APLICACIONES DE BASE DE DATOS II

TPNº8: Matrices, Regresión y SVM

Apellido y Nombre	Email	LU	DNI	ROL
Chauque, Pedro Santiago	44812800@fi.unju.edu.ar	INF9121	44812800	Resp. PowerPoint
Cruz, Facundo Emanuel	41609573@fi.unju.edu.ar	INF7766	41609573	Resp. del video
Orquera, Dario Joaquin	33482732@fi.unju.edu.ar	INF5573	33482732	Project Leader
Ortega, Gonzalo Leandro	41679923@fi.unju.edu.ar	INF8190	41679923	Resp. Informe 1
Vargas, Carlos Francisco	45172279@fi.unju.edu.ar	INF9022	45172279	Resp. Informe 2

Fecha de Presentación: 06-10-25

CONSIGNAS

Preguntas Teóricas

1. ¿Qué es una Matriz de Correlación? Donde se considera necesario aplicar una matriz de correlación.

Una matriz de correlación es una tabla de doble entrada para los atributos del DataSet, que muestra una lista multivariable (la lista de atributos) horizontalmente y la misma lista verticalmente y con el correspondiente coeficiente de correlación llamado r o la relación entre cada pareja en cada celda, expresada con un número que va desde 0 a 1.

Se utiliza cuando queremos analizar relaciones entre variables numéricas, Análisis estadístico exploratorio, Modelos de regresión, Análisis multivariante

2. ¿Qué es una Matriz de Confusión? Nombre las Ventajas y Desventajas. ¿Qué diferencia se tiene con una matriz de correlación?

Una matriz de confusión es una herramienta utilizada para evaluar el rendimiento de un modelo de clasificación (por ejemplo, en Machine Learning).

Muestra de manera tabular cómo el modelo clasifica correctamente o incorrectamente las instancias de cada clase.

Por ejemplo, para un problema binario (positivo/negativo):

	Predicción: Positivo	Predicción: Negativo
Real: Positivo	Verdaderos Positivos (TP)	Falsos Negativos (FN)
Real: Negativo	Falsos Positivos (FP)	Verdaderos Negativos (TN)

Ventajas

1. Permite entender errores específicos del modelo (por ejemplo, si confunde más una clase que otra).

2. Proporciona métricas derivadas como precisión (*precision*), sensibilidad (*recall*), exactitud (*accuracy*) y F1-score.
3. Visualiza el desempeño global y por clase, especialmente útil en datasets desbalanceados.
4. Identifica patrones de confusión entre clases similares.

Desventajas

1. No es útil para problemas de regresión, solo clasificación.
2. Se vuelve compleja con muchas clases, ya que la tabla crece y cuesta interpretarla.
3. No muestra el motivo del error, solo cuántas veces ocurre.
4. Depende del umbral de decisión, por lo que pequeños cambios pueden alterar los valores.

Diferencia con una Matriz de Correlación

Aspecto	Matriz de Confusión	Matriz de Correlación
Uso principal	Evaluar el rendimiento de un modelo de clasificación	Analizar la relación estadística entre variables
Tipo de datos	Etiquetas reales vs. predichas (categóricas)	Variables numéricas (continuas)
Medida	Conteo de aciertos y errores	Coeficientes de correlación (valores entre -1 y 1)
Interpretación	Indica cuántas veces el modelo acierta o se equivoca	Indica el grado y dirección de la relación entre variables

Ejemplo de uso	Saber cuántas imágenes de “gato” se confundieron con “perro”	Ver si la variable “edad” se correlaciona con “ingreso”

3. ¿Cuál es la diferencia entre un modelo de clasificación y un modelo de regresión?

La principal diferencia entre estos modelos radica en que un modelo de clasificación predice valores discretos. Se emplea principalmente para asignar una etiqueta adecuada o para categorizar(es decir, clasificar). Puede tratarse de una clasificación binaria, como “si o no”, “aceptar o rechazar”, o una clasificación múltiple, como un motor de recomendación que sugiere un producto específico entre varios disponibles. Mientras tanto, un modelo de regresión predice valores continuos, como la edad, precio, tamaño o tiempo. Se emplea principalmente para determinar la relación entre una o varias variables independientes (x) y una variable dependiente (y), de modo que, dada una x, el modelo pueda predecir el valor correspondiente de y.

4. Explique que es un modelo de regresión. Detalle las variantes que existen.

Es un modelo estadístico usado para predecir un valor continuo basándose en una o varias variables independientes. Este modelo encuentra la relación entre las variables mediante una función matemática que minimiza el error de predicción.

Variantes:

- **Regresión Lineal:** Modelo que asume una relación lineal entre las variables.
- **Regresión Polinómica:** Extiende la regresión lineal al permitir relaciones de mayor orden.
- **Regresión Logística:** Aunque es clasificación, usa la técnica de regresión para predecir probabilidades de pertenencia a clases.
- **Regresión Ridge y Lasso:** Incluyen penalizaciones para reducir el sobreajuste y seleccionar variables.
- **Regresión No Lineal:** Modelos que capturan relaciones complejas no lineales.

5. Explique los conceptos de Validación Simple y Validación Cruzada.

Validación Simple

- Es una técnica de evaluación de modelos predictivos que divide aleatoriamente el conjunto de datos en dos partes:
    1. Conjunto de entrenamiento: se utiliza para aprender o construir el modelo.
    2. Conjunto de prueba: se usa para evaluar el desempeño del modelo.
- En RapidMiner, el operador de Validación Cruzada puede realizar este tipo de validación simple al dividir el dataset en entrenamiento y prueba, estimando el rendimiento del modelo sobre datos no vistos
- Su objetivo es obtener una estimación de la precisión del modelo sobre datos nuevos, pero tiene la desventaja de depender mucho de cómo se dividieron los datos (si la muestra fue poco representativa, el resultado puede estar sesgado).

Validación Cruzada (Cross-Validation)

- Es una técnica más robusta que repite varias veces el proceso de división del dataset para garantizar independencia entre los datos de entrenamiento y prueba.
- Se usa especialmente cuando el objetivo es predecir y evaluar la capacidad de generalización del modelo.
- Consiste en dividir el conjunto de datos en k particiones o “folds”:
    1. En cada iteración, una partición se usa como prueba y las demás como entrenamiento.
    2. Se repite k veces y se calcula la media aritmética de las precisiones obtenidas.

- En RapidMiner, el operador “Cross Validation” tiene dos subprocesos:
    1. Entrenamiento (donde el modelo aprende).
    2. Prueba (donde se mide el rendimiento).

       Así se obtiene una estimación más estable y confiable del desempeño real del modelo

6. Explique los métodos que existen para medir la precisión de un modelo de regresión.

Los modelos de regresión no clasifican categorías sino que predicen valores numéricos. Por eso, su precisión no se mide con una matriz de confusión, sino con métricas de error o ajuste que comparan los valores predichos con los reales.

Principales métodos de evaluación:

Error Absoluto Medio (MAE - Mean Absolute Error)

- Calcula el promedio de los valores absolutos de los errores.
- Mide cuánto se alejan las predicciones, en promedio, de los valores reales.
- Cuanto menor sea el MAE, mejor el modelo.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Error Cuadrático Medio (MSE - Mean Squared Error)

- Promedio de los errores al cuadrado.
- Penaliza más los errores grandes.

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

Raíz del Error Cuadrático Medio (RMSE - Root Mean Squared Error)

- Es la raíz cuadrada del MSE.
- Devuelve los errores en la misma unidad de la variable predicha.
- Más interpretable que el MSE.

Ejercicios Prácticos

1. Se cuenta con información de locales de venta de ropa y desde la gerencia se solicita realizar un pronóstico de ventas anual para la ropa de mujer basado en la información de la sucursal, las cuales son:

- ropamujer: ventas de prendas para mujer en millones de pesos durante el año.
- correo: número de catálogos enviados durante el año.
- páginas: número de páginas del catálogo.
- teléfono: número promedio de líneas para llamada abiertas para pedidos.

- impresa: cantidad gastada en publicidad impresa.
- servicio: número de representantes del servicio al cliente.
- idmercado: tipo de mercado. Solo se listan números de clases o tipos de mercado que se desea atacar. No se proveerá información adicional.
- tamañomer: tamaño potencial del mercado, proyectado de acuerdo a cifras del área de marketing.
- Idloc: ID de la tienda.
- Edadloc: años transcurridos desde la llegada a la zona.
- promo (promoción): tipo de promoción que se llevó a cabo durante el año. De nuevo, información descriptiva sobre estas promociones es reservada.
- nomina: valor total de la nómina durante el año

a. Realizar dos visualizaciones que permita analizar la relación de las variables con el objetivo del proyecto. Saque conclusiones de las visualizaciones que realizó.

Matriz de correlación correspondiente al dataset usando el operador “Correlation Matrix”:

Attribut...	idloc	edadloc	correo	paginas	telefono	impresa	servicio	nomina	idmerc...	tamamer	promo	ropamu...
idloc	1	0.041	0.030	0.034	0.036	0.005	0.054	0.011	0.016	?	-0.052	-0.056
edadloc	0.041	1	0.033	0.019	0.028	0.028	0.637	-0.009	-0.018	?	0.014	0.349
correo	0.030	0.033	1	0.651	0.752	0.016	0.031	0.026	0.047	?	-0.025	0.416
paginas	0.034	0.019	0.651	1	0.652	0.043	0.012	0.042	0.016	?	-0.032	0.378
telefono	0.036	0.028	0.752	0.652	1	0.039	0.038	0.016	0.052	?	-0.021	0.375
impresa	0.005	0.028	0.016	0.043	0.039	1	0.037	-0.026	-0.003	?	0.013	0.365
servicio	0.054	0.637	0.031	0.012	0.038	0.037	1	0.006	0.040	?	-0.027	0.535
nomina	0.011	-0.009	0.026	0.042	0.016	-0.026	0.006	1	-0.001	?	-0.420	-0.003
idmercado	0.016	-0.018	0.047	0.016	0.052	-0.003	0.040	-0.001	1	?	-0.006	0.035
tamamer	?	?	?	?	?	?	?	?	?	1	?	?
promo	-0.052	0.014	-0.025	-0.032	-0.021	0.013	-0.027	-0.420	-0.006	?	1	-0.007
ropamujer	-0.056	0.349	0.416	0.378	0.375	0.365	0.535	-0.003	0.035	?	-0.007	1

Pesos de cada variable respecto al label usando el operador “Weight by Correlation” :

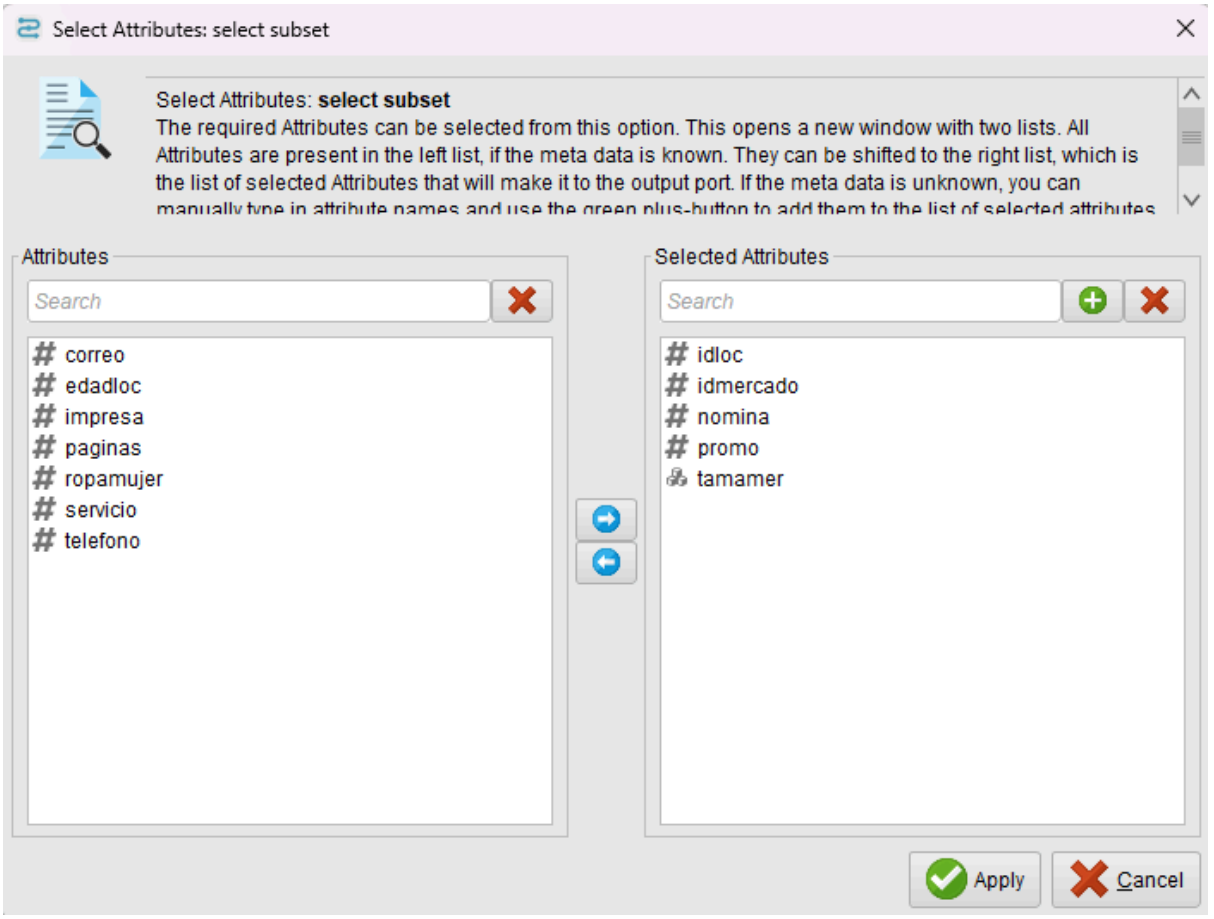
attribute	wei... ↓
servicio	0.535
correo	0.416
paginas	0.378
telefono	0.375
impresa	0.365
edadloc	0.349
idloc	0.056
idmercado	0.035
promo	0.007
nomina	0.003
tamamer	0.000

A partir de las visualizaciones generadas, se observa cómo algunas de las variables presentan alta correlación entre sí, como por ejemplo paginas-correo, ropamujer-correo, mientras que otras muestran baja correlación, como nómina- ropa mujer, impresa-promo. Asimismo mediante el uso del operador “Weight by Correlation” se identificaron las variables con mayor impacto en la determinación del label. Permitiendo sacar del modelo aquellas que no se consideran de gran impacto o con poca influencia en el modelo (por ejemplo, nómina,

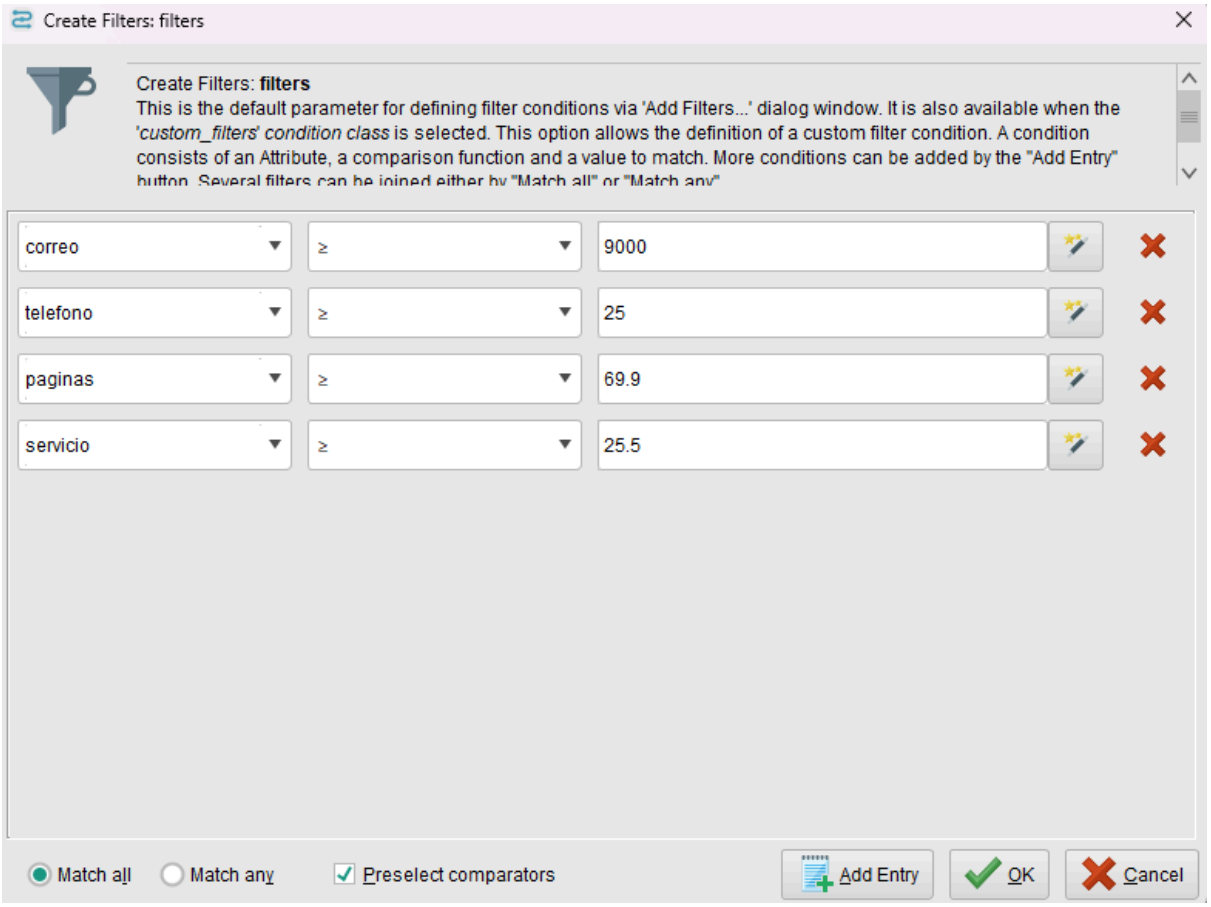
promo, tamamer, entre otros).

Se realizaron ajustes en los datos del dataset para adecuarlos al modelo a implementar.

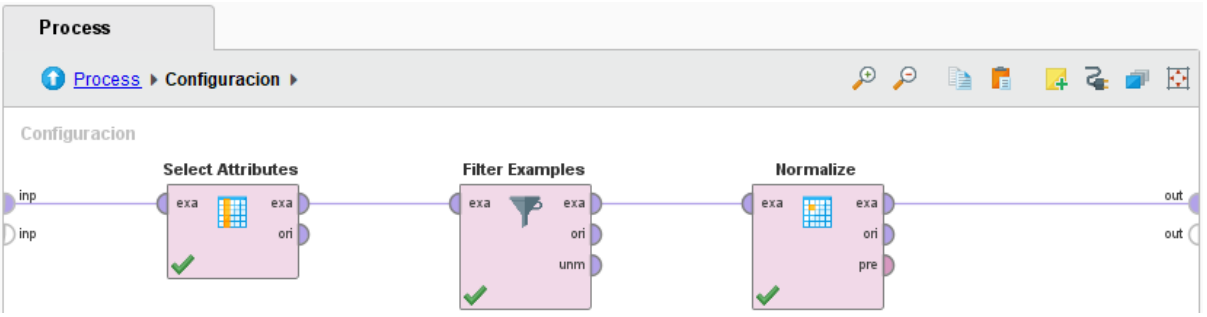
A través del operador “Select attributes” se quitaron las variables que tenían bajo impacto en el label:



Se quitaron valores outlier o atípicos usando el operador “Filter Examples”:

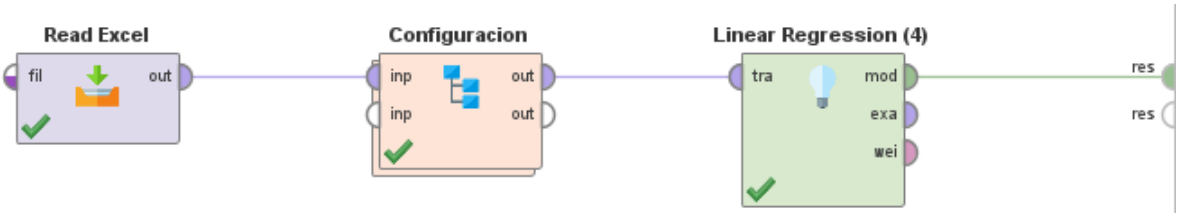


Por último se utilizó el operador “Normalize” para normalizar los datos. Todos estos cambios se los encapsulo en el subproceso “Configuración”:



b. Realice un modelo de Regresión Lineal con los datos brindados.

Implementación del modelo mediante el operador “Linear Regression”:

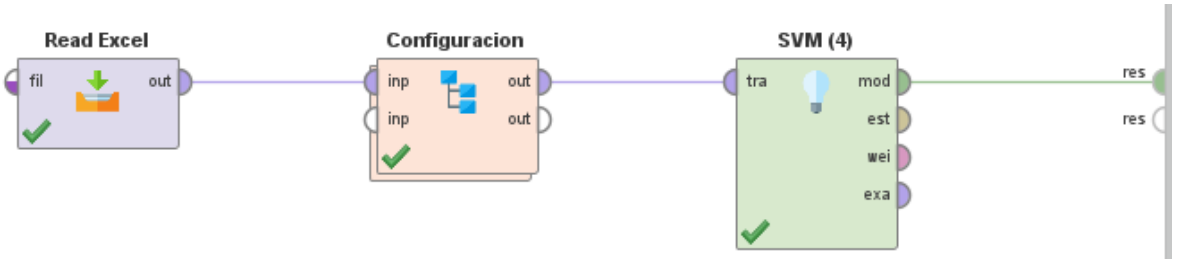


Resultado:

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
correo	13393.527	1888.372	0.222	0.407	7.093	0.000	****
paginas	7478.486	1476.598	0.130	0.604	5.065	0.000	****
telefono	2515.153	1914.152	0.040	0.431	1.314	0.189	
impresa	21177.614	1144.500	0.371	0.994	18.504	0	****
servicio	27447.994	1191.204	0.462	0.995	23.042	0	****
(Intercept)	22152.834	1018.049	?	?	21.760	0	****

c. Realice un modelo de SVM para resolver el problema

Modelo de SVM aplicado a través del operador “SVM (Support Vector Machine)”:



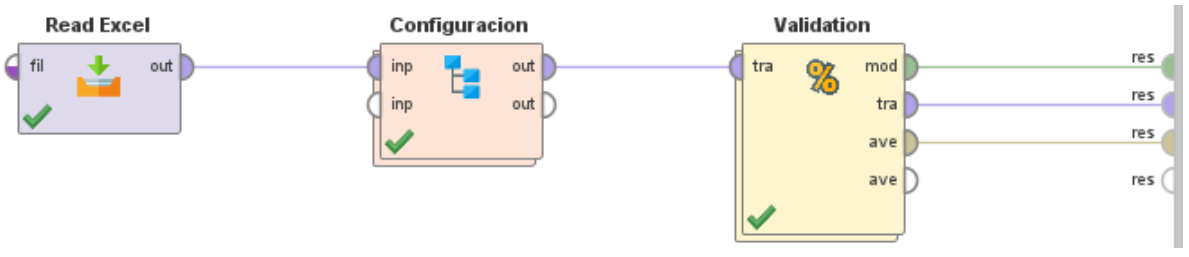
Resultado:

Kernel Model

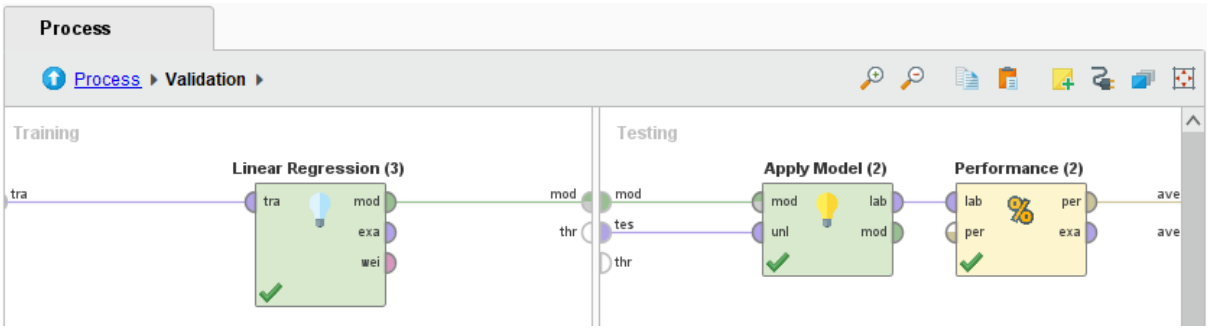
```
Total number of Support Vectors: 1231
Bias (offset): 53500.714

w[edadloc] = 182.581
w[correo] = 2758.219
w[paginas] = 1779.205
w[telefono] = 1311.588
w[impresa] = 4906.128
w[servicio] = 5983.589
```

d. Valide los modelos del punto anterior validación simple y cruzada.

Validación simple(operador Validation):



Para el modelo de regresión lineal:



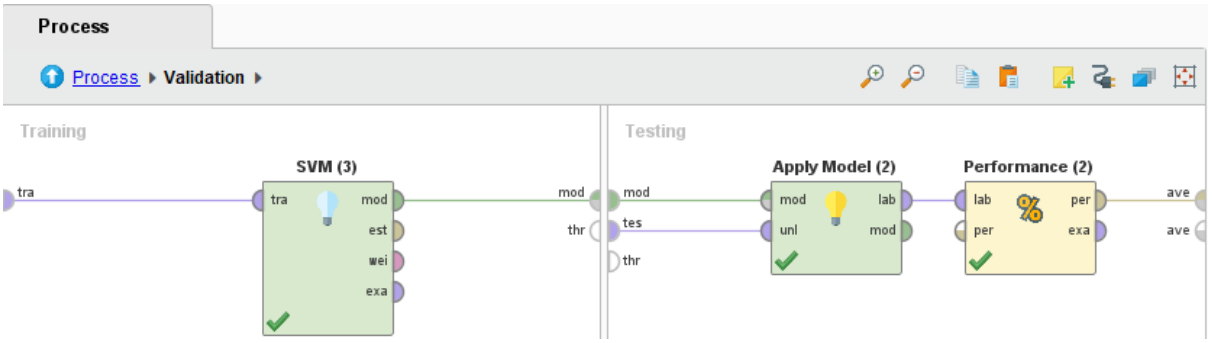
Resultado:

root_mean_squared_error: 11520.318 +/- 0.000

absolute_error: 9121.763 +/- 7036.416

relative_error: 18.88% +/- 17.96%

Para el modelo de SVM:



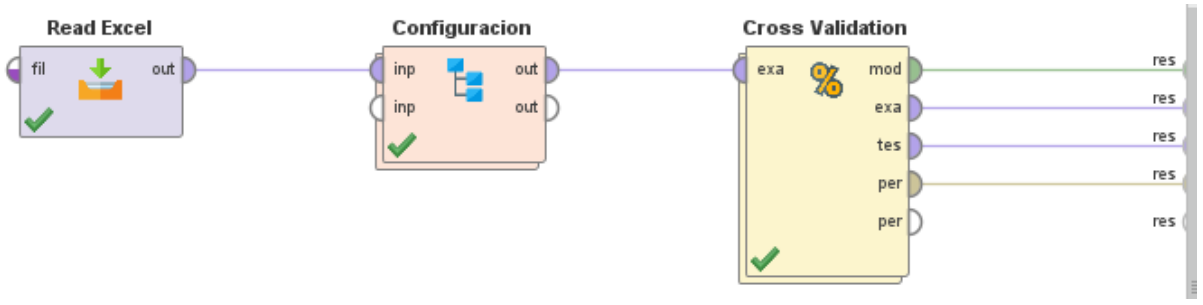
Resultado:

root_mean_squared_error: 11799.608 +/- 0.000

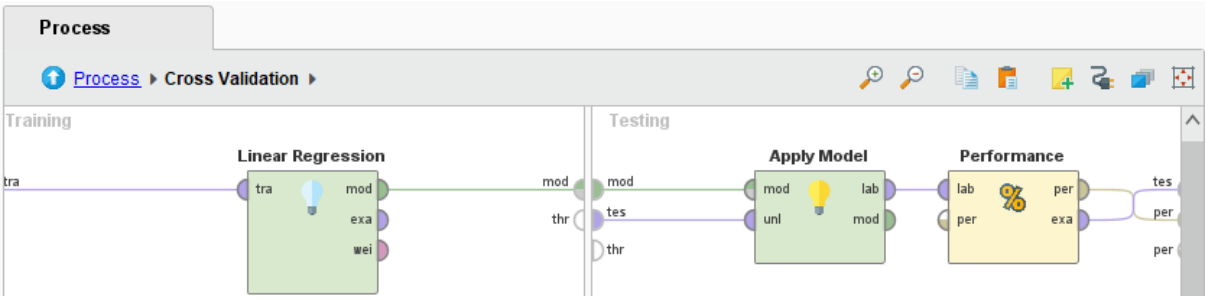
absolute_error: 9257.532 +/- 7316.342

relative_error: 19.21% +/- 19.32%

Validación cruzada:



Para el modelo de regresión lineal:



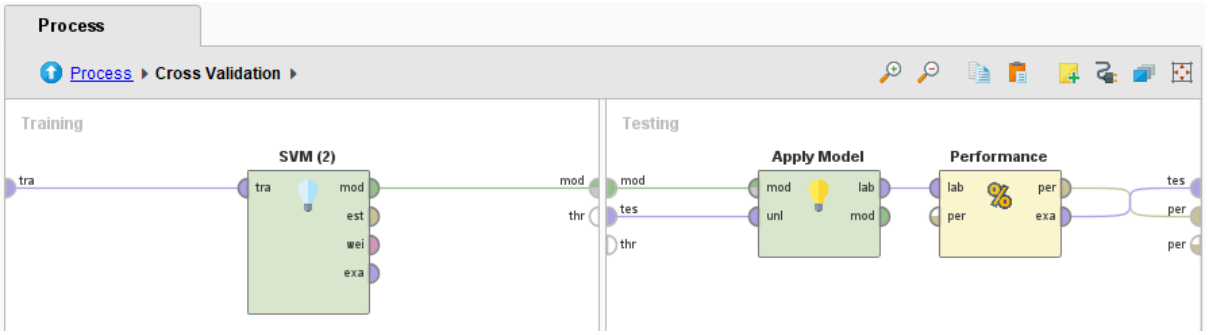
Resultado:

root_mean_squared_error: 11554.770 +/- 1236.793 (micro average: 11615.780 +/- 0.000)

absolute_error: 9143.905 +/- 851.780 (micro average: 9145.156 +/- 7161.875)

relative_error: 18.75% +/- 1.41% (micro average: 18.75% +/- 18.24%)

Para el modelo de SVM:



Resultado:

root_mean_squared_error: 11788.162 +/- 1303.106 (micro average: 11854.805 +/- 0.000)

absolute_error: 9267.662 +/- 955.423 (micro average: 9269.248 +/- 7390.361)

relative_error: 18.91% +/- 1.53% (micro average: 18.91% +/- 18.86%)

e. Compare que modelo se comporta mejor

De los resultados obtenidos tras realizar las validaciones en los modelos anteriores, podemos remarcar que:

En primer lugar, usando el operador “Validation (Split Validation)”, correspondiente a la validación simple, el modelo que obtuvo mejor resultado, es decir que tiene menos error, fue el modelo de regresión lineal.

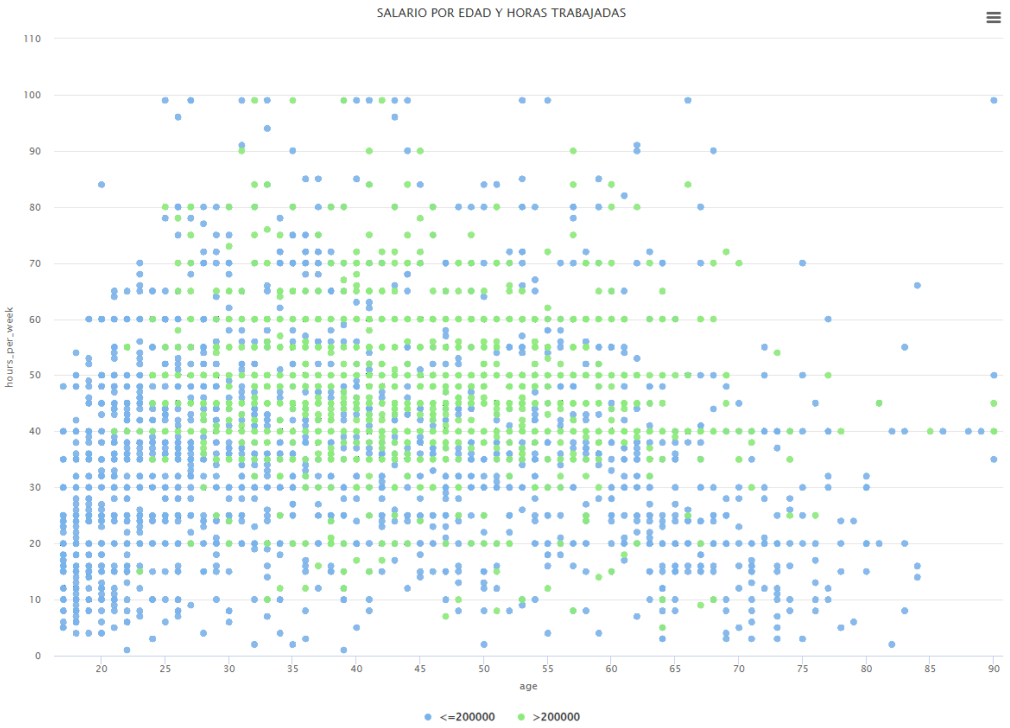
De igual modo, utilizando el operador “Cross Validation” para la validación cruzada, se confirma nuevamente que el modelo de regresión lineal obtuvo mejores resultados que el

modelo de SVM.

De esta forma se concluye que el mejor modelo que se adapta para este conjunto de datos es el modelo de regresión lineal, siendo su mejor desempeño el obtenido mediante la validación cruzada.

2. Se desea clasificar a las personas que ganan por debajo de $200.000 y las que superan esa cantidad. Se trabajará con el dataset adult_data.csv

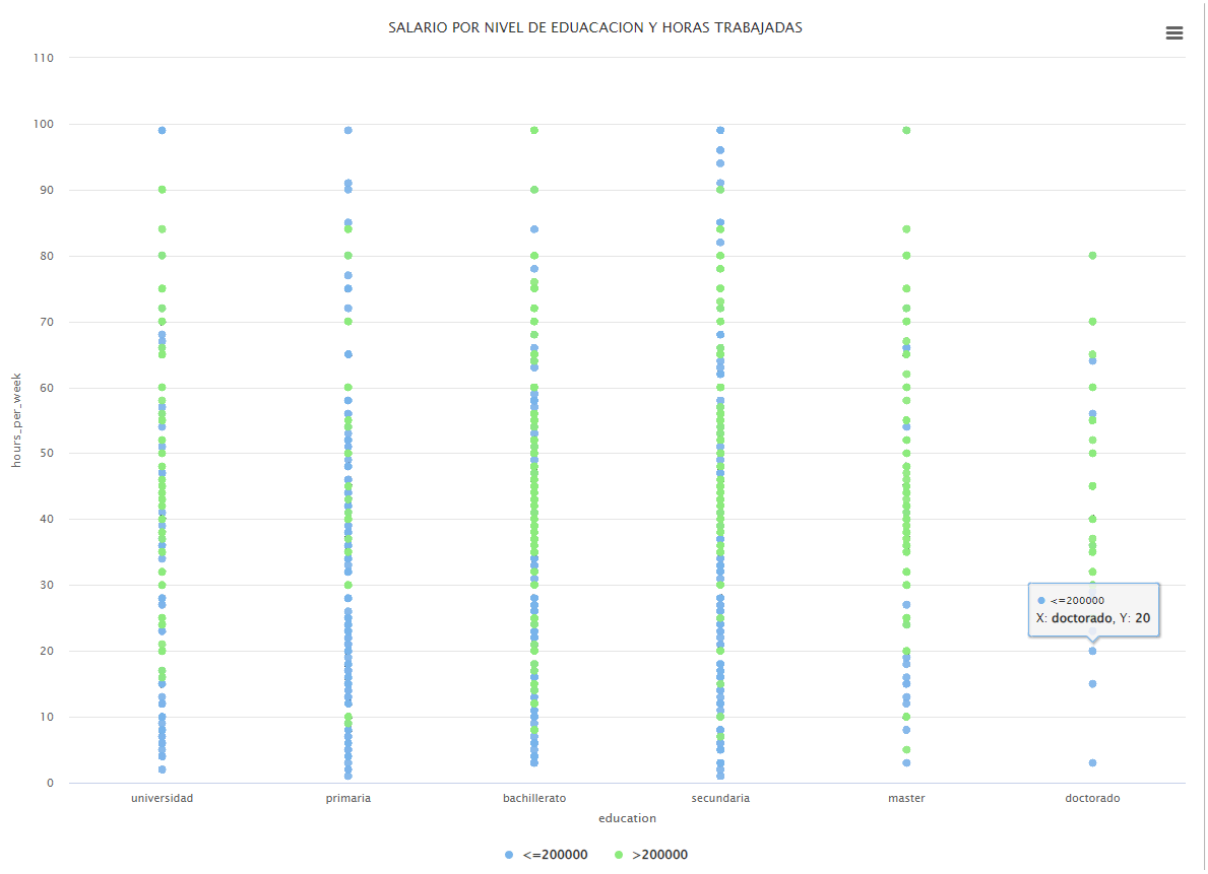
a. Realizar dos visualizaciones que permitan analizar la relación de las variables con el objetivo del proyecto.



En el gráfico se observa una tendencia general donde las personas con mayores ingresos (puntos verdes) tienden a trabajar más horas semanales que las de ingresos menores (puntos azules).

Además, los individuos de mayor edad (entre 40 y 60 años) presentan una mayor concentración de altos salarios, lo que podría estar relacionado con la experiencia laboral o con ocupaciones de mayor responsabilidad.

En cambio, las personas más jóvenes (menores de 30 años) muestran mayor dispersión, pero en su mayoría pertenecen al grupo de menores ingresos.



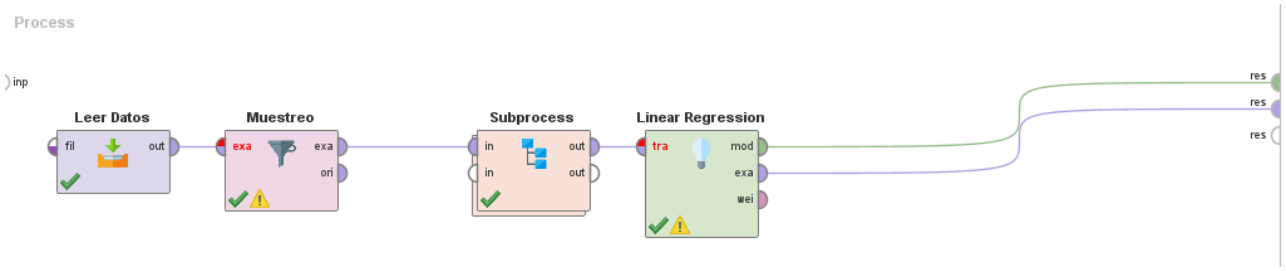
En el gráfico se observa la relación entre el nivel educativo y las horas trabajadas por semana, diferenciando los grupos de salario.

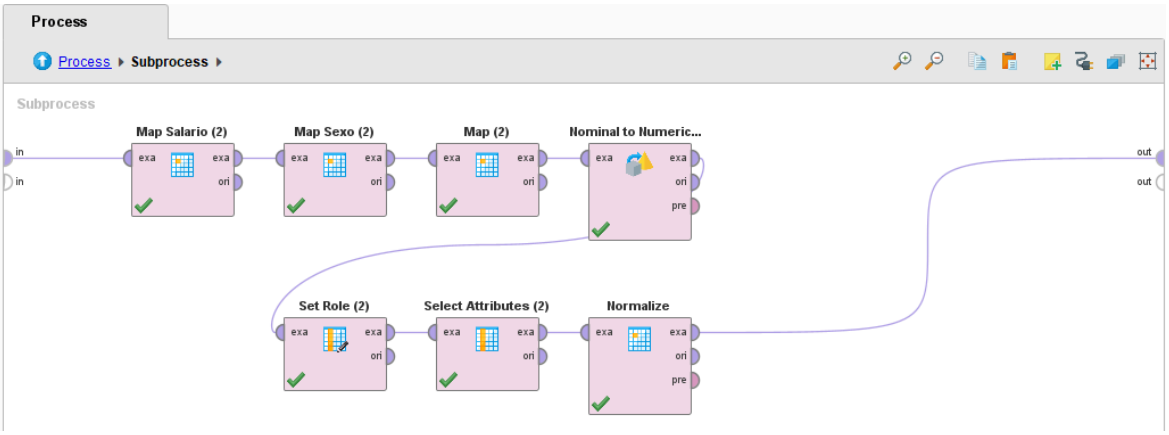
Las personas con niveles educativos más altos (universidad, máster y doctorado) tienden a concentrarse en los rangos de salario más altos (>200000), representados por los puntos verdes.

También se nota que, aunque las horas trabajadas por semana son similares entre los distintos niveles educativos (la mayoría entre 35 y 50 horas), los ingresos aumentan claramente con el nivel de estudios.

Por ejemplo, casi no hay individuos con educación primaria o secundaria en el grupo de mayores salarios, mientras que los niveles universitarios y de posgrado presentan una proporción más alta de puntos verdes.

b. Realice un modelo de Regresión Lineal con los datos brindados.

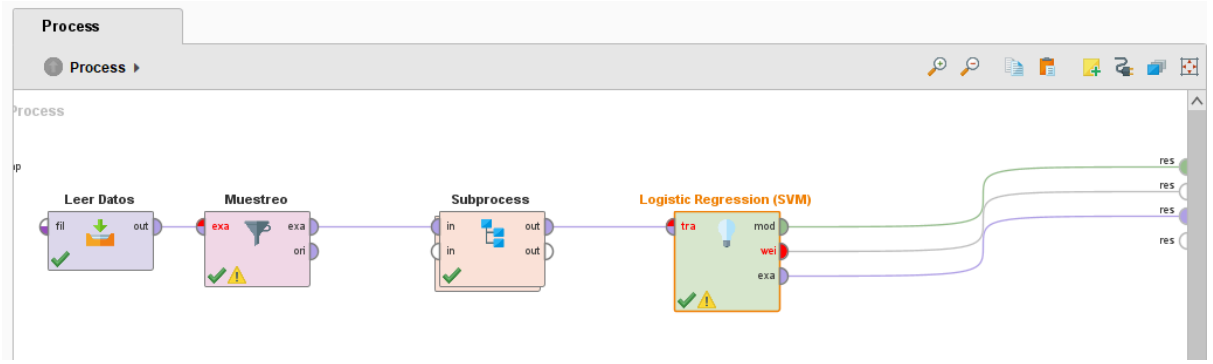




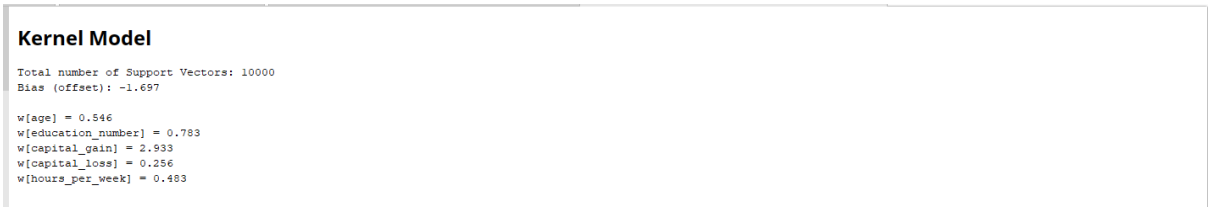
Resultado:

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
age	0.441	0.021	0.185	0.990	20.621	0	****
education_number	0.671	0.023	0.264	0.967	29.105	0	****
capital_gain	0.951	0.052	0.165	0.979	18.345	0	****
capital_loss	0.395	0.036	0.097	0.994	10.876	0	****
hours_per_week	0.548	0.033	0.153	0.958	16.833	0	****
(Intercept)	-0.536	0.019	?	?	-28.594	0	****

c. Realice un modelo de clasificación con SVM.

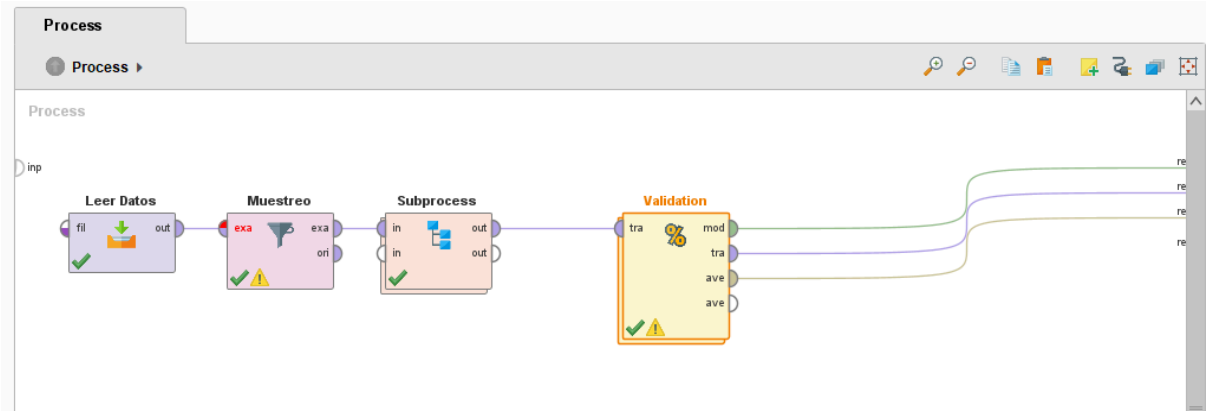


Resultado:



d. Valide los modelos del punto c y d mediante validación simple y cruzada.

Validación Directa::



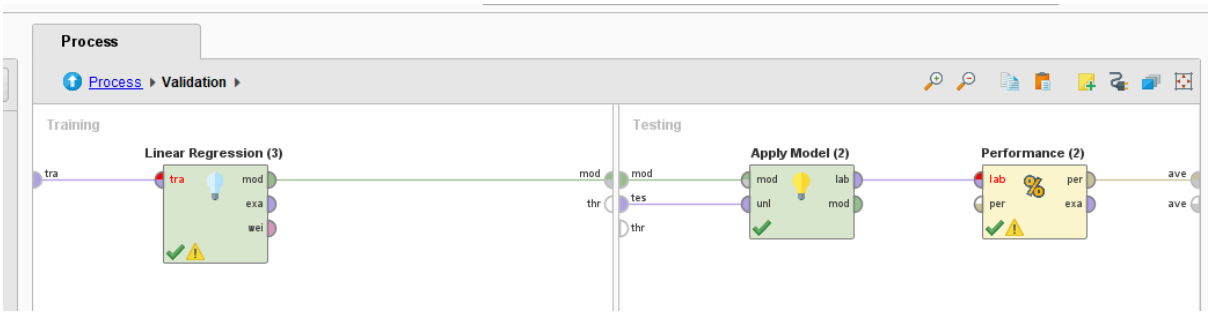
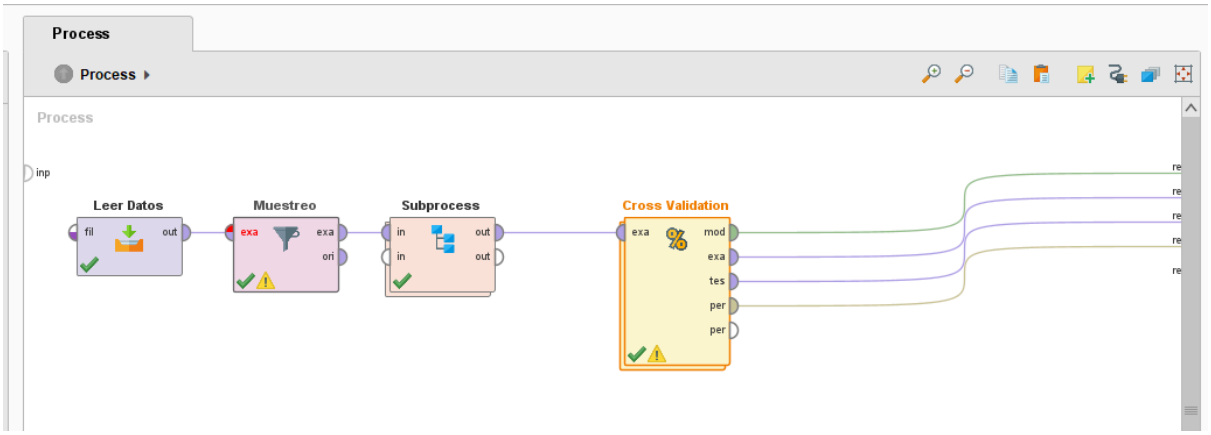


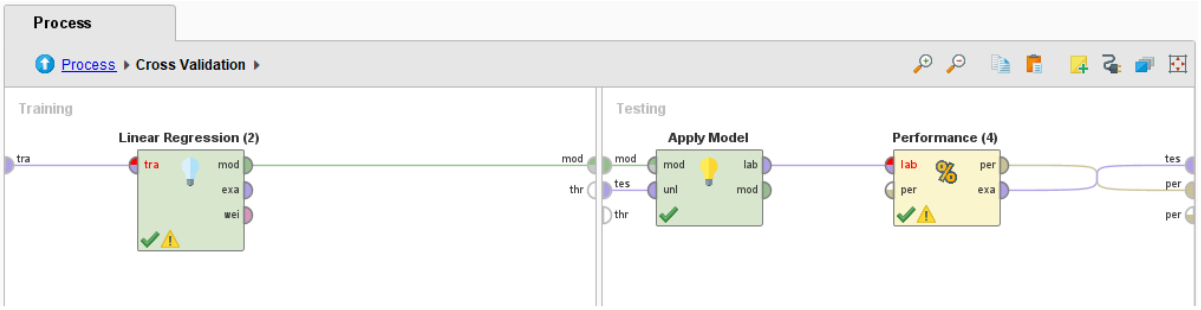
Table View Plot View

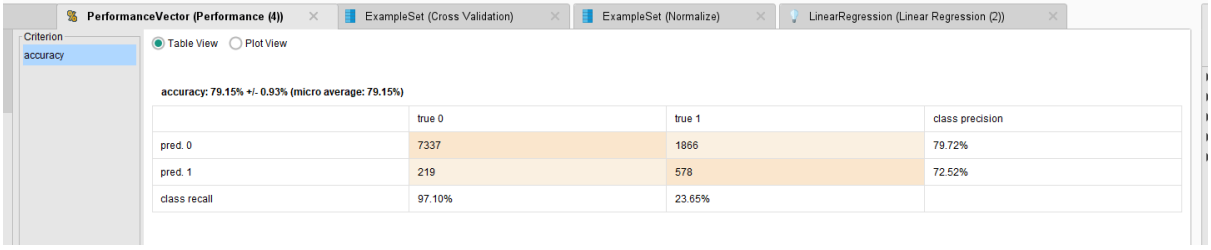
accuracy: 76.74%

	true 0	true 1	class precision
pred. 0	7313	2136	77.39%
pred. 1	167	284	62.97%
class recall	97.77%	11.74%	

Validación Cruzada:







	true 0	true 1	class precision
pred. 0	7337	1866	79.72%
pred. 1	219	578	72.52%
class recall	97.10%	23.65%	

e. **¿Qué se puede concluir sobre la precisión que tiene cada modelo?**

El modelo de regresión lineal aplicado al problema de clasificación presenta una precisión del 76.74 % bajo validación directa y del 73.15 % bajo validación cruzada.

Aunque la validación directa muestra una mayor exactitud, la validación cruzada ofrece una estimación más robusta y menos sesgada, al evaluar el modelo en múltiples particiones del conjunto de datos.

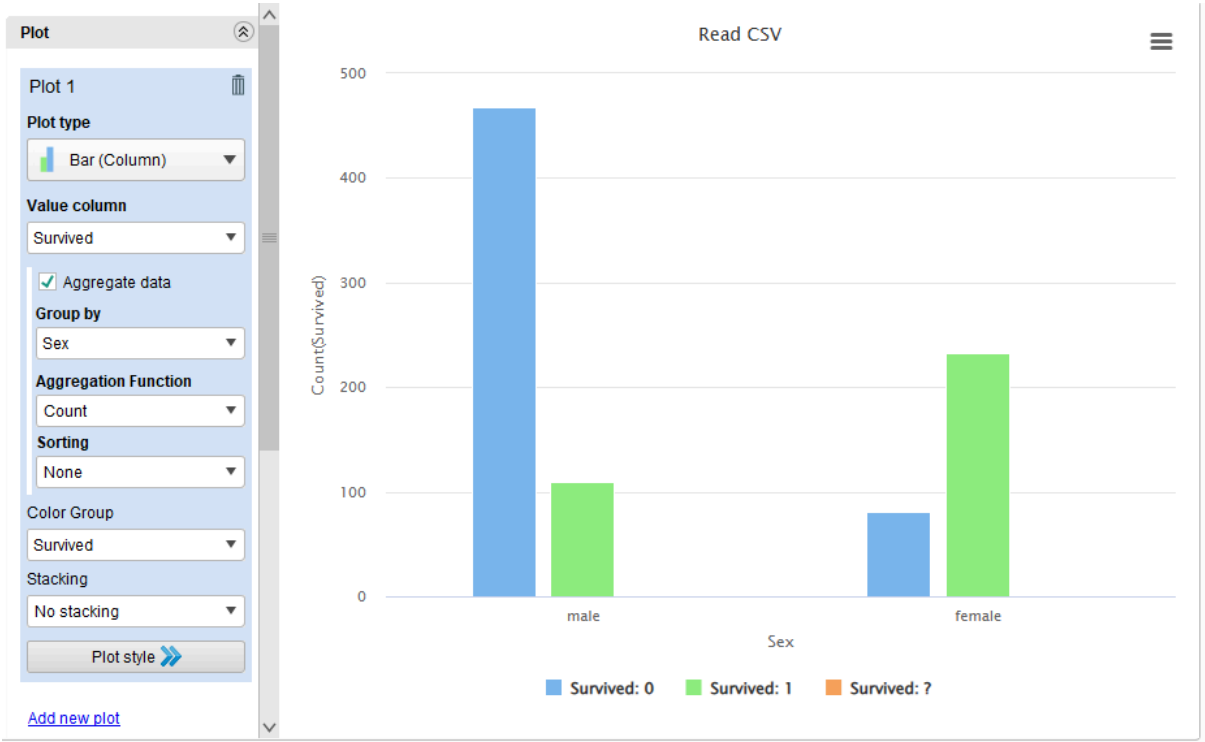
Además, se observa que el modelo tiene mayor capacidad para predecir la clase de menores ingresos, mientras que su desempeño disminuye significativamente en la predicción de la clase de mayores ingresos.

3. Analizaremos datos que se obtuvieron del barco de Titanic, donde se tienen atributos respecto a los pasajeros que estuvieron en el embarque. Utilizaremos el archivo titanic.xlsx.

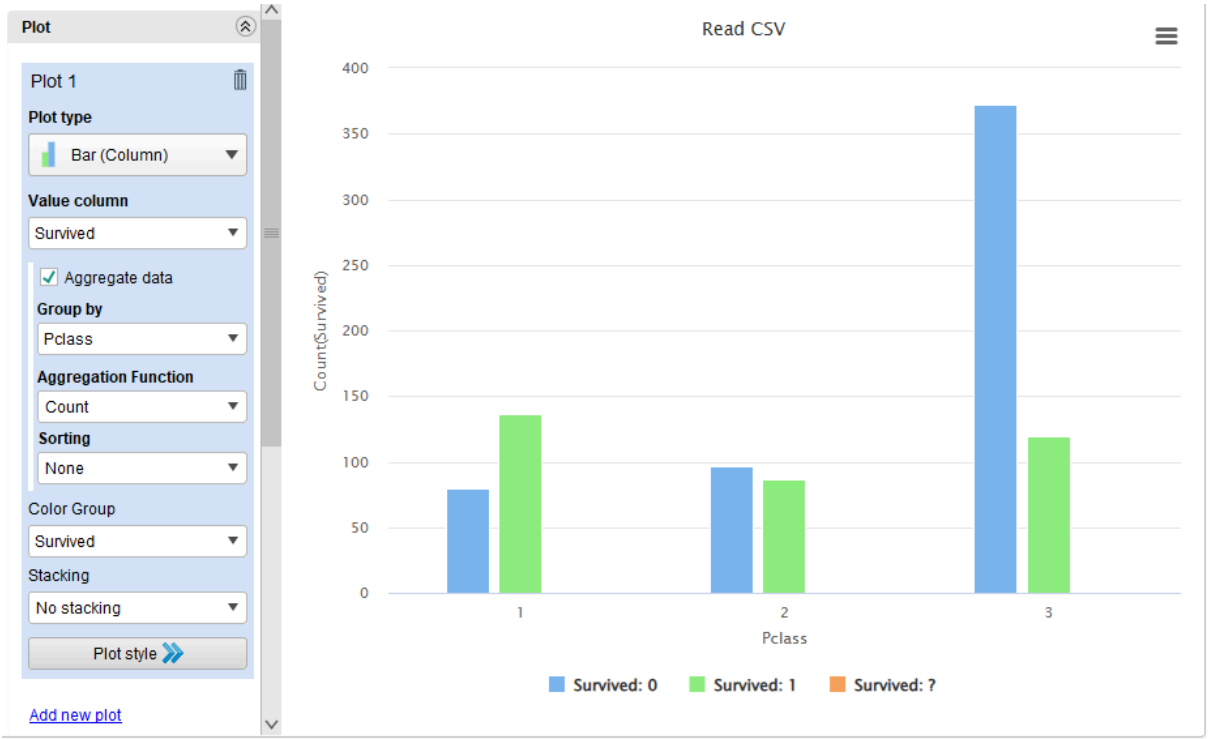
1. Realizar la visualización de datos:

   a. Los sobrevivientes por sexo, clase, por edad.

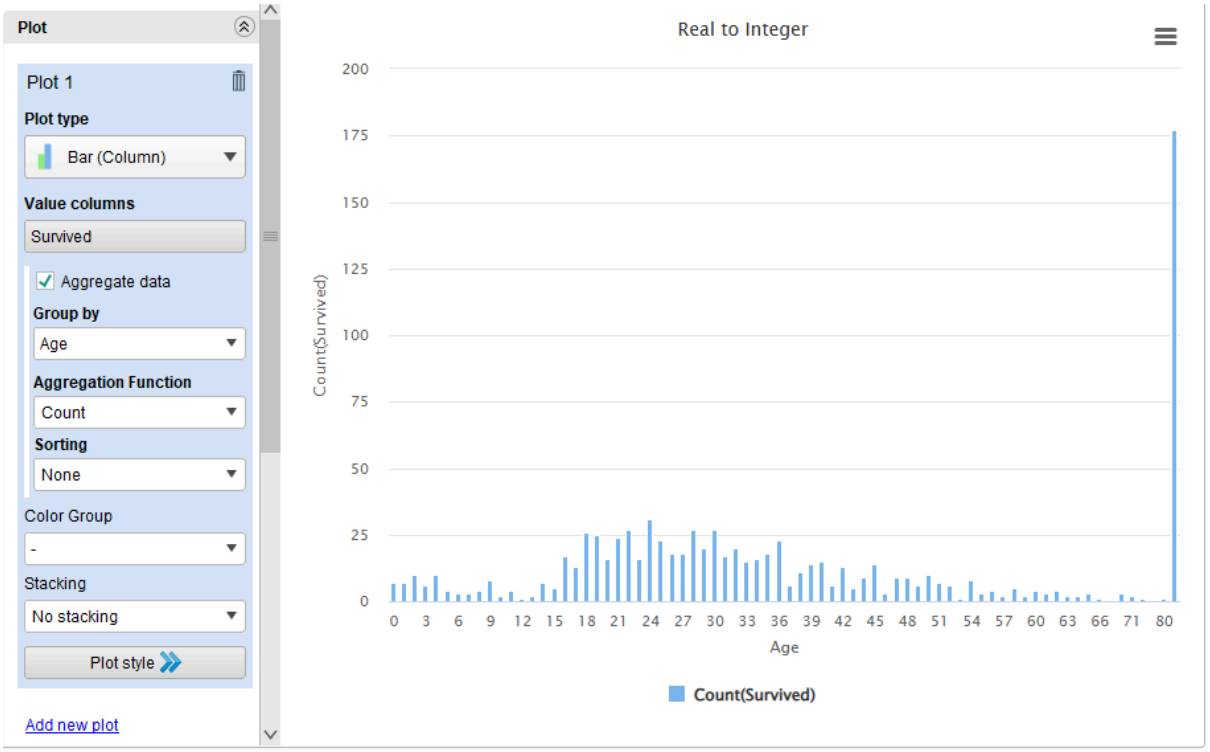
   Sobrevientes por sexo



Sobrevivientes por clase



Sobrevivientes por edad



b. Existe relación de los sobrevivientes respecto a si tenían parientes.

attribute	wei...
Sex	0.543
Pclass	0.338
Fare	0.257
Embarked	0.109
Parch	0.082
Age	0.077
Ticket	0.047
Cabin	0.046
SibSp	0.035
Passeng...	0.005
Name	0.005

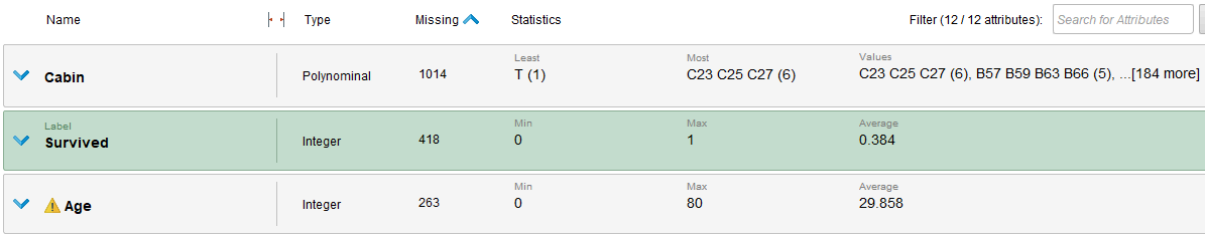
Se puede observar que no hay una fuerte relación entre los sobrevivientes y parientes, es decir entre el atributo “Survived” y “SibSp”/”Parch”.

2. Limpieza de los datos.

   a. Identificar los datos faltantes. ¿Considera necesario rellenar los datos? En caso afirmativo realícelo.

   Existen varios atributos con datos faltantes o nulos, por eso van a ser rellenados con el promedio de sus valores a través del operador "Replace Missing Values". Los atributos son los siguientes:

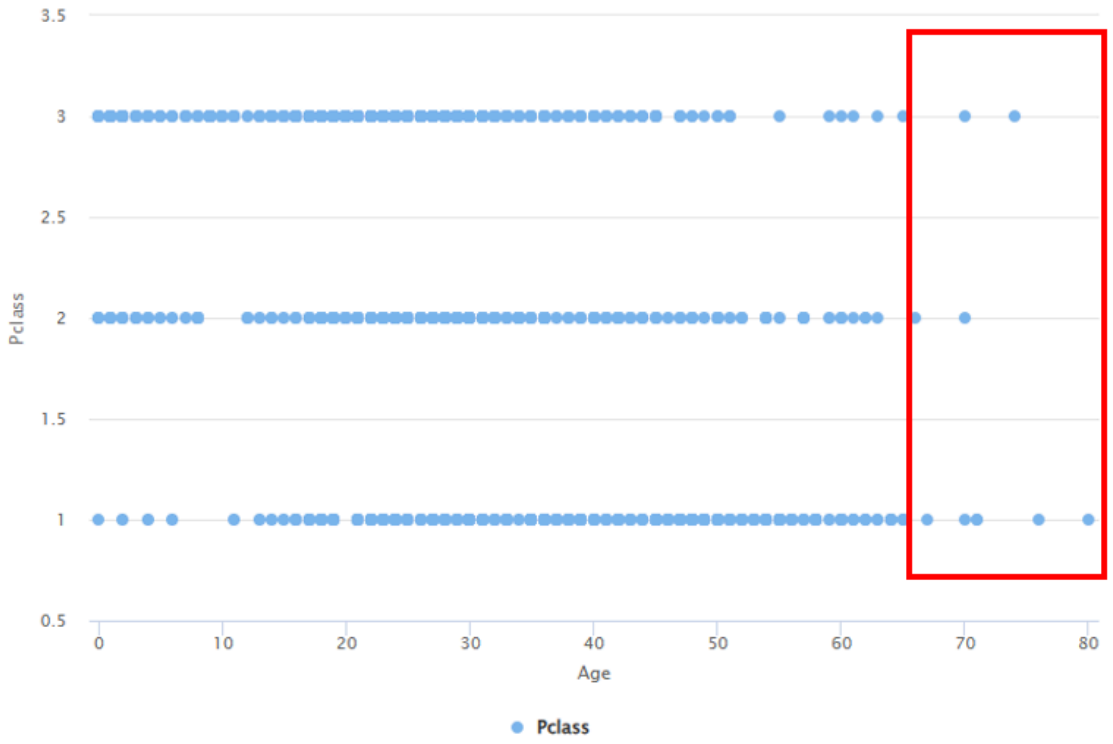
   Antes



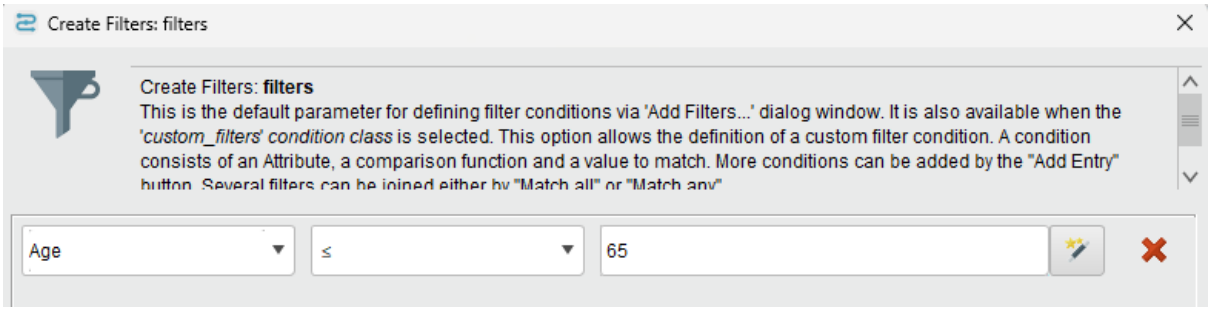
Después

Name	Type	Missing	Statistics			Filter (12 / 12 attributes): Search for Attributes
Age	Integer	0	Min 0	Max 80	Average 29.886	
Survived	Integer	0	Min 0	Max 1	Average 0.261	
Cabin	Polynominal	0	Least T (1)	Most C23 C25 C27 (1020)	Values C23 C25 C27 (1020), B57 B59 B63 B66 (5), ...[184 more]	

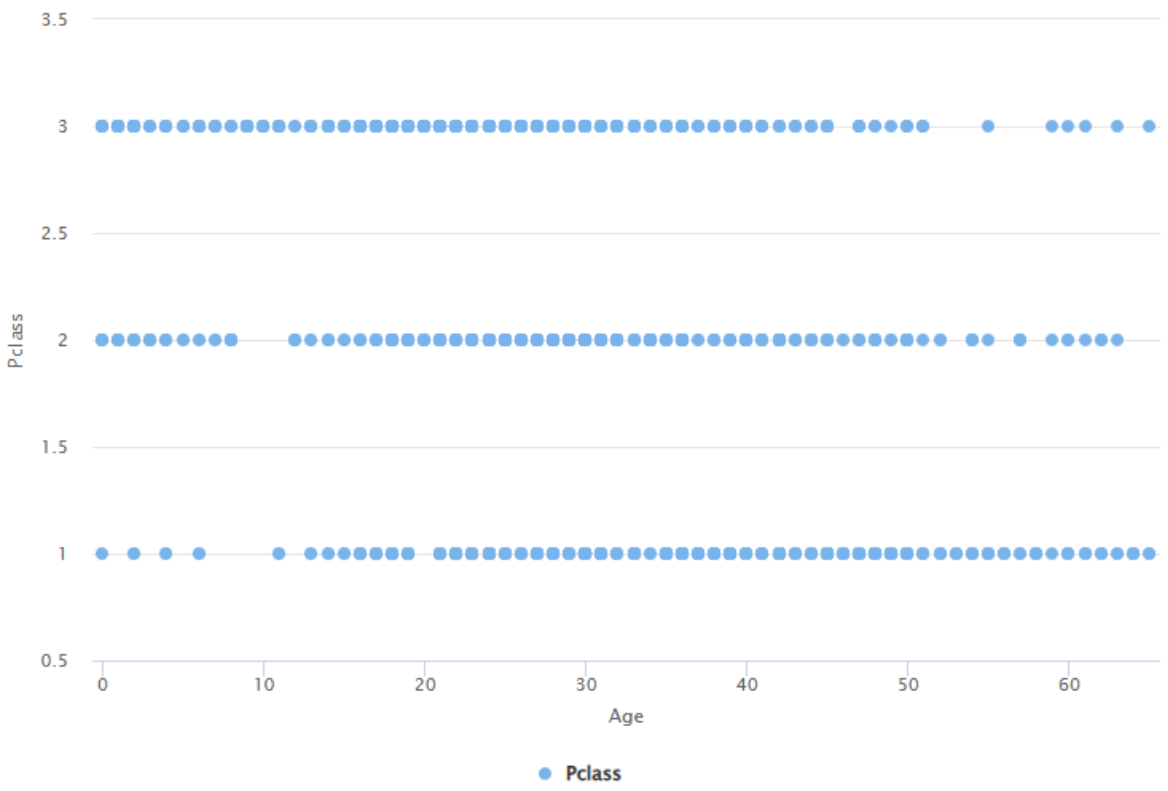
b. Identificar datos anómalos. De acuerdo a la clase del boleto y edad identifique los datos anómalos y descarte esos datos.



Dada la imagen, los datos anómalos son aquellos que superan los 65 años de edad, por lo que van a ser descartados mediante el operador “Filter Examples”.



Resultado



3. Realice una matriz de correlación e identifique cuales son los atributos que tienen mayor relación.

Attributes	Sex = m...	Sex = fe...	Embark...	Embark...	Embark...	Age	Passen...	Pclass	SibSp	Parch	Fare
Sex = male	1	-1	0.121	-0.068	-0.090	0.046	0.016	0.128	-0.108	-0.213	-0.188
Sex = fe...	-1	1	-0.121	0.068	0.090	-0.046	-0.016	-0.128	0.108	0.213	0.188
Embarke...	0.121	-0.121	1	-0.776	-0.490	-0.068	-0.053	0.098	0.074	0.073	-0.177
Embarke...	-0.068	0.068	-0.776	1	-0.164	0.080	0.051	-0.270	-0.048	-0.008	0.290
Embarke...	-0.090	0.090	-0.490	-0.164	1	-0.012	0.015	0.229	-0.048	-0.101	-0.129
Age	0.046	-0.046	-0.068	0.080	-0.012	1	0.032	-0.362	-0.194	-0.129	0.171
Passeng...	0.016	-0.016	-0.053	0.051	0.015	0.032	1	-0.039	-0.058	0.008	0.028
Pclass	0.128	-0.128	0.098	-0.270	0.229	-0.362	-0.039	1	0.062	0.017	-0.558
SibSp	-0.108	0.108	0.074	-0.048	-0.048	-0.194	-0.058	0.062	1	0.373	0.159
Parch	-0.213	0.213	0.073	-0.008	-0.101	-0.129	0.008	0.017	0.373	1	0.224
Fare	-0.188	0.188	-0.177	0.290	-0.129	0.171	0.028	-0.558	0.159	0.224	1

- **Sexo masculino vs femenino**: Como era de esperarse, tienen una correlación perfecta negativa (-1), ya que son mutuamente excluyentes.

- **Embarked = S vs Embarked = C**: Correlación negativa fuerte (-0.776), lo que sugiere que los pasajeros que embarcaron en Southampton (S) rara vez lo hicieron en Cherbourg (C), y viceversa.
- **Pclass vs Fare**: Correlación negativa fuerte (-0.558), lo que indica que los pasajeros de clases más altas (Pclass más bajo) pagaron tarifas más altas.
- **Pclass vs Age**: Correlación positiva moderada (0.408), lo que sugiere que los pasajeros de clases más altas tendían a ser mayores.
- **SibSp y Parch**: Correlación positiva (0.373), lo que tiene sentido: quienes viajaban con hermanos/esposos también solían viajar con padres/hijos.

4. Identifique el atributo dependiente (label) para el dataset. Aplique el operador Weight by Correlation. Identifique los atributos que no son de importancia para el label y descártelos.

attribute	wei... ↓
Sex = male	0.405
Sex = female	0.405
Pclass	0.248
Fare	0.177
Embarked = C	0.098
Embarked = S	0.084
Age	0.055
Parch	0.054
SibSp	0.014
Embarked = Q	0.012

Atributo dependiente (label)

Survived, es la variable objetivo que queremos predecir (0 = no sobrevivió, 1 = sobrevivió).

Atributos con mayor relevancia (según peso)

Estos atributos tienen mayor correlación con Survived y deberían incluirse en el modelo:

Atributo	Peso	Interpretación
Sex = male	0.405	El sexo es altamente relevante. Los hombres tuvieron menor tasa de supervivencia.
Sex = female	0.405	El sexo femenino tiene alta probabilidad de sobrevivir.
Pclass	0.248	Los pasajeros de primera clase sobrevivieron más.
Fare	0.177	Tarifa más alta suele asociarse con mejor clase y mayor supervivencia.
Embarked = C	0.098	Embarcar en Cherbourg tiene cierta relación con supervivencia.
Embarked = S	0.084	Southampton también tiene influencia moderada.
Age	0.055	La edad influye: niños y adultos mayores tienen patrones distintos.
Parch	0.054	Viajar con padres/hijos tiene leve impacto.

Atributos descartables (por baja relevancia)

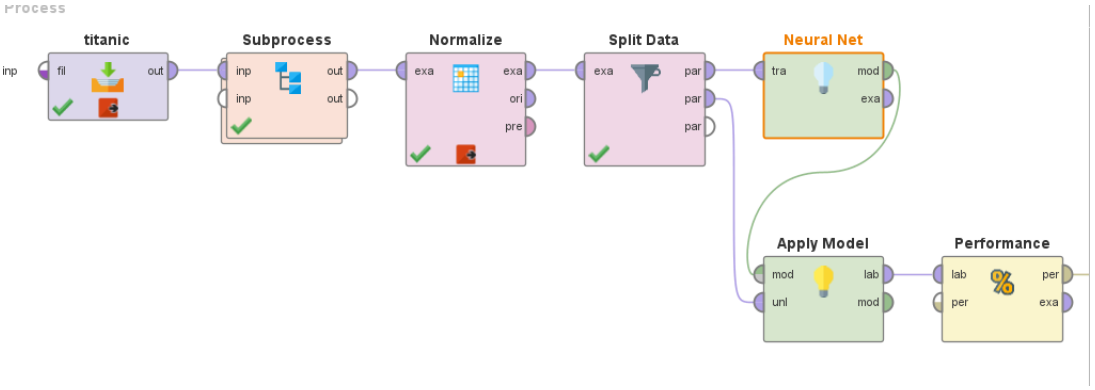
Atributo	Peso	Motivo de descarte
SibSp	0.014	Muy baja correlación con Survived.
Embarked = Q	0.012	Peso insignificante.

5. Realice el modelo de predicción con operadores de redes neuronales (a elección) y SVM. Compare los modelos. ¿Cuál modelo se comporta mejor? Brinde las conclusiones necesarias.

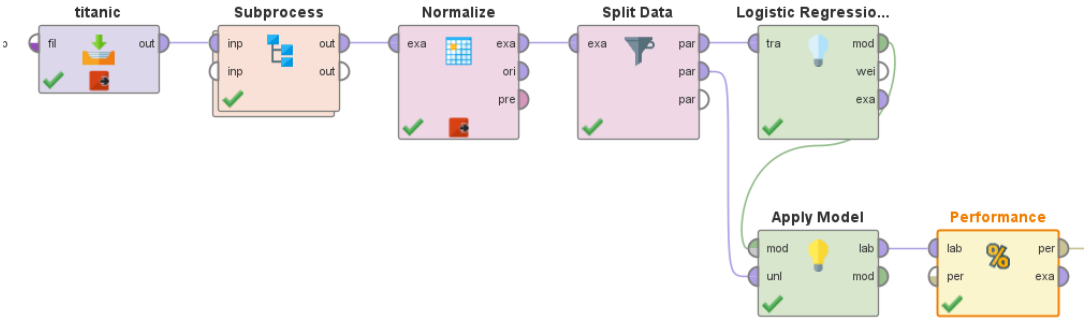
Configuración de los modelos

Característica	**Red Neuronal**	**SVM (Kernel ANOVA)**
Proporción de datos	80% entrenamiento / 20% prueba	80% entrenamiento / 20% prueba
Arquitectura / Kernel	1 capa oculta, 10 neuronas	Kernel ANOVA
Variable objetivo	Survived	Survived

Modelo de Red Neuronal



Modelo SVM



Resultados obtenidos

Modelo de Red Neuronal

PerformanceVector (Performance)

Table View | Plot View

accuracy: 76.45%

	true 0	true 1	class precision
pred. 0	163	33	83.16%
pred. 1	28	35	55.56%
class recall	85.34%	51.47%	

Modelo SVM

PerformanceVector (Performance)

Table View | Plot View

accuracy: 74.13%

	true 0	true 1	class precision
pred. 0	162	38	81.00%
pred. 1	29	30	50.85%
class recall	84.82%	44.12%	

Conclusiones

- Ambos modelos —**Red Neuronal** y **SVM con kernel ANOVA**— presentan un desempeño **similar**, con una diferencia de precisión de apenas **2.32 puntos porcentuales** (76.45% vs. 74.13%).
- Aunque la **Red Neuronal** muestra mejores métricas en precisión, recall y menor tasa de error, **la diferencia no es drásticamente significativa** en términos prácticos.
- En aplicaciones reales, una diferencia del **2% en precisión** puede no justificar una mayor complejidad computacional o tiempo de entrenamiento, especialmente si el modelo más simple (como SVM) ofrece resultados aceptables.

BIBLIOGRAFÍA

IBM. (s. f.). ***¿Qué son las redes neuronales?*** **IBM Think. Recuperado de** https://www.ibm.com/es-es/think/topics/neural-networks

OpenAI. (2025, octubre 6). ***Explicación sobre modelos de regresión y sus variantes*** **[Respuesta generada por inteligencia artificial]. ChatGPT.** https://chat.openai.com/

Google Cloud. (s. f.). ¿Qué es una red neuronal? Google Cloud Discover. Recuperado de: https://cloud.google.com/discover/what-is-a-neural-network?hl=es

Amazon Web Services (AWS). (s. f.). ¿Qué es una red neuronal? AWS. Recuperado de: https://aws.amazon.com/es/what-is/neural-network/#seo-faq-pairs#what-is-a-nn