Aprendizaje automático aplicado a la pandemia del virus Covid-19 en Argentina

Carolina Cardoso[1], Lorena Talame[1], Matias Amor[1]

Grupo de Análisis de Datos /Facultad de Ingeniería e IESIING
[1] Universidad Católica de Salta
Campo Castañares s/n, 4400 Salta, (0387) 426 8536
{acardoso, mltalame, mnamor}@ucasal.edu.ar

RESUMEN

Este proyecto se desarrolló con el interés de analizar los casos registrados sobre Covid-19 en nuestro país, publicados por fuentes oficiales. Se experimentó con redes neuronales con el fin de predecir casos positivos de la enfermedad, y para encontrar similitudes entre algunos distritos de nuestro país se plantearon relaciones difusas.

***Palabras clave*:** Minería de datos, aprendizaje automático, covid-19

CONTEXTO

Este proyecto continúa la línea de investigación que el Grupo de Análisis de Datos de la Facultad de Ingeniería de la Universidad Católica de Salta viene desarrollando en minería de datos.

1. INTRODUCCION

En diciembre de 2019, Wuhan, China, experimentó un brote de una enfermedad respiratoria causado por un nuevo coronavirus (COVID-19). En marzo de 2020 la OMS (Organización Mundial de la Salud) declaró al mundo en estado de pandemia. A la fecha, fueron confirmados más de 100 millones de personas contagiadas en el mundo.

En Argentina, con el objetivo de evitar la propagación del virus y detección de infectados se tomaron diferentes medidas desde el gobierno nacional [1], entre ellas, el aislamiento social, preventivo y obligatorio.

Si bien, el análisis de datos relacionados a diferentes enfermedades fue explorado ampliamente en la comunidad de investigadores de grandes volúmenes de datos, la exploración de los datos de las personas contagiadas con el nuevo coronavirus nos coloca en una situación especial. No solo por los descubrimientos que se van haciendo día a día por parte de los científicos, sino también por la rápida propagación mundial del virus.

En nuestro país encontramos estudios que toman como base los informes diarios del Ministerio de Salud, es decir, se trata de informes estadísticos que muestran la progresión día a día de la cantidad de casos positivos para Covid-19 y cantidad de decesos [2] [3].

En este proyecto se intentó encontrar similitudes entre algunos distritos argentinos a partir de las cantidades de casos diagnosticados con la enfermedad y con el aprendizaje de relaciones difusas. También se intentó construir una red neuronal para predecir la cantidad de casos positivos de Covid-19 en una provincia argentina.

Existen una variedad de herramientas informáticas para el análisis de datos y desarrollo de modelos para extracción de información. En este trabajo se utilizó el lenguaje de programación Python. Se destaca por ser uno de los más aceptados por la comunidad científica [4] y es uno de los lenguajes más potentes por su simplicidad. Además, es de distribución open source, posibilita la integración con múltiples librerías y por la experiencia de uso de los investigadores de este proyecto. Por otro lado, las herramientas de visualización de datos son complementarias a este proceso, facilitando la lectura y entendimiento de la información detectada [5].

2. LINEA DE INVESTIGACION Y DESARROLLO

El proyecto se desarrolló en las siguientes etapas

- Revisión de la literatura referente a las técnicas de aprendizaje automático aplicadas al problema, en reuniones periódicas del equipo.
- Descarga y almacenamiento de los datos, provenientes de fuentes oficiales (Ministerio de Salud de Nación Argentina, INDEC, etc.)
- Pre-procesamiento de los datos: filtrado, limpieza y selección de datos para entrenamiento y para control del modelo.
- Aplicación de técnicas y algoritmos adecuados para el problema del proyecto
- Validación y evaluación de los modelos obtenidos.

Se utilizó el lenguaje de programación Python y diferentes librerías para machine learning como scikit-learn, scikit-fuzzy y keras.

3. RESULTADOS OBTENIDOS/ ESPERADOS

Desde el mes de Marzo del año 2020, el Ministerio de Salud de la Nación Argentina emite reportes de la situación sanitaria del país, en relación a la situación sanitaria relacionada a Covid-19 [6]. Los casos se registran en el Sistema Nacional de Vigilancia Epidemiológica (SISA) y publicados en la página web de Datos Abiertos del Ministerio [7] en un archivo csv. Cada registro contiene información básica de las personas registradas en SISA, con diagnóstico positivo, negativo (descartado) o fallecido por la enfermedad.

Los datos utilizados en este proyecto se obtuvieron del archivo con fecha de actualización el 31 de octubre de 2020. El archivo consta de 1.048.575 registros y 25 atributos con información básica de las personas registradas: edad, sexo, provincia de residencia, entre otros.

Se realizaron dos tipos de tareas, por un lado, se experimentó con los datos de seis distritos argentinos con el objeto de encontrar similitudes a partir de la definición de relaciones difusas, y por otro, con redes neuronales para la predicción del número de casos positivos de Covid en una provincia.

A través del uso de relaciones difusas se puede definir la similitud entre los elementos de un conjunto como el grado de pertenencia de cada uno de esos elementos a ese conjunto. Se probaron distintos grados de pertenencia para los meses estudiados y se determinó la similitud de seis provincias argentinas.

 RedUNCI - UNdeC. ISBN: 978-987-24611-3-3

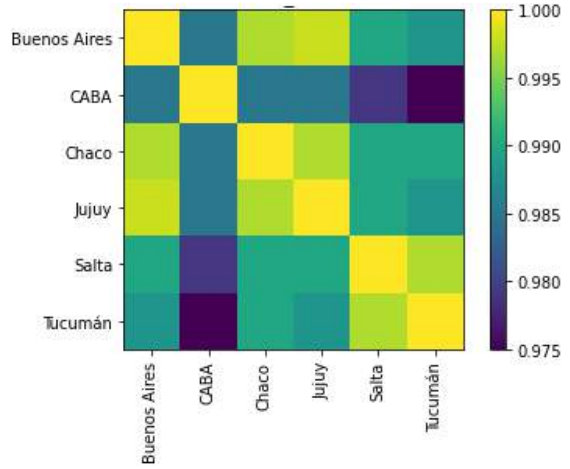

Figura 1. Mapa de calor relaciones de equivalencia mes de Octubre

De esta manera, se detectó por ejemplo, para el mes de octubre 2020, con un grado de pertenencia del 0.990, tres clases de equivalencias donde Buenos Aires y CABA son similares entre sí; Chaco, Jujuy y Tucumán pertenecen a la misma clase y Salta es la única provincia que se distingue del resto (Figura 1).

El modelo de red neuronal para predecir casos de contagios en la provincia de Salta que se probó fue una red secuencial con función de activación hiperbólica. Con esta red y utilizando siete días previos para predecir los casos del octavo se obtuvo un resultado aceptable con el menor error cuadrático medio. tanto en el entrenamiento como en la validación se mantuvo alrededor del 20%. Se pretende probar con otras configuraciones para lograr mejores resultados (Figura 2).

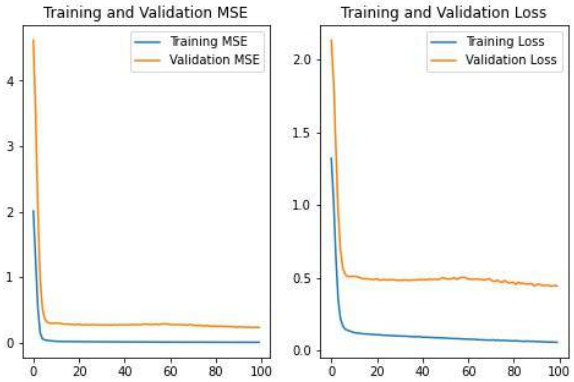

Figura 2. Evaluación

Se espera que esta línea de investigación continúe y amplíe los conocimientos sobre minería de datos. Se pretende que este proyecto anime el interés por la investigación y por esta temática a los alumnos de nuestra Facultad.

4. FORMACION DE RECURSOS HUMANOS

El equipo de trabajo está integrado por tres docentes de la carrera de Ingeniería en Informática. Se espera incorporar al proyecto alumnos interesados en la temática.

5. BIBLIOGRAFÍA

[1] M. d. S. d. l. N. Argentina, «¿Qué medidas está tomando el gobierno?,» 2020. [En línea]. Available: https://www.argentina.gob.ar/coronavirus/medidas-gobierno. [Último acceso: Mayo 2020].

[2] J. Aliaga, «Sobre Covid-19,» 2020. [En línea]. Available: http://www.jorgealiaga.com.ar/?p=1805.

[3] E. Iarussi, «Covid-19 en Argentina,» 2020. [En línea]. Available: https://observablehq.com/@eiarussi/covid-19-en-argentina/3.

[4] G. Piatetsky, «KDnuggets,» Mayo 2019. [En línea]. Available: https://www.kdnuggets.com/2019/05/poll-top-data-science-machine-learning-platforms.html.

[5] J. Hernández Orallo, Introducción a la minería de datos, Madrid: Pearson Prentice Hall, 2004.

[6] M. d. S. d. l. N. Argentina, «Informe diario Covid-19,» 2020. [En línea]. Available: https://www.argentina.gob.ar/coronavirus/informe-diario.

[7] Datos Abiertos, «Ministerio de Salud,» 2020. [En línea]. Available: http://datos.salud.gob.ar/dataset/covid-19-casos-registrados-en-la-republica-argentina.

[8] INDEC, «Instituto Nacional de Estadisticas y Censos - Bases de datos,» 2020. [En línea]. Available: https://www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos.

[9] R. Arias Montoya, J. J. Santa Chávez y J. D. J. Veloza Mora, «Aplicación del aprendizaje automático con árboles de decisión en el diagnóstico médico,» *Cultura del cuidado,* vol. 10, nº 1, pp. 63-72, 2013.

[10] I. Witten y E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 3 ed., Morgan Kaufmann, 2011.