JUAN LUIS ACEBAL RICO

PEC 2 IDS

INDICE

Yo, Juan Luis Acebal Rico, declaro que para realizar esta entrega... he utilizado mis medios sin plagio.

ENUNCIADO 1: PERIODISMO DE DATOS

CONTEXTO

El artículo "Tracking global data on electric vehicles" de Hannah Ritchie, examina la adopción global de vehículos eléctricos (VE) con una mirada puesta a que es una estrategia esencial para descarbonizar el transporte por carretera. Se centra en datos de ventas, el stock de VE (el parqué actual) y el uso de vehículos de combustión vs vehículos híbridos, y eléctricos.
Fundamentalmente intenta ilustrar el nivel de adopción de VE que pueda contribuir a sostenibilidad y mostrando que países lideran la transformación.

LOS DATOS UTILIZADOS.

El análisis se basa en datos proporcionados por la Agencia Internacional de Energía (AIE), que tiene datos anuales sobre las tendencias de ventas y uso de VE. Los datos comienzan en 2010 y llegan hasta 2023 e incluyen VE e híbridos enchufables desglosados. Además, tiene el número de coches totales en unidades, en porcentaje, etc

CONCLUSIONES.

El estudio se observa que desde 2010 ha habido un aumento significativo de ventas de VE, siendo una proporción de 1 de cada cinco coches vendidos (18%) eran eléctricos en 2023. Lo más llamativo del estudio es sin duda Noruega, seguido de Suecia con un 90% y 60% respectivamente, del resto del mundo destaca especialmente china con un 40%. La investigación hablar también de la huella de carbono que genera la fabricación de un VE, que se compensa con rapidez, especialmente en países como Noruega donde tienen una proporción de electricidad mucho más grande que por ejemplo Polonia. [1]

LO MAS LLAMATIVO

Para mí lo que más me ha llamado la atención es la disparidad entre ciertos países, reflejando grandes diferencias en políticas ambientales y la capacidad de su infraestructura. Por ejemplo, me llama la atención la nula implantación en América latina, la disparidad de datos que hay en la Unión Europea y como desde 2021 el crecimiento de ventas se ha estancado a excepción de China. Esto me sugiere que hay obstáculos en el resto del mundo (donde la Unión Europea es

quizás lo más llamativo) donde quizás no se ejercen correctamente las políticas medioambientales para que la adopción se VE sea un hecho.

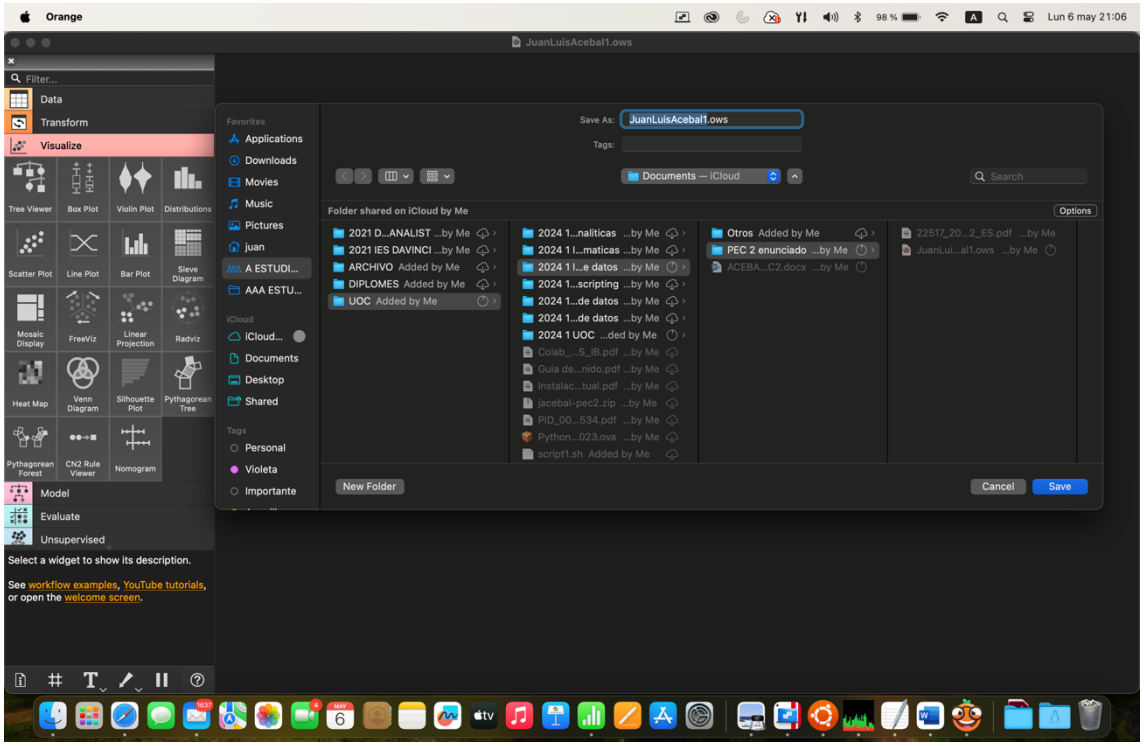
QUE HUBIERA ANALIZADO

Hubiera analizado a fondo el impacto de las políticas y ayudas gubernamentales para la adopción de VE. Claramente se necesita más infraestructura para mejorar la tasa de adopción y que los VE tengan un coste inicial más cercano a los de combustión.

Como vías de estudio/análisis también puedo sugerir el efecto del uso de electricidad sobre cada país, ya que quizás hay países que podrían vender, pero no cargar los vehículos eléctricos por falta de electricidad. Por último, el caso de Polonia donde el uso de un VE no es muy diferente de usar un vehículo de combustión ya que la red eléctrica es suministrada fundamentalmente por combustibles fósiles, y creo que dando visibilidad a esto puede fomentarse el uso y el cambio social necesario en esos países.

ENUNCIADO 2: MACHINE LEARNING I

PREGUNTA 2.1



Aquí podemos ver como guardar un proyecto.

PREGUNTA 2.2.A

Es un dataset con valores sobre pacientes que tienen diabetes, hay la edad, valores en sangre(glucosa), embarazos, presión sanguínea, grosor cutáneo, insulina, IMC, predisposición genética (parece una función de antecedentes familiares), y si tienen o no diabetes.

En el contexto adecuado podría ayudar a diagnosticar y diseñar políticas de salud para detección y prevención precoz de la diabetes.

Las preguntas que podría responder a mi juicio son:

Principales predictores de la diabetes.

Factores de riesgo valorando peso, grosor de la piel, presión sanguínea, IMC.

Factores de riesgo con la edad.

Relación de los embarazos y la diabetes.

Riesgo de diabetes según predisposición genética/antecedentes familiares.

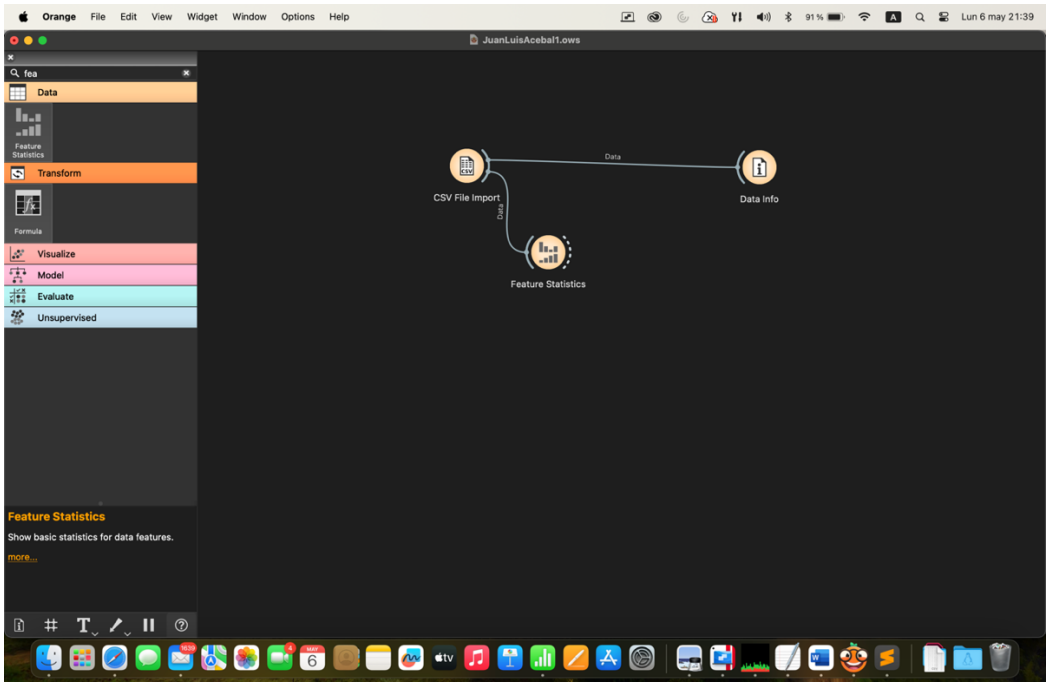
Predicción temprana de diabetes.

Pronóstico de una persona con diabetes según niveles de insulina, glucosa y peso.

Factores de buen pronóstico.

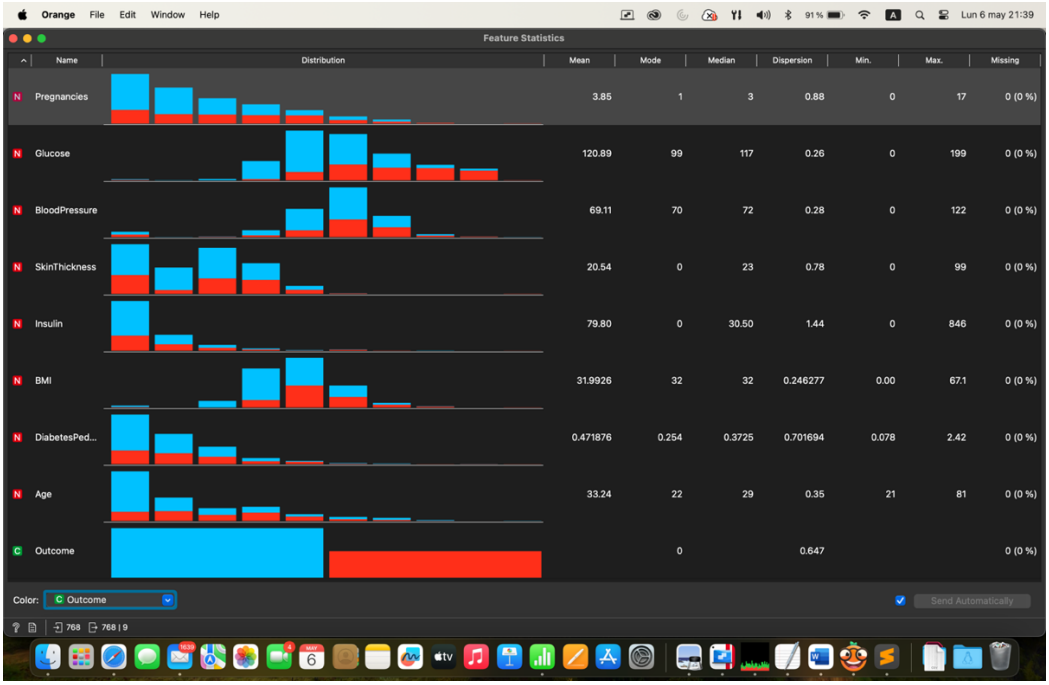
IMC en relación con el peso, grosor piel y desarrollo diabetes.

PREGUNTA 2.2.B



Aquí podemos ver como el diseño empieza a tener forma

SEGUNDA CAPTURA



Aquí podemos ver la estadística descriptiva de cada columna.

ANALIZA LAS ESTADÍSTICAS DE LOS ATRIBUTOS DEL DATASET

Podemos ver que tenemos 9 tipos de dato, son todos simples, no hay valores nulos en principio, de los datos, todos son numéricos a excepción de Outcome que es booleano.

DENTRO DE LAS FEATURE STATISTICS TENEMOS:

NOMBRE DEL DATO

Columna 0: Tipo de dato, numérico o categórico.

Nombre: Nombre del dato.

Distribución: grafico con la distribución separando en 2 colores según si tiene la enfermedad o no (Outcome u otro dato elegido aunque en este caso es la que tiene más sentido)

Media: Es la media de todos los valores

Moda: Es el valor que más se repite

Mediana: Es la mitad de todos los valores

Dispersión: Indica como de juntos o separados están los valores. A más alto, menos cerca están unos valores de otros.

Min y **Max** es el rango de los valores.

Missing son los valores faltantes, en términos absolutos y en porcentaje.

ANÁLISIS RESUMIDO DE CADA ATRIBUTO

PREGNANCIES (EMBARAZOS)

Media: 3.85.

Moda: 1, el valor más común es 1 embarazo.

Mediana: 3, lo que sugiere una distribución relativamente simétrica alrededor del promedio.

Dispersión: 0.88, indica una variabilidad moderada en el número de embarazos.

Rango: 0 a 17.

Valores Faltantes: Ninguno. Valor máximo de 17, para una mujer de 47 años. No sé qué pensar de la veracidad de ese registro, parece un registro erróneo o outlier.

GLUCOSE (GLUCOSA)

Media: 120.89, promedio de glucosa en sangre.

Moda: 99, el valor más común de glucosa en sangre.

Mediana: 117, punto medio de los datos de glucosa.

Dispersión: 0.26, baja dispersión en los niveles de glucosa.

Rango: 0 a 199.

Valores Faltantes: Ninguno. Valores de glucosa en el rango en 0, quizás deban ser considerados valores faltantes, nulos o outliers.

BLOODPRESSURE (PRESIÓN ARTERIAL)

Media: 69.11.

Moda: 70.

Mediana: 72.

Dispersión: 0.28, baja dispersión en la presión arterial.

Rango: 0 a 122

Valores Faltantes: Ninguno. Valores de presión arterial en el rango en 0, quizás deban ser considerados valores faltantes, nulos o outliers.

SKINTHICKNESS (GROSOR DE LA PIEL)

Media: 20.54

Moda: 0, Hay muchos registros tienen un valor de 0, lo que a mi juicio indica datos nulos.

Mediana: 23.

Dispersión: 0.78, moderada.

Rango: 0 a 99.

Valores Faltantes: Ninguno. Valores de grosor de la piel en el rango en 0 y moda de 0, quizás deban ser considerados valores faltantes, nulos o outliers.

INSULIN (INSULINA)

Media: 79.80.

Moda: 0, lo mismo que con el grosor de la piel, muchos valores en 0.

Mediana: 30.5.

Dispersión: 1.44, muy alta dispersión.

Rango: 0 a 846.

Valores Faltantes: Ninguno. Valores de Insulina en el rango en 0 y moda de 0, quizás deban ser considerados valores faltantes, nulos o outliers.

Nota: la alta dispersión junto a la moda min 0 indican diferencias de condiciones de salud altas junto con un problema en la toma de datos y quizás valores nulos.

BMI (ÍNDICE DE MASA CORPORAL (IMC))

Media: 31.9926.

Moda: 32.

Mediana: 32.

Dispersión: 0.246277, baja.

Rango: 0.00 a 67.1.

Valores Faltantes: Ninguno. Valores de IMC en el rango en 0, quizás deban ser considerados valores faltantes, nulos o outliers.

DIABETESPEDIGREEFUNCTION (FUNCIÓN DE PEDIGRÍ DE LA DIABETES, ¿ANTECEDENTES FAMILIARES?)

Media: 0.471876.

Moda: 0.254.

Mediana: 0.3725.

Dispersión: 0.701694, bastante alta en relación con el rango de la medida.

Rango: 0.078 a 2.42.

Valores Faltantes: Ninguno.

AGE (EDAD)

Media: 33.24.

Moda: 22.

Mediana: 29.

Dispersión: 0.35, moderada.

Rango: 21 a 81.

Valores Faltantes: Ninguno.

Nota: Se repite muchas personas jóvenes (moda 22 años) y no está separado por género, lo que me indica que los embarazos no se pueden tomar en cuenta para ciertas perspectivas a la hora de hacer el estudio.

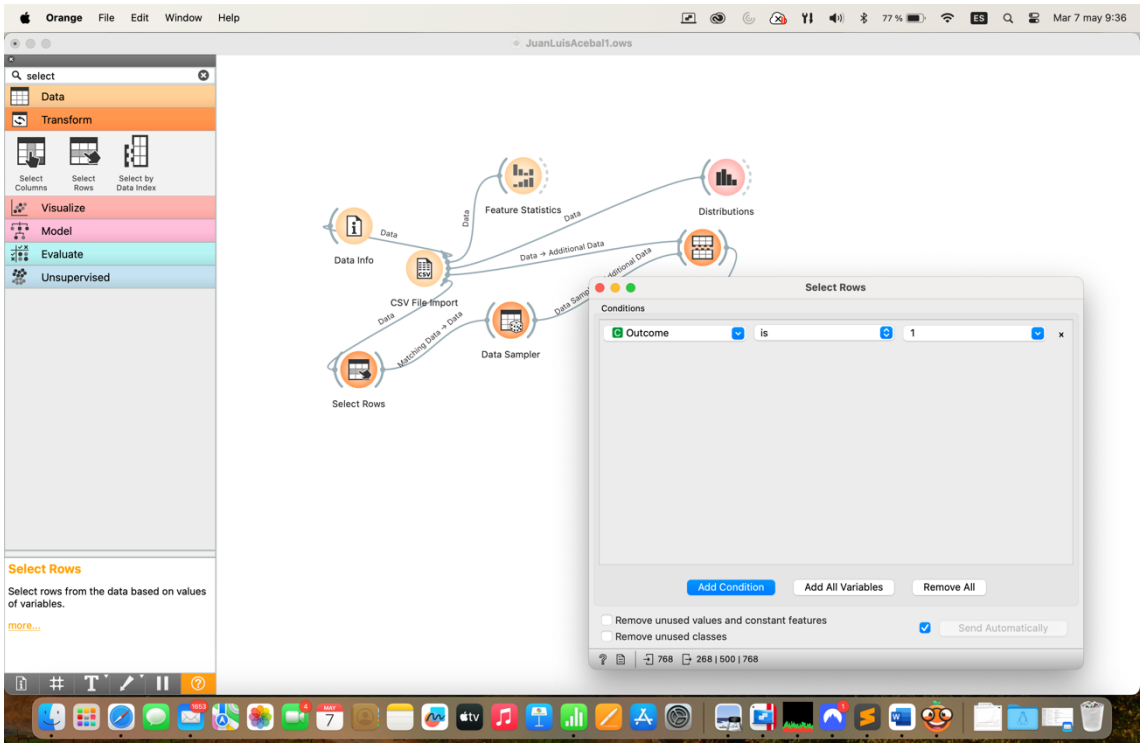
OUTCOME (ENFERMEDAD PRESENTE)

Dispersión: 0.647, un valor alto considerando que es una variable booleana (0 o 1).

Valores Faltantes: Ninguno

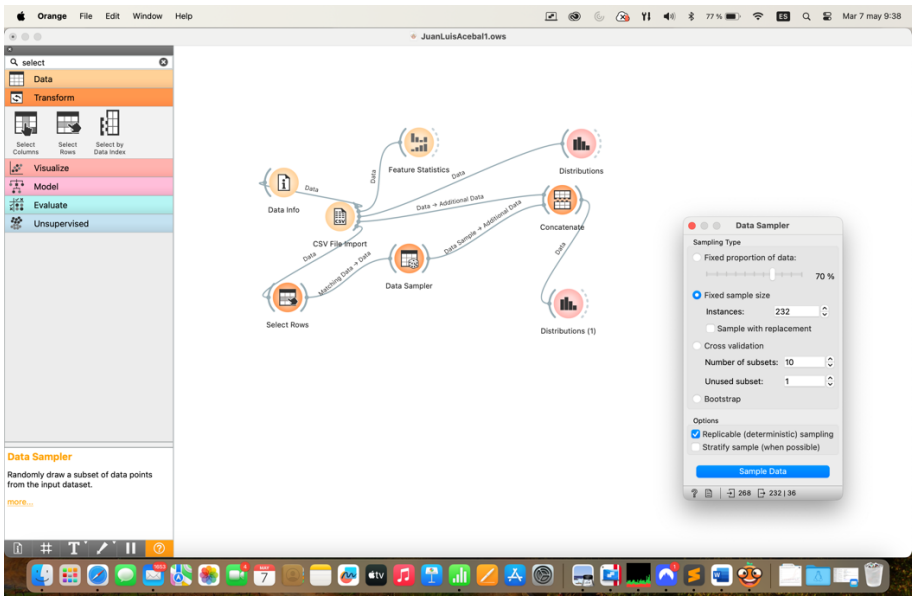
PREGUNTA 2.3.A

SELECT ROWS



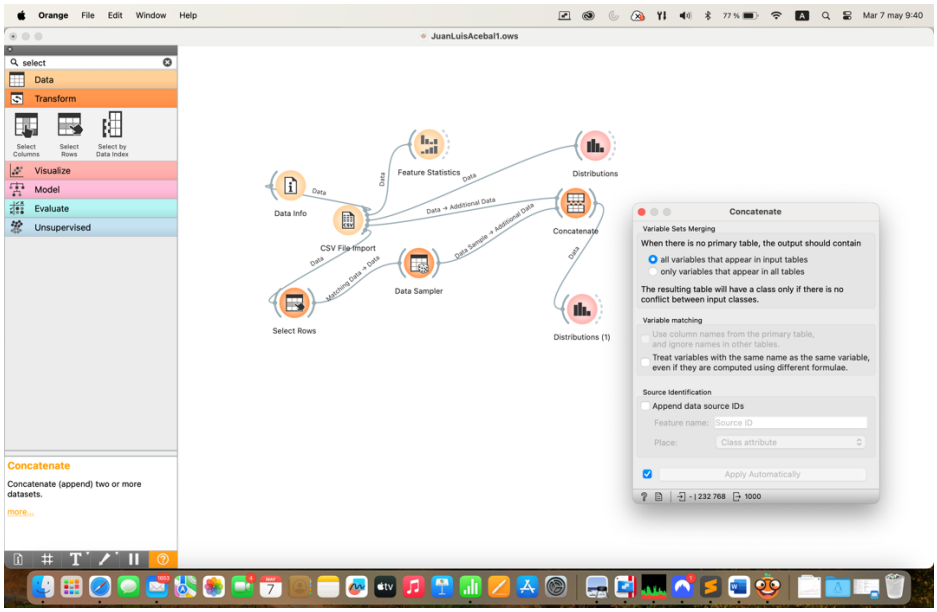
Aquí tenemos la configuración de select rows, donde seleccionamos a personas con diabetes (268 personas) para después trabajar con estos datos.

DATA SAMPLER



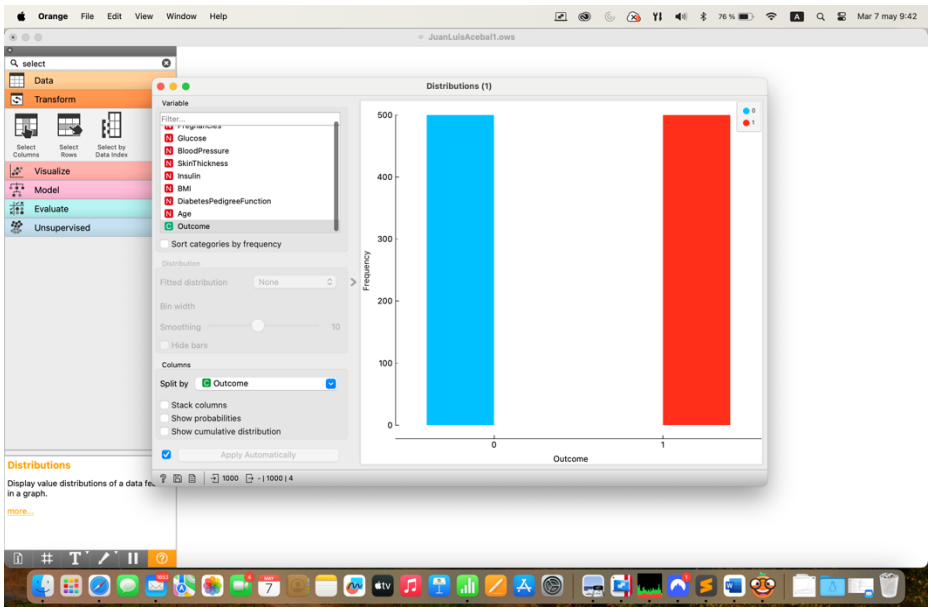
El objetivo en data sampler es añadir una muestra de 500-268=232 registros para igualar los registros sin enfermedad que son 500 (Nuestro dataset son 768 registros en total)

CONCATENATE



Hemos hecho concatenate, o join de los datos creados en Data Sampler con los datos que teníamos originales.

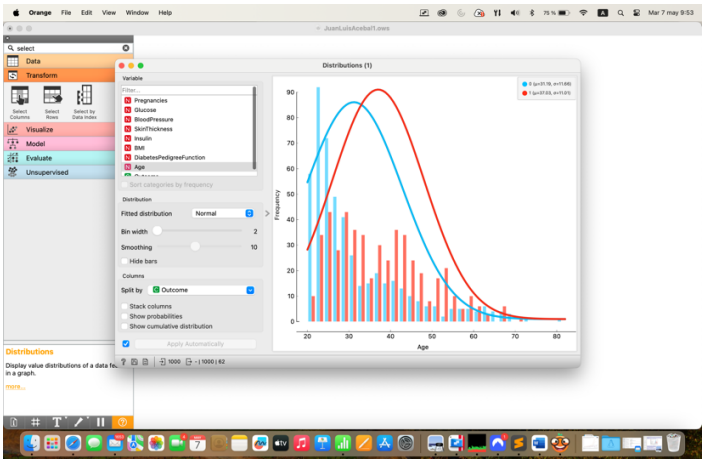
DISTRIBUTIONS



Por último podemos ver cómo, de tener 768 registros no distribuidos uniformemente, hemos pasado a 1000, 500 para si enfermedad y otros 500 a no enfermedad.

PREGUNTA 2.3.B

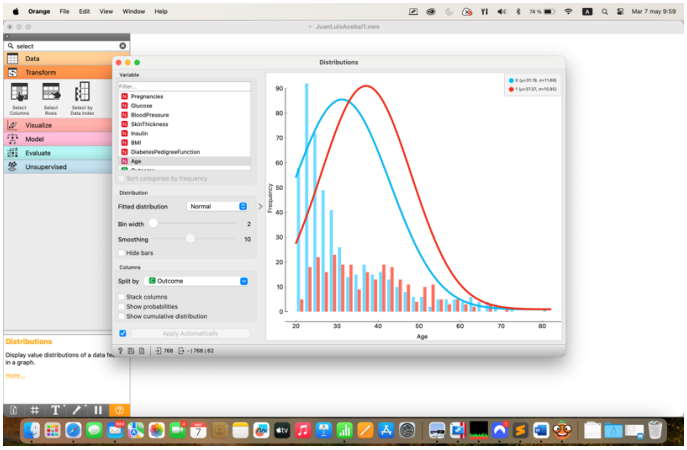
FITTED DISTRIBUTION DE EDAD CON 1000 REGISTROS



Aquí podemos observar cómo hay una distribución parecida pero esa distribución, su curva es un poco más alta en el caso de existir enfermedad y la edad es un poco

mayor. La media es 31,19 en no enfermedad y 37,03 si existe enfermedad. Además la diferencia de desviaciones estándar es 11,66 vs 11,01.

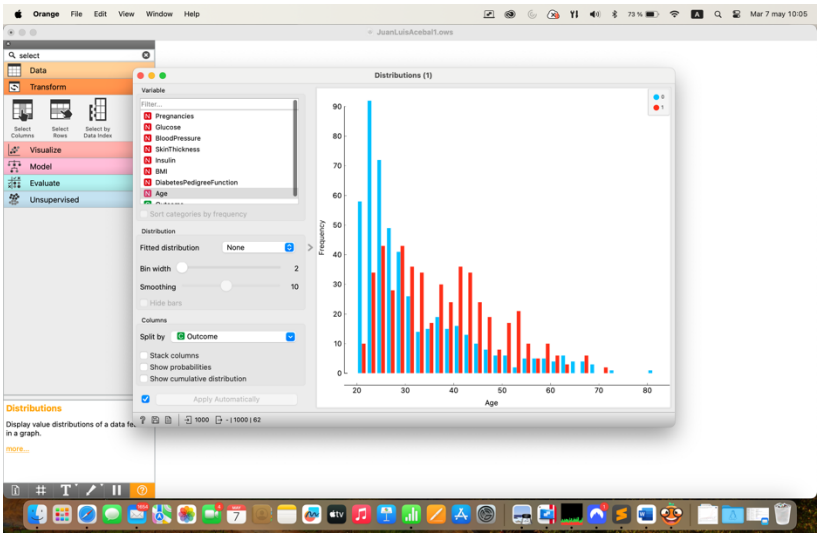
FITTED DISTRIBUTION DE EDAD CON 768 REGISTROS



Aquí vemos lo mismo, pero con una media para en el caso de haber enfermedad de 37,07 y una desviación estándar de 10,95. Aquí trabajamos solamente con 768 registros.

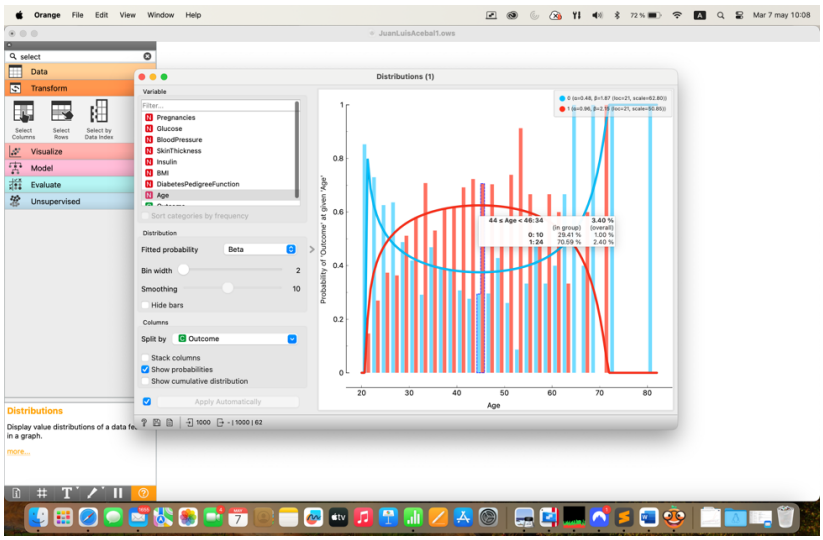
PREGUNTA 2.4.A

CAPTURA 1



Podemos ver el paso previo a filtrarlo por la beta viendo las probabilidades.

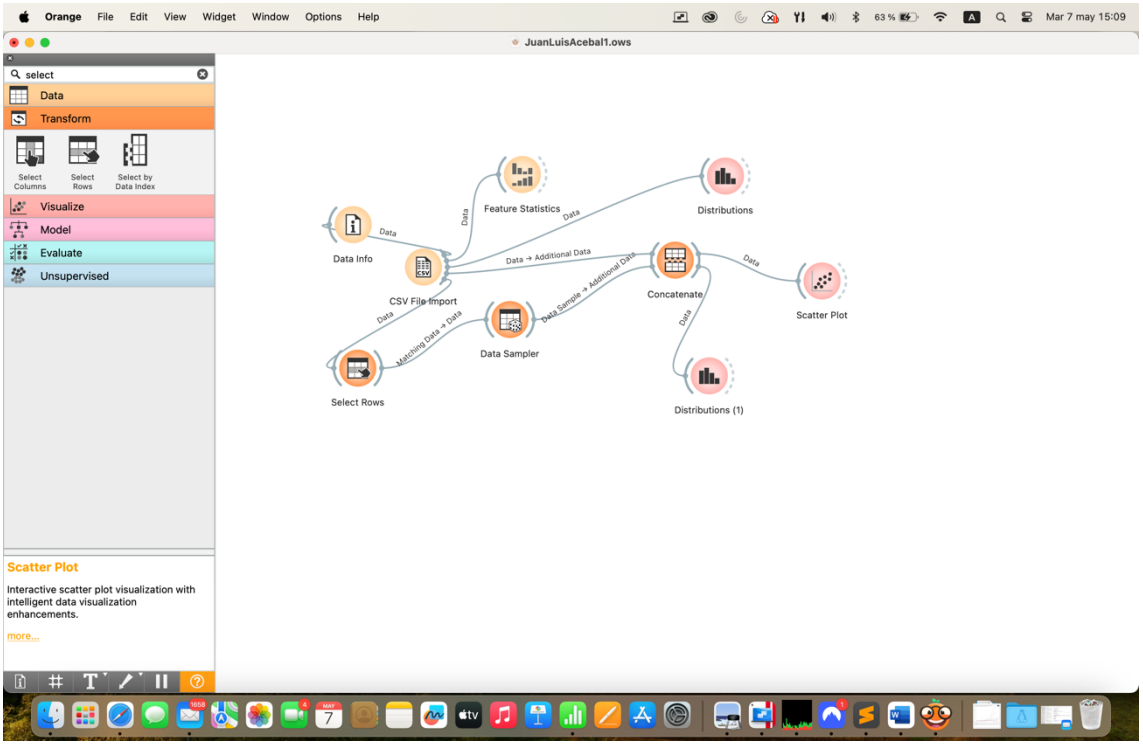
CAPTURA 2 DE LA BETA Y PROBABILIDADES



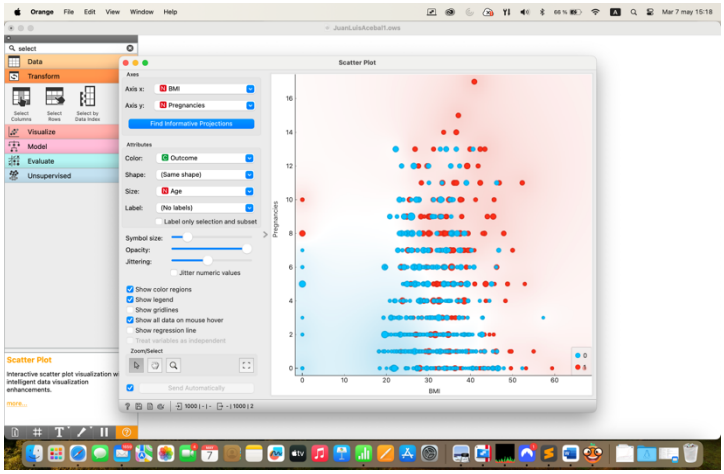
Aquí podemos observar como para por ejemplo la edad entre 44 y 46 años, la probabilidad de tener la enfermedad es de 70,59% y de no tenerla, solamente del 29,41%

Podemos observar también que a partir de los 68 años no hay probabilidad casi de tener la enfermedad, y las personas jóvenes, tienen algo de riesgo menores de 22 años de parecer la enfermedad, pero tampoco es significante. Los puntos de inflexión son >25 años y <63 años para tener la enfermedad con el pico en 44-46 años como ya he comentado.

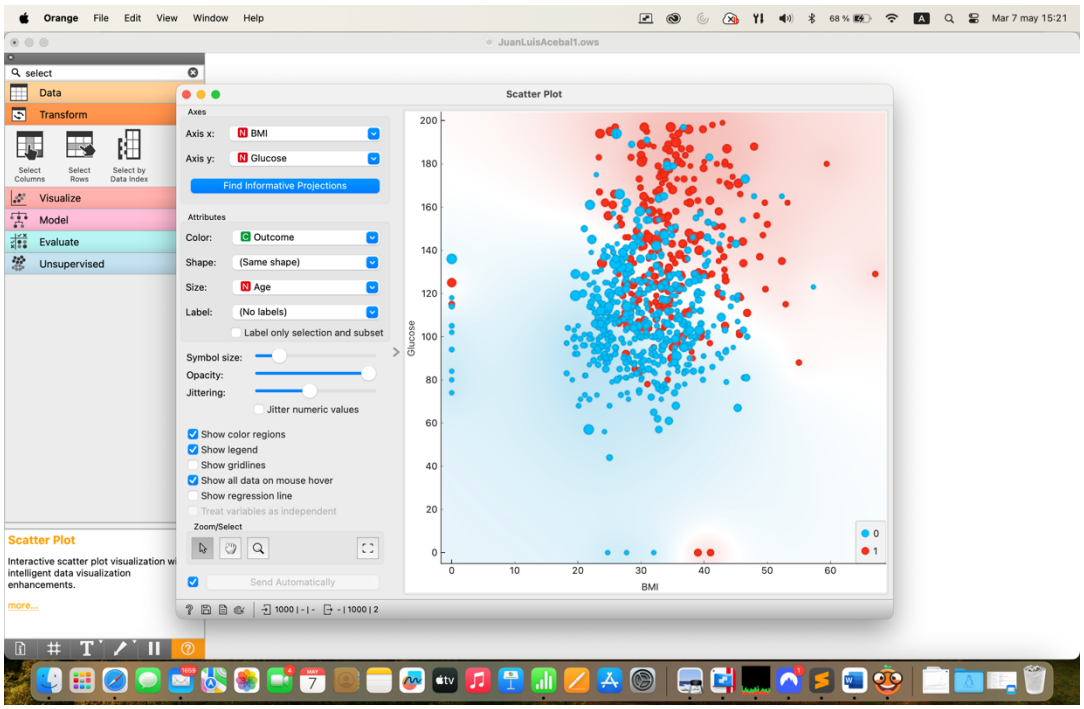
PREGUNTA 2.4.B.



SCATTER PLOT



Mi aportación en la segunda captura es IMC con embarazos. Además los puntos tienen tamaño de la edad. Hubiera elegido el del enunciado que es el más evidente marcando 2 zonas en el grafico. En este lo he elegido por ser un buen ejemplo para hablar de outliers, ya que se ven muchos outliers en el grafico a la izquierda donde IMC (BMI) es cero.

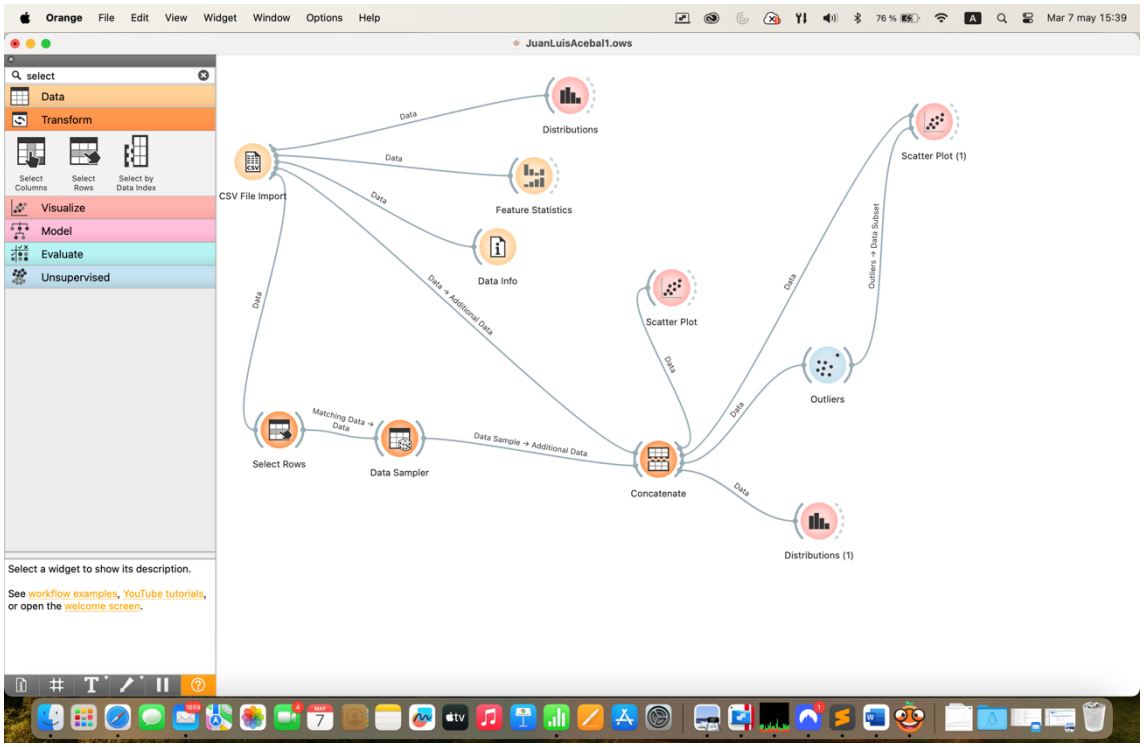


Además, añado las sugerencias de Orange, que es la que voy a elegir en los próximos pasos del ejercicio, la segunda sugerencia por outcome, que vemos que es más útil o más de la que yo he propuesto inicialmente, ya que marca mejor las zonas y para hacer clasificaciones y modelos predictivos es más interesante aun. Tiene outliers también en dos zonas, sobre todo, **IMC igual a 0** (11 registros) y **Glucosa igual a 0** (5 registros).

Además, podríamos valorar en diferentes escenarios si los datos son outliers con IMC mayor que 55-60 (4 registros).

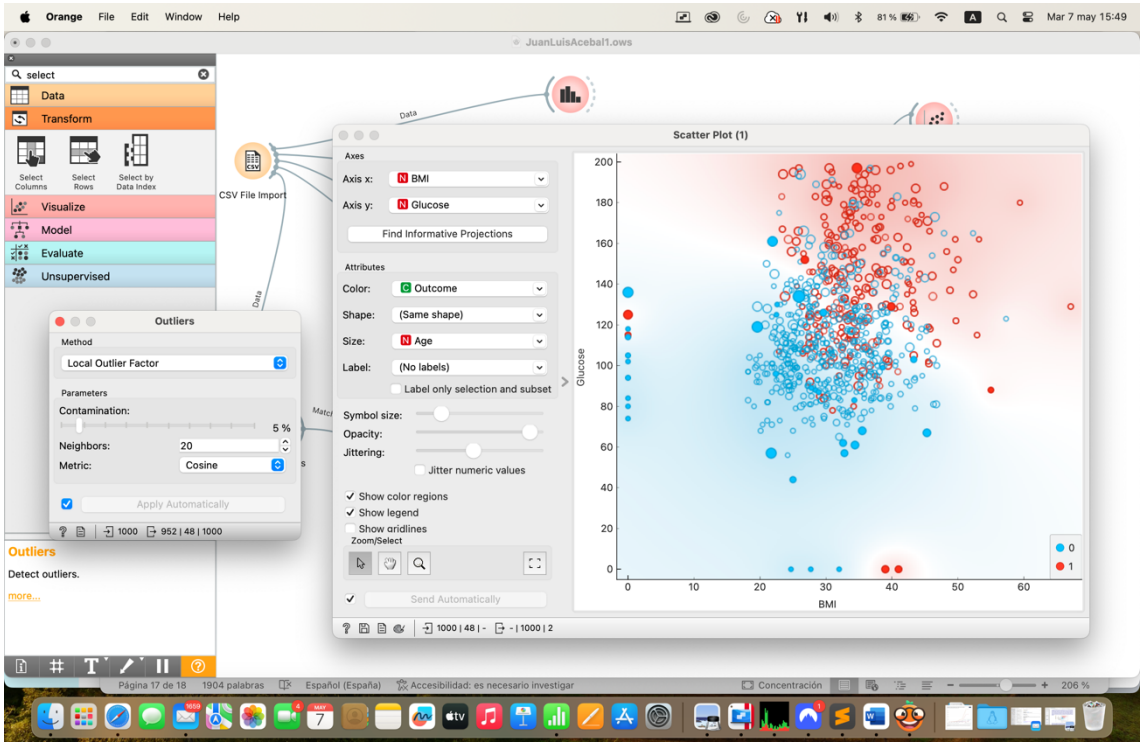
PREGUNTA 2.5.B

CANVAS CON LOS OBJETOS CREADOS



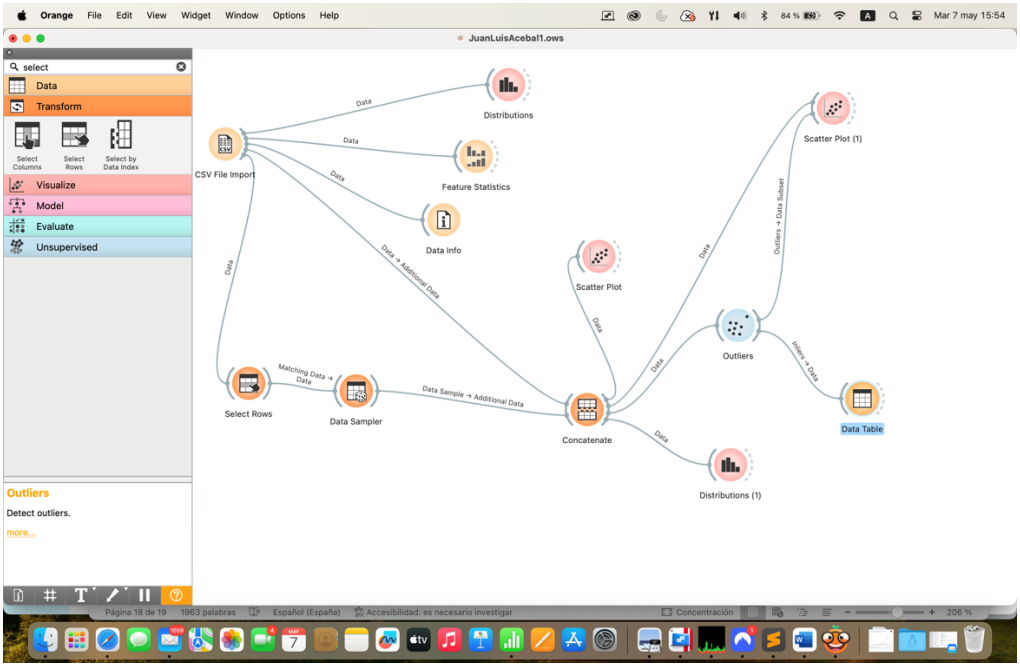
Aquí he creado tal y como indica el enunciado del ejercicio un nuevo objeto llamado Outliers y scarter plot (1) para tratar y visualizar los outliers. He puesto como fuentes de datos Concatenate para Outliers, y Concatenate y Outlier para scarter plot (1)

SCATTER PLOT MARCANDO OUTLIERS



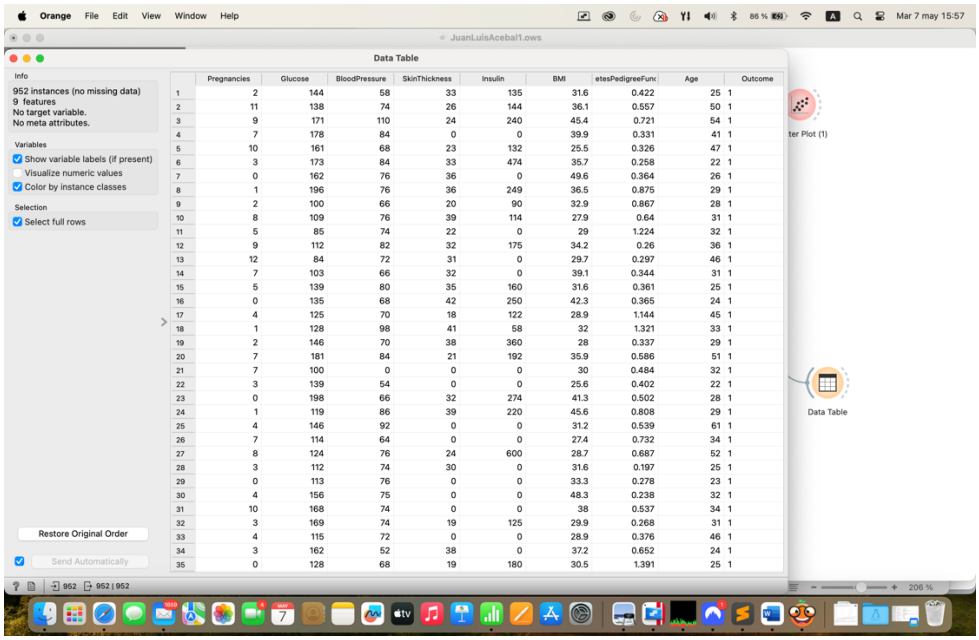
Después de configurar el scatter plot, he ido probando y con **un 5% sería suficiente** para eliminar los outliers que yo he considerado principales. Si quisiera eliminar IMC mayor a 55, tendría que **eliminar el 24%**, demasiados registros a mi parecer, entonces para hacer este primer modelo quizás vale la pena dejar esos registros y después valorar el modelo.

CANVAS AÑADIENDO "DATA TABLE"



Hay que seleccionar como datos inliers+data en el momento de la creación.

DATA TABLE

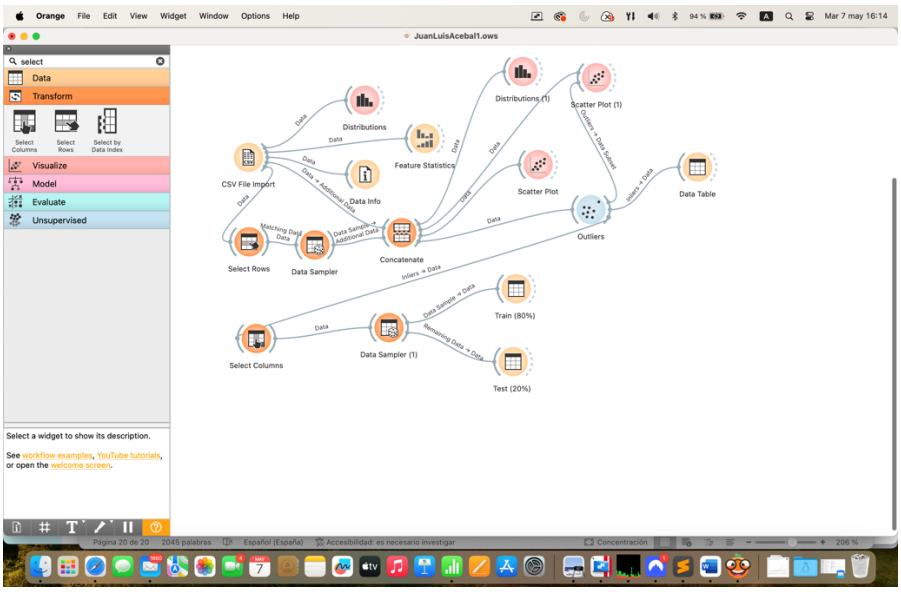


Tengo 952 registros, eso quiere decir que el paso anterior ha eliminado bien 48 registros correspondientes al 5 % de los registros.

ENUNCIADO 3

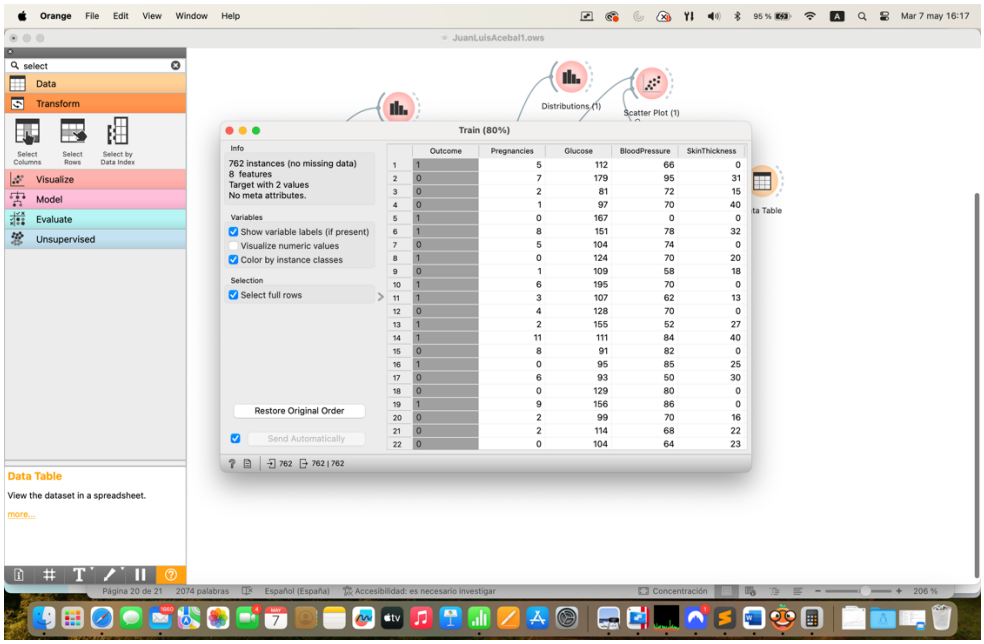
PREGUNTA 3.1

CANVAS CON EL SPLIT ENTRE TRAIN Y TEST.



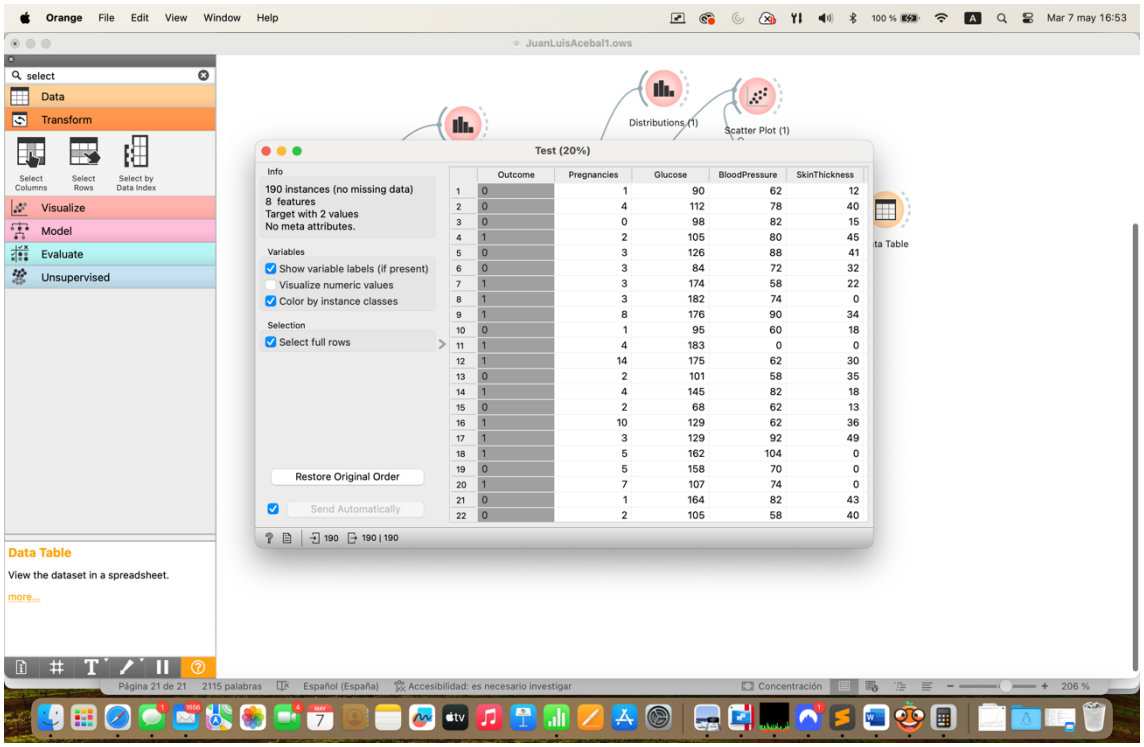
Donde mis 952 registros tendrían que quedar el 80% en un lado (761,6 aprox, 952x0,8) y el 20% en el otro (190,4 aproximadamente).

TRAIN



Aquí vemos que que ha sido 762, tal y como había dicho que debería de ser antes de verlo. Esta muestra va a servir para entrenar nuestro modelo, y que pueda desde predecir valores del resto de componentes a predecir la enfermedad. Ya eso depende de cómo se haga un modelo (hay muchos, desde regresiones, clasificaciones, clustering, etc)

TEST



Aquí tenemos 190 registros. Se van a utilizar para evaluar nuestro modelo, esto quiere decir, que, vamos a crear un modelo que busque y prediga en base al conjunto de datos TRAIN, que debe ser, en este caso Outcome (creo, aún no hemos llegado ahí en la PEC), pero podría usarse también TRAIN para predecir datos faltantes, con regresión linear por ejemplo, y datos boleanos regresión logística por ejemplo. Después de crear nuestro modelo, la validación se hace prediciendo los registros TEST y que sean iguales a los registros de la captura de pantalla. Eso se hace si queremos saber Outcome, creando un nuevo dataset eliminando los valores Outcome, y predecirlos con el modelo, y que la predicción sea la misma que en el conjunto de datos TEST original.

PREGUNTA 3.2.A

CANVAS

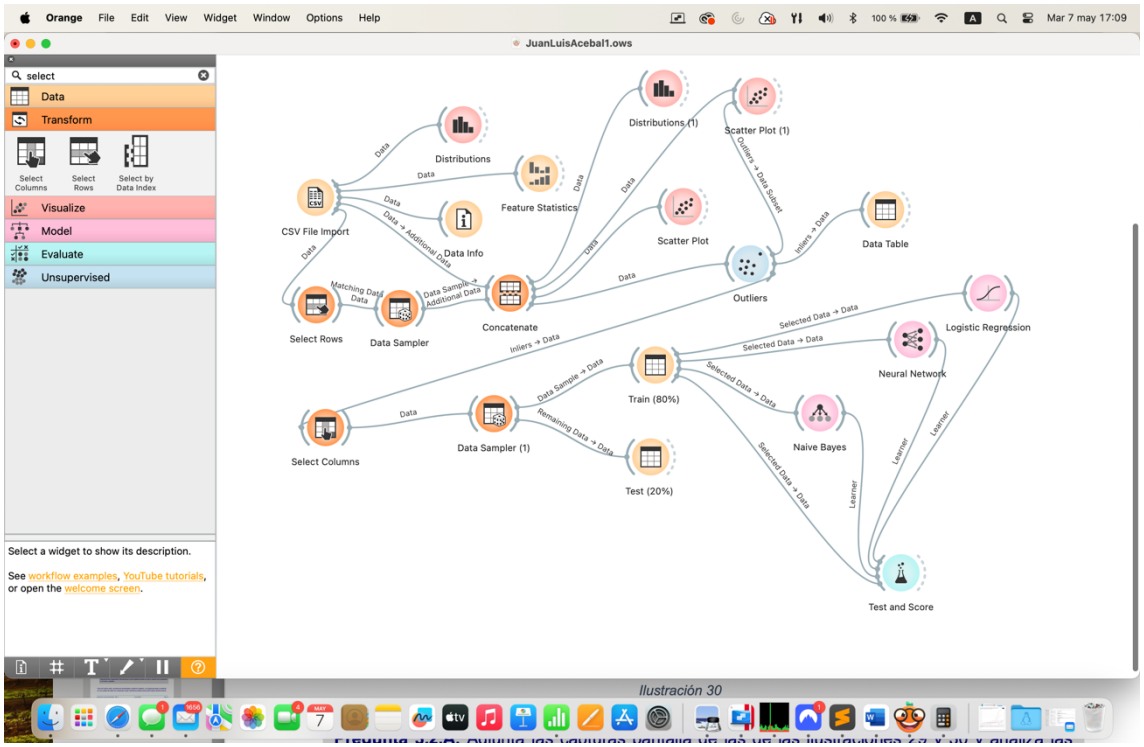
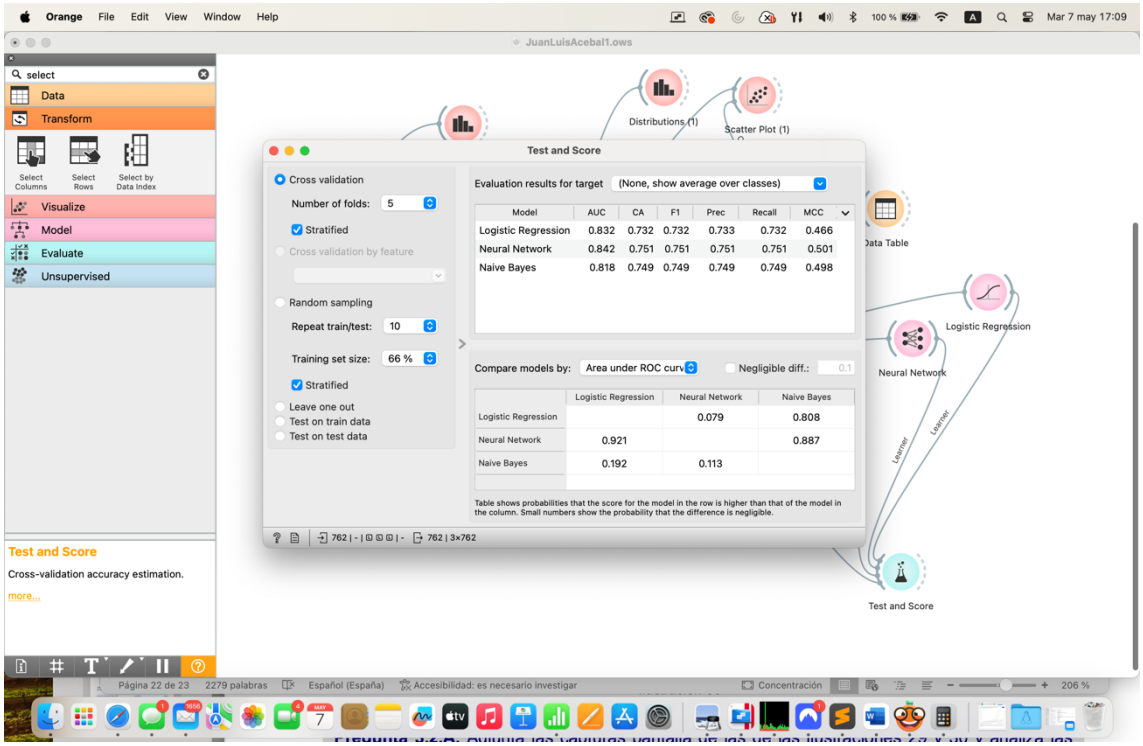


Ilustración 30

Podemos ver la actualización de canvas incluyendo los 3 modelos predictivos más test and score.

METRICAS

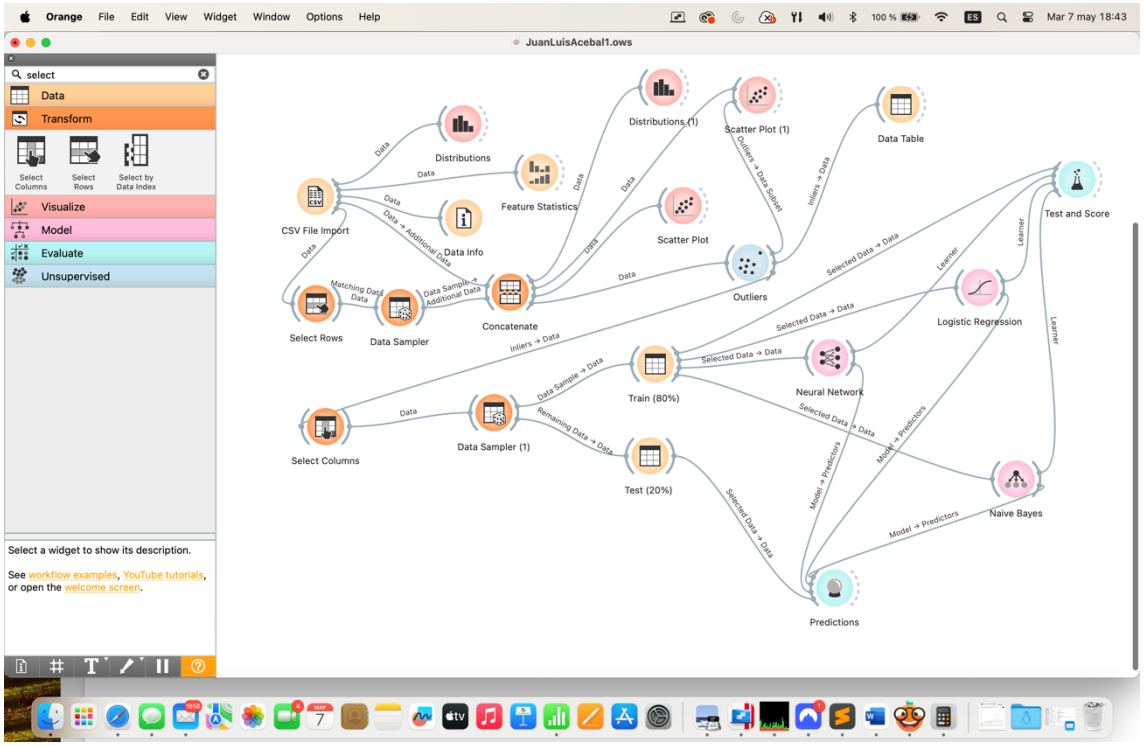


Aquí la conclusión después de mucho leer y no ser la primera vez, no podría decir categóricamente que estoy seguro ya que aun no siendo la primera vez que lo hago, la experiencia hace mucho, y en data science, no es la primera vez que veo que las interpretaciones son contradictorias o diferentes a lo evidente, la cautela importa. Además hasta donde yo sé, a veces tienes que probar y testear decenas de veces modelos con distintos parámetros, eligiendo bien como entrenarlo (Por ejemplo, quizás el grosor de la piel perjudica a la regresión logística, pero si lo eliminamos podríamos obtener unos scores muchos más altos en ella), o para qué usarlo (A veces un modelo es más fuerte en los positivos, y otro es más fuerte en los negativos, existiendo menos falsos positivos en uno y falsos negativos en otro, pudiendo combinar ambos según el propósito).

Dicho eso, yo veo que los 3 modelos están bien, el mejor es red neuronal, y el peor es regresión logística en base a los 3 indicadores clave que son AUC, CA y F1. Prec y Recall forman parte de F1 (también con la curva de ROC de AUC) y MCC que es una correlación. Además, red neuronal es mejor en MCC.

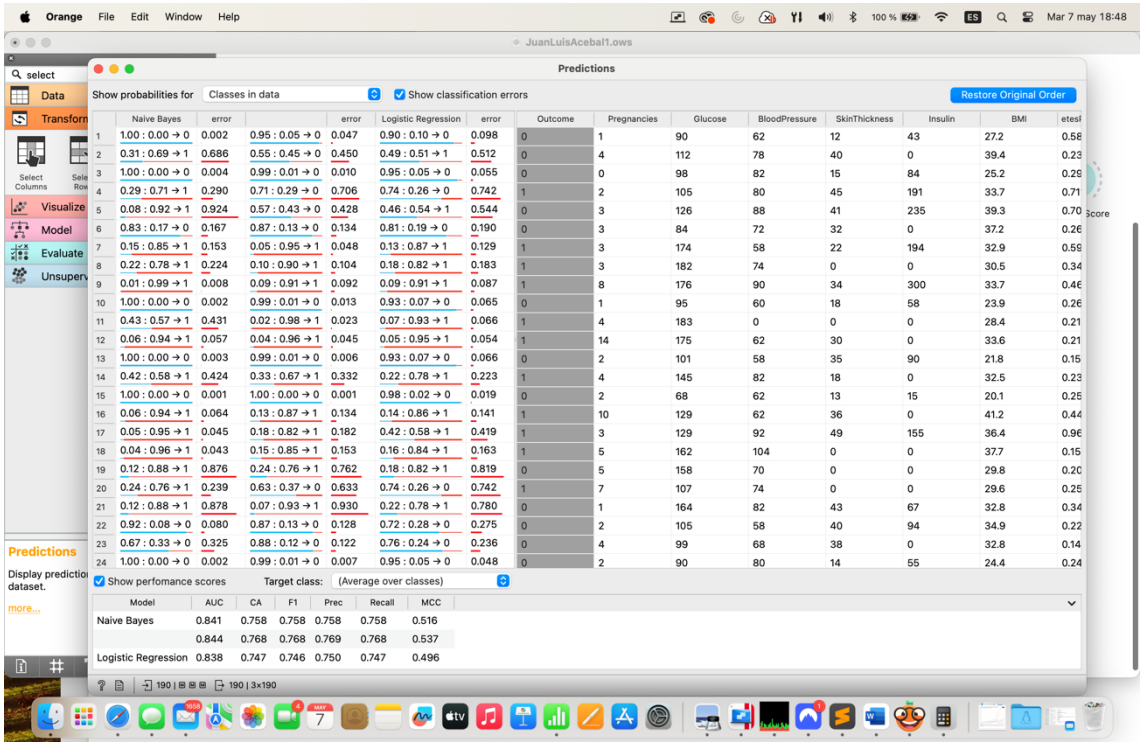
PREGUNTA 3.2.B

CAVAS CON EL RESULTADO DEL APARTADO



Podemos observar cómo, ha quedado el modelo, añadiendo predicciones y conectando los 3 modelos para hacer esas predicciones.

MÉTRICAS CLAVE MODELOS

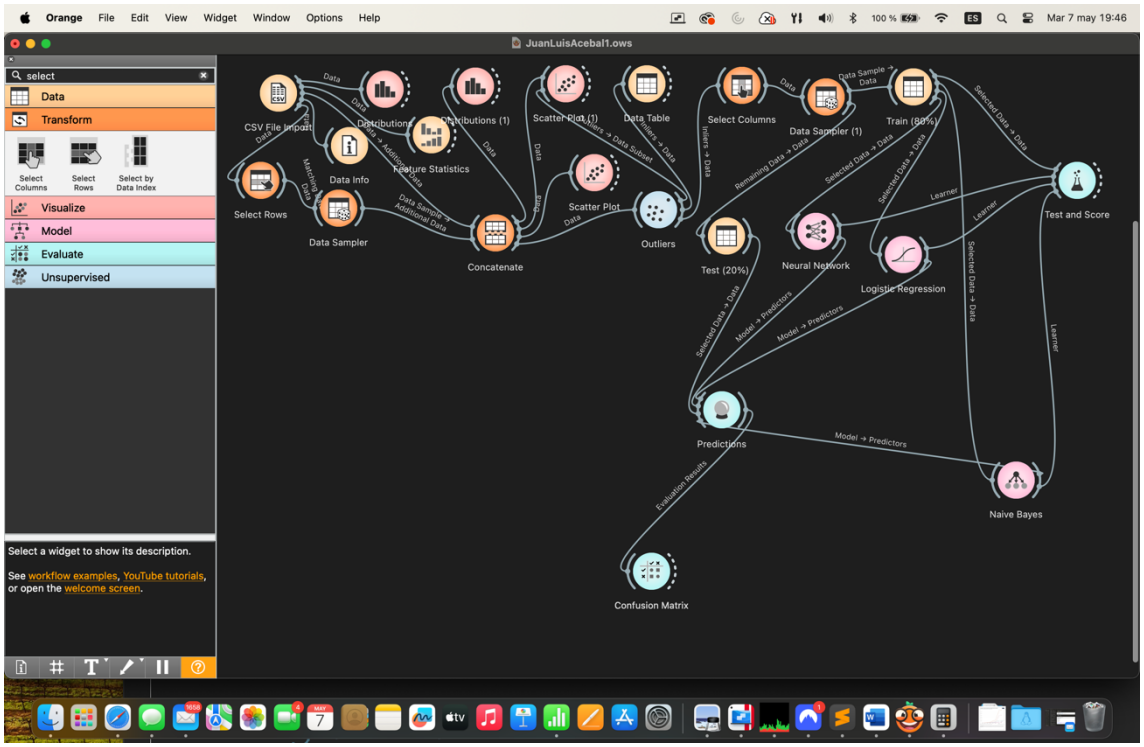


Aquí podemos ver las estadísticas de rendimiento de los 3 modelos, Naive Bayes, la red neuronal (en blanco) y la regresión logística.

Después de hacer la prueba de ver las estadísticas en cualquier clase y como objetivo cuando outcome es 0 y cuando es 1, observo que la red neuronal es la que mejor funciona, en general. Es decir, para clasificar y predecir resultados da mas eficacia y equilibrio. La regresión logística, si bien funciona bien, es un poco peor que al resto, y el modelo de Naive Bayes se sitúa entre ambas métricas.

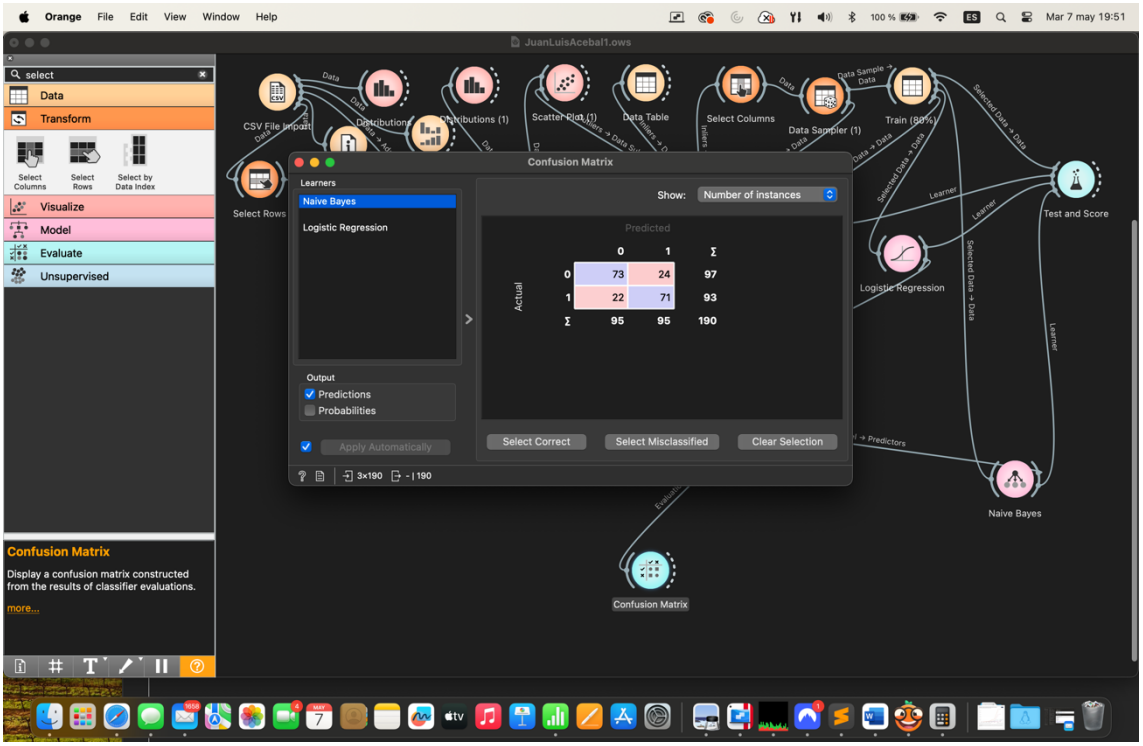
PREGUNTA 3.2.C

MODELO



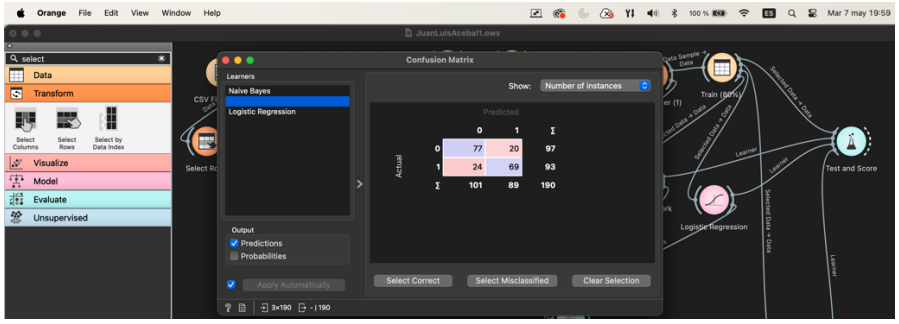
Aquí no hay mucho que comentar a excepción que hemos puesto la matriz de confusión conectada a predictions.

MATRIZ DE CONFUSIÓN DE NAIVE BAYES.



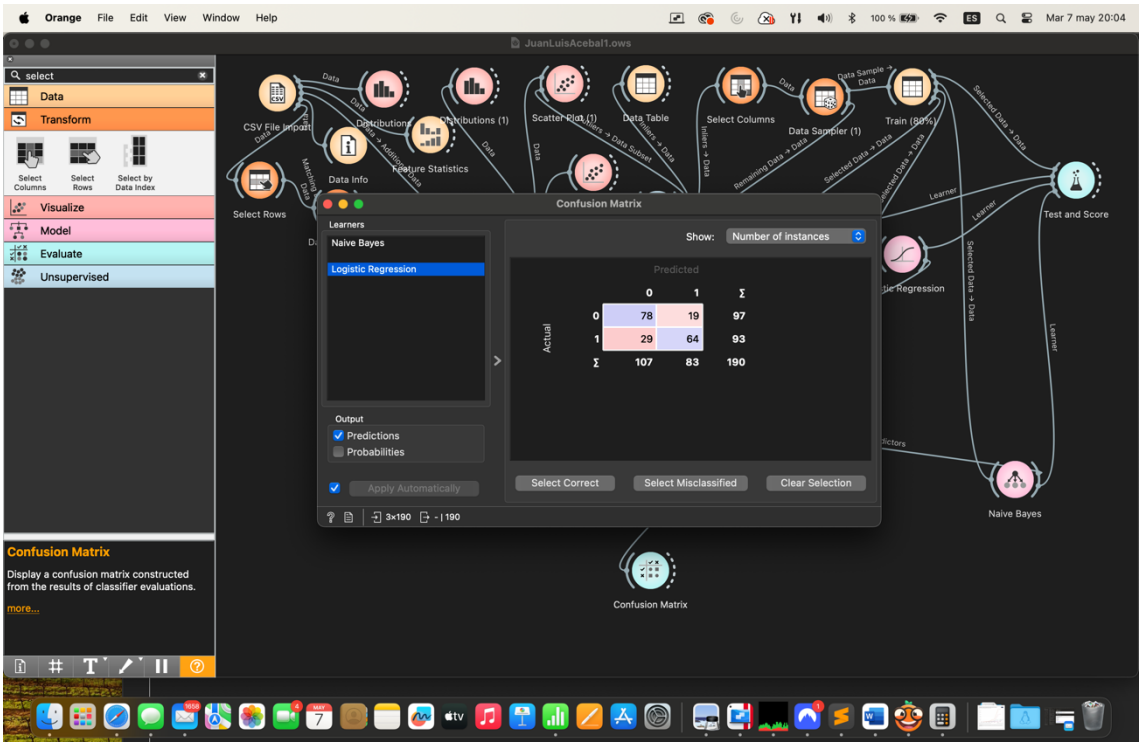
Podemos observar que tiene una tasa de acierto de 73 pacientes que son negativos en diabetes y realmente lo son (Verdaderos negativos), y una tasa de fallo de 24 pacientes que has sido puesto positivo y no lo son (Falsos positivos). Respecto a pacientes diagnosticados con diabetes, tenemos 71 pacientes enfermos bien diagnosticados (Verdaderos positivos), y 22 que se diagnostican como sanos y realmente están enfermos (Falsos negativos).

REDES NEURONALES



Para redes neuronales, la situación es algo mejor en verdaderos negativos (VN), FP, FN y VP, aunque como hemos visto en Naive Bayes, tiene un rendimiento peor que éste en VP se refiere.

REGRESION LOGISTICA



Es mejor que las otras 2 en VN (78 casos), se comporta para cuando no hay enfermedad mejor de igual manera que Naive Bayes se comporta mejor cuando hay enfermedad.

CONCLUSION

Para cuando es negativo (0), la red neuronal sigue siendo en general la mejor, y tiene unos valores más homogéneos que los otros modelos, tanto en el caso general como cuando el objetivo es 0 o 1. Además tiene menos falsos positivos y negativos. Sin embargo, quizás se podría valorar usar la regresión logística para filtrar personas sanas.

Model	AUC	CA	F1	Prec	Recall	MCC
Naive Bayes	0.841	0.758	0.760	0.768	0.753	0.516
	0.844	0.768	0.778	0.762	0.794	0.537
Logistic Regression	0.838	0.747	0.765	0.729	0.804	0.496

Para la clase 0 la regresión logística tiene recall más alto, quiere decir más positivos (mas positivos en cero) pero la precisión es más baja, lo que indica más falsos positivos.

Para la clase 1, tenemos una situación parecida entre Naive Bayes y red neuronal. Siendo mejor en este caso Naive Bayes en recall pero teniendo menos precisión.

Model	AUC	CA	F1	Prec	Recall	MCC
Naive Bayes	0.841	0.758	0.755	0.747	0.763	0.516
	0.844	0.768	0.758	0.775	0.742	0.537
Logistic Regression	0.838	0.747	0.727	0.771	0.688	0.496

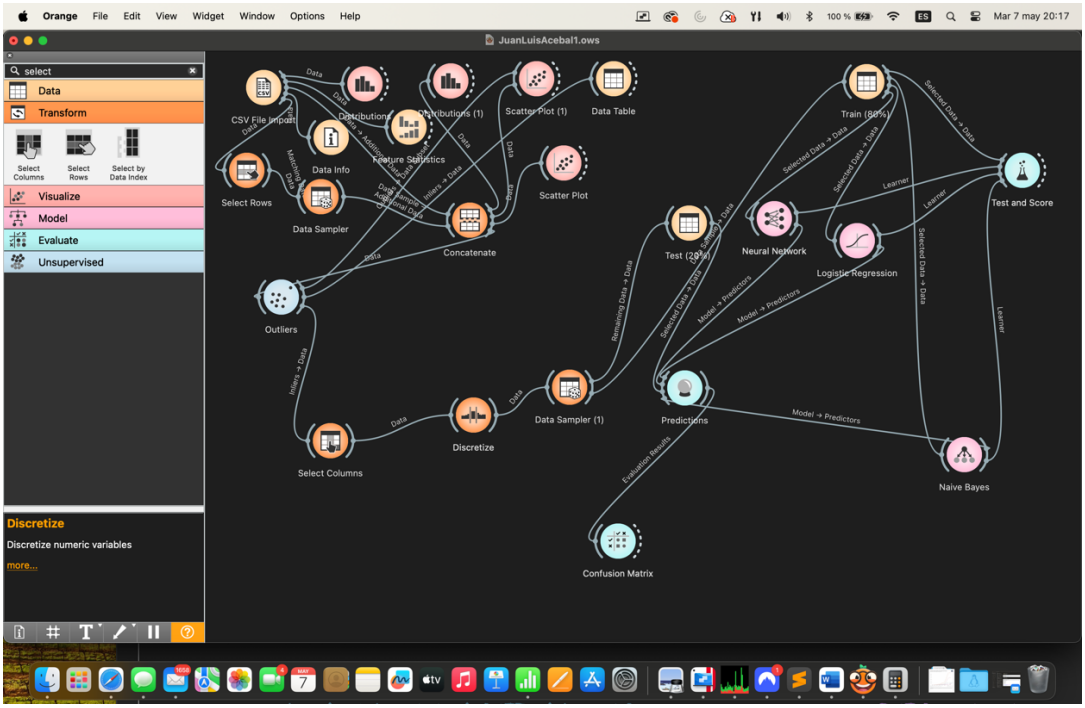
Esto significa que, en este caso, donde quizás la política sanitaria más eficaz pueda ser tener posibles casos de diabetes para hacerles estudios más profundos, interesa más usar en general red neuronal para una primera clasificación y luego con Naive Bayes buscar predicciones positivas (1, enfermedad).

Combinar red neuronal y regresión logística creo que no interesa, ya que el mal menor es predecir equivocadamente un diagnóstico en diabetes respecto al mal mayor, predecir una persona que está sana y que realmente no es así. Un ejemplo

de porque interesa más o menos un modelo, es que no es lo mismo detectar un paciente como negativo cuando es positivo, que sea positivo y sea detectado como negativo. Si esto lo vemos en diferentes sectores, no pasa igual cuando se trata la vida de una persona, de un billete falso, la necesidad de encender la calefacción, o la necesidad de abono de una planta...

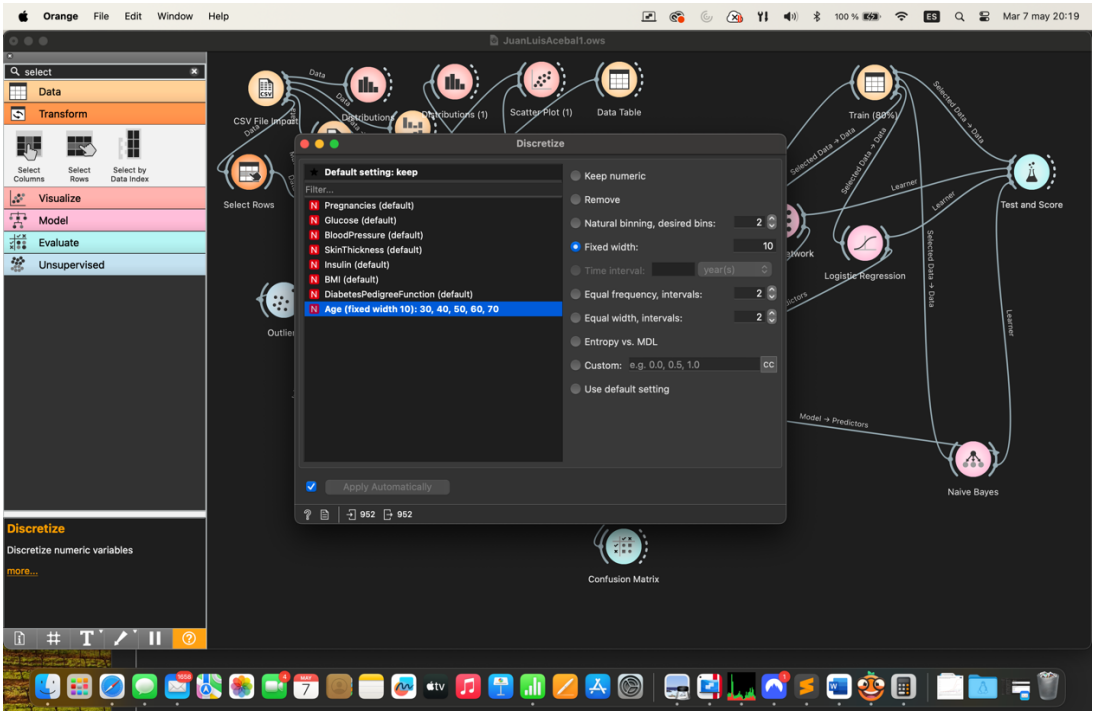
Según el caso de estudio, es necesario o importante centrarse en la predicción positiva o viceversa, y sabiendo que los errores de predicción pueden costar muy caros o baratos según el sector. Aquí, nos importa que los **errores de tipo 2** sean los **menos posibles**.[3]

PREGUNTA 3.3.A



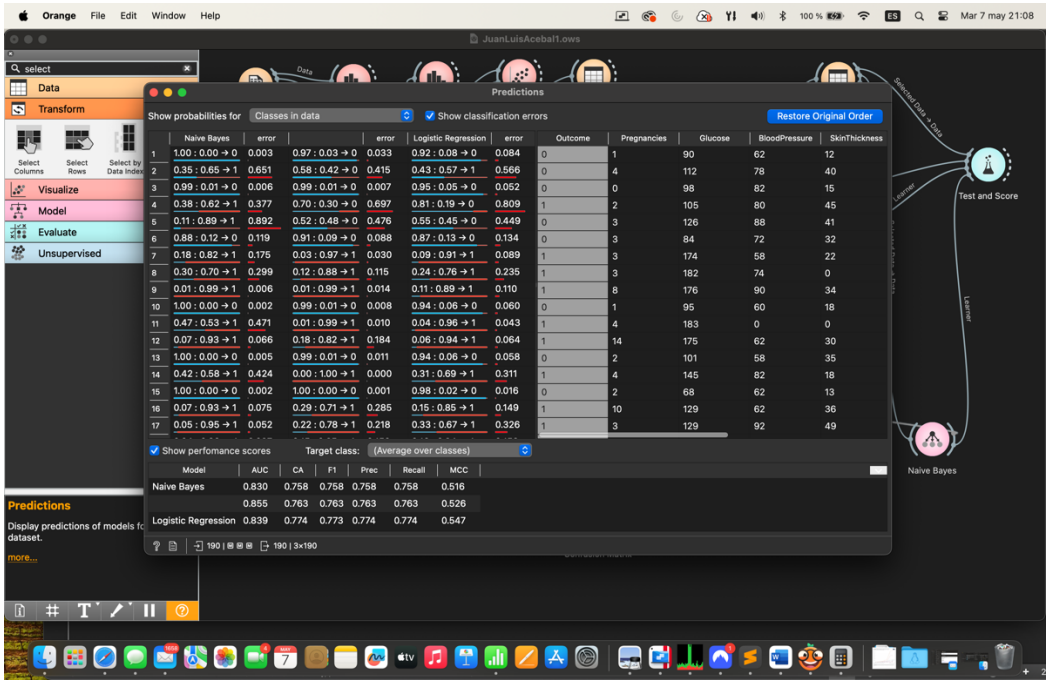
Introduzco Discretize entre Select Columns y Data Sampler para poder optimizar el modelo y categorizar edad u otras variables.

CATEGORIZACION DE AGE



Aquí convertimos en categorías age. Eso quiere decir que no es un numero tener 25 años, sino que pertenecerá a una categoría en la que se incluyen las personas entre 20 y 29 años.

ESTADISTICAS

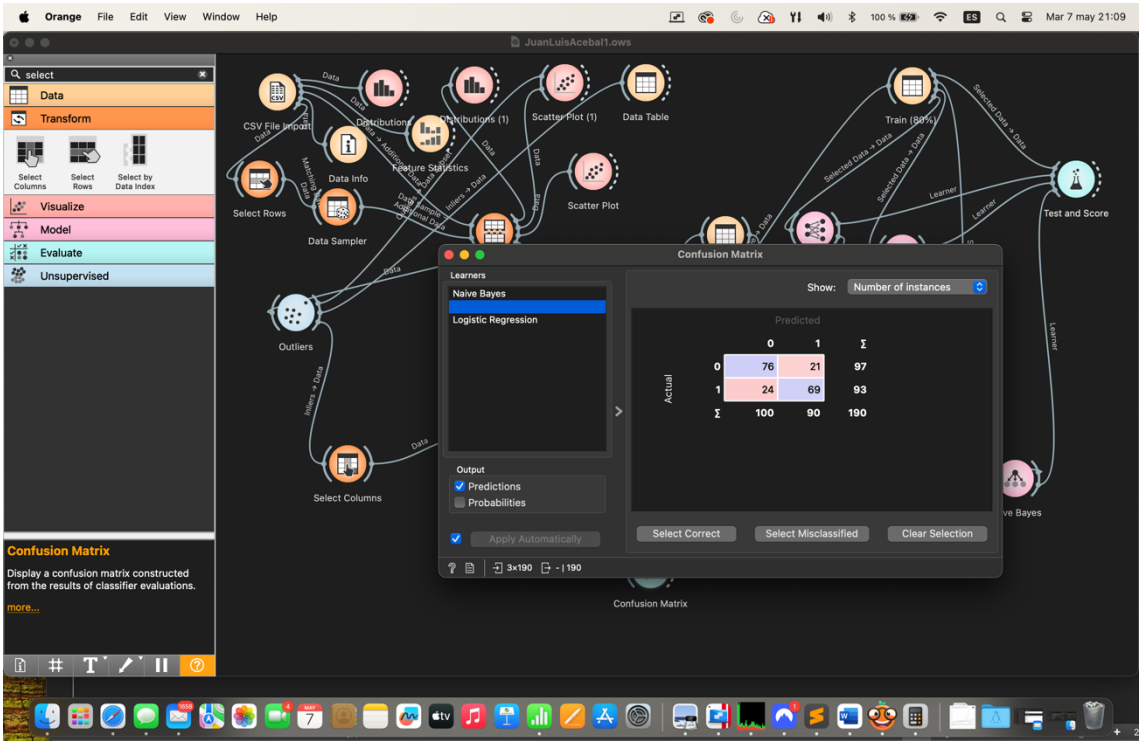


Han cambiado sensiblemente en algunos valores, se ve como el recall ha bajado, se observa si se compara con las estadísticas de antes de la modificación:

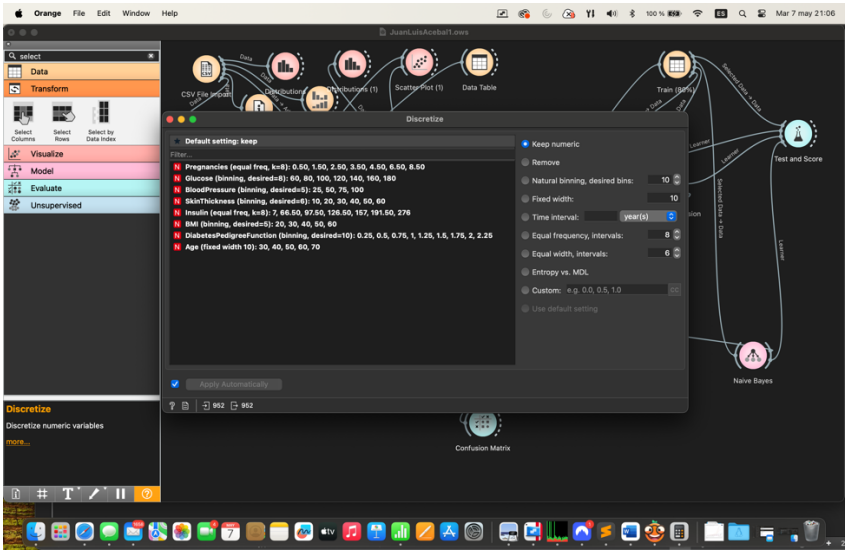
Model	AUC	CA	F1	Prec	Recall	MCC
Naive Bayes	0.841	0.758	0.758	0.758	0.758	0.516
	0.844	0.768	0.768	0.769	0.768	0.537
Logistic Regression	0.838	0.747	0.746	0.750	0.747	0.496

Aunque no son significativos, es decir, cambia, pero no es un gran cambio ni a peor ni a mejor.

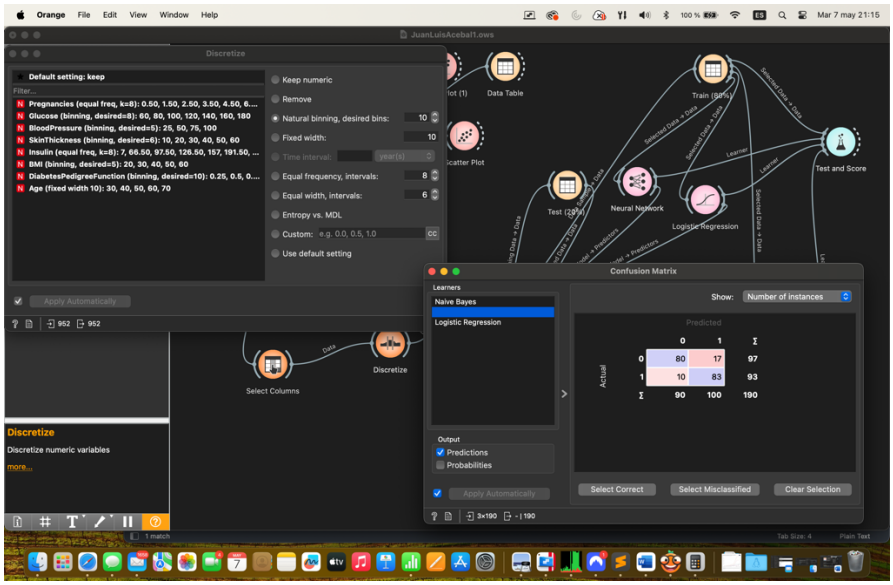
Vemos por ejemplo que ha bajado solamente en un VN:



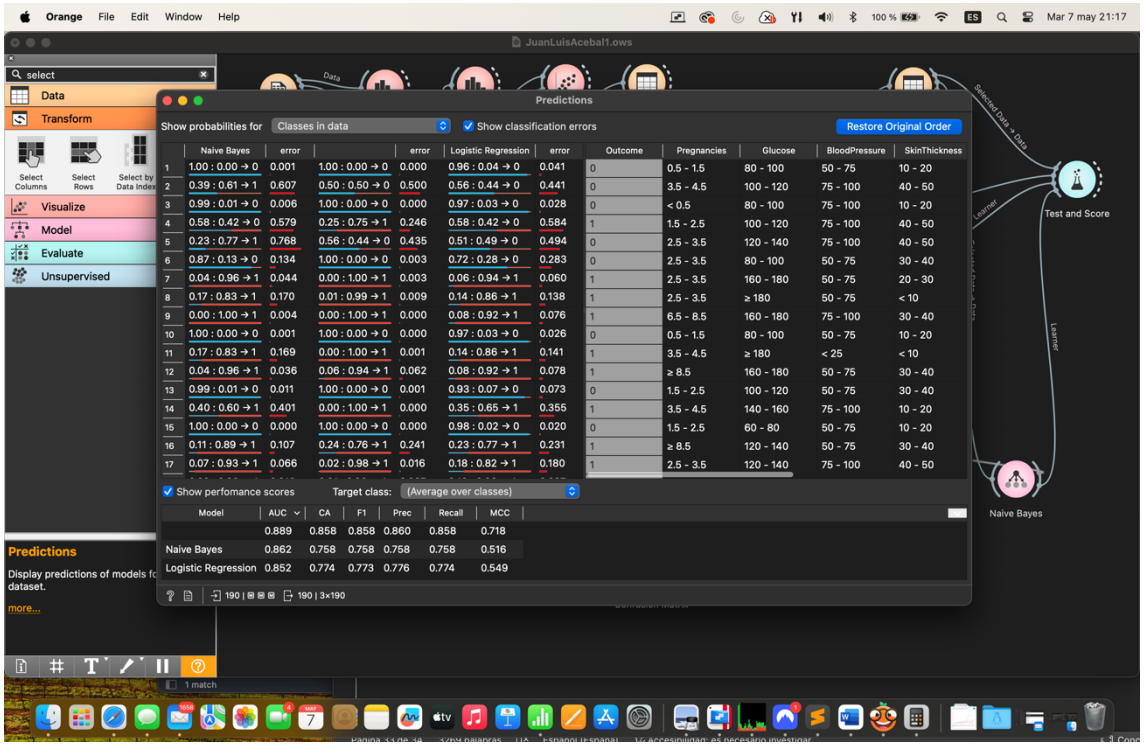
PREGUNTA 3.3.B



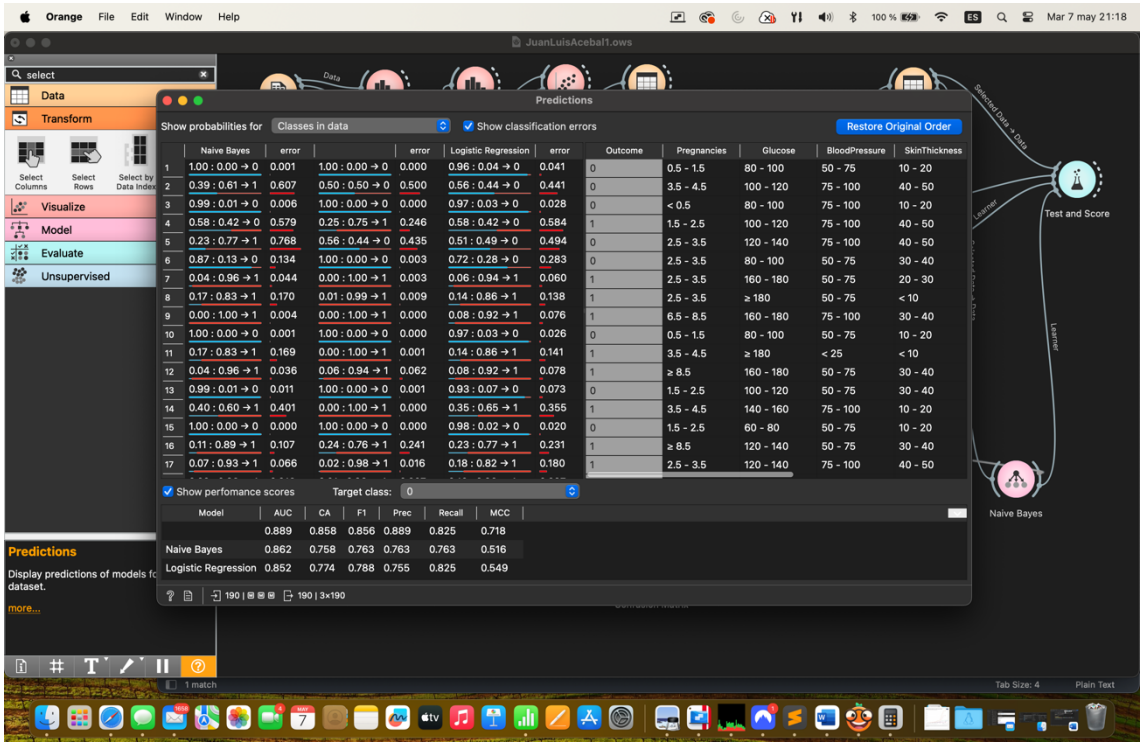
La matriz de confusión:



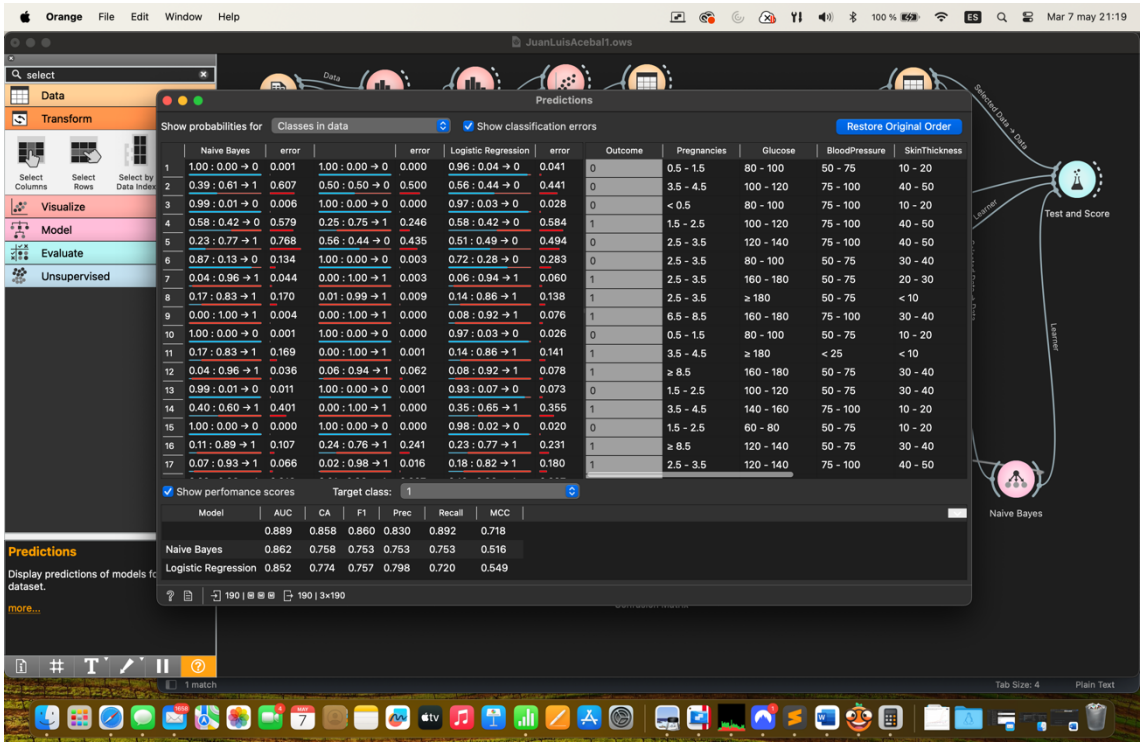
MI MODELO PARA TODAS LAS CLASES



MI MODELO PARA LA CLASE 0



MI MODELO PARA LA CLASE 1



Como se puede observar, he explorado y he categorizado con diferentes argumentos en todas las variables, probando y jugando con decenas de combinaciones y experimentos, eliminando variables incluso.

Después de varios intentos he llegado a esta configuración donde el recall es realmente alto, sin embargo, no solo sobresale el recall, también las métricas AUC, CA, F1, y una precisión alta, pero no alcanzando estos niveles. MCC es alto también, demostrando la ausencia de aleatoriedad y haciendo que sea consistente y fiable la predicción de mi modelo. Destacaría que he intentado hacer muchas combinaciones con los parámetros y finalmente he intentado beneficiar las estadísticas para la clase 1 que para la clase 0 aunque las dos son excepcionales.

Esto es así ya que he beneficia, en cualquier caso, pero particularmente en el caso de Falsos negativos, ya que en la medicina es más dañino para el paciente ser filtrado con un diagnóstico como este que alguien que necesita atención y cuidados, fuera asignado al grupo de personas sanas.

ENUNCIADO 4

PREGUNTA 1.1. LA SENSIBILIDAD O RECALL

Está definida por la proporción de verdaderos positivos (un billete verdadero detectado como verdadero) respecto al total de casos (todos los billetes verdaderos positivos más los billetes falsos negativos)

Sería su fórmula recall=VP/(VP+FN)

PREGUNTA 1.2. ESPECIFICIDAD

Mide la proporción de los VN (billete falso detectado como falso, o dicho de otra forma, billete negativo verdaderamente) que han sido realmente detectados como VN respecto al total de la suma de VN y FP

Especificidad= VN/(VN+FP)

PREGUNTA 1.3. DETECCIÓN DE ENFERMEDADES.

Es muy importante recall, como ya he dicho durante todo el ejercicio, en este caso lo importante es minimizar los falsos negativos, y por tanto recall en este sector es imprescindible.

La especificidad es igualmente importante pero un error de tipo II es más peligroso en el campo de la salud que en otro lugar. Los errores de tipo I que serían los derivados de especificidad más baja, se deben evitar en la medida de lo posible pero no son tan imprescindibles que en el caso de un no diagnostico por ejemplo de una enfermedad terminal.

PREGUNTA 2.1. LA PRECISIÓN

Es la proporción de identificaciones positivas que fueron realmente positivas. Para lo que concierne a la diabetes, es una métrica interesante junto a recall en el contexto médico, ya que dependiendo la escala y el contexto, nos tendremos que afinar más el modelo ya que no es lo mismo llamar al 20% de la población de una ciudad por un riesgo sanitario sabiendo que podría ser un 5% si el modelo estuviera bien ajustado, ya que aquí de lo que habla la precisión es la suma de los “enfermos más los no enfermos categorizados como enfermos”

Precisión=VP/(VP+FP)

PREGUNTA 2.2. COMBINACIÓN DE LA MÉTRICA F1.

F1 score es una media de precisión y sensibilidad, es decir:

Formula desglosada:

F1=2VP/(2VP+FP+FN)

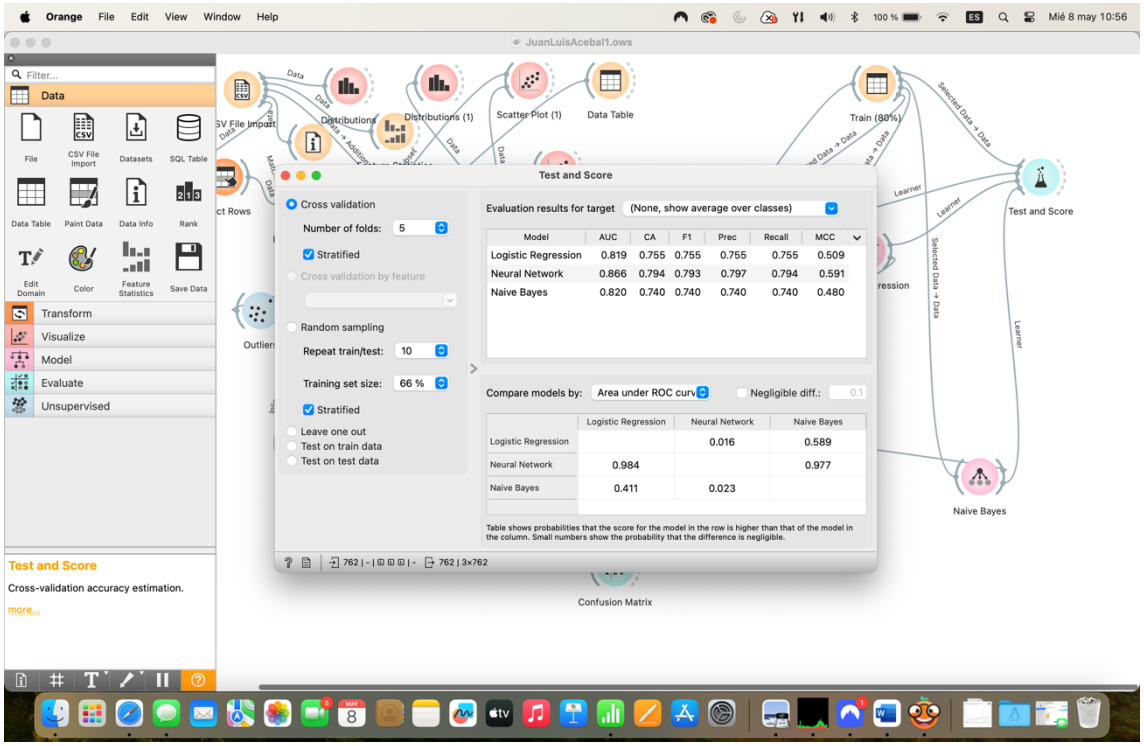
La fórmula se obtiene de:

F1=2(Precisión x Sensibilidad) /(Precisión + Sensibilidad)[2]

Es útil cuando se busca un balance entre precisión y sensibilidad ya que al ser una media armónica, que sin entrar a explicarlo, la media armónica suele ser igual o menor a la media aritmética, es decir, F1 score será cercano al menor de precisión o sensibilidad, entonces si hay un balance entre ambos, F1 score será alto, y si uno de ellos es bajo, F1 score será casi tan bajo, por decirlo así, la media armónica penaliza mucho a los outliers que se utilizan para calcularlo, entonces si no hay un equilibrio entre ambos será baja, es por eso de que es útil cuando se busca un balance entre precisión y sensibilidad.

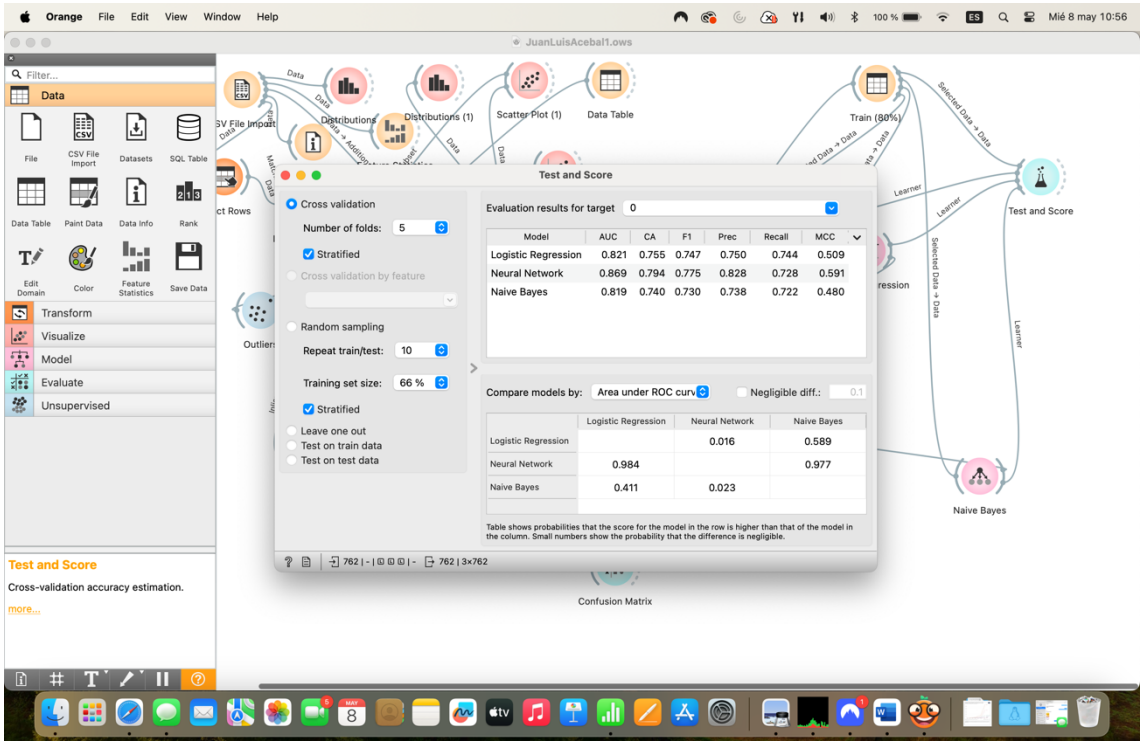
PREGUNTA 3

CLASES 0 Y 1



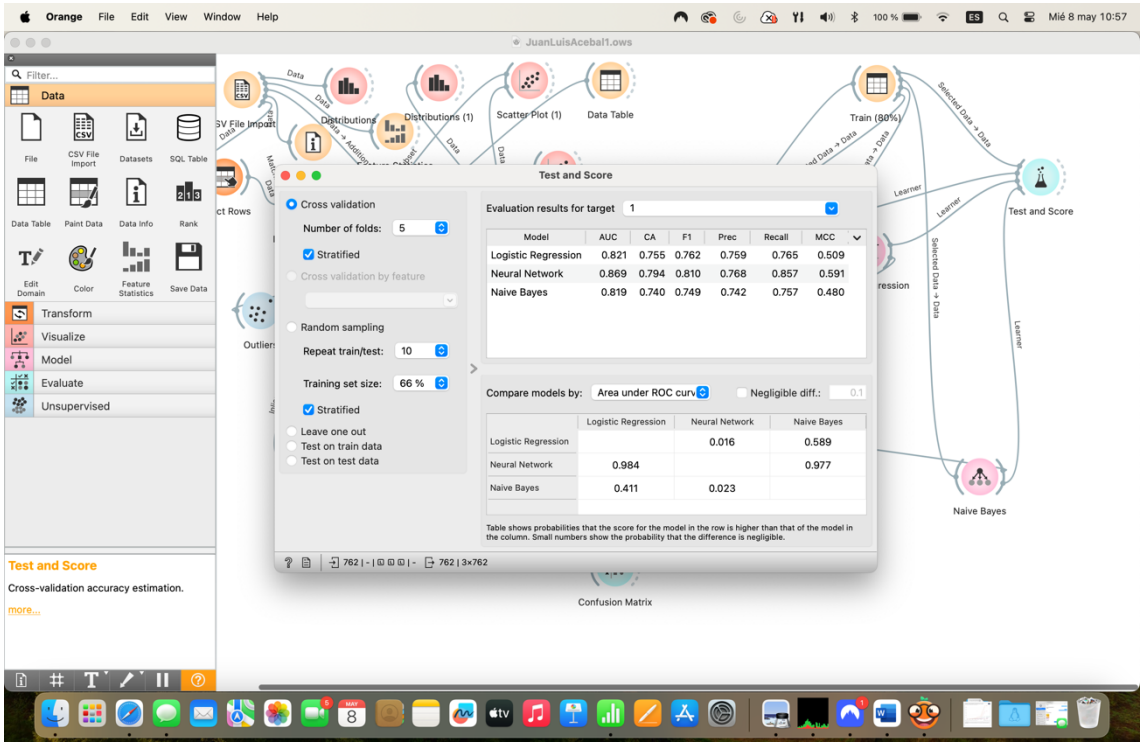
	Las dos clases (0), (1)		
Modelo	**Precisión**	**Recall**	**Análisis**
Regresión logística	0.755	0.755	Alta precisión, alto recall. Buen comportamiento, tiene un equilibrio entre métricas y no es tan bueno como red neuronal, pero lo hace bien.
Red neuronal	0.797	0.794	Alta precisión, alto recall. Se acerca a una tasa muy alta en las métricas, buen equilibrio entre ellas, siendo especialmente bueno en la clase 1, que es la importante en este caso de estudio (sector salud). Además, tiene recall muy alto en la clase 1, eso es una noticia excelente.
Naive Bayes	0.740	0.740	Baja precisión, bajo recall. Es el que peor comportamiento tiene de los 3 modelos, pero aun así tiene unas métricas aceptables

CLASE 0



	Clase negativa (0)		
Modelo	**Precisión**	**Recall**	**Análisis**
Regresión logística	0.750	0.744	Alta precisión, bajo recall. Rendimiento balanceado, adecuado para identificar no eventos.
Red neuronal	0.828	0.728	Alta precisión, bajo recall. Alta precisión, pero un recall justo en comparación, excelente para evitar FP para la clase. (FN para la muestra)
Naive Bayes	0.738	0.722	Baja precisión, bajo recall. El más bajo de los 3 modelos, no es tan confiable para negativos claros.

CLASE 1



	Clase positiva (1)		
Modelo	**Precisión**	**Recall**	**Análisis**
Regresión logística	0.759	0.765	Alta precisión, alto recall. Buen balance entre precisión y recall. Es efectivo, pero no es el más optimo.
Red neuronal	0.768	0.857	Alta precisión, muy alto recall. Muy bueno en recall, el mejor, muy efectivo para identificar positivos, pero con falta de precisión (dentro del equilibrio entre ambas métricas, de base la precisión es alta), detectando también FP
Naive Bayes	0.742	0.757	Baja precisión, alto recall. Balance normal, siendo el peor de los modelos comparando sus métricas.

CONCLUSION

La evaluación de los tres modelos desarrollados en la actividad, utilizando las métricas AUC, CA, F1, Precision, recall y MCC muestra un muy buen desempeño general de red neuronal, sin embargo, para la clase positiva (1), no es solo bueno, es excelente, y para el caso de estudio que trabajamos, es la clase más importante.

La regresión logística, que en sus métricas quedaría como la segunda, tiene un muy buen desempeño general, podría ser usada igualmente si la clase objetivo del estudio hubiera sido la clase negativa (0), ya que tiene un desempeño a nivel recall mejor que el resto para esta clase, tiene un rendimiento bueno, pero no sobresale al nivel de las redes neuronales.

El modelo Naive Bayes, si bien tiene unos resultados entre aceptables y altos, no sirve en el contexto del modelo desarrollado pues queda en última posición, no destaca en ninguna métrica y en ninguna clase concreta respecto al resto de modelos analizados, eso quiere decir que en modelos donde la precisión y recall son críticos, y teniendo ya modelos mejores, no puede ser una opción con el modelo analizado.

Quizás en otros modelos sea mejor, incluso con el mismo dataset, pero en el que yo he desarrollado no ha destacado. Sin embargo, tiene métricas en general altas, su peor métrica es recall 0.722 para la clase 0.

BIBLIOGRAFÍA

(1) https://ourworldindata.org/explorers/energy?facet=none&Total+or+Breakdown=Select+a+source&Energy+or+Electricity=Electricity+only&Metric=Share+of+total&Select+a+source=Low-carbon&country=USA~GBR~CHN~OWID_WRL~IND~BRA~ZAF

(2) https://www.britannica.com/science/F-score

(3) https://datascience.recursos.uoc.edu/es/errores-de-tipo-i-y-ii/?filter=.tecnicas-metodos-algoritmos