Biga Data, Business Intelligence, Análisis de Datos (Data Analytics), Ciencia de Datos (Data Science), etc.

Clarificando conceptos, poniendo orden

Jose Ignacio González Gómez . Departamento de Economía, Contabilidad y Finanzas - Universidad de La Laguna

Índice

1. Presentación

El objetivo de estas notas es poner en orden las palabras clave, etiquetas y conceptos que nos invaden cuando intentamos acercarnos a entender Big Data, Inteligencia Artificial, Machine Learning, etc..

Este mundo es amplio, multidisciplinar y en continua evolución cuyas primeras notas a tener en cuenta son:

- La Ciencia de Datos es una carrera de fondo, imposible de dominar en 1, 3 ó 6 meses.
- Hay una cantidad enorme de documentación, cursos, etc que no nos debe abrumar
- Debemos entender lo mínimo necesario para continuar implementando en nuestro primer contacto y después iremos profundizando, es decir ir de menos a mas.
- Es un maratón donde el aprendizaje se adquiere día a día. No es necesario saberlo todo
- Nos encontraremos con dificultades y frustraciones a lo largo del camino. La persistencia es clave.

Esquema… línea argumental

	Empresas	Asesorías - Consultoras - Retail - Hoteles - PYMES….	Hospitales- Banca - TITSA- Binter - DISA - Amazon - AEAT
Pefiles, habilidades y técnicas			
	Volumen de datos	**DATOS**	**BIGDATA**
	Perfiles Profesionales	Economistas-ADE-Contable-Ofimáticos-Administrativos-Autodidactas……	→ + **Multidisciplinar** Ingenieros-Mamáticos-Físicos-Estadísticos-Autodidactas
	Tipo y técnicas de análisis	**Análisis Datos (Data Analitic)** ***Descriptivo:*** ¿Qué ha pasado? *Técnicas*: Resumen y visualizaciones de datos. Tablas, gráficos, estadística descriptiva ***Diagnóstico:*** ¿Por qué ha ocurrido? *Técnicas*: Inferencia estadística. Técnicas de asociacion-correlación, comparativas de muestras, modelos estadísticos o causales	→ + **Ciencia de Datos (Data Science)** ***Predictivo*** **:** ¿Qué es probable que ocurra en el futuro? *Técnicas*: Modelos estadisticos de probabilidad, marchine learning,… ***Prescriptivo***: ¿Cuál es la mejor estrategia a seguir? *Técnicas*: Tácnicas avanzadas como análisis de grafos, redes neronales, sistemas de recomendación…

2 Concepto de Big Data

2.1 Esquema general

Big Data son soluciones basadas en el manejo y procesado de volúmenes muy grandes de datos, tan grandes que tienen que repartirse entre varios ordenadores.

Estos datos se generan de forma continua y se tienen que poder tratar o almacenar a una velocidad alta y, por si fuera poco, son diversos. Algunos son registros con información estructurada, archivos de correo electrónico, imágenes, vídeos…son datos variados.

Fuente: La guía ninja del Big Data y la Inteligencia Artificial

2.2 Las 3Vs del Big Data

las características del Big Data son Volumen, Velocidad y Variedad, también conocidas como las 3 Vs de Big Data:

- **Volumen**: el volumen de los datos que tenemos es lo suficientemente grande como para que no podamos almacenarlos en una sola máquina ¡Y además no paran de crecer!
- **Velocidad**: El procesamiento y el análisis de los datos se tiene que poder hacer en un tiempo razonable. Si Google tardara media hora en ofrecerme el resultado en el buscador, cambiaría de buscador
- **Variedad**: Los datos pueden ser de todo tipo.

3 Ordenando conceptos: Analítica de Datos, Inteligencia de Negocio y Ciencia de Datos

3.1 Los cuatro tipos de analisis de datos: Descriptivo, Diagnóstico, Predictivo y Prescriptivo

Análisis de Datos (Data Analytics)	• ***Descriptivo "¿Qué ha pasado?":*** Nos cuenta qué ha pasado hasta el momento y nos describe la situación que tenemos en el presente. Da respuesta a: "¿Qué ha pasado?". **Técnicas.** Visualizaciones de datos resumen como tablas y graficos, estadística descriptiva, etc. • ***Diagnóstico "¿Por qué ha ocurrido?":*** Una vez que sabemos lo que ha ocurrido, querremos saber por qué ha ocurrido, entender por qué se está desarrollando una tendencia o por qué se ha producido un problema. Hay que evitar conjeturas inexactas y no confundir correlación y causalidad. **Técnicas.** Inferencia estadística.
Ciencia de Datos (Data Science)	• ***Predictivo "¿Qué es probable que ocurra en el futuro?***": El análisis predictivo examina los patrones de datos actuales e históricos para determinar si es probable que esos patrones vuelvan a surgir. Daremos respuesta a preguntas : ¿Qué es probable que ocurra en el futuro? **Técnicas**. Se fundamenta modelos estadísticos de probabilidad, técnicas de modelado y machine learning para obtener la probabilidad de que se cumpla o no una hipótesis. • ***Prescriptivo "¿Cuál es la mejor estrategia a seguir?***". La analítica prescriptiva tiene en cuenta específicamente la información sobre posibles situaciones o escenarios, los recursos disponibles, el rendimiento pasado y el rendimiento actual, y sugiere una estrategia operativa. Nos dice que tiene que pasar para que consigamos los resultados que queremos. Trata de responder a la pregunta "¿Cuál es la mejor estrategia a seguir?" **Técnicas.** Utiliza técnicas avanzadas como análisis de grafos y modelos de machine learning basados en redes neuronales, sistemas de recomendación, etc..

¡Resumiendo!, contamos con cuatro tipos de análisis de datos:

1. *Descriptivo: "¿Qué ha pasado?"*
2. *Diagnóstico: "¿Por qué ha ocurrido?"*
3. *Predictivo: "¿Qué es probable que ocurra en el futuro?"*
4. *Prescriptivo: "¿Cuál es la mejor estrategia para seguir?"*

3.2 Analítica de Datos (Data Analytics)

Como hemos podido ver dentro del mundo del Big Data se distinguen dos grandes ramas Analítica de Datos y Ciencia de Datos, pero no es necesario estar en un ecosistema de Big Data para realizar análisis de datos.

Para aplicar Analítica de Datos (Data Analytics) no necesitamos muchísimos datos, simplemente necesitamos datos, sin embargo, la posibilidad de disponer de muchos más datos y ser capaces de procesarlos vitamina ese análisis, resultando en una decisión más informada sobre cómo solucionar un problema concreto.

En concreto, la analítica de datos tiene como propósito principal extraer, procesar, agrupar y analizar datos masivos de una fuente específica y, a partir de ellos, generar informes con soluciones para poder sacar conclusiones y optimizar la toma de decisiones de negocio.

Quienes se desempeñen como profesionales de este sector deben saber recopilar datos y analizarlos de forma estadística con facilidad para ofrecer soluciones o asesoría de negocio a partir de los patrones y tendencias identificadas en el comportamiento de los datos.

3.3 Inteligencia de Negocio (Business Intelligence)

Otro campo que suele mezclarse en el ecosistema de Big Data es la Inteligencia de Negocio (Business Intelligence), en este caso, se utilizan los datos para conocer el estado actual de una empresa y asistir en la toma de decisiones estratégicas. Es una manera de asistir a los responsables de la toma de decisiones a la hora de realizar su trabajo, pero no predice que va a suceder, ni indica la decisión a tomar.

Asi la Inteligencia de Negocio es descriptiva y por ello sus herramientas están muy relacionadas con la visualización de los datos.

3.3 Ciencia de Datos (Data Science)

La **Ciencia de Datos** (**Data Science**) es un área multidisciplinar que, a través de diferentes campos como la estadística, informática y matemáticas, extrae de diferentes fuentes una gran cantidad de datos con el objetivo de predecir acciones basándose en patrones del comportamiento de los datos en el pasado, es decir revela las tendencias y métricas.

El ámbito de aplicación de la ciencia de datos es muy amplio, como, por ejemplo:

- **Ciberseguridad:** Con algoritmos es posible detectar comportamientos extraños que, al romper el patrón habitual, permiten identificar posibles amenazas de forma automatizada.
- **Finanzas:** Sucede como en el caso anterior, con los patrones de comportamiento de los usuarios y el día que ocurre algo inesperado, saltan las alarmas por ejemplo cuando te llaman del banco porque has intentado hacer una transacción que no es común para ti, una compra por internet en otro país, el pago de una cantidad muy alta, o cualquier cosa que no hagas con frecuencia. Estos son mecanismos antifraude.
- **Marketing:** Este es el campo que la mayoría de las personas asocian con la Ciencia de Datos. Actualmente es posible clasificar clientes con "lead scoring", una metodología que asigna puntos a los potenciales clientes de una empresa a partir de su comportamiento, características y datos, para identificar quiénes tienen más probabilidad de comprar.
- Lo mismo sucede con la **venta cruzada**. A través de la interpretación de los datos, es posible desarrollar un sistema que, basado en los intereses del usuario, recomiende otros productos que pueda interesarle comprar y, de esta forma, incrementar el ticket medio de compra.

En inteligencia artificial y aprendizaje automático, el científico de datos tiene un gran papel que desempeñar. Para el científico de datos, el conocimiento del aprendizaje automático es imprescindible. El machine learning es el desarrollo más impresionante en el mundo de la tecnología. Él requiere saber qué método de aprendizaje automático lo ayudará exactamente.

4 Diferencia entre Analítica de Datos (Data Analytics) vs Ciencia de Datos (Data Science). Dos perfiles profesionales

4.1 Diferencias conceptuales

- En Data Science se proponen preguntas y en Data Analytics se formulan respuestas.
- En Data Science se convierten los datos en información y en Data Analytics se convierten los datos en insights de negocio (conocimiento que aporta valor para la mejora del negocio).
- En Data Science se recaba la información desde diferentes fuentes, en Data Analytics desde una sola, en general.
- En Data Science se investigan soluciones y se crean estrategias para lo que está por venir y en Data Analytics se buscan soluciones a problemas ya detectados a través de datos y variables conocidas.

4.2 Diferencias según sus labores profesionales

Responsabilidades de un Analista de Datos (Data Analysts)

Identifique cualquier problema de calidad de datos, es decir, asegurar la calidad y fiabilidad de los datos
Documentar los tipos y la estructura de los datos comerciales, financieros, de producción, etc...
Cálculo de métricas e indicadores y análisis estadístico de los datos.
Analizar la información que contienen los datos para la obtención de insights empresariales. Resolver problemas de negocio e ineficiencias en la actividad empresarial. Dar respuesta a preguntas específicas.

Responsabilidades de un Científico de Datos (Data Scientists)

Tratamiento, transformación y limpieza de datos.
Análisis predictivo / Forecasting. Predicción del problema de negocios.
Desarrollo de algoritmos y de métodos analíticos. Desarrollo de modelos de machine learning y deep learning
Data Mining (minería de datos) utilizando métodos de última generación.
Presentar resultados de manera clara y hacer el análisis ad-hoc.

5 El Analista de Datos (Data Analyst)

5.1 Profesionales mixtos, híbridos o balanceados

- Un Analista de Datos o Data Analyst es la persona que analiza e interpreta los datos y los convierte en información relevante. Casi el 90% de los analistas de datos se centran en el Analisis Descriptivo
- En este sector se buscan perfiles balanceados o perfil mixto, es decir, multidisciplinar de áreas técnica que posean no solo habilidades duras (ligadas a los conocimientos teóricos) sino ademas que poseen conocimientos propios de las humanidades y de las ciencias sociales, especialmente habilidades blandas como buena comunicación, organización, empatía, trabajo en equipo, etc.., estas habilidades, se consideran, que complementan a la perfección los perfiles técnicos.
- Estos nuevos profesionales combinan conocimientos de programación con habilidades offline, como ventas, análisis, diseño, marketing, entre otras. Uno de los elementos diferenciadores de estos perfiles es que su formación tampoco es tradicional. No solo están formados en universidades, sino además en formación online como Udemy y otros y con mucho de autoaprendizaje y en programas part-time que permite compaginar su trabajo actual con el aprendizaje, a la vez que adquirieren nuevas habilidades tecnológicas en un entorno muy dinámico y orientado a crear productos digitales desde el primer día

5.2 Por dónde empezar

Para los que quieren convertirse en un Analista de Datos (Data Analyst) es importante, como hemos comentado anteriormente, considerar que esta es una carrera de fondo, imposible de dominar en 6 meses, pero en la que podemos distinguir una serie de fases en el proceso de formación que dependerá su temporalidad de la dedicación y conocimientos previos.

5.2.1 Fase I: Análisis Descriptivo y de Diagnostico (de 0 a 12 meses)

- Nuestro proceso de estudio debe comenzar con el Analisis Descriptivo, que incluye un conocimiento de Excel de tipo medio, trabajo con tablas dinámicas, funciones condicionales como SUMAR SI, CONTAR SI, etc..
- Este conocimiento instrumental debe combinarlo con un conocimiento de estadística descriptiva como son las medidas de centralización (media, mediana, moda, etc) y las medidas dispersión (varianza, desviación típica, etc..) y graficos relacionados como de frecuencia, etc. Este conocimiento básico de estadística es fundamental para una correcta interpretación de los datos y las conclusiones a las que lleguemos sean solidas desde el punto de vista estadístico.
- En esta línea es también importante el conocimiento del proceso de ETL con un editor de consultas como Power Query asi como el modelado de datos con Power Pivot y la creacion de medidas basicas con DAX.
- Finalmente, y para las visualizaciones es recomendable trabajar con Power BI que nos permitirá presentar nuestros trabajos de forma extraordinariamente visual.

Este conocimiento es el más requerido en las Pymes, por tanto, es muy importante consolidar el conocimiento de esta fase y poco a poco afrontar el resto del proceso de formación.

Casi el 90% de los analistas de datos que existen en el mercado se centran en el Analisis Descriptivo y poco a poco se va pidiendo competencias de analisis predictivo

5.2.2 Fase II: Análisis Predictivo (de 12 a 24 meses)

- Una vez hemos afianzado el conocimiento en el analisis descriptivo podemos empezar a diferenciarnos del resto de analista de datos con el analisis predictivo que implica usar software estadístico especializado como SPSS, Stat y/o introducirnos en un lenguaje de programación como R o Python.
- En esta fase tendremos que entrar con Estadística Inferencial que se encarga de hacer deducciones, es decir, inferir propiedades, conclusiones y tendencias, a partir de una muestra del conjunto, y donde calculo probabilístico es fundamental. Su papel es interpretar, hacer proyecciones y comparaciones.
- Las técnicas estadísticas de inferencia son por ejemplo, diferencias de la muestra y la población, como seleccionar muestras de manera correcta, entender cómo se hacen contrastes de hipótesis, entender las distribuciones de probabilidad, etc.,

5.2.3 Fase III: Análisis Prospectivo (de 24 meses a…)

Aquí entramos con un analisis más avanzado con los modelos de machine learning y forcasting es decir de predicción o estimación y el análisis de demanda futura mediante algoritmos que analizan muchas variables que influyen, como son los históricos de venta, estimaciones de marketing, promociones, campañas, estudios de mercado, cuadros de mando, etc… Donde las técnicas de series temporales es la base.

6 Técnicas estadísticas y desarrollo de algoritmos y de métodos analíticos en el Análisis de Datos y Ciencia de Datos

Las técnicas estadísticas y el desarrollo de algoritmos y métodos analíticos (modelos de machine learning y deep learning) las podemos agrupar en las siguientes categorías.

***1. Técnicas de descripción y exploración de los datos*.**

Trata de traducir el conjunto de datos en tablas y gráficos para facilitar su compresión.

2. Técnicas de asociación y correlación

Se trata de ver si dos características están relacionadas de alguna forma.

3. Técnicas comparativas, analisis de la varianza Anova y otras

Trata de comparar grupos, por ejemplo, comparar un grupo control de uno de investigación. comparar países por continentes, comparar ventas por productos, etc.
Se trata de ver si los grupos son diferentes o no y qué diferencias hay. Una de las técnicas más utilizadas son las tablas ANOVA.

4. Modelos estadísticos o modelos causales

Con las técnicas de asociación y correlación, tratamos de analizar si dos características están relacionadas de alguna forma, aunque ***correlación no implica causalidad***. Con modelos estadísticos o causales ***vamos a modelar las relaciones causa-efecto***.

En realidad, las técnicas de asociación y comparación complementan los modelos estadísticos que son los más potentes para sacar jugo a los datos. Algunos modelos estadísticos: Regresiones logísticas, Regresión lineal, Modelos lineales generalizados, etc…

En las técnicas del punto 5, 6 y 7 vamos a dar un salto en la complejidad y el valor que podemos extraer de nuestros datos.

5. Técnicas de reconocimiento de patrones o machine learning I. Segmentación de datos o clustering

Consiste en crear grupos de clientes similares, pacientes similares, etc… similares a unas determinadas características, eso es la segmentación.
Para ello se dispone de diferentes técnicas como clustering, k-means, etc...

6. Técnicas de reconocimiento de patrones o machine learning II. Reducción dimensional

En muchas ocasiones el número de variables a tener en cuenta hace que el problema sea complicado de abordar y para ello contamos con técnicas para reducir el número de variables e incluso es posible usar técnicas para poder representar en dos o tres dimensiones de muchas variables a la vez.

7. Técnicas de reconocimiento de patrones o machine learning III. Algoritmos y modelos predictivos

El último grupo de técnicas a utilizar son los algoritmos o modelos predictivos que nos permitan realizar proyecciones, predecir o estimar. En el fondo, se trata de saber qué va a ocurrir con nuevos datos.
Estas técnicas se basan en aprender de datos pasados para poder estimar lo que va a ocurrir con nuevos datos. Por ejemplo:

- Estimar qué probabilidades tiene un paciente sobreviva a una determinada enfermedad.
- Estimar las ventas del próximo mes, etc..

Estas técnicas nos permitirán realizar los analisis técnicos de los datos, pero lo más importante es interpretar esos analisis, esta es la parte esencial.