UNIVERSIDAD NACIONAL AUTÓNOMA DE HONDURAS
UNAH-TEC Danlí

ASIGNATURA:
IA-251 Gerencia Informática II

ASIGNACION:
Investigación y Defensa sobre Ciencias de Datos

SECCION:
1600

CATEDRÁTICA:
Ing. Cristiana Suyapa Ferrera Duarte

ESTUDIANTES:

Claudia Waleska Molina Salinas	20162530072
Saydi Gissell Calix Ávila	20122500055
Wuilmer Avimael Ávila Cáceres	20162500080

FECHA:
09/3/2022

Introducción

El presente informe tiene como finalidad sustentar la investigación y defensa virtual sobre **Ciencia de Datos** en la asignatura **IA251 Gerencia Informática II**; para el tercer periodo académico en **UNAH-TEC Danlí**.

La ciencia de datos combina múltiples campos, como las estadísticas, los métodos científicos, la inteligencia artificial y el análisis de datos para extraer el valor de los datos y abarca la preparación de los datos para el análisis, incluida la limpieza, la agregación y la manipulación de los datos para realizar análisis avanzados. Las aplicaciones analíticas y los científicos de datos pueden revisar los resultados para descubrir patrones y permitir que los líderes empresariales obtengan información fundamentada.

Los practicantes de la ciencia de datos se llaman científicos de datos y combinan una variedad de conocimientos para analizar los datos recopilados de la web, teléfonos inteligentes, clientes, sensores y otras fuentes para obtener información útil.

Ciencia de Datos

Origen de la ciencia de datos

La Ciencia de Datos tiene su origen en el año 1962, cuando el estadístico estadounidense John W. Tukey, conocido por el desarrollo de complejos algoritmos y el famoso diagrama de caja y bigotes (Box Plot), escribe y se cuestiona el futuro de la estadística como ciencia empírica. Sin embargo, no sería hasta más adelante en 1974 cuando Peter Naur, científico danés conocido por sus trabajos en las ciencias computacionales y ganador del premio Turing en el año 2005, acuñara el término que actualmente conocemos.

Por su parte, William S. Cleveland, informático y estadístico estadounidense, famoso por sus aplicaciones en la visualización de datos, introdujo en 2001 a la Ciencia de Datos como una disciplina unificada y con independencia de lo que hasta ese momento se había conocido como Estadística. Un año más tarde, en 2002, comienzan las publicaciones de la primera revista científica en lo referente a los datos, la conocida como Data Science Journal. Esta revista fue fundada con el fin de promover a través de sus artículos la Ciencia de Datos y su correspondiente aplicación en áreas como las políticas públicas, las prácticas y la gestión de Datos Abiertos (datos accesibles en los que se garantiza su fiabilidad y su estructuración) para contribuir a la eficacia y eficiencia en el conocimiento y el aprendizaje.

Actualmente, es una herramienta estadística y de investigación que se utiliza como plataforma para ampliar, compartir y difundir el conocimiento.

Definición de ciencia de datos

La Ciencia de Datos (**Data Science**) se encarga de analizar grandes volúmenes de información con la ayuda de la inteligencia artificial para la toma de decisiones comerciales, la planificación estratégica y otros usos. Es cada vez más crítico para las empresas: la información que genera la ciencia de datos ayuda a las organizaciones a aumentar la eficiencia operativa, identificar nuevas oportunidades comerciales y mejorar los programas de marketing y ventas, entre otros beneficios.

La ciencia de datos incorpora varias disciplinas tales como: **ingeniería de datos**, **preparación de datos**, **minería de datos**, **análisis predictivo, aprendizaje automático** (machine learning, ML) y **visualización de datos**, así como estadísticas, matemáticas y programación de software. Lo realizan principalmente científicos de datos capacitados, aunque también pueden participar analistas de datos de nivel inferior. Además, muchas organizaciones ahora dependen en parte de los científicos de datos ciudadanos, un grupo que puede incluir profesionales de inteligencia empresarial (BI), analistas empresariales, usuarios empresariales conocedores de datos, ingenieros de datos y otros trabajadores que no tienen una formación formal en ciencia de datos.

Con los sistemas basados en ciencia de datos se puede acceder de forma automatizada a cientos y hasta millones de documentos al mismo tiempo, para analizarlos y extraer de ellos información puntual, ya sea para citar instancias o fallos específicos y preparar casos actuales con base en ello con gran rapidez. Con base en todo lo explicado hasta ahora, podemos decir que la importancia de la ciencia de datos reside en la posibilidad de generar un conocimiento profundo de cualquier proyecto, e incluso del negocio jurídico en general y hasta de los competidores y clientes. A continuación, hablaremos sobre cada uno de sus aspectos.

Proceso y ciclo de vida de la ciencia de datos

Los proyectos de ciencia de datos implican una serie de pasos de recopilación y análisis de datos. Por lo tanto, se describen estos seis pasos principales:

- Identifique una hipótesis relacionada con el negocio para probar.
- Reúna datos y prepárelos para su análisis.
- Experimente con diferentes modelos analíticos.
- Elija el mejor modelo y ejecútelo con los datos.
- Presente los resultados a los ejecutivos de empresas.
- Implemente el modelo para un uso continuo con datos nuevos.

Desafíos en la ciencia de datos

La ciencia de datos es intrínsecamente desafiante debido a la naturaleza avanzada de la analítica que involucra. La gran cantidad de datos que normalmente se analizan se suma a la complejidad y aumenta el tiempo que lleva completar los proyectos. Además, los científicos de datos trabajan con frecuencia con grupos de big data que pueden contener una variedad de datos estructurados, no estructurados y semiestructurados, lo que complica aún más el proceso de análisis.

Las técnicas estadísticas y analíticas comunes que se utilizan en proyectos de ciencia de datos incluyen las siguientes:

- **Clasificación**, que separa los elementos de un conjunto de datos en diferentes categorías;

- **Regresión**, que traza los valores óptimos de las variables de datos relacionadas en una línea o un plano; y

- **Agrupación**, que agrupa puntos de datos con una afinidad o atributos compartidos.

Equipo de ciencia de datos

Muchas organizaciones han creado un equipo separado, o varios equipos, para manejar las actividades de ciencia de datos. También puede incluir los siguientes puestos:

- **Ingeniero de datos**: Las responsabilidades incluyen la configuración de canalizaciones de datos y la ayuda en la preparación de datos y la implementación del modelo, trabajando en estrecha colaboración con los científicos de datos.

- **Analista de datos**: Este es un puesto de nivel inferior para los profesionales de análisis que no tienen el nivel de experiencia o las habilidades avanzadas que tienen los científicos de datos.

- **Ingeniero de aprendizaje automático**: Este trabajo orientado a la programación implica desarrollar los modelos de aprendizaje automático necesarios para las aplicaciones de ciencia de datos.

- **Desarrollador de visualización de datos**: Esta persona trabaja con científicos de datos para crear visualizaciones y cuadros de mando que se utilizan para presentar los resultados de los análisis a los usuarios comerciales.

- **Traductor de datos**: También llamado traductor analítico, es un rol emergente que sirve como enlace con las unidades de negocios y ayuda a planificar proyectos y comunicar resultados.

- **Arquitecto de datos**: Un arquitecto de datos diseña y supervisa la implementación de los sistemas subyacentes utilizados para almacenar y administrar datos para usos analíticos.

Por lo general, el equipo está dirigido por un director de ciencia de datos, un gerente de ciencia de datos o un científico de datos líder, que puede depender del director de datos, el director de análisis o el vicepresidente de análisis; el científico de datos jefe es otro puesto de gestión que ha surgido en algunas organizaciones.

Plataformas y herramientas de ciencia de datos

Hay numerosas herramientas disponibles para que los científicos de datos las utilicen en el proceso de análisis, incluidas opciones comerciales y de código abierto:

- Plataformas de datos y motores de análisis, como bases de datos Spark, Hadoop y NoSQL;
- Lenguajes de programación, como Python, R, Julia, Scala y SQL;
- Herramientas de análisis estadístico como SAS e IBM SPSS;
- Bibliotecas y plataformas de aprendizaje automático, incluidas TensorFlow, Weka, Scikit-learn, Keras y PyTorch;
- Jupyter Notebook, una aplicación web para compartir documentos con código, ecuaciones y otra información; y
- Bibliotecas y herramientas de visualización de datos, como Tableau, D3.js y Matplotlib.

Además, los proveedores de software ofrecen un conjunto diverso de plataformas de ciencia de datos con diferentes características y funcionalidades. Eso incluye plataformas de análisis para científicos de datos capacitados, plataformas de aprendizaje automático automatizadas que también pueden ser utilizadas por científicos de datos ciudadanos y centros de flujo de trabajo y colaboración para equipos de ciencia de datos. La lista de proveedores incluye Alteryx, AWS, Databricks, Dataiku, DataRobot, Domino Data Lab, Google, H2O.ai, IBM, Knime, MathWorks, Microsoft, RapidMiner, SAS Institute, Tibco Software y otros.

Científico de Datos

Las tareas de un científico de datos pueden incluir el desarrollo de estrategias para analizar datos; la preparación de datos para su análisis; explorar, analizar y visualizar datos; construir modelos con datos mediante lenguajes de programación como Python y R; e implementar modelos en aplicaciones.

El científico de datos no trabaja solo. De hecho, la ciencia de datos más efectiva se ejecuta en equipos. Además de un científico de datos, este equipo puede incluir un analista empresarial que define el problema, un ingeniero de datos que prepara los datos y su método de acceso, un arquitecto de tecnología informática que supervisa los procesos subyacentes y la infraestructura, y un desarrollador de aplicaciones que implementa los modelos o las salidas del análisis en aplicaciones y productos.

Desafíos de la implementación de la ciencia de datos

A pesar de la promesa de la ciencia de datos y las grandes inversiones en equipos de ciencia de datos, muchas empresas no materializan todo el valor de sus datos. En su carrera por contratar talento y crear programas de ciencia de datos, algunas empresas han experimentado flujos de trabajo ineficientes para los equipos, donde diferentes personas utilizan diferentes herramientas y procesos que no funcionan bien en conjunto. Sin una administración centralizada más disciplinada, es probable que los ejecutivos no obtengan un retorno completo de sus inversiones.

La plataforma de ciencia de datos

Muchas compañías se percataron de que, si no cuentan con una plataforma integrada, el trabajo de la ciencia de datos es ineficiente, inseguro y difícil de escalar. Esto condujo al desarrollo de plataformas de ciencia de datos. Estas plataformas son centros de software, alrededor de los cuales se lleva a cabo todo el trabajo de ciencia de datos. Una buena plataforma alivia muchos de los desafíos de la implementación de la ciencia de datos y ayuda a las empresas a convertir sus datos en información de forma más rápida y eficiente.

Con una plataforma centralizada de aprendizaje autónomo, los científicos de datos pueden trabajar en un entorno de colaboración con sus herramientas de código abierto favoritas, y donde todo su trabajo se sincroniza mediante un sistema de control de versiones.

Los beneficios de una plataforma de ciencia de datos

Una plataforma de ciencia de datos disminuye las redundancias y fomenta la innovación al permitir que los equipos compartan código, resultados e informes. Elimina los cuellos de botella en el flujo de trabajo al simplificar la administración e incorporar prácticas recomendadas.

En general, las mejores plataformas de ciencia de datos tienen como objetivo:

- Permitir que los científicos de datos sean más productivos al ayudarlos a acelerar y entregar los modelos en forma más rápida y con menos errores.
- Facilitar que los científicos de datos trabajen con grandes volúmenes y variedades de datos.
- Brindar una inteligencia artificial confiable, de categoría empresarial, que esté libre de sesgos, sea auditable y reproducible.

Lo que un científico de datos requiere de una plataforma

Elegir una interfaz de usuario basada en proyectos que fomente la colaboración: La plataforma debe facultar a las personas para que trabajen en conjunto en un modelo, desde la concepción hasta el desarrollo final.

Priorizar la integración y la flexibilidad: Asegurándose de que la plataforma sea compatible con las últimas herramientas de código abierto; proveedores comunes de control de versiones como GitHub, GitLab y Bitbucket; y una estrecha integración con otros recursos.

Incluir funcionalidades de categoría empresarial: Asegurándose de que la plataforma pueda escalar con su negocio a medida que crece su equipo. La plataforma debe contar con un alto grado de disponibilidad, tener controles de acceso robustos y admitir una gran cantidad de usuarios concurrentes.

Permitir que la ciencia de datos se convierta en autoservicio: Buscar una plataforma que reduzca la carga del departamento de Tecnología Informática e Ingeniería y permita que los científicos de datos creen de manera instantánea entornos, realicen un seguimiento de todo su trabajo e implementen fácilmente modelos en la producción.

Garantizar una implementación más sencilla de los modelos: La implementación y puesta en funcionamiento del modelo es uno de los pasos más importantes del ciclo de vida del aprendizaje autónomo.

Big Data vs Data Science

Big data y data science emergieron para transformar y dotar de sentido al panorama digital y tecnológico actual. A continuación, se presentan algunas de las principales diferencias ambos conceptos:

- Los macrodatos se distinguen por su variedad, velocidad y volumen. Mientras que la ciencia de datos proporciona los métodos o técnicas para analizarlos.

- La inteligencia de datos proporciona el potencial de rendimiento. No obstante, es la ciencia de datos la que utiliza enfoques teóricos y experimentales, además del razonamiento deductivo e inductivo.

- El análisis de big data realiza la extracción de información útil de grandes volúmenes de conjuntos de datos. Contrariamente al análisis, la ciencia de datos utiliza algoritmos de aprendizaje automático y métodos estadísticos para entrenar a los ordenadores y obtener predicciones precisas. De este modo, la ciencia de datos no debe confundirse con el análisis de los macrodatos.

- Big Data se relaciona más con la tecnología (Hadoop, Java, Hive, etc.) la computación distribuida y las herramientas y el software de análisis. Esto se opone al otro concepto que se enfoca en estrategias para decisiones de negocios, diseminación de datos utilizando matemáticas, estadísticas, etc.

De las diferencias anteriores se puede observar que el concepto data science se engloba dentro del concepto de big data. En este sentido, la ciencia de datos juega un papel importante en muchas áreas de aplicación. En resumidas cuentas, data science se desenvuelve dentro del ámbito del big data para obtener información útil a través del análisis predictivo, donde los resultados se utilizan para tomar decisiones inteligentes. De esta forma, sin big data no existiría el concepto de data science. Y sin el segundo, el primero no tendría (u obtendría) tanto valor.

Resumen

La ciencia de datos revela tendencias y genera información que las empresas pueden utilizar para tomar mejores decisiones y crear productos y servicios más innovadores. Quizás lo más importante es que permite que los modelos de aprendizaje autónomo administren de una manera más eficiente las grandes cantidades de datos que se les suministran en vez de depender principalmente de los analistas de negocios para ver qué pueden descubrir a partir de los datos.

En conclusión, la ciencia de datos es importante, ya que las empresas disponen un tesoro de datos sin aprovechar. Ahora que la tecnología moderna ha permitido la creación y el almacenamiento de cantidades cada vez mayores de información, el volumen de datos explotó.