




Universidad Tecnológica Metropolitana
Facultad de Ingeniería
Ingeniería Civil en Computación mención Informática

Modelo predictivo de retiro escolar

Proyecto Semestral - Minería de Datos

Javier Galvez
Sebastian Garrido
Benjamin Martinez



- 
1. Introducción
 2. Descripción del Proyecto
 3. Contexto de la Problemática
 4. Objetivos e Hipótesis
 5. Pasos de construcción del modelo
 6. Descripción de Datos
 7. Aplicación estadística descriptiva
 8. Preparación de datos
 9. Detección de Outliers
 10. Tratamiento Outliers
 11. Clustering y agrupamiento de datos
 12. Análisis PCA y Factorial
 13. Modelos Predictivos
 14. Evaluación Modelo
 15. Conclusión

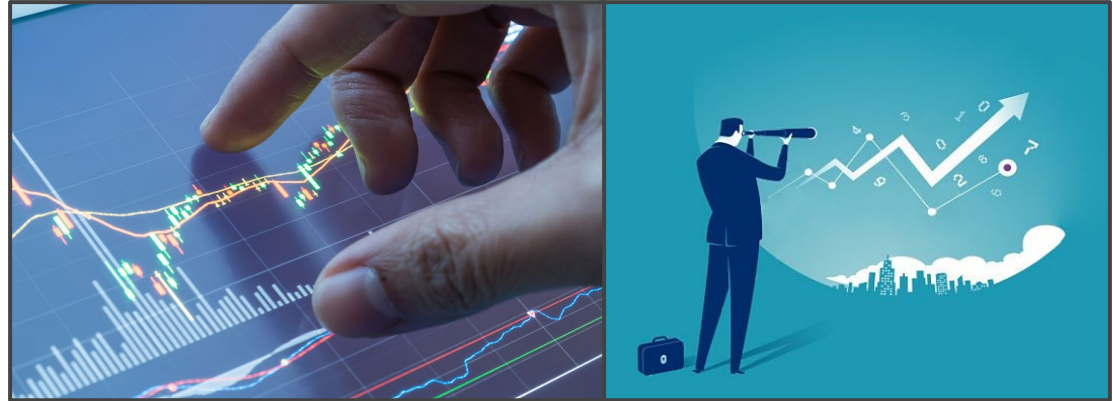
Contenidos



Introducción

Un modelo predictivo es un sistema que emplea datos y estadísticas para predecir resultados a partir de unos modelos de datos. El modelado predictivo es útil porque proporciona información precisa sobre cualquier pregunta y permite a los usuarios crear previsiones de esta, así ir manteniendo una ventaja competitiva tras la obtención de información detallada de eventos y resultados futuros. En el análisis de minería de datos se suele extraer información esencial de diversas fuentes para complementar los modelos predictivos, estos son:

- 1) Datos sobre transacciones.
- 2) Datos de servicio al cliente.
- 3) Datos de encuestas o sondeos.
- 4) Datos de marketing digital y publicidad.
- 5) Datos económicos.
- 6) Datos demográficos.
- 7) Datos geográficos.
- 8) Datos de tráfico web.



Descripción del Proyecto

El proyecto busca predecir la cantidad de retirados de los diferentes establecimientos educacionales de las comunas de Santiago, Región Metropolitana, adquiriendo datos de la cantidad de alumnos retirados, con base a los datos recopilados del año 2018.

Bien se sabe que el rendimiento escolar en los últimos tiempos no ha sido del todo confiable, por diversos motivos como:

- Estallido Social
- Pandemia Covid-19

Y tras esto y varios factores se decidió usar datos de años anteriores a estos ya que trae una mejor apreciación de un año “normal” en el ámbito escolar





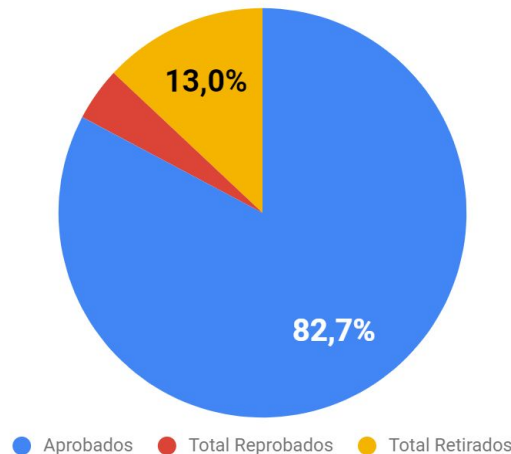
Contexto

En el contexto que existe actualmente, es que en los colegios y liceos existe una cierta cantidad de alumnos que se retira del establecimiento.

A partir de las estadísticas del año 2018 se interpretan porcentajes de acuerdo la situación final de los alumnos, donde:

- 82,7% de alumnos aprobaron.
- 4,23% de alumnos reprobaron.
- 13,03% de alumnos que se retiraron de sus establecimientos.

PORCENTAJES DE ALUMNOS Y ALUMNAS DE APROBADOS, REPROBADOS Y RETIRADOS



Año escolar	Número de Unidades Educativas	Número de Unidades Educativas en establecimientos en funcionamiento	Número de establecimientos	Número de establecimientos en funcionamiento
2018	13.940	13925	9.192	9.179
2017	13.942	13936	9.273	9.268
2016	14.009	13.995	9.349	9.337
2015	14.069	14.041	9.458	9.437



Objetivos e Hipótesis



General: Formular un modelo predictivo capaz de identificar la cantidad de alumnos retirados de las instituciones educacionales de la comuna de santiago, para proporcionar ayuda al desempeño de dichas instituciones.

Específicos:

- Recopilar Información respecto a los alumnos retirados de colegios de santiago en el año 2018.
- Preparación de los datos para generar un modelo
- Generar un modelo predictivo con herramientas de minería de datos

Hipótesis:

¿Es posible predecir el posible retiro de los alumnos para un determinado conjunto de instituciones de una comuna?



Pasos de construcción del modelo



- Aplicación de estadística descriptiva
- Preparación de Datos
- Aplicación de Clustering
- Aplicación de PCA (Reducción de Dimensiones)
- Regresión logística multivariable
- Árboles de Decisión
- Redes Bayesianas y SVM
- Redes Neuronales
- Evaluación del modelo final con base a los pasos anteriores

matplotlib



NumPy





Descripción de datos



- Comuna
- Cantidad total de retirados
- Superficie (Km2)
- Población 2020
- Tasa de pobreza por Ingresos (%)
- Personas en hogares carentes de servicios básicos (%)
- Hogares hacinados (%)
- Cantidad de establecimientos municipales
- Cantidad de establecimientos particulares subvencionados
- Cantidad de establecimientos particulares pagados
- Cantidad total de matrículas por comuna
- Ingresos Educación (M\$)
- Aporte Municipal al Sector Educación (M\$)
- Gastos Educación (M\$)
- Tasa de denuncias por delito de mayor connotación social
- Tasa de denuncias por violencia intrafamiliar



Biblioteca del Congreso
Nacional de Chile / BCN



Aplicación de estadística descriptiva

	CANTIDAD TOTAL DE RETIRADOS	Superficie km2	poblacion 2020	tasa de pobreza por ingresos %	Personas en hogares carentes de servicios básico %	Hogares hacinados %	cantidad de establecimientos municipales	cantidad de establecimientos particular subvencionado	cantidad de establecimientos particular pagado	cantidad de matriculas totales	Ingresos Educación (M\$)	Aporte Municipal al Sector Educación (M\$)
count	26.000000	26.000000	26.000000	26.000000	26.000000	26.000000	26.000000	26.000000	26.000000	26.000000	2.600000e+01	2.600000e+01
mean	378.269231	66.869231	195516.076923	4.930769	7.061154	15.596154	8.384615	43.076923	9.423077	29236.615385	1.881607e+07	2.611578e+06
std	416.096677	197.689346	129042.959903	3.168674	4.589723	5.447236	10.135391	37.333013	11.876609	19675.390914	1.336517e+07	2.614291e+06
min	4.000000	7.000000	86510.000000	0.130000	0.190000	2.900000	0.000000	1.000000	0.000000	7022.000000	6.930984e+06	1.623700e+05
25%	163.000000	10.000000	104849.500000	3.062500	4.200000	12.700000	0.000000	24.250000	0.250000	17897.500000	1.068189e+07	8.755062e+05
50%	286.500000	15.500000	139216.500000	4.800000	6.100000	17.200000	7.000000	35.500000	3.000000	21702.000000	1.393831e+07	1.733466e+06
75%	424.500000	30.025000	239342.000000	6.570000	9.175000	18.425000	14.500000	48.250000	16.000000	34978.750000	2.231339e+07	3.800000e+06
max	2123.000000	1024.000000	578605.000000	14.140000	20.100000	24.800000	44.000000	163.000000	44.000000	81529.000000	6.808654e+07	1.074911e+07

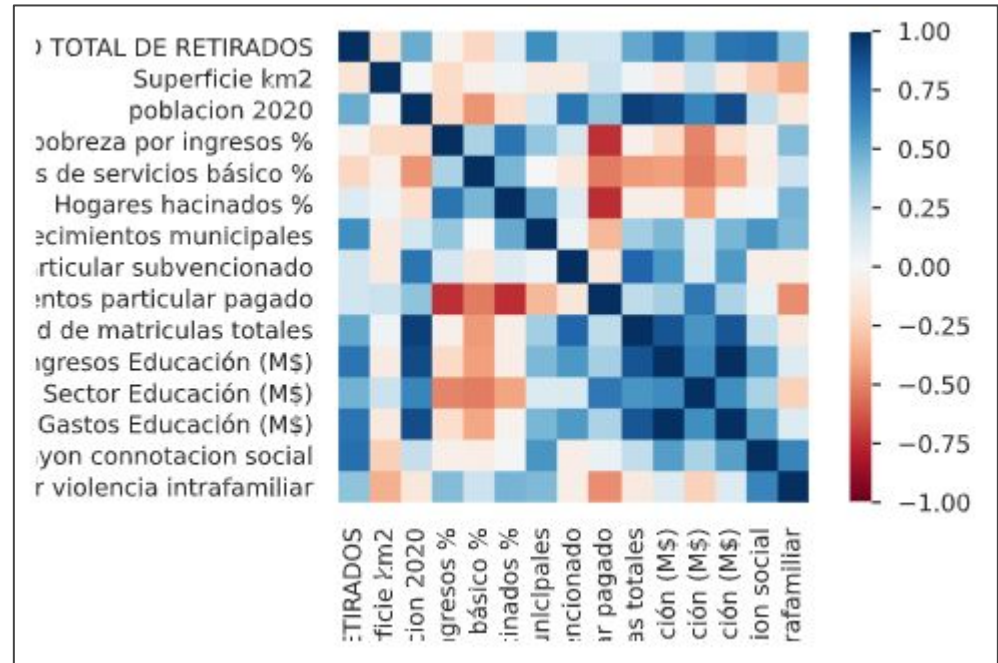


Aplicación de estadística descriptiva cont.

	Gastos Educación (M\$)	tasa de denuncias por delito de mayon connotacion social	tasa de denuncias por violencia intrafamiliar
count	2.600000e+01	26.000000	26.000000
mean	1.895675e+07	4989.334615	644.296154
std	1.337046e+07	4054.570376	330.333837
min	6.658703e+06	868.700000	193.400000
25%	1.072232e+07	2701.500000	421.850000
50%	1.505666e+07	3795.750000	629.750000
75%	2.286784e+07	6256.950000	799.000000
max	6.846788e+07	21169.700000	1520.300000

Por lo tanto, al comparar con la variable objetivo, que es la cantidad de personas que se retiraron, la fuerza de asociación positiva es fuerte para alguna correlación de Pearson como es el caso de

- La tasa de denuncias por delito de mayor connotación social
- Cantidad de establecimientos municipales, Gastos Educación (M\$),
- Ingresos de Educación (M\$).





Preparación de datos

```
b.drop('COMUNA',axis=1, inplace=True)
```

```
variables_independientes = b.iloc[:,1:17]
```

```
variable_objetivo = b.iloc[:,0:1]
```

```
X_std = preprocessing.normalize(variables_independientes)
```

```
X_std0 = preprocessing.normalize(variable_objetivo)
```

```
X_std = preprocessing.minmax_scale(variables_independientes)
```

```
X_std0 = preprocessing.minmax_scale(variable_objetivo)
```

```
X_std = preprocessing.scale(variables_independientes)
```

```
X_std0 = preprocessing.scale(variable_objetivo)
```

Para normalizar los datos, primero se ha extraído las columnas del archivo CSV para generar un Dataframe utilizando la librería Pandas, luego se eliminó la columna “Comuna” para no alterar los resultados por su tipo de datos String. Luego se realizó una división del dataset, quedando en 2 variables, “variable_objetivo” y “variables_independientes”, donde la primera tiene almacenado la “cantidad de retirados totales” y la segunda tiene almacenado todas las columnas restantes. En la normalización de datos, se utilizó la librería “sklearn”, aplicando las funciones como Scale, “preprocessing.scale()”, Normalize, “preprocessing.normalize()”, y MinMaxScale, “preprocessing.minmax_scale()”.

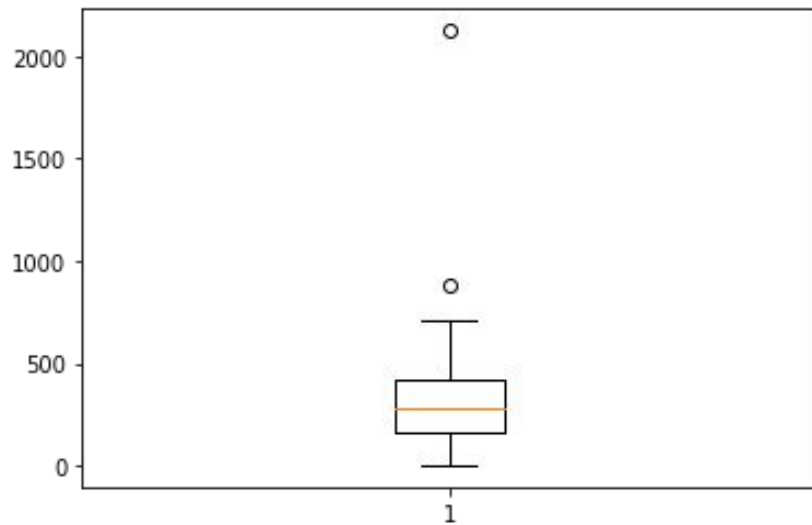


Detección de Outliers

N°	VARIABLES	OUTLIERS	OUTLIERS SEVERO
1	Cantidad total de retirados	-	2123
2	Superficie km2	99 y 135	1024
3	Población 2020	503147 y 578605	-
4	Tasa de pobreza por ingresos %	14,14	-
5	Personas en hogares carentes de servicios básico %	20,1	-
6	Hogares hacinados %	-	-
7	Cantidad de establecimientos municipales	44	-
8	Cantidad de establecimientos particular subvencionado	-	147 y 163
9	Cantidad de establecimientos particular pagado	44	-

N°	VARIABLES	OUTLIERS	OUTLIERS SEVERO
10	Cantidad total de matrículas en los establecimientos	81.529	-
11	Ingresos Educación (M\$)	44285876	68086543
12	Aporte Municipal al Sector Educación (M\$)	8810430 y 10749110	-
13	Gastos Educación (M\$)	44979840	68467876
14	Tasa de denuncias por delito de mayor connotación social	-	21169,7
15	Tasa de denuncias por violencia intrafamiliar	1520,3	-

Boxplot, Cantidad de retirados



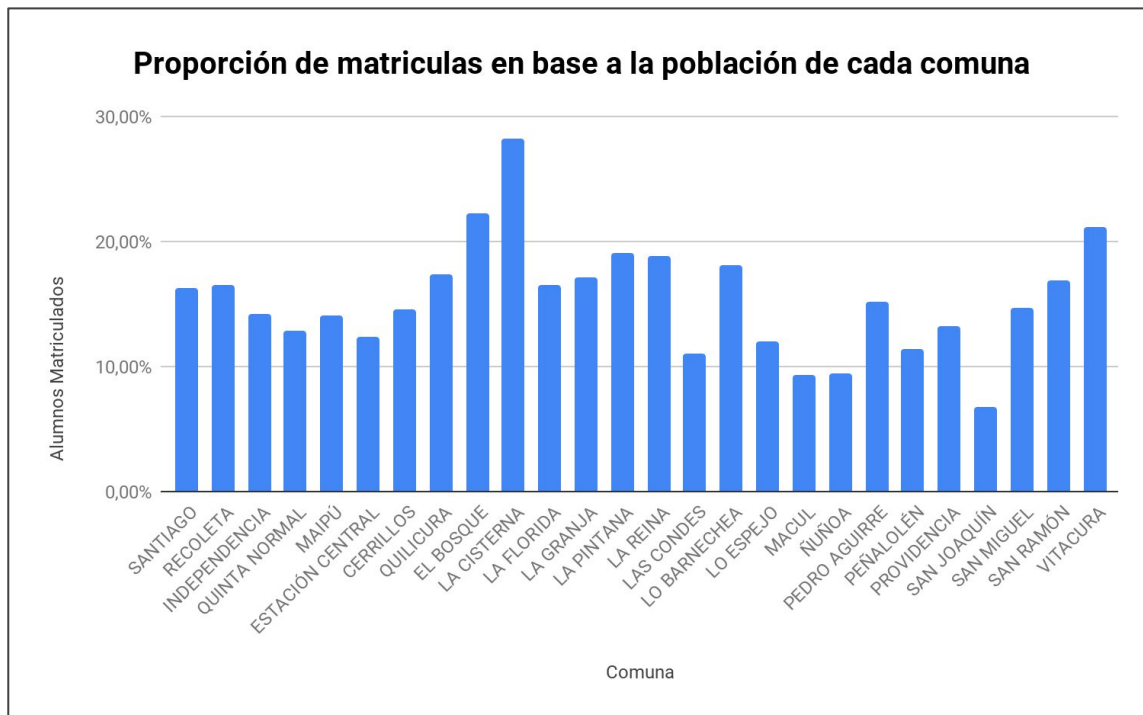
COMUNA	CANTIDAD TOTAL DE RETIRADOS
SANTIAGO	2123
RECOLETA	713
INDEPENDENCIA	402
QUINTA NORMAL	58
MAIPÚ	11
ESTACIÓN CENTRAL	432
CERRILLOS	31
QUILICURA	4
EL BOSQUE	292
LA CISTERNA	119
LA FLORIDA	885
LA GRANJA	281
LA PINTANA	265

COMUNA	CANTIDAD TOTAL DE RETIRADOS
LA REINA	214
LAS CONDES	322
LO BARNECHEA	157
LO ESPEJO	181
MACUL	300
ÑUÑO A	584
PEDRO AGUIRRE CERDA	270
PEÑALOLÉN	531
PROVIDENCIA	296
SAN JOAQUÍN	267
SAN MIGUEL	394
SAN RAMÓN	569
VITACURA	134



Proporción sobre cantidad de matrículas totales en base a población registrada en 2020

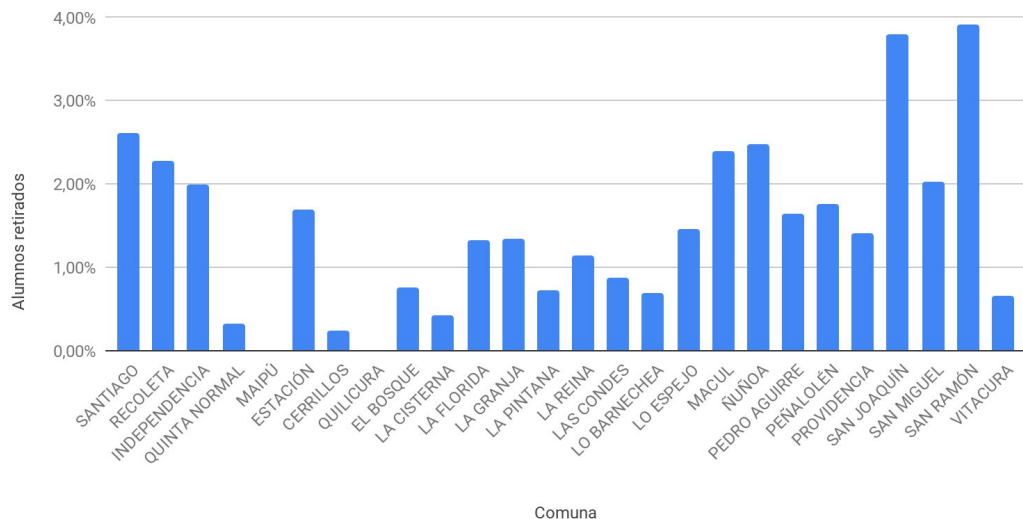
- No todas las comunas tienen las mismas cantidades de habitantes y también los números de alumnos que se han matriculado en sus respectivos establecimientos.
- La comuna con mayor porcentaje de alumnos matriculados es “La Cisterna”,
- La comuna con menor porcentaje de alumnos matriculados es “San Joaquín”.





Proporción sobre cantidad de alumnos retirados en base a la cantidad total de matrículas en los establecimientos

Proporción de cantidad de alumnos retirados en base al número total de matrículas en los establecimientos

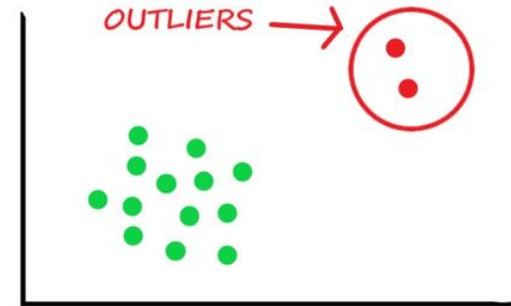


- En los porcentajes, indican una contradicción a la detección de outliers con el método basado en el recorrido intercuartílico y boxplot
- Santiago no presenta como outliers en comparación con el resto a partir de esta proporción.
- La comuna “San Ramón” es el que ha tenido mayor porcentaje sobre el retiro de los alumnos.
- Todos los porcentajes representa una cantidad de retiros muy pequeños.

Tratamiento de Outliers

Los datos atípicos (outliers) pueden ser un problema para precisar mejor los resultados de predicción, y como se busca predecir la cantidad de alumnos retirados a futuro de comunas de la Región Metropolitana, Es importante tener definido una estrategia para reducir su impacto. Cuando los datos no cumplen con estos supuestos disminuye la capacidad de detectar efectos reales, por lo que cualquier interpretación de los datos pueden ser erróneas.

Para el tratamiento de estos outliers los cuales perjudican los resultados de predicción respecto a la variable objetivo (cantidad de alumnos retirados), se procedió con la determinación de promedios para cada columna que contengan presencias de estos outliers que son severos, para después reemplazar su valor y así poder equilibrar el conjunto de datos. Las únicas excepciones son la cantidad de alumnos retirados y la cantidad total de matrículas.



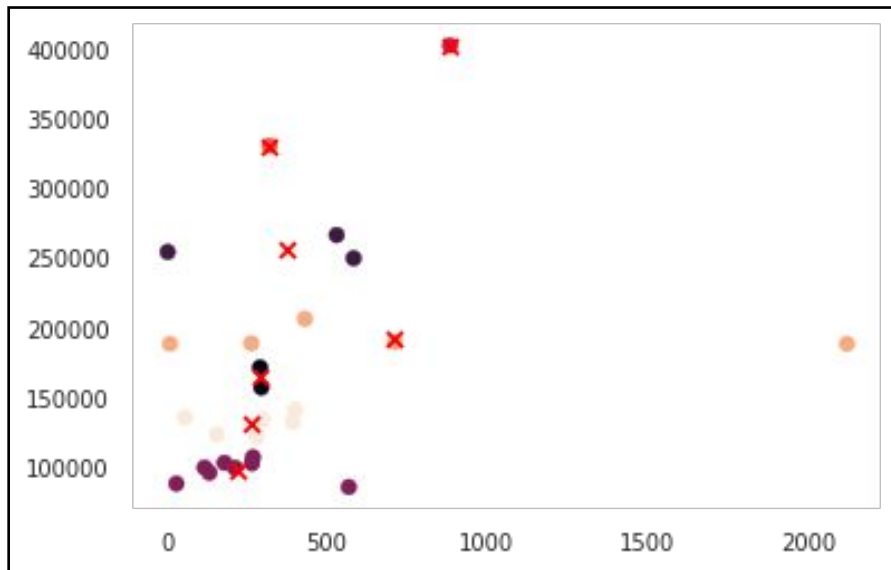
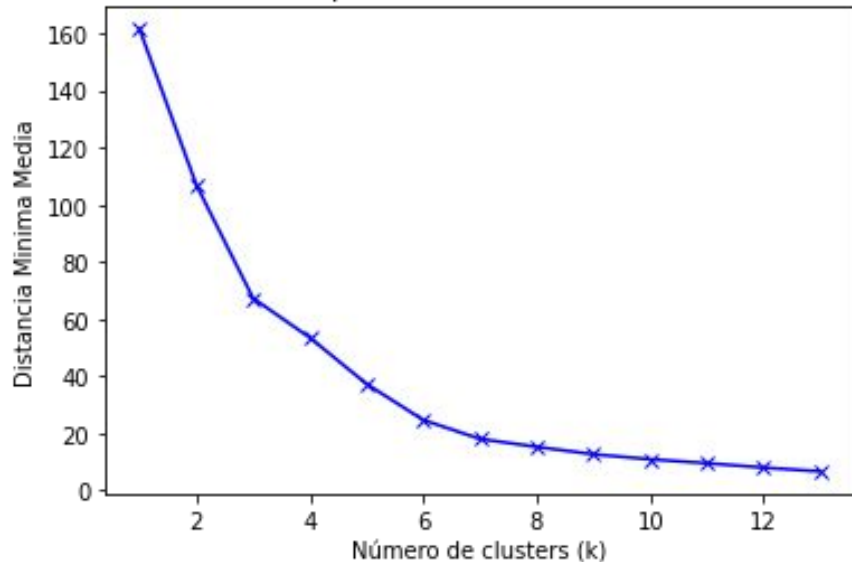
	CANTIDAD TOTAL DE RETIRADOS	Superficie km2	poblacion 2020	tasa de pobreza por ingresos %	Personas en hogares carentes de servicios básico %	Hogares hacinados %	cantidad de establecimientos municipales	cantidad de establecimientos particular subvencionado	cantidad de establecimientos particular pagado	cantidad de matriculas totales	Ingresos Educación (M\$)	Aporte Municipal al Sector Educación (M\$)
count	26.000000	26.000000	26.000000	26.000000	26.000000	26.000000	26.000000	26.000000	26.000000	26.000000	2.600000e+01	2.600000e+01
mean	378.269231	26.184615	168441.538462	4.576538	6.561154	15.596154	7.015385	34.461538	8.076923	29236.615385	1.594145e+07	2.060179e+06
std	416.096677	22.044386	78977.665966	2.552959	3.742346	5.447236	7.073765	17.586883	9.557921	19675.390914	6.854353e+06	1.527255e+06
min	4.000000	7.000000	86510.000000	0.130000	0.190000	2.900000	0.000000	1.000000	0.000000	7022.000000	6.930984e+06	1.623700e+05
25%	163.000000	10.000000	104849.500000	3.062500	4.200000	12.700000	0.000000	24.250000	0.250000	17897.500000	1.068189e+07	8.755062e+05
50%	286.500000	15.500000	139216.500000	4.800000	6.100000	17.200000	7.000000	35.500000	3.000000	21702.000000	1.393831e+07	1.733466e+06
75%	424.500000	30.025000	189890.000000	6.412500	8.875000	18.425000	12.750000	43.000000	15.750000	34978.750000	2.116664e+07	2.876650e+06
max	2123.000000	70.200000	402433.000000	9.580000	14.700000	24.800000	21.000000	71.000000	31.000000	81529.000000	3.288611e+07	5.717647e+06

	Gastos Educación (M\$)	tasa de denuncias por delito de mayon connotacion social	tasa de denuncias por violencia intrafamiliar
count	2.600000e+01	26.000000	26.000000
mean	1.605159e+07	4367.011538	610.603846
std	6.696983e+06	2358.939278	277.929048
min	6.658703e+06	868.700000	193.400000
25%	1.072232e+07	2701.500000	421.850000
50%	1.505666e+07	3795.750000	629.750000
75%	2.059408e+07	6114.850000	726.600000
max	3.156323e+07	8697.200000	1283.000000

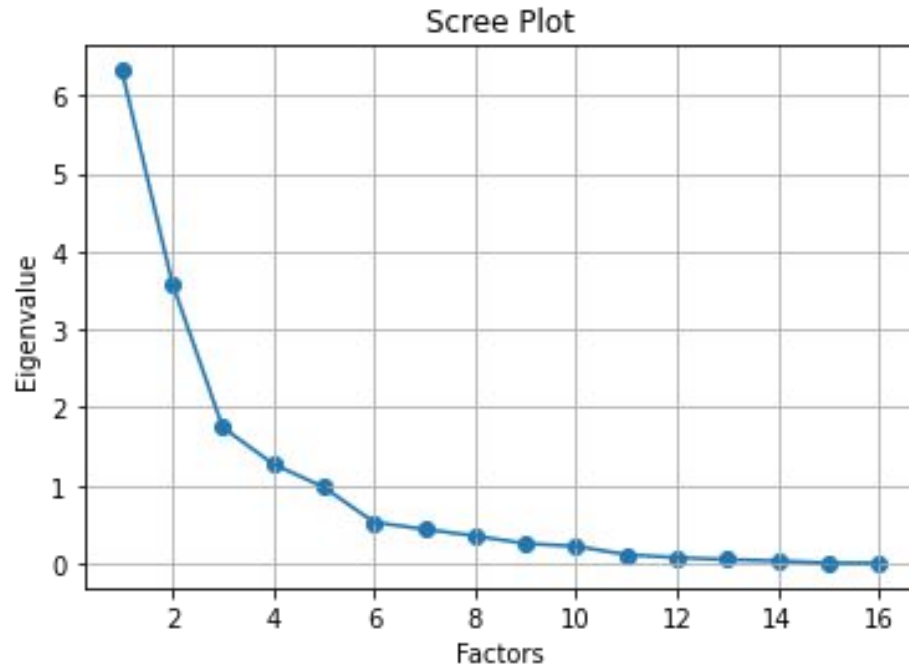
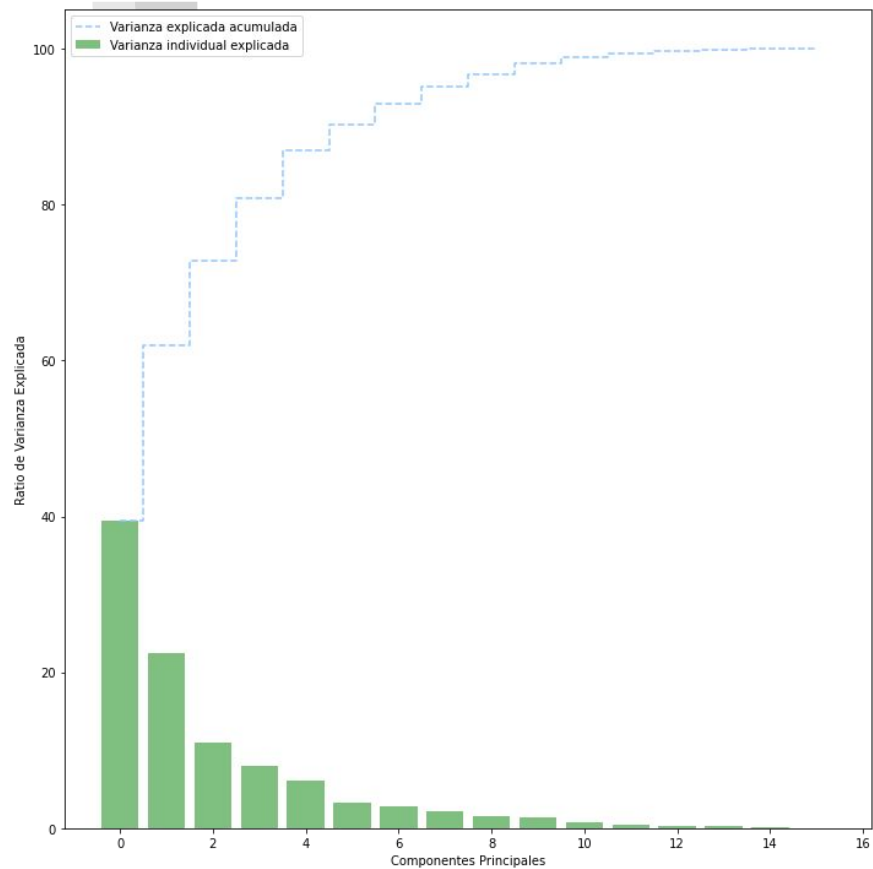


Clustering y agrupamiento de datos

Método del Codo para Determinar el Número de Clusters



Análisis PCA y Factorial

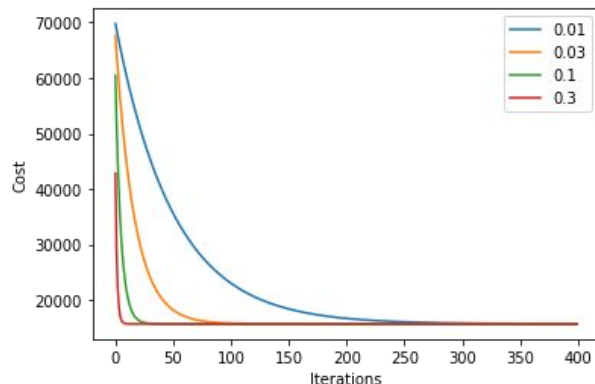


Evaluación de los Modelos Predictivos

Predicción de retiro escolar

Regresión Lineal Multivariable

Se determinó la precisión para cada comuna por separado y no se generalizó, para poder obtener una vista distinta. La rama de normalización con mejor resultado fue Scale, con dataset ya tratados con outliers y sin PCA, con un error cuadrático medio de 22.230,09



```
0 cantidad retirados 1928.0 / % 90.81488459726802
1 cantidad retirados 537.0 / % 75.31556802244039
2 cantidad retirados 384.0 / % 95.5223880597015
3 cantidad retirados 33.0 / % 56.89655172413793
4 cantidad retirados 232.0 / % 2109.090909090909
5 cantidad retirados 578.0 / % 133.7962962962963
6 cantidad retirados 319.0 / % 1029.032258064516
7 cantidad retirados 70.0 / % 1750.0
8 cantidad retirados 147.0 / % 50.342465753424655
9 cantidad retirados 150.0 / % 126.05042016806723
10 cantidad retirados 904.0 / % 102.14689265536722
11 cantidad retirados 320.0 / % 113.87900355871886
12 cantidad retirados 470.0 / % 177.35849056603774
13 cantidad retirados 327.0 / % 152.80373831775702
14 cantidad retirados 204.0 / % 63.35403726708075
15 cantidad retirados 15.0 / % 9.554140127388536
16 cantidad retirados 391.0 / % 216.02209944751382
17 cantidad retirados 300.0 / % 100.0
18 cantidad retirados 836.0 / % 143.15068493150685
19 cantidad retirados 278.0 / % 102.96296296296296
20 cantidad retirados 364.0 / % 68.54990583804143
21 cantidad retirados 222.0 / % 75.0
22 cantidad retirados 130.0 / % 48.68913857677903
23 cantidad retirados 355.0 / % 90.1015228426396
24 cantidad retirados 312.0 / % 54.83304042179262
25 cantidad retirados 30.0 / % 22.388059701492537
```

```
rmse = mean_squared_error(y,prediccioness)
print("El error (rmse) de test es: ", rmse)
```

El error (rmse) de test es: 22230.094606176965

Árboles de Decisión

La rama de normalización con mejor resultado fue Scale, con dataset ya tratados con outliers, una precisión del entrenamiento igual a 95,21%, otra precisión respecto la prueba es 19,58%, y el error cuadrático que se obtiene es de 0,36 lo cual implica que es un mejor algoritmo y modelo para predecir la cantidad de alumnos retirados.



Árbol de decisión con profundidad 3

```
--- Personas en hogares carentes de servicios básico % <= -1.69
|--- value: [4.28]
--- Personas en hogares carentes de servicios básico % > -1.69
|--- Gastos Educación (M$) <= 1.14
|   |--- poblacion 2020 <= -1.04
|   |   |--- value: [0.47]
|   |   |--- poblacion 2020 > -1.04
|   |   |--- value: [-0.44]
|   |--- Gastos Educación (M$) > 1.14
|   |--- Aporte Municipal al Sector Educación (M$) <= 0.12
|   |   |--- value: [0.37]
|   |   |--- Aporte Municipal al Sector Educación (M$) > 0.12
|   |   |--- value: [1.03]
```

```
rmse = mean_squared_error(
    y_true = y_test,
    y_pred = predicciones,
    squared = False
)
print(f"El error (rmse) de test es: {rmse}")
```

El error (rmse) de test es: 0.3646249140957541

```
[715] acc_decision_test = round(modelo.score(X_test, y_test) * 100, 2) #se ob
print("precisión: ", acc_decision_test, "%")
```

precisión: 19.58 %

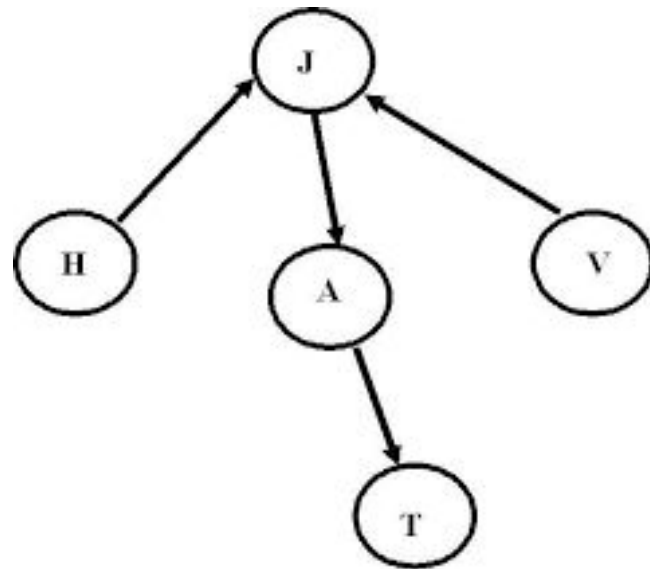
```
[716] acc_decision_train = round(modelo.score(X_train, y_train) * 100, 2) #se
print("precisión: ", acc_decision_train, "%")
```

precisión: 95.21 %



Redes Bayesianas

La rama de normalización con mejor resultado fue Scale, con dataset original y sin PCA, cuyo error cuadrático medio es 64.955,5.



El error (rmse) de test es: 64955.5



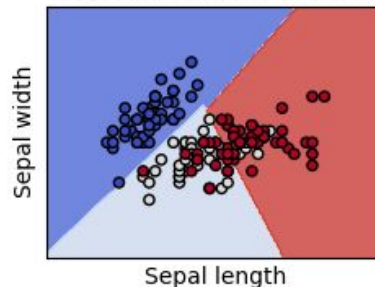
SVM

La rama de normalización con mejor resultado fue Normalize, con dataset original y sin PCA, utilizando el Kernel POLY, cuyo error cuadrático medio es 130,90.

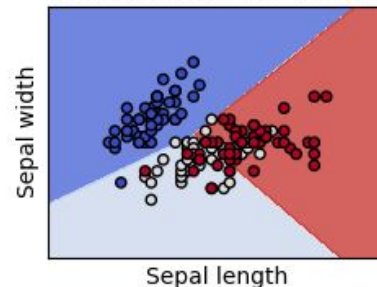
```
print(f"El error (rmse) de test es: {rmse}")
```

El error (rmse) de test es: 130.9005979045888

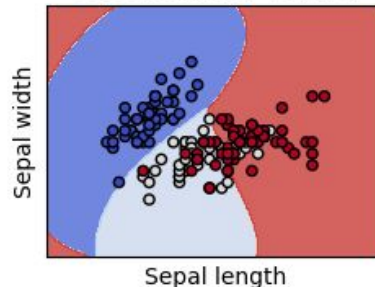
SVC with linear kernel



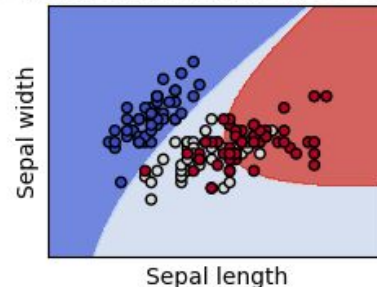
LinearSVC (linear kernel)



SVC with RBF kernel

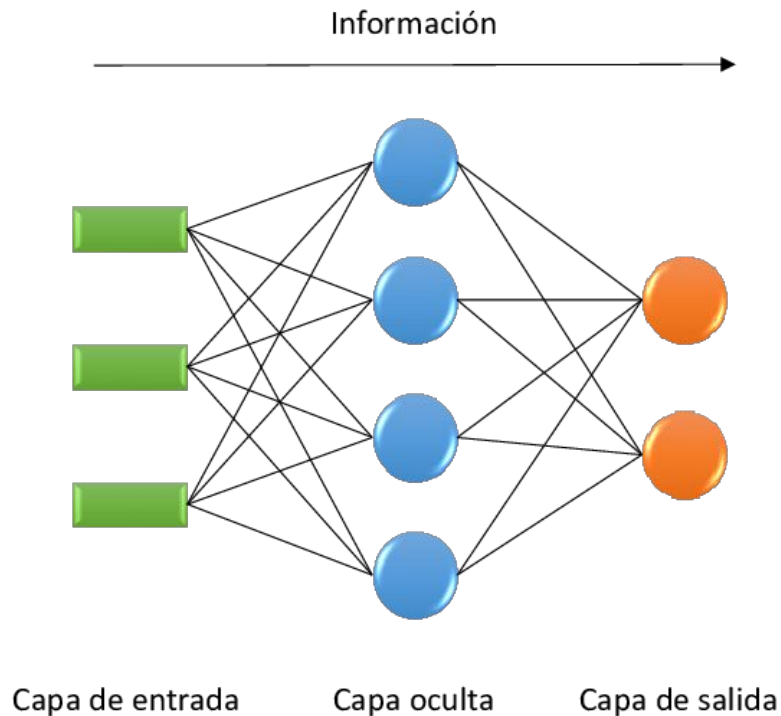


SVC with polynomial (degree 3) kernel





Redes neuronales



La rama de normalización con mejor resultado fue Normalize, con dataset original y sin PCA, cuyo error cuadrático medio es 383,52.

```
print(f"El error (rmse) de test es: {rmse}")
```

El error (rmse) de test es: 383.52147159620637



Observación

La precisión de datos arrojada estos modelos, surgió con base a las diferentes ramificaciones analizadas, como la detección en su totalidad de los Outliers para cada columna del dataset, esto para tomar una decisión con respecto y no estirar la media y varianza para los datos, posteriormente tras otro análisis estadístico descriptivo de los datos se obtuvieron agrupamientos y finalmente se comenzó a trabajar con cada tipo de modelo, para la evaluación de este.

Cabe destacar que, al aplicar el algoritmo sobre árboles de decisión completamente, con la normalización “*Normalize*” no se logra desplegar algún resultado, por lo que estos datos normalizados no funciona para evaluar los datos y así predecir resultados.

Predecir la cantidad de alumnos retirados, en su cercanía, con respecto al año estudiado, 2018, para proporcionar una visión de desempeño de establecimientos de comunas de la región metropolitana. A su vez, se recogieron datos confiables para determinar este número de alumnos, además se generaron varios modelos predictivos, lo cual el más acertado fue árboles de decisión.

Características como:

- Economía
- Violencia
- Hogares hacinados
- Estatus social
- Gastos educacionales

Conclusión