CIENCIAMATRIA
Revista Interdisciplinaria de Humanidades, Educación, Ciencia y Tecnología
Año VIII. Vol. VIII. Nro. 4. Edición Especial 4. 2022
Hecho el depósito de ley: FA2021000002
ISSN-L: 2542-3029; ISSN: 2610-802X
Instituto de Investigación y Estudios Avanzados Koinonía (IIEAK). Santa Ana de Coro. Venezuela

DOI 10.35381/cm.v8i4.912

Machine learning para la extracción de información biomédica en un laboratorio clínico

Machine learning for the extraction of biomedical information in a clinical laboratory

Ivan Mauricio Álvarez-Bonilla
pa.ivanmab44@uniandes.edu.ec
Universidad Regional Autónoma de los Andes, Ambato, Tungurahua
Ecuador
https://orcid.org/0000-0001-7114-0412

Ariel José Romero-Fernández
ua.arielromero@uniandes.edu.ec
Universidad Regional Autónoma de los Andes, Ambato, Tungurahua
Ecuador
https://orcid.org/0000-0002-1464-2587

Gustavo Eduardo Fernández-Villacrés
ua.eduardofernandez@uniandes.edu.ec
Universidad Regional Autónoma de los Andes, Ambato, Tungurahua
Ecuador
https://orcid.org/0000-0003-1028-1224

Luis Rafael Freire-Lescano
ua.luisfreire@uniandes.edu.ec
Universidad Regional Autónoma de los Andes, Ambato, Tungurahua
Ecuador
https://orcid.org/0000-0002-6527-6417

Recibido: 01 de mayo 2022
Revisado: 25 de junio 2022
Aprobado: 01 de agosto 2022
Publicado: 15 de agosto 2022

CIENCIAMATRIA
Revista Interdisciplinaria de Humanidades, Educación, Ciencia y Tecnología
Año VIII. Vol. VIII. Nro. 4. Edición Especial 4. 2022
Hecho el depósito de ley: FA2021000002
ISSN-L: 2542-3029; ISSN: 2610-802X
Instituto de Investigación y Estudios Avanzados Koinonía (IIEAK). Santa Ana de Coro. Venezuela

Ivan Mauricio Álvarez-Bonilla; Ariel José Romero-Fernández; Gustavo Eduardo Fernández-Villacrés
Luis Rafael Freire-Lescano

RESUMEN

El objetivo es aplicar modelos de Machine learning para encontrar patrones e información biomédica escondida que permita mejorar la toma de decisiones y ayudar al diagnóstico clínico, desde un enfoque racionalista de la investigación. El proceso de Knwoledge Discovery in Databases (KDD) se utilizó para descubrir y extraer cocimiento, puesto que es iterativo e interactivo. Al ser un proceso iterativo en cada paso, significa que puede ser necesario volver a los pasos anteriores. Los algoritmos supervisados y no supervisados tienen la capacidad de ayudar en la toma decisiones, debido a que permite una mejor compresión de los datos y puede ayudar a descubrir nuevas interrogantes que puede llevar a otras investigaciones de vital importancia. Las reglas de asociación pueden ayudar a determinar factores que influye en la salud de los humanos y con ello se puede tomar medidas de prevención para mejorar el estado de salud.

Descriptores: Sistema de información de gestión; industria de la información; medicina clínica. (Tesauro UNESCO).

ABSTRACT

The objective is to apply Machine learning models to find patterns and hidden biomedical information to improve decision making and aid clinical diagnosis, from a rationalistic approach to research. The Knwoledge Discovery in Databases (KDD) process was used to discover and extract knowledge, since it is iterative and interactive. Being an iterative process at each step means that it may be necessary to go back to previous steps. Supervised and unsupervised algorithms have the ability to aid in decision making because it allows for better understanding of the data and can help uncover new questions that can lead to further vital research. Association rules can help to determine factors that influence the health of humans and with this, preventive measures can be taken to improve health status.

Descriptors: Management information systems; information industry; clinical medicine. (UNESCO Thesaurus).

CIENCIAMATRIA
Revista Interdisciplinaria de Humanidades, Educación, Ciencia y Tecnología
Año VIII. Vol. VIII. Nro. 4. Edición Especial 4. 2022
Hecho el depósito de ley: FA2021000002
ISSN-L: 2542-3029; ISSN: 2610-802X
Instituto de Investigación y Estudios Avanzados Koinonía (IIEAK). Santa Ana de Coro. Venezuela

INTRODUCCIÓN

El Machine learning ha experimentado un desarrollo significativo durante la última década y se está utilizando con éxito en muchas aplicaciones inteligentes que cubren una amplia gama de problemas relacionados con los datos (Jordan & Mitchell, 2015). En el diagnóstico médico, la información disponible se complementa con la recopilación de datos adicionales, que se pueden obtener a partir de la historia clínica de un paciente, un examen físico y de varias pruebas de diagnóstico, incluidas las pruebas de laboratorio clínico. Las pruebas de laboratorio se utilizan para confirmar, excluir, clasificar o monitorear enfermedades y para guiar él tratamiento (Badrick, 2013).

El Machine learning aplicado a los resultados de las pruebas laboratorio de ferritina de pacientes ambulatorios ofrece un nuevo tipo de apoyo a la decisión clínica, destinado a mejorar el valor diagnóstico de múltiples analitos. Los datos demográficos de los pacientes y los resultados de otras pruebas de laboratorio diferencian los resultados normales de los resultados anormales de ferritina con un alto grado de precisión. Solo los resultados bajos de ferritina se clasificaron como anormales, ya que el objetivo era identificar la deficiencia de hierro, lo que se indica con una ferritina baja, la clasificación se realizó utilizando la regresión logística implementada en el paquete Python Scikit-learn. Las técnicas de regresión utilizadas fueron la regresión lineal, la regresión lineal bayesiana, la regresión aleatoria de bosques (Luo, et al. 2016).

Las redes bayesianas sirven para identificar la Diabetes Mellitus con base en el análisis de algunas variables tales como: número de embarazos, presión arterial diastólica, espesor cutáneo del tríceps, índice de masa corporal, herencia y edad en base a muestras tomadas de pacientes diabéticos y no diabéticos en el 87,69% de los casos, el clasificador bayesiano logra detectar correctamente esta enfermedad con base en las variables antes sugeridas. Sin embargo, al agregar la variable "insulina en suero", el porcentaje aumentó al 98.46% (Castrillón & Sarache, 2017).

El Machine learning es el campo de más rápido crecimiento en informática, y la informática para la salud se encuentra entre los mayores desafíos. La aplicación de métodos de Machine learning en biomedicina y salud puede, por ejemplo, llevar a una mayor toma de decisiones basada en la evidencia y ayudar a la medicina personalizada (Holzinger, 2016).

En general son dos los métodos de análisis de datos para Machine learning y estos son: supervisados y no supervisados. En ambos casos se requiere una muestra de datos, esta información puede denominarse la muestra de entrenamiento o datos conocidos. La muestra de entrenamiento es utilizada por las actividades de minería de datos para aprender los patrones en los datos. El análisis de datos supervisado y sus algoritmos se aplican a datos etiquetados intentado encontrar una función que, dadas las variables de entrada les asigne la etiqueta de salida adecuada de modo de que el algoritmo predice el valor de salida (Ahlemeyer-Stubbe & Coleman, 2014).

La clasificación está dirigido a crear un modelo de dependencia entre variables independientes que describen fenómenos y una variable dependiente en forma de atributo. Los métodos más frecuentes son los árboles de clasificación, las redes neuronales y el clasificador bayesiano. Mientras que los métodos no supervisados representado en este caso por la asociación consiste en encontrar y extraer patrones de las relaciones que hay entre diferentes atributos de la información (Ahlemeyer-Stubbe & Coleman, 2014).

Por tal razón, el objetivo es aplicar modelos de Machine learning para encontrar patrones e información biomédica escondida que permita mejorar la toma de decisiones dentro del área del laboratorio clínico y ayudar al diagnóstico clínico, desde un enfoque racionalista de la investigación.

Propuesta de aplicación de Machine learning

El proceso de Knwoledge Discovery in Databases (KDD) se utilizó para descubrir y extraer cocimiento, puesto que es iterativo e interactivo. Al ser un proceso iterativo en cada paso, significa que puede ser necesario volver a los pasos anteriores. El proceso

Año VIII. Vol. VIII. Nro. 4. Edición Especial 4. 2022
Hecho el depósito de ley: FA2021000002
ISSN-L: 2542-3029; ISSN: 2610-802X
Instituto de Investigación y Estudios Avanzados Koinonía (IIEAK). Santa Ana de Coro. Venezuela

de KDD puede llegar a ser una actividad multidisciplinaria, que engloba técnicas que están al alcance de cualquier disciplina particular, como el Machine learning. Además, hay que destacar que el proceso KDD emplea la minería de datos como uno de sus pasos (Kononenko & Kukar, 2007).

En un proceso típico de descubrimiento de conocimiento, generalmente se consideran las siguientes etapas en el proceso KDD como se muestra en la figura 1 (Oded & Lior, 2014):

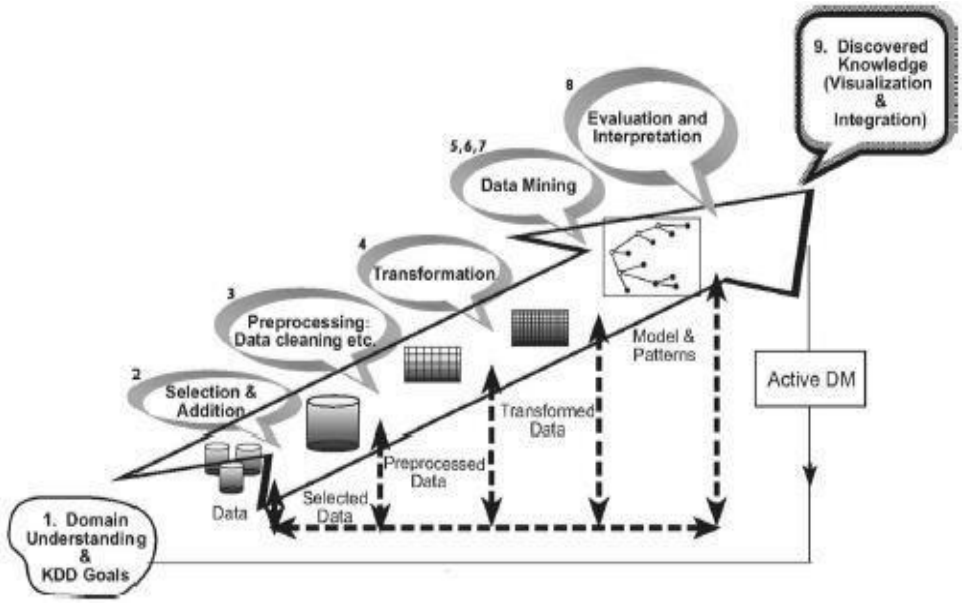

Figura 1. El proceso de extracción de conocimiento KDD.
Fuente: Oded & Lior, (2014).

1 **Comprensión del dominio de la aplicación**: proceso inicial, que tiene la labor de comprender los objetivos y requerimientos, por lo tanto, se aplicó técnicas de Machine learning para buscar conocimiento y patrones biomédicos.

2 **Selección de datos:** se obtiene el primer conjunto de datos que se utilizará en el proceso de KDD. Se realizaron un conjunto de análisis de laboratorio clínico, como: biometrías, bioquímica sanguínea y estudios hormonales, simultáneamente se aplicó una encuesta sociodemográfica

y se realizaron mediciones antropométricas básicas. La información estuvo almacenada en un archivo de extensión sav del programa estadístico SPSS y luego se constituyeron todos los datos en un solo conjunto de datos de tipo csv con 272 registros.

9 **Preprocesamiento y limpieza**: se eliminaron los datos que se relacionan con: valores perdidos, eliminando ruido o valores atípicos, para obtener datos más fiables, para lo cual los datos cualitativos se remplazaron con el valor de más frecuencia y para los datos cuantitativos con el valor de la media.

10 **Transformación de datos:** se preparó y desarrolló para generar mejores datos, para el proceso de KDD. Uno de los métodos que se utilizó fue la reducción de dimensión, así como la selección de características. La discretización de atributos numéricos y transformación funcional fue necesaria en esta etapa. Para la función tiroidea y renal se estableció los siguientes atributos divididos en características y objetivos.

Cuadro 1.
Atributos para la función tiroidea.

CARACTERÍSTICA	**VALOR**	**INTERVALO**
Caida_cabello	si o no	
TSH	Numérico	(0.40 - 4.00) µUI/L
T3	Numérico	(81.00 - 178.00) ng/dL
T4	Numérico	(0.89 - 1.76) ng/dL
Sensacion_frio	si o no	
OBJETIVO	**VALOR**	**CONDICIÓN**

funcion_tiroidea	Normal o anormal	Si se encuentra dentro de los tres intervalos de TSH, T3, T4 es normal, de lo contrario anormal

Elaboración: Los autores.

Cuadro 2.
Atributos para la función renal.

CARACTERÍSTICA	VALOR	INTERVALO
Sexo	Hombre, mujer	
Carne_vaca	Siempre frecuentemente, poco frecuente, nunca	
Carne_cerdo	Siempre frecuentemente, poco frecuente, nunca	
Pollo	Siempre frecuentemente, poco frecuente, nunca	
P_Deporte	Si, no	
Urea	Numérico	(7.0 - 18.0) mg/dL
Acido_úrico	Numérico	(2.60 - 6.00) mg/dL
Creatinina (hombre)	Numérico	(0.70 - 1.30) mg/dL

Creatinina (mujer)	Numérico	(0.55 - 1.02) mg/dL
OBJETIVO	**VALOR**	**CONDICIÓN**
funcion_renal	Normal o anormal	Si se encuentra dentro de los tres intervalos de: urea, ácido úrico y creatinina se considera función normal

Elaboración: Los autores.

La función metabólica de carbohidratos contiene las siguientes características y se procedió a la transformación para la generación de los nuevos atributos.

Cuadro 3.
Atributos para función metabólica.

CARACTERÍSTICA	INTERVALO	ATRIBUTO	CONDICIÓN
IMC	(18.5-24.99) Kg/m^2	Función_IMC	Si 18.5< IMC <24.9= normal Caso contrario = anormal
Glucosa basal	(74.0-106.0) mg/dL	Función_ Glucosa	Si 74.0 < glucosa < 106.0= normal Caso contrario = anormal
Globulina	(1.40–3.40) g/dL	Función _Globulina	Si 1.40 < globulina <3.40 = normal Caso contrario = anormal
	(50.0-200.0)	Función	Si 50.0 < colesterol total

Colesterol total	mg/dL	_Colesterol	<200.0 = normal Caso contrario = anormal
HDL colesterol	(40.0-60.0) mg/dL	Funcion _HDL	Si 40.0 < HDL colesterol <60.0 = normal Caso contrario = anormal
LDL colesterol	(70-130) mg/dL	Función_LDL	Si 70 < LDL colesterol < 130 = normal Caso contrario = anormal
Triglicéridos	(30.0-150.0) mg/dL	Función _Triglicéridos	Si 30.0 <triglicéridos <150.0 = normal Caso contrario = anormal
TGO	(15.0-37.0) U/L	Función _TGO	Si 15.0 < TGO < 37.0 = normal Caso contrario = anormal
TGP	(14.0-59.0) U/L	Función _TGP	Si 14.0 < TGP < 59.0= normal Caso contrario = anormal
GGT	(15.0-85.0) U/L	Función _GGT	Si 15.0 < GGT < 85.0 = normal Caso contrario = anormal

Elaboración: Los autores.

5. **Minería de Datos:** una vez realizadas las anteriores tareas, se elige el modelo de Machine learning que se acopla a las necesidades de la investigación, tales como: clasificación, asociación, regresión y agrupación. Para el presente estudio se aplicaron técnicas de clasificación y asociación:
6. **Elegir el algoritmo de Machine learning:** se procedió a la selección del de los algoritmos, por lo cual se seleccionó lo siguiente:
    a. **Técnicas supervisadas:** Tree, Logistic Regression, CN2 rule inducer y AdaBoost
    b. **Técnicas no supervisadas:** reglas de asociación
7. **Emplear el algoritmo:** los algoritmos fueron empleados con la herramienta Orange Canvas 3.20.1 para las técnicas supervisadas y para la no supervisadas Rapidminer en su versión 9.2.
8. **Evaluación:** este paso se concentra en la comprensibilidad y la utilidad del modelo, por esta razón se utilizó para los algoritmos supervisados el método de cross validation, método común y muy utilizado para evaluar el rendimiento de los sistemas basados en el Machine learning y las estadísticas de rendimiento y estas son (Hutton, 2012)
    a. **El área bajo la curva (AUC):** es el área bajo la curva operativa del receptor.
    b. **Exactitud de clasificación (CA):** es la proporción de ejemplos correctamente clasificados.
    c. **F-1: es un medio armónico ponderado de precisión y recuperación (ver más abajo).**
    d. **La precisión:** es la proporción de verdaderos positivos entre los casos clasificados como positivos.
    e. **Sensibilidad (Recall):** es la proporción de verdaderos positivos entre todas las instancias positivas en los datos**.**

A estas métricas se les acompañó con la matriz de confusión, que permite comprender como un modelo se comporta con las predicciones en comparación con los valores reales de un conjunto de datos (Torres, 2018). En el caso de los algoritmos no supervisados y en especial para la regla de asociación, el nivel de una regla de asociación se cuantifica generalmente por el soporte (support), medidas de confianza (confidence), lift y medidas de convicción (conviction) que pueden ser utilizadas en casos especiales, estas medidas son establecidas con la frecuencia relativa de ocurrencias de un atributo particular establecido en el conjunto de datos (Braulio Gil & Curto Díaz, 2016).

Para la predicción de la función tiroidea, los algoritmos supervisados sobrepasan la sensibilidad, precisión y exactitud en un 85%, sin embargo, AdaBoost presentò una sensibilidad, precisión y exactitud del 98.9%, de acuerdo con la matriz de confusión solo 1.3% de los 224 casos de una función tiroidea normal se equivocó en su predicción. Hay que mencionar que el algoritmo Tree, tomó en cuenta los atributos de las hormonas T4 y TSH para establecer la normalidad de la función tiroidea.

En la predicción de la función renal en mujeres el algoritmo AdaBoost contiene una sensibilidad, precisión y exactitud del 97.1%, de acuerdo con la matriz de confusión, solo en un 2.5% de las 119 mujeres que mantienen su función renal normal se equivocó en la predicción. El algoritmo Tree consideró para predecir la función renal en mujeres los atributos urea y creatinina.

Mientras en la predicción de la función renal en hombres el algoritmo Tree obtuvo una sensibilidad, precisión y exactitud del 98.7%; a la vez el algoritmo Tree considera que el atributo creatina y ácido úrico son los atributos que determinan una función renal como normal o anormal, de acuerdo a la matriz de confusión Tree predijo el 100% de los casos anormales y se equivocó en predecir los casos normales en un 5.4% de 37 casos.

En cuanto a las reglas de asociación, con una confianza del 100 % y con un lift mayor a 1, en los 53.24% de los casos, los valores de colesterol, glucosa y globulina están relacionados en el valor de la gamma glutamil transpeptidasa.

Similares estudios en 2018, utilizaron árboles de decisiones, bosques aleatorios y redes neuronales, que pertenecen a las técnicas supervisadas de Machine learning, para predecir la diabetes mellitus, los resultados mostraron que la predicción con bosque aleatorio podría alcanzar la mayor precisión (ACC = 80.84%) cuando se utilizaron todos los atributos (Zou et al., 2018).

CONCLUSIONES

Los algoritmos de árbol de decisión han demostrado un buen rendimiento para la predicción, esto ayuda en el diagnóstico médico.

Los algoritmos supervisados y no supervisados tienen la capacidad de ayudar en la toma decisiones, debido a que permite una mejor compresión de los datos y puede ayudar a descubrir nuevas interrogantes que puede llevar a otras investigaciones de vital importancia.

Las reglas de asociación pueden ayudar a determinar factores que influye en la salud de los humanos y con ello se puede tomar medidas de prevención para mejorar el estado de salud.

FINANCIAMIENTO

No monetario.

AGRADECIMIENTO

A la Universidad Regional Autónoma de los Andes; por motivar el desarrollo de la investigación.

REFERENCIAS CONSULTADAS

Badrick, T. (2013). Evidence-based laboratory medicine. *The Clinical Biochemist. Reviews*, *34*(2), 43–46. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/24151340

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255 LP – 260. https://doi.org/10.1126/science.aaa8415

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus With Machine Learning Techniques. *Frontiers in Genetics*, *9*, 515. https://doi.org/10.3389/fgene.2018.00515

Braulio Gil, N., & Curto Díaz, J. (2016). *Customer analytics*. Barcelona: Editorial UOC, S.L.

Hutton, J. (2012). Pediatric Biomedical Informatics: Computer Applications in Pediatric Research (1st ed.). New York: Springer Science & Business Media.

Torres, J. (2018). DEEP LEARNING Introducción práctica con Keras. (I. Published, Ed.) (3rd ed.). Barcelona.

Oded, M., & Lior, R. (2014). Data Mining With Decision Trees: Theory And Applications (2nd ed.). London: World Scientific Publishing Company.

Kononenko, I., & Kukar, M. (2007). Machine Learning and Data Mining (1st ed.). Chichester,Uk: Elsevier Science.

Ahlemeyer-Stubbe, A., & Coleman, S. (2014). A Practical Guide to Data Mining for Business and Industry. https://doi.org/10.1002/9781118763704

Holzinger, A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, *3*(2), 119–131. https://doi.org/10.1007/s40708-016-0042-6

Castrillón, O. D., & Sarache, W. (2017). Sistema Bayesiano para la Predicción de la Diabetes Bayesian System for Diabetes Prediction, *28*, 161–168.

Luo, Y., Szolovits, P., Baron, J. M., & Dighe, A. S. (2016). Using Machine Learning to Predict Laboratory Test Results. *American Journal of Clinical Pathology*, *145*(6), 778–788. https://doi.org/10.1093/ajcp/aqw064

CIENCIAMATRIA
Revista Interdisciplinaria de Humanidades, Educación, Ciencia y Tecnología
Año VIII. Vol. VIII. Nro. 4. Edición Especial 4. 2022
Hecho el depósito de ley: FA2021000002
ISSN-L: 2542-3029; ISSN: 2610-802X
Instituto de Investigación y Estudios Avanzados Koinonía (IIEAK). Santa Ana de Coro. Venezuela

Ivan Mauricio Álvarez-Bonilla; Ariel José Romero-Fernández; Gustavo Eduardo Fernández-Villacrés
Luis Rafael Freire-Lescano