



**USO DE MÉTODOS MULTIVARIADOS A PRUEBAS HECHAS EN MANTOS DE
CARBON DE LA FORMACION GUADUAS - CUNDINAMARCA.**

Yully Katherine Araque Nuñez

**Fundación Universitaria Los Libertadores
Departamento de Ciencias Básicas
Especialización en estadística aplicada
Bogotá D.C.
2016**



**USO DE MÉTODOS MULTIVARIADOS A PRUEBAS HECHAS EN MANTOS DE
CARBON DE LA FORMACION GUADUAS - CUNDINAMARCA**

Yully Katherine Araque Nuñez

**Asesor:
Heivar Yesid Rodríguez Pinzón**

**Fundación Universitaria Los Libertadores
Departamento de Ciencias Básicas
Especialización en estadística aplicada
Bogotá D.C.
2016**

Nota de Aceptación

Firma del presidente del jurado

Firma del Jurado

Firma del Jurado

Bogotá, D.C Febrero del 2016

Las Directivas de la Universidad de
Los Libertadores, los jurados calificadores y el cuerpo
Docente no son responsables por los
criterios e ideas expuestas En el presente documento.
Estos corresponde únicamente a los autores

CONTENIDO

CAPITULO 1	10
1. INTRODUCCIÓN.	10
1.1 Planteamiento del problema.....	11
1.2 Objetivos del proyecto	11
1.2.1 Objetivo general.....	11
1.2.2 Objetivos específicos	11
1.3. Justificación	11
1.4 Ubicación de la zona de estudio.....	12
CAPITULO II	12
2. USO DE LA ESTADISTICA EN RESOLUCION DE PROBLEMAS EN GEOLOGIA DEL CARBON.....	12
2.1 Gas asociado al carbón.....	12
2.2 Estadística multivariada implementada en la solución de problemas de estudios descriptivos..	13
2.2.1 Análisis de clúster o análisis de conglomerados	13
2.2.1.1 Método de Ward	13
2.2.1.2 Dendograma	15
2.2.2 Análisis estadístico por Redes Neuronales	16
2.2.2.1 Rede neuronal	16
2.2.2.2 Estructura de red neuronal	17
2.2.2.3 Arquitectura Freedforward con una capa oculta	17
2.2.2.4 Perceptron multicapa	18
CAPITULO III.....	19
3. METODOLOGÍA.	19
3.1 Tipo de estudio.....	19
3.2 Variables	19
3.3 Método estadístico	19
3.3.1 Implementación de análisis de clúster	19
3.3.2 Implementación de Redes Neuronales.....	21
3.4 Unidad de análisis	21

4. PROCEDIMIENTO	21
4.1 Tratamiento de los datos	21
4.2 Aplicación de cluster y redes neuronales	22
4.2.1 Aplicación Cluster	22
4.2.1.1 Resultados de las corridas de Clúster para análisis de roca	22
4.2.1.2 Resultados de las corridas de Clúster para análisis de Agua	27
4.2.1.3 Resultados de las corridas de Clúster para análisis de gas.....	34
4.2.2 Aplicación de redes neuronales.....	40
Tabla 36.Variable dependiente: Clasificación	44
5. ANALISIS DE RESULTADOS	44
5.1 Análisis de resultados de corrida de clúster para Roca	44
5.2 Análisis de resultados de corrida de clúster para Agua.....	45
5.3 Análisis de resultados de corrida de clúster para Gas	45
5.4 Análisis de resultados de corrida de la Red Perceptrón Multicapa	46
CAPITULO IV	48
6. CONCLUSIONES	48
7. BIBLIOGRAFÍA	1

LISTA DE TABLAS

Tabla 1. Variables usadas en el estudio	14
Tabla 2. Corrida de cluster para roca	20
Tabla 3. Corrida de cluster para agua	20
Tabla 4. Corrida de cluster para gas.....	20
Tabla 5. Datos de roca	22
Tabla 6. Resultados estadísticos descriptivos para roca.....	22
Tabla 7. Matriz de proximidad para roca.....	23
Tabla 8. Estadísticos de los nodos para roca.....	23
Tabla 9. Centroides de las clases para roca.....	25
Tabla 10. Distancia entre los centroides de las clases para roca	25
Tabla 11. Objetos centrales para roca	25
Tabla 12. Distancia entre los objetos centrales para roca	26
Tabla 13. Resultados por clase para roca.....	26
Tabla 14. Resultados por objeto para roca.....	26
Tabla 15. Estadísticos descriptivos para agua.....	28
Tabla 16. Matriz de proximidad para agua	28
Tabla 17. Continuidad de la matriz de proximidad para agua	29
Tabla 18. Estadísticos de los nodos para agua	29
Tabla 19. Centroides de las clases para agua	31
Tabla 20. Centroides de las clases para agua	31
Tabla 21. Objetos centrales para agua	32
Tabla 22. Distancia entre los objetos centrales para agua.....	32
Tabla 23. Resultados por clase para agua	32
Tabla 24. Resultados por objeto para agua	33
Tabla 25. Estadísticos descriptivos para gas	34
Tabla 26. Matriz de proximidad para gas	34
Tabla 27. Estadísticos de los nodos para gas	35
Tabla 28. Centroides de las clases para gas	37
Tabla 29. Distancia entre los centroides de las clases para gas	37
Tabla 30. Objetos centrales para gas.....	37
Tabla 31. Distancia entre los objetos centrales para gas.....	38
Tabla 32. Resultados por clase para gas	38
Tabla 33. Resultados por objeto para gas	39
Tabla 34. Información sobre la red(a. Sin incluir la unidad de sesgo).....	41
Tabla 35. Resumen del modelo (variable dependiente: clasificación).....	43
Tabla 36. Variable dependiente: Clasificación	44

LISTA DE FIGURAS

Figura 1.Ubicación de la zona de estudio	12
Figura 2.Red Neuronal.....	17
Figura 3.Grafico de barras de niveles para roca.....	23
Figura 4.Dendograma para roca.....	24
Figura 5.Dendograma para roca.....	25
Figura 6.Perfil de clases para roca	27
Figura 7. Gráfico de barras de los niveles para agua	30
Figura 8.Dendograma para agua	30
Figura 9.Dendograma para agua	31
Figura 10.Perfil de clases para agua.....	33
Figura 11.Gráfico de barras de los niveles para gas.....	35
Figura 12.Dendograma para gas	36
Figura 13.Dendograma para gas	37
Figura 14.Perfil de las clases para gas	39
Figura 15.Arquitectura Freedforward	42

Resumen

En Colombia la industria petrolera por medio de los centros de investigación de las universidades, llevan a cabo investigaciones sobre temas de actualidad como lo es la respuesta a problemas de exploración de hidrocarburo, es el caso de los hidrocarburos no convencionales como lo es el metano asociado a mantos de carbón.

El origen de este gas preocupa a la comunidad petrolera, ya que económicamente es costosa su extracción y es necesario definir su origen para generar estrategias que faciliten su obtención. En el presente trabajo se analiza un grupo de variables que intervienen directamente en la definición de origen del gas metano asociado a los mantos de carbón de la formación Guaduas en el sinclinal Checua Lenguaza que en Cundinamarca, un área exclusiva en reservas de este hidrocarburo.

Mediante métodos multivariados, específicamente el análisis clúster y redes neuronales se evaluará el comportamiento de un grupo de variables pertenecientes a varios análisis hechos en roca, gas y agua asociados a estos mantos de cuatro zonas de explotación de carbón en el sinclinal.

El análisis hecho contribuye a orientar sobre el comportamiento de las variables cuando se tiene gas metano asociado a mantos de carbón, aporte que orienta el área de exploración del hidrocarburo en la industria.

CAPITULO 1

1. INTRODUCCIÓN.

Una de las mayores preocupaciones en la industria del petróleo a nivel mundial, es la disminución de reservas de hidrocarburos convencionales; por este motivo el conocimiento y la tecnología se han puesto a disposición a investigar sobre nuevas alternativas de energía, es por esto que la primer idea de contribuir a reservas de hidrocarburo es la utilización de hidrocarburos no convencionales, los cuales son los más difíciles de obtener ya que por su condición es complicada la obtención y de un costo mucho más elevado, en este caso hablaremos de gas contenido en roca fuente como lo es el carbón.

En Colombia la industria energética, se ha estado preocupando por encontrar alternativas para aumentar sus reservas de hidrocarburos, con el apoyo de las instituciones de investigaciones de las universidades, se ha venido implementando estrategias para atender esta emergencia.

La Universidad industrial de Santander es medio del desarrollo de estas estrategias con el centro de investigación de hidrocarburos y carbones donde se realizan investigaciones sobre hidrocarburos no convencionales como lo es el gas biogenico y termogenico generado en los carbones del sinclinal Chegua-Lenguazaque de la cuenca de Bogotá,; investigaciones pioneras en este tema tan importante.

Los datos utilizados en el desarrollo de esta tesis son facilitados por este centro de investigación, con el principal interés de contribuir a definir el origen del gas generado en los mantos de carbón de esta cuenca mediante el uso de técnicas de estadística aplicada como lo es el análisis Multivariado de las variables como isotopos de carbono, deuterio, calidad del agua, calidad del gas asociados a estos mantos, estado de madurez de la roca, contenido de gas, entre otras.

La implementación de métodos multivariados generara respuesta al problema presentado en discusión, ya que nos permite medir, explicar y predecir el grado de relación que existe entre la variación (combinación lineal ponderada de las variables), siendo esto suficiente para definir el tipo de generación de gas en los mantos de carbón.

1.1 Planteamiento del problema

¿Cómo se comportan las variables (análisis de muestras de carbón, gases y aguas) que definen el tipo de generación de gas metano en la formación Guaduas en el sinclinal Checua-Lenguazaque-Cundinamarca?

1.2 Objetivos del proyecto

1.2.1 Objetivo general

Estudiar por métodos multivariados, específicamente análisis clúster y redes neuronales el comportamiento de las variables que intervienen en los procesos asociados a la generación de gas metano biogenico o termogenico en mantos de carbón de la Formación Guaduas.

1.2.2 Objetivos específicos

- Implementar el método de Ward en el análisis clúster
- Aplicar el análisis nodal en el uso de clúster con las variables de gas, agua y roca
- Medir mediante la distancia euclidea cada observación entre muestras
- Definir los clúster mediante el uso de dendograma
- Caracterizar cada clúster encontrado con sus observaciones
- Presentar una introducción a la aplicación de redes neuronales para clasificar las muestras en estudio.

1.3. Justificación

El conocimiento que se tiene de las cuencas colombianas asociadas a la generación de gas producido por mantos de carbón, no es extenso; por consiguiente el presente estudio, genera información muy valiosa como origen, calidad, presencia del hidrocarburo factores claves para la exploración del mismo en el territorio Colombiano.

En la cuenca de Bogotá se presentan manifestaciones de gas metano en la mayoría de minas que explotan carbón en Cundinamarca, el siguiente proyecto se desarrolla con un grupo de análisis que se realizaron en minas del área carbonífera Checua –Lenguazaque.

Se tienen veintiocho variables, las cuales por ser un número considerable se hace necesario usar métodos estadísticos multivariados que me ayuden a definir grupos o clúster que orienta al investigador a conocer cómo se comporta una muestra asociada a gas contenido en mantos de carbón y el uso de redes neuronales quienes ayuden a determinar la aparición del hidrocarburo en la roca.

1.4 Ubicación de la zona de estudio

El área de estudio se encuentra en la localidad de Sutatausa, en la plancha 209-I-D-2 en distribución de varias minas que explotan carbón en la zona.

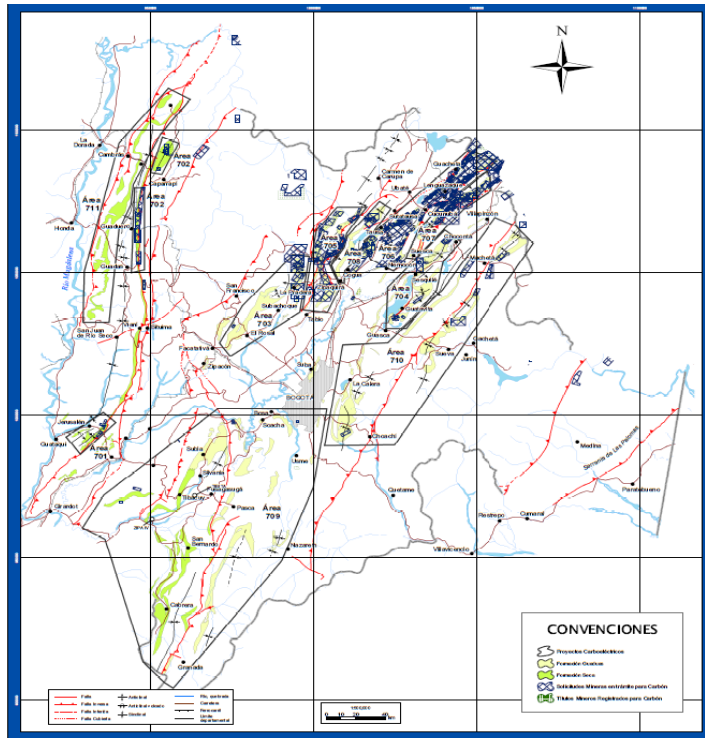


Figura 1. Ubicación de la zona de estudio

CAPITULO II

2. USO DE LA ESTADISTICA EN RESOLUCION DE PROBLEMAS EN GEOLOGIA DEL CARBON

2.1 Gas asociado al carbón

La generación de gas en mantos de carbón es resultado de distintos procesos de acumulación preservación y transformación de la materia orgánica de origen vegetal que da como resultado la formación de depósitos de turba los cuales son transformados a carbones por procesos que se denominan carbonificación llevados a cabo desde la diagénesis hasta el metamorfismo. Durante estos procesos se generan importantes cantidades de gas bacteriano en las etapas iniciales y posteriormente de gases termales por descomposición de la molécula de kerogeno en carbones. La Figura 5 ilustra la cantidad relativa de generación de gas termal durante el proceso de maduración térmica o carbonificación. El gas generado es de composición predominante de metano con cantidades menores de C₂-C₆, CO₂ y CO. Los gases generados se encuentran en tres formas: gas libre presente en poros, y fracturas del carbón, gas adsorbido que corresponde al gas adheridos a

superficies de fracturas o diaclasas del carbón y gas absorbido que se encuentra en espacios intramoleculares dentro del kerogeno.

2.2 Estadística multivariada implementada en la solución de problemas de estudios descriptivos

El análisis multivariante son todas aquellas técnicas estadísticas que simultáneamente analizan múltiples resultados en los individuos u objetos bajo investigación; y su función es representar datos de forma inteligible, esclarecer la distribución real de varias variables, desarrollar un modelo de predicción basado en múltiples variables y hallar las relaciones de causa y efecto entre variables.

De esta forma el análisis multivariante llega a medir, explicar y predecir el grado de relación que existe entre la variación (combinación lineal ponderada de las variables). El carácter multivariante del análisis descansa no solo en el número de variables sino en las múltiples combinaciones existentes entre las variables, la implementación de técnicas estadísticas que manejan un grupo grandes de variables, las cuales se quieren reducir es necesaria la implementación de análisis multivariados de componentes principales.

2.2.1 Análisis de clúster o análisis de conglomerados

Este análisis engloba un conjunto de técnicas multivariantes que tiene como objetivo, agrupar conjunto o casos en grupos, teniendo en cuenta el criterio de individuos parecidos y máxima heterogeneidad de individuos.

Este análisis maneja data cualitativa o cuantitativa y puede transformar las variables para el uso si no se encuentran en las mismas unidades, punto importante con el cual no se desecha información por no presentarse en las mismas características.

Luego de aplicar los algoritmos y agrupar el número de variables, se genera una nueva variable que describe la pertenencia del caso al grupo.

2.2.1.1 Método de Ward

A continuación se describe el uso del método paso a paso

Estadísticos descriptivos:

Nos permiten recolectar, ordenar, analizar y representar un conjunto de datos, con el fin de describir apropiadamente las características de este. Con este estudio se calcula un aserie de medidas de tendencia central para conocer en qué medida los datos se agrupan o dispersan en torno a un valor central.

En este trabajo se usa los parámetros: con Variable, Observaciones, Obs. con datos perdidos, Obs. Sin datos perdidos, mínimo, máximo, media, Desv. Típica.

- **Variable:** representa los análisis realizados a agua, gases y roca, a continuación se muestran en una tabla todas las variables.

Esp acum(m)
Prof. de ext. (m)
Ro
TOC
Volumen cm ³
%Metano
%Butano
%Hexano
Log 10/C1/C3+C6
δ D1/2
δ ¹³ C/2
TSD mg/L
Ph
SAL/ppl
CONDUCTI (μS/cm)
ALCALI (mmol/L)
DUREZA (meq/L) CaCO ₃
CO ₃ /meq/L
K/meq/L
NO ₃ /meq/L
Ca meq/L
Mg meq/L
Na meq/L
Cl meq/L
SO ₄ meq/L
HCO ₃ meq/L

Tabla 1. Variables usadas en el estudio

- **Observaciones:** representa la medida de veces que fue tomado un análisis
- **Observaciones con datos perdidos:** valores perdidos en los datos
- **Observaciones sin datos perdidos:** datos completos para las variables
- **Mínimo:** especifica el menor valor de la variable en el análisis
- **Máximo:** especifica el valor máximo de la variable en el análisis.
- **Media:** La media aritmética de una variable estadística es la suma de todos sus posibles valores, ponderada por las frecuencias de los mismos.
- **Desviación típica:** es una medida de dispersión usada en estadística que nos dice cuánto tienden a alejarse los valores concretos del promedio en una distribución de datos.

Medidas de similitud: Depende de las escalas de medida, se pueden agrupar observaciones según la similitud expresada en términos de distancia .un ejemplo de distancia entre objetos es la distancia euclidea.

Distancia euclidea:

Según los objetos I1 y I2 medidos según dos variable X1 y X2, la distancia euclidea entre ambos es

$$d_{I_1 I_2} = \sqrt{(x_{11} - x_{21})^2 + (x_{12} - x_{22})^2}.$$

Con más dimensiones(o variables que se miden) es equivalente a :

$$d_{I_1 I_2} = \sqrt{\sum_{k=1}^p (x_{1k} - x_{2k})^2}$$

En notación vectorial se expresa como

$$d_{I_i I_j}^2 = (x_i - x_j)'(x_i - x_j).$$

Si se consideran n objetos para i, j E {1,...,n},la distancia total es

$$d = \sum_{i=1}^n \sum_{j=1}^n \left(\sum_{k=1}^p (x_{ik} - x_{jk})^2 \right)^{1/2} .$$

Estadísticos de los nodos:

Un árbol binario es una estructura de datos en la cual cada nodo puede tener un hijo izquierdo y un hijo derecho, un ejemplo común de uso son los arboles binarios de búsqueda, donde se compara los elementos o claves de los mismos, con el fin de encontrar el dato de interés.

Gráficos de barras de los niveles:

Representación gráfica donde se debe mostrar el nivel donde se encuentra cada nodo.

2.2.1.2 Dendograma

Es un gráfico usado en el procedimiento jerárquico que permite visualizar el proceso de agrupamiento de los clúster en los distintos pasos, formando un diagrama de árbol. Esta herramienta ayuda a decidir el número de grupos que podrían representar mejor la estructura del dato teniendo en cuenta la forma en que se van anidándolos clúster y la medida de similitud a la cual lo hacen.

Descomposición de la variación para la clasificación óptima:

Se muestra el valor absoluto y el porcentaje de la intraclase y las interclases y el total de las mismas

Centroide de las clases:

Tabla donde se define las clases y el peso de cada muestra

Distancia entre los centroides de las clases:

Tabla donde se muestra las distancias tomadas a cada clase

Objetos centrales:

Tabla donde se muestra las clases y el peso para cada muestra

Distancia entre los objetos centrales:

Tabla donde se muestra las distancias entre cada clase

Resultados por clase:

Tabla que muestra los objetos de cada clase, la suma de los pesos de las clases, la varianza intraclass, la distancia mínima al centroide, la distancia media al centroide y la distancia máxima al centroide.

Resultados por objeto:

Tabla donde se presenta la observación y a que clase pertenece.

Perfiles de las clases:

Grafico donde se muestra los niveles punto a punto por muestra Vs los pesos

2.2.2 Análisis estadístico por Redes Neuronales

Las redes neuronales utilizadas en las aplicaciones predictivas, como las redes de perceptrones multicapa (MLP) y las de función de base radial (RBF), se supervisan en el sentido de que los resultados pronosticados por el modelo se pueden comparar con los valores conocidos de las variables de destino. La opción de redes neuronales de le permite ajustar las redes MLP y RBF y guardar los modelos resultantes para la puntuación.

2.2.2.1 Rede neuronal

Una red neuronal es un procesador distribuido en paralelo de forma masiva con una propensión natural a almacenar conocimiento experimental y convertirlo en disponible para su uso. A semeja al cerebro en dos aspectos (Haykin, 1998):

- El conocimiento se adquiere por la red mediante un proceso de aprendizaje.
- Las fuerzas de conexión interneuronal, conocidas como ponderaciones sinápticas, se utilizan para almacenar el conocimiento.

Una red neuronal puede aproximar una amplia gama de modelos estadísticos sin que tenga que hipotetizar de antemano determinadas relaciones entre las variables dependientes e independientes. En lugar de eso, la forma de las relaciones se determina durante el proceso de aprendizaje. Si una relación lineal entre las variables dependientes e independientes es adecuada, los resultados de la red neuronal deben aproximarse lo máximo posible a los del modelo de regresión lineal. Si una relación no lineal es más adecuada, la red neuronal se aproximará automáticamente a la estructura del modelo “correcto”

2.2.2.2 Estructura de red neuronal

La red de perceptrones multicapa (MLP) es una función de predictores (denominados también entradas o variables independientes) que minimiza el error de predicción de las variables de destino (también denominadas salidas).

Se usa un ejemplo para entender el uso de la red neuronal:

Se desea poder identificar a los posibles morosos entre un grupo de solicitantes de préstamos. Una red MLP aplicada a este problema es una función de las medidas que minimiza el error al pronosticar la posibilidad de mora.

2.2.2.3 Arquitectura Feedforward con una capa oculta

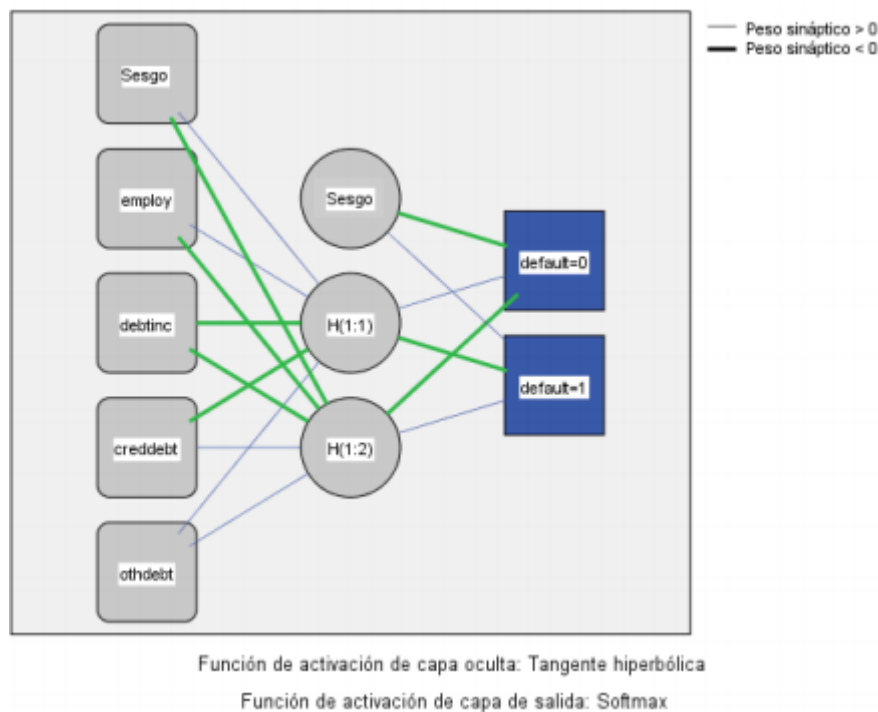


Figura 2.Red Neuronal

Esta estructura se denomina arquitectura feedforward porque las conexiones de la red fluyen unidireccionalmente desde la capa de entrada hasta la capa de salida sin ciclos de retroalimentación. En esta figura:

- La capa de entrada contiene los predictores.
- La capa oculta contiene nodos (o unidades) no observables. El valor de cada unidad oculta es una función de los predictores; la forma exacta de la función depende, por un lado, del tipo de red y, por otro lado, de especificaciones controlables por el usuario.
- La capa de salida contiene las respuestas. Como el historial de moras es una variable categórica con dos categorías, se recodifica como dos variables de indicador. Cada unidad de salida es una función de las entradas ocultas. Nuevamente, la forma exacta de la función depende, por un lado, del tipo de red y, por otro lado, de especificaciones controlables por el usuario

La red MLP permite una segunda capa oculta; en ese caso, cada unidad de la segunda capa oculta es una función de las unidades de la primera capa oculta, y cada respuesta es una función de las unidades de la segunda capa oculta.

2.2.2.4 Perceptron multicapa

El procedimiento Perceptrón multicapa (MLP) genera un modelo predictivo para una o más variables dependientes (de destino) basada en los valores de las variables predictoras.

La red MLP es un algoritmo de aprendizaje supervisado de tipo adaptativo que permite realizar actividades de clasificación y análisis multivariado, además es considerada como un modelo óptimo en tareas de predicción, debido a que puede resolver problemas con unidades de análisis complejas que no han podido ser trabajadas mediante métodos clásicos (Kuncheva 2004), (Jain et al. 2000).

- **Variables dependientes:** Nominales, ordinales o de escala.
- **Variables predictoras:** los predictores se pueden clasificar como factores (categóricos) o covariables (de escala).
- **Replica de los resultados:** si se desea replicar los resultados, utilice el mismo valor de inicialización para el generador de números aleatorios, el mismo orden de datos y el mismo orden de variables además de utilizar la misma configuración del procedimiento.

CAPITULO III

3. METODOLOGÍA.

3.1 Tipo de estudio

El presente estudio " USO DE MÉTODOS MULTIVARIADOS A PRUEBAS HECHAS EN MANTOS DE CARBON DE LA FORMACION GUADUAS - CUNDINAMARCA " corresponde a relación de variables de tipo descriptivo, dado que se presenta un numero de variables las cuales están íntimamente relacionadas con la obtención de la respuesta, y determinación de gas en los mantos de carbón.

3.2 Variables

- ✓ Resultados de análisis en laboratorio de desgasificación de carbones por medio de la técnica uso de canisters.
- ✓ Resultados de análisis a muestras de carbón en laboratorio: Reflectancia de la vitrinita, medición de total de materia orgánica.
- ✓ Resultados de análisis a muestras de gas en laboratorio: Cromatografía de gases, análisis de isotopos de δC_{13} y δD_1 en muestras de gas metano (CH_4).
- ✓ Resultados de análisis a muestras de agua in situ y en laboratorio: análisis químicos in situ, temperatura, pH, Eh conductividad, salinidad, alcalinidad, dureza, nitratos, carbonatos, potasio, sulfatos, calcio, magnesio, sodio, cloruros, bicarbonatos.

3.3 Método estadístico

Se utilizó el método estadístico de análisis multivariado de tipo clúster y redes neuronales.

3.3.1 Implementación de análisis de clúster

Debido a ciertas características se agruparan las variables y se definirán las de más peso o las que explican mucho mejor la determinación de gas metano en los mantos de carbón en el sinclinal Checua-Lenguazaque.

Para la implementación de Clúster, se decidió analizar las variables por grupos de pruebas las cuales son: agua, roca y gases. A continuación se presenta la tabla de método usado en la implementación de clúster para cada grupo de pruebas y análisis realizados.

Clúster para pruebas de roca

Tabla observaciones/variables: Libro = BASE_ARREGLADA2.xlsx / Hoja = Roca / Rango = Roca!\$B\$1:\$E\$11 / 10 filas y 4 columnas									
Etiquetas de las filas: Libro = BASE_ARREGLADA2.xlsx / Hoja = Agua / Rango = Agua!\$A\$1:\$A\$11 / 10 filas y 1 columna									
Agrupar columnas									
Disimilitud: Distancia euclídea									
Método de aglomeración: Método de Ward									
Centrar: Sí									
Reducir: Sí									
Truncamiento: Automático - Inercia									

Tabla 2. Corrida de cluster para roca

Clúster para pruebas de agua

Tabla observaciones/variables: Libro = BASE_ARREGLADA2.xlsx / Hoja = Agua / Rango = Agua!\$B\$1:\$R\$11 / 10 filas y 17 columnas																	
Etiquetas de las filas: Libro = BASE_ARREGLADA2.xlsx / Hoja = Agua / Rango = Agua!\$A\$1:\$A\$11 / 10 filas y 1 columna																	
Agrupar columnas																	
Disimilitud: Distancia euclídea																	
Método de aglomeración: Método de Ward																	
Centrar: Sí																	
Reducir: Sí																	
Truncamiento: Automático - Entropía																	

Tabla 3. Corrida de cluster para agua

Clúster para pruebas de gas

Tabla observaciones/variables: Libro = BASE_ARREGLADA2.xlsx / Hoja = Gas / Rango = Gas!\$B\$1:\$J\$11 / 10 filas y 9 columnas									
Etiquetas de las filas: Libro = BASE_ARREGLADA2.xlsx / Hoja = Agua / Rango = Agua!\$A\$1:\$A\$11 / 10 filas y 1 columna									
Agrupar columnas									
Disimilitud: Distancia euclídea									
Método de aglomeración: Método de Ward									
Centrar: Sí									
Reducir: Sí									
Truncamiento: Automático - Inercia									

Tabla 4. Corrida de cluster para gas

- Estadísticos descriptivos: con Variable, Observaciones, Obs. con datos perdidos, Obs. sin datos perdidos, mínimo, máximo, media, Desv. Típica
- Matriz de proximidad (Distancia euclídea): con las variables (Esp acum, prof de ext, Ro, TOC)
- Estadísticos de los nodos: con el nodo, nivel, peso, objetos, hijo izquierdo, hijo derecho.
- Gráfico de barras de nivel
- Dendograma
- Descomposición de la variación para la clasificación óptima: con las intraclases y interclases
- Centroides de las clases: Para cada muestra, en dos clases
- Distancias entre los centroides de las clases
- Objetos centrales: para las diez muestras en dos clases
- Distancias entre los objetos centrales: 1(esp acum(m)), 2(TOC)

- Resultados por clase: objetivos(2), suma de pesos, varianza intraclase, distancia mínima al centroide, Distancia media al centroide, Distancia máxima al centroide
- Resultados por objetivo: con observación de (esp acum, prof. de ext, Ro, TOC)
- Perfiles de las clases(2)

3.3.2 Implementación de Redes Neuronales

```
*Multilayer Perceptron Network.
MLP Clasificación (MLEVEL=N) WITH Espacumm Profdeextm Ro TOC Volumencm3 Metano
Butano Hexano Log10C1C3C6 D12 @13C2 TSDmgL Ph SALppl CONDUCTIuScm CameqL MgmeqL
NameqL ClmeqL SO4meqL HCO3meqL
/RESCALE COVARIATE=STANDARDIZED
/PARTITION TRAINING=7 TESTING=3 HOLDOUT=0
/ARCHITECTURE AUTOMATIC=YES (MINUNITS=1 MAXUNITS=50)
/CRITERIA TRAINING=BATCH OPTIMIZATION=SCALEDCONJUGATE LAMBDAINITIAL=0.0000005
SIGMAINITIAL=0.00005 INTERVALCENTER=0 INTERVALOFFSET=0.5 MEMSIZE=1000
/PRINT CPS NETWORKINFO SUMMARY CLASSIFICATION SOLUTION
/PLOT NETWORK
/STOPPINGRULES ERRORSTEPS= 1 (DATA=AUTO) TRAININGTIMER=ON (MAXTIME=15)
MAXEPOCHS=AUTO ERRORCHANGE=1.0E-4 ERRORRATIO=0.0010
/MISSING USERMISSING=EXCLUDE .
```

3.4 Unidad de análisis

la información usada, perteneciente a estudios de investigación en el año 2010 y 2011 en el sinclinal Checua-Lenguazaque ,zona minera de extracción de carbón y aparición de fuente de gas metano en evaluación de reservas como hidrocarburo, por el grupo de investigación de hidrocarburos y carbones de la universidad industrial de Santander, facilitada por su director Mario García González. Ver tablas 1, 2,3.

4. PROCEDIMIENTO

4.1 Tratamiento de los datos

Representada por la organización de la data, arreglo de unidades y nombres de los análisis. A continuación se muestra una tabla, donde se presenta los análisis de roca hechos a diez muestras, las cuales fueron renombradas para mayor comodidad en el tratamiento estadístico, el nombre de muestra es equivalente al manto de carbón al que se le hizo el análisis(**Ver tablas de agua y gases como anexos**).

Muestra/variable	Esp acum(m)	Prof. de ext. (m)	Ro	TOC
Muestra 1	372,2	500	0,74	76,8
Muestra 2	350	250	0,7	78,3
Muestra 3	337,03	88	0,69	78
Muestra 4	162,96	460	0,75	86
Muestra 5	153,7	507	0,76	70,7
Muestra 6	116,66	230	0,83	82,5
Muestra 7	116,66	200	0,81	82,5
Muestra 8	98,14	450	0,88	77,7
Muestra 9	44,44	500	0,94	87,8
Muestra 10	22,44	400	0,97	83,8

Tabla 5. Datos de roca

4.2 Aplicación de cluster y redes neuronales

En esta fase se hace el ingreso de los datos al Software Excel y el uso de sus diferentes aplicaciones en análisis clúster y SPSS, de donde se obtienen los resultados.

4.2.1 Aplicación Cluster

4.2.1.1 Resultados de las corridas de Clúster para análisis de roca

Estadísticos descriptivos

Se tiene cuatro observaciones el espesor acumulado, la profundidad de extracción del manto, la Reflectancia de la vitrinita y la cantidad de carbono orgánico, cada análisis tiene diez observaciones, completas es decir sin datos perdidos. La variable espesor acumulado y profundidad de extracción presentan una mayor desviación típica que el Ro y el TOC.

Variable	Mínimo	Máximo	Media	Desv. típica
Esp acum(m)	22,440	372,200	177,423	128,816
Prof. de ext. (m)	88,000	507,000	358,500	152,414
Ro	0,690	0,970	0,807	0,097
TOC	70,700	87,800	80,410	5,066

Tabla 6. Resultados estadísticos descriptivos para roca

Matriz de proximidad (Distancia euclídea)

En esta tabla se observa los valores de la distancia euclídea, de la cual podemos identificar las menores distancias de las cuales saldrán las agrupaciones de clúster. Como se observa en la matriz las menores distancias son la del TOC con el Ro seguido por profundidad de extracción del manto.

	Esp acum(m)	Prof. de ext. (m)	Ro	TOC
Esp acum(m)	0	4,852	5,808	5,170
Prof. de ext. (m)	4,852	0	3,358	4,293
Ro	5,808	3,358	0	2,958
TOC	5,170	4,293	2,958	0

Tabla 7. Matriz de proximidad para roca

Estadísticos de los nodos

La tabla muestra los tres nodos con cada hijo, con lo cual se puede armar el árbol binario, con dos clúster, con casi el mismo peso.

Nodo	Nivel	Peso	Objetos	Hijo izquierdo	Hijo derecho
7	17,795	4	4	1	6
6	8,444	3	3	2	5
5	4,376	2	2	3	4

Tabla 8. Estadísticos de los nodos para roca

Gráfico de barras de los niveles

Con esta grafica se puede apreciar que el nodo siete es el de mayor nivel seguido por el nodo seis y por último el nodo cinco con el menor nivel.

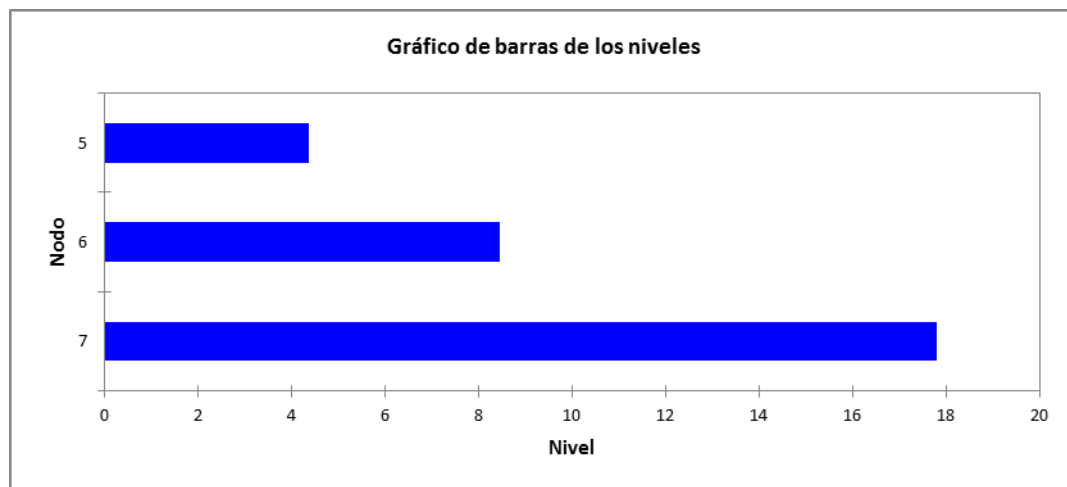


Figura 3. Grafico de barras de niveles para roca

Dendrograma

Se puede observar en la figura que se arman dos clúster, con los análisis anteriores se definieron los nodos y sus pesos, dando lugar a esta estructura de árbol; donde el nodo siete define los dos clúster, el hijo izquierdo será el espesor acumulado y el hijo derecho o nodo seis será dividido a su vez en hijo izquierdo que será la profundidad de extracción y el hijo derecho el nodo cinco que a su vez tendrá de hijo izquierdo el Ro y de hijo derecho el TOC, siendo estas dos variables las de menor distancia euclídea.

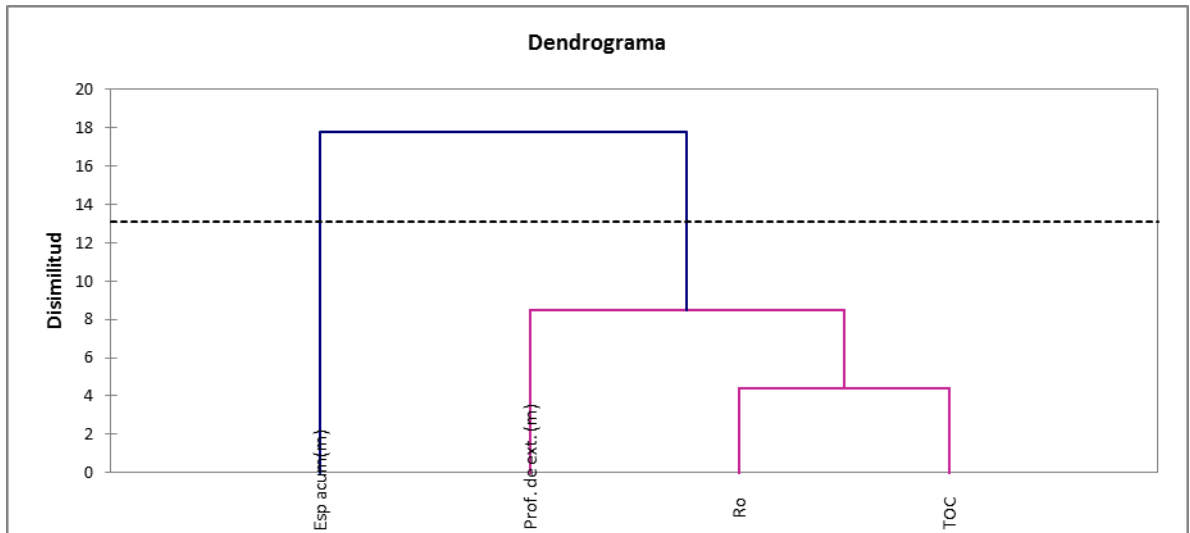


Figura 4.Dendrograma para roca

Dendrograma

Según el análisis del dendrograma, se establecen claramente dos clúster, a continuación se muestra el dendrograma con clases, donde se observa la configuración final.

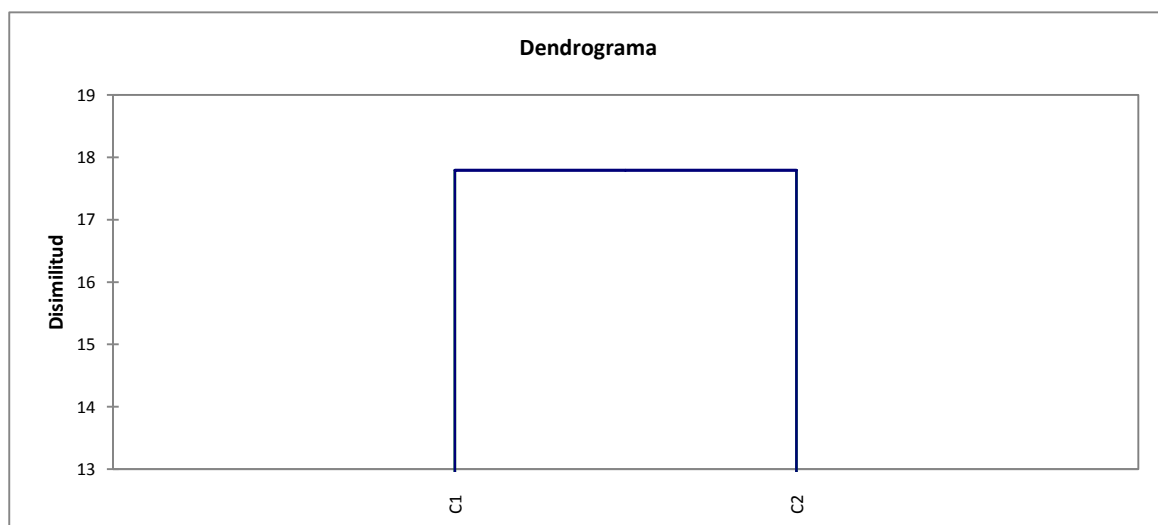


Figura 5. Dendrograma para roca

Centroides de las clases

La siguiente tabla nos muestra las clases establecidas con los valores para cada muestra con respecto a cada clase.

Clase	Muestra 1	Muestra 2	Muestra 3	Muestra 4	Muestra 5	Muestra 6	Muestra 7	Muestra 8	Muestra 9	Muestra 10
1	372,200	350,000	337,030	162,960	153,700	116,660	116,660	98,140	44,440	22,440
2	192,513	109,667	55,563	182,250	192,820	104,443	94,437	176,193	196,247	161,590

Tabla 9. Centroides de las clases para roca

Distancias entre los centroides de las clases

La distancia entre los centroides de la clase uno y dos es 469,379

	1	2
1	0	469,379
2	469,379	0

Tabla 10. Distancia entre los centroides de las clases para roca

Objetos centrales

En la siguiente tabla se presenta los valores de cada muestra en función de cada clase el espesor acumulado y el total de carbono orgánico.

Clase	Muestra 1	Muestra 2	Muestra 3	Muestra 4	Muestra 5	Muestra 6	Muestra 7	Muestra 8	Muestra 9	Muestra 10
1 (Esp acum(m))	372,200	350,000	337,030	162,960	153,700	116,660	116,660	98,140	44,440	22,440
2 (TOC)	76,800	78,300	78,000	86,000	70,700	82,500	82,500	77,700	87,800	83,800

Tabla 11. Objetos centrales para roca

Distancias entre los objetos centrales:

	1 (Esp acum(m))	2 (TOC)
1 (Esp acum(m))	0	499,387
2 (TOC)	499,387	0

Tabla 12. Distancia entre los objetos centrales para roca**Resultados por clase**

En la siguiente tabla se observa que la clase uno o espesor acumulado, tiene un objeto, la suma por ende es uno, no presenta varianza intraclases, ni distancias al centroide; por el contrario la clase dos presenta tres objetos (profundidad de extracción, Ro, TOC) y es donde se presenta toda la varianza intraclases, una distancia mínima al centroide de 259,212 y una distancia máxima al centroide de la clase de dos de 736,296.

Clase	1	2
Objetos	1	3
Suma de los pesos	1	3
Varianza intraclase	0,000	422496,753
Distancia mínima al centroide	0,000	259,212
Distancia media al centroide	0,000	493,656
Distancia máxima al centroide	0,000	736,296
	Esp acum(m)	Prof. de ext. (m)
		Ro
		TOC

Tabla 13. Resultados por clase para roca**Resultados por objeto**

En la siguiente tabla nuevamente se presenta la clasificación de cada observación por clase.

Observación	Clase
Esp acum(m)	1
Prof. de ext. (m)	2
Ro	2
TOC	2

Tabla 14. Resultados por objeto para roca**Perfil de las clases**

Este grafico nos presenta el comportamiento de cada muestra con respecto a cada clase, se observa que las muestras cuatro, cinco, seis y siete, presentan un comportamiento similar en las dos clases. La muestra tres presenta la mayor caída en la clase dos, por tener el valor más bajo en una de las observaciones. La clase uno línea de color rojo, tiene una sola observación y sus valores son en descenso por presentar la muestra uno el mayor valor de espesor acumulado en la ubicación de este

manto de carbón y la muestra diez tiene el menor valor de espesor acumulado por la ubicación de este manto de carbón.

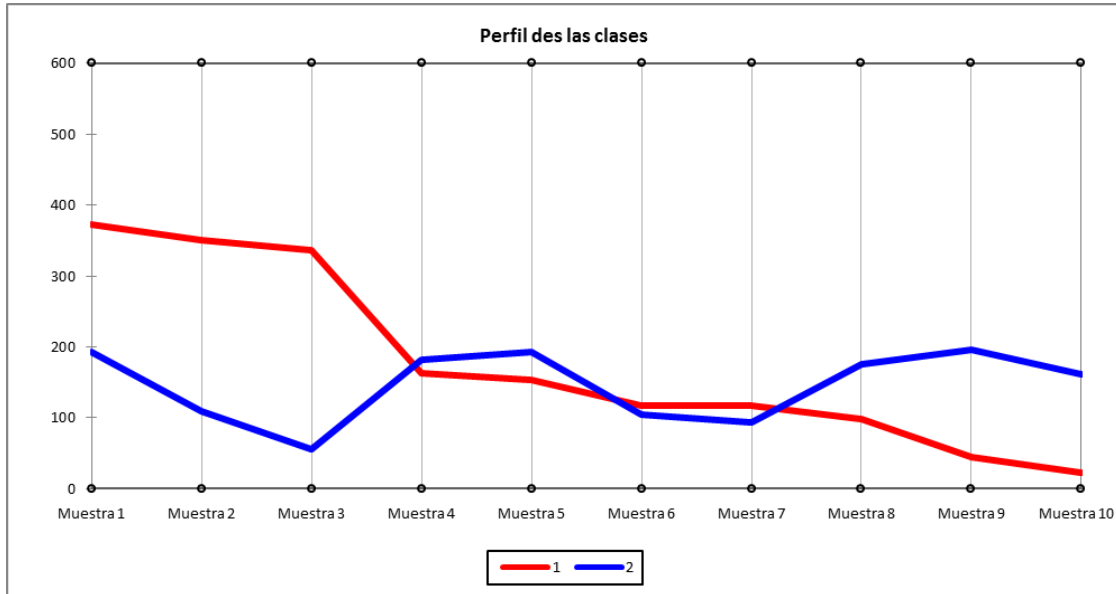


Figura 6. Perfil de clases para roca

4.2.1.2 Resultados de las corridas de Clúster para análisis de Agua

Estadísticos descriptivos

Se tiene diez variables, cada una con diez observaciones, completas es decir sin datos perdidos. La variable conductividad presenta una mayor desviación típica, y el NO₃ la menor desviación típica.

Variable	Mínimo	Máximo	Media	Desv. típica
Esp acum(m)	22,440	372,200	177,423	128,816
Prof. de ext. (m)	88,000	507,000	358,500	152,414
TSD mg/L	249,600	2372,500	1218,500	591,036
Ph	3,790	8,350	7,203	1,295
SAL/ppl	0,000	1,800	0,800	0,514
CONDUCTI (µS/cm)	384,000	3650,000	1874,600	909,295
ALCALI (mmol/L)	0,000	7,500	4,570	2,641
DUREZA (meq/L) CaCO ₃	0,360	16,470	5,348	5,939
CO ₃ /meq/L	0,000	2,410	0,483	0,864
K/meq/L	0,020	0,700	0,223	0,211
NO ₃ /meq/L	0,280	0,380	0,353	0,044
Ca meq/L	0,360	16,440	5,338	5,927
Mg meq/L	0,280	11,580	4,033	4,515
Na meq/L	1,330	22,660	9,323	6,166
Cl meq/L	0,230	2,440	0,792	0,777
SO ₄ meq/L	6,490	47,890	23,547	15,990
HCO ₃ meq/L	0,400	6,850	4,025	2,206

Tabla 15. Estadísticos descriptivos para agua

Matriz de proximidad (Distancia euclídea)

En esta tabla se observa los valores de la distancia euclídea, de la cual podemos identificar las menores distancias de las cuales saldrán las agrupaciones de clúster.

Por ejemplo las variables TSD, Conductividad y sal son agrupadas por presentar una distancia euclídea similar por ende se relacionan en una clase del clúster. De la misma forma las variables al calidad y HCO₃ por presentar una distancia euclídea similar también se agrupan en otra clase del clústeres esa forma se deben analizar todas las distancias.

Variable	Esp acum(m)	Prof. de ext. (m)	TSD mg/L	Ph	SAL/ppl	CONDUCTI (μS/cm)	ALCALI (mmol/L)	DUREZA (meq/L) CaCO ₃
Esp acum(m)	0	4,852	2,736	4,563	2,781	2,736	4,117	3,504
Prof. de ext. (m)	4,852	0	4,651	4,657	4,608	4,651	4,897	4,314
TSD mg/L	2,736	4,651	0	5,129	0,194	0,000	4,801	3,354
Ph	4,563	4,657	5,129	0	5,156	5,129	3,124	5,487
SAL/ppl	2,781	4,608	0,194	5,156	0	0,194	4,803	3,309
CONDUCTI (μS/cm)	2,736	4,651	0,000	5,129	0,194	0	4,801	3,354
ALCALI (mmol/L)	4,117	4,897	4,801	3,124	4,803	4,801	0	5,097
DUREZA (meq/L) CaCO ₃	3,504	4,314	3,354	5,487	3,309	3,354	5,097	0
CO ₃ /meq/L	5,036	4,682	4,709	3,377	4,728	4,709	2,987	5,076
K/meq/L	3,852	3,918	3,601	5,558	3,616	3,601	4,788	2,110
NO ₃ /meq/L	5,070	1,978	4,803	4,614	4,732	4,803	4,675	4,445
Ca meq/L	3,505	4,314	3,354	5,487	3,309	3,354	5,097	0,002
Mg meq/L	3,924	4,066	3,603	4,975	3,556	3,603	5,515	1,702
Na meq/L	2,791	5,041	2,299	4,111	2,394	2,299	3,602	4,795
Cl meq/L	3,716	4,638	2,005	5,691	1,980	2,005	4,959	2,991
SO ₄ meq/L	2,680	4,126	2,634	5,424	2,585	2,634	5,004	1,573
HCO ₃ meq/L	3,221	4,836	3,815	2,890	3,857	3,815	2,388	4,773

Tabla 16. Matriz de proximidad para agua

Variable	CO ₃ /meq/L	K/meq/L	NO ₃ /meq/L	Ca meq/L	Mg meq/L	Na meq/L	Cl meq/L	SO ₄ meq/L	HCO ₃ meq/L
Esp acum(m)	5,036	3,852	5,070	3,505	3,924	2,791	3,716	2,680	3,221
Prof. de ext. (m)	4,682	3,918	1,978	4,314	4,066	5,041	4,638	4,126	4,836
TSD mg/L	4,709	3,601	4,803	3,354	3,603	2,299	2,005	2,634	3,815
Ph	3,377	5,558	4,614	5,487	4,975	4,111	5,691	5,424	2,890
SAL/ppl	4,728	3,616	4,732	3,309	3,556	2,394	1,980	2,585	3,857
CONDUCTI (μS/cm)	4,709	3,601	4,803	3,354	3,603	2,299	2,005	2,634	3,815
ALCALI (mmol/L)	2,987	4,788	4,675	5,097	5,515	3,602	4,959	5,004	2,388
DUREZA (meq/L) CaCO ₃	5,076	2,110	4,445	0,002	1,702	4,795	2,991	1,573	4,773
CO ₃ /meq/L	0	4,363	3,926	5,076	5,101	4,080	4,780	5,339	3,294
K/meq/L	4,363	0	4,199	2,110	3,088	4,600	3,122	2,726	4,506
NO ₃ /meq/L	3,926	4,199	0	4,444	4,163	5,185	4,668	4,491	4,976
Ca meq/L	5,076	2,110	4,444	0	1,701	4,795	2,991	1,574	4,773
Mg meq/L	5,101	3,088	4,163	1,701	0	5,033	3,607	2,424	4,916
Na meq/L	4,080	4,600	5,185	4,795	5,033	0	3,333	4,111	2,743
Cl meq/L	4,780	3,122	4,668	2,991	3,607	3,333	0	2,942	4,835
SO ₄ meq/L	5,339	2,726	4,491	1,574	2,424	4,111	2,942	0	4,359
HCO ₃ meq/L	3,294	4,506	4,976	4,773	4,916	2,743	4,835	4,359	0

Tabla 17. Continuidad de la matriz de proximidad para agua

Estadísticos de los nodos

La tabla muestra los dieciséis nodos con cada hijo, con lo cual se puede armar el árbol binario, con tres clúster, con diferentes pesos. Por ejemplo el nodo 33 tiene el mayor nivel 40,898 y el mayor peso es diecisiete, por ende será de una clase definida del clúster del dendograma.

Nodo	Nivel	Peso	Objetos	Hijo izquierdo	Hijo derecho
33	40,898	17	17	29	32
32	28,554	13	13	30	31
31	22,001	7	7	22	27
30	7,056	6	6	25	26
29	5,810	4	4	9	28
28	5,087	3	3	4	24
27	4,343	5	5	10	23
26	3,894	2	2	1	14
25	2,984	4	4	15	20
24	2,851	2	2	7	17
23	2,504	4	4	13	21
22	1,955	2	2	2	11
21	1,650	3	3	16	19
20	0,025	3	3	5	18
19	0,000	2	2	8	12
18	0,000	2	2	3	6

Tabla 18. Estadísticos de los nodos para agua

Gráfico de las barras de los niveles

En el siguiente gráfico, se observa que los nodos menores tienen menor nivel y los nodos mayores como el nodo 33 tienen el mayor nivel, del cual se desprenderá los clúster del dendograma.

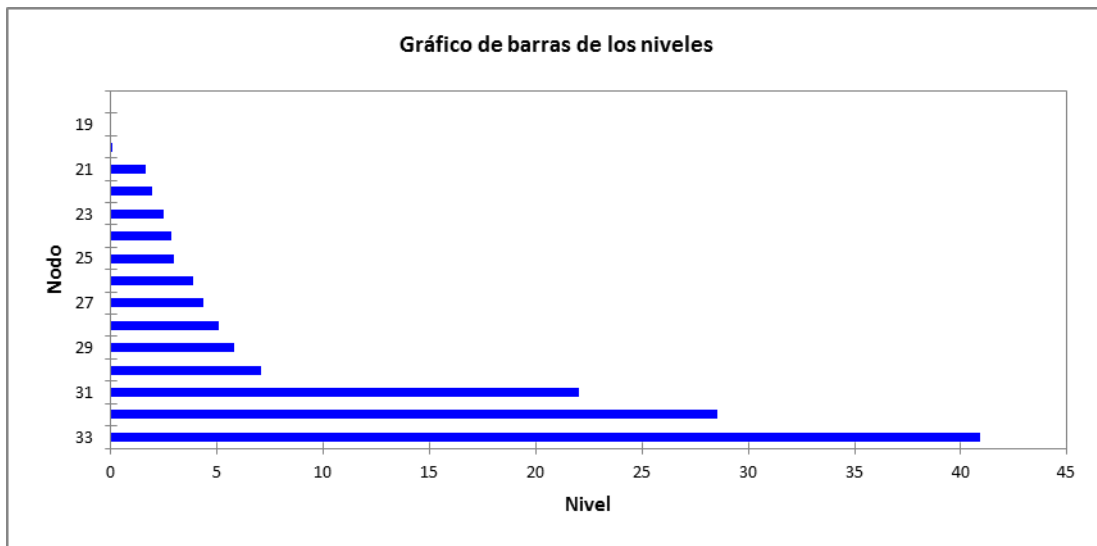


Figura 7. Gráfico de barras de los niveles para agua

Dendograma

Se puede observar en la figura que se arman tres clúster, donde la estructura de árbol la define el nodo treinta tres con un hijo derecho treinta dos y un hijo izquierdo veintinueve.

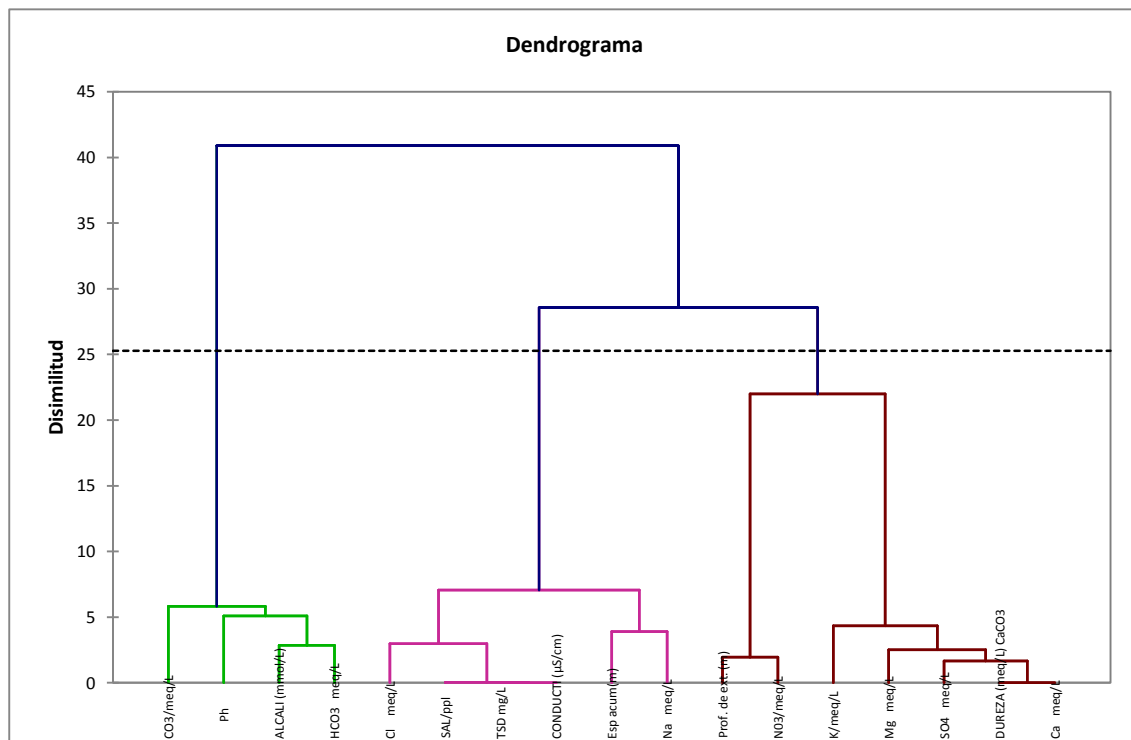


Figura 8. Dendograma para agua

El nodo treinta dos tiene el hijo derecho treinta uno y el hijo izquierdo treinta esta división será el origen de dos clúster, es decir los cluster del lado derecho del dendograma.

En el clúster del nodo veintinueve encontramos un grupo con cuatro variables; CO_3 , pH, Alcalinidad, HCO_3 ; donde la alcalinidad y el HCO_3 presentan la menor distancia de este grupo, seguido por el pH y de último CO_3 .

Dendograma

Según el análisis hecho anteriormente se puede decir que se establecen tres clases en el dendograma

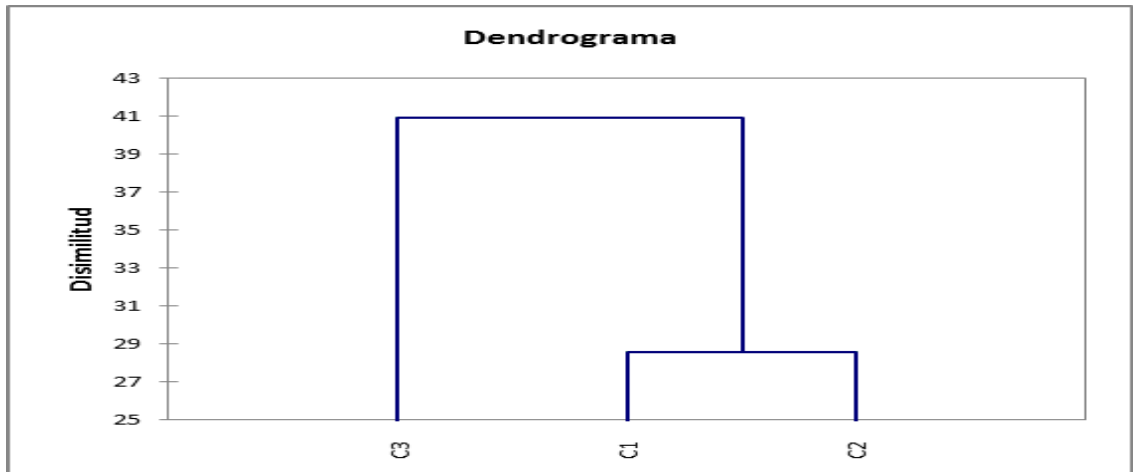


Figura 9.Dendograma para agua

Centroides de las clases

La siguiente tabla nos muestra las clases establecidas con los valores para cada muestra con respecto a cada clase.

Clase	Muestra 1	Muestra 2	Muestra 3	Muestra 4	Muestra 5	Muestra 6	Muestra 7	Muestra 8	Muestra 9	Muestra 10
1	639,443	568,910	1064,337	831,863	488,275	333,392	391,130	476,805	565,257	109,652
2	75,026	48,477	18,840	78,717	78,921	34,267	30,383	70,784	73,287	58,929
3	4,333	5,093	4,328	1,048	3,693	4,598	5,493	2,693	5,960	3,468

Tabla 19.Centroides de las clases para agua

Distancias entre los centroides de las clases

La distancia entre los centroides de la clase uno y dos es 1745.381 y a la clase tres 1890.937, la distancia de la clase tres a dos 180.354, siendo la menor perteneciente a los extremos del dendograma con referencia a los nodos veintisiete a la derecha y veintinueve a la izquierda.

	1	2	3
1	0	1745,381	1890,937
2	1745,381	0	180,354
3	1890,937	180,354	0

Tabla 20.Centroides de las clases para agua

Objetos centrales

En la siguiente tabla se presenta los valores de cada muestra en función de cada clase: el espesor acumulado, SO₄, HCO₃

Clase	Muestra 1	Muestra 2	Muestra 3	Muestra 4	Muestra 5	Muestra 6	Muestra 7	Muestra 8	Muestra 9	Muestra 10
1 (Esp acum(m))	372,200	350,000	337,030	162,960	153,700	116,660	116,660	98,140	44,440	22,440
2 (SO ₄ meq/L)	22,700	47,890	31,230	46,910	33,940	6,490	7,280	20,970	11,400	6,660
3 (HCO ₃ meq/L)	5,230	6,440	5,230	0,400	3,620	4,430	4,830	2,420	6,850	0,800

Tabla 21. Objetos centrales para agua

Distancias entre los objetos centrales

	1 (Esp acum(m))	2 (SO ₄ meq/L)	3 (HCO ₃ meq/L)
1 (Esp acum(m))	0	605,096	669,243
2 (SO ₄ meq/L)	605,096	0	78,684
3 (HCO ₃ meq/L)	669,243	78,684	0

Tabla 22. Distancia entre los objetos centrales para agua

Resultados por clase

En la siguiente tabla se observa que la clase uno o espesor acumulado, tiene seis objetos, la suma por ende es seis, presenta la varianza intraclases más elevada, la clase dos presenta siete objetos y su suma es siete y tiene una varianza intraclases alta, la clase tres presenta cuatro objetos y tiene la menor varianza y la distancia mínima al centroide 94,449, ver tabla.

Clase	1	2	3
Objetos	6	7	4
Suma de los pesos	6	7	4
Varianza intraclase	7865324,055	207786,730	94,449
Distancia mínima al centroide	1323,721	127,576	3,508
Distancia media al centroide	2326,220	295,837	7,623
Distancia máxima al centroide	4624,269	1031,467	11,841
	Esp acum(m)	Prof. de ext. (m)	Ph
	TSD mg/L	DUREZA (meq/L) CaCO ₃	ALCALI (mmol/L)
	SAL/ppl	K/meq/L	CO ₃ /meq/L
	CONDUCTI (μS/cm)	N03/meq/L	HCO ₃ meq/L
	Na meq/L	Ca meq/L	
	Cl meq/L	Mg meq/L	
		SO ₄ meq/L	

Tabla 23. Resultados por clase para agua

Resultados por objeto

A continuación se presenta la tabla con la clasificación de cada observación en la clase, en la clase uno se encuentran seis observaciones (Espesor acumulado, TSD, SAL, CONDUCTI, Na, Cl) en la clase dos (Prof de extracción, dureza, K, NO₃, Ca, Mg, SO₄) en la clase tres (Ph,Alcalinidad,CO₃, HCO₃).

Observación	Clase
Esp acum(m)	1
Prof. de ext. (m)	2
TSD mg/L	1
Ph	3
SAL/ppl	1
CONDUCTI (μS/cm)	1
ALCALI (mmol/L)	3
DUREZA (meq/L) CaCO ₃	2
CO ₃ /meq/L	3
K/meq/L	2
NO ₃ /meq/L	2
Ca meq/L	2
Mg meq/L	2
Na meq/L	1
Cl meq/L	1
SO ₄ meq/L	2
HCO ₃ meq/L	3

Tabla 24.Resultados por objeto para agua

Perfiles de las clases

Este grafico nos presenta el comportamiento de cada muestra con respecto a cada clase, se observa que el comportamiento de todas las muestras se comportan similares según la clasificación de la clase dos línea azul y clase tres línea tres.

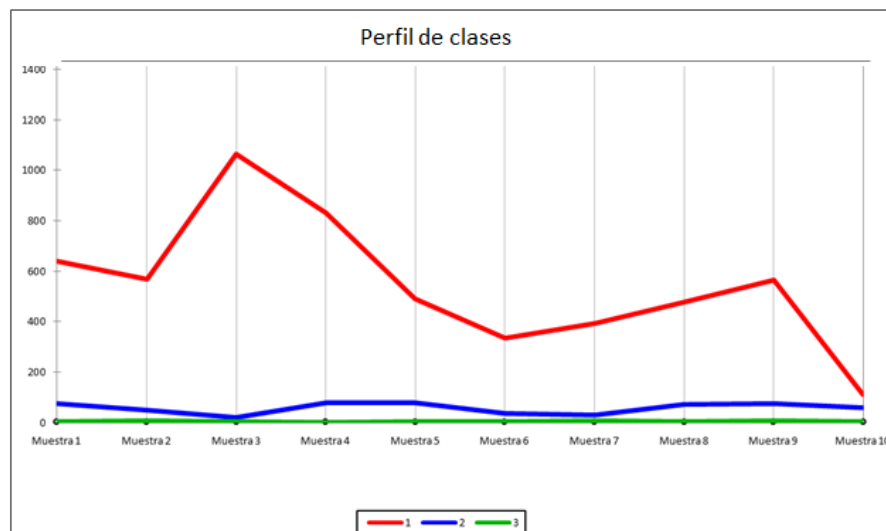


Figura 10.Perfil de clases para agua

Las muestras en la clase uno, presentan una variación considerable con respecto al comportamiento frente a las otras dos clases, el mayor incremento en la clase uno se presenta en la muestra tres ya que según las observaciones de esta muestra son las de mayor valor, la clase descende y vuelve a aumentar en la muestra nueve., la clase uno presenta considerablemente variaciones en la muestras.

4.2.1.3 Resultados de las corridas de Clúster para análisis de gas

Estadísticos descriptivos

Se tiene nueve variables, cada una con diez observaciones, completas es decir sin datos perdidos. La variable volumen en m³ presenta la mayor desviación con 198,017 y la variable %Metano la menor desviación típica con 0,011.

Variable	Mínimo	Máximo	Media	Desv. típica
Esp acum(m)	22,440	372,200	177,423	128,816
Prof. de ext. (m)	88,000	507,000	358,500	152,414
Volumen cm ³	0,000	636,150	105,262	198,017
%Metano	0,971	0,999	0,991	0,011
%Butano	0,000	0,018	0,006	0,008
%Hexano	0,000	0,013	0,003	0,005
Log ₁₀ /C ₁ /C ₃ +C ₆	1,730	3,390	2,595	0,649
δ D _{1/2}	-241,790	-197,170	-227,771	12,428
δ ₁₃ C/2	-62,900	-32,730	-50,621	9,937

Tabla 25. Estadísticos descriptivos para gas

Matriz de proximidad (Distancia euclídea):

	Esp acum(m)	Prof. de ext. (m)	Volumen cm ³	%Metano	%Butano	%Hexano	Log ₁₀ /C ₁ /C ₃ +C ₆	δ D _{1/2}	δ ₁₃ C/2
Esp acum(m)	0	4,852	4,215	4,635	4,083	3,420	4,667	3,924	4,620
Prof. de ext. (m)	4,852	0	3,084	2,744	5,122	5,489	3,002	5,038	4,599
Volumen cm ³	4,215	3,084	0	3,371	4,962	4,840	2,868	4,643	4,440
%Metano	4,635	2,744	3,371	0	5,935	5,825	1,313	5,364	5,516
%Butano	4,083	5,122	4,962	5,935	0	2,276	5,896	3,038	2,573
%Hexano	3,420	5,489	4,840	5,825	2,276	0	5,523	2,447	2,487
Log ₁₀ /C ₁ /C ₃ +C ₆	4,667	3,002	2,868	1,313	5,896	5,523	0	5,117	5,120
δ D _{1/2}	3,924	5,038	4,643	5,364	3,038	2,447	5,117	0	2,562
δ ₁₃ C/2	4,620	4,599	4,440	5,516	2,573	2,487	5,120	2,562	0

Tabla 26. Matriz de proximidad para gas

En esta tabla se observa los valores de la distancia euclídea, de la cual podemos identificar las menores distancias de las cuales saldrán las agrupaciones de clúster.

Por ejemplo las variables butano, hexano, isotopos de deuterio y carbono el espesor acumulado tienen las menores distancias si se analiza solo el Hexano frente a todas las variables. Este grupo formara una clase en el dendograma.

Estadísticos de los nodos

La tabla muestra los nodos con cada hijo, con los que arma tres clúster, con diferentes pesos. Por ejemplo el nodo 17 tiene el mayor nivel 42,191 y el mayor peso es nueve, por ende será el que definirá los clúster del dendograma.

Nodo	Nivel	Peso	Objetos	Hijo izquierdo	Hijo derecho
17	42,191	9	9	15	16
16	11,034	5	5	1	13
15	6,224	4	4	10	14
14	4,756	2	2	2	3
13	4,069	4	4	11	12
12	3,283	2	2	8	9
11	2,590	2	2	5	6
10	0,862	2	2	4	7

Tabla 27. Estadísticos de los nodos para gas

Gráfico de las barras de los niveles

En el siguiente gráfico, se observa que los nodos menores tienen menor nivel y los nodos mayores como el nodo 17 tienen el mayor nivel, y será de donde se desprenderán los clúster.

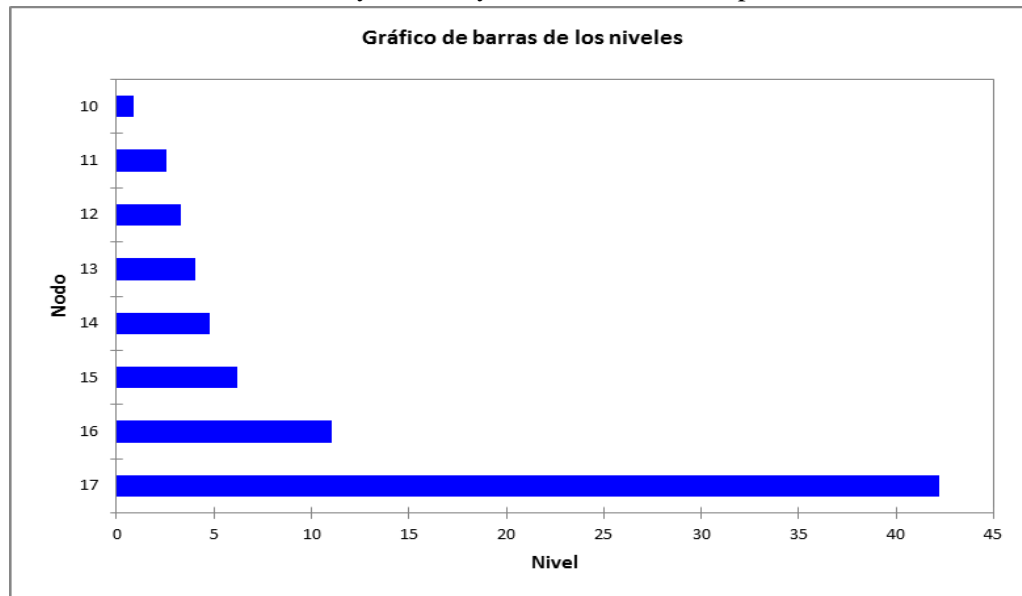


Figura 11. Gráfico de barras de los niveles para gas

Dendograma

Se puede observar en la figura que se arman dos clúster, con los análisis anteriores se definieron los nodos y sus pesos, dando lugar a esta estructura de árbol; de donde el nodo diecisiete se desprenden dos clúster.

El nodo dieciséis es el hijo derecho del nodo treinta y tres y está compuesto por las variables: Espesor acumulado, % Butano, % Hexano, y los isotopos de Deuterio y carbono.

El nodo quince es el hijo izquierdo del nodo treinta y tres y está compuesto por las variables %Metano, $\log_{10} \log_{10} C1/C3+C6$, profundidad de extracción y volumen de gas.

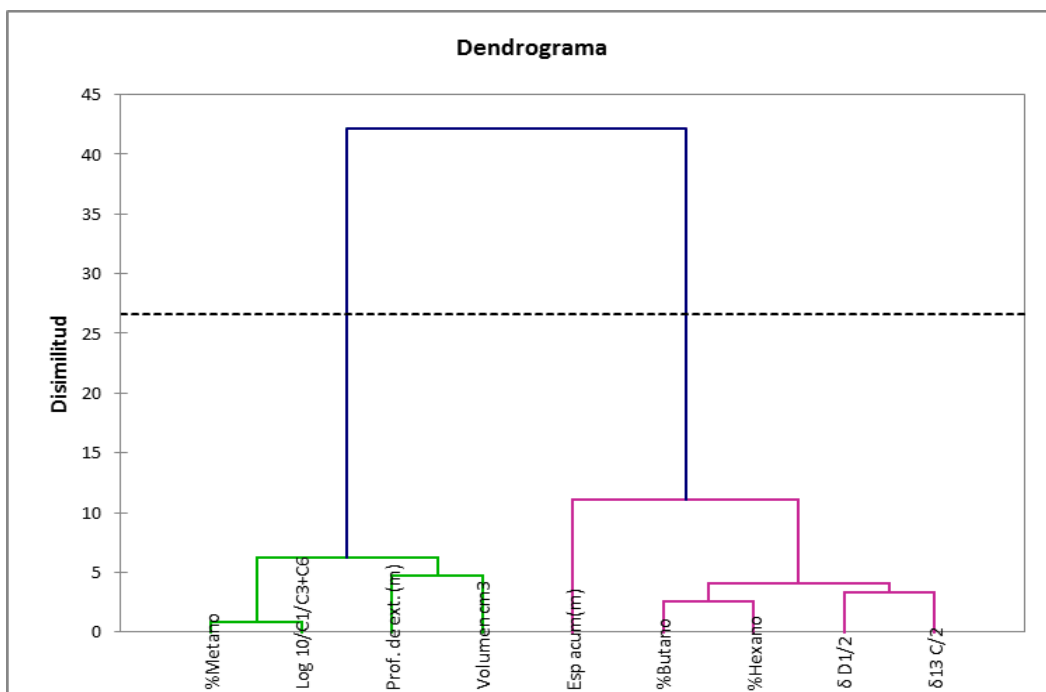


Figura 12. Dendrograma para gas

Dendograma

En la siguiente figura se observa las clases C1 y C2 o dos clúster que se forman en la realización de este análisis que tiene nueve variables.

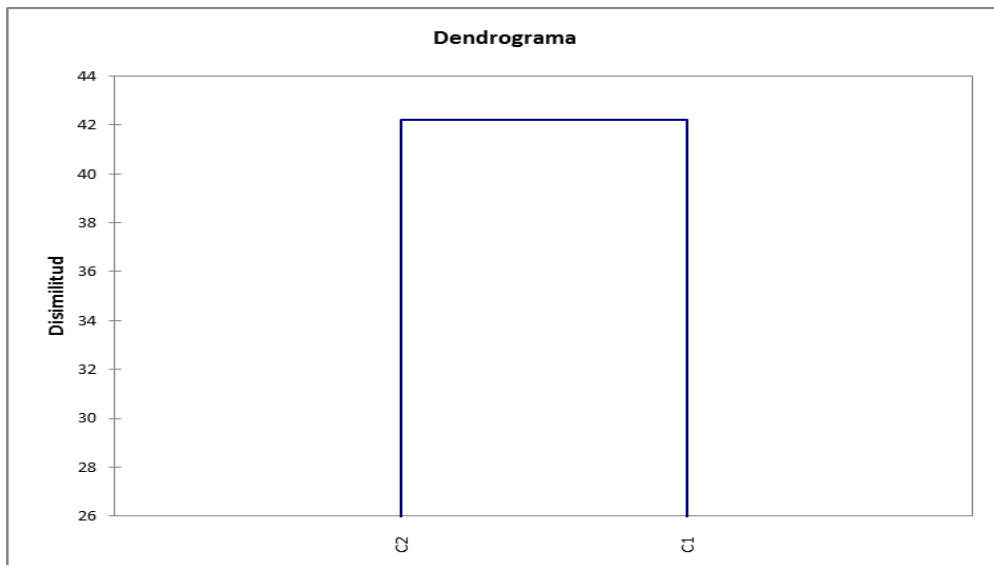


Figura 13. Dendrograma para gas

Centroides de las clases

La siguiente tabla nos muestra las clases establecidas con los valores para cada muestra con respecto al centroide de la clase.

Clase	Muestra 1	Muestra 2	Muestra 3	Muestra 4	Muestra 5	Muestra 6	Muestra 7	Muestra 8	Muestra 9	Muestra 10
1	16,508	10,531	19,906	-27,370	-24,770	-28,934	-36,204	-36,249	-42,920	-52,417
2	156,288	67,029	23,203	167,112	286,885	58,688	59,332	121,853	126,037	101,942

Tabla 28. Centroides de las clases para gas

Distancias entre los centroides de las clases

La distancia entre los centroides de la clase uno y dos es 501,864

	1	2
1	0	501,864
2	501,864	0

Tabla 29. Distancia entre los centroides de las clases para gas

Objetos centrales

En la siguiente tabla se presenta los valores de cada muestra en función de cada clase: el porcentaje de Hexano y el volumen de gas en cm³; se observa que la muestra en todas las muestras presentan un comportamiento muy similar con respecto a la clase uno, mientras en la clase dos las muestras uno, cuatro y cinco presentan un comportamiento muy diferente al grupo..

Clase	Muestra 1	Muestra 2	Muestra 3	Muestra 4	Muestra 5	Muestra 6	Muestra 7	Muestra 8	Muestra 9	Muestra 10
1 (%Hexano)	0,004	0,001	0,013	0,000	0,000	0,011	0,001	0,000	0,001	0,002
2 (Volumen cm ³)	121,740	14,400	2,050	204,060	636,150	2,050	33,410	34,650	0,000	4,110

Tabla 30. Objetos centrales para gas

Distancias entre los objetos centrales

En la siguiente tabla se muestra la distancia que hay en los centroides de la clase uno y el centroide de la clase porcentaje con 680,952.

	1 (%Hexano)	2 (Volumen cm3)
1 (%Hexano)	0	680,952
2 (Volumen cm3)	680,952	0

Tabla 31. Distancia entre los objetos centrales para gas

Resultados por clase

En la siguiente tabla se observa que la clase uno o %Hexano, tiene cinco objetos, la suma por ende es cinco, presenta la varianza intraclases menor, la clase dos presenta cuatro objetos y su suma es cuatro y tiene una varianza intraclases mayor, tanto la distancia mínima, la media y la máxima. al centroide para la clase son las menores comparadas con la clase dos.

Pertencen a la clase uno las variables Espesor acumulado, %Butano, %Hexano, $\delta D_{1/2}$, $\delta_{13} C/2$. las variables de la clase dos la Profundidad de extracción del manto, volumen de gas en cm^3 , %metano y el $\log_{10} C_1/C_3+C_6$.

Clase	1	2
Objetos	5	4
Suma de los pesos	5	4
Varianza intraclase	239965,440	402321,242
Distancia mínima al centroide	100,966	405,836
Distancia media al centroide	337,701	521,035
Distancia máxima al centroide	697,325	821,919
	Esp acum(m)	Prof. de ext. (m)
	%Butano	Volumen cm^3
	%Hexano	%Metano
	$\delta D_{1/2}$	$\log_{10} C_1/C_3+C_6$
	$\delta_{13} C/2$	

Tabla 32. Resultados por clase para gas

Resultados por objeto

A continuación se muestra la tabla de clasificación de observación por clase, donde la clase uno está conformada por Espesor acumulado, %Butano, %Hexano, $\delta D_{1/2}$, $\delta_{13} C/2$. la clase dos está conformada por la Profundidad de extracción del manto, volumen de gas en cm^3 , %metano y el $\log_{10} C_1/C_3+C_6$.

Observación	Clase
Esp acum(m)	1
Prof. de ext. (m)	2
Volumen cm3	2
%Metano	2
%Butano	1
%Hexano	1
Log 10/C1/C3+C6	2
δ D1/2	1
$\delta^{13}C/2$	1

Tabla 33. Resultados por objeto para gas

Perfiles de las clases

Este grafico nos presenta el comportamiento de cada muestra con respecto a cada clase, se observa que el comportamiento de todas las muestras se comportan similares según la clasificación de la clase uno línea roja. La clase dos línea azul presenta un comportamiento diferente en la muestra uno, cuatro y cinco con un aumento.

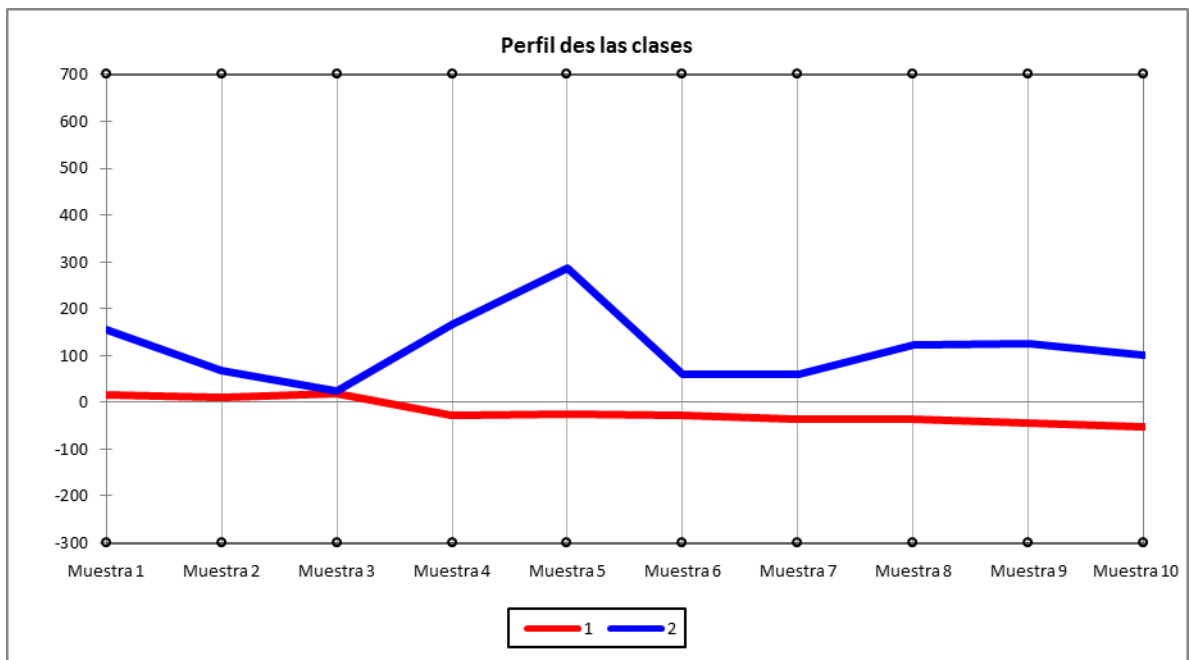


Figura 14. Perfil de las clases para gas

4.2.2 Aplicación de redes neuronales

Resultados de la Red Perceptrón Multicapa

Se utilizó la red neuronal Perceptrón multicapa a todas los análisis aplicados a las muestras de agua, gas y carbón pertenecientes a minas de extracción de carbón en el sinclinal Checua-Lenguazaque.

Se presenta la estructura de la red implementada con una variable dependiente (SI, que significa determinación de presencia de gas metano y NO que significa no determinación de presencia de gas metano), en esta se presenta una capa de entrada con el número de neuronas como factores tiene cada una de las variables trabajadas, y una capa oculta de una neurona con funciones de activación.

Finalmente, en la capa de salida se presentan dos neuronas, los cuales corresponden a los grupos determinados en el análisis clúster presentado, cada una con función de activación Softmax y se analizará el comportamiento de la red como aceptable al observar el valor del error de entropía cruzada con un aprendizaje de tipo Backpropagation.

		1	Espacumm
		2	Profdeextm
		3	Ro
		4	TOC
		5	Volumencm3
		6	Metano
Capa de entrada	Covariables	7	Butano
		8	Hexano
		9	Log10C1C3C6
		10	D12 ?D12
		11	@13C2 ?13C2
		12	TSDmgL
		13	pH

		14	SALppl	
		15	CONDUCTI μ Scm	
		16	CameqL	
		17	MgmeqL	
		18	NameqL	
		19	ClmeqL	
		20	SO4meqL	
		21	HCO3meqL	
	Número de unidades ^a			21
	Método de cambio de escala para las covariables		Tipificados	
	Número de capas ocultas			1
Capas ocultas	Número de unidades de la capa oculta 1 ^a			4
	Función de activación		Tangente hiperbólica	
	Variables dependientes	1	Clasificación	
	Número de unidades			2
Capa de salida	Función de activación		Softmax	
	Función de error		Entropía cruzada	

Tabla 34. Información sobre la red (a. Sin incluir la unidad de sesgo)

La siguiente figura representa a través de grafos la construcción del modelo de la red implementado, donde se evidencia las 21 neuronas en la capa de entrada, la neurona en la capa oculta y las 2 neuronas de la capa de salida.

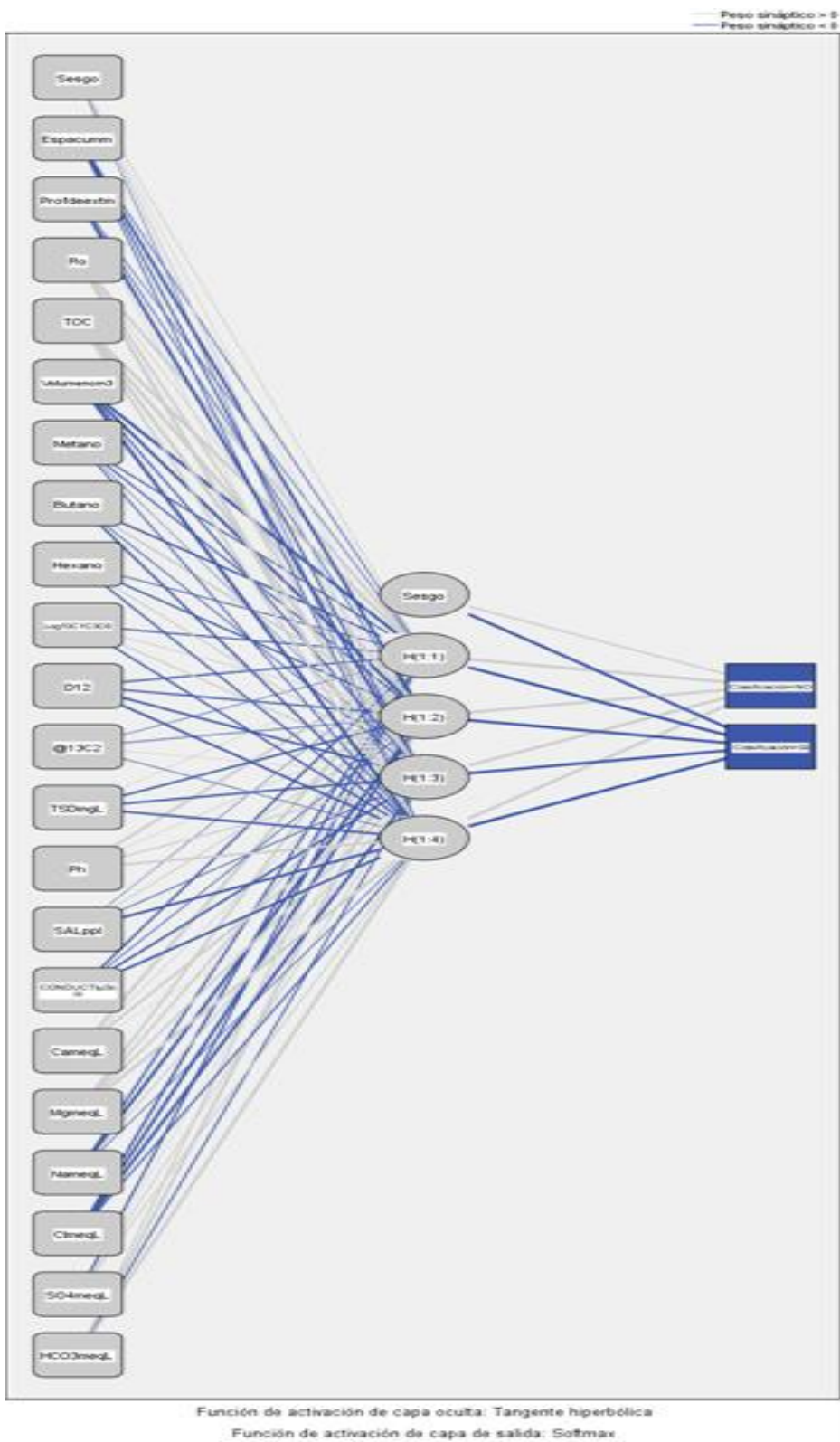


Figura 15. Arquitectura Feedforward

La clasificación de las neuronas de la capa de salida corresponde a dos variables:

- **Variable presencia de SI:**(determinación de presencia de gas metano en mantos de carbón) y está representada por las siguientes variables: Espesor acumulado, Profundidad de extracción, Volumen de gas en cm^3 , %Metano, %Butano, %Hexano, isotopos de D_{12} , isotopos de ^{13}C , TSD mgL, SAL ppl, Conductibilidad, Mg meqL, Cl meqL, $\text{Log}_{10}\text{C}_1\text{C}_3\text{C}_6$.
- **Variabes que NO presentan la presencia**, es decir no forman parte en la determinación de presencia de gas metano en mantos de carbón, gases y agua son: Reflectancia de la vitrinita, Total de carbono orgánico, pH, Ca meqL, S04 meqL y HCO_3meqL .

En la siguiente tabla, se presenta un resumen de los resultados obtenidos por la red, donde se verifica que el error de entropía cruzada en la partición de entrenamiento es mayor que cero y que el porcentaje de individuos mal clasificados es cero. Además se muestra que el conjunto de datos de prueba se encuentra un valor de entropía cruzada fue de 0,000027 y casos incorrectos del 0 %.

Resumen del modelo

Entrenamiento	Error de entropía cruzada	8,247E-005
	Porcentaje de pronósticos incorrectos	0,0%
	Regla de parada utilizada	Criterio de tasa de errores de entrenamiento (,001) alcanzado
	Tiempo de entrenamiento	0:00:00.01
Prueba	Error de entropía cruzada	2,725E-005
	Porcentaje de pronósticos incorrectos	0,0%

Tabla 35.Resumen del modelo (variable dependiente: clasificación)

En la siguiente salida se presenta la clasificación obtenida por la red, en la cual se identifica la correcta clasificación del modelo de 9 casos de entrenamiento, que corresponde al 100%.

Clasificación

Muestra	Observado	Pronosticado		
		NO	SI	Porcentaje correcto
Entrenamiento	NO	6	0	100,0%
	SI	0	3	100,0%
	Porcentaje global	66,7%	33,3%	100,0%
Prueba	NO	0	0	0,0%
	SI	0	1	100,0%
	Porcentaje global	0,0%	100,0%	100,0%

Tabla 36. Variable dependiente: Clasificación

5. ANALISIS DE RESULTADOS

En esta fase se analiza los resultados de las corridas de clúster para cada grupo de muestras: roca, agua y gas.

5.1 Análisis de resultados de corrida de clúster para Roca

La corrida arroja dos clúster bien definidos un clúster con la agrupación de las observaciones TOC y Reflectancia de la vitrinita con la menor distancia euclídea acompañados con la profundidad de extracción de manto de carbón, esta asociación tiene mucho sentido ya que si el manto es profundo, tendrá mayor madurez termal y mayor Total de carbono orgánico, variables que los separa del espesor acumulado o segundo clúster al que no se asocia otra observación.

Según el análisis de perfil de clases se puede decir que la clase uno y dos en las muestras cuatro, cinco, presenta un comportamiento muy similar esto se debe a que es las dos muestras pertenecen al mismo manto pero se encuentran en diferente ubicación de zona de explotación y las muestras seis y siete también con un comportamiento similar en las dos clases son de una misma zona de explotación por este motivo la variación en la respuesta de las clases.

Es decir muestras de diferente minas difieren un poco en las clases del clúster que las muestras que pertenecen a la misma mina de explotación.

5.2 Análisis de resultados de corrida de clúster para Agua

La corrida arroja tres clúster, En el clúster de clase uno se tiene el grupo con las variables conductividad, total de solidos disueltos, salinidad, seguido por cloro y en otro nodo del mismo clúster espesor acumulado y Na. Es importante esta asociación ya que la conductividad es directamente proporcional a la concentración de sal y al total de solidos disueltos en el agua, y su asociación con cloro también depende de que tanta interacción se tenga con medios salinos. El otro nodo que agrupa el sodio y el espesor acumulado nos da una idea de cómo se debe comportar su concentración ya que a mayor espesor acumulado se tiene un registro de mayor concentración de Na.

En el clúster de clase dos, se tiene el grupo con SO_4 , dureza, calcio seguido de magnesio y potasio; los sulfatos SO_4 se encuentran en casi todas las aguas naturales y su combinación con calcio, magnesio y potasio, los componentes más comunes de la dureza del agua la hacen característica de aguas meteóricas no aptas para consumo humano

En el clúster de clase tres, se tiene el grupo con la alcalinidad y HCO_3 seguido de pH y por ultimo CO_3 , tiene mucho sentido esta agrupación ya que la alcalinidad en el agua es causada por la presencia de iones de CO_3 y bicarbonato HCO_3 además el pH se relaciona con estos dos variables de la siguiente manera, el pH entre 4 y 8.3 se relaciona con HCO_3 y pH entre 8.3 y 11 se relaciona con CO_3 .

5.3 Análisis de resultados de corrida de clúster para Gas

Se tiene dos clúster, el clúster de la clase uno está compuesto por las variables Profundidad de la extracción del manto, volumen en cm^3 y en otro brazo del mismo clúster el %Metano y el $\log_{10}(C_1/C_3+C_6)$.

Se hace la relación de las variables a medida que aumenta la profundidad de extracción, se encuentra mayor volumen de gas acumulado en los mantos de carbón. También se hace la relación del logaritmo de los gases con el %Metano pues son directamente proporcionales, con el fin que su aumento sea asociado a gas metano de origen biogenico.

El clúster de la clase dos está conformado por los isotopos de deuterio y carbono, seguido por la agrupación de los porcentajes de los gases butano y Hexano y finalmente el espesor acumulado. Según los análisis se puede decir que el comportamiento de los isotopos de Carbono son directamente proporcionales a los isotopos de deuterio.

La agrupación de los %de Butano y %Hexano son congruentes en este estudio pues la investigación se centra sobre muestras que contienen casi un 100% de metano, por consiguiente la porción de estas dos variables son muy pequeñas en este estudio y están íntimamente relacionadas por el tipo de muestra de gas analizada.

La última variable que es el espesor acumulado se relaciona con las demás de forma proporcional es decir si hay aumento del grupo de variables esta se inclinara también por aumentar o de manera contraria si los grupos disminuyen esta tendera a disminuir.

5.4 Análisis de resultados de corrida de la Red Perceptrón Multicapa

Variable presencia de SI:(determinación de presencia de gas metano en mantos de carbón) y está representada por las siguientes variables:

- **Espesor acumulado, Profundidad de extracción, Volumen de gas en cm^3 :** son variables que se deben usar en toda decisión ya que aportan la ubicación o detección de la muestra de gas por ende hace parte de las variables que si dan presencia de gas en los mantos de carbón.
- **%Metano, %Butano, %Hexano:** El porcentaje de estas variables es muy importante ya que en análisis de cromatografía de gases el porcentaje debe ser mayor que el de butano y hexano, lo cual indicara que efectivamente la muestra de gas que presenta el carbón es metano.
- **Isotopos de D_{12} e Isotopos de ^{13}C :** estos dos isotopos se deben analizar en conjunto para tomar decisión si hay gas biogenico, termogénico o mezcla en las muestras de carbón, es decir es indispensable el analizar estos variables.
- **$\text{Log}_{10}\text{C}_1/\text{C}_3+\text{C}_6$:** El índice de sequedad es una variable muy importante ya que se usa con el isotopo de Isotopos de ^{13}C para determinar el origen del gas encontrado sea biogenico o termogénico.
- **TSD mgL, SAL ppl, Conductibilidad, Cl meqL:** Es importante esta asociación ya que la conductividad es directamente proporcional a la concentración de sal y al total de solidos disueltos en el agua, y su asociación con cloro también depende de que tanta interacción se tenga con medios salinos. Lo anterior ya se había concluido en los análisis de asociación clúster variables indispensables en la determinación de gas en los mantos de carbón ya que se manejan ciertos valores en concentración para esta asociación.
- **Mg meqL:** este catión es uno de los marcadores en la aparición de gas metano asociado a mantos de carbón, con poca concentración en las aguas encontradas en las minas de carbón.

Variables que NO presentan la presencia, es decir no forman parte en la determinación de presencia de gas metano en mantos de carbón, gases y agua son:

- **Reflectancia de la vitrinita:** El promedio de la variable en esta área de estudio es de 0,8, lo cual no ayuda en la determinación de gas en mantos de carbón pues, es necesario tener un promedio inferior de 0,65 para hablar de generación de gas metano por medios biogénicos, estando en 0,8 puede encontrarse el gas en mezcla.

- **Total de carbono orgánico:** El promedio de la variable en el área de estudio es 80.41, lo anterior nos dice que las muestras poseen el suficiente TOC para generar metano, pero no me determinan el contenido del hidrocarburo en la muestra.
- **pH:** El promedio de la variable en el área de estudio es de 7.23, aclarando que la variabilidad del pH se presenta muy marcada en las tres zonas analizadas, es decir las aguas encontradas en la Zona A son más básicas o alcalinas debido a la pobreza de nutrientes disueltos y la abundancia de oxígeno en las mismas; en la Zona B las aguas son más ácidas debido al enriquecimiento de hidrógeno y nutrientes disueltos.
Con lo anterior se concluye que este parámetro no me determina la aparición del metano asociado a los mantos de carbón.
- **Ca meq/L:** El promedio del catión Ca es de 5,69 valor que cambia del manto 1 con 0,36meq/L hasta el manto 7 en la Esperanza con 16,44meq/L, este catión presenta cambios en toda la zona, por ende no es una variable que ayude en la predicción de gas metano en mantos de carbón.
- **SO₄ meq/L:** El promedio de este anión es de 25,42meq/L, teniendo en cuenta que es un anión que presenta cambio muy drástico ya que en el manto 5 tiene un valor de 6,49meq/L y en el manto 12 con 47,89meq/L, con estos valores la variable no es estable en la decisión de aparición de gas en mantos de carbón.
- **HCO₃ meq/L:** El promedio de este anión es de 4,38meq/L, donde en el manto 7 en la Esperanza presenta 0,40meq/L y en el manto 1 con 6,85meq/L, el comportamiento del anión no es estable en el estudio por ende no ayuda en la determinación de aparición de gas metano en los mantos de carbón.

CAPITULO IV

6. CONCLUSIONES

Es importante usar el análisis clúster de forma separada por análisis de roca, agua y gas , ya que su implementación grupal arroja resultados diferentes y no se puede asociar con congruencia las variables implicadas.

La implementación de análisis multivariado clúster método Ward, define grupos de variables según estudios de roca, agua y gases de manera efectiva, ya que asocia muy bien los comportamientos de las mismas orientando al investigador a definir cada variable como se deben comportan en presencia de gases a asociados a mantos de carbón.

El modelo de la red reconoció a la variable dependiente permitiendo identificar similitudes entre las muestras, a partir de las asociaciones encontradas en las variables de entrada. Con lo anterior se definen las variables que determinan y no determinan la aparición de gas metano en mantos de carbón.

Teniendo conocimiento de este grupo de variables y su comportamiento es más rápido y fácil generar conclusiones acerca de zonas de interés de exploración del hidrocarburo, llevando a definir mejor las áreas estudiadas.

7. BIBLIOGRAFÍA

Bosch,A., Rosales,W. Diferenciación química de aguas subterráneas mediante Análisis multivariante. Aplicación a los acuíferos de la sierra de Gádor (Almería). Departamento de Geodinámica, Facultad de Ciencias, Universidad de Granada. 1995

Helsel,C.,Hirsch,R. Statistical Methods in Water Resources. Charter A3U.S.Geological Survey. .2002.

Jain, A. K., Duin, P. W. & Mao, J. `Statistical pattern recognition: a review', IEEE Trans. Pattern Analysis Machine Intell 22(1), 4-372. 2000.

Kuncheva, L. Combining Pattern Classifiers, Wiley- Interscience.2004.

Pope,M.,Diaz,A. Quality-of-water data and statistical summary for selected coal-mined strip pits in Crawford and Cherokee counties, Southeastern Kansas".U.S.Geological Survey.1982

Rionda,A. El gas asociado a los Yacimientos de carbón mineral (GAC)".Academia de ingeniería, A.C. México.2010.

Rodriguez,M., Benavente,J, & Cruz, J. Análisis multivariable aplicado a datos hidrogeoquímicos de las zonas húmedas de la Provincia de Malaga".Instituto del agua. Universidad de Granada. Granada. 2000.

Manrique,L. Análisis estadístico de los factores de riesgo que influyen en la enfermedad angina de pecho. Facultad de ciencias Matematicas E.A.P de Estadística. Universidad Nacional Mayor de San Marcos. Perú. 2002.

Ubeda,J. Coalbed methane (CBM)=Gas metano del carbón (GMC).MFG OIL Y Gas S.A, Buenos Aires. 2008

Warwick, P., Clayton, B.Preliminary results from coal-bed methane drilling in Ouachita Parish, Louisiana. USGS Reston, VA 20192, Report -1239. 2004

Wayne, A., Van, V. Geochemical signature of formation water associated with coalbed methane. Montana Bureau of Mines and Geology, AAPP BULLETIN, V.87, NO 4, p.667-676. 2003