DOI: https://doi.org/10.46480/esj.4.2.83

Modelo computacional de clasificación de aprendizaje de máquina supervisado, para el análisis de datos cardiovasculares y pronóstico médico

Computational model of supervised machine learning classification, for the analysis of cardiovascular data and medical prognosis

Glenda Blanc-Pihuave[1], Lorenzo Cevallos-Torres[2] y José Arteaga-Vera[3]

RESUMEN

Las enfermedades cardiovasculares son un problema de salud pública en Ecuador y todo el mundo, por lo que este trabajo investigativo propone el diseño de un modelo computacional de clasificación a través del uso de técnicas de machine Learning, con el apoyo de modelos probabilísticos que permitan modelar los factores de riesgo de enfermedades cardiovasculares. Este modelo está basado en Redes Bayesianas, que, en base a los factores de riesgo de la enfermedad, mostrará como resultado el porcentaje que tiene el paciente de contraer la misma. Se aplicó la metodología de investigación documental que aporte con el conocimiento necesario para la realización de este proyecto en el cual se realizaron pruebas para verificar el comportamiento de cada una de las variables utilizadas en el modelo probabilístico, el cual brindará resultados eficientes y en un corto periodo de tiempo, siendo así una herramienta de apoyo en la toma de decisiones para los expertos.

Palabras clave: Enfermedades Cardiovasculares, Machine Learning, Redes Bayesianas, diagnostico, Naive Bayes.

ABSTRACT

Cardiovascular diseases are a public health problem in Ecuador and around the world, so this research work proposes the design of a computational model of classification using techniques of ma-chine Learning, with the support of probabilistic models that allow modeling of cardiovascular disease risk factors. This model is based on Bayesian Networks, which, based on the risk factors of the disease, will show the percentage that the patient has of contracting it. The documentary research methodology was applied that provides the necessary knowledge to carry out this project in which tests were carried out to verify the behavior of each of the variables used in the probabilistic model, which will provide efficient results. and in a short period of time, thus being a support tool in decision-making for experts

Keywords: Cardiovascular Diseases, Machine Learning, Bayesian Networks, diagnosis, Naive Bayes.

Fecha de recepción: Marzo 23, 2020.
Fecha de aceptación: Agosto 31, 2020.

Introducción

América Latina, así como la mayoría de los países occidentales sufre de la epidemia silenciosa de las enfermedades cardiovasculares. En nuestra región, el 9% de las personas tiene colesterol elevado y el 20% problemas de hipertensión, esto sin contar otros factores de riesgo como la diabetes, la obesidad, el sedentarismo o el tabaquismo.

Con el desarrollo acelerado de la sociedad, los cambios en el estilo de vida y el incremento de la esperanza de vida, las enfermedades cardiovasculares (ECV) han pasado a ser la primera causa de muerte en el mundo. Según la Organización Panamericana de la salud (OPS), es un hecho reconocido que América Latina y el Caribe presenta las mayores disparidades socioeconómicas dentro de la Región de las Américas, un panorama que inevitablemente se ha traducido en una elevada mortalidad por enfermedades crónicas no transmisibles (ECNT), incluidas las enfermedades cardiovasculares, la diabetes y el cáncer. Moreno, G. A. (2008).

El envejecimiento, la globalización, la urbanización y el aumento de la obesidad y la inactividad física han llevado a que las enfermedades cardiovasculares sean la principal causa de muerte y discapacidad en la Región, representando casi un tercio de la mortalidad total a nivel regional, con un riesgo generalmente mayor en los hombres que en las mujeres. Fernández, E., Sabán, J., Fabregate, M., & Fabregate, R. (2009).

Vega Romero, R. (2009). La Organización Mundial de la Salud (OMS), define a las ECV como un grupo de desórdenes del corazón y de los vasos sanguíneos, entre los que se incluyen:

La cardiopatía coronaria: enfermedad de los vasos

[1] Magister en Administración de Empresas MBA, Facultad de Ing. Sistemas Computacionales, Universidad Ecotec, Ecuador. E-mail: blanc.glenda@gmail.com.
[2] Máster en Modelado Computacional en Ingeniería, Facultad de Ciencias Matemáticas, Universidad de Guayaquil, Ecuador. E-mail: lorenzo.cevallost@ug.edu.ec.
[3] Doctor en Ciencias Pedagógicas, Universidad Laica Eloy Alfaro de Manabí, Ecuador. E-mail: jose.arteaga@uleam.edu.ec.

Como citar: Blanc-Pihuave, G., Cevallos-Torres, L., & Arteaga-Vera, J. (2020). Computational model of supervised machine learning classification, for the analysis of cardiovascular data and medical prognosis. *Ecuadorian Science Journal.* 4(2), 71-79.
DOI: https://doi.org/10.46480/esj.4.2.83

sanguíneos que irrigan el músculo cardiaco. Chávez Domínguez, R., Ramírez Hernández, J. A., & Casanova Garcés, J. M. (2003).

Las enfermedades cerebrovasculares: enfermedades de los vasos sanguíneos que irrigan el cerebro.

Las arteriopatías periféricas: enfermedades de los vasos sanguíneos que irrigan los miembros superiores e inferiores.

La cardiopatía reumática: lesiones del músculo cardiaco y de las válvulas cardíacas debidas a la fiebre reumática, una enfermedad causada por bacterias denominadas estreptococos.

La cardiopatía congénita: malformaciones del corazón presente desde el nacimiento. La trombosis venosa profunda y Embolia pulmonar: coágulos de sangre (trombos) en las venas de las piernas, que pueden desprenderse (émbolos) y alojarse en los vasos del corazón y los pulmones. Volschan, A., Caramelli, B., Gottschall, C. A. M., Blacher, C., Casagrande, E. L., & Manente, E. R. (2004).

En últimos estudios realizados sobre las ECV se han analizado los factores de riesgo cardiovascular. Pero ¿cuáles son los principales factores o factores primarios que ocasionan las ECV?, ¿Cuál de todos estos factores es el más ocasional?, y ¿Cuál de los factores de riesgo desencadenan a otros?, muchos profesionales de la salud dan respuesta a estas interrogantes en base a la experiencia obtenida durante su carrera, pero dentro de su diagnóstico está enfrascada la incertidumbre de saber si el diagnóstico y tratamiento será el adecuado para el paciente. Murillo, A. Z., & Esteban, B. M. (2005).

Los estilos de vida o hábitos son un conjunto de comportamientos que desarrollan las personas que algunas veces son saludables y en otras son nocivos para la salud; entre los nocivos, se encuentran la falta de actividad física, alimentación no saludable y el consumo de sustancias sicoactivas como el cigarrillo y el alcohol, constituyendo factores de riesgo para las ECNT, como la hipertensión arterial, enfermedad coronaria, cerebrovascular, obesidad, diabetes tipo II y el cáncer. Es así como la dieta, el estilo de vida saludable y la detección temprana de las ECV desempeñan roles importantes para elevar la calidad de vida. Salazar ÁLvarez, Y. (2011).

Los ataques cardíacos y accidentes cerebrovasculares (ACV) suelen tener su causa en la presencia de una combinación de factores de riesgo, tales como el tabaquismo, las dietas malsanas y la obesidad, la inactividad física, el consumo nocivo de alcohol, la hipertensión arterial, la diabetes y la hiperlipidemia. Los efectos de los factores de riesgo comportamentales pueden manifestarse en las personas en forma de hipertensión arterial, hiperglucemia, hiperlipidemia y sobrepeso u obesidad. Estos "factores de riesgo intermediarios", que pueden medirse en los centros de atención primaria, son indicativos de un aumento del riesgo de sufrir ataques cardíacos, ACV, insuficiencia cardíaca y otras complicaciones. Guarnaluses, L. J. B., & Ramos, A. P. (2016).

La inteligencia artificial (IA) se puede definir como una rama de las ciencias de la computación que se ocupa de la comprensión, desde el punto de vista informático, de lo que se denomina comúnmente comportamiento inteligente. Incluye distintos campos como el aprendizaje automático o machine Learning (ML), el procesamiento de lenguaje natural, los sistemas expertos, la visión artificial, etc., y es la base de otros como la robótica o big data, dos de las áreas que más están creciendo en la actualidad. Encinas, C., Sacristán, J., Cenamor, D., & Morell, L. (2019).

En los últimos años, la revolución digital proporcionó recursos relativamente económicos y disponibles para la recolección y almacenamiento de datos. Los hospitales recopilan ese gran volumen de datos de los pacientes obtenidos de sus chequeos médicos por medio de sus máquinas de monitoreo y otros dispositivos de recolección de datos, los cuales se comparten en grandes sistemas de información. La implementación de modelos de inteligencia artificial proporciona varias herramientas indispensables para el análisis inteligente de datos. Quesada, Y., Wong, D., & Rosete, A. (2012), De Mitri, M. J. (2019).

El ML, rama de la inteligencia artificial que construye y estudia sistemas capaces de aprender a partir de un conjunto de datos de adiestramiento y de mejorar procesos de clasificación y predicción es la tecnología adecuada para analizar grandes volúmenes de datos médicos y, en particular, el progreso en el ML para facilitar el diagnóstico clínico ha sido constante y existen varios ejemplos disponibles de aplicaciones, así como resultados de esfuerzos previos desde los cuales construir. Aranda Núñez, A. P. (2019), Moncayo, K. C., Sanchez, A. G., Anton, P. R., & Cevallos-Torres, L. (2019).

El aprendizaje automático desempeña un papel esencial en el diagnóstico de ECV y más, ya que permite obtener diagnósticos más precisos debido a la mayor capacidad de procesamiento, comparación y síntesis de la información. El diagnóstico permite brindar ayuda a los expertos de la medicina en la toma de decisiones para obtener tratamientos mejorados, el cuál es el objetivo principal, desarrollar un modelo que permita una evaluación más precisa del pronóstico y diagnóstico o presencia de ECV en el paciente, el modelo propuesto servirá como sistema de soporte a los cardiólogos en el diagnóstico médico general y tratamiento.

Para crear el modelo de aprendizaje automático se utilizó el algoritmo de clasificación bayesiano Naive Bayes (NB) para llevar a cabo el diagnostico de las ECV y las Redes bayesianas como herramienta para identificar los factores de riesgo cardiovascular de mayor incidencia y así establecer prioridades para la prevención primaria. Una vez obtenida la estimación del riesgo esto nos servirá como punto clave de utilidad en la toma de decisiones para el tratamiento de la enfermedad cardiovascular que se vaya a tratar en el paciente. Portugal, R., & Carrasco, M. (2007, January).

Las redes bayesianas organizan los datos mediante un conjunto de variables y analizan la relación entre ellas, es decir, estiman la probabilidad de las variables no analizadas en base a las variables analizadas. Por otro lado, el algoritmo NB, predice la probabilidad de los posibles resultados y utiliza las probabilidades de cada variable para hacer una predicción, en este algoritmo los atributos numéricos se modelan mediante una distribución normal. Pereira-Toledo, A., López-Cabrera, J. D., & Quintero-Domínguez, L. A. (2017).

Materiales y métodos

A continuación, se presentan las técnicas y métodos implementados de forma individual, así como la combinación para el desarrollo del tema propuesto, así mismo se presentan definiciones básicas relacionadas con los problemas cardiovasculares, así mismo se definen conceptos estadísticos, el tipo muestra.

Riesgo cardiovascular

El riesgo cardiovascular (RCV) se define como la probabilidad de padecer un evento cardiovascular en un determinado periodo de tiempo, que habitualmente se establece en entre 5 y 10 años, la mejor herramienta para establecer prioridades en la prevención primaria de las ECV es la estimación del RCV mediante las funciones de riesgo. O'Donnell, C. J., & Elosua, R. (2008).

Factores de riesgo cardiovasculares

El factor de riesgo (FR) se define como un elemento o una característica mensurable que tiene una relación causal con un aumento de frecuencia de una enfermedad y constituye factor predictivo independiente y significativo del riesgo de contraer una enfermedad. Medrano, M. J., Cerrato, E., Boix, R., & Delgado-Rodríguez, M. (2005).

Un factor de riesgo cardiovascular (FRCV) es una característica biológica, condición y/o comportamiento que aumenta la probabilidad de padecer o fallecer a causa de una ECV. Los FRCV se dividen fundamentalmente en dos grupos: modificables o controlables (tabaquismo, sedentarismo, obesidad, diabetes, hipertensión arterial y dislipidemia y no modificables (raza, sexo, edad y antecedentes familiares. Álvarez Cosmea, A. (2001).

Tabla 1. Clasificación de los factores de riesgo de ECV

Modificables	No modificables
Sobrepeso u obesidad	Edad
Inactividad física	Herencia genética
Hipertensión arterial	Género
Hipercolesterolemia	
Tabaquismo	
Consumo de alcohol	
Glucosa.	

Machine Learning

Machine Learning o Aprendizaje Automático es una disciplina de la Inteligencia Artificial que crea sistemas que aprenden automáticamente. Aprender en este contexto quiere decir identificar patrones complejos en millones de datos. La máquina que realmente aprende es un algoritmo que revisa los datos y es capaz de predecir comportamientos futuros. Automáticamente, también en este contexto, implica que estos sistemas se mejoran de forma autónoma con el tiempo, sin intervención humana. Molina Espinoza, C. I. (2014).

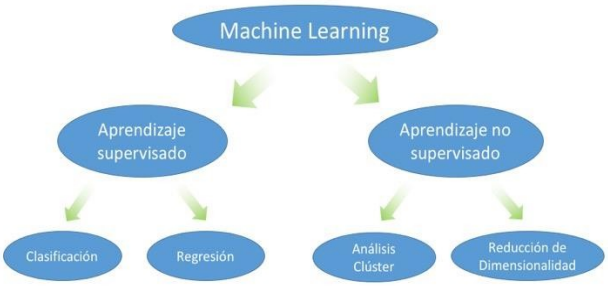

Figura 1. Clasificación Aprendizaje Automático

Tipos de aprendizaje automático

Los algoritmos de aprendizaje automático se organizan taxonómicamente en función del resultado deseado del algoritmo. González, F. A. (2015). Los tipos algoritmos incluyen:

Aprendizaje no supervisado

Aprendizaje no supervisado es un método de Aprendizaje Automático donde un modelo se ajusta a las observaciones. Se distingue del Aprendizaje supervisado por el hecho de que no hay un conocimiento a priori. En el aprendizaje no supervisado, un conjunto de datos de objetos de entrada es tratado, tal como lo indica Rodríguez, C. A., Gallego, J. H., Mora, I. D., Duque, A. O., & Bustamante, J. (2014). En su investigación presenta el desempeño de dos algoritmos basados en aprendizaje de máquina no supervisado para la detección de latidos de contracción ventricular prematura en la señal ECG".

Aprendizaje semi – supervisado

Donde el algoritmo aprende una política como actuar dad una observación del mundo. Cada acción tiene algún impacto en el medio ambiente, y el entorno proporciona retroalimentación que guía el algoritmo de aprendizaje.

Aprendizaje supervisado

Donde el algoritmo genera una función que asigna entradas a las salidas deseadas. Una formulación estándar de la tarea de aprendizaje supervisado es el problema de clasificación: se requiere que el alumno aprenda (para aproximar el comportamiento de) una función que mapea un vector en una de varias clases mirando varios ejemplos de entrada-salida de la función

Redes Bayesianas

Una red bayesiana es un grafo acíclico dirigido en el que cada nodo representa una variable y cada arco una dependencia probabilística; son utilizadas para proveer: una forma compacta de representar el conocimiento y métodos flexibles de razonamiento

Las redes bayesianas constituyen una alternativa a los árboles de decisión por permitir la representación de modelos más complejos de diagnosis o prognosis. Las redes bayesianas basan su estudio en la teoría de la probabilidad y permiten realizar una inferencia al integrar el juicio del experto con las bases de datos disponibles, y realizar inferencia entre cualquier subconjunto de variables.

Las Redes Bayesianas se componen principalmente de dos partes, por un lado, la estructura, el modelo o parte cualitativa: un grafo dirigido a cíclico (DGA), donde cada nodo representa una variable aleatoria y los arcos representan dependencias probabilísticas entre variables. Pero por otra, también se componen de una distribución condicional de probabilidades, esta parte de la red bayesiana se conoce como la parte paramétrica o cuantitativa de la red.

Una red Bayesiana está compuesta por "Nodos", que representan variables generalmente discretas, cuya fuerza de relación entre ellas se cuantifica mediante una distribución de probabilidades condicionales que determinan el valor final de aquellos nodos que no se han cargado como evidencia. La relación entre dichas variables se establece mediante "arcos" que representan una determinación causal entre nodos.

Aprendizaje de las Redes Bayesianas

El aprendizaje en las redes bayesianas consiste en definir la red probabilística a partir de datos almacenados en bases de datos. Este tipo de aprendizaje ofrece la posibilidad de definir la estructura gráfica de la red a partir de los datos observados o de la base de datos y de definir las relaciones entre los nodos basándose también en dichos casos.
El aprendizaje es una de las características que definen a los sistemas basados en inteligencia artificial porque siendo estrictos se puede afirmar que sin aprendizaje no hay inteligencia; es difícil definir el término "aprendizaje", pero la mayoría de las autoridades en el campo coinciden en que es una de las características de los sistemas adaptativos que son capaces de mejorar su

comportamiento en función de su experiencia pasada, por ejemplo, al resolver problemas similares.

Una red Bayesiana proporciona un sistema de inferencia, donde una vez encontradas nuevas evidencias sobre el estado de ciertos nodos, se modifican sus tablas de probabilidad; y a su vez, las nuevas probabilidades son propagadas al resto de los nodos. La propagación de probabilidades se conoce como inferencia probabilística, es decir, la probabilidad de algunas variables puede ser calculada dadas evidencias en otras variables. Las probabilidades antes de introducir evidencias se conocen como probabilidades a priori; una vez introducidas evidencias, las nuevas evidencias propagadas se llaman probabilidades a posteriori.

La ventaja fundamental del uso de la inferencia bayesiana radica en la utilidad que se le da para la toma de decisiones, actualmente su uso es frecuente por que se obtienen resultados más acertados en el contexto de parámetros desconocidos.

El mecanismo de inferencia sobre redes bayesianas permite utilizarlas para construir clasificadores. Para que esto se debe crear una red bayesiana en la que las variables se interrelacionen en el grafo. La clase pertenecerá a la variable desconocida, objetivo de la inferencia. Proporcionada una instancia cualquiera para la que se conozcan todos sus atributos, la clasificación se verificará infiriendo sobre el grafo la probabilidad posterior de cada uno de los valores de la clase, y eligiendo aquelvalor que maximice dicha probabilidad

Caso de Estudio

Este estudio pretende a partir de técnicas basadas en aprendizaje de máquina, analizar los factores de riesgo asociados a las enfermedades cardiovasculares, haciendo uso de algoritmos supervisados tales como son, las redes bayesianas y redes neuronales, esto se hace con la finalidad de poder identificar la presencia o ausencia de una enfermedad cardiovascular permitiéndoles a los especialistas apoyarse en un diagnóstico predictivo basado en computadora. Para este estudio se han considerado una base de datos de la Universidad de Ryerson que tiene dos tipos de características, la información objetiva recopilada de los exámenes médicos y objetiva sobre los hábitos del estilo de vida del paciente Las variables de la base de datos son:

- Edad: Edad del paciente.
- Altura: Signo que es tomado al paciente al momento de consulta médica.
- Peso: Peso en kg del paciente.
- Género: genero del paciente: hombre – mujer.
- Presión arterial sistólica.
- Presión arterial diastólica.
- Colesterol: nivel de colesterol presente en el paciente.
- Glucosa en la sangre: Signo del paciente al momento de la consulta médica.
- Fuma: Identifica si el paciente fuma.
- Consumo de alcohol: Identifica si el paciente consume alcohol.
- Presencia o ausencia de enfermedad cardiovascular: Para definir si el paciente presenta probabilidad de ECV.

Construcción de la Red Bayesiana (RB).

Los FR que incrementan su probabilidad de contraer una ECV en base a los datos bibliográficos y el juicio de expertos incluyen:

Principales factores de riesgo (FR) incluyen:
Sexo: masculino (los hombres tienen mayor riesgo de ataque cardiaco que las mujeres).
Edad: 45 años y más para los hombres, 55 años y más para las mujeres.
Obesidad y tener sobrepeso.
Fumar.
Colesterol total: Nivel deseable < 200. Límite alto 200-239. Alto > 240.

Glucemia: el nivel de glucosa en sangre se mantiene dentro de límites estrechos a lo largo del día (72-145 mg/dl; 4-8 mmol/l). Sin embargo, sube después de las comidas y es más bajo por la mañana antes del desayuno. Las personas con diabetes se caracterizan por tener niveles de glucosa más altos de lo normal.
Estilo de vida sedentario.
Otros factores de riesgo incluyen:
Estrés.
Consumo excesivo de alcohol.
Síndrome metabólico: combinación de presión arterial elevada, obesidad abdominal, y resistencia a la insulina.
Las variables de FR usadas en la RB y sus valores los mostramos a continuación:

Tabla 2. Variable y valores de la Red Bayesiana

variables	Valores
Edad	(edad)
Peso	(peso)
Sexo o género	(0: masculino, 1: femenino)
Presión sanguínea diastólica	(mmHg)
Presión sanguínea sistólica	(mmHg)
Alcohol	(0: no, 1: sí)
Fuma:	(0: no, 1: sí)
Colesterol	(1: normal, 2: Encima de lo normal, 3: elevado)
Glucosa	(1: normal, 2: Encima de lo normal, 3: elevado)
Actividad física	(0: no, 1: sí)
Riesgo Cardiovascular	(0: no, 1: sí)

La red bayesiana consta de 11 nodos que fueron previamente seleccionados con la ayuda de los expertos segmentados de acuerdo con su tipo:

Tabla 3. Signos característicos de problemas cardiovasculares

Grupo	Nombre	Variables Lingüística
Signos	Presión sanguínea sistólica	mmHg
	Presión sanguínea diastólica	mmHg
	Glucosa	Normal Limite

	alto

La estructura de red bayesiana está construida mediante el software Elvira, en él se observan los 11 nodos descritos con sus respectivas relaciones entre sí, para determinar la probabilidad de riesgo cardiovascular en base a los factores de riesgo y signos presentes en el paciente.

Tabla 1V. Descripción de los nodos.

Grupo	Nombre	Variables Lingüística
Factores de riesgo	Edad	Edad en años
	Sexo	Hombre
		Mujer
	Actividad física	Si
		No
	Colesterol	Normal
		Limite alto
		Alto
	Fumador	Si
		No
	Consumo de alcohol	Si
		No
	Peso	kg

A continuación, se representa gráficamente la Red neuronal, acorde a la descripción da en la tabla IV.

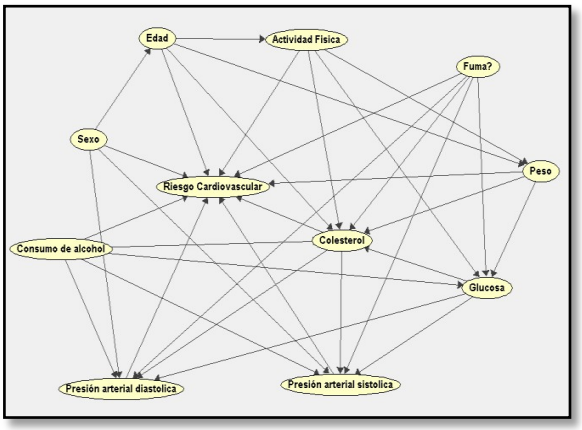

Figura 2. Red Bayesiana de los factores relacionados con las enfermedades cardiovasculares**.**

La vista de inferencia me muestra el peso de los nodos en relación con sus variables dependientes, es decir muestra el estado normal de la red bayesiana al tener los valores presentados como caso a priori, dichos valores fueron obtenidos tras la ejecución de métodos probabilísticos en conjunto con un software de análisis estadístico.

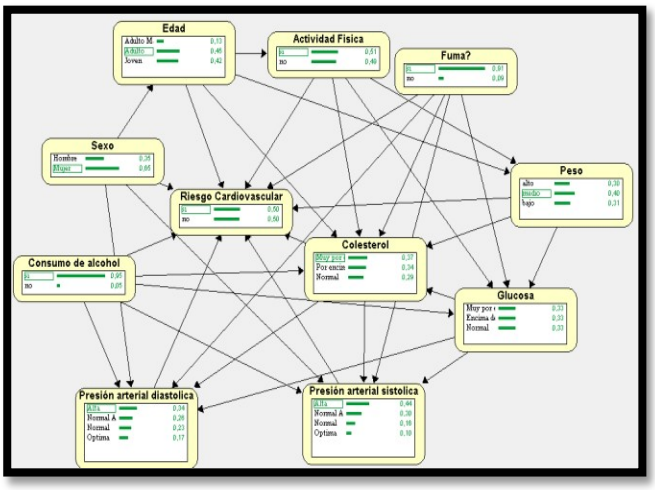

Figura 3. Vista de inferencia de nodos con sus respectivos valores causales a Priori.

Luego de obtener los pesos de cada variable de manera individual, se procede a correlacionar las variables identificadas y que independientemente de su peso son causas o consecuencias de otras, formando un caso probabilístico para cada evento de la variable.

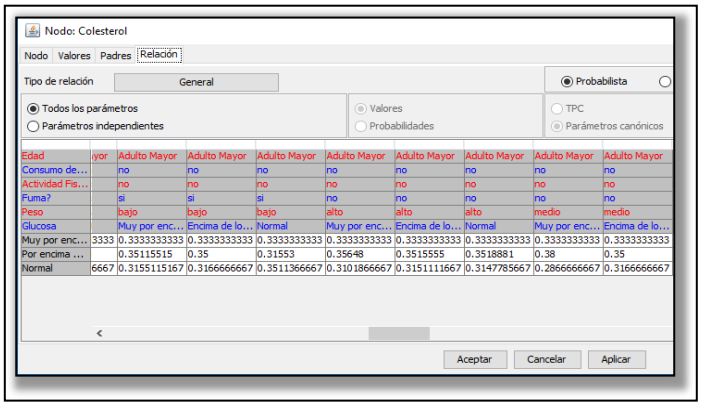

Figura 4. Correlación de los nodos hecho en Programa Elvira

Inferencia de pesos y afectación de eventos en red bayesiana

Como caso de prueba se ejecutarán un posible evento al azar la cual presenta los síntomas y signos en el paciente relacionados con las, haciendo la respectiva simulación, con el fin de obtener resultados de un caso real.

Caso1.- Al encender ciertos nodos al azar vemos variar los porcentajes a priori, esto se debe a la causa que genera esa variación frente a las otras variables, para este evento tenemos el caso de una personas, de edad Adulta, Mujer, que consume alcohol, su nivel de colesterol se encuentra muy por encima de lo normal, que presenta signos de sobrepeso, que realiza actividad física sin embargo presenta una presión arterial alta, dando como resultado mayor del 50% de riesgo cardiovascular, la misma que producirá alguna afección cardiovascular a largo plazo, siendo estos los nodos de causa mayor en las RCV.

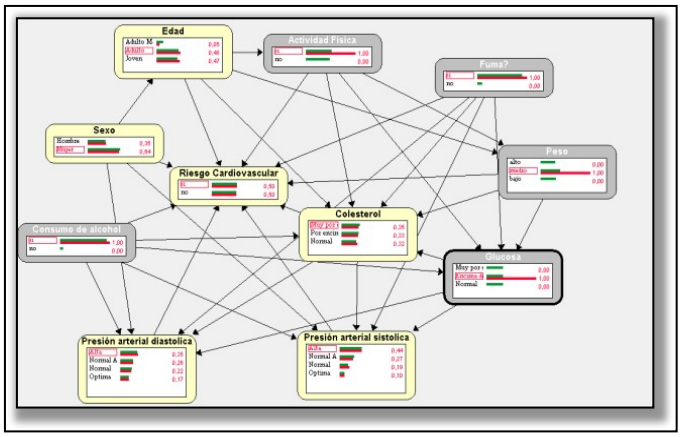

Figura 5. Red Bayesiana con factores de incidencia Caso 1.

Caso 2.- Al encender ciertos nodos al azar vemos variar los porcentajes a priori, esto se debe a la causa que genera esa variación frente a las otras variables, para este evento tenemos el caso de una persona, Hombre, en un rango de edad joven que consume alcohol y fuma, su nivel de colesterol se encuentra muy por encima de lo normal, presenta signos de sobrepeso, que realiza actividad física sin embargo presenta una presión arterial normal alta, dando como resultado del 51% de riesgo cardiovascular, la misma que producirá alguna afección cardiovascular a largo plazo, siendo estos los nodos de causa mayor de RCV.

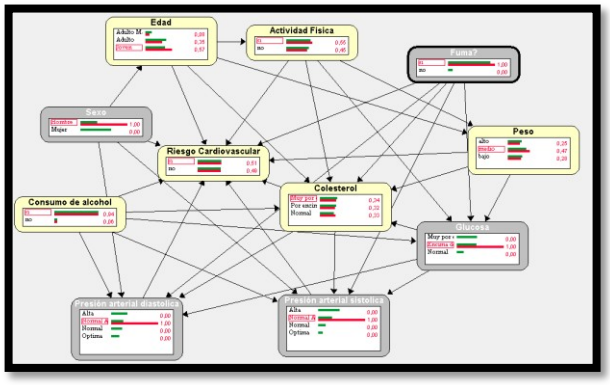

Figura 6. Red Bayesiana con factores de Incidencia Caso 2.

Caso 3.- Al encender ciertos nodos al azar vemos variar los porcentajes a priori, esto se debe a la causa que genera esa variación frente a las otras variables, para este evento tenemos el caso de una persona, una Mujer, en un rango de edad joven que no consume alcohol ni fuma, su nivel de colesterol se encuentra en un nivel normal al igual que presenta un nivel normal de glucosa en la sangre, presenta un peso por debajo de lo normal, realiza actividad física y presenta una presión arterial optima, dando como resultado un 59% de no presentar riesgo cardiovascular. Los nodos de causa mayor de RCV: como colesterol, glucosa, presión arterial, peso y el consumo de alcohol y el fumar.

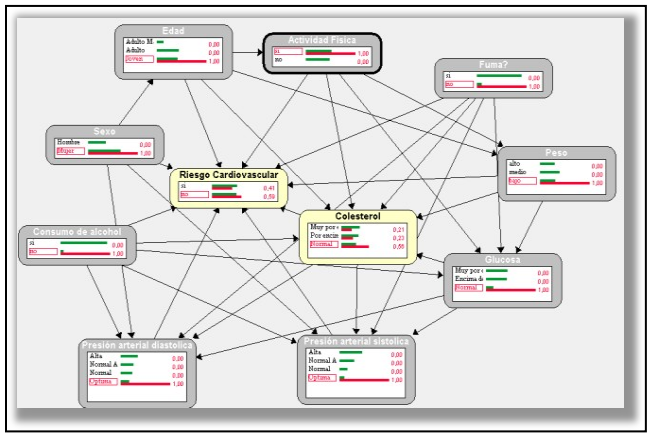

Figura 7. Red Bayesiana con factores de Incidencia Caso 3.

Caso 4.- Al encender ciertos nodos al azar vemos variar los porcentajes a priori, esto se debe a la causa que genera esa variación frente a las otras variables, para este evento tenemos el caso de una persona, un Hombre, en el rango de edad adulta que no consume alcohol, fuma, su nivel de colesterol se encuentra en un nivel normal y presenta un nivel de glucosa en la sangre encima de normal, presenta un peso por encima de lo normal que indica signos de sobrepeso, realiza actividad física, el nivel de presión arterial diastólica es óptimo y mientras que el nivel de presión arterial sistólica es normal, dando como resultado un 52% de no presentar riesgo cardiovascular. Los nodos de causa mayor de RCV son: colesterol, glucosa, presión arterial, peso y el consumo de alcohol y el fumar

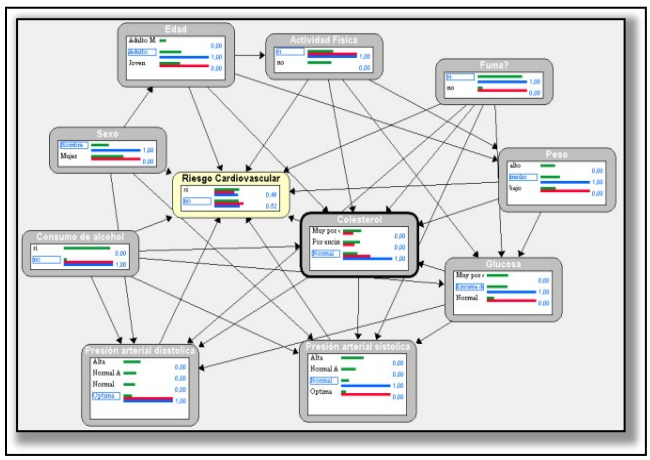

Figura 8. Red Bayesiana con factores de Incidencia Caso 4.

Naive Bayes

La técnica de clasificación Naive Bayes es aplicable principalmente cuando la dimensionalidad de las entradas es alta. A pesar de su simplicidad, Naive Bayes a menudo puede superar más Métodos de clasificación sofisticados. El modelo Naive Bayes reconoce las características de pacientes con ECV. Muestra la probabilidad de cada atributo de entrada para el estado predecible.

La Naive Bayes o la regla de Bayes es la base de muchos métodos de aprendizaje automático y minería de datos. Este se utiliza para crear modelos con capacidades predictivas. Y proporciona nuevas formas de explorar y comprender los datos. Por qué preferir la implementación Naive Bayes:

1. Cuando los datos son altos.
2. Cuando los atributos son independientes entre sí.
3. Cuando esperamos una salida más eficiente, en comparación con otros métodos de salida.

Algoritmo de clasificación Naive Bayes

El Naive Bayes, o clasificador simple bayesiano, funciona de la siguiente manera:

Dado que D sea un conjunto de tuplas de entrenamiento y su clase asociada etiquetas como Ca y Cp. Como de costumbre, cada registro está representado por un vector de atributo n-dimensional, X = ($x1$, $x2$ . . . , xn − 1, xn), representando n mediciones hechas en la tupla de n atributos, es decir, $A1a\ An$.

Suponga que hay m número de clases para predicción, C1, C2 ... Cm. Dado un registro, X, el clasificador predecirá que X pertenece a la clase que tiene la mayor probabilidad posterior, condicionada por X. Es decir, el Naive Bayes predice que la tupla x pertenece a la clase Ci si y solo sí.

$P\ (Ci|X) > P\ (Cj|X)\ for\ 1 \le j \le m\ and\ j \ne i$ Por lo tanto, maximizamos $P\ (Ci|X)$ .La clase Ci para la cual $P\ (Ci|X)$ se maximiza se llama hipótesis del máximo posterior. Por el teorema de Bayes.

Como P (X) es constante para todas las clases, solo $P\ (X|Ci) * P(Ci)$ necesita ser maximizado. Si las probabilidades anteriores de la clase no son conocidas, entonces a menudo se supone que las clases son igualmente probablemente, es decir, P (C1) = P (C2) = . . . P (Cm − 1) = P (Cm) y, por lo tanto, maximizaría $P\ (X|Ci)$. De lo contrario, maximizamos $(X|Ci) * P(Ci)$. Tenga en cuenta que las probabilidades previas de la clase pueden ser estimado por $P(Ci)$ = | Ci, D | / | D |, donde | Ci, D | es el número de tuplas de entrenamiento de clase $Ci\ en\ D$.

Dados conjuntos de datos con muchos atributos, sería extremadamente computacionalmente costoso para calcular $P\ (X|Ci)$. Para reducir el cálculo al evaluar $P\ (X|Ci)$, naive bayes asume la independencia condicional de clase. Esta supone que los valores de los atributos son condicionalmente independientes entre sí, dada la etiqueta de clase de la tupla (es decir, que no hay relaciones de dependencia entre los atributos). Así:

$$P\ (X|C_i) = \prod_{k=1}^{m} P\ (xk|C_i)$$
$$= P\ (x_1|C_i) * P\ (x_2|C_i) * \ldots P\ (x_m|C_i)$$

Podemos estimar fácilmente las probabilidades $P\ (x1|Ci)$, $P\ (x2|Ci)$ … $P\ (xm|Ci)$ de las tuplas de entrenamiento de la base de datos. Hay que recordar que aquí xk se refiere al valor del atributo Ak para la tupla X. Para cada atributo, veremos que si el atributo es categórico o de valor continuo. Por ejemplo, para calcular $P\ (x1|Ci)$, consideramos lo siguiente:

Si Ak es categórico, entonces $P\ (xk|Ci)$ es el número de tuplas de clase Ci en D que tiene el valor xk para Ak, dividido por | Ci, D |, el número de tuplas de la clase Ci en D.

Para predecir la etiqueta de clase de X, $P\ (X|Ci)\ P(Ci)$ es evaluado para cada clase $P(Ci)$. El clasificador predice que la etiqueta de clase de la tupla X es la clase Ci si y solo si $P\ (X|Ci)P(Ci) > P\ (x|Cj)\ P(Cj)\ \ for\ 1 \le j \le m, j \ne i$. En otras palabras, la etiqueta de clase predicha es la clase Ci para cual $P\ (X|Ci)\ P(Ci)$ es el máximo.

Arquitectura del sistema

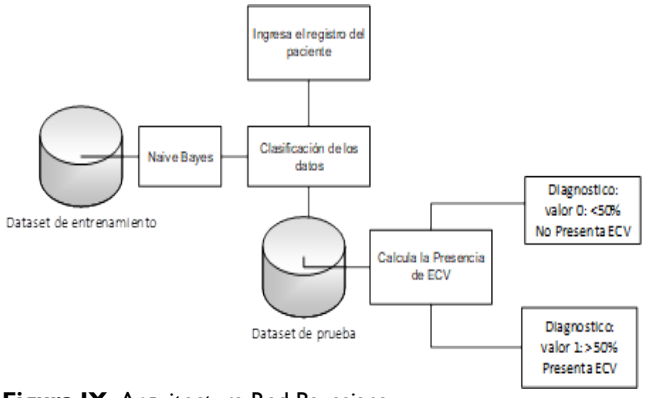

Figura IX. Arquitectura Red Bayesiana

Se propone un prototipo de página web que permita a los usuarios obtener orientación instantánea sobre su ECV a través de un sistema inteligente en línea. La aplicación permite al paciente compartir sus problemas relacionados con el corazón. Luego procesa los detalles específicos del paciente para verificar si hay presencia de una ECV.

En función del resultado, el sistema muestra automáticamente el resultado a médicos específicos para un tratamiento adicional. El sistema se puede usar en caso de emergencia. El objetivo principal de este sistema es predecir la presencia de ECV mediante la técnica de minería de datos, como el algoritmo Naive bayesiano. Se utiliza el conjunto de registros médicos sin procesar, luego se procesan y transforman el conjunto de datos. Luego aplicar la técnica de minería de datos como el algoritmo Naive Bayes en el conjunto de datos transformado. Se predice la ECV y el usuario recibe el resultado en función de la predicción de presencia de ECV y la probabilidad de presencia de esta.

Proceso de entrenamiento

El proceso de clasificación asume datos etiquetados: sabemos cuántas clases hay y tenemos ejemplos para cada clase. La clasificación es supervisada.

Clasifica datos (construye un modelo) según el conjunto de entrenamiento y los valores (etiquetas de clase) en un atributo de clasificación y lo usa para clasificar nuevos datos.

Para el proceso de entrenamiento dividimos el Dataset y utilizamos el 70% de los registros para el entrenamiento para el algoritmo de clasificación.

Fase de prueba

La fase de prueba implica la predicción de una muestra de datos desconocidos. En las pruebas, verificamos los datos que no entrar en el conjunto de datos que hemos considerado (Dataset de pruebas). Después de la predicción, obtendremos las etiquetas de clase.

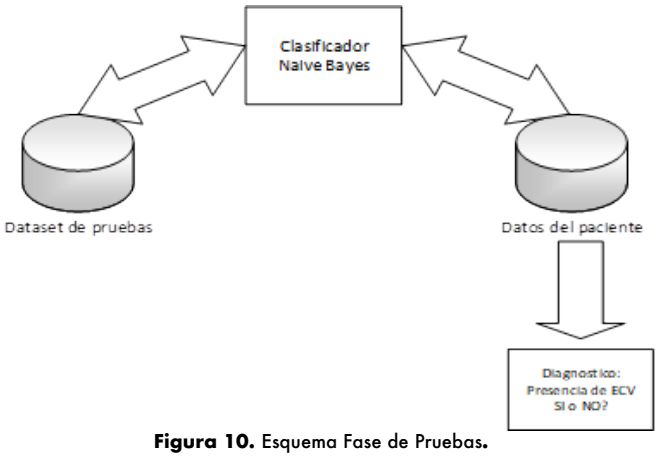

Figura 10. Esquema Fase de Pruebas.

La siguiente tabla muestra la precisión obtenida tomando diferentes bloques de registros del conjunto de datos de entrenamiento.

Tabla 5. Precisión del conjunto de entrenamiento

Registros de entrenamiento	Registros de prueba	Registros correctamente clasificados	Registros incorrectamente te clasificados
500	350	304	46
500	430	382	50
500	490	422	68

Conclusiones y Trabajos futuros.

En el presente trabajo se han utilizado los factores de riesgo más importantes y predecir el riesgo a enfermedades cardiovasculares. El estudio resulta de gran ayuda a los especialistas, ya que a través de esta técnica de aprendizaje automático hemos podido determinar al nivel de riesgo que tiene un individuo de adquirir alguna enfermedad.

Se culminó con éxito el desarrollo de un prototipo de página web dinámica logrando obtener un archivo que nos permite efectuar la clasificación o la predicción de ECV a cualquier paciente que presente factores de riesgo que inciden en ellas.

El Machine Learning es una técnica que está implementando en diversas las áreas alrededor del mundo, sin embargo, para este trabajo se ha considerado su aplicación en el área de la medicina para hacer una predicción de enfermedades cardiovasculares, mediante la implementación del algoritmo Naive Bayes y Elvira nos ha permitido modelar este tipo de enfermedad para realizar predicciones y lograr un entrenamiento con un porcentaje de error del 20%.

Se recomienda hacer uso de varios indicadores a la hora de realizar comparaciones entre algoritmos; ya que de esta manera se puede tener un panorama desde varias perspectivas de la performance de aprendizaje, así mismo que se realicen estudios de la aplicabilidad del Machine Learning en las demás ramas de la ciencia, utilizando otros métodos predictivos con la finalidad de mejorar el proceso del tratamiento masivo de datos.

Se recomienda la implementación de esta investigación en el desarrollo de una aplicación móvil, permitiendo así que la herramienta esté al alcance de todos los ciudadanos que cuenten con un dispositivo inteligente, con la facilidad de la implementación un módulo para que el sistema pueda generar reportes y que estos puedan ser enviados vía correo electrónico.

Referencias Bibliográficas

Moreno, G. A. (2008). La definición de salud de la Organización Mundial de la Salud y la interdisciplinariedad. Sapiens. Revista Universitaria de Investigación, 9(1), 93-107.

Vega Romero, R. (2009). Informe comision Determinantes sociales de la salud de la organizacion mundial de la salud. Revista Gerencia y Políticas de Salud, 8(16), 7-11.

Fernández, E., Sabán, J., Fabregate, M., & Fabregate, R. (2009). Epidemiología de la enfermedad cardiovascular. Control total del riesgo cardiometabólico. Madrid: Díaz de Santos, 31-77.

Chávez Domínguez, R., Ramírez Hernández, J. A., & Casanova Garcés, J. M. (2003). La cardiopatía coronaria en México y su importancia clínica, epidemiológica y preventiva. Archivos de cardiología de México, 73(2), 105-114.

Volschan, A., Caramelli, B., Gottschall, C. A. M., Blacher, C., Casagrande, E. L., & Manente, E. R. (2004). Diretriz de embolia pulmonar. Arq Bras Cardiol, 83(Suppl 1), 1-8.

Salazar ÁLvarez, Y. (2011). Uso de la metformina en la diabetes mellitus tipo II. Revista Cubana de Farmacia, 45(1), 157-166.

Murillo, A. Z., & Esteban, B. M. (2005). Obesidad como factor de riesgo cardiovascular. Hipertensión y riesgo vascular, 22(1), 32-36.

Guarnaluses, L. J. B., & Ramos, A. P. (2016). Factores de riesgo de los accidentes cerebrovasculares durante un bienio. Medisan, 20(5), 666-674.

Encinas, C., Sacristán, J., Cenamor, D., & Morell, L. (2019). SOFIA: Soporte Fármaco Terapéutico Inteligencia Artificial. I+ S: Revista de la Sociedad Española de Informática y Salud, (137), 7-14.

Quesada, Y., Wong, D., & Rosete, A. (2012). Minería de Datos aplicada a la Gestión Hospitalaria. 14 Convención Científica de Ingeniería y Arquitectura, 2-5.

De Mitri, M. J. (2019). Predicción de marcadores cardíacos en pacientes diabéticos e hipertensos medicados por medio de inteligencia artificial (Doctoral dissertation, Universidad Católica de Córdoba).

Aranda Núñez, A. P. (2019). Estudio de la relación del esfuerzo de corte con la presión en aneurismas cerebrales y la predicción del riesgo de ruptura usando herramientas de inteligencia artificial basado en datos morfológicos, fluidodinámicos y estructurales.

Moncayo, K. C., Sanchez, A. G., Anton, P. R., & Cevallos-Torres, L. (2019). Modelo de simulación para la optimización del inventario de una distribuidora, basado en Simulación Monte Carlo y Algoritmo Metaheurístico Genético. Ecuadorian Journal of Science, Research and Innovation, 3(2), 33-38.

Portugal, R., & Carrasco, M. (2007, January). Ensamble de Algoritmos Bayesianos con Árboles de decisión: una alternativa de clasificación. In XVII Congreso Chileno de Control Automático ACCA, Universidad de la Frontera, Chile.

Pereira-Toledo, A., López-Cabrera, J. D., & Quintero-Domínguez, L. A. (2017). Estudio experimental para la comparación del desempeño de Naïve Bayes con otros clasificadores bayesianos. Revista Cubana de Ciencias Informáticas, 11(4), 67-84.

O'Donnell, C. J., & Elosua, R. (2008). Factores de riesgo cardiovascular. Perspectivas derivadas del Framingham Heart Study. Revista española de Cardiología, 61(3), 299-310.

Medrano, M. J., Cerrato, E., Boix, R., & Delgado-Rodríguez, M. (2005). Factores de riesgo cardiovascular en la población española: metaanálisis de estudios transversales. Medicina clínica, 124(16), 606-612.

Álvarez Cosmea, A. (2001). Las tablas de riesgo cardiovascular: Una revisión crítica. Medifam, 11(3), 20-51.

Molina Espinoza, C. I. (2014). Detección temprana de riesgo cardiovascular usando text mining en los campos de texto no estructurado del registro clínico electrónico.

González, F. A. (2015). Machine learning models in rheumatology. Revista Colombiana de Reumatología, 22(2), 77-78.

Rodriguez, C. A., et al. "Clasificación de latidos de contracción ventricular prematura basados en métodos de aprendizaje no supervisado." Revista Ingeniería Biomédica 8.15 (2014): 51-58.

Pedregal, P. (2006). *Introduction to optimization* (Vol. 46). Springer Science & Business Media.