ANÁLISIS DE DATOS

Fase 5

Código: 202016908\_38

Presentado al tutor:

JAVIER LEONARDO PINEDA URIBE

Entregado por:

ÁLVARO CAMILO PEDREROS BENAVIDES

Universidad Nacional Abierta Y A Distancia - UNAD

Escuela de Ciencias Básicas Tecnología e Ingeniería

Diciembre de 2023

Bogotá D.C.

Resumen

En el siguiente documento de quiere poner en práctica lo aprendido de análisis de datos por medio de una base de datos que se encontró en el sitio web kaggle la cual corresponde a distintos datos que ayudan estimar el resultado de una enfermedad de arterias coronarias y este procedimiento se va a pasar a implementar por medio de dos métodos de análisis de datos el cual es el árbol de decisiones y el Naive Bayes para lo que al comienzo se va a hacer en análisis de los datos en bruto sin modificaciones y proceder a ver sus resultados, para después discretizar y depurar la información dada y hacer otra base de datos con esta información y hacer el mismo análisis de datos y el entrenamiento de los algoritmos a implementar para después ver sus resultados y hacer un análisis de lo obtenido

# Solución

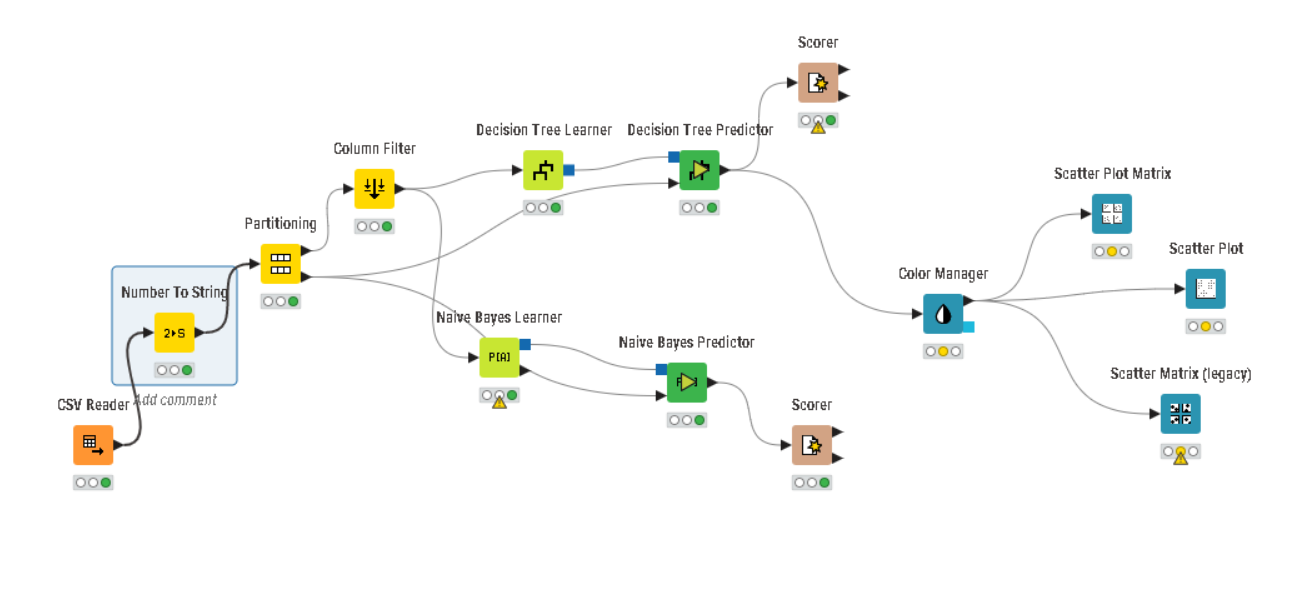
Antes de comenzar con el desarrollo del trabajo pasare a explicar la base de datos que se seleccionó, la cual tiene como nombre **Heart dataset csv** la cual contiene datos de posibles causas de enfermedad de las arterias coronarias, esta se puede encontrar en el siguiente link:

<https://www.kaggle.com/datasets/abhishekbagwan/heart-dataset-csv/data>

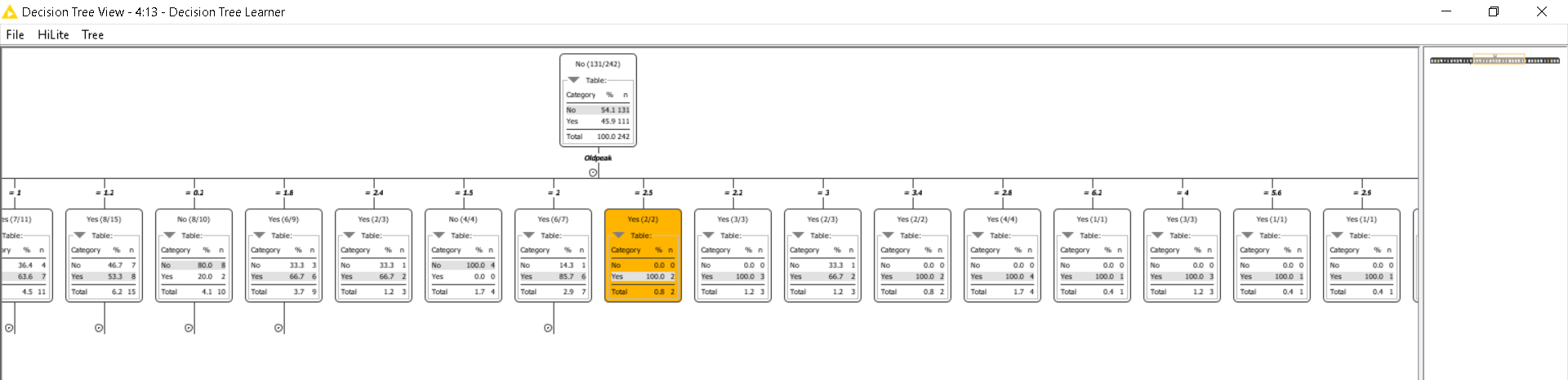
y contiene 303 datos con los siguientes nombres de columnas:

* ChestPain: Dolor en el pecho
* RestBP: Presión arterial en reposo
* Chol: Colesterol
* Fbs: Nivel de azúcar en sangre en ayunas
* RestECG: Electrocardiograma en reposo
* MaxHR: Ritmo cardíaco máximo
* ExAng: Angina inducida por ejercicio
* Oldpeak: Depresión del segmento ST inducida por el ejercicio
* Slope: Pendiente del segmento ST durante el ejercicio
* Ca: Número de vasos sanguíneos principales coloreados por la fluoroscopia
* Thal: Resultado de la prueba de esfuerzo nuclear
* AHD: Enfermedad de las arterias coronarias

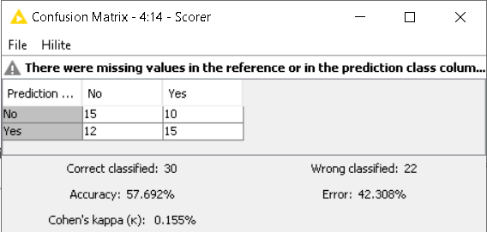
Siendo la columna **AHD** el dato que se quiere procesar como resultado por lo que para comenzar en análisis se va a hacer por medio de 2 algoritmos los cuales son arboles de decisión y Naive Bayes y se procede a hacer un filtro de información ya que para estos dos casos los algoritmos requieren de datos tipo string como mostraremos a continuación



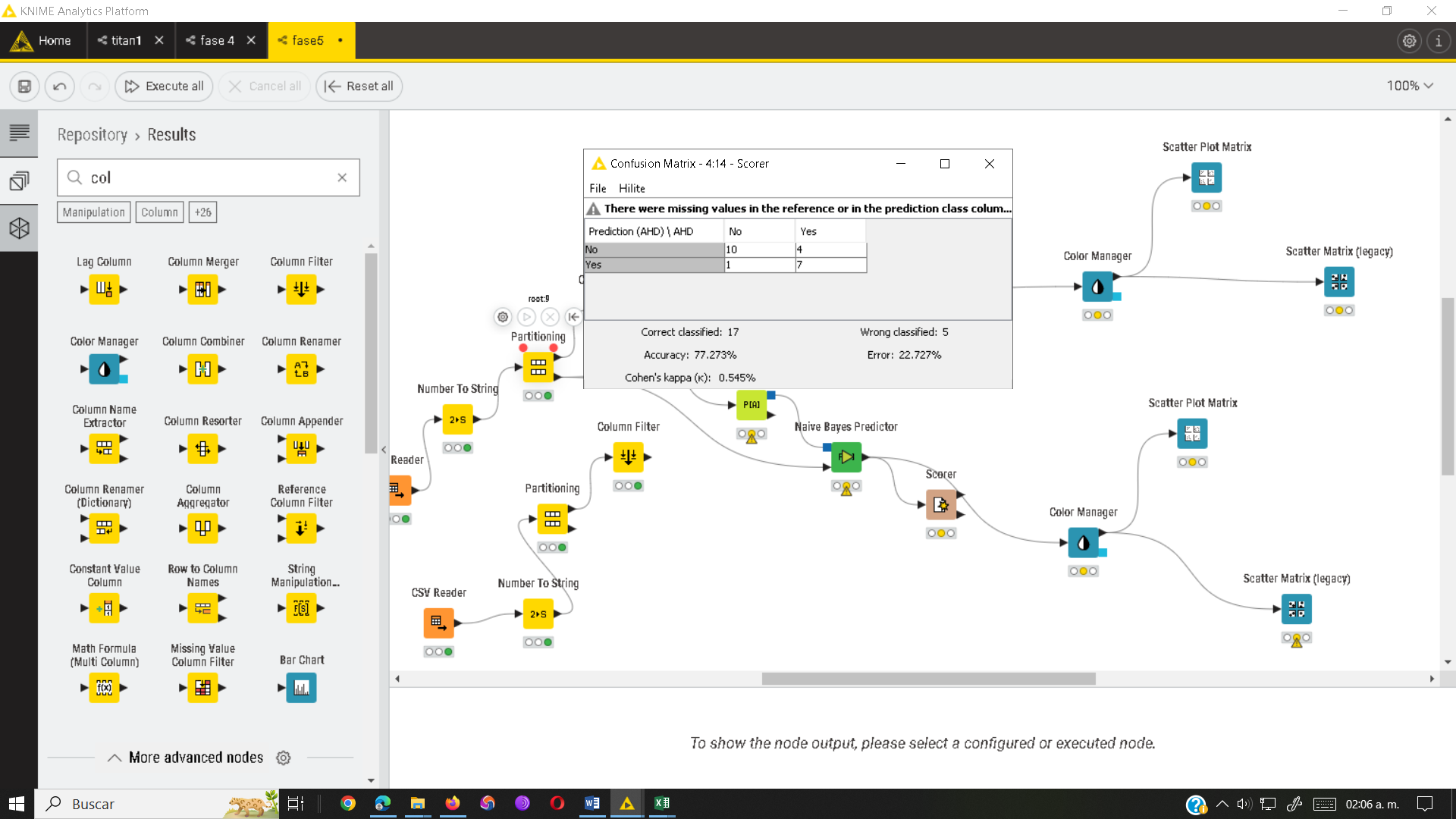
Ya después de pasar los datos a string podemos hacer un filtro para no ingresar los datos por completo para el entrenamiento de los algoritmos por lo que se coloca un partidor usando un 80% de los datos por lo que tenemos los siguientes resultados primero miramos el árbol de decisiones donde tenemos las siguientes respuestas



Donde vemos que según el análisis del árbol una de las principales causas de esta falencia es el Oldpeak ocupando uno de los primeros nodos de la misma para después los siguientes nodos son la edad, el calcio, y la presión baja para después pasar a hacer un análisis del 20% de datos que se quedó para ser analizados y obtenemos como resultado la siguiente imagen

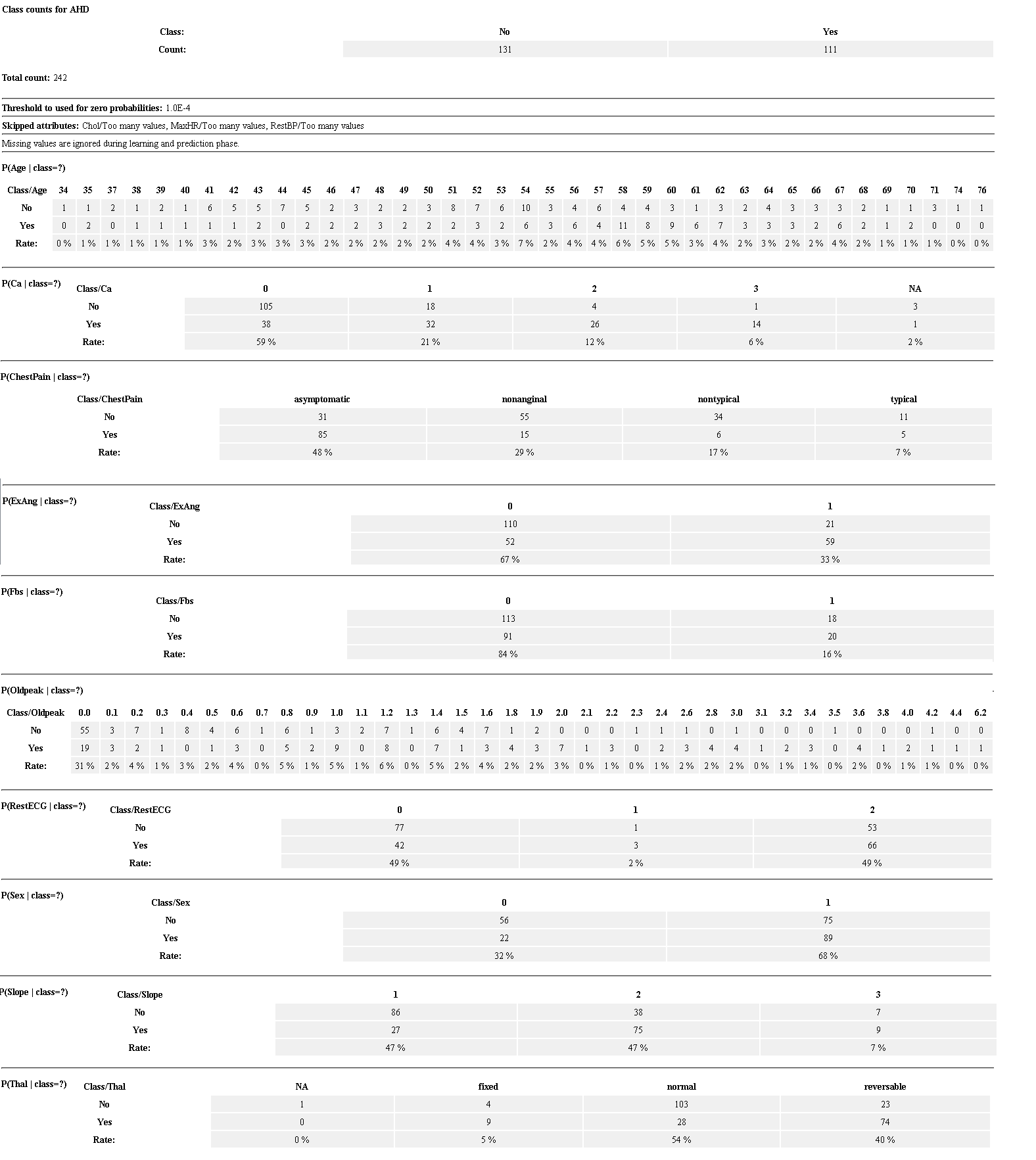


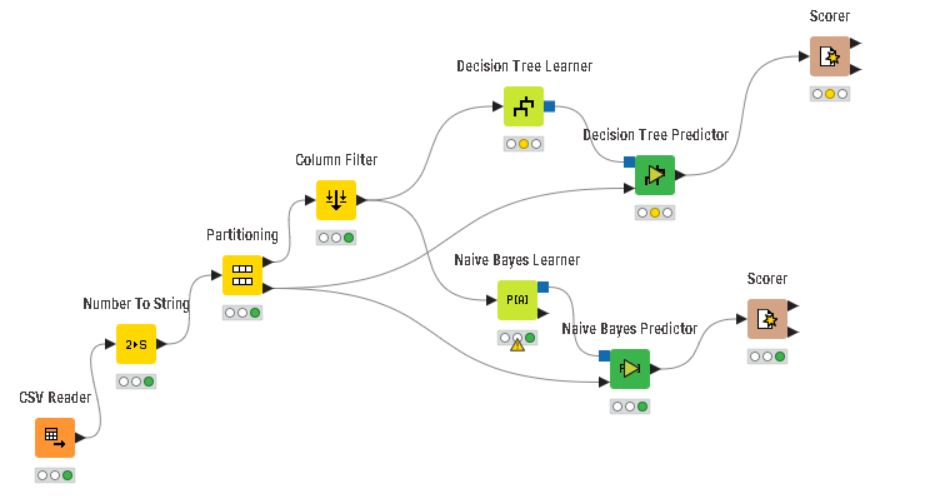
Con el entrenamiento dado y los datos ingresados vemos que tenemos un error del 42.3% y 22 datos fueron mal catalogados y podemos pensar que este error puedes ser causado por el tamaño del árbol ya que tomo uno delos primeros criterios a Oldpeak el cual tiene 33 variables o por el tamaño de los datos usados para el entrenamiento por lo que en el siguiente caso se pasara a entregar en algoritmo con un 90% de los datos dejando 10% para ser analizado



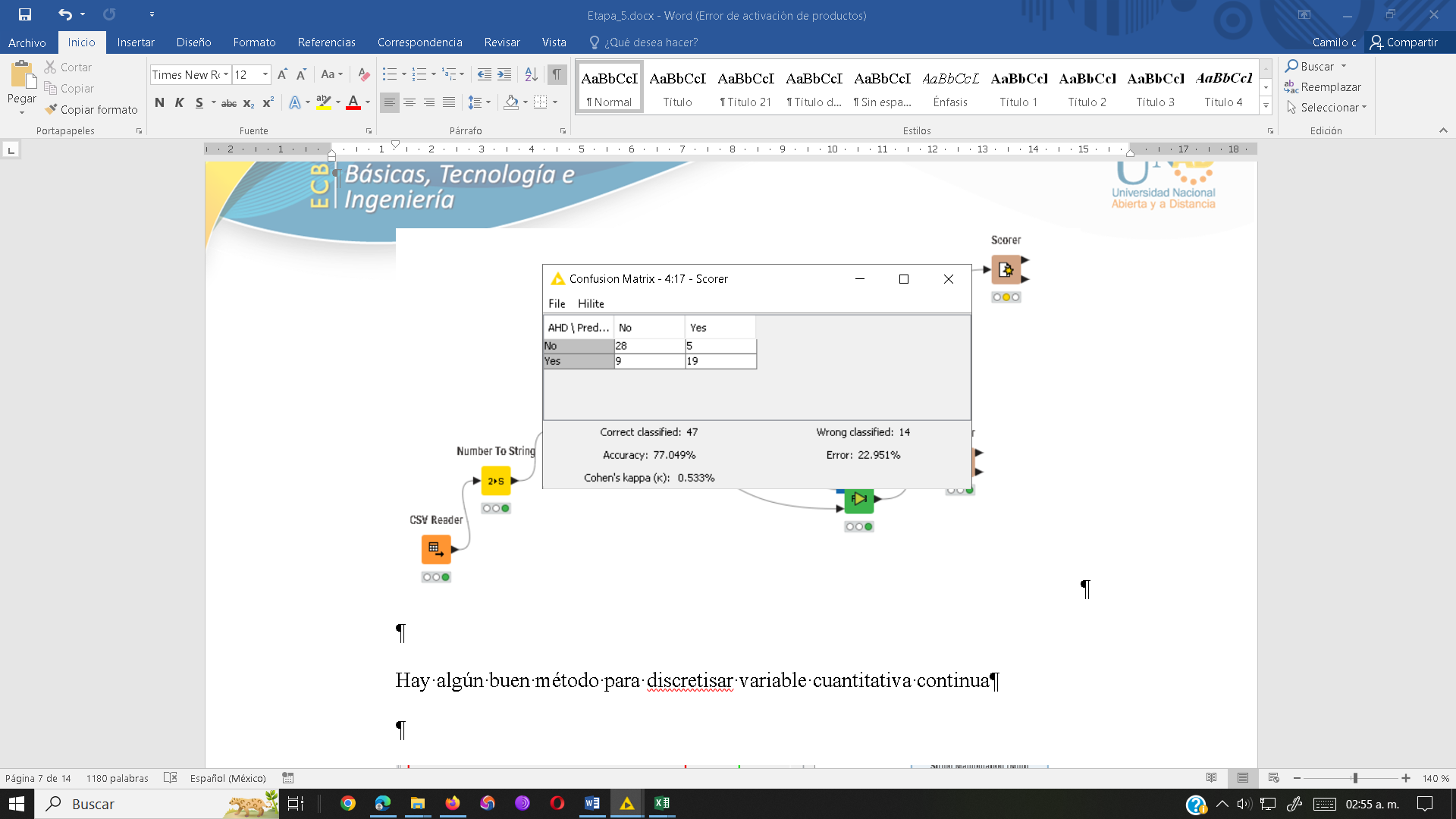
Al momento de hacer el entrenamiento con el 90% de los datos podemos cer que mejoro el método de clasificación obteniendo un error dl 22.7% disminuyendo a la mitad el nivel de error y de igual manera paso con los datos mal clasificados, a un que en este caso se tienen menos datos para ser clasificados

# Naive Bayes

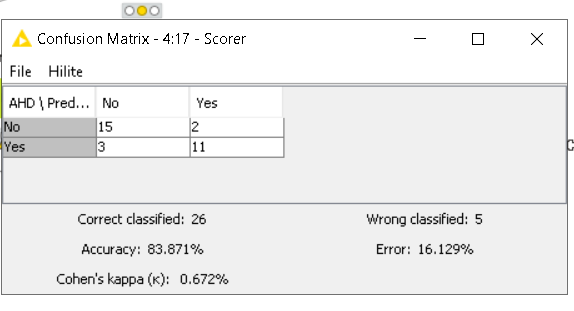
Al momento de hacer en entrenamiento de Naive Bayes podemos obtener el siguiente resultado donde vemos que por cada una de las variables compara con el resultado querido dando un porcentaje a cada variable como se ve a continuación



Ya teniendo el algoritmo de Naive Bayes entrenado pasamos a validar los resultados de igual manera como se realizó el árbol de decisión donde se dejó un 80% de datos a ser analizados donde se obtuvo el siguiente resultado



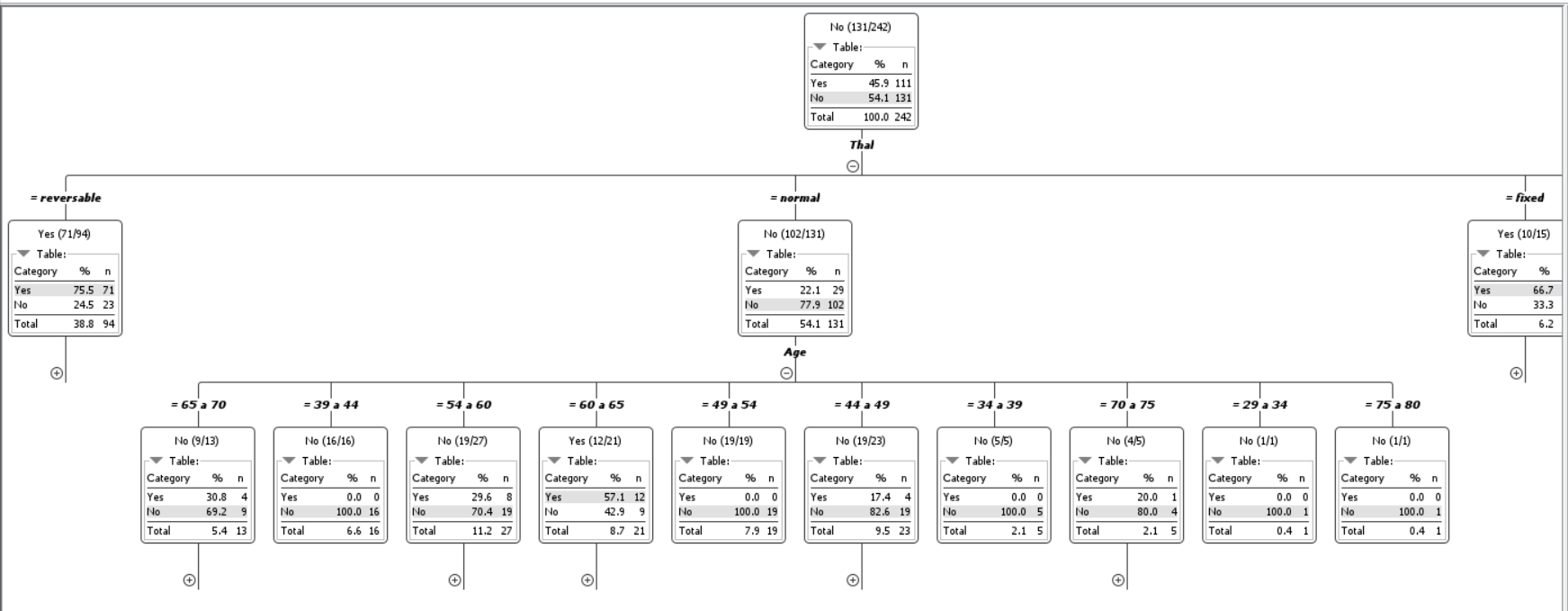
En este caso podemos ver que después de ingresar los datos tenemos una precisión de un 77% y un error del 22.9%, a pesar de que solo se ha ingresado el 80% de los datos para hacer en entrenamiento del algoritmo por lo que podemos pensar que este es mucho más indicado para el tipo de los datos ingresados, por lo que pasaremos a hacer el análisis con el 90% de los datos, dejando el 10% para el análisis como veremos a continuación



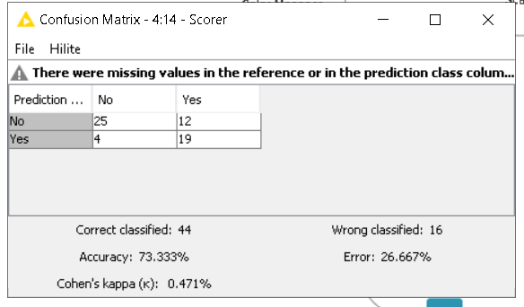
Ya teniendo el 90% de los datos vemos que ha mejorado mucho la precisión de la predicción teniendo un 83.8% precisión y podemos pensar que este método de análisis es el mejor para este tipo de datos ya que en su gran mayoría con datos son variable cuantitativa continua y pues en ese caso se quiere saber qué pasaría si discretizamos los datos creando otra base de datos nueva con la misma información y agrupando los datos

# Árbol de decisión datos discretos

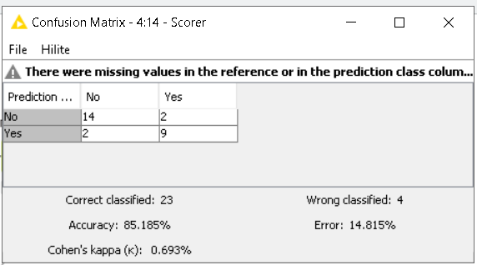
Para el momento de hacer el análisis del árbol de decisión se discretizaron todos los datos con constantes mayores a 10 y se agruparon en 10 grupos por medio del método de agrupación de variable cuantitativa continua, por lo que podemos ver en el árbol es que uno de los primeros nodos ya no es el Oldpeak sino la variable Thal que corresponde a el Resultado de la prueba de esfuerzo nuclear y después de estos se pasa a la edad y a medir el calcio en sangre y podemos notar que el árbol de decisión ya está mucho más compacto y se ve mucho más eficiente como se ve en la siguiente imagen



Ya teniendo el árbol de decisiones entrenado con el 80% de los datos pasamos a hacer la validación de su eficiencia como se ve a continuación



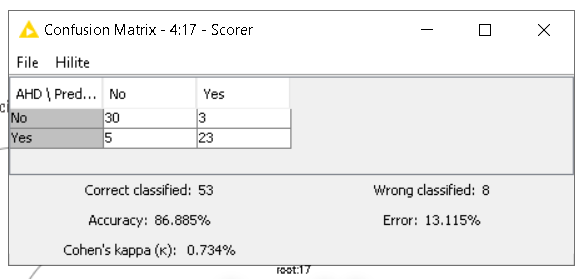
Ya con esto podemos ver que la eficiencia paso de un 57% a un 73.3% con tan solo un 80% de datos para el entrenamiento así que pasaremos a hacer el entrenamiento con un 90% de datos como se ve a continuación



Ya para este momento vemos que ha mejorado la eficiencia del algoritmo de árbol de decisión al discretizar los datos y darle un 90% de los datos para entrenarlo obteniendo un 85% de precisión al momento de leer los datos ingresados

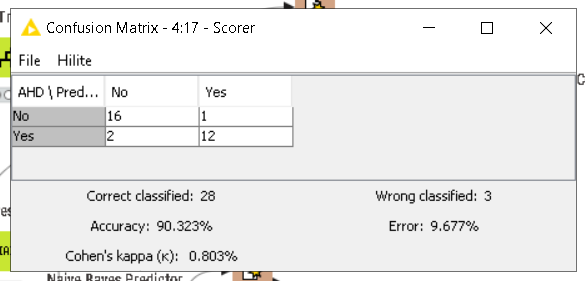
# Naive Bayes datos discretos

Ya que anteriormente pasamos a agrupar los datos de nuestra base de datos para el método de árbol de decisión ahora se pasará a hacer el mismo análisis con un 80% de los datos para el entrenamiento de Naive Bayes como vemos a continuación



Y podemos ver que pasamos de un 77% a un 86.8% de precisión al momento de hacer el análisis de los datos

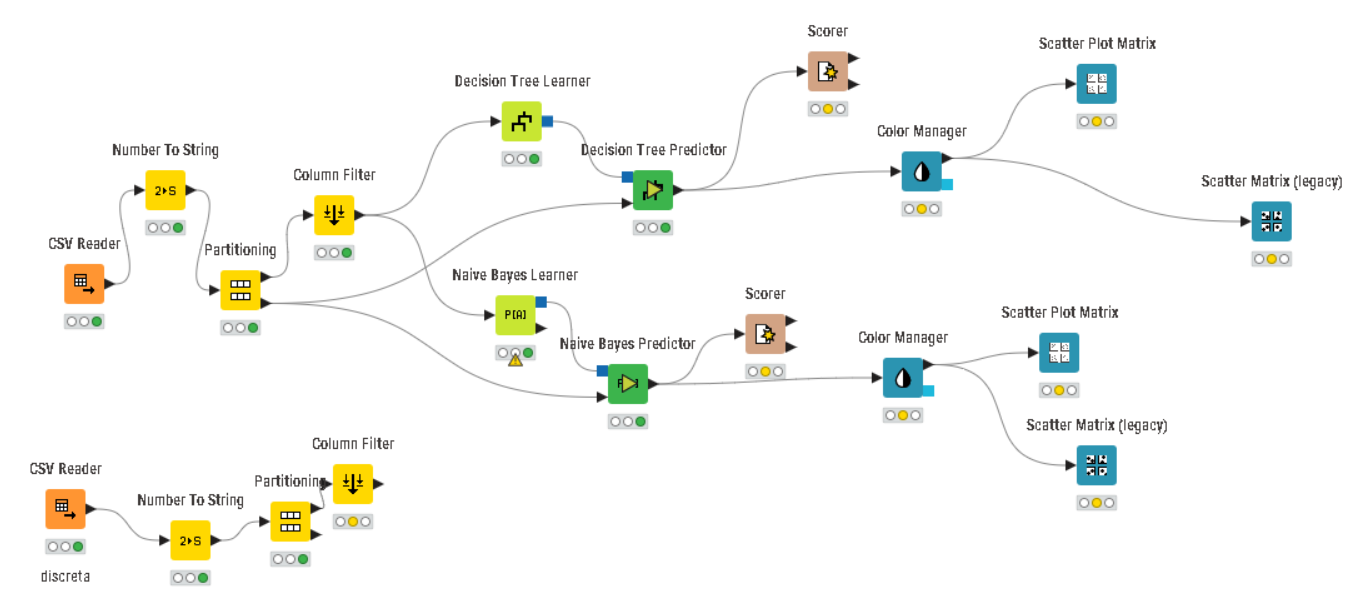
Para continuar se realizó el entrenamiento con un 90% de los datos obteniendo el siguiente resultado



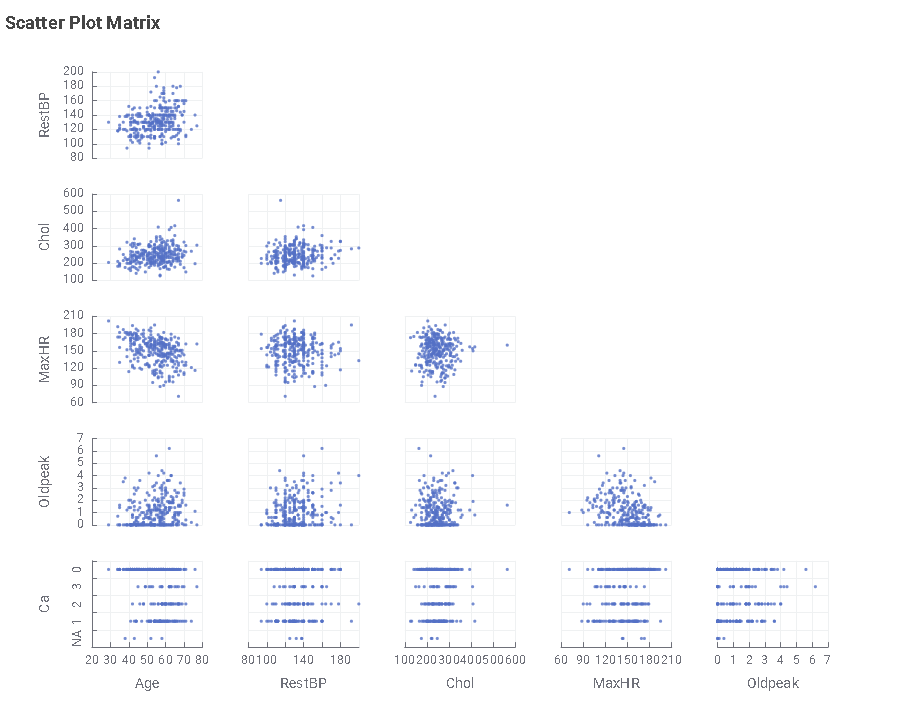
Y pasamos de un 83.8% a un 90% de precisión viendo que la eficiencia de los algoritmos mejora

# Análisis de datos por medio de matriz de puntos

Antes de hacer el análisis los datos con mas correlaciónese tiene la siguiente configuración de los algoritmos ya que para hacer análisis de datos por puntos es mejor usar los datos normales como se ven a continuación

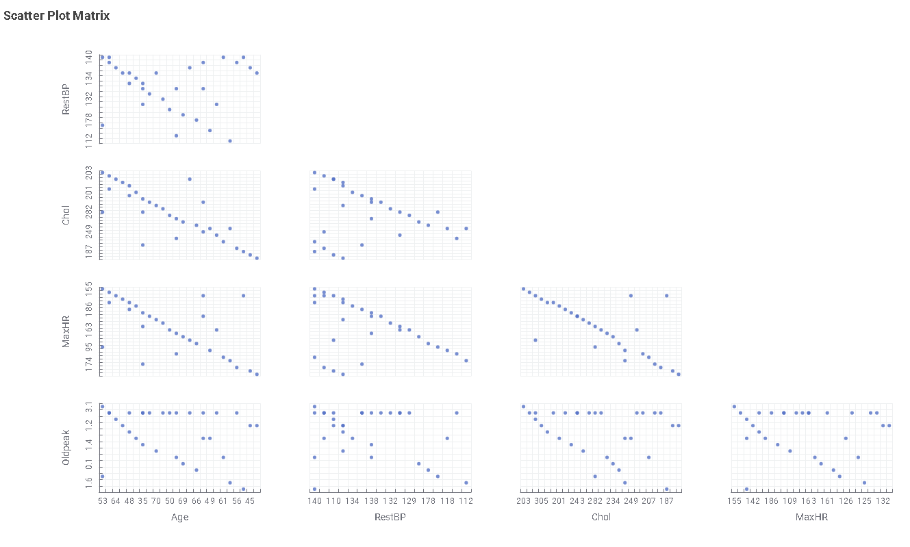


Antes de validar los datos se hace un muestreo de la información sin procesar con las variables cuantitativas continuas de la tabla, se tomaron estas debido a que se tendrán más puntos para graficar obteniendo la siguiente información



* RestBP: Presión arterial en reposo
* Chol: Colesterol
* MaxHR: Ritmo cardíaco máximo
* Oldpeak: Depresión del segmento ST inducida por el ejercicio
* Ca: Número de vasos sanguíneos principales coloreados por la fluoroscopia

Al momento de hacer el análisis de los datos por medio de matriz de puntos se validó si los dos algoritmos tenían resultados distintos, y en este caso los resultados de los datos son exactamente iguales y obtenemos el siguiente resultado con las variables seleccionadas

podemos ver que la edad es uno de los primeros factores correlacionados con el nivel de presión arterial en reposo, con el colesterol con el ritmo cardiaco máximo y el ritmo cardiaco inducido por el ejercicio como vemos tiene una relación linealmente proporcional donde mientras más aumenta la edad aumentan los factores de riesgo, y de igual manera mientras más alto es el ritmo cardiaco en reposo mas se ve aumento en el colesterol y otro dato que denota buen estado físico es el del ritmo cardiaco en reposo puede tener un mayor cardiaco al momento del ejercicio

Conclusiones

Referencias

Arce, J. I. B. (2023). La matriz de confusión y sus métricas. *Juan Barrios*. https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

*Coeficiente de determinación (R cuadrado) - MATLAB & Simulink - MathWorks América Latina*. (s. f.). https://la.mathworks.com/help/stats/coefficient-of-determination-r-squared\_es.html

colaboradores de Wikipedia. (2023a). Coeficiente de Gini. *Wikipedia, la enciclopedia libre*. https://es.wikipedia.org/wiki/Coeficiente\_de\_Gini

colaboradores de Wikipedia. (2023b). Máxima verosimilitud. *Wikipedia, la enciclopedia libre*. https://es.wikipedia.org/wiki/M%C3%A1xima\_verosimilitud

colaboradores de Wikipedia. (2023c). Entropía (información). *Wikipedia, la enciclopedia libre*. https://es.wikipedia.org/wiki/Entrop%C3%ADa\_(informaci%C3%B3n)

Gonzalez, L. (2023). Curvas ROC y área bajo la curva (AUC). *🤖 Aprende IA*. https://aprendeia.com/curvas-roc-y-area-bajo-la-curva-auc-machine-learning/

*IBM documentation*. (s. f.). https://www.ibm.com/docs/es/cognos-analytics/11.2.0?topic=terms-adjusted-r-squared

*La ganancia de información y entropía*. (s. f.). www.yizinet.com. Recuperado 11 de septiembre de 2023, de https://www.yizinet.com/ganancia-de-informacion-y-entropia/

Team, D. S. (2021). Estimación de máxima verosimilitud. *DATA SCIENCE*. <https://datascience.eu/es/programacion/estimacion-de-maxima-verosimilitud/>

Dangeti, P. (2017). Statistics for Machine Learning. Packt Publishing. https://bibliotecavirtual.unad.edu.co/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1560931&lang=es&site=eds-live&scope=site&ebv=EB&ppid=pp\_7 Capítulo 1 y 2.

Garriga Trillo, A. J. (2009). Introducción al análisis de datos. UNED - Universidad Nacional de Educación a Distancia. https://elibro-net.bibliotecavirtual.unad.edu.co/es/ereader/unad/48460?page=130

Carlos Veliz. (2020). Machine learning. Introduction to deep learning . The Editorial Fund of the Pontifical Catholic University of Peru. https://bibliotecavirtual.unad.edu.co/login?url=https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=2600876&lang=es&site=eds-live&scope=site&ebv=EB&ppid=pp\_113 Capítulo 4.

Romero Villafranca, R. y Zúnica Ramajo, L. (2020). Métodos estadísticos para ingenieros. Editorial de la Universidad Politécnica de Valencia. https://elibro-net.bibliotecavirtual.unad.edu.co/es/ereader/unad/129644?page=212

Minguillón, J. Casas, J. y Minguillón, J. (2017). Minería de datos: modelos y algoritmos. Editorial UOC. https://elibro-net.bibliotecavirtual.unad.edu.co/es/ereader/unad/58656?page=208

Sepúlveda, J. F. D., & Morales, J. C. C. (2013). Comparación entre árboles de regresión CART y regresión lineal. Comunicaciones en Estadística, 6(2), 175-195. https://dialnet.unirioja.es/servlet/articulo?codigo=7393722