Semana Tec 1002S

Actividad 4 patrones con K-Means

13/01/2021

Herramientas computacionales: El arte de la analítica

Alison Magie Yáñez Dávila A01423011

Gilberto Huesca Juárez

Campus Cuernavaca

Instrucciones

En esta actividad encontrarás patrones de tus datos utilizando la técnica de clustering k-means. Trabajarás con el conjunto de datos que seleccionaste. Tienes que replicar los pasos vistos durante la clase.

Realiza un código que haga lo siguiente:

1. Carga tus datos
2. Selecciona dos variables que consideres interesantes para este análisis.
3. Determina un valor de k de acuerdo a los datos que tienes, las variables y una pregunta que quieres contestar con este análisis.
4. Utilizando scikitlearn calcula los centros del algoritmo k-means

Realiza un pequeño reporte que muestre lo siguiente. Justifica tus respuestas con base en los centros encontrados.

1. Describa el significado de las variables que seleccionaste.
2. ¿Crees que estos centros puedan ser representativos de los datos? ¿Por qué?
3. ¿Por qué seleccionaste ese valor de k a usar?
4. ¿Los centros serían más representativos si usaras un valor k más alto? ¿Más bajo? ¿En qué cambiaría la interpretación?
5. ¿Qué distancia tienen los centros entre sí? ¿Hay alguno que esté muy cercano a otros?
6. ¿Qué pasaría con los centros si tuviéramos muchos outliers en el análisis de cajas y bigotes?
7. ¿Qué puedes decir de los datos basándose en los centros?

Introducción

Los ríos son corrientes de agua que fluyen desde su nacimiento hacia otro río, lago o el mar. Los ríos son considerados una de las formaciones geográficas más importantes de la tierra, ya que son elementos vitales para todos los seres vivos no sólo por suministrar agua a ciudades o granjas, si no que también por que nos proporcionan alimentos e incluso nos permiten entretenernos y navegar, ya que se consideran poderosos agentes de transportación.

Sin embargo, hoy en día la contaminación de los ríos representa un gran problema a nivel mundial no sólo porque abre paso a enfermedades, sino que también porque la alteración de estos presenta un impacto en el mar, zonas costeras u otros ríos. Es por lo anterior que se han realizado estudios minuciosos sobre la calidad del agua.

A continuación, en el siguiente proyecto se mostrará un análisis estadístico sobre la calidad del agua en diferentes ríos de India, donde la investigación se basará en el promedio de los valores de calidad del agua obtenidos de algunos de los ríos de India.

Variables

- *Station code*
  Variable que representa un código único para cada lugar.
  El tipo de variable es Entero.

- *Locations*
  Localizaciones de los ríos. Nombre de los ríos y dónde se encuentran ubicados.
  El tipo de variable es String.

- *State*
  Estado en donde corre el río. Los estados que se encuentran son Maharashtra, Bihar, Uttar Pradesh, Karnataka y Adhra pradesh.
  El tipo de variable es String.

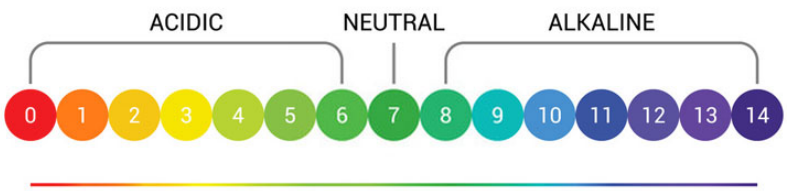
- *Temp*
  Temperatura promedio del río. El valor mínimo registrado es 0 mientras que el máximo es de 33.8°.
  El tipo de variable es Flotante.

- *DO*
  Valor promedio del oxígeno disuelto. Esta variable nos permite dar una idea sobre la contaminación en el agua, debido a que un contenido en oxígeno significativamente menor no permite el desarrollo de especies acuáticas.

[OD] mg/L	Condición	Consecuencias
0	Anoxia	Muerte masiva de organismos aerobios
0-5	Hipoxia	Desaparición de organismos y especies sensibles
5-8	Aceptable	**OD] adecuadas para la vida de la gran mayoría de**
8-12	Buena	**especies de peces y otros organismos acuáticos.**
>12	Sobresaturada	Sistemas en plena producción fotosintética.

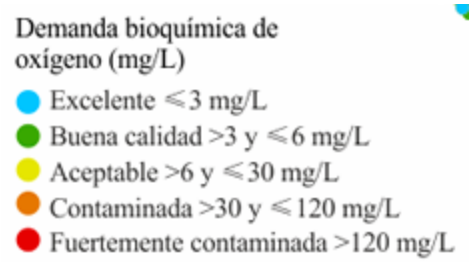
  El valor mínimo registrado es 0.2mg/l mientras que el máximo es 16.3mg/l.
  El tipo de variable es Flotante.

- *pH*
  Valor promedio del ph. El valor mínimo registrado es 6.3 mientras que el máximo es 14.7.
  <u>El tipo de variable es Flotante.</u>

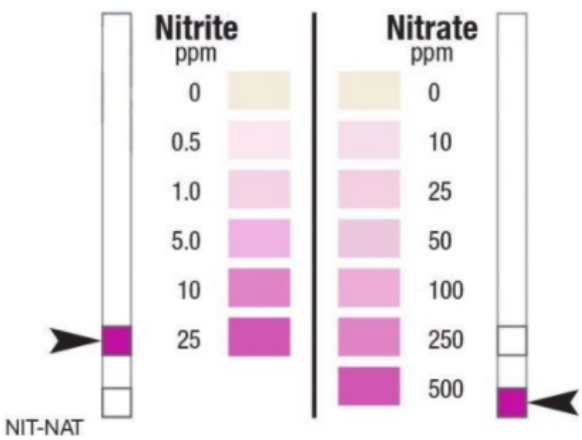


- Conductivity
  Valor promedio de la conductividad eléctrica. Esta variable nos permite saber si las aguas superficiales son potables. El valor mínimo registrado es 39µS/cm, mientras que el máximo es 24062 µS/cm.
  <u>El tipo de variable es Entero.</u>

- *BoD*
  Valor promedio de la demanda bioquímica de oxígeno. Este parámetro nos permite determinar el estado/ calidad del agua de los ríos.
  El valor mínimo es 0.2 mg/l mientras que el máximo es 75.6 mg/l.
  <u>El tipo de variable es Flotante.</u>



- *Nitrate_N_Nitrite_N*
  Valor promedio de nitrato-n y nitrito-n. Los nitratos y nitritos en aguas naturales también nos ayudan a determinar el estado/calidad del agua.
  Los niveles de nitrito superiores a 0,75 ppm pueden provocar estrés en peces, mayores a 5 ppm se considera tóxico para los seres



vivos.
Por otra parte, los niveles de nitrato entre 0 y 40 ppm son seguros, mientras que superior a 80 se considera tóxico.
Los rangos registrados en la dataset varían entre 0 ppm y 45.5ppm.
El tipo de variable es Flotante.

- *Fecal_Coliform*
  Valor promedio de coliformes fecales, es decir, de un subgrupo de bacterias de coliformes. Su presencia indica que el agua está contaminada por excrementos o desechos de alcantarillas.
  Los rangos registrados en la dataset varían entre 0 y 310417.
  El tipo de variable es Flotante.

- *Total_Coliform*
  Valor medio de coliformes. Estas son bacterias que usualmente son encontradas en el medio ambiente.
  Los rangos registrados en la dataset varían entre 1 y 23816667.
  El tipo de variable es Flotante.

Valoración	Agua	
	Coliformes fecales (NMP/100ml)	Coliformes totales (NMP/100ml)
No contaminado	0%-20% > 200	0%-20% > 1000
Contaminación Media	41% - 60 % > 200	41% - 60 % > 1000
Contaminación Alta	61% - 100 % > 200	61% - 100 % > 1000

Tabla 1. Valoración conceptual indicativa del grado de contaminación para los contaminantes microbiológicos. (Tomado y modificado de Marín, Garay *et al.*, 2001).

Análisis - gráfica 1 KMeans

Para el desarrollo del análisis nuevamente se decidió hacer uso de las variables BOD y Nitrate_N_Nitrite_N, que como se observó anteriormente estas variables se encuentran correlacionadas, lo que indica que la falta de presencia de estos elementos en los ríos representan el grado de contaminación en el agua.

- *Demanda Bioquímica de Oxígeno (BOD)*

  El BOD es la cantidad de oxígeno que los microorganismos, hongos y plancton consumen durante la degradación de sustancias orgánicas.

  Mientras más altos se encuentren los valores, más contaminado se encontrará el río.

- Nitrato y Nitrito (Nitrate_N_Nitrite_N)

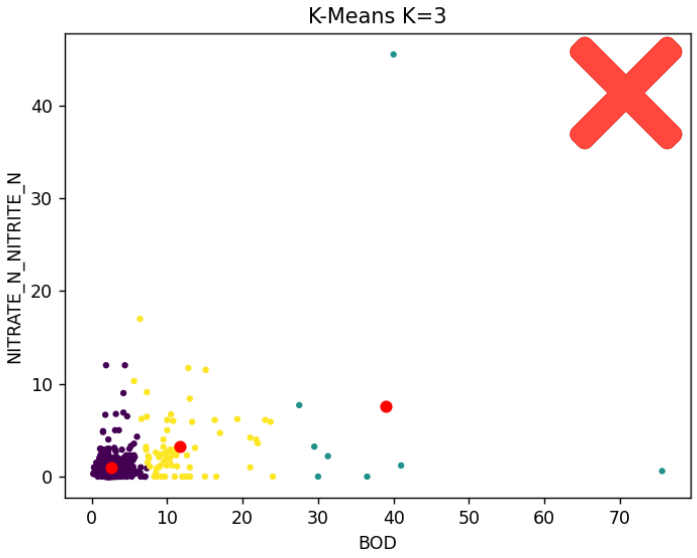
  El Nitrato y nitrito son compuestos solubles que contienen nitrógeno y oxígeno. El estándar del nitrato es 10 ppm sin embargo hasta 40 ppm es considerado seguro, por otra parte el estándar del nitrito es 1 ppm.

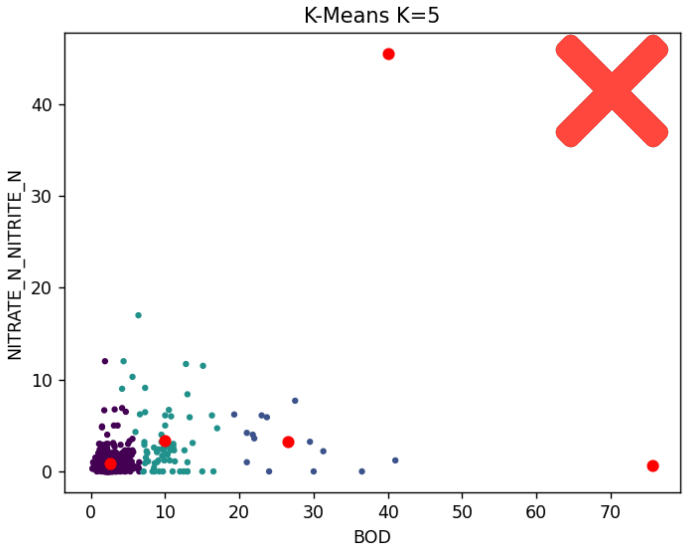
  Un alto contenido de nitrato y nitrito pueden causar problemas de salud crónicos como enfermedades del corazón o pulmones.

Valor de k

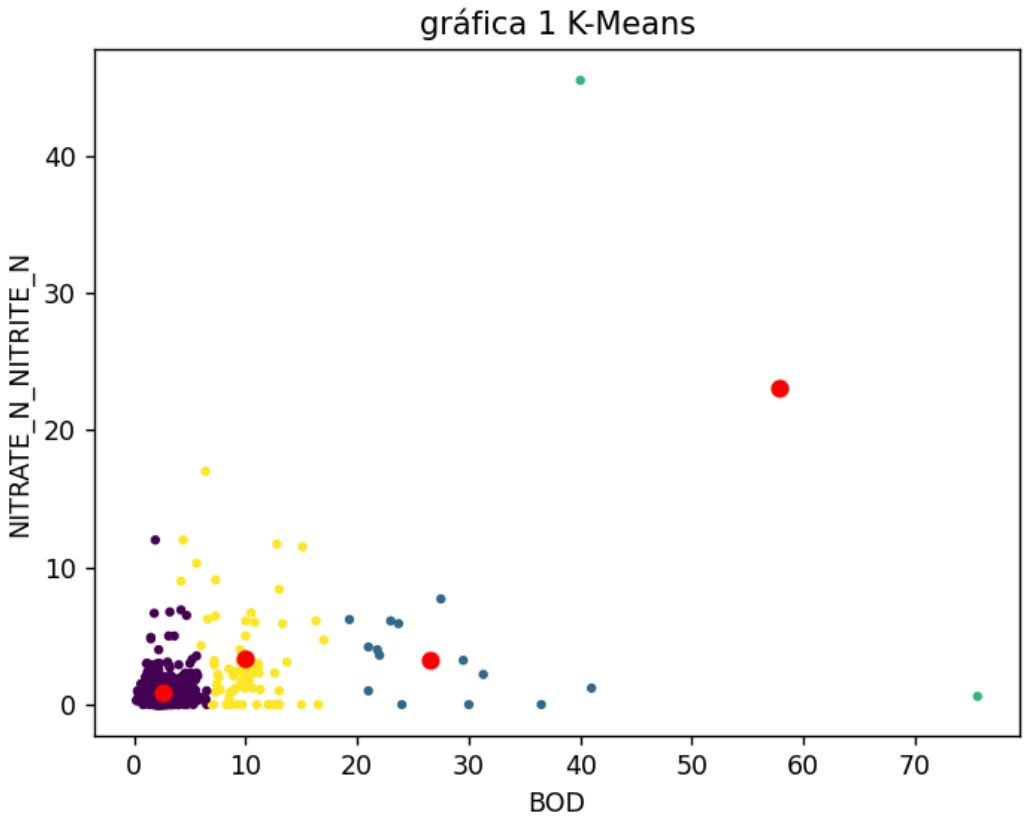
Para el desarrollo de la “gráfica 1 K-Means” se hizo uso del valor de k=4 debido a que de esta forma se representaría de mejor forma los rangos de seguridad establecidos para la detección de los componentes químicos que analizaremos, qué son nitratos y demanda de oxígeno.

Cabe recalcar que el valor de k fue seleccionado tras varias pruebas realizadas en el código. Se observó que este valor al ser menor o mayor a 4 no permitiría una mejor visualización de los datos debido a que se podría interpretar que los rangos de seguridad son diferentes a los que se establecen en las instituciones relacionadas al medio ambiente, lo que puede generar confusión en el usuario:





Gráfica 1 K-Means



Tal como se puede observar en la gráfica generada, los valores de Nitrato y nitrito en su gran mayoría oscilan entre 0 y 12 ppm, mientras que los valores de demanda de oxígeno lo hacen entre 0 y 16.5mg/L a excepción de algunos datos que alcanzan a llegar hasta un límite de 45 ppm en nitratos y 75 mg/L en demanda de oxígeno.

En cuanto a los centros generados, 3 de ellos se encuentran muy cercanos entre sí, que son los puntos 1,4 y 3 , que tienen una distancia de:

Cluster:

- Punto 1 (morado): (2.54547945, 0.88235616)
- Punto 2 (amarillo): (26.54285714 , 3.23785714)
- Punto 3 (azul): (57.8 ,23.05)
- Punto 4 (verde): (9.9530303 ,3.31772727)

Distancias:

- Punto 1 a 4: 7.79
- Punto 4 a 3: 16.59

Este comportamiento resulta ser beneficioso para el análisis debido a que significa que los datos dentro de la dataset no se encuentran muy dispersos, lo que nos puede ayudar a establecer mejores análisis al relacionarlo con los rangos de seguridad. No obstante, en caso de que existieran muchos outliers dentro del análisis de cajas y bigotes, en los centros dentro de la “gráfica 1 K-Means” y sus valores alrededor de él existiría mucha variabilidad.

Conclusiones

Finalmente, haciendo un análisis del significado de los resultados obtenidos en la “gráfica 1 K-Means” basándonos en lo que la OMS define como agua contaminada que refiere a aquella que sufre cambios en su composición donde comúnmente se involucran sustancias como fertilizantes, pesticidas, nitratos, desechos fecales etc. Lo que se puede concluir, es que dentro de la gráfica se indica que la gran mayoría de ríos en India tienen una cantidad de demanda de oxígeno y nitratos en cantidades menores, que quiere decir que gran parte de los ríos se encuentran en un rango de contaminación menor y no tan dañino para la salud, no obstante, también se indica que existe una cantidad considerable de ríos que se encuentran contaminados y un porcentaje pequeño de ríos altamente contaminados.

Referencias

- Utcarsh, A. (2020, 29 mayo). *Water Quality Data*. Kaggle. Recuperado 12 de enero de 2022, de https://www.kaggle.com/utcarshagrawal/water-quality-data/version/1
- *Río - NILSA*. (s. f.). nilsa. Recuperado 12 de enero de 2022, de https://www.nilsa.com/es/como-educamos/oferta-educativa/rio/
- Valdivielso, A. (2020, 19 octubre). *¿Qué es un río?* iAgua. Recuperado 12 de enero de 2022, de https://www.iagua.es/respuestas/que-es-rio
- *analisis_aguas*. (s. f.). . Recuperado 12 de enero de 2022, de http://ficus.pntic.mec.es/ngom0007/analisis_aguas.html
- Andreo, M. (s. f.). *Demanda Biológica de Oxígeno (D.B.O.)*. . Recuperado 12 de enero de 2022, de https://www.mendoza.conicet.gov.ar/portal/enciclopedia/terminos/DBO.htm#:%7E:text=Se%20expresa%20en%20mg%20%2F%20l,microorganismos%20para%20oxidarla%20(degradarla).
- N. (s. f.). *DBO y DQO para caracterizar aguas residuales*. Nihon Kasetsu Europe | Monitoring & Water Clarification. Recuperado 12 de enero de 2022, de https://nihonkasetsu.com/es/dbo-y-dqo-para-caracterizar-aguas-residuales/
- *El medio ambiente en México*. (2014). semarnat.gob. Recuperado 12 de enero de 2022, de https://apps1.semarnat.gob.mx:8443/dgeia/informe_resumen14/06_agua/6_2_1.html
- *nitratos y nitritos-lenntech*. (s. f.). Lenntech. Recuperado 12 de enero de 2022, de https://www.lenntech.es/nitratos-y-nitritos.htm
- Dirección de Salud Pública de Carolina del Norte. (s. f.). *https://epi.dph.ncdhhs.gov/oee/docs/Las_Bacterias_Coliformes_WellWaterFactSt.pdf*. . Recuperado 12 de enero de 2022, de https://epi.dph.ncdhhs.gov/oee/docs/Las_Bacterias_Coliformes_WellWaterFactSt.pdf
- *Oxígeno Disuelto*. (s. f.). . Recuperado 12 de enero de 2022, de https://www.ucm.es/data/cont/docs/952-2015-02-14-Oxigeno%20disuelto%20f.pdf
- *Nitrato y Nitrito*. (s. f.). . Recuperado 12 de enero de 2022, de http://region8water.colostate.edu/PDFs/we_espanol/Nitrate%202012-11-15-SP.pdf