Implementación de una plataforma para análisis de datos un enfoque de big data y datamining

Roberth Figueroa-Díaz[1] , José A. Gutiérrez de Mesa[2]
[1] Universidad Nacional de Loja, Carrera de Ingeniería en Sistemas
Loja, Ecuador
{roberth.figueroa}@unl.edu.ec

[2] Universidad de Alcalá, Ciencias de la Computación
Alcalá, España
{jantonio.gutierrez}@uah.es

Abstract. Los avances tecnológicos e internet en los últimos años ha facilitado que se genere gran cantidad de datos a velocidades sorprendentes en la que los sistemas computacionales se han visto desbordados y limitados. La acumulación de datos sumado a los avances en aprendizaje automático, procesamiento distribuido y minería de datos han sido esenciales para la extracción de nuevo conocimiento desde grandes volúmenes de datos; es por ello que, el presente documento describe la experiencia de implementar una plataforma experimental a través de la integración de Weka con base de datos para la aplicación de algoritmos de minería con el propósito de extraer conocimiento útil a partir de datos almacenados.

Keywords: Big Data, Minería de datos, DataScience, Clasificación, Clustering, KDD,Weka.

1 Introducción

El descubrimiento de conocimiento en base datos KDD [3] como base fundamental para el análisis en grandes conjuntos de datos, ha permitido la generación de nuevas líneas de investigación. Es así como Big Data y Minería de Datos, han evolucionado permanentemente para cubrir nuevos desafíos en búsqueda de conocimiento a partir de los datos. Esto debido principalmente a que internet duplica su tamaña cada 2 años y para 2020 los datos de forma anual se estiman en 44 Zetabytes, la expansión de internet a nuevas personas y empresas que realizan su trabajo a través de internet o en línea [8].

El presente documento recoge la experiencia realizada en la implementación experimental e integración de Weka con base de datos para la aplicación de técnicas de minería de datos a nivel de clasificación y clustering para datos relacionados con paciente de Diabetes cuyo detalle de la fuente se puede ver en [9]. Considerando datos de plasma, presión arterial, índice de masa corporal, número de embarazos, entre otros campos de información. Se describe su diseño, experimentación y finalmente se presenta los resultados obtenidos en conjunto con las conclusiones.

2 Estado del Arte

Fayyad M. y Piatetsky, en 1996, incluyen la definición de KDD como el "Descubrimiento de conocimiento en base de datos", se puede ver ampliamente en [4], como un proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos [2]. La minería de datos aparece como una etapa dentro de un proceso mayor llamado KDD.

Las metodologías para la Minería de datos como CRIS-DM y Two Crows[7] surgen como una alternativa a la necesidad de su aplicación en campos diversos. Aparecen sectores diversos de tecnologías de la información y comunicación que manipula grandes conjuntos de datos con un nuevo enfoque en almacenamiento de grandes datos con su característica de las 3V clásica[1], dando inicio a un nuevo desafío denominado Big Data [9], la que se caracteriza por:

- *Volumen*: alta capacidad de almacenamiento de datos
- *Velocidad*: respuesta en menor tiempo posible
- *Variedad*: diversos tipos de datos (estructurados, no estructurados, semiestructurados)

Esta tecnología presenta como objetivo el proveer de datos a los diversos interesados de un entidad, en el menor tiempo posible y considerando la complejidad existente en el entorno. Surge la era de Big Data, como una combinación de las tecnologías de los últimos 50 años [1]. Dando mayor relevancia al conocimiento que se puede extraer de los datos.

Es así que, la Ciencia de Datos o Data Science, como una tendencia sucesiva de la Analítica se presenta prometedora, donde el dominar técnicas de aprendizaje e integrar grandes conjuntos de datos es esencial. Al considerar nuevos contextos donde los datos son recolectados desde diversos orígenes, sean estos datos estructurados como los ubicados en las bases de datos. Semiestructurados recolectados desde algunos ficheros con cierto criterio de conocimiento y organización, hasta datos no estructurados como los existentes en la web. Sin embargo, el tipo es diverso, sean datos de texto, numéricos, audio, video, y la integración de nuevos datos a partir de sensores dan inicio a la recolección de gran volumen de información que necesita de técnicas avanzadas y algoritmos adecuados para la extracción de conocimiento. Lo que genera un nuevo contexto donde todo lo conectado a la internet, movilidad y su influencia se relaciona al área del Internet de las Cosas o IoT, como una nueva área de investigación y de perfeccionamiento constante de técnicas sobre grandes datos nacen nuevas líneas de investigación que se presentan como útiles áreas de investigación y desarrollo para lo cual, el fortalecer nuestra capacidad en este contexto es esencial y de permanente atención.

3 Trabajos Relacionados

El área de conocimiento relacionado a Big Data [10] y Datamining, esta en constante evolución y por ello existe congresos científicos SIGKDD, IEE,ACM que avalan sus avances e investigaciones con fines de descubrir nuevas soluciones a los problemas existentes. Se puede revisar[1],[11] y además [13].

Es por ello que, en [2] se propone una Plataforma Cloud para análisis de Big Data, (Universidad de Nanjing y Universidad de Hohai, China), donde se presenta una arquitectura experimental e integración de R para análisis de grandes datos con alta capacidad de procesamiento.

Una propuesta en [1] de Software como servicio para Análisis de grandes datos, evalúa el uso de Hadoop, Spark y Map Reduce como herramientas tecnológicas. Sin olvidar la gran discrepancia y resistencia que se presente al momento de implementar este tipo de soluciones a problemas priorizados.

La visión computacional, liderado por la Universidad de Carnegie Mellon [5] como proyecto de gran atención científica, presenta diversas mejoras a los algoritmos de Minería para soportar la diversidad de contenido que cada vez aumenta de forma inquietante. La nube como una alternativa de solución a la movilidad y acceso, ha generado la presencia de diversos servicios como el denominado: Servicio Analítico basado en la nube [11],[13] que genera varios modelos de operatividad como:

- Software de Analítica de datos como Servicio
- Plataforma de Analítica de datos como Servicio
- Infraestructura de Analítica de datos como Servicio

3 Implementación y Experimentación

La propuesta a seguir en el presente proyecto se basa en cuatro elementos a considerar en su desarrollo. El primero es la metodología seguida en la cual se detalla paso a paso lo desarrollado, luego la arquitectura planteada para el proyecto. Como tercer elemento se considera la integración entre SGBD Mysql y Weka como herramienta de aprendizaje automático aprovechando sus ventajas del gran número de técnicas o algoritmos que posee, para al final realizar la etapa de experimentación y puesta en marcha con datos analizados de pacientes con diabetes.

Los datos considerados en el presente proyecto corresponden a 9 atributos para la determinación de diabetes cuya fuente de datos fue proporcionada como donación por la Universidad Jhons Hopkins, exactamente el Instituto Nacional de diabetes y enfermedades digestivas y renales, el cual esta disponible online a través del sitio web de la Universidad de California [9]. Los datos se describen en la Tabla 1.

El proceso metodológico seguido para la obtención de nuevo conocimiento a través de la aplicación de Minería de datos, es una variación del proceso KDD mencionado en [3], por lo que permitirá llegar a detectar nuevas reglas con cierto grado de facilidad. Aunque la interpretación de los patrones requiere de un grado mayor de conocimiento de áreas relacionados como Redes Neuronales, Estadística entre las principales.

El proceso inicia con los datos obtenidos del paciente diabético, a estos datos iniciales pueden sumarse nueva información proveniente de cualquier origen de datos y tipo de información, lo que da origen a Big data o grandes datos a considerarse dentro del descubrimiento de conocimiento. En este caso se trabajó con los atributos antes mencionados correspondientes a información de pacientes diabéticos.

Tabla 1. Descripción de los atributos utilizados como entrada de datos para determinar si un paciente da positivo o negativo a la diabetes.

Descripción del Atributo	Tipo de Dato	Etiqueta usada
Número de embarazos	Numérico	embarazos
Glucosa (plasma)	Numérico	plasma
Presión arterial	Numérico	presion
Espesor del trícep	Numérico	esptricep
Insulina (durante 2 horas)	Numérico	insulina
Índice de masa corporal	Numérico	idc
Función pedigrí Diabetes	Numérico	fpedigri
Edad	Numérico	edad
Clase (positivo/negativo)	Nominal	clase

Luego los datos almacenados en un repositorio o base de datos unificado, SGBD Mysql en este caso, como se lo propone justamente en la arquitectura de la Fig. 1, como una visión conjunta. Después los datos según la necesidad pueden ser transformados o preprocesados según el tipo deseado y el formato establecido por la herramienta de minería de datos que se utilice. Luego pasa a la fase de descubrir conocimiento, donde se hace el preprocesado, posterior se utiliza Weka como herramienta de minería de datos por sus ventajas ya antes descritas, obteniendo nuevos patrones o conocimiento que puede utilizarse en beneficio de la organización o entidad, en nuestro caso, esos patrones permitirán ser considerados para adaptarse por ejemplo a actuales o nuevos sistemas informáticos para mejorar y dar soporte a la toma de decisiones.

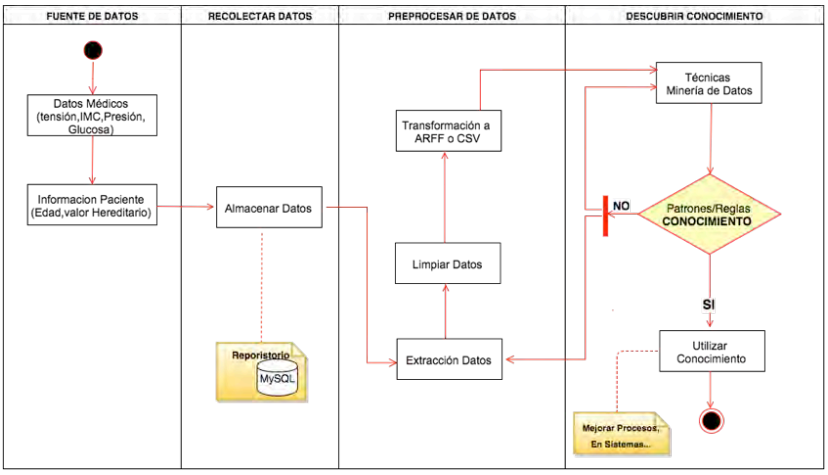

Fig. 1. Metodología aplicada para la extracción de conocimiento a partir de los datos de pacientes diabéticos.

El proceso general a seguir para la obtención de patrones o nuevo conocimiento es descrito en la Fig. 1. y consta de 4 fases principales que detallan su proceso en general.

Respecto a la Arquitectura tecnológica Fig. 2. propuesta [6] a nivel de la solución se plantea la integración de todas la fuentes de datos u orígenes de datos y centralizarlos en la base de datos Mysql como en este caso, para a partir de este origen una vez que ha sido integrado con Weka pueda ir extrayendo los datos necesarios y específicos para aplicar técnicas de Minería de datos según el contexto específico a experimentar.

Sin lugar a dudas este repositorio tiene la finalidad de dar una idea al lector sobre como es el ambiente de Big Data y su alcance con mayor cantidad de información y su funcionamiento global. La aplicación de los diversos algoritmos de clasificación o agrupamiento u otras tareas propias de la Minería de datos facilita su manejo una vez integrado y permite obtener un modelo global que representa el conocimiento que puede integrarse a sistemas dedicados o específicos dentro de la organización. Esto simboliza la obtención de reglas o patrones de conocimiento que incluso pueden integrarse a los procesos para mejorar la organización que impulse el uso de estas técnicas que generan un valor agregado.

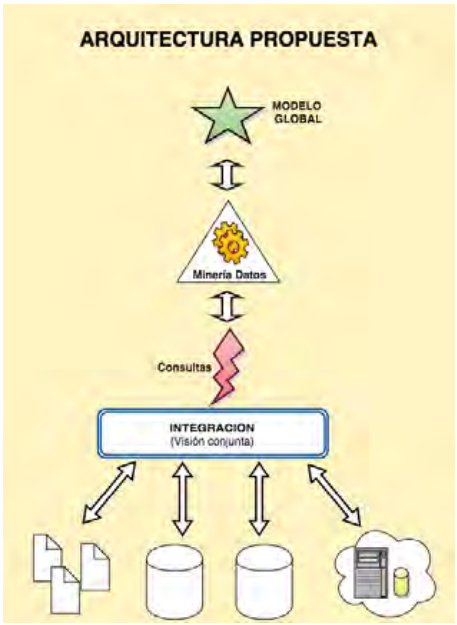

Fig. 2. Arquitectura propuesta para la integración y uso de información relacionada para la extracción de conocimiento a partir de técnicas de Minería de Datos.

En este proyecto se experimento con dos tareas puntuales de extracción de conocimiento, la primera, que es la clasificación, cuyo objetivo es la predicción de nuevo conocimiento a partir de los datos y la segunda tarea que corresponde al agrupamiento o clustering que corresponde a la tarea de describir los datos, es decir encontrar relaciones no obvias dentro de los datos. Para lo cual, se utilizó algunas técnicas especificas como árboles de decisión, redes neuronales, estadísticas, vecino más cercano, cuyas implementaciones en Weka varían y se organizan de tal forma que facilita su operatividad.

Algunas de las reglas que se pueden obtener e interpretar para ser integradas a cualquier sistema informático médico o similar, son las siguientes:

- Si el ***Plasma***(glucosa) es menor a 127 y el ***Índice de masa corporal*** es menor a 26.4 entonces es Negativo a la prueba de Diabetes.
- Si el ***Plasma***(glucosa) es menor a 127 y el ***Índice de masa corporal*** es menor a 26.4 entonces es Negativo a la prueba de Diabetes.
- Si el ***Plasma*** es menor a 127 y el ***Índice de masa corporal*** es mayor a 26.4 y la ***Edad*** es menor a 28, entonces Negativo a la prueba de Diabetes.

De lo anterior se puede mencionar que el criterio importante para determinar es el *Plasma* o glucosa seguido del *Índice de Masa Corporal* como segundo atributo y luego la *Edad* del paciente, son criterios importantes para determinar si una persona da positivo o no a la diabetes.

En la Tabla 2. se describe los algoritmos utilizados tanto en clasificación como en agrupamiento o clustering.

Tabla 2. Algoritmos considerados para la extracción de conocimiento para las tareas de Clasificación y Agrupamiento o clustering.

Algoritmos de Clasificación	**Algoritmos de Agrupamiento**
C4.5 en Weka J48	Cobweb
OneR	Maximización-Expectación o EM
Multilayer Perceptron o MLP	K-medias
Vecino más cercado o NNge	

4 Resultados

Luego de la aplicación de las diversas técnicas o algoritmos de minería de datos sobre la plataforma integrada a nivel de tareas de clasificación y agrupamiento permitió determinar cual es el mejor método que se ajusta con los datos analizados en función de encontrar nuevo conocimiento útil sobre los datos de pacientes diabéticos.

Como resultado se obtuvo que el mejor algoritmo para la tarea de clasificación es el algoritmo MLP como se describe en la Tabla 3. con un 75,26% de confianza y para la tarea de agrupamiento se tiene que el algoritmo que presentó mejores resultados fue el K-medias con 2 clústeres o grupos.

Tabla 3. Resultados obtenidos luego de la ejecución de los algoritmos de Clasificación y Agrupamiento considerados para pacientes con diabetes.

Algoritmos Clasificación	**Resultado Clasificación**	**Algoritmos Agrupamiento**	**Resultado Agrupamiento**
J48	73,44%	Cobweb	2 clúster + ruido
OneR	70,83%	EM	3 clúster
MLP	75,26%	K-medias	2 clúster
NNge	73.95%		

A continuación se indica en la Fig. 3., la ejecución de las técnicas en Weka con la finalidad de obtener el mejor resultado, tanto a nivel de Clasificación como de Agrupamiento utilizando el Multilayer Perceptron basado en redes neuronales.

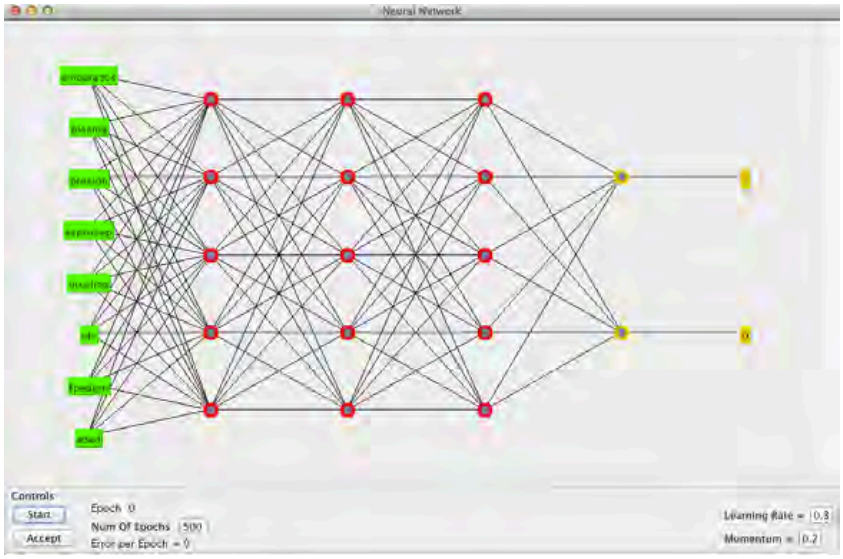

Fig. 3. Resultado obtenido con la ejecución del algoritmo Multilayer Perceptron o MLP, con una arquitectura de una capa de entrada, 3 capas ocultas y una capa de salida.

5 Conclusiones y trabajos futuros

La Ciencia de datos (Data Science) es una evolución natural de la Minería de Datos cuyo propósito es extraer conocimiento de grandes volúmenes de datos o Big Data.

La plataforma de análisis de datos experimental permitió determinar que el mejor algoritmo de clasificación para analizar los datos descritos fue MLP, mientras que para la tarea de clustering fue K-medias; permitiendo demostrar que la plataforma pueda utilizarse en cualquier contexto como por ejemplo información médica, transporte, campo empresarial o cualquier otra problemática donde se puede contar con datos.

Para aprovechar de forma dinámica el conocimiento adquirido con la plataforma de análisis de datos experimental se debería considerar como trabajos futuros la integración de ésta a diversos sistemas informáticos desde los cuales se pueda aprovechar su capacidad de aprendizaje para obtener patrones, reglas o conocimiento de forma automática. Así como también mejorar la integración a base de datos no estructuradas o no SQL y herramientas como hadoop, spark o map reduce para incrementar capacidades de distribución, volumen, alta disponibilidad y mejorar velocidades de procesamiento.

Referencias

1. Zheng, Z., Zhu, J., R. Lyu, M.: Service-generated Big Data and Big Data-as-a-Service: An Overview. 2013 IEEE International Congress on Big Data (2013)
2. Ye, Wang , Z., Zhou , F., Wang , Y., Zhou, : Cloud-based Big Data Mining & Analyzing Services Platform integrating R. 2013 International Conference on Advanced Cloud and Big Data (2014)
3. Fayyad, U. M., Piatetsky-Shapiro, G., Padhraic, S., Ramasamy, U.: Advances in Knowledge Discovery and Datamining. Menlo Park: AAAI Press (1996)
4. Fayyad M., U.: Data Mining and Knowledge Discovery: Making sense out of Data. IEEE Expert, 20-25 (October 1996)
5. University, C.: Carnegie Mellon University-CMU. In: Research-Machine Learning Department. Available at: "http://www.ml.cmu.edu/research/index.html"
6. Hernández Orallo, J., Ramírez Quintana, M. J., Ferri Ramírez, C.: Introducción a la Minería de Datos. Pearson Education S.A., Madrid, España (2004)
7. Edelstein, H. A.: Introduction to Data Mining and Knowledge Discovery Third Edition edn. Two Crows Corporation (1999)
8. Gantz, J., Reinsel, D.: The Digital Universe in 2020:Big Data, Bigger Digital Shadows, and Biggest growth in the Fast East. IDC Analyze the Future , 1-16 (2012)
9. Universidad de California: Machine Learning Repository. In: Machine Learning Repository. (Accessed 1987) Available at: "http://archive.ics.uci.edu/ml/index.html"
10. Zhang , L., Stoffel, Behrisch, Mittelstadt, Schreck, Pompl, R., Weber, Last, Keim, D.: Visual Analytics for the Big Data Era – A Comparative Review of State-of-the-Art Commercial Systems. IEEE Symposium on Visual Analytics Science and Technology 2012 (October 2012)
11. Khan, Anjum, Kiani, S. L.: Cloud based Big Data Analytics for Smart Future Cities. In : 6th International Conference on Utility and Cloud Computing (2013)
12. Inje, B., Patil , U. : Operational Pattern Revealing Technique in Text Mining. In : IEEE Students' Conference on Electrical, Electronics and Computer Science (2014)
13. Zhang, Li, Zhang, Y., Xing, : DataCloud: An Efficient Massive Data Mining and Analysis Framework on Large Clusters. Ninth Web Information Systems and Applications Conference (2012)