Autor:
Iván García Santillán

Fundamentos de la ciencia de datos

Introducción

En esta unidad se revisa los conceptos teóricos relevantes inmersos en el mundo de la ciencia de datos, su importancia para las personas, empresas y sociedad, las metodologías para el desarrollo de proyectos de análisis de datos, los roles existentes y las habilidades blandas necesarias, así como los aspectos éticos en el uso del big data. Además, se realiza el preprocesamiento de los datos y un análisis exploratorio de datos (estadísticas y visualización).

Palabras clave

Fundamentos, metodologías, roles, aspectos éticos, preprocesamiento de datos, análisis exploratorio de datos.

Reto

¿Cuáles son las fases y actividades que se deben desarrollar durante un proyecto de análisis de datos?

Desarrollo

1. Fundamentos del Data Science

1.1 Definición e importancia de la ciencia de datos y Big Data

El Big Data es el análisis masivo de datos. Una cuantía de datos, sumamente grande, que las aplicaciones de software de procesamiento de datos que tradicionalmente se venían usando no son capaces de capturar, tratar y poner en valor en un tiempo razonable; También se refiere a los procedimientos

usados para encontrar patrones/relaciones repetitivos dentro de esos datos (Wikipedia, 2023).
Las dificultades más habituales vinculadas a la gestión de estas cantidades de datos se centran en la recolección y el almacenamiento, búsqueda, compartición, análisis, y visualización.
La tendencia a manipular enormes cantidades de datos se debe a su crecimiento de manera exponencial provenientes de varias fuentes: Internet (buscadores, correo), IoT, dispositivos móviles, redes sociales, comercio electrónico, etc.

Tipos de datos en big data (Hoang, 2016):

- **Datos estructurados** (Structured Data): Datos que tienen bien definidos su longitud y su formato, como las fechas, los números o las cadenas de caracteres. Se almacenan en tablas. Un ejemplo son las bases de datos relacionales y las hojas de cálculo.
- **Datos no estructurados** (Unstructured Data): Datos en el formato tal y como fueron recolectados, que carecen de un formato específico. No se pueden almacenar dentro de una tabla ya que no se puede desgranar su información a tipos básicos de datos. Algunos ejemplos son los PDF, documentos multimedia (imagen, audio, video), e-mails o documentos de texto.
- **Datos semiestructurados** (Semistructured Data): Siguen una especie de estructura pero no es lo suficiente regular como para gestionarla como datos estructurados, Ej. HTML, XML, Json

Características del Big Data: las 7 "V":

- **Velocidad**: con la que actualmente se generan y procesan los datos.
- **Volumen**: la cantidad de datos generados está aumentando exponencialmente.
- **Variedad**: proceden de numerosas fuentes y se encuentran en distintos formatos, cada vez más en forma no estructurada.
- **Veracidad**: fiabilidad y calidad de los datos (ISO/IEC 25012)

- **Valor**: El valor de los datos está en que sean *accionables*, es decir, que permitan tomar una decisión/acción (la mejor) en base a los datos. Hace referencia al beneficio para las empresas.
- **Visualización**: mostrar gráficamente los resultados de forma clara, sencilla. Resumida (dashboard, balacescorecard).
- **Variabilidad**: varían mucho, no son fijos en el tiempo y requieren de un control periódico. Ej. modelos post-covid

Beneficios del Big Data:

- Convertir: Dato → información → conocimiento → decisión → acción
- Permite que una empresa tenga ventajas competitivas.
- Transformación digital basada en datos.
- Podremos sacar conclusiones con una base más sólida y unos conceptos que se orienten a la toma de decisiones/acciones efectivas.
- Todo ello, aplicando no únicamente variables del pasado, sino predicciones a futuro mucho más fundamentadas en una base científica.
- Casos de éxito empresarial: Amazon: recomendación de productos; Netflix: creación de nuevos contenidos basado en preferencias; T-Mobile: retención de clientes. Nike: Fidelización de clientes (promociones, desafíos).
- Otros ámbitos de aplicación: selección de deportistas para clubes de élite (Manchester United), ayuda en campaña electoral (Facebook), detección de fraudes bancarios, sistemas de detección de intrusos IDS, etc.

¿Por qué dominar Big Data?

- Tres razones (Udemy, 2023):
    - Actualmente, el Big Data es el principal destino para la inversión para las empresas.
    - Es la principal fuente de empleo cualificado.
    - Es la mayor causa de creación de empresas de productos y servicios (*startups*) en el ámbito de los sistemas de información.

- Los profesionales dedicados al tratamiento del Big Data, denominados *Data Scientists*, se han convertido en unos de los mejores pagados del sector TIC, en parte por la escasez de profesionales con este perfil.
- Según la encuesta anual *KDNuggets*, el salario de un *Data Scientist* en Estados Unidos oscila entre los 103.000 y los 131.000 dólares mientras que en Europa se encuentra entre los 54.000 y los 82.000 dólares.

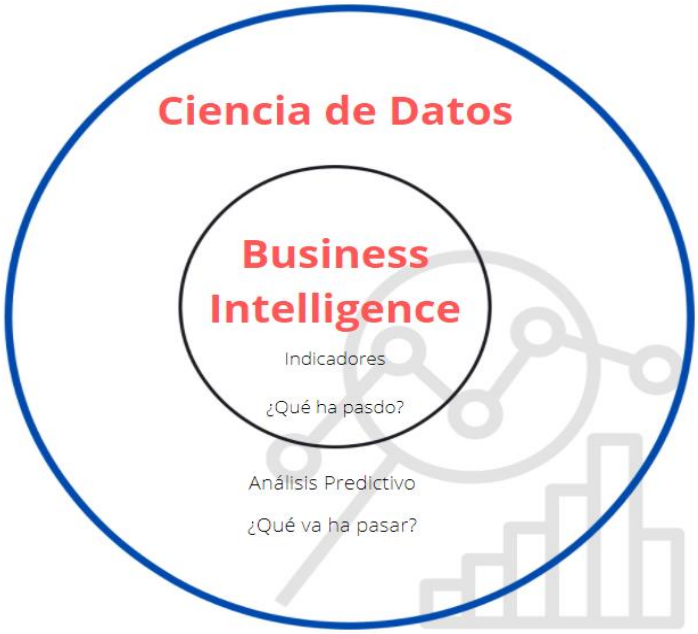
Ciencia de Datos:

La ciencia de datos es un campo interdisciplinario que utiliza estadística, computación científica, métodos, procesos, algoritmos y sistemas para obtener (recolectar o extraer), tratar, analizar y presentar informes a partir de datos ruidosos, estructurados y no estructurados (Wikipedia, 2023b).

La parte medular de la ciencia de datos son los algoritmos de Inteligencia Artificial que manejan datos, ellos forman parte de la minería de datos, y tienen la finalidad de explorar y sacarles el máximo valor. Dentro de este campo, tenemos 2 tipos de aprendizaje: El supervisado y no supervisado. La principal diferencia entre el aprendizaje supervisado y el aprendizaje no supervisado radica en la disponibilidad de etiquetas en los datos de entrenamiento. En el aprendizaje supervisado, el modelo se entrena utilizando datos etiquetados con el objetivo de hacer predicciones precisas, mientras que en el aprendizaje no supervisado, el modelo busca estructuras o patrones en los datos sin la guía de etiquetas. Ambos enfoques son esenciales en el aprendizaje automático y se utilizan en una variedad de aplicaciones.

Además, en la Figura 1 se muestra la relación de la ciencia de datos con la inteligencia de negocios que combina análisis de negocios, minería de datos, visualización de datos, herramientas e infraestructura de datos, y las prácticas recomendadas para ayudar a las organizaciones a tomar decisiones más basadas en los datos.

Figura 1 Ciencia de datos y su relación con la inteligencia de negocios



Fuente: Bantu (2021)

Principalmente, la inteligencia de negocios trabaja con los datos históricos preparados, generados y almacenados por las empresas. La ciencia de datos se enfoca en el futuro, responde preguntas tales como, ¿qué pasaría si se hace tal o cual cosa?, e identifica patrones o tendencias para crear predicciones (Bantu, 2021).

Minería de Datos (Data Mining)

La minería de datos o exploración de datos (es la etapa de análisis de "*Knowledge Discovery in Databases*" o KDD) es un campo de la estadística y las ciencias de la computación referido al proceso que intenta descubrir patrones (relaciones, tendencias) en grandes volúmenes de conjuntos de datos (Lara, 2014).

Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos.

Tenemos 2 grandes grupos: técnicas predictivas y descriptivas.

Técnicas y algoritmos de análisis de datos masivos:

Técnicas predictivas:

- **Clasificación**: El atributo a predecir (target) es de tipo cualitativo (categoría). ej. clasificación de objetos en imágenes: perro, gato, avión, laptop. Algoritmos clásicos: RNA (MLP), SVM, decision tree, K-NN, Naive Bayes, logistic regression, Random Forest (ensemble).
- **Regresión**: similar a la clasificación, pero el atributo a predecir es de tipo cuantitativo. Regresión lineal y múltiple. Algoritmos: linear and multiple regression with least-squares method.

  Nota: la regresión *logística* es una técnica utilizada para clasificación.

Técnicas Descriptivas:

- **Agrupamiento:** segmentación en grupos homogéneos. Ej. Clientes tipo A, B, C. Algoritmos: : k-means, Expectation Maximization (EM).
- **Asociación**: identificación de productos que habitualmente se compran juntos (análisis de la canasta). Ej. pan, azúcar → leche. Reglas antecedente-consecuente. Algoritmos: A priori, FP Growth
- **Detección de atípicos**: localización de objetos que manifiestan características significativamente diferentes al resto y afectan a los modelos. Técnicas estadísticas: scatter-plot, box-plot y Algoritmos: Isolation Forest, Minimum Covariance Determinant, Local Outlier Factor, One-Class SVM.

1.2 Metodologías para desarrollo de proyectos de análisis de datos

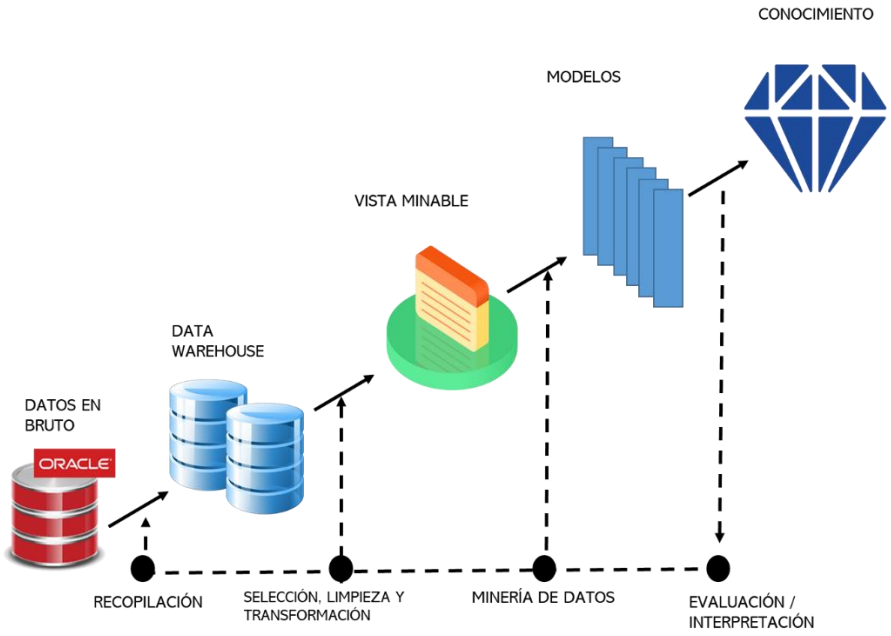
Una **metodología** no solo define las **fases** de un proceso, sino también las **tareas** que deberían realizarse y **cómo** llevar a cabo las mismas. Algunas metodologías muy utilizadas para el análisis de datos son:

- KDD (knowledge Discovery in databases, 1996)
- CRISP-DM (Cross-Industry Standard Process for Data Mining, 2000)

KDD constituyó el primer modelo aceptado en la comunidad científica que estableció las etapas principales de un proyecto de explotación de información. Está más cercano a un modelo de proceso, ya que sólo proponen las fases generales para el proceso de minería de datos y no incorpora

actividades para la gestión del proyecto (como la gestión del tiempo, costo, riesgo). En la Figura 2 se muestra el proceso KDD.

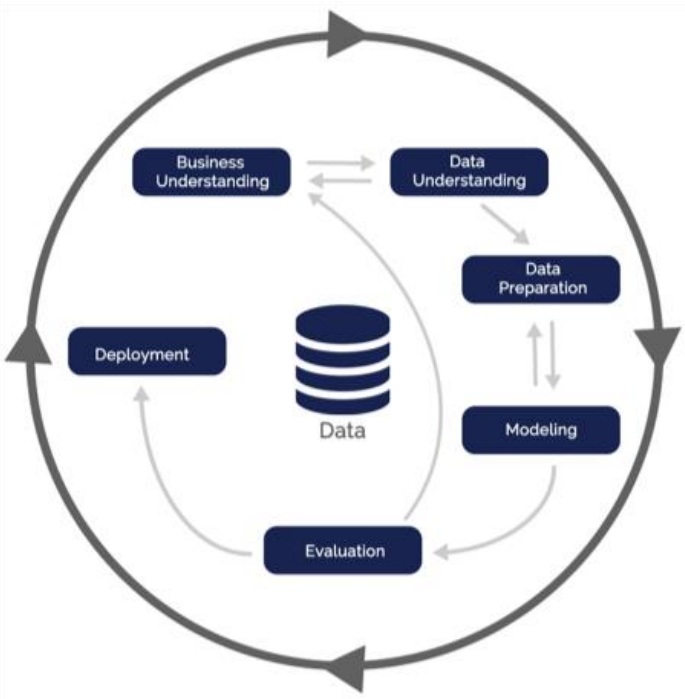
Figura 2 Metodología KDD para de análisis de datos



Fuente: Fayyad, U. (1996).

CRISP-DM podrían ser considerados una metodología, por el nivel de detalle con el que describen las tareas en cada fase del proceso, y porque incorporan actividades para la gestión del proyecto. CRISP-DM es actualmente la guía de referencia más utilizada en el desarrollo de proyectos de minería de datos.

Figura 3. Metodología CRISP-DM para de análisis de datos



Fuente: Chapman et al. (2000)

Las fases del proceso de análisis de datos en cada modelo se muestran en la Tabla 1.

Tabla 1. Fases de las metodologías de análisis de datos

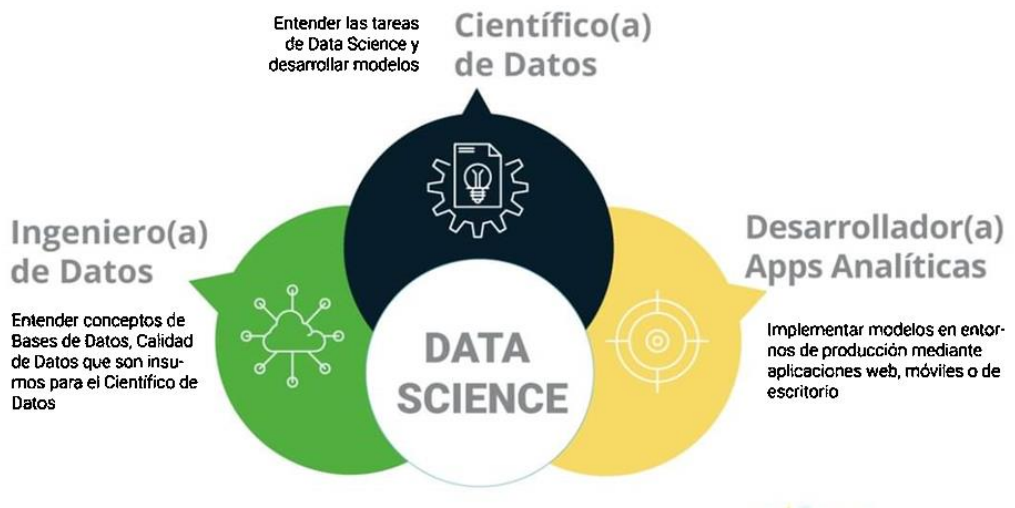
Fases	KDD	CRISP – DM
Análisis y comprensión del negocio	Comprensión del dominio de aplicación	Comprensión del negocio
Selección y preparación de los datos	Crear el conjunto de datos Limpieza y pre-procesamiento de los datos Reducción y proyección de los datos	Entendimiento de los datos Preparación de los datos
Modelado	Determinar la tarea de minería Determinar el algoritmo de minería Minería de datos	Modelado
Evaluación	Interpretación	Evaluación
Implementación	Utilización del nuevo conocimiento	Despliegue

Fuente: Moine et al. (2011)

1.3 Roles y habilidades blandas para ciencia de datos

Los roles identificados en ciencia de datos son: Ingeniero de datos, Científico de datos y Desarrollador de aplicaciones analíticas como se muestra en la Figura 4.

Figura 4. Roles identificados en la ciencia de datos



Fuente: SEE (2023)

En la Figura 5 se muestran las habilidades blandas requeridas para ciencia de datos.

Figura 5. Habilidades blandas para ciencia de datos

Fuente: Data Solutions (2022)

1.4 Aspectos legales y éticos del uso del Big Data.

La privacidad y la seguridad son cuestiones críticas en el mundo de la tecnología y los datos. **La privacidad** se refiere a la protección de la información personal y sensible de las personas. En el contexto del Big Data, existen preocupaciones legales y éticas relacionadas con la recopilación, el almacenamiento y el uso de datos personales. **La seguridad** de los datos en el contexto del Big Data se refiere a la protección contra amenazas cibernéticas y la garantía de que los datos se almacenan y transmiten de manera segura.

El uso del Big Data plantea desafíos legales y éticos importantes en relación con la privacidad y la seguridad de los datos. Las organizaciones que trabajan con Big Data deben ser conscientes de estas preocupaciones y tomar medidas adecuadas para proteger la privacidad de las personas y garantizar la seguridad de los datos. Además, deben estar al tanto de las regulaciones y leyes específicas en sus regiones y cumplir con ellas de manera rigurosa para evitar consecuencias legales y daños a su reputación. Algunas normativas que se deben considerar en Ecuador y la región son las siguientes:

- Ley orgánica de protección de datos personales (LOPDP, mayo 2021.
- Recomendación sobre la Ética de la Inteligencia Artificial (Unesco, Nov. 2021).

El objeto y finalidad de la **ley LOPDP** (2021) es garantizar el ejercicio del derecho a la protección de datos personales, que incluye el acceso y decisión sobre información y datos de este carácter, así como su correspondiente protección, Para dicho efecto regula, prevé y desarrolla principios, derechos, obligaciones y mecanismos de tutela.

La **Recomendación sobre la ética de la inteligencia artificial** es el primer instrumento normativo mundial sobre la ética de la inteligencia artificial propuesto por la Unesco (2021). La ética en la inteligencia artificial se refiere al estudio de los valores y principios éticos que deben guiar la creación y uso de sistemas de inteligencia artificial. Esto incluye la toma de decisiones, la transparencia y la responsabilidad en el desarrollo y uso de la tecnología.

Trabajo autónomo:

Realice un cuadro sinóptico o mapa mental o infografía respecto a la ética, Privacidad y Seguridad de los datos considerando las dos normativas mencionadas y suba a la plataforma.

1.5 Preprocesamiento de datos

Este paso implica la preparación de los datos crudos para su posterior análisis. En general, algunas de las tareas comunes de preprocesamiento de datos incluyen:

- Limpieza de datos para tratar valores faltantes (nulos), valores atípicos y datos inconsistentes.
- Escalado o normalización de datos numéricos.
- Transformación (codificación) de datos categóricos en un formato adecuado para su análisis.
- Reducción de la dimensionalidad si fuese necesario.

En la lectura de profundización se abordará cada una de estas actividades a detalle (InteractiveChaos, 2023).

1.6 Análisis exploratorio de datos (estadísticas y visualización)

El análisis exploratorio de datos (EDA por sus siglas en inglés, Exploratory Data Analysis) es una parte fundamental de cualquier proyecto de análisis de datos que tiene varios propósitos importantes, tales como: Comprender los datos, incluyendo la estructura de los datos, sus características, distribuciones, tendencias y patrones subyacentes); Detectar Anomalías y Errores; Seleccionar variables relevantes; Generar hipótesis iniciales sobre relaciones entre variables; Tomar decisiones iniciales sobre cómo abordar el proyecto de análisis de datos (elección de algoritmos), Optimización de recursos eliminando variables irrelevantes o reducir la dimensionalidad de los datos; Preparación para un modelado avanzado, etc.

Aquí se mostrará cómo realizar un análisis exploratorio de datos básico en el conjunto de datos del *Titanic* utilizando Python, centrándonos en estadísticas descriptivas y visualización de datos. Estas actividades se abordarán a detalle en la lectura de profundización.

Conclusiones

En esta unidad se ha revisado los fundamentos de la ciencia de datos, tales como:

- Definición e importancia de la ciencia de datos y Big Data
- Metodologías para desarrollo de proyectos de análisis de datos
- Roles y habilidades blandas para ciencia de datos
- Aspectos legales y éticos del uso del Big Data
- Preprocesamiento de datos
- Análisis exploratorio de datos (estadísticas y visualización)

La ciencia de datos es una disciplina interdisciplinaria que utiliza estadística, computación científica, métodos, procesos, algoritmos y sistemas para obtener (recolectar o extraer), tratar, analizar y presentar informes a partir de datos ruidosos, estructurados y no estructurados. La parte medular de la ciencia de datos son los algoritmos de Inteligencia Artificial que manejan datos, y tienen la finalidad de explorar y sacarles el máximo valor. Dentro de este campo, tenemos 2 tipos de aprendizaje: El supervisado y no supervisado. La principal diferencia entre el aprendizaje supervisado y el aprendizaje no supervisado radica en la disponibilidad de etiquetas en los datos de entrenamiento. En el aprendizaje supervisado, el modelo se entrena utilizando datos etiquetados con el objetivo de hacer predicciones precisas, mientras que en el aprendizaje no supervisado, el modelo busca estructuras o patrones en los datos sin la guía de etiquetas. Ambos enfoques son esenciales en el aprendizaje automático y se utilizan en una variedad de aplicaciones.

Para el desarrollo de proyectos de análisis de datos se requiere seguir una metodología la cual no solo define las fases de un proceso, sino también las tareas que deberían realizarse y cómo llevar a cabo las mismas. Algunas metodologías muy utilizadas para el análisis de datos son: KDD y CRISP-DM.

También, la privacidad y la seguridad son cuestiones críticas en el mundo de la tecnología y los datos. La privacidad se refiere a la protección de la información personal y sensible de las personas y la seguridad de los datos se refiere a la protección contra amenazas cibernéticas. Para ello hay que considerar la normativa existente en el país y la región, tales como: Ley orgánica de protección

de datos personales (Asamblea Nacional) y las recomendaciones sobre la Ética de la Inteligencia Artificial (Unesco).

Por otro lado, antes de realizar el análisis de los datos, es necesario realizar un preprocesamiento de los datos. Este paso implica principalmente la limpieza de datos, Escalado o normalización de datos numéricos y Transformación (codificación) de datos categóricos en un formato adecuado para su análisis.

Así mismo, el análisis exploratorio de datos es una parte fundamental de cualquier proyecto de análisis de datos que tiene varios propósitos importantes, tales como: Comprender los datos, incluyendo la estructura de los datos, sus características, distribuciones, tendencias y patrones subyacentes; Detectar Anomalías y Errores; Seleccionar variables relevantes; Generar hipótesis iniciales sobre relaciones entre variables; Tomar decisiones iniciales sobre cómo abordar el proyecto de análisis de datos (elección de algoritmos), Optimización de recursos eliminando variables irrelevantes o reducir la dimensionalidad de los datos; Preparación para un modelado avanzado de datos, etc.

ANEXOS

Bibliografía:

Asamblea Nacional del Ecuador. (2021). Ley Orgánica de Protección de Datos Personales de Ecuador (LOPDP). Ecuador.

Bantu. (2021). Ciencia de datos e Inteligencia de Negocios, características y diferencias. https://www.bantugroup.com/blog/ciencia-de-datos-e-inteligencia-de-negocios

Brownlee, J. (2023). Machine Learning Mastery. https://machinelearningmastery.com

Chapman, P. et al. (2000). CRISP-DM 1.0: Step-by-step data mining guide. https://api.semanticscholar.org/CorpusID:59777418

Data Solutions. (2022). Habilidades blandas para ciencia de datos.

Fayyad, U. (1996). Advances in Knowledge Discovery and Data Mining. *MIT Press*.

García-Santillán, I. (2023). Ciencia y analítica de datos. Maestría en Inteligencia Artificial aplicada. Universidad Hemisferios.

Gonzalez, L. (2023). Predecir la supervivencia del Titanic. http://ligdigonzalez.com/predecir-la-supervivencia-del-titanic-utilizando-python/

Hoang, A. (2016). Tipos de datos en Big Data. https://anhngohoang.wordpress.com/2016/05/

InteractiveChaos. (2023). Preprocesamiento de datos. https://interactivechaos.com/en/node/916

Lara, J. (2014). Minería de Datos. Madrid: CEF-Udima.

Lind et al., (2015). Estadística aplicada a los negocios y la economía. 16° ed. McGraw-Hill: México.

Moine, J. (2011). Análisis comparativo de metodologías para la gestión de proyectos de minería de datos. CACIC 2011 - XVII Congreso argentino de ciencias de la computación. https://core.ac.uk/download/pdf/15775611.pdf

SEE. (2023). Roles en un proyecto de ciencia de datos. Sociedad Ecuatoriana de Estadística en Facebook. https://www.facebook.com/socecuest

Unesco. (2021). Recomendación sobre la ética de la inteligencia artificial. https://unesdoc.unesco.org/ark:/48223/pf0000380455_spa

Udemy (2023). Análisis de Big Data. https://www.udemy.com/course/analisis-de-big-data/

Wikipedia. (2023a). Big Data. https://es.wikipedia.org/wiki/Macrodatos

Wikipedia. (2023b). Ciencia de Datos. https://es.wikipedia.org/wiki?curid=7036548