185 ESTADÍSTICA: ¿EL CORAZÓN DE BIG DATA?

Batista Einer
Facultad de Ciencias Económicas, Jurídicas y Sociales. Universidad Nacional de Salta
einerbatista@hotmail.com
Especialidad: Educación Matemática
Palabras Claves: Big Data, Modelización, Regresión, Aprendizaje Supervisado.

Resumen

En estos años ha comenzado a hablarse de Big Data gracias a los avances tecnológicos que permiten procesar grandes volúmenes de datos a un muy bajo costo. En dicho proceso se utilizan herramientas estadísticas que existen y son estudiadas desde hace muchísimos años pero que han tomado una gran revitalización gracias a la democratización de Big Data en estos últimos 10 años. Se define Big Data como todas las herramientas y procesos para gestionar grandes volúmenes de datos que permitan agregar valor para mejorar la toma de decisiones. Esta disciplina ya es utilizada desde hace tiempo en las grandes organizaciones y se está comenzando a implementar en las más pequeñas que en el pasado no tenían los recursos económicos ni los conocimientos técnicos pero que gracias al progreso tecnológico y a la disminución de costos hoy ya pueden acceder a ellos. En este trabajo se expone el proceso real de implementación de Big Data en una compañía de retail y la influencia de las herramientas estadísticas en todo dicho proceso que abarca desde la etapa inicial del conocimiento del negocio, el tratamiento de los datos, la modelización y la consecuente presentación de los resultados y posterior seguimiento del modelo. La finalidad del mismo es generar un marco de discusión sobre el enfoque de la enseñanza actual de la estadística en las Facultades de Ciencias Económicas, principalmente en la parte práctica de la materia y su posible reorientación como consecuencia de los avances tecnológicos que formará el entorno en donde trabajará el futuro profesional de Ciencias Económicas.

1. Introducción

Hoy en día el 80 % de los datos en el mundo (texto, imagen, video, sonido, etc.) son desestructurados o sea que no tienen forma y no están ordenados bajo ninguna estructura pero que por si solos o mediante relaciones entre ellos pueden generar información útil para la toma de decisiones.

Ejemplos de la implementación de Big Data en distintas industrias o procesos:

- Comercio minorista: se puede determinar la variedad y cantidad necesaria de mercadería a reponer, predecir las ventas futuras anticipándose a la demanda, evitar los quiebres de stock, personalizar la oferta en función del conocimiento del cliente, etc.
- Industria Bancaria: establecer planes de acciones para fidelizar a los clientes que tienen renovaciones de productos anuales y evitar su baja, establecer mediante regresión logística si se otorga un crédito a un posible cliente de determinadas características, etc.
- Seguros: establecer patrones de fraudes disminuyendo los riesgos y bajando la siniestralidad.
- Salud: gestionar grandes volúmenes de datos permite acelerar los tiempos en las investigaciones científicas para el desarrollo de nuevos medicamentos, permite la generación de sistemas de alerta inteligentes para el diagnóstico temprano de enfermedades, el pronóstico de su evolución y la planificación de los tratamientos. Posibilita la segmentación de pacientes crónicos y la toma de decisiones proactivas minimizando el gasto sanitario.
- Economía: se puede predecir acontecimientos económicos como recesión, inflación, pobreza o variaciones en tasas de interés y anticiparse en la toma de decisiones de parte de los gobiernos.
- Campañas Electorales: segmentar y profundizar las propuestas electorales en función del perfil ideológico del votante principalmente en aquellos indecisos. El mejor ejemplo de esto se produjo en el referéndum por el Brexit en Gran Bretaña

o en la campaña presidencial en Estados Unidos en 2016 donde intervino Cambridge Analytica con unos 5.000 datos por cada votante aplicando segmentación publicitaria, encontrando preferencias y probabilidad de realizar el voto, con la posterior determinación de los electores a favor de las armas o en contra del aborto y la consecuente campaña focalizada para dichos grupos.

2. Características de Big Data

En Big Data se habla de las 3 V cuando se refiere a:

- Procesar una gran cantidad de información (Volumen).
- Procesar distintos tipos de información (Variedad).
- Realizarlo en un periodo razonable de tiempo (Velocidad).

Cuando se ejecuta un proceso de Big Data intervienen distintas disciplinas:

- Desarrollo de Sistemas y Tratamiento de Base de Datos.
- Infraestructura y Tecnología (IT).
- Aprendizaje Automático (Machine Learning) e Inteligencia Artificial (IA).
- Ciencias de la Comunicación y Diseño Gráfico (Infografías y Visualizaciones de Datos).
- Y por supuesto, Estadística.

Etapas del Proceso de Big Data:

1) Conocimiento del Negocio: se determina el problema y el objetivo.
2) Comprensión de Datos: que datos son necesarios para resolver el problema.
3) Plataforma Tecnológica: que infraestructura y tecnología es necesaria.
4) Tratamiento de Datos: como se procesan los datos.
5) Modelización: se crean modelos que permitan sacarle valor a los datos.
6) Presentación de Resultados: comunicar el conocimiento obtenido.
7) Despliegue: desplegar en la arquitectura el modelo construido.
8) Puesta en Valor: integrar e implementar el modelo construido.
9) Seguimiento: controlar la evolución del modelo

3. La influencia de Estadística en un Proceso de Big Data. Fundamentación.

3.1 Conocimiento del Negocio:

En esta etapa se identifica un problema de negocio y se lo transforma en un problema analítico mediante un análisis integral:

- Análisis Descriptivo: Mostrar mediante estadísticos la realidad capturada.
- Análisis Inferencial: Generalizar conclusiones muéstrales a toda la población, estudiar las relaciones entre las variables y contrastar hipótesis.
- Análisis Predictivo: Determinar datos futuros a través de datos históricos.
- Análisis Prescriptivo: Recomendar la acción adecuada y sus consecuencias.

En dicha etapa se identifican preguntas como:

- ¿Porque tengo quiebre de stock en un determinado rubro?
- ¿Qué cantidad y variedad recomendada de un artículo debo comprar para satisfacer las ventas?
- ¿Cómo disminuyo las bajas de los servicios anuales que brindamos a los clientes?
- ¿Cómo ajusto una campaña publicitaria a un determinado barrio de una ciudad?
- ¿Cómo minimizo la posibilidad de un fraude futuro al otorgar una póliza de seguro?

3.2 Comprensión de Datos:

En esta etapa se identifican las distintas fuentes de información con las que se va a trabajar en el proceso. Dentro de las fuentes de información se agrupan:

- Fuentes Internas.
- Fuentes Externas.
- Redes Sociales.
- Datos disponibles en forma libre (Ej. Indec).

Luego se relaciona la información en función del problema a resolver buscando conectores.

3.3 Plataforma Tecnológica:

En esta etapa se define que plataforma tecnológica se va a usar para la construcción del modelo. Desde la captura, almacenamiento y procesamiento de datos hasta la explotación del modelo definido. Hay que estar actualizado ya que los componentes de Big Data evolucionan constantemente.

En la etapa de explotación de datos se emplean reportes, herramientas de análisis estadísticos y de visualización de datos.

3.4 Tratamiento de Datos:

Se realiza una preparación de datos mediante:

- Adquisición y Registro de los Datos desde la fuente hasta llegar a un Almacén de Datos.
- Se construye un Metadatado (datos de datos).
- Construcción y transformación de variables (Ej. transformar una fecha en mes o día de semana).
- Luego se realiza una Exploración y Análisis de Variables para comprender mejor los datos mediante Histogramas, Indicadores Estadísticos, y Visualización de Nulos.
- Se realiza una limpieza de los datos para mantener la calidad (Ej. tratar campos vacíos)

3.5 Modelización:

En esta etapa se realiza la construcción del modelo analítico utilizando algunas técnicas estadísticas.

Dichas técnicas se agrupan de la siguiente manera:

- Aprendizaje Supervisado: se aplican técnicas con un conocimiento a priori o sea con datos de entrenamiento.
- Aprendizaje no Supervisado: se aplican técnicas sin conocimiento a priori.

Tabla 1. Técnicas utilizadas para la modelización

Aprendizaje Supervisado	Clasificación (para variables categóricas)	Regresión Logística Arboles de Decisión. Redes Bayesianas SVM
	Regresión (para variables cuantitativas)	Regresión Lineal Arboles de Regresión Redes Neuronales SVM
	Recomendación	Varias.
Aprendizaje No Supervisado.	Clustering	Varias.

3.6 Presentación de Resultados:

En esta etapa se transmite los resultados de la etapa anterior a todos los sujetos relacionados. Para realizar esto se usan:

- Informes y Reportes: se usan reportes estadísticos, gráficos, diagramas de caja y sesgo, diagramas de dispersión, Histogramas, etc.
- Infografías: se comunica mediante una combinación de textos, gráficos, imágenes, símbolos, mapas, etc. Se expone de una manera más estática.
- Visualizaciones: similar a la infografía pero dinámica o sea con fines exploratorios donde el usuario mediante filtros puede interactuar o busca observar lo que le interese. Para realizarlos es indispensable el uso de software. Pueden mostrar: Tendencias, Patrones, Comparaciones, Anomalías, Conexiones, Correlaciones, Localizaciones, etc.
- Tableros de Mando: se pueden elegir las principales variables de interés según el usuario mostrándole información dinámica.

3.7 Despliegue:

En esta etapa se integrar en la plataforma tecnológica el modelo construido.

3.8 Puesta en Valor:

Se integra el modelo en las operaciones cotidianas de la organización.

3.9 Seguimiento:

Luego de la puesta en marcha se realiza un seguimiento a la misma evaluando la estabilidad de variables, estabilidad del modelo y su capacidad.

4. Aplicación Práctica de la Modelización. Desarrollo.

A continuación se describirán algunas técnicas estadísticas durante la Modelización dentro de la planificación de una implementación de un proceso de Big Data en una compañía de retail con 46 locales de venta de equipos celulares y los accesorios de estos.

4.1 Regresión Lineal Simple:

Permite construir un modelo (1) que predice la relación entre dos variables cuantitativas.

$$\widehat{y_i} = \beta_1 \, x_i + \beta_0 \qquad (1)$$

Variable Independiente x = Costo de Alquiler por Sucursal

Variable Dependiente y = Cantidad de equipos celulares vendidos por Sucursal

Tabla 2. Costo de Alquileres y Venta de Celulares por Sucursal

Sucursal	Alquiler (x)	Celulares (y)
1	215	578
2	185	562
3	45	292
4	75	236
5	65	125
6	123	351
7	369	767
8	238	889
9	172	801
Total	1.487	4.601

Se estima la recta de regresión con el siguiente modelo lineal (2) y se representa dicha relación y el modelo en el Gráfico 1:

$$\widehat{y_i} = 2{,}1694 \, x_i + 152{,}8 \qquad (2)$$

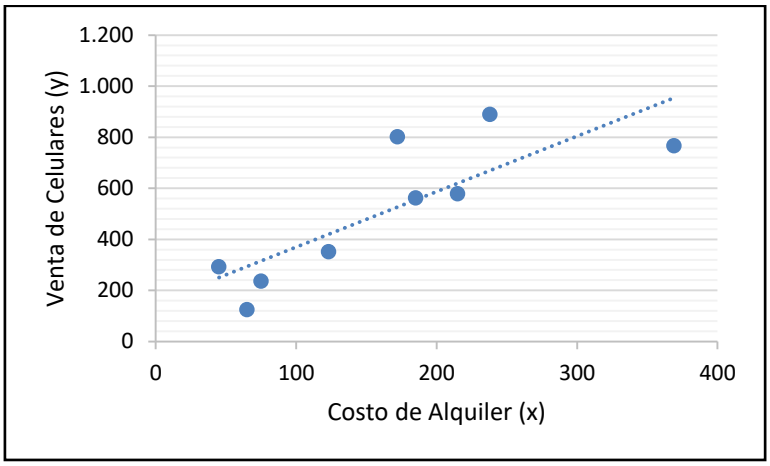

Gráfico 1. Relación entre el Costo de Alquiler y la Venta de Celulares por Sucursal.

Coeficiente de Determinación $R^2 = 0{,}6625$

Error Estándar de Estimación $S_{YX} = 169{,}7636$

4.2 Regresión Múltiple:

Similar a la Regresión Lineal Simple pero existen más variables independientes. En (1) se agrega una variable independiente adicional y se obtiene el siguiente modelo (3)

$$\widehat{y_i} = \beta_2 \, z_i + \beta_1 \, x_i + \beta_0 \qquad (3)$$

Dicha la variable independiente z se define como la cantidad de accesorios (de equipos celulares) vendidos por Sucursal.

Tabla 3. Costo de Alquileres, Venta de Accesorios y Venta de Celulares por Sucursal

Sucursal	Alquiler (x)	Accesorios (z)	Celulares (y)
1	215	3.086	578
2	185	3.232	562
3	45	2.543	292
4	75	1.730	236
5	65	2.310	125
6	123	4.636	351
7	369	3.919	767
8	238	5.444	889
9	172	3.090	801
Total	1487	29.990	4.601

Se obtiene el siguiente modelo (4) de regresión múltiple

$$\widehat{y_i} = 0{,}0646\, z_i + 1{,}7509\, x_i + 6{,}3399 \qquad (4)$$

Coeficiente de Determinación R^2 = 0,7139

Error Estándar residual: 168,8093

4.3 Regresión Logística:

Es una técnica estadística que predice el resultado de una variable categórica. La variable de respuesta es binaria, es decir puede admitir solamente 2 valores: 0 y 1. La especificación del modelo se observa en (5)

Se la emplea en los negocios para:

- Predecir quiebre de stock.
- Medir el riesgo de morosidad de un posible crédito a otorgar.
- Fidelizar a los clientes actuales para que lo sigan siendo en el futuro.

$$f(z) = \frac{1}{1+e^{-z}} \qquad (5)$$

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots\ldots + \beta_k x_k$$

Se sigue trabajando con el mismo ejemplo. En este caso se clasifica a las diferentes sucursales en función de si son rentables o no.

0: No Rentables (generan pérdidas).

1: Rentables (generan ganancias).

Tabla 4 – Costo de Alquileres, Venta de Accesorios y Rentabilidad por Sucursal

Sucursal	Alquiler (x_1)	Accesorios (x_2)	Rentables (y)
1	215	3.086	0
2	185	3.232	1
3	45	2.543	1
4	75	1.730	1
5	65	2.310	0
6	123	4.636	0
7	369	3.919	0

8	238	5.444	1
9	172	3.090	1
Total	1.487	29.990	1

Aplicando dicha técnica se obtiene el siguiente el modelo descripto en (6)

$$f(y) = \frac{1}{1+e^{-(1{,}026131-0{,}005964\,X_1\,+0{,}000056X_2)}} \qquad (6)$$

4.4 Arboles de Clasificación y de Regresión:

Mediante una combinación de las técnicas anteriores permite disminuir el error de estimación particionado el conjunto de datos. Tiene la forma de un esqueleto de un árbol invertido y tiene las siguientes características:

- En cada vértice hay una partición del conjunto de datos.
- Cada nodo terminal representa una parte del modelo.
- Para realizar una predicción hay que recorrer desde la raíz hasta el nodo terminal

En los árboles de clasificación se predicen variables categóricas (Ej. el precio de la acción sube o baja) y en los de regresión variables numéricas (Ej. a cuanto sube o baja el precio de la acción).

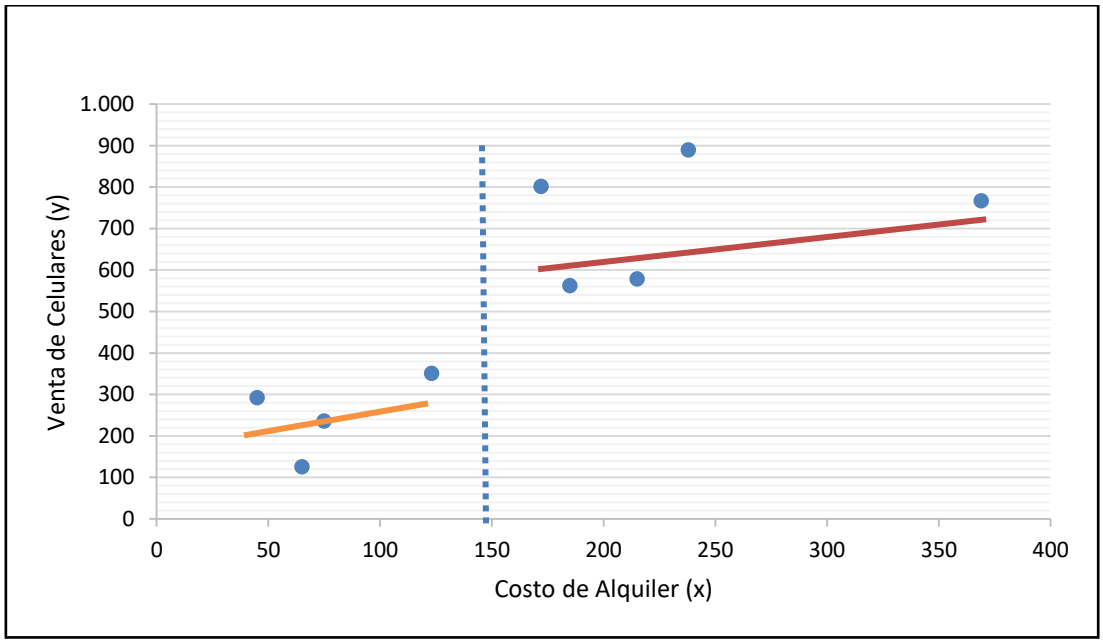

Gráfico 2 – Partición del conjunto de datos de la relación entre el Costo de Alquiler y la Venta de Celulares

En esta segmentación se construye un modelo específico para cada participación con el objetivo de representar mejor el conjunto de datos y disminuir el error de estimación

Para los alquileres menores a 150 (x_i < 150) obtenemos un R^2 = 0,2554 y la recta descripta en (7)

$$\hat{y_i} = 1{,}469\; x_i + 137{,}89 \qquad (7)$$

Para los alquileres mayores a 150 (x_i > 150) obtenemos un R^2 = 0,0756 y la recta descripta en (8)

$$\hat{y_i} = 0{,}5012\; x_i + 601{,}22 \qquad (8)$$

Como se mencionó antes, la principal finalidad de esta partición es disminuir el error de estimación por lo que si se calcula el Error Estándar de Estimación se obtiene $S_{YX} = 117{,}6601$ que comparado con el generado con la Regresion Lineal indica que mediante la aplicación de árboles se puede disminuir la incertidumbre al bajar el S_{YX}.

5. Conclusiones y Futuros Trabajos.

En este trabajo se ha intentado exponer la importancia de la Estadística en la implementación de un proceso de de Big Data en una organización y que gracias a los avances tecnológicos que se producen en una velocidad casi exponencial un proceso de Big Data en poco tiempo se estará ejecutando en casi todas las organizaciones.

Esto va a generar como efecto una revisión desde un punto de vista estructural y pedagógico de los conceptos básicos a transmitir al alumno durante el dictado de la materia no solo por los conocimientos que deberían adquirir los futuros profesionales de ciencias económicas que por su formación terminan ocupando puestos en los niveles más altos de dirección en las distintas organizaciones sino también por la relación que deberían tener estos profesionales con las nuevas carreras que se están creando como los científicos de datos que detectan patrones y analizan datos para maximizar su valor, los ingenieros de Big Data que desarrollan los software y los arquitectos de Big Data que diseñan y construyen las arquitecturas durante una implementación.

La necesidad de estas discusiones pueden dar lugar a la realización de nuevos trabajos que van desde la relación de la Estadística con el machine learning (aprendizaje automático) y la inteligencia artificial (AI), la importancia del uso de los software en las clases prácticas de la materia como la revitalización de Estadística Descriptiva en la Descripción Grafica de los Datos empleando herramientas como las infografías, las visualizaciones y los tableros de mando.

6. Referencias

- OLSHEN RICHARDO. A Conversation with Leo Breiman. Statistical Science 2001, Vol. 16, No. 2, 184–198
- XLSTAT - Software de Estadística para Microsoft Excel. www.xlstat.com Version: Julio 2019.
- BERENSON MARK, LEVINE DAVID. Estadística Básica en Administración – Sexta Edición. Prentice Hall. Año: 1996.
- NEWBOLD PAUL, CARLSON WILLIAM, BETTY THORNE. Estadística para administración y económica. Prentice Hall. Año: 2011.
- Infogram – Software de visualización de datos. www.infogram.com. Acceso: 20/07/2019.
- Programa Especializado Big Data – Introducción al uso práctico de datos masivos. Universidad Autónoma de Barcelona. www.uab.cat. Finalizacion: 31/06/2019.