Análisis de datos de la viruela símica

Kang Cheng Lei Lei
Departamento de Ingeniería de Sistemas
Universidad del Norte
Barranquilla, Colombia
leik@uninorte.edu.co

Julio Manuel Pérez Suárez
Departamento de Ingeniería de Sistemas
Universidad del Norte
Barranquilla, Colombia
juliomp@uninorte.edu.co

Jhon Naider Natera Ariza
Departamento de Ingeniería de Sistemas
Universidad del Norte
Barranquilla, Colombia
jnnatera@uninorte.edu.co

Issa David Dovale Morales
Departamento de Ingeniería de Sistemas
Universidad del Norte
Barranquilla, Colombia
idovale@uninorte.edu.co

Wilson Nieto Bernal (Tutor)
Departamento de Ingeniería de Sistemas
Universidad del Norte
Barranquilla, Colombia
wnieto@uninorte.edu.co

Andrea Cristina Zapata Delgado (Tutora)
Departamento de Ingeniería de Sistemas
Universidad del Norte
Barranquilla, Colombia
aczapata@uninorte.edu.co

***Abstract* — This project seeks to develop an application to perform the analysis of datasets with information about the simian smallpox virus, we intend to use data analysis techniques and pattern recognition (Data Mining). We will analyze how the virus behaves and its spread, and based on that we will obtain statistical indicators, then we will apply different linear regression methods to make predictions about the spread worldwide.**

***Keywords* — Monkeypox, Artificial Intelligence, disease outbreak, data analysis, correlation, predictive analysis, supervised learning, feature extraction, linear regression, Kfold, Overfitting, Polynomial Features, Data Mining, epidemic, pandemic.**

***Resumen* — En este proyecto se busca desarrollar una aplicación para realizar el análisis de datasets con información sobre el virus de la viruela símica, se pretende utilizar técnicas de análisis de datos y de reconocimiento de patrones (Minería de datos). Analizaremos cómo se comporta el virus y su propagación, y con base a eso obtendremos indicadores estadísticos, luego aplicaremos distintos métodos de regresión lineal para realizar predicciones sobre la propagación a nivel mundial.**

***Palabras clave* — Viruela símica, Inteligencia Artificial, brote de enfermedad, análisis de datos, correlación, análisis predictivo, aprendizaje supervisado, extracción de características, regresión lineal, Kfold, Overfitting, Polynomial Features, minería de datos, epidemia, pandemia.**

I. INTRODUCCIÓN

Las pandemias y epidemias han afectado a la humanidad desde hace un buen tiempo, siendo capaces de causar millones de muertes. Por suerte, el avance tecnológico que se da hoy en día es considerable con respecto a décadas pasadas. Esto invita al uso de nuevas herramientas tecnológicas para combatir y/o entender las pandemias y otras enfermedades que se presentan hoy en día. Además, existen estudios que confirman que el Covid-19 ha acelerado el uso de tecnologías digitales e inteligentes (Moosavi et al., 2021).

La epidemia en el que nos enfocaremos se llama la viruela símica, si bien no es tan peligrosa como otras, en algunas personas, las infecciones pueden provocar complicaciones médicas e incluso la muerte (Organización Mundial de la Salud [OMS], 2022). Por ende, no hay que tomar esto a la ligera, para entender mejor el comportamiento de este virus y sacar algunas conclusiones al respecto, será necesario hacer un buen uso de herramientas de análisis.

Para este trabajo, comenzaremos con una descripción detallada del problema, explicando la mayoría de los riesgos que puede traer la viruela símica. Luego, se justificará el porqué de la elección de esta problemática. Más adelante, nos encontraremos con el objetivo general y los objetivos específicos de este proyecto. Adicionalmente, en el marco teórico nos referiremos a acontecimientos similares que pasaron antes para entender la gravedad del trabajo. Después, en el marco conceptual estarán las definiciones de las palabras claves. Seguidamente, está la metodología de desarrollo en donde se tomará como referencia el marco de trabajo CRISP-DM, también está la metodología de investigación de carácter cuantitativo y sus fases para un desarrollo exitoso del proyecto. Luego, en el modelo de requerimientos de analítica se mostrará el listado de requerimientos resueltos en este trabajo. En el modelo de datos se realiza una descripción de los datos y cómo se relacionan. Más adelante, se mostrarán de forma gráfica las arquitecturas lógica y física de la solución. Posteriormente, se hablará acerca de la implementación de la solución, allí se explicará cómo funciona el prototipo. Después, en la revisión sistemática de la literatura compartiremos las tablas con los resultados de búsqueda y filtrado. Luego se mostrará la tabla de valoración del prototipo. Para finalizar, se hará la conclusión de este trabajo, haciendo énfasis en los objetivos planteados.

Para la parte práctica, el desarrollo que proponemos es el de crear una herramienta que analice los casos de viruela símica alrededor del mundo. Para lograrlo es necesario encontrar datasests confiables y que sirvan para el propósito del proyecto, en nuestro caso lo elegiremos de la página de Kaggle. Luego, escogeremos a Python como lenguaje de programación y a Jupyter como herramienta para poder usar Python. Después, se utilizarán solo los datos necesarios de los datasets para crear diversos gráficos que nos ayuden a hacer un análisis de estos. Como los datos se generan a partir de diversas fuentes y también los tipos de datos varían, es indispensable encontrar algoritmos que puedan adaptar las características de los datos. Entonces, eligiendo los algoritmos de minería de datos que más se adapten, podemos identificar patrones de datos y elaborar predicciones, ya que

la precisión de la predicción es alta en este caso (Deepa et al., 2022). Por último, presentaremos los resultados haciendo uso de visualización analítica.

De esta forma, por medio de este trabajo se espera aportar respuestas y conclusiones para que las personas puedan conocer aún más acerca de este virus y que otros trabajos o investigaciones puedan usar a este de referencia. Este trabajo será útil para que los formuladores de políticas tomen acciones inmediatas para mitigar el peligro de la pandemia y mejorar el bienestar humano en las ciudades y, a largo plazo, los ayudará a estar mejor preparados para manejar futuras pandemias (Rahman et al., 2021).

II. DESCRIPCIÓN DEL PROBLEMA

Actualmente, afrontamos una epidemia que puede convertirse en pandemia conocida como la viruela símica, que según la OMS (2022) "*Es una enfermedad causada por el virus de la viruela símica. Se trata de una infección zoonótica vírica, lo que significa que puede propagarse de animales a seres humanos. También puede propagarse de persona a persona.*", esta epidemia una vez infecta a una persona hace que este presente síntomas que pueden llegar a ser letales, "*Los síntomas más comunes de la viruela símica son fiebre, cefalea, dolores musculares, dolor de espalda, falta de energía y ganglios linfáticos inflamados. A estos síntomas les sigue o acompaña una erupción que puede durar de dos a tres semanas. La erupción se puede ubicar en la cara, las palmas de las manos, las plantas de los pies, los ojos, la boca, el cuello, la ingle y las regiones genitales o anales del cuerpo*" (OMS, 2022). Esta epidemia se esparce continuamente por distintos países alrededor del mundo y por distintos medios, según la OMS (2022) "*La viruela símica se propaga de persona a persona mediante contacto directo con alguien que tiene una erupción cutánea de viruela símica, en particular mediante contacto cara con cara, piel con piel, boca con boca o boca con piel, incluido el contacto sexual*".

El propósito de este estudio es desarrollar una aplicación que realice un análisis de datos de la propagación del virus símico en las personas de distintos países. Como instrumento de recolección de datos se utilizarán datasets encontrados en Kaggle. Se realizará una estructuración de los datos haciendo uso de Python, luego se usarán indicadores estadísticos y técnicas de minería de datos para hacer el análisis de datos y predicción de los datos, y una vez que tengamos estos resultados pensamos presentarlos haciendo uso de visualización analítica y distintos gráficos con los indicadores estadísticos.

III. JUSTIFICACIÓN

Para la población en general, es importante mantenerse al tanto de la situación por la que está pasando el mundo respecto al virus de la viruela símica para principalmente, evitar cometer los errores que se cometieron durante la propagación del virus COVID-19, los cuales se dieron principalmente por la inexperiencia que se tenía con pandemias de esta magnitud, y por ende causaron problemas en los sectores de la economía y la salud (Poongodi, M., Malviya, M., Hamdi, M., Rauf, H. T., Kadry, S., & Thinnukool, O., 2021), además de que con los estudios realizados por la experiencia con el COVID-19, se pudo notar la eficacia de las intervenciones no farmacéuticas como un enfoque acertado para reducir propagaciones del virus.

El desarrollo de esta aplicación ayudaría a que los datos, que vayan saliendo sobre el flujo de contagios de la viruela símica, se presten para soluciones eficientes con respecto a posponer e incluso evitar la propagación del virus, ya que como se menciona en la OMS, la viruela símica se presenta principalmente en zonas de selva tropical de África central y occidental y, esporádicamente, se exporta a otras regiones. (OMS, 2022).

El virus es similar en ciertos aspectos al COVID-19 en términos de duración de síntomas, siendo esta de 2 a incluso 4 semanas, la cual la hace una enfermedad autolimitada, pero que puede tener cuadros graves con mayor frecuencia en los niños, y su evolución depende del grado de exposición al virus, el estado de salud del paciente y la naturaleza de las complicaciones, con una tasa de letalidad de un 6% máximo.

Es por esto que se le debe dar gran importancia a estos datos, resultado de trabajos que utilizaron minería de datos e inteligencia artificial, no solo para el presente, sino para el futuro (Moosavi et al., 2021).

IV. OBJETIVOS

A. Objetivo General

Desarrollar una aplicación con base en analítica de datos para analizar el comportamiento de la propagación de la viruela símica en distintos países.

B. Objetivos Específicos

- Identificar los componentes claves asociados con la analítica de datos en el contexto de las pandemias y epidemias, a través de la revisión sistemática de la literatura.
- Desarrollar el modelo y el diseño de una aplicación para hacer análisis de datos sobre la viruela símica.
- Desarrollar el prototipo de la herramienta de análisis de datos de la viruela símica.
- Validar que el prototipo de la solución realice un correcto análisis de los casos de viruela símica.

V. MARCO TEÓRICO

Las pandemias y muchas otras enfermedades condicionan la calidad de vida humana (Avuçlu, 2022), el Coronavirus que sucedió recientemente ha causado desgracias críticas en varios campos como la económica, la de salud, dejando un ambiente de supervivencia para muchas personas (Poongodi et al., 2021), también ha devastado al mundo matando a millones de personas (Chrin & Wang, 2021). Además, reveló fragilidades en varias áreas, incluida la capacidad limitada de los sistemas de salud pública para el seguimiento y la notificación eficiente de casos. Según (Pinheiro, C. A. R., Galati, M., Summerville, N., & Lambrecht, M., 2021) El enfoque más eficaz para reducir la propagación del virus y evitar un colapso sustancial del sistema sanitario, en ausencia de vacunas, son las intervenciones no farmacéuticas (NPI), como la aplicación de restricciones de contención social, el control de la movilidad general de la población, la aplicación de pruebas virales generalizadas y el aumento de las medidas de higiene.

Tanto organizaciones gubernamentales como no-gubernamentales, han estado trabajando para implementar tecnologías de Análisis de Datos para la pandemia actual (Jalil & Ei Leen, 2022). Los avances innovadores son importantes para las pandemias que vienen, ya que pueden facilitar nuevas soluciones a las próximas dificultades (Poongodi et al., 2021). Las teorías sociales y psicológicas de la salud sugieren un conjunto limitado de predictores, por

esto se opta por utilizar técnicas de minería de datos, porque por ejemplo durante epidemia del COVID se utilizaron para clasificar 115 candidatos que se correlacionan con el comportamiento de prevención de infecciones de 56 072 participantes de 28 países, gracias a esto fue predicho el 52% de la varianza del comportamiento de prevención de infecciones. (Van Lissa, C. J., Stroebe, W., Leander, N. P., Agostini, M., Draws, T., Grygoryshyn, A., & Bélanger, J. J., 2022).

A causa del Covid-19 se está acelerando el uso de la inteligencia artificial para hallar soluciones a las nuevas pandemias (Moosavi et al., 2021).

Ahora el virus que está llamando la atención de todos es el de la viruela símica, debido a que está aumentando los casos en distintos países. Aunque, es un tema muy hablado en la actualidad, no es nada nuevo. Según la OMS (2022), este agente patógeno se detectó por primera vez en los seres humanos a principios de la década de los 70 en la República Democrática del Congo, que hasta el día de hoy el virus no ha dejado de propagarse en los alrededores. Sin embargo, en mayo de 2022 se detectaron numerosos casos de viruela símica en varios países no endémicos, siendo esto algo preocupante. Actualmente, se están haciendo estudios para comprender mejor las fuentes de infección, la epidemiología y las características de la transmisión.

Ahora con el virus de la viruela del simio, es importante estar preparados y brindar no solo medidas preventivas, sino también estrategias y soluciones digitales efectivas para defender a nuestra población de futuros brotes (Chet Ng et al., 2022).

VI. MARCO CONCEPTUAL

Viruela símica

La viruela símica es una enfermedad causada por el virus de la viruela símica. Es una infección viral de animal a humano. También se puede propagar de persona a persona (OMS, 2022).

Inteligencia Artificial

Se refiere a las operaciones de la máquina que imitan la inteligencia humana, como la resolución de problemas y el aprendizaje. La inteligencia artificial permite que las máquinas y los sistemas tomen decisiones y actúen como humanos (Moosavi et al., 2021).

Brote de enfermedad

Un brote es un aumento repentino del número de casos de una enfermedad. Un brote puede ocurrir en una comunidad o área geográfica, o puede afectar a varios países. Puede durar unos días o semanas, o incluso varios años (APIC, 2022).

Análisis de datos

Las herramientas de análisis de datos juegan un papel importante en la construcción de conocimientos que son vitales para la toma de decisiones y las medidas de precaución, habiendo una plétora de métodos de análisis de datos a disposición de todos (Jalil, N. A., & Ei Leen, M. W., 2022).

Análisis predictivo

Es una rama del análisis avanzado que realiza predicciones sobre resultados futuros usando datos históricos junto a modelos matemáticos, minería de datos y técnicas de Machine Learning, normalmente se le asocia con Big data y ciencia de datos (IBM, 2022).

Aprendizaje supervisado

Uso de conjuntos de datos marcados para entrenar algoritmos que clasifiquen datos o para predecir resultados con precisión. A medida que los datos de entrada se introducen en el modelo, este ajusta sus cálculos hasta que el modelo se haya ajustado adecuadamente, esto ocurre en partes del proceso de validación cruzada (IBM, 2022).

Extracción de características

Se enfoca en seleccionar un subconjunto de características más relevantes para desarrollar un modelo robusto de aprendizaje automático. En el proceso de selección de características, los datos redundantes y/o irrelevantes se eliminan de la base de datos principal, por lo que el rendimiento del modelo de diagnóstico puede mejorar (Malik, H., Fatema, N., & Iqbal, A., 2021).

Regresión lineal

El análisis de regresión lineal se utiliza para predecir el valor de una variable en función del valor de otra. La variable que se quiere predecir se llama variable dependiente. La variable que se utiliza para predecir el valor de la otra variable se denomina variable independiente. Esta forma de análisis estima los coeficientes de la ecuación lineal, en la que intervienen una o varias variables independientes, que mejor predicen el valor de la variable dependiente (IBM, 2022).

Validación cruzada Kfold

Método estadístico que consiste en dividir un conjunto de datos en k segmentos, luego a cada segmento se le realiza una prueba del modelo utilizando una parte del segmento para entrenar los datos y la otra parte para realizar pruebas, de esta forma se hace con todos los segmentos (AWS, 2022).

Overfitting

Cuando se entrena a un sistema de Machine Learning y se busca que las predicciones sean perfectas, se corre el riesgo de alimentar al modelo con “ruido” o peculiaridades de los datos, esto ocasiona que el modelo se enfoque en estos en lugar de encontrar una regla de predicción general (Ying, X., 2019).

Polynomial Features

Las características polinómicas, como su nombre lo indica, son características polinómicas que se hallan usando las características existentes de nuestro modelo y elevándolas a un exponente. Estas características nos ayudan en el mejoramiento de la precisión de nuestros modelos (Jason Brownlee, 2020).

Minería de datos

La minería de datos (Data Mining) es un proceso técnico, tanto automático como semiautomático, utilizado con el propósito de analizar grandes cantidades de información dispersa para darle sentido y convertirla en datos pertinentes para el proyecto que se esté realizando, ya sea buscando anomalías, patrones o correlaciones entre millones de registros para predecir resultados (Iberdrola, 2022).

Epidemia

Se refiere a la enfermedad que azota un gran número de personas o animales en un mismo lugar en un período determinado (Castañeda & Ramos, 2020).

Pandemia

El término pandemia significa epidemia que se extiende a muchos países y ataca a muchos individuos en una región (Castañeda & Ramos, 2020).

VII. METODOLOGÍA DE DESARROLLO

Para el avance y la resolución del proyecto, se tomó como referencia el marco de trabajo CRISP-DM (Cross Industry Standard Process for Data Mining), el cual "proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos" (Sngular, 2019), asignando tareas concretas y definiendo lo que es deseable obtener tras cada fase, teniendo en cuenta la aplicación al entorno de negocio de los resultados.

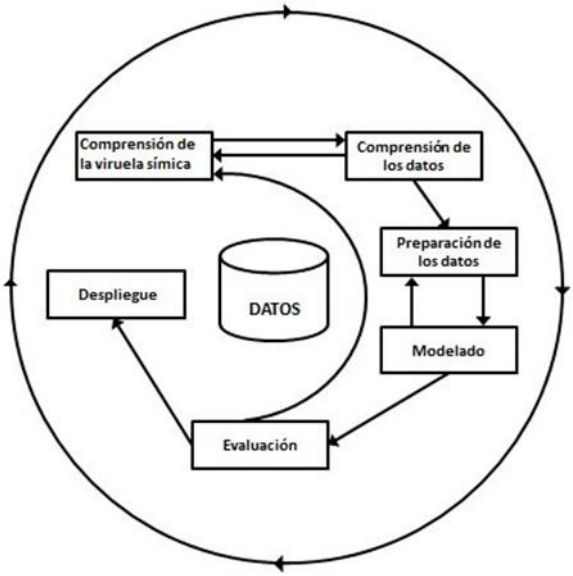


Figura 1. Metodología CRISP-DM

Fase I. Comprensión de la viruela símica: En esta primera fase se enfoca en la comprensión de los objetivos del proyecto. Luego, viene convertir este conocimiento en la definición de un problema y en un plan preliminar diseñado para alcanzar los objetivos anteriormente planteados (Sngular, 2019).

Fase II. Comprensión de los datos: Durante esta fase se inicia con la obtención de los datos, después nos encargamos de realizar tareas que nos ayudan a hallar algún problema presente en los datos, acostumbrarnos a los datos o descubrir otros conjuntos de datos que nos ayuden a obtener otras conclusiones (Sngular, 2019).

Fase III. Preparación de los datos: En esta fase nos enfocamos en realizar todo tipo de actividades que nos van a permitir depurar los datos para obtener el conjunto final que vamos a utilizar. Entre las actividades tenemos elección de tablas, registros y atributos o la transformación y la limpieza de datos (Sngular, 2019).

Fase IV. Modelado: Para esta fase nos enfocamos en seleccionar las técnicas de modelado que vamos a aplicar posteriormente, ya que para los problemas de minería de datos se suelen implementar modelos parecidos, existen algunos que requieren datos específicos sobre la estructura de los datos y es por esto que muchas veces se regresa a la fase de preparación de datos (Sngular, 2019).

Fase V. Evaluación: Al llegar a esta etapa se espera contar con por lo menos un modelo que satisfaga los requisitos desde el punto de vista del análisis de datos. Es importante que antes de pasar al despliegue se realice una comparación entre los objetivos de negocio y si el modelo cumple con todos ellos, también es importante revisar los pasos que nos llevaron a dar con ese modelo (Sngular, 2019).

Fase VI. Despliegue: Aunque considerada la fase final del proyecto en general, esto no es cierto, ya que dependiendo los requisitos, esta fase puede ser muy simple o muy compleja (Sngular, 2019).

Como conclusión, la metodología CRISP-DM establece que el proyecto no solo no acaba una vez se halla el modelo idóneo, ya que después se requiere un despliegue y un mantenimiento, además de estar relacionado con otros proyectos. "Es preciso documentarlo de forma exhaustiva para que otros equipos de desarrollo utilicen el conocimiento adquirido y trabajen a partir de él" (Sngular, 2019).

Con base en esta premisa, el desarrollo del proyecto nos permitiría adaptarnos de forma natural a las condiciones cambiantes y a los requisitos que este requiera, cubriendo las fases del proyecto, sus tareas respectivas, y las relaciones entre estas tareas, a fin de cuentas, CRISP-DM es una metodología que se utiliza de una forma u otra, en los proyectos de análisis de datos que se pretendan abordar con seriedad y asegurando la calidad de los resultados.

VIII. METODOLOGÍA DE INVESTIGACIÓN

Para garantizar resultados válidos y fiables que respondan a las metas y los objetivos del proyecto, necesitamos una metodología de investigación cuantitativa. Estas "son aquellas con las que se pueden obtener datos cuantitativos o medibles. Su importancia es que pueden validarse con modelos y principios científicos, pero pueden llegar a ser inflexibles y frías." (Enago Academy, 2021).

Fase I. Revisión sistemática de la literatura: Esta fase se enfoca en realizar una revisión de los aspectos cuantitativos de los estudios realizados con los datasets, para así resumir la información existente de estos en nuestra investigación con respecto a la viruela símica, de esta forma, se analiza y se compara lo obtenido con otras fuentes.

Fase II. Desarrollo de la arquitectura de la solución: Durante esta fase se abarcará la arquitectura del proyecto, en específico, los sistemas, información, seguridad, aplicaciones y tecnología de esta, para así estudiar las herramientas más viables para el desarrollo de este mismo.

Fase III. Desarrollo del prototipo de la solución: Se comienza a desarrollar la aplicación haciendo uso de la arquitectura seleccionada en la fase anterior y se presenta un prototipo que intente cumplir con los objetivos del proyecto.

Fase IV. Validación del prototipo de la Solución: Hacemos pruebas de validación en el prototipo de la solución y se regresa a la fase III en caso de no cumplir con los objetivos.

Fase V. Conclusiones y resultados: Una vez el prototipo pase todas las pruebas y validaciones, podemos empezar a realizar conclusiones sobre los resultados que obtenemos al hacer uso de esta.

IX. MODELO DE REQUERIMIENTOS DE ANALÍTICA

Lo siguiente es un listado de los requerimientos las cuales se resolvieron en este trabajo:

- Top 8 países con más casos.

- Número total acumulado de casos por país.

- Número de casos por día en países.

- Número global total acumulado de casos de viruela símica

- Predicción número global total acumulado de casos de viruela símica

X. MODELO DE DATOS

Los datasets que se usaron fueron "Data Science with world countries" (TEK, 2022) y "Monkeypox Dataset" (CONTRACTOR, 2022).

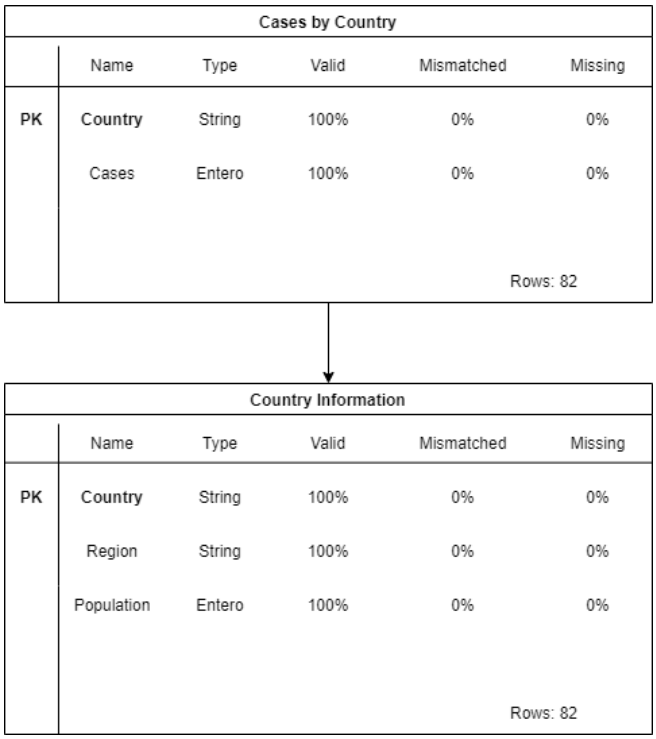


Figura 2. Modelo de datos

XI. ARQUITECTURA LÓGICA DE LA SOLUCIÓN

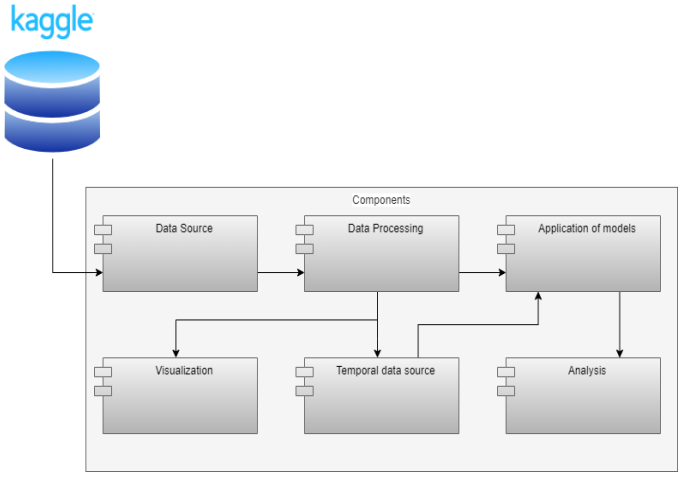


Figura 3. Arquitectura Lógica

Data Source: En esta capa se hace la búsqueda de datasets en Kaggle que sirvan para el propósito del trabajo.

Data Processing: En esta capa se hace una limpieza de los datos de los datasets seleccionados.

Visualization: En esta capa se visualizan los datos por medio de distintos tipos de gráficos.

Appication of models: En esta capa se aplican los modelos que ayudan a predecir el comportamiento de los datos.

Temporal data source: En esta capa se crean tablas temporales derivados del procesamiento de datos, para usarse en la aplicación de modelos.

Analysis: En esta capa se muestran los resultados de forma gráfica de los modelos aplicados.

XII. ARQUITECTURA FÍSICA DE LA SOLUCIÓN

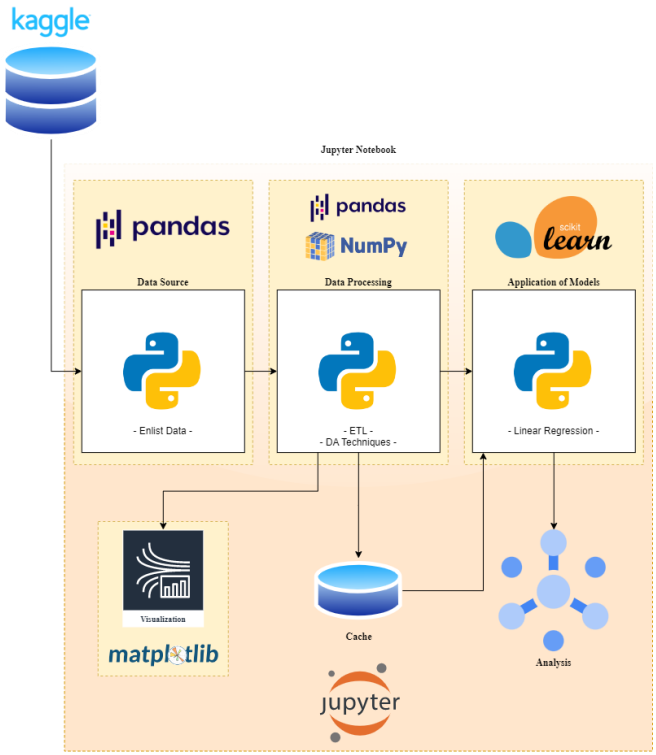


Figura 4. Arquitectura Física

Data Source: En esta capa se hace la captura de los datos del virus de la viruela símica en los repositorios de Kaggle, en donde se usaron los datasets llamados "Data Science with world countries" (TEK, 2022) y "Monkeypox Dataset" (CONTRACTOR, 2022). Se utilizó la librería Pandas de python para traer los datos de los datasets en forma de Dataframe.

Data Processing: En esta capa para el procesamiento de los datos se utilizaron las librerías Pandas y Numpy de Python. Además, se usaron técnicas de ETL y análisis de datos.

Visualization: En esta capa, para que se puedan visualizar las técnicas de análisis de datos, se utilizó la librería Matplotlib de Python.

Analysis of Results: En esta capa se utiliza el modelo de regresión lineal para predecir el comportamiento de los datos procesados.

Caché: En esta capa se utiliza la memoria caché para guardar las tablas temporales.

Analysis: En esta capa se visualizan por medio de la librería Matplotlib los resultados de la regresión lineal.

XIII. IMPLEMENTACIÓN DE LA SOLUCIÓN

La implementación del prototipo se llevó a cabo en Jupyter Notebook, utilizando Python. Para su desarrollo, se hizo uso de librerías como Pandas, Numpy, Matplotlib, entre otras.

```
import pandas as pd
import warnings
warnings.filterwarnings('ignore')
import numpy as np
import matplotlib.pyplot as plt
import plotly.express as px
import math
import seaborn as sns
from sklearn import preprocessing
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.metrics import precision_score
from sklearn.metrics import max_error
from sklearn.metrics import r2_score
from sklearn.metrics import mean_squared_error
from sklearn.model_selection import KFold
from sklearn.preprocessing import PolynomialFeatures
```

Figura 5. Librerías utilizadas

A. Preparación

Previamente a la creación del modelo de regresión lineal, se preparan los datos necesarios para este. Para ello, se toman los datos del primer conjunto de datos (casos por país) y se crea uno nuevo cuyos atributos son “day”, que hace referencia al conteo de fechas con casos registrados y el atributo “cases” que representa el número total de casos mundial (suma del total de casos de todos los países) registrados hasta esa fecha (acumulados); por ejemplo, la fecha “2022-01-31" es la fecha número 0 (primera fecha registrada) y “3” es el número total mundial de casos registrados acumulados hasta esa fecha.

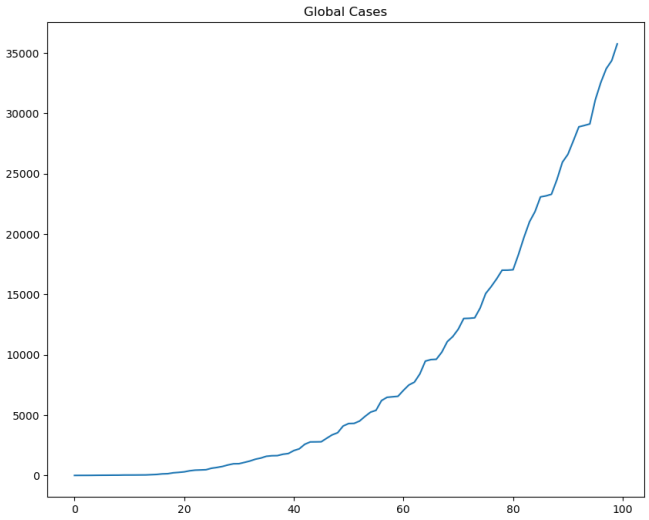


Figura 6. Gráfico de casos globales

B. Implementación del modelo

Posteriormente, se procede con la creación y entrenamiento del modelo de regresión lineal seleccionado (LinearRegression) de la biblioteca Scikit-Learn, para ello se dividieron los datos en datos de prueba y datos de entrenamiento, cuyo porcentaje fue de 30 y 70 por ciento respectivamente.

```
#creating model

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.3, random_state=4)
lr = LinearRegression(fit_intercept=False)

#We train with _train
lr_fit = lr.fit(x_train, y_train)
```

Figura 7. Creación del modelo de regresión lineal

La siguiente gráfica muestra la predicción realizada por el modelo entrenado y los datos originales simultáneamente.

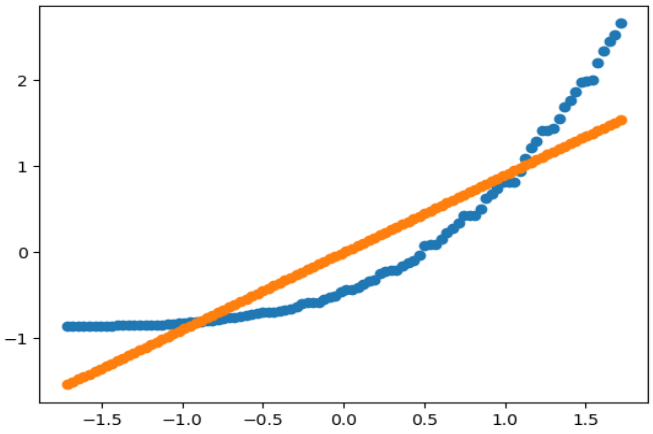


Figura 8. Gráfico de la predicción realizada

C. Evaluación del modelo

Inicialmente, se utilizaron 3 métricas diferentes para la evaluación de nuestro modelo de regresión lineal: error residual máximo, coeficiente de determinación y error cuadrático medio.

```
print('Maximum residual error of linear model for MonkeyPox Cases data set', max_error(y_test, y_pred))
```

```
Maximum residual error of linear model for MonkeyPox Cases data set 0.8759774
```

```
print('coefficient of determination of linear model for MonkeyPox Cases data set', r2_score(y_test, y_pred))
```

```
coefficient of determination of linear model for MonkeyPox Cases data set 0.8183754783059509
```

```
print('Mean squared error of linear model for MonkeyPox Cases data set', mean_squared_error(y_test, y_pred))
```

```
Mean squared error of linear model for MonkeyPox Cases data set 0.19327849
```

Figura 9. Métricas de evaluación del modelo de regresión lineal

Adicionalmente, para evaluar el desempeño del modelo seleccionado, utilizamos el método Kfold de la librería sklearn, el cual tiene como objetivo dividir los datos en segmentos, a los cuales se les aplicará el modelo. Esto nos ayudará a observar su desempeño para cada segmento de los datos.

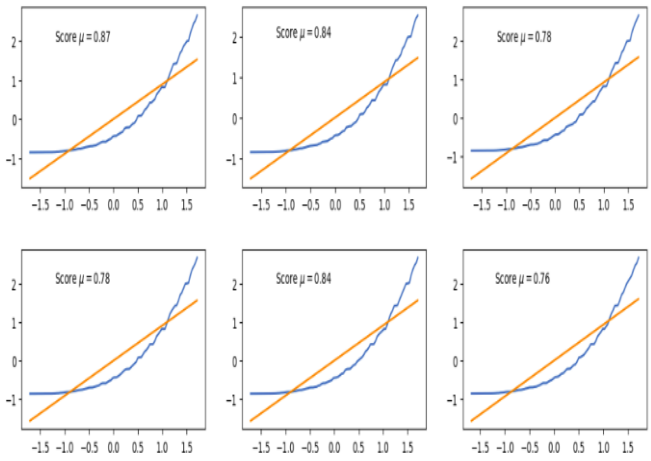
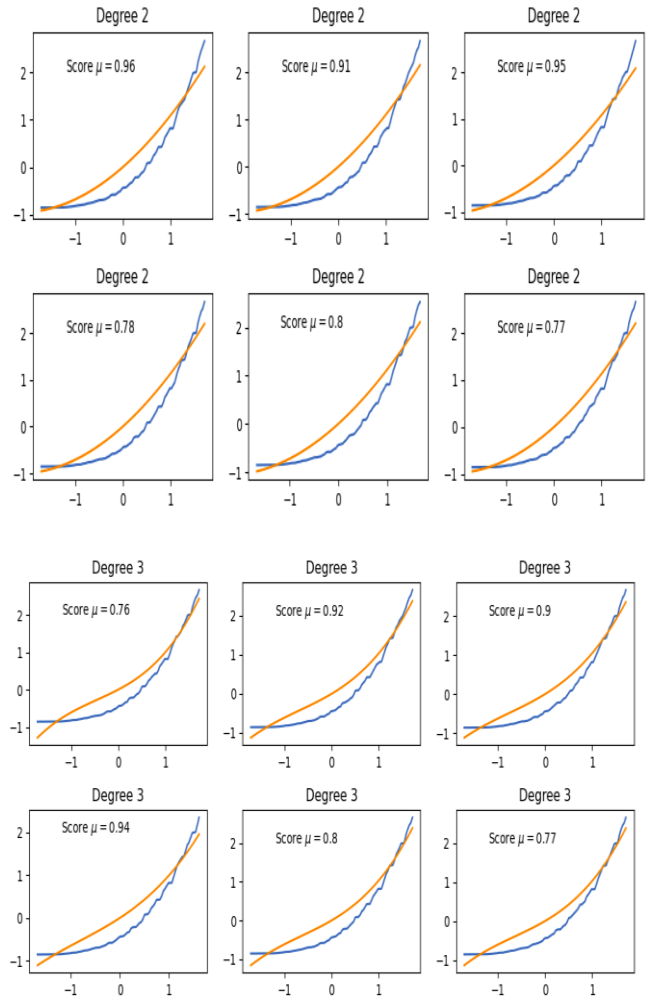


Figura 10. Método Kfold

Además, se utilizó también la función Polynomial Features de la librería sklearn, la cual genera características polinómicas del grado especificado a partir de nuestros datos, esto con el fin de que la predicción realizada por nuestro modelo se "ajuste" mejor a los datos originales.



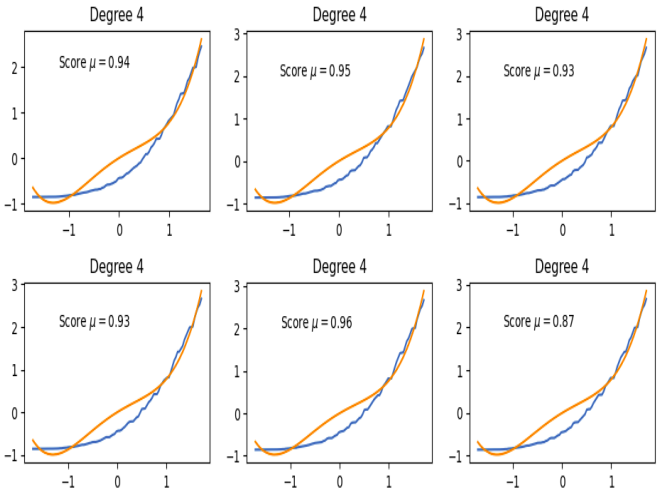


Figura 11. Función Polynomial Features

A pesar de que aumentar el grado de los polinomios generados puede mejorar el "score" del modelo, aumenta la probabilidad de que se presente Overffiting.

XIV. REVISIÓN SISTEMÁTICA DE LA LITERATURA

A. Criterios de búsqueda

- **Filtro: ninguno**

FUENTE DE DATOS	MONKEY-POX VIRUS MACHINE LEARNIG	PANDEMC MACHINE LEARNIG	VIRUS HUMAN MACHINE LEARNIG
Ieee-xplore	0	1,217	235
ScienceDi-rect	31	9,708	12,906
ACM DL	423,803	422,174	495,107

- **Filtro: Años (2021-2022)**

FUENTE DE DATOS	MONKEY-POX VIRUS MACHINE LEARNIG	PANDEMC MACHINE LEARNIG	VIRUS HUMAN MACHINE LEARNIG
Ieee-xplore	0	1,016	115
ScienceDi-rect	12	8,082	5,112

ACM DL	44,738	45,151	47,104

- **Filtro: Revistas / Artículos de investigación**

FUENTE DE DATOS	MONKEY-POX VIRUS MACHINE LEARNIG	PANDEMC MACHINE LEARNIG	VIRUS HUMAN MACHINE LEARNIG
Ieee-xplore	0	65	4
ScienceDi-rect	4	5,110	2,743
ACM DL	35,612	35,771	36,660

- **Filtro: PDF**

FUENTE DE DATOS	MONKEY-POX VIRUS MACHINE LEARNIG	PANDEMC MACHINE LEARNIG	VIRUS HUMAN MACHINE LEARNIG
Ieee-xplore	0	65	4
ScienceDi-rect	4	5,110	2,743
ACM DL	35,444	35,615	36,479

B. Tabla resumen RSL

La siguiente tabla muestra los trabajos de otros autores que más aportaron para la realización de este proyecto:

Título	Autor/es	Palabras clave	Fuente y año
The application of industry 4.0 technologies in pandemic management: Literature review and case study	Javid Moosavi, Javad Bakhshi, Igor Martek	Industry 4.0; Internet of Things (IoT); Blockchain; Artificial Intelligence (AI); Covid-19; Pandemic	APA/ 2021
Towards applying internet of things and machine learning for the risk prediction of COVID-19 in pandemic situation using Naive Bayes classifier for improving accuracy	N. Deepa, J. Sathya Priya, T. Devi	Accuracy; COVID-19; Naive Bayes classifier; Prediction; Random forest	APA / 2021
Machine Learning on the COVID-19 Pandemic, Human Mobility and Air Quality: A Review	M. Rahman, C. Paul, A. Hossain, N. Ali, S. Rahman & C. Thill	COVID-19; air quality; coronavirus; human mobility; machine learning; pandemic; public health; review; spatio-temporal analysis.	APA/ 2021
Big Data in the Era of Pandemic COVID-19 : Application of IoT based data analytics, Machine Learning and Artificial Intelligence	N. Jalil & M. Leen	Big Data, Machine Learning, Artificial Intelligence, IoT based data analytics, COVID-19, Pandemic	APA/ 2022
The recent technologies to curb the second-wave of COVID-19 pandemic	M. Poongodi, M. Malviya, M. Hamdi, H. Rauf, S. Kadry, O. Thinnukool	5G; CT-scan; Epidemic; X-Ray; artificial intelligence; cloud; coronavirus; drone; telemedicine	APA/ 2021

XV. TABLA DE VALORACIÓN DEL PROTOTIPO

Uno de los requisitos para la entrega de este trabajo, fue la valoración de este por parte de otro grupo de estudiantes. La siguiente tabla muestra el resultado de dicha valoración:

Caracteristica	Definición o descripción	1	2	3	4	5
Understandability	¿Fácil de comprender?			3		
Documentation	¿Documentación de usuario completa, apropiada y bien estructurada?			3		
Buildability	¿Fácil de construir en un sistema compatible? (Close-Open)					5
Installability	¿Fácil de instalar en un sistema compatible?					5
Learnability	¿Fácil de aprender a usar sus funciones?					5
Identity	¿La identidad del proyecto / software es clara y única?				4	
Copyright	¿Es fácil ver quién posee el proyecto / software?					5
Licencing	¿Adopción de la licencia apropiada?					5
Governance	¿Fácil de entender cómo se ejecuta el proyecto y cómo se gestiona el desarrollo del software?					5
Community	¿Evidencia de				4	

	comunidad actual / futura?					
Accessibility	¿Evidencia de capacidad de descarga actual / futura?					5
Testability	¿Fácil de probar la corrección de las funciones caja negra?				4	
Portability	¿Utilizable en múltiples plataformas?				4	
Supportability	¿Evidencia de soporte para desarrolladores actuales / futuros?				4	
Analysability	¿Fácil de entender a nivel fuente?				4	
Changeability	¿Fácil de modificar y aportar cambios a los desarrolladores?			3		
Evolvability	¿Evidencia de desarrollo actual / futuro?					5
Interoperability	¿Interoperable con otro software requerido / relaciona					5
	modelo de evaluación basado en el estandar ISO 9126 ISO 15504 + ESCALA DE LIKERT					

Totalmente en desacuerdo	1	Muy bajo
En Desacuerdo	2	Bajo
Ni Acuerdo NI Desacuerdo	3	Medio
De acuerdo	4	Alto
Totalmente de acuerdo	5	Muy alto
Escala Likert / ISO 15504		

XVI. CONCLUSIÓN

Después de varios meses de investigación, con base a la revisión sistemática de la literatura, y por medio de criterios claves, se obtuvieron varios documentos, útiles para nuestro proyecto. Gracias a la metodología de desarrollo seleccionada (CRISP-DM), diseñamos una arquitectura lógica y física para explicar la estructura de nuestro prototipo desde distintos puntos de vista. Además, investigamos las herramientas que nos ayudarían a tener la mejor fase de desarrollo, con base a lo anterior optamos por usar el lenguaje Python y sus librerías, también el uso de fuentes de datos confiables, luego aplicamos técnicas de análisis de datos y predicción de datos (regresión lineal).

Podemos decir que los resultados presentados pueden dar un indicio del crecimiento que ha tenido este virus de la viruela símica en el mundo entero, lo cual refleja lo importante que es para la sociedad un análisis como este, no solo para prevenir futuras pandemias, sino para tomar acciones de la propagación de este. Con base a los resultados, podemos concluir que la realización de este proyecto cumplió los objetivos propuestos inicialmente en este informe, debido a que se logró crear una aplicación que puede analizar el comportamiento general del virus, logramos tener una revisión sistemática de literatura exitosa, logramos el desarrollo del modelo y se validó la precisión del prototipo.

Durante el desarrollo, nos topamos con varios problemas, en cuestión a los datasets y el análisis de la literatura, debido a que la viruela símica empezó a propagarse significativamente a nivel mundial, y como hace unos meses no se encontraba mucha información pertinente en la que basarnos, decidimos ampliar las palabras claves de nuestras búsquedas hasta que pudimos obtener suficientes documentos que apoyaran nuestro proyecto.

Luego de tener nuestro prototipo evaluado por otras personas fuimos capaces de ver que nuestro proyecto, aunque tiene cosas por mejorar, también tiene gran potencial de desarrollo, con implementación de mejores modelos y mayor cantidad de datos podríamos realizar predicciones con un margen de error mucho menor, he incluso podemos llegar a pensar en predecir en que países ocurrirá un aumento de casos y las variables que impulsan esto. Esperamos que nuestro proyecto pueda servir como una base para los proyectos de otras personas.

XVII. REFERENCIAS

- Moosavi, J., Bakhshi, J. y Martek, I. (2021). La aplicación de las tecnologías de la industria 4.0 en la gestión de pandemias: revisión de literatura y estudio de caso. *Análisis de atención médica*, *1*, 100008. https://www.sciencedirect.com/science/article/pii/S2772442521000071
- World Health Organization. (2022, August 4). *Viruela Símica*. World Health Organization. Retrieved August 24, 2022, from https://www.who.int/es/news-room/questions-and-answers/item/monkeypox
- Pinheiro, C. A. R., Galati, M., Summerville, N., & Lambrecht, M. (2021). Using network analysis and machine learning to identify virus spread trends in COVID-19. Big Data Research, 25, 100242. https://www.sciencedirect.com/science/article/pii/S2214579621000599?via%3Dihub
- Lmater, M. A., Eddabbah, M., Elmoussaoui, T., & Boussaa, S. (2021). Modelization of Covid-19 pandemic spreading: A machine learning forecasting with relaxation scenarios of countermeasures. Journal of Infection and Public Health, 14(4), 468-473. https://www.sciencedirect.com/science/article/pii/S1876034121000083?via%3Dihub
- Avuçlu, E. (2022). Un método novedoso que utiliza conjuntos de datos Covid-19 y algoritmos de aprendizaje automático PARA EL DIAGNÓSTICO MÁS PRECISO que se puede obtener en el diagnóstico médico. *Procesamiento y control de señales biomédicas*, 103836. https://www.sciencedirect.com/science/article/pii/S1746809422003585
- Poongodi, M., Malviya, M., Hamdi, M., Rauf, H. T., Kadry, S., & Thinnukool, O. (2021). The recent technologies to curb the second-wave of COVID-19 pandemic. *Ieee Access*, *9*, 97906-97928. https://ieeexplore.ieee.org/document/9471880
- Chrin, R. y Wang, S. (2021, octubre). Análisis y predicción de datos de COVID-19 utilizando modelos de aprendizaje automático. En *2021 10th International Conference on Computing and Pattern Recognition* (págs. 296-301). https://dl.acm.org/doi/10.1145/3497623.3497671
- Ng, P. C., Spachos, P., Gregori, S., & Plataniotis, K. N. (2022). Epidemic Exposure Tracking With Wearables: A Machine Learning Approach to

Contact Tracing. *IEEE Access*, *10*, 14134-14148. https://ieeexplore.ieee.org/document/9698092

- Jalil, N. A., & Ei Leen, M. W. (2022). Big Data in the Era of Pandemic COVID-19 : Application of IoT based data analytics, Machine Learning and Artificial Intelligence. https://dl.acm.org/doi/10.1145/3524383.3524433
- What is predictive analytics? |. (s. f.) IBM. https://www.ibm.com/analytics/predictive-analytics
- Outbreaks, epidemics and pandemics—what you need to know |. (s. f.). Association for Professionals in Infection Control and Epidemiology (A.P.I.C.) https://apic.org/monthly_alerts/outbreaks-epidemics-and-pandemics-what-you-need-to-know/
- van Lissa, C. J. (2022). Van Lissa, C. J., Stroebe, W., Leander, N. P., Agostini, M., Draws, T., Grygoryshyn, A., ... & Bélanger, J. J. (2022). Using machine learning to identify important predictors of COVID-19 infection prevention behaviors during the early phase of the pandemic. *Patterns*, *3*(4), 100482. S. https://www.cell.com/patterns/fulltext/S2666-3899(22)00067-8?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS2666389922000678%3Fshowall%3Dtrue
- Deepa, N., Priya, J. S., & Devi, T. (2022). Towards applying internet of things and machine learning for the risk prediction of COVID-19 in pandemic situation using Naive Bayes classifier for improving accuracy. Materials Today: Proceedings.
- Rahman, M. M., Paul, K. C., Hossain, M. A., Ali, G. M. N., Rahman, M. S., & Thill, J. C. (2021). Machine learning on the COVID-19 pandemic, human mobility and air quality: A review. Ieee Access, 9, 72420-72450.
- Malik, H., Fatema, N., & Iqbal, A. (2021). *Intelligent data-analytics for condition monitoring: smart grid applications*. Academic Press. https://www.sciencedirect.com/book/9780323855105/intelligent-data-analytics-for-condition-monitoring
- CRISP-DM: La Metodología para Poner Orden en los Proyectos. Sngular. (2019, August 28). https://www.sngular.com/es/data-science-crisp-dm-metodologia/
- Academy, E. (2021, October 29). ¿Cómo elegir la mejor metodología de Investigación para su Estudio? Enago Academy Spanish. https://www.enago.com/es/academy/choose-best-research-methodology/#:~:text=La%20metodología%20de%20investigación%20es,el%20rumbo%20de%20la%20investigación.
- Amazon Machine Learning: Guía para desarrolladores - https://docs.aws.amazon.com/es_es/machine-learning/latest/dg/machinelearning-dg.pdf#cross-validation
- Ying, X. (2019, February). An overview of overfitting and its solutions. In *Journal of physics: Conference series* (Vol. 1168, No. 2, p. 022022). IOP Publishing.
- Brownlee, J. (2020) *How to Use Polynomial Feature Transforms for Machine Learning*. Available at: https://machinelearningmastery.com/polynomial-features-transforms-for-machine-learning/ (Accessed: November 19, 2022).
- Iberdrola. (n.d.). Descubre Cómo el 'Data Mining' Predecirá Nuestro comportamiento. Iberdrola. Retrieved September 30, 2022. https://www.iberdrola.com/innovacion/data-mining-definicion-ejemplos-y-aplicaciones
- TEK, M., 2022. Data Science with world countries. [online] Kaggle.com. Available at: <https://www.kaggle.com/code/mehmettek/data-science-with-world-countries> [Accessed 30 September 2022].
- CONTRACTOR, D., 2022. Monkeypox Dataset (Daily Updated). [online] Kaggle.com. Available at: <https://www.kaggle.com/datasets/deepcontractor/monkeypox-dataset-daily-updated> [Accessed 30 September 2022].
- Castañeda Gullot, Carlos, & Ramos Serpa, Gerardo. (2020). Principales pandemias en la historia de la humanidad. *Revista Cubana de Pediatría*, *92*(Supl. 1), e1183. Epub 20 de julio de 2020. Recuperado en 27 de noviembre de 2022, de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0034-75312020000500008&lng=es&tlng=es.