Tecnología de Big Data en el análisis del estado de la pandemia por covid-19 en Colombia

Big Data technology in the analysis of the state of the covid-19 pandemic in Colombia

[1]Jorge Luis Quintero López, [2]Andrés Arismendi Ramírez, [3]Ángela Liceth Pérez Rendón

[1]Universidad Cuauhtémoc Aguascalientes, México
[2,3]Fundación universitaria del Areandina, Colombia

Recibido: 10/15/2021 Aprobado 11/20/2021

Resumen

En la actualidad de la pandemia, se presenta la necesidad de procesar grandes volúmenes de información generados por casos reportados positivos, con el fin de identificar patrones que conlleven a afrontar la emergencia con medidas de contingencia oportunas. En el presente estudio se plantea el tratamiento de un data set de la población general de Colombia, con información comprendida del mes de marzo y abril del 2021, con el fin de caracterizar, georreferenciar y predecir para darle valor a los datos, en busca de una comprensión de la dinámica del virus, para lo que se utilizaron tres modelos Naive Bayes, Random Forest y árboles J-48, buscando identificar aquel con mayor precisión; al usar el aplicativo Weka se llega a la conclusión de que el modelo que mejor se ajusta a la predicción, es el algoritmo de clasificación de árboles J-48 con un nivel de clasificación de instancias correctas de 99.24%, con un valor de Kappa de 0.9266 informando que se aproxima al 100 % de concordancia en la clasificación de las clases, con una cantidad, para este caso, de estudio de 221.583 clases y la predicción con 30 clases tomadas de la base original que consta de aproximadamente 2.774.465 datos. Al aplicar pruebas estadísticas se logra identificar la correlación entre los atributos, que llevan a garantizar el correcto modelado para la predicción. Este proceso se convierte en un insumo potencial para apoyar los procesos de administración de la sociedad y que beneficie las decisiones que se toman en términos de salud pública.

Palabras clave: predicción, machine learning, Sars-Cov-2, cuarentena.

Citación: Quintero López, J. L. ., Arismendi Ramírez, A. ., & Pérez Rendón, Ángela L. . (2021). Tecnología de Big Data en el análisis del estado de la pandemia por covid-19 en Colombia. Publicaciones E Investigación.

[1]Ciencia de los datos, Big Data. jquintero108@areandina.edu.co, https://orcid.org/0000-0002-0816-1592
[2]Ciencias Básicas. aarismendi2@areandina.edu.co, https://orcid.org/0000-0002-6578-5107
[3]Ciencias Básicas. aperez56@areandina.edu.co, https://orcid.org/0000-0002-2363-8782

https://doi.org/10.22490/25394088.5612

Abstract

At the present time of the pandemic, there is a need to process large volumes of information generated by reported positive cases, in order to identify patterns that lead to facing the emergency with timely contingency measures. In the present study, the treatment of a data set of the general population of Colombia is proposed, with information from the month of March and April 2021, in order to characterize, georeference and predict to give value to the data, in search of an understanding of the dynamics of the virus, for which three Naive Bayes, Random Forest and J-48 tree models were used, seeking to identify the virus with greater precision; When using the Weka application, it is concluded that the model that best fits the prediction is the J-48 tree classification algorithm with a classification level of correct instances of 99.24%, with a Kappa value of 0.9266 reporting that there is close to 100% concordance in class classification, with an amount, for this case, of study of 221,583 classes and the prediction with 30 classes taken from the original base consisting of approximately 2,774,465 data. By applying statistical tests, it is possible to identify the correlation between the attributes, which leads to guaranteeing the correct modeling for the prediction. This process becomes a potential input to support the management processes of society and that benefits the decisions that are made in terms of public health.

Keywords: *prediction, machine learning, Sars-Cov-2, quarantine.*

1. Materiales y métodos

Con el desarrollo de la investigación se accede a la base de datos del Ministerio de Salud de Colombia, a través de la página oficial de datos públicos (https://www.datos.gov.co/) la cual permite el acceso a diferentes temáticas de interés para que los investigadores del país generen insumos de análisis en pro de la mejora de las actividades públicas de Colombia. De allí se obtiene el data set del cual se depura y se filtran los atributos de interés para el estudio, quedando 6 ítems: ciudad de origen, edad, sexo, tipo de caso, estado de salud, clasificación por ciclo de vida.

Se extrae información en un periodo de tiempo estipulado para el estudio entre el 10 de mayo y el 9 de abril del año 2021; después de depurar la información se establece una muestra de 221.583 datos de individuos que han sido reportados como positivos en cada departamento, al identificar que el data set está en orden y no posee valores perdidos, se procede a través del software SPSS a realizar una análisis descriptivo de las variables para clasificar cada uno de ellos, seguido a esto se continua con análisis del tipo inferencial, como lo son análisis de correlación a través de pruebas de Chi-cuadrado y Rho de Spearman. Después de la clasificación estadística se procede implementando la aplicación Weka para la predicción de los modelos seleccionados para la presente investigación, identificando el de mayor exactitud para el cumplimento del objetivo propuesto. Finalmente se realiza un análisis a través de georreferenciación con el fin de ubicar características visuales en el estudio que permita un mayor entendimiento de los datos.

2. Resultados y discusión

Se estudiaron 221.174 personas 2,9% (117.101) de sexo femenino y 47,1% (104.073) masculino. La edad promedio de toda la población fue de 40,07 ±18,29. Al estudiar los pacientes de acuerdo a su condición se observó que el 1,3 % se encontraban activos con una edad promedio de 52,64 ± 21,17 años, un 2,7

% de los pacientes fallecieron durante el mes que se realizó el estudio marzo-abril con una edad media 68,61 ±14,82 y un 96 % de los pacientes atendidos en ese tiempo se recuperaron con un promedio de edad 39,11 ± 17,62. En el estudio también se logró identificar que durante los meses de marzo- abril el 23 % (50.799) de los pacientes con COVID- 19 fueron del departamento de Antioquia, un 18 % (39.744) de la capital del país (Bogotá) y un 14,8 % (32.744) de la ciudad de Barranquilla. Identificando que fue el departamento de Antioquia el que presentó mayor número de casos, de estos 12,2 % sexo femenino y 10,8 % masculino, como se representa en la georreferenciación en la Figura 1.

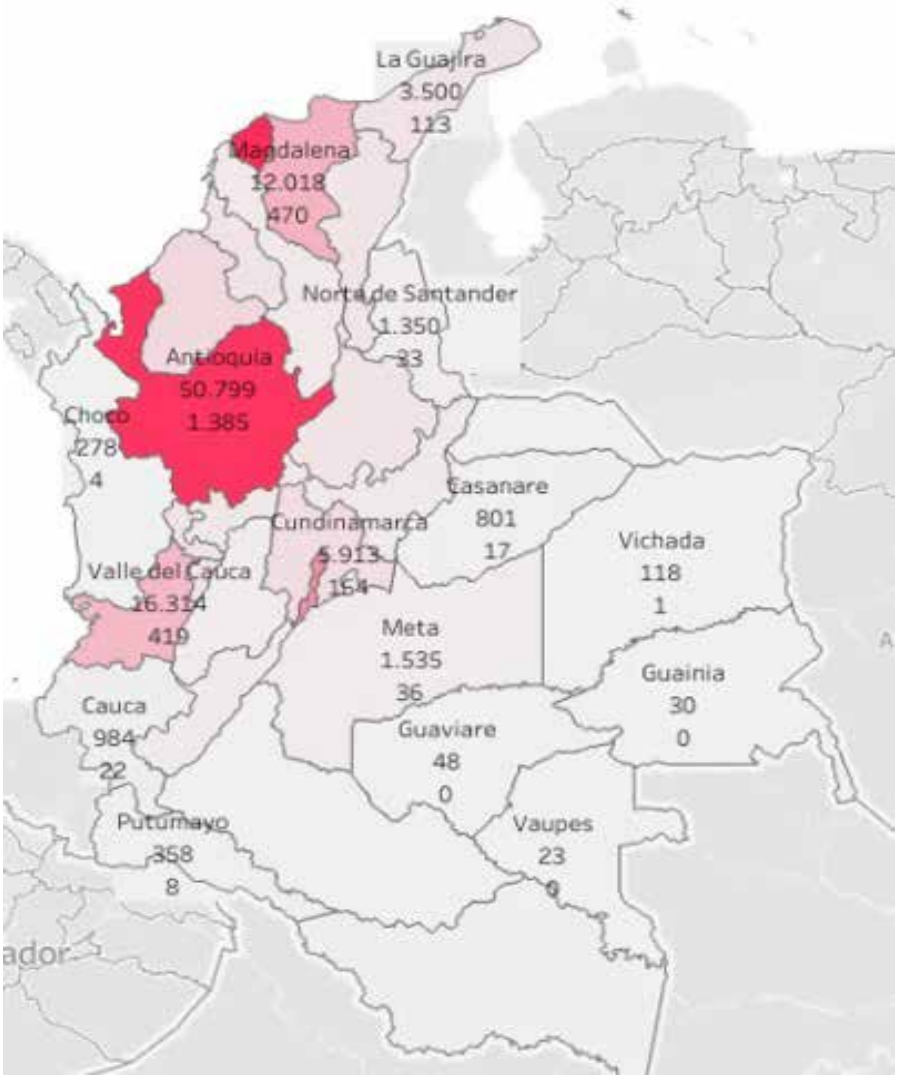

Figura 1. Georreferenciación del estado de la pandemia en el periodo estudiado

Al estudiar el estado de los pacientes se observó que 1,3 % (2.849) se encontraron en estado moderado y 95,7 % (211.554) su estado es leve como se muestra en la Tabla 1.

Tabla 1.

Estado de los pacientes según sexo

			Sexo		
			F	M	TOTAL
Estado	Leve	Recuento	112958	98596	211554
		% del total	51,1%	44,6%	95,7%
	Moderado	Recuento	1295	1554	2849
		% del total	0,6%	0,7%	1,3%
	Grave	Recuento	369	525	894
		% del total	0,2%	0,2%	0,4%

Posteriormente se realizó un modelo de regresión GLM Poisson simple en el cual se observa que a medida que los pacientes tenían mayor edad aumentaba la probabilidad de (RP:1,204, p<0,05), también se observa una mayor tendencia con el aumento de edad en relación de los pacientes de sexo femenino (RP: 1,028, $p < 0,05$). Además, se observa que existe una relación inversa con la edad y el estado del paciente a menor edad la condición del paciente es leve (B=-0,381; [0,666-0,700]).

Tabla 2.

Estimaciones de parámetro del modelo de regresión múltiple

Parámetro sexo	B	RP	IC 95%	Valor p
Femenino	0,028	1,028	[1,024 - 1,032]	0,000
Masculino	0[a]	1		
Estado				
Leve	-0,381	-0,683	[0,666 - 0,700]	0,000
Fallecido	0,186	1,204	[1,173 - 1,236]	0,000
Moderado	-0,082	0,922	[0,895 - 0,949]	0,000
Grave	0[a]	1		
Varable dependiente: Edad Modelo: (Intersección). Sexo, Estado				

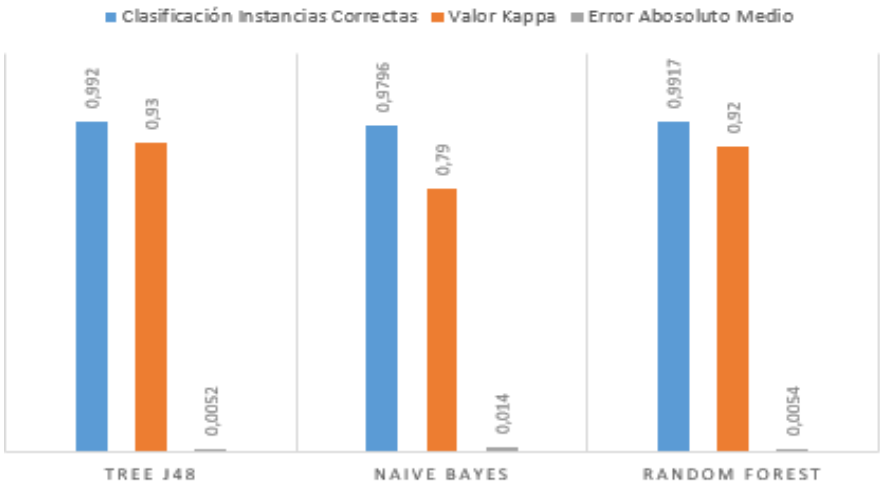

Figura 2. Validación cruzada de los modelos implementados

El resultado obtenido en la Figura 2 indica que el algoritmo de árboles J48 comparado con Naive Bayes y Random Frest obtuvo 99.2 %, de instancias correctamente clasificadas, siendo el algoritmo que mejor se ajusta al data set indicando un valor Kappa de 0.93 el cual es más alto en comparación con los modelos anteriores, aunque se espera un valor de 0,95 se encuentra muy cerca, adicionalmente el valor MAE indica 0,0052 lo cual demuestra que el modelo tiene la menor dispersión.

Así como en la actual investigación en Matilde *et al.* (2020), se analizaron los datos disponibles de pacientes en México hasta el 20 de abril, incluyendo 16 rasgos (físicos y clínicos) sobre cerca de 9.000 casos positivos (más de 700 fallecidos), con el foco en identificar patrones que predigan un desarrollo fatal de la enfermedad. Se emplearon técnicas de preparación y visualización de datos, selección de rasgos e inducción de reglas empleando el algoritmo J48, las redes neuronales y la teoría de los conjuntos aproximados, al igual que los resultados de esta investigación el algoritmo J-48 obtuvo un nivel de concordancia alto.

Por su parte en Medina Mendieta *et al.* (2020), Cuba realiza un análisis de su población, donde se evidencia que existe una adecuación de los modelos presentados con respecto a los valores pronosticados y los reales, lo cual permite una confiabilidad de los mismos para los pronósticos efectuados en dicho país, así mismo aplicando las técnicas de Big Data en esta investigación, se encuentran aciertos relacionados con la tendencia de crecimiento de los casos positivos en las principales ciudades de Colombia, todo esto en dependencia de las políticas aplicadas por las diferentes gobernaciones.

Desde Brasil en Kloeckner *et al.* (2020), a partir de una base de datos de imágenes digitalizadas de portaobjetos histopatológicos representativos de cáncer gástrico, identificamos tres patrones morfológicos de neoplasia, así como patrones de tejido no neoplásico, con los que entrenamos un algoritmo de red neuronal convolucional, diseñado para identificar y categorizar imágenes similares dentro de estos estándares, de forma automatizada. Temática que se puede ampliar en la actual investigación dando resultados a partir de las patologías de los pacientes.

3. Conclusiones

Con el uso de tecnologías del Big Data es posible realizar planteamientos que permitan estar mejor preparados para futuros estados de emergencia por pandemia, o mejor aún, evitar que ocurra. Son las tecnologías de la industria 4.0 como el Internet de las cosas (IoT), la minería de datos y la computación en la nube las que realizan aportes para que el monitoreo sea haga de manera constante en los diferentes entornos donde coexistimos los seres humanos, en la búsqueda de las anomalías biológicas y químicas que puedan presentarse y se consideren un riesgo para los seres vivos.

El margen de éxito en el que las herramientas del Big Data han ayudado a la sociedad desde la administración pública, hasta la reacción en temas de salud es muy alto ya que, con el procesamiento de información proveniente de datos estructurados y no estructurados, permite realizar predicciones con un alto porcentaje de acierto, todo esto basado en el estudio y entrenamiento de los modelos a partir de datos históricos de la pandemia y atributos de impacto en las medidas.

En la actual investigación se hace evidente la necesidad de agregar un mayor número de atributos a los data set de entrenamiento para aumentar la clasificación y obtener valores más elevados en la predicción del estado de salud de un paciente contagiado por covid-19, dichos atributos deben ser relacionados con las patologías de los pacientes positivos, lamentablemente dicha información aún es reservada en las instituciones de salud tratantes de los pacientes en mención.

Como muestran los resultados obtenidos, el entrenamiento de los modelos deja entrever que es posible realizar una predicción del estado de salud positivo para covid-19, teniendo en cuenta atributos como edad, sexo y ubicación geográfica, este último atributo deja en evidencia si la situación de contingencia está siendo tratada de una manera adecuada en cada una de las regiones del país.

Con el avance de los modelos predictivos se hace factible identificar cuales territorios tendrán impactos mayores en el sistema de salud además de identificar aquellos pacientes que se puedan complicar; ayudando a los centros médicos a contar con una mayor eficiencia en la prestación del servicio.

Agradecimiento

Agradecimiento de los autores a la página de datos abierta del Gobierno nacional, por el aporte constante de información para darle valor a los datos y una mejor compresión a la dinámica de la pandemia.

Referencias

Kloeckner, J., Sansonowicz, T. K., Rodrigues, Á. L., & Nunes, T. W. N. (2020). Multi-categorical classification using deep learning applied to the diagnosis of gastric cancer. *Jornal Brasileiro de Patologia e Medicina Laboratorial, 56*, 1–8. https://doi.org/10.5935/1676-2444.20200013

Matilde, M., Lorenzo, G., Ramón-Hernández, A., Bello- García, B., & Caballero, Y. (2020). Adquisición de conocimiento sobre la letalidad de la COVID-19 mediante técnicas de inteligencia artificial. *Anales de la Academia de Ciencias de Cuba, 10*(3), 1–12.

Medina-Mendieta, J., Cortés-Cortés, E., Cortés-Iglesias, M., Pérez-Fernández, A., & Manzano-Cabrera, M. (2020). Estudio sobre modelos predictivos para la COVID-19 en Cuba. *MediSur, 18*(3), 431–442.