Encuentro Internacional de Educación en Ingeniería ACOFI
Una formación de calidad en ingeniería para el futuro
Centro de Convenciones Cartagena de Indias
15 al 18 de Septiembre de 2015

ANÁLISIS DE TÉCNICAS DE MD EN DIAGNÓSTICO DE ENFERMEDADES CARDIOVASCULARES

Jhon Harol Campo Mendoza, Karen Dayana Parra García, Fabio Mendoza Palechor, Alexis De La Hoz Manotas

Universidad de la Costa
Barranquilla, Colombia

Resumen

Las enfermedades cardiovasculares representan una amenaza real para los sistemas de salud de muchos países, debido a que se han convertido en uno de los diagnósticos que cobra un número significativo de vidas en el mundo entero. De acuerdo a los datos emitidos por la Organización Mundial de la Salud (OMS) las enfermedades cardiovasculares son una causa importante de muertes, se estima que 9.4 millones y medio de muertes, es decir, el 16,5% de las muertes anuales, son atribuibles a la hipertensión únicamente. Esto incluye el 51% de las muertes por accidentes cardiovasculares cerebrales (AVC) y el 45% de las muertes por cardiopatía coronaria. De acuerdo a lo anteriormente mencionado, el análisis de este tipo de enfermedades se ha convertido en un factor común de investigación, la aplicación de sistemas informáticos inteligentes brindan la posibilidad de identificar de forma anticipada los pacientes que puedan padecer dicha enfermedad, por lo cual se propone en esta investigación la utilización de distintas técnicas de minería de datos como lo son árboles de decisión, las máquinas de soporte vectorial, la regresión logística, el método de naivebayes, IBk (k vecinos más cercanos) y redes neuronales, implementados utilizando un mismo conjunto de datos "Heart Disease Data Set" alojado en el repositorio Machine Learning UCI y bajo un mismo ambiente de prueba, con la finalidad de establecer cuál de las técnicas antes mencionadas logran un mayor porcentaje de precisión a la hora de identificar pacientes que padezcan la enfermedad objeto de estudio; para la realización de las pruebas se utilizó validación cruzada con el fin de seleccionar un porcentaje del conjunto de datos para realizarlas y otro para entrenamiento. Las técnicas que lograron mejores resultados fueron RegresionLogistica y NaiveBayes las cuales alcanzaron un 84% de precisión, las técnicas de Redes Neuronales, IBK, Máquinas de Soporte Vectorial y Arboles de Decisión obtuvieron porcentajes de precisión inferiores lo cual indica que su desempeño no es el más adecuado para la identificación de este tipo de enfermedad.

Palabras clave: enfermedades cardiovasculares; minería de datos; precisión

Abstract

Cardiovascular diseases have become a real threat for health systems in many countries, since it´s one of the diagnosis with bigger toll deaths around the world. OMS data indicate that cardiovascular diseases are one of the most important causes of death, estimated in more than 9,4 million casualties, represented in 16,5% of annual human losses by hypertension only. This includes 51% of deaths due cerebrovascular accidents (CVA) and 45% due coronary cardiopathy Given this situation, the analysis of these type of diseases has become a popular subject for research, and the use of intelligent computational systems enables the prediction of these conditions in patients, and this is the reason why we propose the use of different data mining techniques as decision trees, vectorial support machines, logistic regression, Naïve Bayes, IBk (k nearest neighbor) and neural networks, implemented in one single data set "Heart Disease Data Set" located in the Machine Learning UCL repository, using the same test environment, looking for the best technique in precision to identify patients that can show the disease subject to study; for testing we used crossed validation to select the samples from the data set and for training. The techniques with best results were Logistic Regression and Naïve Bayes each with 84% of precision, neural networks, IBK,Vectorial Support Machines and Decision Trees obtained lower precision percentages which rules them out for identificaction of this type of disease.

Keywords*: cardiovascular diseases; data mining; precision*

1. Introducción

Las enfermedades del corazón han sido la principal causa de muerte en el mundo durante los últimos diez años según la organización mundial de la salud.

Según HellerChinn et al., Wilson, D'Agostino et al., Simons, Simons et al., Salahuddin y el rabino (citado por Zapata, Mora, y Pérez, 2014) análisis estadísticos han identificado factores de riesgos que están asociados con enfermedades cardiovasculares como la edad, la presión arterial, el hábito de fumar, altos niveles de colesterol, diabetes, hipertensión, antecedentes familiares de enfermedad cardiaca, la obesidad y la falta de actividad física. Todo lo anteriormente mencionado ayuda al profesional de la salud a identificar si en un paciente existe riesgo de enfermedad cardiovascular, sin embargo, los expertos están aplicando diferentes técnicas de minería de datos para obtener una mayor precisión en el diagnóstico de enfermedades del corazón. Los arboles de decisión, las redes neuronales, el método de naivebayes, las máquinas de soporte vectorial, entre otras, son esas técnicas que han sido utilizadas en diversos estudios a parte de los que están relacionados con el corazón, como la predicción de la resistencia a antibióticos y las infecciones nosocomiales de Gerontini, Vazirgiannis, Vatopoulos, y Polemis en 2011 o la predicción para la hospitalización de pacientes en

hemodiálisis de Yeh, Wu, y Tsao en el mismo año, lo que permite afirmar que son altamente eficientes para los diagnósticos en el campo de la salud.

Las técnicas tomadas en cuenta en este artículo serán ampliadas posteriormente, reciben el nombre de: arboles de decisión (J48), máquinas de soporte vectorial (SMO), regresión logística (simple logistic), el método de naivebayes, IBk (k vecinos más cercanos) y redes neuronales (multilayer).

2. Metodología

El presente artículo está basado en el entrenamiento de un conjunto de datos extraídos del repositorio de Machine Learning con el propósito de identificar el método o técnica de minería de datos que permita obtener mayores resultados de precisión al identificar enfermedades cardiovasculares en diferentes pacientes. Los pasos utilizados para la investigación y obtención de la información son los siguientes:

- Descargar los datos referentes a enfermedades coronarias, generados por la Universidad de Cleveland en Estados Unidos. Se encuentran en el repositorio de Machine Learning UCI (Lichman, M. 2013).
- Utilización dela herramienta para minería de datos Weka, para adelantar el entrenamiento yprueba de los clasificadores (árboles de decisión, máquinas de soporte vectorial, regresión logística, naivebayes, IBk y redes neuronales).
- Las técnicas fueron sometidas a pruebas utilizando validación cruzada con el propósito de particionar el dataset utilizando unos registros para entrenamiento y otro para la clasificación.
- Análisis de porcentajes de Tprate, Fprate, Presicion y cobertura para definir la técnica que arroje mejores resultados al momento de clasificar la enfermedad objeto de estudio.
- Realización de tablas y gráficos comparativos enlos que se pueden observar los resultados.

Para comparar estas técnicas de minería de datos se utilizó un conjunto de datos Cleveland que hacen referencia a enfermedades cardiovasculares a través del repositorio de Machine Learning uci. El conjunto de datos cuenta con 14 atributos y 303 registros así:

- Edad: en años

- Sexo: masculino y femenino (tomando como valor numérico el 1 y 0 respectivamente).

- Tipo dolor de pecho:

    - Valor1: Angina típica
    - Valor2: Angina atípica
    - Valor3: Otro dolor

- Valor4: Asintomático

- Presión arterial en reposo: En mmHgen la admisión al hospital.
- Colesterol: mg/dl
- Nivel de azúcar>120mg/dl: verdadero o falso (tomando como valor numérico el 1 y 0 respectivamente).
- Resultado electrocardiograma:
  - Valor0: Normal
  - Valor1: Anomalías
  - Valor2: Hipertrofia ventricular
- Frecuencia cardiaca máxima alcanzada.
- Ejercicio de inducción de angina: sí o no (tomando como valor numérico el 1 y 0 respectivamente)
- Depresión inducida.
- Pendiente curva máxima del ejercicio.
- Número de vasos mayores (0-3).
- Thal:
  - 3: Normal
  - 6: Defecto fijo
  - 7: Defecto reversible
- Diagnóstico de enfermedad cardiaca (estado de enfermedad hagiográfica): menor 50% o mayor 50% (tomando como valor numérico el 0 y 1 respectivamente).

3. Fundamentos teóricos

3.1Árboles de decisión

Los árboles de decisión según Hernández, Ramírez y Ferri (citado por Lizazo, Delfor& Torres, 2011)son un conjunto de condiciones organizadas por medio de una estructura jerárquica, con el fin de que siguiendo las condiciones que se cumplen desde la raíz del árbol hasta algunas de sus hojas se pueda determinar una decisión final.

Los árboles de decisión pertenecen a las tareas predictivas de clasificación en la minería de datos y por ello tienen la función de clasificar datos y predecir el comportamiento de los mismos de manera estadística.

En Weka el algoritmo J48 es uno de los algoritmos de aprendizaje de árboles de decisiones más utilizados por su efectividad, debido a que selecciona la prueba que genera mayor cumulo de información luego de haber dividido el conjunto de datos, es decir, genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente.

3.2 Máquinas de soporte vectorial

Ésta técnica constituye un tipo de aparato de aprendizaje de gran éxito en la resolución de problemas básicos del aprendizaje supervisado: clasificación y regresión. Son máquinas lineales donde las soluciones no se construyen en el espacio de entrada sino en un espacio de mayor dimensionalidad, el espacio de características, donde es muy posible que una función lineal sea suficiente para resolver el problema, dadas las correlaciones de alto orden de los datos que se hacen explícitas.

El modelo al entregarle un conjunto de datos predice, por cada entrada dada, cuál de dos clases forma la salida: es un clasificador lineal binario no probabilístico según (Lugo, Maldonado y Murata, 2014).Teniendo los datos se realiza un modelo de representación como puntos en el espacio en un hiperplano (calculado maximizando la distancia de los patrones más cercanos) con una separación lo más amplia posible.

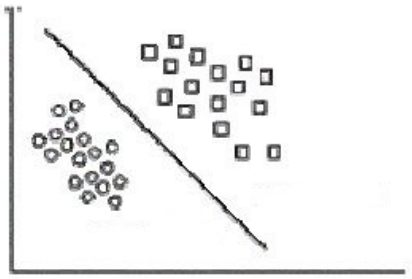

Figura 1. Hiperplano de separación
Fuente: Orozco et al. 2010

3.3 Regresión logística

La predicción de valores continuos puede planificarse por las técnicas estadísticas de regresión. El objetivo esencial de la técnica de regresión es utilizar las variables independientes, cuyos valores son conocidos, para predecir la única variable criterio (dependiente) seleccionada por el que esté utilizando la técnica.

El objetivo primordial de la regresión logística es el de modelar la influencia de las variables regresoras en la probabilidad de ocurrencia de un suceso particular. Sistemáticamente tiene dos objetivos: el primero es investigar cómo influye la probabilidad de ocurrencia de un suceso, la presencia de diversos factores, o no, y el

valor o nivel de estos; el segundo es determinar el modelo más ajustado que describa la relación entre la variable respuesta y un conjunto de variables regresoras. Salcedo (citado por De la Hoz, Martinez& Mendoza, 2013).

3.4 Naivebayes

Este método de clasificación está fundamentados en modelos probabilísticos, concretamente utiliza el teorema de Bayes sobre las probabilidades condicionadas pero no tiene en cuenta las dependencias que puedan existir.

El método de NaiveBayes al ser relativamente fácil de implementar se ha aplicado muchas investigaciones, se ha usado como método de aprendizaje automático debido a su habilidad para combinar hechos o indicios a partir de un conjunto de rasgos. Su mayor problema radica en suponer que la presencia de un rasgo independiente de otro (Martí, 2004). La suposición de independencia puede no cumplirse para algunos atributos: se deben utilizar otras técnicas tales como redes de creencias bayesianas.

3.5 IBk

Es un algoritmo que pertenece a la técnica basada en ejemplos que consiste en la clasificación realizada por medio de una función que mide la proximidad o parecido. Dado un ejemplo para clasificar se le clasifica de acuerdo al ejemplo o ejemplos más próximos. El BIAS (sesgo) que rige este método es la proximidad, es decir, la generalización se guía por la proximidad de un ejemplo a otros. La técnica basada en ejemplos se suele considerar no adecuada para el tratamiento de atributos no numéricos y valores desconocidos.

IBk es el método de los k vecinos más cercanos para regresión, es un método de aproximación sin parámetros, éste, permite resolver problemas de clasificación y regresión. Se basa en la suposición que la clase a la cual corresponde un objeto es la misma a la que pertenecen sus vecinos más cercanos. Moujahid (citado por Zapata, Mora & Pérez, 2014)

3.6 Redes neuronales

Las redes neuronales son redes que están compuestas de una capa de entrada (capa distribuidora de las entradas a la siguiente capa), una capa de salida, y al menos una capa intermedia. En las técnicas de redes neuronales la red contiene elementos procesadores de información de cuyas interacciones locales depende el comportamiento del conjunto del sistema. Tratan de emular el comportamiento del cerebro humano. Éstas constituyen modelos de aprendizaje y procesamiento automático.

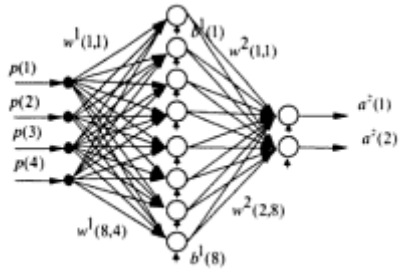

Figura 2. Red neuronal con entradas dos neuronas de salida y ocho neuronas intermedias
Fuente: Valderrama, 1999

Hay que destacar que no todos los modelos neuronales son aptos para todas las tareas de minería de datos. De manera general cada método tiene características específicas por construcción y se destacan en diversas áreas como el control, optimización, visión, entre otras. Lo más probable es que siempre la arquitectura de una red neuronal esté estrechamente relacionada con el algoritmo de aprendizaje de la misma. Siendo así encontramos que puede existir un red estática simple donde las neuronas de una red se organizan en una capa, este tipo de red posee una sola capa ya que todas la unidades de cómputo están en el mismo nivel jerárquico; reciben el mismo tipo de entradas y proveen el mismo tipo de salidas. (CYTED-Conicit, 1999)

4. Resultados

El proceso de clasificaciónfue desarrollado utilizando la herramienta *Weka*versión 3.6. Los datos suministrados por la base de datos de Cleveland alojadas en el repositorio de Machine Learning uci son los siguientes:

- 14 Atributos
- 303 Registros

En la base de datos existían datos faltantes que fueron tomados como valores en ceros para incorporar en los procesos de clasificación. En el estudio realizado se aplicaron seis técnicaspara realizar un comparativo entre estos y observar quien proporcionaba mayor precisión y menor error en las predicciones relacionadas con enfermedades cardiovasculares. En totallos métodos comparados fueron seis: árboles de decisión, regresión logística, naivebayes, IBk, redes neuronales y máquinas de soporte vectorial. Se obtuvo la tabla de indicadores que permitió conocer la tasa de verdaderos y falsos positivos (tp y fp respectivamente), la precisión y la cobertura, lo cual nos permite establecer el mejor método a través de los indicadores de precisión o cobertura, esto último dependiendo de si la precisión no nos arroja resultados tan exactos entonces revisamos la cobertura de cada método. A continuación se muestran los resultados de cada método y sus respectivas gráficas realizadas en Excel para un mejor análisis:

	DIAGNÓSTICO DE ENFERMEDAD CARDIACA (menor 50%)			
	TP Rate	**FP Rate**	**Precision**	**Recall**
J48	0,805	0,281	0,772	0,805
SMO	0,86	0,216	0,825	0,86
SIMPLELOGISTIC	0,878	0,216	0,828	0,878
NAIVEBAYES	0,872	0,209	0,831	0,872
IBk	0,78	0,266	0,776	0,78
MULTILAYER	0,902	0,899	0,542	0,902

Tabla 1. Datos para el diagnóstico de enfermedad cardiaca (menor 50%)

	DIAGNÓSTICO DE ENFERMEDAD CARDIACA (mayor 50%)			
	TP Rate	**FP Rate**	**Precision**	**Recall**
J48	0,719	0,195	0,758	0,719
SMO	0,784	0,14	0,826	0,784
SIMPLELOGISTIC	0,784	0,122	0,845	0,784
NAIVEBAYES	0,791	0,128	0,84	0,791
IBk	0,734	0,22	0,739	0,734
MULTILAYER	0,101	0,098	0,467	0,101

Tabla 2. Datos para el diagnóstico de enfermedad cardiaca (mayor 50%)

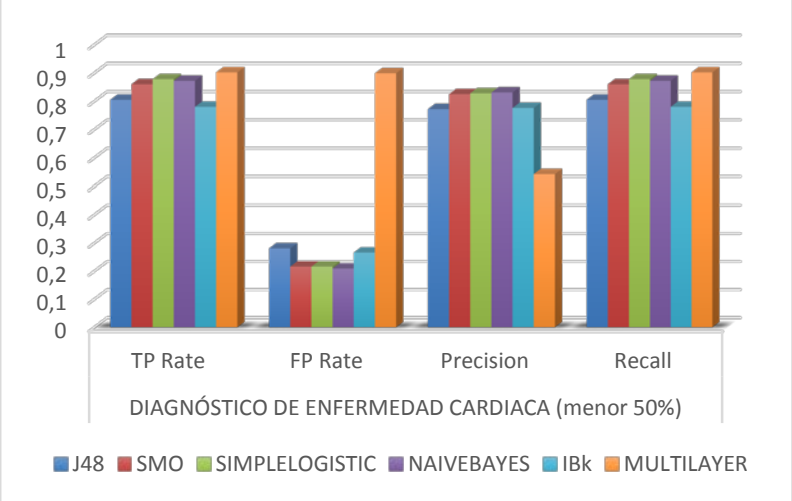

Gráfico 1. Diagnóstico de enfermedad cardiaca (menor 50%)

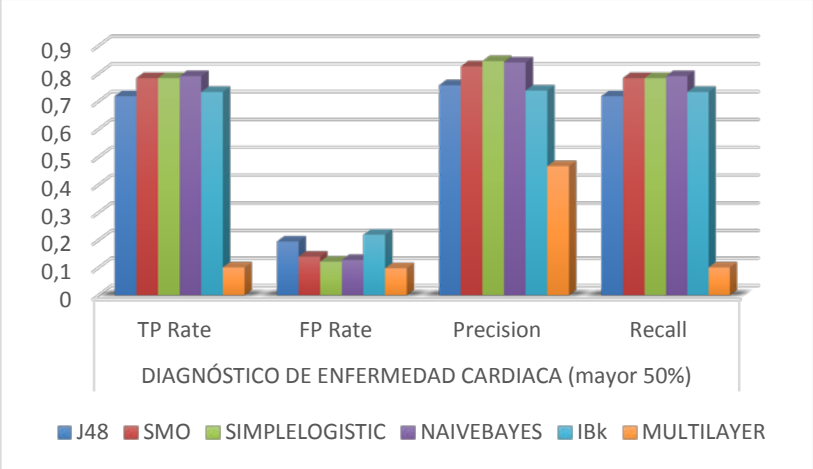

Gráfico 2. Diagnóstico de enfermedad cardiaca (mayor 50%)

Se puede notar que para el diagnóstico de enfermedad cardiaca existen dos técnicas que son precisas, éstas son: SimpleLogistic (regresión logistica) y NaiveBayes, debido a que, para el diagnóstico de enfermedad cardiaca menor del 50% la mejor es SimpleLogistic y para el diagnóstico de enfermedad cardiaca mayor del 50% es NaiveBayes.

5. Conclusión

Las técnicas de Machine Learning permiten el análisis de datos a gran escala para lograr un mayor acierto relacionado con la detección o predicción de enfermedades cardiovasculares, tras someter a pruebas técnicas como son los árboles de decisión, las máquinas de soporte vectorial, la regresión logística, IBk, redes de múltiples capas y naive bayes. Después de validar las técnicas anteriormente mencionadas con los mismos datos (repositorio de Machine Learning uci), la regresión logística consigue un resultado desde el punto de vista de precisión del 84,50%, para la solución de problemas de clasificación de enfermedades cardiovasculares mayor del 50%, mientras que las máquinas de soporte vectorial obtienen un nivel de precisión de 82,60% y el método de naive bayes 84,00%, lo cual indica que son métodos que registran resultados aceptables al momento de ser utilizado como técnica de clasificación, analizando los datos existentes en el repositorio de Machine Learning uci. Por otro lado, los árboles de decisión ofrecen un nivel de precisión de 75,80%, IBk 73,90% y redes neuronales 46,70% lo que permite deducir que esta última es la técnica menos precisa en el momento del análisis de los datos almacenados en el repositorio, agregando que las dos anteriormente mencionadas a redes neuronales también son inaceptables por su poca precisión.

Por otro lado, para la solución de problemas de clasificación de enfermedades cardiovasculares menor del 50%, el método de NaiveBayes consigue un resultado desde el punto de vista de precisión del 83,10% primando entre las demás técnicas puesto que los arboles de decisión tienen una precisión del 77,20%, máquinas de soporte vectorial 82,50%, regresión logística 82,80%, IBk 77,60% y redes neuronales

54,20%, siendo los arboles de decisión, IBk y redes neuronales métodos inaceptables por presentar una baja precisión.

6. Referencias bibliográficas

- CYTED-Conicit, De la fuente J., &Calonge T. (1999). *Aplicaciones de las redes de neuronas en supervisión, diagnosis y control de procesos.* Recuperado de: https://books.google.com.co/books?id=jUHGRXd5xU8C&printsec=frontcover&hl=es#v=onepage&q&f=false
- De la Hoz, A., Martínez, U., & Mendoza, F. (2013).*Técnicas de ML en medicina cardiovascular.* Memorias, 11(20), 41-46.
- García, A. (2012). *Inteligencia artificial: Fundamentos práctica y aplicaciones.*Recuperado de:https://books.google.com.co/books?id=WDuqquRP70UC&printsec=frontcover&hl=es#v=onepage&q&f=false
- Gerontini M., Vazirgiannis M., Vatopoulos A., &Polemis M. (2011). *Predictions in antibiotics resistance and nosocomial infections monitoring.* ACM digital library. Recuperado de: http://dl.acm.org/citation.cfm?id=2190647.2190787&coll=DL&dl=GUIDE
- Gonzales, F. (2007). *Análisis supervisado III. Modelos probabilísticos*. Bogotá: Universidad Nacional de Colombia – UNAL. Recuperado de: http://dis.unal.edu.co/~fgonza/courses/2007-II/mineria/bayesianos.pdf
- Lichman, M. (2013). UCI Machine Learning Repository. Irvine, University of California, School of Information and ComputerScience. Consultado 21 Febrero 2015
- http://archive.ics.uci.edu/ml
- Lizazo, D., Delfor R., & Torres V. (2011). *Minería de datos en la encuesta permanente de hogares 2009, Universidad Nacional del Litoral, Argentina.* Recuperado de: http://web.b.ebscohost.com/ehost/pdfviewer/pdfviewer?sid=6658f722-ef5b-4074-af91-82526a9be75f%40sessionmgr115&vid=4&hid=102
- Lugo S., Maldonado G., &Murata Ch. (2014). *Inteligencia artificial para asistir el diagnóstico clínico en medicina.* Revista Alergia México 2014; 61:110-120. Recuperado de: http://web.b.ebscohost.com/ehost/pdfviewer/pdfviewer?sid=f19f62c6-f445-4ded-8146-66a981b0b960%40sessionmgr113&vid=1&hid=102
- Martí, M., &Llisterri, J. (2004).*Tecnologías del texto y del habla.* Recuperado de:https://books.google.com.co/books?id=hNPMEnqfc44C&printsec=frontcover&hl=es#v=onepage&q&f=false
- Orozco E., Delgado J., Vázquez S., Castro J., Villanueva A., &Gutierrez F. (2010). *Diagnóstico de lesones en la piel a partir de espectros de reflexión difusa empleando algoritmos computacionales: un estudio preliminar.* Revista Cubana de Física. Recuperado de: http://web.b.ebscohost.com/ehost/pdfviewer/pdfviewer?sid=4ae914c5-4cbf-48f4-b583-51d4779356e5%40sessionmgr110&vid=1&hid=102

- Pérez, C. &Santín, D. (2007). *Minería de datos. Técnicas y Herramientas.* Recuperado de: https://books.google.com.co/books?id=wz-D_8uPFCEC&printsec=frontcover&hl=es#v=onepage&q&f=false
- Shouman M., Turner T., & Stocker R. (2011). *Using decision tree for diagnosing heart disease patients.* ACM digital library. Recuperado de: http://dl.acm.org/citation.cfm?id=2483628.2483633&coll=DL&dl=GUIDE&CFID=667338602&CFTOKEN=46629837
- Valderrama, J. (1999) *Información tecnológica.* CIT. Recuperado de: https://books.google.com.co/books?id=EyIJD5tMQ0C&printsec=frontcover&hl=es#v=onepage&q&f=false
- Vieira L., OrtizL., & Ramírez. (2009).*Introducción a la minería de datos.*Recuperado de: https://books.google.com.co/books?id=jlJEhHyESFsC&printsec=frontcover&hl=es#v=onepage&q&f=false
- Yeh, J., Wu T., &Tsao Ch. (2011). *Using data mining techniques to predict hospitalization of hemodialysis patients.* ACM digital library. Recuperado de: http://dl.acm.org/citation.cfm?id=1899564.1899623&coll=DL&dl=GUIDE
- Zapata A., Mora J., & Pérez S. (2014, 29 de Abril). *Metodología híbrida basada en el regresorknn y el clasificador boosting para localizar fallas en sistemas de distribución.* Ingeniería y competitividad. Recuperado de: http://web.b.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=3&sid=0e08fe78-23c8-4e5b-ae30-311ec756a62b%40sessionmgr114&hid=125

Sobre los autores

- **Jhon Harol Campo Mendoza**: Estudiante de Ingeniería de Sistemas, Universidad de la Costa, CUC
- **Karen Dayana Parra García**: Estudiante de Ingeniería de Sistemas, Universidad de la Costa, CUC
- **Fabio Mendoza Palechor**: Ingeniero De Sistemas, Magister en Ingeniería, Universidad Tecnológica de Bolívar, Estudiante de Doctorado en Ingeniería, Universidad Pontificia Bolivariana. Profesor Tiempo Completo, Universidad de la Costa, fmendoza1@cuc.edu.co
- **Alexis De La Hoz Manotas**: Ingeniero de Sistemas, Magister en Ingenieria de Sistemas y Computación, Universidad del Norte. Profesor Tiempo Completo, Universidad de la Costa, CUC, adelahoz6@cuc.edu.co