GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

TRABAJO DE FIN DE GRADO

MACHINE LEARNING PARA EL DIAGNÓSTICO DE COVID-19

Autor: Elvira Bausili Llamas

Director: Ángel González Prieto

(Universidad Autónoma de Madrid)

Madrid, 2021

Declaro, bajo mi responsabilidad, que el Proyecto presentado con el título

"Machine Learning para el diagnóstico de COVID-19"

en la ETS de Ingeniería - ICAI de la Universidad Pontificia Comillas en el

curso académico 2020/21 es de mi autoría, original e inédito y

no ha sido presentado con anterioridad a otros efectos.

El Proyecto no es plagio de otro, ni total ni parcialmente y la información que ha sido

tomada de otros documentos está debidamente referenciada.

Fdo.: Elvira Bausili Llamas Fecha: 23 / 06 / 2021

Autorizada la entrega del proyecto

EL DIRECTOR DEL PROYECTO

Fdo.: Ángel González Prieto Fecha: 23 / 06 / 2021

GRADO EN INGENIERÍA EN TECNOLOGÍAS INDUSTRIALES

TRABAJO DE FIN DE GRADO

MACHINE LEARNING PARA EL DIAGNÓSTICO DE COVID-19

Autor: Elvira Bausili Llamas

Director: Ángel González Prieto

(Universidad Autónoma de Madrid)

Madrid, 2021

Agradecimientos

Principalmente, quería agradecer a mi tutor de proyecto Ángel González Prieto todo el apoyo y tiempo invertido en este trabajo. Además de proporcionar todos los recursos logísticos y académicos necesarios para lograr los excelentes resultados finales, ha sido una gran ayuda e inspiración respecto a la potencia que puede tener nuestro trabajo como ingenieros en la mejora de la sociedad.

Gracias por haber hecho de este Trabajo de Fin de Grado un proyecto con el que me he ilusionado y he crecido, tanto académica como personalmente. Hay ideas con mucho potencial que solo están esperando a que alguien se ponga manos a la obra, y creo que este proyecto podría implementarse y tener un gran impacto en la recuperación final de esta pandemia global que tan radicalmente ha cambiado nuestro mundo.

Por otra parte, agradecer a esos profesores de la facultad de ingeniería ICAI de los que tanto he aprendido en estos años. Y por supuesto a mi familia, en especial a mis padres; que nunca han dejado de apoyarme y creer en mí.

MACHINE LEARNING PARA EL DIAGNÓSTICO DE COVID-19

Autor: Bausili Llamas, Elvira.

Director: González Prieto, Ángel – Universidad Autónoma de Madrid

Entidad Colaboradora: ICAI – Universidad Pontificia Comillas

RESUMEN DEL PROYECTO

Este trabajo analiza el uso de inteligencia artificial como alternativa para la detección de COVID-19 en radiografías. Para ello, se realizará un estudio comparativo aplicando modelos de Machine Learning a tres bases de datos de radiografías diferentes.

Tomando en consideración los resultados iniciales obtenidos, se procederá a la obtención del modelo con desempeño óptimo para la detección de dicha enfermedad según las métricas de rendimiento establecidas, haciendo uso de una base de datos que englobe la totalidad de radiografías disponibles. Finalmente, se llegará la conclusión de que el método más adecuado para la detección de COVID-19 es K-Nearest Neighbors (KNN) aplicado a imágenes en crudo, sin ningún preprocesado previo.

Palabras clave: COVID-19, radiografías, Machine Learning, procesamiento de imágenes, Gradient Boosting, KNN.

6. Introducción

La crisis sanitaria provocada por la aparición y propagación del COVID-19 es un problema de gran actualidad que nos afecta a todos en nuestro día a día. La intención de este Trabajo de Fin de Grado es hacer uso de los conocimientos técnicos de ingeniería para tratar de buscar una solución alternativa a los métodos actuales de detección de dicha enfermedad. De esta forma, se lograría un ahorro en medios económicos y de personal sanitario, que podría reorientarse al cuidado de las personas con sintomatología grave y al desarrollo del plan de vacunación.

7. Definición del Proyecto

La finalidad del presente Proyecto de Fin de Grado es la obtención de un modelo automático de detección de pacientes afectados por COVID-19 a partir de radiografías haciendo uso de

herramientas de Machine Learning. Los pacientes se clasificarán en personas con COVID-19, con neumonía o sin enfermedad.

Este trabajo cuenta con una fase de desarrollo preliminar en la que se lleva a cabo el **minado de datos multifuente y el preprocesamiento de las imágenes** para su utilización posterior en los modelos seleccionados. De acuerdo con la bibliografía consultada, se partirá de la idea de que un procesamiento ligero llevará asociados mejores resultados de clasificación. [1]

A continuación, se llevará a cabo la programación de los **modelos de clasificación** Naïve Bayes, KNN, Random Forest y Gradient Boosting para su posterior aplicación en las tres bases de datos por separado. Con el fin de aportar un mayor realismo y aplicabilidad a nuestro trabajo, se procurará la **optimización del modelo** a utilizar para una base de datos global (Global Database), en la que se aplicarán las conclusiones extraídas del análisis comparativo previo. Asimismo, se considerarán las **métricas de rendimiento**, la proporción de casos correctamente diagnosticados con COVID-19 y la ausencia de falsos negativos para determinar las características del modelo más adecuado para el objetivo final de este proyecto.

8. Descripción del modelo

Nuestro proyecto da comienzo con la **lectura de las radiografías**, asignando la **clase de cada imagen** con un valor numérico. Las categorías utilizadas son: **3 (COVID-19), 2 (neumonía) y 1 (normal).**

Los algoritmos de aprendizaje automático requieren de un conjunto de datos de entrenamiento (train) y otro de comprobación del modelo (test). [2] De esta forma, en nuestro trabajo se asigna una **distribución aleatoria de imágenes para train y test,** constituida por un 80%-20% de las radiografías en las bases de datos en las que no viene separado previamente. En la primera parte del proceso, se hace uso de las radiografías de train para el aprendizaje del modelo, que será evaluado posteriormente con las imágenes de test.

Típicamente, los modelos de Machine Learning se construyen en base a ciertas características que deben determinarse previamente, los denominados **hiperparámetros**. [2] En los programas que hacen uso de estos hiperparámetros, antes de ejecutar el programa se realizará una **búsqueda de los valores óptimos** para los mismos mediante el análisis previo del modelo con GridSearch.

Comparando los resultados obtenidos de clasificación de las imágenes de test con el modelo y la categoría real de las mismas, se obtienen las **métricas de eficiencia** correspondientes. Finalmente, y con el fin de obtener una mayor visibilidad en los resultados obtenidos, se elaboran las **matrices de confusión** en magnitudes reales y unitarias, lo que nos permite discernir entre los casos correctamente asignados, falsos negativos y falsos positivos.

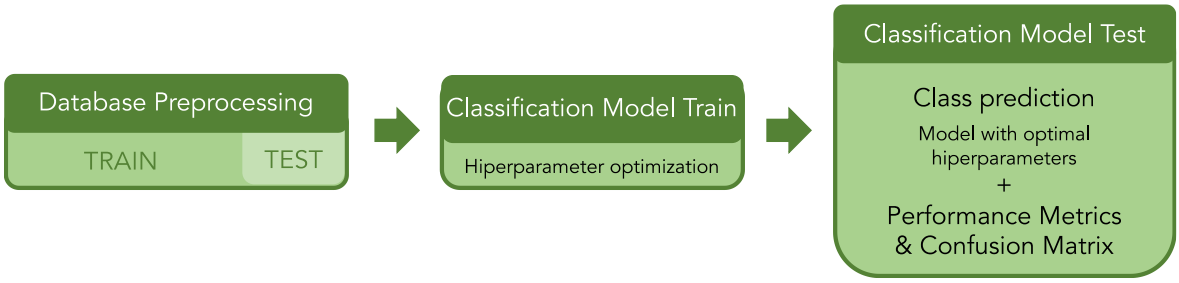


Figura I Representación del modelo de trabajo

9. Resultados

Tras llevar a cabo el proceso explicado en la sección 3, se obtienen unos resultados altamente satisfactorios, tanto por las métricas de rendimiento como por la proporción de pacientes correctamente clasificados.

Base de datos	Métrica	Gradient Boosting	KNN
Global Database	Accuracy	0,966	**0,972**
	F1 score	0,066	**0,972**
Global Database Noise Filter	Accuracy		**0,975**
	F1 score		**0,975**

Tabla I. Métricas de rendimiento para Global Database

Asimismo, podemos ver en la Figura II cómo se muestran los resultados de las matrices de confusión para KNN con la base de datos global sin filtrado de reducción de ruido:

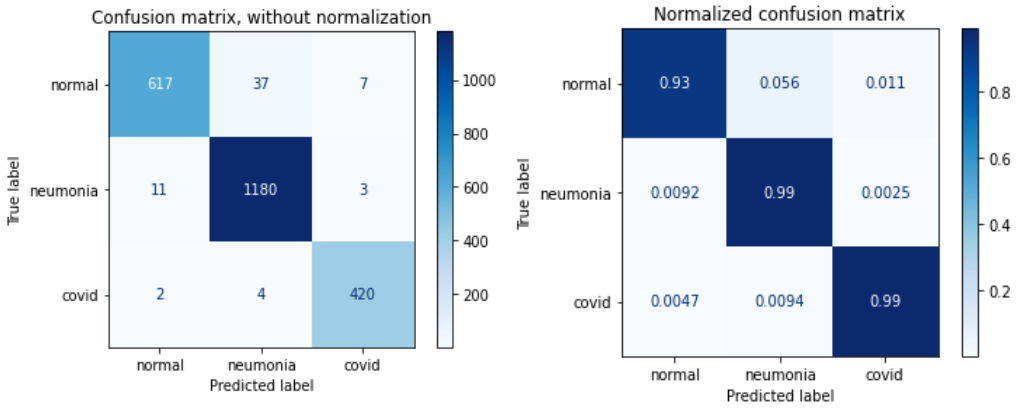


Figura II Matrices de confusión para Global Database con KNN sin reducción de ruido

Tras analizarlo en profundidad, **se propondrá el modelo KNN con las imágenes sin reducción de ruido como modelo definitivo**, ya que presenta unas métricas de rendimiento del 97,2% (Tabla I). Además, como se aprecia en la Figura II, este modelo logra una proporción de falsos negativos inferior al resto de las opciones, y los tiempos de cómputo son muy reducidos.

10. Conclusiones

El estudio desarrollado en el presente trabajo demuestra que el **Machine Learning supone una alternativa viable para la detección de COVID-19.** Los resultados obtenidos son tan competitivos que la sensibilidad obtenida es **comparable** con la de las distintas **pruebas de diagnóstico de COVID que existen actualmente en el mercado** [3], con la ventaja añadida de su reducido coste económico, de personal y de tiempo.

Las conclusiones del presente trabajo abren la puerta a futuros análisis más minuciosos. Por ejemplo, en este estudio no se ha tomado en consideración una diferenciación de los pacientes de COVID-19 según si presentan **síntomas de la enfermedad** o no. De igual forma, como trabajo futuro podría profundizarse en la detección de COVID-19 utilizando modelos de aprendizaje profundo (**Deep Learning**).

11. Referencias

[1] J. D. Arias-Londoño, J. A. Gómez-García, L. Moro-Velázquez and J. I. Godino-Llorente (2020), "Artificial Intelligence Applied to Chest X-Ray Images for the Automatic Detection of COVID-19. A Thoughtful Evaluation Approach", Institute of Electrical Electronics Engineers (IEEE). Recuperado en febrero 2021 de https://doi.org/10.1109/ACCESS.2020.3044858

[2] J. Bobadilla (Febrero 2020), "Machine Learning y Deep Learning usando Python, Scikit y Keras". RA-MA Editorial

[3] Quirón Salud (2021). Recuperado en junio 2021 de: https://www.quironsalud.es/es/informacion-pruebas-covid

MACHINE LEARNING FOR COVID-19 DIAGNOSIS

Author: Bausili Llamas, Elvira.

Supervisor: González Prieto, Ángel.

Collaborating Entity: ICAI – Universidad Pontificia Comillas

ABSTRACT

This paper analyses the use of artificial intelligence as an alternative for the detection of COVID-19 in radiographs. For this purpose, a comparative study will be carried out by applying Machine Learning models to three different radiography databases.

Considering the initial results obtained, we will proceed to obtain the model with optimal performance for the detection of this disease. To this end, we will develop a global model according to the established performance metrics, using a database that includes all the available radiographs. Finally, it will be concluded that the most suitable method for COVID-19 detection is K-Nearest Neighbors (KNN) applied to raw images, without any prior preprocessing.

Keywords: COVID-19, radiography, Machine Learning, image processing, Gradient Boosting, KNN.

1. Introduction

The health crisis caused by the appearance and spread of COVID-19 is a highly topical problem that affects us all in our daily lives. The intention of this Final Degree Project is to make use of engineering know-how to find an alternative solution to the current methods for detecting this disease. If this can be accomplished, savings in economic means and health personnel would be achieved, which could be redirected to the care of people with severe symptoms and to the development of the vaccination plan.

2. Project definition

The aim of this Final Degree Project is to obtain an automatic model for detecting patients affected by COVID-19 from X-rays using Machine Learning tools. Patients will be classified into people with COVID-19, with pneumonia or without any disease.

This work has a preliminary development phase in which **multisource data mining and image preprocessing** are carried out for subsequent use in the selected models. According to the literature, it will be assumed that lightweight processing will lead to better classification results. [1]

Next, Naïve Bayes, KNN, Random Forest and Gradient Boosting **classification models** will be programmed and then applied to the three databases separately. In order to provide greater realism and applicability to our work, we will try to **optimize the model** for a global database, in which the conclusions drawn from the previous comparative analysis will be applied. **Performance metrics**, the proportion of cases correctly diagnosed with COVID-19 and the absence of false negatives will also be considered to determine the characteristics of the most appropriate model for the final objective of this project.

3. Model description

Our project starts with the reading of the radiographs, assigning the class of each image with a numerical value. The **categories used are: 3 (COVID-19), 2 (pneumonia) and 1 (normal**).

Machine learning algorithms require a **training dataset (training) and a model checking dataset (test).** [2] Thus, in our work, we assign a random distribution of images for training and testing, consisting of 80%-20% of the radiographs in the databases that are not previously divided. In the first part of the process, the train X-rays are used to learn the model, which is then evaluated using the test images.

Typically, Machine Learning models are built on the basis of certain characteristics that must be determined beforehand, so-called **hyperparameters**. [2] In the algorithms that require these hyperparameters, we will make a **search for their optimal values** before executing the program, by pre-analysing the model using GridSearch.

Comparing the results obtained from the classification of the test images with the model and their real category, the corresponding **performance metrics** are obtained. Finally, we will develop the confusion matrices to obtain greater visibility in the results obtained, both in real and unitary magnitudes. This allows us to distinguish between correctly assigned cases, false negatives and false positives.

The diagram displayed in Figure I gives a graphical overview of the model used for this project in general terms.

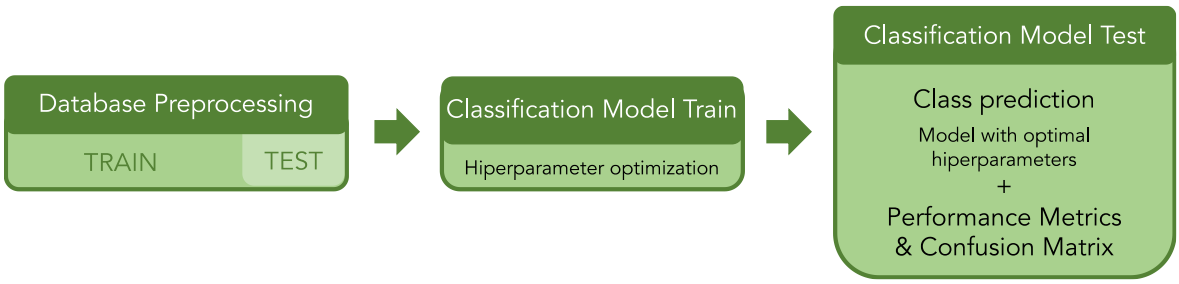


Image I Work model representation

4. Results

Highly satisfactory results are obtained after carrying out the process explained in section 3, both in terms of performance metrics and the proportion of patients correctly classified.

Database	Metric	Gradient Boosting	KNN
Global Database	Accuracy	0,966	**0,972**
	F1 score	0,066	**0,972**
Global Database Noise Filter	Accuracy		**0,975**
	F1 score		**0,975**

Chart I. Performance metrics for Global Database

Besides, Figure II shows the results of the confusion matrices for KNN with the global database without any filtering:

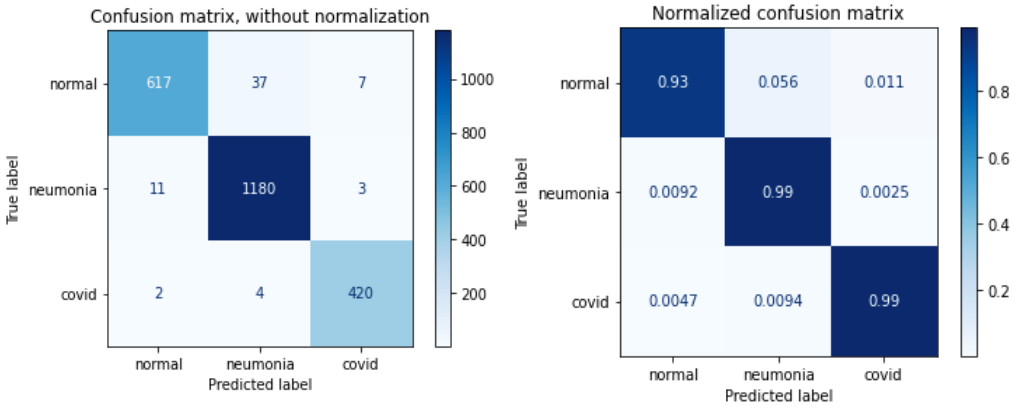


Image II Confusion matrices for Global Database with KNN without noise reduction

After a thorough analysis, **KNN model with the images without noise reduction will be proposed as the definitive model**. It presents performance metrics of 97.2% (Table I) with

very short computation times. Furthermore, as can be seen in Figure II, this model achieves a lower false negative rate than the other options.

5. Conclusions

The study carried out in this paper demonstrates that **Machine Learning is a viable alternative for COVID-19 detection.** The results obtained are so competitive that the efficacy obtained is **comparable** to the reliability of the different COVID diagnostic tests **currently available on the market** [3]. This method also has the added advantage of the reduced use of resources in terms of economic, personnel and time costs.

The findings of the present work open the door to more detailed analyses in the future. For example, this study has not taken into consideration a differentiation of COVID-19 patients according to whether they have **symptoms of the disease or not**. Likewise, future work could further investigate the detection of COVID-19 using **Deep Learning** models.

6. References

[1] J. D. Arias-Londoño, J. A. Gómez-García, L. Moro-Velázquez and J. I. Godino-Llorente (2020), "Artificial Intelligence Applied to Chest X-Ray Images for the Automatic Detection of COVID-19. A Thoughtful Evaluation Approach", Institute of Electrical Electronics Engineers (IEEE). Retreived in February 2021 from: https://doi.org/10.1109/ACCESS.2020.3044858

[2] J. Bobadilla (February 2020), "Machine Learning y Deep Learning usando Python, Scikit y Keras". RA-MA Editorial

[3] Quirón Salud (2021). Retreived in June 2021 from: https://www.quironsalud.es/es/informacion-pruebas-covid

Índice de la memoria

Índice de figuras

Índice de tablas

Capítulo 1. INTRODUCCIÓN

La pandemia de COVID-19 a la que nos enfrentamos hoy en día es una de las crisis de mayor magnitud a nivel mundial de los últimos años. Esta situación ha provocado consecuencias sanitarias, sociales y económicas que se extenderán en el tiempo incluso una vez superemos el estado crítico actual.

COVID-19 hace referencia a la enfermedad infecciosa asociada al coronavirus SARS-COV-2 (Coronavirus Infectious Disease) iniciada a finales del año 2019. [1] Se considera un trastorno respiratorio con sintomatología muy diversa, en la que se incluye fiebre, tos, fatiga, dificultad para respirar y dolor de cabeza, entre otros. [2] Actualmente, el número de personas de las que se tiene constancia de haber sido contagiadas asciende a 179 millones en el mundo, y 3,76 millones en España. De estos, los fallecidos ascienden a 3,87 millones a nivel mundial y 80690 en España, según las cifras actualizadas proporcionadas por Johns Hopkins University [3] a 22 de junio de 2021. De esta forma, podemos establecer una ratio de mortalidad de un 2,16% a nivel mundial y de un 2,14% en nuestro país.

Aunque a priori puedan parecer cifras de bajo impacto, la realidad es que esta enfermedad ha cambiado completamente nuestro estilo de vida. Los dirigentes de todas las comunidades se ven obligados a establecer medidas y restricciones legales que oscilan semanalmente en función de la evolución de las cifras de contagio e intensidad de los afectados por el virus. Estas disposiciones tienen el objetivo de evitar en lo posible el colapso sanitario, pero procurando preservar en lo posible la estabilidad económica.

Asimismo, cabe destacar otro aspecto de vital importancia en el proceso global de eliminación de este virus: la vacuna. Durante los primeros meses de la pandemia, instituciones científicas e industrias farmacéuticas de todo el mundo han dedicado todos sus esfuerzos en la investigación de la vacuna que logre erradicar esta enfermedad. Igual que en el siglo XX tuvo lugar la carrera espacial por la conquista del espacio, en 2020 los países enfocaron sus recursos en la obtención de la mejor manera de dar fin a esta crisis sanitaria.

En España, como en la mayoría de los países de Occidente, se cuenta con las propuestas de Estados Unidos y Reino Unido. Por parte del país estadounidense, las primeras vacunas autorizadas y recomendadas fueron las de Pfizer-BioNTech y la de Moderna, además de dos vacunas aprobadas posteriormente (Janssen y Novavax). Por su parte, Reino Unido también ha aportado su propia vacuna AstraZeneca en colaboración con la Universidad de Oxford. Asimismo, cabe destacar que países como Rusia y China han llevado a cabo sus propios procesos de obtención de la vacuna, como Gamaleya en Rusia y CanSino en el gigante asiático. [4]

La crisis sanitaria provocada por la aparición y propagación de este virus es un asunto de gran actualidad que nos afecta a todos en nuestro día a día. La intención de este Trabajo de Fin de Grado es hacer uso de los conocimientos técnicos de ingeniería adquiridos durante toda la carrera para tratar de buscar una solución alternativa, viable y útil a la detección de COVID-19; con el propósito de agilizar el proceso y reducir el coste de recursos sanitarios en este ámbito. De esta forma, ese ahorro en medios económicos y de personal sanitario podría reorientarse al cuidado de las personas con sintomatología grave y a llevar a cabo el plan de vacunación de manera eficiente.

No obstante, hasta que la vacunación se encuentre ampliamente extendida, los métodos de contención de la enfermedad requieren de un diagnóstico rápido y preciso que permita aislar a los enfermos para evitar la propagación de la enfermedad. Actualmente, este diagnóstico es manual, basado en test PCR, de antígenos y de detección de anticuerpos que deben ser interpretados por un experto médico. [5] Esto supone un coste en términos humanos que impide la destinación de dichos recursos al cuidado de los pacientes que sufren los síntomas más graves de esta enfermedad, así como a la implantación del plan de vacunación de manera eficiente. Por este motivo, resulta crucial proponer nuevos métodos de diagnóstico que liberen a los profesionales sanitarios. En este objetivo reside la justificación de este trabajo.

Capítulo 2. DESCRIPCIÓN DE LAS TECNOLOGÍAS

Para la realización de este estudio, se hará uso del lenguaje de programación Python [6], así como de las utilidades relativas al aprendizaje automático contenidas en la librería sklearn. [7] Esta librería nos proporcionará los modelos necesarios y las herramientas para la obtención de las métricas de medida de la eficiencia cada algoritmo. Asimismo, serán de gran utilidad las librerías de matplotlib [8] para la representación de gráficos que ilustren las conclusiones obtenidas tras la ejecución de cada modelo, así como las librerías skimage [9] y PIL [10] para el procesamiento de imágenes.

Con el fin de reducir los tiempos de cómputo, para la ejecución de los programas se ha hecho uso de los servidores externos Galactica y Columbia, proporcionados por la Universidad Politécnica de Madrid. Galactica nos ha permitido trabajar con CPU (Central Processing Unit) para la ejecución de los modelos de Machine Learning, mientras que Columbia está optimizado para GPU (Graphics Processing Unit), especialmente útil para operaciones matriciales. Esta última tarjeta gráfica tiene un rendimiento superior a una CPU, ya que es específica para gráficos y datos de entrada más complejos, como en nuestro caso, imágenes de radiografías.

Capítulo 3. ESTADO DE LA CUESTIÓN

Dentro del amplio abanico de posibilidades para la detección de COVID-19, podemos encontrar cuatro tipos de pruebas distintas, y presentamos a continuación algunas de sus características:

- Test PCR (Reacción en Cadena de la Polimerasa), es la prueba más sensible para la medición de infección activa mediante la comprobación de existencia de material genético del virus en una muestra de mucosa. La especificidad y sensibilidad analíticas de dicha prueba diagnóstico es de prácticamente un 100%, entendiendo sensibilidad como el porcentaje de verdaderos positivos detectados, y especificidad como la proporción de verdaderos negativos. [11]

- Test de antígenos, que detectan la presencia de una proteína de la superficie el virus también a partir de una muestra de secreción del tracto respiratorio, indicado la presencia actual de la infección. La gran ventaja es el tiempo de respuesta en el resultado (minutos) y la elevada sensibilidad en los primeros días de la infección. La sensibilidad en los primeros días de contagio (0-3 días) es de un 100% y del 4-7 día del 90%. [5]

- Test serológico o de anticuerpos, que analiza la presencia en sangre de inmunoglobulina M (IgM), segregada de los 7 a 10 primeros días tras contraer la infección; y la inmunoglobulina G (IgG), que aparece cuando ha cesado la fase aguda de la infección y puede permanecer a largo plazo en muchos casos. De esta forma, una detección de la IgM indica que la persona puede estar pasando por la enfermedad actualmente; mientras que si posee ambas, IgM e IgG, el paciente puede haber pasado la infección de manera reciente y permanecen restos de la fase aguda. La detección de IgG exclusivamente indicaría que la persona ha superado definitivamente la enfermedad. [12] La sensibilidad de este test serológico rápido es del 88,6% y la especificidad del 90,63%. [13]

- Test rápido. Detecta la presencia en sangre de anticuerpos de forma cualitativa, por lo que presentan una fiabilidad inferior, pero son los únicos que no requieren su realización en un centro médico especializado. Como ejemplo representativo, el nuevo test de autodiagnóstico del laboratorio Cinfa presentaría cuenta con una sensibilidad del 96,77% y una especificidad del 99,20%. [14]

No obstante, paralelamente a la detección de COVID-19 mediante pruebas diagnósticas, recientemente se han llevado a cabo estudios para lograr determinar la presencia del virus en pacientes mediante métodos más innovadores. Es el caso de la detección del virus en radiografías con herramientas tecnológicas de inteligencia artificial, como se pretende llevar a cabo en el presente trabajo.

Asimismo, se realizado un estudio en este mismo ámbito en el trabajo "Artificial Intelligence Applied to Chest X-Ray Images for the Automatic Detection of COVID-19. A Thoughtful Evaluation Approach". [15] En él, se hace uso de 8 fuentes distintas para la recopilación de bases de datos de radiografías, y tras ciertos análisis y procesamientos de dichas imágenes, logran obtener una medida de F1 score de 93% en la detección de COVID, en el experimento de mayor éxito entre los realizados. Las métricas y procesos utilizados en este trabajo nos servirán como comparativa para la evaluación de los resultados en este Trabajo de Fin de Grado.

Observamos que en esta publicación [15] se desarrollan tres experimentos comparativos con distintas estrategias de preprocesamiento de datos, de las que extraemos conclusiones especialmente relevantes para nuestro trabajo. En el Experiment 1 el modelo utiliza los datos en bruto, el Experiment 2 procesa la imagen recortada con una máscara, y el Experiment 3 utiliza las imágenes con una segmentación del pulmón, también aplicando el uso de máscaras. Los resultados de las métricas de eficiencia de este trabajo concluyen que el mejor experimento es el Experiment 1, tanto a nivel de accuracy (91,67 ± 2,56) como de F1 score (93,35 ± 0,68 en neumonía, 97,24 ± 0,23 en normal y 93,00 ± 1,00 en COVID). Teniendo esto en cuenta, nuestro preprocesamiento inicial procurará mantener las imágenes en su estado original tanto como sea posible.

Otro trabajo destacable que hace uso de programas de inteligencia artifical aplicados en la temática de COVID-19 es el citado [16], que procura identificar pacientes de riesgo, su tasa de mortalidad y otras anormalidades hallando patrones y realizando predicciones en base a las mismas. De la misma forma, este otro estudio [17] realiza una investigación en el uso de programas de Machine Learning y Deep Learning para desarrollar modelos matemáticos que permitan analizar la situación de la pandemia actual, con el fin de comprender su comportamiento exponencial junto con la predicción futura de su expansión en el mundo. Otro caso interesante [18] es el análisis de las posibles de combinaciones de proteínas para el desarrollo de la vacuna más eficaz contra esta enfermedad mediante algoritmos de aprendizaje automático.

Cabe señalar que no es la primera vez que se utilizan herramientas de inteligencia artificial para la detección de enfermedades. Ya en 2019, en el artículo [19], en el que se llevaba a cabo una comparativa en la eficacia de métodos tecnológicos y humanos para determinar la presencia de cáncer cerebral. Asimismo, se han utilizado este tipo de tecnologías en otros ámbitos de carácter social, como es el caso de la predicción de reincidencia en casos de violencia de género. [20]

Capítulo 4. DEFINICIÓN DEL TRABAJO

4.1 JUSTIFICACIÓN

A pesar del gran avance que ha supuesto la salida al mercado de las distintas propuestas de vacunación, la situación sigue en estado crítico, principalmente por dos motivos. Por una parte, el ritmo de producción de las vacunas ya aprobadas no logra satisfacer la incipiente demanda de la población, por lo que el proceso de vacunación se ralentiza; teniendo en cuenta también las posibles ineficiencias en la gestión y logística de los planes de vacunación de los distintos países.

Por otro lado, en los últimos meses se han descubierto casos de infección de COVID-19 con ciertas variaciones, mutaciones en el código genético del virus que provocan alteraciones en la intensidad y sintomatología que podrían suponer una dificultad añadida para el trato con las vacunas existentes. De esta forma, se han detectado cinco cepas distintas, las llamadas cepa británica, sudafricana, brasileña, nigeriana y californiana [21]; que se engloban distintas mutaciones caracterizadas por presentar alteraciones en las proteínas del virus y por ser más contagiosas que la cepa original.

El objetivo del presente Trabajo de Fin de Grado es ofrecer una nueva alternativa para la detección de COVID-19 con herramientas de Inteligencia Artificial a partir de radiografías, tratando de incorporar las conclusiones y recomendaciones de otros estudios de investigación relacionados. De este modo, se pretende ofrecer una alternativa viable y eficiente que permita la dedicación de recursos económicos y de personal sanitario al trato de los pacientes con sintomatología más grave y al desarrollo del plan de vacunación.

4.2 OBJETIVOS

Como se explicaba con anterioridad, la finalidad específcia del presente Proyecto de Fin de Grado es la detección de pacientes afectados por COVID-19 a partir de radiografías haciendo uso de herramientas de Machine Learning. En conjunto, se llevará a cabo un estudio

comparativo entre distintos modelos de clasificación para determinar el diagnóstico con la mayor exactitud posible, así como el uso de diversas bases de datos para perfeccionar las fases de entrenamiento y test de los modelos implicados.

Con el propósito de maximizar la exactitud de los modelos definidos, se establecen ciertos objetivos más específicos detallados a continuación:

- **Objetivo 1: Minado de datos multifuente de radiografías**

Se hará uso de tres bases de datos disponibles en Kaggle [22], la comunidad de científicos de datos especializados en aprendizaje automático, con el propósito de combinar y comparar el uso todas las radiografías disponibles para obtener el modelo de Machine Learning más eficiente en la detección de COVID-19.

- **Objetivo 2: Aplicación de técnicas de preprocesamiento de imágenes**

Se partirá de un preprocesamiento inicial que incluye la lectura de imágenes en escala de grises y su redimensionamiento para facilitar la acción de los programas que se utilizarán posteriormente. Esta fase del trabajo se irá perfeccionando de manera iterativa a medida que se obtengan los resultados de la detección en los distintos modelos, con la posibilidad de aplicar filtros, máscaras u otras herramientas que faciliten el entrenamiento y test con los distintos modelos de aprendizaje automático.

- **Objetivo 3: Determinar los modelos de clasificación de Machine Learning más apropiados para lograr la detección de COVID-19**

Una vez se obtenga un conjunto de datos limpio que permita el entrenamiento de modelos de Machine Learning, se utilizarán diversas métricas de calidad estándar para la evaluación de su rendimiento. Para lograr este objetivo, se determinarán los parámetros más adecuados para la medida de exactitud y eficiencia de los modelos. Inicialmente, el estudio se enfocará como un problema de clasificación, por lo que se analizarán los parámetros de precisión, exactitud (accuracy), exhaustividad (recall), valor F (F score) y la matriz de confusión. [23]

La exactitud (accuracy) mide el porcentaje de aciertos del modelo respecto al total de casos. Para ello, también se puede comparar la exactitud al porcentaje de aciertos asumiendo que todos son positivos o que todos son negativos.

En nuestro trabajo, se cuentan con tres clases diferentes: pacientes con COVID-19 (categoría covid), con neumonía (categoría neumonía) y personas sin enfermedad (categoría normal). No obstante, y para mayor simplicidad, se procederá a explicar el cálculo del accuracy contando con dos únicas clases: positivo y negativo en COVID-19, englobando en la categoría negativa los pacientes sanos y con neumonía. De esta forma, podemos hallar esta métrica de la siguiente forma:

$$Accuracy = \frac{VP\_COVID \ + \ VN\_COVID}{VP\_COVID + VN\_COVID + FP\_COVID + FN\_COVID}$$

Donde VP_COVID denota a los Verdaderos Positivos en COVID-19, VN_COVID a los Verdaderos Negativos, FP_COVID a los Falsos Positivos y FN_COVID a los Falsos Negativos.

Obsérvese que la métrica de la exactitud únicamente no resulta suficiente para determinar el rendimiento de un modelo, ya que no realiza el cálculo considerando el peso de cada categoría a clasificar. Es una métrica de gran utilidad pero que no puede considerarse de manera exclusiva, ya que puede inducir a error cuando las clases están desproporcionadas.

La precisión se utiliza como medida de calidad del modelo, y tiene en cuenta el número de verdaderos positivos detectados sobre el total de positivos detectados. En nuestro caso:

$$Precision = \frac{VerdaderoPositivo\_COVID}{VerdaderoPositivo\_COVID + FalsoPositivo\_COVID}$$

La métrica de recall informa sobre la cantidad que el modelo en cuestión es capaz de identificar correctamente, y se puede definir como:

$$Recall = \frac{VerdaderoPositivo\_COVID}{VerdaderoPositivo\_COVID + FalsoNegativo\_COVID}$$

El parámetro de valor de F supone una combinación entre las dos medidas anteriores, ya que se halla mediante la media armónica entre la precisión y exhaustividad. En general, este parámetro se calcula de la siguiente forma:

$$F_\beta = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall}$$

Una elección típica en Machine Learning es tomar $\beta = 1$ obteniendo el denominado F1 score, que considera que se concede el mismo peso en importancia a la precisión y a la exhaustividad.

Por otra parte, las matrices de confusión son una representación en dos dimensiones de la proporción de casos correcta e incorrectamente clasificados de forma gráfica. En el eje horizontal se representarán los valores reales, y en el vertical los resultados predichos o clasificados con cada modelo. Para el caso simplificado de dos categorías, una matriz de confusión tomaría la siguiente forma:

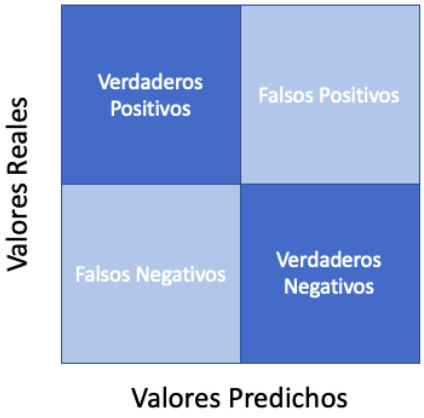


Figura 1. Ejemplo de matriz de confusión para dos clases

Esta representación gráfica nos permite calcular de forma sencilla la sensibilidad del modelo (porcentaje de pacientes correctamente clasificados como positivos en COVID-19) y la especificidad (proporción de radiografías correctamente clasificadas como negativas en COVID-19). [24]

En la evaluación de nuestros modelos nos centraremos en el accuracy y F1 score (ya que engloba precision y recall), para decidir el mejor enfoque en la detección de la enfermedad.

Asimismo, observaremos la especificidad y sensibilidad de cada modelo mediante las matrices de confusión.

- **Objetivo 4: Comparar los resultados obtenidos para proponer un modelo óptimo**

Tras la obtención definitiva de los modelos de clasificación óptimos tras el proceso iterativo de perfeccionamiento de la fase de preprocesamiento, se podrá estudiar el empleo de otras alternativas de modelos como plan de contingencia si los parámetros de eficiencia no presentan resultados suficientemente satisfactorios.

Tras el análisis comparativo de los distintos modelos en cada base de datos por separado, se ha procedido a la evaluación de los modelos con mejores rendimientos para la base de datos conjunta que agrupa a la totalidad de las radiografías. Inicialmente, en esta sección se consideraba la elaboración de modelos de Deep Learning o modelos de regresión haciendo uso de una gradación de las distintas categorías de las radiografías según su gravedad. Sin embargo, los resultados obtenidos han sido de gran calidad, por lo que la optimización se ha centrado finalmente en el análisis de la aplicación de un filtrado de reducción de ruido exclusivamente, dejando como trabajos futuros la implementación de modelos de predicción y el uso de Deep Learning.

4.3 METODOLOGÍA

Una vez definido el problema a resolver y los objetivos concretos en los que se centrará este estudio, procederemos a distribuir las tareas correspondientes en el intervalo de tiempo asignado a este Trabajo de Fin de Grado mediante el diagrama de Gantt expuesto a continuación:

Tarea asignada	Fecha de inicio	Fecha de finalización	Estado	01.01.2021	15.01.2021	01.02.2021	15.02.2021	01.03.2021	15.03.2021	01.04.2021	15.04.2021	01.05.2021	15.05.2021	01.06.2021	15.06.2021
TRABAJO FIN DE GRADO	01.01.2020	30.06.2020	Abierto												
Objetivo 1	01.01.2021	01.03.2021	En progreso												
Tarea 1.1 Minado de datos	01.01.2021	01.02.2021	Terminado												
Tarea 1.2 Elaboración de la memoria correspondiente	01.02.2021	01.03.2021	En progreso												
Objetivo 2	15.02.2021	15.05.2021	En progreso												
Tarea 2.1 Preprocesamiento inicial	15.02.2021	15.03.2021	En progreso												
Tarea 2.2 Perfeccionamiento del preprocesamiento	15.03.2021	15.05.2021	Abierto												
Tarea 2.3 Elaboración de la memoria correspondiente	15.02.2021	15.05.2021	Abierto												
Objetivo 3	15.03.2021	01.05.2021	Abierto												
Tarea 3.1 Modelos ML y evaluación de métricas	15.03.2021	01.05.2021	Abierto												
Tarea 3.2 Elaboración de la memoria correspondiente	15.04.2021	01.05.2021	Abierto												
Objetivo 4	01.05.2021	01.06.2021	Abierto												
Tarea 4.1 Optimización del proyecto	01.05.2021	01.06.2021	Abierto												
Tarea 4.2 Elaboración de la memoria correspondiente	01.05.2021	01.06.2021	Abierto												
Adopción de feedback y últimas correcciones	01.06.2021	15.06.2021	Abierto												
Presentación del TFG	15.06.2021	30.06.2021	Abierto												

Tabla 1 – Diagrama de Gantt

La distribución de trabajo a lo largo del cuatrimestre ha sido bastante equitativa, siguiendo con lo indicado en la planificación representada en la Tabla 1. Se han observado ciertos solapamientos en la ejecución de los programas pertenecientes a los objetivos 3 y 4, mientras que la redacción de la memoria en profundidad se ha llevado a cabo mayoritariamente en la fase final de este proyecto.

Capítulo 5. DESARROLLO DEL PROYECTO

5.1 MINADO DE DATOS

Para la elaboración de este proyecto, se ha llevado a cabo una búsqueda en distintas fuentes de información para la obtención de una base de datos de radiografías más adecuada para su posterior clasificación con modelos de Machine Learning. Finalmente, se ha hecho uso de ciertas bases de datos de la web Kaggle, la comunidad de libre acceso de programadores especializados en tecnologías de inteligencia artificial.

Como se indica en la sección 4.2, a lo largo de este estudio se trabajará con tres bases de datos distintas de forma independiente y posteriormente de forma combinada. Con ello, se pretenden comprobar las posibles variaciones que experimenten las métricas de eficiencia establecidas para cada conjunto de datos, con el fin de lograr el mejor entrenamiento posible para el algoritmo.

El primer conjunto de radiografías [25], denominado Database1 en este estudio, consta de 6432 imágenes, clasificadas en carpetas de entrenamiento y test para cada una de las tres categorías: pacientes sin enfermedad, con neumonía y con COVID-19. Sus características se especifican de manera más detallada en la Tabla 2.

La segunda base de datos Database2 [26] consta de 3890 imágenes divididas en radiografías de pacientes sin enfermedad, con neumonía y con COVID-19. En este caso, no existe una división preestablecida de los archivos de train y test, por lo que se determinará en su procesamiento, asignando una proporción de 80%-20% para cada fin. Asimismo, esta base de datos presenta un certificado de calidad al ser el ganador del COVID-19 Dataset Award by Kaggle Community.

Por último, el Database3 [27] incluye 1088 imágenes diferentes divididas en entrenamiento y test para cada uno de los supuestos de enfermedad. Lo interesante de esta base de datos es que las imágenes se presentan en su archivo original, con una reducción de ruido y con una superposición de los anteriores. De esta forma, el conjunto de radiografías Database 3

asciende a 3264 imágenes, y permitirá hacer la comparativa entre los resultados obtenidos con las imágenes en sus distintos estados para evaluar en qué medida este preprocesamiento afecta al modelo.

Un resumen de las principales características de cada base de datos puede consultarse en en la Tabla 2:

	Nombre	Fuente	Tamaño	Distribución previa
Database 1	Chest X-ray (Covid-19 & Pneumonia)	Kaggle	2,64 GB 6432 imágenes	División train-test 80%-20%
Database 2	COVID-19 Radiography Database	Kaggle	739,97 MB 3886 imágenes	No
Database 3	COVID-19 X-Ray Dataset With Preprocessed Images	Kaggle	2,22 GB 2364 imágenes	Unpreprocessed, Preprocessing Fuzzy Color Technique y Stacking y División train-test

	Clasificación	Dimensión imágenes	Formato
Database 1	COVID, neumonía, normal	Variado Desde 1000x1000 a 2000x2000 aprox	jpg
Database 2	COVID, neumonía, normal	299 x 299	png
Database 3	COVID, neumonía, normal	Variado Desde 1000x1000 a 2000x2000 aprox	png, jpg jpeg

Tabla 2 – Bases de datos

5.2 PREPROCESAMIENTO DE DATOS

La fase de preprocesamiento es esencial para lograr mejores resultados tras la aplicación de modelos de aprendizaje automático, ya que los datos se presentan homogeneizados y enfocados de manera apropiada para facilitar la tarea de clasificación del programa que se utilizará posteriormente. Asimismo, permite reducir los tiempos de computación, lo cual es un aspecto de gran relevancia cuando se trabaja con archivos con de ocupación de memoria elevada, como imágenes, en este caso.

Esta parte del trabajo se presenta como una fase iterativa. Inicialmente, se trabajará con una base de datos reducida para establecer el preprocesamiento inicial de las imágenes de cada

Database mencionado anteriormente. Posteriormente, se procederá a la optimización de los modelos de clasificación y a la evaluación de las métricas de rendimiento. Finalmente, se procederá a adoptar posibles mejoras en el preprocesamiento que se consideren tras evaluar los resultados obtenidos de la clasificación con el preprocesamiento inicial de las imágenes (fase de perfección de preprocesamiento).

5.2.1 PREPROCESAMIENTO INICIAL

Recuérdense las conclusiones extraídas del artículo [15] mencionadas en el Capítulo 3 de este trabajo, que evidenciaban que un procesamiento ligero lleva asociados mejores resultados en clasificación. Teniendo esto en cuenta, nuestro preprocesamiento inicial procurará mantener las imágenes en su estado original tanto como sea posible.

En la primera fase del preprocesado, se leen los archivos contenidos en cada Database y se contienen en una lista. A continuación, se debe establecer la etiqueta de la imagen para cada categoría. En vistas a la posible elaboración de un plan de contingencia basado en modelos de regresión, se establecerán las distintas categorías de las radiografías haciendo uso de una gradación de según la gravedad de la enfermedad para el paciente. De esta forma, se los pacientes sin enfermedad se etiquetarán con el nivel de gravedad más bajo (normal, 1), los que padecen de neumonía con el nivel intermedio (neumonía, 2) y los que sufren de COVID-19 con el valor superior en la escala (COVID-19, 3).

Finalmente, se procede a la reducción y homogenización del tamaño y color de las imágenes, lo que facilitará el uso de los modelos posteriormente y reducirá los tiempos de cómputo. Se establecen unas medidas iniciales de 150x150 píxels, que pueden ser alterados en la fase de perfeccionamiento si se considera necesario. Se convierten las imágenes a escala de grises, y se almacenan los datos de lectura de los píxels de cada imagen (ya en blanco y negro y con el tamaño adecuado) en un fichero csv.

Durante la ejecución del preprocesamiento se ha observado que el software proporcionaba ciertos errores debido a la existencia de ciertos archivos vacíos en la carpeta de Database1 (2 imágenes en la subcarpeta de test y 1 imagen en la de train), por lo que se ha decidido suprimirlos.

A continuación, se incluye una muestra de las radiografías de una de las bases de datos antes (Figura 2) y después (Figura 3) de su preprocesamiento inicial. La lectura de las imágenes en el fichero csv que se utilizará posteriormente para el entrenamiento y testeo de los modelos se llevará a cabo con las imágenes en escala de grises y de tamaño 150x150.

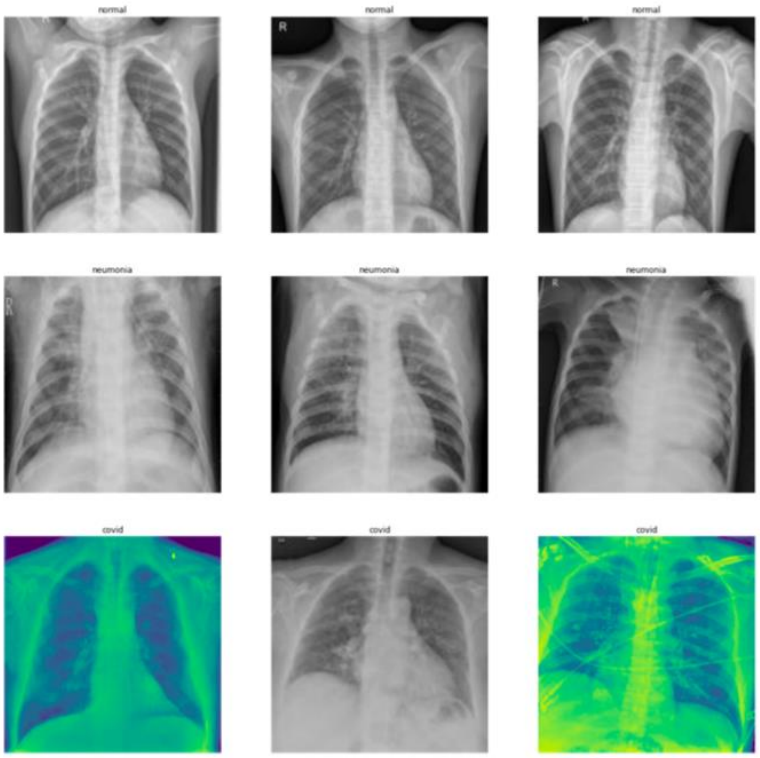


Figura 2. Muestra de radiografías antes del preprocesamiento inicial

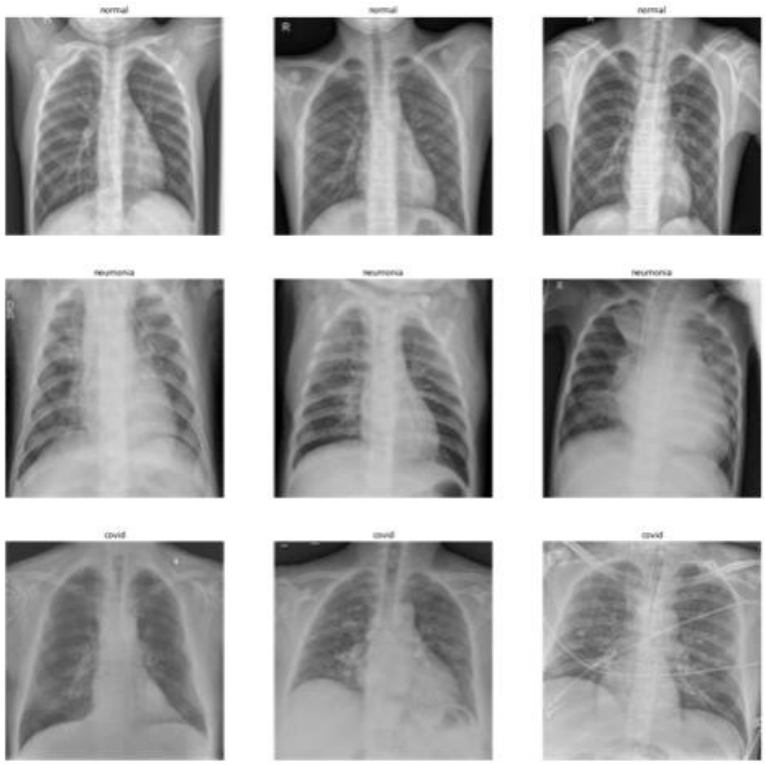


Figura 3. Muestra de radiografías después del preprocesamiento inicial

Como producto del procesamiento obsérvese que las imágenes son ahora en escala de grises, y se ha llevado a cabo una reducción en su tamaño aunque no sea posible su apreciación en las figuras mostradas.

5.3 MODELOS DE CLASIFICACIÓN

Una vez se ha completado el pre-procesado de datos, se procederá al desarrollo de los modelos de clasificación que permitirán discernir entre los tres tipos de radiografías posibles: normal, con neumonía o con COVID-19. La implementación de estos programas se realizará utilizando las librerías de sklearn [7], como se mencionaba en el Capítulo 2 de este trabajo (Descripción de las Tecnologías). Se llevará a cabo la optimización de los hiperparámetros de todos los modelos de clasificación utilizados, haciendo uso de la herramienta de GridSearch propia de esta librería.

GridSearchCV es una herramienta que lleva a cabo la ejecución de un modelo de aprendizaje automático haciendo uso de todas las combinaciones posibles de los hiperparámetros que se especifiquen en un rango o grid. Este algoritmo se combina con validación cruzada (CV) para asegurar mayor solidez en los resultados.

Tras ejecutar este proceso, se proporciona una puntuación de calidad a cada una de las combinaciones (scoring) basada en la métrica de eficiencia del modelo al que se aplica. En nuestro proyecto, los modelos elegidos son Naïve Bayes (no requiere hiperparámetros), KNN, Random Forest y Gradient Boosting; y todos ellos hacen uso de la métrica mean accuracy por defecto, por lo que la búsqueda de los mejores hiperparámetros se basará en esta.

A continuación, se explicará brevemente la lógica teórica subyacente en cada uno de los modelos, centrándonos en el procedimiento seguidos en cada uno y en las métricas de calidad de los resultados obtenidos en la detección de COVID-19.

5.3.1 NAÏVE BAYES

El modelo de clasificación de Naïve Bayes parte de la premisa de que todas las variables que influyen en la clasificación en una u otra categoría son independientes entre sí. Se basa en el teorema de Bayes, que proporciona la probabilidad de que suceda un evento teniendo en cuenta la probabilidad de que tengan lugar otros previamente. Es un modelo muy útil como primera aproximación, ya que resulta muy sencillo de implementar. Sin embargo, la presunción de independencia entre las variables no es un fiel reflejo de la realidad, como observaremos en la comparativa con otros modelos de clasificación de mayor complejidad.

La fórmula en la que se basa este modelo es la siguiente [28]:

$$P(C \mid A) = \frac{P(A \mid C) \cdot P(C)}{P(A)}$$

Siendo P(C | A) la probabilidad a posteriori de que se trate de una clase C dado un atributo A, P(A | C) la probabilidad a posteriori de que se trate de un atributo A sabiendo que pertenece a una clase C, P(A) la probabilidad a priori de que tenga cierto atributo A y P(C)

la probabilidad a priori de que pertenezca a una clase C. Así, la probabilidad resultante para n atributos distintos y una categoría C se hallaría como:

$$P(C \mid A) = P\left(A_1/_C\right) \cdot P\left(A_2/_C\right) \cdot \ldots \cdot P\left(A_n/_C\right) \cdot P(C)$$

A pesar de la posible falta de realismo que conlleva este modelo, los resultados que se han obtenido con este algoritmo resultan considerablemente satisfactorios, y se muestran en la Tabla 3:

		Accuracy	Recall	F1 score
Database 1		0,8	0,8	0,808
Database 2		0,861	0,861	0,86
Database 3	**Raw**	0,812	0,812	0,813
	Noise	0,784	0,784	0,785
	Stacked	0,794	0,794	0,795

Tabla 3. Métricas de rendimiento para Naïve Bayes

Asimismo, se han elaborado las matrices de confusión para las distintas bases de datos: Database 1 (Figura 4), en el Database 2 (Figura 5), en el Database 3 con imágenes en crudo (Db_raw, Figura 6), en el Database 3 con imágenes filtradas para la reducción de ruido (Db_noise, Figura 7) y en el Database 3 con superposición de las imágenes originales y con filtro (Db_stacked, Figura 8)

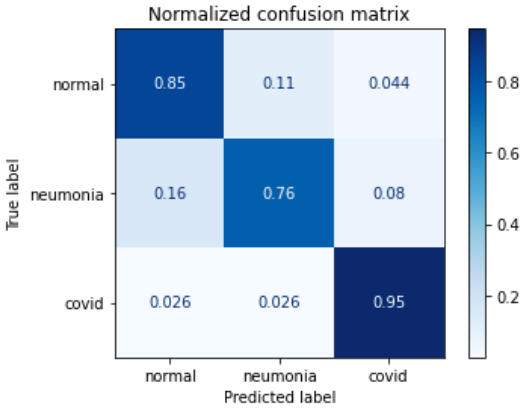


Figura 4. Matriz de confusión NB Database1

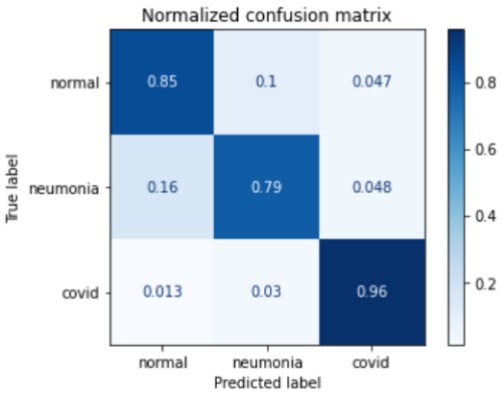


Figura 5. Matriz de confusión NB Database2

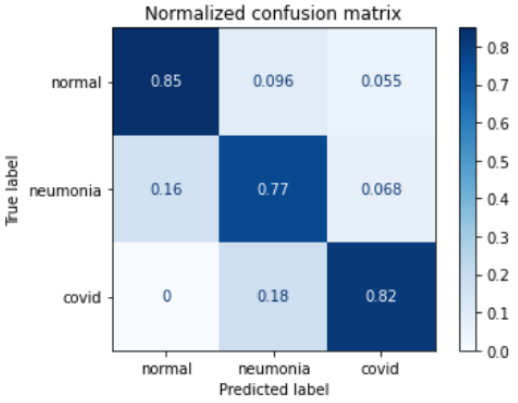


Figura 6. Matriz de confusión NB Db3_raw

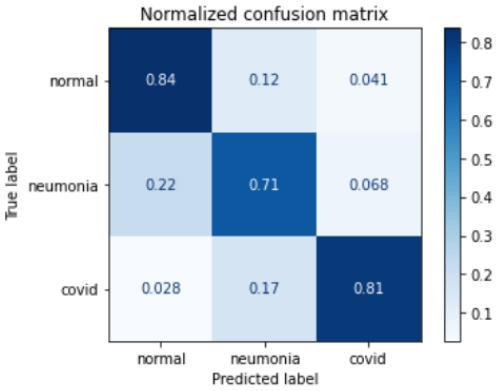


Figura 7. Matriz de confusión NB Db3_noise

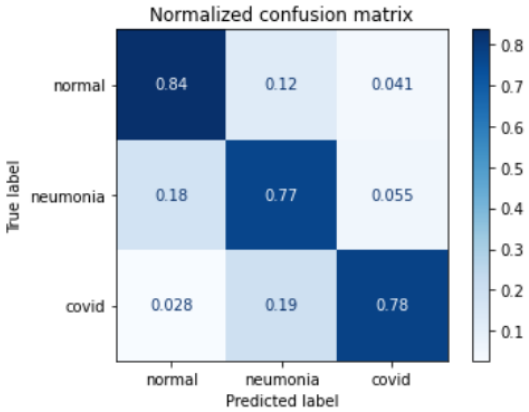


Figura 8. Matriz de confusión NB Db3_stacked

A pesar de la sencillez de este modelo, observamos una **clasificación bastante acertada**, cercana al 100% en la predicción de COVID-19, que es la que interesa en especial a este estudio. El caso de predicción errónea de mayor importancia es el caso del 22% de los pacientes de neumonía, clasificados como sin enfermedad en la base de datos Database 3 noise**.** Un aspecto muy positivo de Naïve Bayes es en el caso del **Database 3 raw,** en el que no hay **ningún paciente de COVID-19 al que se le diagnostique como sin enfermedad.**

Las métricas de eficiencia avalan estos resultados, observando una **mejor calidad** en la detección de COVID mediante el uso del **Database 2** (tanto por accuracy como F1 score), y un **peor rendimiento en la base de datos con preprocesamiento para la reducción de ruido** (Database 3 noise), lo cual va **en línea con la bibliografía** mencionada en la sección

de 5.2.1, respecto a que el preprocesado de imágenes es más eficaz cuanto más se mantengan las fotografías originales.

5.3.2 K NEAREST NEIGHBORS (KNN)

Este modelo busca las K muestras más cercanas al punto que queremos clasificar, y se realiza una predicción basada en la categoría mayoritaria entre sus etiquetas. KNN tiene en cuenta la distancia entre el punto y el dato cercano para ponderar la predicción de la categoría en base a la de ese punto próximo. Este algoritmo puede considerar distintas funciones para el cálculo de la distancia, como son la distancia Manhattan, Euclídea y Minkowski. [29] Para nuestro modelo, haremos uso de la distancia euclídea, ya que es la más utilizada y la que viene por defecto en la librería de sklearn para KNN. [7]

En nuestro programa, logramos optimizar este modelo mediante la obtención de los mejores valores de los hiperparámetros con GridSearch, como se comentaba previamente. Con el fin de hacer posible la escalabilidad de este proyecto, buscamos en un rango del 1 al 3 en saltos de 1 unidad en todos las bases de datos.

De esta forma, obtenemos los valores mostrados en la Tabla 4 para el número más adecuado de vecinos cercanos a considerar para cada uno de los databases, así como los resultados de las métricas de clasificación:

		K	Accuracy	Recall	F1 score
Database 1		1	0,826	0,826	0,811
Database 2		1	0,915	0,915	0,916
Database 3	**Raw**	1	0,853	0,853	0,855
	Noise	1	0,881	0,881	0,882
	Stacked	2	0,872	0,872	0,872

Tabla 4. Hiperparámetros óptimos y métricas de rendimiento para KNN

Por otra parte, cabe considerar las matrices de confusión para el modelo KNN aplicado en el Database 1 (Figura 9), en el Database 2 (Figura 10), en el Database 3 con imágenes en

crudo (Figura 11), en el Database 3 con imágenes filtradas para la reducción de ruido (Figura 12) y en el Database 3 con superposición de ambas (Figura 13):

Figura 9. Matriz de confusión KNN Database1

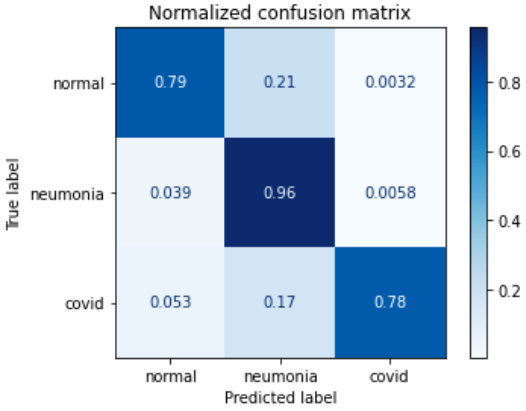


Figura 10. Matriz de confusión KNN Database2

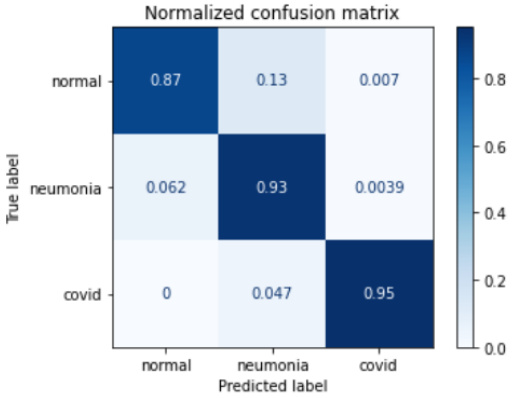


Figura 11. Matriz de confusión KNN Db3_raw

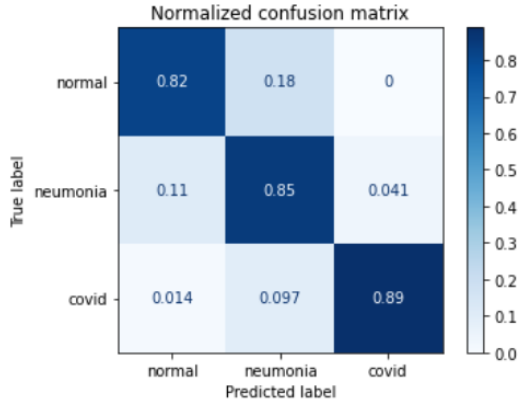


Figura 12. Matriz de confusión KNN Db3_noise

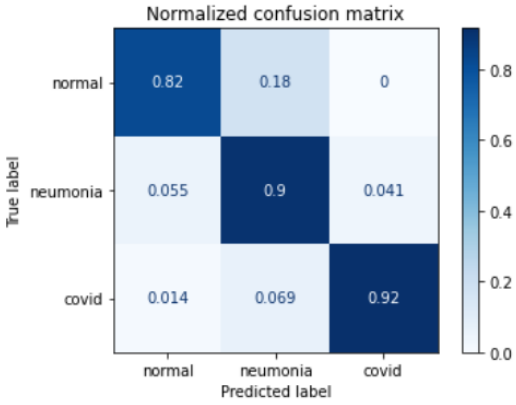
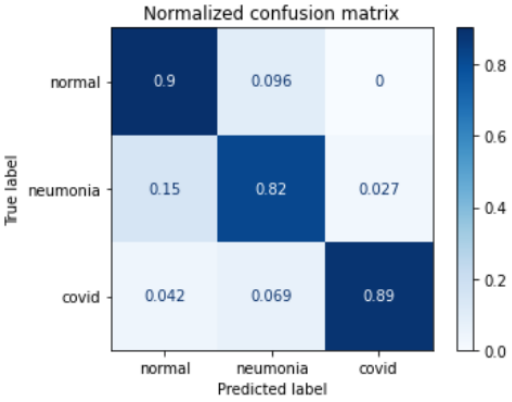


Figura 13. Matriz de confusión KNN Db3_stacked



En este caso, observamos una **mejora de las métricas** de rendimiento para todas las bases de datos, manteniéndose de nuevo el **Database 2** como el mejor de los comparados, con un 91,5% de accuracy y 91,6% de F1 score. Comparando con el modelo anterior, el Database 3 con **reducción de ruido logra la mayor mejora de todos**, con un aumento del 12% en las métricas.

Observando las matrices de confusión, nos sorprende observar que con KNN la primera base de datos logra la correcta clasificación de pacientes con neumonía en el 95% de los casos, más que en el caso de pacientes de COVID. Cabe destacar el gran rendimiento de este modelo con la base de datos 2, ya que nos permite obtener un 0% de falsos negativos diagnosticados sin enfermedad, lo cual supone un factor de vital importancia en la aplicabilidad del modelo en la realidad. En efecto, en virtud de este modelo, no habría ningún paciente con COVID al que se le diagnosticara como si no padeciera ninguna enfermedad, lo cual es especialmente relevante para nuestro ámbito de estudio, ya que se reduce la posibilidad de contagio de COVID-19 por dar un falso negativo.

Database 2 y 3 en crudo y con reducción de ruido logran mejores resultados en la detección de COVID, mientras que el Database 3 stacked detecta mejor los pacientes sin enfermedad. Asimismo, la primera base de datos detecta mejor los pacientes con neumonía que los que padecen de COVID con este modelo. El aspecto que más debería preocuparnos es la detección errónea del 17% de los pacientes del Database 1 con COVID, que serían diagnosticados con neumonía. Aunque en esta misma base de datos se obtenga una detección errónea superior en pacientes sin enfermedad a los que se les asigna neumonía (21%), al tratarse de una asignación más restrictiva no resulta tan alarmante.

5.3.3 RANDOM FOREST

Random Forest es un método de tipo “ensemble” basado en los árboles de decisión. Este tipo de modelos procura ajustar y combinar varios algoritmos para hallar la predicción más acertada entre los distintos métodos empleados. En el caso de Random Forest, se utiliza un mismo algoritmo (árboles de decisión) pero aplicados a distintos subconjuntos aleatorios de los datos del entrenamiento. [9]

En este sentido, siguiendo la notación de sklearn [7], el hiperparámetro "n_estimators" hace referencia al número de árboles de decisión más adecuado para la comparativa en el ensemble, que por defecto se establece en 100. Por otra parte, max_depth hace referencia al máximo número de ramificaciones dentro de un mismo decision tree. Igual que en el caso anterior, se procede al método iterativo de optimización de los hiperparámetros descrito a continuación:

- Database 1 y Database 2. Para max_depth, buscamos en un rango del 1 al 3 con saltos de 1 unidad. Para n_estimators, el rango se establece inicialmente del 1 al 1000 en saltos de 100, para finalmente hallar el óptimo centrando el rango del 500 al 700 en saltos de 10 (Database 1) y un rango de 800 a 900 en saltos de 10 (Database 2)
- Database 3 (raw, noise y stacked). Teniendo en cuenta los valores de n_estimators óptimos en los casos anteriores, decidimos reducir el rango de búsqueda a valores del 500 a 1000 en saltos de 100 unidades inicialmente. Posteriormente acotamos el rango y hacemos uso de escalones de 10 unidades para afinar los resultados. Asimismo, mantenemos los valores de max_depth en un rango del 1 al 3 en saltos unitarios.

La ejecución del modelo con los hiperparámetros obtenidos gracias al proceso iterativo descrito da lugar a los resultados comprendidos en la Tabla 5:

		Hiperparámetros		**Métricas**		
		n_estimators	**max_depth**	**Accuracy**	**Recall**	**F1 score**
Database 1		600	2	0.826	0.826	0.811
Database 2		880	2	0,817	0,817	0,817
Database 3	**Raw**	590	2	0,821	0,821	0,821
	Noise	530	2	0,725	0,725	0,727
	Stacked	869	2	0,812	0,812	0,813

Tabla 5. Hiperparámetros óptimos y métricas de rendimiento para Random Forest

Asimismo, podemos analizar las matrices de confusión para el modelo Random Forest (RF) aplicado en el Database 1 (Figura 14), en el Database 2 (Figura 15), en el Database 3 con imágenes en crudo (Figura 16), en el Database 3 con imágenes filtradas para la reducción de

ruido (Figura 17) y en el Database 3 con superposición de las imágenes originales y con filtro (Figura 18).

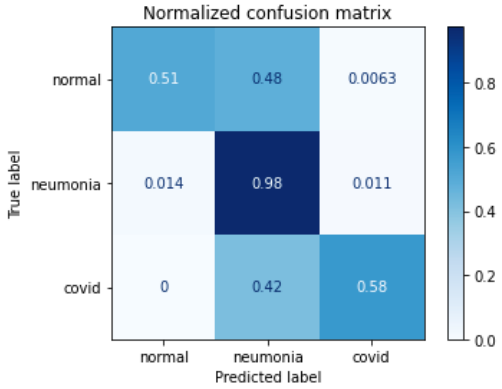


Figura 14. Matriz de confusión RF Database1

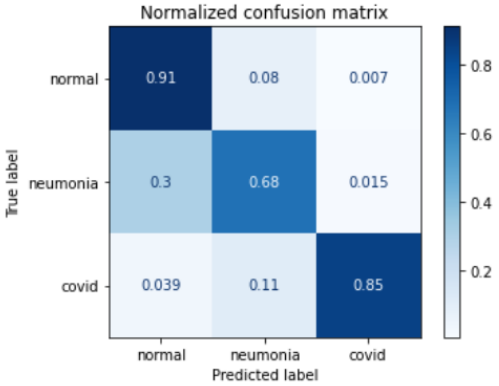


Figura 15. Matriz de confusión RF Database2

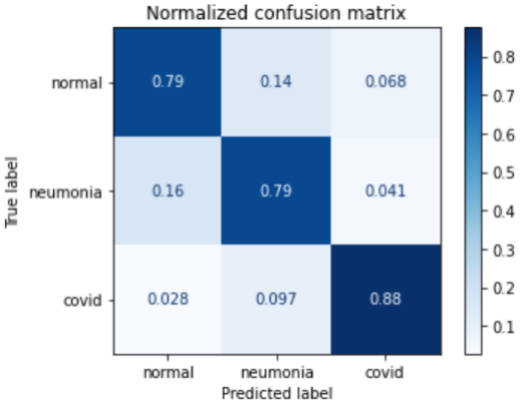


Figura 16. Matriz de confusión RF Db3_raw

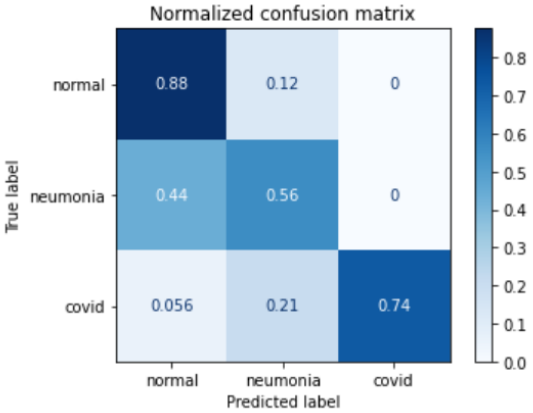


Figura 17. Matriz de confusión RF Db3_noise

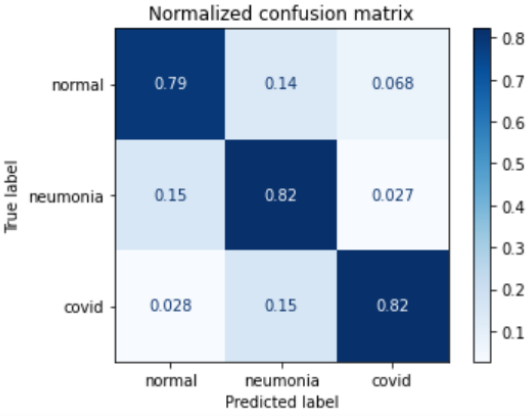


Figura 18. Matriz de confusión RF Db3_stacked

En la ejecución de **Random Forest** se obtienen **resultados algo inferiores a KNN**, y de **mejora variable respecto a Naïve Bayes según la base de datos** de la que se trate. Para la primera base de datos este modelo no resultaría demasiado efectivo en la detección de COVID, ya que como muestra su matriz de confusión sería más adecuado en la detección de neumonía (los pacientes de COVID solo serían detectados correctamente en un 58% de los casos). Por tanto, aunque las métricas de rendimiento del Database 1 sean considerablemente buenas, no debemos dar por hecho que sea el mejor modelo para esta base de datos.

Asimismo, observamos **detecciones erróneas bastante superiores** en los Database 1, 2 y 3 con reducción de ruido, siendo la detección errónea más elevada de un 48% en pacientes sin enfermedad diagnosticados con neumonía y la más alarmante de un 42% de pacientes con COVID diagnosticados con neumonía (Database 1). Observamos errores de diagnóstico más elevados en todas las matrices de confusión, por lo que concluimos que para nuestro estudio este modelo no resultaría ser la mejor opción.

A pesar de que las métricas de rendimiento son bastante elevadas, cabe recordar que en nuestro estudio de clasificación **lo más importante es la detección de COVID y minimizar la asignación de falsos negativos de esta enfermedad**, por lo que la clasificación en categorías tiene una importancia distinta según cada grupo.

5.3.4 GRADIENT BOOSTING CLASSIFIER

Un modelo Gradient Boosting está formado por un conjunto de árboles de decisión individuales, entrenados de forma secuencial, de forma que cada nuevo árbol trata de mejorar los errores de los árboles anteriores. La predicción de una nueva observación se obtiene agregando las predicciones de todos los árboles individuales que forman el modelo. [30]

En nuestro caso, y siguiendo la implementación de sklearn, los hiperparámetros que han sido ajustados son: n_estimators, max_depth, min_samples_split, min_samples_leaf. N_estimators hace referencia al número de árboles de decisión utilizados, y max_depth a la profundidad de los árboles (en este caso, suele tomar un valor bajo). Min_samples_split indica el mínimo número de muestras para desglosar un nodo del árbol (2 por defecto), y

min_samples_leaf determina el mínimo número de muestras para desglosar un nodo de hojas dentro de una misma rama del árbol (predeterminado en 1).

Observando los resultados de los modelos anteriores, procuraremos **perfeccionar** la **clasificación de las radiografías del Database 2** exclusivamente, ya que es la base de datos que ha logrado obtener las métricas de eficiencia superiores (con el modelo KNN). Esto se debe a que el coste computacional de este algoritmo es muy elevado, requiriendo 6 días y 21 horas para la finalización de la búsqueda con GridSearch. De la misma forma que en los modelos de Machine Learning anteriores, detallamos a continuación los rangos de hiperparámetros analizados en la búsqueda de los valores óptimos.

- Database 2:
  - Para hallar n_estimators, se parte de un rango de búsqueda del 500 al 1000 en pasos de 100.
  - Max_depth se fijará al valor 1, con el fin de reducir los tiempos de cada iteración al tener que profundizar menos en cada modelo.
  - Min_samples_leaf inicialización de rango de 1 a 3 en steps de 1 unidad.

Tras seguir este proceso, se obtienen los resultados mostrados en la Tabla 6:

	Hiperparámetros			**Métricas**		
	n_estimators	**max_depth**	**Min_samples_leaf**	**Accuracy**	**Recall**	**F1 score**
Database 2	600	1	1	0,942	0,942	0,942

Tabla 6. Hiperparámetros óptimos y métricas de rendimiento para Gradient Boosting

Mediante el uso de Gradient Boosting en el Database 2, logramos aumentar las métricas de rendimiento en un 2.84%, alcanzando un **94,2% en las métricas de eficiencia**. Estos resultados son muy satisfactorios, ya que **se asemejan** considerablemente a los que se suelen obtener con las **pruebas de detección de COVID-19** realizadas con muestras de mucosa y sangre en centros de salud. [5] No obstante, no parece el modelo más adecuado si se pretende aumentar la base de datos para el entrenamiento continuado del modelo, debido a los elevados tiempos de ejecución mencionados anteriormente.

Asimismo, como se puede observar en la matriz de confusión unitaria (Figura 19). Este algoritmo **detecta COVID-19 correctamente en un 96% de los casos**, con tan solo un 1,9% de negativos mal clasificados a los que se les diagnosticaría sin ninguna enfermedad. No obstante, habría también otro 1,9% de pacientes con COVID-19 que se clasificarían como enfermos de neumonía, dando lugar a un **3,8% de falsos negativos** en total.

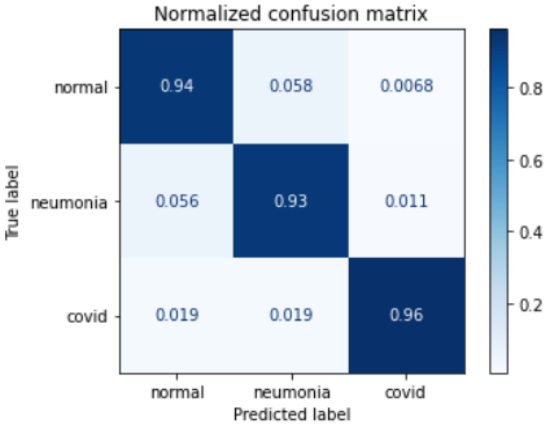


Figura 19. Matriz de confusión Gradient Boosting Database 2

5.4 EVALUACIÓN DE MÉTRICAS DE EFICIENCIA

A modo de resumen, en la Tabla 7 se muestran los resultados de las métricas para la configuración óptima de los hiperparámetros en cada uno de los modelos analizados:

Base de datos		Métrica	Naïve Bayes	Random Forest	KNN	Gradient Boosting
Database 1		Accuracy	0,8	0,826	0,826	
		F1 score	0,808	0,811	0,811	
Database 2		Accuracy	0,861	0,817	**0,915**	**0,942**
		F1 score	0,86	0,817	**0,916**	**0,942**
Database 3	**Raw**	Accuracy	0,812	0,821	0,853	
		F1 score	0,813	0,821	0,855	
	Noise	Accuracy	0,784	0,725	**0,881**	
		F1 score	0,785	0,727	**0,882**	
	Stacked	Accuracy	0,794	0,812	0,872	
		F1 score	0,795	0,813	0,872	

Tabla 7. Comparativa de métricas de rendimiento entre todas las bases de datos

A raíz de estos resultados y el análisis de las matrices de confusión contenidas en el apartado correspondiente a cada modelo (sección 5.3), podemos afirmar que el **modelo que mejores**

métricas nos proporciona es el método ensemble de Gradient Boosting, que se ha utilizado para el perfeccionamiento de la clasificación del Database 2, que gozaba con los resultados más destacables en el resto de los modelos de clasificación.

Sin embargo, como ya se ha comentado, Gradient Boosting es un modelo que requiere **elevados tiempos de cómputo** para la optimización de sus hiperparámetros a pesar un servidor optimizado para CPU, ya que nos ha llevado 6 días obtenerlos. Por este motivo, para los análisis que se presentarán posteriormente se hará uso de los hiperparámetros óptimos ya hallados para el Database 2, tomándose como base de datos representativa.

Paralelamente, consideramos que el **modelo KNN** ha logrado **resultados altamente satisfactorios para todas las bases de datos con una utilización considerablemente inferior de recursos** (tiempos de ejecución para su optimización mucho más manejables), por lo que lo mantendremos en los estudios de las secciones siguientes.

En el caso del modelo KNN se debe tener en cuenta que en la comparativa de las radiografías del Database 3 en bruto (raw), con reducción de ruido (noise) y las dos opciones solapadas (stacked), se obtienen mejores resultados haciendo uso de las radiografías con un **filtrado de reducción de ruido, mejorando el rendimiento en las métricas en un 3,5% respecto a las radiografías en bruto**. Teniendo esto en cuenta, sería de gran interés aplicar un filtro de este tipo como preprocesamiento de las radiografías en análisis posteriores para comparar si afecta positivamente a los resultados en las otras bases de datos.

Capítulo 6. OPTIMIZACIÓN DEL PROYECTO

Una vez se ha completado el estudio comparativo de la clasificación de radiografías para las tres bases de datos por separado y con cada uno de los modelos de aprendizaje automático, se procederá a aportar una mayor dosis de realismo y aplicabilidad a este trabajo de investigación.

Recuérdese que el objetivo final de este Trabajo de Fin de Grado consiste en proporcionar un programa eficiente para la detección de COVID-19 en cualquier tipo de radiografía de la zona pectoral que se introduzca en él. Esto supone que cualquier paciente, radiografiado con cualquier sistema, puede ser sensible de diagnóstico. Sin embargo, las pruebas realizadas hasta ahora han sido considerablemente parciales debido a la evaluación del impacto en pequeños datasets diferenciados.

Por ese motivo, en esta sección se analizará la viabilidad y escalabilidad de los modelos determinado en la sección anterior como más efectivos (modelos **Gradient Boosting y KNN**) para la **totalidad de radiografías comprendidas en las tres bases de datos**, englobadas en un único conjunto al que denominaremos **Global Database**.

Esta nueva base de datos incluirá todas las radiografías de las bases de datos anteriores salvo en el caso de Database 3, del que **solo se incluirán las radiografías en bruto** (Database 3 raw), para **simular el procedimiento real que tendría lugar en un centro de diagnóstico** en el que se utilizara nuestro programa para la detección de COVID-19. en este caso, no habría una reducción de ruido o tratamiento de la radiografía previo, solo el que se incorporase en nuestro programa.

De esta forma, para los análisis finales de este proyecto trabajaremos con dicha base de datos global, que consta de una totalidad de 11403 radiografías, siendo 2134 categorizadas con COVID-19, 5980 con neumonía y 3289 de pacientes sin enfermedad. A efectos de análisis, se realizará una reordenación aleatoria de las imágenes para asegurar imparcialidad en los resultados. Asimismo, se dividirá en una muestra de train del 80% de la base de datos, dejando un 20% para la fase de test, como se mencionaba en la sección 5.1.

6.1 KNN PARA EL MODELO GLOBAL Y CON REDUCCIÓN DE RUIDO

En este apartado, se trabajará con la base de datos Global Database creada por superposición de todas las utilizadas en la sección 5.3. Además del análisis básico contra el dataset en crudo y en vista de la mejora de las métricas obtenidas en el Database 3 con reducción de ruido en el apartado 5.3.2, en esta sección analizaremos el impacto de un filtrado de este tipo.

Se hará uso del filtro de reducción de ruido "median" de la librería skimage [31], que resulta de gran utilidad para reducir ruido en imágenes. Este filtrado logra sustituir el valor del píxel central de una ventana de la radiografía en cuestión por el valor medio de la misma, eliminando así los ruidos que tienden a alejarse de este valor central. [32]

El primer paso, al igual que en las secciones previas, será determinar el valor óptimo del hiperparámetro K, para lo cual se sigue un procedimiento iterativo de búsqueda en un rango de valores del 1 al 3 en saltos de 1 unidad. Aplicando este proceso para las radiografías en bruto y con un filtrado de reducción de ruido, se obtienen los resultados mostrados en la Tabla 8:

	K	Accuracy	Recall	F1 score
Global Database	1	0,972	0,972	0,972
Global Database Noise Filter	1	0,975	0,975	0,975

Tabla 8. Hiperparámetros y métricas de rendimiento para KNN Global Database

En el caso del **modelo KNN sin filtrado para Global Database**, cabe destacar los excelentes resultados (**métricas de rendimiento de más del 97%**) con un uso ínfimo de recursos, ya que partiendo del hiperparámetro K=1, el **tiempo de ejecución** necesario para la obtención del entrenamiento y test de este modelo ha sido de **12 minutos** para la totalidad de las radiografías. Podemos argumentar esta mejora gracias a la ampliación de la base de datos de trabajo, ya que el modelo ha entrenado con el 80% de las 11403 radiografías constituyentes del Global Database.

Logramos una ligera mejora en las métricas de eficiencia en el caso de la aplicación del filtrado para la reducción de ruido explicado con anterioridad, aunque este perfeccionamiento es algo inferior al observado en el análisis del Database 3 raw frente a

Database 3 noise (anteriormente se mejoraba un 3% en las métricas de eficiencia, y con el Global Database apenas un 0,3%).

De igual forma, podemos evaluar el rendimiento del modelo KNN sin el efecto del filtrado (Figura 20) y con él (Figura 21) mediante las matrices de confusión correspondientes:

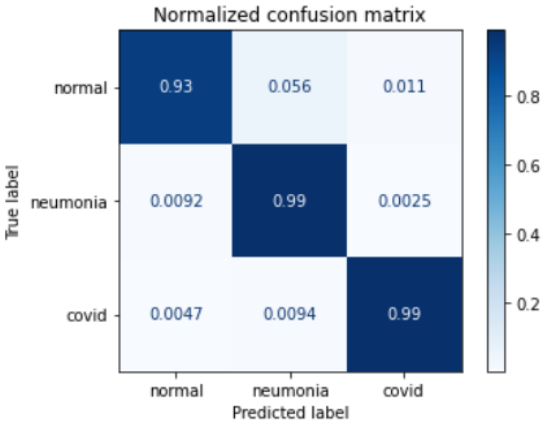


Figura 20. Matriz de confusión KNN Global Db

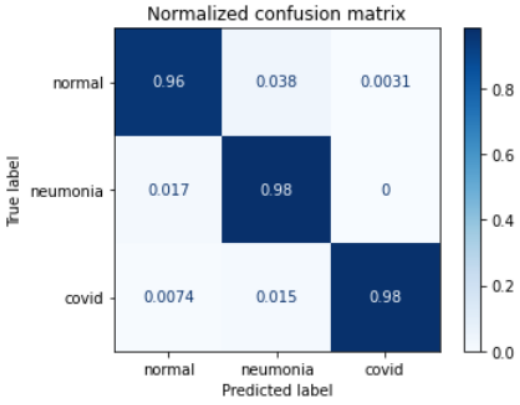


Figura 21. Matriz de confusión KNN Global Db filter

El modelo con las imágenes en bruto resulta de gran utilidad para el diagnóstico de COVID-19 y neumonía, aunque la proporción de personas sanas bien diagnosticadas es algo inferior, sigue siendo de un 93%. **Los falsos negativos de COVID-19 apenas alcanzan el 1,4%.** Teniendo todo esto en cuenta, podemos considerar este modelo como una solución realista implementable al problema de la detección de COVID-19.

Paralelamente, en el análisis llevado a cabo con el filtrado de las radiografías para la reducción de ruido observamos unos resultados igual de satisfactorios, manteniendo unos tiempos de ejecución de 12 minutos una vez localizado el hiperparámetro óptimo. La correcta detección de personas sanas aumenta en un 3% respecto al modelo con las imágenes en bruto, pero disminuye en un 1% en la correcta clasificación de COVID-19 y neumonía. **Los falsos negativos de COVID-19 se localizarían en un 2,2% en este escenario.**

Para nuestro ámbito de estudio **siempre es preferible diagnosticar de manera más restrictiva (evitar falsos negativos más que falsos positivos),** y teniendo en cuenta que nuestro objetivo principal es la detección de COVID podemos razonar que la mejor opción sería la utilización de este modelo con las imágenes en bruto, ya que las métricas de

eficiencia presentan una mejora prácticamente despreciable y la correcta clasificación de las personas con enfermedad disminuiría.

6.2 GRADIENT BOOSTING PARA EL MODELO GLOBAL

A pesar de los excelentes resultados obtenidos en la sección anterior, se procederá a la evaluación de un último modelo para comprobar si existe una mejora sustancial en el rendimiento: Gradient Boosting con la base de datos global. El motivo de esta decisión se sustenta en las conclusiones extraídas del apartado correspondiente a este modelo en la sección 5.4, en la que se ha podido comprobar que es el algoritmo que presenta mejores métricas de rendimiento**.**

Nótese que, como se explicaba en la introducción, el desarrollo del modelo para la base de datos global con Gradient Boosting se realizará utilizando los valores de los hiperparámetros determinados en el capítulo anterior. Esto es debido a los elevados costes computacionales en términos de tiempos de ejecución, ya que esta base de datos es de un tamaño de casi el triple que Database 2, utilizado en la sección 5.3.4. y que había necesitado 6 días de cómputo. Haciendo uso de los hiperparámetros óptimos para el Database 2 como representación del conjunto total de radiografías, se obtienen los resultados de la Tabla 9:

	Hiperparámetros			**Métricas**		
	n_estimators	**max_depth**	**Min_samples_leaf**	**Accuracy**	**Recall**	**F1 score**
Global Database	600	1	1	0,966	0,966	0,966

Tabla 9. Hiperparámetros utilizados y métricas de rendimiento para Gradient Boosting Global Database

Resulta muy gratificante observar las métricas de evaluación del rendimiento del **Gradient Boosting**, alcanzando un **96,6% de eficiencia**. Cabe destacar que este resultado podría ser superior si se hubiera llevado a cabo la búsqueda de hiperparámetros óptimos con GridSeach. Esta metodología no se ha llevado a cabo en este caso por la falta de realismo en la aplicación de este modelo en la detección de COVID-19 en un centro de diagnóstico real, ya que si se pretendiesen actualizar dichos hiperparámetros óptimos al aumentar la base de datos con cada radiografía evaluada conllevaría unos elevados tiempos de ejecución y la necesidad de recursos tecnológicos de los que no se tendría disponibilidad.

Asimismo, los resultados de la clasificación se muestran de manera unitaria en la matriz de confusión reflejada en la Figura 22:

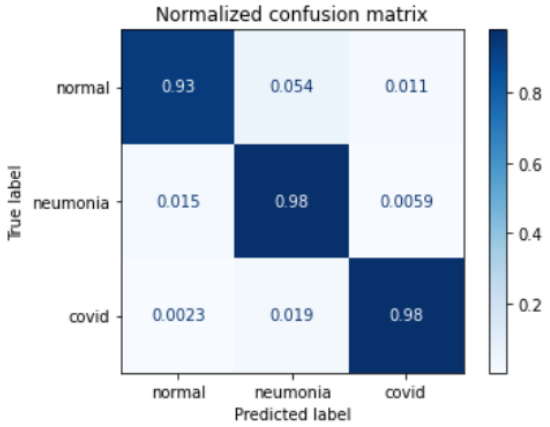


Figura 22. Matriz de confusión Gradient Boosting Global Database

En esta matriz se muestra un **98% de pacientes diagnosticados con COVID-19 correctamente**, y tan **solo un 2,13% de falsos negativos** (siendo un 1,9% diagnosticados con neumonía, y tan solo el 0,23% clasificados como pacientes sanos). Aunque los resultados son algo inferiores a los mostrados en el modelo KNN para el mismo Global Database, podemos confirmar el destacable rendimiento de este modelo mediante la matriz de confusión, por la ínfima cantidad de falsos negativos en cada categoría. El caso de clasificación errónea más elevada es la asignación del 5,4% de los pacientes sin enfermedad como enfermos de neumonía.

Los **tiempos de ejecución** del conjunto de entrenamiento y test con el servidor Galactica para el entrenamiento del modelo **son algo más elevados** (4h y 41 min partiendo del hiperparámetro a utilizar), por lo que **podemos recomendar el modelo KNN por encima de Gradient Boosting, al obtener mejores resultados en menos tiempo.** Cabe destacar que el tiempo de cómputo en el caso de KNN estaría invertido mayoritariamente en la fase de entrenamiento, siendo más rápida la fase de test. Por el contrario, en Gradient Boosting la ejecución es más lenta, tanto en train como en test.

Capítulo 7. ANÁLISIS DE RESULTADOS

Para la detección de COVID-19 con modelos de Machine Learning, se han aplicado los distintos modelos descritos a lo largo de este trabajo (Naïve Bayes, KNN, Random Forest y Gradient Boosting) a las tres bases de datos de partida de este trabajo por separado.

Una vez se han obtenido esos primeros resultados, se ha procedido a la optimización del proyecto haciendo uso de una base de datos que contiene a todas las imágenes en bruto del conjunto de los tres databases, con el fin de perfeccionar el entrenamiento del modelo y aportar mayor realismo a nuestro programa. Esta base de datos global ha sido utilizada para el entrenamiento y test de los modelos comprobados como los más efectivos en las secciones anteriores de este trabajo, evaluando la clasificación de las radiografías para KNN, KNN con reducción de ruido y Gradient Boosting Classifier. De la ejecución de dichos programas se obtienen los resultados de métricas de eficiencia recopilados en la Tabla 10:

Base de datos	Métrica	Gradient Boosting	KNN
Global Database	Accuracy	0,966	**0,972**
	F1 score	0,066	**0,972**
Global Database Noise Filter	Accuracy		**0,975**
	F1 score		**0,975**

Tabla 10. Métricas de rendimiento para Global Database para los modelos finalistas

Como se explicaba en la sección correspondiente a Gradient Boosting, los hiperparámetros de este modelo no han sido optimizados para la totalidad de la base de datos global debido a la extensión de los tiempos de ejecución observados en el análisis previo. Con el fin de testear este modelo, se han tomado como representativos los hiperparámetros óptimos para el Database 2 obtenidos en la sección 5.3.4, que son los mostrados en la tabla superior. Este podría ser el principal motivo por el que KNN obtiene resultados superiores a Gradient Boosting, ya que K sí ha sido hallado por GridSearch para lograr los resultados óptimos.

A pesar de ello, consideramos que para la utilización de este programa de detección de COVID-19 y teniendo en cuenta el consumo de recursos necesario para ejecutar Gradient Boosting, **la opción más recomendable sería implementar KNN sin reducción de ruido**

en los centros de salud o lugares donde se pueda realizar un análisis de detección de COVID-19. Esto nos permite el mayor número de aciertos en la detección de COVID-19 y minimizar los falsos negativos, aspecto clave en este estudio.

Capítulo 8. CONCLUSIONES Y TRABAJOS FUTUROS

En este Trabajo de Fin de Grado se ha abordado el problema de diagnóstico de COVID-19 mediante técnicas de aprendizaje automático. Para ello, se han empleado los modelos de clasificación Naïve Bayes, KNN, Random Forest y Gradient Boosting.

El desarrollo del proyecto ha dado comienzo con una fase de análisis preliminar (Capítulo 5), en la que se ha realizado un estudio exhaustivo de la efectividad de los distintos modelos de aprendizaje automático aplicados en distintas bases de datos, de tamaño y formato diverso. Siguiendo este procedimiento, se han considerado las primeras conclusiones, extraídas del Capítulo 6. En ellas, se observaba que el modelo aplicado que nos conducía a mejores resultados era Gradient Boosting Classifier aplicado al Database 2, ya que es la base de datos con mejores resultados en el resto de los modelos utilizados de forma generalizada.

Asimismo, se destacaba la efectividad del modelo KNN, en especial para el Database 2, ya que proporcionaba resultados muy favorables en unos tiempos de cómputo mucho más reducidos que Gradient Boosting. Una apreciación interesante ha sido que al utilizar dicho modelo (KNN) en la comparativa del Database 3 (imágenes en bruto y con reducción de ruido), se obtenía una mejora del rendimiento del programa con la aplicación de este filtrado.

Teniendo en cuenta estas valiosas deducciones, se ha procedido a la **optimización del modelo** en el Capítulo 6, procurando aportar un mayor realismo y aplicabilidad de nuestro programa al aumentar la base de datos de trabajo a la yuxtaposición de todas las radiografías en bruto aportadas por las tres bases de datos de nuestro proyecto. Esto ha permitido entrenar el modelo de forma más completa (con más de 11000 radiografías de distintos tipos) para que este pudiera ser utilizado posteriormente en la detección de COVID-19 en centros de diagnóstico reales. Gracias a la utilización de esta **base de datos conjunta**, se ha reducido el riesgo derivado de un posible sesgo en las diferencias entre resultados por las calidades de cada base de datos, ya que en la aplicación real de este trabajo cada punto de diagnóstico proporcionaría la radiografía de entrada con especificidades diversas.

De esta forma, el análisis empírico demuestra que logramos alcanzar excelentes resultados en la detección de COVID-19. El modelo global es capaz de alcanzar unas **métricas de rendimiento del 97% (accuracy y F1 score) para KNN sin filtado en las imágenes, y una correcta clasificación de pacientes con COVID-19 y neumonía del 99% de las veces.** A pesar de que las métricas de eficiencia mejoraban ligeramente con la reducción de ruido de las radiografías, finalmente se ha priorizado la correcta detección de la enfermedad, por lo que se ha corroborado la información extraída en primer lugar de la bibliografía. [30]

La Tabla 10 elaborada en el Capítulo 7 muestra un resumen de los resultados globales del proyecto:

Base de datos	Métrica	Gradient Boosting	KNN
Global Database	Accuracy	0,966	**0,972**
	F1 score	0,066	**0,972**
Global Database Noise Filter	Accuracy		**0,975**
	F1 score		**0,975**

Tabla 10. Métricas de rendimiento para Global Database para los modelos finalistas

Teniendo en cuenta el consumo de recursos necesario para ejecutar Gradient Boosting, **la opción más recomendable sería implementar KNN sin reducción de ruido** en los centros de salud o lugares donde se pueda realizar un análisis de detección de COVID-19. Esto nos permite el mayor número de aciertos en la detección de COVID-19 y minimizar los falsos negativos, aspecto clave en este estudio. Asimismo, como recomendación, sería interesante ampliar la base de datos de entrenamiento con cada radiografía que se introdujera en el programa para su clasificación, **manteniendo el aprendizaje del modelo en el tiempo.** Esto sería factible al utilizar KNN pero no con Gradient Boosting, por el elevado coste computacional observado en la ejecución del modelo.

Los resultados y conclusiones extraídas de este proyecto son altamente satisfactorios, y comparables con las fiabilidades de los distintos métodos y tests de detección de esta enfermedad que existen actualmente en el mercado. La sensibilidad en la clasificación de COVID-19 con el uso de inteligencia artificial se ha optimizado hasta alcanzar un 99**%, al nivel de la PCR** (con prácticamente el 100% de casos correctamente diagnosticados [11]). Asimismo, es capaz de **superar** la clasificación del test de **anticuerpos** (con una sensibilidad

del 88.6% [13]), de **antígenos** (de 90% a 100%, dependiendo de los días transcurridos desde la infección [5]) y al **test rápido** (96,77% de sensibilidad [14]).

Otra ventaja competitiva de este método de detección de la enfermedad sería el **bajo coste** económico (tanto el precio de la prueba como los recursos necesarios de personal sanitario) y de tiempo. Asimismo, la detección de COVID-19 en radiografías permitiría analizar si la causa del fallecimiento de un paciente ha sido por este tipo de coronavirus a posteriori, cuando ya no es posible realizar pruebas de análisis de mucosa o sangre.

Sin embargo, la propuesta de detección de COVID-19 a través de radiografías quedaría supeditada al **análisis posterior** de una posible diferencia en los resultados de clasificación **según si las personas presentan síntomas de la enfermedad o no**. Esta idea sería una propuesta interesante como trabajo futuro si se pudiera contar con una nueva categorización de las imágenes según el tipo de paciente (con o sin síntomas), ya que no podemos afirmar que la fiabilidad de nuestro programa sea homogénea para ambos tipos de pacientes. Asimismo, podría realizarse una mayor profundización en la detección de COVID-19 utilizando modelos de aprendizaje profundo (**Deep Learning**) además de los ya estudiados programas de aprendizaje automático (Machine Learning) tradicionales.

BIBLIOGRAFÍA

[1] M. A. Shereen, S. Khan, A. Kazmi, N. Bashir, R. Siddique (15 marzo 2020), "COVID-19 infection: Emergence, transmission, and characteristics of human coronaviruses", Journal of Advanced Research. Recuperado en junio 2021 de https://doi.org/10.1016/j.jare.2020.03.005

[2] Organización Mundial de la Salud (OMS) (10 noviembre 2020), "Información básica sobre la COVID-19. ¿Cuáles son los síntomas de la COVID-19?". Recuperado en junio 2021 de https://www.who.int/es/emergencies/diseases/novel-coronavirus-2019/question-and-answers-hub/q-a-detail/coronavirus-disease-covid-19

[3] E. Dong, H. Du, L. Gardner (19 febrero 2020), "An interactive web-based dashboard to track COVID-19 in real time", Johns Hopkins University. Recuperado en junio 2021 de https://doi.org/10.1016/S1473-3099(20)30120-1

[4] C. Zimmer, J. Corum, S. Wee (junio 2021), "Coronavirus Vaccine Tracker", The New York Times. Recuperado en junio 2021 de https://www.nytimes.com/interactive/2020/science/coronavirus-vaccine-tracker.html

[5] Quirón Salud (2021). Recuperado en junio 2021 de: https://www.quironsalud.es/es/informacion-pruebas-covid

[6] [Lenguaje de programación Python] (s.f.). Python https://www.python.org

[7] [Libería sklearn] (s.f.), "Scikit-Learn: Machine Learning in Python". Recuperado en marzo 2021 de: Scikit-learn https://scikit-learn.org/stable/

[8] [Librería matplotlib] (s.f.), "Matplotlib: Visualization with Python". Recuperado en marzo 2021 de: Matplotlib https://matplotlib.org

[9] [Librería skimage] (s.f.), "Scikit-Image: Image Preprocessing in Python". Recuperado en marzo 2021 de: Scikit-image https://scikit-image.org

[10] [Librería PIL] (s.f.), "Pillow. PIL Fork". Recuperado en marzo 2021 de: Pillow https://pillow.readthedocs.io/en/stable/

[11] The College of American Pathologists (s.f.), "How Good are COVID-19 (SARS-CoV-2) Diagnostic PCR Tests?". Recuperado en junio 2021 de:

https://www.cap.org/member-resources/articles/how-good-are-covid-19-sars-cov-2-diagnostic-pcr-tests

[12] Gallardo Ponce (febrero 2021), "PCR, IgG, IgM y antígenos: ¿Cómo sé si he pasado el coronavirus?". Recuperado en febrero 2021 de: https://cuidateplus.marca.com/bienestar/2020/07/30/pcr-igg-e-igm-como-se-he-pasado-coronavirus-174147.html

[13] Dr. Carlos Rocha (s.f.), "¿Cómo se puede Diagnosticar el COVID-19? ¿Cuál es la Precisión de los Test Diagnósticos?". Recuperado en junio 2021 de: https://sociedadandaluzadeoftalmologia.es/como-se-puede-diagnosticar-el-covid-19-cual-es-la-precision-de-los-test-diagnosticos/

[14] Diario Farma (11 mayo 2021), "Cinfa distribuirá el primer test de antígenos covid-19 sin receta". Recuperado en junio 2021 de: https://www.diariofarma.com/2021/05/11/cinfa-distribuira-el-primer-test-de-antigenos-covid-19-sin-receta

[15] J. D. Arias-Londoño, J. A. Gómez-García, L. Moro-Velázquez and J. I. Godino-Llorente (2020), "Artificial Intelligence Applied to Chest X-Ray Images for the Automatic Detection of COVID-19. A Thoughtful Evaluation Approach", Institute of Electrical Electronics Engineers (IEEE). Recuperado en febrero 2021 de https://doi.org/10.1109/ACCESS.2020.3044858

[16] S. Kushwaha, S. Bahl, A. K. Bagha, K. S. Parmar, M. Javaid, A. Haleem, R. P. Singh (2020), "Significant applications of Machine Learning for COVID-19 pandemic", Journal of Industrial Integration and Management. Recuperado en junio 2021 de: https://doi.org/10.1142/S2424862220500268

[17] N. S. Punn, S. K. Sonbhadra, S. Agarwal (1 junio 2020), "COVID-19 Epidemic Analysis using Machine Learning and Deep Learning Algorithms". Recuperado en junio 2021 de: doi: https://doi.org/10.1101/2020.04.08.20057679

[18] E. Ong, M. U. Wong, A. Huffman, Y. He (3 julio 2020), "COVID-19 Coronavirus Vaccine Design Using Reverse Vaccinology and Machine Learning", University of Michigan. Recuperado en junio 2021 de: https://doi.org/10.3389/fimmu.2020.01581

[19] D. Molina-García, L. Vera-Ramírez, J. Pérez-Beteta et al. (2019) "Prognostic models based on imaging findings in glioblastoma: Human versus Machine", Nature Research (Scientific Reports). Recuperado en febrero 2021 de https://doi.org/10.1038/s41598-019-42326-3

[20] E. Turner, J. Medina, G. Brown (Septiembre 2019), "Dashing Hopes? the Predictive Accuracy of Domestic Abuse Risk Assessment by Police", The British Journal of Criminology. Recuperado en febrero 2021 de: https://doi.org/10.1093/bjc/azy074

[21] "Las 5 cepas de la covid que se han detectado hasta ahora y que preocupan a la OMS" (Febrero 2021), Las Provincias. Recuperado en febrero 2021 de https://www.lasprovincias.es/sociedad/salud/cepas-covid-20210219202828-nt.html

[22] [Web para la obtención de bases de datos y código de libre acceso Kaggle] (s.f.), "Kaggle" https://www.kaggle.com

[23] J. Martínez Heras (Octubre 2020), "Precision, Recall, F1, Accuracy en clasificación". Recuperado en febrero 2021 de https://www.iartificial.net/precision-recall-f1-accuracy-en-clasificacion/

[24] J. I. Barrios Arce (julio 2019), "La matriz de confusión y sus métricas". Recuperado en junio 2021 de: https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/

[25] P. Patel (últ. actualización septiembre 2020), "Chest X-ray (Covid-19 & Pneumonia)". Recuperado en enero 2021 de: https://www.kaggle.com/prashant268/chest-xray-covid19-pneumonia

[26] T. Rahman (Versión 3, enero 2021), "COVID-19 Radiography Database". Recuperado en enero 2021 de: https://www.kaggle.com/tawsifurrahman/covid19-radiography-database

[27] S. Gupta (Versión 1, junio 2020), "COVID-19 X-Ray Dataset With Preprocessed Images". Recuperado en enero 2021 de: https://www.kaggle.com/shreyanshgupta/covid19-xray-dataset-with-preprocessed-images

[28] Dr. Saed Sayad (n.f.), "Naïve Bayesian". Recuperado en febrero 2021 de: https://www.saedsayad.com/naive_bayesian.htm

[29] José Italo (noviembre 2018), "KNN K-Nearest Neighbors". Recuperado en junio 2021 de: https://towardsdatascience.com/knn-k-nearest-neighbors-1-a4707b24bd1d

[30] J. Amat Rodrigo (octubre 2020), "Gradient Boosting con Python". Recuperado en abril 2021 de: https://www.cienciadedatos.net/documentos/py09_gradient_boosting_python.html

[31] [Filtro Median de Scikit-Image] (s.f.). "Module: filters" https://scikit-image.org/docs/dev/api/skimage.filters.html#skimage.filters.median

[32] İmren Dinç, ... Marc L. Pusey, in Emerging Trends in Image Processing, Computer Vision and Pattern Recognition (Diciembre 2014). Recuperado en junio 2021 de https://www.sciencedirect.com/topics/computer-science/median-filter

APÉNDICE I: OBJETIVOS DE DESARROLLO SOSTENIBLE

Los Objetivos de Desarrollo Sostenible (ODS) propuestos por Naciones Unidas están constituidos por diecisiete metas que pretenden asegurar un futuro sostenible para todos, incorporando los desafíos globales a los que nos enfrentamos en nuestro día a día: la pobreza, la desigualdad, el clima, la degradación ambiental, la prosperidad, la paz y la justicia. [1]

El presente Trabajo de Fin de Grado tiene como objetivo ofrecer una nueva alternativa para la detección de COVID-19 con modelos de Inteligencia Artificial a partir de radiografías. Este estudio está íntimamente relacionado con el Objetivo 3 de Desarrollo Sostenible "Salud y bienestar", ya que se presenta una herramienta de gran utilidad para hacer frente a la crítica situación de la pandemia del COVID-19. Mediante la detección de nuevos contagiados, se pueden tomar las medidas adecuadas que dificulten la propagación del virus. Asimismo, se pretende ofrecer una alternativa viable y eficiente que permita la dedicación de recursos económicos y de personal sanitario al trato de los pacientes con sintomatología más grave y al desarrollo del plan de vacunación.

Paralelamente, el uso de este algoritmo para la detección de la enfermedad lleva asociados unos costes económicos muy inferiores a las alternativas de diagnóstico que se utilizan actualmente, lo reduciría la desigualdad entre los pacientes que requieren pruebas de detección según su nivel socioeconómico (ODS 10 "Reducción de las desigualdades"), siendo accesible para todos.

[1] Naciones Unidas (s.f.), "Objetivos de Desarrollo Sostenible". Recuperado en junio 2021 de: https://www.un.org/sustainabledevelopment/es/sustainable-development-goals/