

**Aprenentatge Computacional**

**Pràctica 1: Regressió**

**Guillem Martínez Arjona – 1563566**

[Introducció al dataset a tractar: 2](#_Toc1191285679)

[Definició dels atributs: 4](#_Toc784375927)

[Neteja del dataset: 5](#_Toc450673109)

[Estratègies per omplir valors nuls: 8](#_Toc1299331748)

[Primeres regressions: 13](#_Toc787102765)

[Anàlisi de components principals: 17](#_Toc1046003458)

[Conclusions: 18](#_Toc153044517)

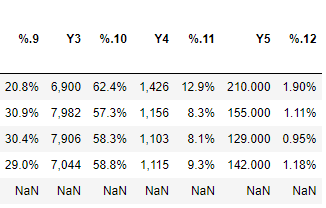
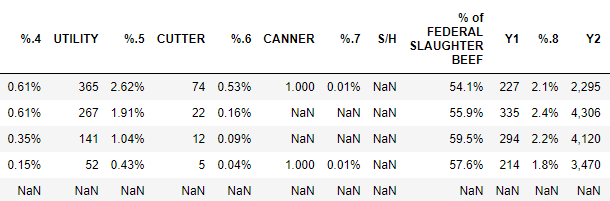
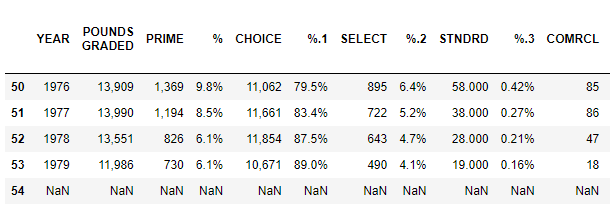
[GitHub: 18](#_Toc2096456678)

# Introducció al dataset a tractar:

El dataset sobre el que treballarem i farem aquesta pràctica, per tal de comprendre i aprofundir en els conceptes de regressió, conté informació sobre la producció i la qualitat de la carn de vedella als Estats Units durant el període de 1930 fins l’any 2015.

Com ja veurem més endavant tindrem molts valors nuls, sobretot a les dades corresponents a l’anterior segle.

Inicialment el dataset té dimensions de 134 fileres i 30 columnes que consten dels següents valors:



## Definició dels atributs:

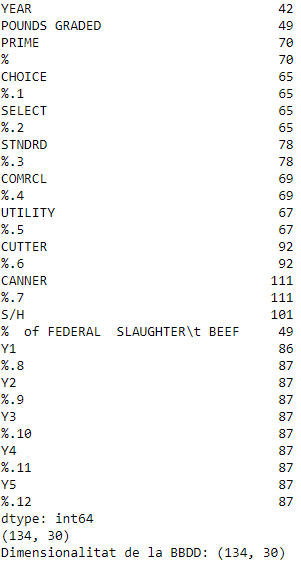
A continuació es llistaran els atributs del dataset amb una petita explicació del que representen:

1. **YEAR**: L’any en què es van recollir les dades.
2. **POUNDS GRADED**: Total de lliures de vedella d’aquell any(en milions).
3. **PRIME**: Total de lliures de carn de tipus prime.
4. **CHOICE**: Total de lliures de carn de tipus choice.
5. **SELECT**: Total de lliures de carn de tipus select.
6. **STNDRD**: Total de lliures de carn de tipus standart.
7. **COMRCL**: Total de lliures de carn de tipus commercial.
8. **UTILITY**: Total de lliures de carn de tipus utility.
9. **CUTTER**: Total de lliures de carn de tipus cutter.
10. **CANNER**: Total de lliures de carn de tipus canner.
11. **S/H**: No estem ben bé segurs però suposem que fa referencia al % de novelles(Steers, Heifers)
12. **% OF FEDERAL SLAUGHTER BEEF**: Representa la porció de carn i esquelet d'un animal en comparació del pes viu.
13. **Y1**: És una estimació del percentatge de rendiment al detall dels quatre talls primaris de carn de vedella (chuck, costella, llom i rodó). També es coneix com a tallabilitat.
14. **Y2**: El mateix que Y1 però amb menor qualitat.
15. **Y3**: El mateix que Y2 però amb menor qualitat.
16. **Y4**: El mateix que Y3 però amb menor qualitat.
17. **Y5**: El mateix que Y4 però amb menor qualitat.
18. **%**: Totes les columnes de percentatge fan referència al percentatge d’aquell tipus de carn respecte el total. També n’hi han per la qualitat(Y1..Y5).

# Neteja del dataset:

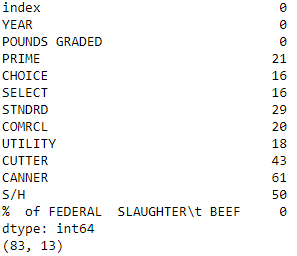
Procedim a veure si tenim valors nuls, i si n’hi ha eliminar-los o tractar-los d’alguna manera.

El resultat que obtenim és el següent:

Com podem veure tenim moltíssims valors nuls per a les dimensions del dataset. Revisant-ho veiem que n’hi han 6 línies al final que informen d’on s’ha extret la informació que no necessitem per fer els exercicis. A més a més, trobem moltes línies buides, aparentment sense sentit, i una línia buida que separa les dades en rangs de 10 anys.   
  
Primerament tractem de fer una aproximació esborrant totes les dades NaN del dataset, però ens queda només una línia amb dades, per tant, no és gens viable i comencem esborrant les que sabem que no tenen cap tipus de dada.  
  
Ja havent eliminat totes aquestes línies buides sense cap dada, passem a mirar realment les dades que tenim al dataset i en quin format les tenim per columnes.  
Veient el tipus de dada per columna ens adonem que YEAR, POUNDS GRADED i % of FEDERAL SLAUGHTER BEEF són de tipus string,. Això és perquè YEAR té dos fileres amb valors estimats marcats amb un ”\*”, que decidim esborrar per evitar possibles outliers. En el cas de POUNDS GRADED es degut a que els americans separen els milers amb ”,” i per tant s’han de treure. I finalment en % OF FEDERAL SLAUGHTER BEEF es posa el signe ”%”.

Decidim esborrar totes les columnes de percentatges referents al tipus de carn i als graus de qualitat, ja que no ens semblen dades necessàries o que aportin gaire, a més que falten moltes dades per a aquestes columnes.

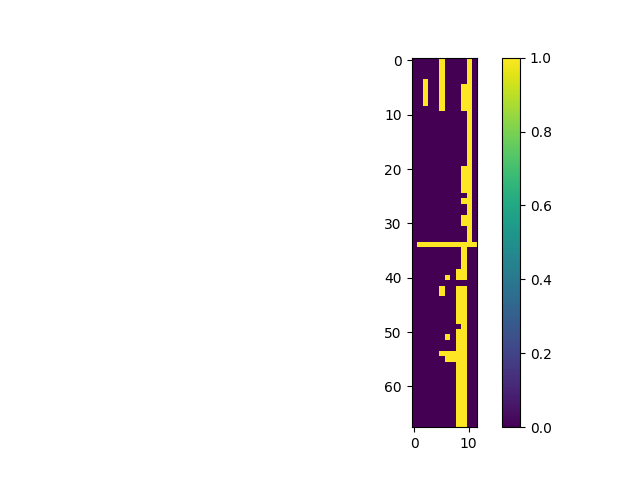
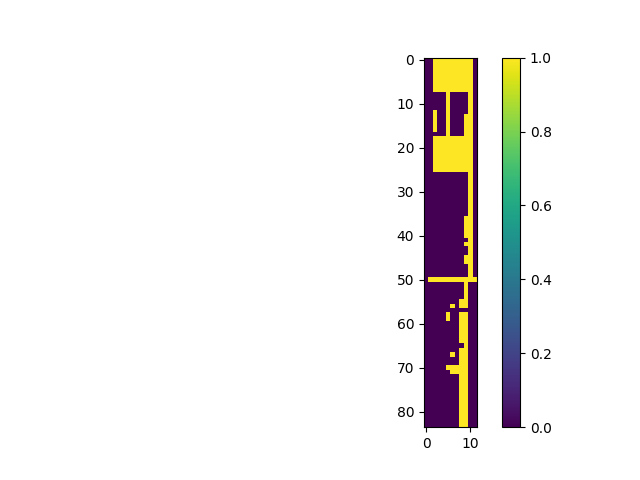
Un cop feta la primera neteja ens queda el següent:



Després decidim mirar quants nuls tenim per files per tal d’obtenir més endavant un millor resultat amb la tercera estratègia per omplir els valors nuls restants, que veurem al següent apartat(KNN Imputer):

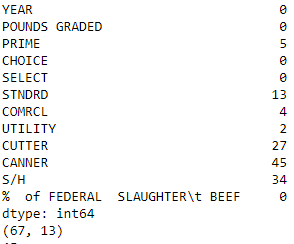
Com podem veure al gràfic de l’esquerra tenim dos grans blocs que contenen molts valors nuls (color groc), que van des de la fila 0 fins la 7 i des de la 18 fins la 25.  
Decidim que esborrarem aquestes files per tractar d’obtenir un millor resultat amb l’estratègia del KNN Imputer.

Al gràfic de la dreta podem veure el resultat després d’haver esborrat aquestes línies.



A continuació per anant acabant de netejar, farem servir un criteri on si de les columnes restants tenim més del 50% de valors nuls, esborrarem la columna, ja que faltarien massa dades.

D’aquesta forma s’eliminen les columnes de CANNER i S/H.



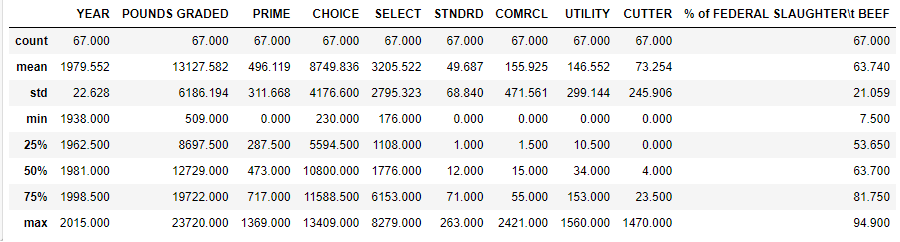
# Estratègies per omplir valors nuls:

Arrel del dataset resultant de la neteja, farem servir tres estratègies per omplir els valors nuls que queden a algunes de les columnes.

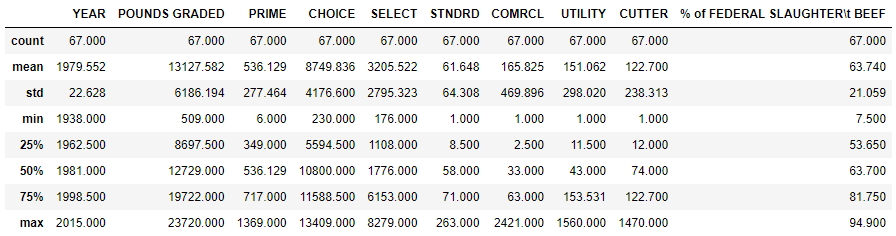
A partir d’aquí tindrem tres datasets diferents amb els que tractarem de veure quin podria donar-nos millors resultats.

* **Primera estratègia (omplir amb zeros):** En aquest primer dataset senzillament substituirem tot valor NaN pel valor 0.
* **Segona estratègia (omplir amb la mitja)**: En el segon dataset en comptes de substituir per 0 ho farem per la mitja extreta d’aquella columna.
* **Tercera estratègia (Nearest neighbors imputation)**: Al tercer dataset i substituirem els valors NaN per el que ens retorna l’algorisme del KNN Imputer de la llibreria sklearn.   
  Per defecte, s'utilitza una mètrica de distància euclidiana que admet els valors que falten, nan\_euclidean\_distances, per trobar els veïns més propers. Cada característica que falta s'imputa utilitzant valors de n\_neighbors veïns més propers que tenen un valor per a la característica. Les característiques dels veïns es fan una mitjana uniforme o ponderada per la distància a cada veí. Si a una mostra li falta més d'una característica, els veïns d'aquesta mostra poden ser diferents segons la característica concreta que s'imputa.

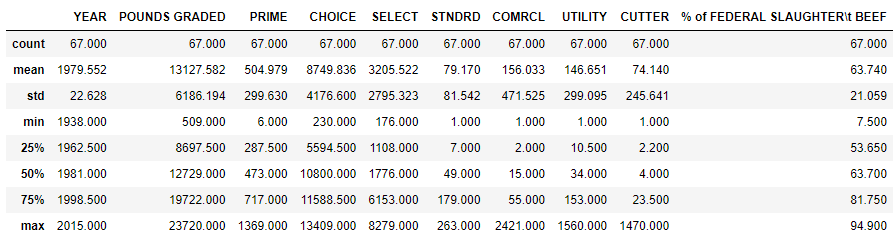
A continuació es mostren els resultats per als tres casos **en el mateix ordre**:



Filled 0



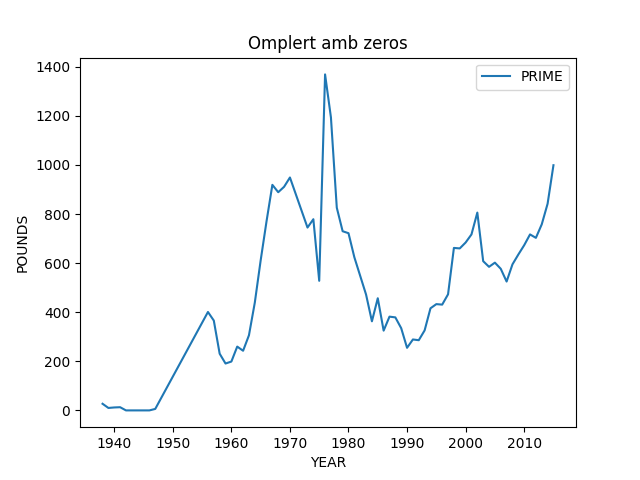
Filled Mean

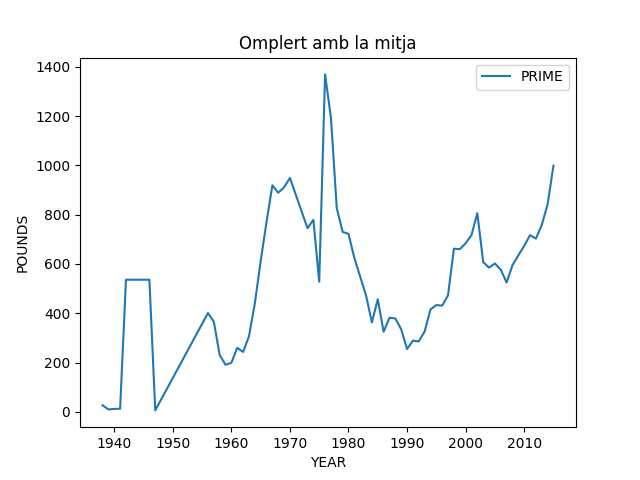


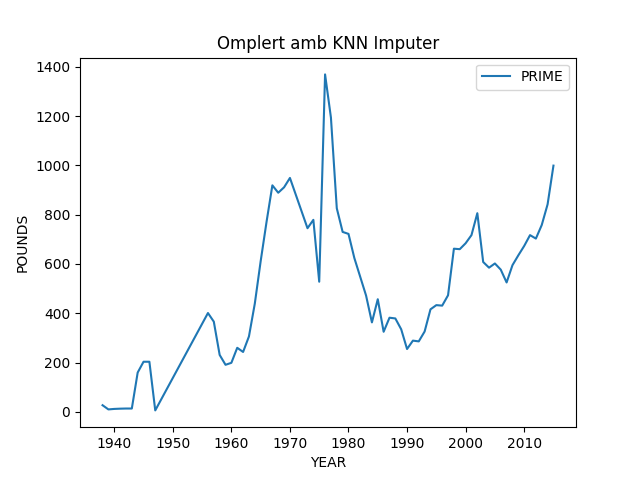
Filled KNN

Com podem observar per exemple a la columna PRIME, pot ser els resultats que més s’assemblen a la realitat deriven de fer servir les estratègies 1 i 3 (omplir amb 0’s i algorisme KNN Imputer), ja que amb la mitja sembla que surten valors més exagerats i que per dir-ho d’alguna manera, ens infla les dades on no correspon.

Ho podem veure d’una manera més visual a les següents gràfiques:

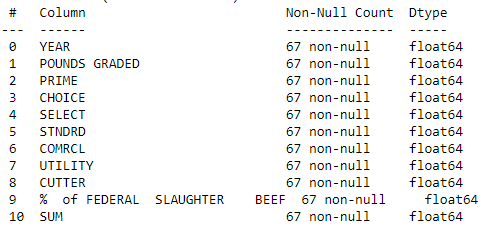






Anàlisi i raonament sobre el nou dataset:

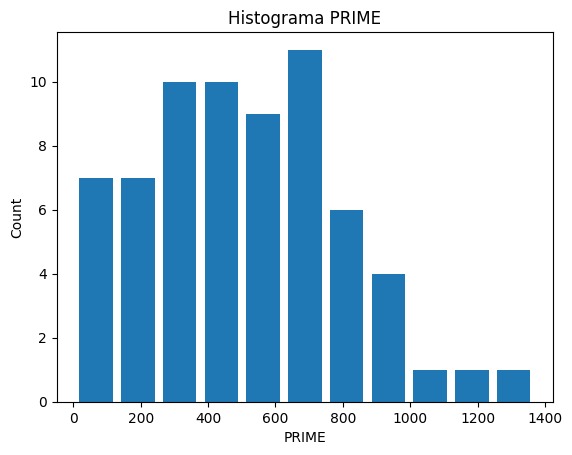
El següent pas és veure de quin tipus són els atributs que tenim fins ara per veure si hem de fer alguna cosa a l’hora de tractar amb ells.

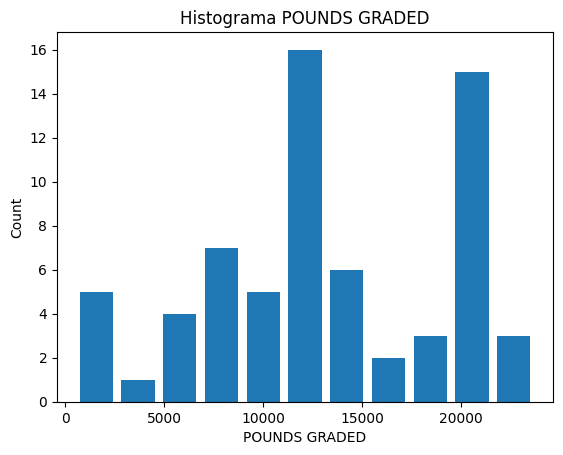


Com podem veure tots els atributs són de tipus float64 per tant no tindrem cap problema en aquest sentit.

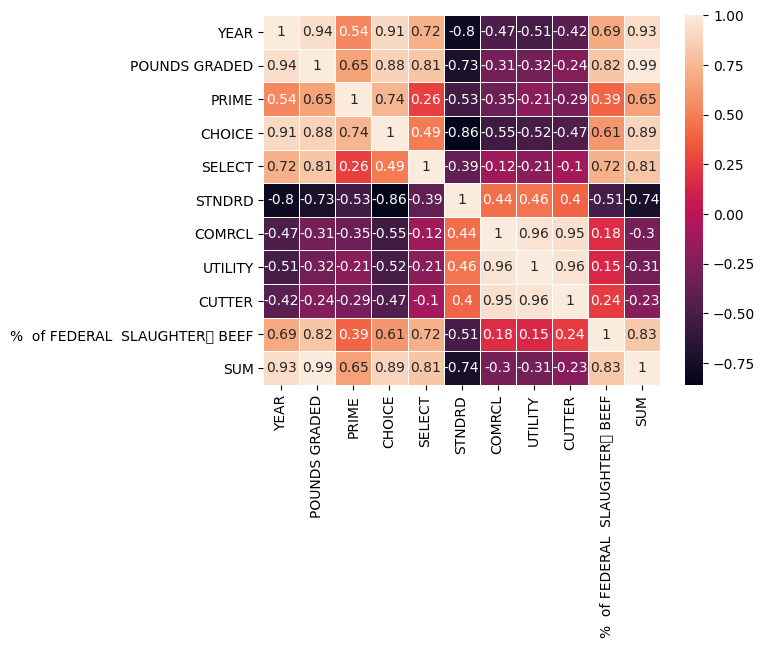
Per altra banda es pot veure que hem creat una nova columna als datasets anomenada SUM, que és la suma de totes les columnes relacionades a un tipus de carn, és a dir, PRIME,CHOICE...CUTTER, que farem servir més endavant.

Pel que fa a la distribució, només podem observar el que més s’assembla a una distribució gaussiana en els valors de PRIME i POUNDS GRADED.





Les correlacions més significatives que s’observen són el total de lliures respecte l’any i la columna SUM respecte la de lliures totals, evidentment ja que aquesta és la suma dels tipus de carn, que no deixen de ser un desglossament del total de lliures.

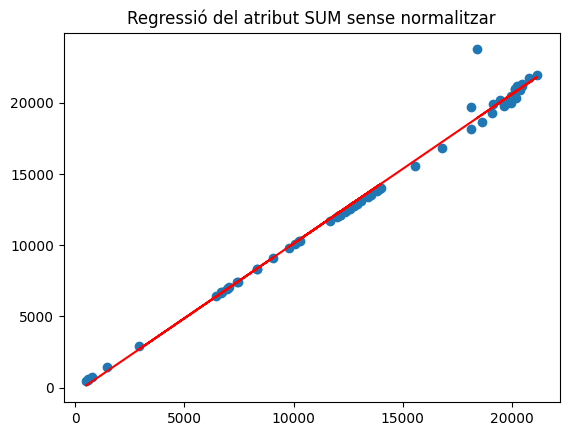


D’aquesta manera definim el total de lliures de vedella com el nostre valor objectiu. Per tant el nostre estudi estarà focalitzat en predir amb la major fiabilitat possible aquest total.

# Primeres regressions:

En aquest apartat ens hem centrat en fer les primeres regressions i a partir d’aquí poder calcular el MSE i la r2. A l’obtenir aquests dos valors podem confirmar i veure quins són els atributs que realment són els millors i si corresponen amb els que hem decidit escollir en l’apartat anterior.

A l’hora de calcular el MSE i la r2 utilitzem el codi que se’ns proporciona en els Jupyter Notebooks de l’assignatura. Hem calculat les dues mètriques per a totes les variables. A continuació es mostra tant la regressió com els càlculs mencionats utilitzant el mètode d’emplenat de Nan amb el KNN Imputer.



Com podem observar la regressió de l’atribut SUM és molt bona ja que com bé ens deia la matriu de correlació prèviament són dos atributs molt correlacionats entre si.

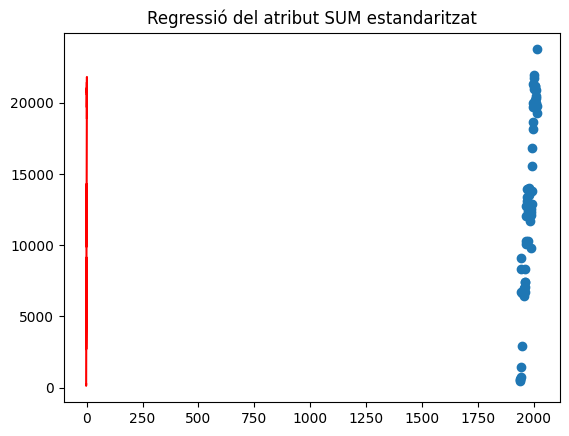
Tot i això hem decidit veure si les mètriques confirmen el que estem veient.

|  |  |  |
| --- | --- | --- |
| **ATRIBUT** | **MEAN SQUARED ERROR** | **R-SQUARE** |
| **YEAR** | 4605950.294898352 | 0.8778191593718118 |
| **POUNDS GRADED** | 1.1279788236430314e-23 | 1.0 |
| **PRIME** | 21701983.512281522 | 0.4243171508458159 |
| **CHOICE** | 8514741.790275076 | 0.7741316681554131 |
| **SELECT** | 12769245.163558425 | 0.6612735682364921 |
| **STANDARD** | 35771280.65717755 | 0.05110457968212667 |
| **COMRCL** | 34175556.75569288 | 0.09343393089328511 |
| **UTILITY** | 33798309.5722154 | 0.1034410684112359 |
| **CUTTER** | 35478391.24225527 | 0.05887398070365746 |
| **% OF FEDERAL SLAUGHTER** | 12570033.243707655 | 0.6665580108101521 |
| **SUM** | **422696.9676677511** | **0.9887872279260513** |

D’aquesta taula podem observar com l’atribut SUM, que és el que estem fent servir per predir la quantitat de lliures de carn és el que es presenta amb millors mètriques, òbviament sense tenir en compte el propi atribut POUNDS GRADED.

La r-square pot oscil·lar entre valors de 0 i 1, sent 1 el millor valor possible. En el nostre cas la r-square en l’atribut SUM és de 0.988 que és molt bona i el MSE és de 427224.776. Al principi pensàvem que el MSE era nefast ja que parlant amb els companys i veient alguns notebooks, el valor era terriblement gran. Després de parlar amb el professor ens va fer entendre millor la mètrica i vam poder arribar a la conclusió que no era tant dolent. Com els nostres valors son milions de lliures de carn, al final el MSE no és excessivament gran.

Després procedim a repetir el procés tant de regressió com de les mètriques però estandarditzant les dades gràcies a la funció que se’ns proporciona al guió de la pràctica. Gràcies a l’estandardització aconseguim que tots els valors dels atributs estiguin en un mateix rang. A continuació mostrem els resultats:



|  |  |  |
| --- | --- | --- |
| **ATRIBUT** | **MEAN SQUARED ERROR** | **R-SQUARE** |
| **YEAR** | 4605950.294898348 | 0.8778191593718118 |
| **POUNDS GRADED** | 5.80261026716303e-25 | 1.0 |
| **PRIME** | 21701983.512281522 | 0.4243171508458159 |
| **CHOICE** | 8514741.790275076 | 0.7741316681554131 |
| **SELECT** | 12769245.163558424 | 0.6612735682364921 |
| **STANDARD** | 35771280.65717755 | 0.05110457968212667 |
| **COMRCL** | 34175556.75569288 | 0.09343393089328511 |
| **UTILITY** | 33798309.5722154 | 0.1034410684112359 |
| **CUTTER** | 35478391.24225527 | 0.05887398070365746 |
| **% OF FEDERAL SLAUGHTER** | 12570033.243707657 | 0.6665580108101521 |
| **SUM** | **422696.96766775066** | **0.9887872279260513** |

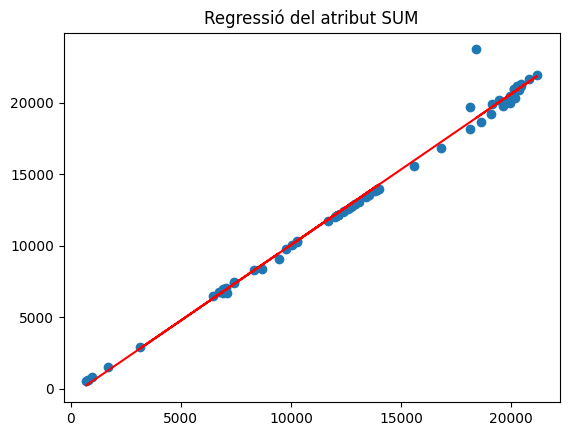
Com podem observar a la taula els valors segueixen sent molt similars als d’abans d’estandarditzar. L’única diferència que podem veure i que era òbvia és la de la regressió.

Aquests passos s’han dut a terme amb les 3 tècniques d’omplir els NaN del dataset. La tècnica que millor resultat ens ha donat tant a les regressions com ens les mètriques ha sigut la tècnica que omple els NaN amb el algorisme KNN. Tot i que la diferència no era desproporcionada hem decidit continuar el que queda de pràctica amb la tècnica del KNN.

En el nostre cas no podem seleccionar atributs a desestimar ja que, a part de que ho fem al principi, tots els atributs restants ens son necessaris. Per entrenar el regressor hem agafat les dades del dataset i hem decidit fer una partició d’aquestes entre dades d'entrenament i dades de test. Això ens servirà per veure com de bo és el nostre regressor i el seu rendiment amb dades “noves”. Hem provat diferents particions i els resultats han sigut els següents:

**90% entrenament i 10% de test:**

Error en atribut SUM: 35960.000031  
R2 score en atribut SUM: 0.998894



**80% entrenament i 20% de test:**

Error en atribut SUM: 57005.179728  
R2 score en atribut SUM: 0.999094

**70% entrenament i 30% de test:**

Error en atribut SUM: 124827.097652  
R2 score en atribut SUM: 0.996377

# Anàlisi de components principals:

L’anàlisi de components principals és una tècnica que ens serveix per poder discernir quins atributs són necessaris per a fer la regressió. Tot i això, el nostre dataset conté una certa particularitat que ens impedeix aplicar un PCA, com a continuació expliquem.

El nostre dataset després de tractar-lo per eliminar els NaN queda amb un nombre de columnes, i en conseqüència, d’atributs els quals tots son útils. La nostra regressió consta de la suma dels valors de tots els atributs menys POUNDS GRADED. Al ser una combinació de **TOTS** els atributs si apliquéssim el PCA diferiríem clarament del resultat de POUNDS GRADED inferint així en un deteriorament de la regressió i predicció.

Podríem dir que en el nostre cas el PCA l’hem aplicat al principi filtrant i eliminant les columnes que tenien molt poca informació. Òbviament no és el mateix, però té una certa semblança.

# Conclusions:

Aquesta primera pràctica ens ha servit per acabar d’entendre tots els aspectes explicats a teoria i veure’ls d’una forma pràctica. El dataset que hem hagut de tractar ha sigut bastant un mal de cap, ja que a diferència d’alguns dels datasets dels nostres companys que hem pogut comprovar, el nostre tenia poca quantitat de dades i, per més inri, estava ple de NaNs com s’ha pogut veure explicat en aquest informe.

Tot i aquest entrebanc inicial hem pogut analitzar i entendre quins són els principals factors d’èxit en l’elaboració de regressions, entendre les correlacions entre les dades, buscar combinacions d’atributs per millorar el dataset, etc.

Malauradament no hem sigut capaços d’aplicar el descens del gradient, per lo que podem afirmar que l’apartat A està incomplet. Tot i així ho hem intentat i, tot i no saber implementar-ho en codi per al nostre dataset, hem tractat de posar tots els nostres esforços per entendre el concepte i crec que ho hem aconseguit.

En conclusió estem contents dels resultats obtinguts, de l’aprenentatge i el saber sobreposar-nos a tots els “problemes” que han sorgit. Per acabar volem agrair l’ajuda dels professors de l’assignatura durant les tutories realitzades.

# GitHub:

<https://github.com/guillem-martinez/Practica1-ApC>