

Declaración de trabajo original (no plagio) del estudiante

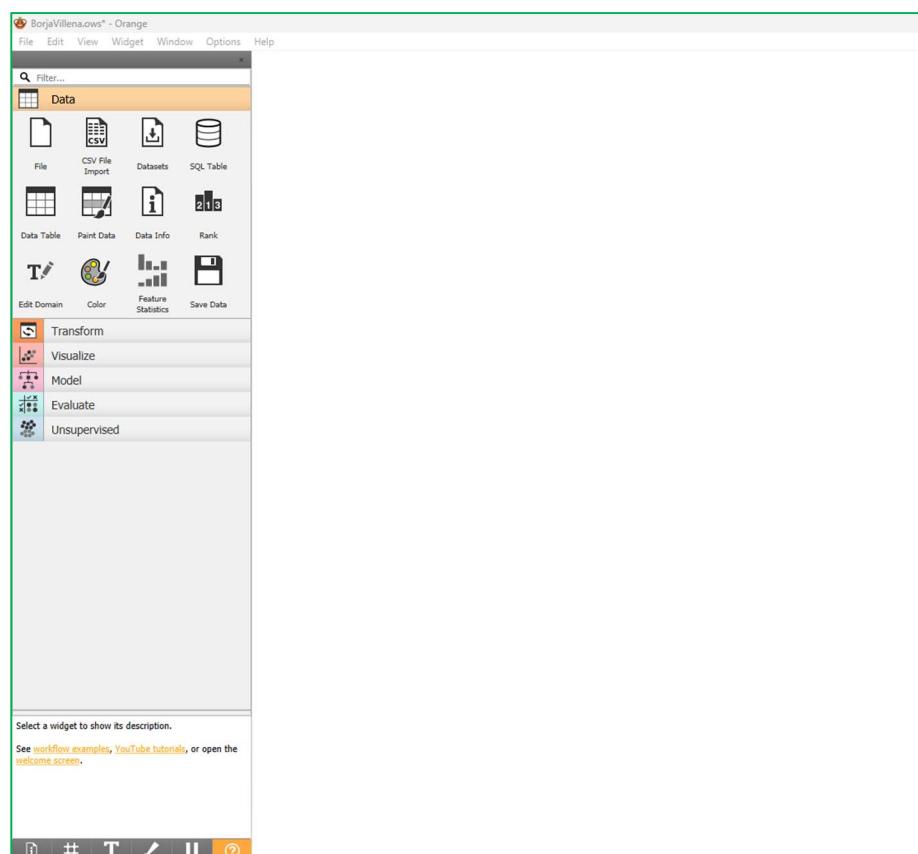
Yo, **Borja Villena Pardo**, declaro que para hacer esta entrega me he basado en mi propio trabajo original y no he realizado acciones que puedan ser consideradas copia o plagio.

Algunas de las fuentes consultadas han sido:

- [https://es.wikipedia.org/wiki/%C3%81rea_bajo_la_curva_\(farmacolog%C3%ADa\)](https://es.wikipedia.org/wiki/%C3%81rea_bajo_la_curva_(farmacolog%C3%ADa))
- <https://www.themachinelearners.com/metricas-de-clasificacion/>
- <https://megharamesh.medium.com>
- <https://www.kaggle.com/datasets/shriyashjagtap/indian-personal-finance-and-spending-habits>
- <https://orangedatamining.com/docs/>
- Temario Asignatura UOC – Análisis Multivariante

Pregunta 1 (35%)

1.1.



1.2.A.

- Fuentes consultadas: *Temario Asignatura Análisis Multivariante UOC*

Tal y como podemos observar en la propia descripción del archivo, se trata de un conjunto de datos que nos facilita información financiera y demográfica detallada sobre 20.000 personas en la India.

Los datos facilitados nos permiten encontrar patrones en las finanzas personales, identificar áreas para la optimización de costos y construir modelos predictivos para la gestión financiera. De esta manera podríamos responder a preguntas relacionadas con los hábitos de gasto o con el potencial de ahorro en el mercado indio.

A continuación, vamos a describir cuatro atributos, uno por cada categoría:

Income & Demographics:

- 1) El atributo **{AGE}** recoge variables independientes tipo *integer*, las cuales representan la edad de los individuos que forman la muestra.

Las variables que recoge este atributo nos permiten por ejemplo aplicarlas en una regresión múltiple, en un análisis de clústeres o en un análisis de factores.

Monthly Expenses:

- 2) El atributo **{RENT}** recoge variables dependientes tipo *float*, las cuales representan el gasto mensual en alquiler de cada individuo que forma la muestra. Estos datos nos ayudan a recoger información sobre cómo afecta el costo de la vivienda al presupuesto general de cada individuo.

Estas variables nos permiten realizar por ejemplo una regresión múltiple o un análisis de componentes principales para identificar la importancia relativa del alquiler dentro de un perfil general de gastos.

Financial Goals & Savings:

- 3) El atributo **{DESIRED_SAVINGS_PERCENTAGE}** recoge variables tipo *float*, las cuales pueden ser dependientes o independientes dependiendo del enfoque que le demos a nuestro análisis. Este atributo nos permite identificar como los ingresos, los gastos y las variables demográficas pueden afectar a los objetivos de ahorro que puede plantearse cada individuo.

Las variables de este atributo nos permiten realizar análisis como la regresión múltiple o el análisis de clústeres, el cual nos permite agrupar a los individuos según objetivos de ahorro similares.

Potential Savings:

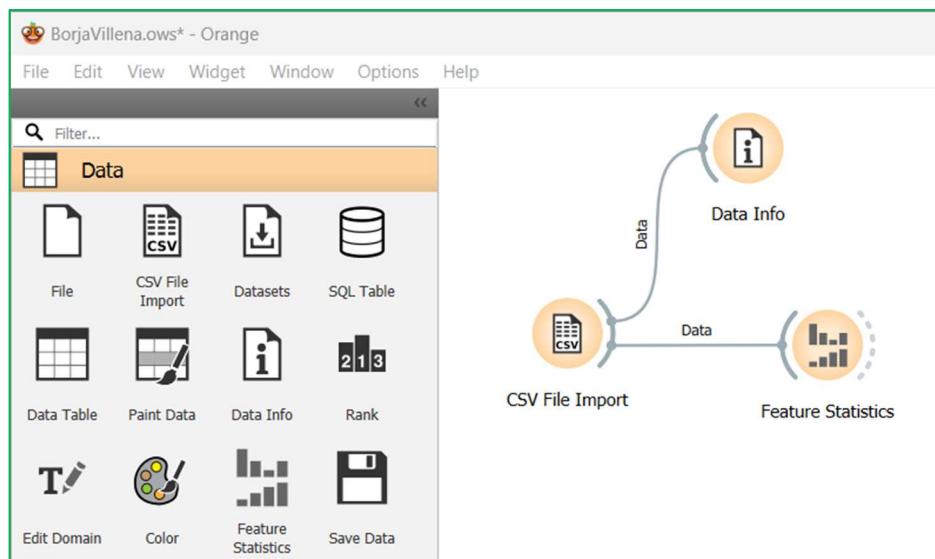
- 4) El atributo **{GROCERIES}** recoge variables dependientes tipo *float*. Este atributo nos facilita información de cada individuo sobre el gasto en alimentos. Esta información puede ser utilizada por ejemplo para el análisis sobre la capacidad de ahorro en comestibles.

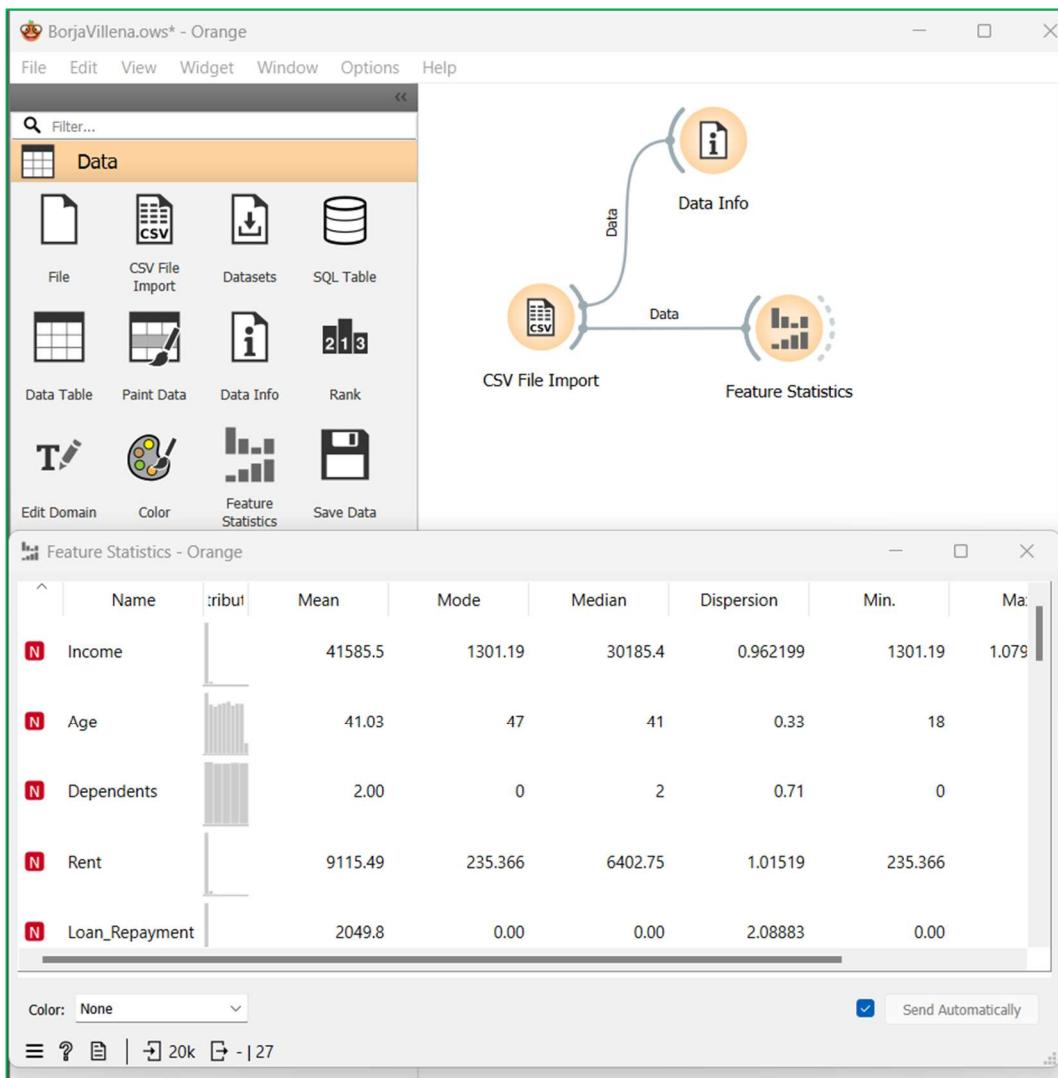
Las variables de este atributo las podríamos utilizar para realizar una regresión múltiple, la cual nos daría información de cómo afectan otros atributos al potencial de ahorro en la compra de alimentos.

1.2.B.

- Fuentes consultadas:

- *Help - Orange3*
- <https://orangedatamining.com/docs/>



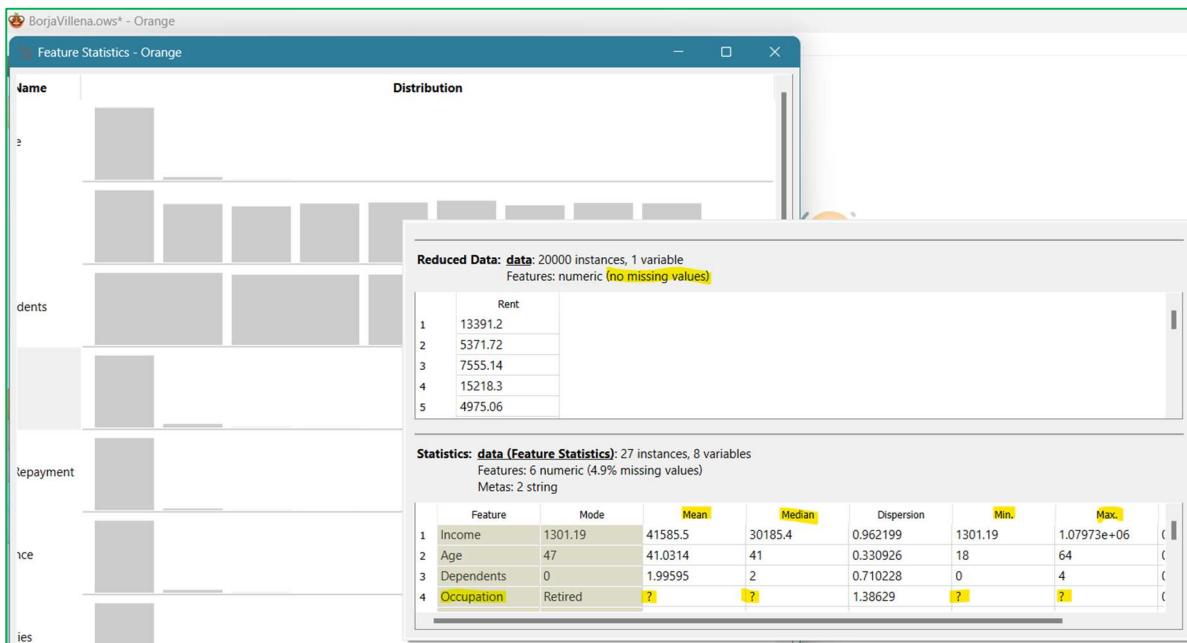


La ventana “Feature Statistics” nos facilita de forma fácil que los atributos categóricos son:

- 5) Occupation
- 6) City_Tier

El resto de atributos del archivo son de tipo numéricos.

Como podemos observar en la siguiente captura, la herramienta nos facilita un resumen de nuestros datos, en los cuales podemos observar que no hay valores que falten en los atributos del dataset:



Analizando las estadísticas, podemos observar que el atributo “Occupation” al tratarse de una variable categórica, no es posible determinar valores de media, mediana, mínimo o máximo. Al igual ocurre con la otra variable categórica del dataset: “Citi_Tier”.

Observando los valores de dispersión de ambos atributos, siendo “Occupation” = 1,39 y “Citi_Tier” = 1,03, podríamos decir que el valor de la dispersión no es elevado al ser menor de 1,5, esto nos podría dar una primera impresión de que la mayoría de los datos de estos atributos están agrupados en una o dos categorías.

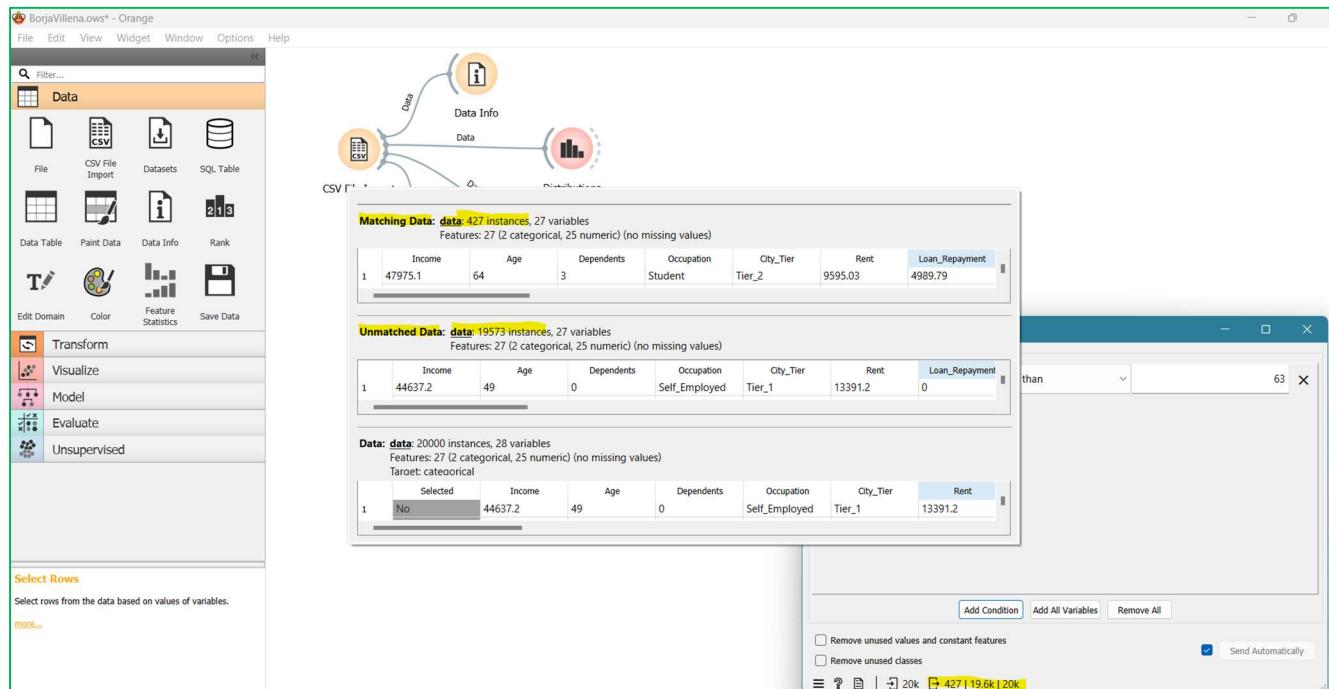
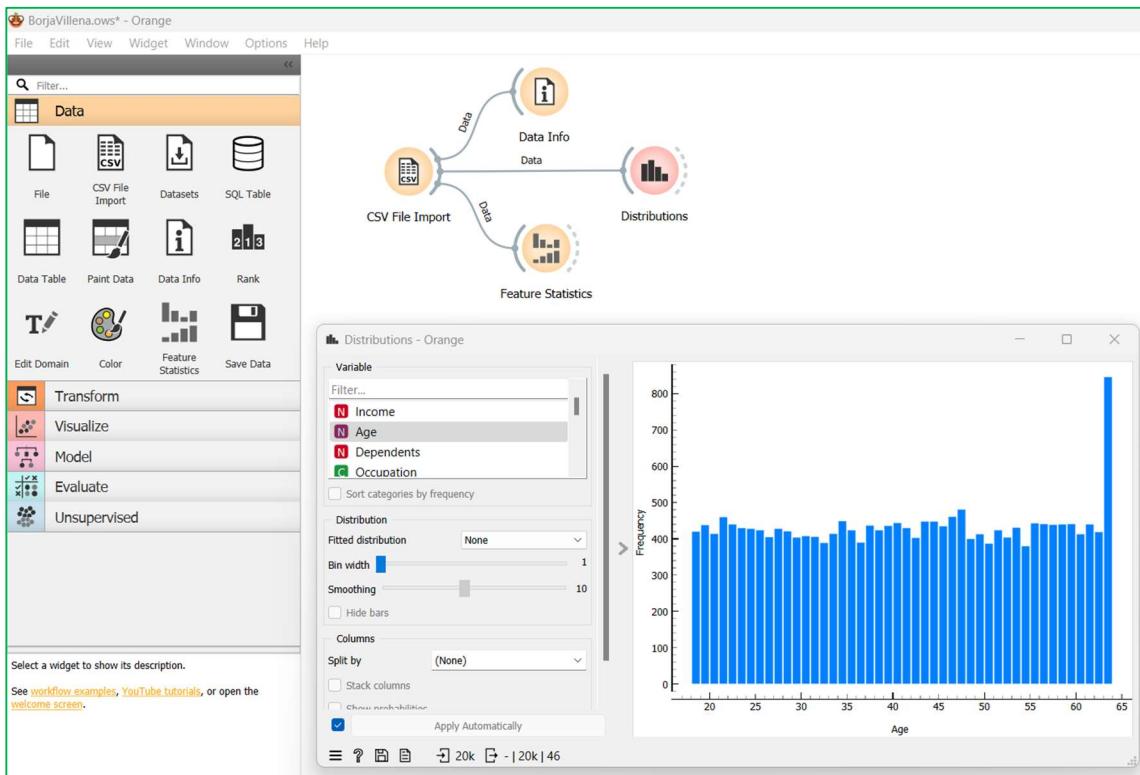
Por otro lado, analizando los atributos numéricos “Age” y “Rent”, podemos observar como la media de edad de los individuos que forman la muestra es de 40 años y que la dispersión es muy baja, siendo del 0.33. A su vez la edad mínima de los individuos encuestados es de 18 años. Esto nos puede indicar que la franja de edad de la mayoría de los encuestados está entorno a los 40 años de edad, es decir, que por ejemplo en el caso de que quisieramos encontrar patrones o diseñar modelos de predicción para casos que engloben población juvenil de la India, este dataset no nos serviría.

Respecto al atributo “Rent”, intuimos que los datos se refieren en valores de INR (rupia india), en tal caso observamos que la diferencia entre el valor mínimo y máximo de renta es muy elevada, siendo la diferencia entre ambos de 215.711,00 INR aproximadamente, lo que equivale a unos 2.440,00 €. La media se sitúa en 9.115,49 INR con una dispersión de los datos de 1,01519 lo que nos puede indicar que no hay una alta variabilidad de los datos, por lo que la mayoría de precios de renta pueden que estén agrupados en uno o dos grupos.

Nuestra impresión la podemos corroborar al observar el histograma de distribución que nos devuelve Orange3, ya que hay una columna que predomina frente a las demás.

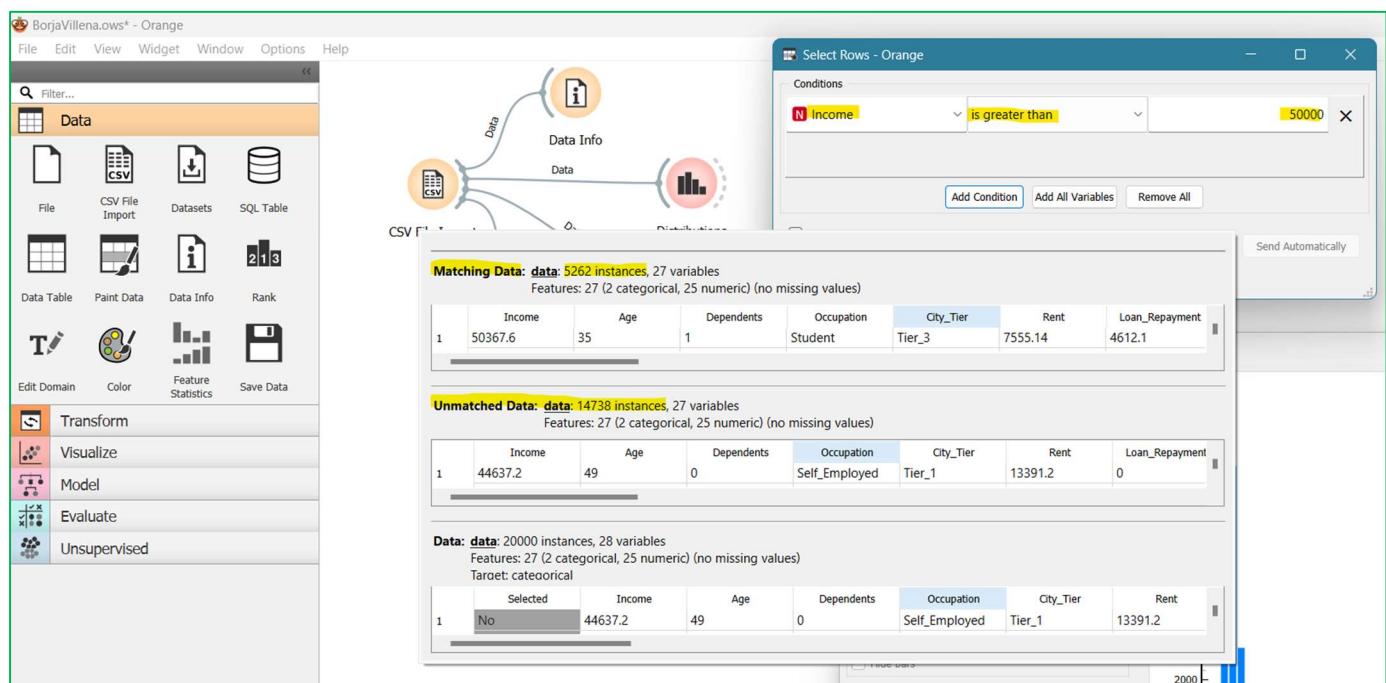
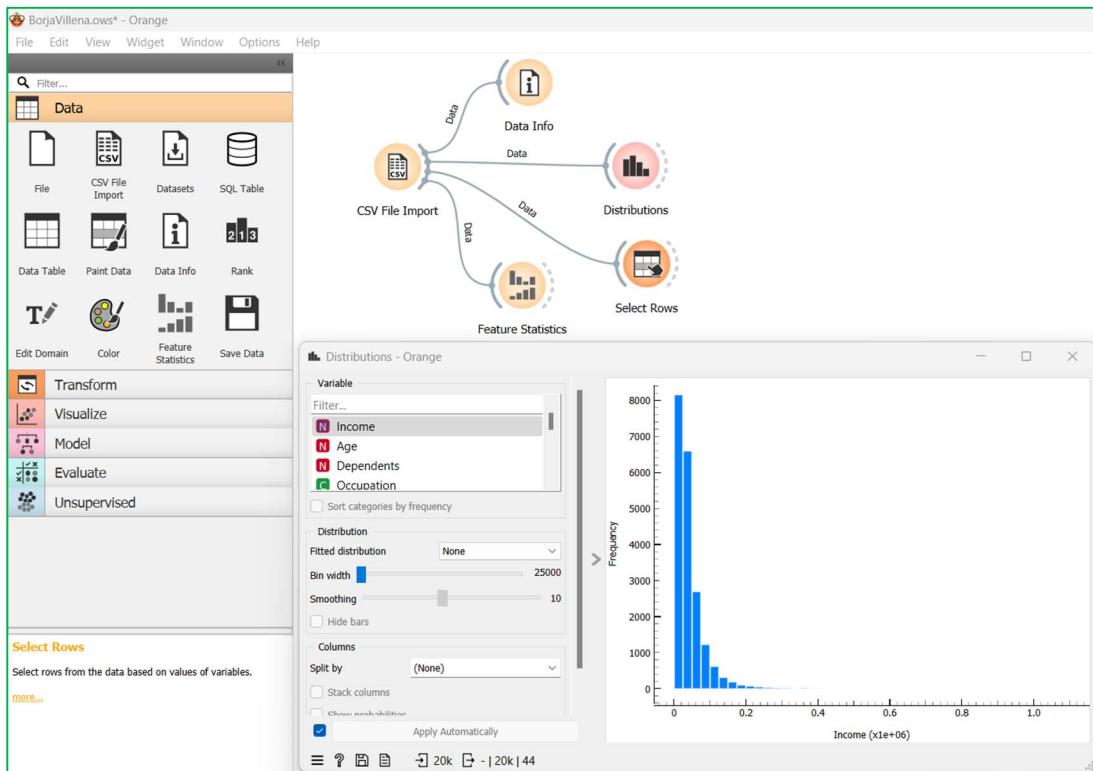
1.3.A.

Variable “AGE”:

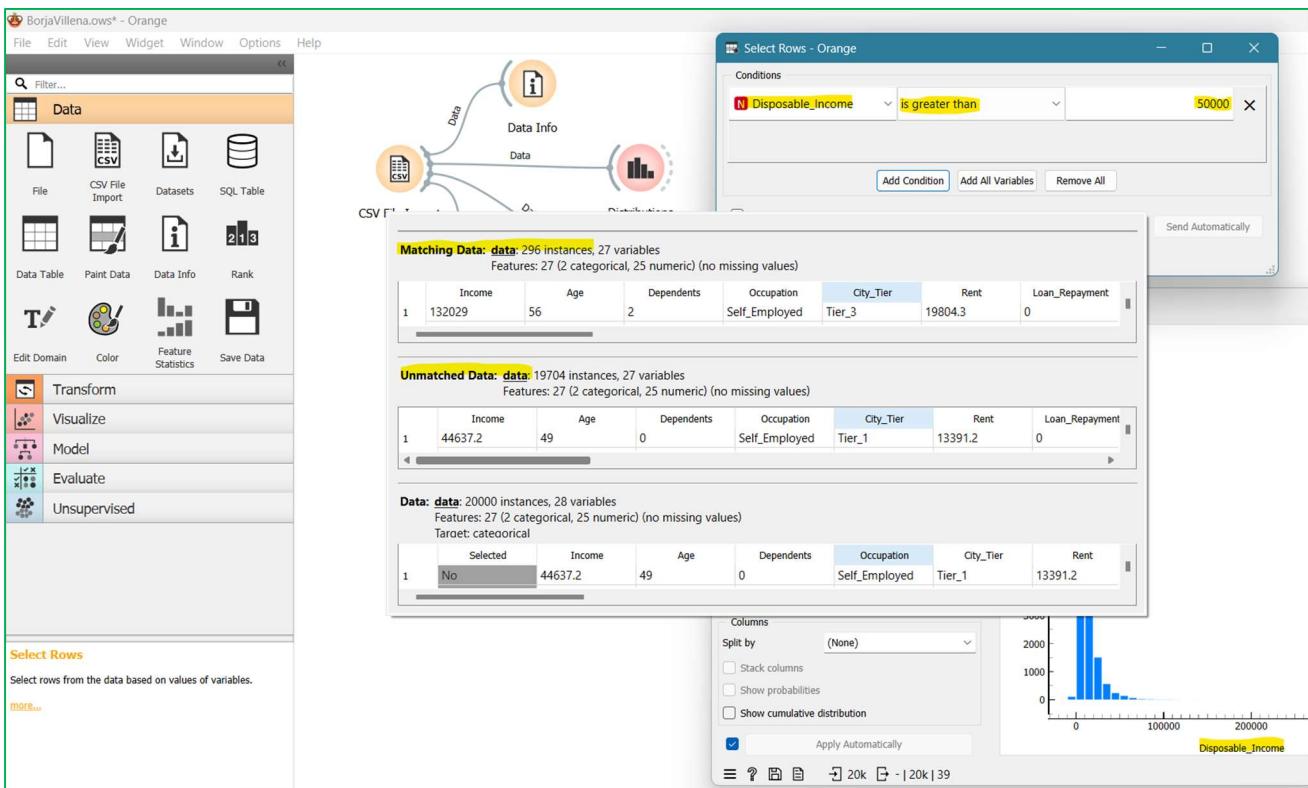
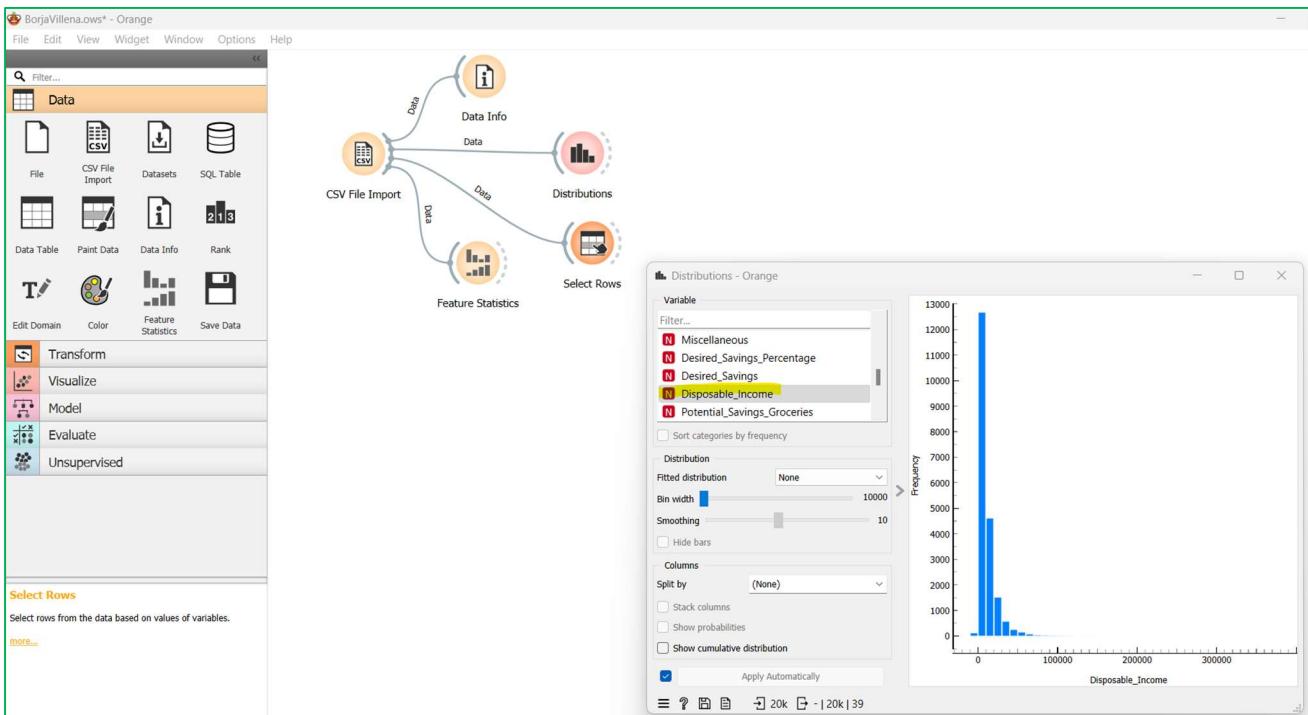


1.3.B.

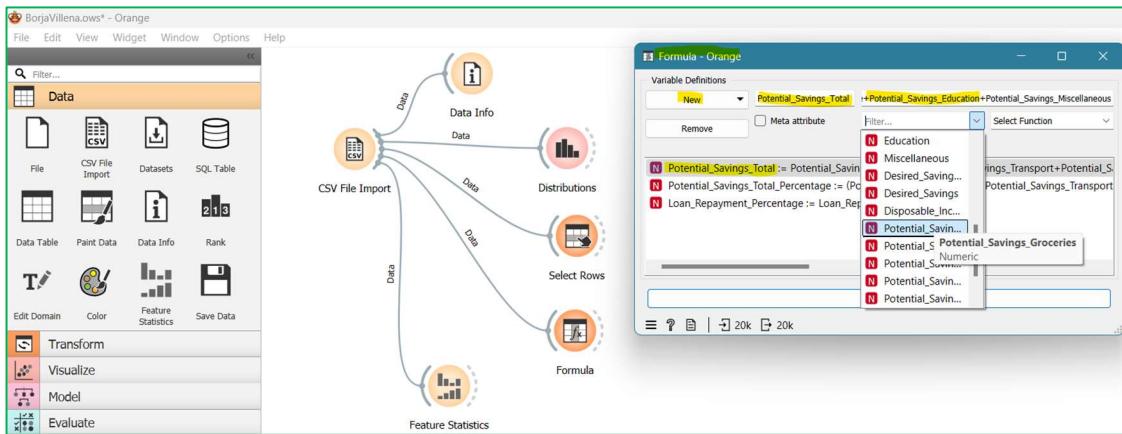
a) Variable "INCOME". Realizamos filtro para valores mayores que 50.000,00 INR.



b) Variable “DISPOSABLE_INCOME”. Realizamos filtro para valores mayores que 50.000,00 INR.

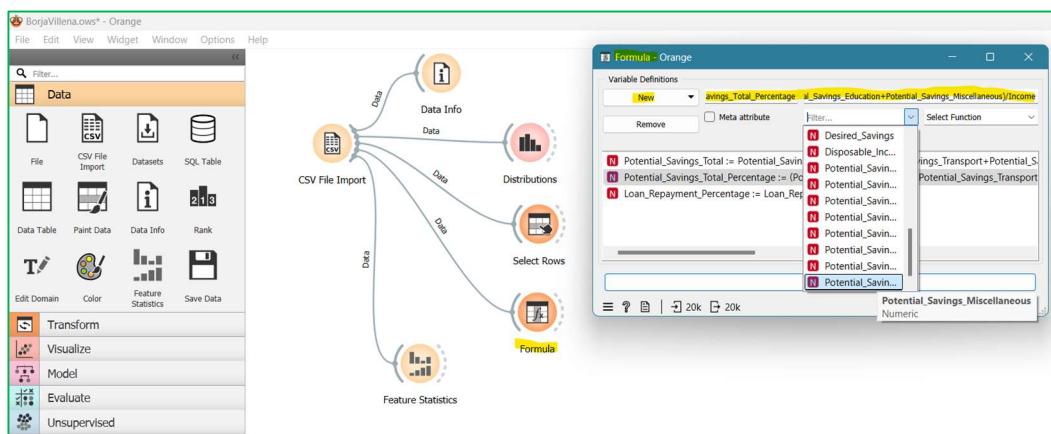


1.3.C.

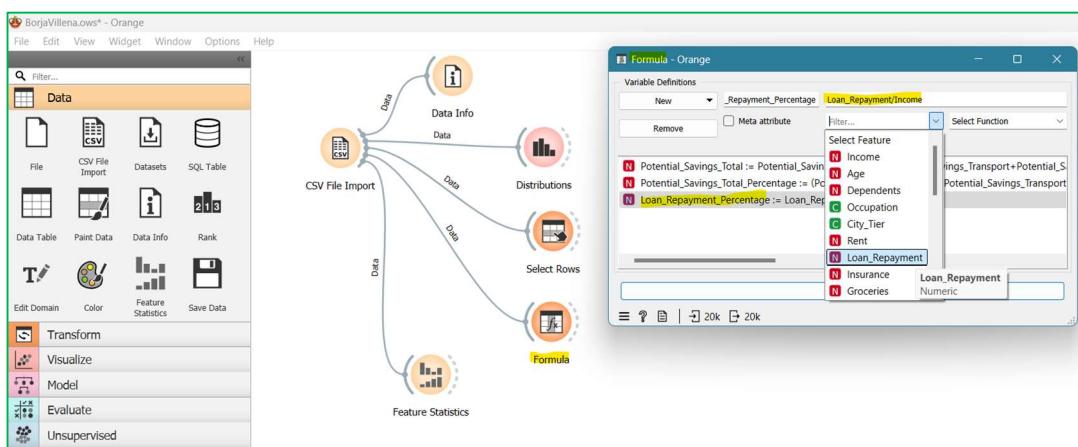


1.3.D.

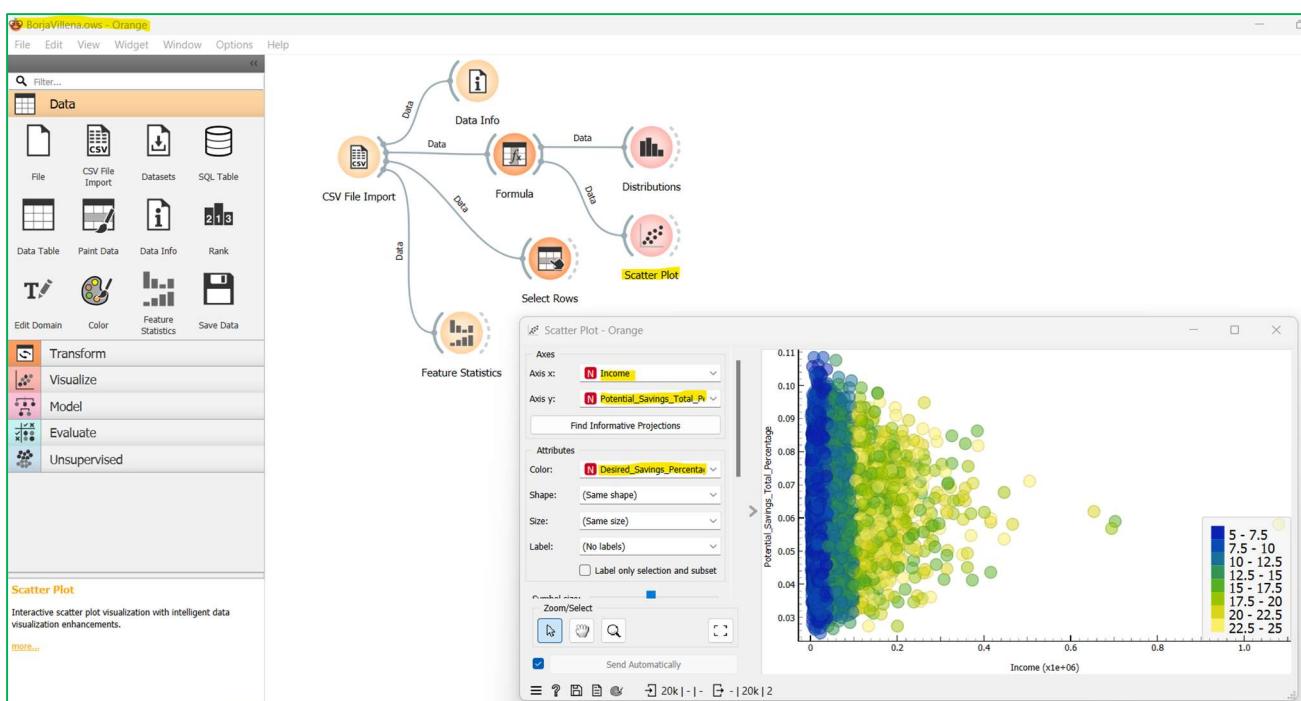
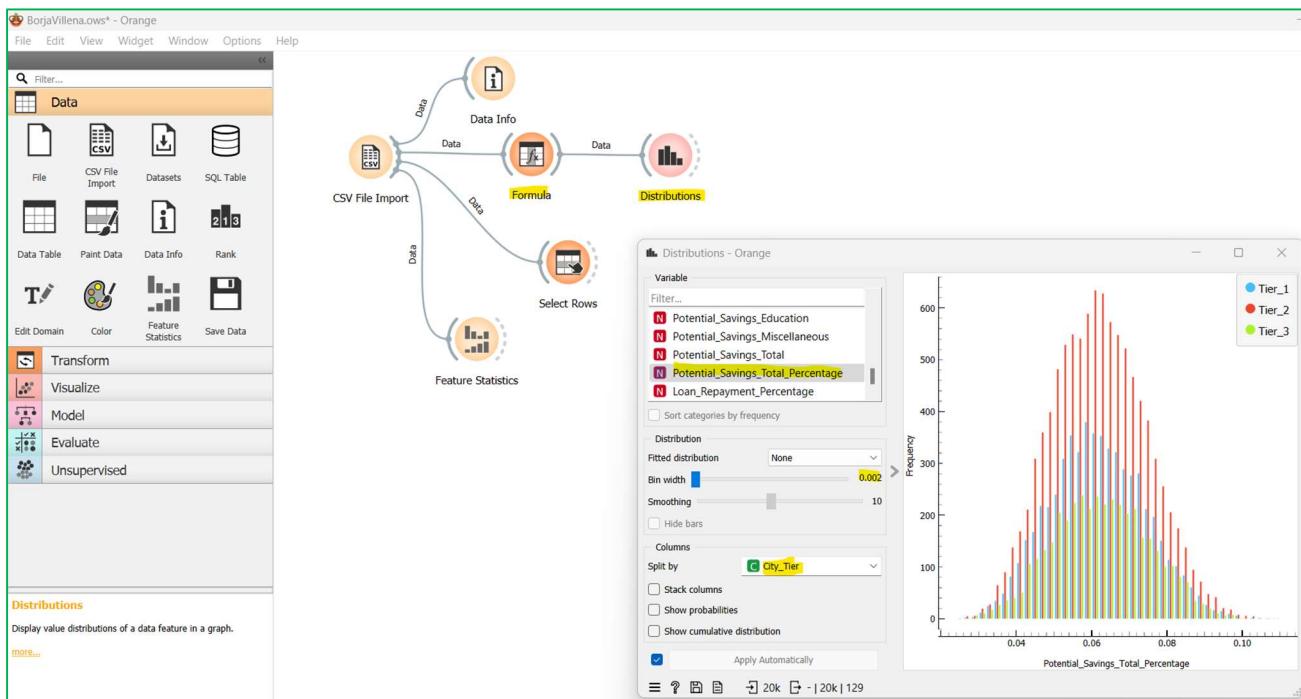
a) Potential_Savings_Total_Percentage:



b) Loan_Repayment_Percentage:



1.3.E.



Pregunta 2 (30%)

2.1. (campo de interés)

Data journalism. La descripción literal que nos facilita el temario es:

"El periodismo de datos es la versión narrativa de la ciencia de datos, es decir, se encarga de explicar historias de relevancia informativa mediante datos y basadas en datos. La inmensa mayoría de las veces se complementa con una visualización de datos impactante y clara."

2.2. (cómo se cataloga el artículo y por qué)

El artículo podría llegar a ser catalogado como Data journalism, ya que nos facilita información sobre patrones de gasto y ahorro en la India, aunque echamos en falta ciertos matices que consideramos de gran relevancia.

Bajo mi criterio, no lo consideraría un ejemplo de artículo de alta calidad, ya que carece de referencias a los datos en los que se basa para tales afirmaciones, desconozco la procedencia de los mismos e incluso echo en falta algún tipo de tabla, histograma o visualización gráfica que sustente los argumentos desarrollados por la escritora.

Sin el soporte visual de los datos, junto al desconocimiento del origen de los mismos, el lector puede tener la impresión de estar leyendo un artículo más bien sensacionalista, basado en opiniones subjetivas, sin sustento técnico que defienda tales afirmaciones.

La mención de la procedencia de los datos en los que se basa el artículo, acompañado de algún tipo de soporte visual, ayudaría en gran medida a evitar sesgos y reafirmar o desmentir tópicos sobre patrones de gasto y ahorro en la India, dándole un grado de mayor fiabilidad al artículo.

2.3. (análisis)

Estamos frente a un artículo periodístico que nos facilita información sobre patrones de gasto y de ahorro en el país indio.

Como podemos leer en el propio artículo al comienzo del mismo, la redactora nos menciona que basa el contenido de sus afirmaciones en la identificación de tendencias y prácticas detectadas por ella misma mediante una "cuidadosa observación" del comportamiento de la sociedad india, gracias a que ella nació y se crio en dicho país.

En este sentido, tal y como hemos comentado en el apartado 2.2 de esta PEC, echamos en falta la referencia y visualización de datos reales procedentes de alguna muestra que sustente las afirmaciones que se llevan a cabo en el artículo. De esta manera evitaríamos sesgos y podríamos reafirmar o desmentir tópicos sobre los temas que se tratan.

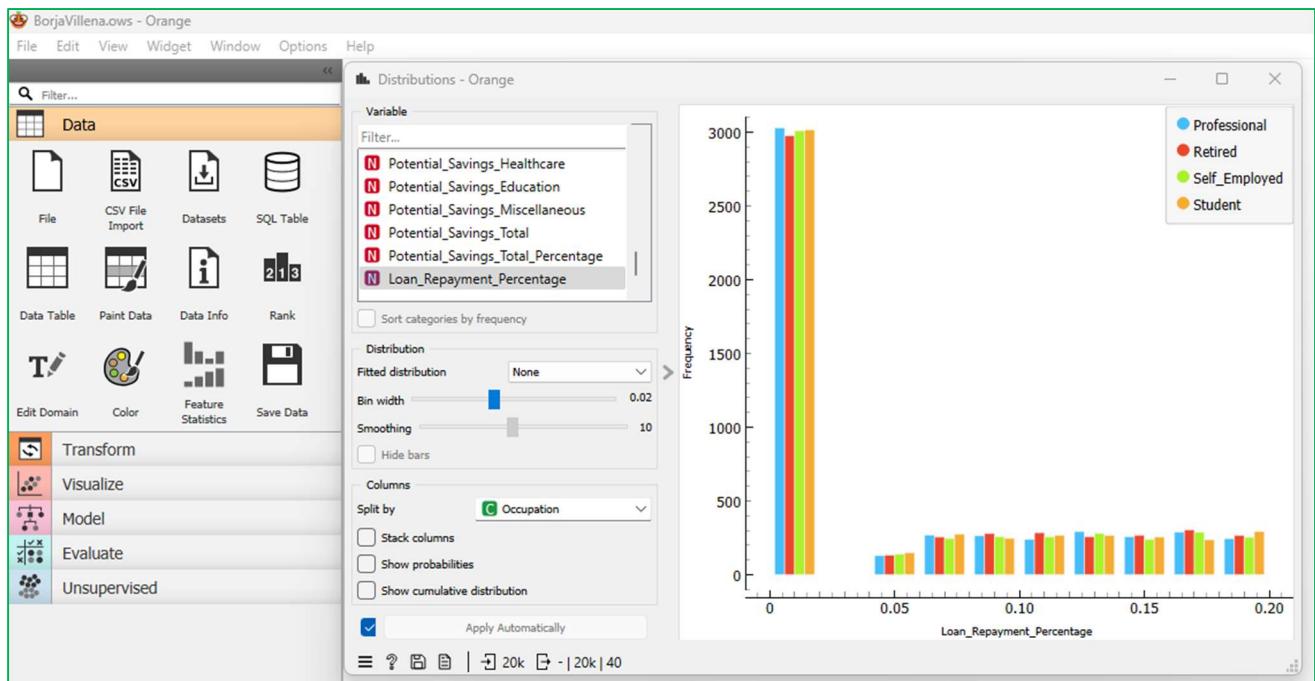
Algunas de las afirmaciones y conclusiones que trata el artículo son:

- 1) “Credit card usage in India has been growing steadily in recent years, but it is still relatively lower compared to some other countries.”

Aunque el uso de tarjetas de crédito esté creciendo en la India, la afirmación de que el uso de estos métodos de pago sigue siendo bajo respecto a otros países nos da una primera señal de que puede darse el caso de que los canales de préstamos más usados en la India sean canales informales alejados de los productos financieros más comunes.

Si nos fijamos en la distribución de la variable “*Loan_Repayment_Percentage*” en función de las ocupaciones de los individuos, podemos observar que las cuantías destinadas al pago de préstamos son muy bajas.

Esto nos sugiere que puede haber una dependencia excesiva de los canales de préstamo informales ya que los datos reflejan un porcentaje bajo de ingresos destinados al pago de deudas.

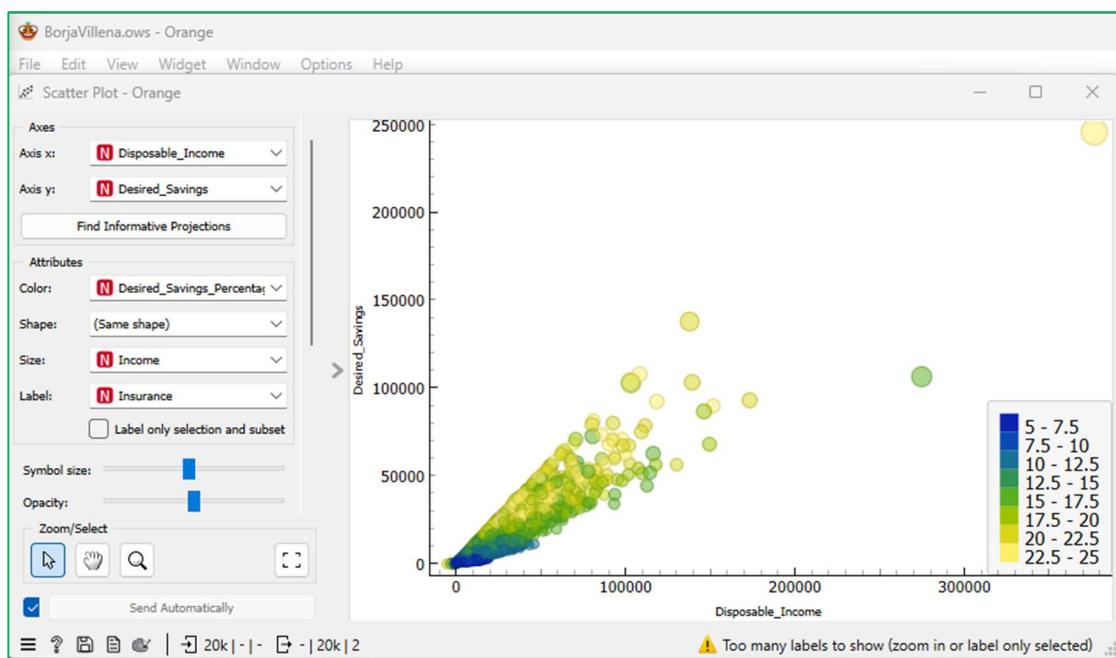


- 2) “The deeply ingrained savings mentality among Indians motivates them to place a high value on financial security, refrain from indulging in excessive expenses, and adopt responsible budgeting practices. This mindset finds its origins in the nation’s modest per capita income, which has fostered a culture of diligently monitoring expenses and optimizing savings.”

La afirmación sobre la mentalidad de ahorro profundamente arraigada, la cual les conlleva a abstenerse de gastos excesivos y sustentada en el modesto ingreso per cápita de la nación, puede ser discutida.

Si nos basamos en los datos que hemos trabajado anteriormente, fijándonos en el resultado del gráfico de dispersión que muestra la relación entre los ingresos y los porcentajes del potencial del ahorro en función de los porcentajes de ahorro deseados, podemos ver por ejemplo que los porcentajes de ahorro deseados aumentan con el aumento de los ingresos recibidos.

En el siguiente gráfico de dispersión hemos relacionado los ingresos que quedan después de contabilizar todos los gastos, con los deseos de ahorro en función de los porcentajes de ahorro deseados. Como podemos observar, claramente hay un deseo de ahorro mayor conforme aumentan los ingresos, y a su vez observamos que los colores más cálidos, los cuales pertenecen a porcentajes de ahorro deseado más elevados, son los colores que predominan en el gráfico. Por lo que podríamos interpretar que efectivamente, los individuos que pertenecen a la muestra si que tienden a tener una mentalidad de ahorro arraigada.



- 3) “*Indians have a long-standing tradition of reusing, repurposing, and recycling items. This is applicable to several areas.*”

Analizar esta afirmación con los datos que tenemos a mano resulta algo complejo, ya que la mayoría de variables disponibles corresponden con tipos de gasto sobre actividades, no sobre artículos u objetos. No obstante, vamos a analizar la variable “*Miscellaneous*” para intentar tener una primera impresión de si dicha afirmación puede ser rebatida o no.

Para analizar este caso, hemos usado un gráfico de dispersión donde hemos relacionado los ingresos, con los gastos de “*Miscellaneous*” en función de los porcentajes de ahorro totales.

Como podemos observar, los gastos en “*Miscellaneous*” son bajos respecto a los ingresos recibidos, lo que nos puede indicar que realmente si puede haber un interés social por la reutilización. A su vez, los colores que predominan en el gráfico corresponden a valores de potencial de ahorro medio-alto, esto fortalece la idea de que la sociedad india si que tienen una tradición de reutilización de objetos.

Por último, como ya hemos confirmado en apartados anteriores, esta relación nos vuelve a demostrar que aquellos individuos cuyos ingresos son menores, también gastan menos y su potencial de ahorro también es menor.



A lo largo del artículo creo que habría sido interesante encontrar alguna referencia respecto a los tipos de consumo y ahorro característicos de cada generación social.

Analizar si hay o no grandes diferencias entre la gente joven, la de mediana edad y los ancianos, puede ser muy interesante para poder encontrar patrones de cambios de mentalidad en el tipo de consumo y ahorro.

Por ejemplo, este tipo de información podría dar fuerza a afirmaciones realizadas en el artículo como: “*Credit card usage in India has been growing steadily in recent years*”. Si analizáramos en profundidad qué rango de edades son las más comunes en el uso de tarjetas de crédito, por ejemplo, podríamos prever un futuro incremento exponencial de estas prácticas de pago, en el caso de que los datos nos sugieran un alto uso por parte de las edades más tempranas.

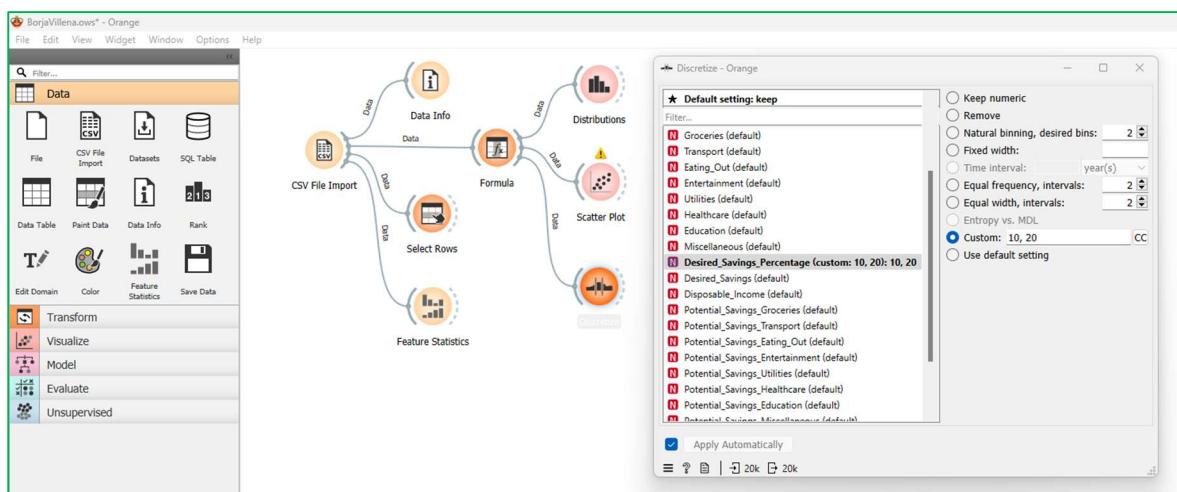
2.4. (otro campo de interés)

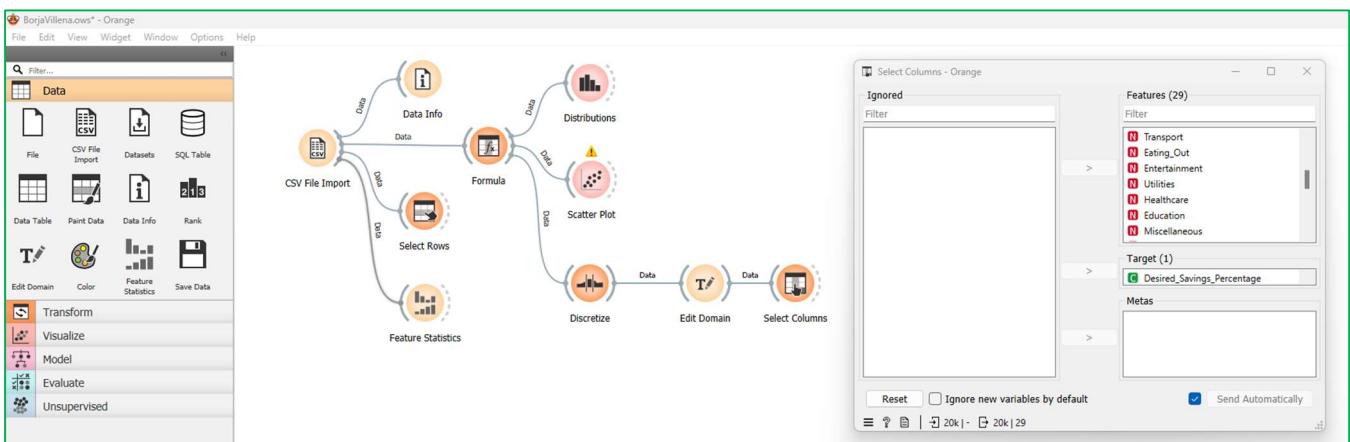
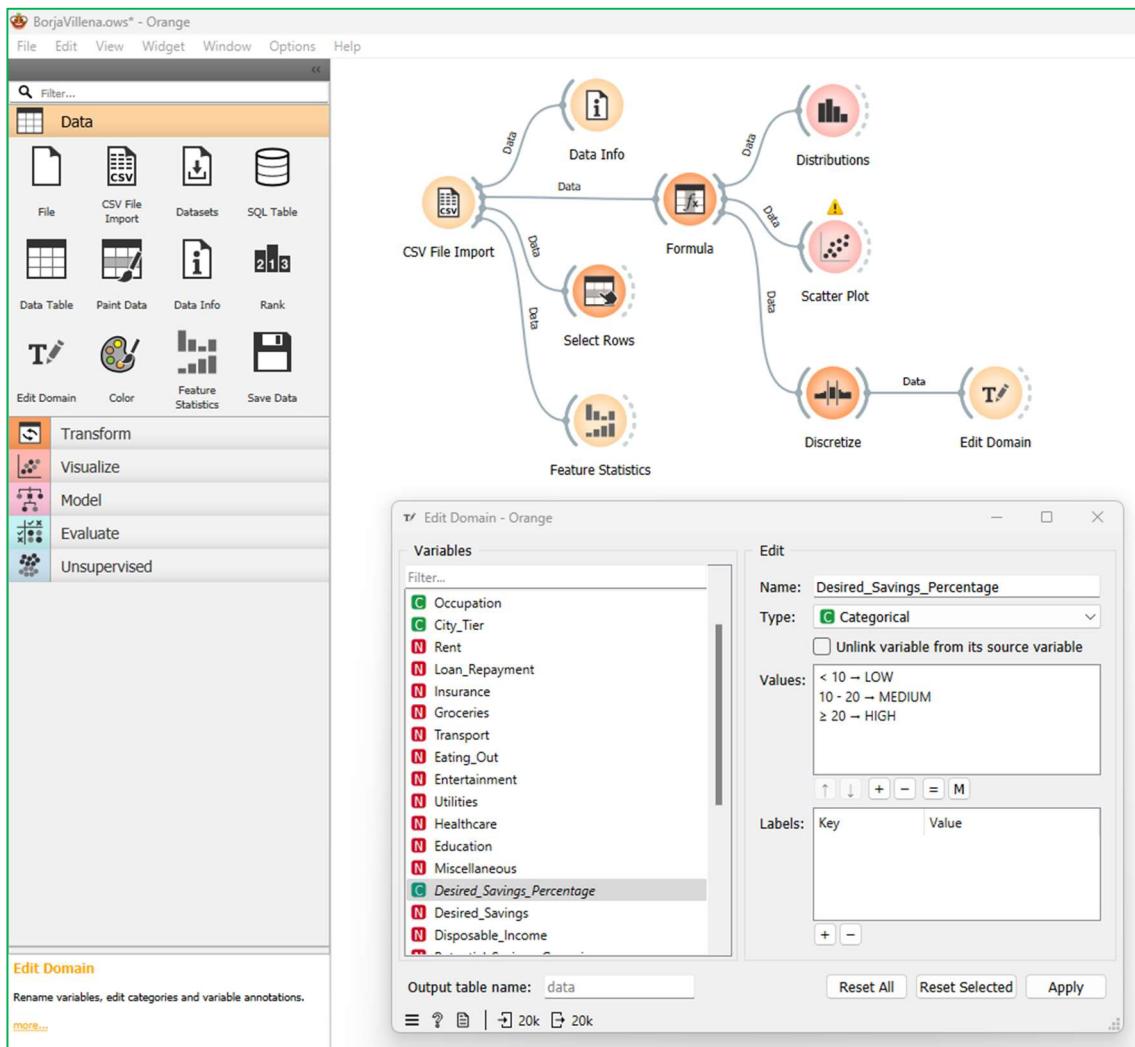
El campo de interés que más me atrae es el de Data science por la amplitud de posibilidades que abarca. Al tratarse de “*la disciplina que utiliza datos y estadística avanzada para hacer predicciones, entender la información y generar conocimiento útil*” el campo de aplicación de esta disciplina se presenta muy amplio.

Personalmente, me atrae poder aplicar los conocimientos de Data Science en dos campos concretos. Por un lado, en el mundo de las Finanzas, y por otro lado en el mundo de la investigación criminal. En ambos casos, la detección de patrones de conducta y comportamiento resultan de un alto valor para prever movimientos y cambios futuros.

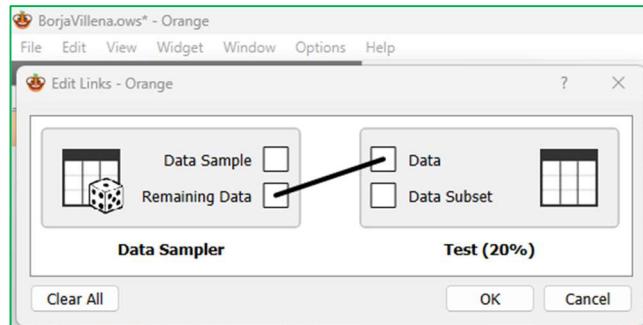
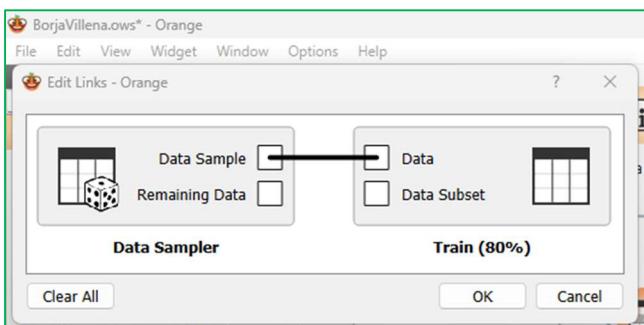
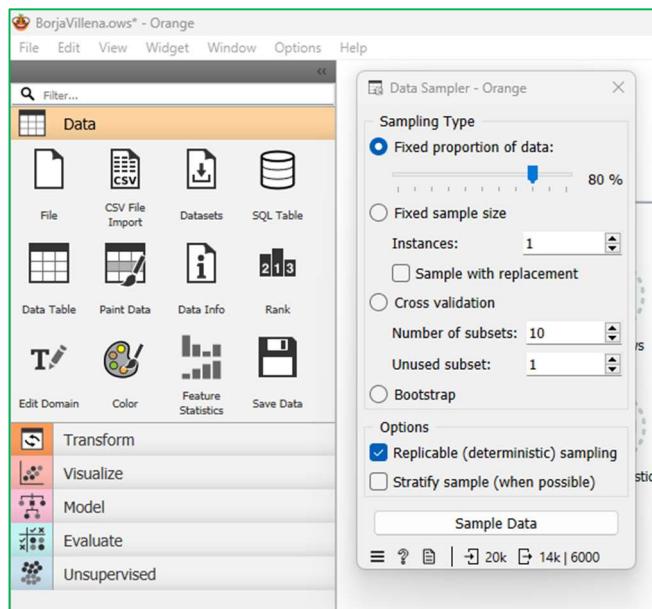
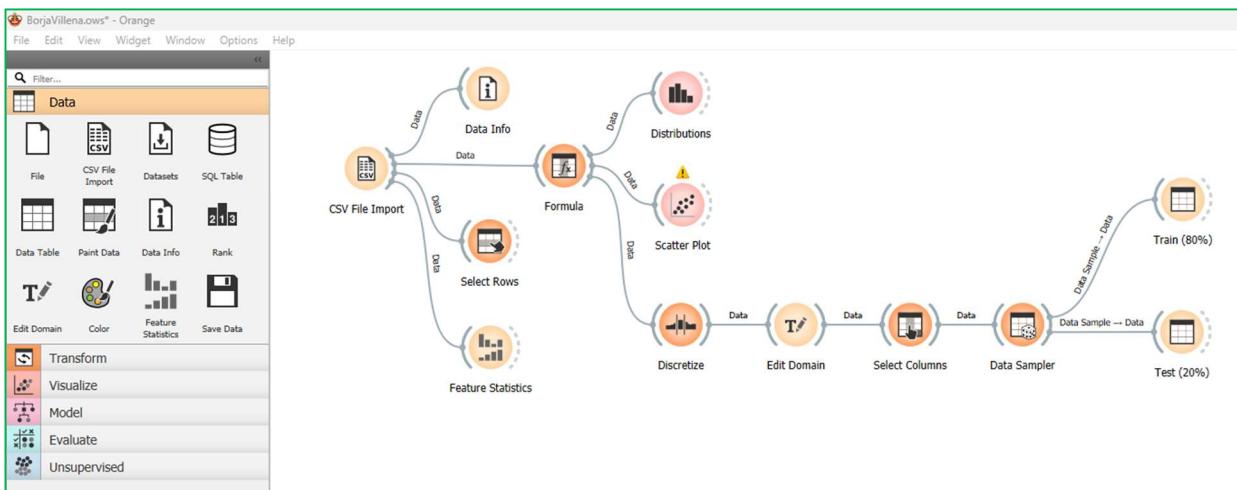
Pregunta 3 (35%)

3.1.A.





3.1.B.



BorjaVillena.ows* - Orange

Data Sampler - Orange

Sampling Type

- Fixed proportion of data: 80 %
- Fixed sample size: Instances: 1
- Sample with replacement
- Cross validation: Number of subsets: 10, Unused subset: 1
- Bootstrap

Options

- Replicable (deterministic)
- Stratify sample (when possible)

Remaining Data: data 16000 instances, 30 variables
Features: 29 (2 categorical, 27 numeric) (no missing values)
Target: categorical

	d_Savings_Perc	Income	Age	Dependents	Occupation	City_Tier	Rent	.loan_Repayment	Insuran
1	LOW	25170.4	23	4	Retired	Tier_2	5034.07	0	571.826
2	LOW	21623.3	48	1	Self_Employed	Tier_2	4324.66	0	468.598
3	MEDIUM	40819.9	30	2	Retired	Tier_2	8163.99	4492.85	1202.68
4	LOW	33836.2	37	4	Retired	Tier_2	6767.25	2446.78	1314.66
5	LOW	28898.7	22	2	Student	Tier_2	5779.75	0	697.131

BorjaVillena.ows* - Orange

Train (80%) - Orange

Info
16000 instances (no missing data)
29 features
Target with 3 values
No meta attributes.

Variables

- Show variable labels (if present)
- Visualize numeric values
- Color by instance classes

Selection

- Select full rows

Restore Original Order

Send Automatically

	d_Savings_Perc	Income	Age	Dependents	Occupation	City_Tier	Rent	.loan_Repayment	Insurance
1	MEDIUM	46164.1	50	4	Retired	Tier_2	9322.82	2984.63	1188
2	LOW	8995.64	37	1	Self_Employed	Tier_1	2688.69	0	430.112
3	LOW	10680.8	43	1	Student	Tier_2	2136.15	0	438.176
4	LOW	66164.9	27	0	Self_Employed	Tier_1	1964.95	867.265	213.095
5	LOW	13525.1	59	1	Retired	Tier_3	2028.76	948.697	572.789

BorjaVillena.ows* - Orange

Test (20%) - Orange

Info
4000 instances (no missing data)
29 features
Target with 3 values
No meta attributes.

Variables

- Show variable labels (if present)
- Visualize numeric values
- Color by instance classes

Selection

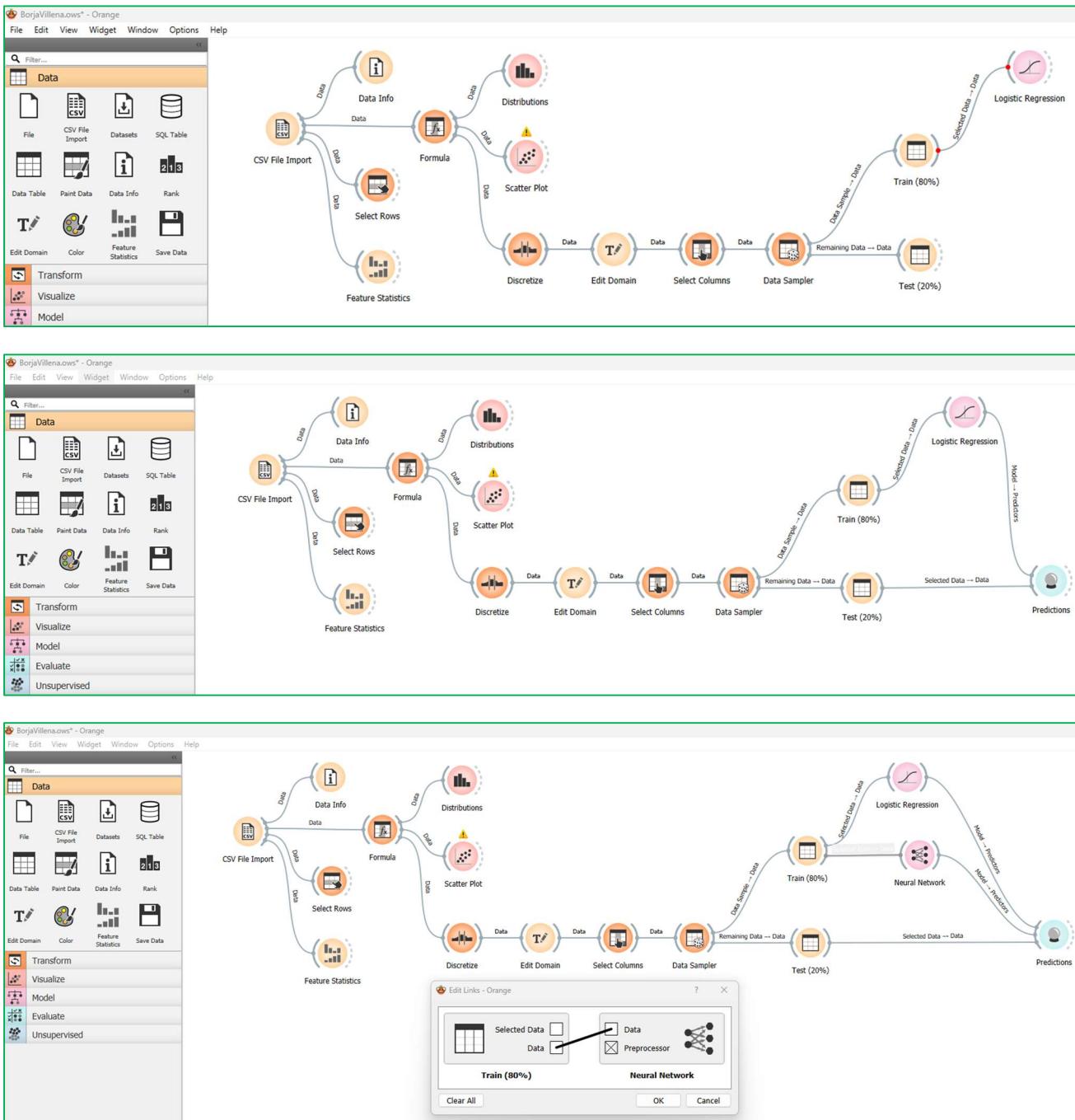
- Select full rows

Restore Original Order

Send Automatically

	d_Savings_Perc	Income	Age	Dependents	Occupation
1	MEDIUM	46164.1	50	1	Retired
2	LOW	8995.64	37	0	Professional
3	LOW	10680.8	43	1	Student
4	LOW	66164.9	27	0	Self_Employed
5	LOW	13525.1	59	1	Retired
6	LOW	17734.1	33	4	Professional
7	LOW	28562.8	61	1	Retired
8	LOW	37349.8	55	1	Retired
9	LOW	33857.8	46	4	Student
10	LOW	24173	50	2	Professional
11	LOW	39790.5	28	4	Student
12	LOW	14322.2	62	3	Professional
13	LOW	7720.9	34	1	Professional
14	LOW	16661.6	45	0	Self_Employed
15	MEDIUM	85549	23	3	Student
16	HIGH	109008	59	4	Retired

3.2.A.



BorjaVillena.ows* - Orange

File Edit View Widget Window Options Help

Predictions - Orange

Show probabilities for (None) Show classification errors Restore Original Order

	Logistic Regression	error	Neural Network	error	ed_Savings_Percei	Income	Age	Dependents	Occupation	City_Tier	Rent	Loan_Repayment	Insurance	
1	MEDIUM	0.027	MEDIUM	0.000	MEDIUM	46164.1	50	1	Retired	Tier_2	9232.82	2984.63	1188	641
2	LOW	0.006	LOW	0.000	LOW	8995.64	37	0	Professional	Tier_1	2698.69	0	430.112	116
3	LOW	0.004	LOW	0.000	LOW	10680.8	43	1	Student	Tier_2	2136.15	0	438.176	118
4	LOW	0.040	LOW	0.000	LOW	6616.49	27	0	Self_Employed	Tier_1	1984.95	867.265	213.095	716
5	LOW	0.001	LOW	0.000	LOW	13525.1	59	1	Retired	Tier_3	2028.76	948.697	572.789	138
6	LOW	0.047	LOW	0.000	LOW	17734.1	33	4	Professional	Tier_3	2660.12	2931.26	366.314	260
7	LOW	0.000	LOW	0.000	LOW	28562.8	61	1	Retired	Tier_3	4284.42	0	1035.22	425
8	LOW	0.001	LOW	0.000	LOW	37349.8	55	1	Retired	Tier_2	7469.95	0	958.395	421
9	LOW	0.009	LOW	0.000	LOW	33857.8	46	4	Student	Tier_2	6771.57	0	1536.23	491
10	LOW	0.001	LOW	0.000	LOW	24173	50	2	Professional	Tier_2	4834.6	0	818.021	284
11	LOW	0.462	LOW	0.003	LOW	39790.5	28	4	Student	Tier_1	11937.2	0	1650.04	565
12	LOW	0.001	LOW	0.000	LOW	14322.2	62	3	Professional	Tier_2	2864.44	2642.99	656.737	204
13	LOW	0.008	LOW	0.000	LOW	7720.9	34	1	Professional	Tier_3	1158.14	0	287.92	834
14	LOW	0.002	LOW	0.000	LOW	16661.6	45	0	Self_Employed	Tier_1	4998.47	0	356.49	216
15	MEDIUM	0.169	MEDIUM	0.000	MEDIUM	85549	23	3	Student	Tier_3	12832.4	13787.8	3906.46	105
16	MEDIUM	0.912	MEDIUM	0.962	HIGH	109008	59	4	Retired	Tier_2	21801.6	0	2554.54	114
17	LOW	0.104	LOW	0.000	LOW	5806.46	19	2	Professional	Tier_2	1161.29	800.373	185.074	842

Show performance scores Target class: HIGH

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.930	0.927	0.735	0.851	0.647	0.849
Neural Network	0.996	0.985	0.789	0.805	0.774	0.969

≡ ? ⌂ | ↗ 4000 | ↘ 4000 | ↗ 4000 | ↘ 4000 | 2×4000

Basándonos en los resultados obtenidos, opinamos que el modelo “*Neural Network*” es mejor, para evaluar este caso, que el modelo “*Logistic Regression*”, ya que los valores de AUC, CA y F1 son más cercanos a 1 en el primer modelo mencionado.

Show performance scores Target class: (Average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Logistic Regression	0.976	0.927	0.927	0.928	0.927	0.849
Neural Network	0.999	0.985	0.985	0.985	0.985	0.969

≡ ? ⌂ | ↗ 4000 | ↘ 4000 | ↗ 4000 | ↘ 4000 | 2×4000

3.2.B.

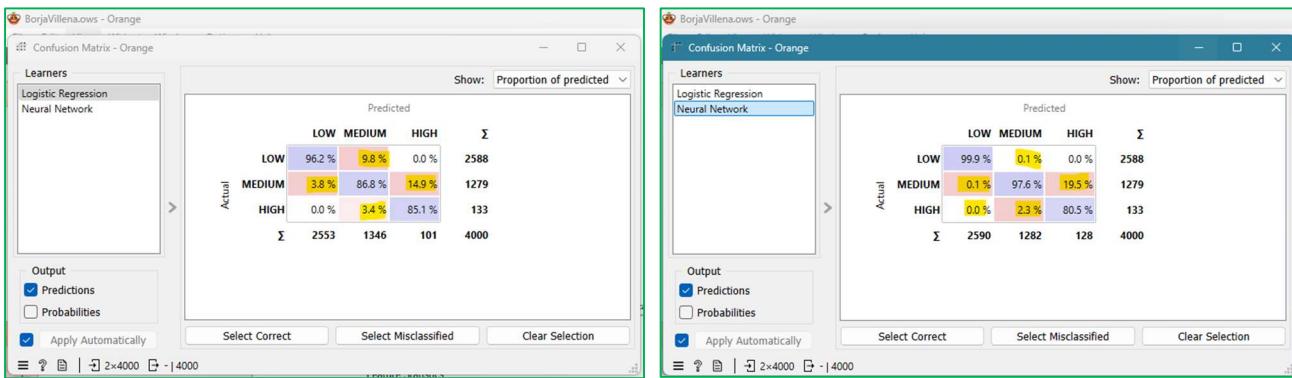
The top screenshot shows a flow diagram in the Orange data mining software. The process starts with 'CSV File Import' (Data tab), followed by 'Data Info' and 'Feature Statistics'. Then, 'Select Rows' and 'Discretize' are used. The data then flows through 'Edit Domain', 'Select Columns', 'Data Sampler' (splitting into 'Train (80%)' and 'Test (20%)'), and finally 'Data' nodes leading to 'Logistic Regression' and 'Neural Network' learners. The outputs from the learners are combined via 'Model -> Predictions' and 'Node -> Predictions' to produce 'Predictions' and 'Evaluation Results' (Confusion Matrix).

The middle and bottom screenshots show the 'Confusion Matrix - Orange' widget. Both screenshots show the same data for two learners: Logistic Regression and Neural Network. The learners are listed in the 'Learners' panel on the left. The output is set to 'Predictions' (checked). The 'Show' dropdown is set to 'Number of instances'. The confusion matrix table is as follows:

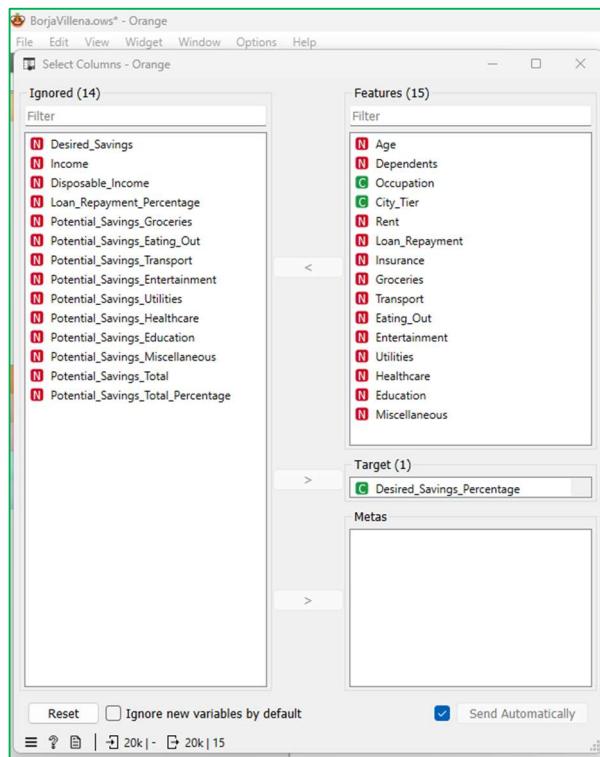
		Predicted			
		LOW	MEDIUM	HIGH	Σ
Actual	LOW	2456	132	0	2588
	MEDIUM	96	1168	15	1279
	HIGH	1	46	86	133
Σ	2553	1346	101	4000	

At the bottom of the middle screenshot, there are three buttons: 'Select Correct', 'Select Misclassified', and 'Clear Selection'. The bottom screenshot shows a slightly different configuration where 'Neural Network' is selected as the learner. The confusion matrix values are slightly different due to the different learner's performance.

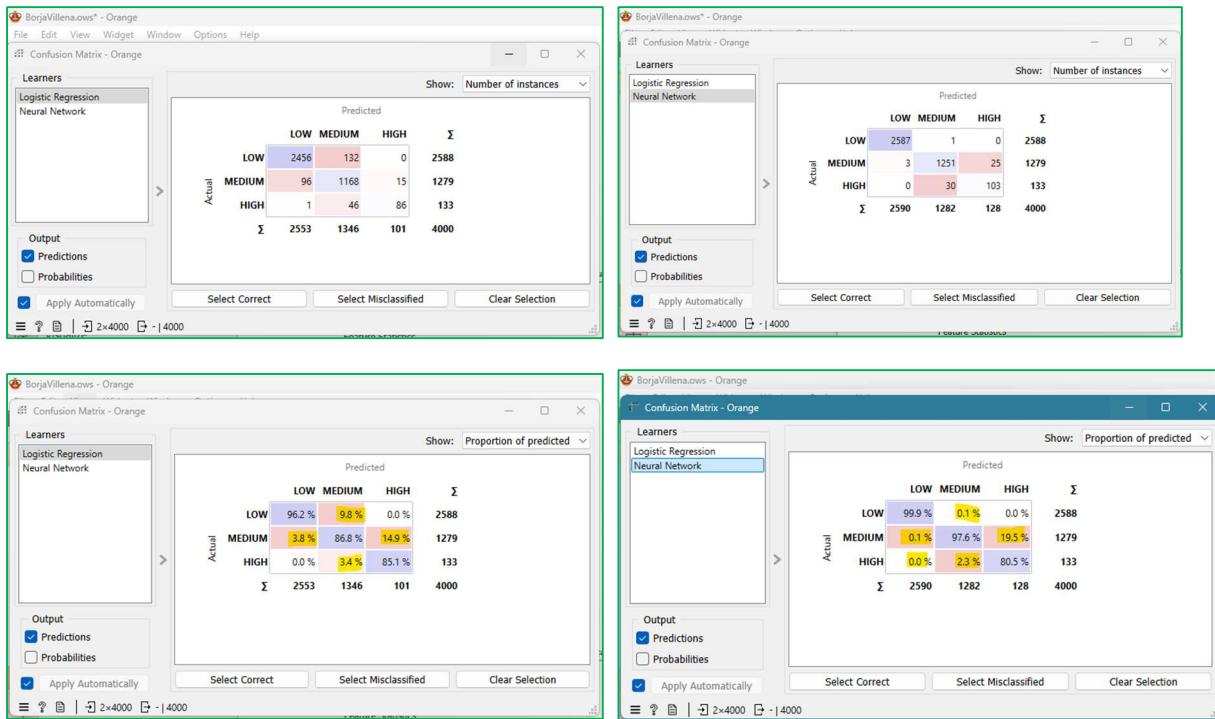
Como podemos observar, el modelo “*Neural Network*” realiza una mejor predicción de las clases **LOW** y **MEDIUM**, con un 99.9% y 97.6% de acierto respectivamente frente al 96.2% y 86.8% que presenta el otro modelo; mientras que, por otro lado, el modelo “*Logistic Regression*” se comporta algo mejor para la predicción de la clase **HIGH** con un 85.1% de acierto frente al 80.5% del modelo anterior.



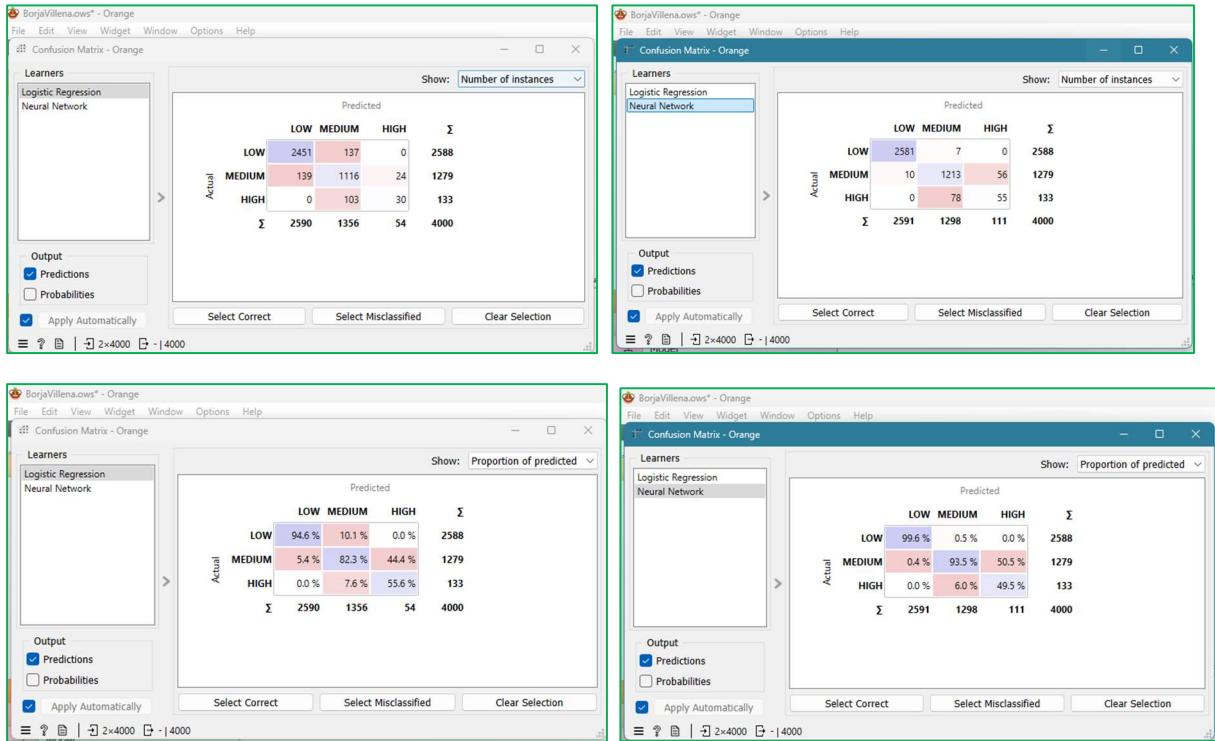
3.3.A.



Predicciones antiguas



Predicciones nuevas



A simple vista podemos observar que las predicciones antiguas obtienen porcentajes más beneficiosos que las nuevas predicciones.

Los resultados son similares, el modelo “*Neural Network*” realiza una mejor predicción de las clases LOW y MEDIUM, con un 99.6% y 93.5% de acierto respectivamente frente al 94.6% y 82.3% que presenta el otro modelo; mientras que, por otro lado, el modelo “*Logistic Regression*” se comporta algo mejor para la predicción de la clase HIGH con un 55.6% de acierto frente al 49.5% del modelo anterior.

3.3.B.

El cambio que se plantea quiere decir que dónde antes el modelo había cometido el error de predecir como MEDIUM 133 casos que realmente eran HIGH, ahora tras el cambio realizado reflejamos que el modelo ha logrado predecir correctamente 133 casos como HIGH.

En este contexto, basándonos en la procedencia de los valores de las métricas, tal y como podemos contrastar con las fórmulas descritas en la siguiente fuente consultada:

<https://www.themachinelearners.com/metricas-de-clasificacion/>

podemos concretar que las métricas se ven afectadas de manera positiva:

- a) AUC → En el caso previo al cambio, al tener la totalidad de los casos HIGH clasificados erróneamente como MEDIUM, la AUC sería muy cercana a 0. Sin embargo, tras el cambio, al clasificarse los casos HIGH correctamente, el valor de AUC aumentará, acercándose al valor de 1.
- b) CA → En este caso ocurre algo similar. Al corregir el error de haber clasificado mal los 133 casos, el número de predicciones correctas aumenta, y por ende la precisión también. En el caso concreto que nos ocupa, el valor de CA sería de 1 ya que habríamos corregido la totalidad de los casos.
- c) F1 → En este caso vuelve a ocurrir lo mismo. El hecho de corregir los 133 casos que estaban mal clasificados por el modelo, la métrica F1-score mejoraría notablemente.