Prediction of breast cancer through biomarkers using machine learning

Andrea Gutiérrez Quintanilla, Bach1, Nicole Mancilla Medina, Bach[1], and Jose Sulla-Torres, Dr[1]
[1]Universidad Católica de Santa María, Arequipa, Perú,
andrea.gutierrez@ucsm.edu.pe, 73219000@ucsm.edu.pe, jsullato@ucsm.edu.pe

Abstract– ***The prediction of breast cancer through biomarkers is proposed through machine learning, in order to minimize the waiting time that exists at the time a cancer is discarded due to the different factors that exist in our national reality. For this, a Neural Network has been used, which is an Automatic Learning algorithm that allowed us to make the prediction. The results showed that with the developed design of the Neural Network an accuracy of 82.76% was obtained, likewise, a prototype was built that allowed validating the proposal, with which it can be concluded that the Neural Network is an adequate algorithm to be used for Complementary to the prediction of breast cancer through biomarkers and that the developed prototype serves those interested in the oncology field.***

Keywords: Prediction, Cancer, Neural Network, Machine Learning, Prototype

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2020.1.1.514
ISBN: 978-958-52071-4-1 ISSN: 2414-6390

Predicción de cáncer de mama a través de biomarcadores mediante aprendizaje automático

Andrea Gutiérrez Quintanilla, Bach[1], Nicole Mancilla Medina, Bach[1], and Jose Sulla-Torres, Dr[1]
[1]Universidad Católica de Santa María, Arequipa, Perú,
andrea.gutierrez@ucsm.edu.pe, 73219000@ucsm.edu.pe, jsullato@ucsm.edu.pe

***Abstract*– The prediction of breast cancer is proposed through biomarkers through machine learning, in order to minimize the waiting time that exists when a cancer is discarded due to the different factors that exist in our national reality. For this, a Neural Network has been used, which is an Automatic Learning algorithm that allowed us to make the prediction. The results showed that with the developed design of the Neural Network an accuracy of 82.76% was obtained, likewise, a prototype was built that allowed validating the proposal, with which it can be concluded that the Neural Network is an adequate algorithm to be used Complementary to the prediction of breast cancer through biomarkers and that the developed prototype serves those interested in the oncology field.**

Keywords — Prediction, Cancer, Neural Network, Machine Learning, Prototype.

***Resumen*– Se propone la predicción de cáncer de mama a través de biomarcadores mediante aprendizaje automático, a fin de poder minimizar el tiempo de espera que existe al momento que se solicita un descarte de cáncer por los diferentes factores que en nuestra realidad nacional existe. Para ello se ha utilizado una Red Neuronal que es un algoritmo de Aprendizaje Automático que nos permitió realizar la predicción. Los resultados mostraron que, con el diseño desarrollado de la Red Neuronal se obtuvo una precisión del 82.76%, asimismo, se construyó un prototipo que permitió validar la propuesta, con lo que se puede concluir que la Red Neuronal es un algoritmo adecuado para ser usado de manera complementaria a la predicción de cáncer de mama a través de biomarcadores y que el prototipo desarrollado sirva a los interesados en el campo oncológico.**

Palabras Clave— Predicción, Cáncer, Red Neuronal, Aprendizaje Automático, Prototipo.

I. INTRODUCCIÓN

El cáncer de mama es una enfermedad oncológica que más afecta a las mujeres en el mundo. Es un problema de salud pública en los países donde los sistemas de salud no tienen organizados programas de prevención y no ofrecen alternativas terapéuticas; en tal escenario la mortalidad por cáncer de mama se eleva dramáticamente.

La base de datos GLOBOCAN 2018, como parte del Observatorio Mundial del Cáncer IARC, proporciona estimaciones de incidencia y mortalidad en 185 países para 36 tipos de cáncer y para todos los sitios de cáncer combinados. Se estima que la carga mundial del cáncer aumentó a 18,1 millones de casos nuevos y 9,6 millones de muertes en 2018. [1].

Según el Boletín epidemiológico del Perú durante el período comprendido entre enero y diciembre del año 2018, se registraron 11836 casos de cáncer; de los cuales, 7627 correspondieron a casos nuevos (64,4 %), los casos procedían de 46 locales de salud [2]. La incidencia anual estimada de cáncer de mama en el Perú es de 28 casos por 100,000 habitantes, la tasa de mortalidad anual es de 8.5 casos por 100, 000 habitantes. Esta neoplasia es la segunda más frecuente en mujeres, con un 16,2% y es la tercera causa de mortalidad por cáncer (8,7%) [3].

Diversas formas de cáncer de mama son detectables tempranamente con una mamografía. Actualmente, la mamografía es el examen recomendado para la detección temprana de cáncer de mama. De acuerdo con los resultados de estudios, de las mujeres de 40 a 59 años, el 15,9% se han realizado el examen de mamografía en los últimos 24 meses, a fin de detectar anormalidades que puedan indicar un cáncer de mama, sin mayor cambio respecto del 2015 (15,7%).

El tratamiento del cáncer de mama ha experimentado varios cambios en las últimas décadas debido al descubrimiento de biomarcadores pronósticos y predictivos específicos que permiten la aplicación de terapias más individualizadas a diferentes subgrupos moleculares [4].

Por otra parte, se han realizado estudios sobre conjuntos de datos y técnicas de aprendizaje automático, precisiones de predicción y estadísticas de frecuencia de uso relacionados a la predicción del cáncer de mama. En función de los resultados obtenidos en los conjuntos de datos disponibles al público, se puede usar como información complementaria en la detección del cáncer de mama [5].

En este contexto, es que se presenta esta propuesta de predicción de cáncer de mama a través de biomarcadores utilizando un algoritmo de aprendizaje automático como las Redes Neuronales, además de un prototipo que será un producto novedoso, ya que por sus características es único en el mercado local y nacional; lo cual beneficiará a los especialistas en el área.

El documento está organizado de la siguiente manera: Se ha presentado la Introducción en el Capítulo I, en el Capítulo II se presenta el Estado del Arte, luego en el Capítulo III se muestra los Materiales y Métodos utilizados en el trabajo, a continuación, en el Capítulo IV se presentan las Pruebas y Resultados obtenidos y finalmente se presenta las conclusiones a las se ha llegado en la investigación.

II. ESTADO DEL ARTE

El cáncer de mama es una enfermedad oncológica que afecta a millones de mujeres, por ese motivo se han realizado

Digital Object Identifier (DOI):
http://dx.doi.org/10.18687/LACCEI2020.1.1.514
ISBN: 978-958-52071-4-1 ISSN: 2414-6390

estudios para la implementación de unidades de mamo vigilancia, para la evaluación y seguimiento de las pacientes con sospecha mamográfica de cáncer en diferentes partes del mundo [6] [7].

Los estudios en los cuales se realizaron la intervención del autoexamen de mama de manera regular demostraron ser efectivos para favorecer el diagnóstico de cáncer de mama. En países de bajos y medianos ingresos, se deben implementar intervenciones educativas para que las mujeres se adhieran a realizarse el autoexamen de mama de manera regular. No obstante, no debe ser la única estrategia de prevención del cáncer de mama, sino que debe estar integrada con mamografía usada de manera racional y acceso a tratamiento oportuno.

En Perú, de acuerdo con la región natural de residencia, se encontró que en Lima Metropolitana el 29,1% de las mujeres reportaron que un médico u otro profesional de la salud les habían realizado un examen físico de mama, en los últimos 12 meses. En mujeres de la Sierra fue el 14,6% y en la Selva 14,0% [8] [9].

En la Fig. 1 se muestra las cifras de cáncer de mama según zona demográfica

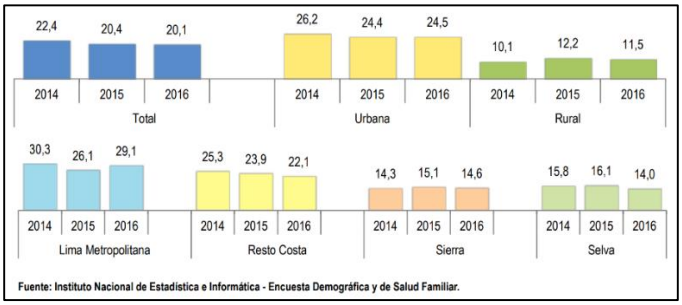

Fig. 1 Cifras de cáncer de mama según zona demográfica.

Según el Ministerio de Salud del Perú. la Vigilancia Epidemiológica de Cáncer notificó un total de 11,836 casos de cáncer de los cuales, 7627 correspondieron a casos nuevos (64,4 %) [2]. El cáncer de mama tiene un 14,9% de los cánceres notificados (superado solo por cérvix y estómago) presentándose con mayor frecuencia en mujeres entre los 40 y 69 años: 30.1% en mujeres de 40 a 49 años y 44.5% de 50 a 69 años. En cuanto a egresos hospitalarios el cáncer de mama fue la segunda causa de hospitalización por cáncer, superado solo por neoplasias hematológicas. La tendencia ha sido ascendente, pasaron de 1,384 en el año 2006 a 2,012 en el año 2011, lo que representa un incremento del 45.4% lo que podría indicar un incremento en el acceso de los pacientes con cáncer de mama a los servicios de salud; pero no necesariamente una atención oportuna.

A. Examen de mamografía

Diversas formas de cáncer de mama son detectables tempranamente con una mamografía. Actualmente, la mamografía es el examen recomendado para la detección temprana de cáncer de mama.

De acuerdo con los resultados de la encuesta, de las mujeres de 40 a 59 años, el 15,9% se han realizado el examen de mamografía en los últimos 24 meses, a fin de detectar anormalidades que puedan indicar un cáncer de mama, sin mayor cambio respecto del 2015 (15,7%).

El 22% de mujeres residentes en el área urbana se realizaron dicho examen. Mientras que el 2015 fue de 21,1%. Este comportamiento en Lima Metropolitana alcanzó el 29,6% en contraste con la Sierra que solo registró el 9,3%.

En la Fig. 2 se muestra las cifras de mamografías según regiones del Perú.

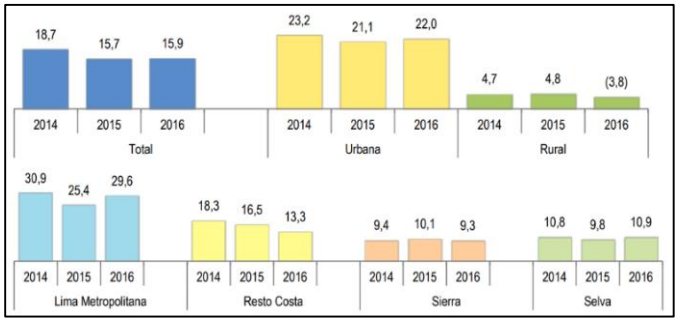

Fig. 2 Cifras de mamografías según región.

B. Brechas de equipamiento de mamografía

En el Perú se registran 202 establecimientos de salud que cuentan con el servicio de mamografía de los cuales 22.3% pertenecen al Ministerio de Salud, 15.3% a ESSALUD, 5.9% a SISOL o Gobierno Municipal; 1.5% a las Fuerzas Armadas y policiales; y 55% pertenecen al sector privado. En total son 19/25 regiones que cuentan con mamógrafo operativo del Ministerio de Salud o Gobierno Regional. Considerando supuestos óptimos de equipamiento instalado y de equipamiento operativo, se pasó a estimar las brechas de equipamiento de mamografía a nivel nacional.

Actualmente se requiere para el cierre de brechas:

a) Comprar equipamiento de mamografía para algunas regiones del Perú.

b) Reparar o reponer prioritariamente el equipamiento de mamografía en algunas regiones del Perú.

c) Reparar o reponer el equipamiento de manera no prioritaria en algunas regiones del Perú que se consideran importantes a nivel nacional.

C. Trabajos Relacionados

En el trabajo desarrollado por Weigel y Dowsett en el año 2010 realizaron estudios sobre los biomarcadores establecidos, como el receptor de estrógenos y el receptor de progesterona desempeñado un papel importante en la selección de pacientes con esta enfermedad [4].

La inteligencia artificial (IA), especialmente el aprendizaje automático y el aprendizaje profundo, ha encontrado aplicaciones populares en la investigación del cáncer de mama en los últimos años, el rendimiento de la predicción del cáncer ha ido aumentando. Huang, Yang, Fong y Zhao han revisado la

literatura sobre la aplicación de la IA al diagnóstico y pronóstico del cáncer, y resume sus ventajas, proporcionando una nueva perspectiva sobre cómo la tecnología de IA puede ayudar a mejorar el diagnóstico y el pronóstico del cáncer, y continuar mejorando la salud humana en el futuro [10].

En el trabajo de Yengec Tasdemir, Kasim Tasdemir y Zafer Aydin, proporcionan un análisis comparativo basado en los métodos de extracción de ROI, los conjuntos de datos y las técnicas de aprendizaje automático empleadas, las precisiones de predicción y las estadísticas de frecuencia de uso [5]. Otros estudios se realizaron utilizando diferentes algoritmos de Aprendizaje automático para predecir [11] y clasificar [12] sobre el cáncer de mama.

Después de haber realizado el respectivo estado del arte respecto al tema, nuestros resultados fueron los siguientes:

TABLA I
ARTÍCULOS REVISADOS PARA EL ESTUDIO

NOMBRE	AUTORES	PAÍS DE ORIGEN
Sistema Predictivo Bayesiano para Detección del Cáncer de Mama [13]	Omar D. Castrillón Eduardo Castaño Luis F. Castillo	Colombia
Redes neuronales: concepto, aplicaciones y utilidad en medicina [14]	N. Sáenz Bajo M. Álvaro Ballesteros	España
Un Modelo para la Predicción de Recidiva de Pacientes Operados de Cáncer de Mama (CMO) Basado en Redes Neuronales [15]	J.A. Gómez Ruiz J.M. Jerez Aragonés J. Muñoz Pérez E. Alba Conejo	España
Breast Cancer Detection using K-nearest Neighbor Machine Learning Algorithm [16]	Moh'd Rasoul Al-hadidi, Abdulsalam, Alarabeyyat Mohannad Alhanahnah	Reino Unido
Inteligencia artificial para asistir el diagnóstico clínico en medicina [17]	Saúl Oswaldo Lugo-Reyes, Guadalupe Maldonado-Colín Chiharu Murata	México
Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis [18]	Hiba Asri a Hajar Mousannif Hassan Al Moatassime Thomas Noel	Marruecos Francia
Molecular Markers for Breast Cancer: Prediction on Tumor Behavior [19]	Bruna Karina Banin Hirata, Julie Massayo Maeda Oda, Roberta Losi Guembarovski Carolina Batista Ariza Carlos Eduardo Coral de Oliveira, Maria Angelica Ehara Watanabe	Brasil
SVM and SVM Ensembles in Breast Cancer Prediction [20]	Min-Wei Huang Chih-Wen Chen Wei-Chao Lin Shih-Wen Ke Chih-Fong Tsai	Taiwán
Lung cancer prediction using machine learning and advanced imaging techniques [21]	Timor Kadir Fergus Gleeson	Reino Unido
Assessment of Diagnostic Values among CA-125, RMI, HE4, and ROMA for Cancer Prediction in Women with Nonfunctional Ovarian Cysts [22]	Shina Oranratanaphan Sinee Wanishpongpan Wichai Termrungruanglert Surang Triratanachat	Tailandia

Evaluando los resultados de la revisión de investigaciones, se concluyó que:

Dentro del Perú no existen estudios sobre el uso de redes neuronales para la predicción de cáncer de mama mediante biomarcadores. En otros países se han realizado estudios dentro de la región América Latina y el Caribe lo mismo que en Brasil. En Europa y Asia se encontraron algunos trabajos relacionados al tema presentado.

III. MATERIALES Y MÉTODOS

A. Datos

Para el desarrollo del aprendizaje automático, se utilizó un conjunto de datos *Breast Cancer Coimbra Data Set* proporcionado por el repositorio de UCI Machine Learning [23], que contiene los siguientes atributos, los cuales son biomarcadores obtenidos de diferentes pacientes que se sometieron a descarte de cáncer de mama. Los biomarcadores son moléculas biológicas que se encuentran en la sangre, otros líquidos o tejidos del cuerpo. [27]:

1) Edad (años),
2) IMC (kg/m2): El índice de masa corporal (IMC) es un método utilizado para estimar la cantidad de grasa corporal que tiene una persona, y determinar por tanto si el peso está dentro del rango normal, o, por el contrario, se tiene sobrepeso o delgadez. Para ello, se pone en relación la estatura y el peso actual del individuo.
3) Glucosa (mg/dL): Es la fuente de energía más importante de las células que componen el organismo. Se conoce como el nivel de azúcar en la sangre, la cual se absorbe de cada uno de los alimentos que la persona consume. Los niveles correctos de la glucosa permiten mantener en buen funcionamiento el calor corporal, la respiración, los latidos del corazón y una buena digestión.
4) Insulina: La insulina es una hormona producida por una glándula denominada páncreas. La insulina ayuda a que los azúcares obtenidos a partir del alimento que ingerimos lleguen a las células del organismo para suministrar energía.
5) HOMA: La evaluación del modelo homeostático o Índice HOMA (homoeostasis model assessment) es un método utilizado para cuantificar la resistencia a la insulina y el porcentaje remanente de células β (beta).
6) Leptina (ng/mL): Es una proteína que se forma principalmente en el tejido adiposo del cuerpo. También puede ser secretada en el hígado, la placenta y la mucosa gástrica. Una vez liberada, la leptina pasa a la sangre y actúa como señal para el cerebro. Los niveles de leptina están relacionados con los de la insulina.
7) Adiponectina (µg/mL): Es una hormona sintetizada mayoritariamente en el tejido graso. Esta interviene en el metabolismo de los lípidos y los hidratos de carbono. Además, tiene función antiinflamatoria y cardioprotector.

8) Resistina (ng/mL): La resistina es una citoquina cuyo papel fisiológico ha sido objeto de mucha controversia en cuando a su relación con la obesidad y la diabetes mellitus tipo II. Se produce y libera a partir de tejido adiposo para servir a las funciones endocrinas probablemente implicadas en la resistencia a la insulina.
9) MCP-1(pg/dL): La proteína quimio atrayente de monocitos 1 (MCP-1), pertenece a la familia de quimio quinas C-C, caracterizadas por tener dos residuos de cisteína adyacentes. está relacionada fundamentalmente con el tránsito de células del sistema inmune.

B. Materiales

Para el preprocesamiento de los datos, se utilizaron las siguientes herramientas:

1) Weka

Waikato Environment for Knowledge Analysis (Weka) de la Universidad de Waikato, es una plataforma de software para el aprendizaje automático y la minería de datos escrito en Java y desarrollado en la Universidad de Waikato. Weka es software libre distribuido bajo la licencia GNU-GPL [24].

El paquete Weka contiene una colección de herramientas de visualización y algoritmos para análisis de datos y modelado predictivo, unidos a una interfaz gráfica de usuario para acceder fácilmente a sus funcionalidades. La versión original de Weka fue un front-end en TCL/TK para modelar algoritmos implementados en otros lenguajes de programación, más unas utilidades para preprocesamiento de datos desarrolladas en C para hacer experimentos de aprendizaje automático. Esta versión original se diseñó inicialmente como herramienta para analizar datos procedentes del dominio de la agricultura, pero la versión más reciente basada en Java (WEKA 3), que empezó a desarrollarse en 1997, se utiliza en muchas y muy diferentes áreas, en particular con finalidades docentes y de investigación.

2) Spyder

Spyder es un potente entorno científico escrito en Python, diseñado por y para científicos, ingenieros y analistas de datos. Ofrece una combinación única de la funcionalidad avanzada de edición, análisis, depuración y generación de perfiles de una herramienta de desarrollo integral con la exploración de datos, la ejecución interactiva, la inspección profunda y las capacidades de visualización de un paquete científico [25].

Más allá de sus muchas características incorporadas, sus habilidades se pueden ampliar aún más a través de su sistema de plugins y API. Además, Spyder también se puede utilizar como una biblioteca de extensión PyQt5, lo que permite a los desarrolladores construir sobre su funcionalidad e incrustar sus componentes, como la consola interactiva, en su propio software PyQt.

C. Métodos: Metodología RUP

RUP divide el proceso en 4 fases, dentro de las cuales se realizan varias iteraciones en número variable según el proyecto y en las que se hace un mayor o menor hincapié en las distintas actividades [26].

- Inicio: Esta fase tiene como propósito definir y acordar el alcance del proyecto con los patrocinadores, identificar los riesgos asociados al proyecto, proponer una visión muy general de la arquitectura de software y producir el plan de las fases y el de iteraciones posteriores.
- Elaboración: En la fase de elaboración se seleccionan los casos de uso que permiten definir la arquitectura base del sistema y se desarrollaran en esta fase, se realiza la especificación de los casos de uso seleccionados y el primer análisis del dominio del problema, se diseña la solución preliminar.
- Construcción: El propósito de esta fase es completar la funcionalidad del sistema, para ello se deben clarificar los requisitos pendientes, administrar los cambios de acuerdo con las evaluaciones realizados por los usuarios y se realizan las mejoras para el proyecto.
- Transición: El propósito de esta fase es asegurar que el software esté disponible para los usuarios finales, ajustar los errores y defectos encontrados en las pruebas de aceptación, capacitar a los usuarios y proveer el soporte técnico necesario. Se debe verificar que el producto cumpla con las especificaciones entregadas por las personas involucradas en el proyecto.

En la actividad de Inicio se realizó la recolección de los datos y luego se procedió a realizar el Preprocesamiento de estos. Para ello se utilizó la opción de Discretización de la herramienta Weka para realizar esta tarea.

Filtro de instancia que discretiza un rango de atributos numéricos del conjunto de datos en atributos nominales.

La discretización es por simple binning. Omite el atributo de clase si se establece.

En la Fig. 3 se muestra los resultados luego de aplicar la discretización a los datos.

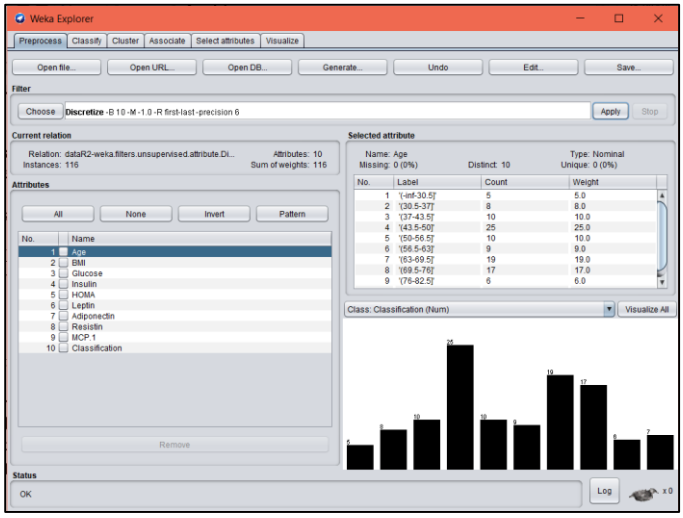

Fig. 3 Datos después de la discretización

Así mismo se realizó la conversión numérica a nominal para ser tratado adecuadamente por el algoritmo.

Se utilizó la opción NumericToNominal de la herramienta Weka. Este es un Filtro para convertir atributos numéricos en nominales. A diferencia de la discretización, solo toma todos los valores numéricos y los agrega a la lista de valores nominales de ese atributo.

Es útil después de las importaciones CSV, para forzar que ciertos atributos se conviertan en nominales, por ejemplo, el atributo de clase, que contiene valores de 1 a 5.

En la Fig. 4 se muestra los resultados luego de aplicar la conversión a los datos.

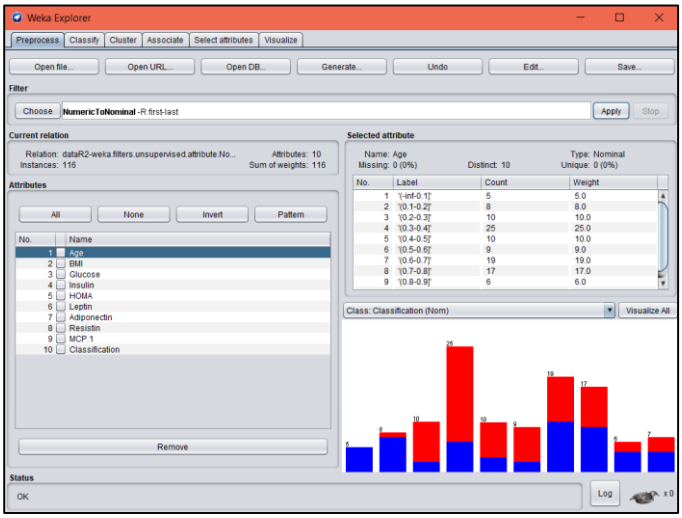

Fig. 4 Datos después de numeric to nominal

Así mismo, se procedió a eliminar las filas incompletas mediante el comando **dropna**, además, se agrupó los datos por edades (0 a 5, 6 a 11, 12 a 17, 18 a 34, 35 a 59, 60 a 99, 100 a más) mediante el comando **cut**. En la Fig. 5 se muestra el resultado de aplicar el preprocesamiento de los datos

```
In [9]: import numpy as np
   ...: import pandas as pd
   ...:
   ...: # Importing the dataset
   ...: dataset = pd.read_csv('C:\\Users\\andre\\Downloads\\datasetwisconsin.csv')
   ...: columnas = ["Edad","IMC", "Insulina","Glucosa","HOMA","Leptina",
"Adiponectina", "Resistina", "MCP-1", "Resultados"]
   ...: dataset.columns = columnas
   ...: #preprocesamiento
   ...: dataset.dropna(axis = 0)
   ...: bins = [0, 5, 12, 18, 35, 60, 100]
   ...: names = ["1", "2", "3", "4", "5", "6"]
   ...: dataset["Edad"] = pd.cut(dataset["Edad"], bins, labels = names)
   ...: dataset
Out[9]:
    Edad    IMC  Insulina  Glucosa  ...  Adiponectina  Resistina   MCP-1  Resultados
0      1   4.00         4     5.00  ...          3.00       2.00    1.00           0
1      1   1.00         1     1.00  ...          3.00       1.00    1.00           0
2      2   8.00         8     1.00  ...          3.00       7.00    1.00           0
3      1   1.00         1     3.00  ...          3.00       1.00    1.00           0
4      2  10.00        10     8.00  ...          9.00       7.00    1.00           1
..   ...    ...       ...      ...  ...           ...        ...     ...         ...
741    5  26.85        92     3.33  ...         12.10      10.96  268.23           1
742    6  26.84       100     4.53  ...         21.42       7.32  330.16           1
743    6  32.05        97     5.73  ...         22.54      10.33  314.05           1
744    6  25.59        82     2.82  ...         33.75       3.27  392.46           1
745    6  27.18       138    19.91  ...         14.11       4.35   90.09           1

[746 rows x 10 columns]
```
Fig. 5 Preprocesamiento de los datos

En la actividad de Elaboración, se realizaron la identificación de los requisitos y su representación mediante los casos de uso y su especificación correspondiente.

Para la actividad de Construcción, se utilizó el algoritmo de red neuronal implementado en el Lenguaje de programación Python junto con sus librerías keras, pandas, numpy y tensorflow. El siguiente bloque de código se muestra una parte de la configuración de la Red Neuronal en Python:

```
dataset = numpy.loadtxt("C:\\Users\\andre\\Downloads\\dataR2.csv", delimiter=",")

X = dataset[:, 0:9]
Y = dataset[:, 9]

model = Sequential()
model.add(Dense(12, input_dim=9, init='normal', activation='relu'))
model.add(Dense(10, init='normal', activation='relu'))
model.add(Dense(8, init='normal', activation='relu'))
model.add(Dense(6, init='normal', activation='relu'))
model.add(Dense(4, init='normal', activation='relu'))
model.add(Dense(1, init='normal', activation='sigmoid'))

model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])

model.fit(X, Y, nb_epoch=100, batch_size=10)

scores = model.evaluate(X, Y)
print("%s: %.2f%%" % (model.metrics_names[1], scores[1]*100))
```

El modelo de red neuronal que se desarrolló se muestra en la Fig. 6:

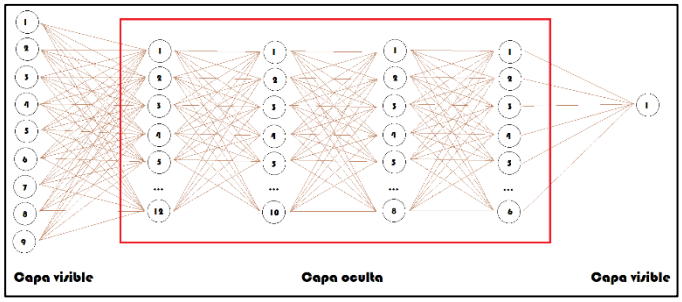

Fig.6: Diagrama de la red neuronal

El modelo de la Red Neuronal consta de la primera capa visible, o, de entrada, con 9 neuronas, seguida de 4 capas ocultas y finalmente una capa visible con 1 neurona que es la salida. En la Tabla II se explica con más detalle las características de las capas.

TABLA II
DISTRIBUCIÓN DE LA RED NEURONAL

N° DE NEURONAS	N° DE ENTRADAS	TIPO DE INICIALIZACIÓN	FUNCIÓN DE ACTIVACIÓN
12	9	Normal	relu
10	-	Normal	relu
8	-	Normal	relu
6	-	Normal	relu
1	-	Normal	sigmoid

IV. RESULTADOS

Las principales funcionalidades que cumple el sistema son las siguientes:

- Autenticación de usuario
- Ingreso de biomarcadores y predicción de datos
- Ingreso de pacientes
- Ingreso de médicos
- Generación de reportes

Dentro de los cuales, también se considera que:

- La predicción tiene una precisión del 90.0%
- Los datos son validados antes de ser mandados a la red neuronal
- La autenticación del usuario se da correctamente
- Antes de ingresar datos en la base de datos, se validan que cumplan todas las condiciones

Las pruebas realizadas y los resultados obtenidos de éstas han sido analizadas y documentadas con el fin de:

- Encontrar errores y posteriormente solucionarlos.
- Evaluar la calidad del software.
- Proporcionar información para la toma de decisiones.

En las pruebas aplicadas a la red neuronal, se logró obtener una precisión del 82.76%.

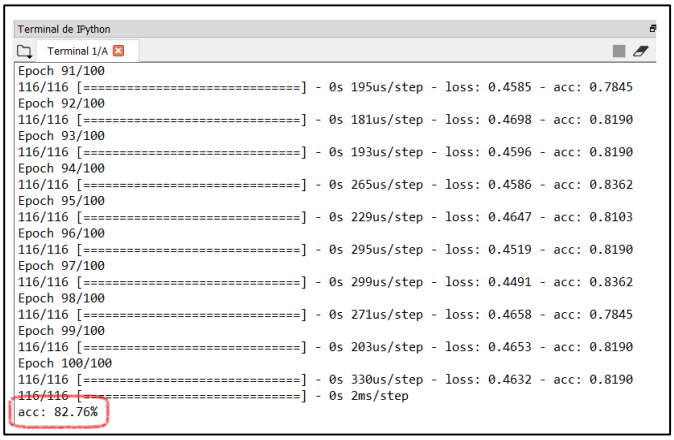

Fig.7: Resultado del entrenamiento a la Red Neuronal

Así mismo, se construyó un prototipo para interactuar con la red neuronal. Se ha creado una interfaz gráfica en Python con soporte en la librería Django. En la Fig. 8 se muestra una interfaz del prototipo.

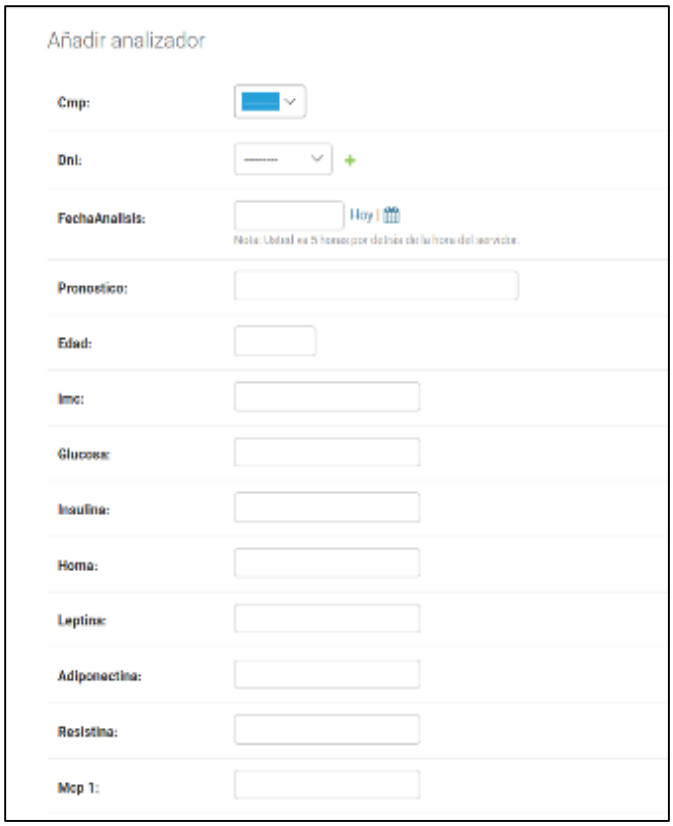

Fig.8: Interfaz de ingreso de biomarcadores

En la Tabla III, se puede apreciar la primera prueba realizada

TABLA III
PRUEBA P01

ID	P01-1
NOMBRE	Comprobación de predicciones de red neuronal
DESCRIPCIÓN	Se comprueba que los resultados (predicciones) de la red neuronal sean correctos basándonos en los registros previos.
ESPECIFICACIÓN DE ENTRADA	Data de testeo que son los mismos registros de la data set sin incluir el atributo de clase ya que es lo que se va a predecir.
RESULTADO ESPERADO	Columna de predicciones (Y_pred) llenos de 0 y 1 (0=Tumor benigno, 1=Tumor maligno) que sea igual a la data de archivo esperado (Y_test)
RESULTADO OBTENIDO	Columna de predicciones con 136 predicciones acertadas de 140 dando un 97.14% de precisión con los resultados esperados
ESTADO	Aprobado

```
#PRUEBA01
#NO
new_prediction01 =
classifier.predict(sc.transform(np.array([[76,29.2184076,83,5.
376,1.1006464,28.562,7.36996,8.04375,698.789]])))
new_prediction01 = (new_prediction01 > 0.5)
print(new_prediction01)
```

```
new_prediction02 =
classifier.predict(sc.transform(np.array([[76,27.2,94,14.07,3.2
62364,35.891,9.34663,8.4156,377.227]])))
new_prediction02 = (new_prediction02 > 0.5)
print(new_prediction02)
new_prediction03 =
classifier.predict(sc.transform(np.array([[75,27.3,85,5.197,1.0
89637667,10.39,9.000805,7.5767,335.393]])))
new_prediction03 = (new_prediction03 > 0.5)
print(new_prediction03)

#SI
new_prediction14 =
classifier.predict(sc.transform(np.array([[45,20.82999519,74,4
.56,0.832352,7.7529,8.237405,28.0323,382.955]])))
new_prediction14 = (new_prediction14 > 0.5)
print(new_prediction14)
new_prediction15 =
classifier.predict(sc.transform(np.array([[49,20.9566075,94,12
.305,2.853119333,11.2406,8.412175,23.1177,573.63]])))
new_prediction15 = (new_prediction15 > 0.5)
print(new_prediction15)
new_prediction16 =
classifier.predict(sc.transform(np.array([[34,24.24242424,92,2
1.699,4.9242264,16.7353,21.823745,12.06534,481.949]])))
new_prediction16 = (new_prediction16 > 0.5)
print(new_prediction16)
```

	Resultado (0 = NO 1 = SI)	Resultado red neuronal	Preciso (SI/NO)
classifier1	0	False	Si
classifier2	0	True	No
classifier3	0	False	Si
classifier14	1	True	Si
classifier15	1	True	Si
classifier16	1	True	Si

Se aplicó la prueba con el 20% de datos del dataset de entrenamiento para poder verificar la precisión de esta. Se puede observar que los datos de color rojo dieron un resultado erróneo en la red neuronal, demostrando así que la red necesita ajustes para poder ser más precisa. Además, demuestra que necesita más datos para ser más precisa

CONCLUSIÓN

Hoy en día la predicción del cáncer de mama está limitado a las mamografías, las cuales mediante imágenes demuestran si es que la paciente está desarrollando un tumor, mas no especifica si es benigno o maligno. Es después de eso que recién se realizan estudios más a profundidad, siendo muchas veces exámenes dolorosos y costosos, los cuales muchas personas no pueden costear.

Por medio de la aplicación de algoritmos de aprendizaje automático puede apoyarse los especialistas para la predicción del cáncer de mama. Uno de estos algoritmos es la red neuronal que con la configuración adecuada puede obtener una aceptable precisión, como fue el caso de lo que se ha propuesto, con un 82.76 % de precisión, lo que puede ser tomado como una herramienta prototipo de apoyo de los médicos oncólogos y pacientes al momento de realizar exámenes de rutina, ya que está comprobado que mediante los niveles de diferentes sustancias que se obtienen del cuerpo (llamados biomarcadores) se puede predecir con un nivel de confiabilidad mayor al 50% si es que la persona cuenta con algún tipo de enfermedad, en este caso, de cáncer de mama.

Los siguientes aportes se han encontrado como resultado de lo desarrollado:

- Ayuda a la detección de cáncer de manera menos dolorosa, más barata y rápida que por los medios convencionales.
- Sugiere tratamientos, y medicinas tomando en cuenta las necesidades y restricciones de los pacientes.
- Presenta informes con resultados del análisis y sugerencias.
- Usa datos reales para mayor exactitud y realismo.
- Maneja una gran cantidad de información de manera flexible y segura, manteniendo en todo momento la confidencialidad.
- Es un gran aporte al tratamiento y detección de cáncer de mama.
- Es de fácil uso.

REFERENCIAS

[1] International Agency for Research on Cancer (IARC), "GLOBOCAN 2018: Latest global cancer data.," *CA. Cancer J. Clin.*, 2018.

[2] MINSA, "Boletín Epidemiológico del Perú," 2018.

[3] S. Berrospi-Reyna, M. Herencia-Souza, and A. Soto Tarazona, "Prevalencia y factores asociados a la sintomatología depresiva en mujeres con cáncer de mama en un hospital público de Lima, Perú," *ACTA MEDICA Peru.*, 2017.

[4] M. T. Weigel and M. Dowsett, "Current and emerging biomarkers in breast cancer: Prognosis and prediction," *Endocrine-Related Cancer*. 2010.

[5] S. B. Yengec Tasdemir, K. Tasdemir, and Z. Aydin, "A review of mammographic region of interest classification," *WILEY Interdiscip. Rev. Min. Knowl. Discov.*

[6] American Cancer Society, "Breast Cancer Facts & Figures 2019-2020," *Am. Cancer Soc.*, 2019.

[7] L. E. Figueroa-Montes, N. E. Chávez-Altamirano, and G. Garc\'\ia-Espinoza, "Implementación de una unidad de mamovigilancia para el diagnóstico de cáncer de mama en una microred de la seguridad social, Lima-Perú," *Acta Médica Peru.*, vol. 36, no. 1, pp. 11–18, 2019.

[8] Instituto Nacional de Estadistica e Informatica, "Perú: Enfermedades No Transmisibles y Transmisibles, 2018," *Inst. Nac. Estadística e Informática*, 2019.

[9] M. de Salud Dirección General de Intervenciones Estratégicas en Salud

Públicae, “Plan Nacional para la prevención y control de cáncer de mama en el Perú. 2017-2021 (RESOLUCIÓN MINISTERIAL N° 442-2017/MINSA).” Ministerio de Salud Lima, 2017.

[10] S. Huang, J. Yang, S. Fong, and Q. Zhao, “Artificial intelligence in cancer diagnosis and prognosis: Opportunities and challenges,” *Cancer Letters*. 2020.

[11] C. H. Shravya, K. Pravalika, and S. Subhani, “Prediction of breast cancer using supervised machine learning techniques,” *Int. J. Innov. Technol. Explor. Eng.*, 2019.

[12] T. Araujo *et al.*, “Classification of breast cancer histology images using convolutional neural networks,” *PLoS One*, 2017.

[13] O. D. Castrillón, E. Castaño, and L. F. Castillo, “Sistema Predictivo Bayesiano para Detección del Cáncer de Mama,” *Inf. tecnológica*, 2018.

[14] N. Sáenz Bajo and M. Álvaro Ballesteros, “Redes neuronales: concepto, aplicaciones y utilidad en medicina,” *Atención Primaria*, 2002.

[15] J. A. . Gomez, J. M. Perez, J. Muñoz, and E. Alba, “Un Modelo para la Prediccion de Recidiva de Pacientes Operados de Cancer de Mama (CMO) Basado en Redes Neuronales,” *Intel. Artif.*, 2000.

[16] M. R. Al-Hadidi, A. Alarabeyyat, and M. Alhanahnah, “Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algorithm,” in *Proceedings - 2016 9th International Conference on Developments in eSystems Engineering, DeSE 2016*, 2017.

[17] S. O. Lugo-Reyes, G. Maldonado-Colín, and C. Murata, “Inteligencia artificial para asistir el diagnóstico clínico en medicina,” *Revista Alergia Mexico*. 2014.

[18] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, “Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis,” in *Procedia Computer Science*, 2016.

[19] B. K. Banin Hirata, J. M. M. Oda, R. Losi Guembarovski, C. B. Ariza, C. E. C. De Oliveira, and M. A. E. Watanabe, “Molecular markers for breast cancer: Prediction on tumor behavior,” *Disease Markers*. 2014.

[20] M. W. Huang, C. W. Chen, W. C. Lin, S. W. Ke, and C. F. Tsai, “SVM and SVM ensembles in breast cancer prediction,” *PLoS One*, 2017.

[21] T. Kadir and F. Gleeson, “Lung cancer prediction using machine learning and advanced imaging techniques,” *Translational Lung Cancer Research*. 2018.

[22] S. Oranratanaphan, S. Wanishpongpan, W. Termrungruanglert, and S. Triratanachat, “Assessment of diagnostic values among CA-125, RMI, HE4, and ROMA for cancer prediction in women with nonfunctional ovarian cysts,” *Obstet. Gynecol. Int.*, 2018.

[23] UCI, “UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set,” *http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%25 28Diagnostic%2529*. 2011.

[24] M. A. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA data mining software: an update,” *SIGKDD Explor.*, vol. 11, no. 1, pp. 10–18, 2009.

[25] Spyder, “SPYDER IDE,” *Spyder Project*, 2018. .

[26] C. Jones, “50 Rational Unified Process (RUP),” in *Software Methodologies A Quantitative Guide*, 2017.

[27] Instituto Nacional de Cáncer, «Diccionario de cáncer: marcador biológico,» Instituto Nacional de Cáncer, [En línea]. Available: https://www.cancer.gov/espanol/publicaciones/diccionario/def/marcador-biologico.