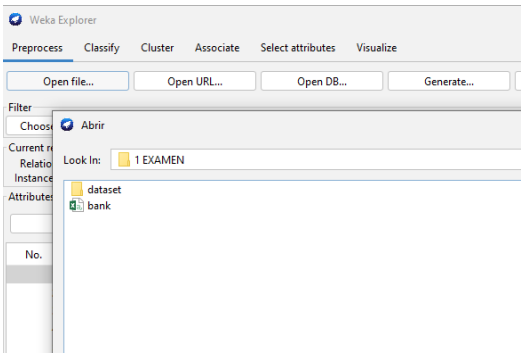
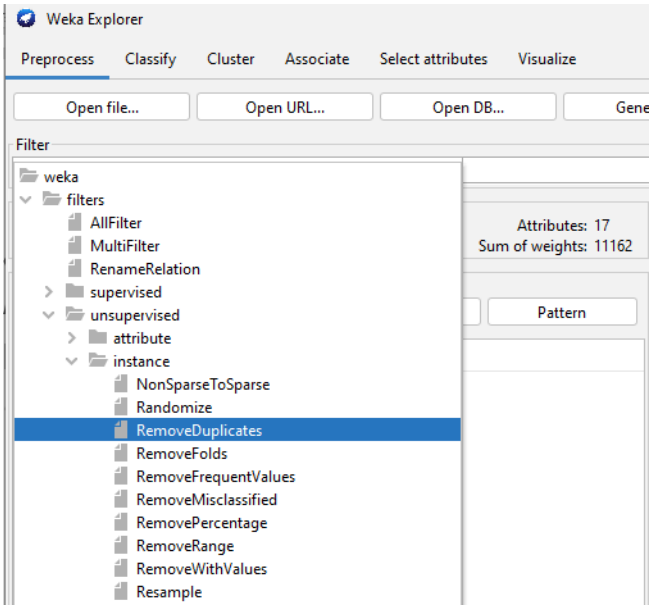
PREPROCESAMIENTO 1: *RemoveDuplicates()*

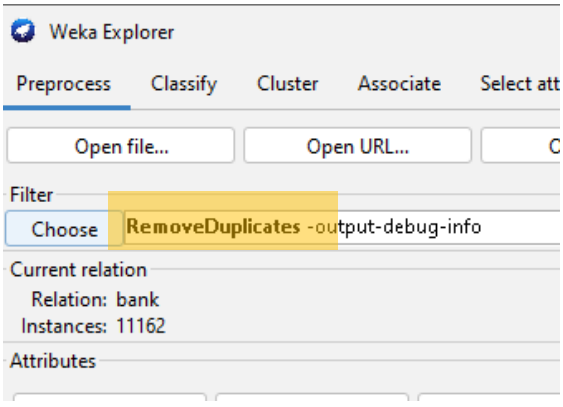
Paso1: Abrimos weka y seleccionamos nuestro archivo.

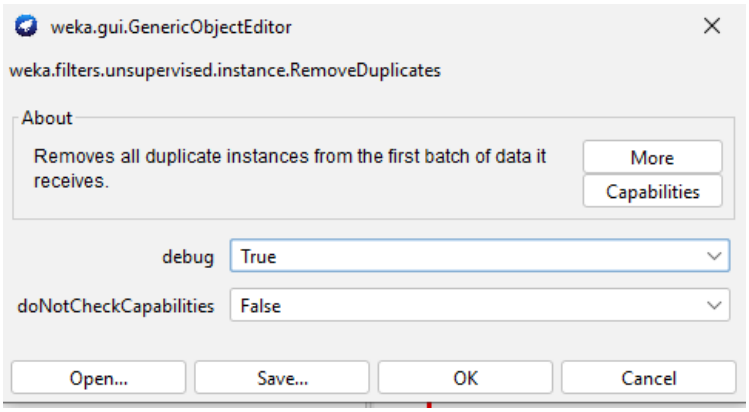


Paso2: Uzaremos un filtro de preprocesamiento llamado ***RemoveDuplicates (Eliminación de Registros Repetidos)*** para eliminar entradas que aparezcan más de una vez en nuestro conjunto de datos. Para ello nos dirigimos a ***Choose -> unsupervised -> instance -> RemoveDuplicates.***

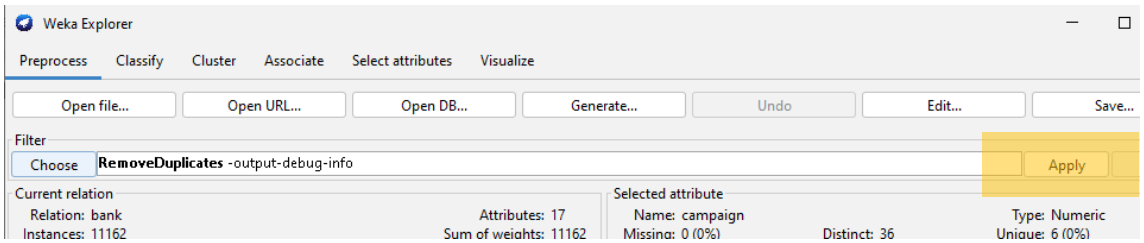


Paso3: En la configuración elegimos Debug:True, para visualizar una información más precisa de la aplicación del filtro.

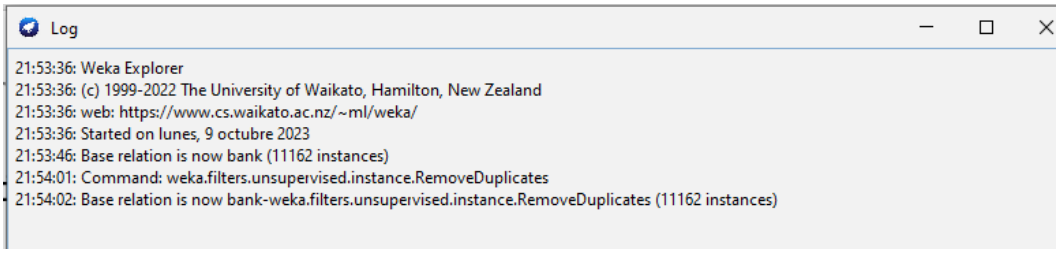




Paso4: Presionamos en Apply para aplicar ***Eliminación de Registros Repetidos.***



Paso5: Para verificar los cambios, nos dirigimos a la sección de ***Log*** y podemos ver como se ejecutó.

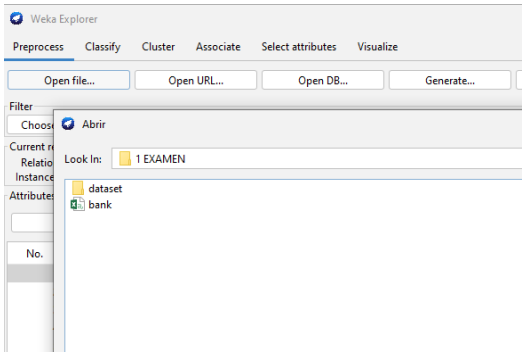


La razón principal para aplicar la técnica de preprocesamiento de "RemoveDuplicates" (eliminar duplicados) en el conjunto de datos es la eliminación de registros duplicados o idénticos que no aportan información adicional y pueden afectar negativamente el rendimiento de los modelos de aprendizaje automático.

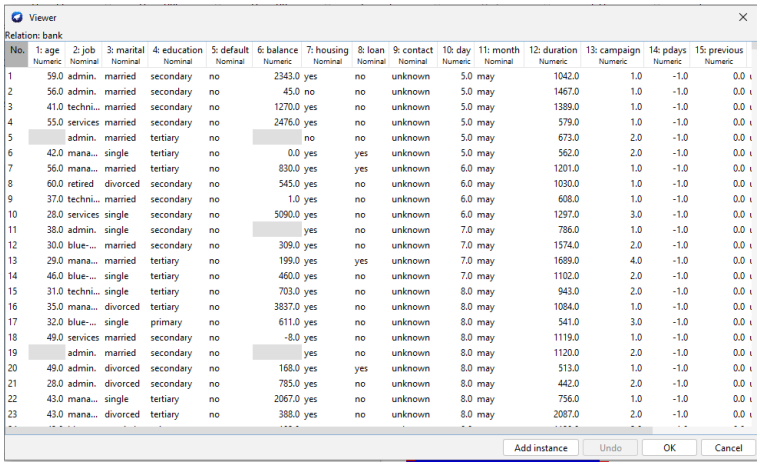
Entonces, al aplicar "RemoveDuplicates," lo que hacemos es como una especie de limpieza para nuestros datos. Nos aseguramos de que cada persona aparezca solo una vez y que nuestros datos sean ordenados y sin repeticiones innecesarias. Esto hace que nuestros análisis y modelos sean más precisos y nos evita problemas más adelante.

PREPROCESAMIENTO 2: *ReplaceMissingValues()*

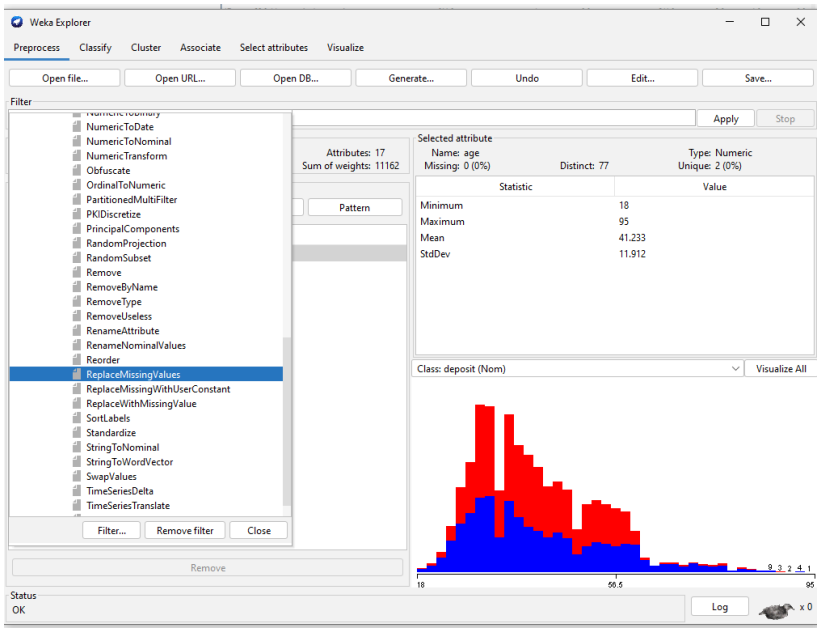
Paso1: Abrimos weka y seleccionamos nuestro archivo.



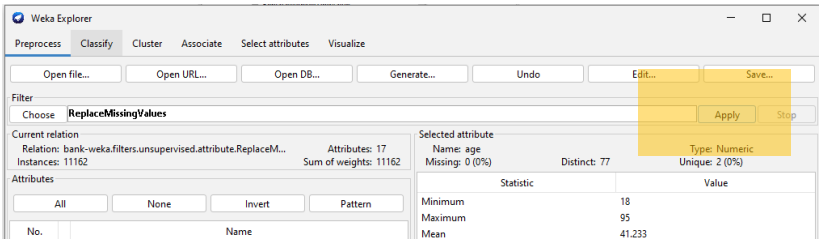
Paso2: Si presionamos sobre ***edit,*** podemos observar campos que no tienen valores por lo cual aremos el uso de la técnica para una mejor interpretación de los datos.

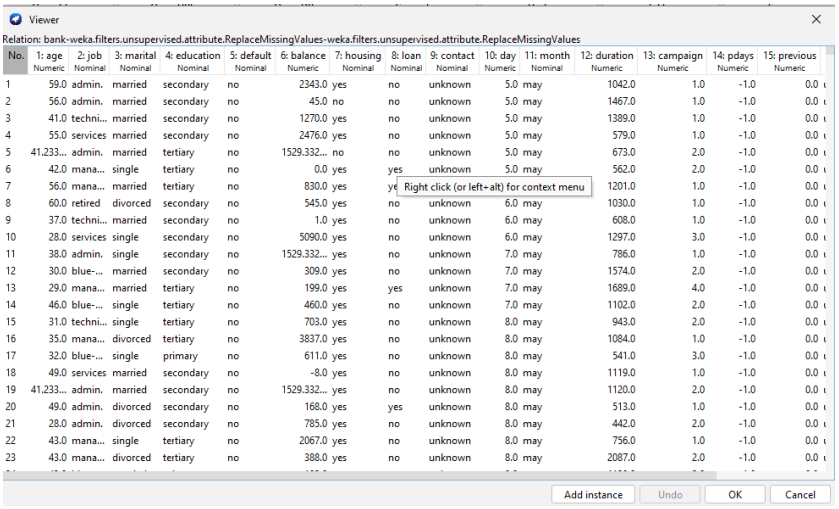


Paso3: Usaremos un filtro de preprocesamiento llamado ***ReplaceMissingValues*** para llenar aquellos valores faltantes, aplicando estrategias como la media, mediana y moda. Para ello nos dirigimos a ***Choose -> unsupervised -> attribute -> ReplaceMissingValues.***



Paso4: Hacemos click en ***apply*** y si volvemos a ingresar a ver los valores en la sección de ***edit*** podemos observar que los valores que estaban vacios ahora tienen un valor.



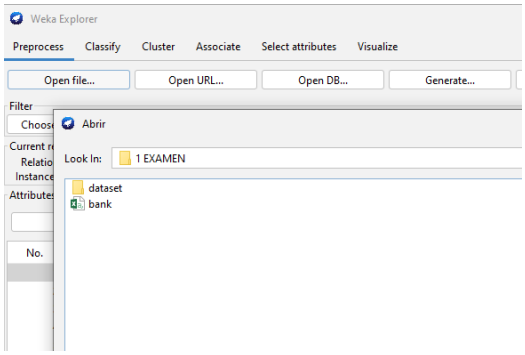


La razón fundamental para utilizar la técnica de reemplazo de valores faltantes, como "ReplaceMissingValues" en WEKA, en el conjunto de datos "Bank Marketing" radica en la necesidad de asegurar que los datos estén completos y listos para un análisis efectivo. En este contexto, los valores faltantes pueden surgir debido a diversos motivos, como errores en la recopilación de datos o la falta de información en ciertos campos. Para obtener resultados precisos en el análisis y modelado de datos, es imperativo abordar estos valores faltantes de manera adecuada.

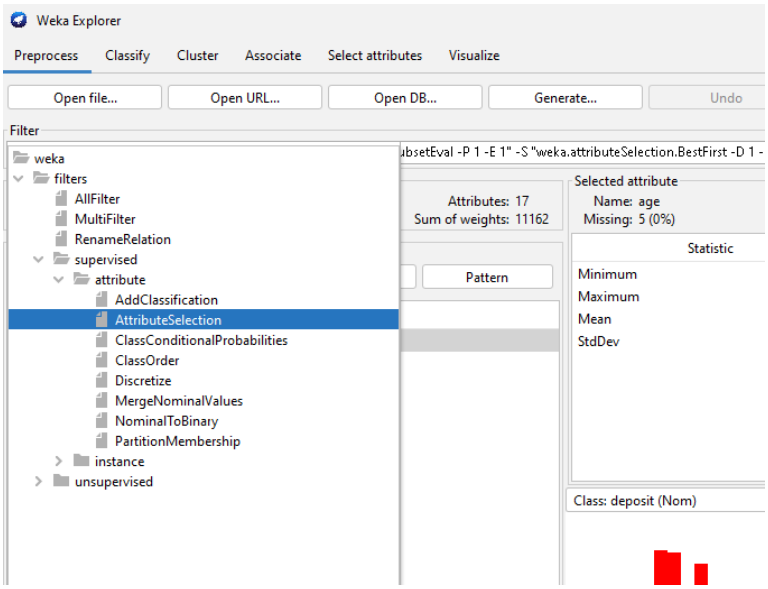
Además, el conjunto de datos "Bank Marketing" contiene información valiosa sobre interacciones de clientes con el banco y sus decisiones de suscripción a productos financieros. Ignorar los registros con valores faltantes podría resultar en la pérdida de información importante y reducir la capacidad de generalización de los modelos de aprendizaje automático.

PREPROCESAMIENTO 3: *AttributeSelection()*

Paso1: Abrimos weka y seleccionamos nuestro archivo.

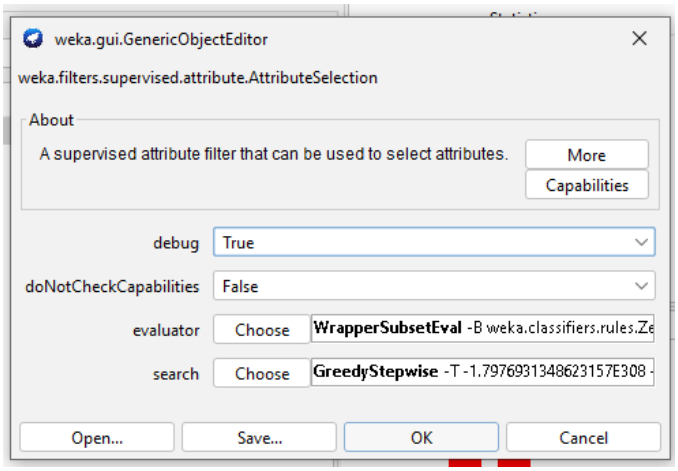


Paso2: Usaremos un filtro de preprocesamiento llamado ***AttributeSelection*** para poder mostrar algunas características importantes de nuestro conjunto de datos. Para ello nos dirigimos a ***Choose -> supervised -> attribute -> AttributeSelection*.**

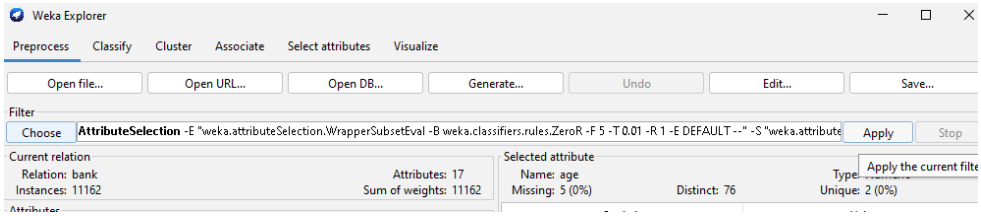


Paso3: Para poder evaluar el rendimiento de una columna de mayor importancia seleccionamos ***WrapperSubsetEval***

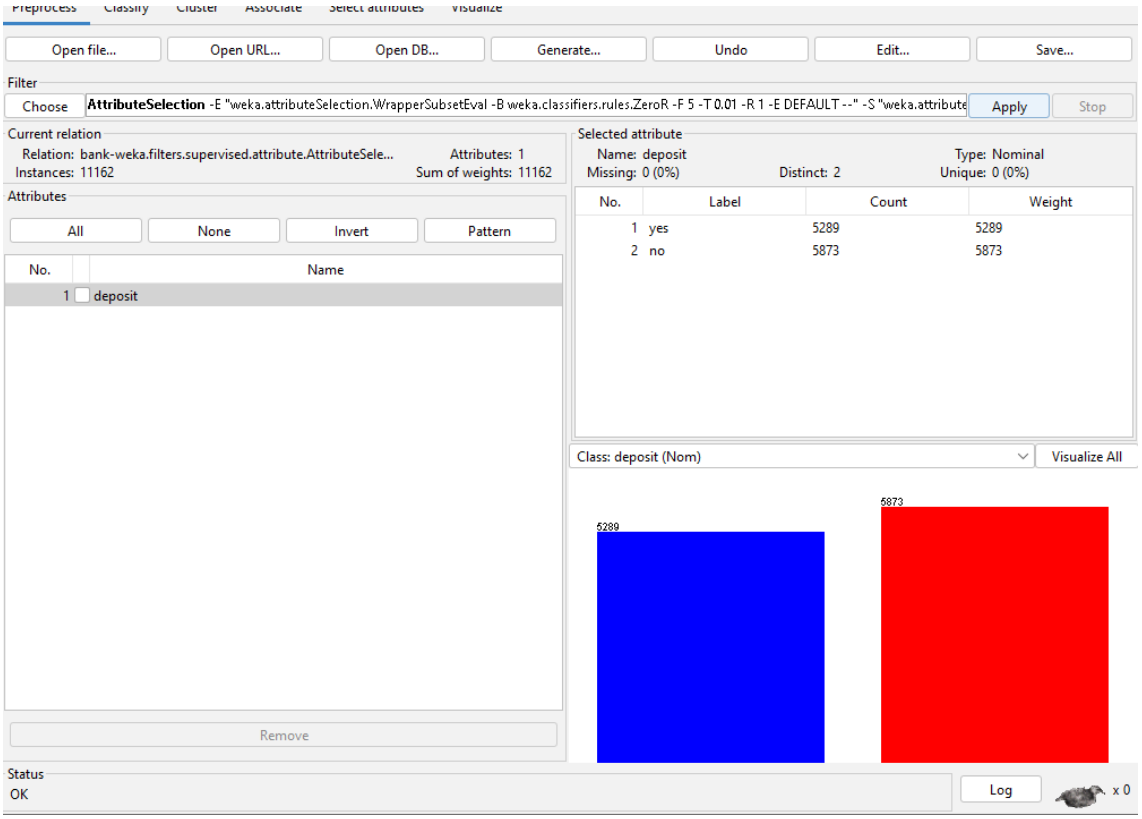
Para evaluar combinaciones posibles eligiendo el mejor subconjunto de características con ***GreedyStepwise***.



Paso4: Aplicamos los cambios hechos en ***Apply***



Paso5: Se puede observar que lo mas relevante en el dataset son los deposit (depósitos), ya que de este depende las predicciones que se pueden aplicar. Esta columna cuenta con valores categóricos de SI o NO donde a través de esta, se puede usar para analizar el dataset, identificar factores que influyen en la decisión del cliente y desarrollar modelos de predicción de clientes.

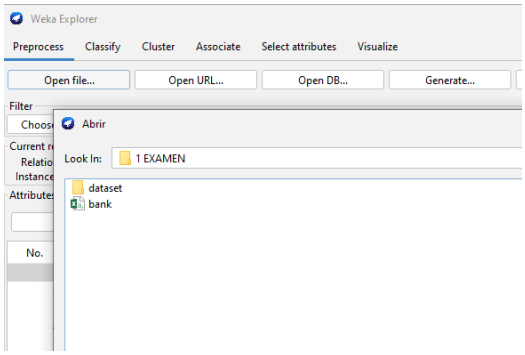


La razón principal para aplicar la técnica de selección de atributos (AttributeSelection) es mejorar la calidad de los modelos y el rendimiento del análisis de datos.

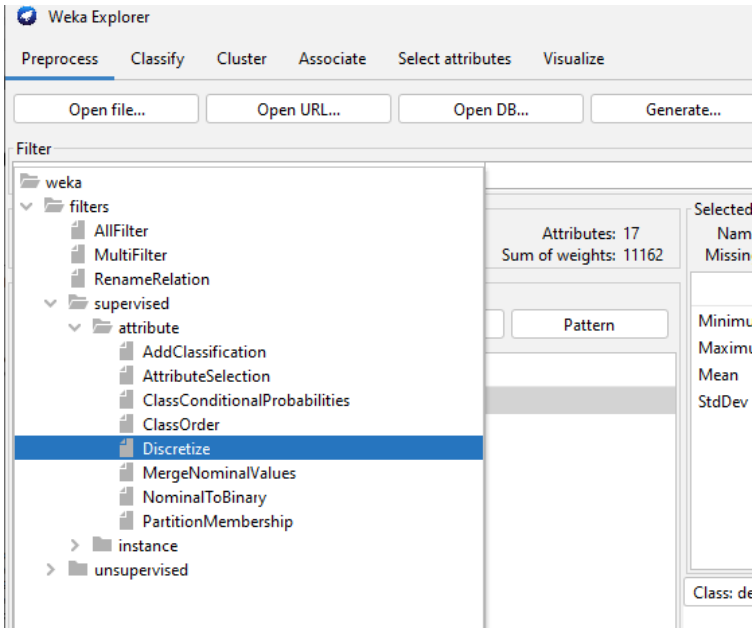
En el conjunto de datos "Bank Marketing," es probable que tengamos una amplia gama de atributos que describen a los clientes, su comportamiento financiero y las campañas de marketing. Algunos de estos atributos pueden ser redundantes o no aportar información significativa para predecir si un cliente se suscribirá a un producto bancario o no. La selección de atributos nos permite identificar y retener solo los atributos más relevantes, lo que simplifica nuestros modelos, mejora la interpretabilidad y reduce el riesgo de sobreajuste.

PREPROCESAMIENTO 4: *discretize()*

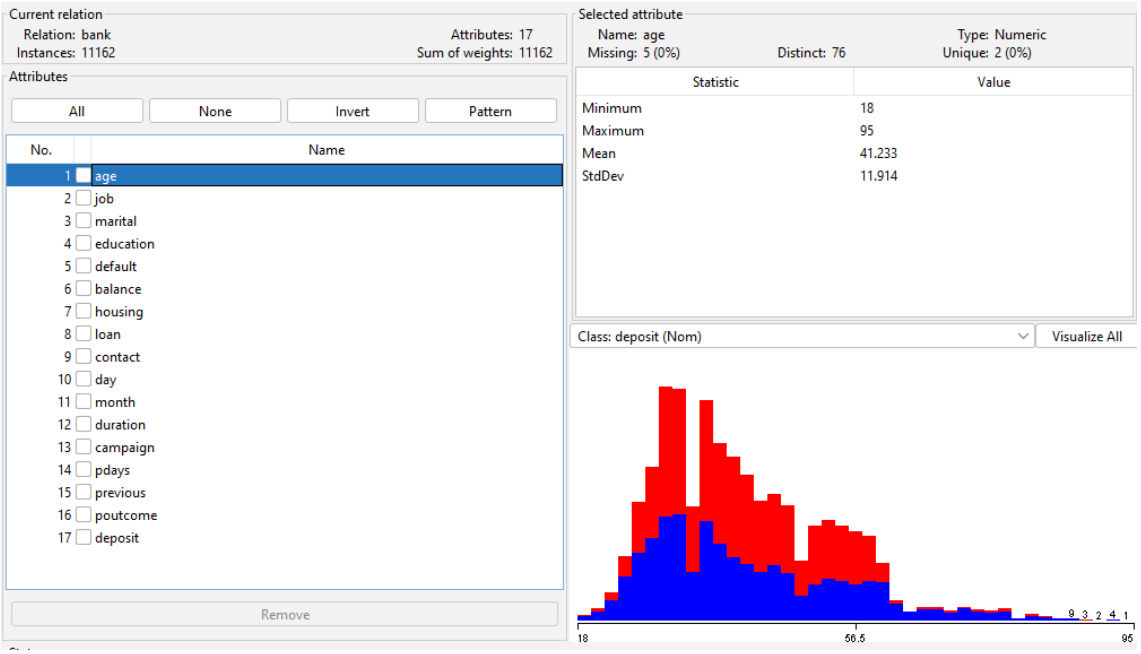
Paso1: Abrimos weka y seleccionamos nuestro archivo.



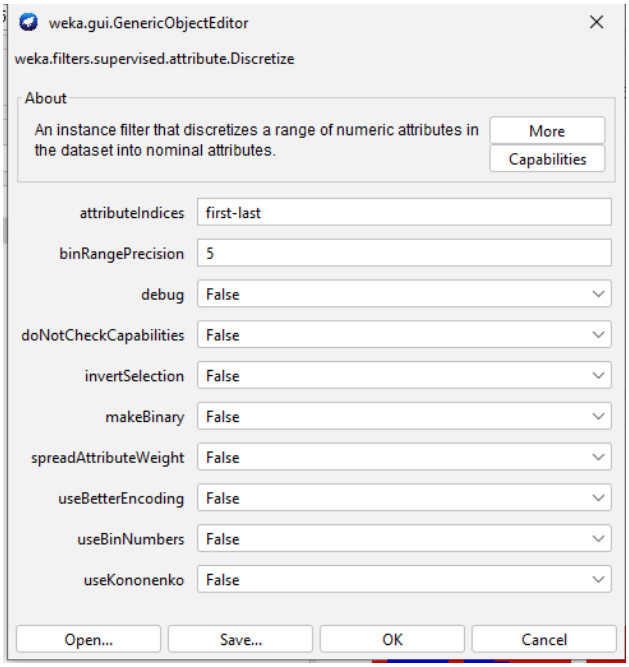
Paso2: Usaremos un filtro de preprocesamiento llamado ***Discretize*** para poder observar e interpretar de la mejor forma los datos. Para ello nos dirigimos a ***Choose -> supervised -> attribute -> Discretize.***



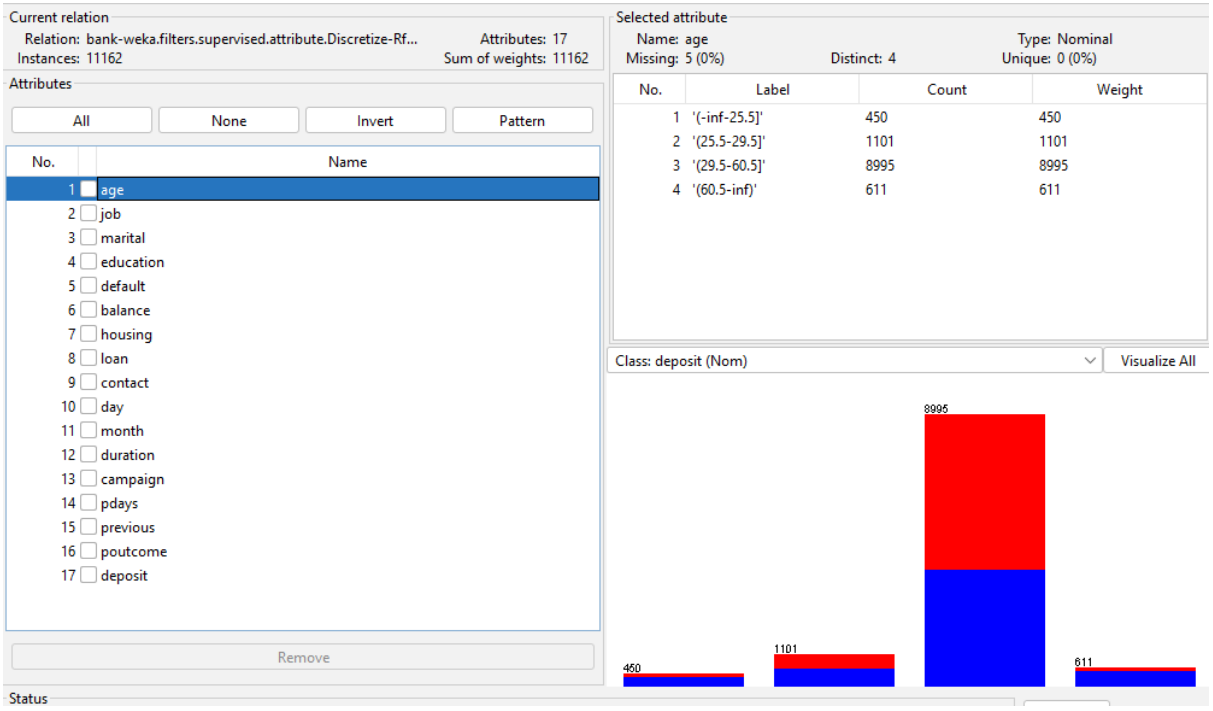
Paso3: Inicialmente los datos no discretizados se lo observan de la siguiente forma, para el caso de la primera columna que es la EDAD.



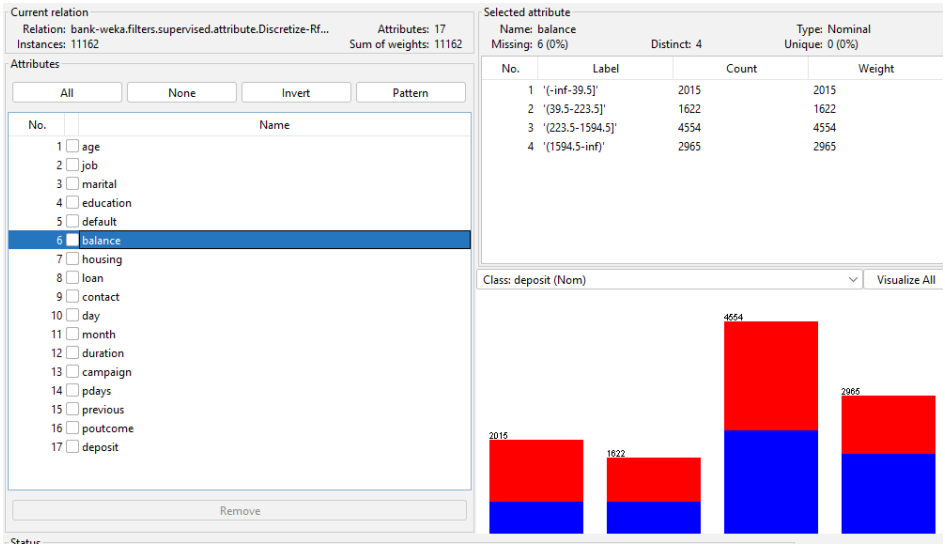
Paso4: En las configuraciones de la discretización, seleccionamos todas las columnas, ponemos en ***True*** la sección del ***debug*** y el número de decimales que se utilizarán para los puntos de corte al generar etiquetas de contenedor será de 5.



Paso5: Hacemos click en ***apply*** y se puede observar que la discretización para la edad lo divide en 4 intervalos.



Para las otras columnas de datos numéricos también lo va discretizando en intervalos.



La razón principal para utilizar la discretización en el análisis de datos es simplificar la representación de variables numéricas al convertirlas en variables categóricas o discretas.

Del conjunto de datos "Bank Marketing," la discretización adquiere una relevancia particular debido a la naturaleza de las variables presentes en este dataset. El conjunto de datos incluye atributos numéricos, como la edad de los clientes, el saldo medio anual, la duración de las llamadas, entre otros. La aplicación de la discretización en este contexto puede proporcionar varias ventajas específicas.

Como ejemplo, la discretización podría ser útil para convertir atributos numéricos, como la edad, en categorías más comprensibles, como grupos de edad. Esto facilitaría la segmentación de clientes en función de su edad y permitiría una mejor comprensión de cómo diferentes grupos demográficos responden a las campañas de marketing bancario.