PROPUESTA DIDÁCTICA INHERENTE AL ÁREA DE CIENCIA DE DATOS

Mag. Raúl Oscar Klenzi, Mag. María Alejandra Malberti, Mag. Graciela Elida Beguerí

Instituto de Informática / Departamento de Informática /Facultad de Ciencias Exactas Físicas y Naturales / Universidad Nacional de San Juan

Av. Ignacio de la Roza 590 (O), Complejo Universitario "Islas Malvinas", Rivadavia, San Juan, Teléfonos: 4260353, 4260355 Fax 0264-4234980, Sitio Web: http://www.exactas.unsj.edu.ar
e-mail: {rauloscarklenzi,amalberti,grabeda}@gmail.com

RESUMEN

En el presente trabajo se expone una propuesta de cómo introducir contenidos inherentes al área de Ciencia de Datos en el marco de las carreras Licenciatura en Ciencias de la Computación y Licenciatura en Sistemas de Información pertenecientes al Departamento de Informática –DI– de la Facultad de Ciencias Exactas Físicas y Naturales de la Universidad Nacional de San Juan –FCEFN–UNSJ–. Se plantea como una experiencia didáctica y surge de la revisión curricular de las carreras de grado. KNIME Analytics es el software de aprendizaje de máquina que utiliza tanto el grupo de investigación como la asignatura Inteligencia Artificial –IA–, marco en la que se realiza la experiencia. La asignatura IA pertenece a ambas carreras. En este caso se expone cómo evaluar distintos tipos de datos, relevados u obtenidos por los alumnos, para la finalidad planteada.

Palabras clave: Ciencia de Datos, Ingeniería de Datos, KNIME Analytics, Software Libre, Visualización, Educación Superior.

CONTEXTO

La temática está inserta en el proyecto "Visualización y Deep Learning en Ciencia de Datos" el cual se enmarca en el Laboratorio de Sistemas Inteligentes para Extracción de Conocimiento en Datos Masivos del Instituto de Informática de la Facultad de Exactas Físicas y Naturales de la Universidad Nacional de San Juan –FCEFN–UNSJ–.

Las líneas de investigación se presentan como una continuidad de los proyectos "Minería de Datos en la Determinación de Patrones de Uso y Perfiles de Usuario" y "Búsqueda de Conocimientos en Datos Masivos" aprobados y subsidiados por CICITCA.

1. INTRODUCCIÓN

La ciencia de datos y la ingeniería de datos son dos ramas diferentes dentro del paradigma de big data, un enfoque en el que se capturan, procesan a velocidades enormes, variedad y volúmenes de datos estructurados, no estructurados y semiestructurados, utilizando un conjunto de técnicas y tecnologías completamente novedosas en comparación con las que se utilizaron en décadas anteriores. Ambas son útiles para derivar conocimiento a partir de datos en bruto. Son elementos esenciales para cualquier sistema integral de apoyo a la toma de decisiones y extremadamente importantes a la hora de formular estrategias sólidas para la gestión empresarial.

Ciencia de Datos. La ciencia de datos se puede pensar como el dominio científico dedicado al descubrimiento de conocimiento mediante análisis de datos. El término específico de dominio, se refiere al sector de

la industria o dominio temático que los métodos de ciencia de datos utilizan para explorar. Los científicos de datos aplican técnicas matemáticas y enfoques algorítmicos para derivar soluciones a problemas empresariales y científicos complejos. Tanto en los negocios como en la ciencia, los métodos de ciencia de datos pueden proporcionar capacidades de toma de decisiones más robustas.

Ingeniería de datos. Es el dominio de ingeniería dedicado a superar los cuellos de botella de procesamiento de datos y problemas de manejo de datos para aplicaciones que utilizan Big Data. Los ingenieros de datos emplean la informática e ingeniería de software para modelar sistemas y resolver problemas con el manejo y la manipulación de grandes conjuntos de datos. Para lo cual se nutre de entornos de procesamiento en tiempo real y plataformas de procesamiento paralelo masivo (MPP), así como de sistemas de administración de bases de datos relacionales. En términos simples, con respecto a la ciencia de datos, el propósito de la ingeniería de datos es diseñar soluciones de big data mediante el desarrollo de plataformas de procesamiento de datos coherentes, modulares y escalables a partir de las cuales los científicos de datos puedan obtener posteriormente información valiosa (Pierson, 2015).

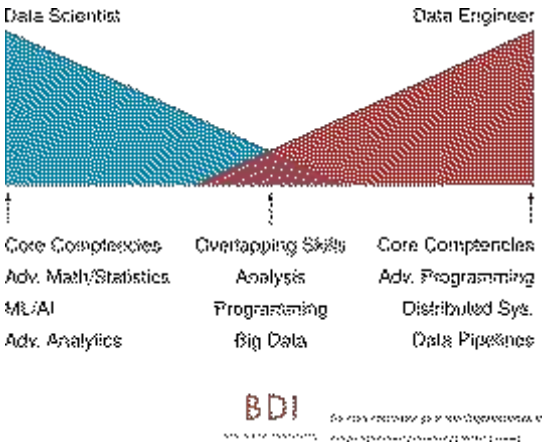

Figura 1. Competencias básicas de los científicos de datos e ingenieros de datos y sus habilidades superpuestas. https://www.oreilly.com

La relación entre un ingeniero de datos y un científico de datos, se observa en la figura 1.

Jesse Anderson, si bien explica también la diferencia entre un científico de datos con un ingeniero de datos, menciona que el profesional que está capacitado para ser competente tanto en ingeniería de datos como en ciencia de datos es el ingeniero en aprendizaje automático, lo que muestra la figura 2.

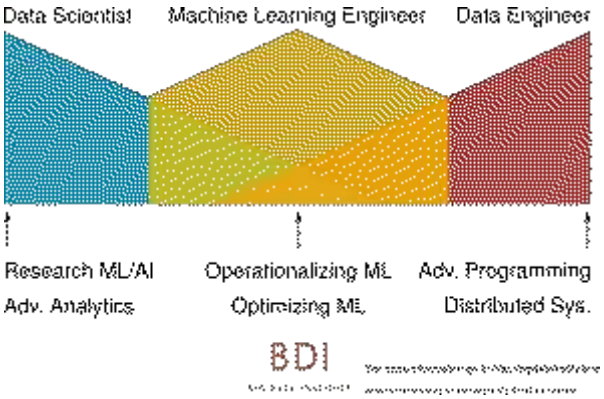

Figura 2. Diagrama que muestra dónde encaja un ingeniero de aprendizaje automático con un científico de datos e ingeniero de datos. https://www.oreilly.com

Si bien existen herramientas de software que pueden involucrar tanto las tareas de un ingeniero como de un científico de datos, el grupo de investigación viene utilizando KNIME Analytics, desde hace varios años, por ser una herramienta de software libre que, en forma modular permite incorporar todos los pasos involucrados en el proceso de ciencia de datos. Además esta plataforma basada en Eclipse, permite la integración tanto con proyectos de Eclipse como BIRT, DTP, etc., como con diferentes herramientas de recopilación, gestión y análisis de información. Proporciona un repositorio de módulos fáciles de usar y adicionalmente a las técnicas estándares de minería de datos, añade los algoritmos más actuales de análisis tales como los de deep learning (Salgueiro, 2016).

Con la intención de que los alumnos experimenten las actividades tanto de un ingeniero de datos como de un científico de

datos, es que en la asignatura IA se planificó y llevó adelante una actividad consistente en estudiar algunos casos provistos en el software y luego realizar una implementación con datos generados u obtenidos por ellos.

En la experiencia práctica se consideraron entre otros, datos georreferenciados, imágenes, texto extraído de redes sociales, interacción de KNIME con otros lenguajes de programación etc. Particularmente se mencionan los referentes a los datos georreferenciados e imágenes: german-bike-routes.gpx y 04_Car_Counting.knwf los que luego de ser estudiados fueron emulados con datos propios.

Es de destacar que para el primero de los ejemplos los alumnos procesaron datos georreferenciados inherentes a diferentes zonas de la provincia de San Juan, rescatados desde sitios de internet o registrados desde caminatas por medio de app's de dispositivos móviles, y luego graficaron los recorridos desde módulos específicos de KNIME.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Entre las líneas más importantes de investigación se mencionan:

- Ciencia de Datos, principalmente con Visualización de Información
- Deep Learning
- Herramientas de software libre para arquitecturas secuenciales, paralelas y distribuidas.

El proyecto tiene contemplado contribuir con el análisis y revisión curricular de las carreras de grado del DI con el fin de vincular las áreas disciplinares e incorporar, integrar y explotar simultáneamente distintas herramientas libres, provenientes de los aportes de las Tecnologías de la Información y la Comunicación, teniendo en cuenta aspectos de Interacción Humano Computador enriquecidos por teorías de la percepción humana.

3. RESULTADOS OBTENIDOS Y ESPERADOS

El proyecto se encuentra en su fase de desarrollo y al momento se está trabajando:

- Con datos provistos por una empresa de mantenimiento del alumbrado público de la ciudad de San Juan y se incursiona en aspectos relativos a geolocalización tales como análisis de coordenadas geográficas, su representación, manipulación e interpretación por parte de herramientas informáticas, así como su incorporación en un proceso de minería de datos (MD).
- En el área de Netmining (área de MD encargada de extraer conocimiento contextualizado a grafos mediante el cual se representa el contenido de redes sociales), se indagan datos provenientes de la Biblioteca de la FCEFN.
- Procesamiento de texto e imágenes mediante la utilización de algoritmos de deep learning.
- Determinación de diferentes métricas, desde series temporales, asociadas a entidades financieras y comerciales sobre datos extraídos de internet.

Como tareas de DS se reconocieron y georreferenciaron diferentes puntos de la ciudad de San Juan, producto de reclamos de usuarios de empresa de mantenimiento de alumbrado público. Con éstos, se emuló el problema del viajante de comercio y mediante Tabú Search fue resuelta la problemática, respetando los sentidos de circulación de cada trayecto en particular.

Como en cada iteración de la búsqueda tabú se debe acceder a los servicios de google a efectos de encontrar el camino mínimo y atento a resolver problemas de accesibilidad, se decidió utilizar los datos del proyecto Open Street Map –OSM-. Con los datos

correspondientes a la ciudad de San Juan - República Argentina, se implementó un servidor propio de mapas utilizando un contenedor (docker) y así tener un acceso permanente y rápido dentro de la intranet de la facultad. Esta alternativa contempla al DI.

Mientras se llevaron adelante las tareas planificadas en la cátedra, se observó un mayor interés por parte de los alumnos de participar en los diferentes proyectos de investigación enmarcados en el área temática de Ciencia de Datos y Big Data.

Así mismo, el equipo de la cátedra Inteligencia Artificial, mantiene hasta el presente esta modalidad como alternativa de enseñanza-aprendizaje referida a la extracción de conocimiento en datos

4. FORMACIÓN DE RECURSOS HUMANOS

Los integrantes del proyecto son docentes-investigadores que se desempeñan en las asignaturas Probabilidad y Estadística, Sistemas de Datos, Bases de Datos, Inteligencia Artificial y Programación Web, entre otras, de las carreras Licenciatura en Ciencias de la Computación y Licenciatura en Sistemas de Información pertenecientes al -DI- de la -FCEFN-UNSJ-.

También se encuentran insertos en el proyecto alumnos, tesistas y becarios de las carreras del DI.

Los alumnos de grado se hallan realizando sus trabajos finales en las líneas de investigación mencionadas.

Miembros del equipo dirigen y asesoran trabajos finales de grado y posgrado; así como becas de Investigación, categoría Alumnos Avanzados CICITCA y CIN.

Es de destacar, en los casos detallados en el presente documento, la importante colaboración brindada por los alumnos Ana Paula Molina, Pablo Gómez, Sergio Quiroga y Juan Olivares, durante el desarrollo de la asignatura IA.

5. **BIBLIOGRAFÍA**

- Alcalde, I. (2015). *Visualización de la información: de los datos al conocimiento*. Editorial UOC.
- Anderson, Jesse (2018) Data engineers vs. data scientists. https://www.oreilly.com/ideas/data-engineers-vs-data-scientists
- Beguerí, G., & Malberti, A. (2017). Minería de Datos y una Aplicación en la Educación Superior. In *XIX Workshop de Investigadores en Ciencias de la Computación (WICC 2017, ITBA, Buenos Aires)*.
- Boulic, R., & Renault, O. (1991). 3d hierarchies for animation. *New Trends in Animation and Visualization, Edited by Nadia Magnenat-Thalmann and Daniel Thalmann, John Wiley & Sons ltd., England*.
- Cervone, G., Lin, J., &Waters, N. (Eds.). (2014). *Data Mining for Geoinformatics*. Springer New York.
- Karimi, H. A. (2014). *Big Data: techniques and technologies in geoinformatics*. Crc Press.
- Klenzi, R. O., Malberti, M. A., & Beguerí, G. E. (2018). Visualization in a Data Mining Environment from a Human Computer Interaction Perspective. *Computación y Sistemas*, *22*(1).
- KNIME (2016). KNIME Analytics versión 3.2.1. Software.
- Ortega, Matías."El negocio de datos en Argentina tiene un potencial de u$s 19.000 millones" https://www.ambito.com/el-negocio-datos-argentina-tiene-un-potencial-us-19000-millones-n4011082 Enero 2018

- Pierson, L. (2015). *Data sciencefordummies*. John Wiley&Sons.
- Salgueiro, A. P. (2016, September). Plataforma de Internet de las cosas para la enseñanza y la investigación. In *XIV Congreso Internacional de Información Info'2016*.
- Smith, A., & Jones, B. (1999). On the complexity of computing. *Advances in Computer Science*, 555-566.