



CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL: EXPERIENCIA EN UNA INVESTIGACIÓN CUALITATIVA

DATA SCIENCE AND ARTIFICIAL INTELLIGENCE: EXPERIENCE IN QUALITATIVE RESEARCH

José Javier Leal Rivero
<https://orcid.org/0000-0002-5030-1640>
 Universidad Politécnica Territorial del Estado Trujillo
 “Mario Briceño Iragorry”
 VENEZUELA

CIENCIA DE DATOS E INTELIGENCIA ARTIFICIAL: EXPERIENCIA EN UNA INVESTIGACIÓN CUALITATIVA

*DATA SCIENCE AND
ARTIFICIAL
INTELLIGENCE:
EXPERIENCE IN
QUALITATIVE
RESEARCH*

Resumen

El presente artículo, se presenta con el propósito fundamental de precisar la aplicabilidad de la denominada “Ciencia de Datos” (Data Science) y sus herramientas de “Inteligencia Artificial” (IA), en el desarrollo de investigaciones cualitativas, aplicadas al estudio de fenómenos sociales complejos. La metodología se corresponde con el Análisis Comparativo Cualitativo; para ello se toman los datos de una investigación cualitativa, referencial, desarrollada usando el método Fenocomplejo, fundamentado en los Modelos de Sistemas Viables de Stafford Beer y la Fenomenología Hermenéutica de Van Manen. A partir de la Clasificación, Predicción y Agrupación de Datos, captados en los registros anecdóticos de la investigación referencial, se logró desarrollar la comparación relacional de los hallazgos con base en la coocurrencia de frases y palabras, así como la valorización de sentimientos. Finalmente se considera que la Ciencia de Datos y la Inteligencia Artificial son perfectamente aplicables, y recomendables, en el desarrollo de estas investigaciones.

Descriptores: ciencia de datos; inteligencia artificial; investigación cualitativa

Abstract

This article is presented with the fundamental purpose of specifying the applicability of the so-called "Data Science" (Data Science) and its "Artificial Intelligence" (AI) tools, in the development of qualitative research, applied to the study of phenomena. complex social. The methodology corresponds to the Qualitative Comparative Analysis; For this, data are taken from a qualitative, referential research, developed using the Phenocomplex method, based on Stafford Beer's Viable Systems Models and Van Manen's Hermeneutic Phenomenology. From the Classification, Prediction and Data Grouping, captured in the anecdotal records of the referential research, it was possible to develop the relational comparison of the findings based on the co-occurrence of phrases and words, as well as the appreciation of feelings. Finally, it is considered that Data Science and Artificial Intelligence are perfectly applicable, and recommended, in the development of these investigations.

Keywords: data science; artificial intelligence; qualitative research

La investigación cualitativa, a diferencia de las desarrolladas bajo el enfoque positivista, o cuantitativas, se caracterizan por su orientación relativista del conocimiento, así como por su pragmatismo metodológico. Para esta, tal como lo señala (Davison, 2006), no existe una verdad universal, persistente e inmutable, siendo lo más resaltante en ella, que todo aquello considerado como real y verdadero, es simplemente un código de representación, donde el conocimiento puede ser alcanzado recorriendo variados caminos, permitiendo con ello situar el significado particular de los fenómenos en un contexto social.

Así, en ella puede observarse como la denominada verdad científica, y la comprensión de la realidad, son constituidas a través del consenso entre los miembros de una determinada comunidad, tomando para sí el discurso ideológico como un proceso de negociación y establecer lo que se considera, y acepta, como verdad (Smith y Deemer, 2000; Hodder, 2000). Por tanto, se aleja de la validez objetiva positivista, decantándose por aquella validez social, instituida por la comunidad y determinada por el momento histórico y espacio-temporal, en el cual se desarrolla.

Ante esas particularidades, durante el desarrollo de las mismas, es común la generación de una enorme cantidad de datos, indiferentemente de la cantidad y cualidad de elementos, o mecanismos, asumidos como generadores de información relevante, durante el abordaje de los fenómenos humanos, o sociales, al concebirlos como una actitud consciente de las personas ante los fenómenos que afectan su vida, proveniente estos bien sea de entrevistas, relatos anecdóticos, observaciones, entre muchos otros, recopilados a través de instrumentos basados en el uso de medios escritos tradicionales, captura de imágenes, audio o video, incluso aprovechando la proliferación de las tecnologías de comunicación social masiva, tomarlos de las conocidas redes sociales tales como Facebook, Instagram, Tik Tok, WhatsApp, Youtube, entre muchísimos otros, que aportan un nivel casi infinito de información personal, o en conjunto, de la sociedad misma y la conciencia social, causada por hechos o situaciones tanto particulares como globales.

Comprensiblemente, este gran caudal de información posteriormente debe ser de alguna forma codificada, categorizada, analizada de manera minuciosa y concienzuda, aplicando métodos específicos acordes a la visión y propósitos planteados inicialmente por el investigador, desarrollando para ello un trabajo arduo, muy laborioso y cíclico, de lectura - relectura, interpretación - reinterpretación, que en algunos casos conlleva a que la investigación deba ser reasumida desde otro punto de vista, re-fundamentada teóricamente y hasta reformulada

metodológicamente. Entonces, la mayor preocupación al enfocarse los trabajos de investigación en este tipo de metodología, puede centrarse en la posibilidad que, de la gestión de esos datos y su interpretación, pueda producirse algún tipo de sesgo, limitaciones o errores, bien sea por parte de los informantes o del propio investigador. (Bertoldi, Fiorito y Álvarez, 2006)

Ante este panorama, se plantea el presente artículo con el propósito fundamental de precisar la aplicabilidad de la denominada “Ciencia de Datos” (Data Science) y sus herramientas de “Inteligencia Artificial” (IA), en el desarrollo de investigaciones cualitativas, aplicadas al estudio de fenómenos sociales complejos, generadoras de una masiva y multivariada cantidad de datos que pueden conllevar a la pérdida del enfoque investigativo, o la omisión de factores relacionantes entre ellos, o entre los elementos de estudio que los generan, tales como individuos, asociaciones, instituciones, entre otros, facilitando una visión panorámica y generalizada de los fenómenos estudiados, el análisis, comprensión y transmisión de sus hallazgos, repercutiendo a su vez en una mayor calidad de la investigación.

El artículo por tanto, es asumido con un criterio crítico y amplio, a partir de la experiencia generada durante el desarrollo de una investigación cualitativa, basada en el estudio de las relaciones existentes entre la praxis universitaria y su consecuente generación de acciones conducentes al logro de la participación ciudadana para el desarrollo de sus comunidades, siendo abordada aplicando el “Método Fenocomplejo”, fundamentado en los Modelos de Sistemas Viables de Stafford Beer (Beer, 1979) y la Fenomenología Hermenéutica de Van Manen (Van Manen, 2003). Para ello se comparan los hallazgos de la misma, como referencia o benchmarking, con los resultados obtenidos a partir de la aplicación de herramientas de Ciencia de Datos e Inteligencia Artificial, a la información no procesada o cruda, captada en registros anecdóticos facilitados por los profesores universitarios participantes.

Argumentación teórica

Aproximación al Concepto de Ciencia de Datos (Data Science)

Para realizar un acercamiento a la conceptualización a la denominada Ciencia de Datos, se hace necesario partir por definir su elemento constitutivo fundamental, es decir, los datos propiamente dichos. Los datos son, en el contexto de la acción humana, “Información concreta sobre hechos y elementos, entre otros, que permite estudiarlos, analizarlos o conocerlos” (Veas, 2021, p.02). Por su parte, García y Cols. (2018, p.19) señalan que:

Los datos recogen un conjunto de hechos (una base de datos, BD) y los patrones son expresiones que describen un subconjunto de los datos (un modelo aplicable a ese subconjunto). BD involucra un proceso iterativo e interactivo de búsqueda de modelos, patrones o parámetros. Los patrones descubiertos han de ser válidos, novedosos para el sistema (para el usuario siempre que sea posible) y potencialmente útiles.

[...] El objetivo final de todo esto es incorporar el conocimiento obtenido en algún sistema real, tomar decisiones a partir de los resultados alcanzados o, simplemente, registrar la información conseguida y suministrársela a quien esté interesado.

Ahora bien, a la par del desarrollo tecnológico moderno, se conforman los denominados entornos digitales, como “una generación de redes desde una plataforma que se apoya en la teoría conectivista y en la tecnología de la información y la comunicación desde escenarios analógicos” (Castro, 2022, p.74), conllevando los mismos a la formación de estructuras virtuales para el intercambio y la socialización masiva, en forma de datos tal como ya se ha señalado anteriormente, en cantidades enormes y de forma casi instantánea, en tiempo real, por lo cual no puede ser procesada totalmente utilizando los métodos tradicionales de análisis. En este contexto, se generan los llamados metadatos, básicamente aquellos “datos que describen la información contenida en otros datos” (Veas, 2021, p.02).

De manera resumida el autor antes citado, precisa que se puede definir a la ciencia de datos como (Veas, 2021, p.04):

[...] un Campo interdisciplinario que involucra métodos, procesos y sistemas de la Computación, Matemáticas, Estadísticas y el área a la cual se quiere aplicar, para investigar y sacar ventaja de la información que dan los datos.

La ciencia de datos se conforma así en la actualidad, como una herramienta fundamental para la explotación de datos y la generación de conocimiento. Entre sus objetivos se encuentran la exploración de modelos descriptivos de patrones y comportamientos a partir de los datos básicos de un fenómeno particular, con la finalidad de facilitar la toma de decisiones, o el desarrollar predicciones respecto a al comportamiento futuro de algunas variables de interés, como podrá ser, por ejemplo, el comportamiento esperado en el nivel de contagios por efecto del COVID-19 entre una determinada población ante la falta de servicios hospitalarios adecuados para ello.

La ciencia de datos ha experimentado un enorme desarrollo al facilitarse el acceso a grandes volúmenes de datos y su posibilidad de procesamiento en tiempo real, requiriendo para ello de

técnicas sofisticadas, al mismo tiempo que abarca a numerosos grupos de investigación de diferentes áreas, tales como la ingeniería, sociología, psicología y medicina, entre otros.

El Trabajo con Ciencia de Datos

La aplicación de la ciencia de datos en el marco de su aplicación fundamental, como lo es la resolución y predicción de problemas, no varía significativamente al enfocarla como herramienta de apoyo para los propósitos de investigaciones cualitativas, conllevando básicamente una serie de etapas que podrían resumirse de la siguiente manera (Díaz, Ventura y García, 2021): La comprensión de los datos, la extracción de sus propiedades, el modelado y análisis del problema, la presentación de resultados y el desarrollo de software para explotar el conocimiento extraído.

De una manera más pragmática, y acorde al propósito del presente artículo, Veas (2021) considera que, a modo general, el proceso de la ciencia de datos consta de cinco (05) etapas que pueden plantearse de la siguiente forma (pp. 06-10):

1- Recopilación de Datos (Data Collection): Aquí se extraen, coleccionan y almacenan los datos que se necesitan estudiar. Tarea que puede conllevar la aplicación de técnicas específicas, tales como:

- Web Scraping: Técnicas informáticas utilizadas para extraer datos de la internet. Usualmente simulan la navegación de un ser humano en páginas webs.

- Big Data: Busca formas de extraer y almacenar información de manera sistemática y lidiar con datos que son demasiado grandes o complejos para ser tratados por las técnicas tradicionales de procesamiento de datos, y;

- Data Mining: También conocida como minería de datos, en el caso particular de textos. Es el proceso para clasificar grandes conjuntos de datos, identificando patrones y estableciendo relaciones para resolver problemas con el análisis de datos.

2- Limpieza de Datos (Data Cleaning): Se limpian los datos, y se les da la estructura adecuada para los objetivos de análisis en los que serán utilizados, por ejemplo: Eliminar datos que no existen, Crear categorías condicionales para clasificar datos, Cambiar el tipo de dato (numérico, decimales, texto) y; Diferenciar entre datos cualitativos y cuantitativos.

3- Análisis Exploratorio de Datos (Exploratory Data Analysis): Se busca conocer y comprender los datos disponibles para elaborar planes de acción que permitan obtener información más relevante para los objetivos del análisis. Se suelen utilizar los dos enfoques fundamentales de la estadística: Estadística Descriptiva y la Estadística Inferencial. Para ambos enfoques se suelen

utilizar gráficos como histogramas, gráficos de barra, línea, punto, o de caja, entre otros, que permiten comprender y explorar mejor los datos a analizar.

4- Construcción e Implementación de Modelos (Model Building and Model Deployment): Se crean los algoritmos necesarios para modelar los datos y efectuar predicciones cuantitativas o cualitativas, dependiendo de los objetivos en los que se quieran utilizar los datos. La generalidad de los modelos se basa en tres enfoques: Clasificación, Predicción y Agrupación de Datos. Es aquí donde se utilizan las áreas de Inteligencia Artificial, dependiendo del enfoque que se quiera modelar de los datos, en las cuales se profundizará más adelante.

5- Reporte Gráfico (Data Visualization): Por lo general esta es la última etapa, y tiene como objetivo generar visualizaciones interactivas y simples para mostrar los resultados producidos de manera posterior al procesamiento de los datos. Con ello se busca facilitar la toma de decisiones y la comunicación de resultados a otras personas.

Áreas de Inteligencia Artificial

La Inteligencia Artificial, como un concepto muy abstracto, y difícil de desarrollar, puede ser definida de diferentes formas. No obstante, en términos generales, se le puede concebir como la capacidad de una máquina para imitar las funciones cognitivas propias de la mente humana, tales como: creatividad, sensibilidad, aprendizaje, entendimiento, percepción del ambiente y uso del lenguaje. También se le considera como una rama de las ciencias computacionales dedicada al diseño y construcción de sistemas capaces de realizar tareas asociadas con la inteligencia humana. (INCYTU,2018)

Ahora bien, las áreas de Inteligencia Artificial, contextualizadas desde el modelado de los datos, pueden ser establecidas de la siguiente manera según Veas (2021, pp. 08-09):

a) Aprendizaje Automático o Machine Learning: Utiliza modelos estadísticos y matemáticas para modelar datos a ser usados en tareas de Predicción, Agrupación o Clasificación.

b) Aprendizaje Profundo o Deep Learning: Va un poco más allá del anterior, aplicando modelos de vanguardia y en constante investigación para desarrollar tareas complejas tales como: Procesamiento del Lenguaje Natural (PLN), donde se busca que el computador comprenda el lenguaje humano; en Procesamiento de Imágenes o Videos, para que comprenda el contenido que hay en una imagen o video; en Sistemas Expertos, para que tenga conocimiento en áreas de alta complejidad y muy específicas, sirviendo muchas veces como un asistente consultor.

Como objeto de interés para el presente artículo, es necesario destacar que el área dedicada al PLN posee cada vez mayor aceptación en disciplinas relacionadas con altos volúmenes de datos, y su incidencia innegable sobre las personas y la sociedad, como lo son la sociología, psicología y la medicina, entre otros. El conjunto de técnicas que comprende el PLN facilita analizar y representar textos naturales mediante software y algoritmos en uno o diferentes niveles de análisis lingüístico, que pueden involucrar sentimientos, con la finalidad de obtener una apariencia humana en el procesamiento de lenguaje para tareas concretas (Liddy, 2001). A través de las técnicas relacionadas con el PLN se busca la aplicación de la inteligencia artificial para el análisis de datos de comportamiento, las cuales se desarrollan mediante el aprendizaje automático integrado, o técnicas de Embedded Machine Learning, tras la recolección de los datos (Rong y Cols., 2020).

c) Internet de las Cosas con IA (Internet of Things): Se enfoca en la conexión, a través del internet, con objetos cotidianos presentes en hogares, oficinas, o similares, los cuales pueden llegar a poseer algún grado de conocimiento del contexto en el que se encuentran y lograr comunicarse entre ellos.

Cabe destacar, que los algoritmos aplicados al Deep Learning, poseen estructuras matemáticas abstractas, las cuales suelen ser representadas como Redes Neuronales Artificiales, definidas como “Modelos complejos que buscan emular el funcionamiento del cerebro y su sistema nervioso”. (Rong y Cols., 2020, p.297)

Por otra parte, debe precisarse que la búsqueda de modelización del comportamiento humano no significa en lo absoluto el encontrar las leyes que lo rigen. Y tal como, la Ciencia de Datos busca patrones en sus unidades fundamentales, los datos per se, las Ciencias Sociales buscan patrones sociales y culturales en las suyas, es decir entre los seres humanos y sus fenómenos sociales; permitiéndose entonces, desarrollar una extrapolación de la interpretación del comportamiento de los datos hacia el comportamiento real de las personas y, por tanto, el de los patrones de datos hacia los patrones culturales de una sociedad.

Aspectos metodológicos

La metodología aplicada en el contexto del presente artículo, se corresponde con el denominado Análisis Comparativo Cualitativo, el cual, tal como lo señala Escott (2018):

[...]es un método enfocado a casos de estudio que habilita el estudio sistemático y formal de la causalidad, el cual fue creado con el propósito de proporcionar instrumentos para

optimizar estudios empíricos que buscan comparar una muestra pequeña de casos, donde la contrastación involucra cierto nivel de complejidad. (p.57)

De manera general, se fundamenta esta metodología en lo que se conoce como las “tres vías”, por las cuales se logra construir una proposición empírica, siendo estas: la experimental, estadística y comparativa (Caramani, 2009). En el caso particular, el artículo se enfoca en la aplicación de la vía comparativa, puesto que se desarrolla como un estudio sistemático, de las observaciones recogidas de entes sociales o momentos históricos de la sociedad, siendo representadas estas por las anécdotas vivenciales facilitadas por los profesores de dos universidades públicas venezolanas en un momento histórico de cambios profundos como el actualmente vivido en Venezuela, con el objetivo de estudiar similitudes y divergencias, así como inquirir sus causas. (Ariza y Gandini, 2012).

Para la aplicación de esta metodología, Escott (2018) precisa que debe abarcar tres momentos, siendo los mismos: Selección de Casos y su descripción, Análisis, y, por último, la Interpretación. Tomando en cuenta lo señalado, a continuación, se desglosan los dos primeros, mientras que la interpretación se aborda más adelante en el contexto de la presentación de los hallazgos del artículo.

Selección de casos

Se tomaron en cuenta los protocolos anecdóticos facilitados por los profesores universitarios participantes en una investigación cualitativa del tipo fenomenológica hermenéutica, señalada anteriormente, en un conjunto de cuatro (04), uno por cada profesor colaborador. Se toman estos protocolos, en lugar de las posteriores entrevistas fenomenológicas realizadas mucho más extensas y profundas, por ser generados a partir de la narrativa espontánea, no condicionada, ni afectada de modo alguno por el investigador, muy representativos de las vivencias y sentimientos de los participantes respecto al fenómeno abordado para ese estudio, básicamente: La Praxis Universitaria en el fomento de Participación Ciudadana.

Análisis

Para este momento, el cual es el más extenso llegando a significar el setenta por ciento (70%) del trabajo total, se inicia con la selección y aplicación de las herramientas relacionadas con la Ciencia de Datos y la Inteligencia Artificial, escogiéndose para esto la potencialidad que ofrece el Lenguaje de Programación “R” por ser un software de acceso libre, permitiendo además la implementación de algoritmos de manipulación y extracción de la información, la generación de

resultados, y su posterior visualización, contando además con librerías usadas en procesos de Inteligencia Artificial, acordes al propósito del artículo.

Este momento se puede desglosar de la siguiente forma:

1- *Recopilación de Datos (Data Collection)*: Se inicia con el desarrollo de los diferentes códigos de programación (Scripts), uno para el análisis individual y otro para el análisis unificado de los protocolos, esto a fin de evitar la confusión y los posibles conflictos durante la compilación de los mismos, enfocándolos a su vez en dos (02) grupos de tareas bien diferenciadas: “Generar relaciones existentes entre frases y palabras” tanto por Universidad como de forma global o unificada, y el “Análisis de Sentimientos” por profesor participante.

Previamente a la tarea de carga de los datos, se instalan y cargan las librerías que serán usadas durante todo el proceso de análisis, siendo estas, y su función, las presentadas en la Tabla 1. Posteriormente se cargan los protocolos desde la Base de Datos generada por la investigación de referencia, utilizando para ello la librería “pdftools”, esto debido a que los datos cualitativos recopilados fueron subidos a la red en formato “pdf” para evitar la pérdida de su estructura y facilitar el acceso o revisión pública de los mismos desde cualquier dispositivo electrónico.

Tabla 1

Librerías utilizadas para el Análisis de los Protocolos Anecdóticos

Librería	Función
tidiverse	Posee como núcleo un conjunto de ocho (08) librerías para la minería de datos, siendo estos:
	ggplot2 Para la gramática de gráficos, cubre los aspectos de exploración de datos y comunicación de conclusiones.
	dplyr Permite mutar o crear nuevas variables, seleccionar, filtrar, sumarizar, acomodar o agrupar las variables de las tablas (Data Frames o Tibles).
	readr Lectura de diferentes fuentes de datos, archivos de texto plano como csv, permitiendo integrarse las dos librerías anteriores, mediante concatenación de órdenes: %>% (pipes).
	purrr Explora una de las grandes funcionalidades de R como lo es el trabajo con vectores y funciones, o vectorización.
	tidyr Transformar filas y columnas de una tabla de datos para hacerlos más eficientes o tidy.
	tibble Transformar datos en un formato de tipo “tibble” siendo más eficiente que “Data Frame”.
	stringr y stringi Permiten tratar con cadenas de caracteres, quitar espacios extra, extrayendo o sustituyendo palabras mayúsculas a minúsculas, remover acentos, entre otras.
	forcats Para manejar factores y variables de tipo categórico.
tidytext	Realizar mineo de texto con el uso de herramientas tidy. Facilita la adecuación, limpieza y visualización de los datos.
pdftools	Extraer el texto de los documentos en formato pdf y guardarlo en un “Corpus de texto”.
stopwords	Ubicar palabras vacías, que no aportan nada al discurso, en varios idiomas.
syuzhet	Librería dedicada al análisis de sentimientos en varios idiomas.
wordcloud	Visualización de nubes de palabras.
scales	Escalar y reescalar las variables para facilitar la implementación de técnicas de Machine Learning.
widyr	Analizar qué palabras tienden a coaparecer en un texto determinado, comparación por pares (pairwise), un método que se emplea mucho en las ciencias sociales.

igraph y ggraph	Genera grafos, un dibujo que tiene vértices o nodos y aristas que unen los vértices entre sí.
quanteda	Facilita funciones para el análisis cuantitativo de texto.
topicmodels	Ajusta modelos temáticos (topic models) para detectar discursos particulares en el corpus.
cvTools	Escribir funciones para validación cruzada con mínimo esfuerzo de programación y asistencia del usuario en la selección de modelos.
RColorBrewer	Facilita el manejo de la paleta de colores en R

Fuente: Elaboración Propia

Una vez cargados los protocolos, se procede a unir entre si todos los párrafos de ellos para que conformen una sola unidad de texto, en el caso del análisis unificado, o global, se unieron previamente los cuatro (04) protocolos individuales. Seguidamente, se llevó a cabo la conversión del resultado anterior en un vector, que facilita su posterior uso como un Data Frame por frases de texto.

2- Limpieza de Datos (Data Cleaning): El paso anterior permite la limpieza del protocolo y su tokenización. Así se procede a remover palabras vacías y tokenizar por palabras usando las funciones “stopwords” y “unnest_tokens”. Las palabras vacías son aquellas que no son útiles para el análisis y usualmente incluyen artículos, pronombres, preposiciones, entre otras, mientras que usando la función “stopwords” se logra cargar un diccionario de palabras vacías preexistentes, en español para el presente caso. La tokenización en cambio se refiere a la estructuración de los datos de texto en filas, donde cada fila será un token, tal token puede ser una palabra, un n-grama, una oración, entre otras.

En este paso, también se genera el denominado “Lexicon de stopwords” en español, el cual consiste en el uso del diccionario anteriormente señalado, pero anexándole aquellas palabras, se observen no presenten ningún tipo de relación o contribución al análisis. Usando este Lexicon, se genera un nuevo Data Frame pero con un Corpus de Texto ya limpio y bien estructurado.

3- Análisis Exploratorio de Datos (Exploratory Data Analysis): Ya en este punto se puede proceder a realizar los primeros análisis básicos que permiten comenzar a depurar aún más la data, agregando palabras al Lexicon de stopwords, o reconfigurando los propios códigos de programa. Ahora bien, debido a que, como se mencionó anteriormente, se desarrollan para el estudio dos tareas distintas, como lo son primeramente encontrar las relaciones entre las frases y las palabras, se genera un conteo de palabras y su escalamiento para hallar su posterior coocurrencia, palabras que tienden a coaparecer en un texto determinado. Mientras que, para la segunda, relacionada con la ejecución del análisis de sentimientos, se hace necesario contar en el texto cuántas palabras tienen un valor positivo y cuántas negativo, para con eso establecer una valencia o valoración, que

puede ser positiva o negativa tomándose el valor de “cero” como neutra. En este último caso se debe primeramente cargar el diccionario de sentimientos adecuado.

No obstante, en español, tal como lo señala Fradejas (2018, s/p) “no se ha aplicado el análisis de sentimientos a textos literarios por el sencillo hecho de que no hay diccionarios disponibles ni investigadores interesados en el tema”, razón por la cual el autor adaptó al español los propios de la librería “tidytext”, creando uno personalizado al cual denominó “uva”, mismo que es usado en el presente estudio. Por otra parte, para evitar que se generen errores, al no ser parte de la librería original, se hace necesario cargar una “pequeña modificación de una de las funciones de {tidytext}...esta orden es leer un script e interpretarlo como una función diseñada por el usuario...get_sentiments” (Ibidem).

4- Construcción e Implementación de Modelos (Model Building and Model Deployment): Finalizada la etapa anterior, se lleva a cabo la Clasificación, Predicción y Agrupación de Datos, por medio de modelos matemáticos adecuados, usando para ello los algoritmos requeridos tanto para la coocurrencia de frases y palabras, así como para la valorización de sentimientos.

Respecto al desarrollo de los algoritmos requeridos para hallar la coocurrencia, Fradejas (2018) puntualiza que esto se logra aplicando el principio de Firth (1957), refiriendo en ese sentido lo siguiente (Fradejas, 2018, s/p):

[...]«conocerás una palabra por sus compañeras». Esto último es la base teórica del análisis de colocaciones: el significado de una palabra se basa en el significado de las palabras que la flanquean, puesto que las palabras tienen sus preferencias y suelen elegir como compañeras otras con las que forman combinaciones frecuentes.

De esta forma se crearon las líneas de código que permiten generar las relaciones existentes entre frases y palabras por los denominados Bigramas, o 2-gramas. Cualquier sistema de predicción como el usado en este estudio, se encuentra basado en modelos de lenguaje de N-gramas, que “...son secuencias de palabras de longitud N y la predicción se realiza a través de la probabilidad de cada N-grama en el modelo de lenguaje”. (Abadía, 2016, p.06)

Para el análisis de los sentimientos, las líneas de código desarrolladas cumplen con lo expuesto por Robinson y Silge (2019), cuando establecen que el ser humano al leer un texto, entiende la intención emocional de una palabra para inferir si una porción de texto es positiva o negativa, o incluso podría reconocer miedo o disgusto, y por tanto este mismo acercamiento puede realizarse aplicando técnicas de mineo de texto para la calificación y agrupación de datos. Aquí el resultado

depende en gran medida de la riqueza del diccionario, puesto que cuantas más palabras posea, más fina y precisa será la calidad del análisis, y es en este punto donde se facilita la tarea usando la librería “syuzhet”, aplicando de manera automatizada, por llamarlo de manera sencilla, una transformación matemática empleada para navegar entre el dominio del tiempo o espacio y el dominio de la frecuencia que se conoce como Transformada de Fourier.

5- Reporte Gráfico (Data Visualization): Aplicados los modelos desarrollados en la etapa previa, es procedente generar la visualización de los datos procesados según los propósitos planteados para el estudio, los cuales se presentan en forma de diagramas de barra, nube de datos y grafos de nodos y aristas en la siguiente sección como parte de los hallazgos del estudio.

Presentación de hallazgos

En esta sección se desarrolla el proceso comparativo de los resultados, más relevantes, obtenidos a partir del proceso metódico anteriormente desarrollado, junto a su respectivo análisis, confrontándolos relacionamente con los hallazgos emergentes encontrados por la investigación cualitativa de referencia, o de benchmarking, resaltándose datos significativos de las tablas, gráficos o figuras generadas desde ambos estudios. Este proceso se enfoca en las relaciones existentes entre frases y palabras, así como a la valoración de sentimientos, concernientes al fenómeno de la praxis universitaria y el fomento de la participación ciudadana para el desarrollo local, manifestados por los profesores universitarios participantes a través de sus respectivos protocolos anecdóticos.

Análisis relacional del texto protocolar

El análisis de los protocolos anecdóticos, exhibe de manera clara una diferencia de enfoque institucional, respecto al fenómeno, existente entre las universidades en estudio, siendo estas: Universidad Nacional Experimental “Simón Rodríguez”, Núcleo Valera y la Universidad Politécnica Territorial del Estado Trujillo “Mario Briceño Iragorry”, mismas que para efectos del trabajo analítico son denominadas como UNESR y UPTT respectivamente, mientras que los profesores participantes fueron identificados, codificados, como: UNESRS01, UNESRS02, UPTT01 y UPTT02.

Estas diferencias pueden identificarse en la grafica 1, donde se representan tanto la nube de palabras (Wordclouds) que presentan mayor frecuencia en el discurso de los profesores y su respectiva coocurrencia dentro de los protocolos como grafos de nodos y aristas, relacionados por Bigramas.

Grafica 2.

Palabras mas utilizadas

Stopwords retiradas

universidad
proyecto
realmente
servicio
comunidades
educación
proyectos
ley
genie
formación
desarrollo
verdad
comunitario
comunidad
unear
trabajo
muchas
extensión
años
universitaria
trujillo
siempre
c
participación
pala
momento
ejemplo
actividades
superior
participantes
nivel
niños
estudiantes
espacios
cultura
caso
atención

Numero de veces usada

Bigramas: Protocolos Anecdóticos Unificados

ciudadana
local
marco
participación
elementos
municipio
socia
ciudadano
hermoso
herramientas
hoy
sentido
decia
evaluación
tiempo
aprendizajes
objetivos
canima
educativo
momento
tecnológica
plan
cientifico
convenzo
ciencia

legislación
nominales
desarrollo
educación
proyectos
servicio
comunidad
participación
trabajo
muchas
extensión
años
universitaria
trujillo
siempre
c
participación
pala
momento
ejemplo
actividades
superior
participantes
nivel
niños
estudiantes
espacios
cultura
caso
atención

Numero de veces usada

Bigramas: Protocolos Anecdóticos Unificados

ciudadana
local
marco
participación
elementos
municipio
socia
ciudadano
hermoso
herramientas
hoy
sentido
decia
evaluación
tiempo
aprendizajes
objetivos
canima
educativo
momento
tecnológica
plan
cientifico
convenzo
ciencia

legislación
nominales
desarrollo
educación
proyectos
servicio
comunidad
participación
trabajo
muchas
extensión
años
universitaria
trujillo
siempre
c
participación
pala
momento
ejemplo
actividades
superior
participantes
nivel
niños
estudiantes
espacios
cultura
caso
atención

Numero de veces usada

En el Dashboard anterior, observamos cuales son las palabras más utilizadas por los profesores universitarios participantes en el estudio referencial, una vez que la totalidad de los textos protocolares fueron “limpiados”, en otras palabras, luego de ser depurados de palabras, signos gramaticales y espacios, entre otros, que no aportaban elementos relacionales importantes para el fenómeno en estudio (Stopwords). A partir de ellas se generó la respectiva nube de palabras unificada (Wordcloud) y su respectivo grafo de relación por Bigramas.

Revista *educare*

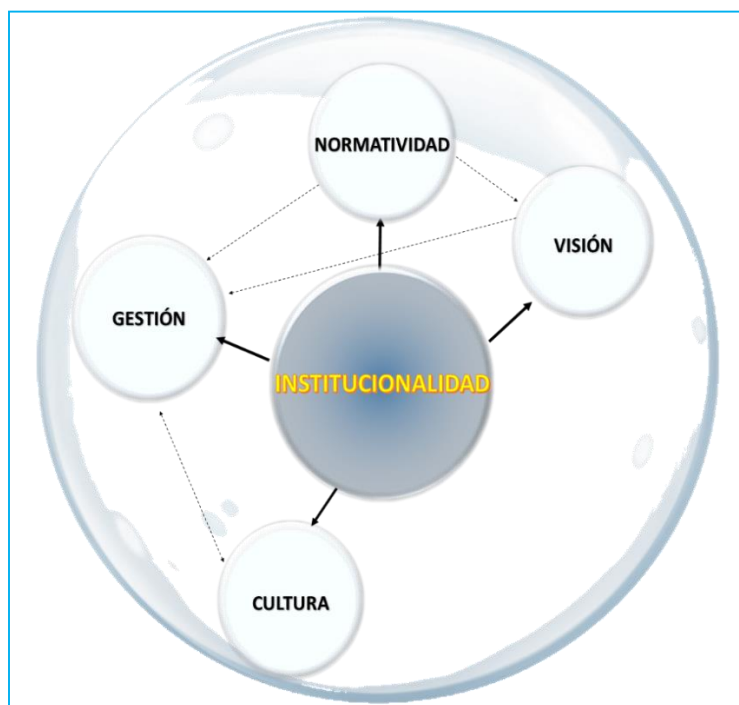
ISSN: 2244-7296

Depósito Legal: pp001002.A3674

facilidad observando la denominada “Red de Relaciones/Incidencia de Aspectos Institucionales”, en la gráfica 3, generada por la investigación cualitativa de referencia como parte de sus conclusiones.

Grafica 3.

Red de Relaciones/Incidencia de Aspectos Institucionales en la Investigación de Referencia



Fuente: Investigación referencial. Leal (2021)

Entonces, a partir de esta comparativa de aspectos cualitativos, resaltan los resultados coincidentes, por demás sorprendentes, al realizar un ejercicio de superposición de imágenes, lo cual puede corroborarse con una simple observación “gráfica a gráfica” por parte del lector, entre la gráfica 3 y el grafo de la gráfica 2. Esto apunta hacia la constitución de la universidad, como ente público complejo y de amplio impacto social, y sus diferentes procesos institucionales, conformándose como un núcleo central denso, “la institucionalidad”, justamente hacia la parte media de ambas imágenes.

Mientras que los aspectos relacionados con los temas de cultura organizacional, gestión, visión y normatividad, se hallan a su vez nucleados, cumpliéndose el sentido de “Rekursividad” planteado desde los Sistema Viables y el método Fenocomplejo, a su alrededor, y en comunicación bidireccional, respecto a la institucionalidad, casi en la misma ubicación espacial, de relevancia, en uno y otro caso. Es importante destacar aquí, la posición coincidente “casi perfecta”, en la parte

superior de cada gráfica, para el aspecto “Normatividad” denotando que los datos reflejan un mismo nivel de jerarquización y causalidad para ambos análisis.

Análisis de sentimientos del texto protocolar

Si bien en la investigación cualitativa referencial, no se ubicaron registros precisos de los sentimientos, o manifestaciones relacionadas a estos, entre los profesores universitarios participantes, se considera importante presentar el análisis de sentimiento arrojado por las herramientas que facilita la Ciencia de Datos usadas en la elaboración del presente artículo, como aporte a la comunidad científico – académica y la discusión crítica de la aplicabilidad de las mismas.

Para ello se presenta, primeramente, un resumen del Data Frame mejorado como tibble de veintitrés (23) filas por once (11) columnas (Tabla 2), que exterioriza las diferentes variables cualitativas relacionadas con los sentimientos a ser valorados según el diccionario, modificado para el español, utilizado en este caso de estudio, siendo nueve (09) en total, así como la valencia asignada a cada página del texto protocolar depurado para cada uno de ellos, con su respectiva valoración, negativa o positiva, por página. Debe notarse, tal como se lee al final del tibble, que no se muestran otras 13 filas de datos, ni las columnas correspondientes a las variables “premonición” y “tristeza”.

Tabla 2.

Resumen de los Datos Protocolares Depurados y Valorizados respecto a los Sentimientos

```
## # A tibble: 23 x 11
##   pagina alegría asombro confianza disgusto ira miedo negativo positivo
##   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl>   <dbl>   <dbl>
## 1     1     4     1     24     1     2     0     -3     29
## 2     2     5     3     17     1     4     4     -3     25
## 3     3     3     2     21     0     0     3     -2     26
## 4     4     2     3     18     0     3     0     -2     26
## 5     5     6     6     9     0     2     2     -3     17
## 6     6     7     3     23     0     0     1     -3     33
## 7     7    10    11     17     1     4     2     -6     22
## 8     8     7     4     22     2     2     3     -4     31
## 9     9    18     6     19     1     0     3     -4     36
## 10    10     3     3     15     3     4     6     -6     27
## # ... with 13 more rows, and 2 more variables: premonición <dbl>,
## #   tristeza <dbl>
```

Fuente: Elaboración Propia. Usando herramientas de Ciencia de Datos.

A partir de los datos resultantes, se puede presentar el siguiente Dashboard (Grafica 4), para su análisis global

responsabilidad con la que actúan las autoridades, en sus distintos niveles, afectando el registro, control, seguimiento, prosecución y éxito de los proyectos.

Otra particularidad observada, y que podría prestarse erróneamente a confusión, es que pareciera entonces no encontrarse manifestaciones negativas en los textos, tal como se observa en el cuadro inferior izquierdo que refleja la valoración del sentimiento por profesor dentro de su propio tiempo narrativo. Esto puede explicarse fácilmente, al observar más detalladamente los datos de la Tabla 2, allí se plasma que la cantidad de valoraciones como sentimientos positivos, en cada página, es mayor que la de los negativos.

Uno de los aspectos más representativos para este análisis, se encuentra en el recuadro inferior derecho, donde la Inteligencia Artificial facilita la “Historia Visual” de los sentimientos a lo largo de la narrativa completa, en forma de “Transformación Valorada del Sentimiento” respecto al “Tiempo Narrativo Total”, en otras palabras, como fluyó el participante. No obstante, aunque podría ser realizado un análisis comparativo cualitativo más profundo, cotejando estos datos con cada una de las páginas del texto protocolar, revisando concienzudamente las expresiones allí plasmadas, esto escapa a la intencionalidad del artículo.

Consideraciones finales

Finalizado el complejo proceso de análisis de datos cualitativos utilizando herramientas propias de la Ciencia de Datos y la Inteligencia Artificial, pueden desarrollarse algunas consideraciones generadas por la experiencia dual vivida, por una parte, la del investigador seducido por el fascinante mundo de las ciencias sociales aplicadas a los fenómenos complejos relacionados con el ser humano y la sociedad, y por otra, la del investigador formado en las ciencias ingenieriles amante de las tecnologías que favorecen el desarrollo humano.

Así, desde las facilidades que permiten la experiencia del trabajo académico y científico en ambos mundos, se podría sintetizar el resultado del presente estudio en sólo pocas palabras: La Ciencia de Datos y la Inteligencia Artificial son perfectamente aplicables, y recomendables, como apoyo al desarrollo de Investigaciones Cualitativas centradas en el estudio de Fenómenos Sociales Complejos.

Quienes se hayan dedicado a la aplicación de métodos de investigación cualitativa, en especial bajo el enfoque de la Fenomenología Hermenéutica o los Sistemas Viabes, entienden que estos se utilizan cada vez más durante la búsqueda del conocimiento en todas las disciplinas del saber, por

su capacidad para ayudar a los investigadores a comprender las múltiples perspectivas de los participantes a través de sus propias palabras, desde sus propias experiencias de vida y el entramado de relaciones que se suceden, y les afectan, desde el entorno inmediato, más próximo a sus personas, y el más externo que influye de forma cuasi intangible en el desenvolvimiento de sus propias vidas.

Sin embargo, estos también podrán dar fe que el análisis cualitativo es un proceso arduo y laborioso, demandante de mucha dedicación, disciplina y recursos. Situación esta, que conlleva a la limitación de los tamaños de muestra, el abordaje de casos más pequeños, en especial cuando conlleva a la adquisición de grandes volúmenes de datos, y en diferentes formatos, para lograr procesarlos como datos de texto, asegurando a su vez la calidad y profundidad de la investigación. Además de cumplir con la rigurosidad del método científico.

En ese sentido, se observa el potencial positivo que aporta el abordaje de estos análisis aplicando técnicas asociadas con el Procesamiento del Lenguaje Natural (PLN). El análisis de texto cualitativo comprende a los investigadores la obligación de leer y releer los datos para asignar códigos y desarrollar categorizaciones temáticas, de forma iterativa, siendo aquí donde el PLN ofrece el potencial de automatizar parte de este proceso al proporcionar una base para codificar cualitativamente de forma más rápida y eficiente, además de ofrecer un método para validar sus hallazgos cualitativos, obtenidos a través de los métodos y técnicas comúnmente aplicados.

Por otra parte, la magnitud de los datos que se manejan y la complejidad a la hora de revelar sus relaciones, hace necesaria la introducción de técnicas de visualización, mucho más precisas y potentes para facilitar la interpretación de estos y la identificación de patrones de comportamiento, que normalmente se escapan durante el proceso de lectura y relectura, por el agotamiento propio de la persona o las posibles distracciones del entorno.

Al respecto, debe recalcar el hecho que los datos son construcciones humanas y la forma de recoger y categorizar la información es fundamental para su correcta interpretación. Por otra parte, la disposición visual de los datos se corresponde a una intencionalidad, lo cual puede encubrir otras posibilidades de lectura. De allí, que sea pertinente ser muy críticos frente a la idea de un automatismo total, de la indiscutible objetividad o facticidad de los datos, así como de la neutralidad de sus visualizaciones. Por tanto, existe el riesgo de sobrevalorar la aplicación de la Ciencia de Datos y la Inteligencia Artificial en el análisis de fenómenos sociales y culturales al aceptar irreflexivamente estos procedimientos, detenerse a examinar y reformular, si es necesario,

la perspectiva teórica y metodológica asumida para su abordaje, o evitar en la medida de lo posible la importación acrítica de valores y supuestos que depositan toda la capacidad de decisión en procesos automatizados.

Vinculando los resultados con la discusión, este argumento apunta a que la Ciencia de Datos funciona perfectamente como un método que, basado en datos cuantitativos, puede inspirar, o coincidir con, los propósitos de un enfoque cualitativo, Fenomenológico Hermenéutico, o cualquier otro que confluya en Sistemas Complejos y de Sistemas Viables. Por lo general, los enfoques cualitativos no buscan probar teorías, sino producir entendimientos a partir de la interpretación de datos, siendo el proceso más inductivo que deductivo, allí confluyen también la Ciencia de Datos y el enfoque Hermenéutico al coincidir en que no se trata de demostrar teorías, sino de hallar patrones relevantes para la investigación.

Se debe tener precaución en la consideración de los resultados obtenidos a través de las técnicas facilitadas por la Ciencia de Datos, puesto que su objeto de estudio no son las personas, ni el comportamiento humano, sino el comportamiento y naturaleza de los datos. Por tanto, y por definición, no puede develar absolutamente nada acerca de lo humano, lo social o lo cultural, sólo puede aportar respecto a los datos y su comportamiento. Una cosa es reconocer en los datos patrones, y otra muy distinta es que estos puedan traducirse a patrones sociales o culturales, es decir, no se trata de que el análisis socio cultural explique los patrones identificados mediante la Ciencia de Datos, sino que éstos tengan sentido desde un análisis socio cultural.

Reflexiones finales

Ante un mundo que se mueve vertiginosamente hacia la digitalización y virtualización de todas sus actividades, por muy escéptico que se pueda ser ante esa realidad, los investigadores sociales se encuentran en la imperante necesidad de repensarse, y asumir el reto de formarse en el uso y aplicación eficiente de las herramientas que facilitan la Ciencia de Datos y la Inteligencia Artificial para la adquisición, el procesamiento, análisis y la presentación de los datos relacionados con sus estudios, tanto de carácter cuantitativo como cualitativo. Aprender a cómo sacarles el mayor provecho en aras de desarrollar trabajos de mayor calidad y profundidad, buscando que al mismo tiempo sus hallazgos puedan ser más seductores y comprensibles al lector común, sin perder su validez y confiabilidad científica, o su utilidad académico-práctica.

No obstante, no se puede pretender abandonar por completo las prácticas y procedimientos propios de los métodos cualitativos, por muy complejos y agobiantes que estos sean, puesto que

son los que permiten el contacto con la realidad humana, con el fenómeno social observado y su complejidad. Ninguna máquina, software o tecnología, por muy sofisticada que sea, podrá jamás sustituir la capacidad de análisis, reflexión, abstracción, y adaptación del conocimiento, entre otras, facultades propias del ser humano, por muy impresionantes que se muestren los avances alcanzados a la fecha.

En este sentido las herramientas de Ciencia de Datos e Inteligencia Artificial, como las planteadas en el presente artículo, deben ser usadas como guías, o facilitadoras, de una visión panorámica inicial de los fenómenos complejos bajo estudio, para la generación de categorías y temas de análisis inicial, la profundización de aspectos o patrones emergentes, así como para la observación de las diferentes aristas y relaciones necesarias a ser tomadas en cuenta en el contexto de la complejidad fenomenológica, entre otras: Son entonces sólo eso, “Herramientas” a la disposición del investigador.

Referencias

- Abadía, M. (2016). *Sistema de diálogo para la Comunicación Aumentativa y Alternativa con Pictogramas*. Tesis de Maestría. Publicada. Universidad de Zaragoza. <https://zaguan.unizar.es/record/60479?ln=es>
- Ariza, M., y Gandini, L. (2012). *El análisis comparativo cualitativo como estrategia metodológica*. Ariza, Marina y Velasco, Laura (Coords.), Métodos cualitativos y su aplicación empírica. Por los caminos de la investigación sobre la migración internacional. México: Instituto de Investigaciones Sociales y Colegio de la Frontera Norte. <https://www.iis.unam.mx/metodos-cualitativos-y-su-aplicacion-empirica-por-los-caminos-de-la-investigacion-sobre-migracion-internacional/>.
- Beer, S. (1979). *Brain of the firm*. Chichester, UK: John Wiley & Sons Ltda.
- Bertoldi, S., Fiorito, M., y Álvarez, M. (2006). Grupo Focal y Desarrollo local: Aportes para una articulación teórico-metodológica. *Ciencia, docencia y tecnología*. 17(33),111-131. <https://www.redalyc.org/pdf/145/14503304.pdf>
- Caramani, D. (2009). *Quantitative Applications in the Social Sciences: Introduction to the comparative method with Boolean algebra* Thousand Oaks, CA: SAGE Publications Ltd. doi:10.4135/9781412984690

- Castro, V. (2022). Las inteligencias múltiples en los entornos digitales de la unidad educativa fiscomisional “Sathya Sai”. *Revista EDUCARE*. 26(1),70-83.
<https://doi.org/10.46498/reduipb.v26i1>
- Davison, K. (2006): Dialectical imagery and postmodern research. *International Journal of Qualitative Studies in Education*. 19 (2), 133-146. <https://eric.ed.gov/?id=EJ734624>
- Díaz, M., Ventura, S. y García, S. (2021). *Ciencia de Datos y Big Data DaSCI*. Documento en línea. Disponible en: <https://dasci.es/es/investigacion/unidades/ciencia-de-datos-y-big-data/>. Consulta: Agosto 10, 2021.
- Escott, M. (2018). Introducción al Análisis Cualitativo Comparativo como técnica de investigación. *Revista DIGITAL CIENCIA@UAQRO*. 11(1), 57-66.
https://www.uaq.mx/investigacion/revista_ciencia@uaq/ArchivosPDF/v11-n1/art4_numerada-VF.pdf
- Fradejas, J. (2018). *Análisis de textos y estilometría con R, (AnaText)*. Documento en línea. Disponible en: <https://github.com/7PartidasDigital/AnaText>.
- García, J. Molina, J., Berlanga, A., Patricio, M., Bustamante, A. y Padilla, W. (2018). *Ciencia de datos. Técnicas analíticas y aprendizaje estadístico*. Bogotá: Alfaomega Colombiana S.A.
- Hodder, I. (2000). *The interpretation of documents and material culture*. En N. K. Denzin y Y. S. Lincoln (Eds.). *Handbook of Qualitative Research*. London: Sage Publications.
- INCYTU (2018). *Inteligencia Artificial*. Documento en línea. Disponible en: https://www.foroconsultivo.org.mx/INCYTU/documentos/Completa/INCYTU_18-012.pdf.
- Leal, J. (2021). *Holopráctica universitaria para la participación ciudadana en el desarrollo local*. Tesis Doctoral. Publicada. Universidad Nacional Experimental “Simón Rodríguez”. Valera.
<https://unesrdsb.wixsite.com/bibliovirtual-unesr/publicaciones-de-la-unesr>
- Liddy, E. (2001). *Natural language processing*. En *Encyclopedia of Library and Information Science*, 2nd Ed. NY: Marcel Decker, Inc.
- Robinson, D. y Silge, J. (2019). *Text Mining with R: A Tidy Approach*. 1st Ed. O'Reilly Media
- Rong, G., Mendez, A., Assi, E., Zhao, B., y Sawan, M. (2019). Artificial Intelligence in Healthcare: Review and Prediction Case Studies. *Engineering*. (6)3,291-301.
<https://doi.org/10.1016/j.eng.2019.08.015>
- Smith, J. y Deemer, D. (2000). *The problem of criteria on the age of relativism*. En N. K. Denzin y Y. S. Lincoln (Eds.). *Handbook of Qualitative Research*. London: Sage Publications.

<http://www.daneshnamehicsa.ir/userfiles/files/1/9->

[%20The%20SAGE%20Handbook%20of%20Qualitative%20Research.pdf](#)

Van Manen, M. (2003). *Investigación Educativa y Experiencia vivida. Ciencia humana para una pedagogía de la acción y de la sensibilidad*. Barcelona: Idea Books

Veas, C. (2021). *Data science para todos!* <https://medium.com/datos-y-ciencia/data-science-para-todos-4cb84264bb5f>.