PRÁCTICA ETL + EDA + MODELING

Modulo III – FUNDAMENTOS TECNOLOGICOS

Jorge Morales Mercado

Link repositorio: https://github.com/jmorales190/M3_Msc_2.git

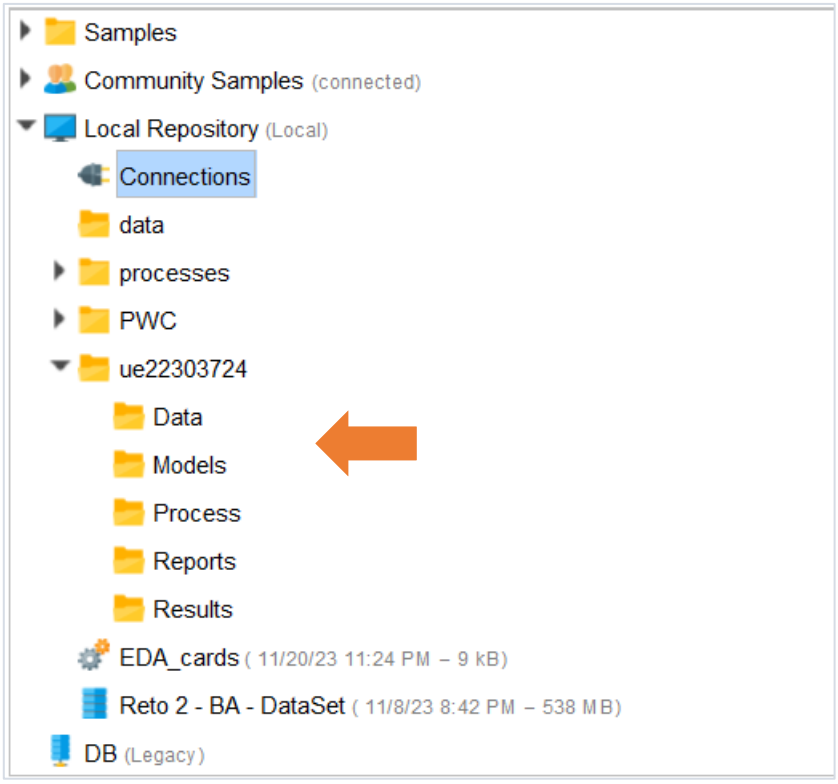
Tabla de contenido

Contenido

Objetivos de la actividad :

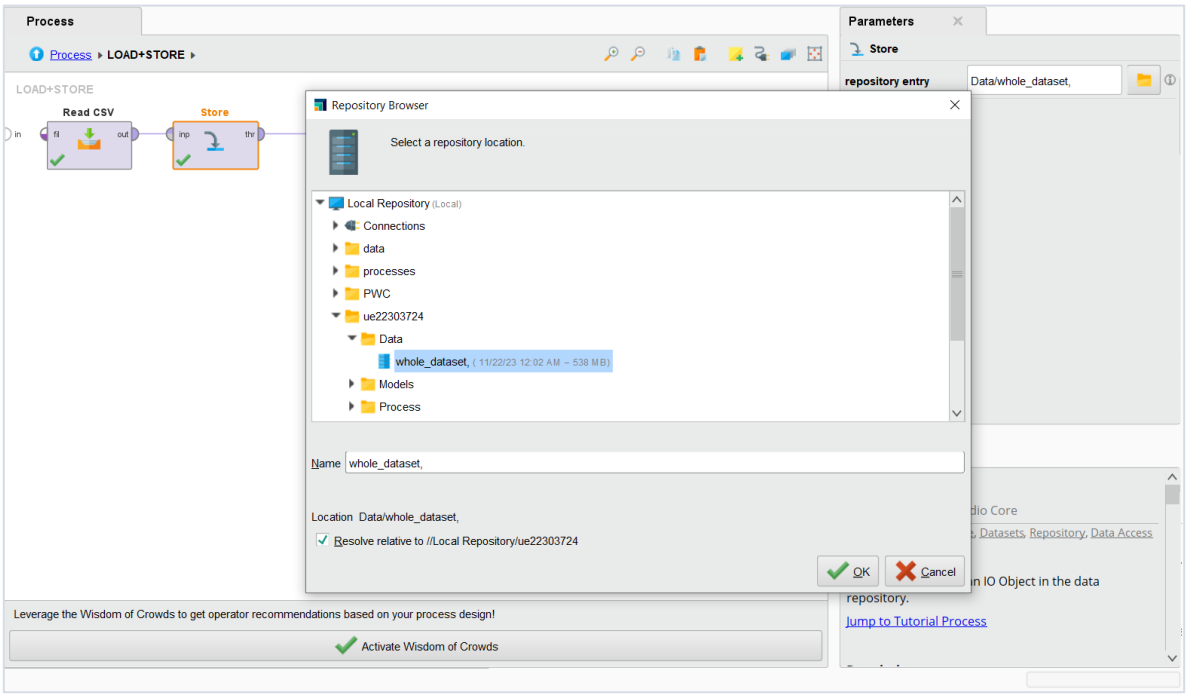
1. Conocer cómo se realiza un ETL + EDA + Modeling
2. Aplicar la plataforma RapidMiner para resolver preguntas de negocio.
3. Entender cómo funcionan los modelos y elegir el de mejor rendimiento
4. Generar el documento que deberá ser subido al Canvas

Task 00 – Definir la estructura del repositorio

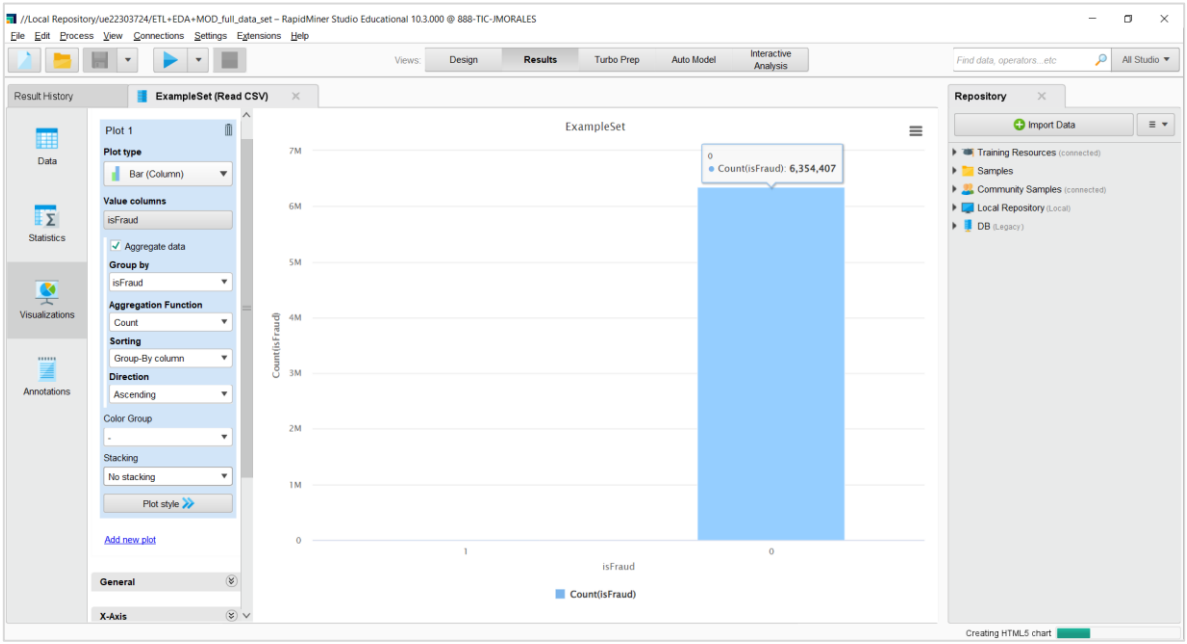


Task 01 Lectura de full dataset

Se realiza la lectura del dataset completo y se almacena en la carpeta data con el nombre ***whole_dataset.***



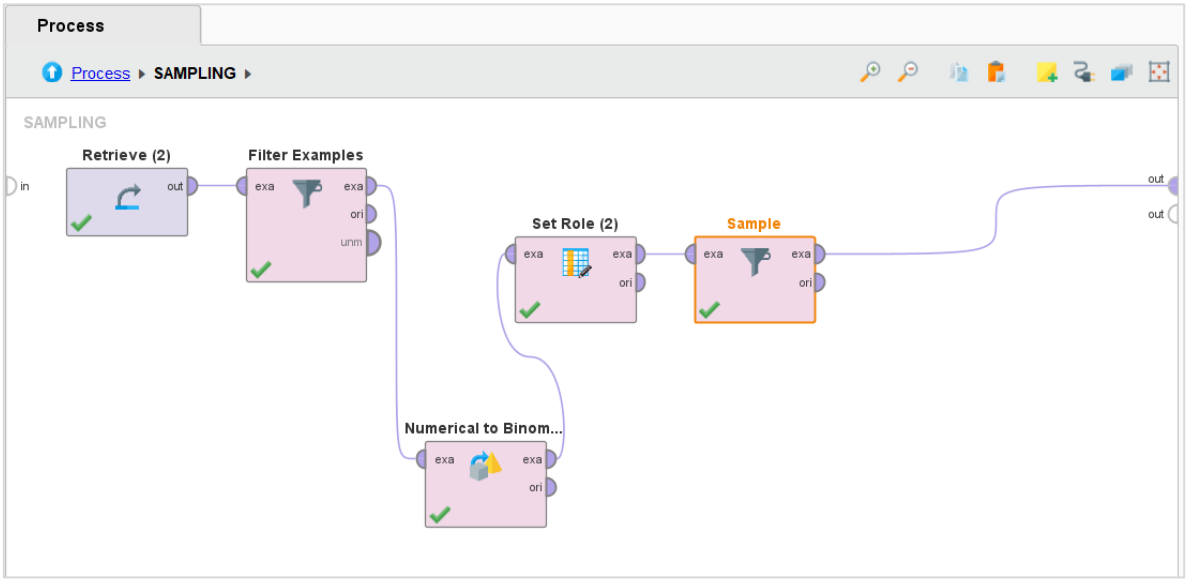
Task 02 – Balanceo del dataset



Se hace el balanceo del full dataset siguiendo los siguientes pasos:

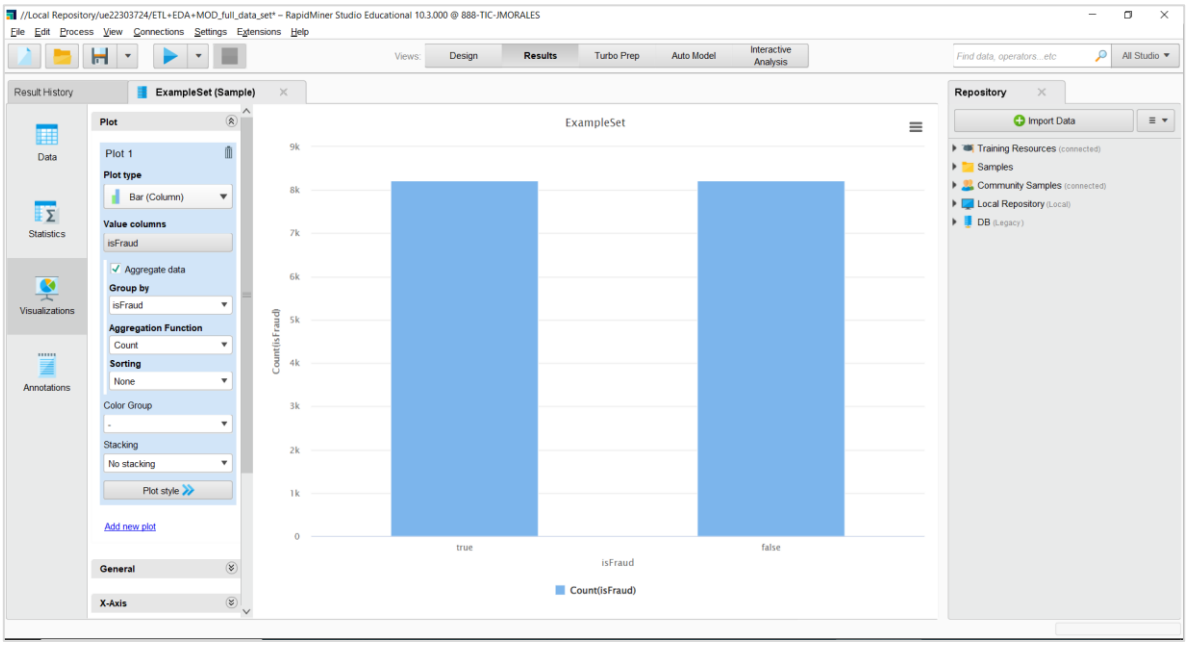
- Mediante retrieve leemos el dataset completo

- Aplicamos un operador de filtro para que tenga en cuenta todos los valores de la variable **isFraud**
- Convertimos el tipo de dato de numerical a binomial
- Fijamos la variable **isFraud** como rol para el muestreo
- Balanceamos la muestra tomando como base el número de observaciones marcadas como true en la variable **isfraud.**

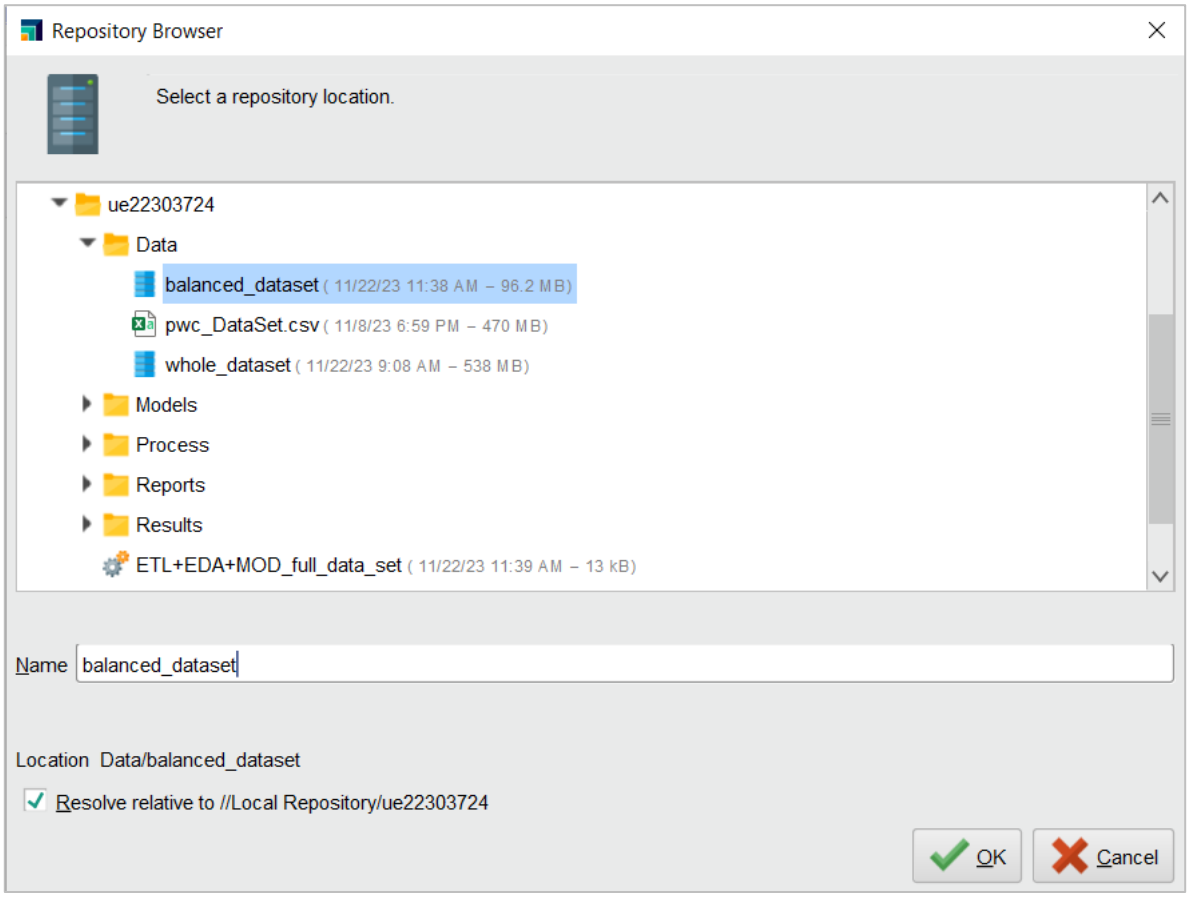


Task 03 – dataset balanceado

Obtenemos el dataset balanceado



Luego de esto teneos dos archivos, el **whole_datset** y el **balanced_dataset**

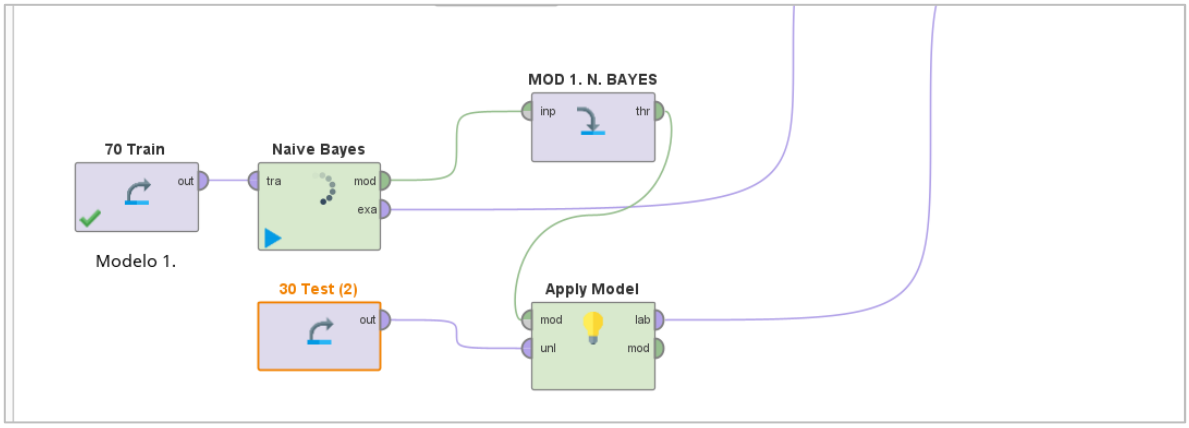


Task 04 prueba y análisis de modelo

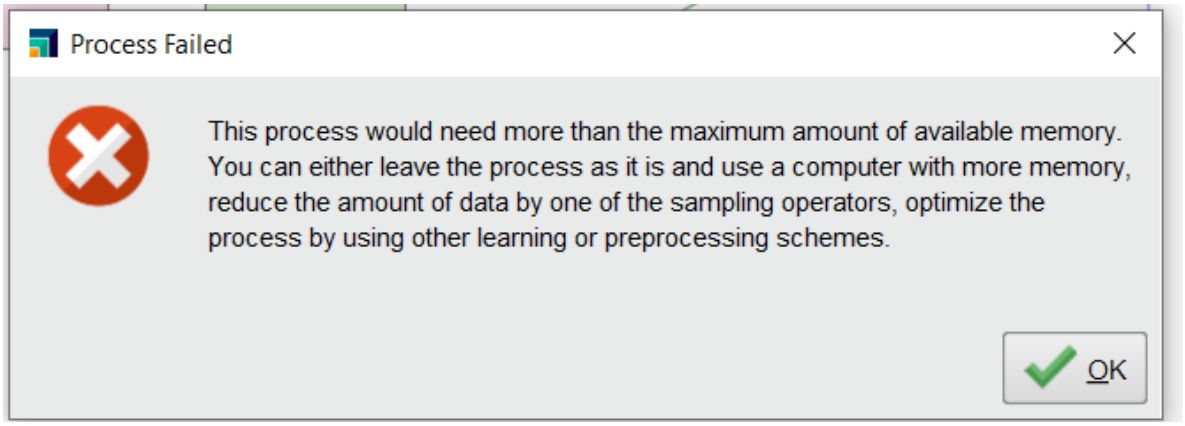
Pruebe al menos **tres modelos** y elija el de mejor rendimiento.

Modelo 1.

Para el primero modelo aplicamos el modelo probabilístico de clasificación de Naive Bayes



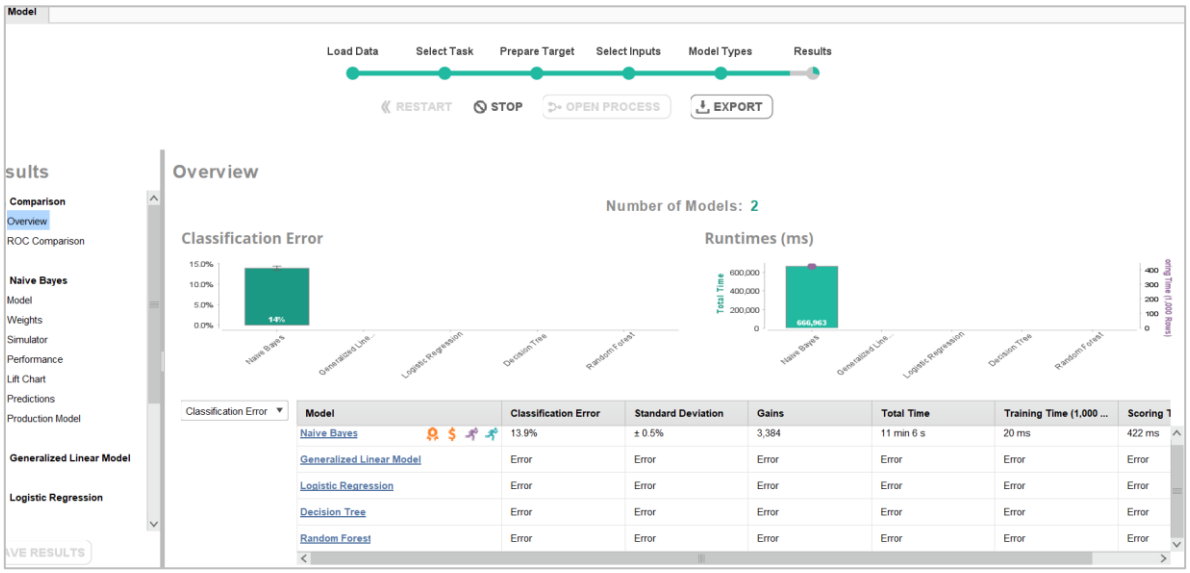
Dado a los recursos de memoria limitados del equipo local no se puede ejecutar el modelo, ya que arroja el siguiente error.



El mismo error se presenta al momento de ejecutar el modelo **K-NN** y **Arbol de decisión.**

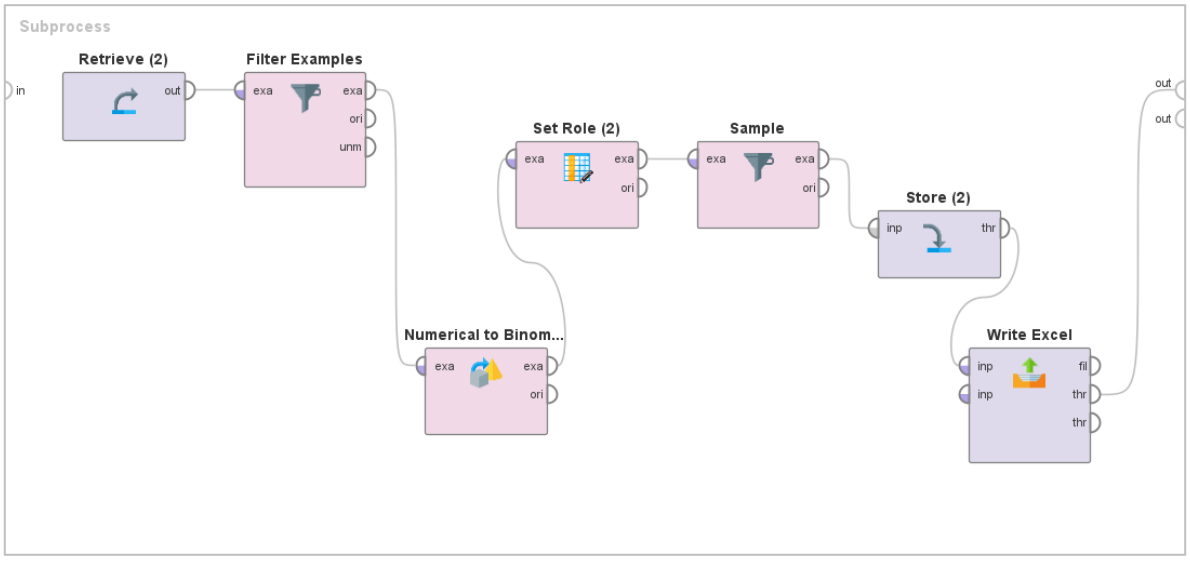
Debido a este error se procedió con la opción **Automodel,** mediante la cual rapidminer analiza cual es el modelo que mas se adapta o mejor precisión tiene.

Tomamos el archivo del dataset balanceado y ejecutamos el proceso.



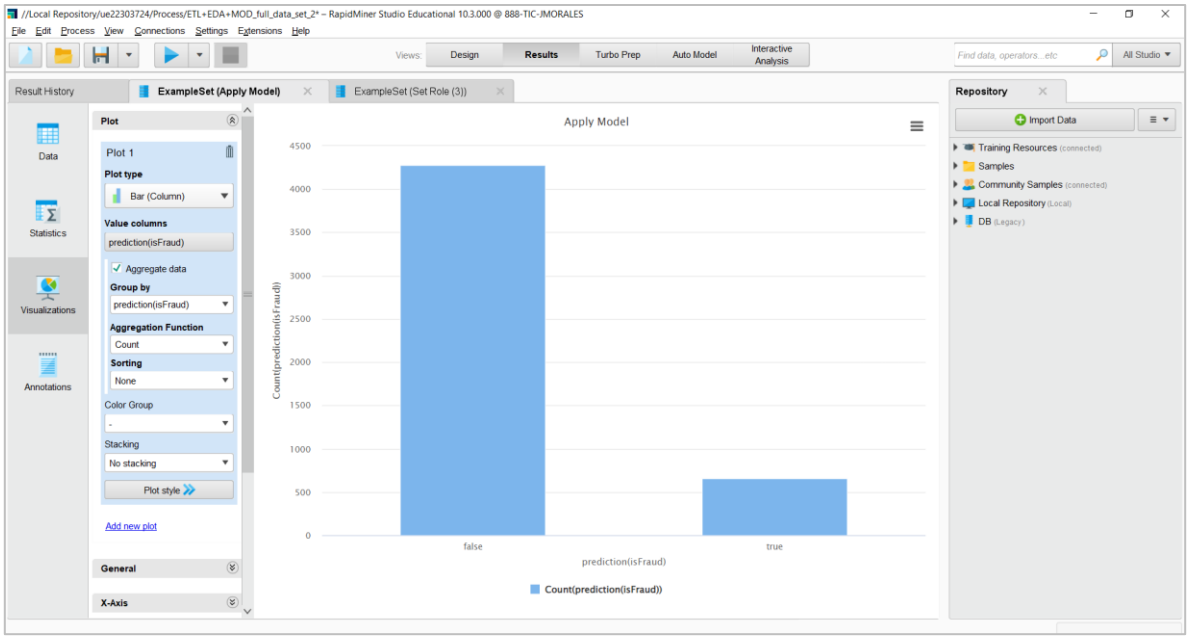
Luego de mas de 3 horas de espera no se obtiene respuesta, lo anterior debido a los recursos que consume rapidminer y que no se tiene la capacidad local para ejecutarlos. Por lo tanto no se tiene una conclusión de cual modelo es el más preciso o adecuado.

Para subsanar lo anterior se guardó el dataset balanceado en formato csv usando el operador writing, de tal modo que tuviera menor tamaño y pudiera ser leído.



Con el modelo Bayes se obtiene el siguiente resultado

Name	Type	Missing	Statistics		
Prediction **prediction(isFraud)**	Polynominal	0	Least: true (654)	Most: false (4274)	Values: false (4274), true (654)
Confidence_true **confidence(true)**	Real	0	Min: 0	Max: 1	Average: 0.155
Confidence_false **confidence(false)**	Real	0	Min: 0	Max: 1	Average: 0.845
isFraud	Polynominal	0	Least: false (2421)	Most: true (2507)	Values: true (2507), false (2421)



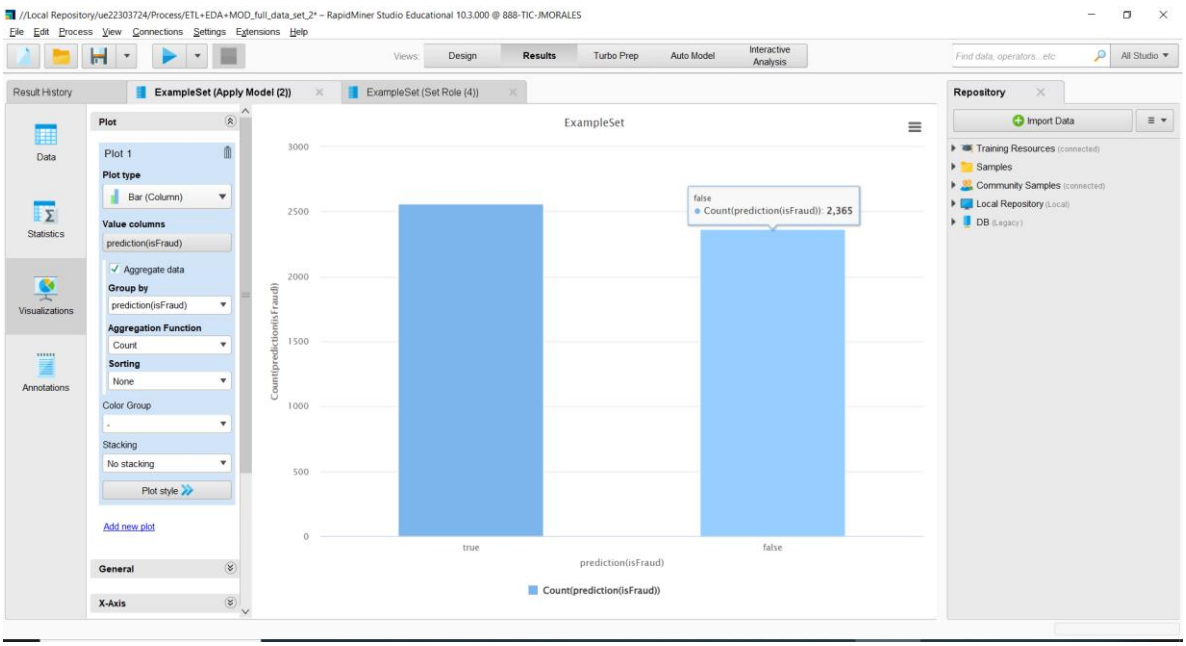
Nominal values

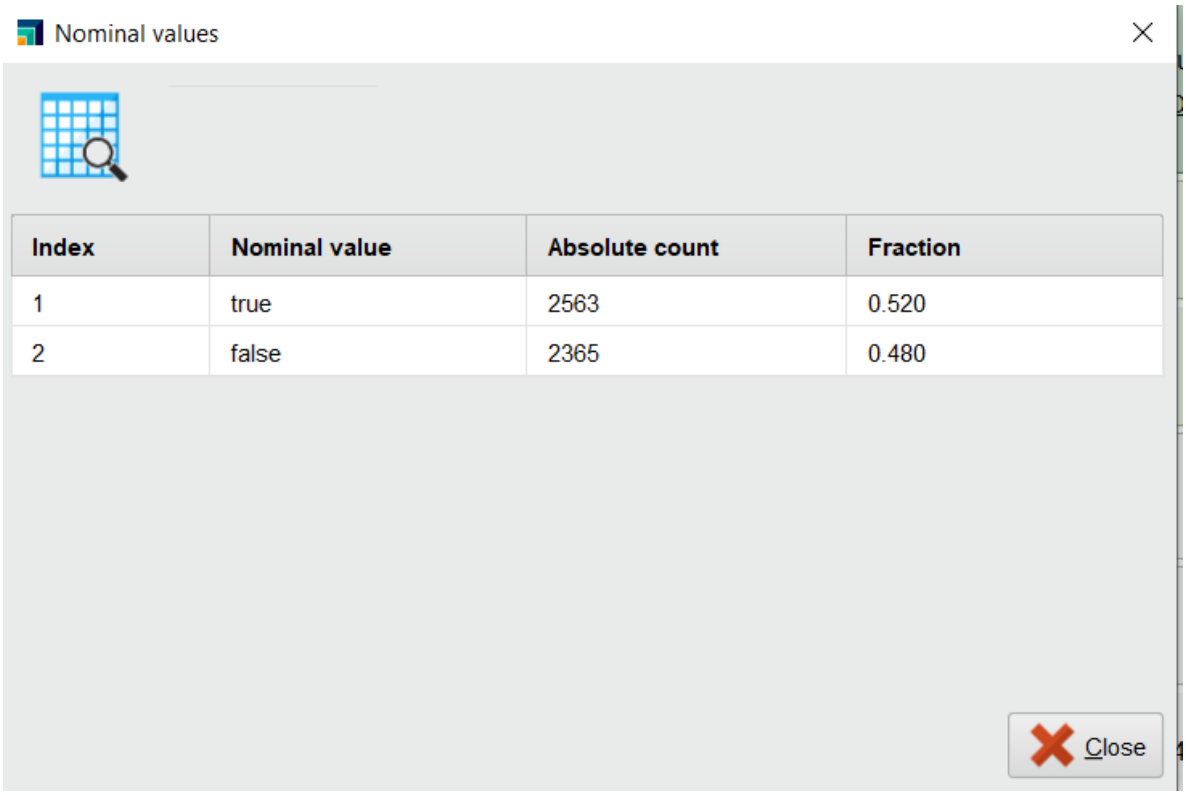
Index	Nominal value	Absolute count	Fraction
1	false	4274	0.867
2	true	654	0.133

Close

Usando el modelo **KNN** se obtienen los siguientes resultados

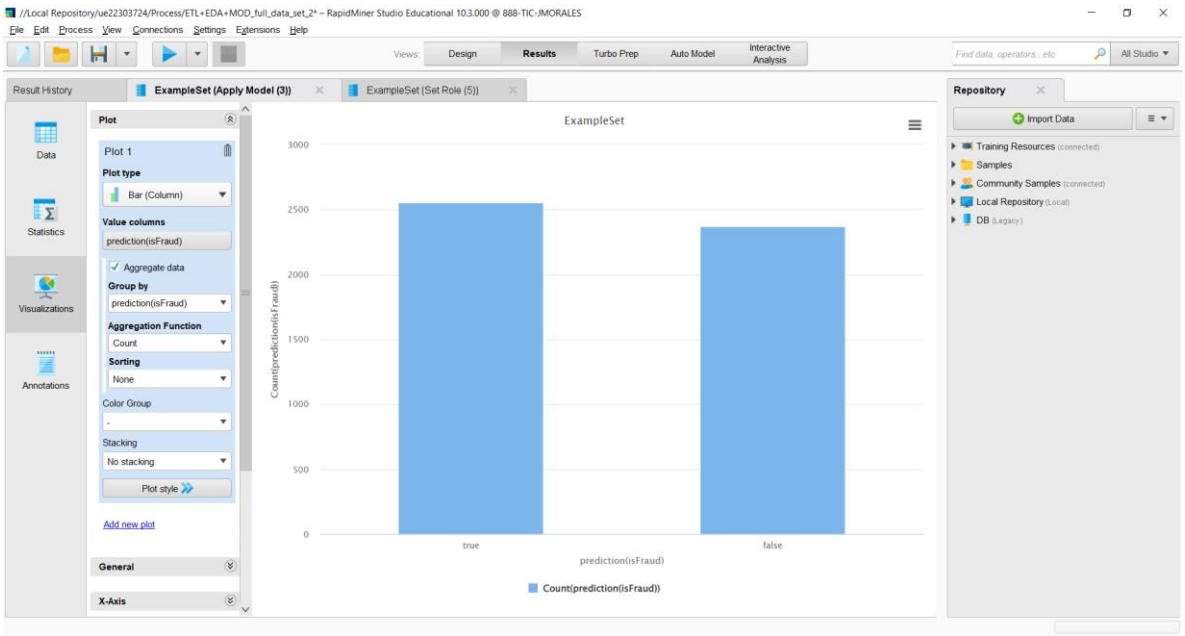
Name	Type	Missing	Statistics		Filter (14 / 14 attributes): Search for Attributes	
Prediction **prediction(isFraud)**	Polynominal	0	Least false (2365)	Most true (2563)	Values true (2563), false (2365)	
Confidence_true **confidence(true)**	Real	0	Min 0	Max 1.000	Average 0.519	
Confidence_false **confidence(false)**	Real	0	Min 0	Max 1.000	Average 0.481	





Por ultimo se usó el modelo **random forest** obteniendo lo siguientes resultados

Name	Type	Missing	Statistics		Filter (14 / 14 attributes): Search for Attributes
Prediction **prediction(isFraud)**	Polynominal	0	Least false (2371)	Most true (2557)	Values true (2557), false (2371)
Confidence_true **confidence(true)**	Real	0	Min 0	Max 1	Average 0.509
Confidence_false **confidence(false)**	Real	0	Min 0	Max 1	Average 0.491



Nominal values

Index	Nominal value	Absolute count	Fraction
1	true	2557	0.519
2	false	2371	0.481

Close

Se concluye que el modelo knn es el de mas alta precisión de los 3 usados, sin embargo esta muy distante del objetivo mínimo el cual debe estar por encima del 97%