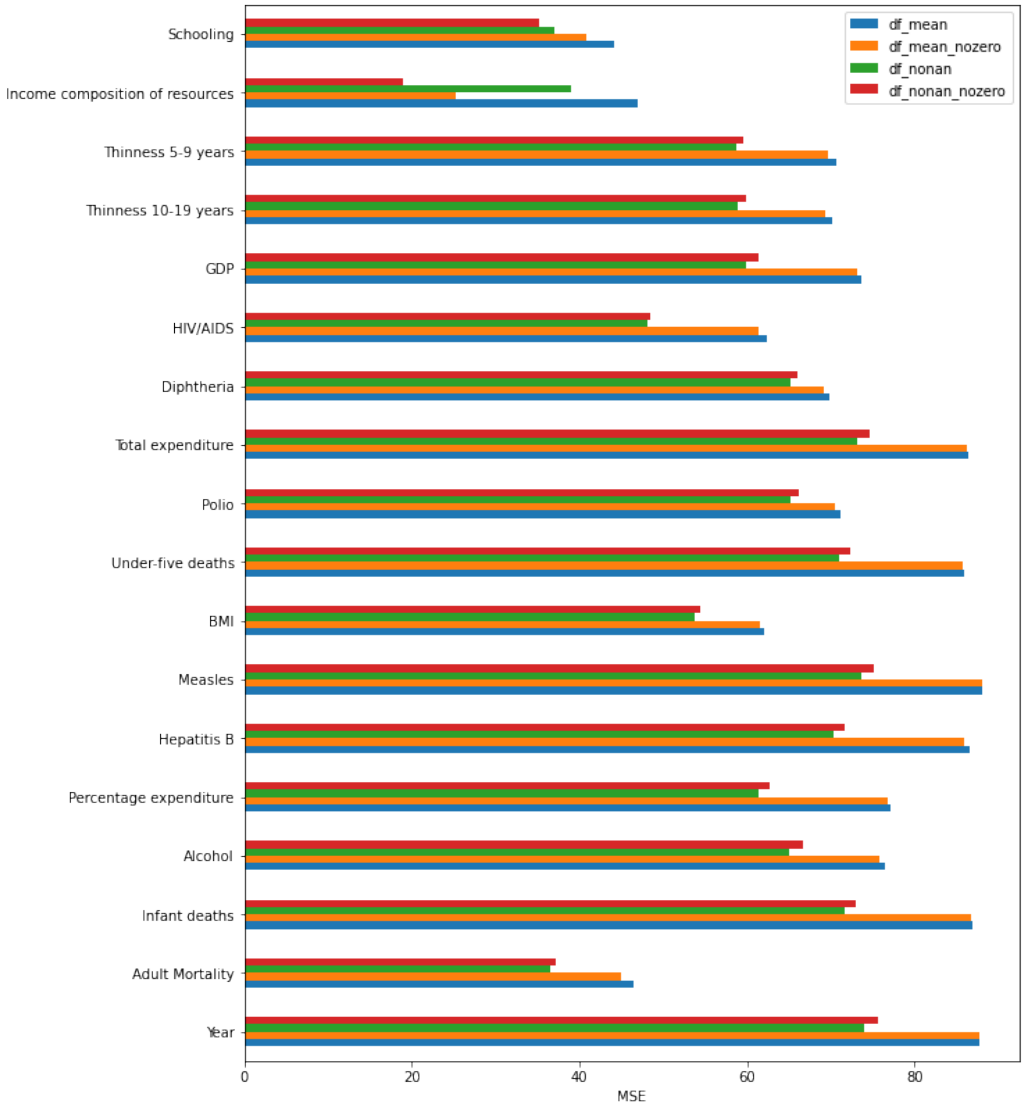

Figura 5: Valors de MSE de cada regressor per els diferents datasets amb les dades sense normalitzar

Observem que els valors de l'error quadràtic mitjà solen ser menors en els datasets `df_nonan_nozero` i `df_nonan` ja que és on no hem fet canvis en el `dataset` original, sinó que simplement ens hem quedat amb les mostres que realment sabiem que eren fiables. Els atributs que tenen menor error quadràtic mitjà en els regressors són *Schooling, Income Composition of Resources i Adult Mortality.*

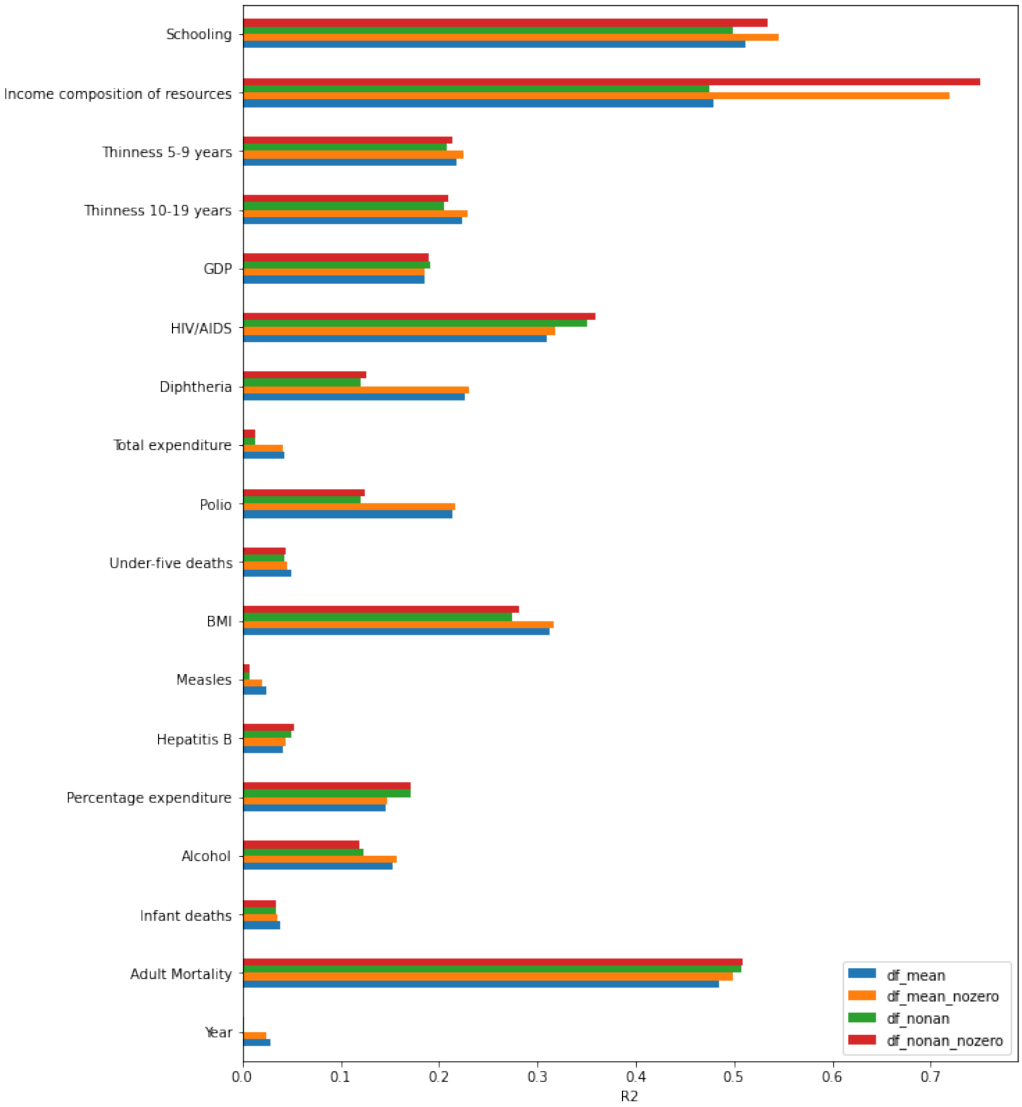

Figura 6: Valors de R^2 de cada regressor per els diferents datasets amb les dades sense normalitzar

Tal i com sabem, els valors de MSE tenen una relació amb el R^2 del regressor, per aquesta raó amb la figura 5 acabem de confirmar que els atributs més relacionats amb *Life expectancy* són *Schooling, Income Composition of Resources i Adult Mortality* al voltant d'un valor de 0.5 en el regressor. Observem que arriba fins i tot a un R^2 de 0.7 en el cas de *Income Composition of Resources* fent servir el dataset `df_nonan_nozero` i el `df_mean_nozero`.

3.1.2 Dades normalitzades

Ara observem els resultats tenint en compte que s'han normalitzat les dades, esperant així, uns millors resultats.

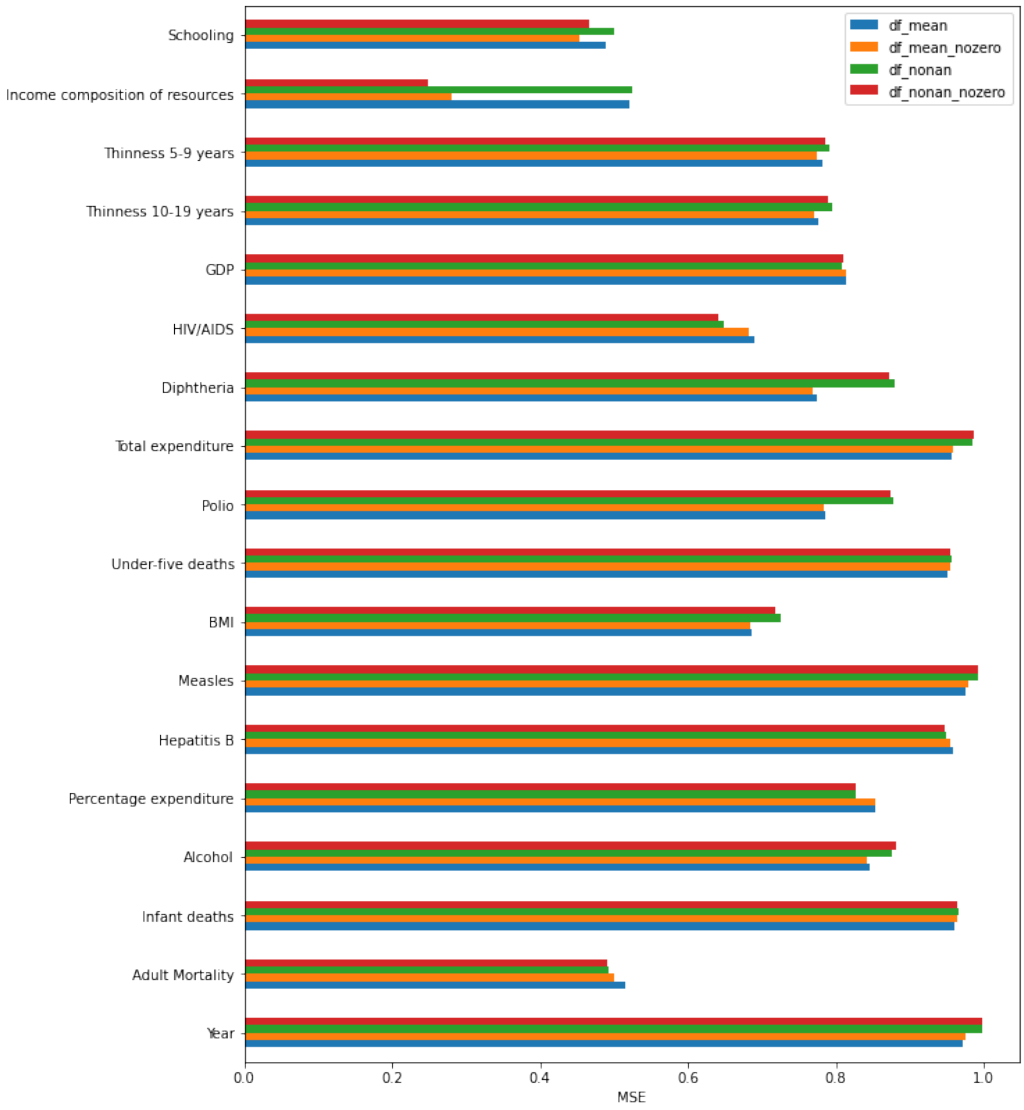

Figura 7: Valors de MSE de cada regressor per els diferents datasets amb les dades normalitzades

Observem que els atributs amb menor error quadràtic mitjà segueixen sent els mateixos que anteriorment, però veiem que ara el rang de valors és molt menor, tant sols està entre 0 i 1 degut a la normalització de les nostres dades. En canvi, ara podem veure que no hi ha cap `dataset` que destaqui amb millors resultats, tots retornen valors bastant similars.

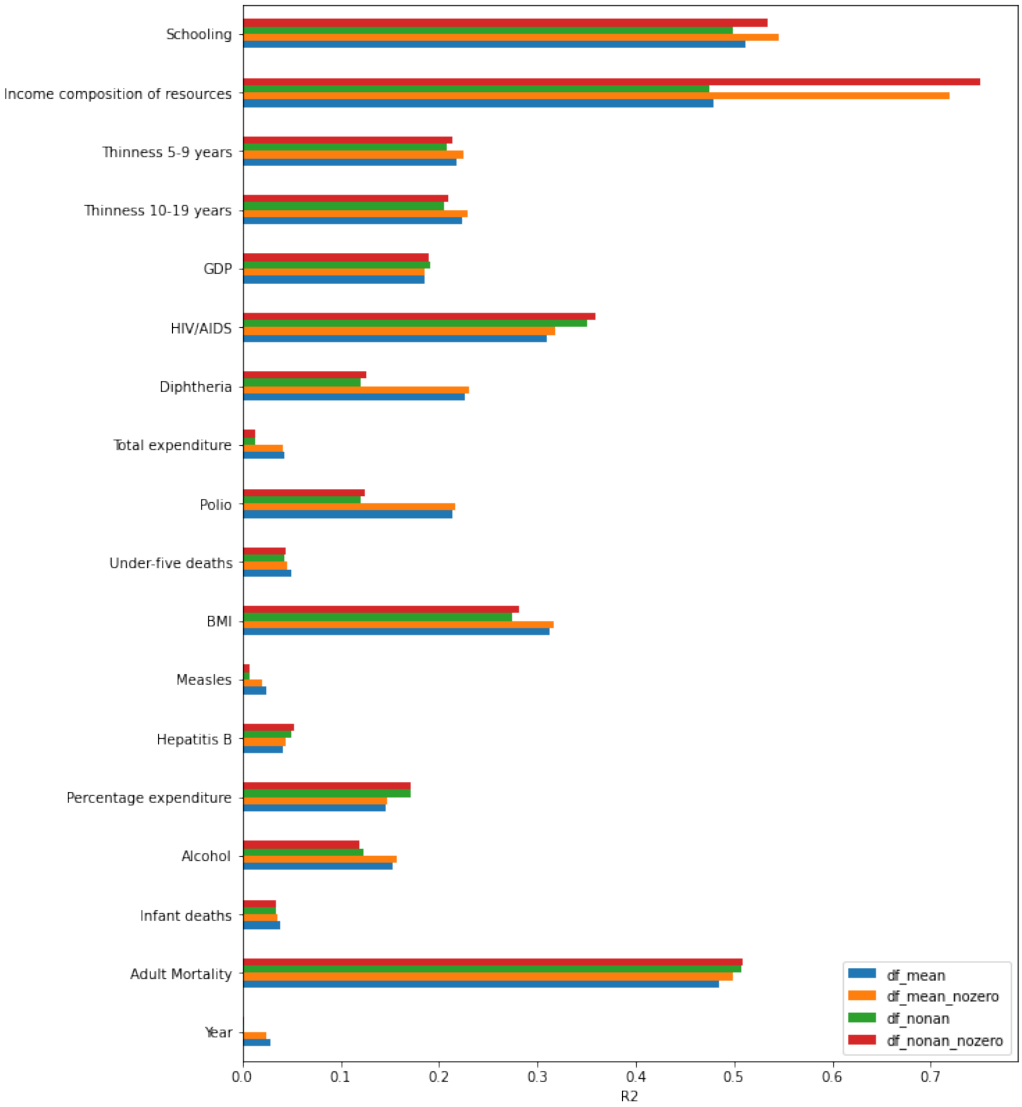

Figura 8: Valors de R^2 de cada regressor per els diferents datasets amb les dades normalitzades

Finalment, en el cas dels valors de R^2 observem que tots segueixen sent bastant similars als obtinguts sense normalitzar les dades. Observem que el millor atribut segueix sent *Income Composition of Resources* fent servir el dataset `df_nonan_nozero` i el `df_mean_nozero` arribant així, fins a un 0.7 de R^2.

3.2 Models lineals amb diverses variables

Seguidament, hem estudiat diferents models lineals per veure quins són els atributs més importants per tal de predir el nostre atribut objectiu. En aquest apartat es mostren els millors models lineals trobats per cadascun dels datasets estudiats tenint en compte les dades quan no estan normalitzades i quan ho estan. Els models mostrats a continuació s'han creat fent servir Regressió Lineal amb 3 atributs per tal de predir *Life Expectancy*.

3.2.1 Dades sense normalitzar

En la següent taula els atributs fets servir en el model apareixen escrits amb les sigles mencionades anteriorment (2.1).

Dataset	**Millor model lineal**
`df_mean`	- 0.0278(AM) - 0.4495(HIV) + 1.4653(SCH)
`df_mean_nozero`	- 0.0197(AM) - 0.3509(HIV) + 38.5631(ICR)
`df_nonan`	- 0.0239(AM) - 0.4117(HIV) + 1.5639(SCH)
`df_nonan_nozero`	- 0.0134(AM) - 0.3944(HIV) + 38.2523(ICR)

Taula 2: Millors models lineals amb dades no normalitzades

Dataset	R^2	**MSE**
`df_mean`	0.70475636	27.58822314
`df_mean_nozero`	0.81618795	16.10351736
`df_nonan`	0.73200805	19.12957691
`df_nonan_nozero`	0.87487528	9.76780489

Taula 3: R^2 i error quadràtic mitjà dels models de la Taula1.

En les taules anteriors podem observar com en tots els models s'utilitzen quasi sempre els mateixos atributs, els quals són els més importants per tal de predir el nostre atribut objectiu *Life Expectancy*. Aquests atributs són: *Adult Mortality, HIV/AIDS, Schooling, Income Composition of Resources*.
Mirant els resultats obtinguts, veiem que el model que dona l'error quadràtic mitjà més baix és el creat a partir del dataset `df_nonan_nozero`, això pot ser degut a que no estem creant dades que no tenim a partir d'altres, sinó que directament les hem esborrat i hem fet servir aquelles que de veritat sabem que són correctes.

3.2.2 Dades normalitzades

Dataset	**Millor model lineal**
`df_mean`	- 0.3112(AM) - 0.2829(HIV) + 0.5154(SCH)
`df_mean_nozero`	- 0.2579(AM) - 0.2074(HIV) + 0.6356(ICR)
`df_nonan`	- 0.3400(AM) - 0.2907(HIV) + 0.4958(SCH)
`df_nonan_nozero`	- 0.2064(AM) - 0.2515(HIV) + 0.6593(ICR)

Taula 4: Millors models lineals amb dades normalitzades

Dataset	R^2	MSE
`df_mean`	0.77395118	0.22371597
`df_mean_nozero`	0.83477990	0.15527841
`df_nonan`	0.75902434	0.24143469
`df_nonan_nozero`	0.88819048	0.11286617

Taula 5: R^2 i error quadràtic mitjà dels models de la Taula3.

En les taules anteriors es pot veure que encara que haguem normalitzat les dades, els atributs que s'han fet servir en els models no han canviat, segueixen sent els més importants i el dataset que té menor error quadràtic segueix sent el `df_nonan_nozero`. Per altra banda, podem observar com en tots els casos el valor de R^2 ha augmentat, per tant sabem que els models creats amb dades normalitzades són millors. Un aspecte important que és causat per les nostres dades és el valor dels pesos dels nostres atributs, aquests són molt més similars entre ells quan tenim dades normalitzades ja que estan equilibrats i no hi ha cap que sigui molt més gran com podem observar que passava en la taula1. Degut a això, també veiem que els valors de MSE han reduït molt, ja que aquest també està equilibrat amb els pesos i dades.

3.3 Models quadràtics

Els models polinòmics poden aportar a la regressió una dimensionalitat extra. Hem estudiat el comportament de la regressió quadràtica per predir el paràmetre *Life expectancy*. Ho hem fet usant la classe `PolynomialFeatures` de la llibreria *sklearn*. Aquesta classe permet transformar la matriu X creant noves columnes. Aquestes nous atributs són combinacions quadràtiques dels atributs del dataset. Per exemple, es crea la columna BMI $\times$ BMI o Schooling $\times$ Measles.

Un cop tenim el nou conjunt de dades amb els nous atributs no lineals, procedim com en la secció anterior. Ara hem seleccionat els atributs més discriminants usant un `SequentialFeatureSelector` indicant que seleccioni 15 atributs.

El resultat ha estat que l'algorisme de selecció dels atributs més discriminants per fer un entrenament posteriorment, no selecciona[2] cap atribut quadràtic. Per tant, considerem que un model quadràtic no és viable per al nostre conjunt de dades i atribut objectiu.

3.4 PCA

Aquests conjunt de dades és força adient per aplicar un PCA. Hem aplicat el PCA de la llibreria *skelearn* amb 3 components. Amb la nova matriu X proporcionada pel PCA hem entrenat un model per comparar-lo amb els models sense transformar el conjunt de dades. El model ha estat

$$y = 0.002 - 0.784x_1 + 0.278x_2 - 0.134x_3$$

L'error quadràtic mitja d'aquest model és 0.303 i el $R^2 = 0.712$. Concloem que, encara que era prometedor, el PCA no potencia les dades per a obtenir un model millor. Les mètriques són pitjors que la resta de models sense PCA.

[2]L'atribut no lineal més discriminant que resulta és ICR $\times$ Schooling.

4 Apartat A: Descens del gradient

El descens de gradient és un algorisme d'optimització que ens permet trobar el mínim local d'una funció de cost donada. En el nostre cas ens ajuda a trobar els pesos més òptims d'un regressor multivariat. Per fer-ho la funció de cost que utilitzem és l'error quadràtic mitjà. Després d'haver fet varies proves, el model que ens va donar millor resultats el vam trobar reduint el nostre `dataset` per tal de predir *Life Expectancy* als atributs que després del nostre estudi hem vist que tenien més importancia: *Adult Mortality, HIV/AIDS i Income composition of Resources.* Aquest ens va donar els següents pesos que corresponen amb els atributs ja mencionats:

$$w0 : 0.004045936726299325, w1 : -0.198, w2 : -0.249, w3 : 0.672$$

L'error amb aquesta funció és de 0.12784488469729643 i el seu valor de R^2 és de 0.8584877770174699.

El *learning rate* que vam utilitzar en la majoria de les nostres proves és de 0.05, ja que vam veure que ens donava els millors resultats; però vam deixar per defecte un valor de 0.01 a la nostra funció `Regressor` ja que vam pensar que seria un valor adequat per començar a fer anàl·lisi quan tinguessim dades noves.

4.1 L2 regularitzador

Per tal d'evitar l'overfitting utilitzem la tècnica de *l2-regularitzation.* El que fem és afegir a la funció de cost el valor dels pesos al quadrat multiplicats per un coeficient que nosaltres escollim que varia entre 0 i 1. Com més proper a 1 sigui, hi ha més possibilitats de tenir underfitting; contràriament si és proper a 0 hi ha més possibilitats de tenir overfitting. Després de fer algunes proves amb diferents `l` hem vist que amb un valor de 0.5 ens dona molt bons resultats.

4.2 Visualització del Descens de Gradient en dos i tres dimensions

Per tal de poder validar el regressor trobat pel descens del gradient visualment, hem estudiat dos models: de recta i de pla.

El model de recta l'hem trobat utilitzant l'atribut *Income Composition of Resources* ja que és el millor atribut per predir *Life Expectancy* com ja hem pogut veure durant la nostra pràctica. Si apliquem un descens de gradient ens dona la següent funció:

$$y = -0.0010263067198609229 + 0.887x$$

Té un error de 0.22653777558099855 i un valor de R^2 de 0.7483696339878237. A continuació hem visualitzat el model amb les dades:

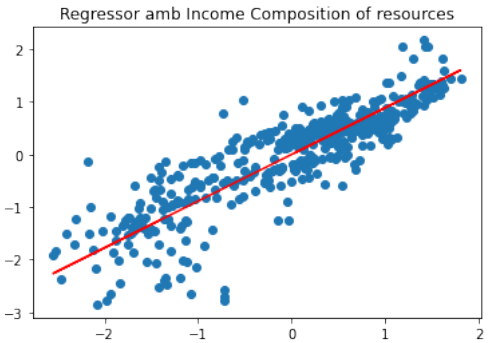

Figura 9: Visualització del model de recta trobat pel nostre descens del gradient

Per altra banda, el model de pla l'hem trobat utilitzant els dos atributs que hem vist que eren millors atributs: *Adult Mortality*(x_1) i *Income composition of Resources*(x_2) per predir el nostre atribut objectiu. Quan hem aplicat el descens de gradient ens ha trobat el següent regressor:

$$y = 0.0002802095223600 43 - 0.344x_1 + 0.674x_2$$

Aquest ens dona un error de 0.1812821943740389 i un valor R^2 de 0.8125350287848616. Com podem comprovar els nostres resultats milloren respecte el model de recta. Si visualitzem les dades podem veure com el model de pla s'adapta millor a les nostres dades:

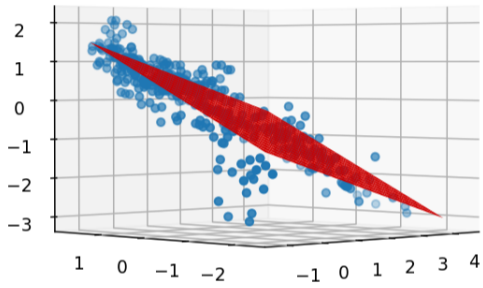

Figura 10: Visualització del model de pla trobat pel nostre descens del gradient

4.3 Descens de gradient amb Polynomial Features

Sigui a = Adult Mortality, h = HIV/AIDS, i = Income composition of resources; el descens del gradient amb atributs no lineals resulta en el següent model (aproximació amb un decimal):

$$y = -0.2a + 0.4h + 0.8i - 0.17a^2 + 0.08ah - 0.08ai - 0.03h^2 + 0.3hi + 0.09i$$

Veiem que és un model excessivament complex.

4.4 Comparativa amb *sklearn*

La nostra implementació de l'algorisme d'optimització del descens del gradient no té res a envejar a la implementació de la llibreria *sklearn*.

Quant al rendiment, és a dir, uns models capaços de predir efectivament, el nostre descens del gradient computa models quasi iguals als de la classe `SGDRegressor`.

Ara bé, quant al temps de còmput la llibreria surt vencedora. La nostra implementació no té gaire capacitat d'emmotllar-se a les dades d'entrada, és a dir, la convergència del nostre descens del gradient falla sovint. Cal trobar els hyper-parametres adequats, ja que si no pot passar que no convergeixi. A més a més, la implementació de la llibreria és ben segur que estigui optimitzada amb parts de codi en C.

Hem fet que els dos models de regressió entrenin un conjunt de dades reduït amb sols tres variables. Si observem les mètriques veiem que la nostra implementació supera per poca la de la llibreria. L'error quadràtic mitjà i el R^2 de la nostra implementació dona:

$$\text{MSE} = 0.128 \qquad\qquad R^2 = 0.858$$

En canvi, en la de la llibreria:

$$\text{MSE} = 0.156 \qquad\qquad R^2 = 0.848$$

No és una millora significativa com per a decantar-se per un dels dos ja que el conjunt de dades podria determinar en gran mesura la capacitat dels regressors.

5 Conclusions

Els models de regressió són una de les eines més fonamentals tant de l'estadística com de l'aprenentatge computacional. Resol una de les tasques més recurrents en la vida real: la predicció. Un bon model de regressió és capaç de predir una variable dependent a partir d'altres variables. També permet inferir relacions causals entre variables.

En aquesta pràctica hem aplicat els models de regressió a un conjunt de dades complex. El *dataset* conté mostres de diversos països en uns anys concrets amb diferents indicadors. L'indicador més interessant que hem trobat, i que posa nom al conjunt de dades, és l'esperança de vida (*Life expectancy*). Durant l'anàlisi hem tractat de predir aquesta variable a partir d'altres com l'escolarització o la mortalitat infantil.

En el conjunt de dades original hem trobat moltes dades buides o nul·les que o bé hem reemplaçat per la mitja de la variable, o bé hem eliminat. També hem trobat que la variable població no té una correlació significativa amb l'esperança de vida. Per aquest motiu, l'hem eliminat abans de dur a terme la regressió. En general, el *dataset* original té una bona predisposició a ser analitzat, ja que hem trobat algunes variables amb una distribució gaussiana.

Abans d'entrenar els models de regressió hem creat quatre variacions del conjunt de dades original tractant de manera diferent les dades buides. Ara bé, les variables tenen escales molt variades i les hem normalitzat.

Un cop hem posat a punt el conjunt de dades, hem fet servir el model de regressió lineal (`LinearRegression`) de la llibreria *sklearn*. Hem entrenat un model per a cada variable, és a dir, hem desenvolupat tantes regressions com variables independents. D'aquesta manera, el model és una recta i el podem visualitzar i entendre millor. El conjunt de dades en el qual hem substituït les dades buides per la mitja de la variable ha estat el millor.

També hem entrenat un model de regressió quadràtic el qual no ha tingut èxit i un model de regressió amb tres variables independents. Hem seleccionat les que més correlació tenen amb l'esperança de vida: *Adult Mortality*, *HIV/AIDS* i *Income composition of resources*. Quan en comptes de fer-ho manualment, hem usat un selector seqüencial de variables hem obtingut les mateixes variables. Altrament, hem provat d'efectuar un PCA amb tres components abans d'entrenar la regressió lineal, però no ha millorat els resultats.

Per últim, hem implementat l'algorisme del descens del gradient per fer regressió de manera manual. Hem obtingut resultats molt satisfactoris i similars a l'implementació de la regressió lineal de la llibreria *sklearn*. En el *plot* final es pot apreciar la gran precisió del model de regressió final.

Podem concloure que hem estat capaços de desenvolupar un model de regressió amb un alt rendiment que permet predir l'esperança de vida a partir de variables com l'escolarització.