IMPLEMENTACIÓN DE TÉCNICAS ESPECÍFICAS DE MINERÍA DE DATOS EN APLICACIONES WEB CON MOTORES DE BASE DE DATOS RELACIONALES

VARGAS, Luis Alejandro; FARFAN, José Humberto; RODRIGUEZ, Mariela Ester; TAPIA, Marcela Alejandra; PAREDES, Julio Cesar RAMOS, Pablo Nicolás; LLAMPA, Alvaro Facundo; MONTES, Leonardo Ezequiel; MOGRO, Nelson Ariel & CORDOBA, Irma Rafaela Mercedes;
Ingeniería + Software (I+S), Área de Ingeniería Informática, Facultad de Ingeniería, Universidad Nacional de Jujuy (U.N.Ju).

ARAMAYO, Fernando Rubén & SPADONI, Gustavo Fernando
Facultad de Ciencias Agrarias, Universidad Nacional de Jujuy (U.N.Ju.)

RESUMEN

El presente proyecto pretende implementar técnicas de Datamining o también denominado Minería de Datos en aplicaciones web de sistemas informáticos y procesan la información con motores de Base de Datos Relacionales, es decir efectuar "Análisis Inteligente de Datos". En el desarrollo de tecnologías web genera la necesidad de contar con herramientas, carencia que se ha analizado y detectado en investigaciones anteriores, por lo que se pretende implementarlo en lenguajes específicos de programación web.

Implica obtener ventajass y beneficios de la Minería de Datos ya sea aplicando técnicas específicas de descripción o de aprendizaje automático. Se puede mencionar Redes Neuronales, reglas de asociación, clustering o similares, en cualquier problema genérico de un Sistema Informático desarrollado en un lenguaje de programación web. El objetivo es realizar en los datos un "Análisis Inteligente de Datos".

Para lograr este objetivo es necesario clasificar y estudiar en profundidad las técnicas que representan a la Minería de Datos, aplicándolas en lenguajes de programación web e implementarlos con motores de Base de Datos. Es obligatorio y necesario conocer en detalle los lenguajes involucrados, estudiar la implementación correcta de las técnicas de Minería de Datos y la conectividad con el motor de Base de Datos.

Palabras Clave: Datamining, Análisis Inteligente, Programación Web

CONTEXTO

El proyecto se encuentra inserto dentro de las siguientes Líneas Prioritarias de Investigación de la Facultad de Ingeniería (LIPIFI) - UNJu:

- Ingeniería del Software
- Ingeniería de Procesos

Es un proyecto aprobado de categoría B (Código D/B030)
Financiamiento: Secretaría de Ciencia y Técnica y Estudios Regionales (SeCTER) de la UNJu.
Vigencia del Proyecto: 01/01/2018 al 31/12/2019

1. INTRODUCCIÓN

Se entiende que la Minería de Datos es "un campo de la estadística y las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos" [1], el actual Grupo de Investigación se inició en el año 2.016 con el proyecto, aprobado, de categoría B denominado "Data Mining aplicado a análisis telefónico", el mismo ha finalizado en diciembre de 2017, siendo importante destacar que se aplicaron técnicas de Minería de Datos exclusivas para encontrar soluciones al problema en cuestión. Este último es un problema genérico en el que se realizó un Análisis de Datos de una Red de Comunicaciones con el formato actual que proveen las compañías telefónicas denominado "Sábanas de llamadas" para ser aplicados en el ámbito de la Seguridad de Organismos Gubernamentales Provinciales y/o Nacionales, siendo necesario por ello examinar y describir técnicas y herramientas que emergen en esa área de investigación aplicadas a la toma de decisiones [2]. El descubrimiento de la información oculta es de importancia estratégica, y es posible por las características de la Minería de Datos, pero es el descubrimiento del conocimiento (KDD, por sus siglas en inglés) el que se encarga de la preparación de los datos y de la interpretación de los resultados obtenidos, los cuales darán significado a los patrones encontrados [3]. Se destaca que KDD es producto del rápido desarrollo de la minería de datos y la aplicación de tecnologías de información y bases de datos. Zhang et al [4] formulan el proceso basado de la extracción de conocimiento (KDD) en una secuencia iterativa de cuatro pasos: la definición del problema, el pre - procesamiento de datos (que incluye la preparación de datos), data mining, y el post data mining.

Como se mencionó anteriormente el proyecto de investigación "Data Mining aplicado a análisis telefónico" se encuentra en su última fase, centrado en el análisis y aplicación de técnicas específicas de minería de datos al problema de estudio, abarcando y estudiando los resultados que son de utilidad a los usuarios finales que hacen uso de la misma, principalmente en el área de seguridad gubernamental.

Sin embargo existe una cantidad de técnicas que son estudiadas por la Minería de Datos y que es necesario profundizar para fortalecer un grupo de investigación destinado a esta ciencia. Una clasificación sumamente interesante y bastante completa es la planteada por Orallo et al [5] en donde se observa en forma más detallada las técnicas particulares que pueden aplicarse según se trate de problemas Predictivos o Descriptivos (Tabla 1).

Nombre	PREDICTIVO		DESCRIPTIVO		
	Clasificación	Regresión	Agrupamiento	Reglas de asociación	Correlaciones / Factorizaciones
Redes neuronales	✓	✓	✓		
Arboles de decisión ID3, C4.5, C5.0	✓				
Arboles de decisión CART	✓	✓			
Otros árboles de decisión	✓	✓	✓	✓	
Redes de Kohonen			✓		
Regresión lineal y logarítmica		✓			✓
Regresión logística	✓			✓	
Kmeans			✓		
Apriori				✓	
Naive Bayes	✓				
Vecinos más próximos	✓	✓	✓		
Análisis factorial y de comp. ppales.					✓
Twostep, Cobweb			✓		
Algoritmos genéticos y evolutivos	✓	✓	✓	✓	✓
Máquinas de vectores soporte	✓	✓	✓		
CN2 rules (cobertura)	✓			✓	
Análisis discriminante multivariante	✓				

Tabla 1: Aplicación de técnicas de Data Mining según el tipo de Problema

El trabajo realizado precedentemente por el grupo de investigación se enfocó en la

aplicación de la minería de datos en el ámbito de la seguridad. Sin embargo, se pudo determinar que el conjunto de herramientas que la caracterizan es amplio y surgieron problemas en la integración con los sistemas de donde se extraían la información, la mayoría de ellos sistemas informáticos de plataforma web. Es objetivo del equipo de trabajo investigar el procedimiento de incluir las técnicas de minería de datos en los sistemas que hoy cuentan con la información necesaria para tomar decisiones.

La Inteligencia de negocios o BI, por sus siglas en inglés, según el Data Warehouse Institute, lo define como la combinación de tecnología, herramientas y procesos que permiten transformar los datos almacenados en información, esta información es conocimiento y este conocimiento debe estar dirigido a un plan o tener una estrategia comercial. La inteligencia de negocios debe ser parte de la estrategia empresarial, que permite optimizar la utilización de recursos, monitorear el cumplimiento de los objetivos de la empresa y la capacidad de tomar buenas decisiones para obtener mejores resultados, tal como puede observar en la imagen1.[8]

Imagen 1: Pasos de la Inteligencia de Negocios

Los pasos que detalla el The Data Warehousing Institute se refiere al proceso de convertir datos en conocimiento, posteriormente reflejar en acciones competitivas. Los datos deben ser procesados para requerir la información de ella, tarea que las empresas se han visto obligadas a requerir y les permite ser competitivos en el mercado actual. En estas últimas décadas los usuarios consumen productos y servicios por la web y esta modalidad de mercado exige que las empresas deban extraer conocimientos de la información con la que cuentan y puedan predecir acciones. Es sumamente importante la integración de los sistemas con la extracción del conocimiento que se realiza. Se puede afirmar en consecuencia que la etapa de requerir y hacer uso de la información tiene un desafío mayor y consistente en la extracción del conocimiento o análisis inteligente.

Una falencia o una situación a considerar que se presentó en la investigación realizada previamente por este grupo, es la falta de implementación o interrelación de las técnicas empleadas en estas herramientas de Minería de Datos en lenguajes de programación. Existen herramientas en el mercado actual para Minería de Datos, tal como se puede observar en el llamado “Cuadrante Mágico” para plataformas de Análisis Predictivo del año 2.016 (Imagen 2), de la empresa Gartner, organización de investigación de las tecnologías de la información reconocida mundialmente, en donde se muestra la comparación de las principales herramientas para Plataformas Analítica Avanzadas [6], con características y funcionalidades totalmente

diferentes entre ellas.

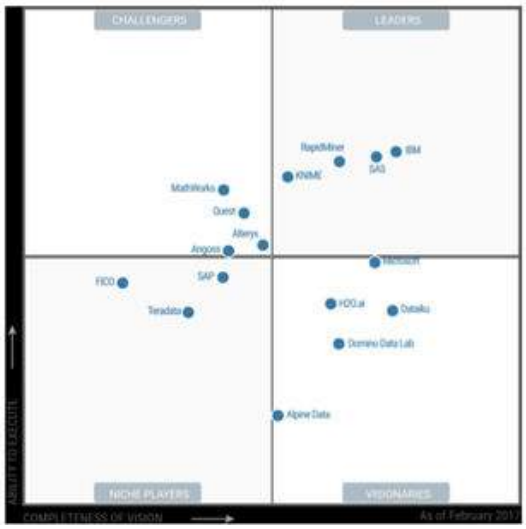

Imagen 2: "Cuadrante Mágico" para plataformas de Análisis Predictivo 2.016 de Gartner

Hay también, lenguajes de programación que son utiizados mundialmente, donde se suele tomar como medida el Índice "TIOBE Quality Indicator", el cual es un índice elaborado por una empresa de software holandesa que se especializa en la evaluación y seguimiento de la calidad de los programas informáticos[7], tal como se observa en la imagen 3, revisando actualmente en tiempo real más de 300 millones de códigos de diversos programas.

Aug 2017	Aug 2016	Change	Programming Language	Ratings	Change
1	1		Java	12.961%	-6.60%
2	2		C	6.477%	-4.80%
3	3		C++	5.550%	-0.25%
4	4		C#	4.795%	-0.71%
5	5		Python	3.692%	-0.71%
6	8	^	Visual Basic .NET	2.969%	+0.08%
7	6	v	PHP	2.255%	-0.60%
8	7	v	JavaScript	2.056%	-0.61%
9	9		Perl	1.955%	-0.52%
10	12	^	Ruby	1.965%	-0.51%

Imagen 3: Comparación de los principales lenguajes, índice TIOBE, agosto de 2.017

De acuerdo a dicho cuadro comparativo se debe hacer hincapié en los lenguajes que son frecuentemente utilizados para desarrollo de sistemas, considerando que en los primeros lugares se puede visualizar que los lenguajes son de caracteristica web y la integración que se debe desarrollar con la herramientas de la imagen 2, deben estar en esta línea de trabajo. La empresa o entidad que necesita tomar decisiones debe por lo tanto integrar la información que posee y el modelo de extracción del conocimiento, siendo para ello necesario relacionar el modelo de conocimiento con el sistema informático, para que personas de los distintos ámbitos de la empresa puedan realizar el análisis con los datos de la organización.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Dentro de las Líneas Prioritarias de Investigación de la Facultad de Ingeniería (LIPIFI) - UNJu los temas que se investigarán y desarrollarán serian:

- Técnicas Descriptivas de Minería de Datos
- Técnicas Predictivas de Minería de Datos
- Herramientas de Software Libre para Datamining
- Análisis Inteligente de Datos

3. RESULTADOS OBTENIDOS / ESPERADOS

El presente proyecto se encuentra en su fase inicial, y busca cumplir el objetivo general: Implementar técnicas específicas de herramientas de Minería de datos en aplicaciones web de sistemas informáticos con Base de Datos Relacionales; y presenta los siguientes Objetivos específicos:

- Estudiar y profundizar en las principales técnicas de Minería de Datos,
- Analizar y establecer las principales características de las herramientas de Minería de Datos disponibles en la actualidad que permitan su implementación en lenguajes de programación web.
- Estudiar y comparar los principales lenguajes de programación web disponibles en el mercado actual.
- Estudiar e implementar la conectividad de los lenguaje/s de programación web seleccionados con un motor de base de datos relacional.
- Implementar las técnicas de las herramientas selecionadas en los lenguajes de programación web estudiados.
- Realizar prácticas de Análisis Inteligente en problemas específicos de Minería de Datos.

4. FORMACIÓN DE RECURSOS HUMANOS

Apellido/s y Nombre/s	Formación y Unidad Académica	Rol Dentro del Proyecto
Vargas, Luis Alejandro	Ingeniero en Informática - Fac.Ing.UNJu	Director
Farfán, José Humberto	Esp.Doc.Sup - Ingeniero en Informática - Fac.Ing.UNJu	Co Director
Rodriguez, Mariela Ester	Lic.Sistemas - Ingeniero en Informática - Fac.Ing.UNJu	Investigador
Aramayo, Fernando Ruben	Lic-Sistemas-Ingeniero en Informática - Fac.Cs.Agr. UNJu	Investigador
Ramos, Pablo Nicolás	Estudiante Ingeniería en Informática - Fac.Ing.UNJu	Investigador
Llampa, Alvaro Facundo	Estudiante Ingeniería en Informática - Fac.Ing.UNJu	Investigador
Montes, Leonardo Ezequiel	Estudiante Ingeniería Informática - Fac.Ing.UNJu	Investigador
Mogro, Nelson Ariel	Estudiante Ingeniería en Informática, Fac.Ing.UNJu	Investigador
Córdoba, Irma Rafaela Mercedes	Ingeniera en Informática - Fac.Ing.UNJu	Investigador
Tapia, Marcela Alejandra	Ingeniera en Sistemas de Información Fac.Ing.UNJu	Investigador
Paredes, Julio Cesar	Programador Universitario Fac.Ing.UNJu	Investigador
Spadoni, Gustavo Fernando	Médico Veterinario Fac.Cs.Agr. UNJu	Investigador

Los alumnos Leonardo Ezequiel Montes y Nelson Ariel Mogro se encuentran desarrollando la tesis de grado para la obtención del título de Ingeniero en Informática denominado "Minería de datos para soporte a decisiones de planificación educativa" cuyo Director es el Esp. Ing. José Farfán, su Codirector Ing. Mariela Rodriguez.

5. BIBLIOGRAFIA

[1] Maimon, O., & Rokac, L., 2010. ¨Data Mining and Knowledge Discovery Handbook", "O. Maimon, & L. Rokac, Data Mining and Knowledge Discovery Handbook", Nueva York, Springer, 2010, págs. 1-18.

[2] Yelitza, J., Marcano, A., & Rosalba Talavera, P. , "Minería de Datos como soporte a la toma de decisiones empresariales" , Obtenido de http://www.scielo.org.ve/scielo.php?script=sci_arttext&pid=S1012-15872007000100008 en Agosto de 2.017, Universidad de Zulia, Maracaibo, 2007

[3] Vallejos, S. J. . ¨Minería de Datos¨, Obtenido de http://exa.unne.edu.ar/informatica/SO/Mineria_Datos_Vallejos.pdf, Universidad Nacional del Nordeste, 2006

[4] Zhang, S., Zhang, C., & Yang, Q. , "Data preparation for data mining. Applied Artificial Intelligence", San Francisco, 2003, págs. 375-381.

[5] Hernández Orallo J., Ramírez Quintana J., Ramirez C.F., "Introducción a la Minería de Datos", Universidad Politécnica de Valencia, España, Ed.Pearson, 2004, págs 19-42, 137-164, 237-252

[6] Gartner, "Magic Quadrant for Advanced Analytics Platforms". Obtenido de http://www.kdnuggets.com/2017/02/gartner-2017-mq-data-science-platforms-gainers-losers.html en Agosto de 2.017.

[7] TIOBE Quality Indicator. The Importance of Being Earnest Quality Indicator. Obtenido de https://www.tiobe.com/tiobe-index/ en Septiembre de 2.017.

[8] Oracle, Inteligencia de Negocios. Obtenido de http://www.oracle.com/ocom/groups/public/@otn/documents/webcontent/317529_esa.pdf en setiembre de 2017