Predicción del salario en jóvenes recién graduados

Bartomeu Vargas Alarcón 1568113
Harold Gonzales Contreras 1571485
Martí Simón Rojas 1568180

Índice

Introducción

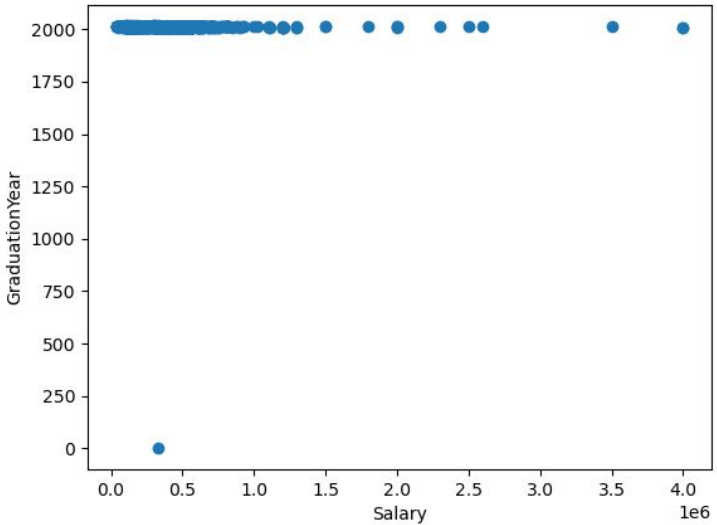
- Recogida de datos de estudiantes Indios.
  - Datos personales
  - Datos académicos
  - Datos financieros
- Sistema educativo Indio.
  - Grados 10 y 12.
  - Pruebas AMCAT
  - Universidad

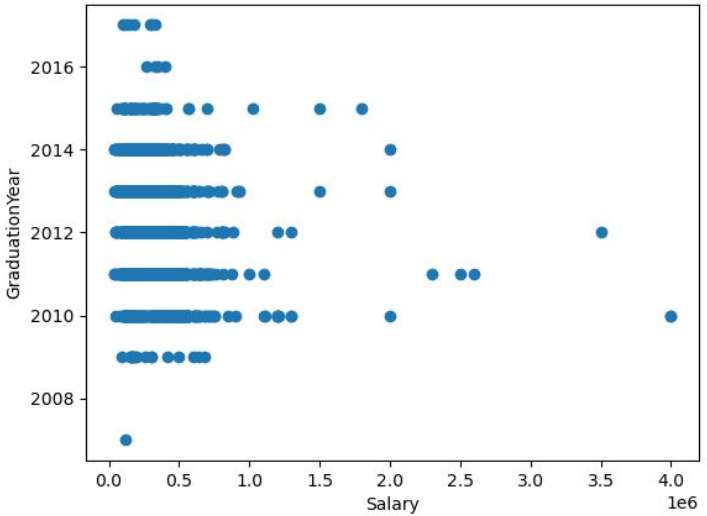
Apartado C: Analizando datos -Tratamiento de datos

- De Strings a valores discretos
  - DOB (yyyy-mm-dd => yy.ymm.dd / 1990-10-23 => 19.901.013)
  - Gender (m=0, f=1)
  - Degree (B.Tech/B.E.=1, MCA=2, M.Tech./M.E.=3, M.Sc. (Tech.)=4, Otros=0)
  - Specialization (electronics and communication engineering=1, computer science & engineering=2, information technology=3, computer engineering=4, computer application=5, Otros=0)
  - CollegeState (Uttar Pradesh=1, Karnataka=2, Tamil Nadu=3, Telangana=4, Maharashtra=5, Otros=0)

Apartado C: Analizando datos -Tratamiento de datos

- Retirada de atributos irrelevantes
  - 10board y 12board.
- Aplicación de Replace
  - GraduationYear
    (if dataset[“GraduationYear”]<2000: dataset[“GraduationYear”]=2010)

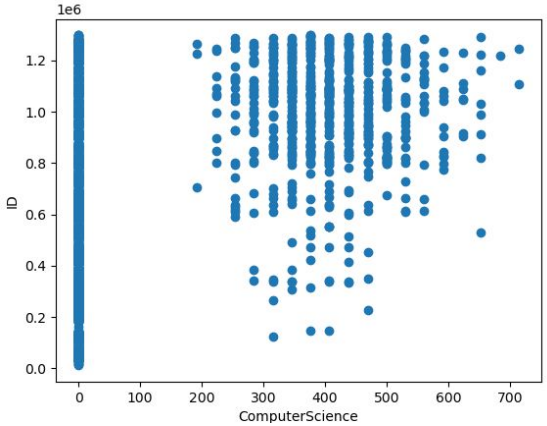


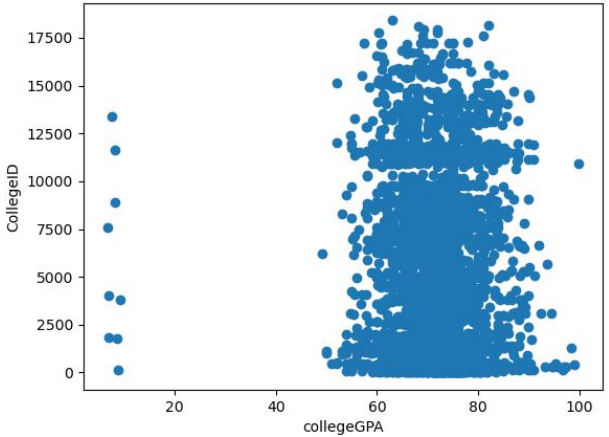


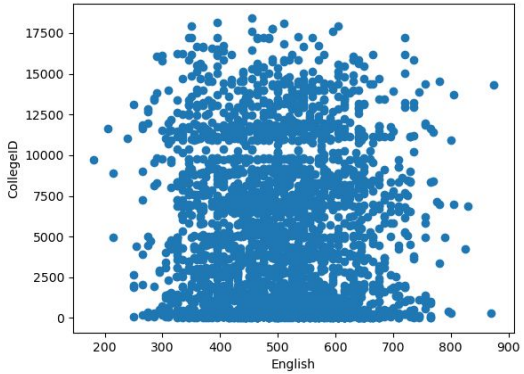
Apartado C: Analizando datos -Tipo de los atributos

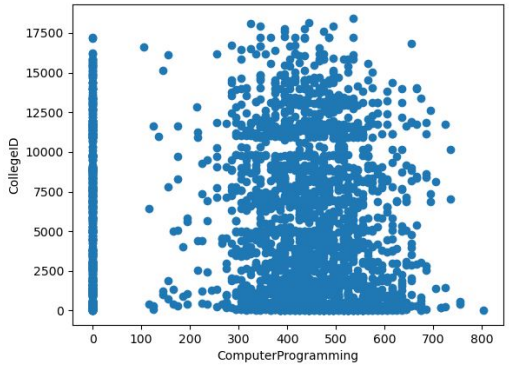
ID	Gender	DOB	10percentatge
10board	12graduation	12percentage	12board
CollegeID	CollegeTier	Degree	Specialization
CollegeGPA	CollegeCityID	CollegeCityTier	CollegeState
GraduationYear	English	Logical	Quant
Domain	ComputerProgramming	ElectronicsAndSemicon	ComputerScience
MechanicalEngg	ElectricalEngg	TelecomEngg	CivilEngg
conscientiousness	agreeableness	extraversion	extraversion
nueroticism	nueroticism	-	-

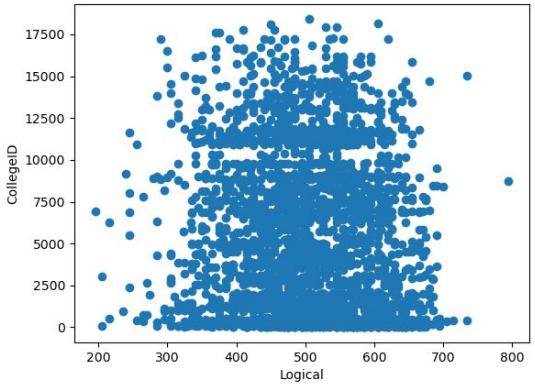
Apartado C: Analizando datos
-Atributos con distribución Gaussiana

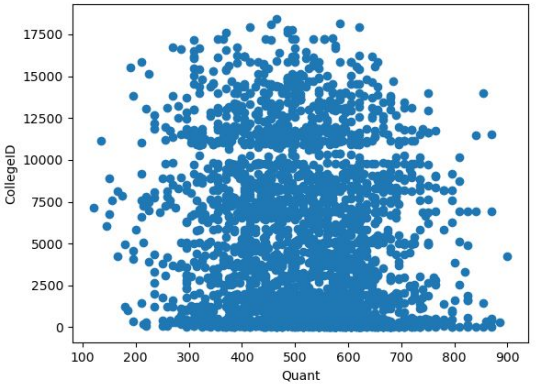




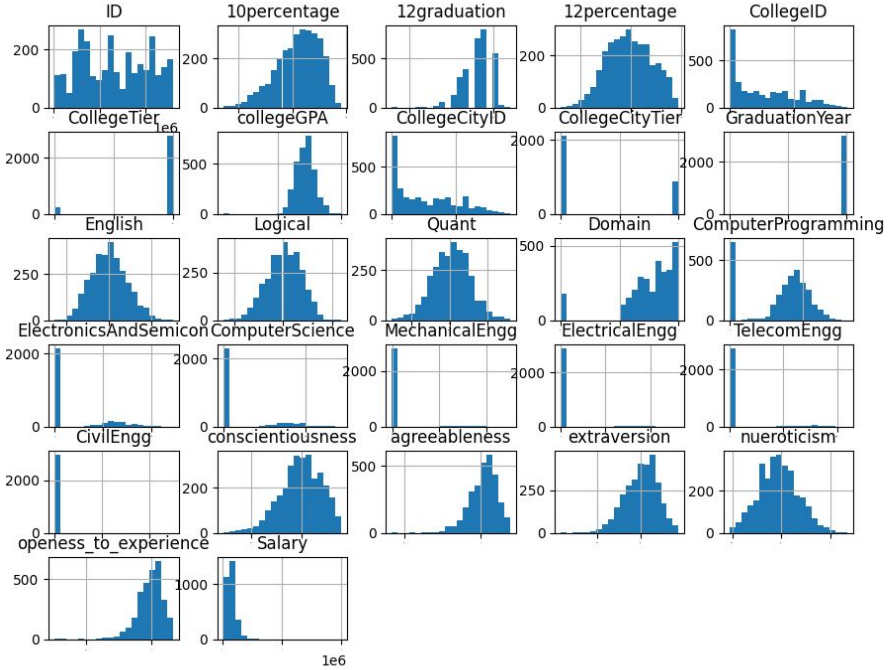




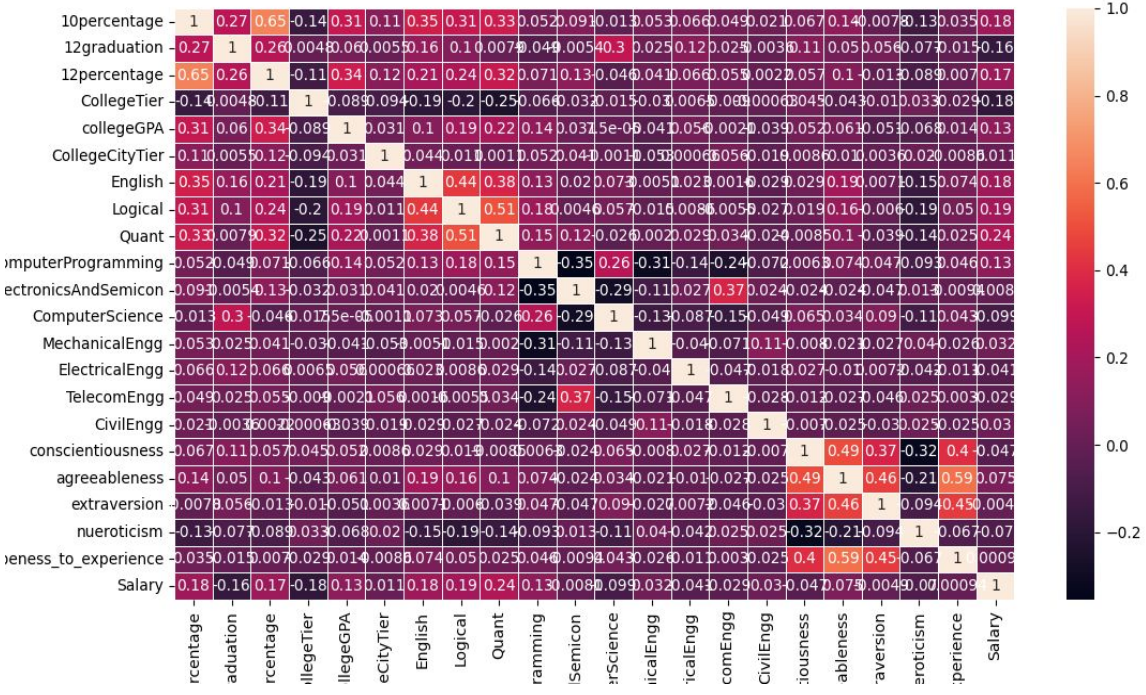




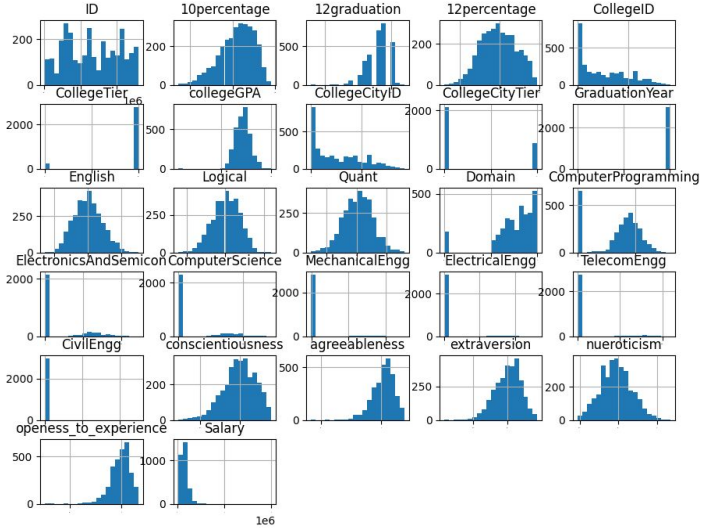
Apartado C: Analizando datos
-Atributos con distribución Gaussiana II

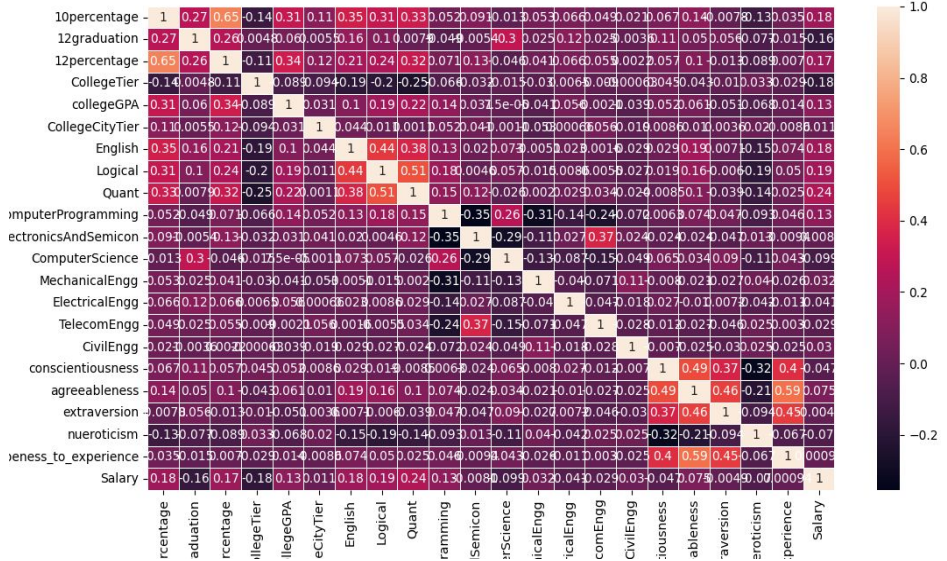


Apartado C: Analizando datos -Atributos con distribución Gaussiana II



Apartado C: Analizando datos -Atributo objetivo: Salario





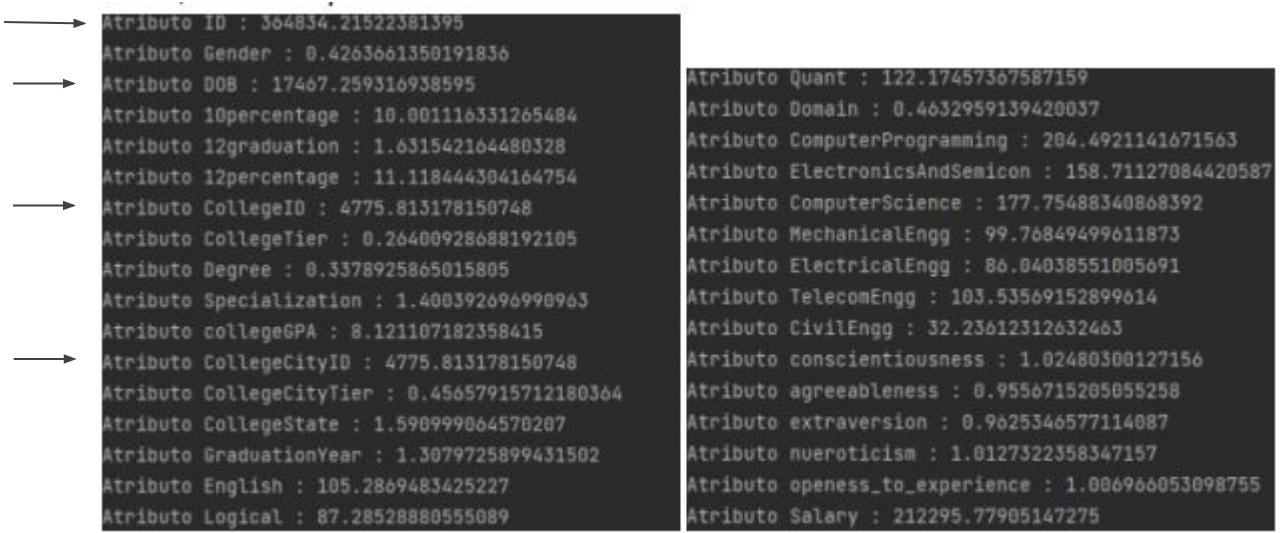
Apartado B: Primeras regresiones -Error cuadrático medio

MSE	SIN NORMALIZAR	NORMALIZADO
Atributo 1	1.5512	0.7963
Atributo 2	1.9261	0.8336
Atributo 3	1.7712	0.8274
Atributo 4	1.6313	0.8062
Atributo 5	1.7281	0.8211
Atributo 6	1.6709	0.8128
Atributo 7	1.8684	0.8349
Atributo 8	1.7726	0.8349
Atributo 9	2.1413	0.8346
Atributo 10	1.6776	0.8161
Atributo 11	1.6725	0.8140
Atributo 12	1.8684	0.8349
Atributo 13	2.0047	0.8065
Atributo 14	1.8861	0.8329
Atributo 15	1.6177	0.8065
Atributo 16	1.6639	0.8123
Atributo 17	1.6801	0.8166
Atributo 18	1.5649	0.7978
Atributo 19	1.7581	0.8237
Atributo 20	1.8002	0.8302
Atributo 21	2.1077	0.8347
Atributo 22	1.9322	0.8395
Atributo 23	1.9345	0.8335
Atributo 24	1.9719	0.8352
Atributo 25	1.9301	0.8335
Atributo 26	1.8990	0.8329
Atributo 27	1.9103	0.8330
Atributo 28	1.8870	0.8325
Atributo 29	2.0636	0.8345
Atributo 30	1.9284	0.8357
Atributo 31	2.0920	0.8358

~~10BOARD~~

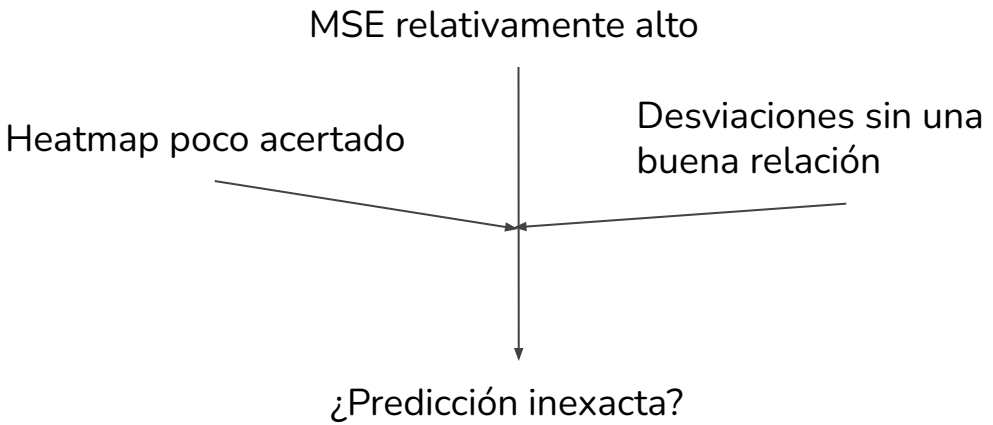
~~12BOARD~~

Apartado B: Primeras regresiones -Desviaciones

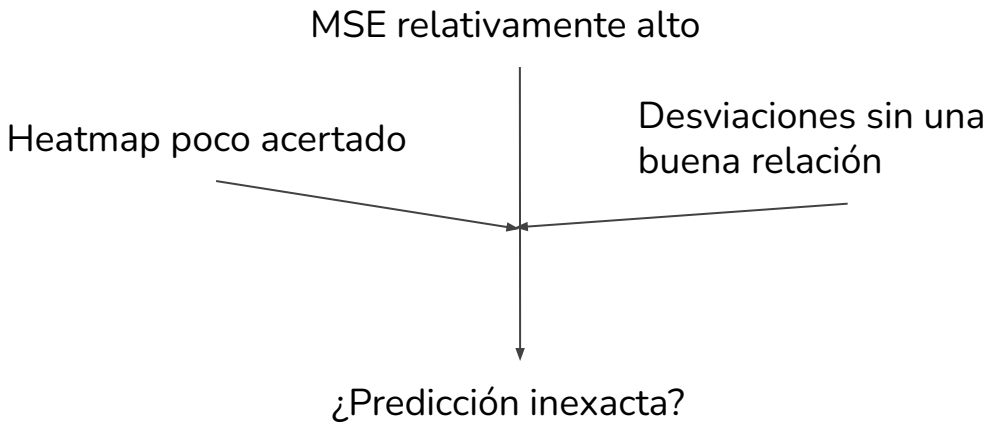


Apartado B: Primeras regresiones -¿Mejores atributos?

- **Quant**
- Logical
- English
- 10percentage
- 12percentage



Apartado B: Primeras regresiones -¿Mejores atributos?



Conclusiones

- **NO** predicción fiable
- **NINGÚN** atributo para una distribución gaussiana
- Heat Map sin una **RELACIÓN** fuerte
- Errores cuadráticos medios prácticamente **SIMILARES** y **GRANDES** -> **NINGÚN** atributo fiable

↓

Poca predicción fiable para la obtención del salario

Problemas encontrados

- Hemos tenido que cambiar los valores string por valores enteros.
- Al cambiar valores daba error y tuvimos que utilizar replace() de panda.
- Hemos ajustado algunos atributos obviando los valores incorrectos o extraordinarios.
- Hemos tenido problemas guardando los datos resultantes

Muchas gracias!