CIENCIA DE DATOS O DATA ANALISIS: TÉRMINOS Y CONCEPTOS BÁSICOS

La ciencia de datos es un campo de la ciencia que busca extraer el conocimiento o el mejor entendimiento posible de datos estructurados o no estructurados involucrando múltiples disciplinas, métodos científicos y sistemas. Iniciar en esto, puede ser sumamente abrumador, y mas aun cuando no se manejan los términos básicos que son utilizados comúnmente en el campo. Para ayudar a aquellos que quieran iniciarse en el mundo de la Ciencia de datos, se presentará a continuación un glosario con los términos básicos necesarios para tener un buen entendimiento de la documentación que se consulte.

Términos Básicos

Big Data

Big Data es un termino que se refiere a un conjunto de datos de gran tamaño (volumen), con gran diversidad de tipos de datos (variedad) y con una velocidad de crecimiento elevada. El Volumen, la Variedad y Velocidad, conforman lo que se conoce como “Las Tres V de Big Data”.

Entrenamiento y Pruebas

Es parte del *Machine Learning*. El entrenamiento es cuando se le suministran datos de entrenamiento al dispositivo para que sea construido el entendimiento inicial. Las pruebas vienen luego del entrenamiento, cuando se pasa el modelo a un conjunto de datos de prueba, en donde se aplica el entendimiento que se adquirió durante el entrenamiento para generar predicciones.

Estadística

Es un valor que se calcula o infiere a través del análisis de un dato o grupo de datos. Un ejemplo puede ser el promedio de notas de un salón de clases.

Estadísticas

Grupo de datos, herramientas y métodos utilizados para analizar otros datos.

Front End

El Front End es la interfaz con la que el usuario tiene interacción. Este, sirve de canal para que el usuario se comunique con el Back End.

Lógica Difusa

Es una abstracción de la lógica booleana, en la cual los valores de las variables son cualquier numero real entre 1 y 0 incluyéndolos. En este tipo de lógica existen afirmaciones “medio ciertas” o “medio falsas”.

Machine Learning

Es el proceso con el cual una computadora obtiene conocimientos o inteligencia a través de un conjunto de datos, para luego hacer predicciones basada en su entendimiento.

Algoritmo

Un algoritmo es un conjunto de instrucciones bien definidas, ordenadas y finitas que permiten llevar a cabo una actividad mediante pasos sucesivos que no generen dudas a quien deba hacer dicha actividad.

Algoritmo Difuso

Algoritmos que implementan la lógica difusa para disminuir el tiempo de ejecución de un programa. Comparándolos con algoritmos que implementan la lógica booleana, son mas rápidos pero menos precisos.

Algoritmo Voraz

Un algoritmo voraz, es aquel que construye la solución de un problema pieza por pieza, escogiendo siempre la pieza que ofrezca la solución mas prometedora o que genere el resultado mas óptimo.

Almacén de Datos

Un almacén de datos (data warehouse) es un sistema utilizado para generar reportes y para el análisis de datos. Consisten en un repositorio central de datos integrados por una o mas fuentes y almacenan datos actuales e históricos, los cuales son analizados y luego utilizados para generar reportes e informes.

Back End

Back End es la parte “oculta” de un sistema. Es la que se encarga de que la parte “visible” (front end) funcione de manera adecuada. Si tomamos de ejemplo una pagina web, el back end podría ser la base de datos, el sistema de autenticación, los servidores, etc.

Base de Datos

Visto de la forma mas sencilla posible, una base de datos es un espacio utilizado para el almacenamiento de datos. Para manejar las bases de datos se utilizan sistemas de manejo de bases de datos (DBMS por sus siglas en ingles) como MySQL o PostgreSQL.

Overfitting

Overfitting es el efecto de proveer demasiada información sobre un modelo con datos de los cuales ya se conoce el resultado deseado. Los algoritmos de *Machine Learning* deben alcanzar un estado en el cual puedan predecir el

resultado en otros casos a partir de lo aprendido con los datos de entrenamiento, generalizando para poder resolver situaciones distintas de las proporcionadas con los datos de entrenamiento. Cuando ocurre el *Overfitting*, el sistema queda ajustado con unas características muy especificas que evitan que ocurra la generalización necesaria para poder actuar en situaciones nuevas.

Regresión

Es cuando un valor cambia a consecuencia del cambio de otro valor dentro de un grupo de datos. Generalmente ocurre con variables continuas, como cuando el kilometraje y el año de un vehículo afectan su precio de venta.

Underfitting

Es cuando no se suministra suficiente información a un modelo que está siendo entrenado, y por lo tanto el modelo no es preciso y no logra generar las predicciones.

Campos de Enfoque de la Ciencia de Datos

A continuación, algunos de los campos de especialización de la ciencia de datos:

Análisis de Datos

Como su nombre lo dice, se enfoca en analizar datos para obtener respuestas sobre el pasado y el presente de un conjunto de datos. Utiliza estadísticas poco complejas y generalmente intenta identificar patrones que puedan ser mejorados.

Análisis Cuantitativo

Este campo se enfoca en la utilización de algoritmos para ganar beneficios en el sector financiero. Estos algoritmos hacen recomendaciones de inversiones basados en grandes cantidades de datos.

Ingeniería de Datos

Consiste en el Back End. Está enfocado en construir y mejorar sistemas para que los analistas de datos hagan su trabajo de forma mas sencilla. En términos sencillos, un analista de datos también puede ser un ingeniero. En grupos grandes, los ingenieros tienen permitido enfocarse únicamente en la aceleración de los análisis y en almacenar los datos de forma organizada y que el acceso a ellos sea sencillo

Inteligencia Artificial (IA o AI por sus siglas en ingles)

Es el campo que busca el desarrollo de máquinas que posean inteligencia similar a la de los humanos. La mayoría del trabajo en el campo de la IA se enfoca en el desarrollo de máquinas que resolver problemas y completar tareas de la misma forma en la que lo haría un humano.

Inteligencia Empresarial (BI por sus siglas en ingles)

Análisis de datos enfocado en medidas y estadísticas de negocios. Involucra el aprendizaje de técnicas para implementar software en el análisis de los datos para generar reportes de manera efectiva y encontrar tendencias importantes.

Periodismo de Datos

Es una especialidad del periodismo que consiste en recabar y analizar grandes cantidades de datos mediante software especializado y hacer comprensible la información a la audiencia a través de artículos, infografías, reportes, historias, entre otros.

Visualización de Datos

Consiste en buscar, interpretar, contrastar y comparar datos para conocer de forma detallada los mismos y luego transformarlos de manera que sean comprensibles para usuarios. Generalmente se implementan graficos, infografías, tablas de datos, entre otros.

Herramientas de Estadísticas

A continuación algunas de las herramientas mas utilizadas por los profesionales en el campo de las estadísticas.

Correlación

Mire la fuerza y dirección de una relación lineal y la proporcionalidad entre dos variables estadísticas distintas. Si ambas variables aumentan juntas, están correlacionadas de forma positiva. Si una crece y la otra disminuye, están correlacionados de forma negativa. No existe correlación si el cambio en una de las variables no tiene conexión alguna con el cambio en la otra variable.

Desviación Típica

Es una medida que se utiliza para cuantificar la variación de un conjunto de datos.

Error

El error es la diferencia entre el valor real de un dato y la estimación que se hizo del valor que tendría dicho dato.

Importancia estadística

Un resultado es estadisticamente importante o significativo cuando se concluye que no ocurrió por casualidad.

Mediana

En una lista de valores ordenados, la mediana es el valor que se encuentra en la posición central de la lista.

Muestra

Es el conjunto de datos al cual se tiene acceso y será utilizado en el estudio.

Normalización

Consiste en ajustar valores medidos en escalas distintas a una escala común, antes de utilizar los datos.

Promedio o Media

Es un calculo que nos da una noción del valor mas común en un grupo de números. Es la suma de los números de una lista dividida entre la cantidad de valores en dicha lista.

Resumen Estadístico

Reúne y resume la información resaltante de la muestra de datos analizada. Es utilizado a la hora de comunicar la información y las conclusiones obtenidas del análisis realizado.

Series de Tiempo

Una serie de tiempo es un conjunto de datos medidos en determinados momentos y ordenados de forma cronológica.

Valor Atípico

Es un punto en el conjunto de datos que esta numéricamente distante del resto de los valores. Generalmente son producto de casos excepcionales o errores en las mediciones.

Varianza

La varianza, mide que tan dispersos están los valores de un conjunto de datos. Matemáticamente, es el promedio de la diferencia que hay entre valores individuales y el promedio del conjunto de valores. La raíz cuadrada de la varianza de un conjunto, nos da como resultado la desviación típica.

Flujo de Trabajo

Para poder extraer la información útil de un conjunto de datos es necesario procesarlos de forma adecuada. A continuación algunos de los procesos, procedimientos y elementos mas importantes con los que se obtiene información en el área de la ciencia de datos.

Data Pipelines

Una colección de scripts o funciones a través de la cual pasan los datos de forma ordenada. La salida del primer método se convierte en la entrada del segundo. Esto continúa hasta que los datos se limpian y transforman adecuadamente para cualquier tarea en la que esté trabajando un equipo.

Data Wrangling (Discusión de datos)

Consiste en el mapeo, procesamiento y conversión de datos en su forma mas básica a un formato distinto con la intención de que sean mas legibles para ser analizados de forma mas sencilla.

ETL (Extract, Transform, Load)

Este proceso es clave para los almacenes de datos. Describe las tres etapas de extraer datos de numerosos lugares, sin formato, procesarlos o transformarlos para que puedan ser utilizados, analizados y estudiados, y para luego ser cargados en bases de datos o almacenes de datos.

Exploración de Datos

La parte inicial del estudio de datos en el cual se hacen preguntas básicas que ayudarán a entender el contexto de un conjunto de datos. Lo que se aprende durante la fase de exploración conducirá a un análisis de mayor profundidad más adelante.

Minería de Datos

Consiste en extraer información procesable de un conjunto de datos para darles uso adecuado. Esto incluye todo desde la limpieza y organización de los datos hasta el análisis para hallar patrones y conexiones significativas.

Web Scraping

Se trata del proceso de extracción de datos del de uno o varios sitios web. Por lo general, se implementa el uso de bots que identifican la información que se necesita, la extrae y la lleva a un archivo para su posterior análisis.

Técnicas de Machine Learning

Con el pasar del tiempo, el *Machine Learning* ha evolucionado de formas muy distintas. Esto ha traído como consecuencia que se hayan desarrollado distintas técnicas o ramificaciones de este campo. A continuación los distintos métodos y técnicas que son utilizados en la actualidad.

Clustering

Consiste en recopilar y agrupar conjuntos de puntos que son “suficientemente similares” o “cercanos” entre sí. “Cercano” varía según la forma con la que se elija medir la distancia.

Deep Learning

Los modelos de deep learning implementan estructuras lógicas que asemejan la organización del sistema nervioso de los seres vivos, teniendo capas de unidades de proceso que se especializan en detectar determinadas características existentes en los objetos, para resolver problemas complejos, como el reconocimiento facial. Las capas en un modelo comienzan con la identificación de patrones muy simples y luego crecen en complejidad. Al final, dichas estructuras tienen una comprensión matizada que puede clasificar o predecir valores con precisión.

Ingeniería de Características

El proceso de tomar conocimiento que tenemos como seres humanos y traducirlo en un valor cuantitativo que una computadora puede entender. Por ejemplo, podemos traducir nuestra comprensión visual de la imagen de una taza en una representación de intensidades de píxeles.

Redes Neuronales

Un método de aprendizaje automático inspirado en las conexiones neuronales del cerebro. Consiste en un conjunto de unidades, llamadas neuronas artificiales, conectadas entre sí para transmitirse información. La entrada atraviesa la red neuronal (es interpretada y procesada) produciendo distintos valores de salida. Tiene como objetivo el resolver problemas de la misma forma en la que lo haría el cerebro humano.

Machine Learning Supervisado

Es una técnica con la que se le enseña un conocimiento (modelo) a una computadora a partir de datos de entrenamiento, los cuales consisten en pares de objetos (datos de entrada y resultados deseados). La salida de la función puede ser un valor numérico o una etiqueta de clase. El objetivo del Machine Learning supervisado es el de crear una función capaz de predecir

el valor correspondiente a cualquier objeto de entrada válido luego de haber procesado los datos de entrenamiento y aprendido.

Machine Learning no Supervisado

En técnicas de aprendizaje no supervisadas, la computadora construye su propia comprensión de un conjunto de datos sin etiquetar. Las técnicas de Machine Learning no supervisadas buscan patrones dentro de los datos y, a menudo, lidian con la clasificación de elementos basados en rasgos compartidos.