REVISTA INVESTIGACIÓN, TRANSCOMPLEJIDAD Y CIENCIA
(Revista ITC)

Volumen 4 Nro. 1 Año 2023

PARTE III.
CIENCIA, TECNOLOGÁ Y SOCIEDAD

Imagen 20: Ciencia, tecnología y sociedad
Fuente: www.facebook.com/CTSVcbta35/

DATA SCIENCE COMO HERRAMIENTA DE INVESTIGACIÓN

Carlos Alfredo La Roche Tomasini

Resumen

En el siguiente escrito se describe brevemente el proceso de investigación por el cual se realizó el análisis de los casos de covid-19 en distintos países de Europa. Siendo este un objetivo complejo, se describe, sobre todo, cómo se empleó el Data Science, como una herramienta de análisis para el procesamiento masivo de datos. ¿Se explica, además, la relación existente entre el diseño propio e implementado para el desarrollo de software, y el diseño de una investigación típica. A lo largo del siguiente artículo se demuestra, además, la capacidad del concepto de Data Science para abordar investigaciones en donde se requiere un mayor alcance para aumentar su fiabilidad. De igual manera, muestra cómo este concepto, así como el desarrollo del software propuesto en la investigación, puede ser sencillamente extrapolado a cualquier ámbito en el que se requiera aumentar el alcance de un análisis de datos específicos.

Palabras clave: Data Science, Covid-19

DATA SCIENCE AS A RESEARCH TOOL

Abstract

The following paper briefly describes the research process by which the analysis of covid-19 cases in different European countries was carried out. This being a complex objective, it describes, above all, how Data Science was used as an analysis tool for massive data processing. In addition, the relationship between the design and implementation of the software development and the design of a typical investigation is explained. The following article also demonstrates the ability of the Data Science concept to address research where a larger scope is required to increase its reliability. Likewise, it shows how this concept, as well as the software development proposed in the research, can be simply extrapolated to any field where the scope of a specific data analysis needs to be increased.

Keywords: Data Science, Covid-19

Introducción

En muchas ocasiones un investigador debe lidiar con un entorno en el que existen demasiadas variables para ser analizadas, o la relación entre estas, es en extremo compleja; en ambos contextos la investigación pudiera verse retrasada en mayor o menor medida. En la actualidad, un ejemplo claro de este tipo de contratiempos sería la pandemia ocasionada por el virus SARS-CoV-2; este es un escenario que conlleva un gran impacto negativo, en distintos aspectos de la sociedad, y en el que, adicionalmente, existen diversas variables interactuando entre sí que limitan enormemente las capacidades de análisis de un investigador.

Pese a esto, existen áreas de las matemáticas como la estadística que permiten abordar correctamente investigaciones complejas. Popularizada entre los años 20 y 80, la estadística es busca llegar a conclusiones sobre una población basándose en el muestreo representativo de una pequeña sección de esta; este concepto amplía las capacidades de investigación. Sin embargo, aún existe una limitante: para que las inferencias realizadas sobre la población, en base a la muestra, sean correctas, se requiere una muestra de un tamaño adecuado; mientras mayor sea el tamaño de la muestra, en relación a la población total, más acertados serán los resultados.

Una de las herramientas actualmente disponibles para enfrentar dicha limitante, es la programación. El Data Science es una rama de la informática que se basa en la combinación de la programación con los conceptos de la estadística y la probabilidad, en distintos niveles. Para ello, se emplea el procesamiento de datos masivos, o Big Data, concepto popularizado mayormente por IBM que sirvió como base para el Data Science y otras ramas de la informática. Siendo la pandemia del covid-19 un evento que cuenta con diversas variables y, por una, una alta complejidad, este representará la

situación inicial, entorno a la cual se centrará la investigación.

En el presente artículo se explica y analiza el proceso de investigación realizado sobre los casos de covid-19, empleando el Data Science como herramienta de investigación, con Python como lenguaje de programación. El artículo que se presenta a continuación está dispuesto de la siguiente manera:

Metodología: Explica el proceso de investigación que se llevó a cabo para el análisis de los casos de covid-19, haciendo énfasis en los pasos con los cuales se desarrolló y aplicó el concepto de Data Science.

Resultados: Se presentan las conclusiones obtenidas de la investigación de manera detallada. Conclusiones: Se realiza un análisis sobre el proceso de investigación, así como una reflexión sobre este y los resultados obtenidos.

Metodología

Durante el proceso de investigación se combinaron dos metodologías: una referente al método científico y otra enfocada al diseño del sistema para la propuesta de solución. Al igual que en toda investigación, proyecto, o plan de acción, el primer paso consiste en evaluar el entorno y cuáles son los elementos que tienen algún tipo de influencia en el proceso planteado.

En este sentido, durante la investigación se plantea, primeramente, cómo se percibe la realidad. Dentro de esa sección se describen detalladamente el contexto que engloba a la situación inicial: la pandemia del covid-19 y las consecuencias sociales que conlleva, incluyendo el desconocimiento científico de su propagación y medios de prevención. Basándose en dichas características, se presentan los síntomas y aparentes causas que contribuyen a la situación inicial y, por supuesto, la propuesta planteada para su abordaje.

En cuanto al diseño del sistema, tratándose de una investigación que emplea el Data Science como herramienta, la descripción de las entidades se centra en aquellas que colaboran a la recolección y

organización masiva de los datos. Entre las entidades principales se pueden destacar: El Centro Europeo para la Prevención y Control de Enfermedades, El Banco Mundial, y Eurostat; representando estas tres entidades aquellas en las que la información es registrada y organizada de forma masiva, pudiendo procesar más de 6700 datos estadísticos de diversos países. Dichas entidades se presentan en el gráfico 1.

Por otra parte, con un enfoque más centrado las bases de las metodologías de investigación en general, y contando con las características de la situación actual bien definidas, el siguiente paso es establecer las bases teóricas que sustentan la investigación. En esta sección se plantean, tanto conceptos relacionados con la programación y las metodologías empleadas para el desarrollo del sistema, incluyendo aquellas que sean propias del

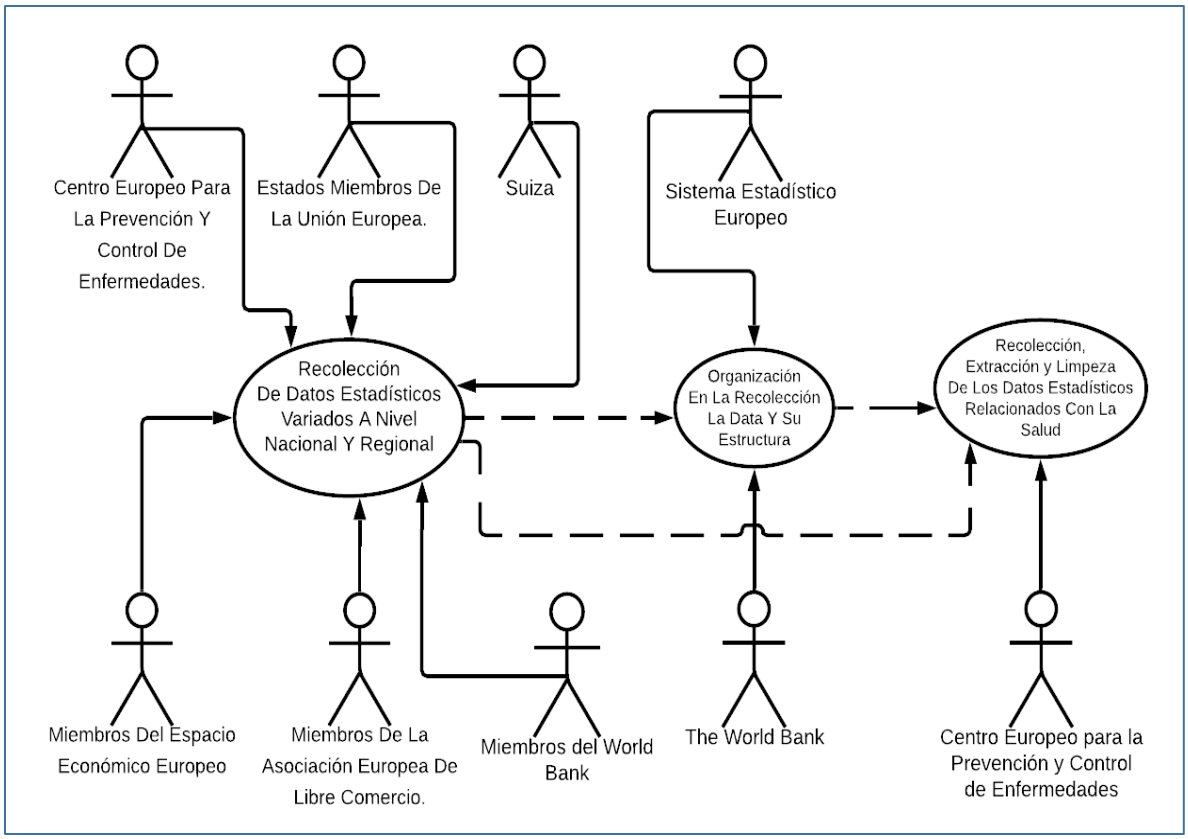


Gráfico 1. Diagrama de Caso de Uso
Fuente: Elaboración propia.

lenguaje de programación Python; como aquellos conceptos relacionados con la estadística y probabilidad que sean pertinentes para la investigación.

Se hace especial énfasis en los conceptos de programación y lineamiento recomendados para el uso de Python. De igual manera, se presenta un apartado para la definición de conceptos relacionados con el Data Science. Adicionalmente, se presentan los requisitos que debe cumplir el equipo destino para que el programa pueda ser correctamente implantado, así como los requisitos que este debe cumplir para adecuarse a la propuesta planteada.

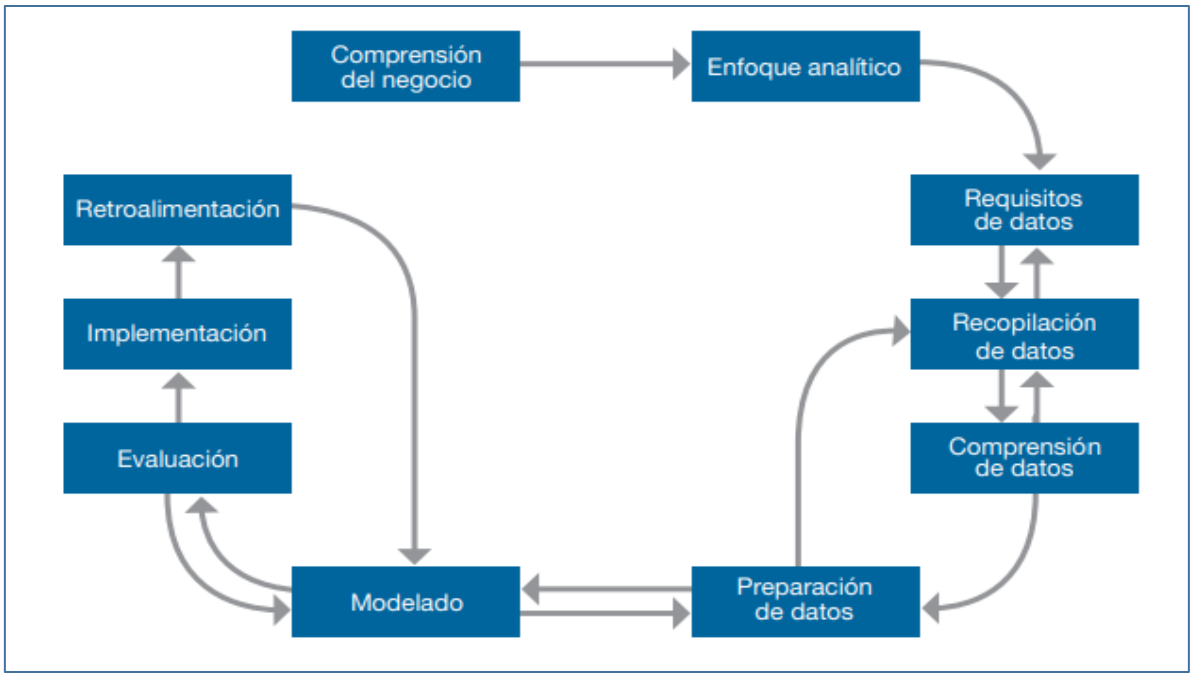
En cuanto a la estructura metodológica que abarca, tanto la investigación, como el desarrolla del programa para aplicar correctamente el concepto del Data Science, así como las metodologías adecuadas y recomendadas en el campo, en una investigación, estas deben ser adecuadas a las metodologías de investigación, en general. Para dicho propósito se plantea La Metodología Fundamental Para La Ciencia De Datos, de IBM. En ella se proponen las prácticas recomendadas profesionales experimentados en los campos del Data Science y Big Data, dichos conceptos pueden ser perfectamente adecuados para cualquier desarrollo de sistemas en estos campos, sin importar el lenguaje de programación empleado.

Es importante destacar que, a pesar de que se emplee una metodología de IBM, los conceptos que se presentan en el método científico, son extrapolados a la metodología fundamental para la ciencia de datos a lo largo de sus fases. Entre estas se pueden percibir ideas similares a las de la “observación y análisis” o la de generar expectativas o hipótesis previamente a la investigación, así como la obvia interpretación de los resultados. Esta sección representa una base para cualquier tipo de desarrollo en el campo del Data Science, y fue aplicada en profundidad, tanto para el desarrollo del programa propuesto como para el

de la investigación como se muestra el gráfico 2.

Gráfico 2. Etapas de la metodología para ciencia de datos propuesta por IBM
Fuente: Metodología Fundamental para la Ciencia de Datos, por IBM.

investigación formal aún se requiere un recolectar información que sustente o facilite el desarrollo e implementación de la propuesta. De igual manera, se determina el proceso de construcción y determinación del programa, como muestra el gráfico 2.



Por otra parte, una implementación más directa de las metodologías de investigación tradicionales se observaría en el proceso de recopilación de información. A pesar de que el Data Science como indica el gráfico 2 propone herramientas y metodologías poco típicas, al tratarse de una

En específico esta sección describe, de manera técnica y a profundidad, todo el proceso de desarrollo del software propuesto; se muestra, principalmente, el diseño, tanto lógico, como visual, del sistema. Para ello se emplean estándares generales para el desarrollo de sistemas de información o softwares,

en específico, mediante los diagramas UML; estos son un punto en común en todo proyecto.

Un punto que se considera necesario destacar es que la definición de las pruebas a las que el programa será expuesto. En particular para la investigación, dichas pruebas representan una pieza fundamental para el desarrollo del software, pues se emplea el Desarrollo Dirigido por Prueba, dicha metodología se base en la creación de las pruebas *antes* del programa en sí. Existen múltiples ventajas: más fiabilidad y confianza por parte de los usuarios, mayor estabilidad en el sistema y, sobre todo, una gran reducción en el tiempo de desarrollo.

Resultados

En general, durante la investigación se obtuvieron todos los objetivos propuestos; se logró el obtener una herramienta que procesara ampliamente en los casos del covid-19, disminuyendo en gran medida las limitaciones que podría suponer una investigación tal contexto. Adicionalmente, se obtuvieron diversas formas y herramientas para expresar e interpretar los datos; no solo se logró desarrollar la propuesta en su totalidad, sino que se obtuvieron datos fiables y fácilmente interpretables con los que desarrollar las conclusiones obtenidas para el tópico abortado.

Existen diversas investigaciones que se ven enormemente limitadas por las capacidades o recursos disponibles para el investigador, destacando, sobre todo, la recolección de datos que sirvan de muestra para alguna demostración de hipótesis o pruebas realizadas. Esto representa un problema en distintos campos; una investigación puede, incluso, llegar a no ser factible por el simple hecho de que le investigador no puede realizar un análisis lo suficientemente extenso.

Conclusiones

En ese sentido, y dentro del contexto de la investigación, el Data Science es un campo que ha permitido evaluar una cantidad de datos considerablemente mayor a la que un

investigador, por cuenta propia y de manera manual, podría procesar.

Con este tipo de herramientas se extiende el alcance de cualquier investigación en la que existan suficientes datos ¿Cuáles podrían ser las posibles limitaciones en este sentido? Mientras que el procesamiento de la data se facilita y se pone a la disposición de cualquier investigador con los suficientes conocimientos técnicos, la misma obtención de dicha data está sujeta a la colaboración de terceros; una cantidad masiva de datos, de una manera u otra, corresponderá a una cantidad masiva de evaluaciones. Sin embargo, la diferencia entre la información disponible para una investigación que emplea el Data Science como herramienta, y una que emplea los métodos tradicionales de recolección de datos, es considerable y, evidentemente, favorable para quienes emplean dichas tecnologías.

Referencias

Ainsworth, Q. (8 de marzo de 2021). **Data Collection Methods. Jot Form Education.** https://www.jotform.com/data-collection-methods/ (1-4-2021)

Beck, K., Beedle, M., Van Bennekum, A., y Cockburn, A. (s.f.). **Manifesto for Agile Software Development**. (30-6-2021) https://agilemanifesto.org/

Downey, A. B. (2014). **Think Stats: Exploratory Data Analysis** (Vol. 2). Needham, Massachusetts: Green Tea Press. (20 -9-2021)

Farmer, D. (18 de diciembre de 2022). **The data science process: 6 key steps on analytics applications.** https://searchbusinessanalytics.techtarget.com/feature/The-data-science-process-6-key-steps-on-analytics-applications (13-6 2021)

FormPlus Blog. (4 de diciembre de 2020). **7 Data Collection Methods & Tools For Research**. FormPlus. https://www.formpl.us/blog/data-collection-method (1 -4-2021)

Han Lau, C. (3 de enero de 2019). **5 Steps of a Data Science Project Lifecycle**. Towards Data Science. https://towardsdatascience.com/5-steps-of-a-data-scienceproject-lifecycle-26c50372b492 (26-2- 2021)

Hotz, N. (21 de febrero de 2021). **What is a Data Science Life Cycle? de Data Science Process Alliance**. https://www.datascience-pm.com/data-science-life-cycle/ (15 -6- 2021)

IBM Analytics. (4 de enero de 2019). **Metodología Fundamental para la Ciencia de Datos**. IBM. https://www.ibm.com/downloads/cas/WKK9DX51 (29 -9- 2021)

Mayo, M. (2016). **The Data Science Process, Rediscovered**. KDnuggets. https://www.kdnuggets.com/2016/03/data-science-process-rediscovered.html (13-7-2021)

Object Management Group. (2015). **OMG Unified Modeling Language TM (OMG UML).** https://www.omg.org/spec/UML/2.5/PDF (29-9-2021)

Wheelan, C. (2013). **Naked statistics: Stripping the Dread from the Data**