CAPÍTULO 5

UNA EXPERIENCIA DE ANÁLISIS DE LA CALIDAD DATOS EN EL CAMPO DE LA SALUD PÚBLICA*

Ricardo Ruiz Cortés[1], Kevin Andrés Prieto Cruz[2], Walter Hugo Arboleda Mazo[3], Raquel Anaya Hernandez[4], Arturo Jesús Laflor[5]

Resumen

Disponer de información confiable y precisa en el campo de la salud pública, es esencial para monitorear la salud y para evaluar y mejorar la prestación de servicios y los programas de atención de salud.

*Capítulo de libro de investigación resultado del proyecto titulado "Una experiencia de análisis de la calidad de datos en el campo de la salud pública"

1 Especialista en Big Data e Inteligencia de Negocios. Ingeniero de Sistemas. Investigador, Grupo de Investigación en Ingeniería Aplicada GI2A, Facultad de Ingeniería, Ingeniería de Sistemas, Corporación Universitaria Adventista. Correo electrónico: ricardo.ruiz@unac.edu.co, Orcid: https://orcid.org/0000-0002-5378-6823

2 Ingeniero de Sistemas, Ingeniería de Sistemas, Corporación Universitaria Adventista. Correo electrónico: kaprieto@unac.edu.co

3 Estudiante de Doctorado en Filosofía en Tecnología de la Información, Magíster en Ingeniería, Especialista en Teleinformática, Ingeniero de Sistemas. Investigador, Grupo de Investigación en Ingeniería Aplicada GI2A, Facultad de Ingeniería, Ingeniería de Sistemas, Corporación Universitaria Adventista. Correo electrónico: warboleda@unac.edu.co, Orcid: https://orcid.org/0000-0003-4937-5359

4 Doctora en Ingeniería de la Programación Artificial, Ingeniera de Sistemas, Investigadora, Grupo de Investigación en Ingeniería Aplicada GI2A, Facultad de Ingeniería, Ingeniería de Sistemas, Corporación Universitaria Adventista. Correo electrónico: raquel.anaya.hdez@gmail.com, https://orcid.org/0000-0002-9187-7427

5 Maestro en Ingeniería, Ciencias de la Computación. Ingeniero de Sistemas Computacionales, Investigador. Correo electrónico: arturo.laflor@uabc.edu.mx

Este proyecto de investigación aplicada tiene como finalidad analizar la calidad de los datos de un programa de salud pública de una secretaria de salud y aplicar mecanismos reconocidos para la depuración de los datos de tal manera que puedan ser utilizados para aplicar procesos automáticos de análisis de datos.

La metodología aplicada es CRISP-DM (Cross Industry Standard Process for Data Mining) que proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos.

Para el entendimiento y preparación de los datos se utilizaron unas plantillas propuestas por un grupo interdisciplinar de investigación las cuales permitieron identificar las características de los datos y su nivel de calidad, a través de las siguientes tareas: Identificar el tamaño del conjunto de datos, realizar el análisis de las propiedades de cada característica que compone el conjunto de datos y definir el tratamiento de los datos faltantes.

Para la caracterización inicial de los datos se utilizó Excel y para realizar el proceso de perfilamiento de datos y variables se utilizó el software DQ Analyzer, posteriormente para la visualización de los datos se utilizó el software Microstrategy.

Como resultados de la investigación se encontraron datos desbalanceados para las variables objeto de estudio, así como datos faltantes, cardinalidad y falta de precisión en el análisis estadístico como la moda, la mediana y cuartiles.

Palabras clave: Big Data, promoción de la salud, prevención de la salud, enfermedades no transmisibles, ciencia de datos, inteligencia de negocios, aprendizaje automático.

Abstract

Having reliable and accurate information in the field of public health is essential to monitor health and to evaluate and improve the delivery of

services and health care programs.

This applied research project aims to analyze the quality of data from a public health program run by a secretary of health and apply recognized mechanisms for data purification in such a way that they can be used to apply automatic data analysis processes.

The applied methodology is CRISP-DM (Cross Industry Standard Process for Data Mining) which provides a standardized description of the life cycle of a standard data analysis project.

For the understanding and preparation of the data, some templates proposed by an interdisciplinary research group were used, which allowed identifying the characteristics of the data and its level of quality, through the following tasks: Identify the size of the data set, carry out the analysis of the properties of each characteristic that makes up the data set and define the treatment of the missing data.

For the initial characterization of the data, Excel was used and the DQ Analyzer software was used to perform the data and variable profiling process, later for the data visualization, the Microstrategy software was used.

As results of the research, unbalanced data were found for the variables under study, as well as missing data, cardinality and lack of precision in the statistical analysis such as mode, median and quartiles.

Keywords: Big Data, health promotion, health prevention, Noncommunicable Diseases, data science, business intelligence, machine learning.

Introducción

Los datos para la promoción y prevención, provienen de las Entidades Generadoras de Salud, estas se encargan de estandarizar indicadores de salud (Strome, 2014).

Aunque existan muchos datos, se encuentra que están principalmente enfocados en aspectos epidemiológicos, dejando de lado los factores de estilo de vida (Imran, 2018).

Para que los datos puedan servir como elementos valiosos para analizar tendencias y establecer políticas de salud, utilizando herramientas informáticas, se hace necesario enfocar esfuerzos para evitar mal tratamiento de los datos y lograr su calidad (Nelson, 2018).

DESAROLLO

Planteamiento del problema y justificación

Un correcto diagnóstico para la creación de programas y proyectos de intervención de la comunidad a nivel municipal (Tawalbeh, Mehmood, Benkhlifa, & Songs, 2016), depende de la calidad de los datos recolectados sobre hábitos saludables en las campañas de salud (Zhang, Qiu, Tsai, Hassan, & Alamri, 2017), es así como realizar una correcta revisión de la forma como son recolectados los datos es de gran importancia en la salud pública.

Las entidades y organismos de salud a nivel regional y nacional e Internacional (EPS, IPS, secretarias de Salud, Ministerio de Salud y Protección Social, OPS, OMS), reconocen hoy en día la importancia de promover entre la población un estilo de vida saludable, como una de las estrategias claves para prevenir las Enfermedades No Trasmisibles. El Plan Decenal de Salud Pública (PDSP) para el período (2012 a 2021) tiene como visión “la salud como un derecho fundamental, dimensión central del desarrollo humano” y una de sus dimensiones es la de Vida Saludable y Condiciones no Trasmisible con dos componentes básicos: modos, condiciones y estilos de vida saludable y condiciones crónicas prevalentes, dentro del

cual ya se tienen resultados preliminares, la intención es descubrir más información de valor para este tipo de poblaciones.

El público objetivo son aquellas personas que trabajan en el campo de la salud pública interesadas en analizar datos para el cuidado de la salud y mejoramiento del estilo de vida.

Contexto

Entidad: Secretaria de Salud de Bello, Unidad: Promoción y prevención en salud.

Fuente de datos

Los datos con los que se cuentan actualmente en el proyecto se obtuvieron de una fuente primaria, almacenados y procesados por el Sistema GENESIS en su versión 1.0, desarrollo propio de la Secretaria de Salud el cual está desarrollado en ASP.NET 4.7 – C#, base de datos: MS SQL Server 2012 SP4, para equipos de escritorio, alojado en un servidor Windows Server 2012, la información inicial se obtiene usando un instrumento de medición como lo es una encuesta de forma manual en las actividades llamadas barriadas, luego se lleva a cabo el proceso de digitación de la información, de esta manera queda almacenada y dispuesta para ser exportada en formato Excel.

Entendimiento del problema

En los resultados preliminares de la investigación se logró realizar una fase donde se estableció una entrevista con la persona encargada del área y se definieron las variables de interés objeto de estudio la siguiente actividad se realizó un primer filtro a los datos iniciales de la encuesta para solo dejar los datos de interés: Datos generales y datos con respecto al riesgo cardiovascular los cuales fueron: barrio, zona, comuna, fecha de nacimiento, sexo, grupo poblacional: indígena, afrocolombiano, Gitano, raizal, mestizo, condición especial: víctima, mujer cabeza de familia, desplazado, LGTBI, desmovilizado, red unidos, discapacidad, riesgo

cardiovascular: hipertensión, diabetes, actividad física, licor, fuma, peso, talla, IMC, presión arterial, perímetro abdominal.

Es por este motivo que se plantean las siguientes preguntas de investigación: ¿Es posible que a partir del análisis de los datos se obtenga un conocimiento más profundo de la población objeto de estudio? ¿Se puede establecer un plan de estilo de vida saludable con base en el conocimiento producto del análisis de los datos?

Objetivo general

Aplicar Técnicas de Minería de Datos que provean la toma de decisiones con los datos recolectados en las jornadas de promoción y prevención en la Secretaría de Salud de Bello.

Objetivos específicos

- Identificar la(s) fuente(s) y estructura(s) de los datos clínicos de la población objeto de estudio.
- Realizar el montaje de los datos clínicos y datos del estilo de vida en el entorno de análisis de datos seleccionado.
- Realizar el análisis de los resultados obtenidos.
- Plantear una infraestructura tecnológica para el análisis de los datos.

Metodología

Enfoque de investigación aplicada usando técnicas de minería de datos sobre la población objeto de estudio que fue un universo limitado al número de encuestados específicamente mayores de 18 años y que tienen los datos de interés, en este aspecto se trata específicamente el riesgo cardiovascular como uno de los factores asociados a enfermedades crónicas no transmisibles y correlacionado con el estilo de vida al tiempo

que se pueden establecer asociaciones y hacer un perfilamiento de los encuestados.

Se aplicó la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) que proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos.

Figura 1. Metodología CRISP-DM

En el desarrollo del proyecto se plantearon las siguientes fases

Fase 1: Conocimiento del negocio.
Fase 2: Conocimiento de los datos.
Fase 3: Preparación de los datos sobre riesgo cardiovascular
Fase 4: Elaboración de reporte final sobre calidad de los datos.

En el marco del convenio realizado con investigadores de la Universidad Autónoma de Baja California se tiene en cuenta el aporte a través del proyecto "Estrategia de estudio preliminar de la viabilidad de la aplicación

de modelos de Machine Learning a bases de datos según la naturaleza de los datos" (Díaz-Valladares, 2016) en el cual se proponen cinco pasos para la implementación de un descubrimiento de conocimiento de datos (KDD).

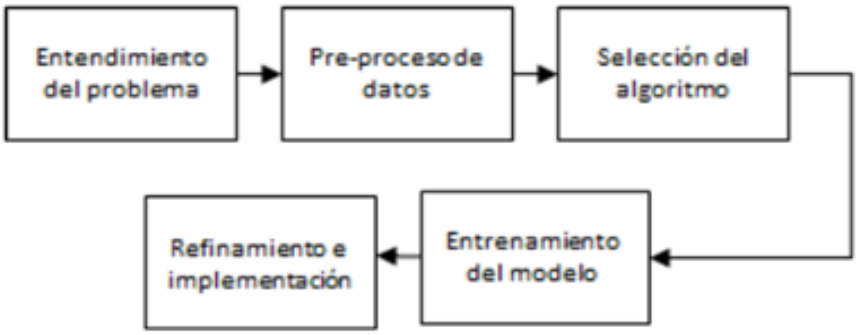

Figura 2. Esquema de pasos para aplicación de algoritmo ML.

Método

El pre-proceso de datos comprende las siguientes tareas asociadas: el tamaño del conjunto de datos, el análisis de las propiedades de cada característica que compone el conjunto de datos y el tratamiento de los datos faltantes.

Tabla 1.
Datos cantidad de encuestados

Año	Cantidad
2014	7.391
2015	14.041
2016 – hasta 6/29/2017	15.221
Total	36.653

Análisis de propiedades de las características

Tabla 2.
Variables objeto de estudio

N°	Categoria	Nombre	Tipo de dato	Valor que puede tomar	Descripción en el contexto	Total
1	Demográfica	Barrio	String / Integer	Lugar	Parte de una población de extensión relativamente grande, que contiene un agrupamiento social.	6
2	Demográfica: Grupo Poblacional	Indígena-Afrocolombiano-Gitano-Raizal-Mestizo	Booleano	Verdadero / Falso	Grupo Poblacional	5
3	Demográfica: Condición especial	Victima - Mujer cabeza de flia - Desplazado - LGTBI - desmovilizado	Booleano	Verdadero / Falso	Condición especial	7
4	Enfermedades	Hipertensión - Diabetes	Booleano	Verdadero / Falso	Riesgo cardiovascular	2
5	Estilo de vida	Actividad física - licor - fuma	Booleano	Verdadero / Falso	Factor protector - Factor de riesgo	3
6	Fisiologica	Perimetro abdominal	Integer	102 cm	Factores que inciden en el estilo de vida de la persona	5

Las siguientes tablas donde se muestran las variables continuas y categóricas se realizaron con Microsoft Excel, donde se realizó el filtrado de los datos a usar, además del uso de fórmulas aplicando estadística descriptiva.

Tabla 3.
Variables continuas

Variable	Cantidad	Faltantes	Cardinalidad	Mínimo	Cuartil-1	Mediana	Cuartil-3	Media	σ
Peso kg	6217	1295	231	2	37	52	64	55,19	79,24
Talla cm	6217	1227	112	14	141	153	160	126,51	62,17
IMC	6217	942	481	1	16,4	21	26	25,73	37,53
Presión Arterial Sistole	6217	3425	195	10	110	120	130	119,95	15,39
Presión Arterial Diastole	6217	4790	218	20	70	80	80	74,68	10,31
Perimetro abdominal	6217	1520	83	3	0	0	83	49,60	1461,42

Tabla 4.
Variables categóricas

Variable	Cantidad	Faltantes	Cardinalidad	Moda	Frecuencia de la moda	Porcentaje de la moda	2da. Moda	Frecuencia de la 2da moda	Porcentaje de la 2da moda
Hipertensión	7391	322	2	FALSO	4878	65,99	VERDADERO	2190	29,62
Diabetes	7391	2	2	FALSO	6512	88,09	VERDADERO	878	11,87
Actividad fisica	7391	2	2	VERDADERO	4889	66,13	FALSO	2502	33,84
Licor	7391	2	2	FALSO	6502	87,95	VERDADERO	889	12,06
fuma	7391	2	2	FALSO	6964	94,20	VERDADERO	427	5,77

Análisis de datos faltantes

La ausencia de datos, es un tema no exclusivo de ML, ha sido tema de estudio de los procesos estadísticos por las implicaciones que esto tiene, en este caso para el proyecto se realiza un análisis de los datos faltantes y se evalúa el impacto que tiene al ser aislados del universo de los datos a trabajar, obteniendo de esta manera que el impacto es menor dado que se tiene una muestra significativa de los datos objeto de estudio con la cual se pueden obtener resultados relevantes.

Limpieza de los datos

El peso, la talla y el índice de masa corporal (IMC) son exportados por el sistema como números enteros, los cuales se convirtieron a decimales para el correcto análisis. Esta limpieza se hizo directamente en Excel a través de macros para generar el valor correcto de las variables.

Resultados

- Aplicación de técnica de Minería de Datos (Análisis Exploratorio de Datos, Ingeniería de características, Análisis Estadístico de Datos)
- Separación de los datos en variables continuas y variables categóricas para generar un reporte de calidad de datos por cada tipo.
- Diseño de una infraestructura tecnológica para el análisis de los datos objeto de estudio.
- Visualización de datos con el Software MicroStrategy para obtener gráficos de correlaciones u otras relaciones entre los datos objeto de estudio.

Tabla 5.
Análisis de datos faltantes de variables categóricas (año 2014)

Variable	Observaciones
Hipertensión	De los 7391 datos objeto de estudio, se hallaron 7069 que poseen el booleano de Hipertensión, Diabetes, Actividad fisica, Licor y Fumar, correspondiente al 95.6 % del total de los datos objeto de estudio.
Diabetes	
Actividad fisica	
Licor	
fuma	

Tabla 7.
Análisis de datos faltantes de variables continuas (año 2015)

Variable	Observaciones
Peso kg	De los 6217 datos objeto de estudio, se hallaron 4595 que poseen valor en el peso, talla, IMC, presión arterial (PAS/PAD) y perímetro abdominal, correspondiente al 73.9 % del total de los datos objeto de estudio.
Talla cm	
Imc	
Presión Arterial Sistólica	
Presión Arterial Diastólica	
Perímetro abdominal	

Discusión

A partir de los resultados obtenidos es necesario desarrollar un plan de acción para la mejora en la calidad de los datos, y esto permitirá dar respuesta a la segunda pregunta ¿Cuáles son los principales análisis

obtenidos de este conjunto de datos?, pues determina la manera de aplicar algoritmos de Machine Learning estableciendo modelos que pueden facilitar la comprensión y entendimiento de patrones a partir de los datos objeto de estudio.

Tabla 8.
Resultados y plan de mejora

Variable	Irregularidad	Plan para atender la irregularidad
Peso	Valores fuera de rango.	Eliminar dichos valores presto que la cantidad es relativamente baja. En el software agregar validaciones para admisión solamente de campos enteros y agregar límites equivalentes a la variable, se deben eliminar estos valores cuando se realiza la preparación de los datos , llamando a este proceso reformateo de los datos, y se procede a hacer nuevamente un análisis de los datos para conocer su comportamiento.
Talla	Valores fuera de rango.	Eliminar dichos valores presto que la cantidad es relativamente baja. En el software agregar validaciones para admisión solamente de campos enteros y agregar límites equivalentes a la variable, se deben eliminar estos valores cuando se realiza la preparación de los datos , llamando a este proceso reformateo de los datos, y se procede a hacer nuevamente un análisis de los datos para conocer su comportamiento.
Perímetro abdominal	Valores fuera de rango.	Eliminar dichos valores presto que la cantidad es relativamente baja. En el software agregar validaciones para admisión solamente de campos enteros y agregar límites equivalentes a la variable, se deben eliminar estos valores cuando se realiza la preparación de los datos , llamando a este proceso reformateo de los datos, y se procede a hacer nuevamente un análisis de los datos para conocer su comportamiento.
Presión arterial sistólica	Valores fuera de rango.	Eliminar dichos valores presto que la cantidad es relativamente baja. En el software agregar validaciones para admisión solamente de campos enteros y agregar límites equivalentes a la variable, se deben eliminar estos valores cuando se realiza la preparación de los datos , llamando a este proceso reformateo de los datos, y se procede a hacer nuevamente un análisis de los datos para conocer su comportamiento.

Conclusiones

El uso de técnicas de minería de datos específicamente la analítica se basa en el hecho de que es posible inferir comportamiento o caracterizar fenómenos de la realidad a partir de la observación y el análisis de datos asociados a dicho fenómeno.

La experiencia en la realización de este trabajo ha aportado y enriquecido el conocimiento de los autores y se obtienen las siguientes

conclusiones con el propósito de continuar con las siguientes fases del proyecto:

Mejorar la captura de la información al momento de realizar la jornada, en lo posible lograr que la captura en el sistema se haga directamente en el sitio de la encuesta. (Usar dispositivos digitales (Hardware - Software))

Mejorar el formato para obtener información con mayor precisión y disponible para su medición (Ejemplo: Fuma: entre 1 - 5, 6 - 10; cigarrillos diarios).

Mejorar la digitación de la información (Debe ser información que esté en el formato y digitarla de manera completa); hacer procesos de validación iniciales para evitar valores fuera de rango y datos faltantes.

Mejorar la exportación de los datos del sistema (ejemplo: uso de decimales para la talla, IMC, peso, etc)

El hallazgo en los datos objeto de estudio se evidencian las posibles enfermedades asociadas al riesgo cardiovascular lo cual conlleva a generar soluciones para mejorar el estilo de vida de la población en estudio.

REFERENCIAS

Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease Prediction by Machine Learning over Big Data from Healthcare Communities. IEEE Access, 3536(c), 1–10. https://doi.org/10.1109/ACCESS.2017.2694446

Imran, M. (2018). Big Data Analytics Tools and Platform in Big Data Landscape. Health Journal, 10(2), 23-27

Kupwade Patil, H., & Seshadri, R. (2014). Big Data Security and Privacy Issues in Healthcare. 2014 IEEE International Congress on Big Data, (November), 762–765. https://doi.org/10.1109/BigData.Congress.2014.112

Nambiar, R., Bhardwaj, R., Sethi, A., & Vargheese, R. (2013). A look at challenges and opportunities of Big Data analytics in healthcare. Proceedings - 2013 IEEE International Conference on Big Data, Big Data 2013, 17–22. https://doi.org/10.1109/BigData.2013.6691753

Nelson, R. (2018). Health informatics: An interprofessional approach. Health Journal, 23(4), 32-45

Strome, T. (2014). Healthcare analytics for quality and performance improvement. Journal in Systems and Health,20(1), 56-78.

Tawalbeh, L. A., Mehmood, R., Benkhlifa, E., & Songs, H. (2016). Mobile Cloud Computing Model and Big Data Analysis for Healthcare Applications. IEEE Access, 4, 6171–6180. https://doi.org/10.1109/ACCESS.2016.2613278

Zhang, Y., Qiu, M., Tsai, C. W., Hassan, M. M., & Alamri, A. (2017). Health-CPS: Healthcare cyber-physical system assisted by cloud and big data. IEEE Systems Journal, 11(1), 88–95. https://doi.org/10.1109/JSYST.2015.2460747

Heinrich, A. (2016). Big Data Technologies in Healthcare Needs, opportunities and challenges. Big Data Technologies in Healthcare Needs, opportunities and challenges, 31. Recuperado a partir de http://www.bdva.eu/sites/default/files/Big Data Technologies in Healthcare.

pdf

Díaz-Valladares, R. A. (2016). Oportunidades para la investigación e innovación. Nuevo León, México: Publicaciones Universidad de Montemorelos.