Universidad Estatal a Distancia

Vicerrectoría Académica

Escuelas Ciencias Exactas y Natural

Cátedra de Ingeniería de Software

Base de Datos

Código: 00826 Créditos: 3

Grado académico: Diplomado

Tarea 1

Estudiante: Andrés Cano Barboza

Modalidad: Virtual Nivel de virtualidad: Avanzada

PRIMER CUATRIMESTRE

2021

Realice una investigación sobre el concepto y aspectos generales sobre data science, data analytics, data mining, big data, machine learning, y conteste las siguientes preguntas.

1. ¿Qué es y cuál es la principal característica de cada una de las 5 disciplinas o prácticas mencionadas?

Observando estos conceptos desde una perspectiva más amplia se puede identificar como estos forman parte de un todo y algunos de estos pueden ser englobados por otros. Por ejemplo, se podría mencionar que el data science puede absorber un tema como el data analytics pero, no así en viceversa, estos se podrá observar a través del análisis de las definiciones y principal propósito siendo así una percepción conceptual de los leído. Por tal se tiene la siguiente información:

Data Science

Según lo que menciona Longbing (2017) los principios de la aparición de la palabra data science se puede indicar en el prefacio del libro “Concise Survey of Computer Methods” del año 1974, aquí se puede observar el término en la literatura. Este la define como "la ciencia de tratar con datos, una vez han sido establecidos” (p. 43). Se puede definir como:

> Nuevo campo interdisciplinario que sintetiza y se basa en estadísticas, informática, comunicación, gestión y sociología para estudiar los datos y sus entornos (incluidos los dominios y otros aspectos contextuales, como organizacionales y aspectos sociales) con el fin de transformar los datos en conocimientos y decisiones siguiendo una metodología y pensamiento de datos a conocimiento y sabiduría. (Longbing, 2017. p 8).

De esta definición se puede observar su principal característica la cual tiene que ver con la utilización de información de gran volumen para transformarla y convertirla en información valiosa para quien la requiera. Hoy en día para las grandes empresas, por ejemplo.

Data Analytics

Este hace referencia al manejo de los datos centrado en el entendimiento del mismo, por tal aquí se verán herramientas o técnicas para su agrupación e interpretación. Por lo que se puede tomar desde la exploración de datos (mediante análisis descriptivo y predictivo) hasta la entrega de conocimientos y decisiones procesables a través de análisis prescriptivos y entrega de conocimiento procesable. Para Longbing (2017) el data analytics se refiere a "las teorías, tecnologías, herramientas y procesos que permiten una comprensión profunda y descubrimiento de información útil sobre los datos. Datos La analítica consiste en analítica descriptiva, analítica predictiva y analítica prescriptiva" (p. 4).

Data Mining

La minería de datos se podría entender como una extensión del data science el cual se puede definir como una técnica en el análisis de grandes volúmenes de datos. Se puede definir como "el análisis masivo de la información contenida en una base de datos o datawarehouse para extraer relaciones, patrones de comportamiento, tendencias, ciclos estacionales, anomalías, etc." (Vicente, 2019. p. 22). Este mismo autor menciona ámbitos de la minería de datos como: Web Mining, Text Mining, Análisis de Sentimientos, Business Intelligence, etc.

Ferri (2004) presenta una amplia lista en la que se puede nombrar aplicaciones de la minería de datos:

- Aplicaciones financieras y banca: Obtención de patrones de uso fraudulento de tarjetas de crédito, etc.
- Análisis de mercado distribución y comercio: Evaluación de campañas publicitarias, etc.
- Seguros y salud privada: Predicción de clientes que contratan nuevos clientes, etc.
- Educación: Detección de abandonos y de fracaso.
- Medicina: Detección de pacientes con algún riesgo, etc.

- Telecomunicaciones: Modelos de carga en redes, etc.
- Otras áreas.

Con esto se podría entender que la minería tiene como característica principal el manejo de gran cantidad de datos, con técnicas específicas con el objeto de analizar comportamientos, tendencias o comprobar hipótesis. Básicamente es una herramienta que ayuda a comprender el contenido de un repositorio de datos.

Big data

Vicente (2019) define el Big Data como “un gran conjunto de datos o información generados por un gran número de diversas fuentes y de forma muy rápida.” (p. 18). Esta información puede venir de cualquier fuente ya sea de manera directa o indirecta. Debido al volumen de los datos estos deben ser manipulados por sistemas informáticos no habituales. Además, su característica principal está enfocada en la gestión de los datos y no así, en la aplicación de técnicas o pruebas para la interpretación del dato. Existen cincos componentes importantes del big data:

- Volumen: ¿Cuántos datos?
- Velocidad: ¿Con qué rapidez esos datos se procesan?
- Variedad: ¿Los diferentes tipos de datos o su heterogeneidad?
- Veracidad: ¿Fiabilidad de los datos utilizados en el proceso de toma de decisiones?
- Valor: ¿Obtener información de los grandes datos de manera rentable?

Machine Learning

Urcuqui (2018), menciona que se puede hablar de dos dimensiones, la primera está orientada a los procesos de pensamiento y razonamiento mientras que la segunda se encamina al comportamiento. Ambas enfocadas en el estudio de la fidelidad del rendimiento humanos o concepto de racionalidad.

"*Machine learning* es la parte de la inteligencia artificial que busca que un sistema tenga la capacidad de aprender en entornos variables, sin que sea programado de forma explícita." (Urcuqui, 2018. p. 27). Haciéndose así desde tres técnicas; aprendizaje supervisado, aprendizaje no supervisado y aprendizaje por refuerzo. Su propósito viene a ser muy similar al de la minería de datos, con técnicas de agrupación, clasificación o regresión con la diferencia de que aquí se trata de inteligencia artificial.

2. ¿Cuáles son las principales diferencias entre data science y data analytics? Describa un escenario donde se utilizaría una y un escenario donde se utilizaría la otra.

Desde las definiciones planteadas en este documento se puede observar que existe una relación muy íntima entre ambos conceptos, sin embargo, todo tiene algún tipo de límite y gracias a estos límites se pueden platear diferencias.

Se puede decir que la principal diferencia entre ambos se da en el enfoque de que área. Ambas trabajan con estrategias y técnicas matemáticas, pero, el Data Science está centrado en alto volumen de datos, mientras que el Data Analytics pertenece al manejo de información más específica, para algo determinado y muy definido. Además, el Data Science utiliza algoritmos matemáticos de más alta complejidad (ej. Machine learning) que el Data Analytics.

El Data Science se podría utilizar en escenarios donde se necesita de agrupación de alto volumen de datos y generar predicciones sobre esos datos, de las variables que se analizan, por ejemplo en el área de la salud se podría utilizar para crear sistemas para la optimización de los diagnósticos médicos, análisis de bases de datos clínicas, modelos de detección temprana de enfermedades, inteligencia artificial y machine learning para telemedicina, etc.

Mientras que el Data Analytics con propósitos más específicos se podría utilizar en empresas comerciales, de mercadeo o políticos como por ejemplo en política

utilizando, manejando e interpretando datos, extrayendo y generando estrategias en función de las preferencias de sus votantes y sus canales de comunicación preferidos.

En el blog perteneciente a Iron Hack llamado *Data science vs. data analytics* se mencionan las siguientes diferencias:

Data science	Data analytics
Creación de modelos y algoritmos predictivos.	Saca conclusiones de diferentes fuentes de datos.
Campo de actividad más amplio y diverso.	Campo de actividad limitado al sector empresarial.
Experto en estadística y matemáticas.	Familiarizado con el almacén de datos, las herramientas ETL y la inteligencia empresarial.
Experiencia con SQL.	Fuerte dominio de Python y R.
Experto en Python, R, SAS y Scala.	Experto en disputa de datos.
Conocimientos avanzados de machine learning.	Experto en visualización de datos.
Tiende a trabajar con datos no estructurados.	Conocimientos empresariales y habilidades para la toma de decisiones.
Aplicaciones en sectores como inteligencia artificial, salud, blockchain o motores de búsqueda de sitios web.	Aplicaciones en sectores como retail, viajes, sanidad o marketing.

3. ¿Cuál es la diferencia entre data mining y machine Learning?

Tanto Machine Learning como Data Mining vienen con propósitos muy similares por lo que a veces distinguir entre ambas se puede volver un poco complicado, debido a sus similitudes, ambas trabajan con herramientas matemáticas para lograr los resultados deseados, incluso utilizan algoritmos para su desarrollo.

Sin embargo, existe una diferencia muy marcada que se puede mencionar y es que el Data Mining se basa en extraer toda la información posible que se tiene a disposición, en esta línea se depende en gran medida del factor humano, sin este el proceo no puede empezar, funcionar o terminar, esto también le da un enfoque al Data Mining de centrarse en el descubrimiento del conocimiento. En el caso de Machine Learning, el factor humano solo aparece en el momento de definir los algoritmos para que este funcione, ya que Machine Learning es inteligencia artificial, la cual se encarga de aprender automáticamente parámetros ara sus modelos, a partir del dato. Este utiliza el autoaprendizaje para mejorar su rendimiento y se encuentra enfocado en el resultado.

4. ¿Por qué los lenguajes R y Python son utilizados en Data analytics? ¿Cuáles son las diferencias y ventajas entre estos y un lenguaje como Java o C++?

Vicente (2019) menciona que el R, es un lenguaje especializado en todos los ámbitos de la estadística y en la presentación de resultados, además, tiene amplia popularidad por la mayoría de científicos del mundo universitario, en tanto que el lenguaje de programación orientado a objetos denominado Python, se ha incorporado al trabajo científico por ingenieros e informáticos que trabajan, es fácil de entender y cómodo para escribir cdigo, comparado con otros lenguajes. “Ambos programas interactúan entre ellos y se complementan; se observa que cada vez más investigadores utilizan ambos softwares.” (p. 40). Razones por las cuáles se podría pensar que son utilizados ya que, permiten mucha facilidad de uso, son compatibles con el manejo de bases de datos, además, de acceso gratuito. Por otro lado, estos lenguajes en conjunto pueden otorgar

a IBM SPSS (programa de estadística) más recursos analísticos y dar más potencia al programa.

A pesar que también se puede encontrar C++ y Java en el desarrollo de data analytics y sus otras áreas como la minería de datos y otros aspectos propios del data science, se puede observar como cada vez gana más terreno el Python o el R (el Python con fuerza). La diferencia entre ambas agrupaciones de programas radica en la simplicidad de su uso. C++ y Java requieren de mucho código para lograr una sola función, en otras palabras, hay que darle una mayor cantidad de instrucciones al programa para que realice la función que deseamos, a pesar de ser programas muy sólidos. Por otro lado, R y Python son programas de alto nivel que requieren de una menor cantidad de código para ejecutar una instrucción, sin dejar por fuera que su aprendizaje sobre cómo manejarlo y programar es más rápido y sencillo, cosa que también lo hace de preferencia por los responsables de realizar data analytics.

Referencias Bibliográficas

Ferri Ramírez, C. Ramírez Quintana, M. J. y Ferri Ramírez, C. (2004). Introducción a la minería de datos. Pearson Educación. https://elibro-net.cidreb.uned.ac.cr/es/ereader/uned/45314?page=34

Iron Hack. (19 de marzo de 2021). Data science vs. data analytics. Available at: <https://www.ironhack.com/en/data-analytics/data-science-data-analytics> [Accessed 21 March 2021].

Longbing Cao. 2017. Data Science: A Comprehensive Overview. ACM Comput. Surv. 50, 3, Article 43 (October 2017), 42 pages. DOI:https://doi.org/10.1145/3076253

P. C. F. Alarcón and S. L. R. Martínez, "Influencia del pre-procesamiento de datos dentro del desempeño de modelos de perfilamiento de clientes elaborados con herramientas de minería de datos," 2016 IEEE 11th Colombian Computing Conference (CCC), Popayan, Colombia, 2016, pp. 1-8, doi: 10.1109/ColumbianCC.2016.7750778

Pulido Romero, E. Escobar Domínguez, Ó. y Núñez Pérez, J. Á. (2019). Base de datos. Grupo Editorial Patria. https://elibro-net.cidreb.uned.ac.cr/es/ereader/uned/121283?page=188

S. M. Sulaiman, P. A. Jeyanthy and D. Devaraj, "Smart Meter Data Analysis Issues: A Data Analytics Perspective," 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS), Tamilnadu, India, 2019, pp. 1-5, doi: 10.1109/INCOS45849.2019.8951377.

Urcuqui L. C. C. García P. M. y Osorio Q. J. L. (2018). Ciberseguridad: un enfoque desde la ciencia de datos. Editorial Universidad Icesi. https://elibro-net.cidreb.uned.ac.cr/es/ereader/uned/120435?page=27

Vicente Vírseda, J. A. González Arias, J. y Parra Rodríguez, F. J. (2019). Métodos de Data Science aplicados a la Economía y a la Dirección y Administración de Empresas. UNED - Universidad Nacional de Educación a Distancia. https://elibro-net.cidreb.uned.ac.cr/es/ereader/uned/122249?page=18