# Analysis model of the most important factors in Covid-19 through data mining, descriptive statistics and random forest

1st Remigio Ismael Hurtado Ortiz
*Universidad Politécnica Salesiana*
Cuenca, Ecuador
rhurtadoo@ups.edu.ec

2nd Juan Carlos Barrera Barrera
*Universidad Politécnica Salesiana*
Cuenca, Ecuador
jbarrerab1@est.ups.edu.ec

3rd Katherine Michelle Barrera Barrera
*Universidad Politécnica Salesiana*
Cuenca, Ecuador
kbarrerab1@est.ups.edu.ec

***Abstract*—The Covid19 pandemic has had a great impact worldwide, it has become a major problem due to the demand for care in hospitals and clinics despite the low level of mortality. This is because the disease has spread rapidly as the spread between people is accelerated. So in this document we propose using a classification-oriented machine learning method, we do a classic data science process so that we can perform noise cleaning and data processing to do descriptive statistical analysis in such a way that the most important variables or factors are identified through unsupervised learning. And with this it is appreciated that the most important variables for the risk of infection and mortality that Covid-19 disease can have are diseases that affect the immune system, such as diabetes, heart disease, hypertension and also kidney disease. They can cause serious kidney problems. And the evaluation of our method will be carried out through quality measures. Finally, this work opens the door to other investigations with the aim of conducting centralized investigations on each variable related to Covid-19, in order to find relevant information that can promote an improvement in the current situation.**

***Index Terms*—Covid19, coronavirus, clustering, Random Forest, PCA, algoritmo, precisión, análisis.**

## I. Introducción

En esta sección se exponen los fundamentos y motivaciones principales del artículo.

Desde diciembre del 2019 el brote de la enfermedad por corona virus 19 conocido como COVID-19, es una enfermedad respiratoria originada en Wuhan, China que se transmite de persona a persona. (1) (2). Se han registrado 4,18 millones de casos y 286.000 muertes en más de 200 países hasta el 12 de mayo de 2020. (3) (4). La Organización Mundial de la salud OMS declaró el COVID-19 como una emergencia sanitaria mundial el 30 de enero del 2020. El rápido crecimiento de pacientes con COVID-19 provoca una escasez de medicamentos y las instalaciones médicas en su mayoría, no logra dar atención a todas las personas que tienen COVID-19 o son potencialmente portadores de la enfermedad.

### A. Fundamentos

La IA se aplica para la detección, prevención e incluso la predicción para contrarrestar la enfermedad de COVID-19. (5)Se espera desarrollar métodos automáticos que promueven ayuda y asistencia a personal médico. Artículos presentan modelos basados en red neuronal convolucional (4) (6) para la detección de pacientes con COVID-19 que utilizan imágenes de rayos X, igualmente del tórax CXR. (7) (8)

Los artículos relacionados con imágenes TC de tórax,(4) tienen gran impacto debido a que los métodos emplean regresión logística para la clasificación de COVID-19,(9) basados en características clínicas y de laboratorio. Al igual que el presente trabajo, usan un modelo de bosque aleatorio o Random Forest y presentan aprendizajes profundos para el diagnóstico de la enfermedad. Existen información sobre análisis de importancia variable en conjuntos de datos desequilibrados.(10)

De igual manera, se busca encontrar patrones, variables o síntomas destacables que tenga peso sobre un contagio o una mayor probabilidad de mortalidad en personas infectas.

### B. Motivación y método propuesto

Los síntomas predominantes para el COVID-19 son la fiebre y tos, mientras que los síntomas gastrointestinales se presentan con menos frecuencia, sin embargo, puede presentarse casos pacientes infectados por COVID-19 con la ausencia de fiebre, de modo que, un mecanismo de vigilancia puede pasar por alto a pacientes que están contagiados, porque la atención está dirigida hacia fiebre y no se da importancia a otros síntomas menos comunes.

Por lo que, el fin de esta investigación es analizar cuáles son las variables que pueden estar relacionadas con la enfermedad. El presente proyecto, busca agrupaciones relevantes entre los datos mediante un método de aprendizaje supervisado para clasificación conocida como Random Forest estadística descriptiva del conjunto de datos.

A continuación, se expone la estructura del artículo, en la sección II se presenta el trabajo relacionado, en la sección III se presenta el método propuesto, en la sección IV se expone el diseño de experimentos partiendo de la descripción

general del conjunto de datos y propiedades de las medidas de calidad para la evaluación. En la sección V se presentan los resultados y análisis correspondiente y finalmente, en la sección VI se encuentran las conclusiones y trabajos futuros de esta investigación.

Se presenta continuación el estado del arte y se expone sobre los trabajos relacionados, tanto generales como específicos.

## II. Estado del arte

En esta sección se ofrece una revisión de proyectos que buscan solventar el problema del COVID-19.

### *A. Trabajo relacionado con el COVID-19 en general*

Para el área de la ciencia de datos, el COVID-19 es una oportunidad para poder emplear estrategias que permiten obtener información relevante de entre grandes cantidades de datos, actualmente la cantidad de posibles variables que afectan al riesgo de contraer COVID-19 son muy diversas, existen múltiples trabajos que se realizan con aprendizaje profundo, el aprendizaje automático (11) , Big Data y la ciencia de datos como trabajo colaborativo para poder hallar una solución lo antes posible. Con estas herramientas se puede dar soluciones aceptables en el monitoreo, detección, diagnóstico y tratamiento de las enfermedades asociadas con el virus (12), El COVID-19 se manifiesta con características clínicas que van desde el estado asintomático a respiratorio agudo (13).

Según el resultado del estudio realizado por la OMS en colaboración con China, de los 55,924 casos de COVID-19 confirmados por laboratorio que fueron examinados, la mayoría exhibe características clínicas como fiebre, tos seca, fatiga y producción de esputo. Al mismo tiempo, solo una porción de los pacientes había mostrado síntomas como dolor de garganta, dolor de cabeza, mialgia y dificultad para respirar, mientras que síntomas como náuseas, congestión nasal, hemoptisis, diarrea y congestión conjuntival no se presentaban con mayor frecuencia. (13).

### *B. Trabajos relevantes de Machine Learning*

La mayor concentración de información se encuentra relaciona a las redes neuronales convoluciones (CNN) y pretenden en la mayoría de los trabajos citados, con las anomalías que se presentaban en las imágenes radiográficas después de la aparición de los síntomas por la enfermedad del Covid-19, buscan usar la radiografía de tórax como un principal método de detección. Estando relacionados, otra investigación expone mediante redes neuronales de convolución que buscan la segmentación del lóbulo pulmonar en la tomografía computarizada, que proponen un enfoque relacional (RTSU-Net) con el fin de que su modulo aprenda relaciones visuales y geométricas para tener características de convolución que producen pesos de atención propia. (14)

Entre otras investigaciones presentan el aprendizaje (ML) como técnica para predecir un evento con precisión que se basa en la experiencia, y proponen mediante su modelo conseguir una predicción importante para casos de Covid-19.(11). Debido a la falta de imágenes radiográficas que lo presentan como su problemática, presentan varias soluciones las cuales están asociadas a generación de imágenes sintéticas de rayos X de tórax (CXR), mediante Red Adversaria Generativa de Clasificadores Axiales (ACGAN) (3). Y sus resultados mejoran en un 10% de precisión.

Existen estudios basados en la sensibilidad de un individuo que pueden resultar de una combinación de terapia y Polimorfismo ACE2. Sugieren que los pacientes con enfermedades cardíacas, hipertensión o diabetes, que son tratados con ACE2, están en mayor riesgo para infección grave por COVID-19 (15). El diagnóstico rápido y preciso de COVID-19 en los casos sospechosos juegan un papel crucial en la cuarentena oportuna y tratamiento médico, que lo exponen en el articulo. Desarrollaron un aprendizaje profundo basado en un modelo para diagnóstico automático de COVID-19 en TC de tórax, afirman que es útil para contrarrestar el brote de SARS-CoV-2 (4). Debido a que la segmentación de lesiones de neumonía a partir de tomografías computarizadas de los pacientes con COVID-19 proponen que es importante para un diagnóstico preciso y seguimiento. Afirman que el aprendizaje profundo tiene el potencial de automatizar esta tarea.(6)

### *C. Trabajos relevantes en Clustering y Clasificación*

El artículo (16) propone un modelo de bosque aleatorio ajustado por el algoritmo AdaBoost. El modelo presenta una precisión del 94% y una puntuación F1 de 0,86 en el conjunto de datos utilizados. El análisis de los datos revela una correlación positiva entre el sexo de los pacientes y las muertes, y también indica que la mayoría de los pacientes tienen entre 20 y 70 años. La documentación acerca de enfermedades que aumentan la mortalidad en las personas incluye adultos mayores (personas mayores de 60 años) y personas con afecciones médicas existentes, como diabetes, hipertensión, asma, enfermedades cardiovasculares y enfermedad al riñón. (13)

El estudio relacionado presenta como problemática la información limitada, sobre la enfermedad renal en pacientes con COVID-19, por lo que en base a datos que obtienen de un estudio a 701 pacientes con COVID-19 ingresados en un hospital universitario terciario, obtienen resultados de prevalencia alta de lesión renal aguda (IRA) en pacientes hospitalizados por COVID-19 y el desarrollo de LRA durante la hospitalización y aseguran la asociación con mortalidad intrahospitalaria (17)

Se presenta a continuación el método propuesto, que busca mediante una estadística descriptiva, dar a conocer variables que afectan a potenciales pacientes y personas contagiadas, mediante aprendizaje no supervisado y predicción de posibles contagios a través de un modelo de bosque aleatorio (Random Forest)

## III. Método propuesto

En esta sección se presenta el método propuesto, se parte de una tabla de símbolos y parámetros y luego se describe el proceso mediante un diagrama y finalmente se especifica el algoritmo que tiene gran relevancia para el método.

El conjunto de datos completo se presenta en el repositorio GitHub, en donde se encuentra un cuaderno de Jupyter que contiene la experimentación del método, incluye el conjunto de datos, GitHub: https://github.com/Juancarlos56/Analisis-por-Clasificacion-Binaria-Covid19

### A. Tabla de parámetros y Algoritmo

En la siguiente tabla I se pude apreciar los parámetros o símbolos que se utilizan en el algoritmo III-C.

TABLE I
PARÁMETROS Y SÍMBOLOS USADOS EN EL ALGORITMO

| Variable | Descripción |
|---|---|
| pca | Análisis de componentes principales |
| kmeans | Método de agrupamiento |
| ac | accuracy, % elementos clasificados correctamente |
| f1 | Puntuación F1, compara rápidamente dos clasificadores |
| pr | Recall, elementos identificados correctamente |
| K | Número de grupos seleccionados |
| RFC | Método de clasificación random forest |
| dfP | Datos de personas encuestadas con pre-procesamiento |
| mae | Error absoluto medio |
| mae | Datos de personas encuestadas con PCA |
| mse | Desviación cuadrática media |
| rmse | Desviación cuadrática media cuadrática |
| distorsión | suma de errores cuadrados (SSE) |

### B. Diagrama general del proceso

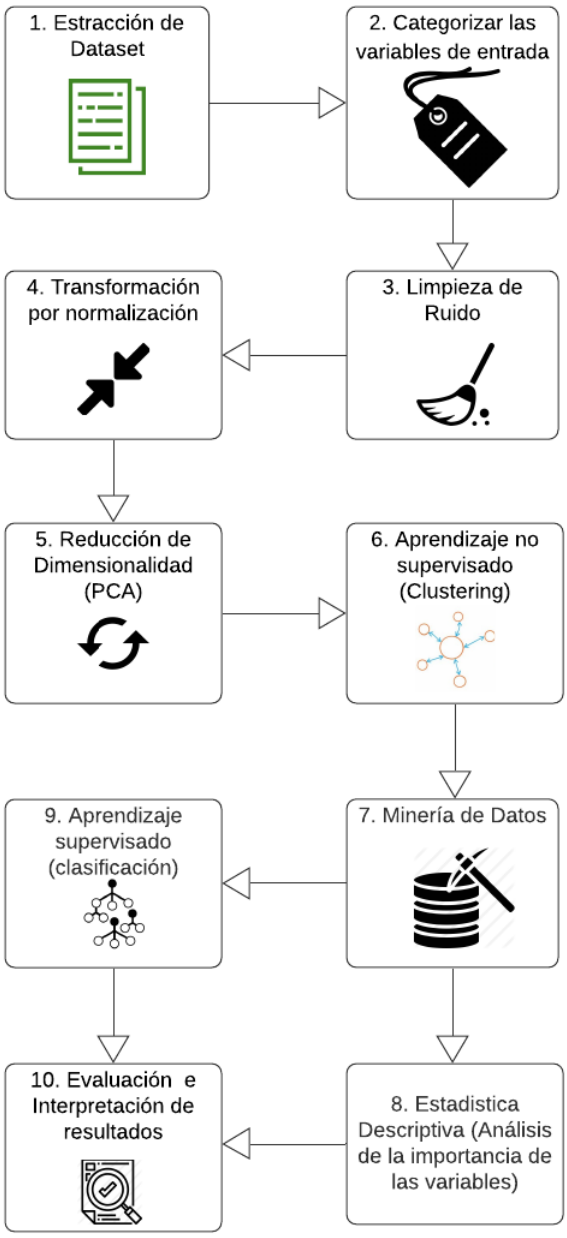


Fig. 1. Diagrama general del proceso

1) Seleccionar un dataset que cuenta con variables que pueden ser de importancia para un estudio, de modo que aporte información importante y crucial para el estudio.
2) La categorización de variables de entrada, con el fin de cambiar las variables categóricas a numéricas para obtener valores numéricos que representen la misma cantidad de información.
3) Para limpiar los datos, que pueden ser nulos, vacíos o blancos, fueron sustituidos por los cálculos de media o moda, dependiendo del resultado buscado y lo que se crea más acertado para el dataset.
4) El procedimiento de transformación se realiza mediante el método de escalamiento para variables del conjunto de datos.
5) La reducción de dimensionalidad de datos es importante porque es importante identificar y eliminar aquellas variables que no son relevantes, provocando un mejor rendimiento computacional.
6) Se aplica el método del codo para elegir el número de clusters adecuados para los datos que se presentan. Con el algoritmo de Clustering k-mean se encuentran e identifican aquellas relaciones de grupos ocultos a simple vista.
7) Una vez que el algoritmo identifica en los registros las características similares, secciona en grupos con una cantidad de registros diferentes en cada uno.

8) Con los clusters (grupos) que se obtiene, se realiza un análisis de cada una de las variables que intervienen en el diagnóstico de COVID-19.
9) De los clusters que se obtienen se selecciona aquel que presente el mayor número de casos positivos de COVID-19, para determinar las variables más influyentes en este grupo específico de personas, que permite desarrollar un método de aprendizaje supervisado para clasificación conocido como Random Forest que permite determinar si una persona puede llegar a contraer o no COVID-19.
10) Finalmente, se visualiza los resultados que se obtienen después de seguir estos pasos, mediante gráficas y tablas explicativas de los procesos.

### C. Algoritmo pseudocódigo

Se presenta un algoritmo que especifica el proceso del método, para obtener los resultados de este articulo.

**Algorithm 1** Clasificación Binaria de Covid-19

**Input:** $df, K, dfP, df_PCA$

**Output:** ac, f1, rc, pr, mae, mse, rmse, distorsión, IDC (medidas de calidad)

Carga de Datos de personas encuestadas (df)
Clasificación en Variables numéricas y categóricas
Limpieza de ruido: Selección de columnas, cambio por medio o moda
Transformación: normalización (0,1) y estandarización (media en 0, varianza en 1)
Reducción de dimensionalidad (PCA)
Clustering mediante kMeans (**return** IDC, distorsión = kMeans(df_PCA, K))
Estadística descriptiva de cada grupo obtenida mediante el paso anterior
Clasificación mediante random forest: RFC($X_{train}, y_{train}$)
**return** ac, f1, rc, pr, mae, mse, rmse
Visualización de resultados: correlaciones y variables más influyentes. =0

A continuación se expone el diseño de experimento que se lleva a cabo durante esta investigación.

## IV. Diseño de experimentos

En esta sección se explica la descripción general de nuestro dataset, parámetros seleccionados y medidas de calidad.

### A. Descripción general del Dataset

El método de estudio utiliza un dataset con 820580 de los cuales un porcentaje del 0.23% que pertenece a personas con Covid positivó, mientras que el 99.97% presenta un diagnóstico negativo de Covid. El dataset original que se encuentra en la página "Nexoid" (18). En la tabla, se especifica características del conjunto de datos. IV

Para trabajar con los datos a una misma escala, se hace una transformación de variables, en donde a variables numéricas y categóricas ordinales se aplica una normalización con la media en 0 y varianza en 1, mientras que, para las variables categóricas nominales, se implementa una columna para cada valor distinto que exista en la columna original; con esta transformación se obtiene un dataset con 64 columnas.

TABLE II
Datos de entrada de covid-19

| DataSet | Original |
|---|---|
| Número de Variables | 57 |
| Número de Registros | 820580 |

### B. Parámetros seleccionados

Con los datos a una misma escala, se puede trabajar con reducción de dimensionalidad (PCA (19)) que permite minimizar el esparcimiento de los datos y trabajar con una matriz con menos variables correlacionadas, debido a esto se experimenta con el 85% de la varianza del dataset.

Posterior al proceso de reducción de dimensionalidad, se lleva a cabo el algoritmo de aprendizaje no supervisado (k-Means). En donde, se trabaja desde 2 hasta 12 clúster, cada clúster devuelve su distorsión, el cual ayuda a determinar el K(20) adecuado para el algoritmo. Se toma el clúster con el mayor número de infectados por el Covid-19.

Con este clúster se realiza un submuestreo para tener una semejanza entre las personas que dieron negativo y positivo, y así evitar un desbalance de los datos, se tiene en cuenta que se secciona el conjunto de datos en un 80% que son destinados para entrenar el algoritmo, mientras que el otro 20% está destinado para realizar y comprobar las predicciones del algoritmo, ver tabla III

TABLE III
Conjunto de datos para Aprendizaje

| Datos para predicción RFC | |
|---|---|
| Datos para Entrenamiento | 2488 |
| Datos para Prueba | 623 |

De igual manera se trabaja con PCA, pero se enfoca en probar desde 2 hasta 50 componentes principales y se observa el comportamiento del porcentaje total de elementos clasificados correctamente y posterior se toma el número de componentes que ofrecen el mayor porcentaje de acierto. Con este número de componentes principales, se llega a entrenar al algoritmo de aprendizaje supervisado (Random Forest).

### C. Medidas de calidad

Para las métricas de evaluación se toman como referentes

1) accuracy: Se define como el número total de predicciones correctas dividido por el número total de predicciones. Se refiere a lo cerca que está el resultado de

una medición del valor verdadero. Se representa por la proporción entre los positivos reales predichos por el algoritmo y todos los casos positivos.

$$accuracy = \frac{\text{True Positive} + \text{True Negative}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

2) Recall: También se conoce como Tasa de Verdaderos Positivos. Es la proporción de casos positivos que fueron correctamente identificadas por el algoritmo.

$$Recall = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

3) F1 score: Esta es otra métrica muy empleada porque resume la precisión y sensibilidad Precisión y Recall en una sola métrica por ello es de gran utilidad cuando la distribución de las clases es desigual.

$$F_1 = 2 \times \frac{\text{Precisión} \times \text{Recall}}{\text{Precisión} + \text{Recall}}$$

4) Precisión: Se refiere a la dispersión del conjunto de valores obtenidos a partir de mediciones repetidas de una magnitud. Cuanto menor es la dispersión mayor la precisión. Se representa por la proporción entre el número de predicciones correctas (tanto positivas como negativas) y el total de predicciones.

$$\text{Precisión} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

En la siguiente sección se da a conocer los parámetros seleccionados, las variables más influyentes, distribución de usuarios por cada clúster, variables más correlacionadas con la variable de salida (covid-positive), obtención de medidas de calidad del algoritmo, estadística descriptiva del dataset como también del clúster con el que se trabaja.

## V. Resultados y discusión

En este apartado se sigue el método del propuesto y se presentan los resultados.

### A. Procesamiento de datos

En esta sección se presenta los resultados del procesamiento de datos. Se realiza la limpieza del ruido, si la variable presenta un 82% de valores de vacíos, se elimina esta variable. Por otro lado, si la variable presenta alrededor de 15% se reemplaza los valores vacíos por la media, y si la variable tiene una salida binaria, se reemplaza el valor por la moda. En la tabla IV se visualiza el número de datos modificados.

TABLE IV
Variables y atributos del dataset procesado

| DataSet | Original | Modificado |
|---|---|---|
| Número de Variables | 58 | 47 |
| Número de Filas | 820580 | 820580 |
| Variables Numéricas | 41 | 47 |
| Variables Categóricas | 6 | 0 |

### B. Reducción de dimensionalidad

En esta sección se presenta los resultados al aplicar reducción de dimensionalidad (PCA). Para la elección del número de componentes principales se calcula la proporción de variación de cada columna hasta tener un 85% de variabilidad, la cantidad óptima es de 29 componentes principales, como se muestra en la siguiente figura 2.

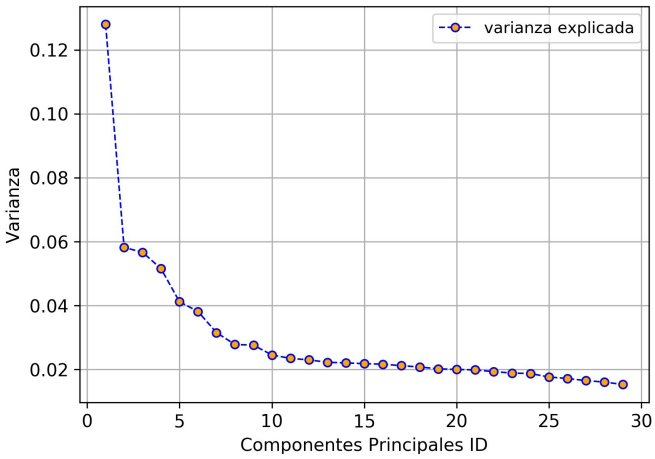


Fig. 2. Varianza de PCA con el 0.85% de la varianza

### C. Aprendizaje no supervisado

Posterior al proceso de reducción de dimensionalidad, se lleva a cabo el algoritmo de aprendizaje no supervisado (k-Means), con el propósito de obtener grupos equilibrados que tienen características similares, en base al método del codo, se selecciona el mejor número de grupos. Ver figura 3, que en este caso el resultado número 8 la cantidad adecuada.

En la tabla V, se presenta el número de registros que se agruparon por cada clúster. Se puede apreciar que el número de personas por cada grupo es balanceado. Agrupando 8 clusters por la similitud entre los datos.

### D. Aprendizaje supervisado

En esta sección, se presenta los resultados que se obtienen después de selección al clúster número 7, porque presenta mayor número de infectados por COVID-19. En la siguiente tabla VI, se muestra información que se obtuvo del clúster seleccionado.

TABLE V
PARAMETERS USED IN THE EXPERIMENTS TO RUN BASELINE METHODS.

| Clúster | # Personas por Grupo | Covid-19 Negativo | Covid-19 - Positivo |
|---|---|---|---|
| 0 | 93142 | 92984 | 158 |
| 1 | 95156 | 95137 | 19 |
| 2 | 172403 | 172382 | 21 |
| 3 | 10741 | 10694 | 47 |
| 4 | 388773 | 388773 | 0 |
| 5 | 12376 | 12276 | 100 |
| 6 | 10060 | 9918 | 142 |
| 7 | 37729 | 36546 | 1383 |

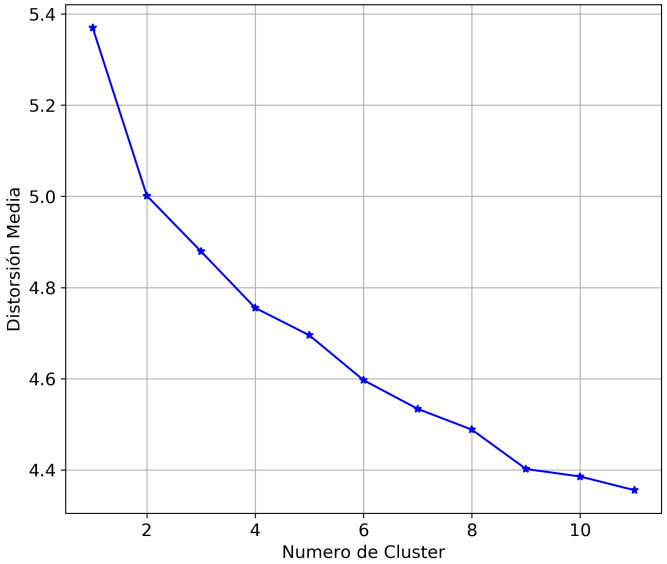


Fig. 3. Distorsión de cada Clúster

TABLE VI
INFORMACIÓN DEL CLUSTER CON MAYOR CANTIDAD DE PACIENTES CON COVID-19

| Información de Clúster seleccionado | |
|---|---|
| Clúster número | 7 |
| Número de filas | 37929 |
| Número de columnas | 65 |
| Cantidad de Pacientes sin covid-19 | 36546 |
| Cantidad de Pacientes con covid-19 | 1383 |

.

Posteriormente el método de aprendizaje supervisado para clasificación aplicado es Random Forest. Mediante análisis de componentes principales (PCA), con 24 componentes principales se obtiene el mayor porcentaje total de elementos clasificados correctamente. Como se puede visualizar mediante la gráfica de la figura V-D.

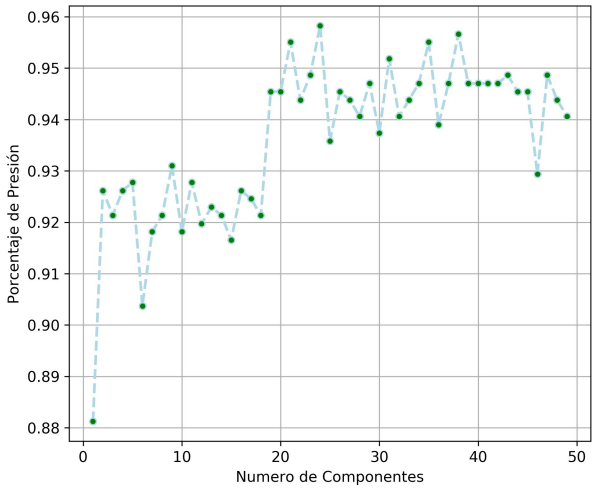


Fig. 4. Porcentaje de elementos clasificados correctamente (accuracy)

### *E. Resultados por medidas de calidad*

En la Tabla VII se exponen los resultados de las medidas de calidad en porcentaje, los aciertos y los errores que se obtuvieron según cada métrica.

TABLE VII
MEDIDAS DE CALIDAD, PARA MÉTODO DE RANDOM FOREST

| Covid-19 | Precisión | Recall | F1 | accuracy |
|---|---|---|---|---|
| 0 | 0.95 | 0.97 | 0.96 | 0.95 |
| 1 | 0.96 | 0.93 | 0.94 | 0.95 |

Como se observa en la tabla VII vemos que el método posee una alta precisión y alto recall, tanto en la predicción de personas que no tienen COVID-19 como también con las que lo pueden contraerlo, por lo que se puede decir que el método aprende muy bien a identificar si una persona puede o no llegar a contraer esta enfermedad.

### *F. Estadística Descriptiva sobre el conjunto de datos completo*

A continuación, se exponen los resultados que se obtienen mediante una estadística descriptiva. Se presentan explica-

ciones breves sobre las variables, que están relacionas con la enfermedad del COVID-19.

El estudio está centrado en variables, como la opinión de mortalidad que influye en un posible contagio. Una manera de controlar la propagación del COVID-19 es mantener a una persona informada, sobre los riesgos y medidas que se puede tomar ante la enfermedad, como la evolución que puede tener la misma.

Se muestra un rango de edad de 10 años que a pesar de contraer la enfermedad su posibilidad de mortalidad es baja, a diferencia de los rangos de edad entre 80 a 110 años, que son altamente vulnerables hacia el contagio de la enfermedad COVID-19 y una probabilidad mayor de riesgo de mortalidad. Se recomienda optar por una enfermera/o en casa con las medidas de protección, con el fin de realizar cuidados adecuados al paciente. Esta medida puede reducir el riesgo de contagio o el riesgo de mortalidad por COVID-19, según los datos que se obtienen del estudio.

Si la persona padece de las siguientes enfermedades de hígado y de riñón que puede ser causado por el uso de analgésicos en periodos prolongados. Y al ser uno de los órganos altamente afectado por el virus COVID-19, que puede provocar una lesión renal en los pacientes, aumentando considerablemente su riesgo de mortalidad.

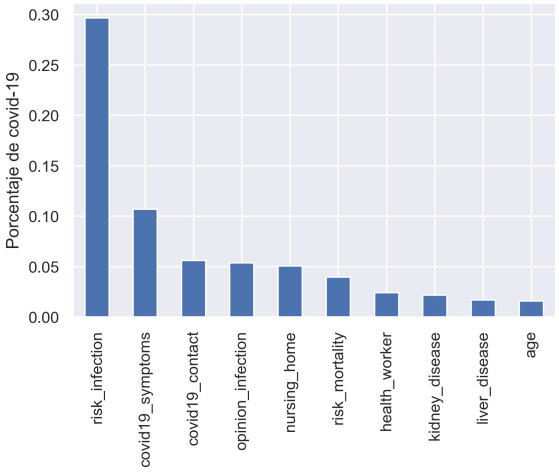


Fig. 5. Variables influyentes en COVID-19

En la siguiente gráfica 6, se presenta las enfermedades que aumentan el riesgo de mortalidad de una persona con COVID-19. La diabetes contribuye a un riesgo de mortalidad mayor porque el sistema inmunitario se ve comprometido, y el periodo de recuperación puede ser más largo, luego se muestra la hipertensión y enfermedades cardiacas. La siguiente variable es el género, los hombres tienen más posibilidad de morir por COVID-19 una vez contagiados. La variable peso, se supone que está comprometido, porque tiene relevancia para pacientes con hipertensión, diabetes o en algunos casos para enfermedades del corazón.

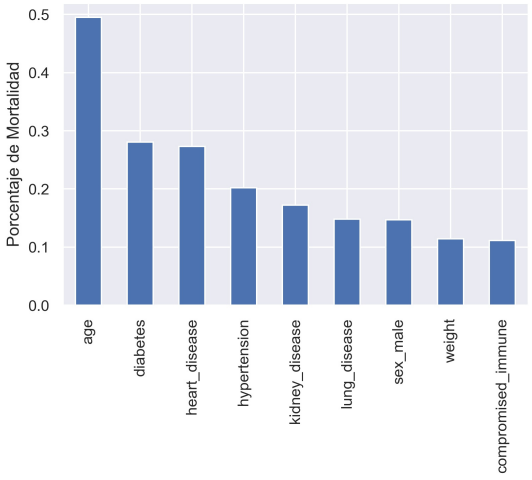


Fig. 6. Riesgo mortalidad en el clúster 7

En la siguiente sección se expone una conclusión general sobre el presente proyecto y los trabajos futuros al que se presta esta investigación.

## Conclusiones

Se logra en el artículo obtener las variables con mayor peso sobre un COVID-19 positivo, en base al conjunto de datos que se estudia. Y las enfermedades que afectan el riesgo de mortalidad en los pacientes, las enfermedades relacionas con problemas en el sistema inmunológico, como la diabetes, enfermedades del corazón, hipertensión y enfermedades del riñón. El método propuesto ha sido eficiente para el conjunto de datos seleccionado.Para trabajos futuros, se propone tomar el presente trabajo como una investigación base. Las variables como enfermedad de riñón necesitan de estudios avanzados, para realizar una observación a los medicamentos que pueden afectar a la persona que tiene COVID-19 o cuales son los componentes de la medicina que provoca inestabilidad en el sistema inmunitario o la deficiencia renal en pacientes de Covid-19.

## References

[1] R. Pung, C. J. Chiew, B. E. Young, S. Chin, M. I. Chen, H. E. Clapham, A. R. Cook, S. Maurer-Stroh, M. P. Toh, C. Poh *et al.*, "Investigation of three clusters of covid-19 in singapore: implications for surveillance and response measures," *The Lancet*, 2020.

[2] J. Cobb and M. Seale, "Examining the effect of social distancing on the compound growth rate of sars-cov-2 at the county level (united states) using statistical analyses and a random forest machine learning model," *Public Health*, 2020.

[3] A. Waheed, M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. Pinheiro, "Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection," *IEEE Access*, vol. 8, pp. 91 916–91 923, 2020.

[4] X. Wang, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, and C. Zheng, "A weakly-supervised framework for covid-19 classification and lesion localization from chest ct," *IEEE Transactions on Medical Imaging*, pp. 1–1, 2020.

[5] M. A. Mohammed, K. H. Abdulkareem, A. S. Al-Waisy, S. A. Mostafa, S. Al-Fahdawi, A. M. Dinar, W. Alhakami, A. BAZ, M. N. Al-Mhiqani, H. Alhakami, N. Arbaiy, M. S. Maashi, A. A. Mutlag, B. García-Zapirain, and I. D. L. T. De La Torre Díez, "Benchmarking methodology for selection of optimal covid-19 diagnostic model based on entropy and topsis methods," *IEEE Access*, vol. 8, pp. 99 115–99 131, 2020.

[6] G. Wang, X. Liu, C. Li, Z. Xu, J. Ruan, H. Zhu, T. Meng, K. Li, N. Huang, and S. Zhang, "A noise-robust framework for automatic segmentation of covid-19 pneumonia lesions from ct images," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2653–2663, 2020.

[7] H. A. Rothan and S. N. Byrareddy, "The epidemiology and pathogenesis of coronavirus disease (covid-19) outbreak," *Journal of autoimmunity*, p. 102433, 2020.

[8] I. Gabriella, S. A. Kamarga, and A. W. Setiawan, "Early detection of tuberculosis using chest x-ray (cxr) with computer-aided diagnosis," in *2018 2nd International Conference on Biomedical Engineering (IBIOMED)*, 2018, pp. 76–79.

[9] F. Shi, L. Xia, F. Shan, D. Wu, Y. Wei, H. Yuan, H. Jiang, Y. Gao, H. Sui, and D. Shen, "Large-scale screening of covid-19 from community acquired pneumonia using infection size-aware classification," *arXiv preprint arXiv:2003.09860*, 2020.

[10] I. Ahrazem Dfuf, J. Forte Pérez-Minayo, J. M. Mira Mcwilliams, and C. González Fernández, "Variable importance analysis in imbalanced datasets: A new approach," *IEEE Access*, vol. 8, pp. 127 404–127 430, 2020.

[11] F. Rustam, A. A. Reshi, A. Mehmood, S. Ullah, B. On, W. Aslam, and G. S. Choi, "Covid-19 future forecasting using supervised machine learning models," *IEEE Access*, vol. 8, pp. 101 489–101 499, 2020.

[12] J. M. Díaz, "Inteligencia artificial y big data como soluciones frente al covid-19," *Revista de Bioética y Derecho*, vol. 0, no. 50, pp. 315–331, 2020. [Online]. Available: https://revistes.ub.edu/index.php/RBD/article/view/31643

[13] V. Chamola, V. Hassija, V. Gupta, and M. Guizani, "A comprehensive review of the covid-19 pandemic and the role of iot, drones, ai, blockchain, and 5g in managing its impact," *IEEE Access*, vol. 8, pp. 90 225–90 265, 2020.

[14] W. Xie, C. Jacobs, J. Charbonnier, and B. van Ginneken, "Relational modeling for robust and efficient pulmonary lobe segmentation in ct scans," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2664–2675, 2020.

[15] L. Fang, G. Karakiulakis, and M. Roth, "Are patients with hypertension and diabetes mellitus at increased risk for covid-19 infection?" *The Lancet. Respiratory Medicine*, vol. 8, no. 4, p. e21, 2020.

[16] 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.

[17] Y. Cheng, R. Luo, K. Wang, M. Zhang, Z. Wang, L. Dong, J. Li, Y. Yao, S. Ge, and G. Xu, "Kidney disease is associated with in-hospital death of patients with covid-19," *Kidney international*, 2020.

[18] Nexoid, "Dataset covid-19," 2020. [Online]. Available: https://www.covid19survivalcalculator.com/es/research

[19] B. D. V. Moreno, R. I. H. Ortiz, and D. A. M. Rivera, "A new approach hybrid recommender system of item bundles for group of users," in *2019 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*. IEEE, 2019, pp. 1–7.

[20] R. I. H. Ortiz, D. A. M. García, and A. S. E. Mancheno, "A new way of finding better neighbors in recommendation systems based on collaborative filtering," in *2019 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC)*. IEEE, 2019, pp. 1–6.