RESUMEN ANALÍTICO EN EDUCACIÓN - RAE -	UNIVERSIDAD CATÓLICA de Colombia Vigilada Mineducación RIUCaC

FACULTAD INGENIERÍA
PROGRAMA DE DE INGENIERÍA DE SISTEMAS
BOGOTÁ D.C.

AÑO DE ELABORACIÓN: 2017

TÍTULO: Minería de datos para el descubrimiento de patrones en enfermedades respiratorias en Bogotá, Colombia

AUTOR (ES): Aguilar,Juan Sebastian y Rojas Gutierrez, Erika Andrea

DIRECTOR(ES)/ASESOR(ES): Velandia, John Alexander

MODALIDAD: Trabajo de investigación

PÁGINAS: 80 **TABLAS:** 3 **CUADROS:** **FIGURAS:** 19 **ANEXOS:** 2

CONTENIDO:

13. REFERENCIAS BIBLIOGRÁFICAS
14. ANEXOS

ANEXOS

RESUMEN: El presente proyecto se basa en la aplicación de minería de datos mediante el algoritmo de clustering K- means que permita la generación de un modelo descriptivo con el análisis de los datos y con el objetivo de identificar posibles comportamientos en enfermedades respiratorias en la ciudad de Bogotá. El conjunto de clústeres generados por la herramienta RapidMiner es la recopilación de datos de un periodo de cinco años de 2012 a 2016, en donde se contemplan el número de casos asociados a 184 diagnósticos de enfermedades respiratorias y la edad de los pacientes corresponde de 0 a 5 años.

METODOLOGÍA: La metodología consta de 7 pasos los cuales se muestran a continuación:

1. **Obtener las bases de datos:** Se buscan las fuentes de extracción, posteriormente se califican las fuentes de acuerdo a algunos criterios de selección que están asociados a calidad de datos, para la finalmente realizar la definición de las variables.
2. **Obtener la herramienta de análisis de datos a utilizar:** Se busca la herramienta de análisis de datos con la cual se va a trabajar, para el caso se escogió la herramienta RapidMiner.
3. **Aplicar las bases de datos en la herramienta de análisis:** Se aplican los datos obtenidos de las fuentes de extracción dentro de la herramienta de análisis de datos.
4. **Aplicar los algoritmos de búsqueda:** Con el objetivo que permita identificar patrones de comportamiento dentro de las bases de datos aplicadas en la herramienta de análisis, por tal motivo, se selecciona un algoritmo de Clustering para ejecutarla en RapidMiner.
5. **Realizar la búsqueda de datos que permitan identificar patrones de comportamiento en dichas enfermedades:** Se ejecuta el algoritmo sobre los datos a analizar dentro de la herramienta de análisis de datos.

6. **Obtener los resultados de las búsquedas:** Se extraen los resultados obtenidos de la búsqueda en la herramienta de análisis de datos para los patrones identificados.
7. **Realizar la documentación sobre resultados obtenidos:** Se realiza la documentación de acuerdo con los resultados obtenidos de la herramienta de análisis de datos.

PALABRAS CLAVE: ARQUITECTURA DE DATOS, CLUSTERING, CALIDAD DE DATOS, DIAGNÓSTICOS, FUENTES DE EXTRACCIÓN, K-MEANS, TÉCNICA DE MINERÍA DE DATOS.

CONCLUSIONES:

1. Como resultado de las fuentes de extracción se obtuvieron los datos correspondientes a cada una de las variables definidas; sin embargo, los datos de las fuentes que se tuvieron en cuenta inicialmente, no contaban con suficientes registros para aplicar minería de datos, ya que algunas variables de las fuentes de extracción presentaban datos resumidos.
No obstante, se tuvo acceso finalmente a la fuente del cubo SISPRO, que permitió realizar la extracción de manera integral de las variables que se acercaban al objetivo general. Las variables extraídas de esta fuente de extracción son: tipo de enfermedad, año, edad, número de caso y género para la ciudad de Bogotá.

2. La gestión de los datos permitió darle a los mismos una estructura y posteriormente convertirlos en información mediante la aplicación de minería de datos RapidMiner, y finalmente realizar un análisis sobre un volumen de 10.000 registros para la aplicación de clustering. Por otro lado, las arquitecturas de datos permitieron tener una idea general de las fuentes de información, conocer donde se puede extraer datos para la realización de futuras investigaciones y las variables correspondientes a cada fuente de datos.
A las variables tipo de enfermedad y género se realizó normalización, debido a que el modelo solo recibe datos tipo numéricos.

3. La selección e implementación de algoritmos permitió conocer algoritmos de acuerdo a sus funcionalidades con respecto a la clasificación de algoritmos de clustering, en el que se tomaron diferentes métricas para evaluar dichos algoritmos, ayudando en la búsqueda y elección de estos. Esta selección tiene como aprendizajes el conocimiento de clasificación en los algoritmos de clustering, conocer sus funcionalidades y ver cual puede llegar a ser el adecuado para la aplicación de un caso de análisis.
También permitió conocer las herramientas que aplican análisis de datos donde permiten la ejecución de varios algoritmos de búsqueda de datos, a través de un proceso que se diseña en la herramienta para que esta pueda ejecutar de acuerdo al orden establecido en el diseño del proceso y dar resultados

4. El reconocimiento de patrones se puede llegar a realizar a través del análisis de datos por medio de los resultados obtenidos los cuales dan a conocer diferentes comportamientos a través de la agrupación obtenida en los clústeres generados, ya que cada clúster contiene un caso con relación de las variables trabajadas. Reconociendo como un comportamiento con respecto a los resultados obtenidos que la mayor parte de los casos en todas las edades presentaron enfermedades con códigos entre el rango de 0 a 60 aproximadamente.

Los clústeres obtenidos realizaron la correspondiente agrupación de acuerdo a los parámetros iniciales que pide el algoritmo, donde se pudo evidenciar una agrupación que busco la relación posible entre las variables y donde su cantidad también se distribuyó en los clústeres de manera equitativa. También se pudo observar en cada clúster el manejo de las variables que realizo cada uno de estos, donde para las variables de género, año se presentan agrupaciones separadas por valor, eso da a demostrar que el algoritmo si agrupa de acuerdo a los valores de las variables.

Respondiendo a la pregunta ¿Cómo la técnica de minería de datos puede identificar patrones que permitan mejorar los programas de prevención para las enfermedades respiratorias en Bogotá?, se puede concluir que a través de la minería de datos y sus aplicaciones se pueden identificar patrones dentro del análisis de los resultados obtenidos por los clústeres, en el cual se da a conocer por medio de representaciones graficas el comportamiento que tiene cada uno de los datos de las variables con respecto a los demás, en donde posteriormente se

podrán tomar decisiones con base en los resultados analizados generando posiblemente programas de prevención y promoción.

Respondiendo a la pregunta ¿Qué comportamiento han tenido las enfermedades respiratorias en los últimos cinco años de la ciudad de Bogotá? Se identificaron cuatro tipos de comportamientos, asociados a la generación de clústeres, los cuales están más especificados en la sección.

FUENTES:

(DANE), D. A. N. D. E. (2017). DANE. Retrieved September 10, 2017, from http://www.dane.gov.co/

Alvarez, F. M., Troncoso, A., & Riquelme, J. C. (n.d.). Reconocimiento de patrones aplicado a la predicción de series temporales.

Anaya, J. J. (2015). Jhon Jairo Anaya Díaz.

Benítez, I., & Díez, J. L. (2005). Técnicas de Agrupamiento para el Análisis de Datos Cuantitativos y Cualitativos, (September 2017), 1–48.

Berkhin, P. (2002). Survey Of Clustering Data Mining Techniques. *Accrue Software, San Jose, CA*, 1–56.

Campos, G. (2009). Aplicación de técnicas de clustering para la mejora del aprendizaje.

COMPUTERWORD. (2017). Retrieved May 20, 2017, from www.computerworld.com

Cumming, G., Fidler, F., Vaux, D. L., Ioannidis, J. P. A., R Development Core Team, R., Hanahan, D., … Gersbach, C. a. (2011). Graph Kernels. *Bioinformatics (Oxford, England)*. https://doi.org/10.1038/sdata.2014.31

Daniel, P. L. C. y S. G. (2006). *Data Mining Soluciones con Enterprise Miner*. (A. R.- Ma, Ed.).

Díaz Arévalo, J. L., & Pérez García, R. (2002). Estado Del Arte En La Utilización De Tecnicas Avanzadas Para La Busqueda De Información No Trivial a Partir

De Datos En Los Sistemas De Abastecimiento De Agua Potable. *Departamento de Ingeniería Hidráulica Y Medio Ambiente*.

Eduardo, J., & Medina, T. (2014). Facultad De Ciencias Físicas Y Matemáticas. Retrieved from http://repositorio.uchile.cl/bitstream/handle/2250/103717/Separacion-de-renio-por-electrodialisis-a-partir-de-soluciones-acidas-con-presencia.pdf?sequence=3

Elkan, C. (2010). *Predictive analytics and data mining*. https://doi.org/10.4018/978-1-4666-9562-7.ch019

Frsf, C. U. T. N., & Conicet, I. U.-. (n.d.-a). Minería de Datos en Base de Datos de Servicios de Salud. Retrieved from http://conaiisi.unsl.edu.ar/2013/132-505-1-DR.pdf

Frsf, C. U. T. N., & Conicet, I. U.-. (n.d.-b). Minería de Datos en Base de Datos de Servicios de Salud.

Galvis, B., & Rojas, N. Y. (n.d.). ciudad de Bogotá, *3*, 336–353.

Gutiérrez, J. (2016). Líneas de investigación en minería de datos en aplicaciones en ciencia e ingeniería: Estado del arte y perspectivas. *Pdfs.Semanticscholar.Org*, (1), 1–17.

Hernández-Flórez, L. J., Aristizabal-Duque, G., Quiroz, L., Medina, K., Rodríguez-Moreno, N., Sarmiento, R., & Osorio-García, S. D. (2013). Contaminación del aire y enfermedad respiratoria en menores de cinco años de Bogotá, 2007. *Revista de Salud Pública*, *15*(4), 503–516. Retrieved from http://www.revistas.unal.edu.co/index.php/revsaludpublica/article/view/38719/44829

Individuales, L. R., General, S., Social, S., Frecuentes, P., Individual, R., & Versi, R. (2000). Preguntas frecuentes, 1–17.

Instituto de Hidrologia, M. y estudios ambientales. (2010). *Calidad del Aire*.

López, C. P. (2017). *Minería de datos: técnicas y herramientas*. España.

Magdaleno, D., Miranda, Y., Fuentes, I. E., & García, M. M. (2015). Comparative Study of Clustering Algorithms using OverallSimSUX Similarity Function for XML Documents. https://doi.org/10.4114/ia.v18i55.1098

Manning, C. D., Raghavan, P., & Schütze, H. (2009). Hierarchical clustering. *Introduction to Information Retrieval*, (c), 377–401. https://doi.org/10.1017/CBO9780511809071.017

Marchán E, Salcedo J, Aza T, Figuera L, Martínez de Pisón F, G. P. (2011). Reglas de asociación para determinar factores de riesgo epidemiológico de transmisión de la enfermedad de Chagas. *Revista de Ciencia E Ingeniería*, (January), 55–60.

Molina, J., & García, J. (2008). Técnicas de Minería de Datos basadas en Aprendizaje Automático. *Técnicas de Análisis de Datos*, 96–266.

Molina, L. C. (2002). Data mining: torturando a los datos hasta que confiesen. *Fuoc*, 1–11. Retrieved from http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html

Oms, & Ops. (2014). Modulo 4: Análisis de la calidad del dato. *Herramientas Para El Monitoreo de Coberturas de Intervenciones Integradas En Salud Pública.*

Organization, W. H. (2017). FluNet. Retrieved September 9, 2017, from http://www.who.int/influenza/gisrs_laboratory/flunet/en/f

Pérez, M. (2014). *MINERÍA DE DATOS A TRAVÉS DE EJEMPLOS*.

Rodríguez, A. O. (2013). Guía práctica para Arquitecturas de Datos Empresariales, 1–9.

Rodríguez, D., Cuadrado, J., & Sicilia, M. (2007a). Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software. *Ingeniería Del Software*, *3*(1), 6–22. Retrieved from http://en.scientificcommons.org/44226406

Rodríguez, D., Cuadrado, J., & Sicilia, M. (2007b). Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de

software. *Ingeniería Del Software*, *3*(1), 6–22.

Salud, I. N. de. (n.d.). SIVIGILA. Retrieved from http://www.ins.gov.co/lineas-de-accion/Subdireccion-Vigilancia/sivigila/Paginas/sivigila.aspx

Salud, M. de. (2017). SISPRO. Retrieved October 15, 2017, from http://www.sispro.gov.co/

Secretaría Distrital de Salud. (2015). Diagnóstico sectorial de salud, 62. Retrieved from http://www.saludcapital.gov.co/Empalme del Sector Salud 20122016/DIRECTIVA 09 DE 2015/1 DIAGNOSTICO SECTORIAL DE SALUD.pdf

Standard, O. G., & Group, T. O. (2011). *Open Group Standard The Open Group*.

Uiaf, D. (n.d.). DETECCIÓN Y PREVENCIÓN Y LA FINANCIACIÓN.

Viera, L. P. (n.d.). *Introducción a la Minería de Datos*.

Wang, R. Y., Reddy, M. P., & Kon, H. B. (1995). Sunx Toward quality data : An attribute-based approach, *13*, 349–372.

Wang, T., Chan-yeung, M., Lam, W. K., Wong, P. C., Lam, B., Ip, M. S., … Sc, D. (2003). A Cluster of Cases of Severe Acute Respiratory Syndrome in Hong Kong, 1977–1985.

Wu, X., Kumar, V., Ross, Q. J., Ghosh, J., Yang, Q., Motoda, H., … Steinberg, D. (2008). *Top 10 algorithms in data mining*. *Knowledge and Information Systems* (Vol. 14). https://doi.org/10.1007/s10115-007-0114-2

LISTA DE ANEXOS:

1. Anexo A Asignación de códigos para variable Enfermedad
2. Anexo B Asignación de códigos para variable Género