Diego Ortiz Barbosa 201224923

Paula Ramirez Gonzalez 201328023

Inteligencia de negocios

Proyecto 1

1.Selección de los datos a trabajar justificado en términos del problema de negocio que va a resolver y la cantidad de datos disponible. La selección debe incluir por lo menos uno de los archivos disponibles en la página www.datos.gov.co y una fuente asociada al sector (e.g., International Civil Aviation Organization (ICAO) si se trata de un proyecto en transporte aéreo). Dependiendo del problema a resolver podrá complementar las fuentes con otros datos disponibles en páginas relacionadas con entidades regulatorias (e.g. aeronáutica civil –aerocivil.gov.co- si se trata de un proyecto en transporte aéreo) que supervisan la organización para la que está diseñando el proyecto.

Los datos seleccionados fueron los de comercio exterior de Colombia según región o país del mundo y los diferentes productos que se exportan a estas. Los datos seleccionados permiten saber desde el año 1991 hasta el 2015 el valor de las exportaciones en dólares. La cantidad de entradas que se encontraron fue de 874 con 44 atributos diferentes que permiten su clasificación.

El problema de negocio planteado es poder proveer al gobierno colombiano información útil acerca de los productos que se exportan con el objetivo de ayudar a la establecer en donde se deberían centrar los esfuerzos para mantener el crecimiento de la economía reforzando a los sectores más importante y ofreciendo tarifas diferenciales a los paises que mas importan ciertos productos. Con la información obtenida se pretende detectar tendencias y clasificar los productos por zonas para poder obtener la información requerida.

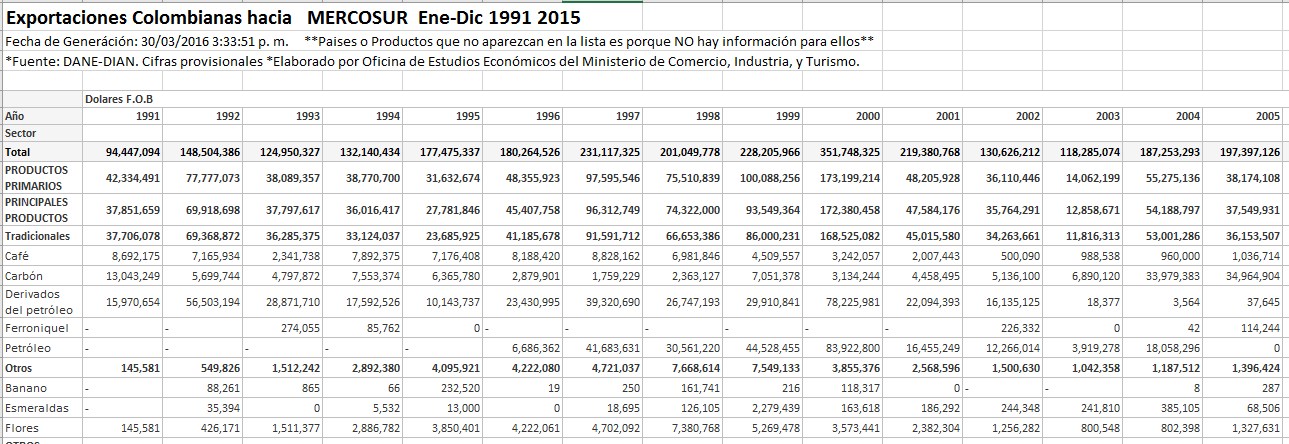
2. Descripción y preparación de los datos usados en los modelos. En este punto debe indicar los archivos utilizados, su estructura, volumen de datos. Adicionalmente, debe describir el proceso que siguió para definir el dataset que utilizará en el proyecto y su estructura.

Para la elaboración del dataset, se obtuvo información de diferentes archivos de zonas que tenían los valores en dólares de las exportaciones de los diferentes productos a dicha zona. Como se dijo anteriormente, la información de exportaciones se tiene de los años 1991 a 2015 y cada archivo con máximo los 44 atributos mencionados,, daba un volumen del dataset de 38379 datos. Desafortunadamente no todos los archivos contaban con todos los diferentes atributos por lo que mas que unir los archivos, se debió reorganizar los datos existentes para que cuadraran con la columna apropiada.

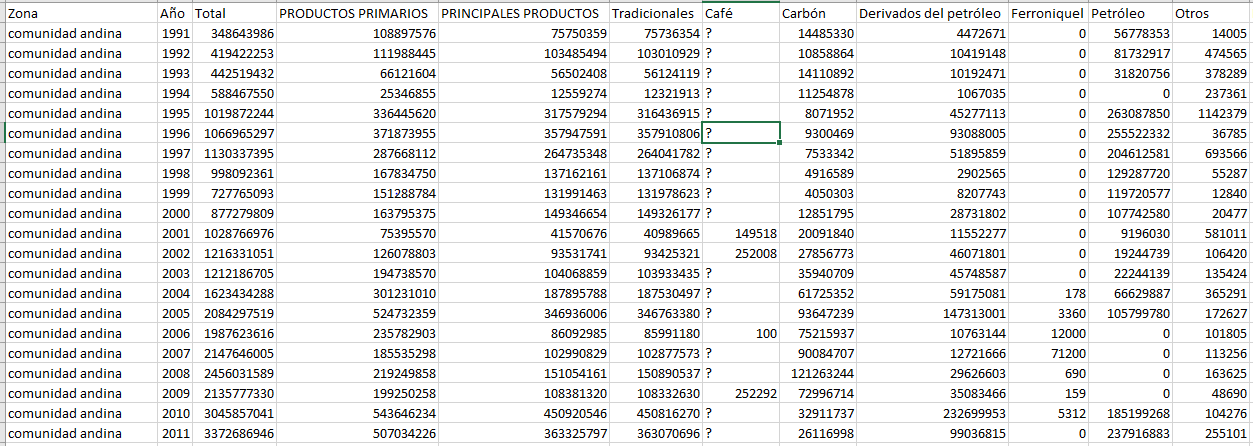
3. El resultado del análisis de calidad de los datos utilizados en los modelos propuestos, acompañado de un análisis del mismo.

Teniendo en cuenta que la calidad de los datos precisamente por la naturalidad del negocio, no se hacen exportaciones de todos los elementos a todas las zonas por lo que en estos casos se cambiaron los valores por “0”. Por ejemplo, las exportaciones de ferroniquel se dirigen a zonas como TLC o Asia pero no se dirigen a Argentina o Bolivia por lo que los registros de ferroníquel de esta zona sn 0

Presentación de los archivos al momento de descarga



Presentación de los datos después del procesamiento



4. Descripción de cómo el requerimiento de negocio es resuelto con el o los requerimientos de minería de datos propuestos, para lo cual debe utilizar la tabla que se presenta a continuación.

|  |  |  |
| --- | --- | --- |
| Oportunidad/problema de negocio | Se quiere proveer al gobierno la información de los productos de exportación que se hacen a las diferentes zonas para saber en que se deben enfocar con el fin de mantener una buena economía | |
| Descripción del requerimiento desde el punto de vista de minería de datos | Se busca conocer el comportamiento de cada zona conforme a sus exportaciones y clasificar las zonas en grupos para conocer su mercado y saber que productos es más viable exportar | |
| Detalles de la actividad de minería de datos | | |
| Tarea | Tecnica | Algoritmo y parámetros utilizados |
| * Predicción * Clasificación | * Arbol de desicion * Clustering | * J4.8 * KMeans, 5 Clusters |

5. El resultado de mínimo 2 modelos realizados, que de forma conjunta permitan resolver el requerimiento de negocio. Para cada modelo debe incluir la selección del algoritmo/método para definir los atributos que va a utilizar y el algoritmo a utilizar en el modelo. Si en los modelos propuestos tiene uno predictivo, debe incluir, la selección del modelo de clasificación, la selección del método de prueba del modelo y las medidas resultado del modelo. Adicionalmente, debe mostrar y explicar los parámetros de configuración más importantes, utilizados en los modelos.

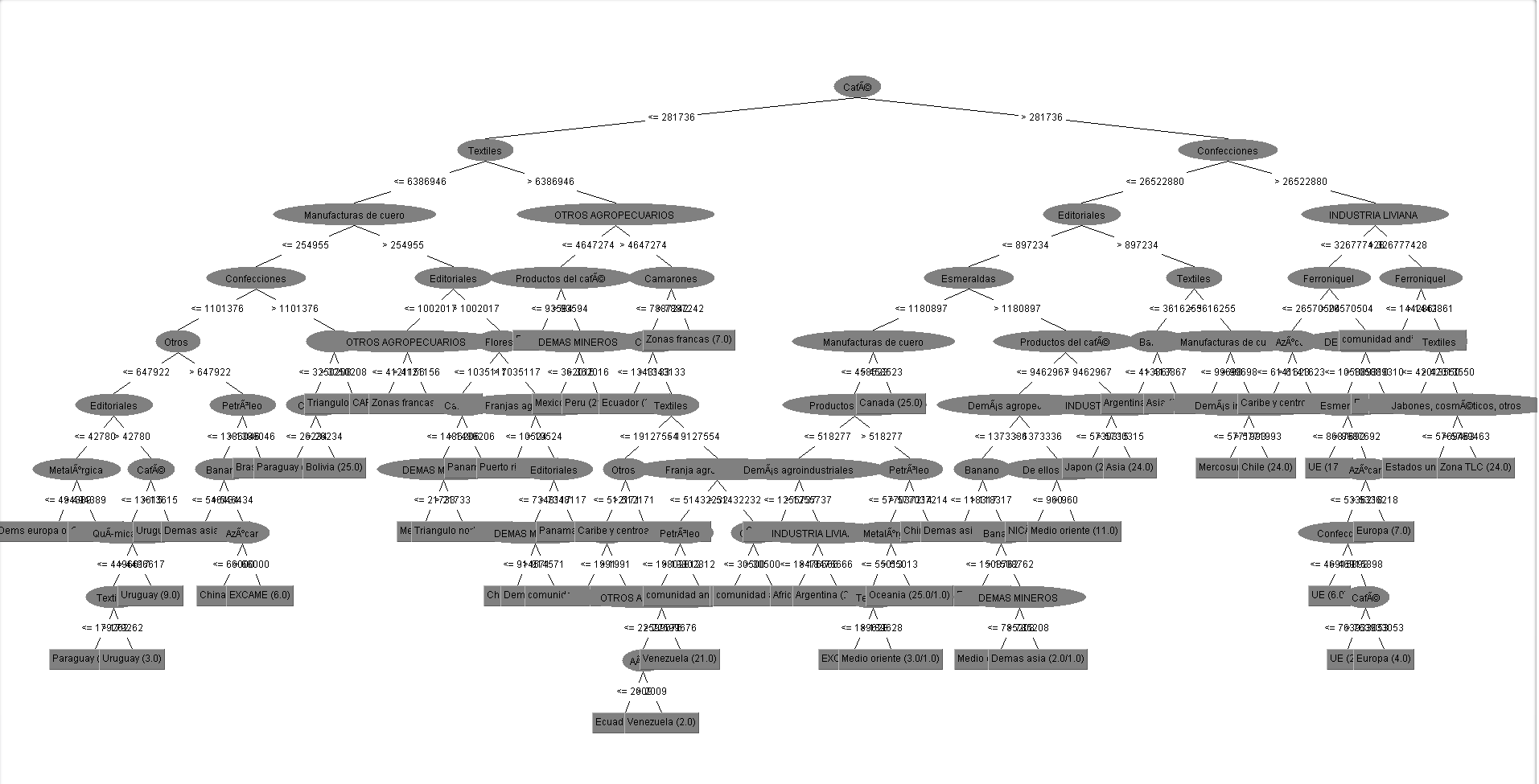
Variable objetivo del arbol de desiciones: Zona.

Se usó la variable zona como variable objetivo ya que lo que se pretende es observar el comportamiento de las diferentes zonas en respecto a su consumo de exportaciones colombianas.

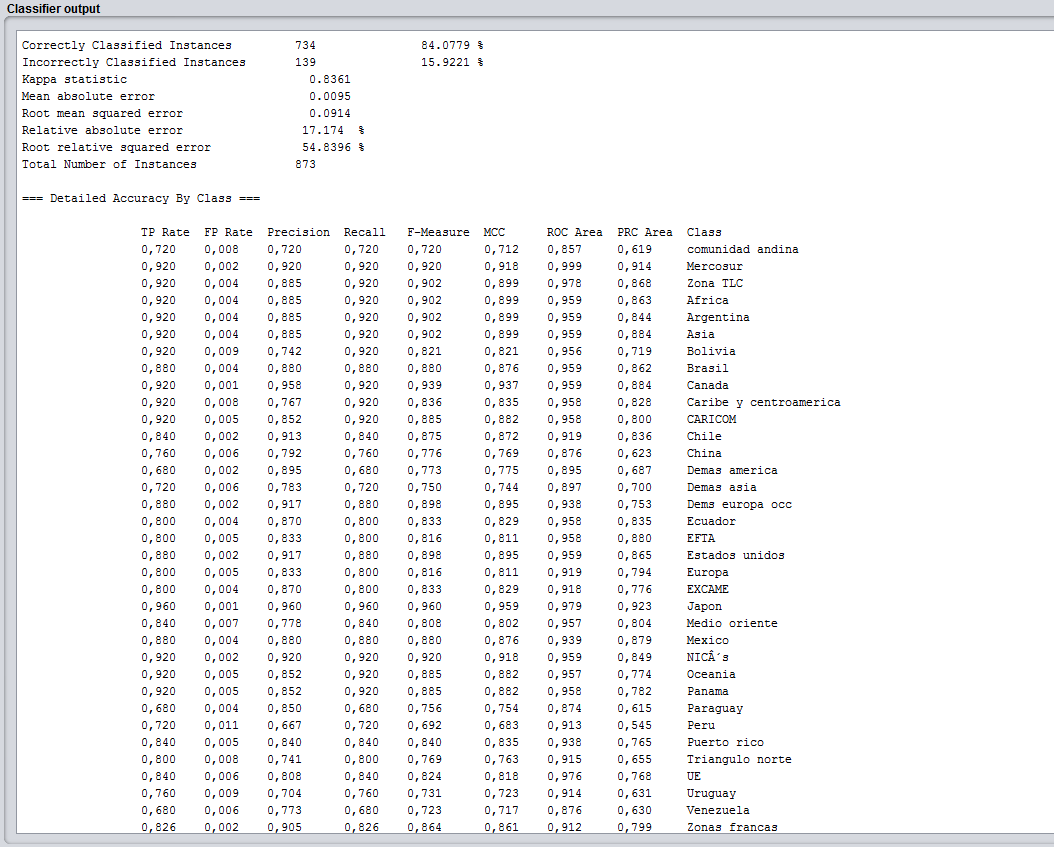
Se usaron 10 Cross-validations teniendo en cuenta que el dataset no supera las mil entradas este número es suficiente para entrenar al algoritmo de manera efectiva y como se verá en las proximas imagenes logra clasificar de manera efectiva el 86% de los datos.

Para el cluster se hizo una minería semidirigida teniendo en cuenta primordialmente las zonas al generar los clusters.

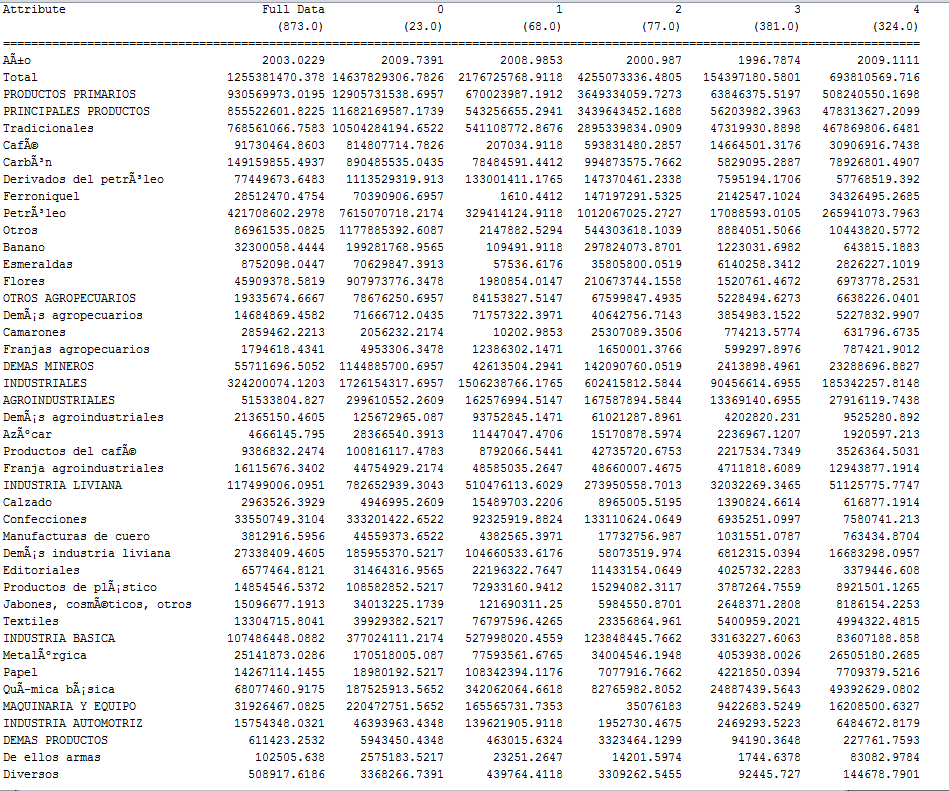
Árbol de decisión:

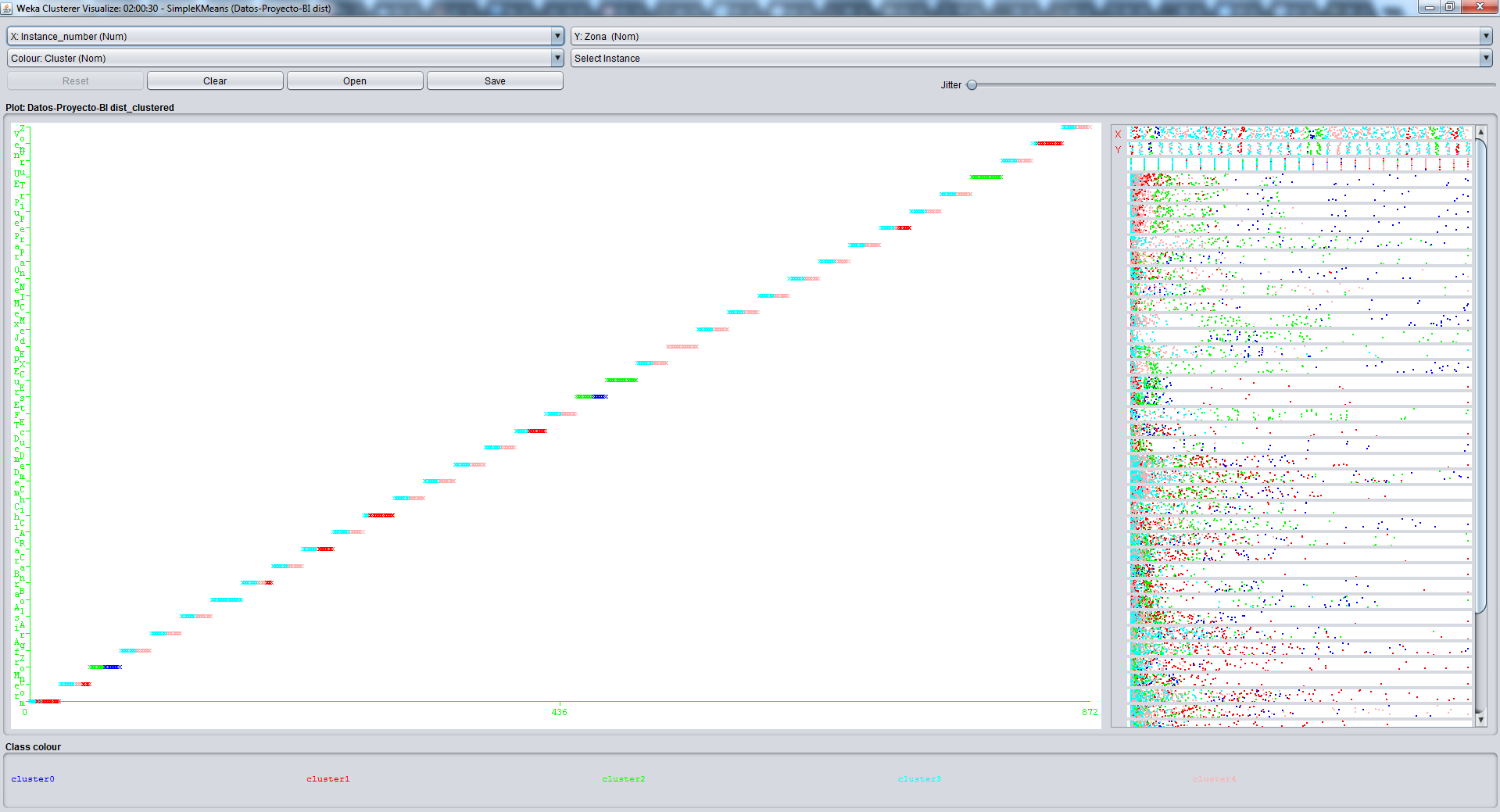


Matriz:



Cluster





6. Análisis del resultado de los modelos. En este punto se debe incluir una descripción de los resultados obtenidos que permita a la organización comprenderlos y un énfasis en el análisis de las medidas arrojadas por los modelos utilizados.

Del árbol que se obtuvo, podemos determinar cuales de los productos son exportados a las diferentes zonas generando una mayor ganancia para el país. Por ejemplo, se puede ver que a zonas como Venezuela se exportan productos de otros agropecuarios, a Europa se exporta café, a chile se le exporta manufacturas de cuero, entre otros.

El cluster se realizó para tener en cuenta las relaciones de productos exportados a la misma zona. Es decir se hacen grupos de productos que se exporten a una misma zona y que tengan una cierta relación de ganancias para el país con el fin de generar posibles proyectos de productos para el tratado de libre comercio con las diferentes zonas lo que mejoraría las relaciones internacionales del país.

7. Las estrategias que la organización debe plantear relacionadas con los resultados obtenidos en los modelos y una justificación de porqué esa información es útil para ellos.

La información anteriormente presentada le permitirá al gobierno nacional priorizar el comercio con los países o regiones en los cuales se genera más movimiento de capital y de exportaciones colombianas. Además ayudar a asignar subsidios para apoyar los diferentes sectores que tengan un mayor potencial a futuro siguiendo las tendencias, finalmente se puede ver un panorama más claro de los mercados internacionales para así poder dirigir las mercancías a los clientes más probables y aplicar incentivos recíprocos como tarifas diferenciales para generar más movimiento de capital.

Referencias:

* <https://www.nube-mcit.gov.co/public.php?service=files&t=e64a662ddfce73cffdea1b0e5e306f26&path=%2F>
* <https://www.datos.gov.co/browse?sortBy=newest&utf8=%E2%9C%93>