Anàlisi de dades i

regressió

GRUP: GPA304-1030

Víctor Bosch Pueyo 1566583

Álvaro Caravaca Hernández 1566685

Juan Carlos Martínez Moreno 1566936

Contenido

[Introducció 3](#_Toc85877650)

[Apartat C: Analitzant Dades 3](#_Toc85877651)

[Explicació de la base de dades 3](#_Toc85877652)

[Apartat B: Primeres Regressions 3](#_Toc85877653)

[Correlació entre dades 3](#_Toc85877654)

[Error quadràtic 3](#_Toc85877655)

[Atribut escollit 3](#_Toc85877656)

[Apartat A: El descens del gradient 3](#_Toc85877657)

[Conclusions 4](#_Toc85877658)

# Introducció

#De que va la pràctica

# Apartat C: Analitzant Dades

Aquest apartat el dedicarem a analitzar la base de dades que tenim entre mans per a entendre el problema al que ens enfrontem. És important saber que no treballem amb un conjunt de números sense sentit, sinó que el que tenim és una base de dades real que hem d’entendre per a no agafar com a atribut més important un que, per molt que hi hagi una correlació forta entre les dades, podria ser pura coincidència.

## Explicació de la base de dades

Com ja hem dit a l’apartat d’introducció, la nostra base de dades ens mostra el producte interior brut en diferents sectors de 2005 fins a 2016, dividit en quadrimestres. L’objectiu de treballar amb aquestes dades es poder aconseguir predir és el número de turistes que arriben a l’Índia. Els trobem en 4 sub-datasets de la següent manera:

* **q1.csv** : corresponent al 1r quadrimestre de 2005 fins a 2016
* **q2.csv** : corresponent al 2n quadrimestre de 2005 fins a 2016
* **q3.csv** : corresponent al 3r quadrimestre de 2005 fins a 2016
* **q4.csv** : corresponent al 4t quadrimestre de 2005 fins a 2016

Hem seguit les indicacions donades pel professor de treballar amb totes les dades juntes a l’hora de fer el descens de gradient i la predicció. Ens hem trobat amb el problema que les dades dels diferents quadrimestres varien dins un llindar, cosa totalment normal ja que en el primer i en l’últim, el número de turistes que arriben és molt més elevat que en els altres dos. Les dades d’aquests dos trimestres amb valors més alts comencen el primer d’Octubre fins el 31 de Març. Buscant informació per internet hem descobert que les dates més recomanades per a visitar l’Índia, per a raons climatològiques, són al nadal. Podem dir llavors que les nostres dades segurament siguin correctes. Vist això, hem decidit seguir amb la pràctica com ens va indicar el professor però, dedicarem un apartat a fer la regressió per a un sol quadrimestre.

Els sectors sobre els que treballem són els següents, expressats en bilions de rupies índies:

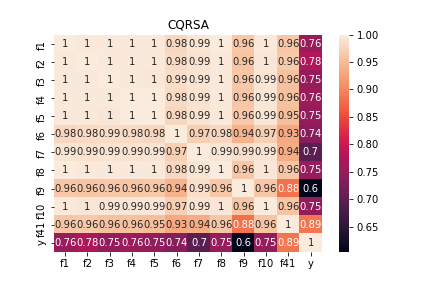
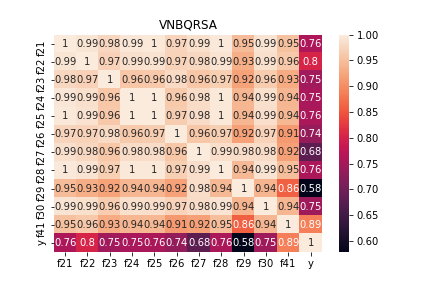
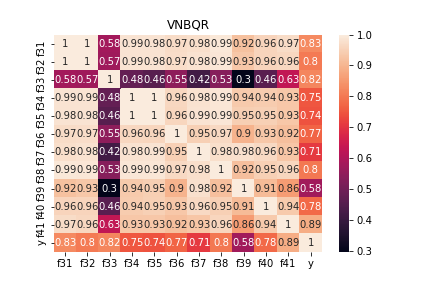
* PIB als preus de mercat (enfoc des de la producció)
* Valor afegit brut a preus bàsics (activitat total)
* Agricultura, silvicultura i pesca
* La indústria (inclosa l'energia)
* Fabricació
* Construcció
* Serveis
* Transport, allotjament i activitat de serveis alimentaris
* Activitats immobiliàries
* Administracions públiques (seguretat social, educació, etc.)

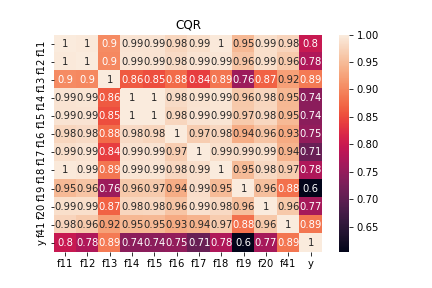
Tots aquests valors els veiem representats 4 cops segons el tipus de mesura en el que s’ha fet; aquestes són:

* **CQRSA**: Moneda nacional, preus corrents, nivells trimestrals, ajustos estacionals
* **CQR**: Moneda nacional, preus corrents, nivells trimestrals
* **VNBQRSA**: Moneda nacional, preus constants, any base nacional, nivells trimestrals, ajustos estacionals
* **VNBQR**: Moneda nacional, preus constants, any base nacional, nivells trimestrals

Per cada atribut hem dibuixat el seu histograma i la seva gràfica de punts per a veure quina distribució segueixen. També hem aplicat el test de Shapiro per a determinar quines variables no segueixen una distribució normal.

##tocho del alvaro

També hem calculat la correlació entre els diferents atributs per tal de saber si estan relacionats entre ells. El que més ens interessa es si estan relacionats amb l’atribut que volem predir. Els mapes de calor obtinguts són els següents:



Aquests tipus de mapa son molt útils ja que, gràficament podem observar molt fàcilment quins atributs tenen una correlació més alta i més baixa segons el color que presenten. Com hem dit abans, el que més ens interessa és la seva relació amb l’atribut a predir. Hem descartat tots aquells valors que, comparats amb y no superen el 0.75, tots els que si que ho fan els hem guardat com a possibles candidats.

# Apartat B: Primeres Regressions

#explicar mes o menys apartat B

## Correlació entre dades

## Error quadràtic

## Atribut escollit

# Apartat A: El descens del gradient

# Conclusions