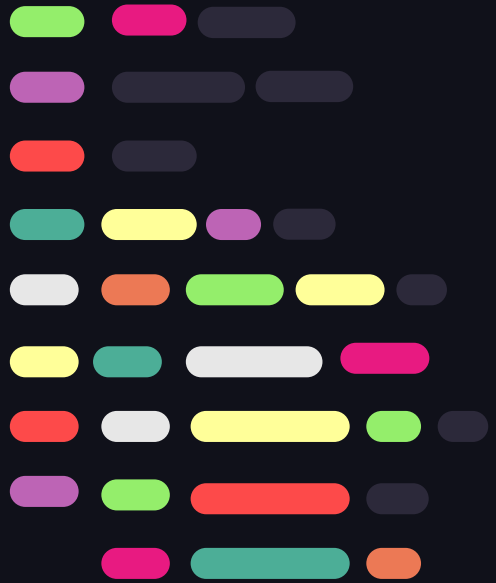
Modelado predictivo y clustering aplicado al mercado de coches usados



< Javier Vázquez Zambrano >

< Juan Noguerol Tirado >

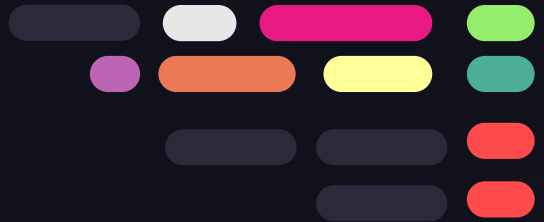
< José Manuel Sánchez Ruiz >

< Gabriel López Bellido >

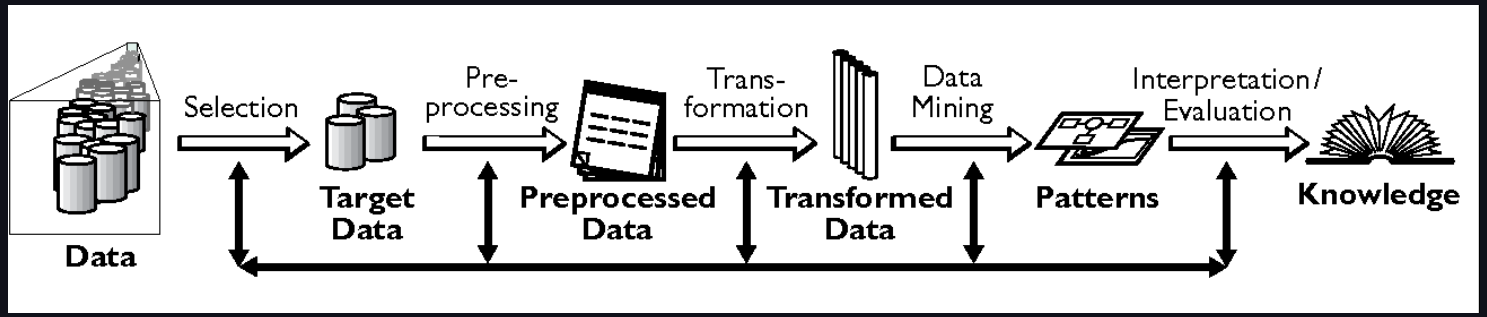
< Celia Sánchez Gaitán >

Grupo 8

ÍNDICE

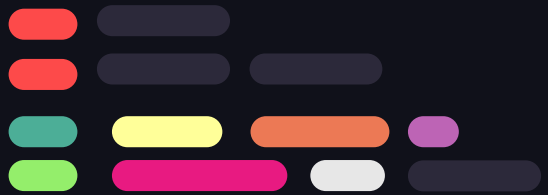


01 INTRODUCCIÓN



kaggle

< Vehicle dataset >



02 DATASET: car_data_1

1	Car_Name	Year	Selling_Price	Present_Price	Kms_Driven	Fuel_Type	Seller_Type	Transmission	Owner
2	ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
3	sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
4	ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
5	wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
6	swift	2014	4.6	6.87	42450	Diesel	Dealer	Manual	0
7	vitara brezza	2018	9.25	9.83	2071	Diesel	Dealer	Manual	0
8	ciaz	2015	6.75	8.12	18796	Petrol	Dealer	Manual	0
9	s cross	2015	6.5	8.61	33429	Diesel	Dealer	Manual	0
10	ciaz	2016	8.75	8.89	20273	Diesel	Dealer	Manual	0

02 DATASET: car_data_2

1	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner
2	Maruti 800 AC	2007	60000	70000	Petrol	Individual	Manual	First Owner
3	Maruti Wagon R LXI Minor	2007	135000	50000	Petrol	Individual	Manual	First Owner
4	Hyundai Verna 1.6 SX	2012	600000	100000	Diesel	Individual	Manual	First Owner
5	Datsun RediGO T Option	2017	250000	46000	Petrol	Individual	Manual	First Owner
6	Honda Amaze VX i-DTEC	2014	450000	141000	Diesel	Individual	Manual	Second Owner
7	Maruti Alto LX BSIII	2007	140000	125000	Petrol	Individual	Manual	First Owner
8	Hyundai Xcent 1.2 Kappa S	2016	550000	25000	Petrol	Individual	Manual	First Owner
9	Tata Indigo Grand Petrol	2014	240000	60000	Petrol	Individual	Manual	Second Owner
10	Hyundai Creta 1.6 VTVT S	2015	850000	25000	Petrol	Individual	Manual	First Owner

02 DATASET: car_data_3

	name	year	selling_price	km_driven	fuel	seller_type	transmission	owner	mileage	engine	max_power	torque
1	Maruti Swift Dzire VDI	2014	450000	145500	Diesel	Individual	Manual	First Owner	23.4 kmpl	1248 CC	74 bhp	190Nm@ 2000rpm
2	Skoda Rapid 1.5 TDI Ambition	2014	370000	120000	Diesel	Individual	Manual	Second Owner	21.14 kmpl	1498 CC	103.52 bhp	250Nm@ 1500-2500rpm
3	Honda City 2017-2020 EXi	2006	158000	140000	Petrol	Individual	Manual	Third Owner	17.7 kmpl	1497 CC	78 bhp	12.7@ 2,700(kgm@ rpm)
4	Hyundai i20 Sportz Diesel	2010	225000	127000	Diesel	Individual	Manual	First Owner	23.0 kmpl	1396 CC	90 bhp	22.4 kgm at 1750-2750rpm
5	Maruti Swift VXI BSIII	2007	130000	120000	Petrol	Individual	Manual	First Owner	16.1 kmpl	1298 CC	88.2 bhp	11.5@ 4,500(kgm@ rpm)
6	Hyundai Xcent 1.2 VTVT E Plus	2017	440000	45000	Petrol	Individual	Manual	First Owner	20.14 kmpl	1197 CC	81.86 bhp	113.75nm@ 4000rpm
7	Maruti Wagon R LXI DUO BSIII	2007	96000	175000	LPG	Individual	Manual	First Owner	17.3 km/kg	1061 CC	57.5 bhp	7.8@ 4,500(kgm@ rpm)
8	Maruti 800 DX BSII	2001	45000	5000	Petrol	Individual	Manual	Second Owner	16.1 kmpl	796 CC	37 bhp	59Nm@ 2500rpm
9	Toyota Etios VXD	2011	350000	90000	Diesel	Individual	Manual	First Owner	23.59 kmpl	1364 CC	67.1 bhp	170Nm@ 1800-2400rpm

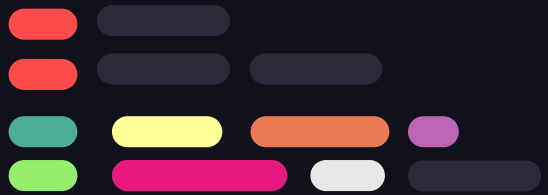
02 DATASET: car_data_4

1	Make	Model	Price	Year	Kilometer	Fuel Type	Transmissi	Location	Owner	Seller Type	Engine	Max Power
2	Honda	Amaze 1.2 VX i-VTEC	505000	2017	87150	Petrol	Manual	Pune	First	Corporate	1198 cc	87 bhp @ 6000 rpm
3	Maruti Suzuki	Swift DZire VDI	450000	2014	75000	Diesel	Manual	Ludhiana	Second	Individual	1248 cc	74 bhp @ 4000 rpm
4	Hyundai	i10 Magna 1.2 Kappa2	220000	2011	67000	Petrol	Manual	Lucknow	First	Individual	1197 cc	79 bhp @ 6000 rpm
5	Toyota	Glanza G	799000	2019	37500	Petrol	Manual	Mangalore	First	Individual	1197 cc	82 bhp @ 6000 rpm
6	Toyota	Innova 2.4 VX 7 STR [2016-2020]	1950000	2018	69000	Diesel	Manual	Mumbai	First	Individual	2393 cc	148 bhp @ 3400 rpm
7	Maruti Suzuki	Ciaz ZXi	675000	2017	73315	Petrol	Manual	Pune	First	Individual	1373 cc	91 bhp @ 6000 rpm
8	Mercedes-Benz	CLA 200 Petrol Sport	1898999	2015	47000	Petrol	Automatic	Mumbai	Second	Individual	1991 cc	181 bhp @ 5500 rpm
9	BMW	X1 xDrive20d M Sport	2650000	2017	75000	Diesel	Automatic	Coimbatore	Second	Individual	1995 cc	188 bhp @ 4000 rpm
10	Skoda	Octavia 1.8 TSI Style Plus AT [2017]	1390000	2017	56000	Petrol	Automatic	Mumbai	First	Individual	1798 cc	177 bhp @ 5100 rpm

Max Torque	Drivetrain	Length	Width	Height	Seating Capacity	Fuel Tank Capacity
109 Nm @ 4500 rpm	FWD	3990.0	1680.0	1505.0	5.0	35.0
190 Nm @ 2000 rpm	FWD	3995.0	1695.0	1555.0	5.0	42.0
112.7619 Nm @ 4000 rpm	FWD	3585.0	1595.0	1550.0	5.0	35.0
113 Nm @ 4200 rpm	FWD	3995.0	1745.0	1510.0	5.0	37.0
343 Nm @ 1400 rpm	RWD	4735.0	1830.0	1795.0	7.0	55.0
130 Nm @ 4000 rpm	FWD	4490.0	1730.0	1485.0	5.0	43.0
300 Nm @ 1200 rpm	FWD	4630.0	1777.0	1432.0	5.0	
400 Nm @ 1750 rpm	AWD	4439.0	1821.0	1612.0	5.0	51.0
250 Nm @ 1250 rpm	FWD	4670.0	1814.0	1476.0	5.0	50.0

Resumiendo…

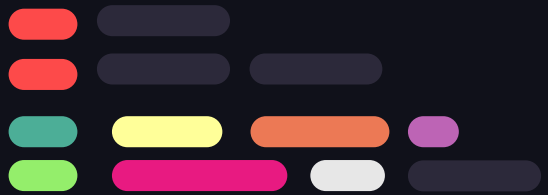
Necesitamos Preprocesar



2.1 PREPROCESADO

```
car_data_1 + car_data_2 + car_data_3 + car_data_4
```

Datos muy diferentes…



2.1 PREPROCESADO

¿Qué tienen en común?


car_data_1

car_data_2 ≡ car_data_3 ≡ car_data_4

2.1 NORMALIZACIÓN

Normalizados → make, model, year, selling_price y km_driven

División de Columnas → name = make y model

Conversión Precio →  a Euros (factor: 0.011)

Ajustes de Valores → fuel, transmission, seller_type y owner

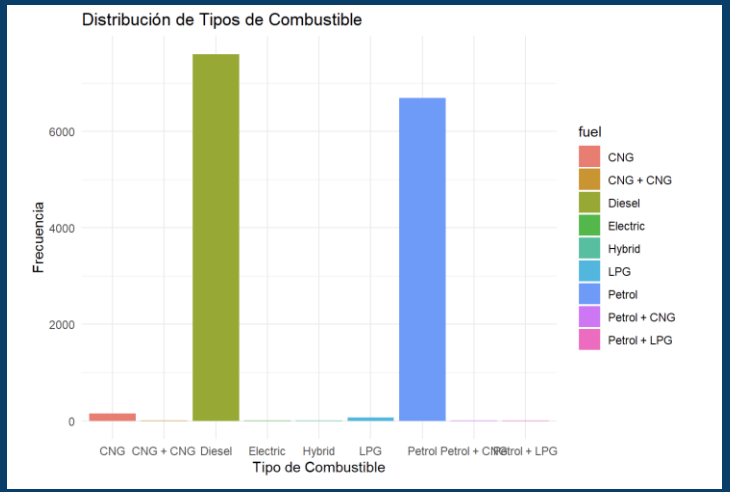
2.1 NORMALIZACIÓN

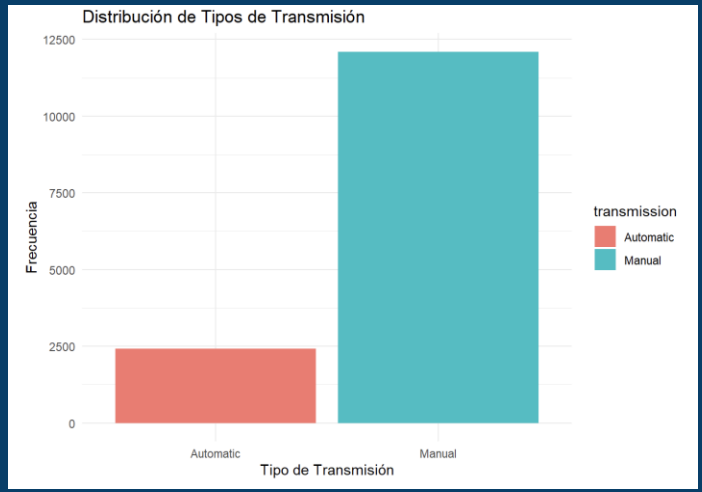
```
car_data_2 + car_data_3 + car_data_4
```

```
car_dataset_total
```

2.2 VISUALIZACIÓN

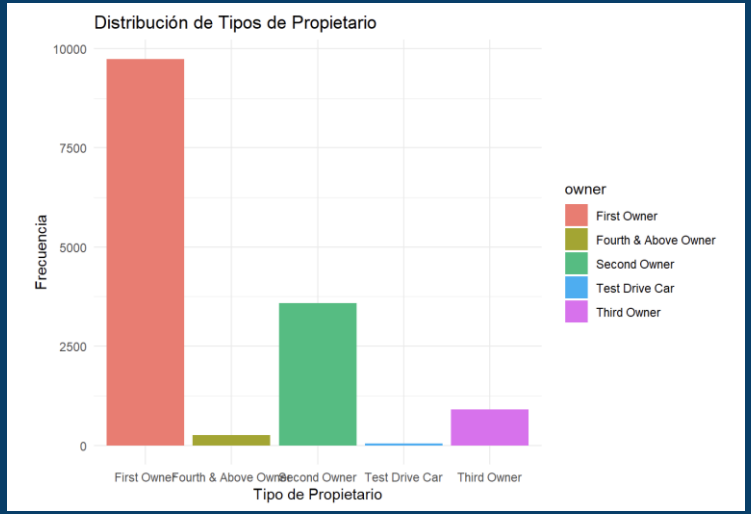
¿Cómo son los coches en este dataset?

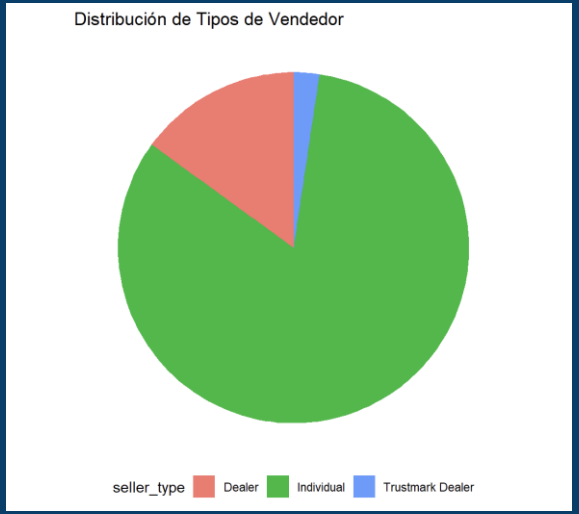




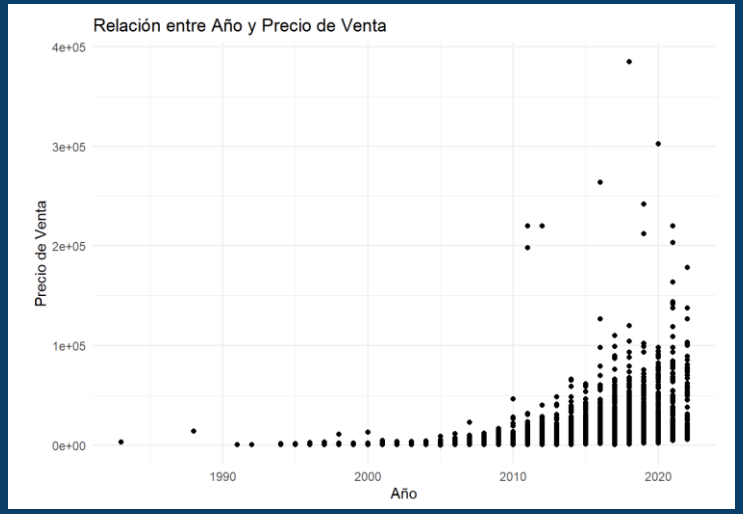
2.2 VISUALIZACIÓN

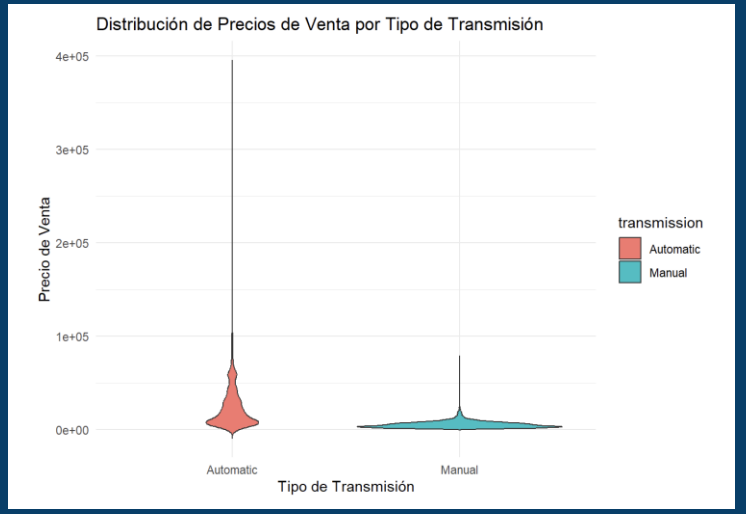
¿Y en relación a las ventas?





2.2 VISUALIZACIÓN

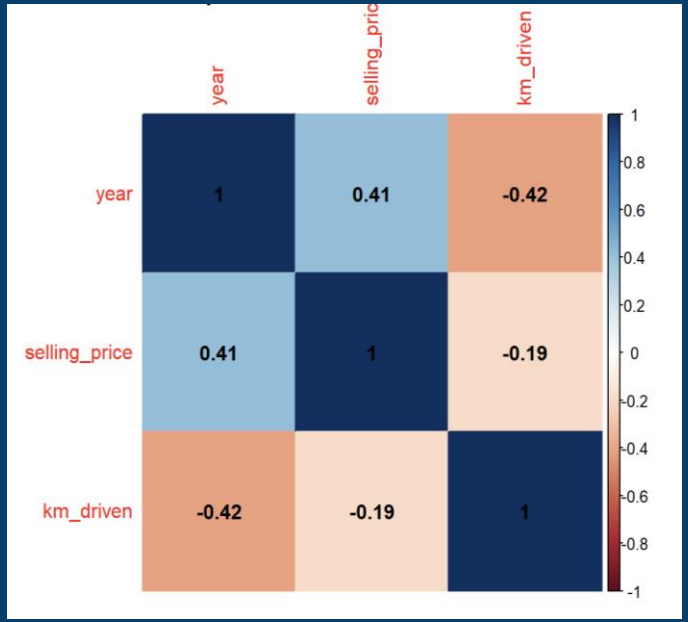




2.2 VISUALIZACIÓN

Mapa de calor

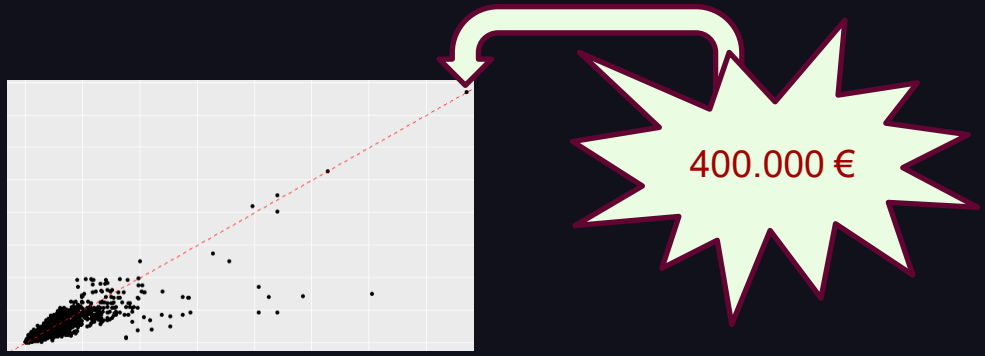
Correlación



2.3 MODELO PREDICTIVO

Tratamiento de los datos

- Eliminación de posibles outliers con un precio mayor o igual a 50,000 €



2.3 MODELO PREDICTIVO

Algoritmos utilizados

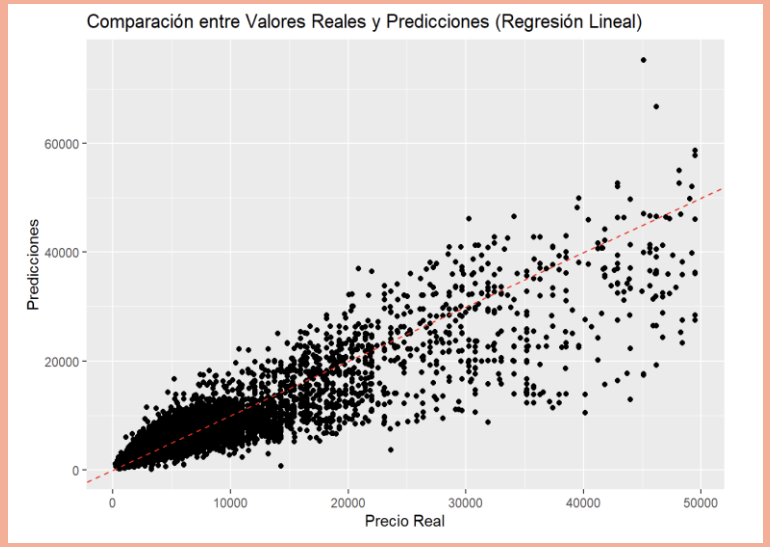
- Random Forest
- Regresión Lineal
- KNN

+

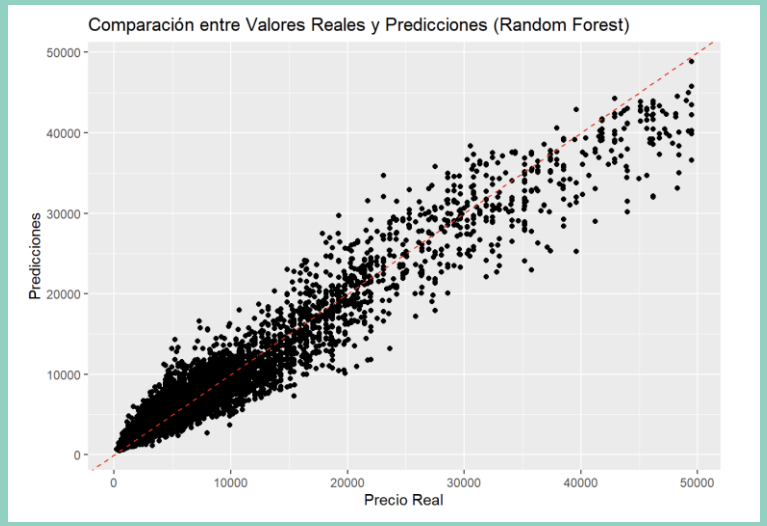
K-fold Cross Validation

2.3.1 REGRESIÓN LINEAL

- MSE: 98,221,616
- R^2: 0.61
- MAE: 6,827.26



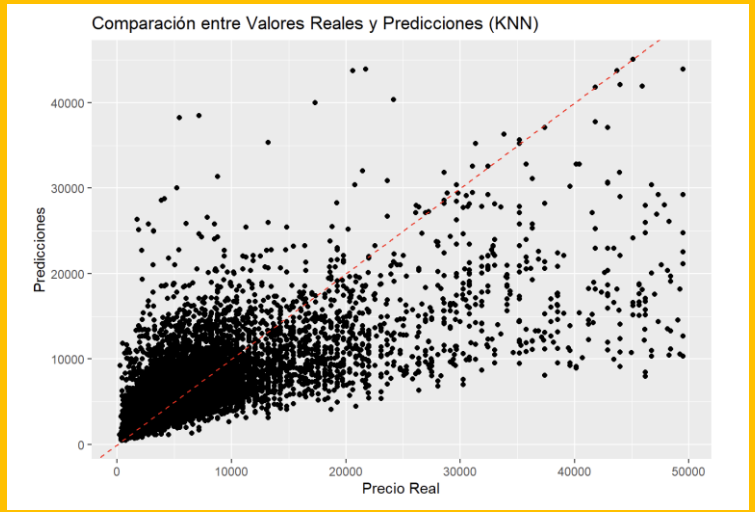
2.3.2 RANDOM FOREST



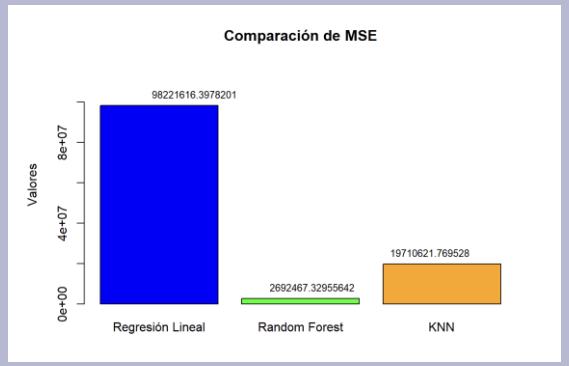
- MSE: 2,692,467
- R^2: 0.95
- MAE: 1,004.29

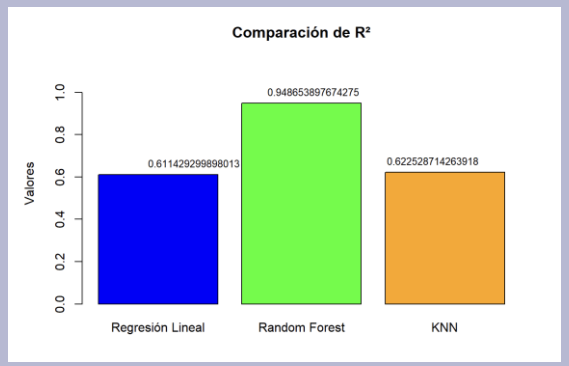
2.3.3 KNN

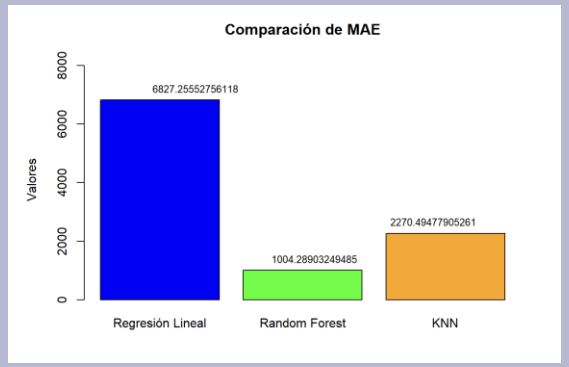
- MSE: 19,710,622
- R^2: 0.62
- MAE: 2,270.50



2.3 COMPARACIÓN







Mejor algoritmo → RANDOM FOREST

2.4 MODELO DESCRIPTIVO: clustering

Algoritmos utilizados

K-medias

Clustering Jerárquico

2.4.1 K-MEDIAS

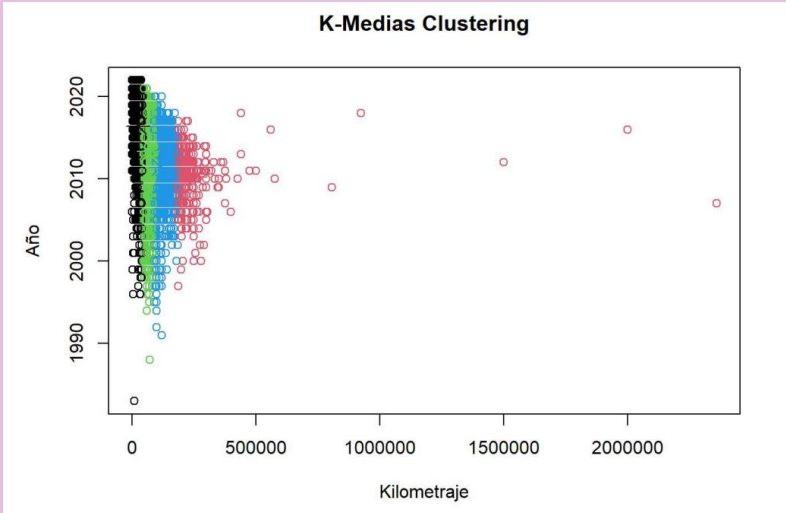
Clusters Identificados:

#1 4957 — Bajo: Promedio 2016

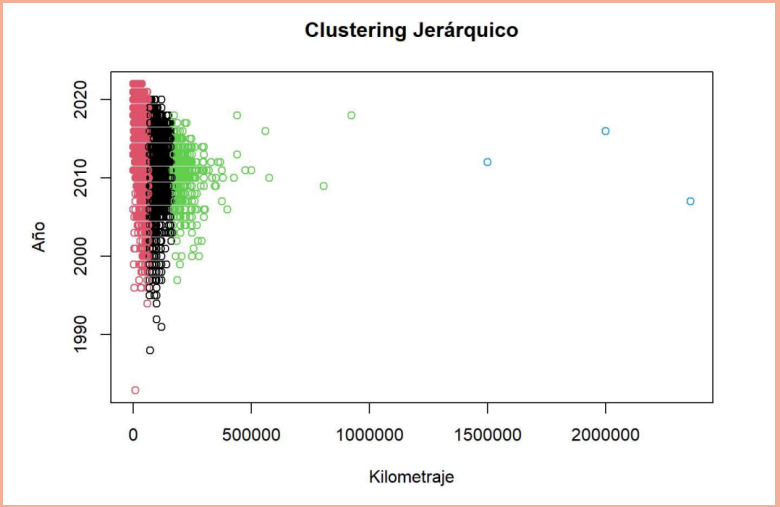
#2 5523 — Medio: Promedio 2013

#3 3532 — Alto: Promedio 2011

#4 312 — Muy alto: Promedio 2010

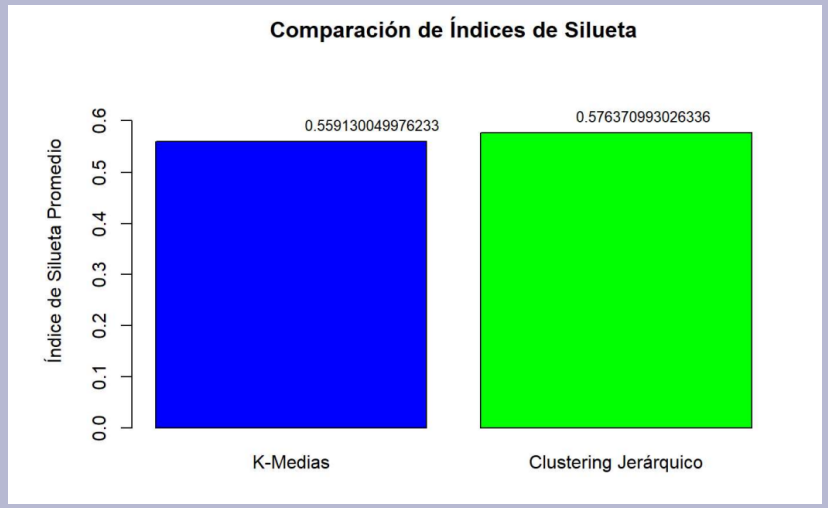


2.4.2 CLUSTERING JERÁRQUICO



Estructura Jerárquica:
Tres grandes clusters y un cuarto con valores alejados

2.4 COMPARACIÓN: Índices de silueta



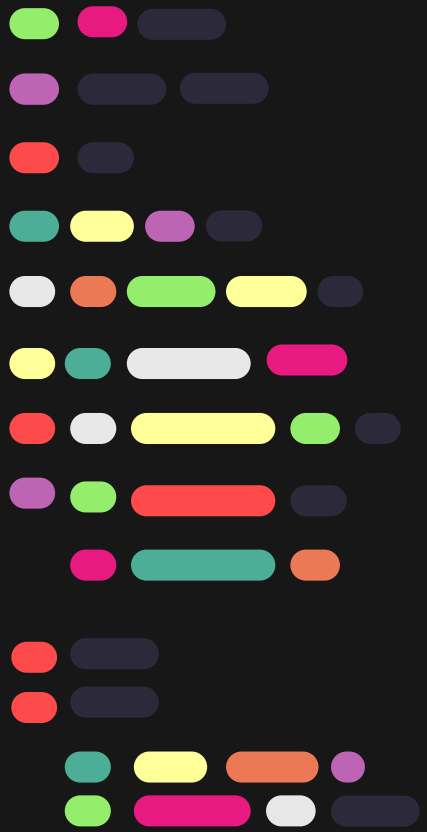
03 CONCLUSIÓN

Predicción → RANDOM FOREST

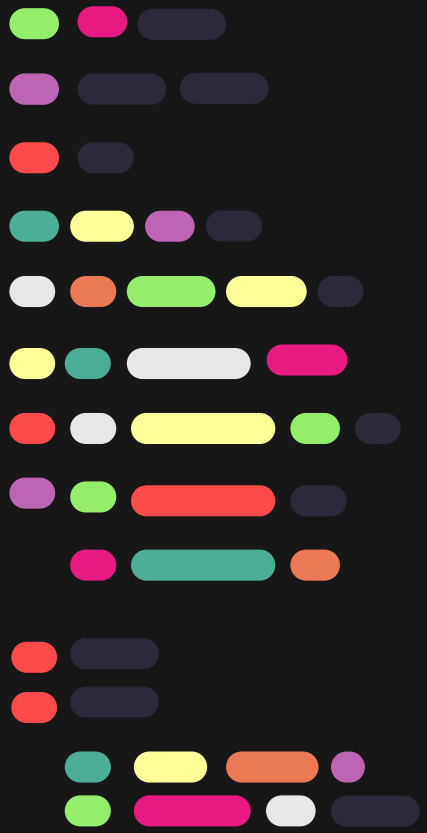
↓

> tiempo de ejecución

Clustering → RESULTADOS SIMILARES



¿PREGUNTAS?



¡MUCHAS GRACIAS POR VUESTRA ATENCIÓN! ;)