**INF 354**

**INTELIGENCIA ARTIFICIAL**

**GUTIERREZ MACHICADO GERMAN**

**CI: 6072590**

**12-04-2022**

**PRIMER EXAMEN PARCIAL**

**Enlace github:** [**https://github.com/German-G-M/procesamiento\_de\_datos**](https://github.com/German-G-M/procesamiento_de_datos)

**PREGUNTAS**

1. Seleccione un dataset del área médica o de abogacía (datos tabulares). Realice lo siguiente:

   a. El calculo del 1er cuartil de datos, el percentil 50 por columna; explique qué significa en cada caso mediante Python sin uso de librerías

   b. Realice lo mismo del inciso (a) con el uso de numpy y pandas

   c. Grafique los datos y explique su comportamiento (PYTHON)

2. Del dataset anterior realice en WEKA, tres algoritmos de pre procesamiento.

3. Del dataset anterior realice en PYTHON, tres algoritmos de pre procesamiento.

4. Con el uso de librerías en PYTHON, construya la dependencia de Abuelos, tíos, padres, primos e hijos de su familia.

5. En PYTHON grafique el árbol de decisión del dataset seleccionado.

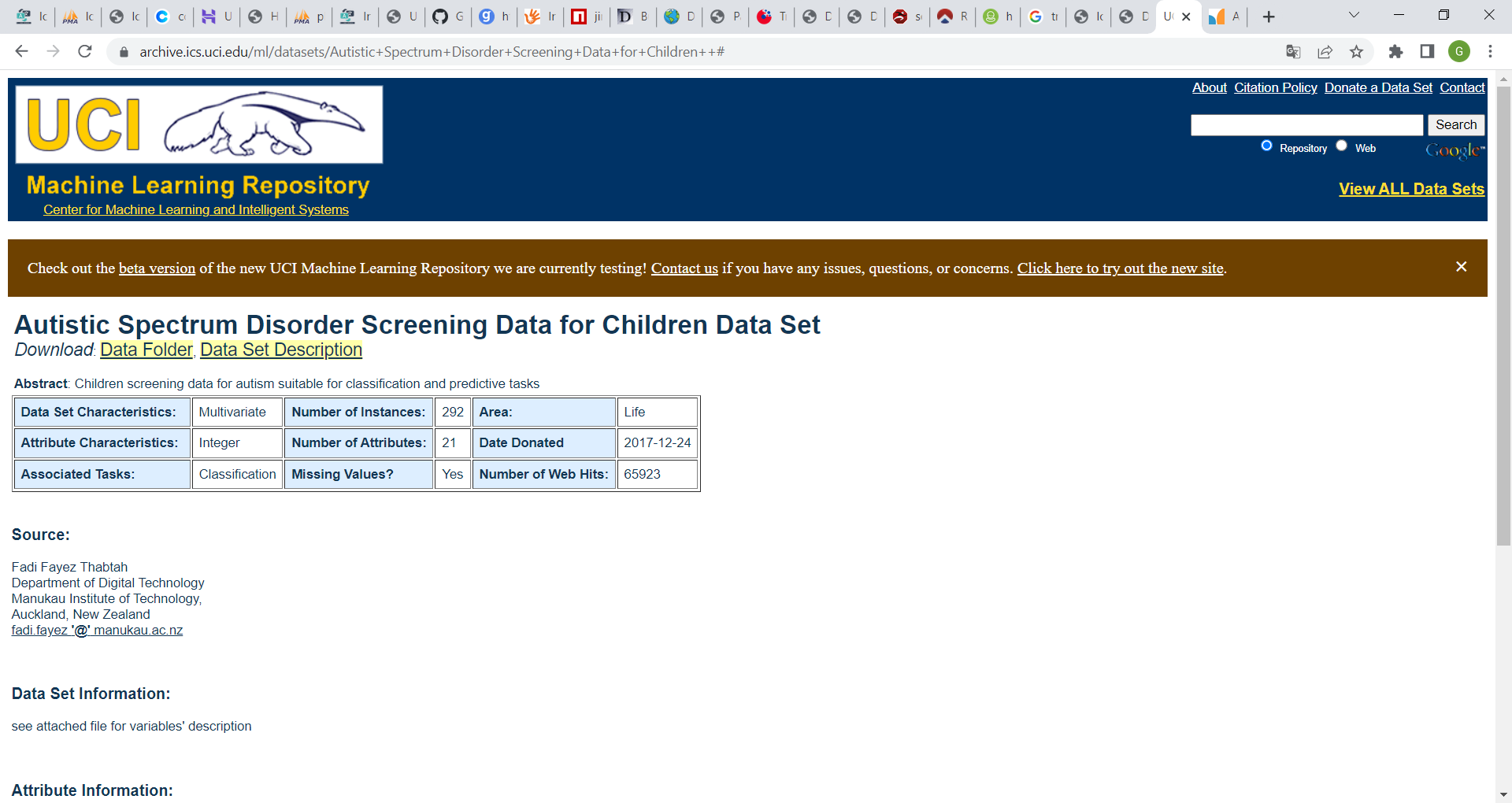
Cada pregunta debe ser almacenada en Github, la misma permitir su acceso mínimamente a [msilva@fcpn.edu.bo](mailto:msilva@fcpn.edu.bo). Adjuntar el link por pregunta en un PDF o Word y enviarlo para su revisión.

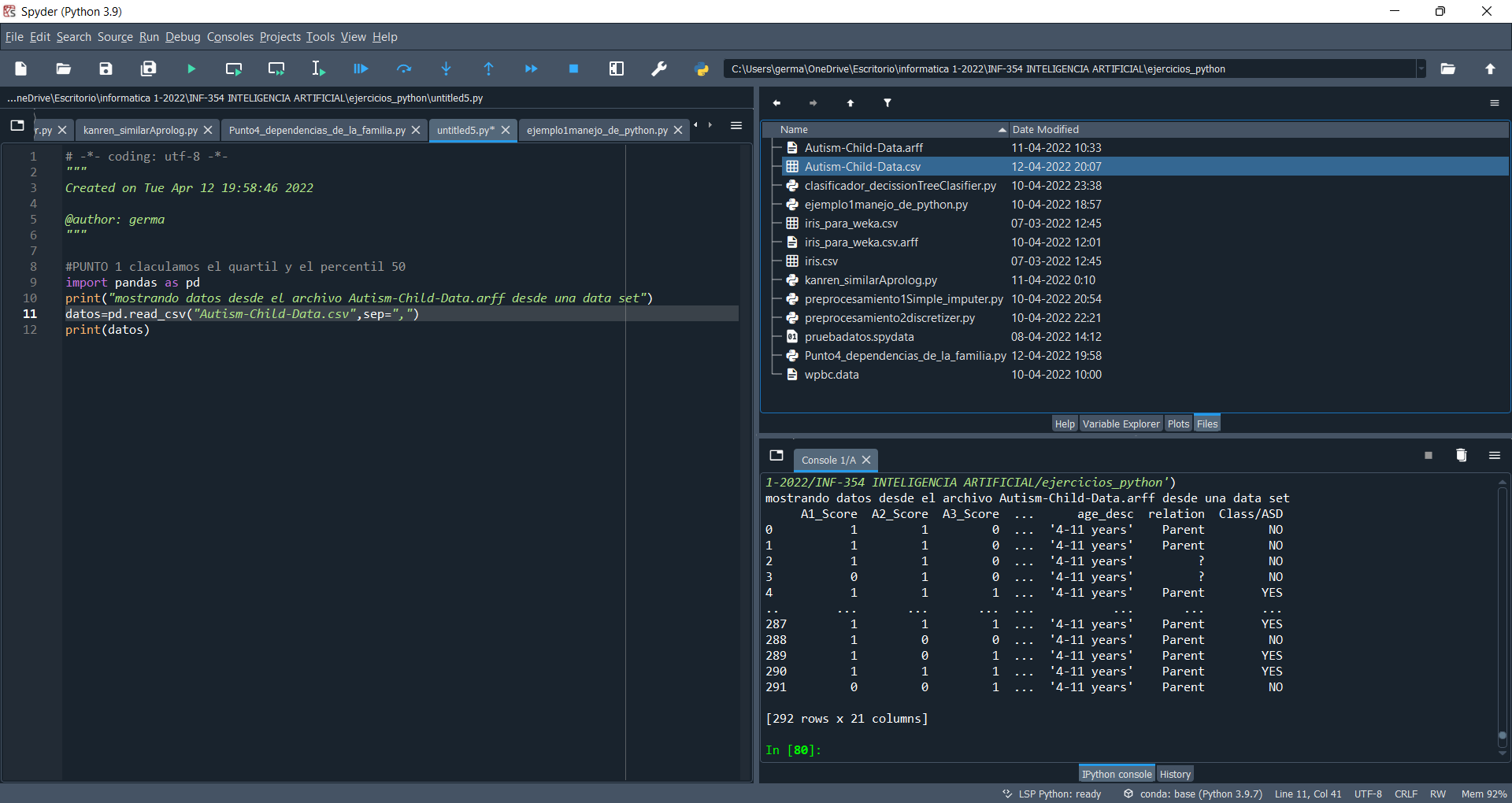
La revisión del examen se hará el día siguiente de finalizado el examen de manera presencial o virtual.

**SOLUCIONES**

**1.** Seleccione un dataset del área medica o de abogacía (datos tabulares). Realice lo siguiente:

<https://archive.ics.uci.edu/ml/datasets/Autistic+Spectrum+Disorder+Screening+Data+for+Children>++





   a. El **cálculo del 1er cuartil** de datos, el **percentil 50** por columna; explique qué significa en cada caso mediante Python sin uso de librerías

**El quartil 1 (Q1**) es **el percentil 25** (P25). El 25% de los datos son menores o iguales a Q1

Formula del 1er cuartil:

(N+1)/4

Nuestro N=292

(N+1)/4 = (292+1)/4 = 293/4 = 73.25

Como es un número decimal el cuartil será un número entre: **X73 y X74**

Q1 = X73 + 0.25 \*(X74 –X73)

El quartil 2 (Q2) es el **percentil 50** (P50) y **es la mediana**. El 50% de los datos son menores o iguales a Q2

La fórmula del percentil 50 o lo lo que es lo mismo del quartil 2 o la mediana es:

N/2 + 1

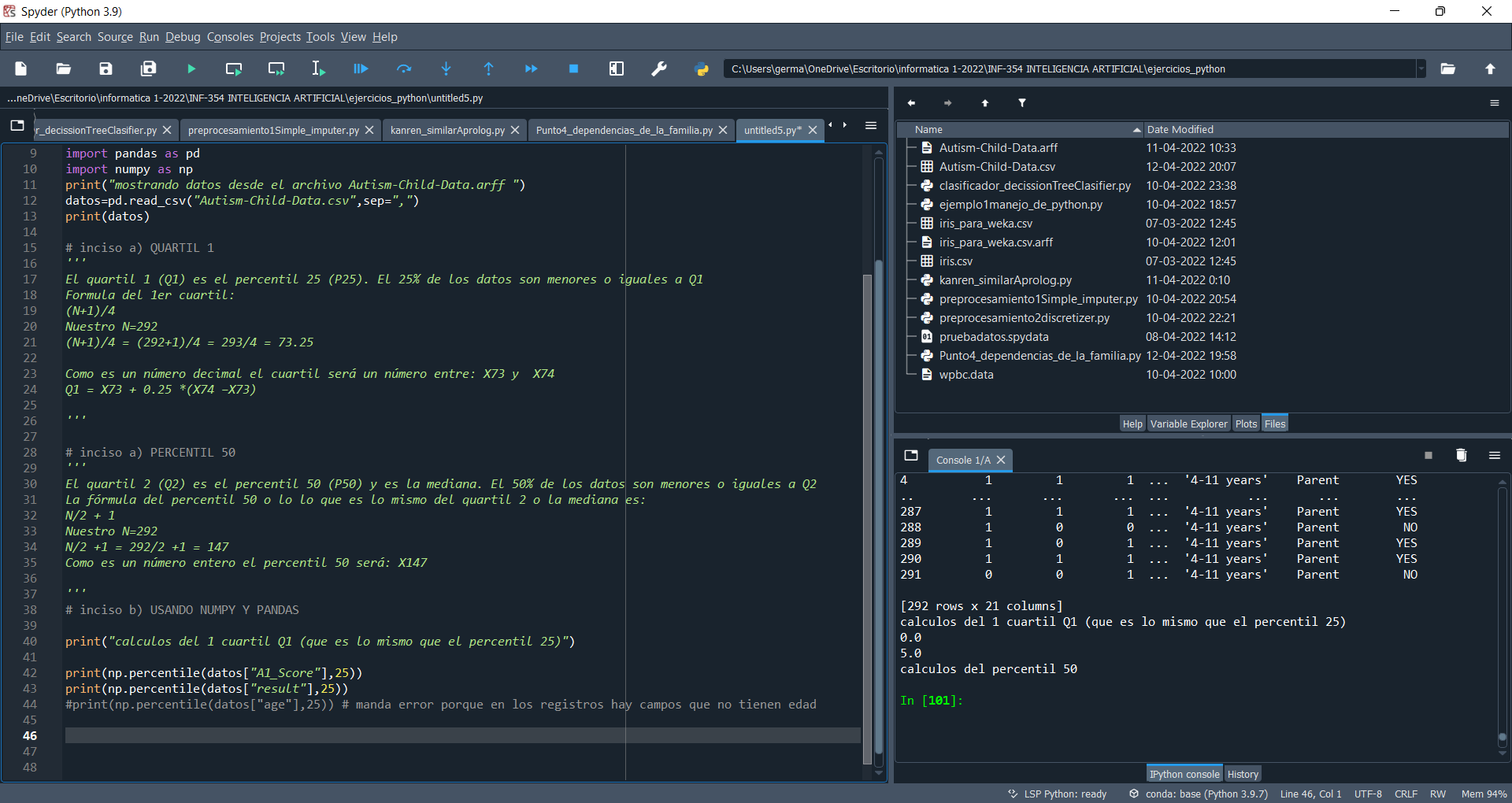
Nuestro N=292

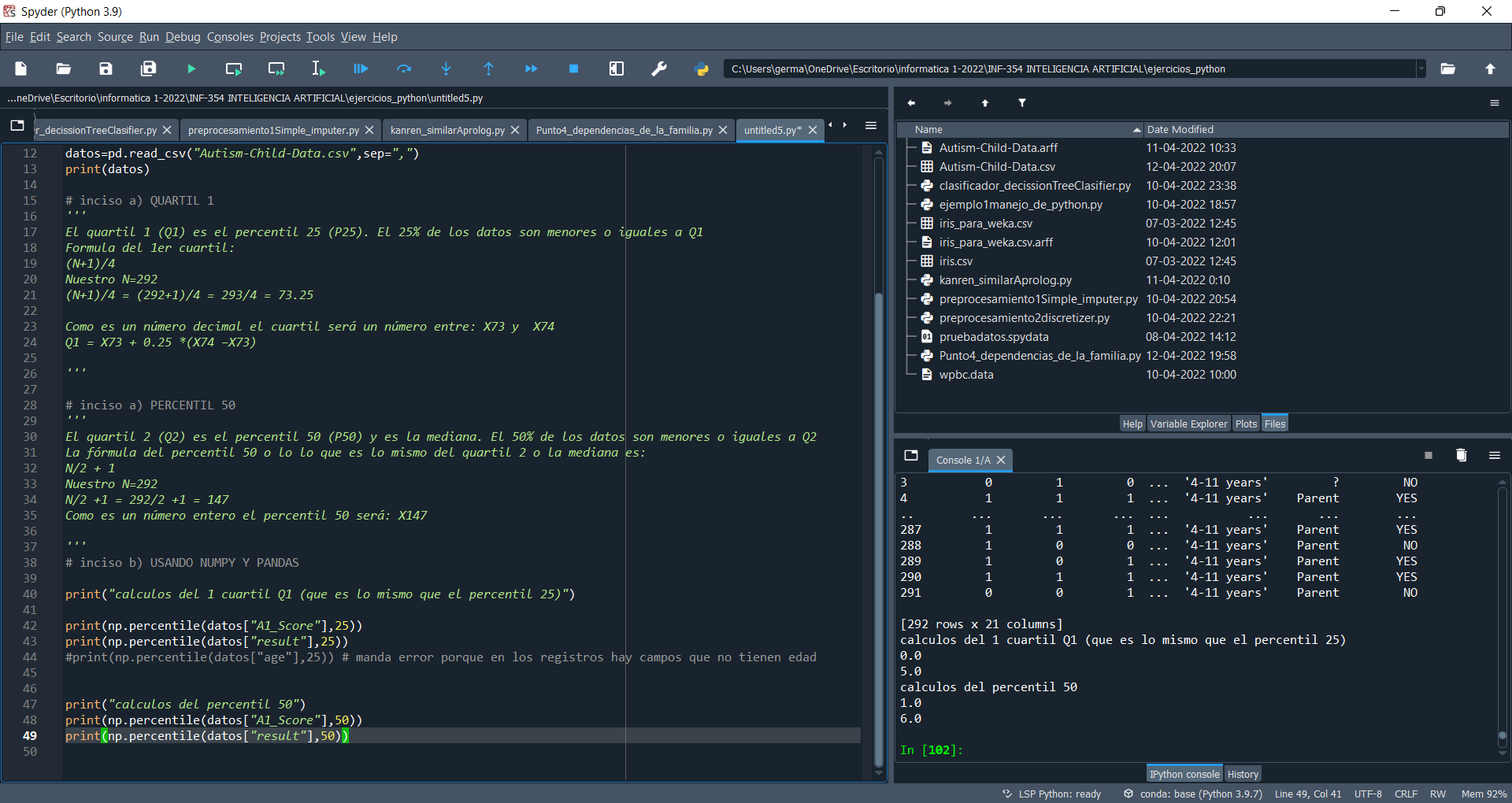
N/2 +1 = 292/2 +1 = 147

Como es un número entero el percentil 50 será: **X147**

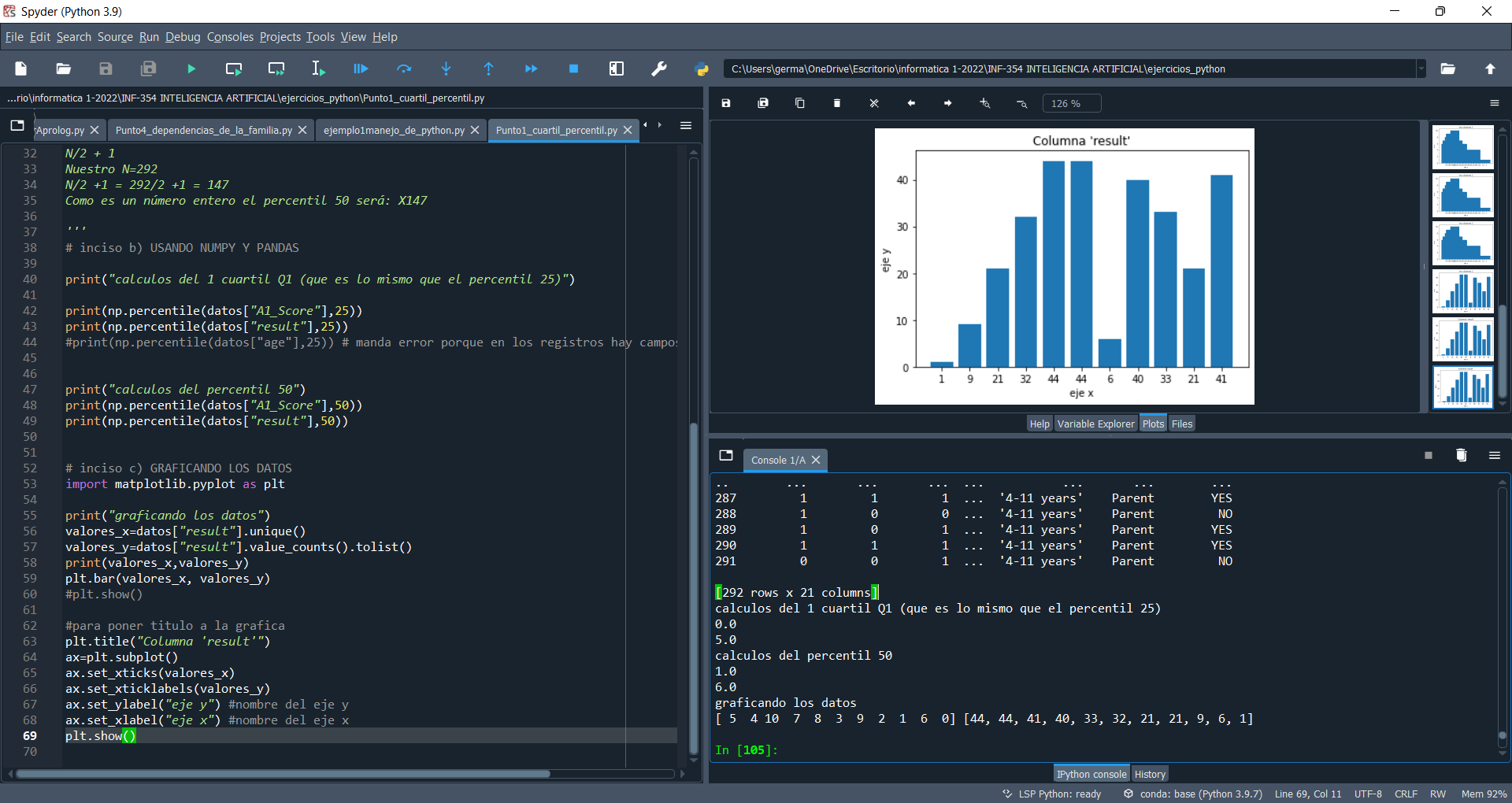
El quartil 3 (Q3) es el percentil 75 (P75). El 75% de los datos son menores o iguales a Q3

   b. Realice lo mismo del inciso (a) con el uso de **numpy y pandas**





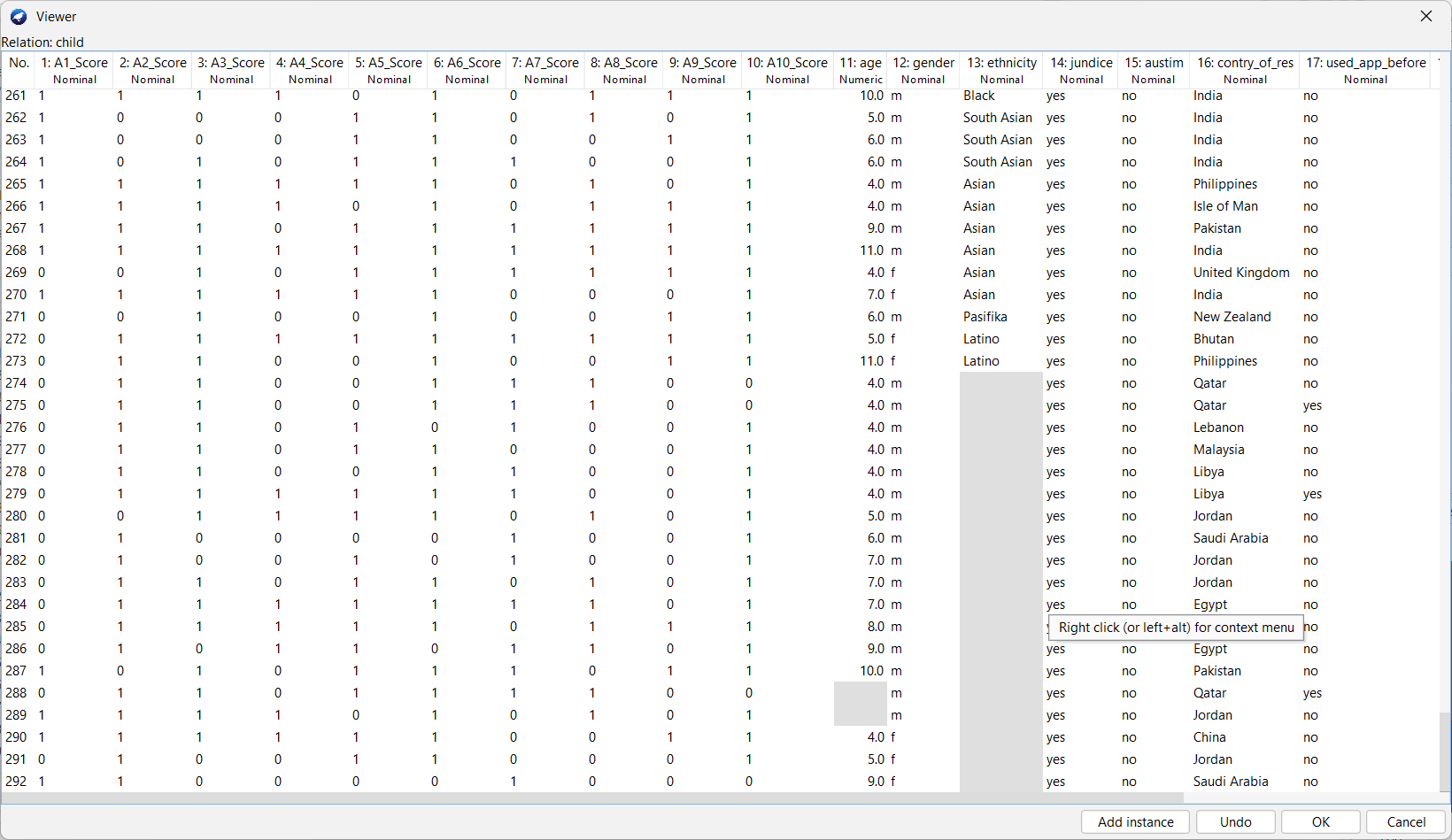
   c. **Grafique los datos** y explique su comportamiento (PYTHON)



**2.** Del dataset anterior realice en **WEKA, tres algoritmos de preprocesamiento**.

**1 primer algoritmo de pre procesamiento: ReplaceMissingValues**





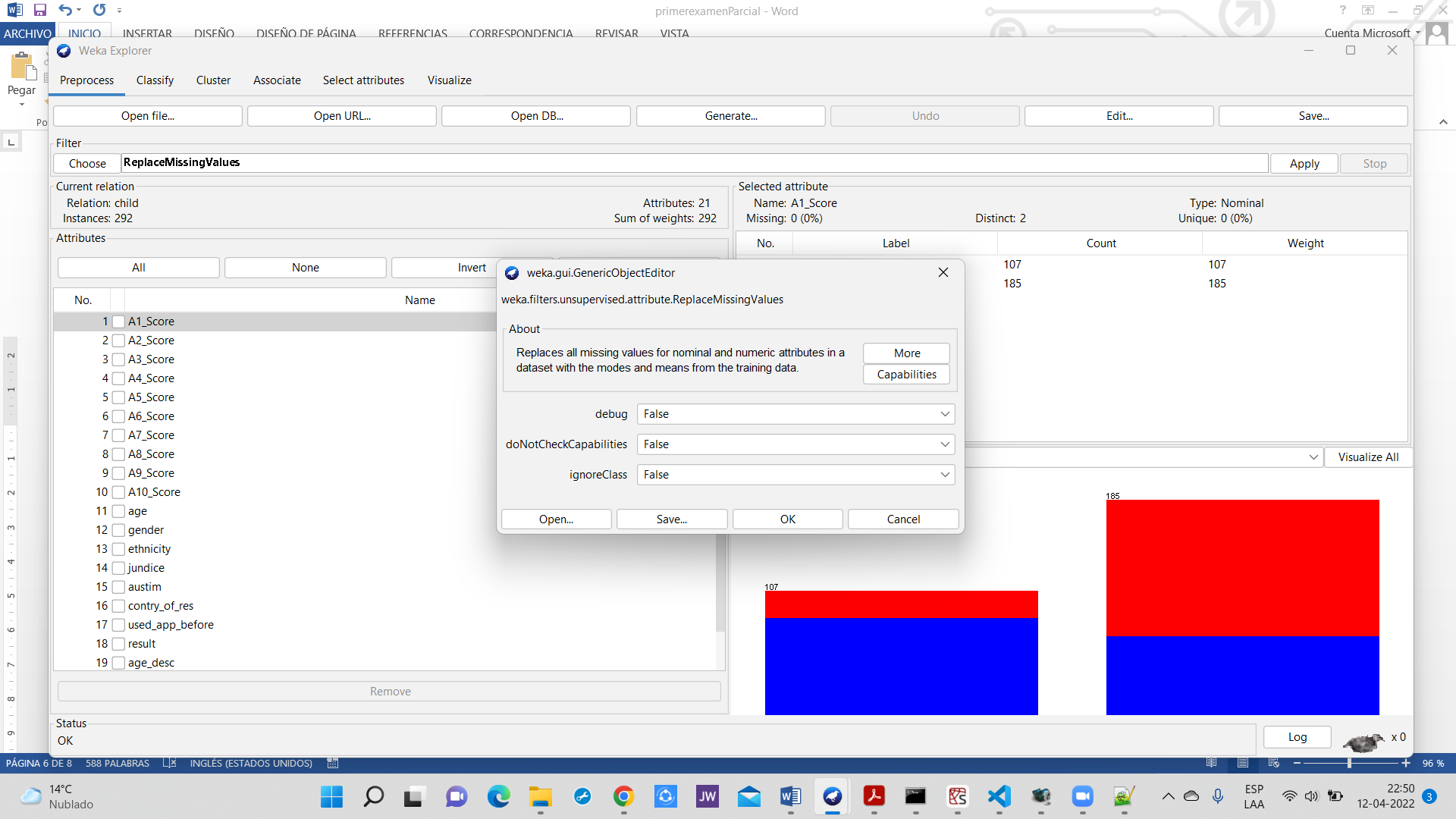
En la columna “age” se pueden observar 4 campos vacíos

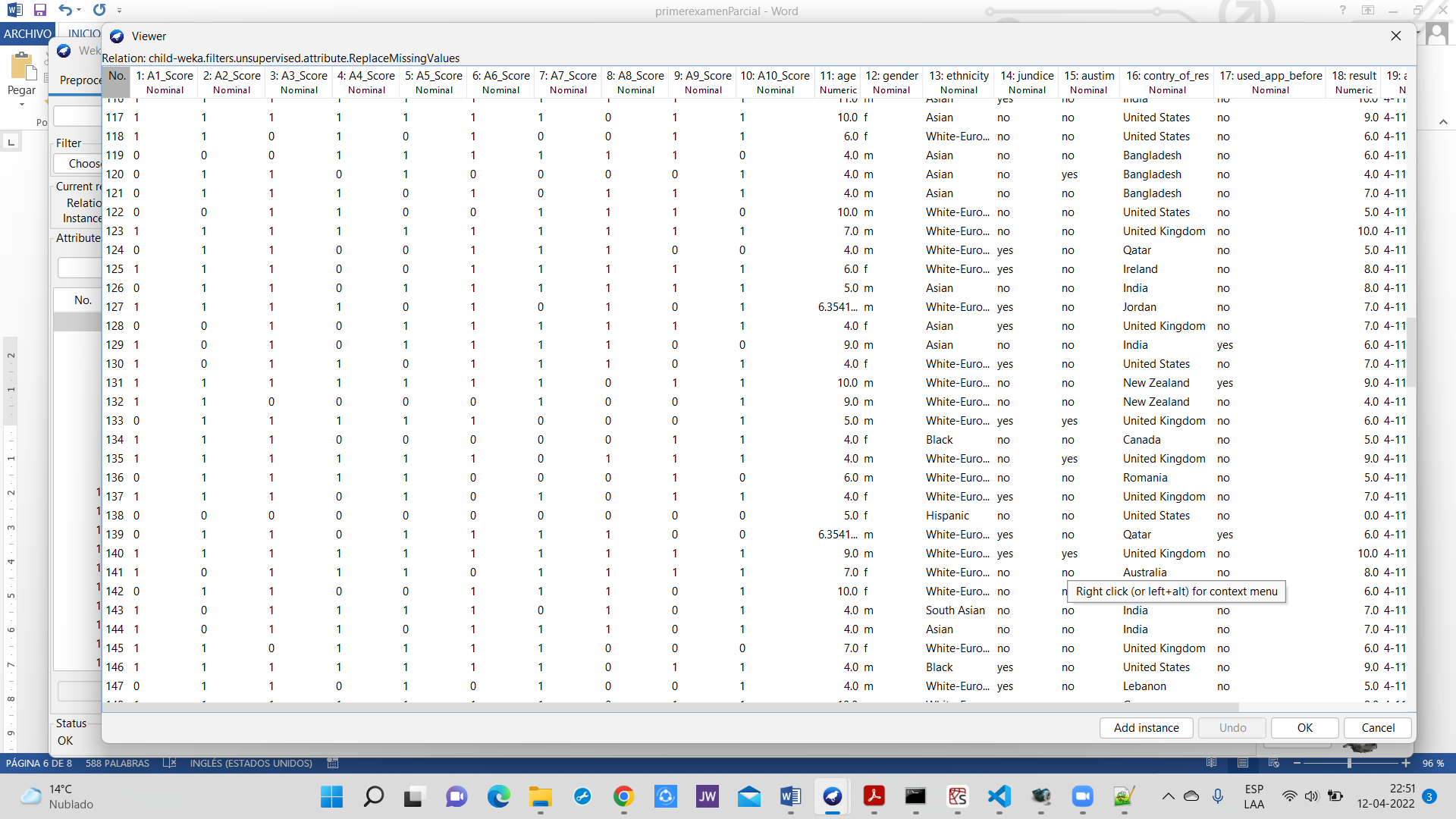
En la columna “ethnicity” se pueden observar 43 campos vacíos

En weka procedemos de la siguiente manera:

En la pestaña “**preprocessing**”

Filter > choose > filters > unsupervised > attribute > **replaceMissingValues** > apply





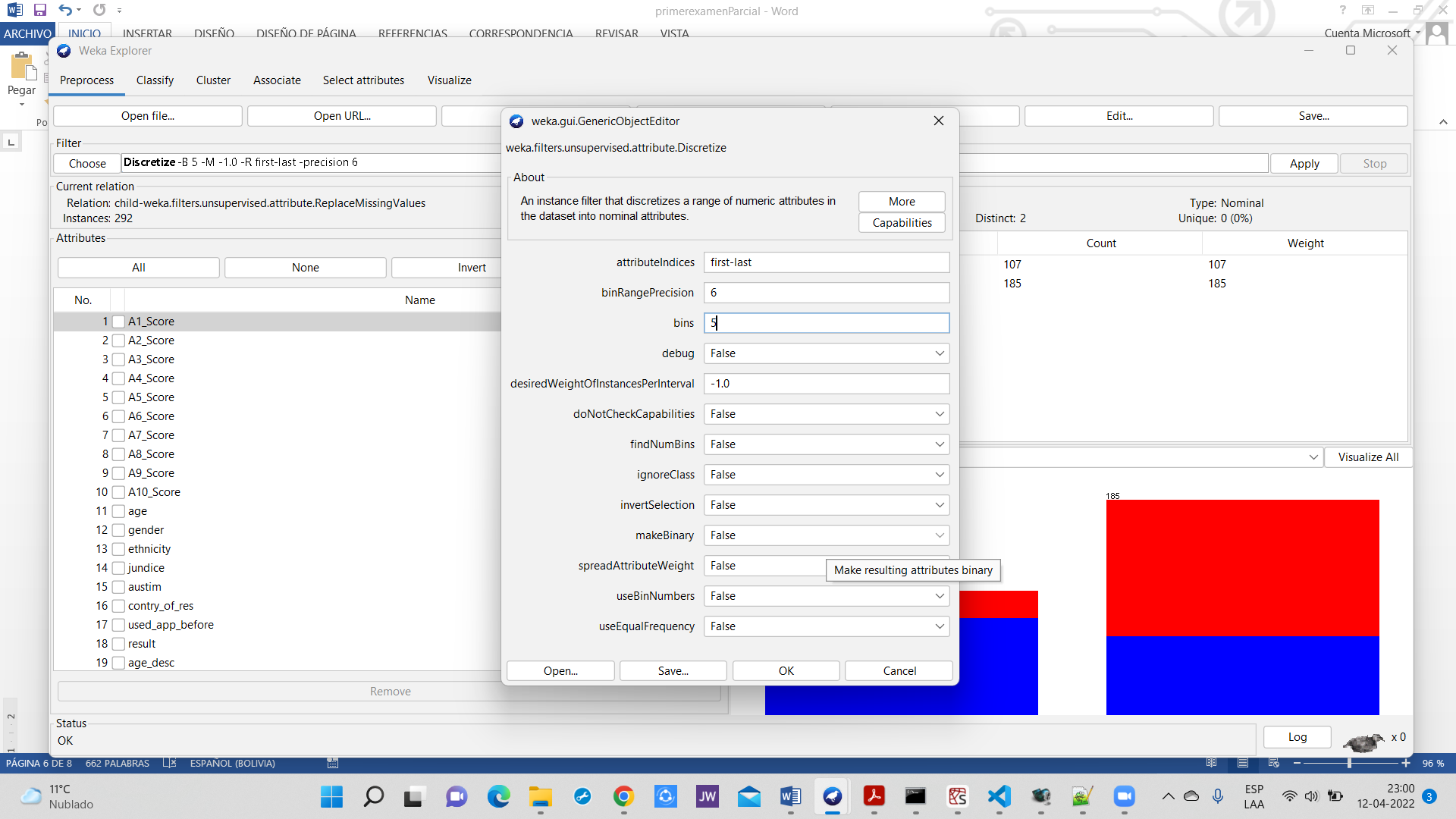
Podemos ver que todos los campos han sido rellenados, los 4 campos de “age” y los 43 campos de “ethnicity”

**2 segundo algoritmo de pre procesamiento: Discretización**

En la pestaña “**preprocessing**”

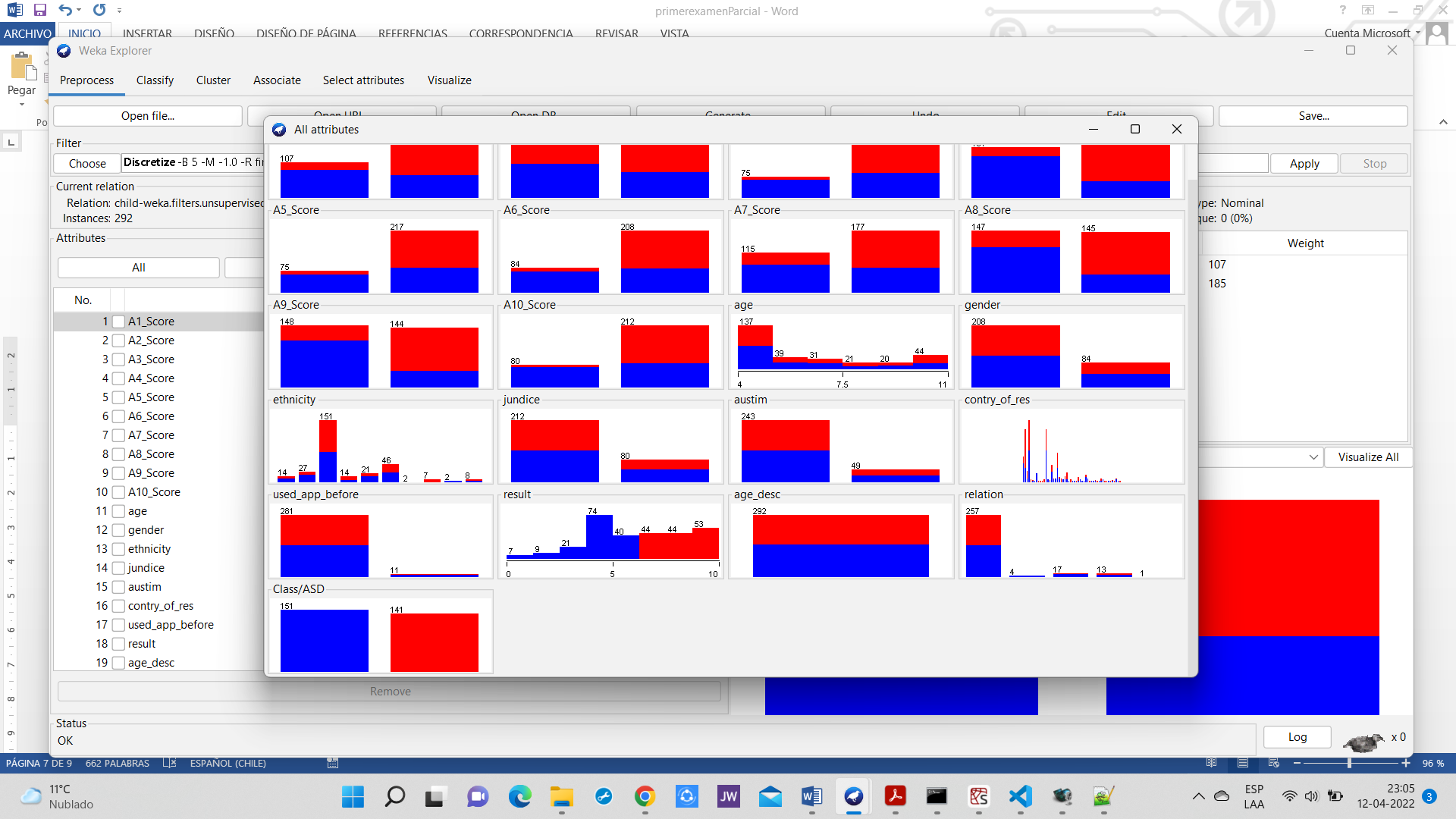
Filter > choose > filters > unsupervised > attribute > **Discretize** > apply

En “bins” colocamos el número de columnas iguales para todos. Dividimos los datos en grupos



Si ponemos el “useEqualFrequency” en “true” se intentará igual las frecuencias. Es decir que se intentrá igualar la altura de las barras

Antes:



Después:

Con bins=5



Vemos que solamente las columnas con números fueron afectadas

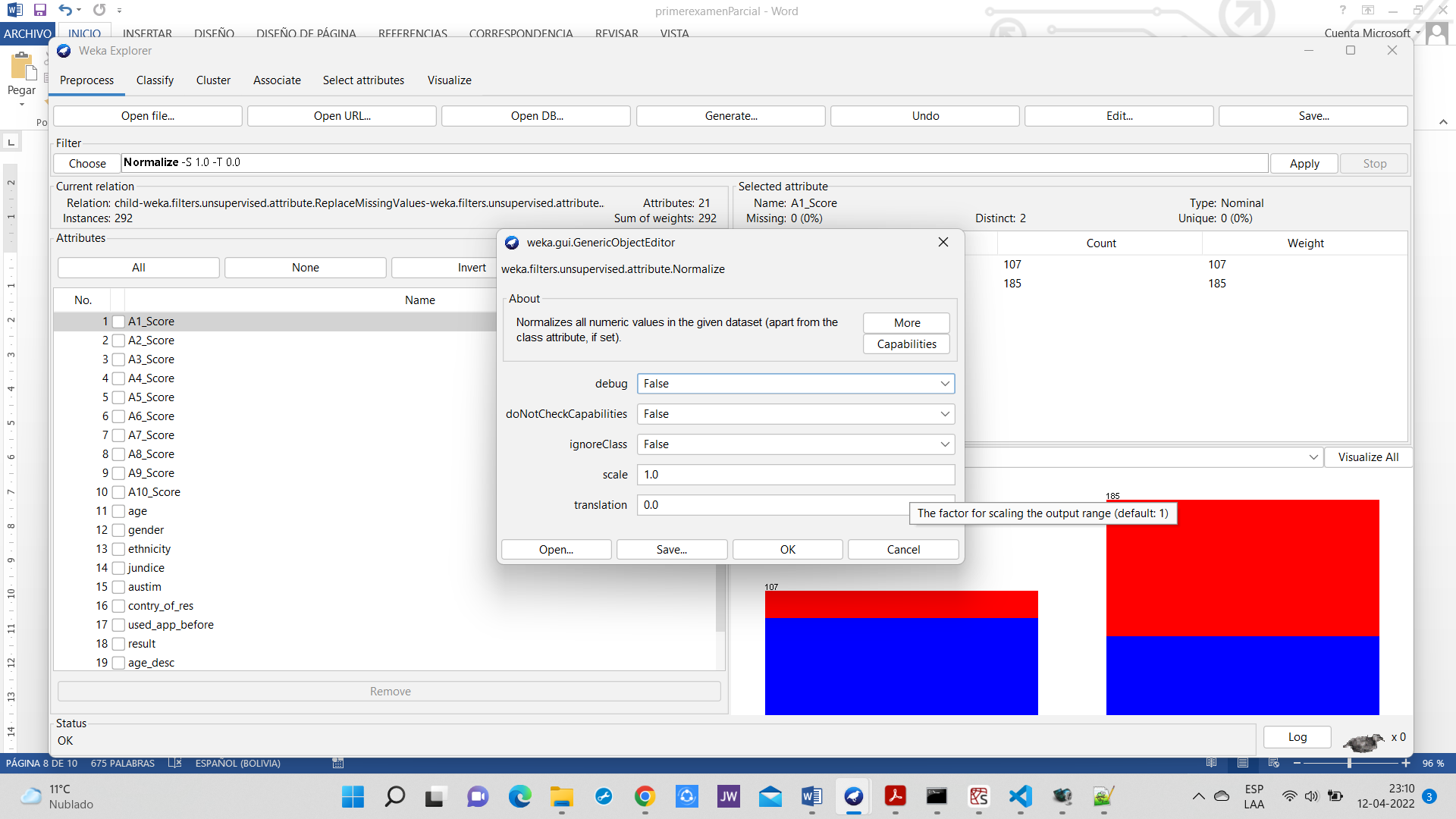
**3 tercer algoritmo de pre procesamiento: Normalización**

En la pestaña “**preprocessing**”

Filter > choose > filters > unsupervised > attribute > **Normalize** > apply

La normalización de los datos es es el proceso de reescalar uno o más atributos a un rango entre 0 y 1.

La normalización es una buena técnica a utilizar cuando no se sabe la distribución de los datos o cuando sabes que la distribución es no Gaussiana (la curva campana)

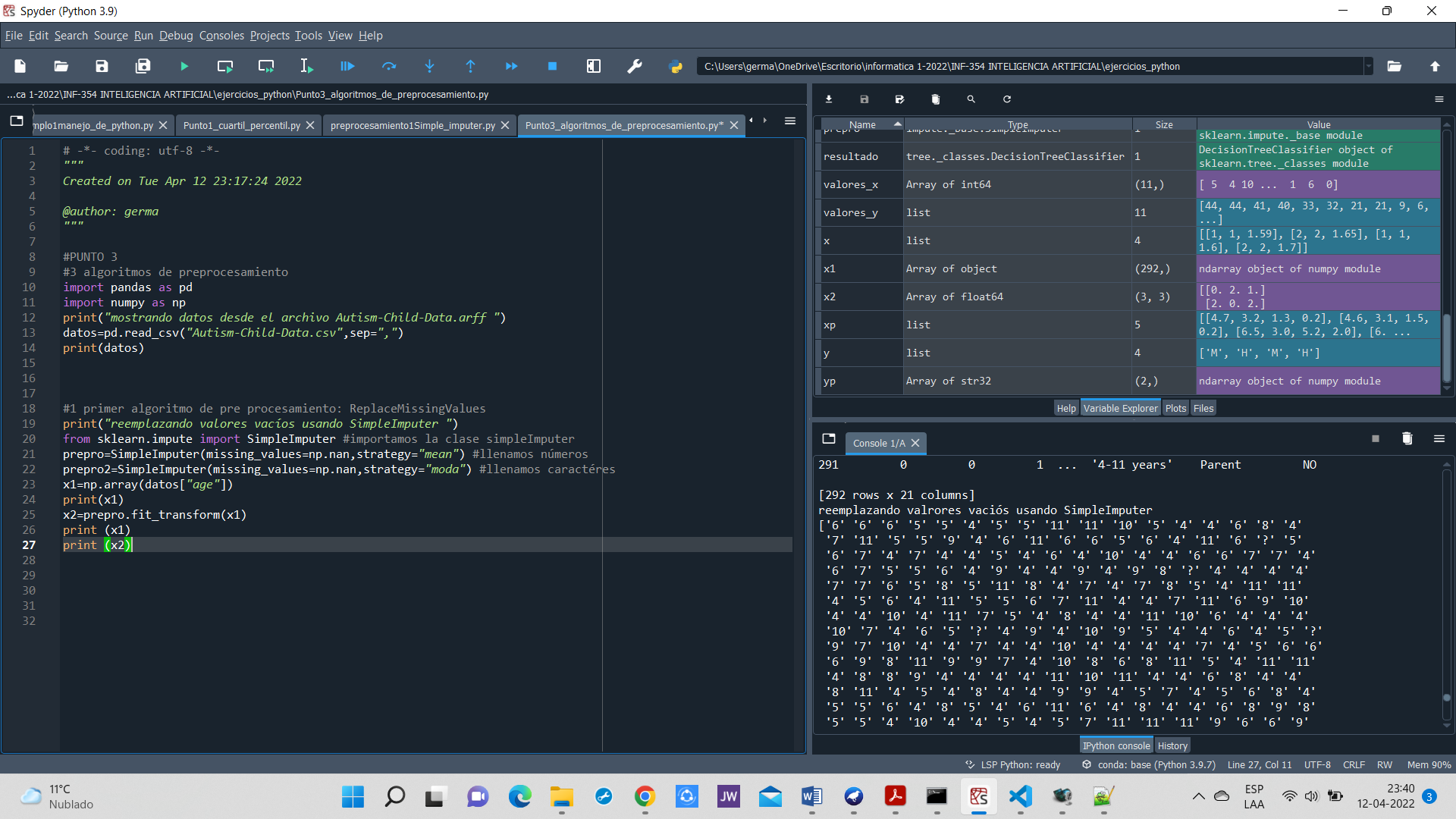




**3.** Del dataset anterior realice en **PYTHON, tres algoritmos de preprocesamiento**.

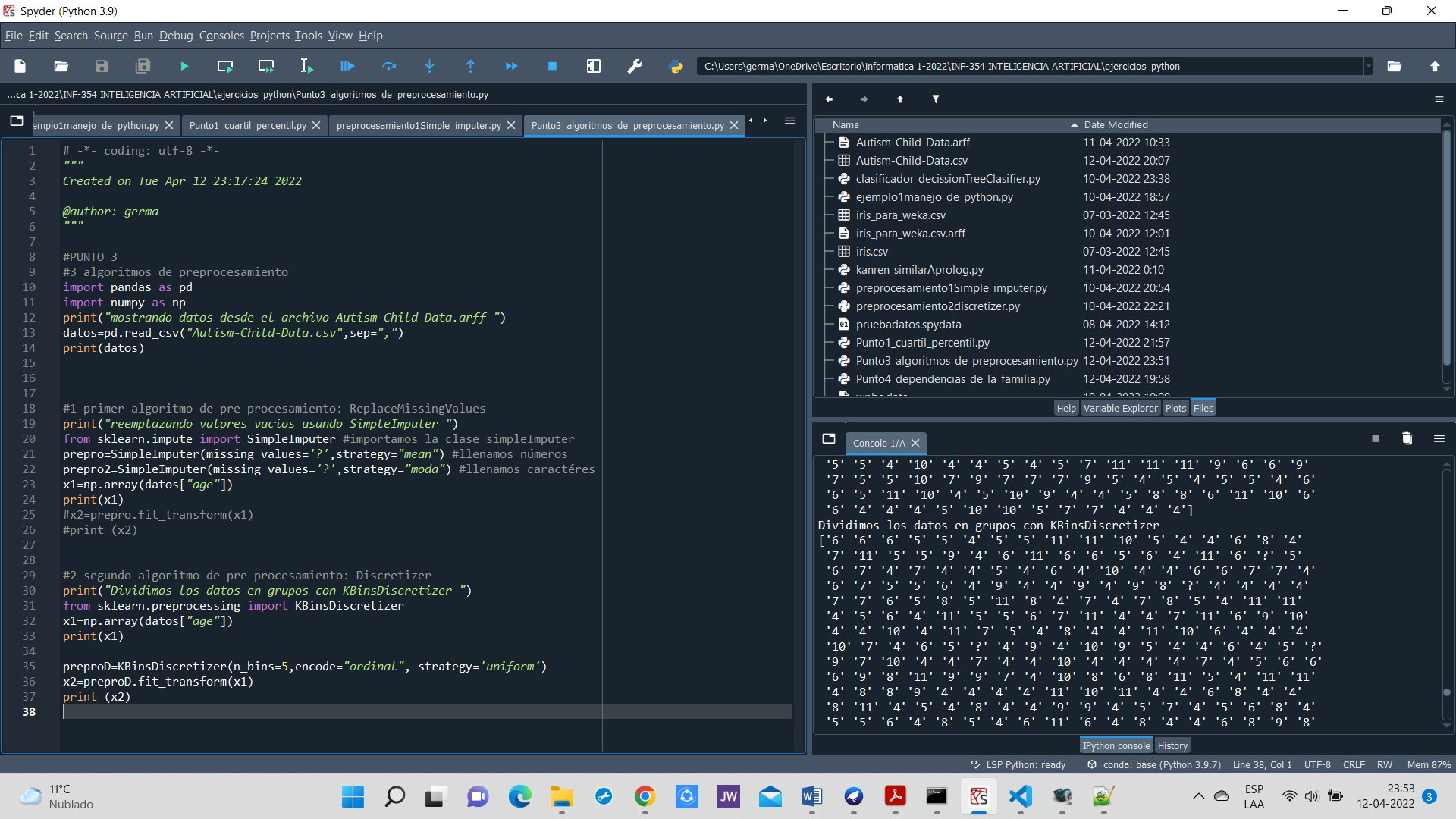
1 primer algoritmo de pre procesamiento: ReplaceMissingValues

Reemplazando valores vacíos usando **SimpleImputer**



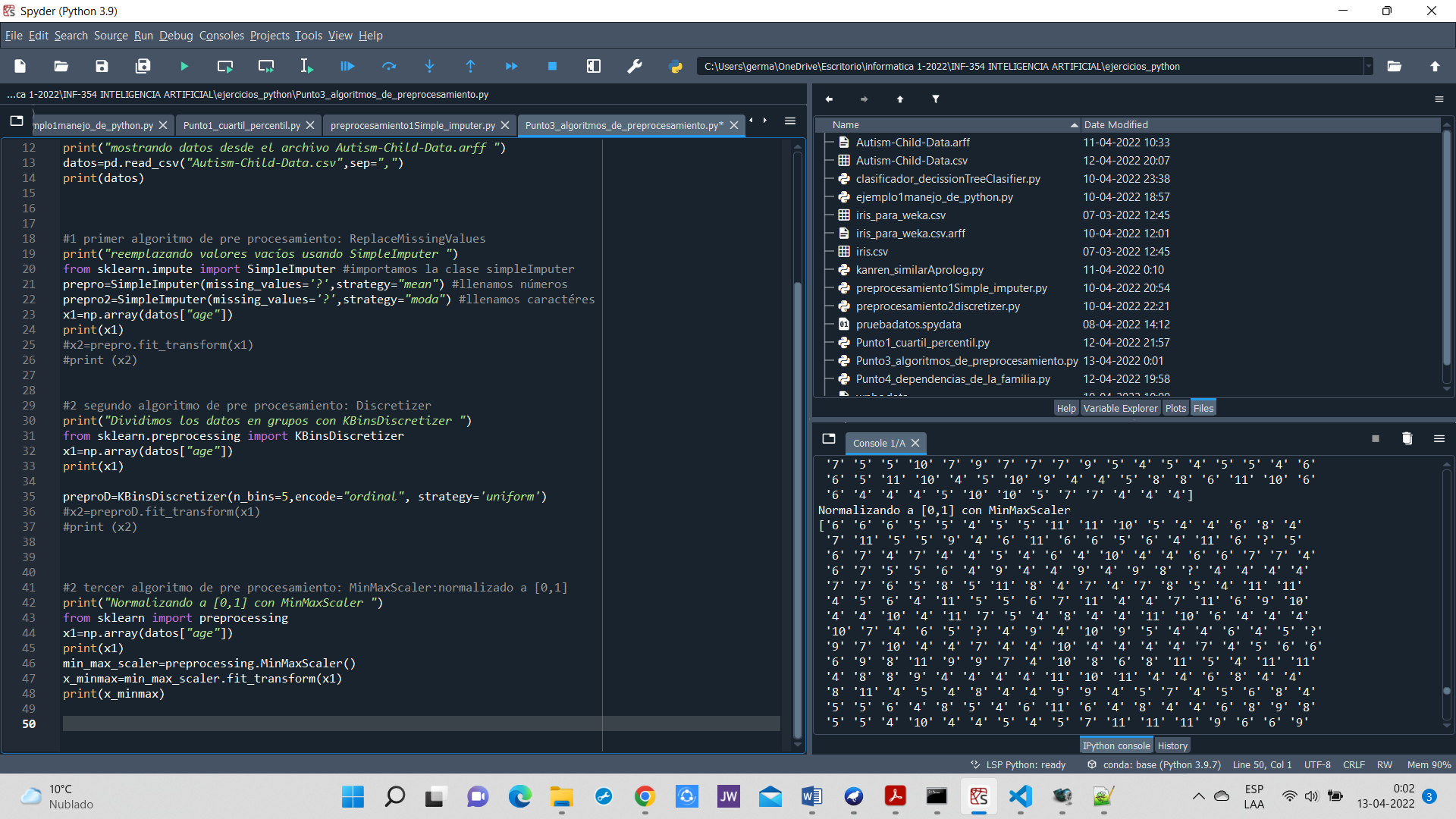
2 segundo algoritmo de pre procesamiento: Discretización

Discretizando con **KBinsDiscretizer**

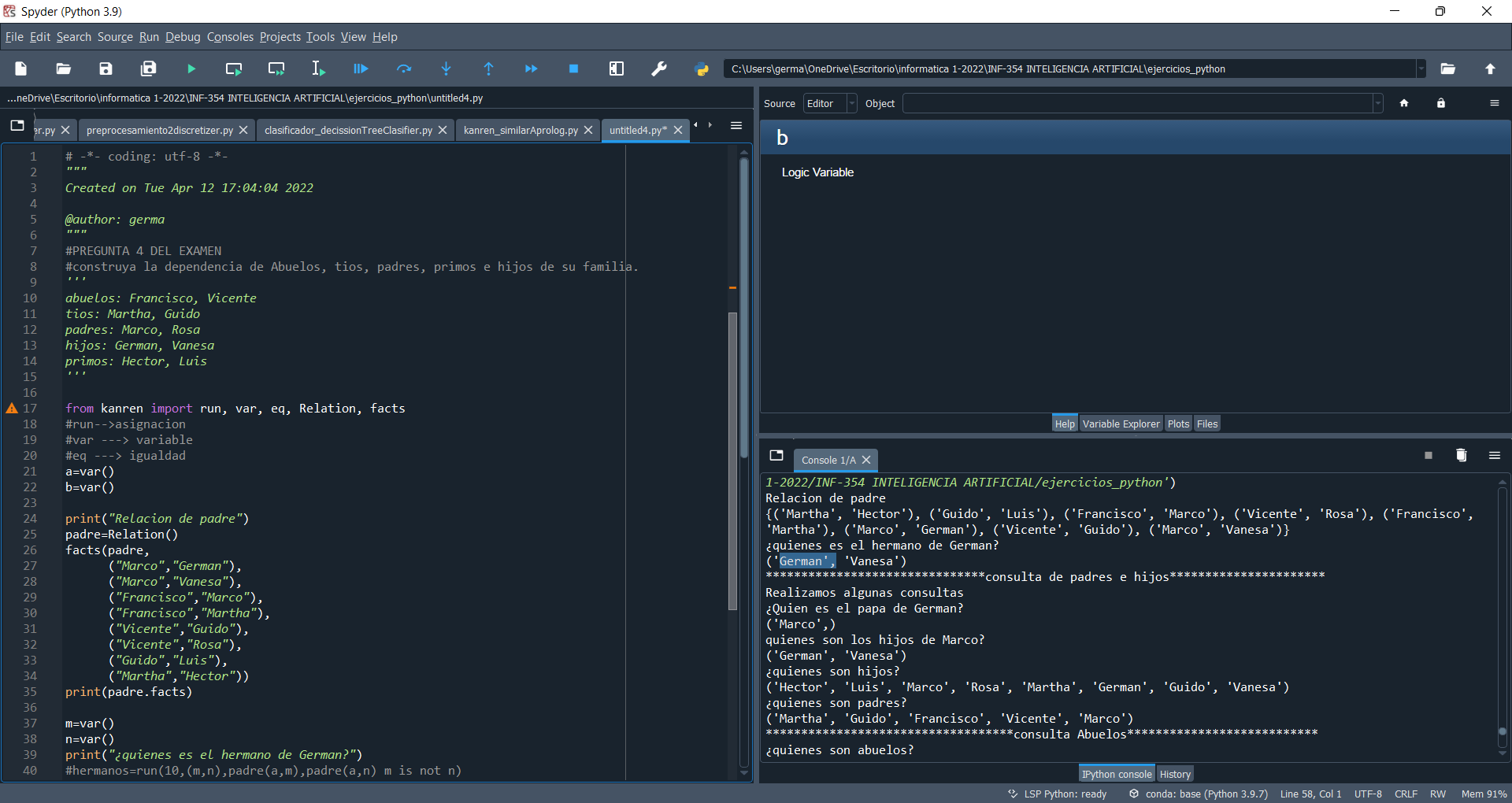


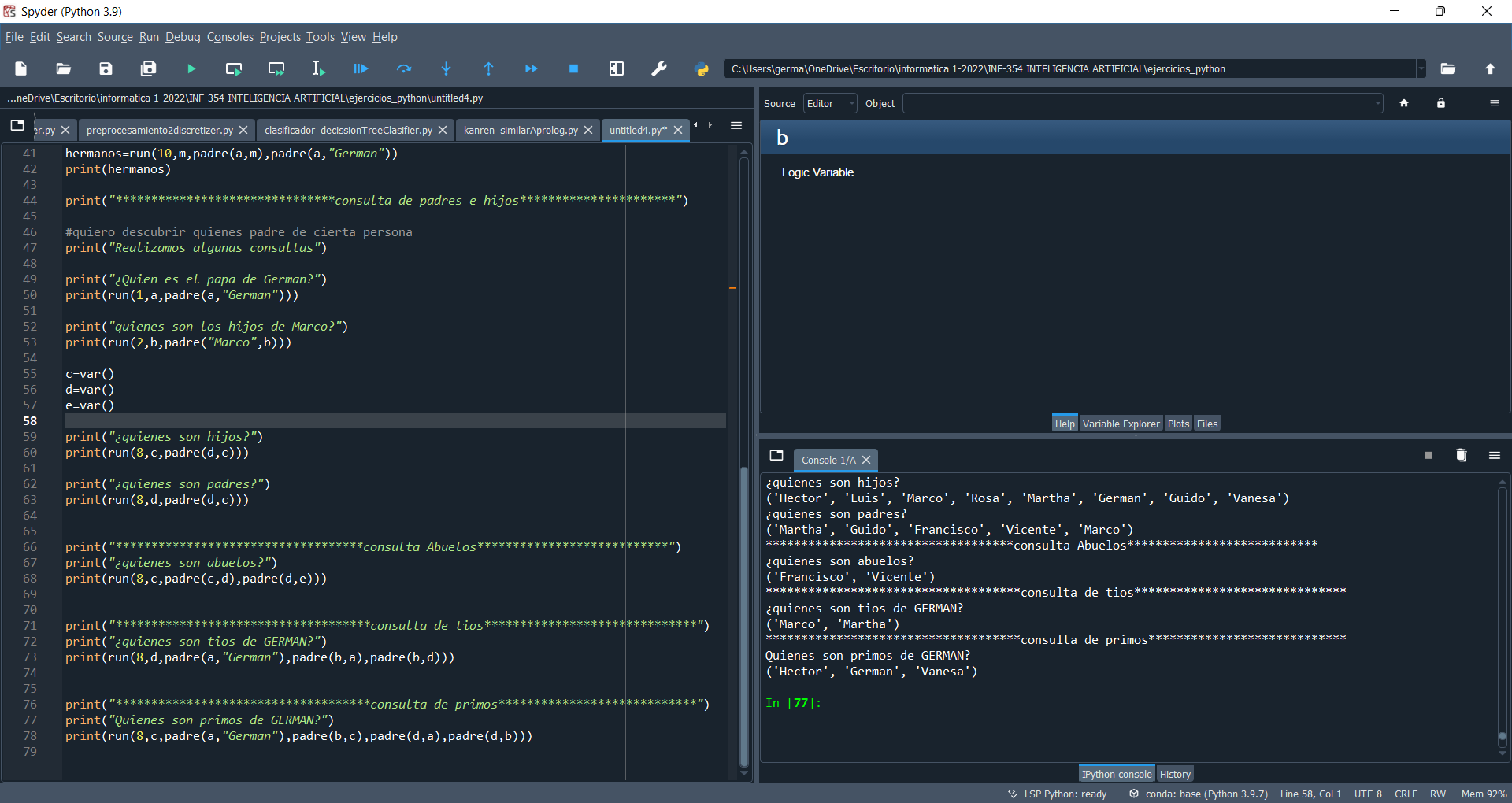
3 tercer algoritmo de pre procesamiento: Normalización

Normalizando a [0,1] con **MinMaxScaler**



**4.** Con el uso de librerías en PYTHON, construya la **dependencia de Abuelos**, tios, padres, primos e hijos de su familia.

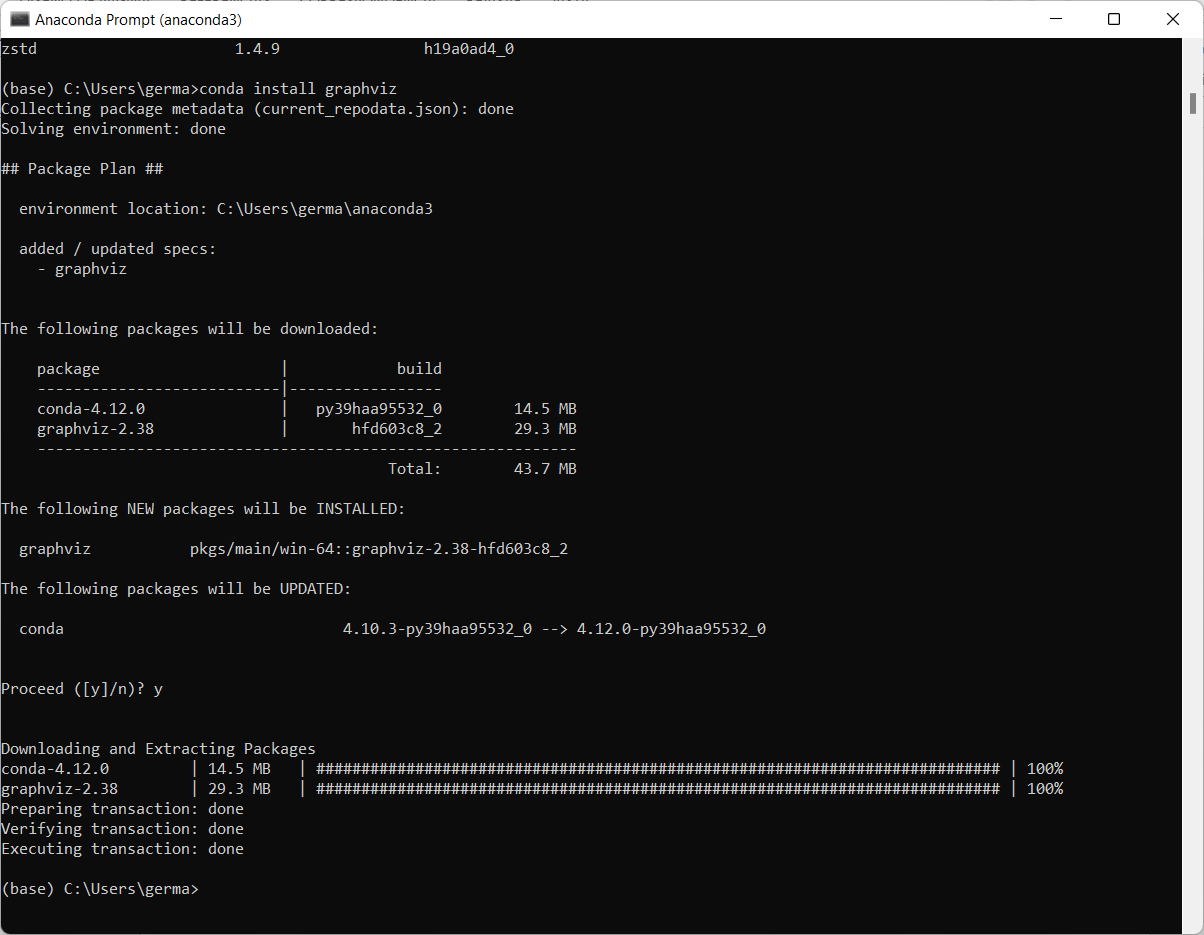




**5.** En PYTHON grafique el **árbol de decisión del dataset** seleccionado.

Para esto vamos a necesitar: **“graphviz” y “pydotplus”**

Instalamos “graphviz”



Instalamos “pydotplus”

