Un Proceso de Big Data aplicado a datos del COVID-19*

Guido Canevello, Nahir Saddi, Gastón Vidart, Carolina Villegas, Sebastián Yaupe, Agustina Buccella
Facultad de Informática
Universidad Nacional del Comahue
Neuquén, Argentina

***Abstract*—Dentro del contexto actual de pandemia declarada por la OMS debido al virus del COVID-19, han surgido muchos repositorios con grandes volúmenes de datos acerca de diferentes aspectos de la enfermedad en cada uno de los países, como cantidad de infectados, síntomas, muertes, etc. Estos datos, en general se han liberado a toda la comunidad de forma tal de promover esfuerzos a nivel mundial en cuanto acciones a tener en cuenta para lidiar con dicha enfermedad. En el campo de la informática, los datos son un valioso recurso que permite ser analizado en busca de información que sea útil y de difícil extracción para ser reconocida manualmente. Es por eso, que en este trabajo describimos la aplicación de un proceso de Big Data que toma dos grandes conjuntos de datos y realiza los pasos necesarios para obtener información útil que permita (1) conocer las medidas tomadas por los gobiernos de 3 países (Argentina, Corea del Sur y España) y analizarlas con respecto a la cantidad de infectados, y (2) predecir la progresión de cantidad de infectados en un rango de 30 días futuros. Siguiendo estos dos objetivos, describimos el proceso aplicado junto con los resultados obtenidos.**

I. Introducción

Big Data es un campo dedicado a tareas de análisis, procesamiento y almacenamiento de grandes conjuntos de datos que se originan desde diferentes fuentes. Generalmente se utilizan las soluciones de Big Data cuando las técnicas tradicionales son insuficientes para estas tareas [1], [2].

A pesar de que el concepto de Big Data parezca relativamente nuevo, la realidad es que muchas de sus bases se han venido aplicando hace muchos años. Existen diferentes técnicas que se han centrando tanto en el análisis de los datos, como el almacenamiento de los mismos y encuentran en Big Data el lugar para relacionarse y combinarse. Un ejemplo de esto son los sistemas federados, los cuales están dirigidos a la integración de diferentes fuentes de información, buscando principalmente la interoperabilidad. Lo mismo sucede con las técnicas de minería de datos las cuales se han centrado en las forma de analizar y extraer información útil. Todos estos avances previos, obviamente no han dejado de existir sino que convergen en el concepto de Big Data y a su vez son extendidos para funcionar en entornos de grandes volúmenes de datos y grandes exigencias de procesamiento. Incluso en los últimos años han surgido una serie de estándares respecto a su arquitectura de referencia, interoperabilidad, terminología, etc. Estos esfuerzos de estandarización han sido llevados a cabo por el comité ISO/IEC JTC 1/SC 42 (Artificial intelligence)[1] y por el grupo de trabajo del NIST (Big Data Public Working Group - NBD-PWG)[2]

Existen en la actualidad muchas metodologías propuestas para llevar a cabo un proceso de desarrollo para Big Data [1], [2]. En general, se converge en tres grandes campos de trabajo involucrando la limpieza y/o preprocesamiento, almacenamiento y la analítica de datos. Los dos primeros pasos pueden ser intercambiados dependiendo del enfoque utilizado. Es decir, para realizar un desarrollo orientado al análisis de los datos, se debe primero seleccionar la/s fuentes de información que puedan resultar útiles, procesarlas para que tengan sentido, y luego almacenarlas en algún tipo de repositorio, en la forma de depósitos de datos [3] o lagos de datos [4]. La característica principal de estos repositorios es que actúan como centros de información en donde se vuelcan, en algún formato específico, los datos de todas las fuentes que se deseen explotar en un proceso de extracción, transformación y carga (ETL)[5], [6], [7]. Luego, la analítica de datos (data analytics) se dedica al proceso de crear información desde los datos fuente por medio de la contextualización, análisis y gobernanza de datos. Las características principales de Big Data son [1]: el *volumen*, el cual es inmenso ya que puede generarse a partir de redes sociales, sistemas bancarios, sensores, internet de las cosas (IoT), etc.; la *velocidad*, en cuanto a la rapidez en que los datos se acumulan o generan; la *variedad*, ya que los datos pueden ser estructurados, semi-estructurados, y sin estructura; en formato de texto, imagen, video, audio, etc.; la *veracidad*, en cuanto a la calidad o fiabilidad de los datos; y por último el *valor*, el cual se enfoca en la utilidad del dato para el propósito para el que va a ser utilizado.

Se han presentado muchas aplicaciones de Big Data en la actualidad en dominios específicos, todas ellas realizando análisis interesantes que permitan explotar la información de manera de clasificarla, relacionarla y/o predecir nuevos comportamientos o sucesos. En particular en el contexto actual en el que nos encontramos a nivel mundial, en el marco de la pandemia declarada por la OMS debido al COVID-19, muchas investigaciones y aplicaciones se han abocado a la tarea de analizar la información que se va recabando en las diferentes regiones del mundo de forma tal de poder predecir contagios, determinar grupos de riesgo, propagaciones futuras, etc.

Considerando este contexto, en el presente trabajo hemos aplicado un proceso de Big Data particular aplicado a dos

(*) Este trabajo es presentado en el marco de la Materia Electiva *Almacenamiento y Análisis para Big Data* perteneciente al 5to año de la Carrera Licenciatura en Sistemas de Información.

[1] https://www.iso.org/committee/6794475.html
[2] https://bigdatawg.nist.gov/

fuentes de datos que contienen información actualizada diaria o semanalmente de los casos de COVID-19 en varios paises y las medidas tomadas por cada uno de esos gobiernos. A partir de estas fuentes, el proceso aplicado predice contagios futuros considerando cantidad de habitantes, medidas tomadas y curvas de infecciones actuales.

Este trabajo se organiza de la siguiente manera. En la sección siguiente describimos los trabajos relacionados en donde se aplican análisis sobre datos de la pandemia y resaltamos a su vez, las diferencias con este trabajo. En la Sección III describimos el proceso de Big Data que utilizamos a nivel general, para luego en la Sección IV aplicarlo al contexto de las fuentes de datos seleccionadas y así lograr los objetivos propuestos. Finalmente se describen las conclusiones y trabajos futuros.

II. Trabajos Relacionados

En este último tiempo, desde el inicio de la pandemia, se han recolectado una gran cantidad de datos acerca de varios aspectos de la enfermedad COVID-19. Así han habido esfuerzos nacionales y también mundiales para que la información recopilada sea completa, coherente y sirva para entender la forma en que este virus actúa en la sociedad. Muchos de estos datos se han dejado disponibles a toda la comunidad incentivando el surgimiento de trabajos de investigación que apliquen técnicas de análisis de datos para clasificar, categorizar, predecir, recomendar (entre otras) comportamientos y acciones acerca de la enfermedad. Todos estos esfuerzos de la ciencia en analizar los datos son de gran utilidad ya que con ellos se puede comprender mas sobre la forma en que la enfermedad se expande, los síntomas más comunes, rango de edades mas afectadas, etc., facilitando a los gobiernos tomar acciones para minimizar sus efectos.

Un trabajo interesante se presenta en [8] donde los autores realizaron un estado del arte sobre la aplicación de técnicas de predicción en datos del COVID-19. Se han analizado mas de 14 trabajos enfocados en Big Data y otros 7 que aplican análisis predictivos basados en aprendizaje automatizado (machine learning) junto con los resultados obtenidos en cada caso. A su vez se definen desafíos para este tipo de análisis como la realización de modelos para la movilidad de las personas, la falta de datos apropiados para realizar los estudios, abundancia de datos que violan la seguridad, complejidad de los modelos, etc. Entre los trabajos similares al nuestro sobre predicción de datos, en [8] los autores han aplicado siete técnicas de aprendizaje automatizado y análisis estadístico sobre 1182 pacientes hospitalizados. El objetivo fue realizar una predicción de supervivencia de los pacientes y evaluar el impacto de la edad y el género como dos factores de riesgo principales. De este estudio se obtuvieron conclusiones como *la probabilidad de recuperación y alta para los pacientes varones hospitalizados en los primeros 15 días a partir de mostrar los síntomas es mayor que en las mujeres*, *después del día 15 hasta casi los 40 días de mostrar los síntomas, la probabilidad de recuperación en las mujeres es ligeramente mayor*, o *las mujeres tienen aproximadamente un 5% más de probabilidades que los hombres de ser dadas de alta del hospital* entre otras.

Un trabajo similar fue presentado en [9] en el que se aplicó la técnica random forest [10] sobre un conjunto de datos disponible en la plataforma Kaggle[3] sobre síntomas, muertes, recuperaciones y si las personas han visitado o viven en Wuhan, China. Los resultados de este estudio mostraron que las tasas de mortalidad eran más altas entre los nativos de Wuhan en comparación con los no nativos, y que los pacientes masculinos tuvieron una mayor tasa de mortalidad. En [11] los autores aplicaron dos técnicas de aprendizaje automatizado (SEIR [12] y modelo de regresión) que ya han sido utilizados para la predicción de otras enfermedades infecciosas. En este caso, estas técnicas pudieron predecir, en forma bastante precisa, en número de casos en un futuro cercano. También, en el trabajo presentado en [13] se utilizaron las técnicas de regresión lineal y de memoria a largo-corto plazo (LSTM) para estimar el número de casos positivos. Estos modelos de predicción fueron aplicados a datos de Irán usando datos de Google Trends[4].

Por último podemos citar trabajos, como [14], [15], que aplicaron, al igual que nosotros, métodos de predicción sobre series de tiempo utilizando el modelo Prophet[5]. En [14] han utilizado Prophet para analizar los países con más contagios desde enero al 15 de junio del corriente año. Se analizaron a nivel mundial y luego para 5 países en particular. Las variables analizadas fueron la fecha y el número de infecciones por país, y en la mayoría de los casos se intentó predecir las fechas en que ocurrirán los picos de contagio. En [15] se presentaron 6 diferentes métodos de predicción, el cual uno de ellos fue Prophet. Estos métodos fueron comparados según rendimiento y predicciones realizadas en varias países según cantidad de infectados.

En este trabajo, a diferencia de los anteriores, nos enfocamos en la aplicación de un proceso completo de Big Data aplicado a datos del COVID-19 aportando las decisiones tomadas dentro de cada actividad del mismo que pueden ayudar a extender el mismo a mas fuentes de información e incluso a ampliar los requerimientos sobre los datos que ayuden aún mas en el análisis de esta pandemia.

III. Proceso de Big Data

En este trabajo usamos el proceso de big Data presentado en [1] el cual hemos adaptado para utilizar un enfoque basado en el uso de Lago de Datos. Es decir, las fuentes de información se almacenan en un repositorio y luego dependiendo del análisis a realizar se procesan, siguiendo así el enfoque ELT (extracción, carga y transformación) [4]. En la Figura 1 podemos ver dicho proceso el cual se divide en 6 actividades:

1) *Evaluación del Caso del Negocio*: aquí se definen los objetivos del trabajo a realizar con Big Data, es decir, cuáles son los resultados analíticos que se quieren obtener y que van a tener a su vez incidencia sobre el negocio o área de estudio.
2) *Identificación y Recolección de Datos*: se deben identificar los datos requeridos para analizar y las fuentes confiables de donde obtenerlos. A su vez, se debe definir la forma en la que se realizará la recolección de los mismos.

[3]https://www.kaggle.com/

[4]https://trends.google.com.ar/trends/?geo=AR

[5]https://facebook.github.io/prophet/

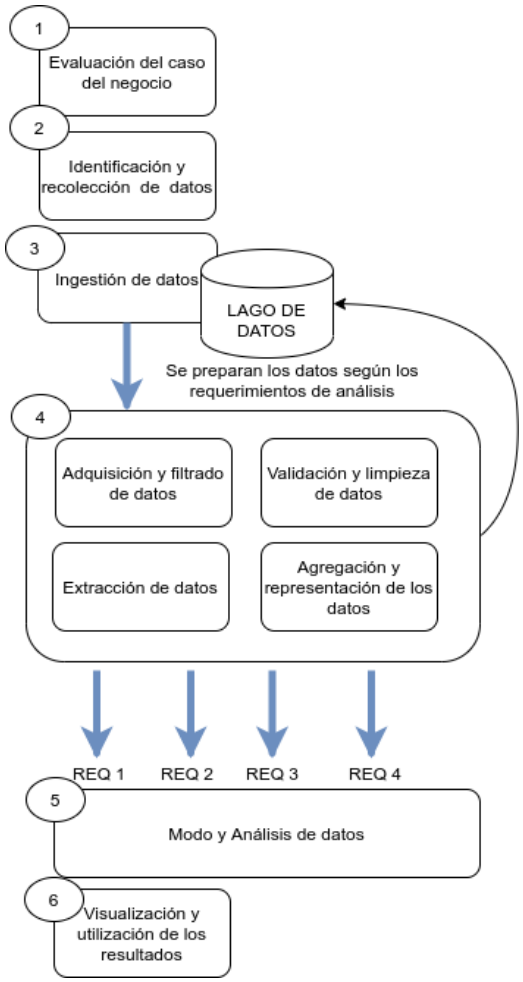

Fig. 1: Proceso de Big Data según un enfoque ELT

3) *Ingestión de Datos*: se refiere a la conexión de varias fuentes de datos, la extracción de los datos de esas fuentes y la detección del momento en que esos datos son modificados. Se trata de mover datos, especialmente los datos no estructurados, desde donde se originan, a un sistema donde se puedan almacenar y analizar (un lago de datos).
4) *Preparación de los datos*: consiste en la manipulación de los datos recibidos para que estos sean útiles a los sistemas que deben usarlos para realizar análisis sobre ellos. Existen múltiples técnicas que se pueden utilizar como normalización, de-duplicación, limpieza, muestreo, etc. Esta tarea puede involucrar volver a almacenar los datos ya procesados nuevamente en el lago.
5) *Modo y Análisis de Datos*: aquí se determina de qué modo se procesarán los datos para realizar el análisis incluyendo por ejemplo si se realizará por lotes, tiempo real o interactivo. Esto dependerá de la naturaleza de las fuentes de información. A su vez, basado en el modo y de acuerdo a los objetivos planteados, se deben realizar tareas de análisis de los datos que incluyen técnicas como regresión, clasificación, clustering, estadísticas básicas y análisis de series de tiempo, etc. [16], [17], [18].
6) *Visualización y Utilización de los Datos*: se deben mostrar los resultados de los análisis anteriores en forma sencilla y fácil de comprender; involucrando la aplicación de técnicas de visualización.

IV. Caso de Estudio

A continuación describimos la aplicación del proceso presentado en la Figura 1 aplicado a un caso de estudio de fuentes de información que poseen datos relacionados con el COVID-19.

1) Evaluación del Caso del Negocio. Los objetivos definidos para la aplicación del proceso de Big Data para este dominio son (1) *analizar las medidas tomadas por los gobiernos para frenar el avance de la pandemia* y (2) *realizar predicciones de curvas de infectados en los próximos 30 días a partir del comportamiento de la información con respecto a cantidad de habitantes, casos y medidas tomadas por países con características similares.*

Al mismo tiempo, en este paso, hemos seleccionado las herramientas a utilizar en cada una de las actividades del proceso. En la Figura 2 podemos ver dichas herramientas en las que se detalla **Apache Hadoop**, específicamente el componente de **HDFS (Hadoop Distributed File System)**[6] como lago de datos. **Optimus**[7] como ayuda a las librerías que ya provee **Apache Spark**[8] para la preparación de los datos. Luego, para el procesamiento por lotes y el análisis de los datos, además de utilizar Apache Spark con sus librerías de machine learning (MLib)[9], utilizamos la API **Prophet**[10]. Prophet es un framework de código abierto de Facebook para el pronóstico de series de tiempo basado en un modelo aditivo. Por último, para la visualización, también utilizamos la librería **Matplotlib**[11]. Todas estas herramientas son de código abierto y pueden descargarse y usarse en forma gratuita.

2) Identificación y Recolección de Datos. Con los objetivos planteados, se recolectó información de dos fuentes. La primera proviene de un sitio que provee datos humanitarios[12] de uso general. Este conjunto de datos (al que llamamos **Medidas Tomadas**)[13] registra las medidas que tomaron los países para combatir al COVID-19[14]. De las medidas se registra principalmente cuándo se aplicaron las mismas y la categoría a la que corresponden. Las categorías pueden ser sociales, económicas, gubernamentales o sanitarias. La segunda fuente de datos corresponde al sitio de *Our World in Data*, del cual extrajimos un conjunto de datos (al que llamamos **Casos Diarios**)[15] que registra la cantidad de casos confirmados por día almacenando la fecha, origen, cantidad

[6]https://hadoop.apache.org/docs/r3.3.0/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html#Introduction
[7]https://hi-optimus.com/
[8]https://spark.apache.org/
[9]https://spark.apache.org/mllib/
[10]https://pypi.org/project/fbprophet/
[11]https://matplotlib.org/3.1.0/index.html
[12]https://centre.humdata.org/
[13]https://data.humdata.org/dataset/acaps-covid19-government-measures-dataset
[14]Cantidad de registros: 15665
[15]https://ourworldindata.org/coronavirus-source-data

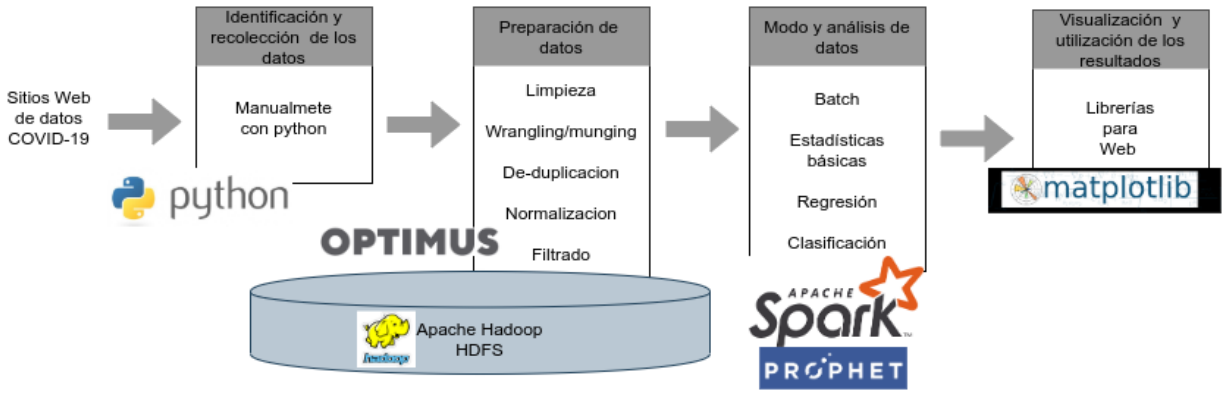

Fig. 2: Herramientas de Software seleccionadas para cada actividad del proceso de Big Data

de muertes o contagiados por millón, el código ISO del país, etc.[16].

3) Ingestión de Datos. Estos conjuntos de datos fueron almacenados manualmente en HDFS. Sin embargo se procedió a crear los scripts necesarios para poder automatizar este proceso en un futuro. Como estamos siguiendo un proceso basado en lago de datos, éstos se almacenan tal cual son obtenidos, dejando la preparación de los datos para las etapas siguientes. De esta forma los datos se pueden *preparar* de acuerdo a los requerimientos que surjan en diferentes momentos.

4) Preparación de los datos. Los conjuntos de datos seleccionados fueron procesados mediante técnicas de limpieza, filtrado, validación y completitud para contener de forma segura y coherente los datos necesarios para realizar los objetivos previstos.

Para el conjunto de datos de **Medidas Tomadas** la preparación incluyó eliminar y/o corregir tuplas con datos inconsistentes, atributos que no proveían beneficios para el análisis planteado, atributos con muchos valores nulos, eliminación de duplicados, etc. En la Tabla I podemos observar como quedaron los atributos junto con su descripción.

Para el conjunto de datos de **Casos diarios** se eliminaron también las columnas que no eran de interés para el trabajo a realizar y se re-formatearon todas las fechas para que queden consistentes con el formato *yyyy/mm/dd*. Se renombraron atributos, se filtraron las filas que se corresponden a los 3 países elegidos (Argentina, Corea del Sur y España) y se ordenaron las tuplas (en cada país) por sus fechas, para luego analizar cada país por separado. Es importante resaltar aquí que se eligieron, además de Argentina, esos otros dos países porque tienen tamaños de población similares, lo que nos permite realizar los análisis sobre datos comparables. A su vez, estos 3 países pasaron por situaciones diferentes; Corea tuvo pocos casos y España al contrario, tuvo muchos mas casos que Argentina al comienzo de la pandemia. En la Tabla II podemos observar los atributos finales junto con sus descripciones.

En cuanto a otros cambios de formato y nombres de las columnas en ambos conjuntos de datos, fueron realizados porque son un requerimiento de Prophet para realizar las predicciones.

[16]Cantidad de registros: 36900

Medidas Tomadas	
Atributo	**Descripción**
country	Nombre del País
iso	Código ISO del País
region	Nombre de la Región continental
category	Categoría de la medida tomada
measure	Descripción de la medida tomada
targeted_pop_group	Fue dirigido a un grupo especifico de la Población (SI/NO)
comments	Comentarios adicionales de la medida tomada
date_implemented	Fecha en la que se implemento la medida
source	Fuente de donde fue recuperada la medida
source_type	Tipo de Fuente
entry_date	Fecha en la que se ingreso la medida en el conjunto de datos
alternative_source	Fuente alternativa

TABLE I: Atributos finales de **Medidas Tomadas**

5) Modo y Análisis de Datos

A. Modo de Análisis. Para cumplir con los objetivos planteados realizamos un tipo de análisis exploratorio, que nos permitiera analizar los datos y poder descubrir patrones o anomalías. Por otra parte, utilizamos el modo de análisis de tipo *por lotes* (batch), el cual consiste en analizar los datos por bloques y diferido a su producción. Luego con la librería Prophet realizamos las predicciones.

B. Análisis de los Datos. El análisis de los datos se realizó considerando los dos objetivos planteados en el primer paso del proceso.

Obj 1) Analizar las medidas tomadas por los gobiernos para frenar el avance de la pandemia.

Casos Diarios	
Atributo	**Descripción**
iso_code	Identificador ISO del país
continent	Nombre del continente del país
location	Nombre del país
date	Fecha de registro
total_cases	Cantidad de confirmados de COVID-19
total_deaths	Cantidad de muertos por COVID-19
total_cases_per_million	Cantidad de confirmados de COVID-19 por millón de habitantes
total_deaths_per_million	Cantidad de muertos de COVID-19 por millón de habitantes
population	Cantidad de Población del País
population_density	Cantidad de personas por m^2
median_age	Edad mediana de la población del país
aged_65_older	Proporción de la población de 65 años o más
aged_70_older	Proporción de la población de 70 años o más
new_cases	Cantidad de nuevos casos de COVID-19
new_deaths	Cantidad de nuevas muertes de COVID-19

TABLE II: Atributos finales de **Casos Diarios**

Para este objetivo se realizó un análisis exploratorio considerando las medidas implementadas por cada país seleccionado. En la Figura 3 se puede observar el caso de España, en donde se consideraron las medidas de distanciamiento social (con color verde), de salud (con color negro) y restricción de movimiento (en color rojo). Como se puede observar, el aumento en la cantidad de casos desde fines de marzo y todo el mes de abril creció de manera considerable. Aunque se tomaron medidas en las etapas iniciales, se puede apreciar que las medidas de salud y restricción de movimiento se incrementaron durante el mes de abril. Estas podrían ser la causa de que se observa un descenso en la cantidad de nuevos casos a partir de mediados de marzo y hasta fines de junio. Luego, a pesar de la estabilidad lograda, se observa también un rebrote en la cantidad de casos que volvió a causar otro pico en el gráfico, período donde se tomaron menos medidas que en etapas previas[17].

Con respecto a Corea del Sur, hay que destacar que las únicas medidas presentes en el conjunto de datos utilizado en éste análisis eran de distanciamiento social (no contenían medidas de salud o restricción de movimiento). De esta forma, en la Figura 4 podemos observar que la cantidad de casos se mantuvo con cierta estabilidad en los primeros dos meses. Luego, al comienzo de marzo se advierte un pico pronunciado, ya que en solo un mes la cantidad de casos quintuplicaron su valor, de 200 a 1000 casos. Luego de este suceso, la cantidad de casos se estabilizó hasta el rebrote ocurrido en el mes de septiembre. Aún así podemos destacar que la cantidad de casos en este país no fue tan alta comparada al resto de los países, como en España.

Por último, en la Figura 5 vemos el caso de Argentina. Nuestro país es el que primero tomó una medida de salud, específicamente una campaña, el 14 de abril. También se observa que las medidas están distribuidas en el tiempo y que existe un crecimiento exponencial de nuevos casos con COVID-19 a partir de fines de Abril hasta hoy. El impacto de las medidas tomadas de distanciamiento social, salud y restricción de movimiento a partir de Marzo refleja una contención efectiva de los casos sólo hasta Abril. A partir de ese momento la cantidad de casos aumentó considerablemente pero no así la cantidad de medidas.

Si comparamos lo ocurrido en España y Argentina, se puede ver fácilmente que tomar las medidas ayudó a mantener reducidos los contagios solamente por un tiempo, ya que Argentina había comenzado a tomar medidas a fines de Marzo y su curva recién empieza a crecer a fines de Abril; y en España, luego de haberse reducido, volvió a crecer a partir de fines de Junio. Luego en Corea del Sur se puede ver que, a comparación de Argentina y España, solo se tomaron medidas de distanciamiento social y, si se observa el eje de nuevos casos en la Figura 4, nunca se superaron los 1000 casos diarios.

Por supuesto que es difícil con solo estos datos poder realizar un análisis completo para comprender las razones de esos comportamientos, ya que no poseemos información acerca de la efectividad de las medidas. Es decir, el hecho de que en Corea Del Sur la cantidad de casos diarios sea baja tomando solo la medida de distanciamiento social, se puede deber a que en ese país dicha medida ha sido cumplida en un muy alto porcentaje. Hecho que posiblemente en la Argentina y España no ha sucedido de ese modo y a pesar de poseer las medidas vigentes, la cantidad de casos siguió incrementándose.

Obj 2) Realizar predicciones de infectados en los próximos 30 días a partir del comportamiento de la información con respecto a cantidad casos y medidas tomadas por países con características similares.

A su vez, realizamos un análisis de serie de tiempo utilizando la librería Prophet para realizar pronósticos a futuro de la cantidad de infectados diarios dentro de un tiempo determinado. El objetivo esperado con este proceso fue poder generar una predicción que determine el nivel de incremento de casos, establecido según el ángulo creciente de las pendientes proyectadas. Para poder realizar este objetivo, la librería permite ingresar el período a futuro que se desea predecir, aunque también permite agregar una frecuencia que puede ser diaria, mensual o por hora, que afecta la forma en la que aprende. Este tiempo fue fijado en 30 días, a partir de la última fecha registrada en el conjunto de datos de **Casos Diarios**, ya que fue el tiempo en la que la predicción arrojó valores analizables.

En las figuras siguientes realizamos las predicciones para cada país. Para comprender las mismas veremos que todas ellas se componen por un conjunto de puntos negros superpuestos a una línea celeste. Estos puntos negros hacen referencia a los datos presentes en la fuente de datos, es decir, son las

[17]es posible que existan medidas tomadas en fechas posteriores al 23/07/2020, de las cuales no hay información

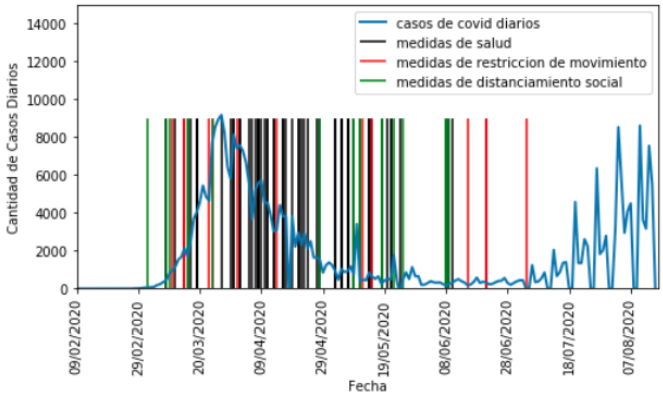

Fig. 3: Medidas en España

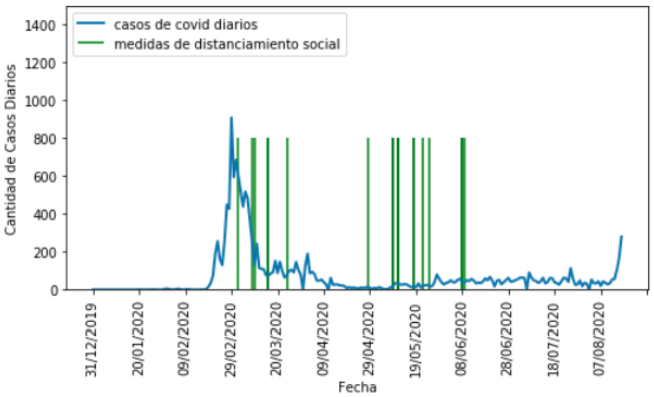

Fig. 4: Medidas en Corea

cantidades de casos diarios en cada fecha. Por otra parte, la linea celeste superpuesta a los puntos negros es el modelo que crea la herramienta al analizar los datos. Cuando la linea deja de tener puntos negros significa que la herramienta crea una predicción.

Para el caso de España, en la Figura 6 el pronóstico de los siguientes 30 días indica que las cifras diarias descenderán

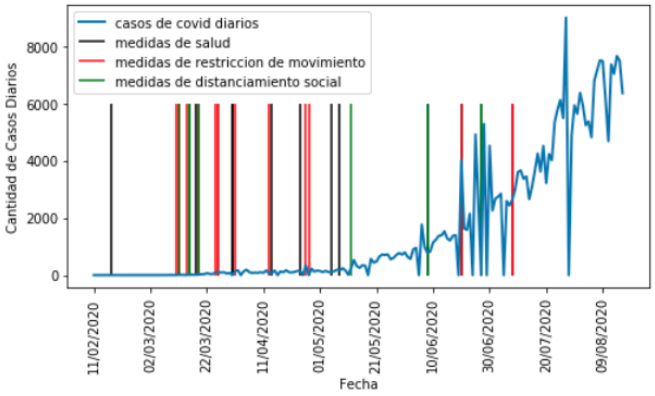

Fig. 5: Medidas en Argentina

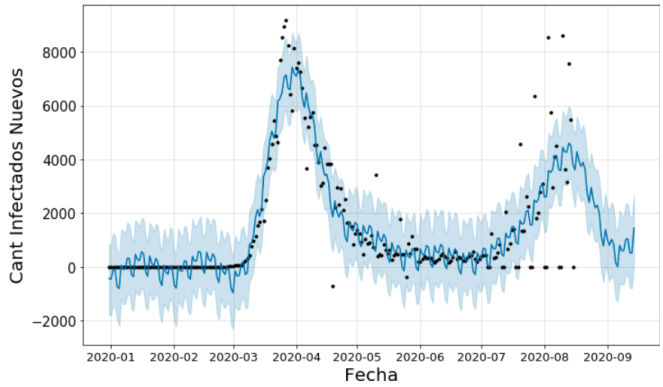

Fig. 6: Predicción de casos a 30 días de total de casos en España

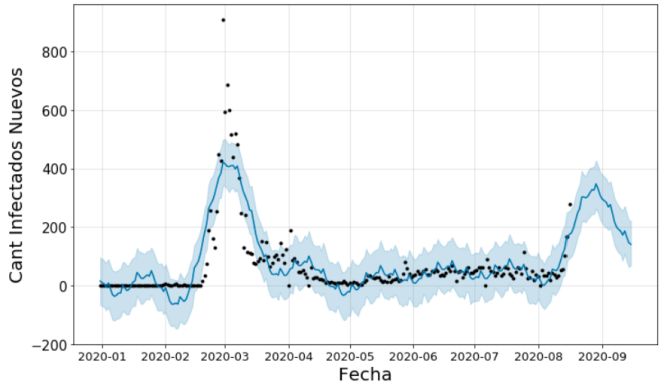

Fig. 7: Predicción de casos a 30 días de total de casos en Corea del Sur

a partir de mediados de agosto[18], pero habrá un rebrote en los últimos días del pronóstico, es decir para mediados de septiembre.

Con respecto a Corea del Sur, en la Figura 7, podemos observar que a partir de agosto hasta septiembre aumentan considerablemente la cantidad de casos diarios. Sin embargo, luego de ese suceso, la cantidad de casos van a ir disminuyendo de a poco.

Finalmente en la Figura 8 vemos el caso de Argentina donde la predicción determina que se va a estabilizar la curva tendiendo a un descenso en la cantidad de casos diarios a partir del mes de Agosto.

Por último podemos realizar un análisis, al menos informal, acerca de los resultados previos. De los datos obtenidos en la actualidad[19] en cuanto a la cantidad de infectados reales (no proyectados) durante agosto-septiembre, notamos que en el caso de Corea del Sur se pronosticaba un leve aumento en el mes de Septiembre de aproximadamente 380 casos diarios que, si lo comparamos con la realidad, fueron de 267. Luego de eso, se pronosticó una reducción a alrededor de 180 casos y, comparado con los datos reales a la fecha 10/09/2020, fueron 155. Luego, respecto a la cantidad de casos reales en

[18]última fecha de actualización del conjunto de datos
[19]https://bit.ly/2Ft1D80

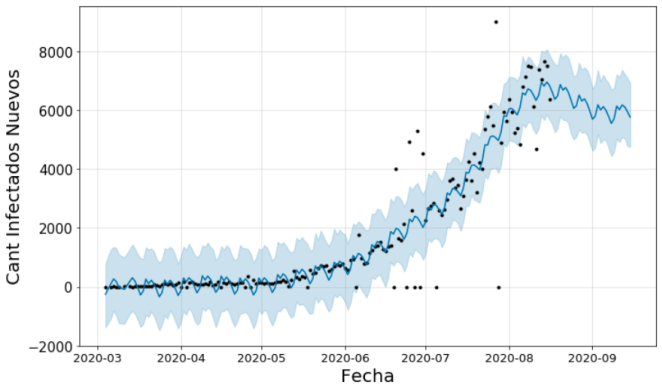

Fig. 8: Predicción de casos a 30 días de total de casos en Argentina

España para el mismo período (agosto-septiembre), se puede observar que luego de un rebrote de entre 4000 y 6000 casos, se pronosticaba una baja llegando a 0 casos. Sin embargo, en la actualidad para la fecha del 10/09/2020 hubieron 10764 casos nuevos. Finalmente, en el caso de Argentina, si bien se predecía que la cantidad de casos se iba a estabilizar tendiendo a la baja, esto ocurrió solo en cortos periodos de tiempo.

5) Visualización y Utilización de los Datos. La técnica de visualización que utilizamos fue estática, ya que nos basamos en valores obtenidos de nuestro lago de datos y observamos los resultados en modelos no dinámicos. Las librerías utilizadas fueron **numpy**[20], que agrega mayor soporte para vectores y matrices, y Matplotlib. A su vez, durante de la predicción se utilizaron las herramientas de visualización provistas por Prophet, para lo cual se tuvo que importar la funcionalidad de graficación *plot_plotly* que es proporcionada por dicha API.

V. Conclusiones y Trabajo Futuro

En este trabajo hemos descripto las actividades realizadas sobre un proceso completo de Big Data aplicado a fuentes de información sobre datos del COVID-19. En particular nos hemos centrado en dos objetivos principales y en base a ellos hemos descripto el proceso realizado destacando las actividades más importantes que nos permitieran obtener resultados interesantes en cuento a las medidas tomadas por tres países y la predicción de contagios a 30 días.

Como trabajo futuro, usando las fuentes aplicadas aquí y posiblemente extendiendo a otras, se plantea organizar y ampliar los objetivos/requerimientos en este trabajo para realizar nuevas predicciones basados en nuevas entradas de datos como síntomas, edad, condiciones de salud, factor de riesgo crítico, etc. Dichos requerimientos pueden surgir de las nuevas medidas adoptadas y de cómo se va comportando la población a medida que esta pandemia se mantiene.

Acknowledgment

References

[1] A. Bahga and V. Madisetti, *Big data science analytics : a hands-on approach*. Bahga, A. and Madisetti, V., 2016.

[2] T. Erl, W. Khattak, and P. Buhler, *Big Data Fundamentals: Concepts, Drivers Techniques*. Prentice Hall Press, 2016.

[3] A. Vaisman and E. Zimnyi, *Data Warehouse Systems: Design and Implementation*, 1st ed. Springer Publishing Company, Incorporated, 2016.

[4] C. Quix and R. Hai, *Data Lake*. Cham: Springer International Publishing, 2018, pp. 1–8. [Online]. Available: https://doi.org/10.1007/978-3-319-63962-8_7-1

[5] S. Luján-Mora, P. Vassiliadis, and J. Trujillo, "Data mapping diagrams for data warehouse design with uml," in *Conceptual Modeling – ER 2004*, P. Atzeni, W. Chu, H. Lu, S. Zhou, and T.-W. Ling, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 191–204.

[6] A. Simitsis, D. Skoutas, and M. Castellanos, "Natural language reporting for etl processes," in *Proceedings of the ACM 11th International Workshop on Data Warehousing and OLAP*, ser. DOLAP '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 65–72. [Online]. Available: https://doi.org/10.1145/1458432.1458444

[7] J. Trujillo and S. Luján-Mora, "A uml based approach for modeling etl processes in data warehouses," in *Conceptual Modeling - ER 2003*, I.-Y. Song, S. W. Liddle, T.-W. Ling, and P. Scheuermann, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 307–320.

[8] G. Shinde, A. Kalamkar, P. Mahalle, N. Dey, J. Chaki, and A. Hassanien, "Forecasting models for coronavirus disease (covid-19): A survey of the state-of-the-art," *SN COMPUT. SCI.*, vol. 1, no. 197, 2020.

[9] C. Iwendi, A. K. Bashir, A. Peshkar, R. Sujatha, J. M. Chatterjee, S. Pasupuleti, R. Mishra, S. Pillai, and O. Jo, "Covid-19 patient health prediction using boosted random forest algorithm," *Frontiers in Public Health*, vol. 8, p. 357, 2020. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpubh.2020.00357

[10] J. Han, M. Kamber, and J. Pei, "8 - classification: Basic concepts," in *Data Mining (Third Edition)*, third edition ed., ser. The Morgan Kaufmann Series in Data Management Systems, J. Han, M. Kamber, and J. Pei, Eds. Boston: Morgan Kaufmann, 2012, pp. 327 – 391.

[11] R. Gupta, G. Pandey, P. Chaudhary, and S. K. Pal, "Machine learning models for government to predict covid-19 outbreak," *Digit. Gov.: Res. Pract.*, vol. 1, no. 4, Aug. 2020. [Online]. Available: https://doi.org/10.1145/3411761

[12] M. Y. Li and J. S. Muldowney, "Global stability for the seir model in epidemiology," *Mathematical Biosciences*, vol. 125, no. 2, pp. 155 – 164, 1995. [Online]. Available: http://www.sciencedirect.com/science/article/pii/0025556495927565

[13] S. M. Ayyoubzadeh, S. M. Ayyoubzadeh, H. Zahedi, M. Ahmadi, and S. R Niakan Kalhori, "Predicting covid-19 incidence through analysis of google trends data in iran: Data mining and deep learning pilot study," *JMIR Public Health Surveill*, vol. 6, no. 2, Apr 2020.

[14] L. J. Z. B. Wang P, Zheng X, "Prediction of epidemic trends in covid-19 with logistic model and machine learning technics," *Chaos Solitons Fractals*, 2020.

[15] P. K. S. Papastefanopoulos, V.; Linardatos, "Covid-19: A comparison of time series methods to forecast percentage of active cases per population," *Appl. Sci.*, vol. 10, 2020.

[16] J. P. Bigus, *Data mining with neural networks: solving business problems from application development to decision support*. Hightstown, NJ, USA: McGraw-Hill, Inc., 1996.

[17] P. Cabena, P. Hadjinian, R. Stadler, J. Verhees, and A. Zanasi, Eds., *Discovering Data Mining: From Concept to Implementation*. IBM Books, 1998.

[18] J. H. Orallo, M. J. R. Quintana, and C. F. Ramírez, Eds., *Introducción a la Minería de Datos*. Prentice Hall - Pearson Education, 2004.

[20] https://numpy.org/