Revista de investigación de sistemas e informática
15(2): 5-12 (2022)
DOI: https://doi.org/10.15381/risi.v15i2.23402

ISSN: 1815-0268 (Impreso) / ISSN: 1816-3823 (Electrónico)
Facultad de Ingeniería de Sistemas e Informática - UNMSM

Investigación sobre el método de predicción del cáncer de mama basado en aprendizaje automático

Research on breast cancer prediction method based on machine learning

Jhelly Reynaluz Pérez Núñez
https://orcid.org/0000-0003-0717-8277
jhelly.perez@unmsm.edu.pe

Universidad Nacional Mayor de San Marcos,
Facultad de Ingeniería de Sistemas e
Informática. Lima, Perú

Recibido: 17/08/2022 - Aceptado: 12/10/2022 - Publicado: 30/12/2022

RESUMEN

Cada año, el número de muertes aumenta extremadamente debido al cáncer de mama. En 2020, la Organización Mundial de la Salud informó que el 25% de las mujeres fueron diagnosticadas con cáncer de mama. En Perú, el cáncer de mama fue la principal causa de muerte por cáncer, donde cada día mueren 5 mujeres por esta enfermedad. La detección temprana del cáncer de mama es facilitada por computadora tecnologías de detección y diagnóstico (CAD), que pueden ayudar a las personas a vivir vidas más largas. Este artículo busca analizar múltiples modelos de aprendizaje automático para identificar el cáncer de mama. Para esto, se trabaja con dos bases de datos de Wisconsin donados por el Dr. William H. Wolberg, el primer conjunto de base de datos es del año 1995 y el segundo es del año 1992 ambos informes es en base a sus casos clínicos. El principal objetivo de este trabajo es aprovechar los últimos desarrollos en sistemas CAD y metodologías relacionadas para las predicciones, los modelos de Decisión Tree, K-vecinos más cercanos (K-NN), Naive Bayes (NB), Regresión logística, Support Vector Machine, Multi-layer Perceptrón y Random Forest se utilizaron. Cuando se comparan los resultados, se encuentra que, para la base de datos del año 1995, el algoritmo Multi-layer Perceptron ofrece el mejor resultado, logrando una precisión de 97,20% y para la segunda base de datos de 1992, el algoritmo que mejor predice es el Support Vector Machine con una precisión de 97,20%.

Palabras clave: decision tree, K-Nearest neighbors, naïve bayes, random forest, multi-layer perceptron and logistic regression.

ABSTRACT

Every year, the number of deaths increases extremely due to breast cancer. In 2020, the World Health Organization reported that 25% of women were diagnosed with breast cancer. In Peru, breast cancer was the main cause of death from cancer, where 5 women die from this disease every day. Early detection of breast cancer is facilitated by computerized detection and diagnostic (CAD) technologies, which can help people live longer lives. This article seeks to analyze multiple machine learning models to identify breast cancer. For this, we work with two Wisconsin databases donated by Dr. William H. Wolberg, the first set of databases is from 1995 and the second is from 1992, both reports are based on their clinical cases. The main objective of this work is to take advantage of the latest developments in CAD systems and related methodologies for predictions, Decision Tree models, K-Nearest Neighbors (K-NN), Naive Bayes (NB), Logistic Regression, Support Vector Machine, Multi-layer Perceptron and Random Forest were used. When the results are compared, it is found that for the database of the year 1995, the Multi-layer Perceptron algorithm offers the best results, achieving an accuracy of 97.20% and for the second database of 1992, the algorithm that best predicts is the Support Vector Machine with an accuracy of 97.20%.

Keywords: decision tree, K-Nearest neighbors, naïve bayes, random forest, multi-layer perceptron and logistic regression.

I. INTRODUCCIÓN

El número de casos de Cáncer de Mama en el Perú es muy alto, durante el 2020, 5 mujeres fallecen cada día por esta enfermedad. Según el informe de la Organización Mundial de la Salud, Latinoamérica ocupa el tercer lugar como causa de muerte. La OMS señaló que el año pasado se diagnosticaron 6; 860 casos de este tipo de cáncer, que representaron el 9, 8% del total de casos de cáncer detectados en el país y fueron la causa principal de muerte relacionadas con esa enfermedad. En promedio, el 95% de casos de cáncer de mama detectados a tiempo se curan. Sin embargo, en nuestro país el 85% de casos son detectados en estados avanzados debido a que las personas se no se han sus chequeos anuales. [1]

El Cáncer de mama ocupa el primer lugar con mayor número de muertes para el género femenino. El pronóstico anticipado de un tipo de cáncer se ha convertido en una necesidad de investigación ya que puede facilitar el tratamiento preventivo para evitar su letalidad en un estado avanzado. Esto refuerza la necesidad invertir tanto en la lucha contra el cáncer como en la prevención del cáncer. La introducción exitosa de información y las tecnologías de la comunicación (TIC) en la práctica médica es una apuesta importante en la renovación del sistema de salud y más precisamente en la atención del cáncer. En 2018 se diagnosticaron más de 2; 1 millones de casos nuevos, y una de cada ocho mujeres será diagnosticada con cáncer de mama invasivo a lo largo de su vida [2]. Además, se han desarrollado varias técnicas de imagen y se utilizan ampliamente para detectar el cáncer de mama temprano. La mamografía, la ecografía mamaria, la resonancia magnética nuclear (RMN), la tomografía por emisión de positrones (PET) y la tomografía computarizada (TC) son algunas de las técnicas de imagen [3], [4], [5]. La ecografía mamaria se puede dividir en categorías diagnósticas y terapéuticas. La ecografía diagnóstica se consideraba no invasiva y la ecografía terapéutica no producía imágenes. En realidad, Big Data ha revolucionado el tamaño de los datos y también ha creado valor a partir de esto Big data ha hecho un gran cambio en Business Intelligence al analizar una gran cantidad de datos no estructurados, heterogéneos, no estándar y datos sanitarios incompletos. No solo pronostica, sino que también ayuda en la toma de decisiones y se nota cada vez más como gran avance en el avance continuo con el objetivo de mejorar la calidad de la atención al paciente y reducir los costos sanitario.

Para el año 2019 en el ámbito tecnológico y de investigación el aprendizaje automático o Machine Learning (ML) que se encuentra dentro de la rama de inteligencia artificial proporciona herramientas y métodos que permiten analizar una gran cantidad de datos y así poder llegar a un resultado eficiente. Estas técnicas han sido utilizadas en diferentes investigaciones para modelar el tratamiento de afecciones cancerosas debido a su capacidad para detectar características significativas en conjuntos de datos complejos. [6]

Los algoritmos de minería de datos aplicados en la industria de la salud juegan un papel importante debido a su alto desempeño en la predicción, diagnóstico de enfermedades, reducción de costos de medicamentos, toma de decisiones en tiempo real para ahorrar la vida de la gente. Los objetivos de modelado de minería de datos más comunes son la clasificación y la predicción, que utiliza varios Algoritmos para la predicción del cáncer de mama. Este documento proporciona principalmente una comparación entre el rendimiento de cuatro clasificadores: Decision Tree models, K-Nearest Neighbors (K-NN), Naive Bayes (NB), Logistic Regression, Support Vector Machine, Multi-layer Perceptron and Random Forest que según la comunidad de investigación se encuentran entre los algoritmos de minería datos más influyentes. [3]

Nuestro objetivo es predecir y diagnosticar mama cáncer, utilizando algoritmos de aprendizaje automático, y encontrar los más efectivo en función del rendimiento de cada clasificador en términos de la matriz de precisión. El resto de este documento está organizado como sigue. La sección 2 presenta los métodos de investigaciones sobre el diagnóstico del cáncer de mama. Sección 3 describe la metodología propuesta para este trabajo. La sección 4 presenta y explica los resultados de los experimentos y la sección 5 concluye el documento.

II. ALGORITMOS DE APRENDIZAJE AUTOMATICO

El aprendizaje esta diseñados para aprender de conjuntos de datos anteriores [4], ingresamos una gran cantidad de datos, el modelo de aprendizaje automático analiza eso datos y sobre la base de ese modelo de entrenamiento podemos hacer una predicción sobre el futuro [5], [6], [7]. Para las predicciones de cáncer de mama, los principales algoritmos de aprendizaje automático son los siguientes:

A. K-Nearest Neighbor (KNN): Se le conoce como el K-Vecino más Cercano, este algoritmo se utiliza en el reconocimiento de patrones. Es

un buen enfoque para la predicción del cáncer de mama. Para reconoce el patrón, a cada clase se le ha dado una importancia igual. K más cercano Neighbor [8] extrae los datos destacados similares de un gran conjunto de datos. Sobre la base de la similitud de las características clasificamos un gran conjunto de datos [9].

B. Decisión Tree (DT): Un árbol de decisión es una especie de mapa en que se muestra cada una de las opciones de decisión posibles y sus resultados en otras palabras es un diagrama en forma de árbol que muestra la probabilidad estadística o determina un curso de acción [10]. Muestra a los analistas y, a los que toman las decisiones, qué pasos deben tomar y cómo las diferentes elecciones podrían afectar todo el proceso. Todo ello soportado en datos. Es una herramienta muy ´util en cualquier organización regida por los datos o Data Driven. [11] [12].

C. Algoritmo Naive Bayes (NB): Este modelo se utiliza para hacer una suposición de grandes entrenamientos de conjunto de datos. El algoritmo se utiliza para calcular la probabilidad a través del método bayesiano [13]. Proporciona la mayor precisión al calcular las probabilidades de datos ruidosos que se utilizan como entrada. Es un clasificador de analogía que se usa para comparar un conjunto de datos de entrenamiento con una tupla de entrenamiento [14].

D. Regresión Logística: La regresión logística es un método estadístico que trata de modelar la probabilidad de una variable cualitativa binaria (dos posibles valores) en función de una o más variables independientes. La principal aplicación de la regresión logística es la creación de modelos de clasificación binaria. Se llama regresión logística siempre y cuando solo hay una variable independiente y regresión logística múltiple cuando hay más de una. Dependiendo del contexto, a la variable modelada se le conoce como variable dependiente o variable respuesta, y a las variables independientes como regresores, predictores o features. [11]

E. Perceptrón Multicapa (MLP): Los Perceptrones multicapa fue desarrollado para hacer frente a esta limitación. Es una red neuronal donde el mapeo entre entradas y salidas no es lineal. Un perceptrón multicapa tiene capas de entrada y salida, y una o más capas ocultas con muchas neuronas apiladas juntas. Y mientras que en el Perceptrón la neurona debe tener una función de activación que imponga un umbral, como ReLU o sigmoides, las neuronas en un Perceptrón multicapa pueden usar cualquier función de activación arbitraria.

F. Random Forest (RF): Un Random Forest es un conjunto (ensemble) de árboles de decisión combinados con bagging. Al usar bagging, lo que en realidad está pasando, es que distintos árboles ven distintas porciones de los datos. Ningún árbol ve todos los datos de entrenamiento. Esto hace que cada árbol se entrene con distintas muestras de datos para un mismo problema. De esta forma, al combinar sus resultados, unos errores se compensan con otros y tenemos una predicción que generaliza mejor.

G. Support Vector Machine (SVM): SVM funciona correlacionando datos a un espacio de características de grandes dimensiones de forma que los puntos de datos se puedan categorizar, incluso si los datos no se puedan separar linealmente de otro modo. Se detecta un separador entre las categorías y los datos se transforman de forma que el separador se puede extraer como un hiperplano. Tras ello, las características de los nuevos datos se pueden utilizar para predecir el grupo al que pertenece el nuevo registro.

III. METODOLOGÍA

El principal objetivo de este trabajo es identificar el algoritmo efectivo y predictivo para la detección del cáncer de mama, por lo que aplicamos clasificadores de aprendizaje automático como: Decision Tree models, K-Nearest Neighbors (K-NN), Naive Bayes (NB), Logistic Regression, Support Vector Machine, Multi-layer Perceptron and Random Forest sobre el conjunto de datos de cáncer de mama de Madison de los hospitales de la Universidad de Wisconsin del año 1995 y 1992 y así evaluar los resultados obtenidos para definir qué modelo proporciona una mayor precisión. La metodología se divide en cuatro pasos: Selección de la base de datos que han sido descargados desde los repositorios de Kaggle y UC Irvine, son dos Data Sets a analizar, seguida del preprocesamiento, que consta de cuatro pasos como: limpieza de datos, selección de atributos, establecimiento de roles objetivo y extracción de características.

Los datos preparados se utilizan para construir algoritmos de aprendizaje automático que pueden predecir el cáncer de mama para un nuevo conjunto de mediciones. Para evaluar el rendimiento de los algoritmos, mostramos al modelo nuevos datos para los que tenemos etiquetas. Esto generalmente se hace dividiendo los datos etiquetados que hemos

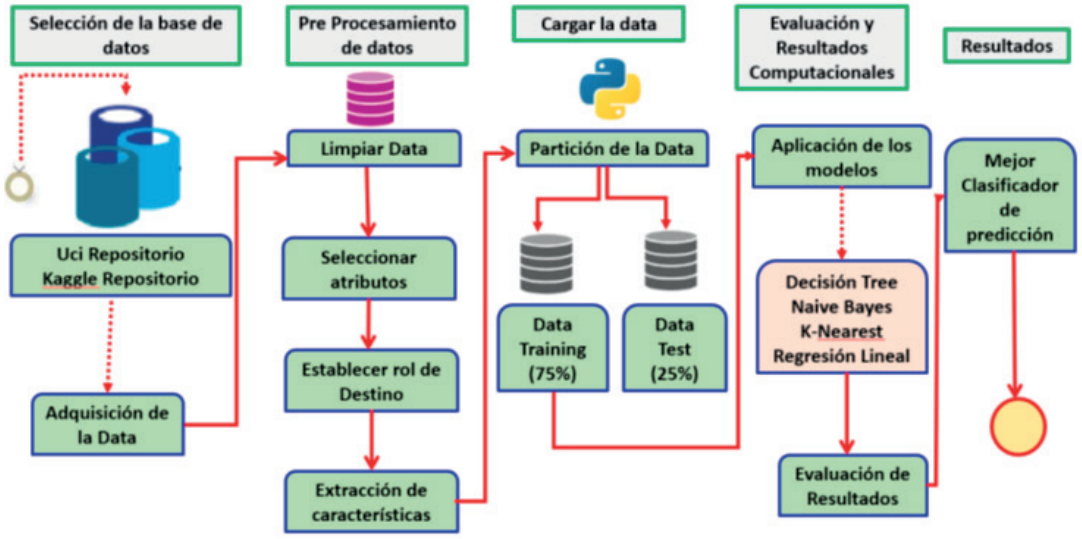
recopilado en dos particiones. El 75% de los datos se utiliza para construir nuestro modelo de aprendizaje automático y se denomina datos de entrenamiento (Data Training) o conjunto de entrenamiento. El 25% de los datos se utilizará para acceder a qué tan bien funciona el modelo y se denomina datos de prueba (Data Test) o conjunto de prueba. Después de probar los modelos, comparamos los resultados obtenidos para seleccionar el algoritmo que proporciona la más alta precisión (Accuracy) e identificamos el algoritmo más predictivo para la detección del cáncer de mama (ver Figura 1).

IV. ADQUISICIÓN DEL DATA SET

El conjunto de datos de este artículo proviene del sitio web de conjunto de datos de prueba estándar de código abierto Kaggle y UCI. En nuestro estudio, utilizamos el conjunto de datos de diagnóstico de cáncer de mama de Wisconsin de la base de datos de cáncer de mama de Madison de los hospitales de la Universidad de Wisconsin [17]. Las características del conjunto de datos se calculan a partir de una imagen mama obtenida por aspiración con aguja fina (FNA). Las características de los núcleos celulares presentes en la imagen se determinan a partir de estas características. El diagnóstico del Cáncer de mama Wisconsin de la data del año 1995 tiene 569 muestras (benigno: 357 y maligno: 212) y 32 características. El Data Set del año 1992 tiene 685 muestras (benigno: 444 y maligno: 239) y 11 características. [18] (ver Figura 2).

Figura 1

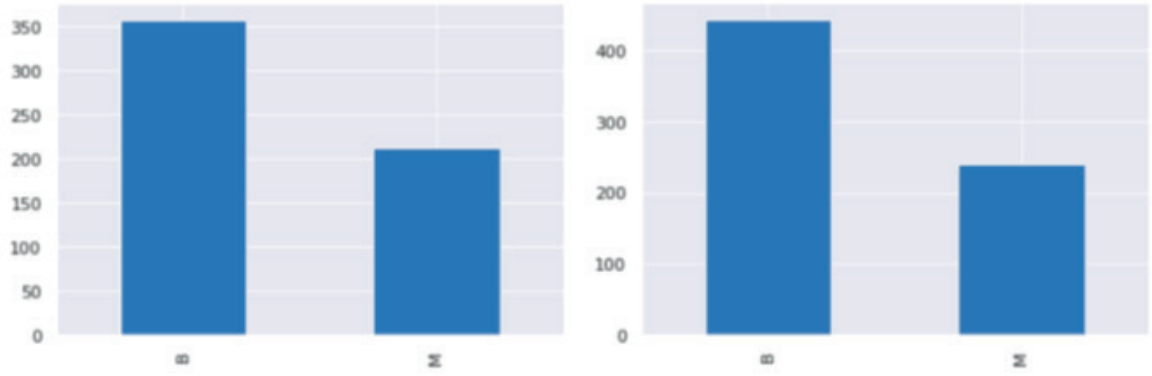
Detalle de la arquitectura usada para los diferentes modelos de predicción.



Fuente. Elaboración propia

Figura 2

Conjunto de datos de diagnóstico de cáncer de mama.



Fuente. Elaboración propia

V. RESULTADO Y DISCUSIÓN

En esta sección colocaremos los pros y los con tras de los modelos de clasificación, es necesario proporcionar métricas para evaluar el rendimiento de los modelos. Aquí dividimos la muestra en cuatro clases: Verdadero positivo

(VP), falso positivo (FP), verdadero negativo (VN) y falso negativo (FN). Sean $TP + FP + TN + FN = n$, donde n es el tamaño de la muestra y la matriz de confusión de los modelos realizados se muestran a continuación.

Tabla 1

Matriz de confusión

	Predicciones	
	Positivo	**Negativo**
Positivo	Verdadero Positivo (VP)	Falso Positivo (FP)
Negativo	Falso Negativo (FN)	Verdadero Negativo (VN)

Fuente. Elaboración propia

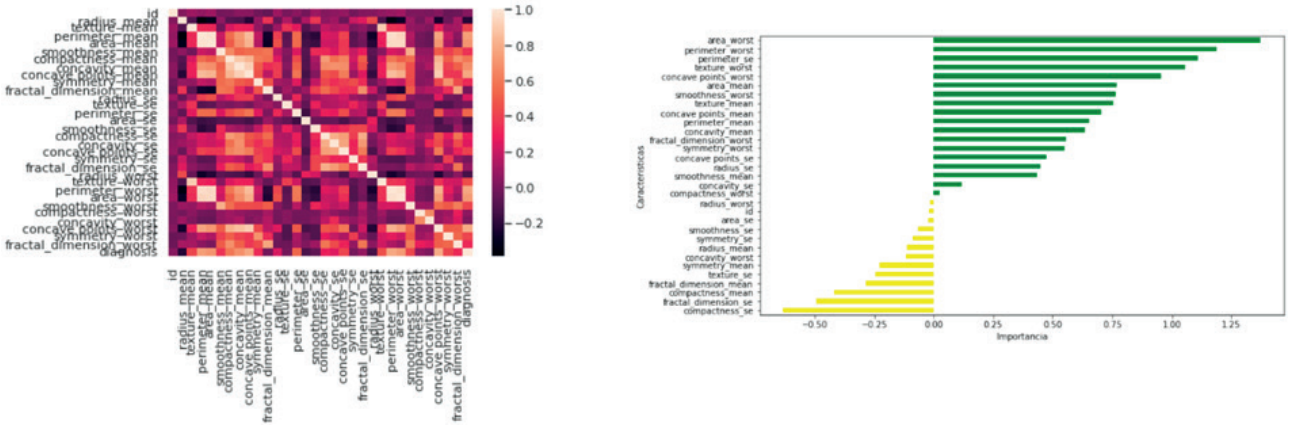
En Machine Learning, la medición del rendimiento es una tarea esencial. Entonces, cuando se trata de un problema de clasificación, podemos contar con una curva AUC - ROC. Cuando necesitamos comprobar o visualizar el rendimiento del problema de clasificación multiclase, utilizamos la curva AUC (Área bajo la curva) y ROC (Características operativa del receptor). Es una de las métricas de evaluación más importantes para comprobar el rendimiento de cualquier modelo de clasificación. También se escribe como AUROC (Área bajo las características operativas del receptor). La curva AUC - ROC es una medida de rendimiento para los problemas de clasificación en varias configuraciones de umbral. ROC es una curva de probabilidad y AUC representa el grado o medida de separabilidad. Indica cuánto es capaz el modelo de distinguir entre clases. Cuanto mayor sea el AUC, mejor será el modelo para predecir 0 clases como 0 y 1 clases como 1. Por analogía, cuanto mayor sea el AUC, mejor será el modelo para distinguir entre pacientes con la enfermedad y sin enfermedad. La

curva ROC se traza con VP (Verdaderos Positivos) contra FP (Falsos Positivos), donde VP está en el eje y y FP está en el eje x. En la base de datos de 1995 se tiene que el modelo Multi-layer Perceptron tiene mejor accuracy con un porcentaje de 97,20%, seguido del modelo Random Forest con un porcentaje de 96,50%. Mientras que la Data del año 1992 se tiene que el modelo Support Vector Machine tiene mejor predicción con 97,20% seguido de Naive Bayes con un porcentaje de 96,49%. La máxima correlación de la variable Diagnosis con sus atributos la tiene concave-points-worst y su correlación es de 0,793566. Mientras que, en relación con la importancia de los atributos con la variable clase Diagnosis en la base de datos de Wisconsin mediante el uso de barras en el Modelo de Regresión Logística, se tiene que los atributos area-Worst y Perimeter-Worst son los más importantes, pues influyen de manera positiva en el desarrollo del cáncer de mama, mientras que las características compactness-se y fractal-dimensión se influyen de manera negativa (ver Figura 3).

De la data de 1992 que proporciona la universidad de Wisconsin se tiene que la máxima correlación de la variable con su atributo lo tiene Uniformity-of-Cell Shape seguida de Uniformity-of-Cell-Size con una correlación de 0,821891 y 0,820801 respectivamente. En cuanto a que atributo es más importante

Figura 3

Matriz de correlación y relación del atributo más importante para la Data 1 (1995).



Fuente. Elaboración propia

mediante el uso de barras en el método de regresión lineal se tiene que son: Clump-Thickness y Bare-Nuclei ellos influyen positivamente al desarrollo del cáncer y Uniformity-of-Cell-Size influye de manera negativa (ver Figura 4).

En la siguiente figura podemos observar la comparación de las curvas ROC (Receiver Operating Characteristic) y AUC (área bajo la curva ROC) de los siete modelos aplicados a las bases de Datos 1 y 2: K-Nearest Neighbors (KNN), Decision Tree, Naive Bayes, Multilayer Perceptron (RPropMLP), Support Vector Machine (SVM), Random Forest y Logistic Regression (ver Figura 5).

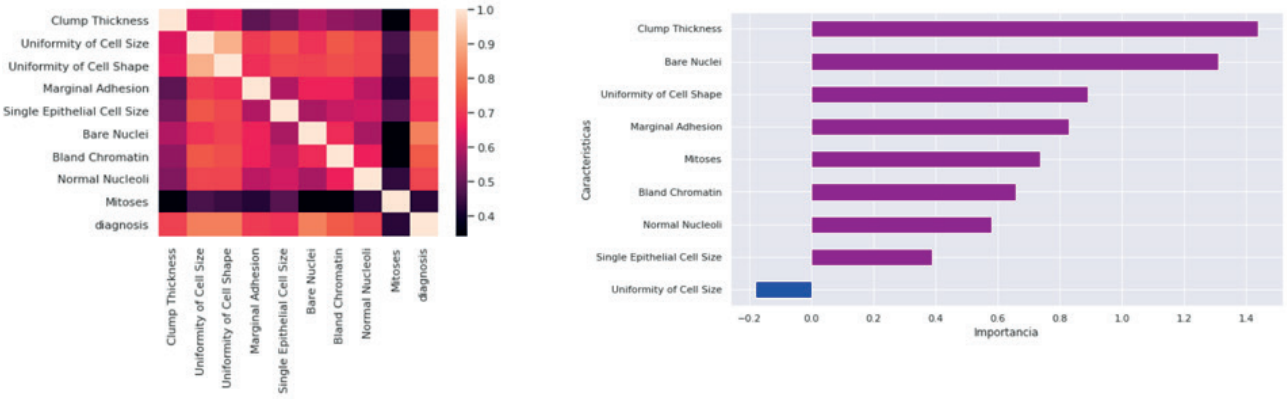
Con respecto a las curvas ROC aplicado a la base de datos Data 1 para medir la eficacia de la predicción del modelo, el que mejor se comporta es el Modelo Random Forest y Multi-layer perceptrón con un AUC (Área bajo la curva) de 1 el cual nos da una idea de que el modelo funciona bien en diferentes particiones. La siguiente curva ROC mejor comportada es la del modelo Decision Tree con un área bajo la curca (AUC) de 0,94 (ver Figura 6).

VI. CONCLUSIONES

En el conjunto de datos de diagnóstico de cáncer de mama de Wisconsin (WBCD) del año 1995 y los proporcionados por el Dr.Wolberg del año 1992, aplicamos 7 algoritmos principales que son: Decisión Tree, K-vecinos más cercanos (K-NN), Naive Bayes (NB), Regresión logística, Support Vector Machine, Multi-layer Perceptrón y Random Forest para calcular, comparar y evaluar diferentes resultados obtenido en base a matriz de confusión para identificar el mejor aprendizaje automático de los

Figura 4

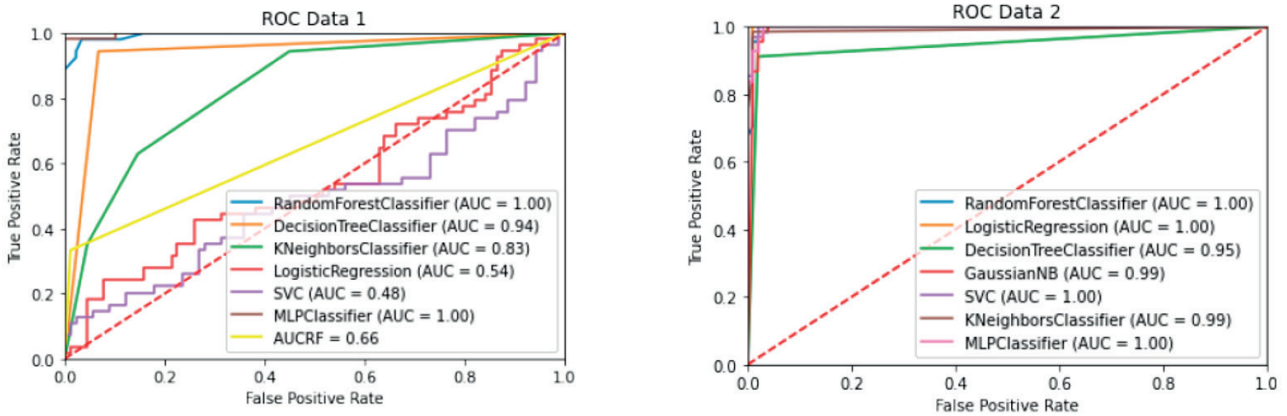
Matriz de correlación y relación del atributo más importante para la Data 2 (1992).



Fuente. Elaboración propia

Figura 5

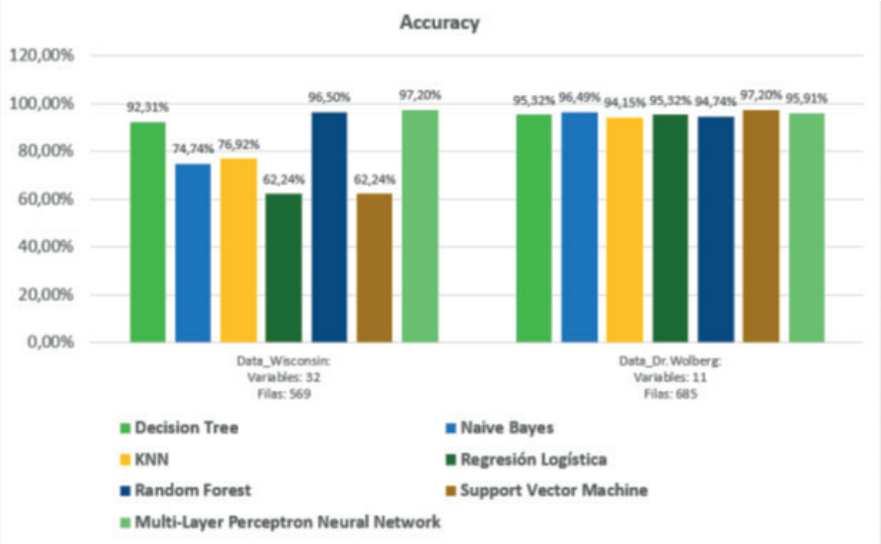
Comparación de los Accuracy de los diferentes modelos realizados para la Data 1 y 2.



Fuente. Elaboración propia

Figura 6

Diagrama de barras para comparar los Accuracy de la Data 1 y 2.



Fuente. Elaboración propia

algoritmos y que tan precisos, confiables y sobre todo que precisión lograr obtener. Encontramos que Multi-layer Perceptron Neural Network logró una mayor precisión del 97,20% para la data 1 (1995) y para la data 2 (1992) el modelo que predice mejor es el Support Vector Machine 97,20% demostrado su eficacia, predicción para el diagnóstico de Cáncer de Mama. Además, si comparamos el nivel de precisión con cada uno de los siete modelos aplicados en las cuatro bases de datos, en promedio el mejor predictor es el modelo SVM, seguidos de los modelos: Random Forest y Decisión Tree.

REFERENCES BIBLIOGRAFICAS

[1] Revilla, T. Situación del cancer en Perú. 2021.

[2] Xindong, Wu. Vipin Kumar, J. Ross, Quinlan. Joydeep, Ghosh. Top 10 Data Mining Algorithms 2015. Knowl Inf Syst. 14, PP.1–3.

[3] Mayer, Z. El cáncer como problema de salud pública en el Perú. Revista Peruana de Medicina Experimental y Salud Publica 2013,30(1).

[4] Gomez,J; Fajardo,J.Modelo En Machine Learning Para El Diagnostico Del Cáncer De Mama.

[5] L. Tuggener, M. Amirian, K. Rombach, S. Lorwald, A. Varlet, C. Westermann, and T. Stadelmann. Automated machine learning in practice: State of the art and recent result. the 6th Swiss Conference on Data Science 2019, pp. 31-36.

[6] A, Dey. Machine learning algorithms: A review 2016. Int. J. Comput. Sci.Inf. Technol. 7(03), pp. 1174-1179.

[7] M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, and S. K. Dhillon.Predicting factors for survival of breast cancer patients using machine learning techniques 2019. BMC Med. Inform. Decis. Making. 19(01).

[8] S. D. Borde and K. R. Joshi. Enhanced signal detection algorithm using trained neural network for cognitive radio receiver 2019. J. Electr. Comput. Eng. 9(1), pp. 323.

[9] S. B. Imandoust and M. Bolandraftar. Application of k-nearest neighbor (KNN) approach for predicting economic events: Theoretical background 2013. 3, pp. 605-610.

[10] H. Sharma and S. Kumar.A survey on decision tree algorithms of classification in data mining 2016.Int. J. Sci. Res. 5(4), pp. 2094-2097.

[11] H. Tran. A survey of machine learning and data mining techniques used in multimedia system 2019. Dept. Comput. Sci., Univ. Texas Dallas Richardson, Richardson, TX, USA, Tech. Rep.

[12] Y.-Y. Song and L. Ying. Decision tree methods: Applications for classification and prediction 2015.Shanghai Arch. Psychiatry. 27(2), pp.130.

[13] W. Wu, S. Nagarajan, and Z. Bayesian machine learning: EEGMEG signal processing measurements 2015.IEEE Signal Process. Mag. 33(1), pp. 14-36.

[14] A. A. Ibrahim, A. I. Hashad, and N. E. M. Shawky. A comparison of open-source data mining tools for breast cancer classification 2015. Handbook of Research on Machine Learning Innovations and Trends. Hershey, PA, USA: IGI Global 2017. pp. 636-651.

[15] T. Evgeniou and M. Pontil. Shawky. Support vector machines: Theory and applications 2005. Advanced Course on Artificial Intelligence. Berlin, Germany: Springer. pp. 249-257.

[16] Y. Yang, J. Li, and Y. Yang. The research of the fast SVM classier method 2015. Proc. 12th Int. Comput. Conf. Wavelet Act. Media Technol. Inf. Process. (ICCWAMTIP). pp. 121-124.

[17] "UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set."

[18] "Kaggle Predicción del cáncer de mama (Diagnostic) Data Set".

Fuentes de financiamiento:

Propia.

Conflictos de interés:

El autor declara no tener conflictos de interés.