Quántica
Ciencia con impacto social
ISSN: 2711-4600

Las TIC en el sector salud, machine learning para el diagnóstico y prevención de enfermedades

Direction of educational projects applied to information and communication technologies

Recibido Julio 2019 – Aceptado Noviembre 2019
Quántica. Ciencia con impacto social
Vol – 1 No. 2, Julio - Diciembre 2020
e-ISSN: 2711-4600
Pgs 1-32

Edgar Olmedo Cruz Micán
Doctor en Administración de Negocios, con estudios en Alta Investigación posdoctoral en Educación, Ciencias Sociales e Interculturalidad
Docente Investigador – Corporación Universitaria Minuto de Dios – UNIMINUTO
ecruz@uniminuto.edu

Fernando Augusto Poveda Aguja
Doctor en Education in Educational Technology
Docente Investigador – Corporación Universitaria Minuto de Dios – UNIMINUTO
Fernando.poveda@uniminuto.edu

Leidy Marcela Buitrago Márquez
© Magíster en Gerencia de TIC
Universidad ECCI
leidym.buitragom@ecci.edu.co

RESUMEN

Las tecnologías de la información y comunicación (TIC), desde su aparición han desempeñado un papel fundamental en la sociedad, desde que el hombre aplico su ingenio y sus diferentes habilidades de invención e innovación, la tecnología ha sido siempre una gran aliada para el continuo avance y desarrollo del mismo ser humano, es por esto que no es de extrañar que la tecnología este tan intrínsecamente asociada a diversos campos y aspectos de la salud.

Quántica
Ciencia con impacto social
ISSN: 2711-4600

El aprendizaje automatizado o de maquina es un campo de aplicación e investigación, derivado del desarrollo de las ciencias de la información y la computación combinados, denominado usualmente como machine learning (ML), los métodos de los que se componen el ML permiten de manera avanzada por medio de recursos computacionales, el tratado de datos a gran escala, con el fin de generar modelos de aprendizaje que retornen información relevante, datos estadísticos y posibles caminos de solución o aplicación para cambiar o predecir eventos o resultados.

Aplicando una metodología de revisión sistemática de literatura (SLR) se propone explicar cómo el machine learning puede ser usado para la prevención, diagnóstico e inclusive el tratamiento de enfermedades, describiéndolo como una potente herramienta tecnológica para mejorar el bienestar, tratamiento y promoción de la salud en el ser humano.

Palabras clave: Algoritmo, aprendizaje, predicción, enfermedad, base de datos

ABSTRACT

Information and communication technologies (ICT), since their appearance have played a fundamental role in society, since man applied his ingenuity and his different skills of invention and innovation, technology has always been a great ally for the continuum advancement and development of the human being himself, this is why it is not surprising that technology is so intricately associated with various fields and aspects of health.

Automated or machine learning (ML) is a field of application and research, derived from the development of information sciences and computing combined, the methods that make up the ML allow in an advanced way through computational resources, the treatment of large-scale data, in order to generate learning models that return relevant information, statistical data and possible ways of solution or application to change or predict events or results.

Applying a Methodology Systematic Literature Review (SLR) aims to explain how machine learning can be used for the prevention, diagnosis and even treatment of diseases, describing it as a powerful technological tool to improve well-being, treatment and health promotion in the human being

Keywords: Algorithm, learning, prediction, disease, database

1. Introducción

La medicina moderna en sus diferentes campos de investigación y aplicación de la salud humana, se vale de tecnologías de la ciencia de la computación para desarrollar métodos más confiables y seguros de diagnóstico de enfermedades, con estos avances la medicina usa métodos estadísticos que en base a un significativo grupo de datos pueden establecer patrones de comportamiento que ayuden a crear modelos predictivos para detectar enfermedades o patologías de determinado grupo poblacional, con estas mismas técnicas de detección se podrán trabajar en tratamientos preventivos con mayor eficacia, que ayuden a que el o los individuos no desarrollen las enfermedades, con mayor probabilidad de aparición, o que ayuden a bajar el nivel de impacto o afectación de estas en la salud de los pacientes; con la capacidad de procesamiento de los computadores actuales la medicina puede aplicar modelos estadísticos que permiten crear deducciones como pruebas de estimación, pruebas de hipótesis, pruebas paramétricas y no paramétricas, términos que más adelante serán explicados de forma simple, para dar a entender como a partir de bases de datos y algoritmos de aprendizajes, el machine learning transforma datos en conocimiento, y este conocimiento aplicado al diagnóstico y prevención de enfermedades, genera resultados altamente positivos y concretos en el sector salud.

Conceptos importantes y como se relacionan con el Machine Learning

Las tecnologías de la información y la comunicación (TIC) se forman de un gran número de ciencias, recursos, herramientas, equipos, programas informáticos, redes de información y comunicación, medios, aplicaciones, etc., que permiten a todas las escalas, actividades de compilación, procesamiento, almacenamiento, transmisión e información de voz, datos, texto, video e imágenes (Art. 6 Ley Colombiana 1341 de 2009) que orientadas a determinados campos de acciones pueden aportar grandes beneficios a la sociedad y aportes significativos a la ciencia.
Con lo anterior sabemos que las TIC en su diversidad de componentes ayudan al adelanto de investigaciones que en este caso en la medicina y en la salud tienen como propósito el diagnóstico y prevención de enfermedades, para esto una de esas muchas herramientas tecnológicas abarcadas en las TIC, el Machine Learning, combinado con diferentes fuentes de información establece un servicio predictivo que permita el diagnóstico o pronóstico.

Medicina

Según la organización mundial de la salud (OMS, 2017) la medicina es la suma total de conocimientos, habilidades y prácticas basados en teorías, creencias y experiencias

oriundos de las diferentes culturas, usados en el mantenimiento de la salud, así como en la prevención, diagnosis o tratamiento de las enfermedades físicas o mentales.

La medicina es entonces la ciencia de prevenir, cuidar y asistir en la curación de enfermedades, como uno de los principales objetivos de la medicina es la prevención, el machine learning se acopla con este propósito ya que a partir de modelos matemático puede predecir y prevenir a partir de datos genéticos, cito-genéticos (estudia el material hereditario principalmente por medio del ADN), clínicos, imágenes médicas (tales como: mamografías, ecografía, radiografía, resonancia magnética, tomografía axial computarizada, etc.), entre otros, enfermedades.

Salud

Según la definición que la OMS, hace del término desde el año 1946, es un estado de completo bienestar físico, mental y social, por tanto, la salud es la condición positiva de un individuo; mientras que se comprende como sector salud, la atención a las personas, un conjunto de valores, normas, instituciones y actores que desarrollan actividades de producción, distribución y consumo de bienes y servicios cuyos objetivos principales o exclusivos son promover la salud de individuos o grupos de población, también se puede ver al sector salud como un grupo de instituciones estatales, educativas, investigativas, públicas o privadas como por ejemplos corporaciones, fundaciones y universidades con y sin fines de lucro que aportan servicios, productos y atención que promueven la salud (Ministerio de Salud Colombiano, 2020).

Prevención (Concepto)

La prevención de la enfermedad es una estrategia de la atención primaria, que se hace efectiva en la atención integral de las personas, Por lo anterior se dice que la prevención implica promover la salud, así como diagnosticar y tratar oportunamente a un enfermo, (Escalante, P. 2004)

“Medidas destinadas no solamente a prevenir la aparición de la enfermedad, tales como la reducción de factores de riesgo, sino también a detener su avance y atenuar sus consecuencias una vez establecida” (OMS,1998).

Diagnóstico (Concepto)

Es el proceso de reconocimiento, análisis y evaluación de una cosa o situación para determinar sus tendencias, solucionar un problema o remediar un mal, en la salud se considera el diagnostico como el proceso mediante el cual son evaluadas, analizadas e identificadas las diferentes variables que influyen en los procesos salud y enfermedad de la población. (Escalante, P. 2004)

Esto se traduce en que a partir de datos o fuentes de información se analizan para evaluar cierta condición, con el machine learning se genera un proceso ordenado, sistemáticos para establecer de manera clara por medio de patrones las causas de algún padecimiento

o enfermedad y con esto conocer el estado de salud de un individuo o población. (Raffino M. 2020)

Factores de riesgo (Concepto)

Se denomina factor de riesgo a ciertas variables asociadas con la probabilidad del desarrollo de una enfermedad, pero que no son suficientes para provocarlas, estos factores se pueden relacionar a la edad, genero, herencia, etc., y existen otros factores susceptibles al medio o al comportamiento del individuo o población, como tabaquismo, obesidad, actividad física regular, etc. (Escalante, P. 2004). Los factores de riesgos son un insumo que alimentan las bases con las que interactuar lo sistemas de aprendizaje y en base a estos se puede crear modelos y patrones. (Escalante, P. 2004)

Conducta de riesgo (Concepto)

La OMS (1998) la define como la forma específica de conducta de la cual se conoce su relación con una susceptibilidad incrementada para una enfermedad específica o para un estado de salud deficiente, estas conductas de riesgo también son insumos para las bases de datos, que pueden arrojar indicadores claves de que conductas aumentan más el riesgo de enfermedad, como por ejemplo hábitos de alimentación de productos inadecuados, que podrían generar obesidad o diabetes, estas conductas sin una enfermedad asociada no darían un nivel de riesgo apropiado, por lo que es una relación de dos aspectos, la enfermedad y la conducta, en este sentido es más fiable el uso de herramientas de análisis como el ML, para generar todas esta relaciones y reducir posibilidad de omisión si se usara un análisis tradicional con apoyo humano al sistema, por lo que la aplicación del ML , cambia el enfoque del apoyo del sistema automatizado al análisis humano.

Estadística

De acuerdo al autor Cabria S. en su libro “La filosofía de la estadística” (1994), la estadística estudia el comportamiento de los fenómenos de los colectivos, o también denominado poblaciones, la estadística se caracteriza por una información acerca de un colectivo o universo, lo que crea un objeto, un modo propio de razonamiento, el método estadístico busca unas previsiones de cara al futuro, se basa en un ambiente de incertidumbre, es una rama de la matemáticas y utiliza el cálculo de las probabilidades, la estadística estudia fenómenos aleatorios intentando deducir leyes sobre los mismos.

Esta es una de las bases funcionales del machine learning, ya que aplica partes de modelos estadísticos para determinar relación entre los datos, como por destacar algunos muy comunes: porcentajes, tasas, proporciones, y algunas otras medidas de tendencia central como mediana moda promedio o desviaciones estándar.

Probabilidad

Se puede definir la probabilidad como un método por el cual se obtiene la frecuencia de un acontecimiento determinado, la probabilidad se maneja con eventos aleatorios, la

ISSN: 2711-4600

probabilidad constituye un importante parámetro en la determinación de las diversas casualidades obtenidas tras una serie de eventos esperados dentro de un rango estadístico, el ML apropia los métodos de probabilidad y trabaja sobre ellos para indicar que tan posible en escala de tiempo y número de veces en ocurrir puede presentarse determinado evento, esto contribuye con el modelo predictivo que se busca para la prevención de enfermedades, (Murphy, K. 2012).

Algoritmo

Se puede definir de forma simple, como un método conformado por una serie de pasos finitos para resolver una tarea o problema, la definición y construcción de un algoritmo es un proceso meticuloso que define cada paso sin generar ninguna ambigüedad en datos de entradas, procesos (funciones, operaciones, métodos, etc.) o datos de salidas, en informática un algoritmo se plantea por medio de un lenguaje de programación que se ejecuta en una máquina, el autor mexicano Juan Bernardo Vázquez, en su libro "Análisis y diseño de algoritmos" (2012) resalta una propiedad relevante de cualquier algoritmo en el campo de la informática y es que: "En cada problema el algoritmo se puede expresar en un lenguaje diferente de programación y ejecutarse en una computadora distinta; sin embargo, el algoritmo será siempre el mismo."

Es entonces que uno o un conjunto combinado de algoritmos define como va hacer el funcionamiento de terminado programa, los algoritmos son lo más importante en la programación de software y por tanto, al momento de desarrollar un sistema de machine learning es fundamental, que los algoritmos estén bien definidos y que sean en cierto modo capaces de aportar inteligencia al sistema, con esto último el ML posee una serie de Métodos de aprendizajes que se traducen con algoritmos y aportan al gran propósito del aprendizaje automatizado, más adelante en el desarrollo de este documento, se enumeraran y se describirán los métodos de aprendizaje pilares del ML.

Base de datos

Las bases de datos (BD), son repositorios o almacenes de datos que guardan relación entre sí, las bases de datos, son una forma organizada de almacenar datos para posteriormente entablar relación entre ellos, la función principal de una base de datos es mantener la integridad y seguridad de los datos, es decir que las bases de datos están en función de los datos, los datos son segmentos de información como por ejemplo, edad, dirección o nombre, y es hasta que los datos se unen que pueden brindar información, y con una análisis más profundo de esta información se puede generar conocimiento (A. Silberschatz, H. F. Korth & S. Sudarshan 2002).

Modelado de datos

En el manejo de grandes volúmenes de datos, como lo debe de hacer el machine learning, y más cuando se aplica aun área de investigación tan extensa como lo es la

salud y la medicina, el modelado de datos es crucial, ya que reduce el tiempo de búsqueda de determinado dato en una base de datos, el modelado de datos es una forma de describir el proceso de diseño de las BD (bases de datos).

Un modelo de datos es un conjunto de conceptos y prácticas utilizadas para organizar los datos de interés y describir su estructura en forma comprensible para un sistema informático, aporta significativamente en el rendimiento y agilidad de los procesos de asociación y consulta que debe realizar el ML, el modelamiento de datos ayudara para la tarea de agrupación y clasificación, en este sentido los algoritmos de ML, al extraer datos de la base de datos, podrán de forma más simple, hacer agrupaciones de poblacionales y clasificarlas de acuerdo a los factores y conductas de riesgo en una categoría de probabilidad de ser diagnosticados con una enfermedad y con esos resultados almacenarlos de acuerdo al modelo definido en la BD, y así ir aumentado su banco de datos, (Ordoñez, M., Tapia, J. & Asanza, W. 2015).

Minería de Datos

Con el fin de desarrollar máquinas de aprendizaje automático, los datos son importantes para el proceso de enseñanza, por lo que la minería de datos es fundamental para poder determinar cómo explotar la información de grandes bases de datos; los algoritmos para la minería de datos llevan a cabo tareas de tipo descriptivas es decir, con fin de hacer reconcomiendo de los datos es decir descifrar la relación entre ellos y su significado individual y en conjunto, como el descubrimiento de patrones, o tareas predictivas, como la clasificación o el ajuste de modelos que permitan predecir el comportamiento, los algoritmos para minería de datos se forman principalmente de: un modelo de datos, reglas o criterios de selección, criterios de satisfacción que indica que es correcta la selección, o funciones de optimización del modelo, todos estos componentes se suman en un algoritmo de búsqueda, cuyo primer paso suele ser el de la iniciación de todos los parámetros (campos de datos) del modelo a valores de estado inicial, normalmente aleatorios, para luego ajustar iterativamente el modelo según el criterio de satisfacción. (Benítez, I. & Diez, J. 2005)

En la siguiente tabla se podrán visualizar los objetivos de la minería de datos en el ML

Tabla 1 *Tabla de objetivos de la Minería de Datos*

Objetivos		
Análisis de secuencia	**Clasificación**	**Agrupación**

ISSN: 2711-4600

Tratamiento a las secuencias de datos, análisis y comparación de valores característicos en secuencias de datos, como los valores medios, medianas, desviación típica, varianza, etc.	Clásica los objetos, entre un rango de categorías, algoritmos que generar reglas de clasificación de datos	Agrupan los objetos según similitud de características, formando conjuntos o clases, los algoritmos de clasificación no agrupan los datos, sino que los clasifican uno a uno
Asociación	**Dependencia de modelo**	**Predicción**
El objetivo es el de descubrir relaciones ocultas entre los objetos, o incluso entre los propios atributos de los objetos, de los cuales se puede extraer una base de reglas, con estructura condicional	Describen relaciones de dependencia entre variable y el grado de relación entre dichas variables	Entrenamiento de los modelos, con el fin de validar hipótesis de comportamientos futuros
Regresión	**Summarización**	**Visualización de modelo**

A partir de muestras de datos, se espera estimar un modelo que pueda establecer relaciones de dependencia de ciertas variables respecto de otras, con el fin de poder predecir valores a partir de nuevos datos	El objetivo es generar descripciones globales de conjuntos de datos. en algunos casos estas descripciones son cualitativas. Se suelen usar para la extracción de información de textos	Adecuar y reinterpretar los datos para que sean visual-mente entendibles y se puedan extraer conclusiones de un vistazo. Comprende todo tipo de gráficos, como histogramas, gráficos en coordenadas
Análisis exploratorio de datos		
Dado un conjunto de datos del cual se desconocen sus posibles interdependencias y relaciones de similitud, estas técnicas tratan de identificar patrones de forma visual y sin ninguna estructura de búsqueda o semejanza preconcebida o sin conocimiento previo de los datos.		

Fuente: Adaptado de Benítez, I. & Diez, J. (2005). Técnicas de agrupamiento o reconocimiento de patrones

Inteligencia artificial

El aprendizaje de maquina o machine learning, se considera como una rama de la inteligencia artificial (IA), La inteligencia artificial se basa a los medios de simular las capacidades de inteligencia del cerebro humano, y a las conductas del ser humano. (Badaró, Ibañez, Agüero, 2013). Grandes Corporaciones a nivel mundial y con un fuerte posicionamiento en el sector TIC, han aportado grandes desarrollos en el campo de sistemas de IA con fuertes componentes como análisis de datos como, IBM, Microsoft, Amazon, Facebook, Google y entre otras grandes empresas. (Marr. B. 2016)

ISSN: 2711-4600

¿Qué es el Machine Learning?

El aprendizaje automático o Machine Learning es un método científico que nos permite usar los ordenadores y otros dispositivos con capacidad computacional para que aprendan a extraer los patrones y relaciones que hay en nuestros datos por sí solos. Esos patrones se pueden usar luego para predecir comportamientos yen la toma de decisiones.

De acuerdo a Hurwitz y a Kirsch (2018) se describe al aprendizaje automatizado (en inglés al Machine Learning - ML) como una forma o derivación de la inteligencia artificial, que permite que un sistema aprenda a partir de datos en lugar de una programación directa y descriptiva, que detalla a exactitud lo que se desea, para esto, el aprendizaje automático utiliza una gran cantidad de algoritmos que se ejecutan de forma reiterativa.

Otra definición dada por Murphy (2012) en su libro "Machine Learning: A Probabilistic Perspective" indica que el ML es un conjunto de métodos capaces de detectar automáticamente patrones en los datos y usarlos para predecir sobre datos futuros o para ayudar en la toma de decisiones en un entorno de incertidumbre, en su libro Murphy resalta ciertos componentes del aprendizaje automático definidos por primera vez por el profesor Tom Mitchell (1997):

- **Fuentes de información**: indican la experiencia, que en el modelo básico del ML se denomina con la letra "E", estas fuentes se alimentan de:
  - Datos estructurados: Datos configurados de tal forma que extraer información de ellos es simple, se presentan de forma ordenada y es fácil

su análisis, algunos ejemplos son: bases de datos relacionales y sistemas de ficheros como los encontrados en archivos de extensión xml o json

- Datos no estructurados: son datos que no siguen un modelo de organización claro, no son fáciles de relacionar entre si e implican mayor complejidad de asociación y clasificación, ejemplo: voz, imágenes, videos, textos, etc.

- **Técnicas y algoritmos**: indican las acciones o tareas a realizar, para el tratamiento de la experiencia, se denomina con la letra "T", se resaltan algunas técnicas de ejemplo:
  - Técnicas para el tratamiento de la información no estructurada: según Magnus Stensmo y Mikael Thorson (2003), se cita textualmente "La Gestión de la Información no Estructurada consiste en las herramientas y métodos necesarios para almacenar, acceder, recuperar, navegar y generar conocimiento primordialmente de la información basada en texto",
  - Métodos basados en la cantidad de supervisión humana en el proceso de aprendizaje como por ejemplo los modelos supervisados y no supervisados, se podrían describir como algoritmos de asociación entre las muestras de datos de entrada y sus salidas correspondientes, sea de forma automática o tradicional.
  - Métodos basados en la capacidad de aprender a partir de muestras de datos incrementales.
- **Capacidad de autoaprendizaje**: es la medida de rendimiento o desempeño, se expresa con la letra "P", que se podría medir a determinados indicadores, como el entretenimiento automático y cíclico del sistema o maquina a parir de nueva información, y como en sí mismo el modelo se reajusta o calibra para disminuir la probabilidad de error en los resultados.

Un componente adicional aplicado a las tecnologías de la información, es el uso de sistemas y software, que por medio de algún lenguaje de programación se crean los algoritmos y así se pueden visualizar los resultados, por ejemplo, el uso de lenguajes como: R, Python, scala, SQL, Matlab, etc., y de plataformas de visualización como Power BI, Tableau, SAS visual Analytics, etc.

Tendencias en Aprendizaje Automático (ML)

Con el uso de diferentes fuentes de información junto con la variedad existentes de técnicas y algoritmos capaces de aprender a partir de los datos, se pueden obtener varios beneficios como (Utrera, R. 2017):

- Generación de nuevos datos que permiten aumentar el tamaño de la base original de información

- Métodos más eficientes de procesamiento de información que ayudan a la toma de decisiones.
- Detección de patrones
- Proceso automatizado de modelamiento de información y aprendizaje del sistema o máquina, que permiten un mayor nivel de predicción, con esto se aumenta la ventaja frente a los sistemas tradicionales de información que requieren de una intervención humana constante y especializada en el campo de estudio para generar resultados.

Antecedentes

En la revisión histórica de la evolución de la TIC y más propiamente de la Inteligencia Artificial y del Machine Learning, se puede ver como pequeños y grandes avances han contribuido a los sistemas inteligentes que hoy en día se ponen a pruebas en área como la salud, para no remontarnos tan atrás, un antecedente importante fue generado por el matemático inglés Alan Turing (1912-1954) quien propuso una prueba con el objetivo de demostrar la existencia de "inteligencia" en un dispositivos no biológico, la prueba es conocida como el "test de Turing".

El Test de Turing se basa en la hipótesis de que "si una máquina se comporta en todos aspectos como inteligente, entonces debe ser inteligente" (Alan Turing, 1950), esta prueba llamo el interés de la comunidad científica hacia el concepto de máquinas inteligentes. Dos de las contribuciones más importantes de Alan Turing son el diseño de la primera computadora capaz de jugar al ajedrez y el establecimiento de la naturaleza simbólica de la computación.

El test de Turing: en el Libro Inteligencia Artificial (2014) indica que el test intenta ofrecer una situación de Inteligencia Artificial que se pueda evaluar. Para que un ser o máquina se considere inteligente debe lograr engañar a un evaluador de que este ser o máquina se trata de un humano, evaluando todas las actividades de tipo cognoscitivo que puede realizar el ser humano, es decir que una persona debe establecer si sostuvo una charla con una máquina o con otra persona.

ELIZA: En el año 1965 Joseph Weizenbaum construyo el primer programa interactivo el cual consistía en que un usuario podía sostener una conversación en ingles con una computadora utilizando una comunicación por escrito.

ChatBots: Son una representación más sutil de la IA combinado en un menor grado con el ML, son asistentes virtuales, que pretenden entablar una conversación humana a través de una plataforma de mensajes instantáneos (ejemplo Whatsapp), un bot en es un software que automatiza una tarea sencilla, de recibir un mensaje y devolver una respuesta, aplican algoritmos de ML para aprender las diferentes posibilidades de mensajes de entrada y mensajes de salida que deben devolver en la conversación (Murphy, K. 2012). El primer chatbot se estima que fue desarrollado en la década de los

60 por el alemán Joseph Wiezenbaum en el laboratorio de inteligencia artificial del Instituto Tecnológico de Massachusetts (MIT).

Escuelas de la IA: se formaron dos grandes "escuelas" de IA, el primero liderado por Newell y Simon de la Universidad de Carnegie-Mellon, proponiéndose desarrollar modelos de comportamiento humano con aparatos cuya estructura se pareciese lo más posible a la del cerebro (Newell, A. & Simon H. A. 1961); el segundo formado por McCarthy y Minsky en el Instituto Tecnológico de Massachusett (MIT), centrándose más en que los productos del procesamiento tengan el carácter de inteligente, sin preocuparse por que el funcionamiento o la estructura de los componentes sean parecidas a los del ser humano (McCarthy J. 1960).

Métodos de Machine Learning

Se pueden establecer 3 grandes métodos, que enmarcan los algoritmos que se pueden emplear, es decir que un proyecto de ML, puede utilizar uno o varios métodos y así uno o varios algoritmos de aprendizaje, para poder analizar los datos consumidos de las Bases, en la Figura 5 se listan los métodos conocidos.

Método de Clasificación: La clasificación es una subcategoría de aprendizaje supervisado donde el objetivo es predecir las etiquetas de clases categóricas de las nuevas observaciones, basada en observaciones pasadas.

Método de Regresión: En el Análisis de Regresión, dada una serie de variables predictorias (explicativas) y una variable de respuesta continua, se requiere encontrar una relación entre esas variables que nos permita predecir un resultado. Con este conocimiento, se puede predecir respuestas para observaciones nuevas no conocidas, proceso que es similar a la clasificación, pero con salidas numéricas continuas.

Método de Agrupación o Clustering: en la investigación de Benítez, Ignacio & Diez, Jose Luis. (2005) sobre técnicas de agrupamiento de datos, se indica que son métodos de aprendizaje automático que intentan encontrar patrones de similitud y relaciones entre muestras de un conjunto de datos, y luego agrupan estas muestras en varios grupos, de modo que cada cluster o grupo de muestras de observaciones tiene cierta similitud, según los atributos o características inherentes. Este método posee algunos submétodos, que permiten elegir que algoritmo es mejor para el análisis de la población objeto, esto de acuerdo al tipo de datos que se posean.

- Métodos basados en centroides: se emplea el K-means (es un método que tiene como objeto la partición de un conjunto de n observaciones en K grupos, en el que cada observación pertenece al grupo cuyo valor medio es más cercano)
- Métodos de agrupamiento jerárquico: Aglomeración y División (algoritmos de Ward, propagación de afinidad).
- Métodos de agrupación basados en distribuciones: Modelos de mezclas gaussianas.

- Métodos basados en densidades: se basa en la detección de en qué áreas existen concentraciones de puntos y dónde están separados por áreas vacías o con escasos puntos. Los puntos que no forman parte de un clúster se etiquetan como ruido

Método de Aprendizaje Reforzado: se utiliza cuando no hay muchos datos disponibles, pero existe un entorno con el cual el sistema puede interactuar, aquí el concepto de Agente toma mayor relevancia, en el libro "Inteligencia Artificial" (2014) los autores definen que una agente es una entidad física o virtual, es capaz de percibir el entorno y tener una representación parcial del mismo; es capaz de actuar sobre el entorno; puede comunicarse con los otros agentes (pueden ser humanos o no) Según Norvig & Russel (Russel & Norvig 2009) "un agente inteligente es aquél que puede percibir su ambiente mediante sensores y actuar sobre ese mundo mediante efectores (o actuadores)".

Aprendizaje por Lotes: también llamado aprendizaje offline, el modelo se entrena con todos los datos disponibles, lo que se hace antes de realizar ningún tipo de predicción. Si los datos de entrenamiento cambian -por ejemplo, si se reciben nuevos datos-, el modelo deberá ser reentrenado desde cero.

Aprendizaje en línea: El modelo se entrena incrementalmente con cada una de las nuevas muestras que se reciban, o en grupos pequeños de muestras llamados mini-batches. Estos sistemas son más adecuados en entornos en los que los datos a partir de los que se entrena el algoritmo cambian con cierta rapidez. También son especialmente útiles en escenarios en los que no es posible cargar todos los datos en la memoria del ordenador.

Aprendizaje basado en instancias: En este tipo de aprendizaje, se almacenan los ejemplos
de entrenamiento y cuando se quiere clasificar un nuevo objeto, se extraen los objetos más parecidos y se usa su clasificación para clasificar al nuevo objeto.

Aprendizaje basado en Modelos: En el aprendizaje basado en modelo el objetivo es la creación de un modelo a partir de los datos de entrenamiento, modelo que servirá para realizar predicciones posteriormente. En este enfoque, una vez creado el modelo, los datos de entrenamiento son descartados. Un ejemplo de esto es una regresión lineal: a partir de los datos de entrenamiento se crea el modelo (por ejemplo, una recta de regresión con la fórmula matemática básica: ax + b), y será esta recta (y no los datos de entrenamiento) la que se utilizará para predecir las etiquetas de las nuevas muestras.

Arboles de decisión

Los arboles de decisión son clasificadores, que utilizan una estructura de árbol para modelar la relación entre las características del modelo y los posibles resultados, a partir de los árboles de decisión, generados por expertos, se puede crear un algoritmo potente de ML. Los expertos identifican algunas ventajas y desventajas de los árboles de

decisión, como ventajas se tienen que son sencillo de armar y de entender, no requieren de una base de datos extensa, puede manejar datos numéricos y categóricos, mientras que algunas desventajas son: se puede llegar a crear árboles que no generalicen ni clasifiquen bien los datos, variaciones en los datos de entrada pueden generar grandes cambios en los resultados esperados, se requieren métodos heurístico para formarlos, es decir del apoyo de un experto en el tema a tratar, como por ejemplo de un médico cardiólogo que ayude en la formación del árbol de decisión para saber de acuerdo a varios criterios que tratamientos se deben seguir para patologías cardiacas.

Desarrollo de la investigación – Machine learning en la salud

Caso Práctico

Con bases clara de cómo funciona el ML, y en que consiste, se pueden plantear un escenario ficticio del uso del ML en el sector salud, para esto se utilizara el ejemplo dado por el consultor en informática medica Juan Ignacio Barrios Arce en su artículo "Inteligencia Artificial y salud… un caso práctico" (2019), miembro editor de la plataforma Health Big Data.

Escenario plantado: Un grupo de 80 pacientes con características demográficas variadas (Sexo, Edad, Presencia de Hipertensión Arterial, Diabetes, Alcoholismo, tabaquismo y Obesidad) los pacientes habían sido clasificados de previo por haber sufrido o no isquemias del miocardio o bien Infartos.

Definir Variables categóricas, como se ven en la Tabla 2.

Tabla 2 Definición de variables categórica – caso practico

Categorías	Variables
Cardiopatía	1 sano o leve; 2 grave o moderado
Sexo	1 = mujer; 2 = hombre
Fuma	1= SI; 2 = no; 3 = ex fumador; 0 = no se sabe
Hipertensión arterial	0= no consta; 1 = sí; 2= no
Diabetes	0= no consta; 1 = sí; 2= no
Obesidad	0= no consta; 1 = sí; 2= no
Edad	Edad en numero

Grupo de edad	1= Joven; 2 = edad fértil; 3 = adulto; 4= persona mayor

Fuente: Adaptado de https://www.juanbarrios.com/inteligencia-artificial-y-salud-un-caso-practico-borrador/ (Barrios, J. 2019)

Con el set de datos cargados en el sistema, se deben de escoger los algoritmos de ML, para analizar la información, para este caso será el Algoritmo de árbol de decisión, que hace parte del modelo de aprendizaje supervisado.
Con este algoritmo se deben de ajustar 4 parámetros S:
min_samples _split: Es el número de muestras antes de dividirse el árbol en una nueva rama.
min_samples _leaf: Es el número de muestras minino que debe tener cada hoja al final = 5 para este caso tratándose de que son muy pocos casos.
max_depth: Es el número de niveles que tendrá el árbol = 2 para este caso, tratándose de que son muy pocos casos.
class_weight: Se refiere al desbalance que hubiese que corregir por desbalance entre las clases en nuestro caso 1:1.633, (esto significa el porcentaje de desviación de resultados por un numero de datos muy bajo).
Cuando todo lo anterior se aplica en una máquina, utilizando un lenguaje de programación, se deberán de invocar recursos de programación (como librerías y funciones) para que el algoritmo de aprendizaje quede bien definido y sea funcional, en este caso el resultado de la aplicación del Machine Learning, es el siguiente diagrama de decisión, que muestra gráficamente cómo será el procedimiento de decisión que haría la máquina. Ver Figura 7.
Explicación de resultado: Como se aprecia en el diagrama que simboliza el procesamiento de la máquina, a partir de los datos de una determinada población, se determinan ciertos datos relevantes y que serían factores de riesgo y conductas de riesgo para determinar por agrupación, la posibilidad de que los individuos posean un infarto agudo, de esta forma se simula la detección temprana de afecciones cardiacas.

Métodos de validación de sistemas con ML
Como todo proyecto de investigación que se desarrolla y se prueba en un ambiente de laboratorio, los sistemas de ML, deben ser validados, para certificar su funcionamiento, para lo cual existen diferentes técnicas, usadas también por lo general en pruebas de software de cualquier tipo, descritas como:

- **Pruebas de Estimación**: Estimar qué va a ocurrir respecto a algo o qué está ocurriendo, o qué ocurrió en el pasado, a pesar de ser un elemento muy claramente

estadístico, es un proceso complejo, en el análisis de datos, las pruebas se soportan con la estadística, es por esto, que estas pruebas validan los resultados del sistema de ML, estimar es establecer conclusiones sobre características poblacionales a partir de resultados muéstrales, y que dichos resultados sean semejantes o iguales a los que el sistema genero por medio de los algoritmos de aprendizaje. (Gardner, M., & Altman, D. 1993)

- **Pruebas de Hipótesis**: De acuerdo al libro "Metodología de la investigación, bioestadística y bioinformática en ciencias médicas y de la salud, 2e" sus autores indican que la hipótesis de investigación que se genera en todo proyecto, se define como la postulación de lo que se busca o se trata de probar, la decisión relaciona la elección entre dos enunciados o afirmaciones competitivas y mutuamente excluyentes, respecto de uno o más parámetros de la población.

Para pasar una prueba de hipótesis, el resultado generado por el sistema ML debe ser lo suficientemente cercano al valor hipotético, como para aceptar la hipótesis planteada, de lo contrario se rechazará.

Se resaltan los atributos principales que debe poseer una hipótesis:

1. Debe situarse en una situación real.
2. Las variables que se presentan en la hipótesis deben ser precisas, comprensibles y concretas.
3. Las relaciones entre las variables deben ser claras, admisibles y lógicas.
4. Los términos y las relaciones planteadas deben ser observables y medibles.

- **Pruebas Paramétricas**: (Murphy 2012) son una herramienta de la probabilidad, que se usa para analizar todos los factores de una población, mientras más grande sea la muestra más exacta será la estimación, mientras más pequeña, más distorsionada será la media de las muestras, sus ventajas es que da estimaciones probabilísticas bastante exactas, más eficientes y menos posibilidad de error, pero su desventaja en que son pruebas más complejas y solo se pueden aplicar a ciertos grupos de datos, la idea de esta prueba es validar que las asociaciones, clasificaciones o agrupaciones y que hace el sistema de ML respecto a una población de acuerdo a sus características (como factores de riesgos o conductas de riesgo) sean las más acertadas de acuerdo a los resultados paralelos de esta prueba.
- **Pruebas no Paramétricas**: (Murphy 2012) Las pruebas no paramétricas permiten analizar datos en escala nominal (se refiere a que las variables se diferencian por sus nombres) u ordinal (se refiere al orden de las variables en este

ISSN: 2711-4600

sentido las primeras a analizar son más críticas que las ultimas, o en sentido contrario), se usa en apoyo a las pruebas de hipótesis, su ventaja es que se pueden aplicar aun cuando no se den todas las condiciones para validar con exactitud una variable o parámetro poblacional, y al igual que las paramétricas buscan medir la eficiencia del sistema ML para genera asociaciones.

Estadísticas de apropiación del Machine Learning

El machine Learning desde su inicio a tenido gran acogida en varios sectores no solo de investigación, sino también en mercados laborales, por aplicación en diferentes áreas de aplicación, como medicina, finanzas, seguridad informática, marketing, entre otras, en la Figura 8, se evidencia Según el Artificial Intelligence Index Report 2019, informe anual del Human-Centered Artificial Intelligence Institute de la Universidad de Stanford, un interés importante en la contratación de profesionales con conocimiento y manejo del Machine Learning en Estados Unidos, donde la contratación aumento del 0,07% en el 2010 a más de un 0,51% en el 2019

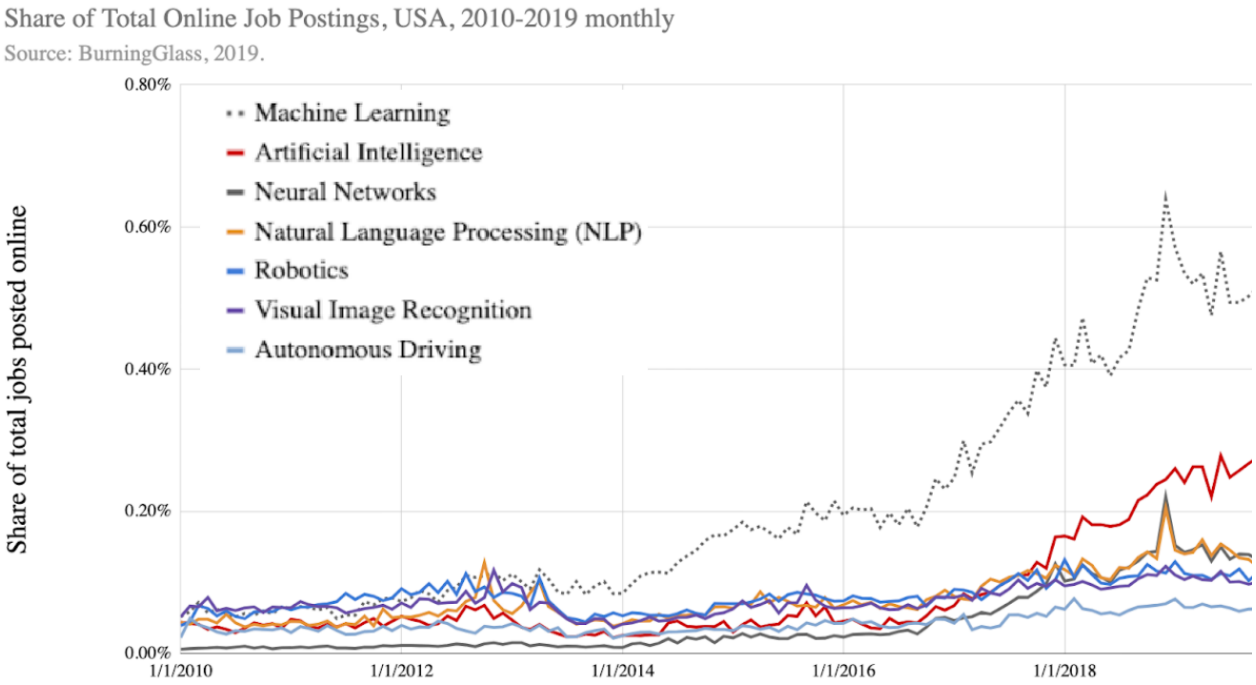

Figura 1 *Crecimiento de las ofertas de empleo relacionadas con inteligencia artificial en Estados Unidos desde 2010*

Fuente: *adaptado de Artificial Intelligence Index Report 2019*

Adicional en reporte del 2019 de la ACCA (the Association of Chartered Certified Accountants), que estudia el impacto y conocimiento del Machine learning en varios países, en el reporte se indica los resultados de una encuesta generada en el 2019 del estado de la adopción del ML, los encuestados expresaron diferentes niveles de comodidad (Figura 9) al hacer decisiones basadas en ML en áreas como: clasificación de transacciones (53%), Medición contable (47%), pruebas de auditoría (43%) y detección de fraude (41%). Sin embargo, hubo menos comodidad en ciertas aplicaciones donde su uso es más amplio, como con datos médicos o de finanzas personales.

Y siguiendo la tendencia anterior, el reporte más reciente de la universidad de Stanford en su "Artificial Intelligence Index Report 2021" indica que existen cada vez más profesionales que estudian y se especializan en ML y en IA, entre todos los graduados de doctorado en ciencias de la computación en 2019, aquellos que se especializaron en inteligencia artificial / aprendizaje automático (22,8%), teoría y algoritmos (8,0%) y robótica / la visión (7,3%) encabeza la lista. La especialidad AI / ML ha sido el más popular en la última década.

Proyectos nivel internacional y nacional

En el sector salud, el machine learning se aplica en proyectos tales:

- **Transformación digital del sistema de salud colombiano**, de acuerdo a Christian Peter Clausen, asesor de Transformación Digital del Ministro de Salud y Protección Social de Colombia (2020), el ministerio de salud busca adaptar el sistema de salud a un mejor esquema de gestión apoyado en varias tendencias tecnológicas actuales, el proyecto denominado Agenta TD-Salud, busca tecnologías en salud orientadas al ciudadano, con el objetivo de integrar las siguientes tecnologías por etapas: Interoperabilidad de la Historia Clínica, Facturación Electrónica, Telesalud, análisis de Big Data, Ventanilla Única Electrónica, "AI Hospital" – inteligencia artificial hospitalaria, Autorizaciones Electrónicas, y desarrollo de capacidades de modelos de análisis soportados en Machine Learning, con esto busca atreves de los datos prestar un mejor servicios a los colombianos, crear procesos más eficientes, óptimos y confiables.

Adicional el impacto de esta transformación ayudara a la prevención de la Salud de personas mayores, de acuerdo a González M. (2019) la importancia de la transformación digital en el desarrollo de un diagnóstico médico de forma predictiva permitiría un tratamiento y seguimiento a tiempo, los servicios que brindan actualmente los sistemas de salud, deben ser un proceso de innovación constante, los estudios de genética serán necesarios para evaluar los pacientes mayores, y debe ser posible que con anticipación se establezca un diagnóstico temprano en la prevención de enfermedades crónicas y cáncer.

ISSN: 2711-4600

- **Plataforma de análisis y predicción de datos del Covid-19 en Colombia**: (Universidad Católica de Colombia, 2020) el proyecto desarrollo por un grupo de investigación colombiano, han desarrollado el servicio tecnológico para el análisis y visualización de datos de salud pública y registros hospitalarios en la pandemia COVID-19 apoyados en el Machine Learning y en la Inteligencia Artificial, "El cual permite la construcción en tiempo real de modelos predictivos para la detección temprana de casos sintomáticos y asintomáticos, monitoreo e identificación de medidas de mitigación del contagio y formulación de posibles medidas adicionales susceptibles de optimizarse a partir de la analítica predictiva", el proyecto hizo uso de los portales de datos públicos de la ciudad de Bogotá como son SALUDATA (plataforma web de información abierta, acerca de la situación en salud de Bogotá).

- **La empresa internacional SAAS**, es un ejemplo de utilización del ML no solo como investigación sino como herramienta para un modelo de negocio rentable, la empresa de origen estadounidense fundada en 1966, líder en analítica con presencia en 149 países, cuyo servicio es apoyar a las empresas de diferentes sectores, entre ellos el sector salud, a analizar por medio de los datos soportados en tecnologías de minería de datos, business intelligence, y con el machine learning, fortalecer los procesos de las empresas para obtener mejores resultados, en la industria de la medicina aporta soluciones para: Salud de las poblaciones (Integra datos de salud y no relacionados con la salud para guiar la atención integral de las personas, así como los programas comunitarios que reducen las disparidades en la salud. Mejora en base a los datos la calidad de la atención y los resultados de salud integrando análisis dentro de su vista de 360 grados de pacientes, miembros y clientes) posee 3 ejes de análisis: datos clínicos de los pacientes, seguridad de los pacientes (aplicación de predicción de medicamentos, riesgos de infección y predicción y prevenciones de enfermedades evitables) y cuidado general de las personas (pronóstico de demanda de servicios de atención en poblaciones de alto riesgo).
- **En el Hospital Infantil de Cincinnati** (Ohio, EEUU), trabajaron en un proyecto de investigación del ML aplicado al fenómeno del suicidio, donde por medio de algoritmos de clasificación y regresión que aplicaron a los datos obtenidos por medio de encuestas para clasificar a los tres grupos que existían: personas que habían cometido (o con riesgo de) suicidio, personas con trastornos mentales sin riesgo de suicidio y un grupo control (formado por personas sin riesgo de suicidio ni enfermedades mentales) de identificar hasta con un 93% de precisión si una persona posee riesgo de suicidio o no (Jiménez J. 2016).
- **Medicina Intensiva y atención médica**, en el artículo de revisión "Big Data Analysis y Machine Learning en Unidades de Cuidados Intensivos" (Núñez A.,

Armengol M.A. & Sánchez M. 20219) se resalta el uso del ML y el análisis de datos como herramienta para la mejora de investigación clínica y dirigir de manera más precisa las terapias proporcionadas a los diferentes pacientes, al usar esta tecnología se pueden generar nuevas estrategias de atención en base a los datos clínicos de los pacientes y como el análisis retrospectivo puede apoyar los ensayos clínicos en poblaciones de pacientes que comparten santos y enfermedades similares para hallar las causas de las enfermedades.

- **Diagnóstico de enfermedades cancerígenas**: como se puede resaltar en la investigación denominada "Clasificación asistida por computadora de nódulos pulmonares en imágenes de tomografía computarizada mediante una técnica de Deep learning" (Hua K., Hsu C., Hidayati S., Cheng W., Chen Y., 2015), técnicas de aprendizaje automatizado pueden ser empleadas para el pronóstico de enfermedades tales como el cáncer, por medio de racionamiento de patrones en tomografías, en el estudio realizado en el 2015, se diseñó un esquema de diagnóstico asistido por computadora convencional (CAD - conventional computer-aided diagnosis), que requirió de varios pasos de procesamiento de imágenes y reconocimiento de patrones para lograr un resultado de diferenciación tumoral cuantitativa, Los resultados experimentales sugieren que los métodos de aprendizaje profundo podrían lograr mejores resultados discriminativos y ser prometedores en el dominio de la aplicación CAD. Para el estudio se usaron gran cantidad de tomografías computarizadas, para hacer el diagnostico de nódulo pulmonar (que pueden ser formaciones anómalas que podrían ser cancerígenas) los cuales pueden ocurrir en cualquier parte del pulmón, lo que hace que el diagnóstico clínico sea una tarea complicada, La clasificación de los nódulos pulmonares con técnicas de aprendizaje se probó en el conjunto de datos del Lung Image Database Consortium, que incluye 1010 pacientes recopilados de Weill Cornell Medical College, University of California en Los Ángeles, University of Chicago, University of Chicago, University of Iowa y University de Michigan.

Los anteriores son solo algunos ejemplos de la aplicación del machine learning en el sector salud, ya que realmente existen cientos de estudios publicados en las diferentes bases y revisaste de medicina, como por ejemplo los cerca de 62,629 resultados en la búsqueda de la base de consulta PubMed, desde comportamientos predictivos de virus como el COVID-19, en anestesiología, probabilidad de obesidad, diagnostico de patologías, oftalmología, psicología, descubrimiento de fármacos, propagación de enfermedades, etc..

ISSN: 2711-4600

Metodología de investigación de los proyectos de ML

En revisión de múltiples proyectos de investigación de Machine, se pueden detectar ciertos principios fundamentales en los proyectos de ML que deben de cumplirse, de acuerdo a (González M. 2019) es su tesis "El impacto de la transformación digital en la prevención de la salud
de personas mayores en Colombia", indica que los efectos de las aplicaciones tecnológicas en Medicina deben ser:

- Eficaces: Efecto procedente en la variable a evaluar cuando la intervención se aplica en condiciones ideales.
- Efectivos: El resultado obtenido es favorable por la aplicabilidad del proyecto, como casos diagnosticados, vidas salvadas, años de vida ganados, calidad de vida del paciente, entre otros.
- Utilidad: es la aplicabilidad del proyecto en la realidad, como al aplicar la tecnología la calidad de vida del paciente mejora
- Beneficio: se relaciona al beneficio económico que el desarrollo genera, sea en nuevos ingresos o reducción de costos y gastos.
- Excelencia: Es la obtención de los mejores resultados con el mínimo de gastos posibles para satisfacción, tanto del paciente como del personal de salud, al realizar correctamente la tarea que corresponde y ahorrar recursos (tiempo, dinero, esfuerzos, etc.) que puedan emplearse en producir nuevos servicios.

Una metodología de investigación y desarrollo es la usada por la Fundación Pfizer, a través del proyecto "Mendelian", identifica todas las enfermedades genéticas inusuales (centrarse en el problema o necesidad), selecciona las bases de datos en donde obtendrán los datos para alimentar el sistema (recopilación de datos), realiza un mapeo (modelado de datos) y a través de algoritmos (selección, diseño y programación de algoritmos de ML), búsqueda de patrones (enseñanza del sistema, procesos reiterativos de análisis) para buscar variantes para identificar el diagnostico de forma acelerada en aproximadamente 7 años (resultados, evidenciar aquellos criterios o condiciones para generar diagnósticos predictivos), esta es una tarea compleja, que los equipos de investigación deben realizar paso a paso y de manera íntegra entre diferentes profesionales, realizar múltiples pruebas para que el sistema aprenda y se disminuyan las probabilidades de error.

Como lo destaca (Bhatt M. 2016) los proyectos de investigación de ML, traen consigo varios retos y desafíos que hay que contemplar en todas las fases de planeación y desarrollo, que son:

1. Heterogeneidad de los datos: muchos algoritmos, requieren que los datos sean de tipo numérico y a su vez estén normalizados y homogéneos (por lo que al momento de trabajar con una fuente o una base de datos lo primero será hacer una limpieza y tratamiento de dichos datos). Los algoritmos que emplean métricas de distancia son muy sensibles a esto, los algoritmos de árboles de decisión pueden manejar datos heterogéneos con mucha facilidad.
2. Redundancia de datos: si los datos contienen información redundante, es decir, contienen valores altamente correlacionados, entonces es inútil utilizar métodos basados en la distancia debido a la inestabilidad numérica. En este caso, se puede emplear algún tipo de norma o regla de los datos para prevenir esta situación.
3. Características dependientes: si existe alguna dependencia entre los vectores de características (relaciones entre los campos de la BD), los algoritmos que monitorean interacciones complejas como las redes neuronales y los árboles de decisión funcionan mejor que otros algoritmos, por lo que se deberá hacer una selección cuidadosa de que método de enseñanza o algoritmo de aprendizaje, para hacer un sistema optimo y que no consuma recursos excesivos de la máquina.
4. Compensación de sesgo-varianza: un algoritmo de aprendizaje está sesgado para una entrada particular, si cuando se entrena el sistema en cada uno de estos conjuntos de datos de entrada, es sistemáticamente y continuamente errada la información, al predecir la salida se tendrán datos incorrectos, una característica clave de los algoritmos de aprendizaje automático es que pueden ajustar el equilibrio entre el sesgo y la varianza automáticamente, o mediante un ajuste manual utilizando parámetros de sesgo, y el uso de dichos algoritmos resolverá esta situación, entonces en la fase de entrenamiento del sistema se debe prestar minuciosa atención a los datos de entrada con única o múltiples salidas para que los errores se minimicen o se corrijan.
5. Maldición de la dimensionaldad: si el problema tiene un espacio de entrada que tiene una gran cantidad de dimensiones es decir capas de aplicación para un solo set de datos pequeño, el algoritmo de aprendizaje automático puede confundirse por la gran cantidad de dimensiones y, por lo tanto, la varianza del algoritmo puede ser alta. En la práctica, si el científico de datos puede eliminar manualmente características irrelevantes de los datos de entrada, es probable que esto mejore la precisión de la función aprendida. Además, existen muchos algoritmos para la selección de características que buscan identificar las características relevantes y descartar las irrelevantes (es decir que el sistema este programado para que determine que datos tener en cuanta y cuáles no, o también

asociar a los campos o características una escala de relevancia, para que al sistema le quede más fácil decidir), por ejemplo, el análisis de componentes principales para el aprendizaje no supervisado. Esto reduce la dimensionaldad.

6. Sobreajuste: anuqué el proceso de entrenamiento es fundamental en el ML, un sobre exposición, causar que el sistema responda perfectamente a los escenarios de entrenamiento, pero para casos no detallas, el sistema tendrá alta probabilidad de error, Una forma práctica de prevenir esto es detener prematuramente el proceso de aprendizaje, así como aplicar filtros a los datos en la fase de pre-aprendizaje para eliminar ruidos.

Como se indica en el libro "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" (2001), solo después de considerar todos estos factores se puede elegir un algoritmo de aprendizaje supervisado que funcione para el conjunto de datos en el que se está trabajando. Por ejemplo, si se estuviera trabajando con un conjunto de datos que consta de datos heterogéneos, entonces los árboles de decisión funcionarían mejor que otros algoritmos. Si el espacio de entrada del conjunto de datos en el que se está trabajando tiene N dimensiones, entonces es mejor realizar primero un tratamiento de datos y un modelamiento en los datos antes de usar un algoritmo de aprendizaje supervisado en ellos.

recolección de información

La metodología que se aplicó para la investigación sobre la influencia de las TIC en el sector salud y como se puede emplear el machine learning para el diagnóstico y prevención de enfermedades, se denomina Metodología Revisión sistemática de la literatura (SLR), es un método que ayuda al estudio de determinada materia para que sea un estudio selectivo y crítico, sustentado por múltiples autores, este estudio aunque no propone una investigación original, si recopila la información más relevante de los temas a tratar, aquí se resume y analiza la información disponible sobre un tema específico basado en una búsqueda cuidadosa de la literatura, para llegar a conclusiones acerca de la evidencia científica para la prevención, diagnóstico o tratamiento de una enfermedad específica apoyadas en herramientas tecnológicas.

Para la recolección de referencias de la metodología SLR se consultaron las bases de datos IEEE explore, Google scholar, Research gate y arXiv, entre muchas otras orientadas al sector salud como MEDLINE (NLM), INDEX MEDICUS, LILACS, EMBASE. Con el fin de clasificarlas según su impacto en este estudio y poder descartar las que menos relevancia o incidencia aportaban.

Otra parte importante que se aplicó de esta metodología fue la definición clara de las preguntas que se buscan responder con este estudio, la principal es:

- ¿Cómo el Machine Learning, combinado con diferentes tecnologías, prácticas y métodos establece un servicio predictivo que permita el diagnóstico y prevención de enfermedades?

Y la definición de las preguntas secundarias que aportan a la repuesta de la pregunta principal:

- ¿Qué es Machine learning?
- ¿Cómo funciona el Machine learning?
- ¿Cómo pude ser usado en el sector salud?
- ¿Cuáles son los aspectos que componen el Machine learning?

Recomendaciones

Como indican los autores Beunza, J., Puertas, E. & Condés E. en su libro "Manual práctico de inteligencia artificial en entornos sanitarios", hay algunas claves importantes, para que el machine learning dentro de un sistema inteligente tenga éxito en proyectos de la salud, los cual se describen como:

1. Definir preguntas muy concretas y específicas (las denominadas tarea "T") que den respuesta a una necesidad específica, para evitar saturaciones en los sistemas, pero para aprovechar las capacidades estas tareas pueden ser combinadas y secuenciadas, para que arrojen varios resultados de un tema en particular, como por ejemplo determinar las poblaciones más propensas a padecer diabetes.
2. Usar preferiblemente "datos duros" es decir datos de alta confiabilidad y calidad, para que los resultados sean lo más asertivos y veraces posibles, datos provenientes de fuentes seguras y por demás confiables.
3. Que las variantes y características de la población final u objetivo estén incluidas en los datos de la población de entrenamiento, a lo que esto se refiere es que mientras se trabaja en el desarrollo del sistema de ML, las pruebas se realicen con datos de poblaciones reales, es decir, que el entrenamiento del sistema sea con datos reales y no ficticios, clave relacionada a ítem anterior.
4. Incluir profesionales técnicos con conocimientos en ciencia de datos, más allá del desarrollo de algoritmos de machine learning, es decir que el equipo de trabajo en el proyecto, no solo se centre en profesionales capacitados en la programación de algoritmos, sino también en personas con conocimiento en base de datos, modelamiento y conocimiento en Big Data y Minería de datos.
5. Incluir a los profesionales sanitarios desde las primeras fases del desarrollo, es decir que, en todas las fases del desarrollo, se acompañe con el conocimiento de un profesional en salud, biológica y medicina como farmacéuticos, nutricionistas, médicos generales, especialistas, etc. ya que ellos

ISSN: 2711-4600

son los que aportan el conocimiento fundamental del sector salud hacia donde esta encaminados las investigaciones, y pueden indicar como deberían ser los resultados esperados.

6. Incluir profesionales sanitarios con conocimientos de machine learning, adicional al ítem anterior, es importante que el equipo de trabajo tenga apoyo de profesionales con experiencia o conocimiento teórico o practico del ML, así será más fácil articular todas las áreas de trabajo.
7. Los resultados de inteligencia artificial y el machine learning aporte valor de manera evidente, es decir que el impacto de las investigaciones pueda ser reconocido, para así incentivar la inversión.
8. Los autores recalcan comenzar con proyectos piloto y diseñar un éxito prematuro.
9. Definir los puntos de referencia del algoritmo antes de comenzar el entrenamiento de aprendizaje de la máquina.
10. Validar el algoritmo en la población sobre la que se va a aplicar en producción, es decir que si se buscar prevenir la diabetes, se debería trabajar sobre las poblaciones que de acuerdo a sus factores de riesgo son más propensas a sufrirlos y así generar mejores prácticas de entrenamiento y cuando su uso sea aplicado fuera de pruebas, prueba predecir en poblaciones generales esta enfermedad.

Conclusiones

Como se ha podido evidenciar en el transcurso de este capítulo, el Machine Learning es una herramienta altamente tecnológica, que se basa en muchos ramas de la tecnología tanto de hardware (servidores de almacenamiento, servidores de aplicación, equipos de cómputo equipos de red, etc.), software (programas licenciados y no licenciados, lenguajes de programación, interfaces de usuario, programación de algoritmos), ciencias de la información (big data, minería de datos, modelamiento de base de datos, bodegas de datos, lagos de datos, etc.), tecnologías de redes de comunicación (servicios de interconexión de plataformas o también llamados web services, conexiones de internet o intranet), tecnologías de servicio en la nube (para modalidades de proyectos de investigación compartidos entre varias entidades, que requieran acceso al sistema de ML desde la web) entre otra serie de componentes tecnológicos que hagan de los sistemas de ML realizables, seguros y confiables, y que a su vez no puedan ser vulnerados o hackeados, esta tecnología se ha desarrollado desde hace ya varias décadas atrás fácilmente remontándose al siglo XVIII, con el desarrollo de modelos estadísticos y probabilísticos (The Royal Society. 2017), mejorando y evolucionando en lo que en la actualidad conocemos como uno de los ejes fundamentales de la Inteligencia Artificial, que busca simular capacidades humanas de aprendizaje y razonamiento

lógico, aunque para eso aún falta mucho camino por recorrer, aunque por ahora y como se ha tratado de explicar desde el inicio del capítulo con la recopilación confiable de muchas fuentes y autores, el ML cuenta con gran potencial de desarrollo, no solo en el sector salud, sino en incontables áreas, como en el sector bancario, educación, atención al cliente, predicciones climáticas, sector agrícola, de transporte de mercancía, ventas, marketing, industria de juegos y el entretenimiento, control de tráfico, seguridad policial y nacional, entre muchas otras más.

Se evidencia que para que los proyectos de ML tenga éxito, es necesario que cuenten con un objetivo claro, que se planee desde el inicio el objetivo del aprendizaje automatizado y entorno a él se desarrolló, se programe y se ejecuten pruebas orientadas a que los resultados sean acorde al objetivo propuesto, sea una necesidad o un problema que se quiera predecir y resulte beneficioso para la población, como en este caso en la detección y prevención de enfermedades, los proyectos se deberán centrar en una enfermedad especifica sea cáncer pulmonar, depresión u contagio de un virus, es necesario acotar el sistema, ya que un solo sistema de ML que detecte todas la enfermedades conocidas, sería un proyecto desproporcionado y que se saldría de las capacidades que hoy en día se manejan, pero si se inicia con una necesidad y se entrena al sistema con el objetivo por el que se creó, las probabilidades de éxito son más altas, adicional que es fundamental contar con los recursos de personal experto e idóneo de varias áreas de conocimiento, de tiempo de programación, pruebas y entrenamiento para ver los primeros resultados éxitos, acceso a importantes volúmenes de información y también recursos financieros.

Aunque es claro que proyectos de este tipo y aún más aplicados al sector salud, un área de tanta importancia a nivel mundial, no es fácil, se destacan proyectos de ML nacionales e internacionales que han tenido éxito y que se siguen adaptando a las necesidades crecientes y a las tecnologías que día a día aparecen para ser proyectos más confiables, de mayor relevancia y que perduren más allá de la actual tendencia o moda que posee el Machine Learning; se demuestra cada día que el análisis de datos es fundamental para la toma de decisiones, acción que hace a cabalidad el ML junto con acciones de predicción y detección o prevención de eventos, para este caso estudio de enfermedades.

Referencias

ACCA (the Association of Chartered Certified Accountants). (2019). N2. Navigating the terminology. En Machine learning More science than fiction (52). Reino Unido: The Association of Chartered Certified Accountants. P 18.

A. Silberschatz, H. F. Korth & S. Sudarshan, (2002) Fundamentos de Bases de Datos, España: MacGraw Hill.

Badaró, S., Ibañez, L., & Agüero, M. (2013, Diciembre). SISTEMAS EXPERTOS: Fundamentos, Metodologías y Aplicaciones. Revista Ciencia y Tecnología, Vol 13, 15. 2021, Mayo 01, De Universidad de Palermo Base de datos.

Benítez, I. & Diez, J..\ (2005). Técnicas de agrupamiento o reconocimiento de patrones (clustering) . En Técnicas de Agrupamiento para el Análisis de Datos Cuantitativos y Cualitativos. Valencia, España: Universidad Politécnica de Valencia. Pp 16 – 23

Beunza, J., Puertas, E. & Condés E. (2020). Libro, Manual práctico de inteligencia artificial en entornos sanitarios. Mayo 02, 2021, de Elsevier Sitio web: https://www.elsevier.com/es-es/connect/ehealth/claves-del-exito-de-un-programa-de-inteligencia-artificial-en-salud

Bhatt M. (Mayo 2016). Machine Learning Project. University of New Orleans, 1, 19. 2021, Mayo 30, De ResearchGate Base de datos.

Cabría, S. (1994). Filosofía de la estadística. España: Publicacions de la Universitat de València. pp 37-54.

Calvo, J., Guzman, M., & Ramos, D. (2018). Machine Learning, una pieza clave en la transformación de los modelos de negocio. abril 10, 2021, de Management Solutions Sitio web: https://www.managementsolutions.com/sites/default/files/publicaciones/esp/machine-learning.pdfCarmona L, Rozo C, Mogollón A. (2005). La salud y la promoción de la salud: una aproximación a su desarrollo histórico y social. Revista Ciencias de la Salud, pp.3, 62-77.

Carrasco, O., (2009). Cómo escribir artículos de revisión. Revista Médica La Paz, Vol 15, 7. 2021, Abril 10, De Scielo Base de datos.

Clausen, C. (2020). Transformación digital del sistema de salud colombiano. 2021, mayo 29, de Ministerio de Salud y Protección Social

Cruz, E., et al. (2018). Aplicación de las Tic en los sectores económicos (productivo, comercial y servicios). Editorial Scientometrics E Researching Consulting Group SAS.

Cruz, E., et al. (2019). Importancia de las TIC en los sectores económicos. Editorial Scientometrics E Researching Consulting Group SAS.

Dodson, T., Mattei, N., & Goldsmith, J. (2011). A natural language argumentation interface for explanation generation in Markov decision processes. In International Conference on Algorithmic. Decision Theory, pp. 42-55.

Escalante, P. (2004). Prevención de la enfermedad. En Curso de Gestión Local de Salud para Técnicos del Primer Nivel de Atención. Uruguay: CENDEISSS y Universidad de Costa Rica. pp. 7-10

Galvez, R. (2018). Extracción y Análisis de Datos No Estructurados: Aplicaciones usando texto, audio, imágenes y video. abril 23, 2021, de Universidad Nacional de Córdoba Sitio web: https://ecodev.eco.unc.edu.ar/files/ief/workshops/2018/Galvez_Extraccin_y_Anlisis_de_Datos_No_Estructurados__Aplicaciones_usando_texto_audio_imgenes_y_video.pdf

Gardner, M. J., & Altman, D. G. (1993). Intervalos de confianza y no valores p: estimación en vez de pruebas de hipótesis. Boletín de la Oficina Sanitaria Panamericana (OSP). De google scholar.

García J., López J., Jiménez F., Ramírez T., Lino L. & Reding A. (2014). Metodología de la investigación, bioestadística y bioinformática en ciencias médicas y de la salud, 2e. México, D.F.: McGRAW-HILL INTERAMERICANA EDITORES. Capítulo 25

González, M. C. (2019). El impacto de la transformación digital en la prevención de la salud de personas mayores en Colombia. Repositorio Instituciona UMNG. Universidad Militar Nueva Granada. Pp 14 - 16

Hastie, Trevor, Robert Tibshirani, and J. H. Friedman (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction: With 200 Full-color Illustrations. New York: Springer

Hua K., Hsu C., Hidayati S., Cheng W., Chen Y. (2015, Agosto 4). Computer-aided classification of lung nodules on computed tomography images via deep learning technique. Onco Targets Therapy, 2021, Mayo 29, De NCBI Base de datos.

Hurwitz, J., & Kirsch D. (2018). Machine Learning For Dummies®, IBM Limited Edition. Hoboken: John Wiley & Sons, Inc.

Interactive Chaos, Curso online: Machine Learning, Mayo 02, 2021, Sitio Web: https://interactivechaos.com/es/manual/tutorial-de-machine-learning/presentacionJiménez, J. (2016, Noviembre 15). Frente al suicidio, tecnología: así es cómo el "machine learning" puede ayudarnos a luchar contra la gran epidemia silenciosa. Xataka, 2021, Mayo 29.

Joyanes, L. (2008). Fundamentos de la Programación. Ed. McGraw-Hill. 2ª Edición.

Joyanes Aguilar, L. (2003) Fundamentos de programación. España: Mc Graw Hill. p. 53.

Lantz, B. (2015). Machine Learning with R Second Edition. Michigan: Packt Publishing.

Linthicum KP, Schafer KM, Ribeiro JD. Machine learning in suicide science: Applications and ethics. Behav Sci Law. 2019 Mayo. doi: 10.1002/bsl.2392. Epub 2019 enero 4.

Marr. B. (2016, Febrero 20). A Short History of Machine Learning -- Every Manager Should Read. Revista Forbes

Monterrey, P. & Gómez-Restrepo, C. (2007). Aplicación de las pruebas de hipótesis en la investigación en salud: ¿estamos en lo correcto?. Universitas Medica, 48(3),193-206.[fecha de Consulta 26 de Abril de 2021]. ISSN: 0041-9095. Disponible en: https://www.redalyc.org/articulo.oa?id=231018668002

Morales, E. & Escalante H.. (2019). Aprendizaje Basado en Instancias. Abril 01, 2021, de INAOE - Gobierno de México Sitio web: https://ccc.inaoep.mx/~emorales/Cursos/NvoAprend/Acetatos/ibl2019.pdf

Murphy, K. (2012). Machine Learning: A Probabilistic Perspective. Canadá: MIT Press; Illustrated edition.

Núñez A., Armengol M.A. & Sánchez M. (octubre 2019). Big Data Analysis y Machine Learning en Unidades de Cuidados Intensivos. Medicina Intensiva, Volume 43, Pp 416-426.

Oliveros, A. (2020). Interoperabilidad de la Historia Clínica, 2021, mayo 30, de Ministerio de Salud y Protección Social

Ordoñez, M., Tapia, J. & Asanza, W.. (2015). Fundamentos de base de datos. Ecuador: Ediciones Utmach. pp 23-43.

Pacheco, J. (2017, Diciembre). METODOLOGÍA PARA ELABORAR EL MODELO CONCEPTUAL DE DATOS. Ediciones Universidad Cooperativa de Colombia, N.° 37, 22. 2021, Mayo 01, De Course Work Base de datos.

Perrault, R., Shoham. Y., Brynjolfsson, E., Clark, J., Etchemendy J., Grosz b, Lyons T., Manyika J., Mishra S. & Niebles J. (2019). Artificial Intelligence Index Report 2019. Stanford University, Edición 2019, 291. 2021, mayo 29, De HAI, Stanford University Base de datos. P 74.

Ponce, Julio & Torres, Aurora & Aguilera, Fátima & Silva Sprock, Antonio & Flor, Ember & Casali, Ana & Scheihing, Eliana & Tupac, Yvan & Torres, Dolores & Ornelas, Francisco & Hernández[1], José-Alberto & D., Crizpín & Vakhnia, Nodari & Pedreño, Oswaldo. (2014). Inteligencia Artificial, Vol 1, Iniciativa Latinoamericana de Libros de Texto Abiertos. P 20.

Poveda, F., et al. (2019). Lineamientos y orientaciones investigativas desde la disciplina del derecho. Colombia: Editorial Scientometrics E Researching Consulting Group SAS.

Poveda, F., et al. (2020). Research, Artificial Intelligence And Tools For Researchers. Colombia: Editorial Scientometrics E Researching Consulting Group SAS.

Raffino, M. (2020). Concepto de Diagnostico. abril 12, 2021, de Concepto.de Sitio web: https://concepto.de/diagnostico/Serna, C.. (2018). Una Pequeña Introducción a Machine Learning. abril 4, 2021, de Universidad Central Sitio web: http://hpclab.ucentral.edu.co/~hfranco/data_analysis/SessionV/VSession.pdf

SAS. Análisis de atención médica (2021), Mayo 29, Sitio Web: https://www.sas.com/es_co/industry/health-care/solution/population-health.htmlStensmo, M., & Thorson M. (2003). Unstructured Information Management. Suecia: Infosphere AB.

The Royal Society. (2017). Machine learning: the power and promise of computers that learn by example. The Royal Society, 4, 128. 2021, Mayo 30, De ResearchGate Base de datos.

Tumbay, M. R. S. (2018). fatherhood's subjective experience in the face of adolescent children'depressive symptomatology and suicide attempt. Studia Universitatis Babes-Bolyai-Philosophia, 63(1), Pp 119-135.

Universidad Católica de Colombia. (2020). Facultad de Ingeniería desarrolla plataforma de análisis y predicción de datos del Covid-19 en Colombia. 2021, Mayo 29, de Universidad Católica de Colombia Sitio web:

https://www.ucatolica.edu.co/portal/facultad-de-ingenieria-desarrolla-plataforma-de-analisis-y-prediccion-de-datos-del-covid-19-en-colombia/Utrera, R.. (2017). Uso de algoritmos de aprendizaje automático aplicados a bases de datos genéticos. abril 15, 2021, de Universitat Oberta de Catalunya Sitio web: http://openaccess.uoc.edu/webapps/o2/bitstream/10609/65426/6/rgagoTFM0617memoria.pdfVázquez, J. (2012). Análisis Y Diseño De Algoritmos. México: Red Tercer Milenio S.C. pp. 16-18.

Valdez, A. (2019). Machine Learning para Todos. 4° Congreso Nacional de profesionales de Computación, información y Tecnologías, Congreso llevado a cabo en Perú.

Velásquez, J., Una Guía Corta para Escribir Revisiones Sistemáticas de Literatura Parte 3, Revista DYNA - Facultad de Minas, vol. 82, no. 189, pp. 9-12, 2015