

Entropy-SelApp

Manual de Usuario

Tabla de contenido

1. Descripción	2
2. Disponibilidad	2
3. Pantalla de inicio	2
4. Manuales e instrucciones	3
5. Área de carga de datos	4
6. Área de resumen	5
6.1. Previsualización de datos	5
6.2. Panel de métricas	6
6.3. Panel de acciones a tomar	7
7. Área de resultados.....	8
7.1. Panel de proceso	8
7.2. Reporte de entropía y sugerencias	9
8. Detección de errores	9
8.1. Cargar conjunto de datos incorrecto	10

1. Descripción

Actualmente, los conjuntos de datos que se utilizan en diferentes industrias en los procesos de ciencia de datos contienen millones de atributos, donde algunos atributos aportan un mayor grado de información al proceso de generación de conocimiento. Durante este proceso, toma lugar la etapa de preprocesamiento de datos, la cual se enfoca en manipular y transformar el conjunto de dato, haciendo que la información sea más entendible, coherente y productiva. El proceso de preprocesamiento de datos implica elegir una medida de resultado para evaluar, posibles variables influyentes, limpiar los datos y crear características. El preprocesamiento de datos es un paso importante en el proceso de descubrimiento de conocimientos, porque las decisiones de calidad deben basarse en datos de calidad. Dentro de las principales tareas del preprocesamiento de datos, se encuentra la selección de atributo o características.

Un algoritmo de selección de atributos, es una solución computacional que tiene como finalidad determinar la relevancia de un atributo en el proceso de descubrimiento de conocimiento, a partir de una definición de una métrica la cual servirá como referencia para determinar dicha relevancia. En pocas palabras, estos algoritmos se centran en encontrar características relevantes y de esta manera mejorar el conjunto de datos utilizado en el descubrimiento de conocimiento. Una de estas técnicas es la selección de atributos basada en Entropía. La entropía como medida para ranquear características, se basa en eliminar N atributos, sin perder las características básicas del conjunto de datos, tomando como referencia una medida de similitud S que es inversamente proporcional a la distancia D entre dos instancias de N dimensiones. Estos cálculos se fundamentan en obtener la entropía del conjunto de datos eliminando cada uno de los atributos y calcular la diferencia frente a la entropía global del conjunto de datos y las entropías particulares obtenidas. Cuando todos los atributos son de tipo numérico, la medida de similitud S se obtiene mediante un método euclidiano, cuando todos los atributos son de tipo nominal se aplica el método de Hamming y en caso de tener un conjunto de datos heterogéneo es necesario aplicar métodos de discretización o numerización.

A partir de lo anterior, este producto de software realizará el proceso de selección de atributos mediante el método de entropía a un conjunto de datos seleccionado. De acuerdo a esto, permite cargar el conjunto de datos, automáticamente lo analizará y determinará cual distancia se debe utilizar si Euclideana o Hamming, o en caso de tener un conjunto de datos heterogéneo, realizará discretización de datos mediante la técnica de ChiMerge para posteriormente aplicar el método de Hamming. Al final de este proceso, el entorno expondrá al usuario cual atributo debería eliminarse de acuerdo a su valor de entropía.

2. Disponibilidad

El código fuente del software se encuentra alojado en el siguiente repositorio de GitHub: <https://github.com/AndresRestrepoRodriguez/Entropy-SelApp>

3. Pantalla de inicio

Al ingresar al portal web por medio del navegador, se observa la pantalla inicial, como se muestra en la figura 1. En ésta se puede visualizar de manera general una explicación breve de su funcionalidad.

En la barra de navegación de la parte superior, se encuentra la opción de inicio, cada vez que ésta sea oprimida se mostrará la pantalla de inicio.



Figura 1. Pantalla de inicio de la aplicación.

4. Manuales e instrucciones

En la barra de navegación también se encuentra la opción de Manuales, al ser seleccionada, se direcciona a una nueva página, donde se halla el manual de instalación, el manual de usuario y el repositorio en GitHub (véase Figura 2). Para descargar los manuales, basta con dar clic sobre el ícono del manual, que se encuentran en formato PDF.



Figura 2. Pantalla de manuales e instrucciones de la aplicación.

5. Área de carga de datos

Esta área se puede visualizar al hacer scroll down en la interfaz principal de la aplicación (Ver Figura 3), en esta zona el usuario podrá cargar el conjunto de datos al cual desea aplicar la selección de atributos. Este conjunto de datos debe ser tener como extensión **csv** o **arff**.

Cargar Conjunto de Datos

Esta sección tiene como propósito, permitir la carga del conjunto de datos deseado.

Nota: El archivo de carga debe ser de extensión **.arff** o **.csv**

Seleccionar archivo Ningún archivo seleccionado
Or Drag It Here.

Cargar

Figura 3. Carga de conjunto de datos.

Al dar clic en “Seleccionar archivo” se desplegará un explorador de archivos, donde se podrá seleccionar el conjunto de datos a utilizar. En este caso, seleccionaremos un conjunto de datos que contiene datos relacionados con las variantes rojas del vino portugués” Vinho Verde”, consolidando características fisicoquímicas. Lo anterior, se presenta en la Figura 4.

Esta sección tiene como propósito, permitir la carga del conjunto de datos deseado.

Nota: El archivo de carga debe ser de extensión **.arff** o **.csv**

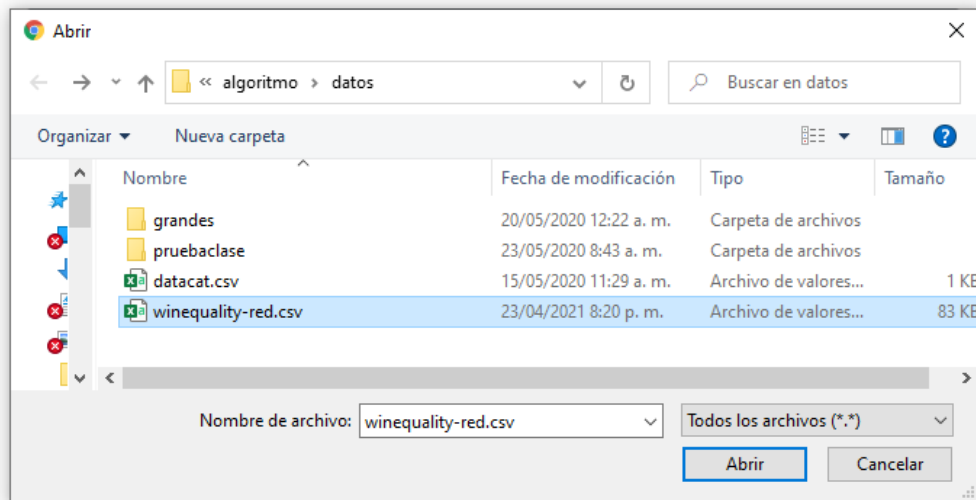


Figura 4. Explorador de archivo en carga de datos.

6. Área de resumen

Esta área se despliega una vez se da clic en el botón “Cargar” en el área de carga de datos presentada anteriormente. En esta zona de resumen el usuario podrá realizar una previsualización de sus datos. Además de esto, podrá observar algunas métricas como la cantidad de registros por cada atributo y el tipo de dato que contiene. En este punto, es importante aclarar que para que el conjunto de datos sea procesado por la aplicación, se debe haber controlado previamente los valores perdidos o missing values. Finalmente, también se podrá visualizar un área de acciones a tomar, donde se le comunicará al usuario el tipo de algoritmo que se aplicará (Euclidiana, Hamming o ChiMerge-Hamming). A continuación, se presentan las subáreas mencionadas anteriormente.

6.1. Previsualización de datos

Una vez se han cargado los datos, el usuario podrá visualizarlos como se presenta en la fig xxx. Se podrán ver todos los registros en su totalidad, contando con 10 entradas por defecto por hoja de visualización. Lo anterior, se expone en la Figura 5.

Overview											
Esta sección expone el conjunto de datos cargado, la cantidad de registros por atributo y el tipo de cada uno de estos											
Show 10 entries											
fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	qual
7.4	0.7	0.0	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
7.8	0.88	0.0	2.6	0.098	25.0	67.0	0.9968	3.2	0.68	9.8	5
7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.997	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.998	3.16	0.58	9.8	6
7.4	0.7	0.0	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
7.4	0.66	0.0	1.8	0.075	13.0	40.0	0.9978	3.51	0.56	9.4	5
7.9	0.6	0.06	1.6	0.069	15.0	59.0	0.9964	3.3	0.46	9.4	5
7.3	0.65	0.0	1.2	0.065	15.0	21.0	0.9946	3.39	0.47	10.0	7
7.8	0.58	0.02	2.0	0.073	9.0	18.0	0.9968	3.36	0.57	9.5	7
7.5	0.5	0.36	6.1	0.071	17.0	102.0	0.9978	3.35	0.8	10.5	5
Showing 1 to 10 of 1,599 entries											
Previous 1 2 3 4 5 ... 160 Next											

Figura 5. Panel de importación de modelo propio del usuario

6.2. Panel de métricas

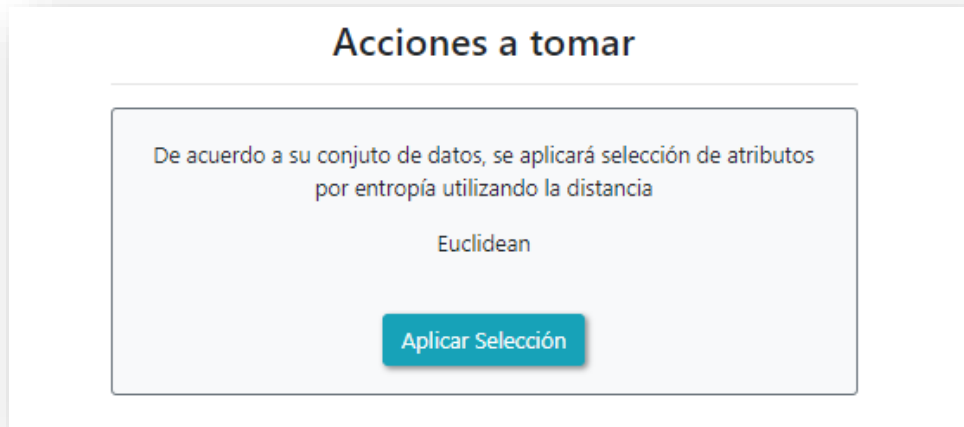
El usuario podrá visualizar cada uno de los atributos con su respectivo número de registros con valores y su respectivo tipo de atributo. En esta sección se podrán ver lo tipo de datos numérico y nominal. Lo anterior lo presenta la Figura 6.

Cantidad de registros		Tipos de atributos	
Atributo	Registros	Atributo	Tipo
fixed acidity	1599	fixed acidity	numeric
volatile acidity	1599	volatile acidity	numeric
citric acid	1599	citric acid	numeric

Figura 6. Panel de métricas

6.3. Panel de acciones a tomar

Una vez la plataforma ha realizado un análisis preliminar del conjunto de datos que el usuario ha cargado y determina los tipos de datos que tienen los atributos, se le proporciona al usuario las acciones a tomar. En este caso, como lo muestra la Figura 7, se determina aplicar la selección de atributos por entropía utilizando la distancia Euclidiana, dado que, todos los atributos son de tipo numérico.



Acciones a tomar

De acuerdo a su conjunto de datos, se aplicará selección de atributos por entropía utilizando la distancia

Euclidean

Aplicar Selección

Figura 7. Panel de acciones a tomar

En el caso que, el conjunto de datos sea heterogéneo, es decir, contenga atributos tanto de tipo numérico como nominal, el panel de acciones a tomar le expondrá al usuario, la decisión de realizar un proceso de discretización mediante el método de ChiMerge. Adicional a ello, se le solicitará al usuario determinar el atributo de referencia de tipo nominal para realizar este proceso, además de, el nivel de confianza, el cual se debe encontrar en un rango de $[0.1, 0.99]$. Lo anterior, se evidencia en la Figura 8.



Acciones a tomar

De acuerdo a su conjunto de datos, se aplicará discretización por el método

Chimerge

Para posteriormente aplicar selección de atributos por entropía utilizando la distancia Hamming

Seleccione el atributo referencia

x9

Seleccione su nivel de confianza

0,9

Aplicar Selección

Figura 8. Panel de acciones a tomar con conjunto heterogéneo

Una vez se ha visualizado el panel de acciones a tomar y proporcionado los datos solicitados, cuando sea requerido. Acto seguido se debe dar clic en el botón “Aplicar Selección”, lo cuál provocará que se presente el siguiente cuadro de espera (Ver Figura 9).

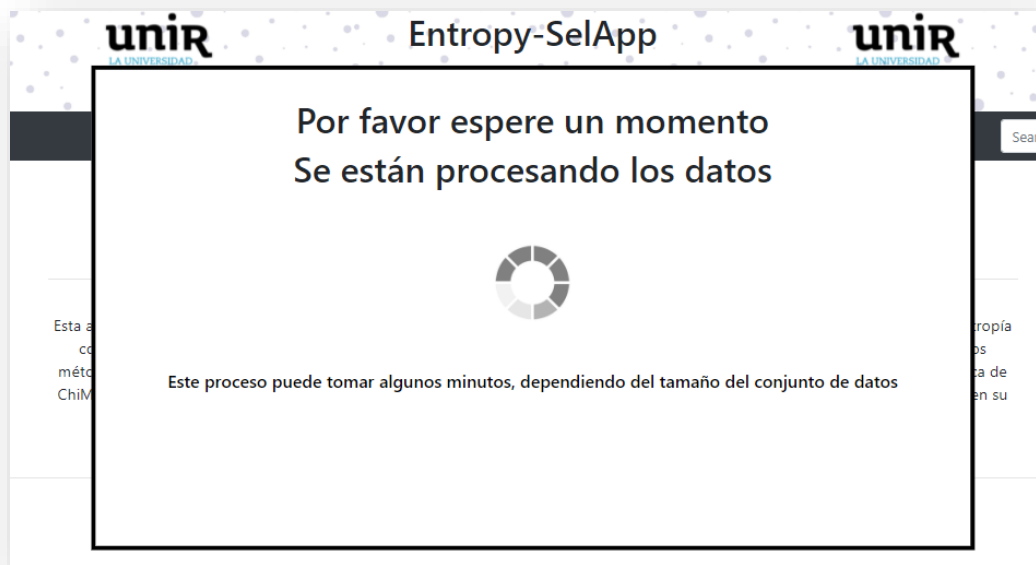


Figura 9. Pop-up de espera durante el procesamiento de datos

7. Área de resultados

Finalmente, una vez se ha aplicado el algoritmo de selección de atributos por entropía por la distancia determinada y apropiado, se presenta al usuario tres secciones diferentes. Estas secciones son: el área de proceso, área de entropía y área de sugerencia.

7.1. Panel de proceso

En este panel se presenta al usuario todo el proceso realizado. En este caso al haber determinado la distancia Euclidiana como método para el cálculo de la distancia, se va presentando el conjunto de pasos realizados por el algoritmo, lo anterior se puede ver en la Figura 10. En el caso de tener un dataset heterogéneo, se presenta los pasos e intervalos generados en el proceso de discretización por ChiMerge y adicionalmente, los pasos para el método por entropía.

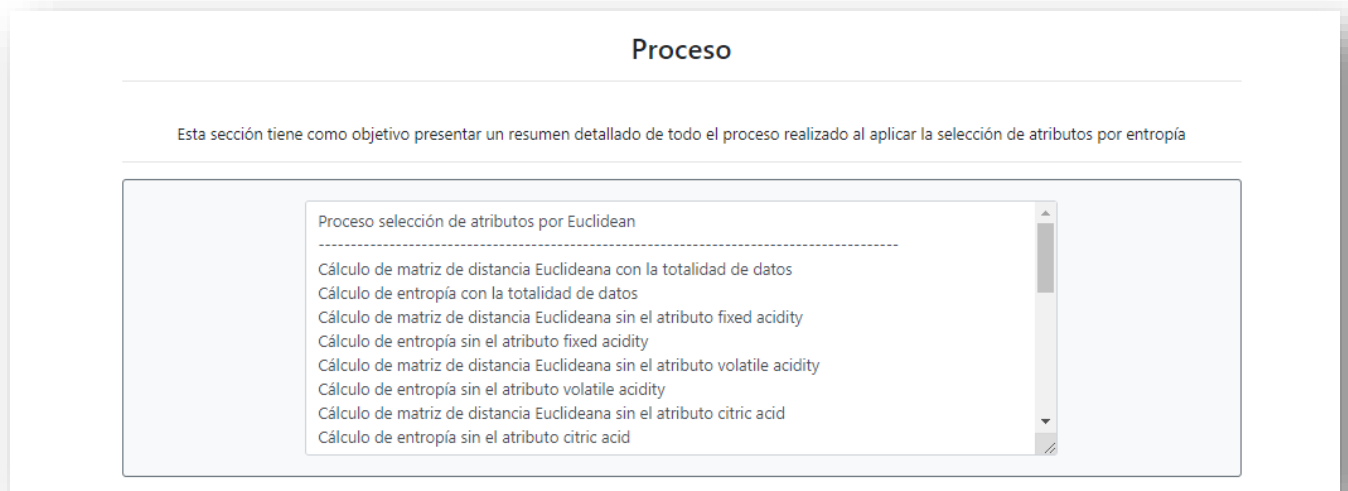


Figura 10. Ejemplo de grafica de curva de ROC a nivel micro

7.2. Reporte de entropía y sugerencias

Finalmente, en esta sección se presentan los resultados y la sugerencia obtenida a partir de la aplicación del algoritmo de selección de atributos por entropía. En primera instancia, tenemos los resultados, en donde se consolidan los valores de entropía con su respectiva diferente, tomando como valor de referencia la entropía general. Adicionalmente, se presenta la sugerencia, donde este cuadro de dialogo, expone cual o cuales atributos del conjunto de datos debería eliminarse, lo anterior obteniendo la o las menores diferentes entre las entropías específicas y la general. Ver Figura 11.

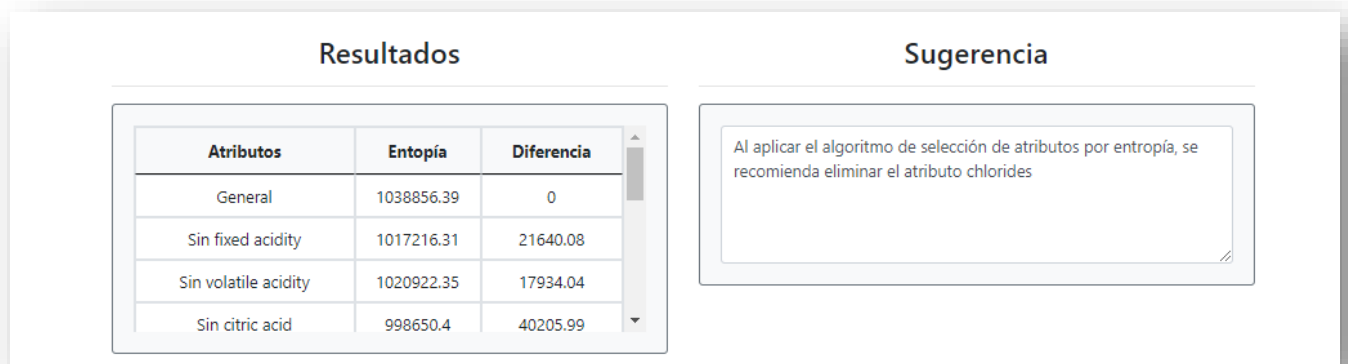


Figura 11. Reporte de entropía y sugerencias

8. Detección de errores

La detección de errores hace referencia a aquellas situaciones en las que las funcionalidades del sistema no se ejecutarán si no se cumplen con ciertas condiciones.

8.1. Cargar conjunto de datos incorrecto

Como se ha mencionado anteriormente, el conjunto de datos que se debe cargar debe cumplir con dos características fundamentales:

1. El conjunto de datos debe tener extensión. arff o .csv
2. El conjunto de datos debe haber pasado previamente por un proceso de imputación de valores perdidos o missing values.

En caso de que el conjunto de datos cumpla con las condiciones anteriormente plasmada, se presenta un mensaje de éxito en la carga de datos, tal como expone la fig xxx. De otra manera, se presentará al usuario la notificación de fallo en el proceso de carga de datos. Ver figura xxx.

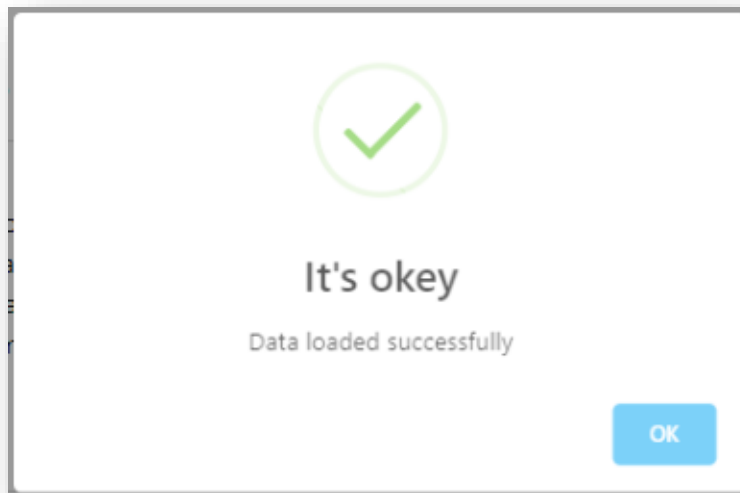


Figura 18. Panel con muestra de éxito al cargar el modelo propio del usuario.

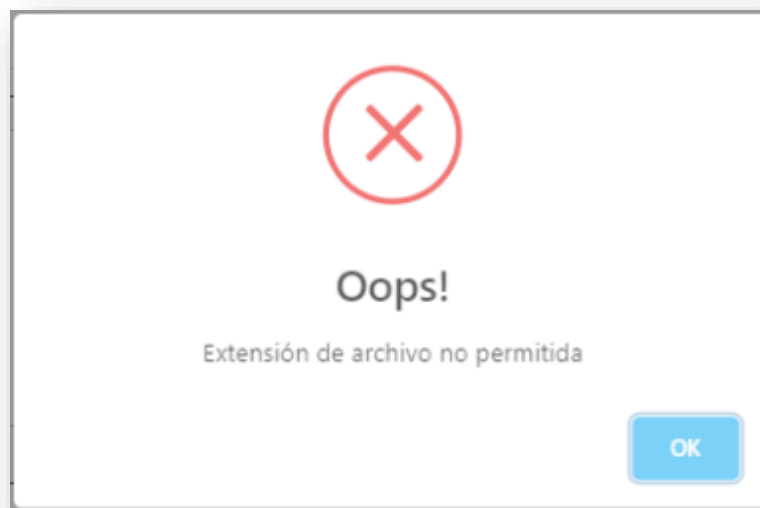


Figura 19. Panel con muestra de fallo al cargar el modelo propio del usuario.

