# Investigación sobre el Cáncer de mama basado en métodos predictivos

Andrade Reyes, Luis Angel; Julca Cano,Julio Angel; Orellana Leon, Ricardo Miguel

[aluis.andrade4@unmsm.edu.pe](mailto:aluis.andrade4@unmsm.edu.pe),

[bjulio.julca@unmsm.edu.pe](mailto:bjulio.julca@unmsm.edu.pe),

[cricardo.orellana1@unmsm.edu.pe](mailto:cricardo.orellana1@unmsm.edu.pe)

# Resumen: El Cáncer de mama (CM) es una enfermedad degenerativa la cual es el tipo de Cáncer más común teniendo 2,2 millones de casos en el año 2020. La OMS confirma que una de cada 12 mujeres padece de esta enfermedad, además el porcentaje de casos de hombres afectados por dicho mal es de 1% con respecto al número de casos totales. Para la prevención del CM se usarán datos de un examen de chequeo general, dichos datos serán analizados con los modelos de machine learning, naive bayes, árboles de decisión y KNN. El dataset contiene datos de 569 pacientes la cual serán analizados con los modelos previamente dichos y a su vez comparados para ver cual modelo tiene es el mejor modelo predictivo para nuestra dataset. Palabras clave: Cáncer de mama, modelos predictivos, machine learning, naive bayes, KNN, árboles de decisión.

# Abstract: Breast Cancer (BC) is a degenerative disease which is the most common type of Cancer, with 2.2 million cases in 2020. The WHO confirms that one in 12 women suffers from this disease, in addition to the percentage of cases of men affected by this disease is 1% with respect to the number of total cases. For the prevention of CM, data from a general check-up exam will be used, said data will be analyzed with machine learning models, naive bayes, decision trees and KNN. The dataset contains data from 569 patients which will be analyzed with the previously mentioned models and in turn compared to see which model is the best predictive model for our dataset. Keywords: Breast cancer, predictive models, machine learning, naive bayes, KNN, decision trees.

# Introducción

El Cáncer es una enfermedad degenerativa que tiene su origen en el ADN de la persona afectada; por ende, hasta ahora no se encuentra una cura para dicha enfermedad y variantes. Los afectados se llevan terapias constantemente, las cuales son para evitar que la degeneración siga progresando. Los tratamientos para amortiguar el avance de dicho mal son diversos siendo los siguientes los más usados: Cirugía a la zona afectada, radioterapia, quimioterapia, inmunoterapia, trasplante de células madre.

El tipo de Cáncer más común es el de mama con más de 2,2 millones de casos en 2020. Este se origina en la capa de revestimiento (epitelio) de los conductos (85%) o lóbulos (15%) del tejido glandular de los senos. Según la sentencia anterior, es una falacia afirmar que dicho tipo de cáncer es exclusivo del género femenino. Pero no cabe negar que el porcentaje de hombres (1%) es muy bajo.

Actualmente con el uso del Big Data, las enfermedades se perfilan a ser diagnosticadas más temprano. Esto no excluye al cáncer de mama, que si bien es cierto que para su detección y confirmación se realiza un examen en específico (mamografía), la cual, mediante algunos datos obtenidos en un chequeo general, la cual se debe realizar periódicamente, se utilizarán para la predicción y detección del cáncer de mama. Este documento evaluará y comparará ciertos modelos frente a una data set la cual se usará para la predicción del Cáncer de mama.

# Estado del Arte

Según la OMS (26 de marzo del 2021), el CM es el tipo de cáncer más común en el mundo con un total de 2,2 millones en el año 2020, además de tener un total 685 000 de fallecimientos en ese mismo año. El CM como todo tipo de cáncer en incurable pero no intratable en el 2018 se usó el *machine learning* para la mejora en el tratamiento del CM detectando con mayor facilidad las células cancerosas además de su próxima eliminación para evitar el avance a las células aledañas.

Los especialistas del National Center for Tumor Diseases están trabajando para implementar tecnología en la Nube el procesamiento en la información SAP HANA (tecnología en base de datos en memoria) Recuperado de la pag. Web Big Data Magazine (2018), el uso de la nube llegaría facilitar para la recopilación de datos externos a dicho centro de investigación. Además de facilitar la exploración de datos basados en informes clínicos.

La Facultad de Ingeniería de la Universidad Nacional de Asunción (03/11/2020), determinó que el 27% de los casos totales de mujeres afectadas con cáncer son detectados tempranamente, además se prevé que para el 2030 dicho porcentaje aumentará a un 34%. Su trabajo de investigación está compuesto por un modelo de red neuronal convolucional llamado U-Net que segmenta las masas benignas (tejido epitelial mayormente piel) y malignas (células cancerígenas).

1. Marco Teórico

## 3.1. Clasificación KNN

Según Ignacio (2018):

“K-Nearest-Neighbor es un algoritmo [basado en instancia](https://www.aprendemachinelearning.com/principales-algoritmos-usados-en-machine-learning/#instancia) de tipo [supervisado](https://www.aprendemachinelearning.com/aplicaciones-del-machine-learning/#supervisado) de Machine Learning. Puede usarse para clasificar nuevas muestras (valores discretos) o para predecir (regresión, valores continuos).

Funciona de la siguiente manera:

1. Calcular la distancia entre el ítem y el resto de los ítems del dataset de entrenamiento.
2. Seleccionar los “k” elementos más cercanos (con menor distancia. según la función que se use).
3. Realizar una “votación de mayoría” entre los k puntos: los de una clase/etiqueta que <<dominen>> decidirán su clasificación final.

Las formas más populares de “medir la cercanía” entre puntos son la distancia Euclidiana (“la “de siempre”) o la Cosine Similarity (mide el ángulo de los vectores, cuantos menores, serán similares).”

Fórmula Euclidiana:

Fórmula Cosine Similarity:

# 

## 3.2. Árbol de decisiones

Alpaydin (2018) explica que los árboles de decisiones de la siguiente forma:

“El árbol de decisiones busca las instancias de entrenamiento más similares mediante una secuencia de pruebas en diferentes atributos de entrada. El árbol está compuesto por nodos de decisión y hojas; a partir de la raíz, cada nodo de decisión aplica una prueba de división a la entrada y, según el resultado, tomamos una de las ramas. Cuando llegamos a una hoja, la búsqueda se detiene y entendemos que hemos encontrado las instancias de entrenamiento más similares.

Cada ruta desde la raíz hasta una hoja corresponde a una conjunción de condiciones de prueba en los nodos de decisión de la ruta y dicha ruta se puede escribir como una regla si-entonces. El árbol se entrena con datos de entrenamiento dados donde se colocan divisiones para delimitar las regiones que tienen la “pureza” más alta, en el sentido de que cada región contiene instancias que son similares en términos de su salida.”

## 3.3. Clasificación Naive Bayes

Rendon (2018) nos aclara los siguiente respecto a la estadística bayesiana:

“La estadística bayesiana se basa en la probabilidad subjetiva, trabaja con la actualización de la evidencia considerando los conocimientos adquiridos previos a una investigación, más la evidencia obtenida con esta. La interpretación de los resultados requiere la especificación de las hipótesis por contrastar y su probabilidad a priori antes del estudio. La evidencia del estudio se mide con el factor Bayes (razón de la compatibilidad de los datos bajo las hipótesis propuestas). La conjunción de las probabilidades a priori de las hipótesis con el factor Bayes permite calcular la probabilidad a posteriori de cada una. La hipótesis con mayor grado de certidumbre en su actualización es la aceptada para la toma de la decisión.”

Factor Bayes:

Donde:

P : es la probabilidad de casos malignos en cáncer de mama

P : es la probabilidad de casos benignos en cáncer de mama

## 3.4. ANN

Es una red compuesta por neuronas artificiales y conexiones sinápticas, donde cada neurona tiene un valor de activación, y una conexión de la neurona A a la neurona B tiene un peso que define el efecto de A sobre B. Si la sinapsis es excitadora, cuando A es activa también intenta activar B; si la sinapsis es inhibitoria, cuando A está activo, intenta suprimir B.

Durante la operación, cada neurona suma las activaciones de todas las neuronas que hacen sinapsis con ella, ponderadas por sus pesos sinápticos, y si la activación total es mayor que un valor umbral, la neurona "dispara" y su salida corresponde al valor de esta activación; de lo contrario, la neurona está en silencio. Si la neurona dispara, envía su valor de activación a todas las neuronas con las que hace una sinapsis como se puede apreciar en la figura 1.

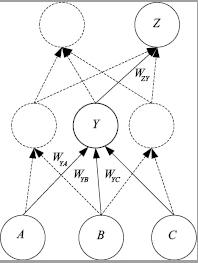


Figura 1. Ejemplo de una red neuronal.

El perceptrón básicamente calcula una suma ponderada antes de tomar una decisión, y esto puede verse como una forma de implementar una variante del modelo lineal que discutimos antes. Estas neuronas se pueden organizar como capas donde todas las neuronas de una capa toman información de todas las neuronas de la capa anterior y calculan su valor en paralelo, y estos valores juntos se alimentan a todas las neuronas de la capa siguiente; esto se llama un perceptrón multicapa.

## 3.5. Matriz de Confusión

La matriz de confusión de un problema de *clase n* es una *matriz nxn* en la que las filas se nombran según las clases reales y las columnas, según las clases previstas por el modelo.

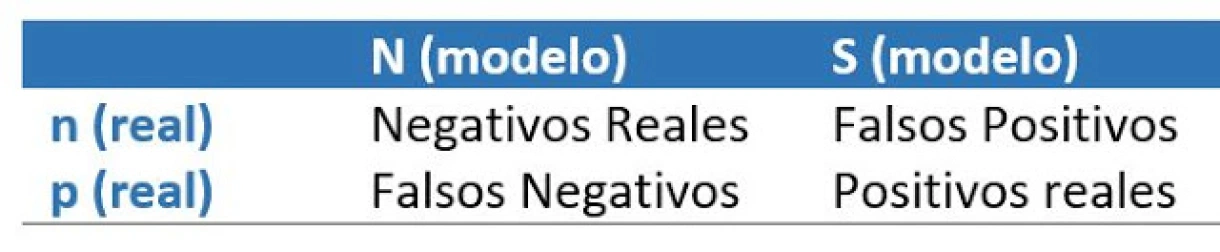


Figura 2. Matriz de confusión binaria

Dicha matriz nos ayuda a obtener la precisión de cada predicción mediante las siguientes fórmulas:

# Metodología

El proceso algorítmico para la realización de este artículo será el que se muestra en Figura 3. La data será estandarizada con datos numéricos para luego ser tratado por los modelos de predicción y a su vez usar dicha data como data a predecir para la cual se usará la metodología de *partition*. Se compararon los resultados con la data a predecir para ver la eficiencia de predicción de cada modelo. Finalmente se escogerá el modelo que alcance una mayor predicción.

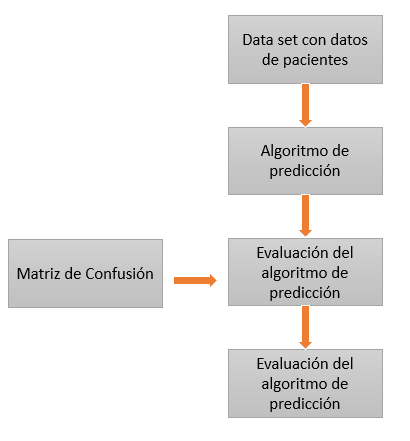


Figura 3. Proceso Algorítmico

## 4.1. Dataset

El dataset en este artículo tiene como fuente la página web Kaggle. Cristopher Gutiérrez (2021), Detection Cancer, hizo uso del dataset que recopiló datos de pacientes de CM en la ciudad de Wisconsin teniendo en cuenta los atributos que se muestran en la Tabla 1.

El acceso al dataset se da mediante el siguiente enlace: [DetecciónCancer | Kaggle](https://www.kaggle.com/keopscris/detecci-ncancer/data).

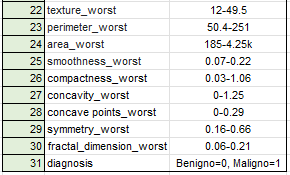
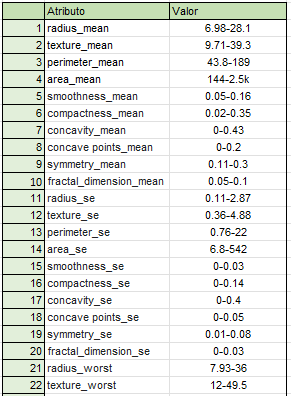


Tabla 1. Descripción de atributos

## Pruebas

Dados los modelos previamente descritos se evaluarán sus resultados individualmente para después comparar mediante su precisión. El codigo python originalmente hecho en Colab se puede acceder mediante el siguiente enlace: https://colab.research.google.com/drive/1g6hw2lDnHOfqvekoVuvqugsA0DPvmq7s?usp=sharing

## 5.1. K-Nearest-Neighbor

El código que utilizamos utiliza la distancia Euclidiana para realizar la predicción y una vez obtenida la distancia se compara con los 5 primeros valores más cercanos a 0 y mediante probabilidad se predice si la persona tiene cáncer benigno o maligno.

Una vez ejecutado el código nos da como resultado el vector de predicciones la cual se puede apreciar en la figura 4.

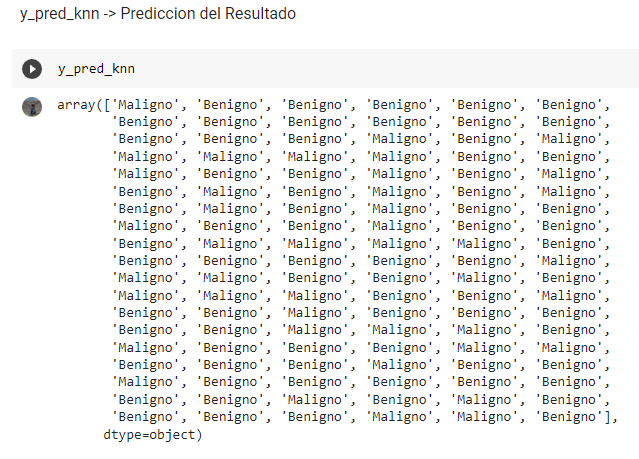


Figura 4. Vector de Predicción KNN

La cual comparando con los verdaderos resultados se obtiene la matriz de confusión que se puede observar en la figura 5 la cual contiene los siguientes valores: verdaderos positivos (67), falsos positivos (0), falsos negativos (5) y verdaderos negativos (42).

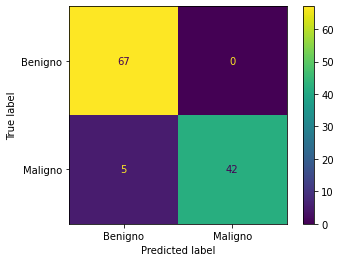


Figura 5. Matriz de confusión KNN

Dicha matriz nos ayudará para poder ver la precisión de la clasificación KNN utilizando la distancia Euclidiana en el diagnóstico de cáncer de mama (CN).

## 5.2. Árbol de decisión:

Ejecutando el algoritmo del árbol de decisión podremos tener el *ruleset* la cuál vendría a ser el conjunto de reglas a utilizar para tener una predicción en este caso se muestra un árbol de decisión como se ve en la figura 6.

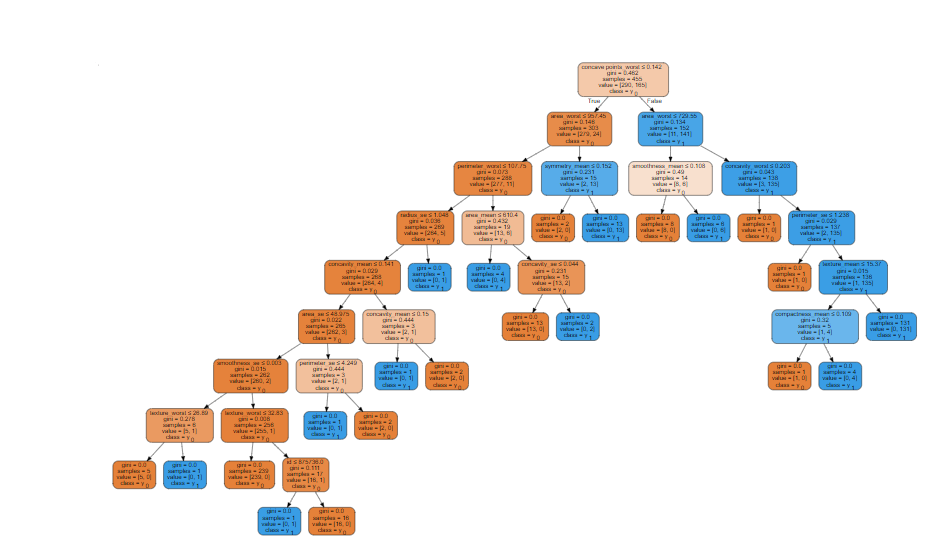


Figura 6. Árbol de decisión

Utilizando el método de *partition* compararemos los datos originales con los mismos datos aplicando el método predictivo la cual serán comparadas y encontrar la precisión. Para esto se usará una matriz de confusión (Figura 7) la cual contiene los siguientes datos: verdaderos positivos (62), falsos positivos (5), falsos negativos (3) y verdaderos negativos (44).

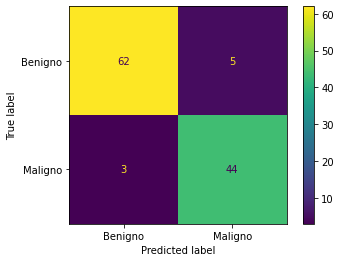


Figura 7. Matriz de Confusión

Dados dichos valores nos ayudará para poder calcular la precisión de la *ruleset* como la precisión de acertar si CM es benigno o maligno.

## 5.3. Naive bayes

Para la clasificación Naive Bayes, utilizamos la predicción de Bosques aleatorios o Random Forest para obtener nuestro vector de predicciones la cual se puede apreciar en la Figura 8.

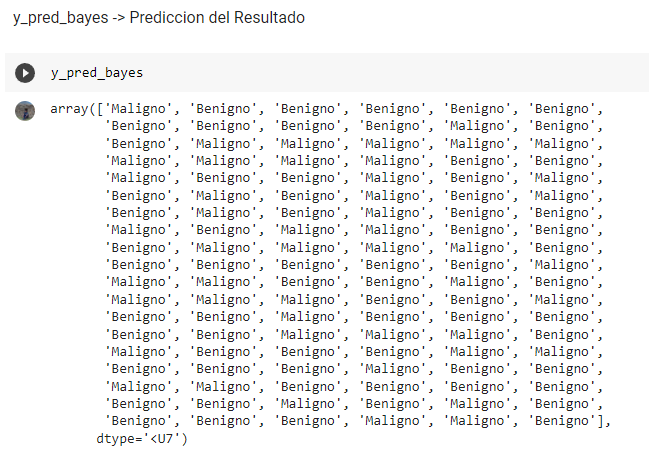


Figura 8. Vector de predicción Naive Bayes

Comparando el vector de resultados verdaderos con el vector de predicciones se obtiene la matriz de confusión que se puede apreciar en la Figura 9 la cual contiene los siguientes datos: verdaderos positivos (61), falsos positivos (6), falsos negativos (5) y verdaderos negativos (42).

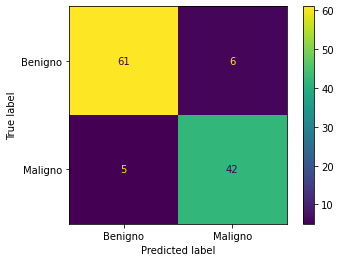


Figura 9. Matriz de confusión Naive Bayes.

Dicha matriz nos ayudará para poder ver la precisión de la clasificación Naive Bayes utilizando los Random Forest en la diagnosticación de cáncer de mama (CN).

## 5.4.Artificial Neural Networks

Usando la clasificación ANN o Artificial Neural Networks, más que usar este tipo de clasificación entrenamos una red neuronal para que dicha red pueda predecir los resultados.

Para llevarlo a cabo, la columna del Diagnóstico se tuvo que cambiar a números (1: Maligno y 0: Benigno) ya que el modelo ANN no trabaja con variables nominales.

La construcción de la red neuronal está dada por tres capas:

**Capa de entrada:** En esta capa se determinarán las variables de entradas (30) y las variables que pasarán a la siguiente capa, en la cual, por recomendación, se elige la mitad de la suma entre las variables independientes más la variable dependiente (15). También se inicializan los pesos con valores cercanos a 0, esto con la ayuda de “uniform”. A esta capa se le conoce como el espacio donde se realiza la sinapsis entre las neuronas.

**Capa oculta:** En esta capa ya no es necesario poner la cantidad de variables de entrada ya que, en la anterior capa, se especificó el número de variables (15).

Capa de salida: En esta capa se especifica el número de variables de salida, que en nuestro caso será 1 (Maligno o Benigno). Además de ello, se especifica la función de activación en la cual, por recomendación, se utilizará la función sigmoide.

Luego de crear la red neuronal, se pasará a ejecutar. Para ello, requerimos un modelo de optimización en la cual se eligió “Adam” y como métrica me mostrará la precisión del modelo.

Una vez construida la red neuronal y compilada, ahora se necesita entrenarla con los datos ya preparados para dicho proceso y para realizarlo se utiliza el siguiente código que se puede apreciar en la Figura 10 donde agrupará cada 10 registros y los validará mediante 100 iteraciones como mínimo.

.

Figura 10. Código para el entrenamiento de una red neuronal.

Con la red neuronal lista para predecir si el cáncer de mama es benigno o maligno pasaremos los datos restantes de la dataset para ver la precisión que tiene nuestra red neuronal.

Los datos que predijo la red neuronal al compararlos con los valores verdaderos se obtiene la siguiente matriz de confusión que se puede apreciar en la figura 11 la cual contiene los siguientes valores: verdaderos positivos (67), falsos positivos (0), falsos negativos (0) y verdaderos negativos (47).

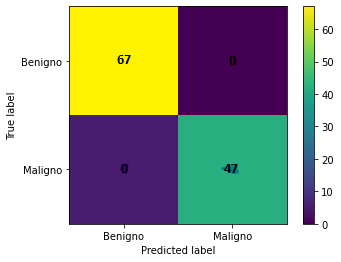


Figura 11. Matriz de confusión ANN

Dicha matriz nos ayudará para poder ver la precisión de la clasificación ANN en la diagnosticación de cáncer de mama (CN).

# Evaluación de resultados

Para comparar los pros y los contras de los modelos de clasificación, es necesario proporcionar métricas para evaluar el desempeño de los modelos. Aquí dividimos la muestra en cuatro clases, cada una con su modelo.

En el caso de KNN se obtiene la siguiente matriz de confusión con una precisión de 95.61% y un margen de error de 4.39%.

|  |  |
| --- | --- |
| Verdadero Positivo | Falso Positivo |
| 67 | 0 |
| Falso Negativo | Verdadero Negativo |
| 5 | 42 |

En el caso de Árbol de Decisión se obtiene la siguiente matriz de confusión con una precisión de 92.98% y un margen de error de 7.02%.

|  |  |
| --- | --- |
| Verdadero Positivo | Falso Positivo |
| 62 | 5 |
| Falso Negativo | Verdadero Negativo |
| 3 | 44 |

En el caso de Naive Bayes se obtiene la siguiente matriz de confusión con una precisión de 90.35% y un margen de error de 9.65%.

|  |  |
| --- | --- |
| Verdadero Positivo | Falso Positivo |
| 61 | 6 |
| Falso Negativo | Verdadero Negativo |
| 5 | 42 |

En el caso de la Red Neuronal(ANN) se obtiene la siguiente matriz de confusión con una precisión de 99.998% y un margen de error de 0.21%.

|  |  |
| --- | --- |
| Verdadero Positivo | Falso Positivo |
| 67 | 0 |
| Falso Negativo | Verdadero Negativo |
| 0 | 47 |

Este artículo divide los resultados característicos en dos categorías, Maligno y Benigno, y en la parte de ANN se ha utilizado "1" para resultados Maligno y "0" para resultados Benigno. Primero, dividimos los datos en dos partes. En este experimento, la relación entre el conjunto de entrenamiento y el conjunto de predicción es 80:20. Se utilizó los datos del conjunto de entrenamiento para que el modelo se entrene y luego usar el modelo entrenado y el conjunto de predicciones como entrada en el componente de predicción.

Resumimos los resultados de los cuatro algoritmos de clasificación anteriores como se muestra en la Tabla 2. En la que se observa las tasas de precisión final en nuestro conjunto

de datos. Se tiene la tasa de precisión más alta, con una tasa de precisión del 99.998% por parte del modelo de red neuronal, seguida de KNN con una tasa de precisión del 95.61%, luego el modelo de árbol de decisión con una tasa de precisión del 92.98%, y por último, el modelo de Naive Bayes con una tasa de precisión del 90.35%.

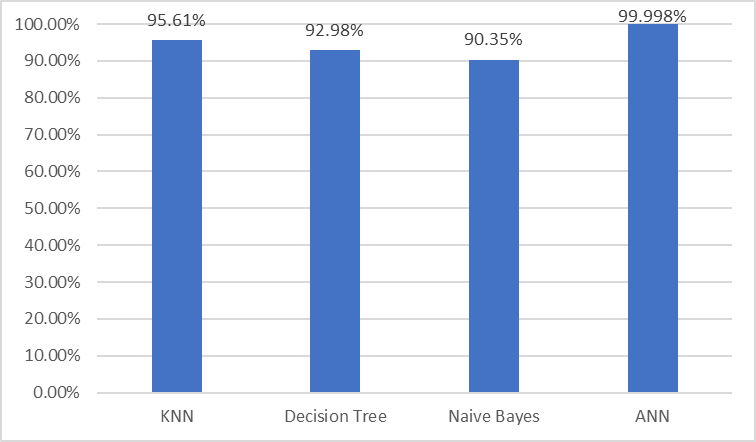


Tabla 2: Precisión de los algoritmos de clasificación

# Aplicación

## 7.1. Aplicación Web

Para la construcción de la página web sobre la predicción del cáncer de mama, se utilizó la aplicación de KNIME Analytics Platform en la cual se pudo obtener a través de una estructura el archivo PMML del árbol de decisión, como se observa en la figura 12.

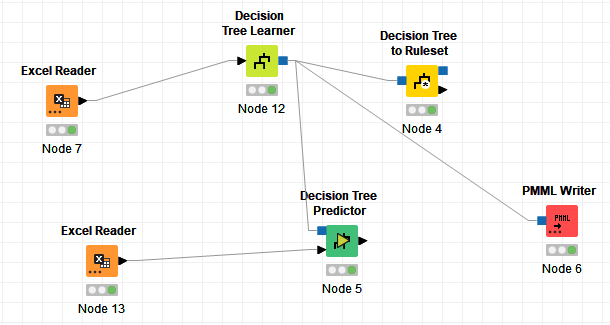


Figura 12: Diagrama para la obtención del PMML

Luego, el archivo PMML se logró subir a la plataforma Clouderizer ya que este me permite crear páginas web con este tipo de extensión, quedando la aplicación web como se observa en la figura 13.

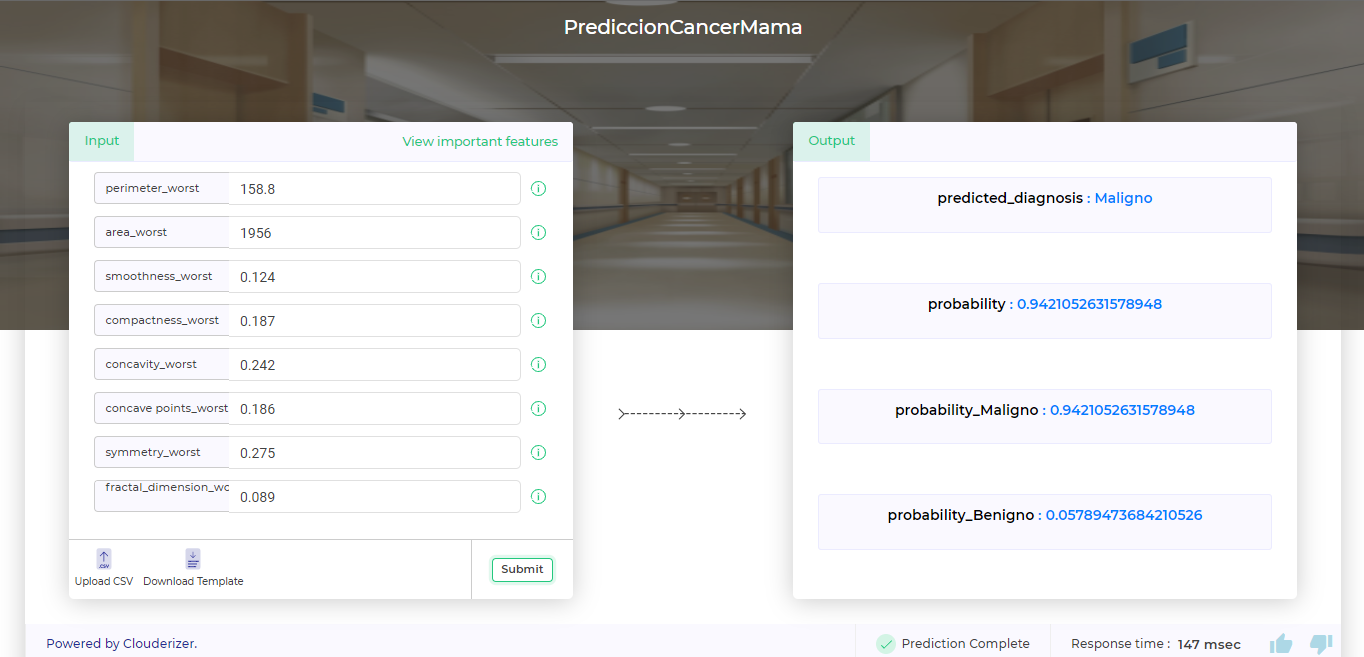


Figura 13: Página Web de la Predicción del cáncer de mama

## 7.2. Aplicación Móvil

Para la realización de predicciones masivas y dar a conocer el valor de este artículo se empleará el uso de una aplicación móvil (apk) la cual se podrán registrar los valores requeridos la cuál después de usar el algoritmo escogido nos dará una predicción.

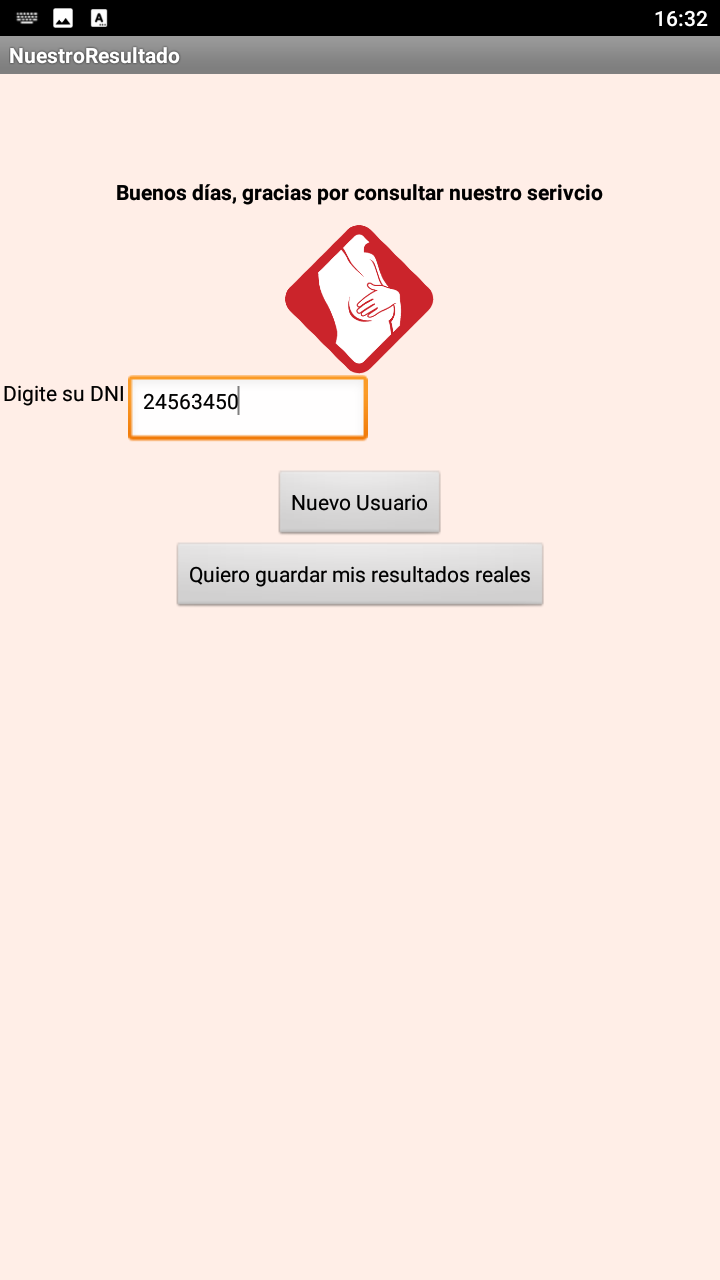


Figura 14. Login de aplicación móvil

Si bien es cierto dicha predicción no es 100% certera y se ha realizado una mamografía diagnosticada debidamente podrá ser de apoyo para futuras actualizaciones.

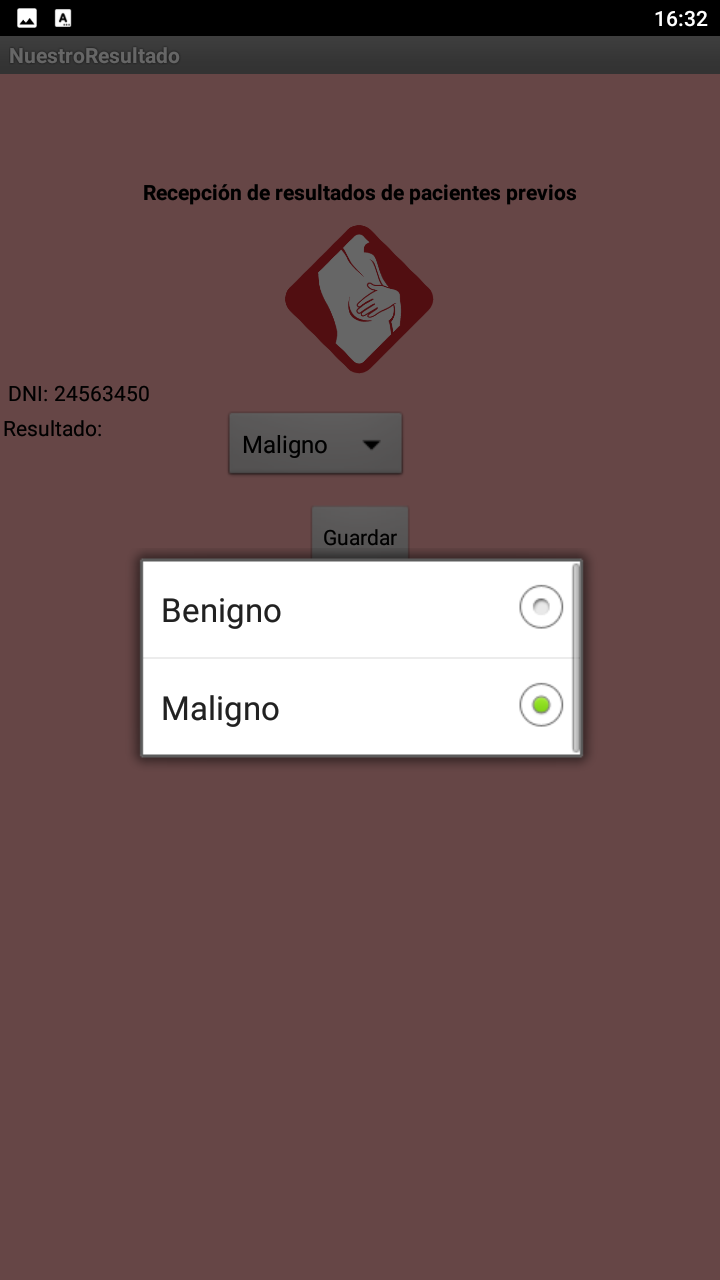


Figura 15. Vista de corrección de resultados

En caso de que quiera registrar una nueva predicción se le mostrará una nueva interfaz como se aprecia en la figura 16 y 17. Esta predicción se realizó en base a las reglas que se pudo obtener del árbol de decisión en la aplicación de KNIME Analytics Platform.

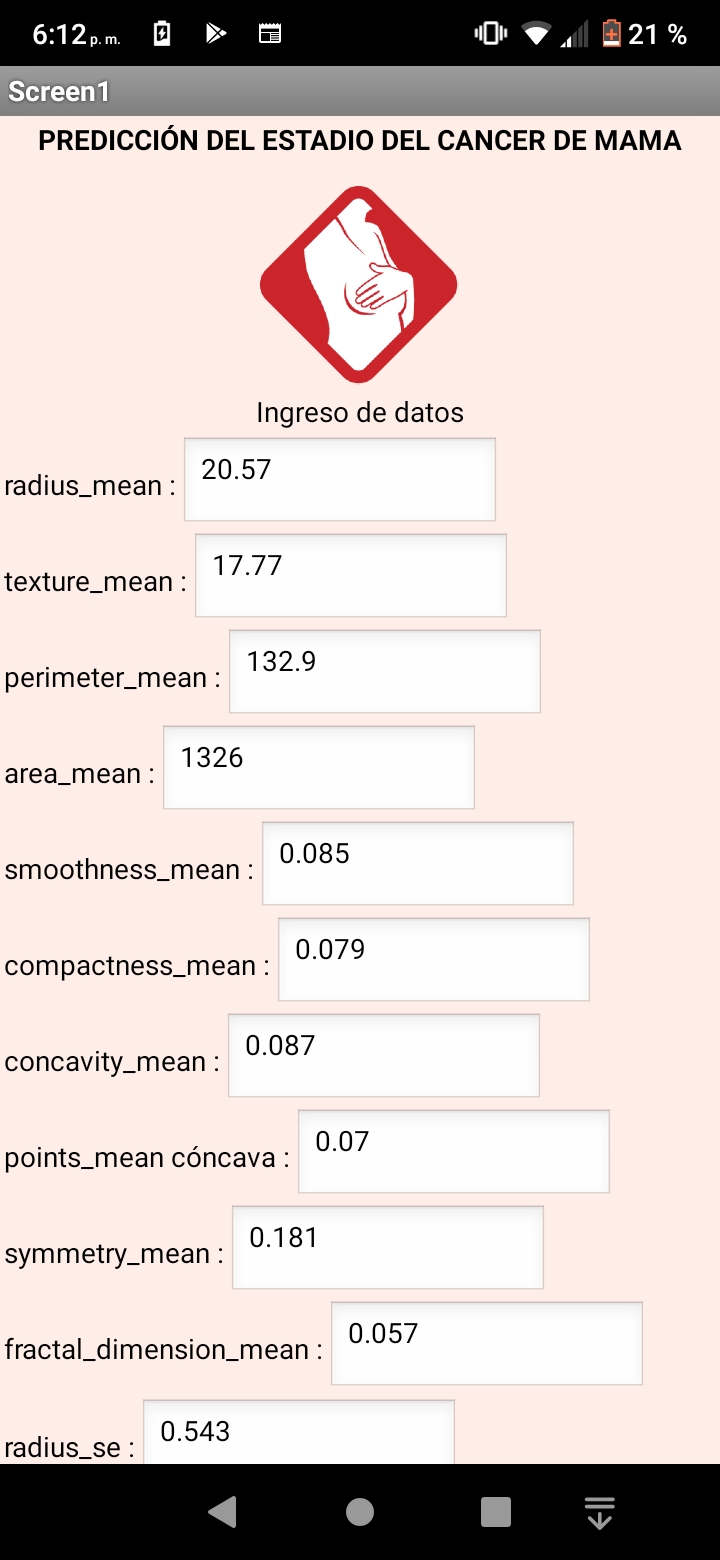
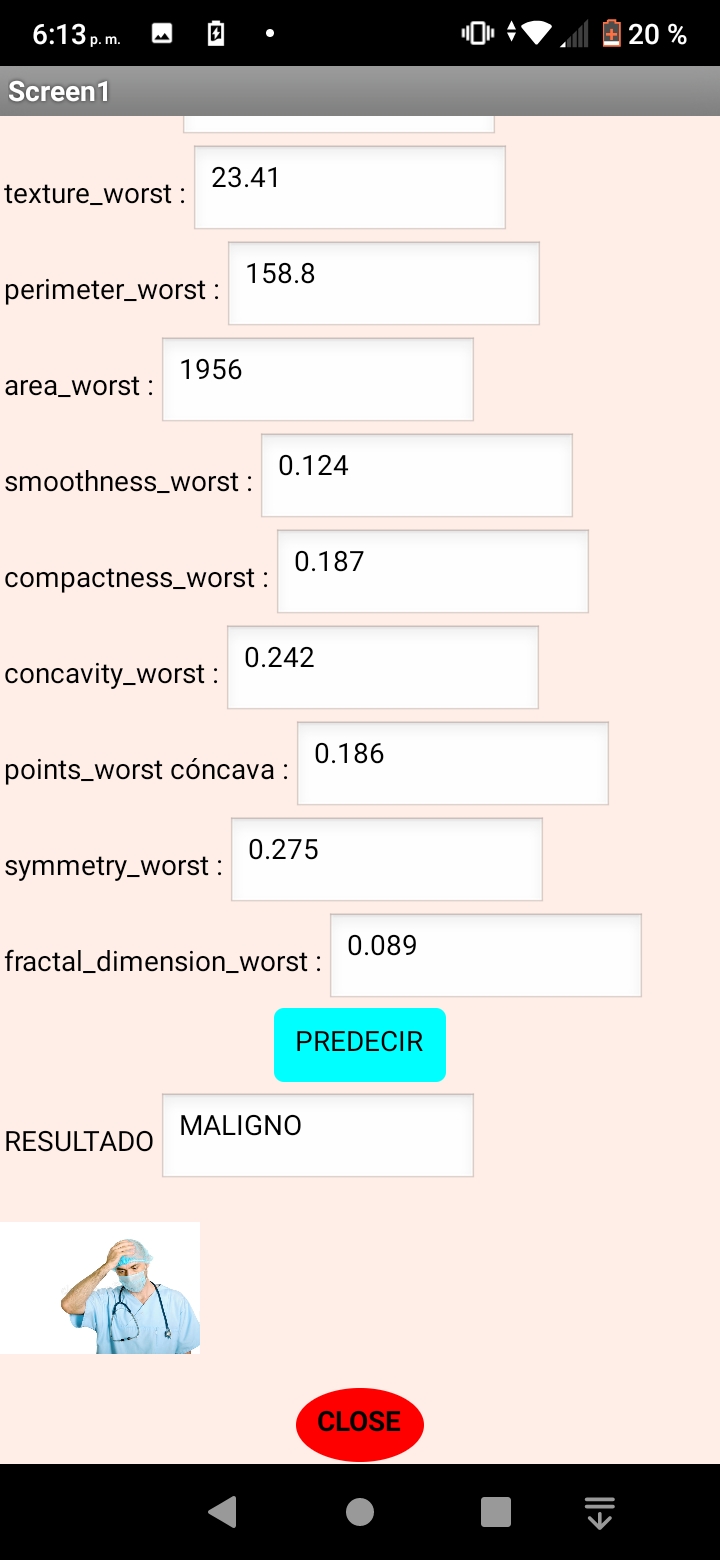
 

figura 16: ingreso de datos figura 17: resultado en base a los datos

## 7.2.1 Base Datos

Los datos obtenidos previamente son almacenados en una base de datos virtual de Firebase (Google). Con la cuál en un futuro se podrá usar como dataset propia para la mejora en el modelo predictivo.

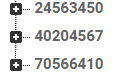


Figura 18. Imagen de la BD DNI registrados



Figura 19. Extendido ejemplar (datos ficticios) de nuestra BD

# Conclusiones

La detección temprana del cáncer de mama en base al aprendizaje automático ha revolucionado el estudio de la predicción del riesgo. Con el avance continuo de los métodos de minería de datos, hemos estudiado varios métodos para diagnosticar el cáncer de mama. Se descubrió que la red neuronal tiene la mayor precisión a través de la prueba de evaluación de la matriz de confusión.

Sin embargo, este tipo de investigación debe actualizarse periódicamente con más conjuntos de datos de instancia. Se espera que pueda ayudar a los médicos a hacer mejores juicios sobre el estado de la enfermedad.

# Recomendaciones

Se recomienda para este estudio aumentar la cantidad de datos de entrenamiento de las clasificaciones ya que al tener una cantidad de data menor a 1000 no se puede confiar mucho en la precisión. Cuando la data aumenta en el dataset de entrenamiento la precisión de los modelos pueden subir o bajar, una vez aumentada la data se podrá verificar si los campos que se usan para predecir el diagnóstico benigno o maligno de cáncer de mama son correctos o se debería agregar o quitar un campo.

# Referencias bibliográficas

* + Recuero, P.. (2018). Machine Learning a tu alcance: La matriz de confusión. noviembre 30, 2021, de Telefonica Tech Sitio web: <https://empresas.blogthinkbig.com/ml-a-tu-alcance-matriz-confusion/>
  + Ignacio, J.. (2018). Clasificar con K-Nearest-Neighbor ejemplo en Python. 29/11/2021, de Aprendiendo Machine Learning Sitio web: <https://www.aprendemachinelearning.com/clasificar-con-k-nearest-neighbor-ejemplo-en-python/#:~:text=K%2DNearest%2DNeighbor%20es%20un,el%20mundo%20del%20Aprendizaje%20Autom%C3%A1tico>.
  + Alpaydin, E.. (2016). Machine learning. EE.UU: Massachusetts Institute of technology.
  + Rendón, M., Riojas, A., Contreras, D. & Martínez, J.. (2018). Análisis bayesiano. Conceptos básicos y prácticos para su interpretación y uso. Revista Alergia México, 65, 285-298.
  + (S/f-a). Who.int. Recuperado el 30 de noviembre de 2021, de [https://www.who.int/es/news-room/fact-sheets/detail/breast-canc](https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer)er.
  + (S/f-b). Bigdatamagazine.es. Recuperado el 30 de noviembre de 2021, de <https://bigdatamagazine.es/el-big-data-contra-el-cancer-de-mama>.
  + (S/f-c). Una.py. Recuperado el 30 de noviembre de 2021, de <http://www.ing.una.py/?p=43778>