Rev. cient. sist. inform. 2(1): e225; ene-jun, 2022
e-ISSN: 2709-992X
Universidad Nacional de San Martín https://doi.org/10.51252/rcsi.v2i1.225

Artículo original / Original article

Discriminación de masas mamográficas mediante K-Nearest Neighbor y atributos BIRADS

Mammographic mass discrimination using K-Nearest Neighbor and BIRADS attribute

Lévano-Rodriguez, Daniel [0000-0001-5652-0601][1]; Cerdán-León, Flor [0000-0001-6747-6335][1]

[1]Universidad Nacional Tecnológica de Lima Sur, Lima, Perú

✉ dlevano@untels.edu.pe

Recibido: 26/10/2021; Aceptado: 27/11/2021; Publicado: 20/01/2022

Resumen: La mamografía es el método más eficaz para la detección del cáncer de mama; sin embargo, el bajo valor predictivo, puede conducir a biopsias innecesarias. Esta investigación tiene como objetivo desarrollar un modelo predictivo para la discriminación de masas mamográficas mediante KNN y el atributo BIRADS con un nivel aceptable de exactitud, precisión, sensibilidad y puntaje-F1. Para ello, realizamos las siguientes fases: limpieza de los datos, entrenamiento del algoritmo KNN y selección del modelo. El resultado obtenido fue un modelo de discriminación de masas mamográficas con una exactitud del 85% así como niveles aceptables de precisión, sensibilidad y puntaje F1. Se concluye que es posible utilizar este modelo como un elemento de juicio para el diagnóstico de cáncer de mama; asimismo, a través de la tasa de error es posible encontrar modelos óptimos en KNN.

Palabras clave: cáncer de mama; knn; machine learning; pronóstico

Abstract: The mammography is the most effective method for the detection of breast cancer, however, the predictive value is low, and it can lead to unnecessary biopsies with benign results. This research aims to develop a predictive model for discrimination of mammographic masses using KNN and BIRADS attributes with an acceptable level of Accuracy, Precision, Recall and F1-Score. For this, we carried out the following phases: Data cleaning, KNN algorithm training and selection. The result obtained was a mammographic mass discrimination model with an accuracy=85% and acceptable levels of precision, sensitivity and F1-score. We concluded that it is possible to use this model as an element of judgment for the diagnosis of breast cancer; also that through the error rate it is possible to find optimal KNN models.

Keywords: breast cancer; knn; machine learning; prognostic

Cómo citar / Citation: Lévano-Rodriguez, D. & Cerdán-León, F. (2021). Discriminación de masas mamográficas mediante K-Nearest Neighbor y atributos BIRADS. *Revista científica de sistemas e informática*, 2(1), e225. https://doi.org/10.51252/rcsi.v2i1.225

© Los autores. Este artículo es publicado por la **Revista científica de sistemas e informática** de la Universidad Nacional de San Martín. Este es un artículo de acceso abierto, distribuido bajo los términos de la Licencia Creative Commons Atribución 4.0 Internacional (https://creativecommons.org/licenses/by/4.0/deed.es) que permite Compartir (copiar y redistribuir el material en cualquier medio o formato), Adaptar (remezclar, transformar y construir a partir del material) para cualquier propósito, incluso comercialmente.

1. Introducción

El cáncer es una de las principales causas de muerte femenina en todo el mundo. La Agencia Internacional para la Investigación del Cáncer (IARC) y la Sociedad Estadounidense del Cáncer informan que se registraron 17,1 millones de nuevos casos de cáncer en 2018 en todo el mundo; se estima que la incidencia de cáncer podría aumentar a 27,5 millones para 2040, con un estimado de 16,3 millones de muertes esperadas como resultado del cáncer (Gardezi et al., 2019).

Cáncer de mama

El cáncer de mama es una enfermedad neoplásica con transformación de células que se proliferan de manera anormal e incontrolada, ocasionando un crecimiento anormal de las zonas de la mama, secreciones, edemas, úlcera y enrojecimiento (Villavicencio Romero et al., 2019). Según reporta Estadísticas Globales del Cáncer (Globocan 2018), el cáncer de mama representó el 11,6% de todos los cánceres, lo que coloca a esta enfermedad como el segundo cáncer más comúnmente diagnosticado después del cáncer de pulmón, y causó el 6,6% del total de muertes por cáncer en 2018 (García Aranda & Redondo, 2019).

En Estados Unidos, representa el segundo más diagnosticado y la segunda causa de muerte por cáncer entre las mujeres; según la Asociación Estadounidense del Cáncer, hubo 268 600 mujeres recién diagnosticadas con cáncer de mama en el 2019, de las cuales 41 760 murieron a causa de la enfermedad (Al-Azzam & Shatnawi, 2021; Wu & Hicks, 2021).

En Europa es un importante problema de salud pública siendo la neoplasia diagnosticada con más frecuencia en mujeres y es un tercio de todos los casos nuevos de cáncer entre mujeres en los 31 países europeos en el 2018, así como la principal causa de muerte en mujeres europeas (Zielonke et al., 2021).

En el Perú el cáncer de mama, representa el segundo tipo de cáncer más frecuente en mujeres (Chachaima-Mar et al., 2021). El nivel de incidencia anual es de 28 casos por cada 100 mil habitantes con una tasa anual de 9.2 casos por cada 100 mil habitantes (Ministerio de Salud, 2017).

El método más aceptado para la clasificación y diagnóstico es el análisis de expresión genética mediante plataformas moleculares; sin embargo, debido a su alto costo no se encuentra disponible (Chachaima-Mar et al., 2021).

El método más utilizado por la mayoría de los médicos es la mamografía, el cual consiste en tomar imágenes con un equipo de rayos X, el mismo que utiliza pequeñas dosis de radiación; sin embargo, la cobertura y la calidad del tamizaje es baja, incluso no puede utilizarse en mujeres lactantes o en pacientes que estén en su periodo menstrual; y lamentablemente este examen, no confirma el resultado de tener cáncer o no, ya que el tejido mamario puede ocultar un cáncer o quiste, entonces existe la necesidad de realizar otros estudios como RMI, biopsia entre otros (Alegría Delgado & Huamani Navarro, 2019)(Villavicencio Romero et al., 2019).

La incidencia, la mortalidad y la prevalencia del cáncer, permiten cuantificar la magnitud de esta patología y orientan las políticas públicas en relación a la prevención y los servicios de salud (Vallejos Sologuren et al., 2020).

La detección temprana del cáncer es necesaria para que se puedan llevar a cabo esfuerzos en su tratamiento, por lo tanto, se requiere una tecnología que permita detectar el cáncer con alta precisión (Naufal et al., 2020).

En el Perú se ha definido el gobierno ha formulado el Plan Nacional para la Prevención y Control del Cáncer de Mama 2017 – 2019, en ella se definen políticas que permita reducir la morbi mortalidad por esta enfermedad, utilizando como elemento de juicio las lesiones BIRADS, que muestra una clasificación de los resultados en categorías numeradas (Ministerio de Salud, 2017).

Machine Learning

Machine Learning, conocido en español como aprendizaje de máquina, es un área dedicada al desarrollo y aplicación de técnicas y algoritmos computacionales capaces de aprender a través de grandes conjuntos de datos (Franco & Ramos, 2019), sin la necesidad de ser programados explícitamente (Aguilar et al., 2018). Se trata del proceso de programar para aprender en lugar de programar para una única salida (Eedi & Kolla, 2020), el método de aprendizaje comienza con datos o un conjunto de datos, como experiencias o instrucciones, para que luego puedan descubrir un patrón o mejorar ese patrón en un futuro cercano, si es necesario (Jean Sunny et al., 2020).

Emplean una variedad de métodos estadísticos, probabilísticos y de optimización para aprender de la experiencia pasada y detectar patrones útiles a partir de conjuntos de datos grandes, no estructurados y complejos (Uddin et al., 2019), asimismo a presentado un constante crecimiento de aplicación en el diagnóstico clínico (Blanc Pihuave et al., 2020).

Los algoritmos de machine learning se pueden dividir en tres categorías amplias de acuerdo con sus propósitos: algoritmos supervisado, algoritmos no supervisado y algoritmos semisupervisado (Uddin et al., 2019).

Los algoritmos supervisados buscan deducir una función a partir de un conjunto de datos (data set) de entrenamiento, estos son pares de objetos que constituyen datos de entrada y los resultados deseados (Rejón Herrera et al., 2021); asimismo, tiene como objetivo aprender a partir de datos etiquetados y luego clasificar nuevos datos en función al conocimiento adquirido (Comarela et al., 2019), este algoritmo puede generalizar la función para predecir lo oculto con precisión (Al-Azzam & Shatnawi, 2021).

K-Nearest Neighbor

K-Nearest Neighbor (KNN) es uno de los algoritmos supervisados de machine learning basado en las similitudes y ofrece en algunos contextos un rendimiento interesante, asimismo, es una generalización para las reglas del vecino más cercano (Cherif, 2018; Xing & Bei, 2020).

KNN es un algoritmo simple, robusto y versátil para clasificación y se ha utilizado para diferentes tipos de problemas como: reconocimiento de patrones, clasificación de modelos, categorización de texto, bioinformática incluso en la medicina entre otros; también es un clasificador no paramétrico y de aprendizaje perezoso; al ser no paramétrico significa que está libre de suposiciones sobre las propiedades de los datos subyacentes por lo que no es necesario tener conocimientos previos sobre los datos; y es de aprendizaje perezoso, porque cualquier generalización de los datos de entrenamiento se pospone hasta que los datos de la prueba se

presentan al sistema (Ehsani & Drabløs, 2020). KNN puede ser usado eficientemente para resolver problemas de clasificación (Arslan & Arslan, 2021)(Rajaguru & Sannasi Chakravarthy, 2019).

KNN está considerado dentro de los 10 mejores algoritmos debido a su simplicidad, efectividad e implementación y se puede aplicar eficazmente en varias tareas de clasificación prácticas del mundo real (Xu et al., 2019).

La idea de este método está en el espacio de características, si la mayoría de las k muestras más cercanas (es decir, los vecinos más cercanos en el espacio de características) a una muestra dada pertenecen a una determinada categoría, esa muestra también pertenecerá a esta categoría (Wu & Hicks, 2021); la clase más cercana se identificará utilizando la distancia euclidiana (Pathanjali et al., 2018); este es el enfoque de predicción también es conocida como la regla de la mayoría (Zhang, 2021); en consecuencia el valor de K juega un papel importante en el desempeño de este algoritmo (Thanh Noi & Kappas, 2017)(Xu et al., 2019).

Medición del desempeño

Para medir el rendimiento de un algoritmo supervisado se utiliza la precisión (precision), sensibilidad (recall), el Puntaje-F1 (F1-Score) y la exactitud (accuracy).

La precisión es la proporción de los ejemplos que realmente pertenecen a la clase X entre todos los que fueron asignados a la clase, la fórmula es: Precisión= VP/(VP+FP), donde VP=Verdadero Positivo y FP=Falso Positivo (Arslan & Arslan, 2021)(Dhahri et al., 2019)(Mercaldo et al., 2017)(Mohana Priya & Punithavalli, 2019)

La sensibilidad es la proporción de ejemplos que se asignaron a la clase X, entre todos los ejemplos que realmente pertenecen a la clase, es decir, qué parte de la clase se capturó, la fórmula es: Sensibilidad = VP/(VP+FN), donde VP=Verdadero Positivo y FN=Falso Negativo (Arslan & Arslan, 2021)(Dhahri et al., 2019)(Mercaldo et al., 2017)

El puntaje-F1 puede interpretarse como un promedio ponderado de la precisión y su fórmula es: Puntaje-F1=2*(Precisión * Sensibilidad) / (Precisión + Sensibilidad) (Arslan & Arslan, 2021)(Dhahri et al., 2019)(Mercaldo et al., 2017)(Mohana Priya & Punithavalli, 2019).

La exactitud es la medida de las predicciones correctas que hizo el clasificador, su fórmula es: Exactitud=(VP+VN)/(VP+VN+FP+FN) (Dhahri et al., 2019)

A continuación, se muestra en la Tabla 1 trabajos realizados para la predicción de cáncer de mamas con machine learning utilizando algoritmos de machine learning:

Tabla 1. Predicción de cáncer de mamas con machine learning

Referencia	Trabajo	Exactitud
(Rajaguru & Sannasi Chakravarthy, 2019)	Clasificación del cáncer de mama con árboles de decisión	91.23%
(Rajaguru & Sannasi Chakravarthy, 2019)	Clasificación del cáncer de mama con KNN	95.61%
(Wu & Hicks, 2021)	KNN	87%
(Wu & Hicks, 2021)	Árboles de decisión	87%

El objetivo de esta investigación fue lograr un nivel aceptable de precisión, sensibilidad, puntaje-F1 y exactitud en la discriminación de masas mamográficas utilizando el algoritmo KNN y los atributos BIRADS.

2. Materiales y métodos

El conjunto de datos (data set) fue obtenido a partir de (Elter, 2007) en el repositorio Irvine Machine Learning Repository de la Universidad de California, con el nombre de Mammographic Mass Data Set, cuenta con 6 atributos y 961 instancias con valores numéricos de los cuales 516 corresponden a casos benignos y 445 a casos malignos; sin embargo, se eliminaros aquellos que presentaban datos vacíos, quedando 829 registros, de los cuales 427 son casos benignos y 402 a casos malignos

En la Tabla 2, se presenta información de los atributos contenidos en el conjunto de datos, el tipo de dato, así como el rango de valores, al finalizar este paso, se creó el archivo: mammographic_masses_sin_datosperdidos.csv

Tabla 2. Información de los atributos del conjunto de datos

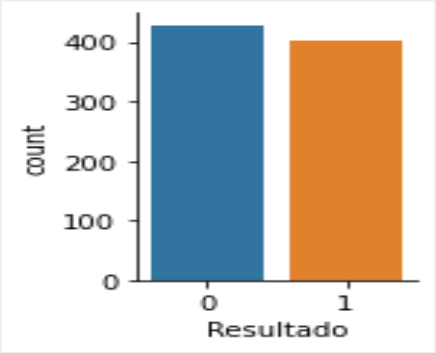
Atributo	Tipo	Rango de Valores
Evaluación BIRADS	Ordinal	[1 , 6]
Edad	Número entero	[1 , Más >
Forma	Nominal	Redonda = 1, Ovalada = 2 Lobular = 3 Irregular = 4
Margen	Nominal	Circunscrito = 1 Microlobulado = 2 Oscurecido = 3 Mal definido = 4 Espiculado = 5
Densidad	Ordinal	Densidad de masa alta = 1 Estándar = 2 Baja = 3 Contiene grasa = 4 (ordinal)
Resultado	Binomial	Benigno = 0 Maligno = 1

Entretenimiento del algoritmo KNN

Para realizar el entrenamiento del algoritmo se utilizó el algoritmo de clasificación KNN y su implementación en Python importando las librerías ScikitLearn, de la siguiente manera:

Primero, se cargó el conjunto de datos a un dataframe utilizando el comando pd.read_csv ('mammographic_masses_sin_datosperdidos.csv',sep=','), en la Fig.1, se muestra la distribución de datos agrupados por el resultado, donde Benigno= 0 y Maligno=1

Figura 1. Distribución de datos por el resultado



Segundo, importar las librerías from sklearn.model_selection import train_test_split; seleccionar aleatoriamente el 70% de los datos para entrenamiento y el 30% restante para las pruebas; tomar como valor inicial n_neighbors=1 y un random_state=30, entrenar el algoritmo o modelo llamado: knn_mm utilizando el método fit(x_train, y_train), hacer predicciones con los datos de prueba utilizando knn_mm.predict() y generar la matriz de confusión inicial como se aprecia en la Tabla 3 con el comando confusion_matrix() y el reporte de clasificación inicial como se observa en la Tabla 4 utilizando el comando classification_report.

En la matriz de confusión de la Tabla 3, se aprecia los valores: Verdadero Positivo (VP), Verdadero Negativo (VN), Falso Positivo (FP) y Falso Negativo (FN).

Tabla 3. Matriz de confusión inicial con n_neighbors=1

		Benigno	***Maligno***
Valores reales	*Benigno*	VN=93	FN=24
	Maligno	FN=34	VP=98
		Predicción	

En el reporte de clasificación inicial con n_neighbors=1 de la Tabla 4 se observa una Precisión(Precision)=73%, sensibilidad(Recall)=79% y un Puntaje-F1(F1-score)=76% para la clase de Benigno, asimismo para la clase de Maligno una Precisión(Precision)=80%, sensibilidad(Recall)=74% y un Puntaje-F1(F1-score)= 77% y finalmente se observa que el nivel de exactitud(Accuracy) del modelo es del 77%.

Tabla 4. Reporte de clasificación inicial con n_neighbors=1

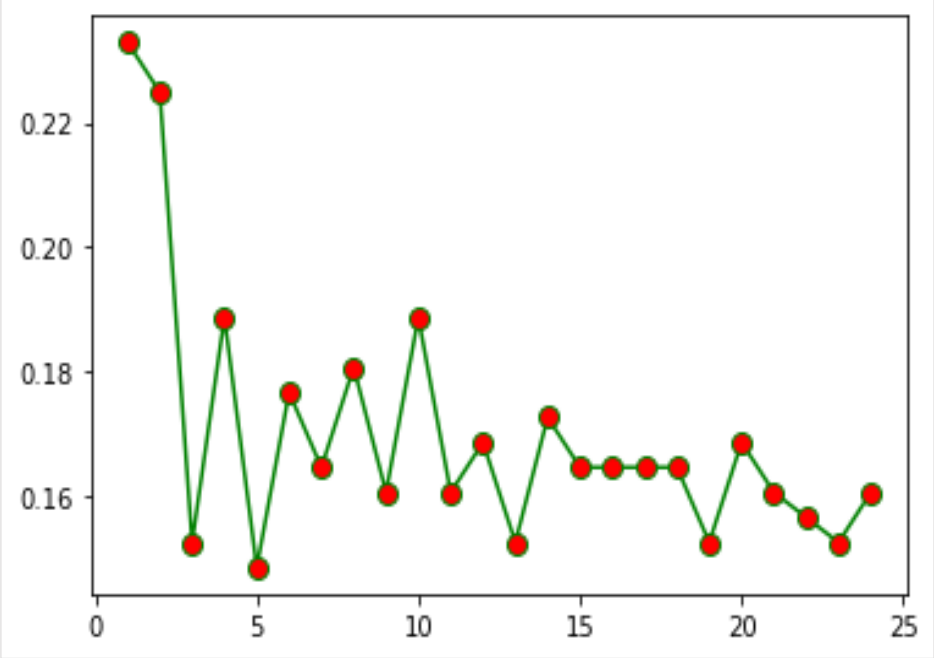
	Precision	**Recall**	**F1-score**	**Support**
Benigno	0.73	0.79	0.76	134
Maligno	0.80	0.74	0.77	115
Accuracy			0.77	249
Macro AVG	0.77	0.77	0.77	249
Weighted AVG	0.77	0.77	0.77	249

Selección del modelo

Para seleccionar el modelo óptimo es necesario buscar el mejor valor para n_neighbors, en ese sentido, se diseñó una estructura repetitiva que muestre un gráfico con las diferentes tasas de

error como se observa en la Figura 2; para este caso, se observa que el mejor valor para n_neighbors = 5.

Figura 2. Reporte de la tasa de error



De esta manera vuelve a generar el modelo knn_mm utilizando el método fit(x_train, y_train)y con el n_neighbors=5, y luego generar la matriz de confusión del modelo óptimo utilizando el comando: confusion_matrix() según se observa en la Tabla 5.

Tabla 5. Matriz de confusión óptimo

		Benigno	***Maligno***
Valores reales	*Benigno*	VN=103	FN=14
	Maligno	FN=23	VP=109
		Predicción	

Utilizando el comando classification_report que ofrece la libraría ScikitLearn, se generó la Tabla 6. Reporte de clasificación óptimo; en ella se observa una Precisión(Precision)=82%, sensibilidad(Recall)=88% y un Puntaje-F1(F1-score)= 85% para la clase de Benigno, asimismo se observa una Precisión(Precision)=89%, sensibilidad(Recall)=83% y un Puntaje-F1(F1-score)= 85% para la clase de Maligno y finalmente se observa que el nivel de exactitud(Accuracy) del modelo knn_mm es del 85%.

Tabla 6. Reporte de Clasificación óptimo

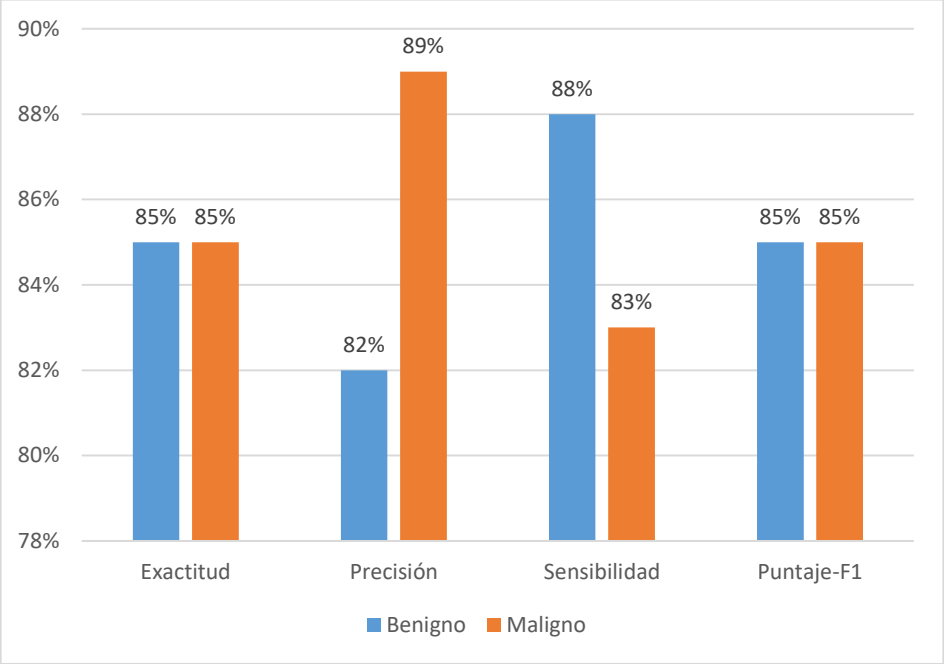
	Precision	**Recall**	**F1-score**	**Support**
Benigno	0.82	0.88	0.85	134
Maligno	0.89	0.83	0.85	115
Accuracy			0.85	249
Macro AVG	0.85	0.85	0.85	249
Weighted AVG	0.85	0.85	0.85	249

3. Resultados

3.1. Desempeño del modelo

El modelo de machine learning llamado KNN_MM, como se observa en la Figura 3 tiene una exactitud=85%; asimismo tiene una precisión=82%, sensibilidad=88% y un puntaje-F1=85% para la clase de Benigno y una precisión=89%, sensibilidad=83% y un puntaje-F1=85% para la clase de Maligno.

Figura 3. Desempeño del modelo KNN_MM



3.2. Predicciones del modelo

En relación a este modelo, como se observa en la Tabla 7, se desarrolló predicciones utilizando para ello el comando knn_mm.predict([[]]):

Tabla 7. Predicciones con el modelo KNN_MM desarrollado

Comando	Resultado
knn_mm.predict([[3,30,2,4,3]])	Benigno
knn_mm.predict([[5,20,2,1,2]])	Benigno
knn_mm.predict([[2,70,1,2,4]])	Maligno
knn_mm.predict([[4,65,1,2,3]])	Benigno

En relación al modelo desarrollado, se muestra en la Tabla 8, la probabilidad de discriminar benigno o maligno utilizando el comando knn_mm.predict_proba([[]]):

Tabla 8. Probabilidad de clasificar benigno o maligno

Comando	Probabilidad de benigno	Probabilidad de maligno
knn_mm.predict_proba([[3,30,2,4,3]])	80%	20%
knn_mm.predict_proba([[5,20,2,1,2]])	100%	0%
knn_mm.predict_proba([[2,70,1,2,4]])	20%	80%
knn_mm.predict_proba([[4,65,1,2,3]])	80%	20%

4. Discusión

El nivel de exactitud=85% del modelo KNN_MM, guarda una estrecha relación con el resultado obtenido por (Wu & Hicks, 2021) quienes obtuvieron un nivel de exactitud=86% y con (Rajaguru & Sannasi Chakravarthy, 2019) quienes obtuvieron un nivel de exactitud=95.61%.

Al tener niveles aceptables de exactitud y al definirse en el Plan Nacional para la Prevención y Control del Cáncer de Mama 2017 – 2019, políticas que permita reducir la morbi mortalidad por esta enfermedad, utilizando como elemento de juicio las lesiones BIRADS. (Ministerio de Salud, 2017); al diseñarse un modelo basado en los atributos BIRADS, esto puede convertirse en una herramienta aprobada para su uso en el diagnóstico de cáncer de mama.

5. Conclusiones

Se logró diseñar un modelo predictivo llamado KNN_MM con niveles de exactitud, precisión, sensibilidad y puntaje-F1 aceptables, con Python, librerías ScikitLearn y el resultado de la evaluación BI-RADS.

El modelo creado, puede ser tomando por personal médico y usado como un elemento de juicio, que sumado a los resultados de la mamografía y su expertis, pueda diagnosticar con mayor precisión si una masa mamográfica representa a peligro de cáncer o es un tumor benigno.

Es útil, revisar la tasa de error, para esta investigación según Figura 2 fue el obtenido cuando el valor de n_neighbors=5, como una alternativa válida o un medio que permita a otros investigadores mejorar sus resultados durante el entrenamiento del algoritmo.

Agradecimiento

Los investigadores brindan su agradecimiento a la Universidad Nacional Tecnológica de Lima Sur, por generar el espacio y tiempo, para la realización de este proyecto.

Referencias bibliográficas

Aguilar, R. M., Torres, J. M., & Martín, C. A. (2018). Aprendizaje Automático en la Identificación de Sistemas. Un Caso de Estudio en la Predicción de la Generación Eléctrica de un Parque Eólico. *Revista Iberoamericana de Automática e Informática Industrial*, *16*(1), 114. https://doi.org/10.4995/riai.2018.9421

Al-Azzam, N., & Shatnawi, I. (2021). Comparing supervised and semi-supervised Machine

Learning Models on Diagnosing Breast Cancer. *Annals of Medicine and Surgery*, 53–64. https://doi.org/10.1016/j.amsu.2020.12.043

Alegría Delgado, D., & Huamani Navarro, M. (2019). Factores asociados a la toma de mamografía en mujeres peruanas: análisis de la Encuesta Demográfica de Salud Familiar, 2015. *Anales de La Facultad de Medicina*, *80*(3), 327–331. https://doi.org/10.15381/anales.803.16204

Arslan, H., & Arslan, H. (2021). A new COVID-19 detection method from human genome sequences using CpG island features and KNN classifier. *Engineering Science and Technology, an International Journal*, *24*(4), 839–847. https://doi.org/10.1016/j.jestch.2020.12.026

Blanc Pihuave, G., Cevallos Torres, L., & Arteaga Vera, J. (2020). Modelo computacional de clasificación de aprendizaje de máquina supervisado, para el análisis de datos cardiovasculares y pronóstico médico. *Ecuadorian Science Journal*, *4*(2), 71–79. https://doi.org/10.46480/esj.4.2.83

Pathanjali, C., Vimuktha, E. S., Jalaja, G., & Latha, A. (2018). A Comparative Study of Indian Food Image Classification Using K-Nearest-Neighbour and Support-Vector-Machines. *International Journal of Engineering & Technology*, 7(3.12), 521. https://doi.org/10.14419/ijet.v7i3.12.16171

Chachaima-Mar, J. E., Pineda-Reyes, J., Marin, R., Lozano-Miranda, Z., & Chian-García, C. (2021). Perfil inmunofenotípico de cáncer de mama de pacientes atendidas en un hospital general de Lima, Perú. *Revista Medica Herediana*, *31*(4), 235–241. https://doi.org/10.20453/rmh.v31i4.3855

Cherif, W. (2018). Optimization of K-NN algorithm by clustering and reliability coefficients: Application to breast-cancer diagnosis. *Procedia Computer Science*, *127*, 293–299. https://doi.org/10.1016/j.procs.2018.01.125

Comarela, G., Franco, G., Trois, C., Liberato, A., Martinello, M., Corrêa, J. H., & Villaça, R. (2019). Introdução à Ciência de Dados: Uma Visão Pragmática utilizando Python, Aplicações e Oportunidades em Redes de Computadores. In *Minicursos do XXXVII Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (2019)*. https://doi.org/10.5753/sbc.6555.9.6

Dhahri, H., Al Maghayreh, E., Mahmood, A., Elkilani, W., & Faisal Nagi, M. (2019). Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms. *Journal of Healthcare Engineering*, *2019*. https://doi.org/10.1155/2019/4253641

Eedi, H., & Kolla, M. (2020). Machine learning approaches for healthcare data analysis. *Journal of Critical Reviews*, *7*(4), 806–811. https://doi.org/10.31838/jcr.07.04.149

Ehsani, R., & Drabløs, F. (2020). Robust Distance Measures for kNN Classification of Cancer Data. *Cancer Informatics*, *19*. https://doi.org/10.1177/1176935120965542

Elter, M. (2007). *Mammographic Mass Data Set*. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml/datasets/mammographic+mass

Franco, E. F., & Ramos, R. J. (2019). Aprendizaje de Máquina y Aprendizaje Profundo en Biotecnología: Aplicaciones, impactos y desafíos. *Ciencia, Ambiente y Clima*, *2*(2), 7–26.

https://doi.org/10.22206/cac.2019.v2i2.pp7-26

García Aranda, M., & Redondo, M. (2019). Immunotherapy: A challenge of breast cancer treatment. *Cancers*, *11*(12), 1–18. https://doi.org/10.3390/cancers11121822

Gardezi, S. J. S., Elazab, A., Lei, B., & Wang, T. (2019). Breast cancer detection and diagnosis using mammographic data: Systematic Review. *Journal of Medical Internet Research*, *21*(7), 1–22. https://doi.org/10.2196/14464

Jean Sunny, Nikita Rane, Rucha Kanade, & Sulochana Devi. (2020). Breast Cancer Classification and Prediction using Machine Learning. *International Journal of Engineering Research And*, *V9*(02), 576–580. https://doi.org/10.17577/ijertv9is020280

Mercaldo, F., Nardone, V., & Santone, A. (2017). Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques. *Procedia Computer Science*, *112*, 2519–2528. https://doi.org/10.1016/j.procs.2017.08.193

Resolución Ministerial N° 442-2017/MINSA[Ministerio de Salud]. P*lan Nacional de Prevención y control de cáncer de mama en el Perú 2017-2021. 01 de enero 2017*

Mohana Priya, T., & Punithavalli, M. (2019). An efficient data mining techniques - Multi-objective KNN algorithm to predict breast cancer. *International Journal of Recent Technology and Engineering*, *8*(8), 986–990. https://doi.org/10.35940/ijrte.B1188.0882S819

Naufal, S. A., Adiwijaya, A., & Astuti, W. (2020). Analisis Perbandingan Klasifikasi Support Vector Machine (SVM) dan K-Nearest Neighbors (KNN) untuk Deteksi Kanker dengan Data Microarray. *JURIKOM (Jurnal Riset Komputer)*, *7*(1), 162. https://doi.org/10.30865/jurikom.v7i1.2014

Rajaguru, H., & Sannasi Chakravarthy, S. R. (2019). Analysis of decision tree and k-nearest neighbor algorithm in the classification of breast cancer. *Asian Pacific Journal of Cancer Prevention*, *20*(12), 3777–3781. https://doi.org/10.31557/APJCP.2019.20.12.3777

Rejón Herrera, E. G., Esparza Sánchez, R., Pasos Ruiz, A., & Moreno Caballero, E. (2021). Clasificación de Indicadores de Interacción del uso de la plataforma Moodle para cursos de modalidad B-learning. *Tecnología Educativa Revista CONAIC*, *2*(1), 78–86. https://doi.org/10.32671/terc.v2i1.170

Thanh Noi, P., & Kappas, M. (2017). Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors (Basel, Switzerland)*, *18*(1). https://doi.org/10.3390/s18010018

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1), 1–16. https://doi.org/10.1186/s12911-019-1004-8

Vallejos Sologuren, C. S., Aguilar cartagena, A., & Flores Flores, C. J. (2020). Situación del Cáncer en el Perú. *Diagnóstico*, *52*(2), 77–85. https://doi.org/10.33734/diagnostico.v59i2.221

Villavicencio Romero, M. E., Moreno Daza, G. A., Ordóñez Andrade, G. E., & Paredes Colcha, L. M. (2019). Diagnóstico por imágenes de cáncer de mamas. Comparación entre técnica ecográfica y mamografía. *Dominio de Las Ciencias*, *5*(3), 647.

https://doi.org/10.23857/dc.v5i3.957

Wu, J., & Hicks, C. (2021). Breast cancer type classification using machine learning. *Journal of Personalized Medicine*, *11*(2), 1–12. https://doi.org/10.3390/jpm11020061

Xing, W., & Bei, Y. (2020). Medical Health Big Data Classification Based on KNN Classification Algorithm. *IEEE Access*, *8*, 28808–28819. https://doi.org/10.1109/ACCESS.2019.2955754

Xu, H., Zhou, J., Asteris, P. G., Armaghani, D. J., & Tahir, M. M. (2019). Supervised machine learning techniques to the prediction of tunnel boring machine penetration rate. *Applied Sciences (Switzerland)*, *9*(18), 1–19. https://doi.org/10.3390/app9183715

Zhang, S. (2021). Challenges in KNN Classification. *IEEE Transactions on Knowledge and Data Engineering*, 1–13. https://doi.org/10.1109/TKDE.2021.3049250

Zielonke, N., Kregting, L. M., Heijnsdijk, E. A. M., Veerus, P., Heinävaara, S., McKee, M., de Kok, I. M. C. M., de Koning, H. J., van Ravesteyn, N. T., Gredinger, G., De Brabander, I., Arbyn, M., Simoens, C., Martens, P., Candeur, M., Arbyn, M., Simoens, C., Burrion, J. B., Dimitrov, P., … Latinovic, R. (2021). The potential of breast cancer screening in Europe. *International Journal of Cancer*, *148*(2), 406–418. https://doi.org/10.1002/ijc.33204

Financiamiento

Ninguno.

Conflicto de intereses

El artículo no presenta conflicto de intereses.

Contribución de autores

Lévano-Rodriguez, Daniel: Investigador y redactor del artículo.

Cerdán-León, Flor Elizabeth: Investigadora y redactora del artículo.