INGENIERÍA EN SISTEMAS COMPUTACIONALES

“Estado del arte de la Ciencia de datos”

Investigación

Mtra. Larissa J. Peniche Ruiz

Materia: Inteligencia Artificial

8SI

Autor:

Br. Carmen Fernanda Medina Andrade

Br. Omar Emmanuel Miranda Fernández

Br. Diego Armando Martinez Ruiz

Br. Angel Stefania Gomez Gonzalez

28/04/2022

Ciencia de Datos

Definición

La Ciencia de Datos es un conjunto de conocimientos, que se combina en múltiples campos como es el de matemáticas y estadísticas, programación en R y Python, Negocio o Academia y por último en el campo de la comunicación, para tener la información de algo concreto, es decir, aportar valor con información esto haciendo referencia en que todo lo que vamos a aportar será basado en datos sin guiarnos por intuiciones o por alguna teoría.

La ciencia de datos abarca toda la preparación de los datos para crear un análisis de esto, es decir, desde la limpieza, la agregación y la manipulación de los mismos datos para realizar un análisis avanzado. De esta manera y con este análisis podemos obtener una amplia información sobre el tema del cual estamos recolectando datos para una mejor utilización del mismo. (Oracle, 2020)

Ciencia de datos es el campo de la aplicación de técnicas analíticas avanzadas y principios científicos para extraer información valiosa de los datos para la toma de decisiones comerciales, el uso de los datos que recolectan y los análisis a partir de ellos son cada vez más críticos para la empresas, ya que ayuda a estas mismas a identificar nuevas oportunidades comerciales, mejorar sus estrategias de ventas, crecer su marketing, y en algunos casos sacar ventaja sobre sus rivales comerciales. En última instancia, pueden generar ventajas competitivas sobre los rivales comerciales. (Craig Stedman, Septiembre 2021)

Modelo Matemático o Metodología

¿Qué es una metodología?

Antes de comenzar a analizar cuál es la metodología que sigue la ciencia de los datos, debemos cuestionarnos qué es una metodología, y bien, una metodología es una estrategia general que sirve de guía para los procesos y actividades que están dentro de un dominio determinado. La metodología no depende de tecnologías, ni tampoco es un conjunto de técnicas, sino más bien es un marco sobre cómo proceder con los métodos y argumentos que se usarán para obtener resultados.

Etapa 1: Comprensión del negocio

En esta primera etapa de la metodología se dice que todos los proyectos deben empezar con la comprensión del negocio, que nos dice esto, que se debe de definir el problema que se planea resolver, los objetivos que va a alcanzar el proyecto, los requisitos que se necesitan para poder llegar a una solución desde una perspectiva empresarial. La primera parte o esta primera etapa es importante para que el problema del negocio se pueda resolver exitosamente.

Etapa 2: Enfoque analítico

En la segunda etapa de la metodología, en este caso el científico de datos prodra definir el enfoque analitico que utilizara para resolver el problema, siempre y cuando este último esté establecido claramente. El objetivo de esta etapa implica expresar el problema bajo el contexto de las técnicas estadísticas y de aprendizaje automático, para que la organización pueda identificar las más adecuadas para el resultado deseado.

Etapa 3: Requisitos de datos

En la tercera etapa podemos ver que según la etapa anterior el enfoque analítico elegido determina los requisitos de datos. Más concretamente, los métodos analíticos a utilizar requieren de determinados contenidos de datos, formatos y representaciones, orientados por el conocimiento en el dominio.

Etapa 4: Recopilación de datos

Para la etapa de la recopilación de los datos, el que esté a cargo de este proyecto, tiene la labor de identificar y reunir los datos disponibles y relevantes para el dominio del problema. Si hay algunas lagunas en la recopilación de datos, es posible que el científico tenga que revisar los requisitos de datos y recopilar más datos o nuevos datos. Aunque el muestreo sigue siendo importante, las plataformas de alto rendimiento, y la funcionalidad analítica en la base de datos permiten que los científicos de datos utilicen conjuntos de datos mucho más grandes que contienen gran parte de los datos disponibles, o incluso

todos. Esto al agregar más datos, los modelos pueden tener eventos raros, como puede ser un fallo del sistema.

Etapa 5: Comprensión de datos

En la quinta etapa de la metodología, después de la recopilación de los datos, para esta etapa los científicos de datos suelen utilizar estadísticas descriptivas y técnicas de visualización para comprender el contenido de los datos, evaluar su calidad, y tener una perspectiva de los datos. Pero para llenar si en dado caso existen espacios vacíos, lo más probable es que sea necesario volver a recopilar datos.

Etapa 6: Preparación de datos

Para esta sexta etapa, se deberá abarcar todas las actividades para construir el conjunto de datos que se utilizará en la siguiente etapa de la metodología. El objetivo de esta etapa de la metodología es el de la limpieza de los datos, que consiste en eliminar si los datos se encuentran duplicados o darles un formato, de igual manera tenemos combinar los datos fuentes, o sea en este caso tablas, archivos y plataformas, y por último transformar los datos en variables aún más útiles. Los científicos de datos utilizan el proceso llamado ingeniería de características con la finalidad de crear variables explicativas, o también llamadas características. Cuando hay disponibles datos en texto, como los registros del centro de atención al cliente o las observaciones de los médicos en forma no estructurada o semiestructurada, la analítica de texto se puede utilizar para derivar nuevas variables estructuradas y, así, enriquecer el conjunto de indicadores y mejorar la precisión del modelo. Esta etapa se dice que es la más larga del proyecto de ciencia de datos.

Etapa 7: Modelado

En esta séptima etapa llamada etapa de modelado, se utiliza el conjunto de los datos ya preparados y se enfoca en desarrollar modelos descriptivos según el enfoque analítico que ya habíamos previamente definido. En los modelos predictivos, los científicos de datos utilizan un conjunto de capacitación para de esta manera construir el modelo. El proceso de modelado normalmente es muy repetitivo, ya que las organizaciones van adquiriendo perspectivas que hace que los datos tengan ajustes y esto influya en el modelado. Luego de

varias pruebas con múltiples algoritmos los científicos de los datos buscan encontrar el mejor modelo para las variables disponibles.

Etapa 8: Evaluación

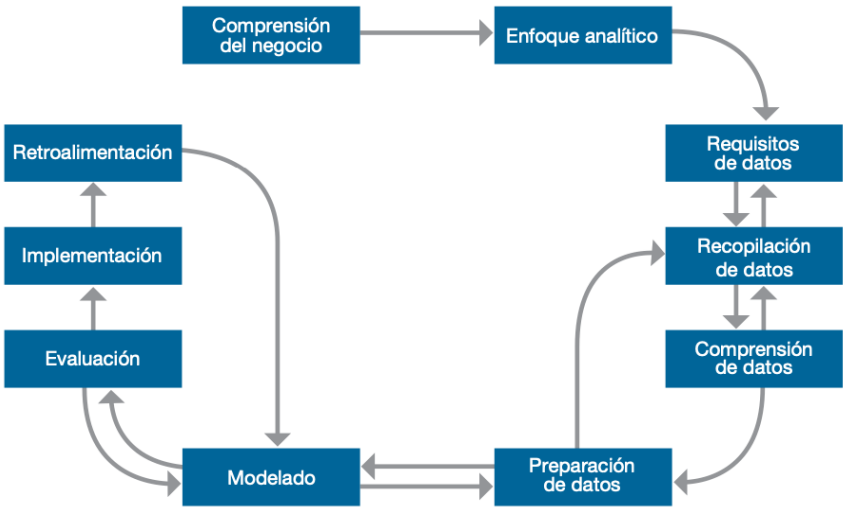
Durante la octava etapa se va dando el desarrollo del modelo y para antes de que se implemente el modelo, el científico de datos en este caso debe evaluar el modelo para estar de acuerdo en la calidad y ver que aborda el problema inicial de manera adecuada y completa. Durante esa evaluación el científico de datos puede utilizar tablas o gráficos de los datos para que pueda ver como va la eficacia del modelo en cuestión de resolver el problema. De igual manera las herramientas que se utilizan para evaluar el modelo se pueden ajustar según las necesidades para que el modelo pueda ser lo más exitoso posible.

Etapa 9: Implementación

Ahora en la novena etapa, luego de que ya el modelo ha sido desarrollado y aprobado por los promotores del negocio, se empieza a implementar en el entorno de producción o en un entorno de pruebas comparable. La implementación del modelo generalmente abarca grupos, habilidades y tecnologías adicionales dentro de la empresa. Por ejemplo, un grupo de ventas puede implementar un modelo de propensión a la respuesta a través de un proceso de administración de campañas creado por un equipo de desarrollo y administrado por un grupo de marketing.

Etapa 10: Retroalimentación

Por último en la etapa de retroalimentación es solamente una etapa en donde se recopilan todos los resultados que obtuvimos del modelo que implementamos en el proyecto, entonces a partir de esos resultados el científico de datos puede ajustar el modelo para mejorar la precisión de los resultados y de esa manera obtener un mayor éxito dentro del proyecto.



Características Relevantes

La principal característica de la ciencia de datos, es la capacidad de recopilar información de cualquier tipo, como imágenes, carácteres numéricos, texto, videos, entre muchos otros datos que se pueden interpretar. Dichos datos se analizan de forma optimizada, buscando siempre la manera sencilla en la que se pueda visualizar, creando patrones de comportamiento de los datos y formando reglas que puedan determinar o predecir nueva información.

Estas características son los cimientos de la ciencia de datos, ya que de igual forma se puede interpretar como la base del análisis de la información. Esto quiere decir que la ciencia de datos es multidisciplinar, ya que es utilizada en diferentes áreas del conocimiento, como el área industrial, farmacéutica, el mundo de los videojuegos, la economía, etc.

Ventajas

La ciencia de datos trajo consigo varias oportunidades de optimización y eficacia en los procesos que se lleven a cabo en los diferentes ámbitos en una empresa u organización, ya que, las herramientas para procesar estas grandes cantidades de datos, lo hacen a gran velocidad, haciendo que las tomas de decisiones sean más ágiles, aumentando el margen de operación y reduciendo al mínimo las potenciales pérdidas en las estrategias que se implementen.

Gracias al big data, se puede tener un seguimiento de los clientes o el público en general mucho más preciso, hablamos de cosas como gustos o hábitos, necesidades y tendencias, entre otras cosas, y de esta forma, crear mercadotecnia personalizada y mejor relacionada con los clientes.

Por lo general estos bancos de información se van actualizando y cambiando en tiempo real por lo que es fácil automatizar los análisis que reflejaran resultados completamente actuales lo que hace que la retroalimentación también sea inmediata.

La ciencia de datos ha dado buenos resultados entre las empresas, haciendo que estas tengan un mayor éxito que las que no usan estrategias estructuradas con big data, esto implica que cada vez más empresas invierten más en el desarrollo de su ciencia de datos, haciendo que crezca su participación en el mercado actual y mejorando cada vez más esta ciencia.

Desventajas

Hay dos problemas fundamentales a la hora de hablar de la generación de la información, el primero es que el volumen de los datos siempre está en constante aumento, lo que hace que la capacidad de los softwares para el procesamiento de esta información se vea superada y se vuelva obsoleta por lo que se necesitan constantes mejoras a los programas. Y el segundo problema es que los datos se recopilan de distintas partes y no todos son de utilidad para el propósito que se definió, esto los vuelve datos basura que solo ocupan espacio, por lo cual se tienen que eliminar.

La ciencia de datos es relativamente nueva y una ciencia que está en constante expansión, por lo cual, hay una escasez de personal que esté bien capacitado para el manejo de las grandes cantidades de información que conlleva el big data.

Como ya se dijo, la ciencia de datos está en constante evolución y siempre busca la manera de procesar cantidades más y más grandes de información, haciendo que implementarlo en una empresa sea costoso, por lo que hoy en día, casi siempre son las grandes empresas las que implementan técnicas de big data ya que pueden darse el lujo de mantener la inversión.

Clasificación de los datos

En la recopilación de datos pueden ser paramétricos y no paramétricos, siendo los paramétricos los datos que requieren que cumplan con ciertas propiedades, mientras que para los no paramétricos no necesitan cumplir con propiedades específicas al momento de manejar los datos, y por consiguiente su clasificación.
Nos permite el uso de recursos gráficos para una mejor visualización y comprensión de los datos, estos dependiendo de los datos que nos estén pidiendo; se acompañan mediante test para comparar los datos obtenidos contra los esperados.

Aplicaciones y Usos Recientes

El uso de los datos siempre ha extraído nuevo conocimiento, por lo que la ciencia de datos tiene ese mismo concepto pero llevado a una extracción de conocimiento masivo que tiene el fin de analizar y solucionar problemáticas que se muestran en diversas áreas, tales como:

- Salud
  La medicina es una de las áreas con mayor aplicación de ciencia de datos, desde sistemas programados para diagnósticos médicos, hasta el uso de la I.A. para innovaciones médicas.

- Procesos productivos
  Las empresas han ido adoptando la tecnología para la ejecución de procesos específicos de manera virtual, así como el monitoreo y control de los mismos, automatizando actividades diarias, optimizando los procesos en cadenas de abastecimiento, control de calidad, sistemas de mantenimiento predictivo y desarrollos que se efectúan gracias a las técnicas de ciencia de datos.

- Comunicación y conocimiento
  Las empresas recopilan diariamente los datos generados de los usuarios que navegan en internet. Sacando provecho de la información revelada en redes sociales y otros medios de comunicación, analizando los gustos, rutas, listas de amigos y otros datos

importantes, para que a partir de este análisis se pueda estudiar el comportamiento del usuario y arrojarle publicidad que le pueda interesar.

De igual forma, es importante saber que no todos los datos son iguales ni tienen el mismo valor o la misma calidad, y la empresa debe definir los datos que le pueden resultar útiles. Ésto significa también que el científico de datos debe conocer muy bien la empresa y sus fines y trabajar en relación estrecha con quienes definen los objetivos de los diversos proyectos. Algunos datos pueden ser útiles para un proyecto en particular pero resultan inservibles para otros. Así, las operaciones de análisis también pueden variar según los objetivos.

Avances y planes a futuro.

La ciencia de datos ha tenido muchos cambios desde a su creación hasta la fecha actual, buscando que todos los campos que esta abarca, mejoren y evolucionen, para la optimización de los datos; hasta la fecha se busca una mejor filtración de datos, esto para extraer los datos relevantes que la empresa considere de los clientes.
Con sus constantes cambios de igual forma se busca una manera ágil en la recopilación de información, de igual forma una profundización en su uso en las IA, ya que a pesar de que trabaja en conjunto, no ha sido completamente explotado; esto se dará con el uso de algoritmos más avanzados.

Avances publicados en revistas científicas.

Local:
CENTROGEO (CONACYT).

Centrogeo es una organización pública que se dedica a la realización de investigaciones científicas, formación de recursos humanos de alto nivel, desarrollo tecnológico e innovación entre otras cosas, desde la perspectiva de la ciencia de información geoespacial, y pertenece a la red de centros de investigación del CONACYT.

Centrogeo cuenta con una subsede en Mérida, Yucatán la cual fue anfitriona del programa ELCIR (Engineering Learning Community Introduction to Research) el cual es impartido por la Texas A&M University, este programa le enseña a jóvenes estudiantes los fundamentos de la construcción de la ciencia mediante prácticas de campo y en laboratorios llevadas de la mano por investigadores en Yucatán de múltiples áreas científicas, entre ellas la ciencia de datos en el área de geointeligencia computacional.

Los alumnos que participaron en esta área, tuvieron la oportunidad de colaborar con el proyecto llamado “Misoginia en Pocas Palabras”. Primero se familiarizaron con el uso del Autómata Geointeligente en Internet (AGEI), la cual fue desarrollada en CentroGeo y que permite la extracción, análisis y visualización de información pública en internet. Cada estudiante identificó twitts con contenido misógino, posteriormente se unieron los conjuntos de datos etiquetados de todos los alumnos y se dividieron en un conjunto de entrenamiento y un conjunto de prueba, y utilizaron el clasificador EvoMSA, desarrollado también por investigadores de CentroGeo, para crear un modelo que dada una publicación en la red social de Twitter, se pueda identificar si el contenido es misógino o no.

Regional:

Peñoles, empresa minera.

La Industria Peñoles, empresa minera coahuilense, ha creado un chaleco inteligente que monitorea el estado de salud de sus trabajadores. Específicamente fue creado para el personal que padece hipertensión y diabetes, pero puede ser usado por cualquier miembro de la organización.

Con esta prenda de vestir inteligente, se podrá monitorear la condición médica y física del empleado, detectando incidentes o complicaciones que se puedan presentar, mandando una cuadrilla de rescate y retirando a la persona en riesgo de la operación industrial en la que se encuentre.

Esta industria ha impartido innumerables charlas sobre la digitalización, ciberseguridad, minería y ciencia de los datos, participando en conferencias internacionales y compartiendo sus conocimientos sobre la minería industrial llevada a la era tecnológica.

Nacional:

CONACYT.

Especialistas exponen avances del Pronaii de Ciencia de Datos y Salud

Entre los avances que presenta la ciencia de datos; el CONACYT (Consejo Nacional de ciencia y tecnología), estipula que el Pronaii en ciencia de datos y salud, lo implemente en instancias públicas para la recaudación de información sobre enfermedades crónicas no transmisibles, dando a conocer nuevas estrategias para la mejora del servicio de salud.

Internacional:

Microsoft Power BI.

Microsoft, siendo la gran empresa multinacional del mundo de la tecnología, utiliza la ciencia de datos en varios de sus softwares y proyectos. Una de las aplicaciones más recientes que ha desarrollado el es el software llamado Power Bi.

La visualización y el análisis de datos basados en la web son cada vez más importantes y necesarios para las empresas, independientemente de su industria o tamaño y es lo que Microsoft quiere cubrir con Power Bi. Esta aplicación es una herramienta que recopila datos y los transforma en información inteligible, utiliza gráficos atractivos y fáciles de procesar, esto permite a los usuarios generar y compartir instantáneas claras y útiles de lo que está sucediendo en la empresa y puede conectarse a una amplia gama de fuentes de datos, desde hojas de cálculo básicas de Excel hasta bases de datos y aplicaciones en la nube.

Está construido usando aprendizaje automático y ciencia de los datos, lo que significa que puede detectar patrones en estos datos y usar esos patrones para hacer predicciones informadas y escenarios hipotéticos, Estas estimaciones permiten a los usuarios generar pronósticos para prepararse para satisfacer la demanda futura en otras métricas clave.

Referencias.

-Oracle. (2020). *¿Qué es la ciencia de datos?* Recuperado 27 de abril de 2022, de https://www.oracle.com/mx/data-science/what-is-data-science/

-Stedman, C. (2021, 8 septiembre). *Ciencia de datos*. ComputerWeekly.es. Recuperado 27 de abril de 2022, de https://www.computerweekly.com/es/definicion/Ciencia-de-datos

-Lemus-Delgado, D., & Pérez Navarro, R. (2020, 1 abril). *Ciencia de datos y estudios globales: Aportaciones y desafíos metodológicos*. Universidad de Los Andes: Revistas. Recuperado 27 de abril de 2022, de https://revistas.uniandes.edu.co/doi/full/10.7440/colombiaint102.2020.03

-IBM. (2015, junio). *Metodología fundamental para la ciencia de datos*. IBM.com. Recuperado 27 de abril de 2022, de https://www.ibm.com/downloads/cas/6RZMKDN8

-Menoyo Ros, D. García López, E. & García Cabot, A. (2021). *Fundamentos de la ciencia de datos.*. Editorial Universidad de Alcalá. https://elibro.net/es/ereader/biblioitmerida/177631?page=1

-Webedia Brand Services. (2020, 26 agosto). *Nueve industrias que aplican la Ciencia de Datos para solucionar problemas reales*. Xataka. Recuperado 27 de abril de 2022, de https://www.xataka.com/n/nueve-industrias-que-aplican-ciencia-datos-para-solucionar-problemas-reales

-Solutions, A. (2021, 1 junio). *Las distintas aplicaciones de Ciencia de Datos*. Argo Solutions. Recuperado 27 de abril de 2022, de https://useargo.com/es/blog/las-distintas-aplicaciones-de-ciencia-de-datos/

-Torralba, P. P. (2021, 20 diciembre). *Tendencias big data 2022 para que el futuro no te pille de sorpresa*. Thinking for Innovation. Recuperado 27 de abril de 2022, de https://www.iebschool.com/blog/tendencias-big-data/

-Thapaliya, R. (2021, 30 septiembre). *¿Habrá demanda de ciencia de datos en el futuro?* Entrepreneur. Recuperado 27 de abril de 2022, de https://www.entrepreneur.com/article/408312

-Consejo Nacional de Ciencia y Tecnología. (2022, 25 febrero). *Especialistas exponen avances del pronaii de ciencia de datos y salud*. CONACYT. Recuperado 27 de abril de 2022, de https://conacyt.mx/especialistas-exponen-avances-del-pronaii-de-ciencia-de-datos-y-salud/

-Bello, E. (2022, 4 febrero). *¿Qué es microsoft power BI? Todo lo que tienes que saber*. Thinking for Innovation. Recuperado 27 de abril de 2022, de https://www.iebschool.com/blog/microsoft-power-bi-analitica-usabilidad/

-Microsoft. (s. f.). *¿Por qué power BI: Funciones y ventajas | microsoft power BI*. powerbi.microsoft.com. Recuperado 27 de abril de 2022, de https://powerbi.microsoft.com/es-es/why-power-bi/

-BBVA. (2021, 2 noviembre). *Ventajas y desventajas del Big Data*. BBVA.CH. Recuperado 5 de mayo de 2022, de https://www.bbva.ch/noticia/ventajas-y-desventajas-del-big-data/

-Ganhi, H., Elena, M., & V, J. C. (2019). *GeoINT - research website*. GEOINT. Recuperado 5 de mayo de 2022, de http://www.geoint.mx/site/publicacion/id/76.html