

UNIVERSIDAD NACIONAL
AUTÓNOMA DE MÉXICO

Proyecto Final

INTELIGENCIA ARTIFICIAL

GRUPO 03

HANNA SOPHIA CERES MARTÍNEZ

09-DICIEMBRE-2021

Índice

1. Objetivo	2
2. Introducción	3
3. Análisis de requisitos	7
3.1. Requerimientos Funcionales	7
3.2. Requerimientos No Funcionales	7
4. Tecnología Utilizada	8
5. Descripción del funcionamiento de la aplicación	8
5.1. Ejecutar la aplicación	8
5.2. Ventana Principal	9
5.3. Importar Datos	10
5.4. Mostrar Datos	11
5.5. Selección de Características	12
6. Descripción del funcionamiento de los algoritmos	15
6.1. Apriori	15
6.2. Métricas de Distancia	17
6.3. Clustering Particional	19
6.4. Clustering Jerárquico	22
6.5. Regresión Logística	24
6.6. Árboles Aleatorios - Pronóstico	28
6.7. Árboles Aleatorios - Clasificación	33
Referencias	38

1. Objetivo

Elaborar una aplicación donde se implementen los algoritmos de aprendizaje automático vistos durante el semestre, con el propósito de que cualquier usuario pueda utilizarla para analizar sus datos.

2. Introducción

Hoy en día los datos se han vuelto una herramienta poderosa. Las empresas suelen analizar los datos de sus usuarios o de una población en específica con diferentes fines, por ejemplo: sistemas de recomendaciones, marketing digital, medir similitudes entre un grupo o sacar diferencias entre objetos, creación de grupos de elementos con características en común, etc.

El Machine Learning nos facilita realizar esta tarea por medio de algoritmos de aprendizaje no supervisado, supervisado, reforzado y profundo.

En la aplicación *CERES* encontrarás algunos de estos algoritmos con los que podrás analizar datos de tu interés. A continuación, se describirán brevemente los algoritmos para que el usuario que lo utilice tenga un conocimiento mayor de ellos.

Apriori

Es un algoritmo de aprendizaje no supervisado basado en reglas, estas reglas encuentran relaciones ocultas en los datos.

Para obtener las reglas se deben ingresar las siguientes mediciones:

- Soporte: Indica cuan importante es una regla en el total de transacciones.
- Confianza: Indica que tan fiable es una regla.
- Lift: Indica el aumento de probabilidad entre el antecedente y consecuente de la regla.

Métricas de Distancia

Las métricas de distancia nos proporciona la diferencia que existe entre dos objetos, por ejemplo: compras, personas, ventas,

etc. Son importantes para identificar objetos con características similares y no similares.

- Distancia Euclidiana: es utilizada para calcular la distancia entre dos puntos. Hace uso del teorema de Pitágoras.
- Distancia Chebyshev: es el valor máximo absoluto de las diferencias entre las coordenadas de un par de elementos.
- Distancia Manhattan: es utilizada para calcular la distancia entre dos puntos en una ruta similar a una cuadrícula.
- Distancia Minkowski: es una distancia entre dos puntos en un espacio n-dimensional.

Clustering Jerárquico

Este algoritmo de aprendizaje no supervisado organiza los elementos en una estructura de árbol. Lo que el árbol representa es la relación de similitud entre los objetos, para esto se apoya de las métricas de distancia.

Clustering Particional

Es un algoritmo de aprendizaje no supervisado que organiza los elementos similares dentro de k clústeres.

Nuestro modulo utiliza el algoritmo K-means para crear los clústeres a partir de un conjunto de elementos.

Clasificación

La clasificación predice clases de tipo discretas (1,2,3) o nominales (a,b,c).

El modelo se construye a través de un conjunto de datos de entrenamiento y se evalúa con un conjunto de datos prueba.

Matriz de clasificación

La matriz de clasificación se utiliza para evaluar una clasificación. La variable clase de entrenamiento toma dos valores, ya sea un dato nominal o discreto el dato será falso o verdadero.

Los valores positivos y negativos que se predicen correctamente se conocen como verdaderos positivos (VP) y verdaderos negativos (VN), respectivamente. Contrario a lo anterior, los valores clasificados incorrectamente se denominan falsos positivos (FP) y falsos negativos (FN).

Mediciones

- Exactitud: es el porcentaje de los datos clasificados correctamente.
- Precisión: es el porcentaje de clasificación positiva.
- Tasa de error: es el porcentaje de los datos clasificados incorrectamente.
- Sensibilidad: es el porcentaje de clasificación del total positivos.
- Especificidad: es el porcentaje de clasificación del total negativos.
- Exactitud: es el grado de conformidad.
- Precisión: es el grado de reproducibilidad.

Regresión Logística

Es un algoritmo de aprendizaje supervisado que predice valores binarios. Se basa en un proceso de clasificación.

Los resultados obtenidos serán una matriz de clasificación y las mediciones de este método (exactitud, tasa de error, precisión, sensibilidad y especificidad).

Árboles de Decisión

Es un algoritmo de aprendizaje supervisado que permite resolver problemas de pronóstico y clasificación. Los datos ingresados pueden ser valores numéricos y nominales.

Parámetros del árbol de decisión

- `max_depth`: indica la profundidad máxima del árbol.
- `min_samples_leaf`: indica la cantidad mínima de datos que debe tener un nodo hoja.
- `min_samples_split`: indica la cantidad mínima de datos para que un nodo de decisión se pueda dividir.

Árboles de decisión (Regresión)

Este árbol es aplicado a problemas de pronóstico. Admite valores continuos.

Criterio de regresión

- `MSE`: establece el valor pronosticado de los nodos terminales con respecto al valor medio aprendido.
- `MAE`: establece el valor pronosticado de los nodos terminales con respecto a la mediana.
- `RMSE`: es la raíz del error cuadrático medio.

Árboles de decisión (Clasificación)

Este árbol es aplicado a problemas de clasificación. Admite valores continuos.

Al seguir la metodología de clasificación los resultados serán una matriz de clasificación y las mediciones de este método (exactitud, tasa de error, precisión, sensibilidad y especificidad).

3. Análisis de requisitos

3.1. Requerimientos Funcionales

- La aplicación tiene un modulo por cada algoritmo
- El usuario podrá subir su archivo de datos, con una extensión .CSV
- En los algoritmos que se requiera el usuario podrá ingresar los valores en los parámetros para el funcionamiento del algoritmo
- La aplicación mostrará los resultados de cada operación realizada

3.2. Requerimientos No Funcionales

Lenguaje de programación:

- La aplicación será desarrollada en el lenguaje de programación Python

Acceso:

- El usuario tendrá acceso a la aplicación desde cualquier dispositivo que cuente con dispositivo con acceso a un navegador

Usabilidad:

- La aplicación debe tener una interfaz gráfica de usuario intuitiva y clara
- La aplicación mostrará “warnings” en caso de errores

Rendimiento:

- Los resultados de los algoritmos serán generados en tiempos no mayores a 3 minutos

4. Tecnología Utilizada

Es proyecto será implementado en StreamLit.

¿Qué es StreamLit?

StreamLit es una herramienta que puede convertir los scripts de datos en aplicaciones web automáticamente.

Se eligió esta herramienta porque facilita la manera de crear una interfaz para el usuario, tiene una amplia documentación y además está enfocada en el aprendizaje automático y la ciencia de datos.

5. Descripción del funcionamiento de la aplicación

5.1. Ejecutar la aplicación

En esta primera versión la aplicación solo puede ejecutarse de manera local, para hacerlo se siguen los siguientes pasos:

- 1 Descarga el archivo llamado “CMHS_proyFinal.py.”
- 2 Sitúate en la carpeta donde esta guardado el archivo.
- 3 Escribe el comando “streamlit run CMHS_proyFinal.py.”

Se abrirá una ventana en tu navegador y la aplicación estará lista para utilizarse.

5.2. Ventana Principal

Al abrir la aplicación se tendrá una ventana principal donde se le dará la bienvenida al usuario:

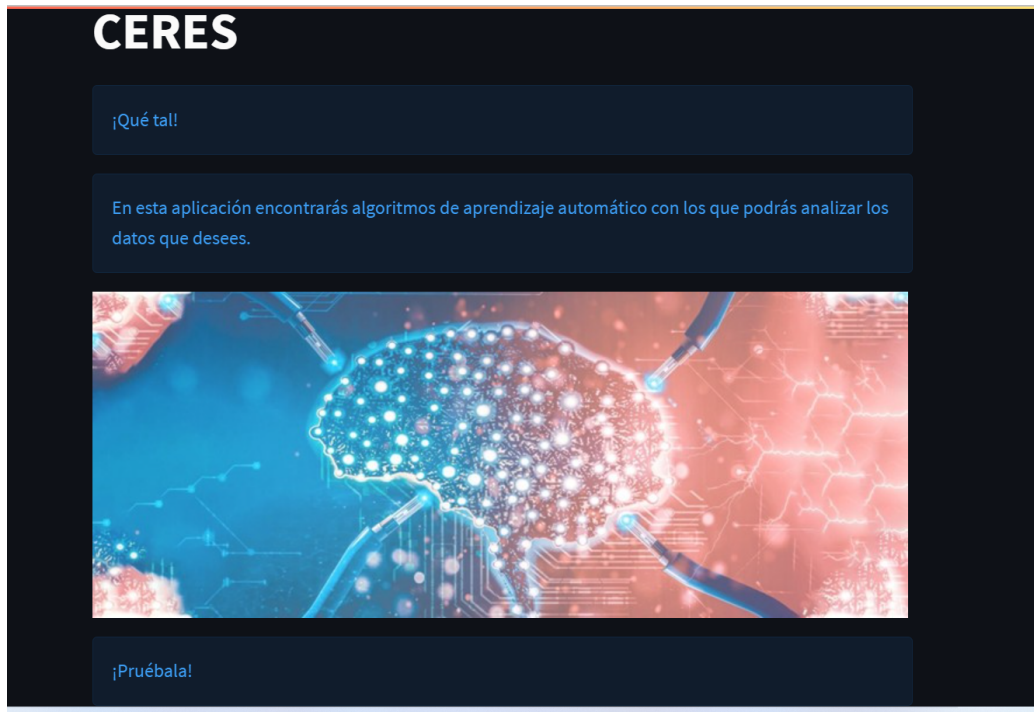


Figura 1: Página de inicio de la aplicación.

El usuario podrá elegir el algoritmo que quiera utilizar en la sección de pestañas que se encuentra en la parte izquierda:

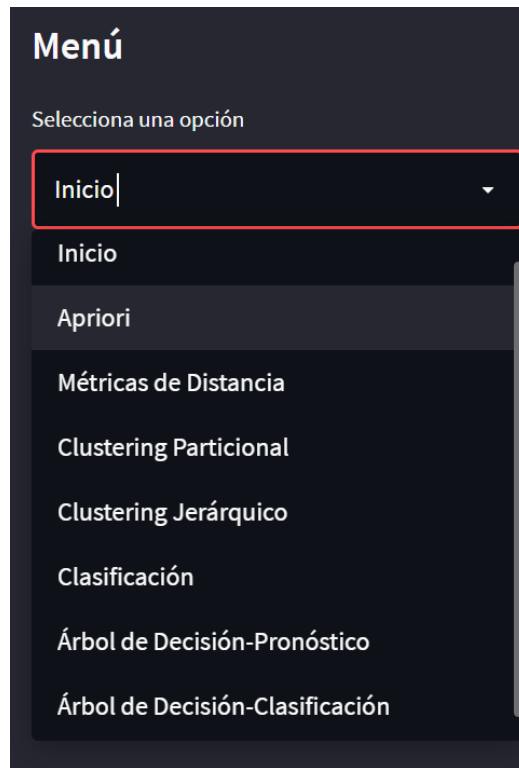


Figura 2: Menú de algoritmos.

5.3. Importar Datos

En cada pestaña el usuario tendrá la opción de subir el archivo con los datos que requiera cargar:

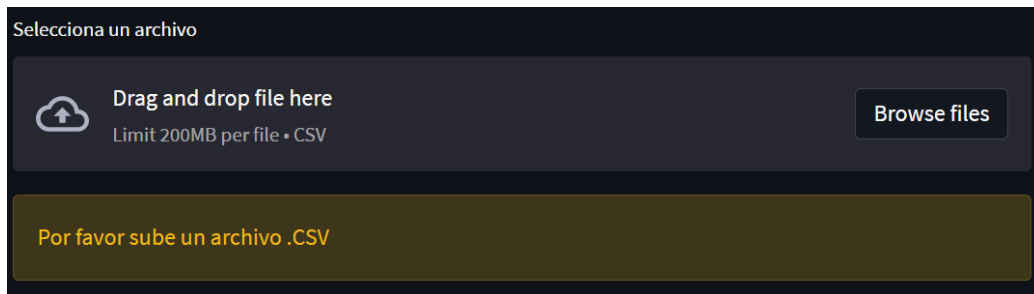



Figura 3: Importación de datos.

Los datos deberán tener una extensión .CSV. No se admite ningún otro tipo de datos.

Los datos cargados en un módulo permanecerán guardados en todos los módulos. Si en un módulo desea trabajar con nuevos datos solo tendrá que subir el nuevo archivo.

5.4. Mostrar Datos

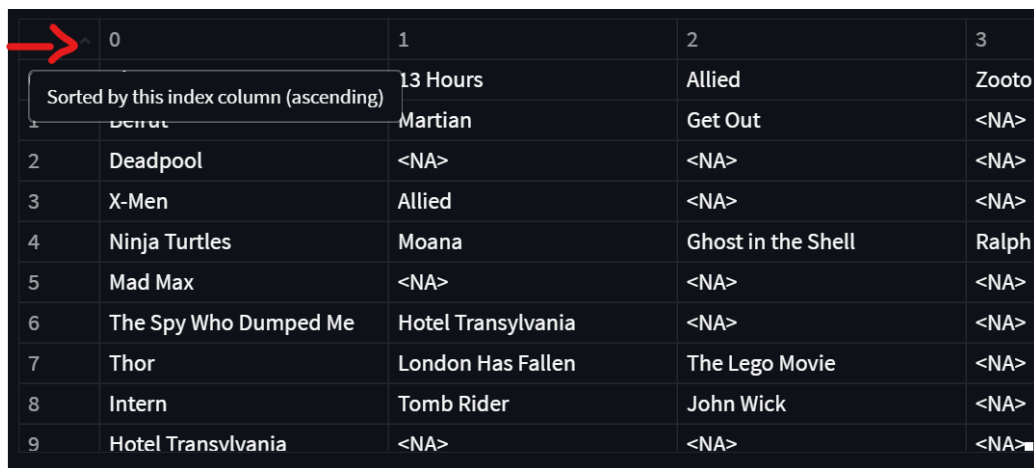
Las tablas de datos o gráficas se pueden expandir posicionando el cursor en la parte superior derecha.



	0	1	2	3	
0	The Revenant	13 Hours	Allied	Zooto	
1	Beirut	Martian	Get Out	<NA>	
2	Deadpool	<NA>	<NA>	<NA>	
3	X-Men	Allied	<NA>	<NA>	
4	Ninja Turtles	Moana	Ghost in the Shell	Ralph	
5	Mad Max	<NA>	<NA>	<NA>	
6	The Spy Who Dumped Me	Hotel Transylvania	<NA>	<NA>	
7	Thor	London Has Fallen	The Lego Movie	<NA>	
8	Intern	Tomb Rider	John Wick	<NA>	
9	Hotel Transylvania	<NA>	<NA>	<NA>	

Figura 4: Expansión de la tabla o gráfica.

Los datos en ellos podrán ser ordenados de manera ascendente o descendente.



A screenshot of a data table interface. A red arrow points to the header of the first column, which is labeled '0'. A tooltip box is visible over the first row, stating 'Sorted by this index column (ascending)'. The table contains 10 rows of data, with the first row being the header and the subsequent 9 rows containing movie titles and other information. The first column is the index, and the other columns contain movie titles and other data.

0	1	2	3
1	Deadpool	13 Hours	Allied
2	X-Men	Martian	Get Out
3	Ninja Turtles	<NA>	<NA>
4	Mad Max	Allied	<NA>
5	The Spy Who Dumped Me	Moana	Ghost in the Shell
6	Thor	<NA>	<NA>
7	Intern	Hotel Transylvania	<NA>
8	Hotel Transylvania	London Has Fallen	The Lego Movie
9		Tomb Rider	John Wick
10		<NA>	<NA>

Figura 5: Ordenamiento de la información.

5.5. Selección de Características

El usuario podrá hacer la selección de características para su modelo de tres formas:

Gráfica de dispersión

El usuario podrá ingresar cualquier par de variables de sus datos y se generará la gráfica correspondiente.

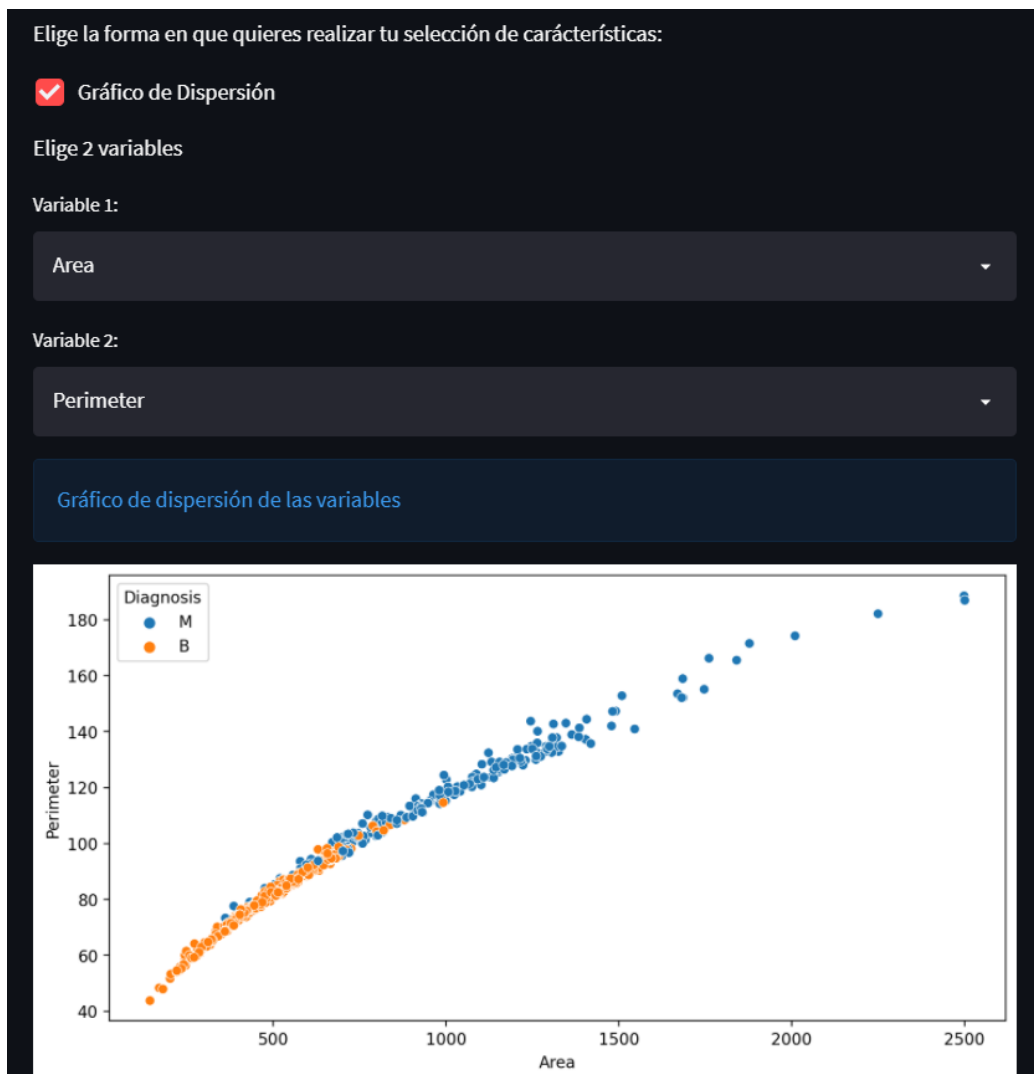


Figura 6: Gráfica de dispersión.

Matriz de correlaciones

La matriz de correlaciones será mostrada automáticamente al dar clic.

✓ Matriz de Correlaciones

Matriz de correlaciones de tus datos

	Radius	Texture	Perimeter	Area	Smoothness	Compactness
Radius	1.0000	0.3238	0.9979	0.9874	0.1706	0.5061
Texture	0.3238	1.0000	0.3295	0.3211	-0.0234	0.2367
Perimeter	0.9979	0.3295	1.0000	0.9865	0.2073	0.5569
Area	0.9874	0.3211	0.9865	1.0000	0.1770	0.4985
Smoothness	0.1706	-0.0234	0.2073	0.1770	1.0000	0.6591
Compactness	0.5061	0.2367	0.5569	0.4985	0.6591	1.0000
Concavity	0.6768	0.3024	0.7161	0.6860	0.5220	0.8831
ConcavePoints	0.8225	0.2935	0.8510	0.8233	0.5537	0.8311
Symmetry	0.1477	0.0714	0.1830	0.1513	0.5578	0.6026
FractalDimension	-0.3116	-0.0764	-0.2615	-0.2831	0.5848	0.5654

Figura 7: Matriz de correlaciones.

Mapa de calor

El mapa de calor será mostrado automáticamente al dar clic.

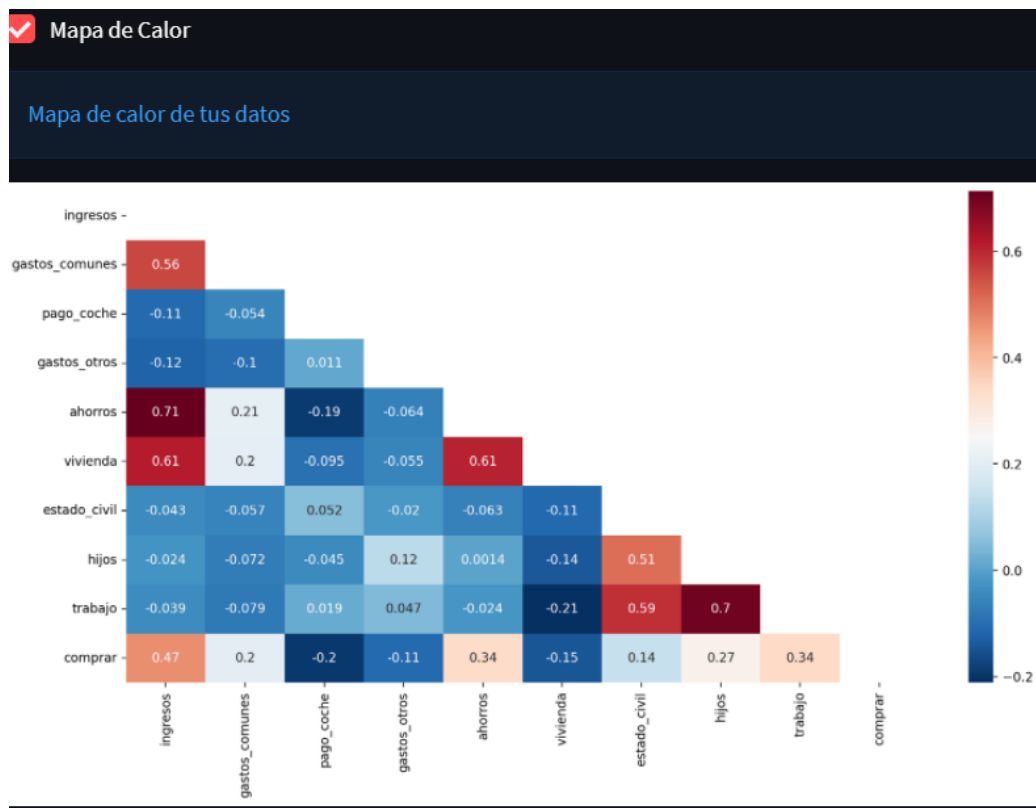


Figura 8: Mapa de calor.

6. Descripción del funcionamiento de los algoritmos

6.1. Apriori

Una vez cargados nuestros datos en la aplicación se desplegará la visualización de los mismos, su tabla de frecuencias y gráfica de frecuencias.

Después el usuarios podrá ingresar el valor para cada una de las reglas y obtendrá como resultado el número de reglas ingresadas y la información de cada una de ellas.

Ingresa el valor para tus reglas:

Soporte:

0.01 - +

Confianza:

0.30 - +

Lift:

2.00 - +

Figura 9: Ingresar reglas para algoritmo apriori.

Número de reglas encontradas:

9

Reglas

Regla: frozenset({'Kung Fu Panda', 'Jumanji'})

Soporte: 0.0160857908847185

Confianza: 0.3234501347708895

Lift: 3.2784483768897226

Figura 10: Resultados de algoritmo apriori.

6.2. Métricas de Distancia

Una vez cargados los datos el usuario puede elegir el tipo de métrica que desea usar.

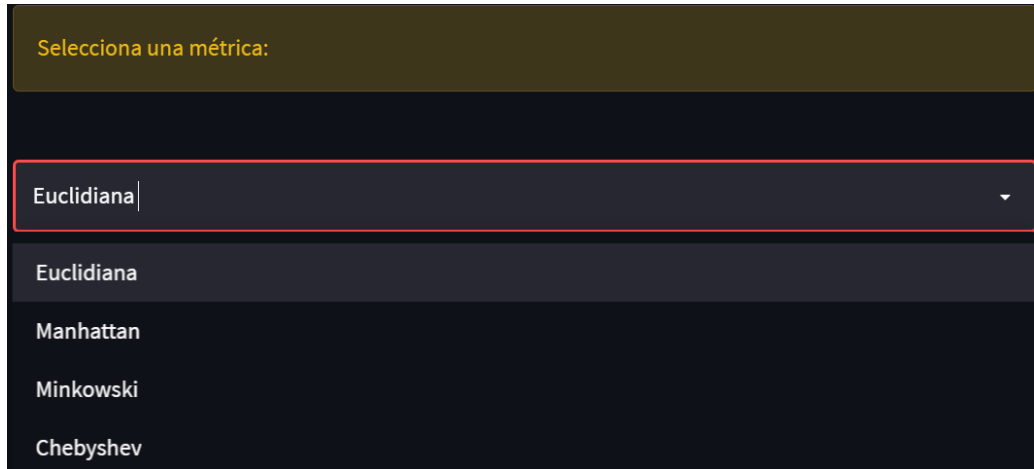


Figura 11: Elegir métrica de distancia.

Si quiere medir un par de elementos deberá seleccionar la opción “Sí” y ingresar el ID de ellos, como resultado tendrá la distancia entre estos dos, si no, seleccionará la opción “No” y obtendrá la matriz de distancias.

¿Deseas medir un par de registros de tus datos?

☒ Sí

Selecciona los registros:

Primer registro:

100.00 - +

Segundo registro:

200.00 - +

Distancia Euclidiana entre los registros

37413.775925452916

Figura 12: Medir un par de objetos.

☒ No

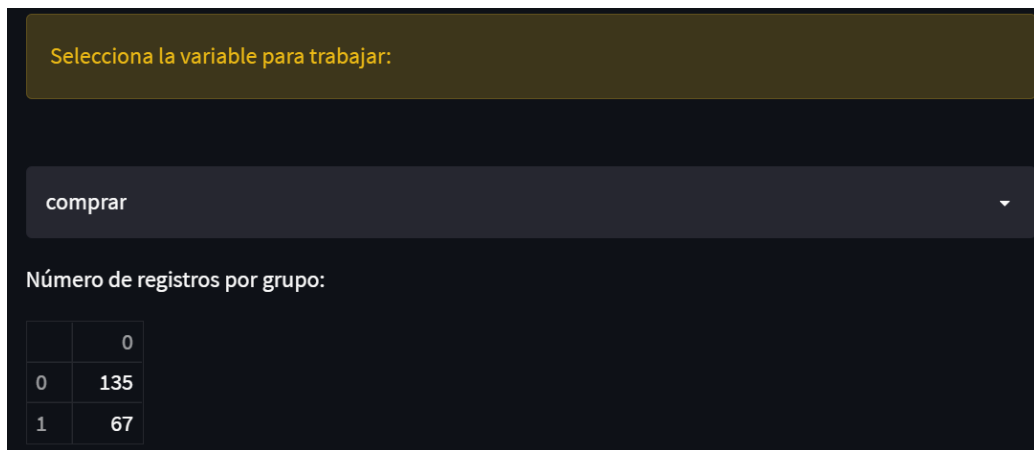
Matriz de Distancias

	0	1	2	3	4	
0	0.0000	236,994.7020	78,577.8403	260,974.5914	51,769.5814	39,
1	236,994.7020	0.0000	315,439.1768	26,550.5278	287,970.8078	276,
2	78,577.8403	315,439.1768	0.0000	339,168.0301	31,494.8080	39,
3	260,974.5914	26,550.5278	339,168.0301	0.0000	312,273.3115	300,
4	51,769.5814	287,970.8078	31,494.8080	312,273.3115	0.0000	15,
5	39,149.0605	276,141.6224	39,645.7607	300,095.4942	15,176.4768	
6	30,003.7979	207,115.4048	108,564.1283	231,251.1315	81,054.3929	69,
7	206,425.7062	33,742.4724	284,512.0079	54,727.3556	257,851.7949	245,
8	108,991.9407	345,963.7744	31,548.7590	369,945.8153	58,617.0264	69,
9	76.488.5430	312.810.3798	17.030.1947	337.121.5764	24.868.5397	38,

Figura 13: Matriz de distancias.

6.3. Clustering Particional

Una vez cargados los datos en la aplicación el usuario podrá elegir la variable de su interés y obtendrá el número de registros por grupo según su variable.



Selecciona la variable para trabajar:

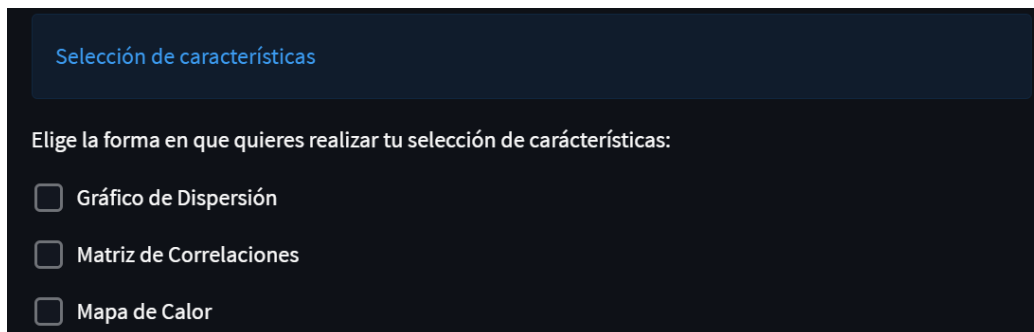
comprar

Número de registros por grupo:

	0
0	135
1	67

Figura 14: Seleccionar la variable para trabajar (C. Jerárquico).

Después tiene la libertad de elegir la forma en que quiere realizar la selección de sus características.



Selección de características

Elige la forma en que quieres realizar tu selección de características:

☐ Gráfico de Dispersión

☐ Matriz de Correlaciones

☐ Mapa de Calor

Figura 15: Selección de características.

Hecho su análisis podrá ingresar las características a usar en su modelo.

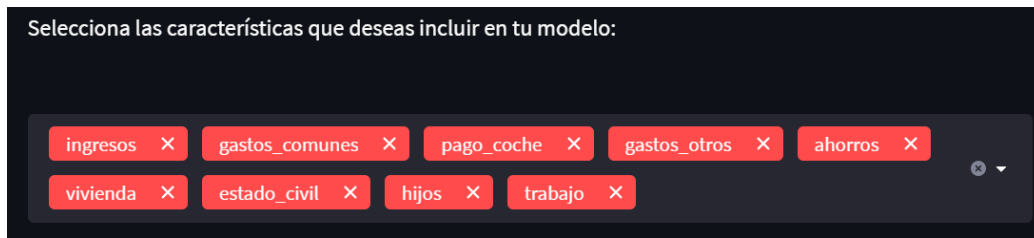


Figura 16: Ingresar las variables seleccionadas (C. Jerárquico).

Automáticamente se desplegarán las gráficas correspondientes al “Elbow method” y “Knee Point”, como el número de clústeres.

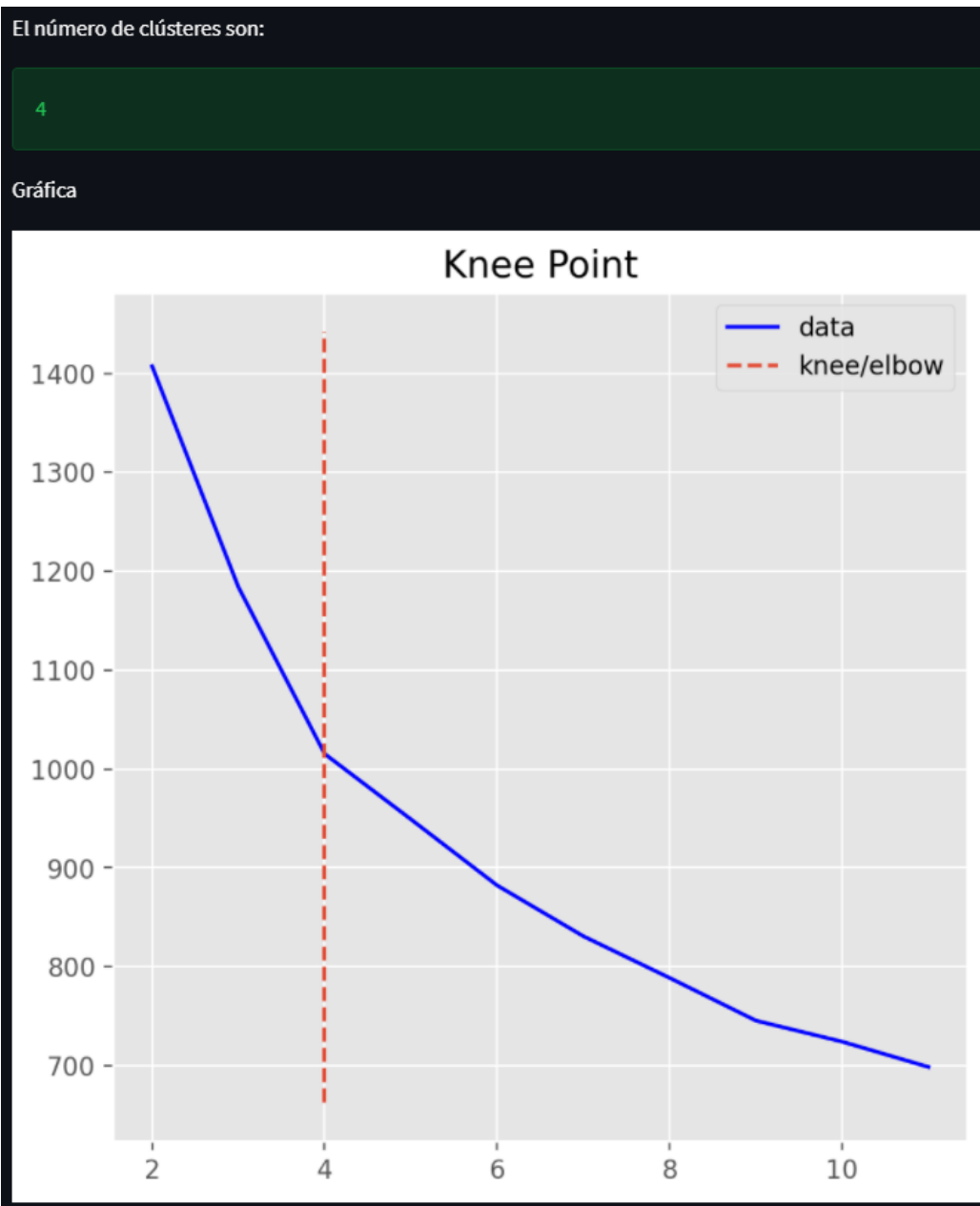


Figura 17: Gráfica Knee Point.

Finalmente se generarán otras tablas para analizar los resultados y la información de cada uno de sus clústeres.

Información de cada uno de tus clústeres						
	ingresos	gastos_comunes	pago_coche	gastos_otros	ahorros	vivier
0	3,502.9302	857.2093	245.7907	533.6279	24,129.1395	291,900.9€
1	3,472.4821	905.6071	224.7321	536.5893	23,957.6429	272,010.5€
2	6,389.6852	998.8519	190.2037	524.1481	54,899.7222	430,860.0€
3	6,358.9592	1,117.3061	190.7551	465.6531	50,687.0816	497,262.2€

Figura 18: Información de clústeres (C. Particional).

6.4. Clustering Jerárquico

Una vez cargados los datos en la aplicación el usuario podrá elegir la variable de su interés y obtendrá el número de registros por grupo según su variable.

Después tiene la libertad de elegir la forma en que quiere realizar la selección de sus características.

Hecho su análisis podrá ingresar las características a usar en su modelo.

Automáticamente se desplegarán el árbol donde los diferentes colores indica el número de clústeres formados.

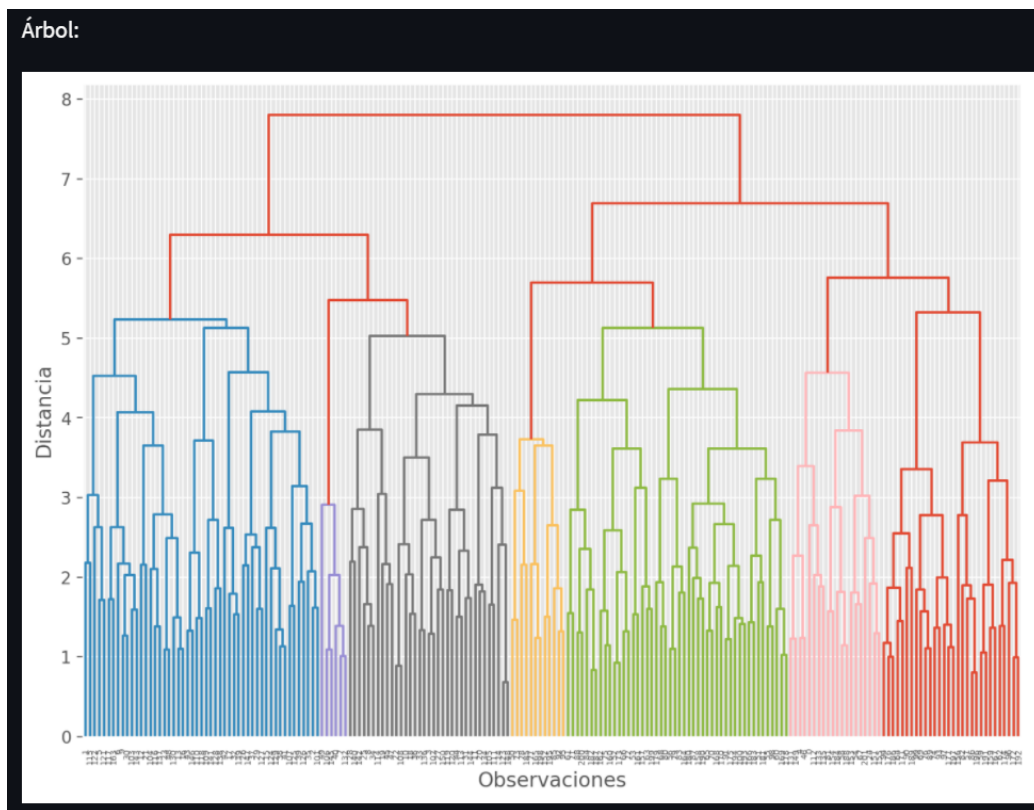


Figura 19: Árbol Clustering Jerárquico.

El usuario podrá ingresar el número de grupos para generar los últimos reportes.

Grupos

7.00 - +

Figura 20: Ingreso de clústeres.

Finalmente se generarán otras tablas para analizar los resultados y la información de cada uno de sus clústeres.

Información de cada uno de tus clústeres						
	ingresos	gastos_comunes	pago_coche	gastos_otros	ahorros	vivier
0	3,421.1333	846.4667	309.9333	527.2333	24,289.6333	295,590.70
1	6,394.0196	1,021.6275	192.2745	533.0392	54,382.5294	421,178.76
2	6,599.5429	1,087.4286	204.7714	362.6000	51,863.0286	515,494.25
3	3,189.6875	785.0208	243.2083	548.2708	23,616.8542	277,066.66
4	4,843.7500	1,009.2000	122.2000	572.8500	36,340.6500	337,164.85
5	4,466.4167	1,315.0833	114.4167	502.7500	23,276.1667	269,429.91
6	6,404.5000	1,176.1667	168.3333	769.3333	61,715.5000	625,138.83

Figura 21: Información de clústeres (C. Jerárquico).

6.5. Regresión Logística

Una vez cargados los datos en la aplicación el usuario podrá elegir la variable de su interés (variable clase) y obtendrá el número de registros por grupo según su variable.

Después tiene la libertad de elegir la forma en que quiere realizar la selección de sus características.

Tendrá la opción de seleccionar el tipo de de datos clasificar.

Por favor indica el tipo de dato de tu variable clase

☒ Categóricas

Figura 22: Elección del tipo de dato.

Hecho lo anterior podrá ingresar las características (variables predictoras) a usar en su modelo.

El usuario tiene la oportunidad de elegir el tamaño de los datos de prueba en un rango de 10 % - 90 %.

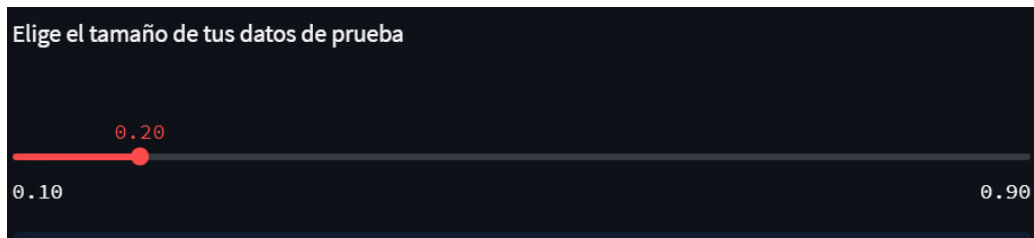


Figura 23: Ingresar tamaño de prueba de los datos).

Se generarán automáticamente otras tablas para analizar los resultados y la validación de su modelo.

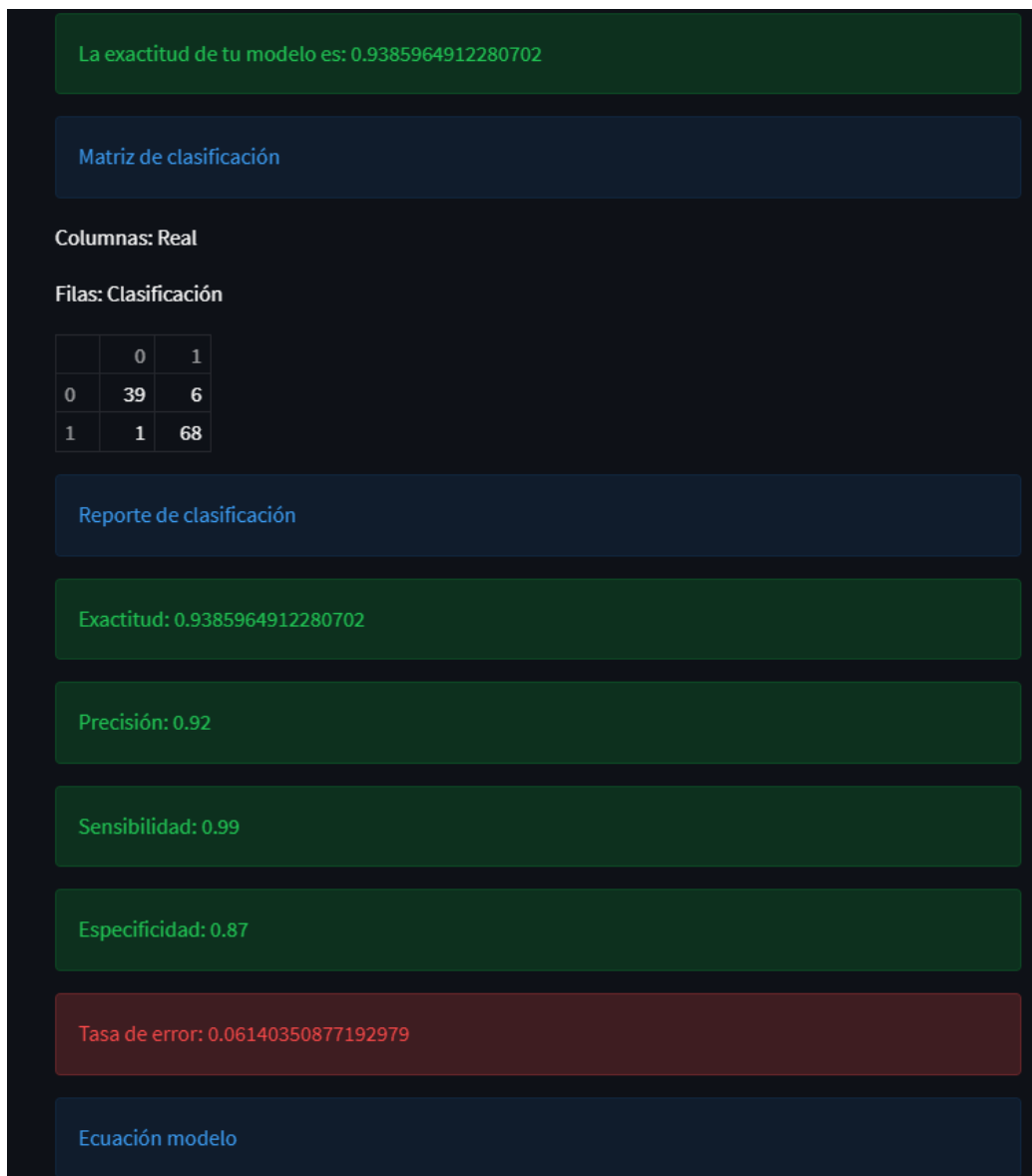


Figura 24: Resultados del algoritmo de regresión logística.

Al final se tiene una ventana en la cual puede realizar una nueva clasificación.

Prueba tu modelo

Nuevos pronósticos

Texture

10.38

-

+

Area

1001.00

-

+

Smoothness

0.11

-

+

Compactness

0.27

-

+

Symmetry

0.24

-

+

FractalDimension

0.07

-

+

El pronóstico fue: [0]

Interpretación de tus resultados:

0:M1:B

Figura 25: Prueba del algoritmo de regresión logística.

6.6. Árboles Aleatorios - Pronóstico

Una vez cargados los datos en la aplicación el usuario tendrá la posibilidad de ingresar dos variable para observar en una gráfica.



Figura 26: Gráfica de los datos ingresados (Á. Pronóstico).

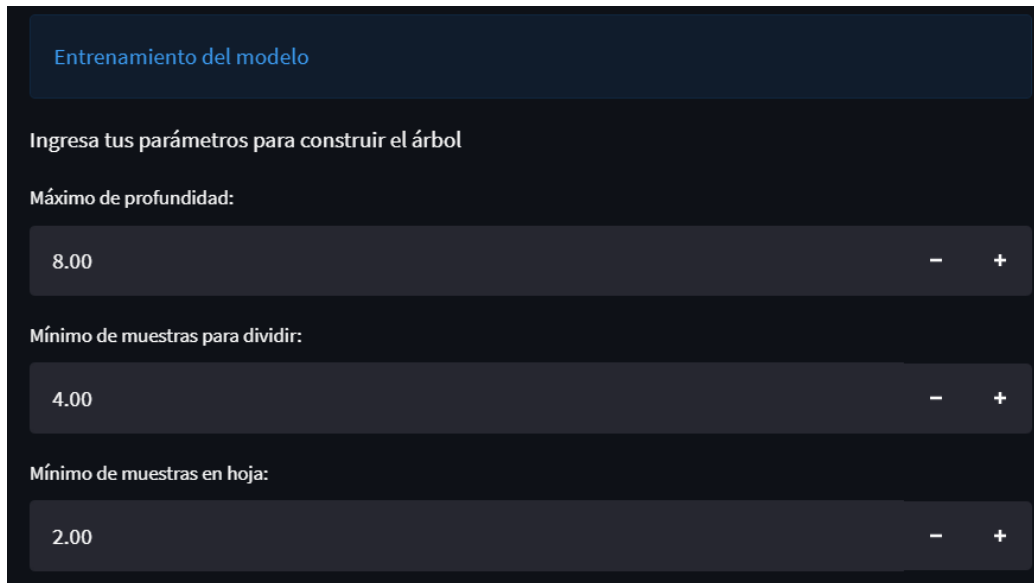
Luego de elegir la variable para trabajar (variable pronóstico), tiene la libertad de elegir la forma en que quiere realizar la

selección de sus características.

Hecho lo anterior podrá ingresar las características (variables predictoras) a usar en su modelo.

El usuario tiene la oportunidad de elegir el tamaño de los datos de prueba en un rango de 10 % - 90 %.

Para generar el árbol el usuario podrá ingresar los datos de: máximo de profundidad, mínimo de muestras para dividir y mínimo de muestras por hoja.



Entrenamiento del modelo

Ingresa tus parámetros para construir el árbol

Máximo de profundidad:

8.00 - +

Mínimo de muestras para dividir:

4.00 - +

Mínimo de muestras en hoja:

2.00 - +

Figura 27: Ingresar los parámetros del algoritmo (Á. Pronóstico).

Se generarán automáticamente otras tablas para analizar los resultados y la validación de su modelo.

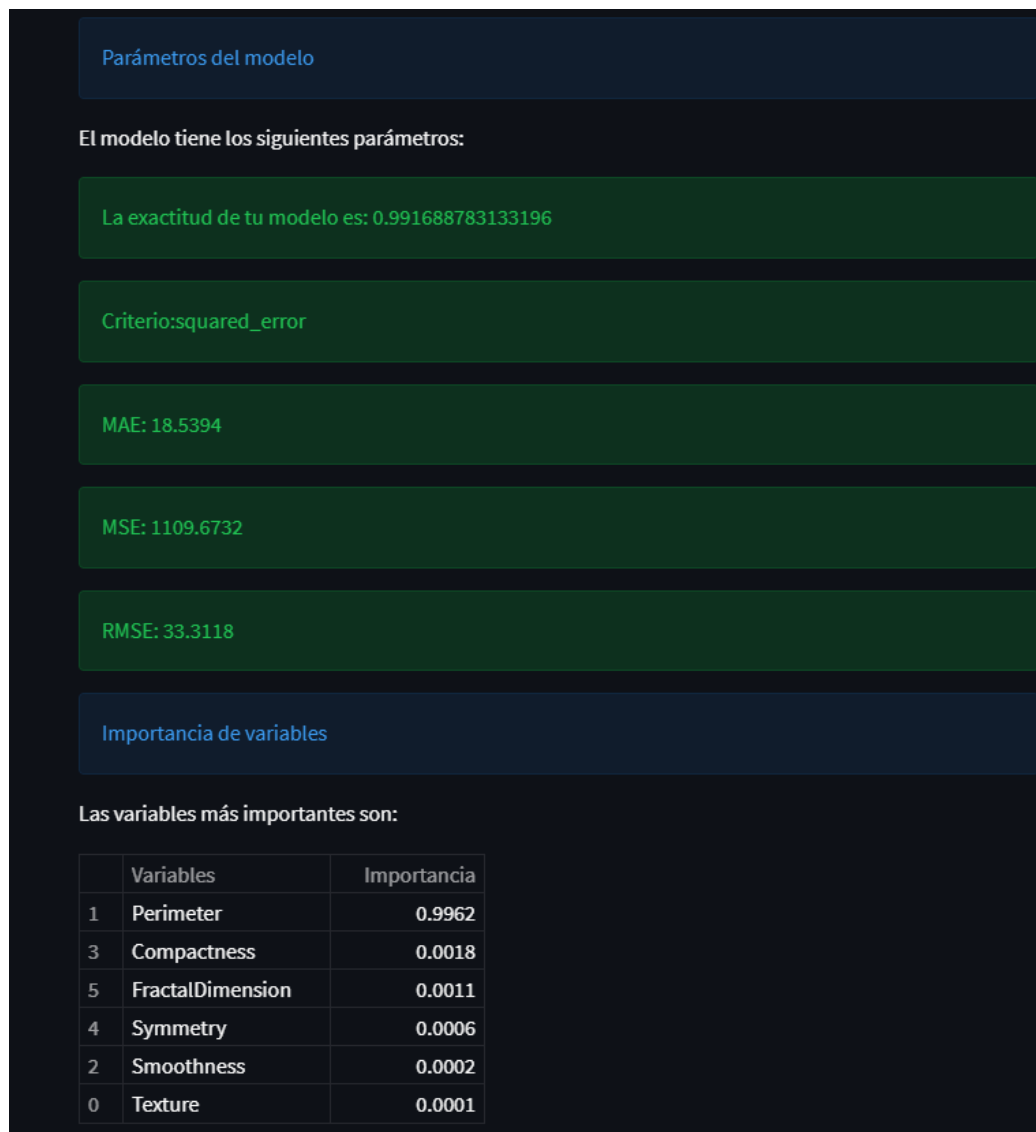


Figura 28: Resultados del algoritmo (Á. Pronóstico).

Si el usuario quiere ver el árbol formado podrá descargarlo en formato .SVG haciendo clic en el botón “Descarga tu Árbol de Decisión (Pronóstico)”.

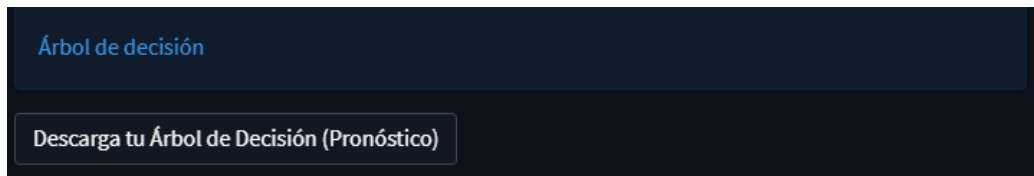


Figura 29: Descargar imagen del árbol (Á. Pronóstico).

También se tiene la opción de ver el reporte del árbol de decisión.



Figura 30: Reporte del árbol de decisión (Á. Pronóstico).

Al final se tiene una ventana en la cual puede realizar un nuevo pronóstico.

Prueba tu modelo

Nuevos pronósticos

Texture

10.38

-

+

Perimeter

122.80

-

+

Smoothness

0.11

-

+

Compactness

0.27

-

+

Symmetry

0.24

-

+

FractalDimension

0.07

-

+

El pronóstico fue: [991.83333333]

Figura 31: Prueba del algoritmo (Á. Pronóstico).

6.7. Árboles Aleatorios - Clasificación

Una vez cargados los datos en la aplicación el usuario podrá elegir la variable de su interés (variable clase) y obtendrá el

número de registros por grupo según su variable.

Después tiene la libertad de elegir la forma en que quiere realizar la selección de sus características.

Hecho lo anterior podrá ingresar las características (variables predictoras) a usar en su modelo.

El usuario tiene la oportunidad de elegir el tamaño de los datos de prueba en un rango de 10 % - 90 %.

Para generar el árbol el usuario podrá ingresar los datos de: máximo de profundidad, mínimo de muestras para dividir y mínimo de muestras por hoja.

Se generarán automáticamente otras tablas para analizar los resultados y la validación de su modelo.

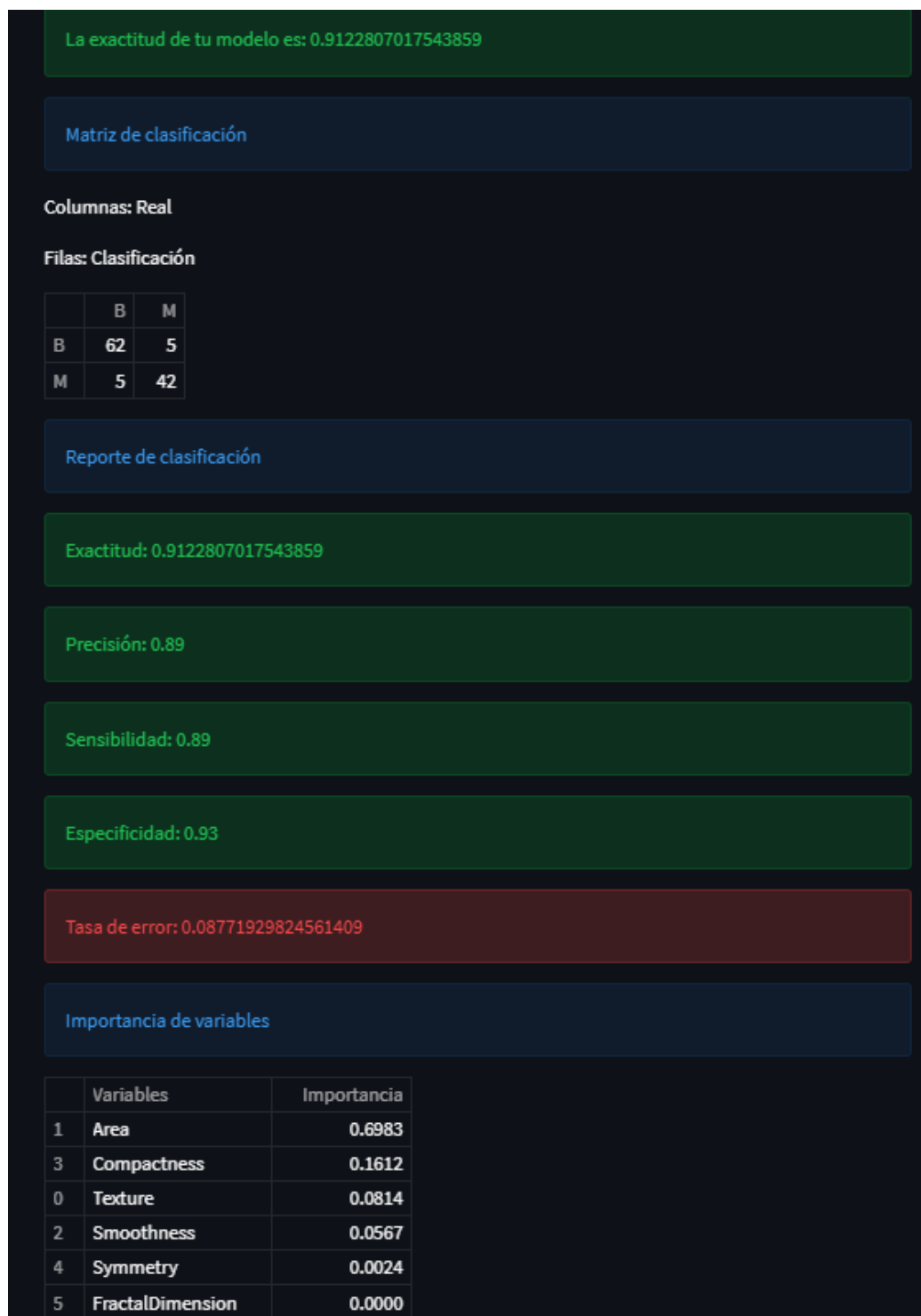


Figura 32: Resultados del algoritmo (Á. Clasificación).

Si el usuario quiere ver el árbol formado podrá descargarlo en el botón “Descarga tu Árbol de Decisión (Clasificación)”.

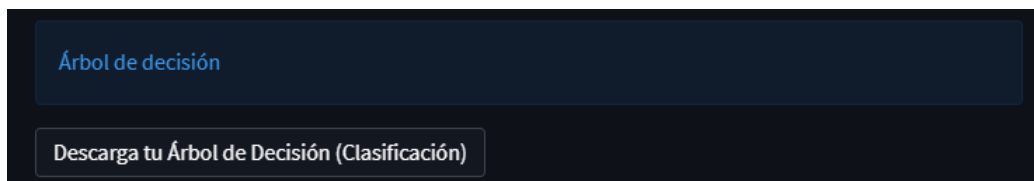


Figura 33: Descargar imagen del árbol (Clasificación).

También se tiene la opción de ver el reporte del árbol de decisión.

```
Haz click para ver el reporte de tu árbol de decisión

Reporte del Árbol de Decisión

|--- Area <= 694.15
| |--- Compactness <= 0.12
| | |--- Texture <= 19.61
| | | |--- Texture <= 18.12
| | | | |--- class: B
| | | |--- Texture > 18.12
| | | | |--- Texture <= 18.19
| | | | | |--- class: B
| | | | |--- Texture > 18.19
| | | | | |--- class: B
| | |--- Texture > 19.61
| | | |--- Area <= 562.55
| | | | |--- Area <= 501.35
| | | | | |--- class: B
| | | | |--- Area > 501.35
| | | | | |--- Smoothness <= 0.09
| | | | | | |--- class: B
| | | | | |--- Smoothness > 0.09
| | | | | | |--- Texture <= 22.14
| | | | | | | |--- class: B
| | | | | | |--- Texture > 22.14
| | | | | | | |--- class: B
| | |--- Area > 562.55
```

Figura 34: Reporte del árbol de decisión (Á. Clasificación).

Al final se tiene una ventana en la cual puede realizar una nueva clasificación.

Prueba tu modelo

Nuevos pronósticos

Texture

24.00
-
+

Area

181.00
-
+

Smoothness

0.05
-
+

Compactness

0.04
-
+

Symmetry

0.15
-
+

FractalDimension

0.05
-
+

El pronóstico fue: ['B']

Figura 35: Prueba del algoritmo (Á. Clasificación).

Referencias

Molero,G. (2021). *Presentaciones, de drive sitio web:*
<https://drive.google.com/drive/u/0/folders/1po22vmWeASavJzTrS7TtDzU6Ccc12myI>

,03 de Diciembre de 2021.

Streamlit. (s.f). *Streamlit*, de <https://streamlit.io> ,19 de Octubre de 2021.