MODELOS DE ATRACCION-GENERACION

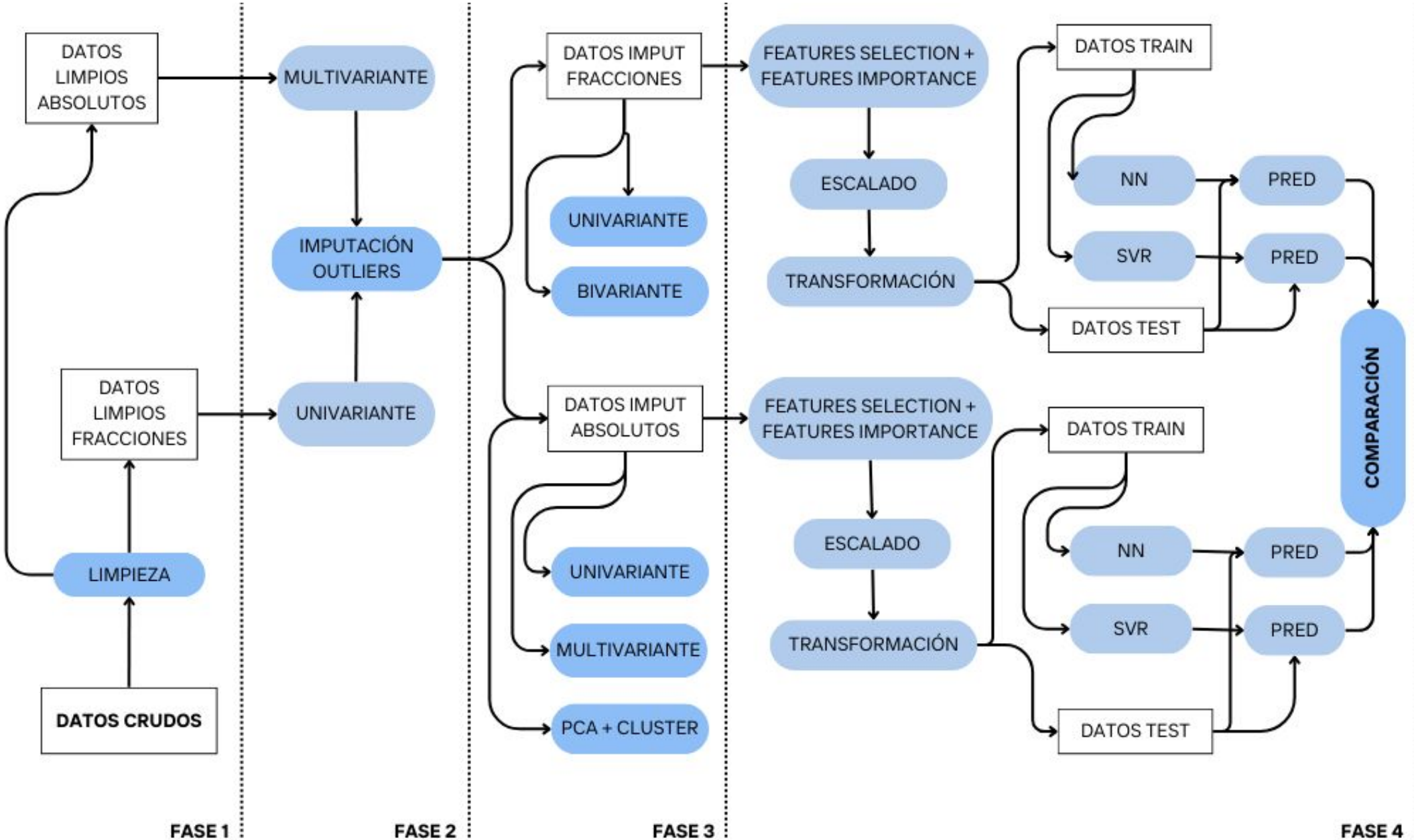
GRUPO 10

Joao Paulo Scabora | Juan Carlos Rubio | Julen Larranaga | Lucas Ezequiel | Lucia Lopez

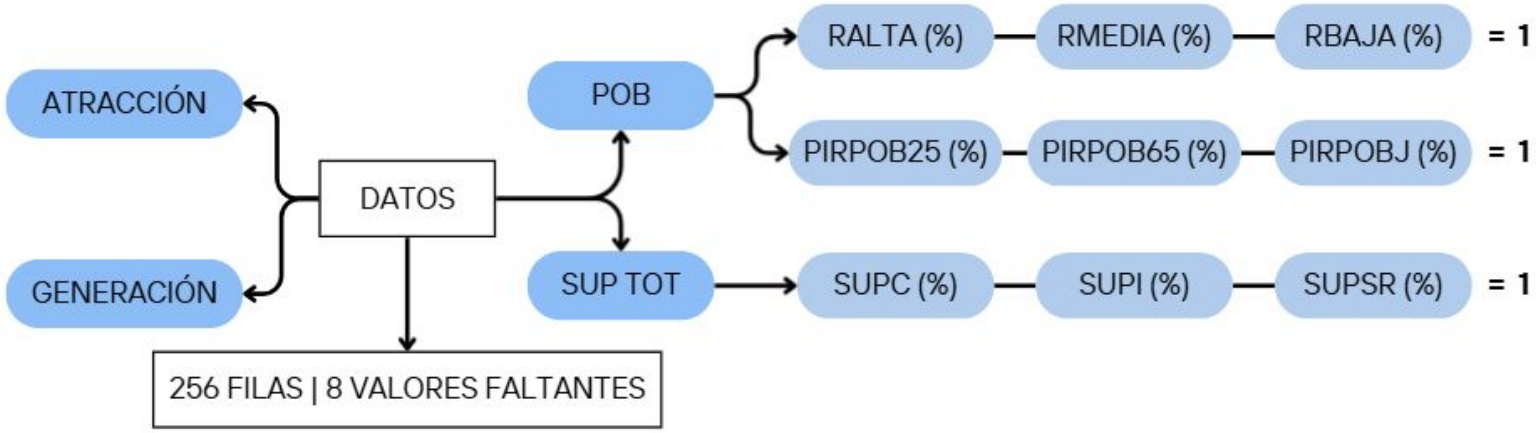
OBJETIVOS DEL PROYECTO

COMPARAR LA EFECTIVIDAD DE SVR Y REDES NEURONALÆS EN LA PREDICCION.

MEJORAR LA PLANIFICACION URBANA Y LA TOMA DE DECISIONES EN POLITICAS PUBLICAS.


DATOS LIMPIOS ABSOLUTOS
MULTIVARIANTE
IMPUTACIÓN OUTLIERS
UNIVARIANTE
DATOS LIMPIOS FRACCIONES
UNIVARIANTE
LIMPIEZA
DATOS CRUDOS
DATOS IMPUT FRACCIONES
UNIVARIANTE
BIVARIANTE
DATOS IMPUT ABSOLUTOS
UNIVARIANTE
MULTIVARIANTE
PCA + CLUSTER
FEATURES SELECTION + FEATURES IMPORTANCE
ESCALADO
TRANSFORMACIÓN
FEATURES SELECTION + FEATURES IMPORTANCE
ESCALADO
TRANSFORMACIÓN
DATOS TRAIN
NN
SVR
DATOS TEST
PRED
PRED
DATOS TRAIN
NN
SVR
DATOS TEST
PRED
PRED
COMPARACIÓN
FASE 1
FASE 2
FASE 3
FASE 4

FASE 1: LIMPIEZA



dataset original:

NZ	POB	RALTA	RMEDIA	RBAJA	PIRPOB25	PIRPOB65	PIRPOBJ	SUPTOT	SUPSR	SUPC	SUPI	A	G
1	12356	0.354	0.1103	0.5357	0.2135	0.6154	0.1711	8.17	0.9161	0.0696	0.0144	119310	121922
2	13198	0.4309	0.2903	0.2788	0.2655	0.5852	0.1492	7.861	0.8559	0.0864	0.0577	150816	148812
29	13905	0.676	0.1269	******	0.259	0.6004	0.1406	9.202	0.8046	0.1222	0.0732	97962	98231
53	30574	******	0.3693	0.57	0.2527	0.5833	0.164	8.38	0.6846	0.2404	0.075	154867	152287

nuevos nombres columnas:

NZ	POB	R_ALTA	R_MEDIA	R_BAJA	POB_25	POB_65	POB_2565	SUP_TOT	SR_SUP	C_SUP	I_SUP	ATRACCION	GENERACION

1 CALCULAR FALTANTES

Ralta = 1 - Rmedia - Rbaja

2 ENCONTRAR OUTLIERS

OUT1: ∑ grupos de fracciones ≠ 1

OUT2: Valores de fraccion < 0

3 IMPUTAR OUTLIERS

MICE

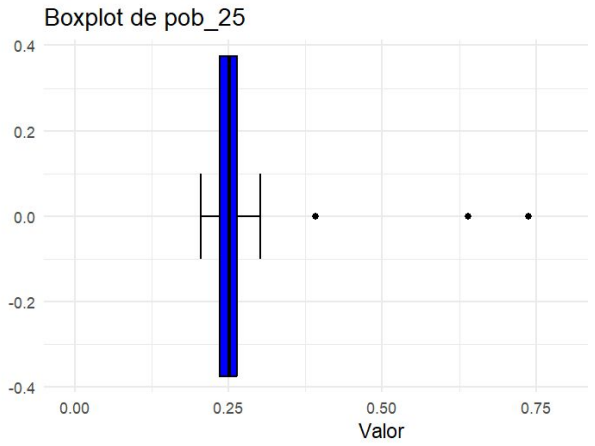
Reajuste para que ∑ gdf=1

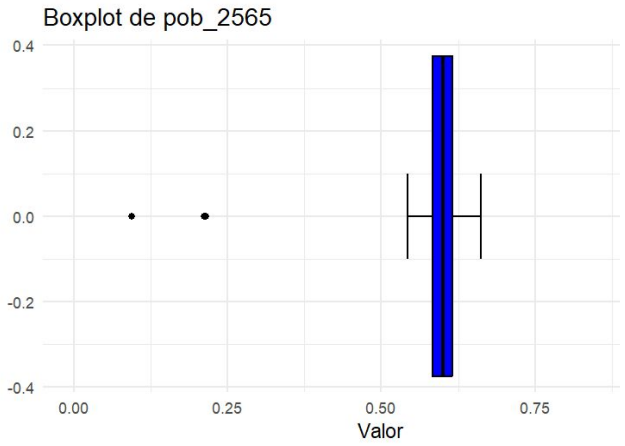
↓

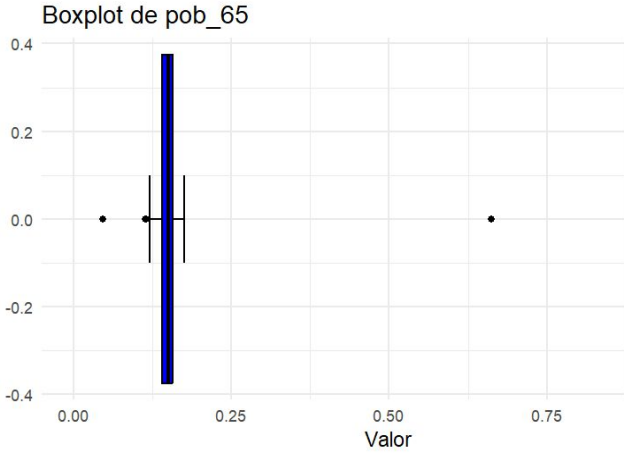
8 valores imputados

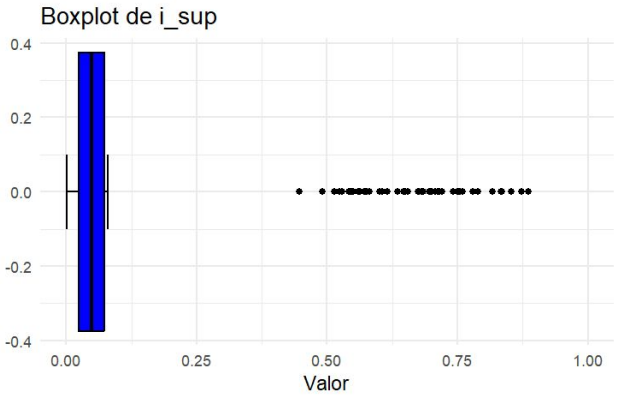
FASE 2: OUTLIERS

UNIVARIANTE

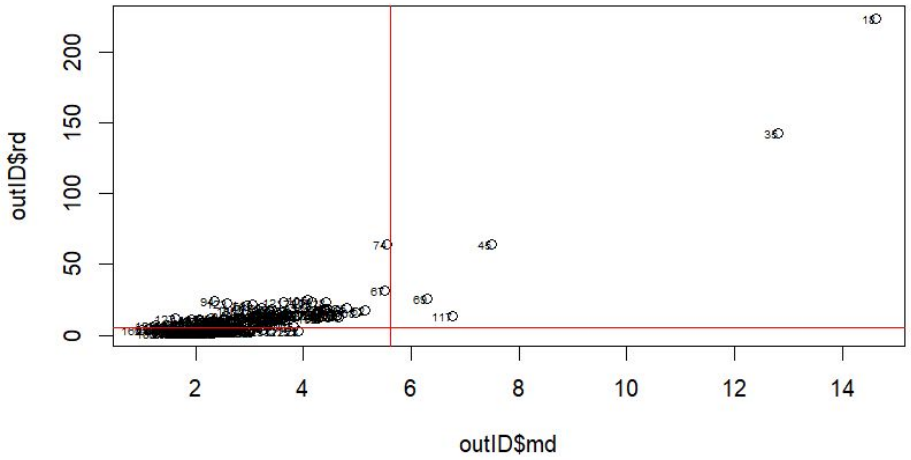








MULTIVARIANTE: DISTANCIA DE MAHALANOBIS

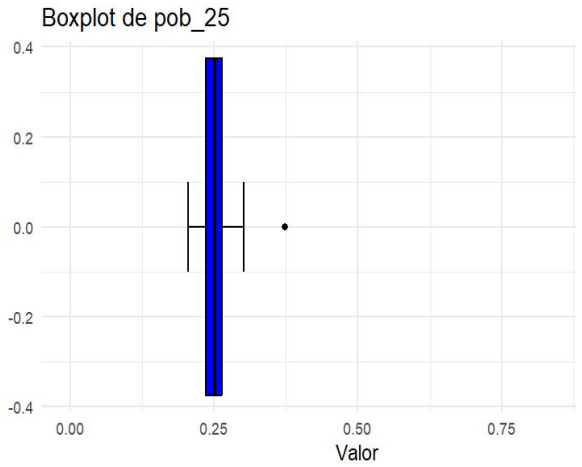


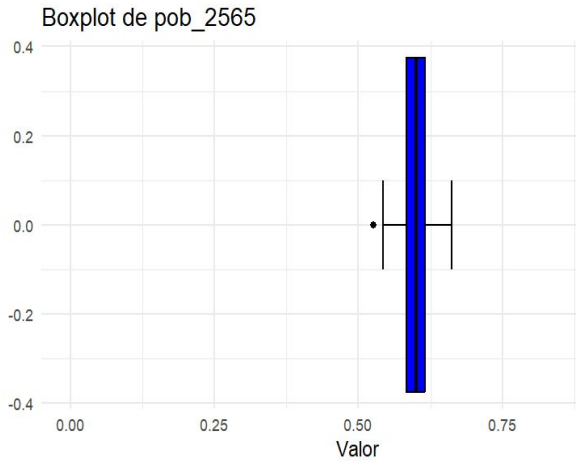
MICE

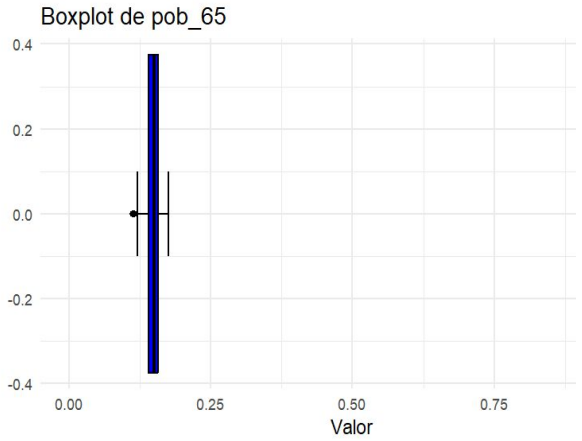
FILA (ID)	OUTLIER
18	Univariante + Mahalanobis
35	Univariante + Mahalanobis
45	Univariante + Mahalanobis
69	Mahalanobis (Cerca del límite)
74	Univariante (Cerca del límite)
94	Univariante (Cerca del límite)
117	Mahalanobis (Cerca del límite)

FASE 3: ANALISIS EXPLORATORIO DE DATOS

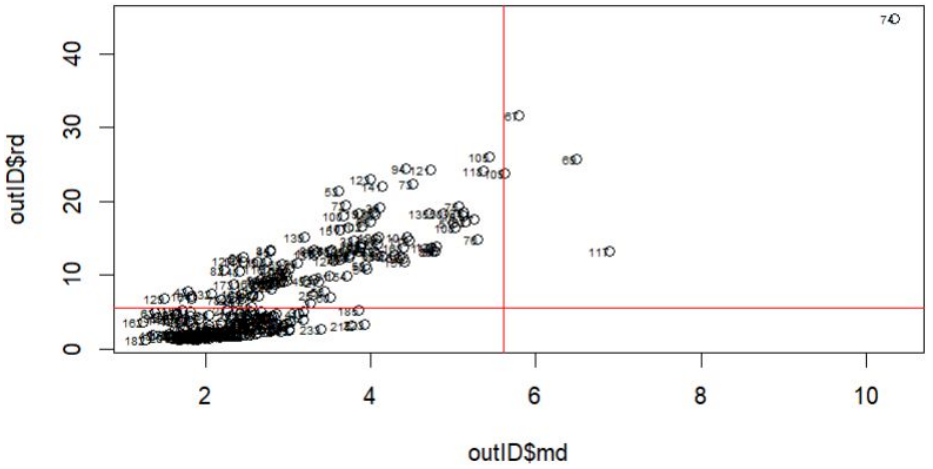
UNIVARIANTE | COMPROBACION ELIMINACION DE OUTLIERS |

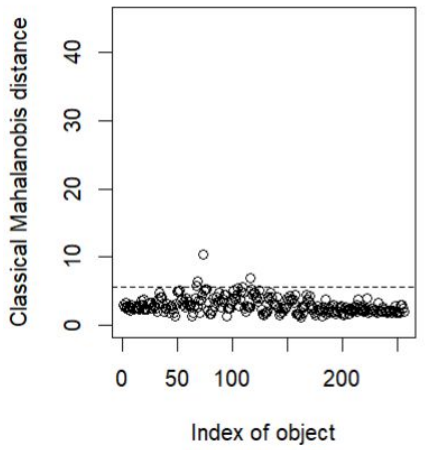


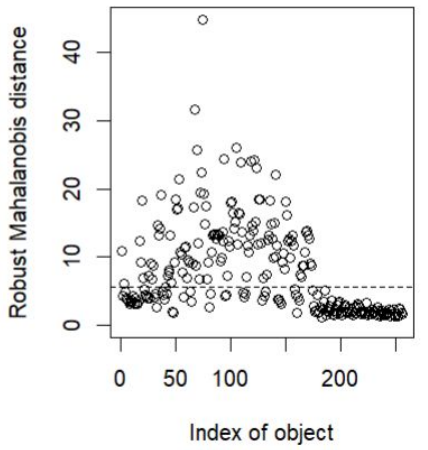




MULTIVARIANTE: DISTANCIA DE MAHALANOBIS

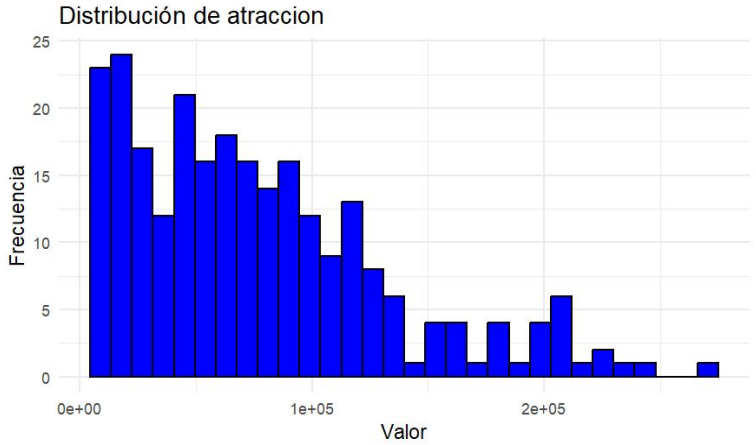


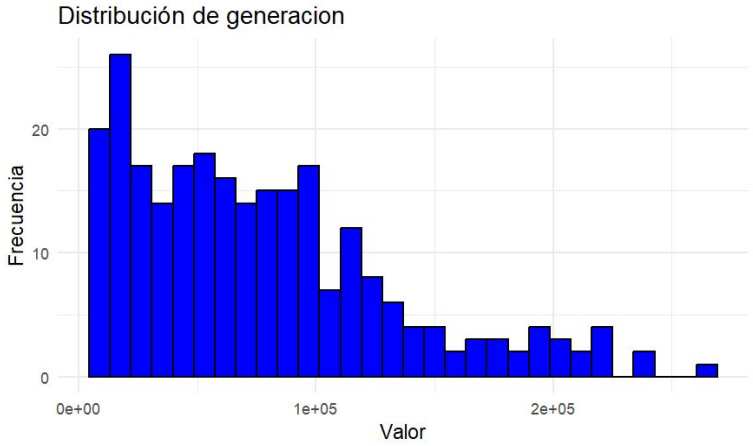


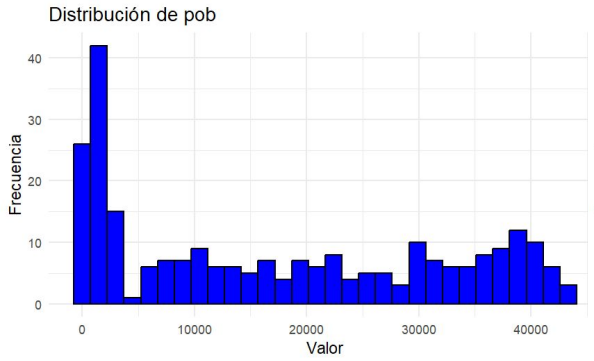


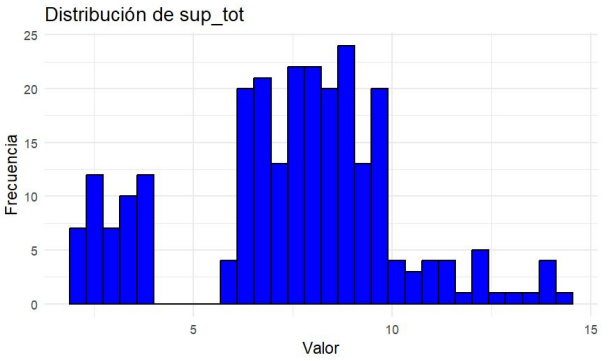
FASE 3: ANALISIS EXPLORATORIO DE DATOS

UNIVARIANTE









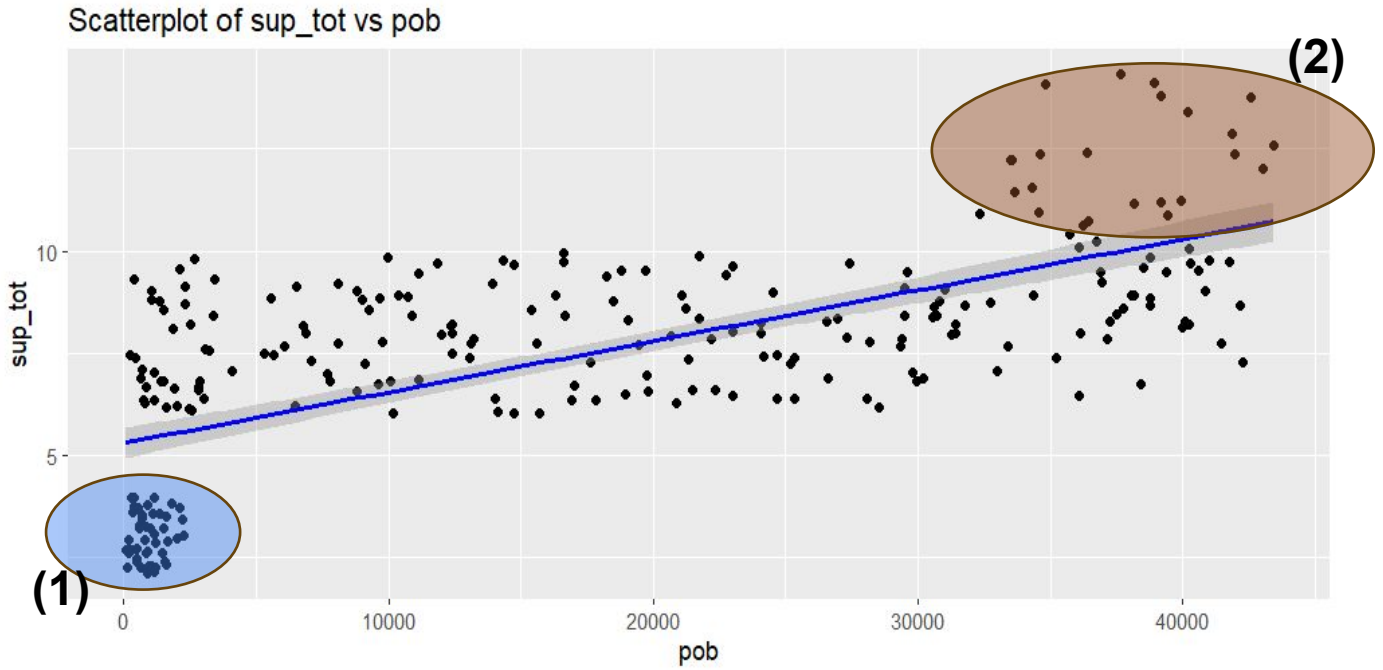
VARIABLE	MEDIA A	SD	MIN	P25	MEDIANA	P75	MAX
pob	17.407	14.660	95	2.002	15.064	31.058	43.416
r_alta	5.256	5.463	16	468	3.694	8.174	22.914
r_media	5.608	5.269	32	588	4.587	8.708	20.120
r_baja	6.543	6.142	48	721	4.600	11.296	27.400
pob_25	4.355	3.704	19	494	3.774	7.874	11.499
pob_2565	10.461	8.816	59	1.216	9.068	18.776	26.258
pob_65	2.593	2.203	17	302	2.174	4.503	7.089
sup_tot	7,48	2,67	2,11	6,38	7,81	9,02	14,3
sr_sup	4,98	2,62	0,0167	4,34	5,53	6,76	10,2
c_sup	1,83	0,997	0,255	1,04	1,78	2,4	5,03
i_sup	0.662	0,725	0,00153	0,204	0,415	0,645	3,22
atraccion	76.593	55.881	5.522	31.231	65.265	107.264	267.001
generacion	76.593	55.642	5.266	31.182	66.168	106.121	261.332

FASE 3: ANALISIS EXPLORATORIO DE DATOS

BIVARIANTE

- Al comparar la superficie total y poblacion, es posible visualizar zonas escasamente pobladas, con una superficie menor en km². **(1)**

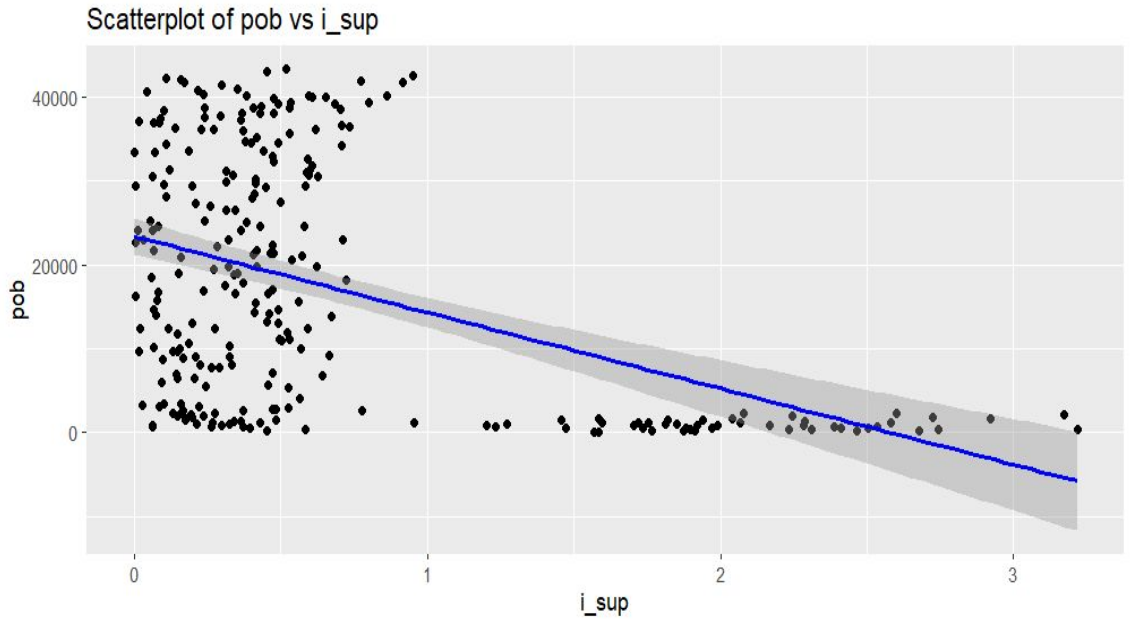
- Dentro del segundo grupo mas grande, tambien se observa un subgrupo caracterizado por altas poblaciones y superficies totales. **(2)**

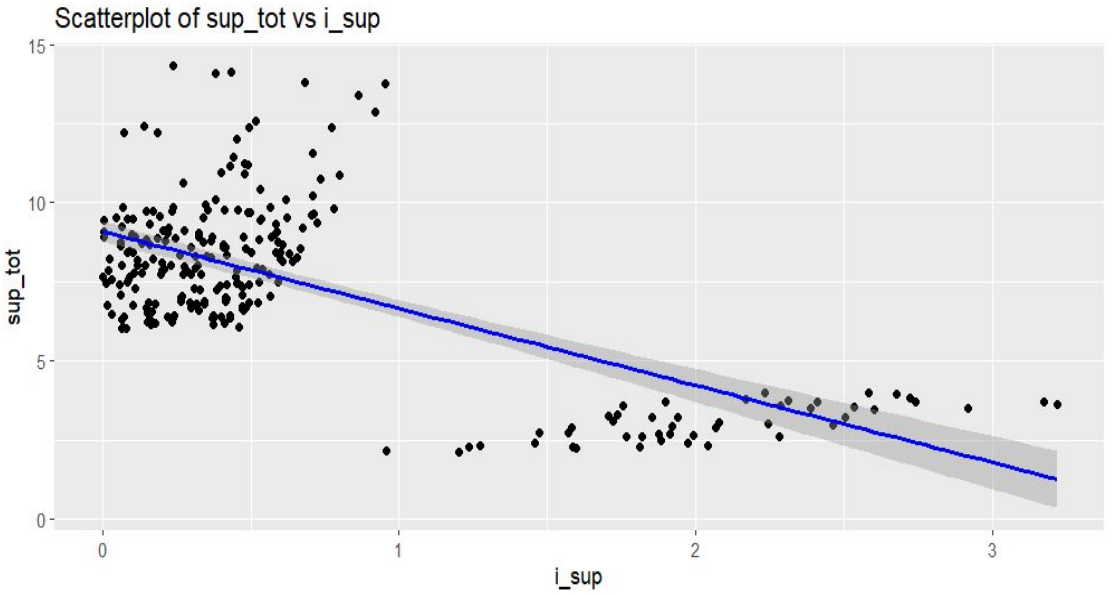


FASE 3: ANALISIS EXPLORATORIO DE DATOS

BIVARIANTE

- **Indicacion de Zona industriales**: areas con una gran concentracion de km² destinado a actividad industrial, reconocidas por ser areas con baja poblacion, altamente concentradas y, en este caso, un tamano menor que el resto





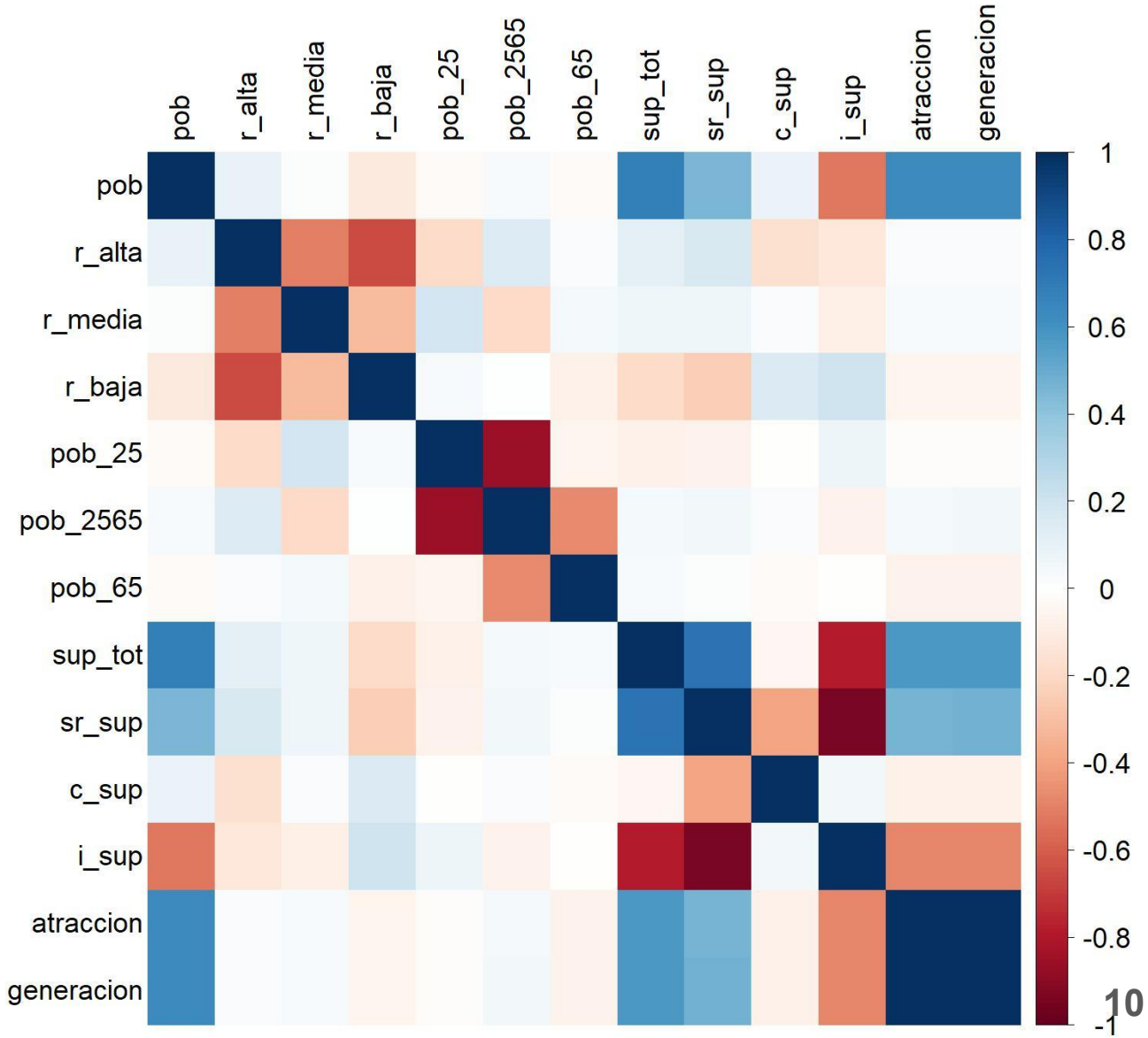
FASE 3: ANALISIS EXPLORATORIO DE DATOS

BIVARIANTE

Heatmap con datos fraccionales

- Cuanto más ocupada está la superficie con actividades industriales, menor es la población presente . La atracción y generación de viajes se ven negativamente afectadas en estas áreas

- La variable población también muestra altas correlaciones, pero positivas, con la superficie total (sup_tot), las zonas residenciales(sr_sup), y la atracción y generación de viajes

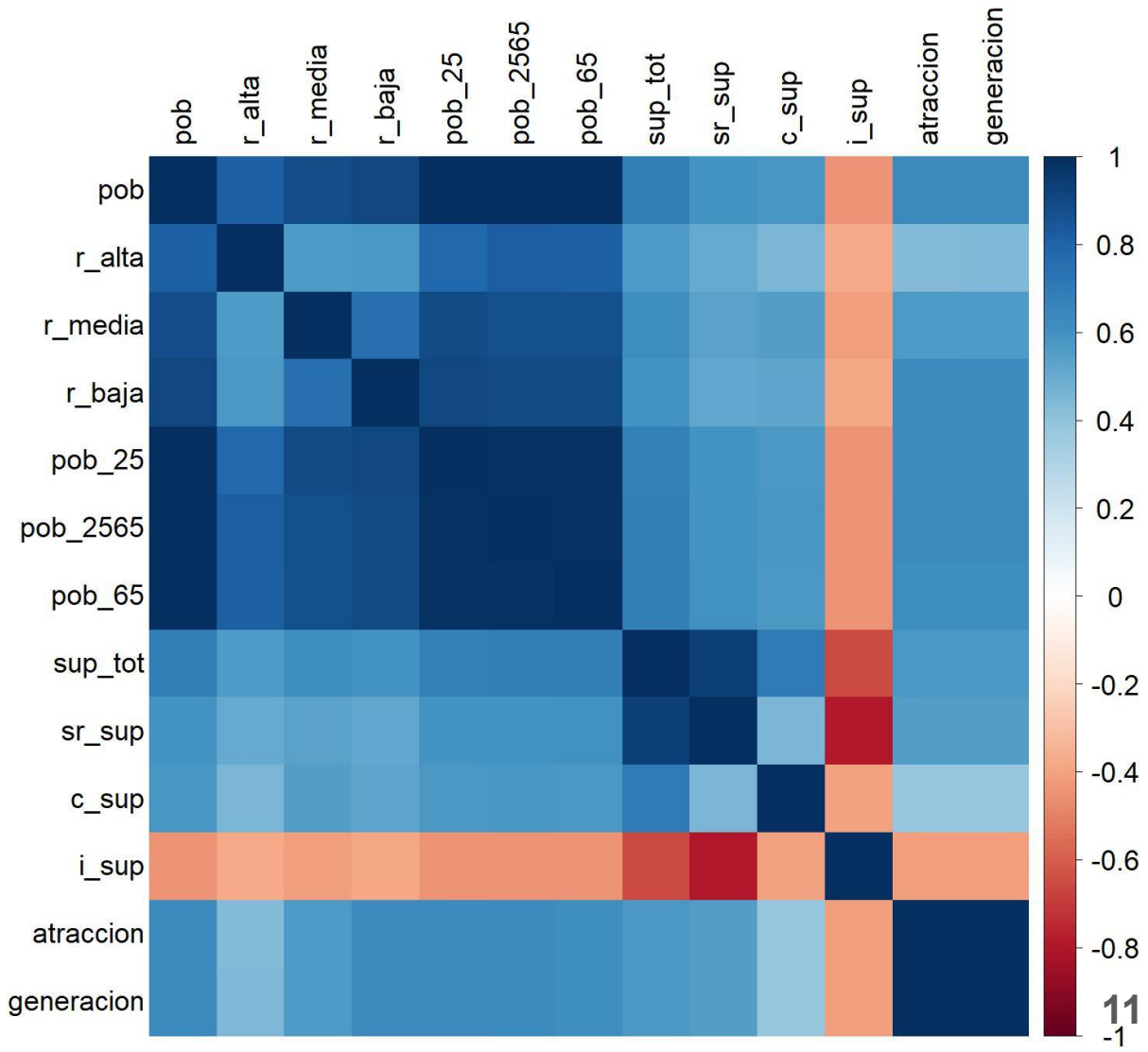


FASE 3: ANALISIS EXPLORATORIO DE DATOS

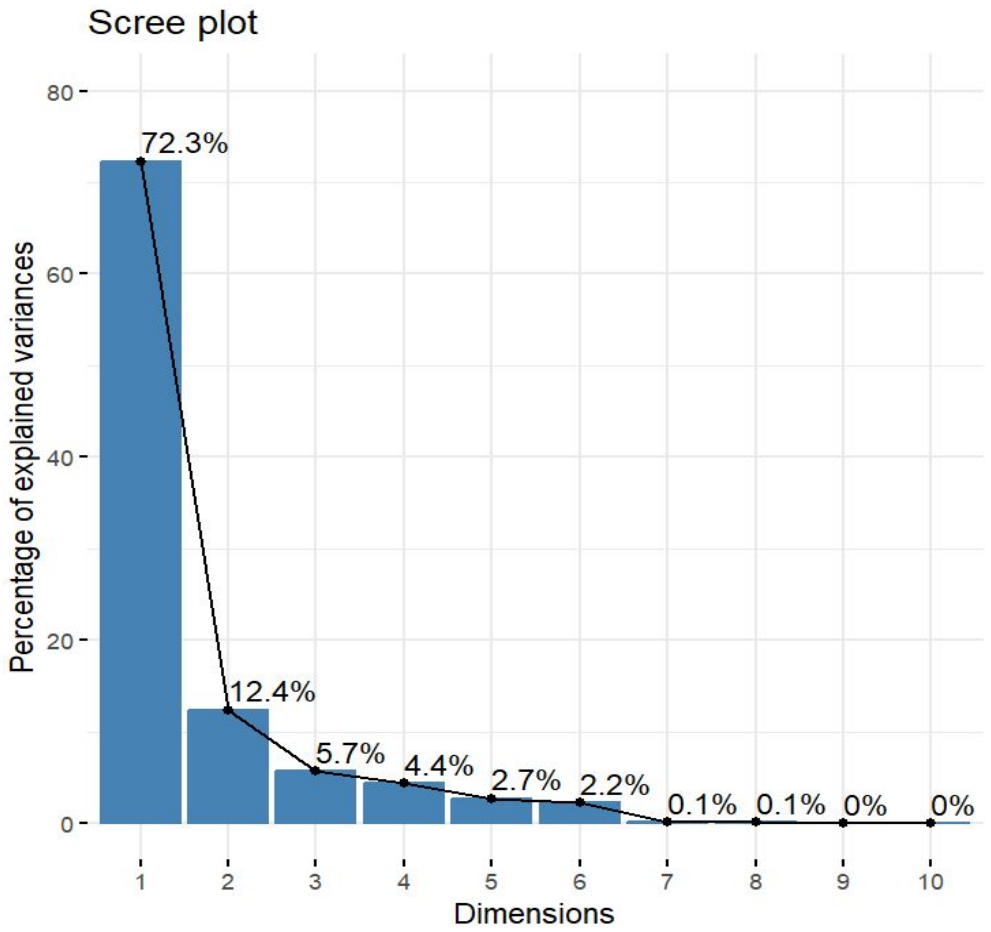
BIVARIANTE

Heatmap con datos absolutos

Elegimos por utilizar el dataset con fracciones, ya que al considerar la base con valores absolutos se ve que todo el conjunto tiene correlacion con la salvedad que destaca de la correlacion negativa de la superficie industrial (i_sup) con las demas variables.

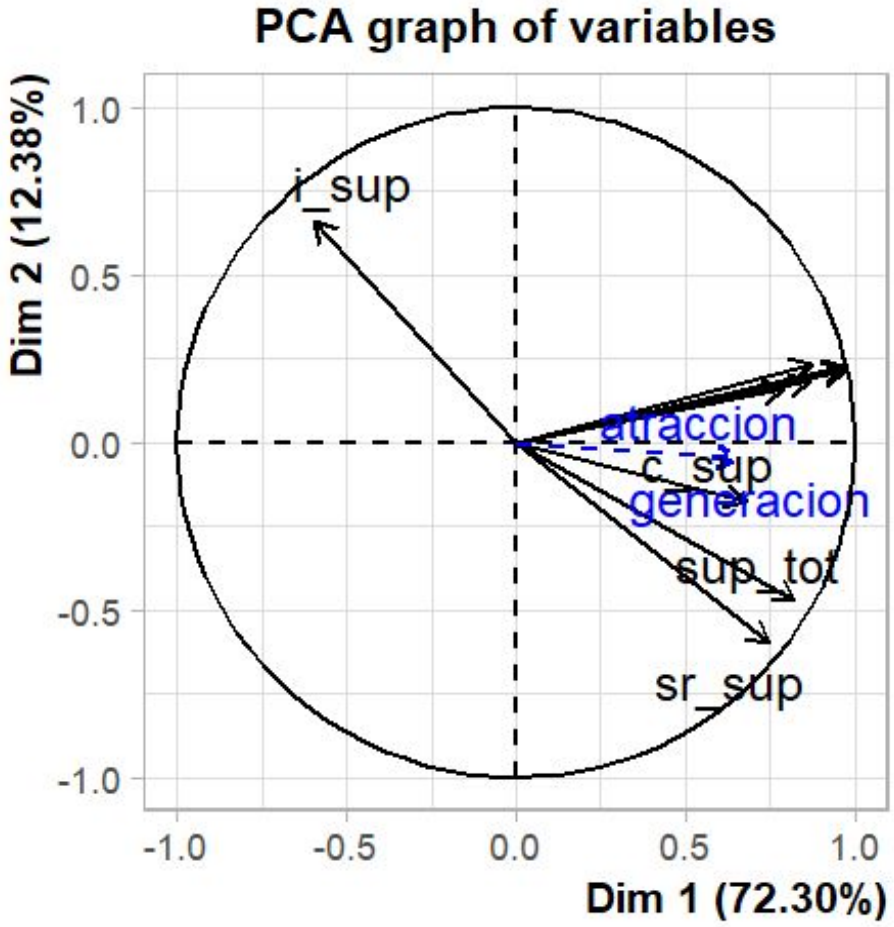


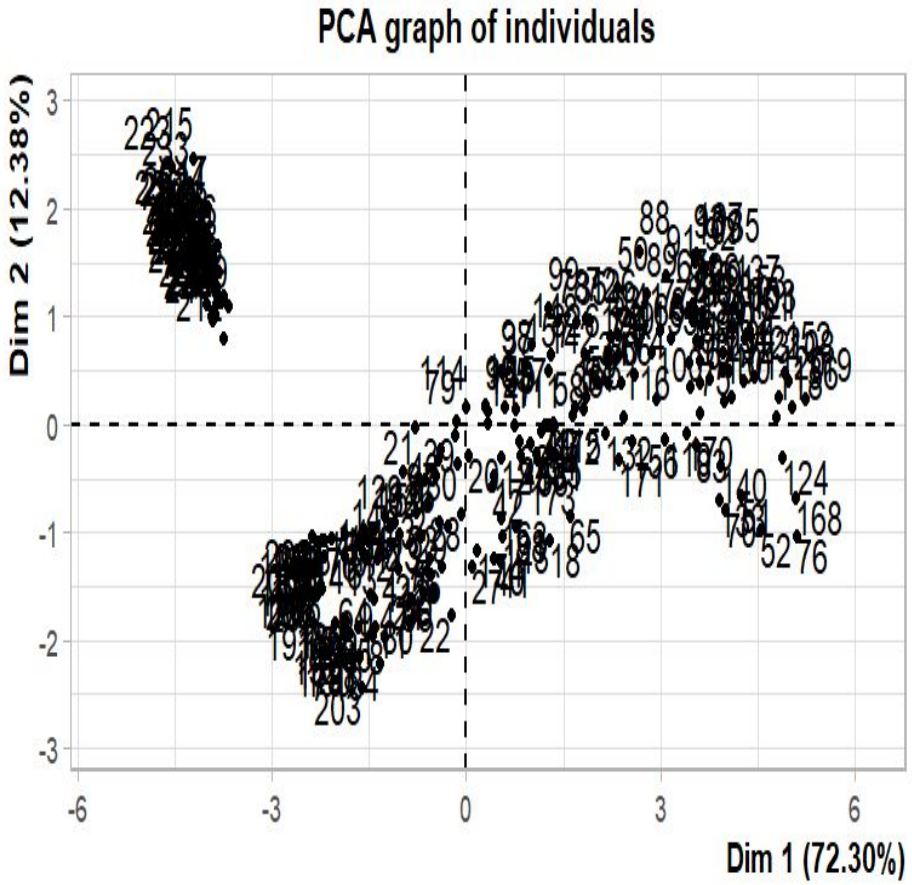
FASE 3: COMPONENTES PRINCIPALES



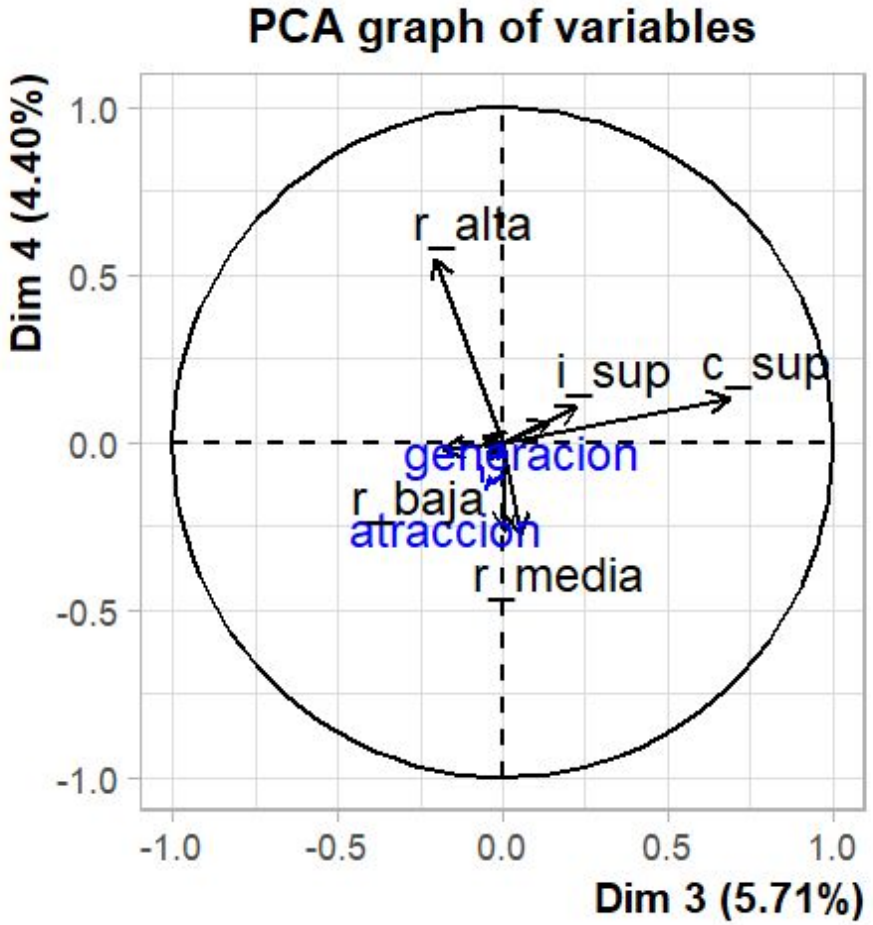
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8
Variance	7,953	1,363	0,628	0,484	0,297	0,247	0,016	0,014
% of var.	72,299	12,380	5,709	4,398	2,701	2,243	0,147	0,124
Cumulative % of var.	72,299	84,678	90,387	94,785	97,486	99,729	99,876	100

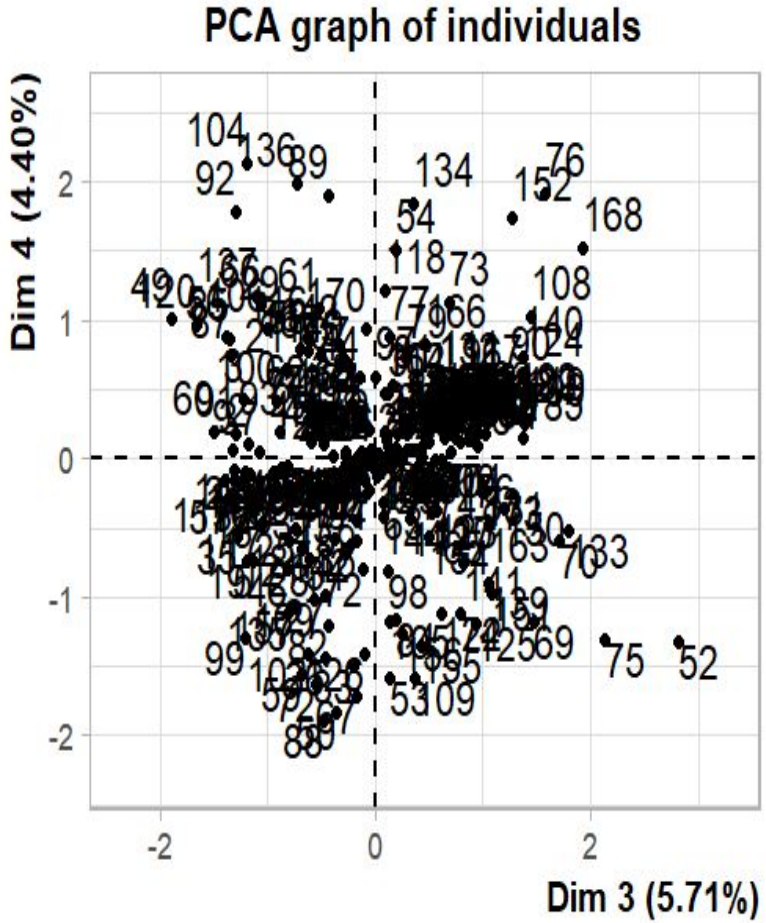
FASE 3: COMPONENTES PRINCIPALES





FASE 3: COMPONENTES PRINCIPALES

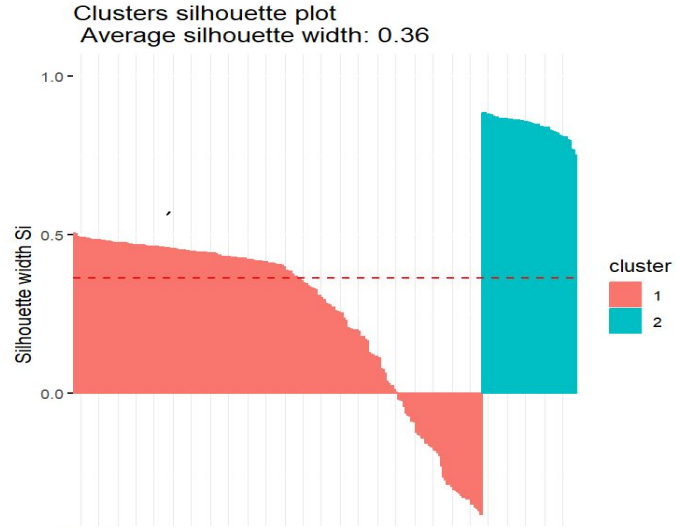


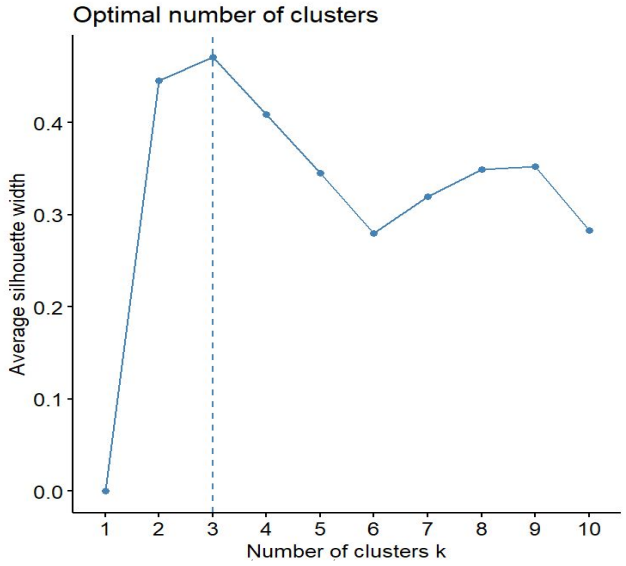


FASE 3: EDA

CLUSTERING

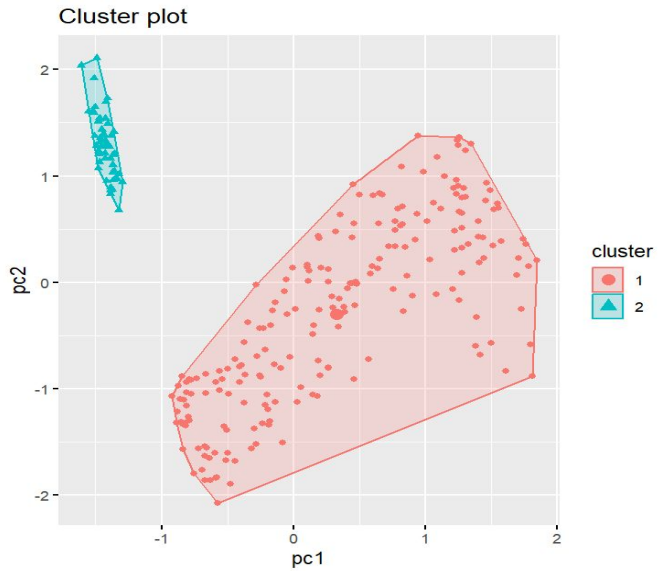
Numero optimo de clusters del metodo de **clustering no-jerarquico dbscan**





Numero optimo de clusters por la metrica **del ancho de la silueta**

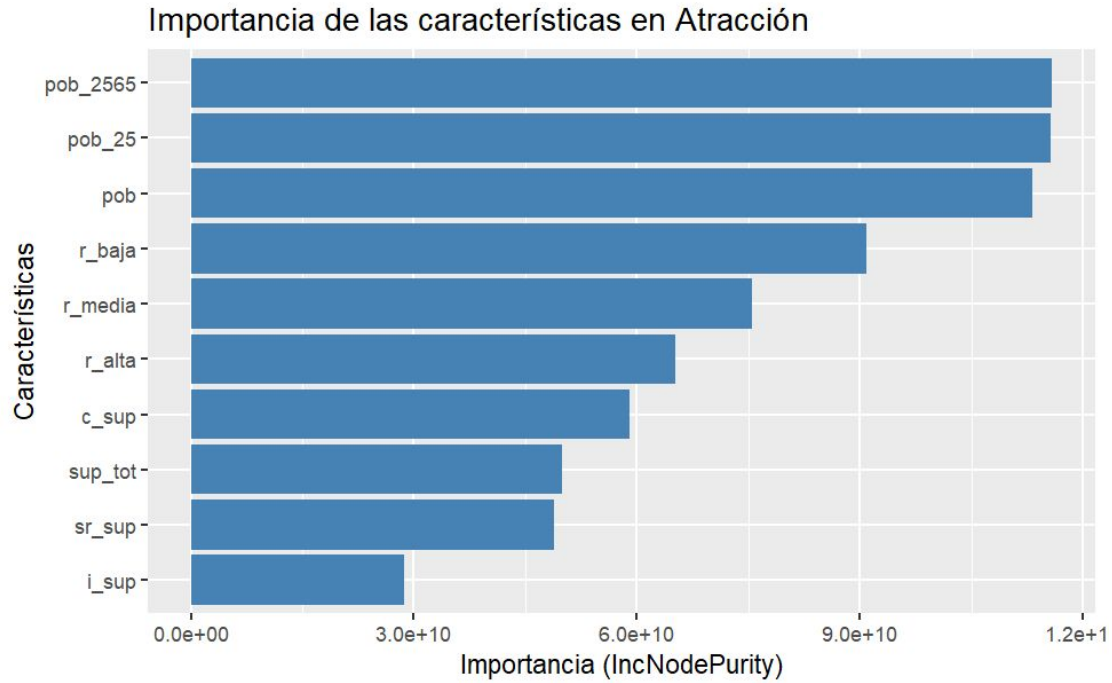
Clusterizacion **jerarquica HCPC**



Clusterizacion **no-jerarquica dbscan**

FASE 4: GENERACION DE MODELOS

FEATURE SELECTION Y FEATURE IMPORTANCE



Los calculos de importancia de caracterısticas se han realizado sobre modelos de **Random Forest** usando la medida de *IncNodePurity*, que se calcula directamente a traves del ındice de Gini.

Random forest reduce el riesgo de overfitting, captura relaciones no lineales, considera las interacciones de las caracterısticas, es menos sensible a outlayers.

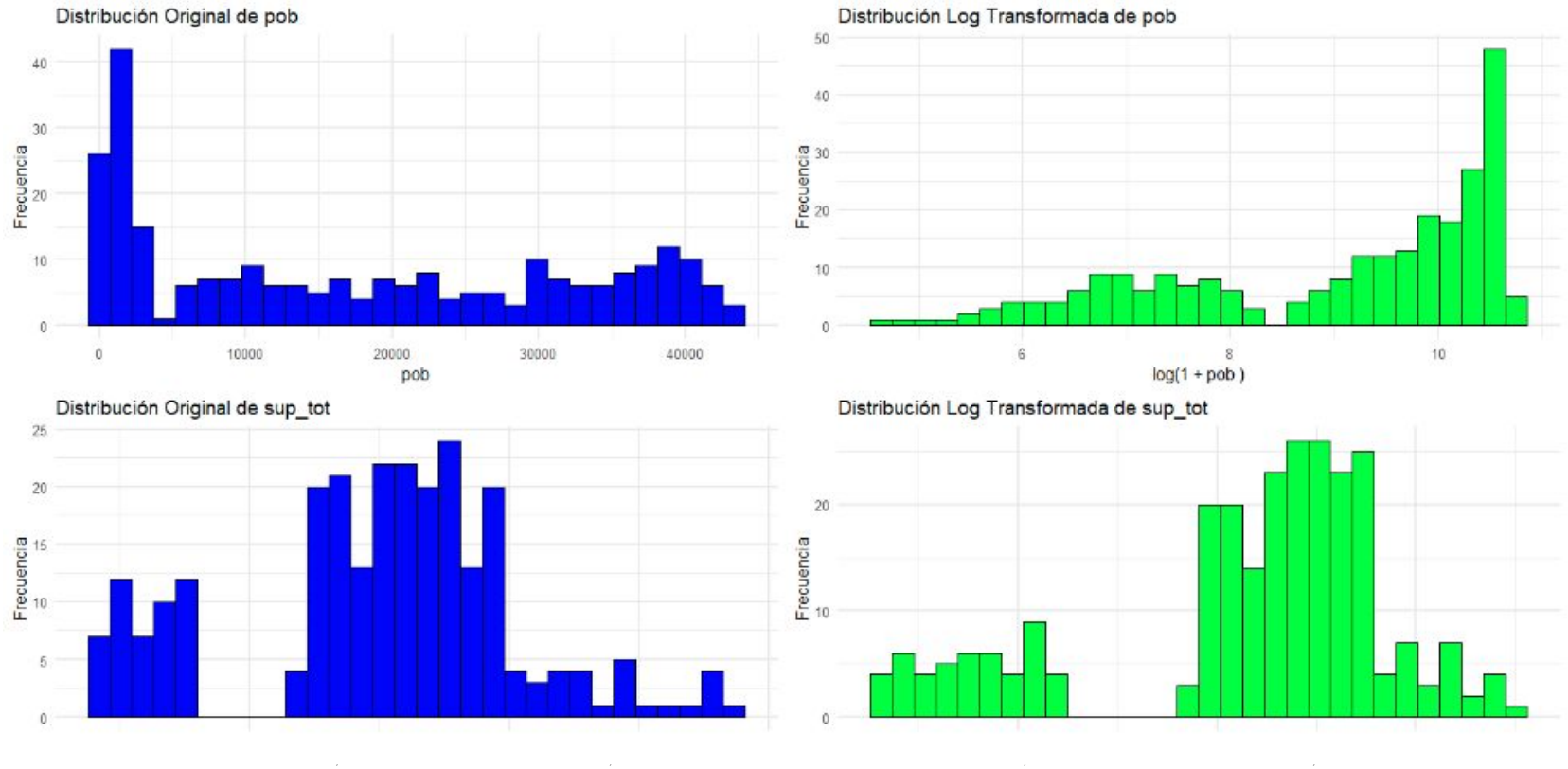
SELECIONADAS

- De los conjuntos de columnas que reflejan fracciones de datos demograficos nos quedamos solo con dos de las tres caracterısticas para mejorar el rendimiento del modelo, evitar problemas de multicolinealidad, reducir el riesgo de sobreajuste y mejorar la eficiencia
- Segun la *feature importance* todas las caracterısticas podrıan tener importancia a la hora de hacer las predicciones, ası que las mantendremos todas
- Eventualmente anadiremos la clusterizacion como una caracterıstica adicional.

pob, *r_media*, *r_alta*, *pob_25*, *pob_2565*, *sup_tot*, *sr_sup*, *i_sup*, *cluster_1*, *cluster_2* y *cluster_3*

FASE 4: GENERACION DE MODELOS

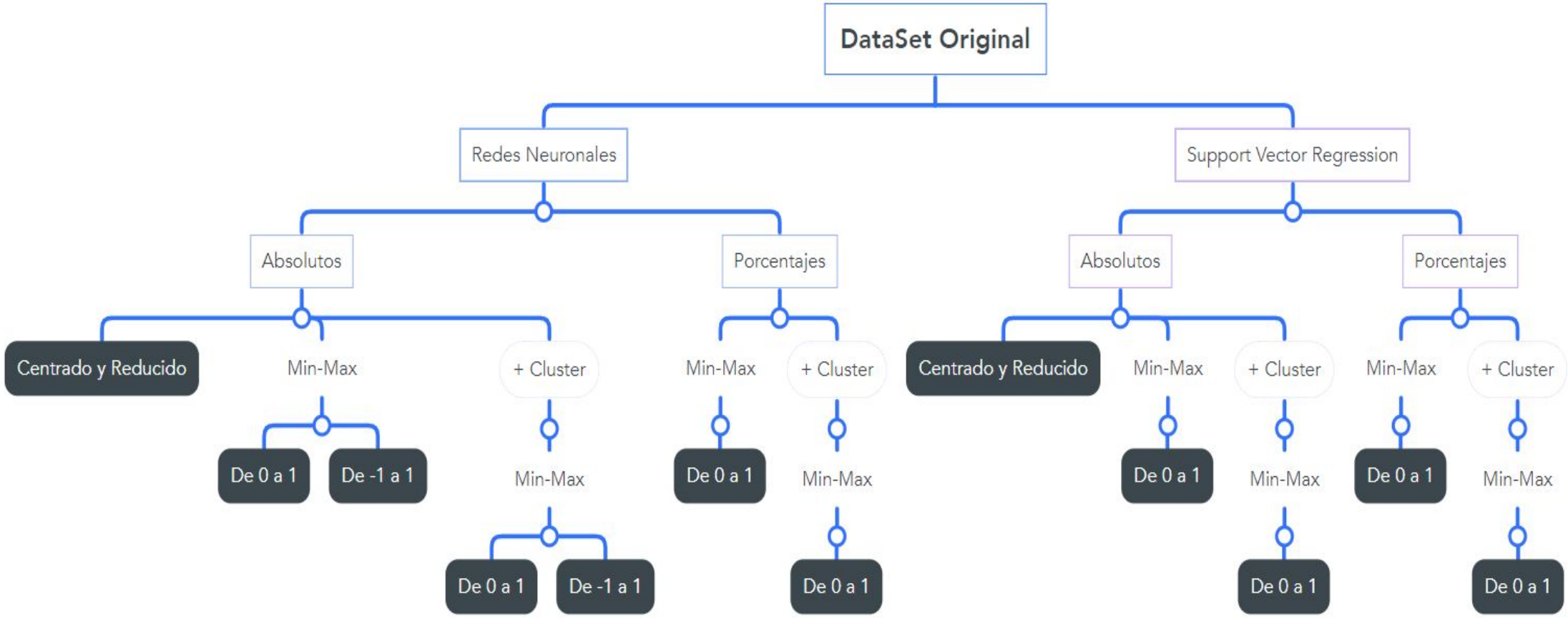
TRANSFORMACIONES



Valoramos aplicar transformaciones logarıtmicas a algunas caracterısticas, pero comprobamos que no tenıa el efecto de normalizacion deseado

FASE 4: GENERACION DE MODELOS

ESCALADO E INGESTA DE DATOS



FASE 4: GENERACION DE MODELOS

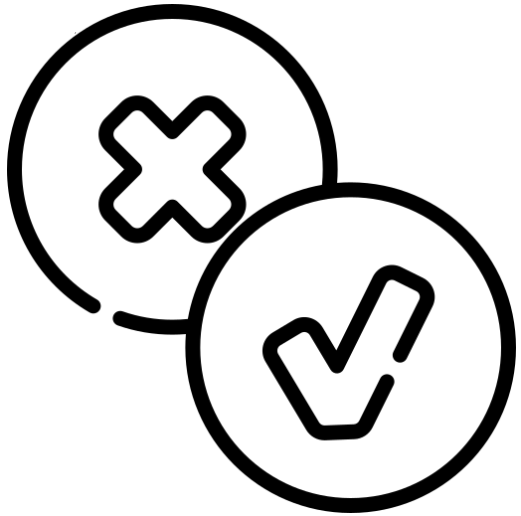
METRICAS DE EVALUACION

RMSLE (Root Mean Squared Logarithmic Error): mide la diferencia entre los valores predichos y los valores reales utilizando logaritmos. Penaliza mas los errores relativos grandes y es util cuando los errores relativos son mas importantes. Un valor mas bajo de RMSLE indica un mejor rendimiento del modelo.

Slope: Es la pendiente de la linea de regresion ajustada entre los valores reales y los valores predichos. Una pendiente cercana a 1 indica que los valores predichos aumentan en una proporcion similar a los valores reales.

Intercept: Es el valor de y para x = 0 en la linea de regresion ajustada. Idealmente, deberia ser cercano a cero si los valores estan bien centrados.

R2 (Coeficiente de Determinacion): Mide la proporcion de la varianza en la variable dependiente que es predecible a partir de las variables independientes. Un R2 cercano a 1 indica que el modelo explica bien la variabilidad de los datos.



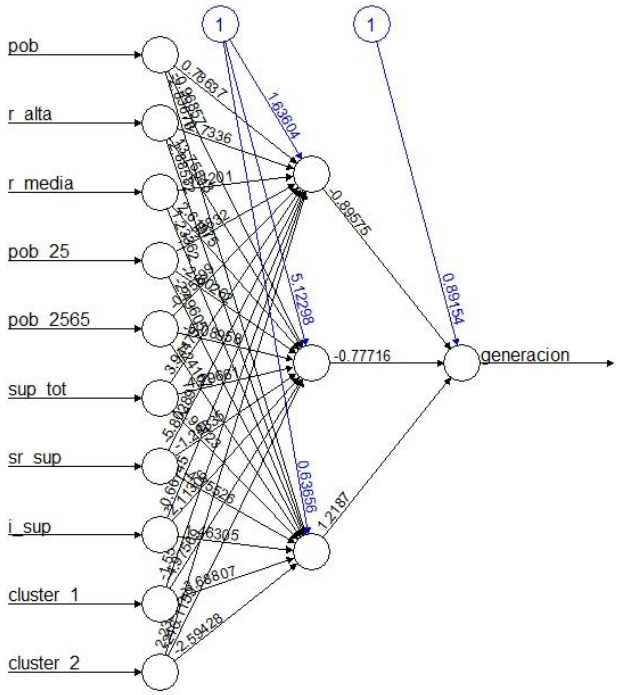
FASE 4: GENERACION DE MODELOS (NN)

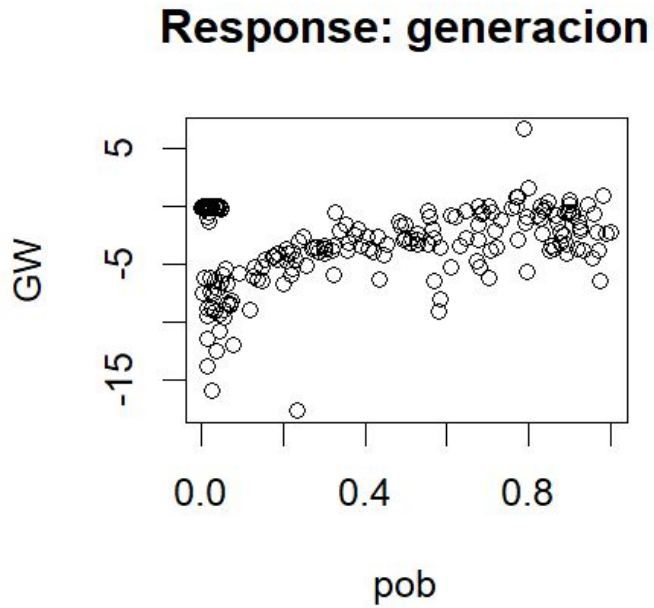
REDES NEURONALES

```
  # Train the neural network model
  nn <- neuralnet(
  generacion ~ pob + r_alta + r_media + pob_25 + pob_2565 + sup_tot + sr_sup +
i_sup + cluster_1 + cluster_2,
  data = dsRE_A_Train_absolutos,
  hidden = j,
  rep=1,
  linear.output=TRUE,
  err.fct='sse',
  threshold=0.01,
  stepmax=1e6,
  lifesign='minimal')
```

FASE 4: GENERACION DE MODELOS (NN)

REDES NEURONALES





Redes de una única capa intermedia de 3,5,7 y 9 neuronas

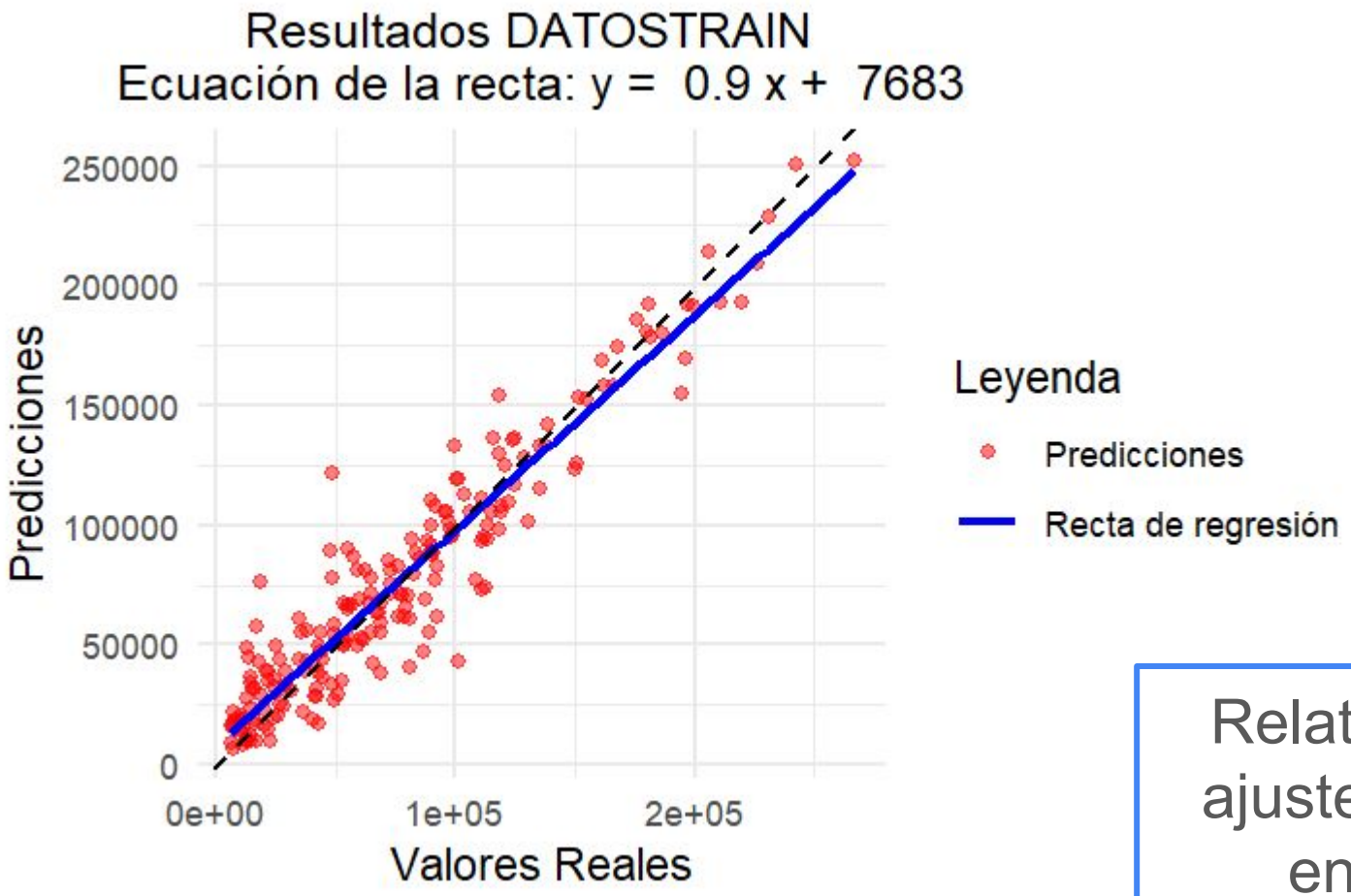
Comprobación varianza pesos

FASE 4: GENERACION DE MODELOS (NN)

ATRACCION							
5 de cada	ABSOLUTOS					PORCENTAJES	
	CR	range01	range11	C. range 01	C. range11	range01	C. range 01
Num. Neuronas optimo	9	9	9	9	9	9	9
RMSLE	2,403	0,868	**0,600**	2,185	1,482	0,992	0,647

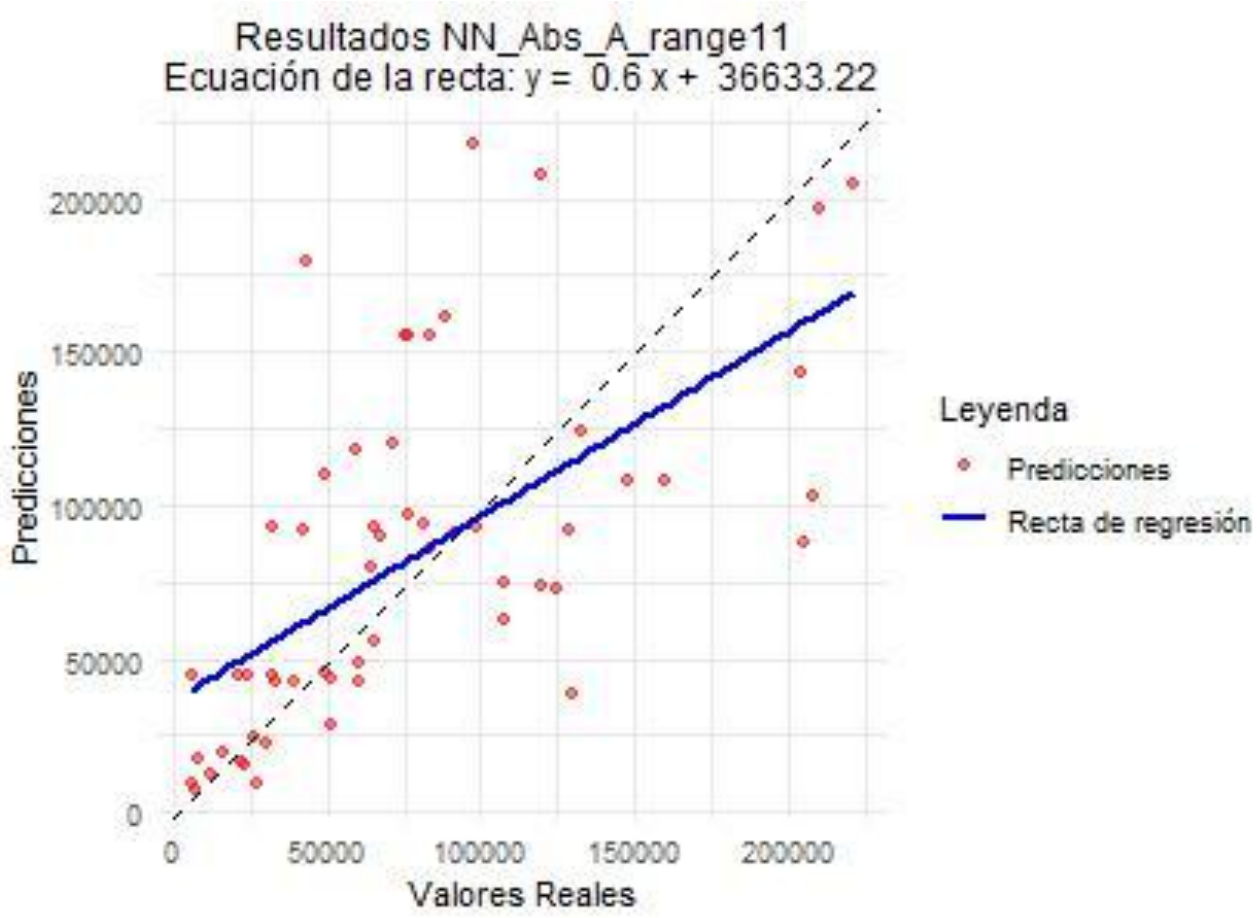
GENERACION							
5 de cada	ABSOLUTOS					PORCENTAJES	
	CR	range01	range11	C. range 01	C. range11	range01	C. range 01
Num. Neuronas optimo	9	9	9	9	9	9	9
RMSLE	2,027	0,606	0,761	0,588	0,626	1,234	**0,584**

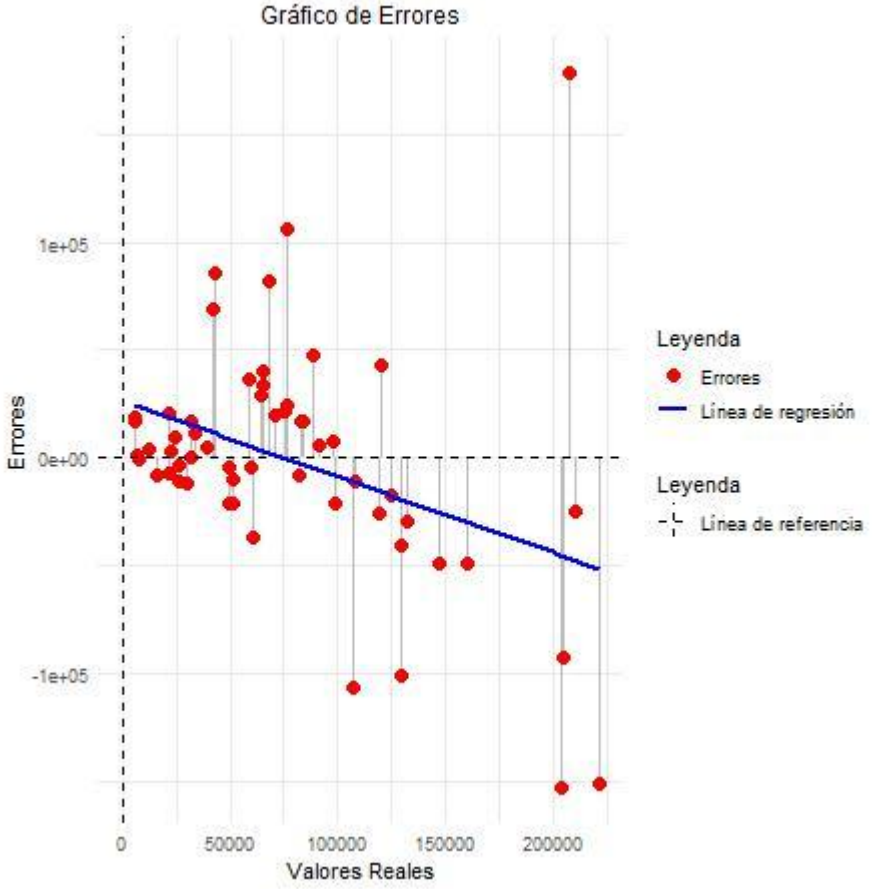
FASE 4: GENERACION DE MODELOS (NN)



Relativamente buen ajuste a los datos de entrenamiento

FASE 4: GENERACION DE MODELOS (NN)





FASE 4: GENERACION DE MODELOS

SVR

```
model <- svm(
  as.formula(paste(target, "~ .")),
  data = df_train,
  type = "nu-regression",
  kernel = "radial",
  gamma = tuneGrid$gamma[i],
  nu = tuneGrid$nu[i],
  cost = tuneGrid$cost[i],
  epsilon = tuneGrid$epsilon[i],
  tolerance = 0.001,
  shrinking = TRUE,
  cross = 3
)
```

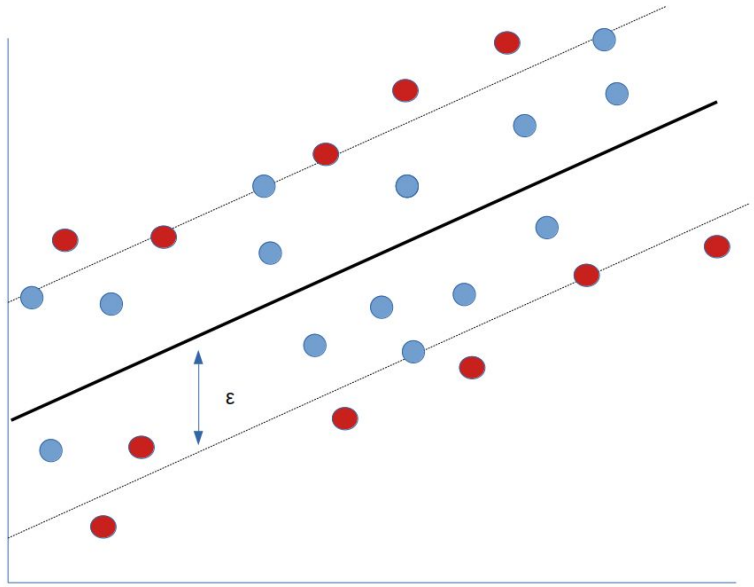
```
tuneGrid <- expand.grid(
  nu = c(0.05, 0.1, 0.15, 0.2),
  gamma = c(0.5, 1, 1.5, 2),
  cost = c(1, 10, 100),
  epsilon = c(0.1, 0.01)
)
```

- Entrenamiento
- Prediccion
- Evaluacion

- Optimizacion

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

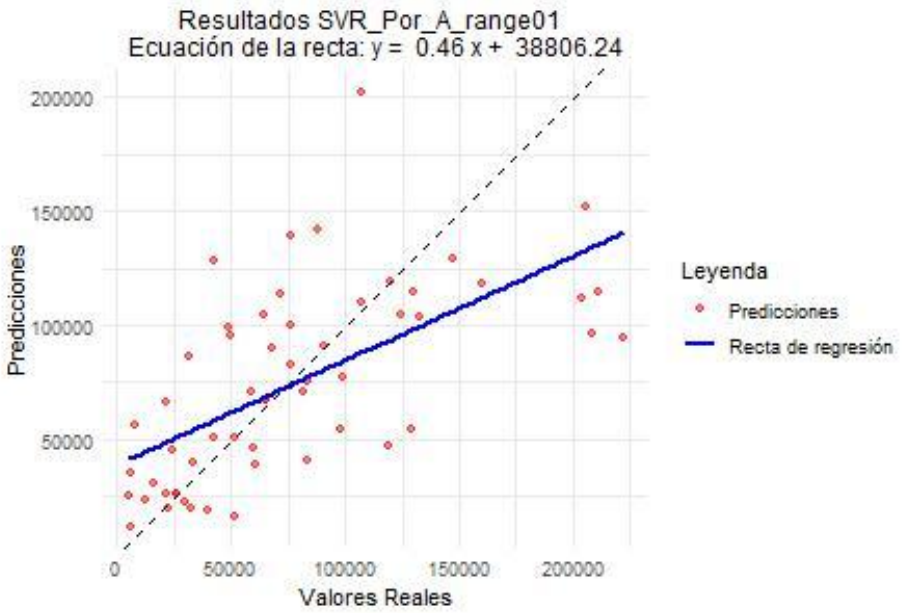
$$\frac{1}{2}\|w\|^2 + C\left(\nu\xi + \frac{1}{l}\sum_{i=1}^{l}(\xi_i + \xi_i^*)\right)$$

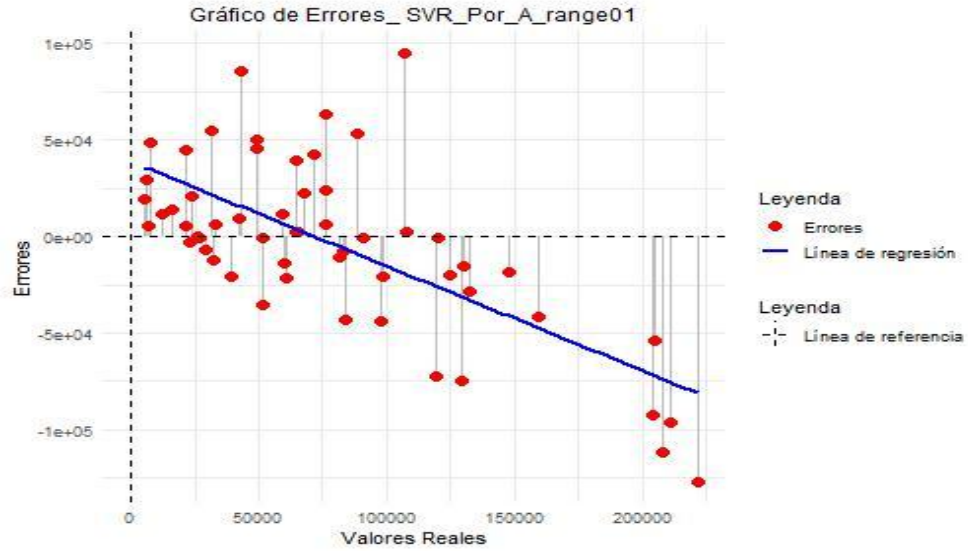


FASE 4: GENERACION DE MODELOS

SVR

Model	RMSLE	Slope	Intercept	R2
SVR_Por_A_range01	0,659	0,459	38806,24	0,388
SVR_Por_G_range01	0,772	0,207	56997,66	0,326
SVR_Abs_A_cr	2,378	0,535	29078,62	0,489
SVR_Abs_A_range01	2,378	0,535	29088,79	0,489
SVRClust_Abs_A_range01	2,379	0,524	29580,77	0,480
SVRClust_Por_A_range01	2,379	0,524	29580,77	0,480
SVR_Abs_G_cr	2,650	0,541	28833,16	0,485
SVR_Abs_G_range01	2,650	0,541	28831,48	0,485
SVRClust_Abs_G_range01	2,650	0,530	29358,83	0,476
SVRClust_Por_G_range01	2,650	0,530	29358,83	0,476



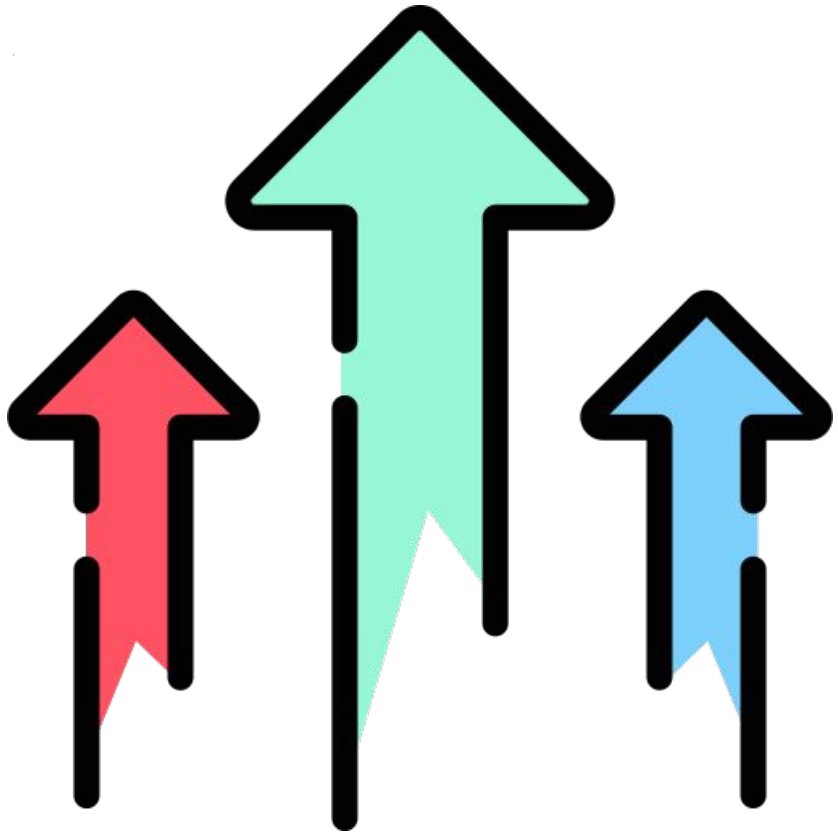


COMPARACION DE MODELOS

Model	RMSLE	Slope	Intercept	R2
NNClus_Por_G_range01	0,584	0,417	41.510	0,303
NN_Abs_A_range11	0,600	*0,600*	36.633	0,378
NN_Abs_G_range01	0,604	*0,449*	39.823	0,400
NNClus_Abs_G_range11	0,626	*0,433*	45.092	0,248
NNClust_Por_A_range01	0,647	*0,359*	43.727	0,214
SVR_Por_A_range01	0,659	0,459	38.806	0,388
NN_Abs_G_range11	0,761	*0,514*	35.325	0,340
SVR_Por_G_range01	0,772	0,207	56.998	0,326
NN_Abs_A_range01	0,868	*0,649*	26.322	0,347
NN_Por_A_range01	0,992	0,388	43.549	0,212
NN_Por_G_range01	1,234	0,450	38.653	0,369
NNClust_Abs_A_range11	1,482	0,605	29.511	0,281
NN_Abs_G_cr	2,027	0,485	30.643	0,298
NNClust_Abs_A_range01	2,185	0,392	33.533	0,223
SVR_Abs_A_cr	2,378	0,535	29.079	0,489
SVR_Abs_A_range01	2,378	0,535	29.089	0,489
SVRClust_Abs_A_range01	2,379	0,524	29.581	0,480
SVRClust_Por_A_range01	2,379	0,524	29.581	0,480
NN_Abs_A_cr	2,403	0,564	24.567	0,442
SVR_Abs_G_cr	2,650	0,541	28.833	0,485
SVR_Abs_G_range01	2,650	0,541	28.831	0,485
SVRClust_Abs_G_range01	2,650	0,530	29.359	0,476
SVRClust_Por_G_range01	2,650	0,530	29.359	0,476

CAMINOS DE MEJORA

- Tratar de entender las caracterısticas de los valores que presentan mayor error. Deep diving.
- Buscar datos reales ayudarıa a tener un modelo que pudiera predecir con un poco mas de exactitud. Ademas de una mayor cantidad de datos y el tipo de datos.
- Probar mas arquitecturas de NN o un grid mas grande en los hiperparametros de SVR tambien serıa de utilidad. Hay que tener en cuenta la carga computacional que esto puede conllevar.
- Utilizar metodos de ensamblaje de modelos o transfer learning para mejorar las predicciones.
- Estratificacion en la validacion cruzada.



APLICACIONES

- INFRAESTRUCTURA: Planificar rutas de transporte publico mas eficientes y ajustar la oferta de servicios de transporte en funcion de la demanda esperada.
- USO DEL SUELO identificar areas con alto potencial de desarrollo o regeneracion urbana.
- DISENO URBANO: Simular posibles impactos en la movilidad y el uso del suelo, mitigando riesgos potenciales.
- POLITICAS URBANAS: Identificar areas urbanas con caracterısticas similares, facilitando la implementacion de polıticas urbanas especıficas y adaptadas a las necesidades particulares de cada cluster
- SERVICIOS PUBLICOS: Asistir en la planificacion y distribucion de servicios publicos.