RESEARCH

Evaluación diagnóstica diabética con aprendizaje automático

Joel Alejandro Rodarte Rivera

Abstract

The Abstract should not exceed 350 words. Please minimize the use of abbreviations and do not cite references in the abstract.

Background: the context and purpose of the study

Methods: how the study was performed and statistical tests used

Results: the main findings

Conclusions: brief summary and potential implications

Keywords: sample; article; author

Introducción

Desde su acuñamiento a mediados de siglo XX el aprendizaje automático , rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender y mejorar su desempeño en una tarea específica a medida que se les proporciona más datos, ha evolucionado para solucionar problemas del mundo real en diferentes áreas. La medicina es una de las que se ha visto ampliamente beneficiada para complementar el esfuerzo de los trabajadores de la salud. El aprendizaje automático puede ser utilizado en la medicina para explorar datos, imágenes o sonidos para así encontrar patrones difíciles de identificar. Algunos ejemplos son el análisis de imágenes radiográficas para detectar tumores o evaluar la presencia o ausencia de una enfermedad dado unos síntomas, para así tomar decisiones más acertadas y atacar la enfermedad con rapidez o evitar cometer un error clínico.

Justamente este es el objetivo del escrito, el cual busca determinar si un paciente, dado ciertos síntomas, es diagnosticado con diabetes. El alcance de este texto se limita a determinar si existe o no diabetes, ya que por el momento determinar si es diabetes tipo I o II no se realizará. El conjunto de datos cuenta con 4,920 pacientes para los cuales, para cada uno de ellos se tiene documentado los síntomas que presentó y el dictamen de la enfermedad que se le diagnosticó. En total, cada paciente cuenta con 132 posibles síntomas los cuáles pueden derivar en 42 enfermedades. Para el presente, los esfuerzos están dedicados a predecir diabetes pero la metodología y base de datos puede ser utilizada para cada una de las enfermedades. El conjunto de datos cuenta únicamente con información binaria, es decir ceros y unos siendo que cuando algo sea verdad es decir, que existe un síntoma o enfermedad se le coloca un uno. Todas las variables están en inglés; como comentario prognosis es una manera de llamar a una enfermedad. El grupo de datos fue encontrado en la base de datos pública de fuente abierta en Kaggle [REFERENCIA]. Se investigó pero no se indica su procedencia o autoría por lo que los resultados de este escrito deben ser tomados como un ejercicio académico.

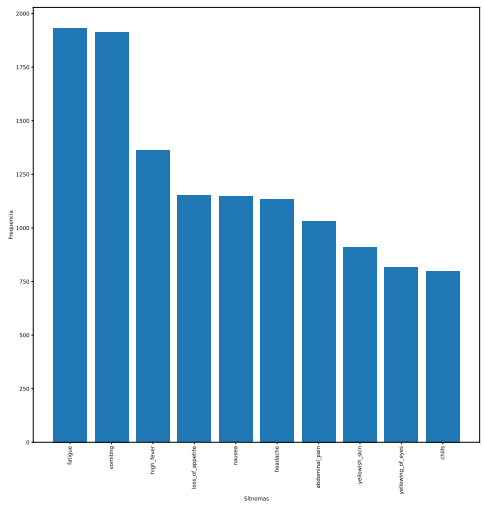
Otro comentario importante por el cuál esta publicación debe ser tomada sólo como un ejercicio académico es que muy avanzada el desarrollo del mismo, y dado que los métricos de desempeño de los algoritmos eran perfectos, se comenzó a sospechar que había riesgo de sobreajuste dado que algo en el conjunto de datos estaba fuera de lugar. Investigando en los foros de la base de datos se encontró que efectivamente existía algo que podía estar generando dichos resultados. De los 4,920 registros solo se cuenta con 304 registro únicos. Esto lo que genera es que los datos sean muy homogéneos y la dispersión de los mismos no sea suficiente para generar buenos modelos de aprendizaje automático para nuevos datos. Esto también se ve reflejado en las siguientes observaciones en donde pareciera ser que todos los síntomas tienen frecuencias de aparición similares. A continuación un pequeño análisis exploratorio de los datos para entender qué es lo que contiene.

De entre los 4,920 registros, la siguiente tabla de frecuencias muestra los diez síntomas más comunes. La intención de tener en la mira a estos síntomas es que son los que menos información podrían aportar a la predicción de diabetes o en general de cualquier

Correspondence: joel.rodarter@uanl.edu.mx
Facultad de Ciencias Fisicomatemática,UANL, Monterrey, Nuevo Léon, México
Full list of author information is available at the end of the article

enfermedad dado que son tan comunes es decir, que están presentes en múltiples enfermedades que incluirlos en los algoritmos podrían llevar a tomar decisión incorrectas por confusión de síntomas. Los 10 síntomas más común de mayor a menor fueron fatiga, vómito, fiebre alta, pérdida de apetito, náusea, dolor de cabeza, dolor abdominal, piel amarillenta, ojos amarillentos y escalofríos.

Figure 1 Frecuencias de síntomas más comunes entre los 4920 pacientes.



Ahora que se sabe cuales son los síntomas más comunes considerando todas las enfermedades, también es importante entender cuáles son los más comunes y la frecuencia del mismo para los pacientes con diabetes. Se sabe que el conjunto de datos de entrenamiento cuenta con 120 pacientes diabéticos. A continuación los diez síntomas más comunes. Son justamente estos síntomas a los que hay que prestar atención cuando se quiere predecir diabetes.

increasedappetite	120
polyuria	120
weightloss	114
$irregularsugar_level$	114
blurredanddistortedvision	114
fatigue	114
lethargy	114
restlessness	114
excessivehunger	114
obesity	114

Por último es importante entender en qué proporción, por cada variable de síntomas de diabetes, se encuentra también en otra enfermedad. La siguiente tabla tiene como objetivo explicar cuantas veces se presentan los síntomas en el conjunto completo y cuantas veces se presenta considerando solo los pacientes con diabetes. Si la proporción es 1, significa que únicamente los pacientes con diabetes son los que cuentan con dicho síntoma. A medida que baja la proporción significa que el síntoma a pesar de ser común para la diabetes, también está presente en algunas otras enfermedades.

Síntoma	*Frec.Total*	*Frec.diabetes*	*Proporción*
increasedappetite	120	120	1.00
polyuria	120	120	1.00
weightloss	456	114	0.25
irregularsugarlevel	114	114	1.00
blurredvision	342	114	0.33
fatigue	1932	114	0.05
lethargy	456	114	0.25
restlessness	228	114	0.50
$excessive_hunger$	462	114	0.24
obesity	228	114	0.50

Sin realizar ningún algoritmo de aprendizaje automático, se comienza a sospechar que las variables que serán clave para distinguir si un paciente tiene diabetes o no son apetito potenciado, poliuria y niveles irregulares de azúcar ya que cuenta con una proporción de 1. El resto de las variables tienen proporciones menores a .5 por lo cual quiere decir que también aparecen en otras enfermedades, sin embargo no deben ser descartados.

Métodos

El presente trabajo utilizará las siguientes técnicas de selección de características, aprendizaje no supervisado, aprendizaje supervisado y clasificación para cumplir con el objetivo del mismo.

Selección de características

Sabiendo que se cuenta con 132 variables de síntomas, y que probablemente no todas sean relevantes para predecir la diabetes. El área médica regularmente usa técnicas basadas en filtros con las pruebas t o Anova, métodos de envoltura o incrustación. La forma de determinar cuál utilizar depende de la naturaleza de los datos. En este caso se tienen datos binarios para los cuáles se podría utilizar un método de selección lineal y otro no lineal. En este caso se utilizaron selección de características basadas en correlación y selección de características basada en información mutua.

- Selección de características basadas en correlación

Para esta primera técnica se requerirá comprender la interacción lineal entre variables. Para ello se utiliza la matríz de correlación con lo siguientes pasos:

1. Calcular la matriz de correlación: se calcula la matriz de correlación entre las características seleccionadas. La matriz de correlación muestra la fuerza y dirección de la relación entre cada par de características.
2. Identificar características altamente correlacionadas: se identifican las características que están altamente correlacionadas entre sí. La alta correlación entre características significa que una de ellas puede ser redundante y se puede eliminar sin afectar significativamente la precisión del modelo.
3. Eliminar características redundantes: las características que se identificaron como redundantes se eliminan del conjunto de características. Esto se hace para reducir la dimensionalidad del conjunto de datos
4. Evaluar el modelo: después de la selección de características, se evalúa el modelo resultante para determinar si ha mejorado la precisión del modelo.

- Selección de características basada en información mutua

El objetivo de esta técnica es comprender el nivel de interacción de dos o maás variables en conjunto y relacionarla con la variable de respuesta, contraria a la téncica anterior la cuál sólo consideraba una variable.

1. Calcular la información mutua: se calcula la información mutua entre cada característica y la variable objetivo. La información mutua mide la dependencia entre dos variables y se utiliza para evaluar la importancia de cada característica.
2. Seleccionar características más importantes: se seleccionan las características que tienen la mayor información mutua con la variable objetivo. Estas características se consideran más importantes y se utilizan para construir el modelo.
3. Eliminar características redundantes: las características que están altamente correlacionadas entre sí se consideran redundantes y se eliminan del conjunto de características. Esto se hace para reducir la dimensionalidad del conjunto de datos y mejorar la precisión del modelo.
4. Evaluar el modelo: después de la selección de características, se evalúa el modelo resultante para determinar si ha mejorado la precisión del modelo. Esto se hace comparando la precisión del modelo antes y después de la selección de características.

Aprendizaje No supervisado

Durante el desarrollo de este escrito se utilizó el algoritmo de K Medias para la exploración de patrones subyacentes que sean difíciles de encontrar con estadística básica. Sin embargo, a medida que el alumno avanzó con el entendimiento del mismo, se comprendió que el algoritmo no era tan adecuado para el conjunto de datos, dado que los algoritmos no supervisados funciona mejor con información no categorizada y el conjunto de datos tienen etiquetas de unos y ceros, los cuáles ya son unas etiquetas implícitas. Además el hecho que solo se cuente con información binaria hace que la representación de grupos no sea muy valiosa dado las representaciones gráficas solo entregan 4 posibles localizaciones del centroides. En los resultados se explica más a detalle el punto anterior y el porqué este algoritmo junto con DBSCAN quedó fuera de los planes para esta aplicación. Esto sirvió como lección aprendida para entender que no siempre se debe correr algoritmos por el hecho de correos, se tiene que comprender el contexto de los mismos para ver si van a funcionar o si funcionan, si entregan información de valor de vuelta. Por esta razón la metodología de estos algoritmos será omitida.

Aprendizaje Supervisado

- Regresión Logística

La regresión logística es una algoritmos supervisado común para problemas que involucra con clasificación binaria. Se basa en modelar la probabilidad de que una variable de respuesta tome el valor de 1 en función a variables predictoras. Se dice que tome el valor de 1 dado que la variable de respuesta sólo puede ser 0 o 1. Aplicado al conjunto de datos, 1 representaría la presencia de una enfermedad mientras que 0 sería la ausencia de la misma. El algoritmo calcula la suma ponderada de las variables predictoras para producir una calificación. Esta calificación después se evalúa con una función logística para obtener la probabilidad estimada para que la variables de salida tome el valor de 1. Si la probabilidad se encuentra entre 0 y .5 la variable de respuesta se estima sea 0 (ausencia de enfermedad). De lo contraria una variable con probabilidad entre .5 y 1 haría que la variable respuesta tome el valor de 1 (presencia de enfermedad) La función logística tiene una forma de S, como la figura 2 lo cual quiere decir que la respuesta de la función es 0 cuando el valor de entrada es negativo y es 1 cuando el valor de entrada es positivo. La función logística se asegura que las probabilidades siempre están entre 0 y 1 por lo que se presta a utilizarla en problemas de clasificación binaria. Una vez que se entrena el modelo, se puede utilizar para predecir la probabilidad que la variable de salida tome el valor de 1. Una vez que se obtiene dicha probabilidad cualquier valor por encima de .5 se le clasifica como 1, presencia de enfermedad y 0 cuando es menor de .5, o ausencia de enfermedad.

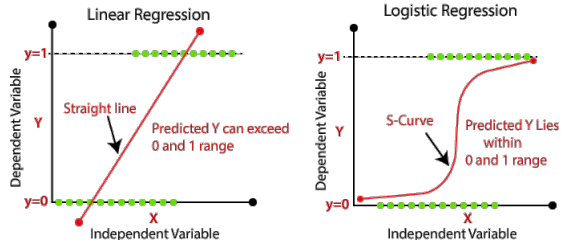
- Bayes Ingenuo

Algoritmos Clasificación

- Árboles de desición

Los árboles de decisión es un método de clasificación que se asemeja a un árbol, en donde en la parte superior, la punta del árbol, también llamada nodo raíz se

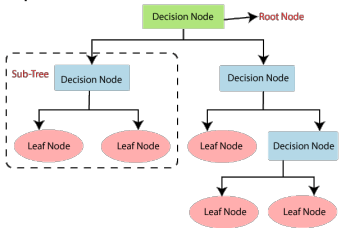
Figure 2 Regresión Linear vs Regresión Logística.



divide en ramas y hojas. Las ramas se dividen en más ramas o hojas, es decir que el árbol se hace más grande. Mientras que las hojas ya no cuentan con otra división disponible, el árbol ha terminado y se ha clasificado algo. El objetivo principal de los árboles es llegar a un punto en donde ya no se divida en más ramas y llegar a una hoja clasificatoria. Una vez que se llega ahí el algoritmo ha terminado.

Un término importante para comprender cómo iniciar el árbol y que servirá como raíz de la misma es utilizando la impureza de sus hojas, la pregunta que ayude a determinar si existe una enfermedad y que tenga la menor impureza servirá para ser la siguiente rama. La impureza de una hoja se refiere a la mezcla o variedad de las clases en la misma. Es decir, en una hoja del árbol, la impureza mide cuánto las instancias de diferentes clases están mezcladas. Cuanto menor sea la impureza de la hoja, más pura será, lo que significa que contiene principalmente instancias de una sola clase. La impureza se utiliza en la selección de la mejor división en un árbol de decisión, ya que el objetivo es minimizar la impureza de las hojas. Hay varios métodos para medir la impureza de una hoja; este escrito utilizará la impureza de gini la cual mide la probabilidad de clasificar erróneamente una instancia de forma aleatoria. Cuanto mayor sea la probabilidad, mayor será la impureza de la hoja. Lo deseado es tener la menor impureza posible cuanto antes para lograr hacer los algoritmos de clasificación veloz. Un árbol de decisión se verá como la figura 3 a continuación. Para esta aplicación dado que se está prediciendo categorías, el árbol generado será un árbol de clasificación.

Figure 3 Ejemplo árboles de desición



Métricos de Desempeño

Results

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Massa tempor nec feugiat nisl pretium fusce id. Lorem sed risus ultricies tristique nulla. Nibh tortor id aliquet lectus proin nibh nisl condimentum id. Ornare arcu dui vivamus arcu felis bibendum ut tristique et. Scelerisque viverra mauris in aliquam sem fringilla ut. Molestie at elementum eu facilisis sed. Diam ut venenatis tellus in metus vulputate eu. Nibh venenatis cras sed felis eget velit aliquet. Egestas tellus rutrum tellus pellentesque eu tincidunt tortor. Nunc sed id semper risus in. Non tellus orci ac auctor augue mauris augue neque gravida. Libero enim sed faucibus turpis in eu mi bibendum neque. Id ornare arcu odio ut sem nulla pharetra. Ultrices tincidunt arcu non sodales neque sodales ut etiam.

Discussion

Algoritmos no supervisados

IMPORTANTE: Por el momento, considero que la explicación del modelo No me queda muy claro dado existen conceptos de probabilidad con los que no estoy familiarizado. Sin emabargo, me es importante al menos saber de su existencia para dominarlo más adelante en el transcurso del curso.

Akaike Information CriterioBenroulli Mixture Model

¿Qué es?

- La distribución Bernoulli es una distribución discreta que predice la probablidad de éxito (1) o fracaso (0) de un evento. El término "mixture" se incluye dado que este método combina multiples distribuciónes de probabilidad para modelar una distribucón más compleja.
- Respecto al punto anterior, este algoritmo asumme que la información se genera de K mezclas de distribuciónes de Bernoulli, en donde cada componente de la mezcla representa a un cluster.
- Cada cluster es caracterizado por un vector de probabilidades que predice la probabilidad de encontrar un 1 en ese cluster. La distribución de la mezcla general es entonces una suma ponderada de las distribuciones de K Bernoulli, donde los pesos especifican la proporción de los datos que pertenecen a cada grupo. (?)
- Este último punto es el que no me queda muy claro.

Clustering para datos binarios

Akaike Information Criterion

El Akaike Information Criterion es un un solo numero que se utiliza para determinar cual dentro de múltiples opciónes de modelos mejor se ajusta para un set de datos. Este numero se puede utilizar en combinación con el Bernoulli Mixture Model para determinar la cantidad de cluster óptima.

Tarea 5

Se determina que para los 4920 pacientes los síntomas más comunes son los siguientes con su respectiva incidencia [1].

Table 1 Sample table title. This is where the description of the table should go.

fatigue	0.393
vomiting	0.389
high fever	0.277
loss of appetite	0.234
nausea	0.233

Kmeans

Se utilizó los 2 síntomas más comunes (fatiga y vómito) para ver la combinación de su aparición en todos los pacientes. Dado se sabe que sólo hay 4 combinaciones posibles de resultados para x enfermedad, se utilizó el algoritmo no supervisado de k medias con 4 clusters.

Los 4 resultados posibles son

- Que paciente no tenga ninguno de los dos síntomas
- Que paciente tenga fatiga pero no vómito
- Que paciente tenga vómito pero no fatiga
- Que paciente tenga ambos síntomas

Los centroides de los 4 clusters son los siguientes:

	x	y
0	0.000	0.000
1	1.000	0.000
2	-0.000	1.000
3	1.000	1.000

La cantidad de pacientes por cada centroide es la siguiente

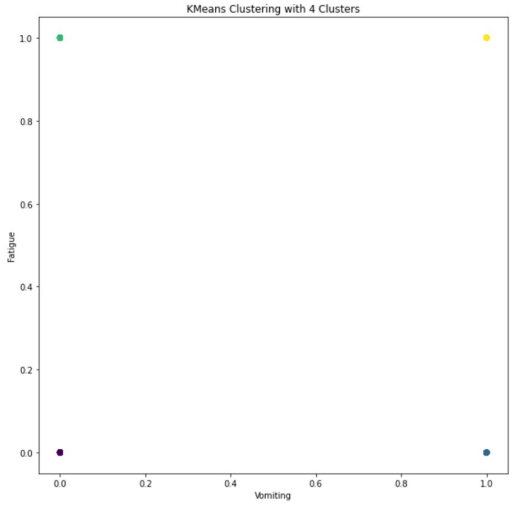
	ck	nk
0	0	1836
1	1	1170
2	2	1152
3	3	762

Interpretación

- Se tienen 1826 pacientes que para alguna enfermedad NO tienen ni vomito ni fatiga
- Se tienen 1152 pacientes que para alguna enfermedad NO tienen fatiga pero SI vómito
- Se tienen 1170 pacientes que para alguna enfermedad SI tienen fatiga pero NO vómito
- Que QSe tienen 762 pacientes que para alguna enfermedad SI tienen fatiga y SI vómito

Dado se cuenta con información binaria se espera que el plot de k medias sea un cuadrado por que sólo hay 4 posibles resultados.

Figure 4 Sample figure title. Figure legend text.



Selección número de clusters adecuado

Por el método del codo se calcula la inercia para entender que sucedería con la dispersión de los datos considerando 2 a 8 clusters.

	nclusters	inertia
0	2	584.658
1	3	152.878
2	4	0.000
3	5	0.000
4	6	0.000
5	7	0.000
6	8	0.000

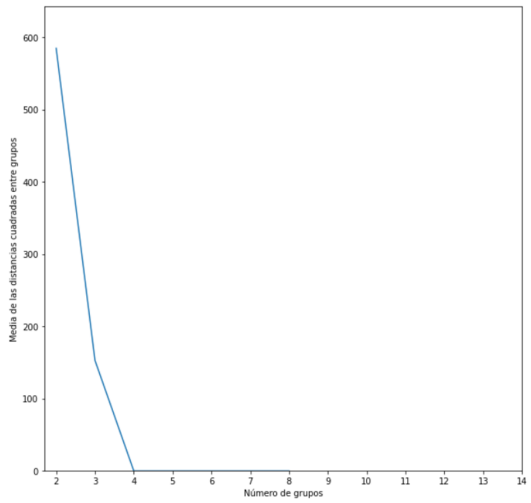
Dada la información es binaria se sabe que debería ser suficiente con 4, pero se quiere entender como cambiaría añadiendo hasta 4 más, se espera que la diferencia sea pequeña o nula despues de cuatro. Además, se espera que sea muy grande menos de 4.

Tarea 6

¿Qué es el la regresión logistica y porqué funciona con data frames de datos binarios?

La regresión logística es una algoritmos supervisado común para problemas que involucracon clasificiación binaria. Se basa en modelar la probabilidad de que una variable de respuesta tome el valor de 1 en función a variables predictoras. Se dice que tome el valor de 1 dado que la variable de respuesta solo puede ser 0 o 1. Aplicado al set de datos trabajado este tetramestre

Figure 5 Sample figure title. Figure legend text.



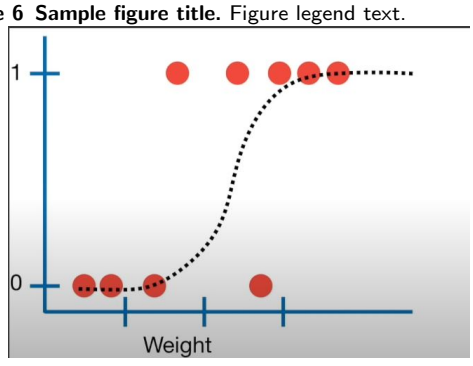
1 representaría la presencia de una enfermedad mientaras que 0 sería la ausencia de la misma.

El algoritmo calcula la suma ponderada de las variables predictoras para producir un calificación. Esta calificación después se evalua con una función logistica para obtener la probabilidad estimada para que la varibales de salida tome el valor de 1.

La función logística tiene una forma de S, como la siguiente imágen lo cual quiere decir que la respuesta de la función es 0 cuando el valor de entrada es negativo y es 1 cuando el valor de entrada es positivo. La función logisitica se asegura que las probabilidades siempre esten entre 0 y 1 por lo que se presta a utilizarla en problemas de clasificación binaria.

Una vez que se entrena el modelo, se puede utilizar para predecir la probabilidad que la variable de salida tome el valor de 1. Una vez que se obtiene dicha probabilidad cualquier valor por encima de .5 se le clasifica como 1, presencia de enfermedad y 0 cuando es menor de .5, o ausencia de enfermedad.

Figure 6 Sample figure title. Figure legend text.



Métricas para analizar performance Logistic Regression

Para evaluar el desempeño de este algoritmo se utiliza técnas como accuracy y ROC (Receiver Operating Characteristic). Accuracy no es más que la division de los casos correctamente predecidos entre los casos totales. ROC es una manera gráfica de interpretar una matriz de confusión para diferenciar de manera gráfica cuantos faltos positivos y verdaderos positivos hay. Más adelante en esta tarea se muestra como realizarlo en python.

Estrategia para tarea 05

Se cuenta con un set de datos de 4920 pacientes. Voy a dividir esos datos en 70

Tarea 8 - Métricas de Desempeño

Se ha particionado los datos en un set de entrenamiento y prueba para poder evaluar el desempeño de los modelos de arbol de desición y regresión logística.

Para ello es importante definir como se calculan los métricos de desempeño y cómo se interpretan:

- Regresión Logística

  Matriz Confusión: Matriz dos por dos que muestra la cantidad de los cuatro escenarios posible de predicción. Verdaderos positivos, verdaderos falsos, falsos positivos y falsos negativos.

  Curva ROC:
- Árboles de desición

  Accuracy:Proporción de clasificaciónes positivas que fueron correctas

$$(VP + VN)/(VP + VN + FP + FN) \tag{1}$$

  Precision: proporción de clasificaciones verdaderas positivas entre las clasificaciones verdaderas

$$VP/(VP + FP) \tag{2}$$

  Recall: proporción de verdaderos positivos dados todos los positivos.

$$VN/(VN + FP) \tag{3}$$

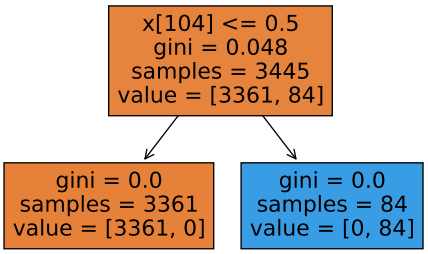
En el algoritmo de regresión logística se obtuvo la siguiente matríz de confusión

1439 0
0 36

En el algoritmo de árboles de desicion se obtuvieron lo siguientes resultados:

- Accuracy: 1.0
- Precision: 1.0
- Recall: 1.0
- F1-score: 1.0 []

Figure 7 Sample figure title. Figure legend text.



Conclusion

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Massa tempor nec feugiat nisl pretium fusce id. Lorem sed risus ultricies tristique nulla. Nibh tortor id aliquet lectus proin nibh nisl condimentum id. Ornare arcu dui vivamus arcu felis bibendum ut tristique et. Scelerisque viverra mauris in aliquam sem fringilla ut. Molestie at elementum eu facilisis sed. Diam ut venenatis tellus in metus vulputate eu. Nibh venenatis cras sed felis eget velit aliquet. Egestas tellus rutrum tellus pellentesque eu tincidunt tortor. Nunc sed id semper risus in. Non tellus orci ac auctor augue mauris augue neque gravida. Libero enim sed faucibus turpis in eu mi bibendum neque. Id ornare arcu odio ut sem nulla pharetra. Ultrices tincidunt arcu non sodales neque sodales ut etiam.

List of abbreviations

If abbreviations are used in the text they should be defined in the text at first use, and a list of abbreviations should be provided.

Declarations

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

JRR was a major contributor in writing the manuscript.

Acknowledgements

Universidad Autónoma de Nuevo León

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

References

1. Hussain, M., Cifci, M.A., Sehar, T., Nabi, S., Cheikhrouhou, O., Maqsood, H., Ibrahim, M., Mohammad, F.: Machine learning based efficient prediction of positive cases of waterborne diseases. Statquest (2023). doi:www.youtube.com

Figures

Figure 8 Sample figure title. A short description of the figure content should go here.

Figure 9 Sample figure title. Figure legend text.

Tables

Table 2 Sample table title. This is where the description of the table should go.

	B1	B2	B3
A1	0.1	0.2	0.3
A2	...	..	.
A3	..	.	.

Additional Files

Additional file 1 — Sample additional file title

Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title

Additional file descriptions text.