PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

ESCUELA DE POSGRADO

Propuesta de un Modelo de Predicción de Cáncer de Mama Utilizando Deep Learning

TESIS PARA OBTENER EL GRADO DE MAGÍSTER EN GERENCIA DE TECNOLOGÍAS DE INFORMACIÓN OTORGADO POR LA PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

PRESENTADA POR

Jorge Antonio Páez Cumpa, 43461755

Henry Edward Palomino Delgado, 71479229

Christian Paul Rosado Farfán, 43071779

Elmer Ronald Salazar Huamanjulca, 45352290

ASESOR

Hobber Arístides Siccha Ayvar, DNI 10140192

ORCID 0000-0002-1670-9730

JURADO

O'brien Cáceres, Juan

Salcedo Huarcaya, Marco Antonio

Hobber Arístides Siccha Ayvar

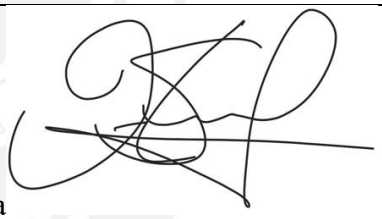
Santiago de Surco, junio, 2023

Declaración Jurada de Autenticidad

Yo, Hobber A. Siccha Ayvar, docente del Departamento Académico de Posgrado en Negocios de la Pontificia Universidad Católica del Perú, asesor(a) de la tesis/el trabajo de investigación titulado Propuesta de un Modelo de Predicción de Cáncer de Mama utilizando Deep Learning, de los autores Jorge Antonio Páez Cumpa, Henry Edward Palomino Delgado, Christian Rosado Farfán, Elmer Ronald Salazar Huamanjulca, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 20%. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el 03/03/2023.
- He revisado con detalle dicho reporte y confirmo que cada una de las coincidencias detectadas no constituyen plagio alguno.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha: Lima, 19 de Octubre del 2023

Apellidos y nombres del asesor / de la asesora: Siccha Ayvar Hobber Arístides	
DNI:10140192	 Firma
ORCID: 0000-0002-1670-9730	

Agradecimientos

Queremos agradecer a todos los profesores de CENTRUM PUCP BUSINESS SCHOOL correspondientes a la Maestría de Gerencia en Tecnologías de la Información, que nos brindaron su enseñanza, experiencias y consejos este largo camino, compartiendo y transmitiendo su pasión por nuestra profesión, y sembrar en nosotros el deseo y pasión de aprendizaje continuo.

A nuestros compañeros de la maestría, que compartieron sus experiencias, apoyaron en todo momento y siempre compartiendo sus energías positivas en cada clase.

A nuestro asesor por acompañarnos en todo momento, siempre dispuesto a brindarnos su tiempo para absolver nuestras dudas y dándonos la motivación necesaria para culminar la presente tesis.

Al médico genetista Yasser Sullcahuaman quien nos brindó el conocimiento, las herramientas y su tiempo para poder entender la información sobre alteraciones genéticas y como estas podrían generar algún tipo de Cáncer.

Finalmente agradecemos a nuestras familias, que son la inspiración para el camino de superación constante de cada uno de nosotros, nuestra razón de seguir adelante y quienes siempre nos dan esas palabras de aliento para continuar.

Dedicatorias

A mi madre Martha, quien me brindo el gran consejo de superación personal y motivo a tomar este camino de aprendizaje, quien siempre se preocupó de que tenga todos los recursos necesarios para poder seguir adelante sin dejar de darme su amor para nunca rendirme. A mi querido hermano Johan, que siempre estuvo pendiente a que no pierda mi norte y a esforzarme el doble para cumplir mis metas. A mi amada novia y futura esposa Jenny, que siempre me motivo con las palabras necesarias en cada momento difícil, siendo el soporte e inspiración que necesito. A mi querida abuela Lorenza que, con su gran sonrisa y alegría, siempre llenó mi corazón para esforzarme tanto como ella lo hizo para tener un futuro. Finalmente, a Dios ya que por su gracia todo es posible.

Henry Edward Palomino Delgado

Nada de esto sería posible sin el gran apoyo de mi querida esposa Jessica, quien siempre fue la que me dio las energías para continuar, a mi adorada madre que siempre hizo todo para apoyarme y supo darme soporte en mis momentos más oscuros, a mis adorados hijos (Facundo, Annie, Jeremy, Manuel y Nicole) para que esto sirva de fuente de inspiración y por último y no menos importante a mi hermana Magaly, ya que sin sus enseñanzas y guías no estaría donde estoy ahora.

Christian Paul Rosado Farfán

A mi hermano Juan Daniel, quién me hizo entender el sentido de la vida y sé que desde donde estés siempre tendrás una sonrisa y alguna ocurrencia para hacerme sentir feliz, a Margarita Alejandra y Karina, madre esposa y hermana, a mis hijos, Santiago y Joaquín motor de mi vida.

Elmer Ronald Salazar Huamanjulca

En primer lugar agradezco el apoyo constante y ánimos recibidos por parte de mi querida esposa Karen y de mi hijita Lucianita, también agradezco el soporte de mis padres Wilfredo y Mirtha que me animan siempre a superarme a mí mismo y dar lo mejor de mí para ser un mejor profesional, de igual manera a mi hermana Claudia que siempre es un ejemplo de estudio y de perseverancia y por último a mi querida abuelita Elva que todos los días me dejaba palabras de aliento para que sea siempre una buena persona.

Jorge Antonio Páez Cumpa

Resumen Ejecutivo

En la presente tesis, queremos demostrar y proponer como la tecnología puede ser utilizada por los genetistas y especialistas en oncología como una herramienta para agilizar la detección de cáncer de mama, siendo este el más común en Perú. El diagnóstico temprano es un mecanismo efectivo que ayuda a la reducción de la mortalidad en este tipo de cáncer de tal manera que se pueda seguir un tratamiento adecuado.

Actualmente una forma de detectarlo es a través de una prueba genética para identificar mutaciones en los genes BRCA 1 y BRCA 2, sin embargo, este camino contiene pruebas que son difíciles, costosas y lentas, que a su vez requieren una carga de trabajo excesiva por parte de un biólogo o genetista. por tal motivo se tiene como objetivo combinar los factores de riesgo asociados con el cáncer de mamá, incluidas las variaciones genéticas para diseñar un modelo predictivo basados en la inteligencia artificial para determinar si el tumor asociado al cáncer es benigno o maligno. El modelo se diseñó utilizando un algoritmo de redes neuronales logrando obtener un rendimiento de 92% precisión con datos de prueba en tan solo unos minutos.

Esta propuesta de modelo de predicción es única en el Perú y puede ser ofrecida por una Gerencia de TI dentro de una organización del sector salud para que posteriormente pueda ser implementada y desplegada por un equipo de científicos de datos.

Palabras clave: Aprendizaje profundo, redes neuronales, BRCA1, BRCA2, cáncer de mama

Abstract

In the present thesis, we are looking for a demonstration and proposal how the technology can be so useful for the genetic and oncology Scientifics as a tool for quick detection of the breast cancer, which ones is the most common in Peru. Early diagnosis is the most effective way for a treatment to help people to prevent the mortality in this kind of cancer.

At this moment, the best way for an early detection is a genetical test to look for mutations in BRCA 1 and BRCA 2 gen, however this way is so hard, because this requires a lot of difficult, expensive, and slowly tests remark a lot of work of the genetic and oncology Scientifics. That is the reason our thesis has as the principal goal to combine all the risk factors associated with breast cancer, including genetical mutations, for generate a predictive model based in artificial intelligence for determinate if a kind of tumor is associated with benign or pathogenic. This designed model has a 92% of precision with open-source test data in a few minutes.

This predictive model is unique in Peru and can be offered by an IT Management within a health sector organization so that it can later be implemented and deployed by a team of data scientists.

Key words: Deep learning, neural network, BRCA1, BRCA2, breast cancer

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Capítulo 1 Introducción

Antecedentes

El cáncer de mama es el tipo de cáncer más común entre las mujeres peruanas; informó el Ministerio de Salud, en el año 2020, “Al año se realizan más de 6000 diagnósticos, pero el 90% de los casos detectados tienen alta posibilidad de cura si se detecta en las etapas tempranas, mejorando así la calidad de vida de las personas con esta enfermedad neoplásica” (Ministerio de Salud, 2020)

Actualmente, en Perú, la única forma de poder obtener un diagnóstico anticipado de este se da mediante las pruebas a los genes BRCA1 y BRCA2, el cual consiste en un examen de sangre para la obtención del código genético, con el fin de detectar los genes que presenten mutaciones correspondientes a los genes BRCA1 y BRCA 2, esto puede indicar si el paciente tiene un mayor riesgo de padecer cáncer. Estas pruebas moleculares para el diagnóstico genético y molecular del cáncer son desarrolladas por los especialistas del Instituto Nacional de Enfermedades Neoplásicas (INEN)

Este proceso toma un promedio de 2 semanas, por el cual el médico genetista hace la revisión manual de cada cromosoma, comparándola con la de sus padres, para determinar así una irregularidad genética y determinar qué tipo de enfermedades puede desarrollar el paciente.

El NIH (Instituto Nacional del Cáncer) explica que los genes BRCA1 y BRCA2 tienen la función de reparar el daño al ADN, esto mediante la codifican de proteínas supresoras de tumores, estas proteínas codificadas cumplen el papel de reparar el ADN dañado, manteniendo así una estabilidad en el código genético.

Cuando un gen sufre una mutación asociada a la pérdida de su función, el paciente presenta un riesgo muy alto de padecer cáncer, en caso de que una mujer presente mutaciones en

los genes BRCA 1 y BRCA 2, el paciente tiene una alta probabilidad de presentar cáncer de mama y ovario. Caso contrario, en el varón, las mutaciones en BRCA1 y BRCA2, están relacionados con un aumento de riesgo de próstata.

En ambos sexos, las mutaciones en BRCA1 y BRCA2 se han revelado como causantes de un aumento en el riesgo de cáncer de páncreas. Las mutaciones en BRCA2 (conocidas como FANCD1), cuando se heredan de ambos padres, pueden provocar un subtipo de anemia de Fanconi, un síndrome que se asocia a tumores sólidos en los niños y a Leucemia Mieloide Aguda. Por su parte, las mutaciones en BRCA1 (también denominadas FACS), cuando se heredan también de ambos padres, están relacionadas con otro subtipo de anemia de Fanconi.

Problema de la Investigación

El tiempo de entrega de resultados de análisis de los genes BRCA1 y BRCA2, el cual determina si una persona podría o no desarrollar un tumor maligno o benigno, demora hasta 10 días, esto incluye una prueba de sangre para extraer la información genética, lo cual puede durar en el mejor de los casos 12 horas, luego sigue un proceso de interpretación de la variación que puede tomar hasta 6 días, donde se realiza búsquedas a bases de datos de alteraciones genéticas para determinar si hay una relación de una alteración con una enfermedad.

Figura 1

Flujo AS IS consulta médica genética

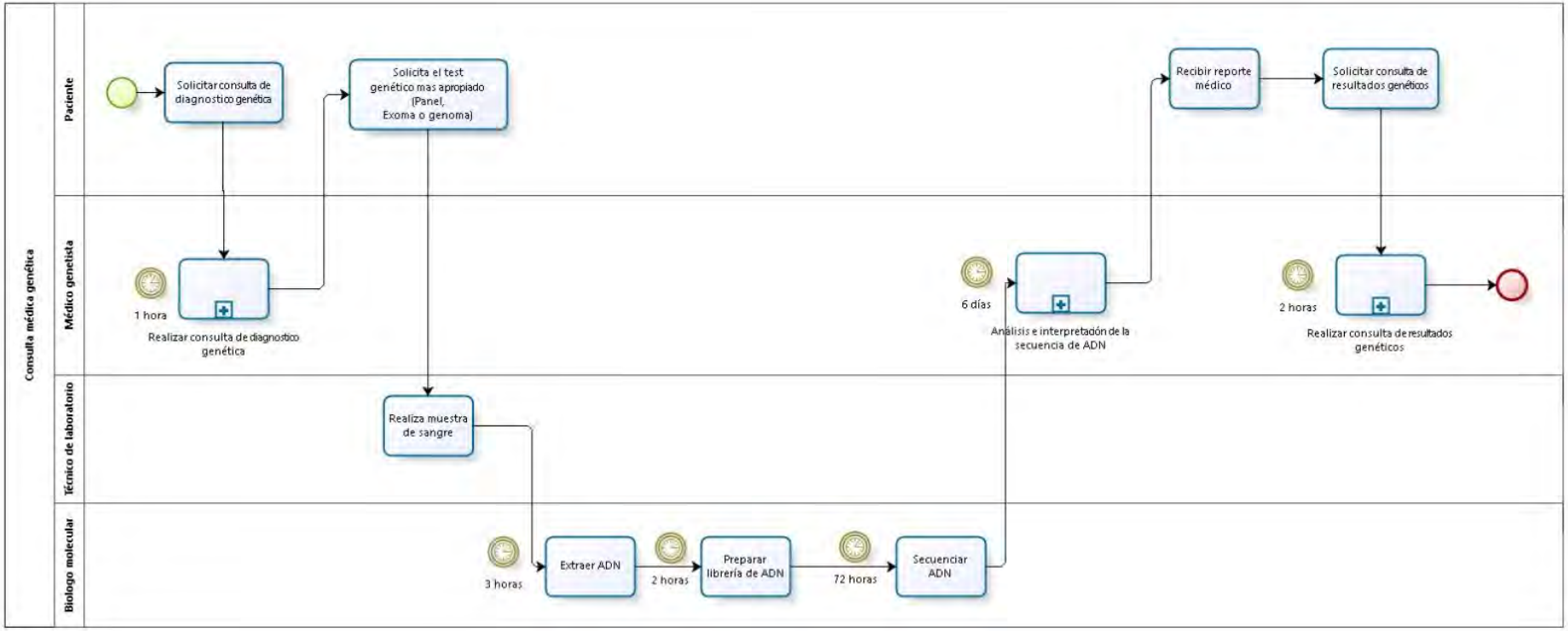


Figura 2

Realizar consulta de diagnóstico genética

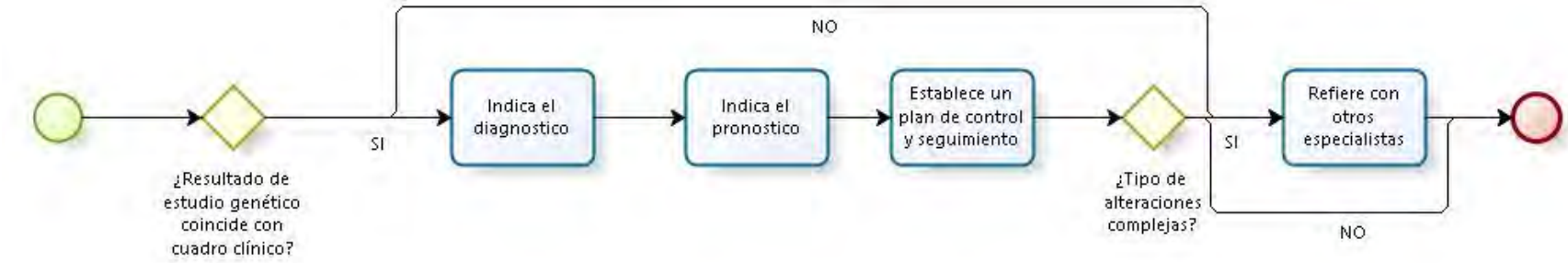


Figura 3

Análisis e interpretación de la secuencia de ADN en el AS IS

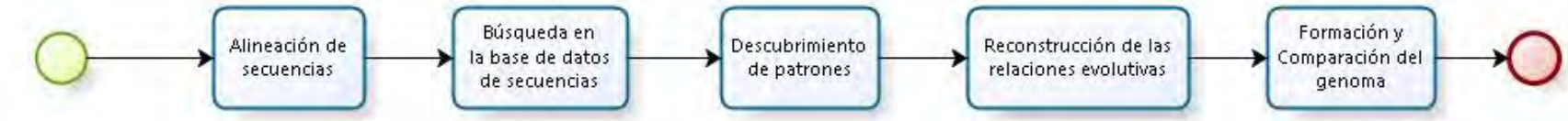
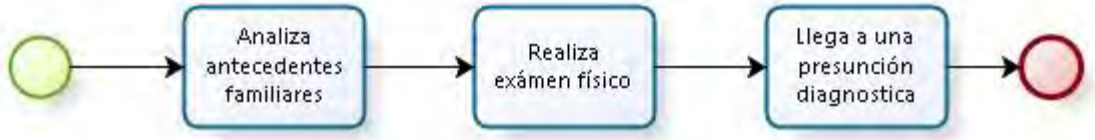


Figura 4

Realizar consulta de resultados genéticos


Analiza antecedentes familiares
Realiza exámen físico
Llega a una presunción diagnostica

Propósito de la Investigación

Objetivo General

Diseñar un modelo analítico para determinar si un tumor de seno es benigno o maligno en base a las mutaciones de los genes BRACA1 y BRCA2 utilizando Deep Learning.

Objetivos Específicos

1. Obtener, en colaboración con un médico genetista, los requerimientos necesarios para proponer un modelo analítico que agilice la detección del cáncer de mama
2. Proponer un modelo analítico que permita agilizar la detección temprana de los tipos de cáncer de mama que podría tener un determinado paciente
3. Revisar las limitaciones legislativas existentes para realizar este tipo de análisis.
4. El modelo analítico debe de contar con un 75% de probabilidad de acierto en sus primeras ejecuciones (según modelos similares revisados en la literatura).

Beneficios

La ejecución del modelo propuesto tiene una duración de minutos lo cual trae beneficios a los pacientes relacionados a la reducción de estrés al evitar pérdida de tiempo y de dinero.

Al identificarse un cáncer de mamá con origen en una alteración genética se puede brindar un tratamiento personalizado y menos invasivo para mejorar la esperanza de vida del paciente. Actualmente, la quimioterapia sigue siendo la primera opción de tratamiento de un cáncer de mamá, ahora se tendría como alternativa el tratamiento por inhibidor PARP.

Preguntas de la Investigación

- ¿Es posible utilizar un enfoque computacional, basado en Deep Learning para analizar datos obtenidos de secuenciadores genéticos, realizando así un correcto

análisis genético e identificar las variantes patógenas, en el cual no se tenga dedicación exclusiva de un médico genetista?

- ¿Puede un modelo de predicción basado en Deep Learning, ayudar a reducir el tiempo del análisis genético?
- ¿En qué consiste la secuenciación genética?
- ¿La población peruana tiene conocimiento de los beneficios de los análisis genéticos?
- ¿Qué iniciativas se tomaron en Perú para promover los análisis genéticos?
- ¿De qué manera la variable de población afectará la distribución de los genes?
- ¿Qué tipo de alteraciones genéticas tienen mas probabilidad de que un tumor sea maligno?
- ¿Existen bases de datos públicas en el Perú que pueden ser utilizadas para realizar análisis genéticos?
- ¿Qué métodos de obtención de probabilidad de resultados de cáncer, que utilicen secuenciación genética, existen actualmente, que sean rápidos, de costo accesible y confiables?
- ¿Siempre es necesario contar con un médico genetista para un correcto análisis genético?

Justificación de la Investigación

Actualmente una forma de detectar este tipo de cáncer es a través de una prueba genética para identificar mutaciones del gen BRCA1/2, sin embargo, es una prueba difícil, costosa y lenta que requiere una carga de trabajo excesiva por parte de un biólogo o genetista.

Por tal motivo se tiene como objetivo combinar los factores de riesgo asociados con el cáncer de mamá, incluidas las variaciones genéticas para diseñar un modelo predictivo basados en la inteligencia artificial para determinar si el tumor asociado al cáncer es benigno o maligno.

La tarea más desafiante para cualquier médico es predecir el resultado de cualquier enfermedad con mayor precisión. Para ello, las técnicas de Deep Learning se han convertido en una herramienta importante y popular entre muchos investigadores. Estas técnicas pueden predecir el futuro resultado de cualquier tipo de cáncer de manera muy efectiva al descubrir e identificar patrones de cualquier conjunto de datos complejo, la mayoría de los autores han recomendado la máquina de vectores de estado (SVM) como un clasificador para la detección de células cancerosas.

A diferencia de Machine Learning, cuya habilidad de aprendizaje es limitada independiente a la cantidad de datos adquiridos, los algoritmos de Deep Learning pueden mejorar cada vez su rendimiento si tienen acceso a más datos, uno de esos es el algoritmo de redes neuronales, el cual puede aprender automáticamente adaptándose y corrigiéndose para ajustarse a los patrones observados en los datos a partir del tuneo de hiper parámetros como son el número de capas y el número de nodos por capas.

Marco Teórico Conceptual

Tumor

Según la revista JAMA Network Journals of the American Medical Association, se define al tumor de la siguiente forma: “Es una masa anormal de células en el cuerpo. Es causado por células que se dividen más de lo normal o que no mueren cuando deberían. Los tumores se pueden clasificar en benignos o malignos”. (Patel, 2020)

Tipos de tumores.

Existen 3 tipos de tumores.

- **Benigno:** estos tumores no son cancerosos. No invaden el tejido cercano ni se propagan a otras partes del cuerpo. Si un médico los extrae, generalmente no regresan.
- **Premaligno:** en estos tumores, las células aún no son cancerosas, pero pueden potencialmente volverse malignas.
- **Maligno:** Los tumores malignos son cancerosos. Las células pueden crecer y propagarse a otras partes del cuerpo.

(Brazier & Rush, 2022).

Tabla 1

Tipos y ejemplos de tumores

Benigno	Pre Maligno	Maligno
La mayoría de los tumores benignos no son dañinos y es poco probable que afecten otras partes del cuerpo. Sin embargo, pueden causar dolor u otros problemas si presionan los nervios o los vasos sanguíneos o desencadenan la sobreproducción de hormonas, como en el sistema endocrino.	Este tipo de tumor no es canceroso, pero los médicos los controlarán de cerca para detectar cambios.	Los tumores malignos son cancerosos. Se desarrollan cuando las células crecen sin control. Si las células continúan creciendo y propagándose, la enfermedad puede convertirse en una amenaza para la vida. Los tumores malignos pueden crecer rápidamente y diseminarse a otras partes del cuerpo en un proceso llamado metástasis. Sin embargo, no todos los tumores malignos crecen rápidamente; algunos pueden crecer mucho más lentamente con el tiempo.
	Ejemplos	
o Adenomas o Fibromas o Hemangiomas o Lipomas	o Queratosis actínica o Displasia cervical o Metaplasia del pulmón o Leucoplasia	o Carcinoma o Sarcoma o Tumor de células germinales o Blastema o Meningioma

Nota. Los datos de la tabla se obtuvieron de la página de “Medical News Suporter” - https://www.medicalnewstoday.com/articles/249141 (Brazier & Rush, 2022).

Cáncer

Es una enfermedad que se produce cuando las células del cuerpo crecen y se multiplican de forma descontrolada y de forma autónoma, invadiendo de forma física a otras regiones de tejidos, ya que las células cancerosas pueden desprenderse del sitio donde comenzó el cáncer, viajando a otras partes del cuerpo, pudiendo terminar en los ganglios linfáticos u otros órganos del cuerpo causando problemas con las funciones normales.

Es posible que el cáncer comience en cualquier parte del cuerpo humano, formado por billones de células. En condiciones normales, las células humanas se forman y se multiplican (mediante un proceso que se llama división celular) para formar células nuevas a medida que el cuerpo las necesita. Cuando las células envejecen o se dañan, mueren y las células nuevas las reemplazan.

A veces el proceso no sigue este orden y las células anormales o células dañadas se forman y se multiplican cuando no deberían. Estas células tal vez formen tumores, que son bultos de tejido. Los tumores son cancerosos (malignos) o no cancerosos (benignos). (Instituto Nacional del Cáncer de EE.UU., 2021)

Interpretando los resultados de la revista Scientific American, Inc. (1996) sugiere que “Casi todos los tejidos del cuerpo pueden generar tumores malignos, y algunos de estos tejidos producen más de un tipo de estos tumores. Cada tipo de cáncer tiene características únicas.” (Weinberg, pág. 1).

Causas del cáncer.

Las células cancerosas se desarrollan debido a múltiples variaciones en los genes, estas variaciones pueden tener muchas causas posibles como, por ejemplo:

- **Fumar**: Fumar cigarrillos está directamente relacionado como la causa del cáncer en humanos. Un tercio de las muertes en los Estados Unidos se han atribuido al cáncer causado por fumar cigarrillos. El cáncer de pulmón ocurre principalmente debido a una combinación de factores que dañan la funcionalidad de múltiples órganos como la laringe, la cavidad oral y el esófago.
- **Estilo de vida:** Los estilos de vida que dependen en gran medida de la dieta juegan un papel importante en la causa del cáncer. Numerosos estudios validan el hecho de que un tercio de las muertes por cáncer en los Estados Unidos ocurren debido a una dieta inadecuada y factores relacionados con el estilo de vida. Los factores contribuyentes son los tipos de alimentos, los volúmenes, las variedades, incluidos los alimentos procesados, y un grave desequilibrio en calorías.
- **Genética:** Según la definición, el cáncer es una enfermedad genética. Los genes son moléculas muy pequeñas en las células del cuerpo, que determinan todo acerca de un ser humano. Los genes están controlados por la genética y la herencia de cada célula. En los tumores cancerosos, varias células genéticas son anormales y estas células anormales se fomentan debido a factores como virus, división celular incontrolable, etc. Los cánceres como el de mama, cerebro y endometrio son causados por estos factores identificados.
- **El ambiente alrededor:** La vinculación de las condiciones de salud humana con el medio ambiente se ha identificado como otra causa. Las personas que han trabajado en un entorno de fumadores de cigarrillos tienen un mayor riesgo de desarrollar cáncer de pulmón. Numerosos productos químicos identificados por

científicos e investigadores que se sabe que causan cáncer ahora están prohibidos en todo el mundo.

- **Agentes infecciosos:** Los virus causan cáncer en el cuerpo humano. Los virus cambian la funcionalidad de las células y generan anomalías. Por ejemplo, el virus de Epstein-Barr crea tumores de linfoma de Burkitt en niños africanos.
- El virus de la hepatitis B es responsable del cáncer de hígado a nivel mundial.
- Muchas veces, no hay una causa obvia.

(Shaikh, Krishnan, & Thanki, Artifcial Intelligence in Breast Cancer Early Detection and Diagnosis, 2021, pág. 6).

Propagación del cáncer.

El cáncer puede diseminarse desde donde comenzó (el sitio primario) a otras partes del cuerpo.

Cuando las células cancerosas se desprenden de un tumor, pueden viajar a otras áreas del cuerpo a través del torrente sanguíneo o del sistema linfático. Las células cancerosas que viajan a través del torrente sanguíneo pueden llegar a órganos distantes. Si viajan a través del sistema linfático, las células cancerosas pueden terminar en los ganglios linfáticos. De cualquier manera, la mayoría de las células cancerosas que escapan mueren o mueren antes de que puedan comenzar a crecer en otro lugar. Pero uno o dos pueden asentarse en una nueva área, comenzar a crecer y formar nuevos tumores. Esta propagación del cáncer a una nueva parte del cuerpo se llama metástasis. (The, 2020, pág. 3)

El cáncer metastásico se llama igual que el cáncer primario. Por ejemplo, el cáncer de seno (mama) que se diseminó al pulmón se llama cáncer de seno metastásico,

no cáncer de pulmón. El tratamiento es para el cáncer de seno en estadio IV, no para el cáncer de pulmón.

A veces cuando las personas reciben un diagnóstico de cáncer metastásico, los médicos no logran saber dónde se formó. Este tipo de cáncer se llama cáncer de origen primario desconocido (CPD). (Instituto Nacional del Cancer, 2020)

Dimensión en Perú.

El Instituto Nacional de Enfermedades Neoplásicas (INEN) registró durante el 2021 un promedio de 17,500 nuevos casos de cáncer entre varones y mujeres, cifra que aumentó en un 40% en comparación con la estadística del 2020; la mayoría de los pacientes son del interior del país.

Asimismo, el año pasado, a pesar de los inconvenientes ocasionados por la pandemia, se realizaron un promedio de 362,000 atenciones en consultorios externos; 4,700 cirugías mayores; 46,600 quimioterapias; y 67,000 radioterapias. En relación con las atenciones para diagnósticos y controles de la enfermedad, el INEN reportó que el departamento de patología realizó 4 millones 500,000 pruebas, mientras que el de radiodiagnóstico, 154,000. En la actualidad, los pacientes entre nuevos y continuadores que visitan esta institución especializada provienen en un 57% de Lima y Callao y el 43% de provincias. (El Peruano, 2022)

A continuación, se lista las estadísticas de crecimiento de casos de cáncer entre el año 2000 y 2019 según el Instituto Nacional de Enfermedades Neoplásicas del Perú.

Tabla 2

Casos nuevos de cáncer registrados en INEN - ambos sexos del 2000 al 2019

Localización	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Cérvix	1319	1360	1402	1337	1379	1359	1532	1500	1621	1593	1568	1611	1639	1600	1485	1585	1632	1415	1499	1505
Mama	1027	1008	1019	1014	1023	1035	1163	1111	1111	1199	1240	1275	1349	1274	1221	1441	1491	1305	1374	1391
Estomago	561	609	615	563	632	626	676	715	754	802	778	786	906	904	925	1010	978	870	901	823
Linfoma no hodgkin	456	442	477	447	503	531	554	531	515	535	553	514	568	652	582	591	629	597	610	616
Próstata	264	307	310	351	365	436	474	536	511	512	509	491	608	585	634	697	624	596	575	614
Piel no melanoma	344	309	341	332	361	408	400	395	395	454	464	482	526	519	575	712	662	560	530	469
Tiroides	240	211	227	265	300	288	287	267	308	300	356	404	480	496	568	620	637	643	649	662
Pulmón	336	310	328	348	364	377	406	420	464	446	411	428	401	412	453	406	461	434	449	374
Leucemia linfoide	247	271	276	272	281	292	320	314	323	310	295	339	350	359	365	380	411	371	374	404
Sts. Nervioso central	161	204	185	223	208	178	209	201	181	199	205	241	258	309	329	381	389	379	392	356
Cavidad oral	155	167	206	189	244	230	228	248	244	223	277	285	302	279	310	303	294	308	290	307
Colon	144	128	168	177	168	193	226	221	226	257	302	273	286	310	282	372	355	296	334	338
Primario desconocido	187	180	244	210	239	175	188	200	207	226	168	234	214	243	211	244	271	243	223	233
Riñón	127	127	118	116	141	132	189	199	208	248	190	208	248	246	300	330	297	243	256	240
Tej. blandos y peritoritoneo	127	160	187	166	198	192	223	213	205	209	217	229	232	225	259	228	233	214	217	217
Recto	128	115	126	121	165	146	159	174	197	229	211	210	238	251	225	242	337	252	271	282
Leucemia mieloide	149	149	167	166	194	184	200	177	195	196	187	217	178	241	212	233	245	223	263	240
Ovario	182	157	167	187	156	180	198	163	198	203	240	209	200	204	199	185	211	191	198	158
Hígado	144	123	152	144	146	155	159	189	185	167	187	199	195	184	212	224	254	234	241	159
Vesicular biliar	94	108	97	111	128	117	130	146	135	179	156	173	193	189	185	211	193	207	190	181
Melanoma de piel	103	116	117	115	127	132	164	166	134	155	128	148	204	194	204	169	194	153	185	156
Páncreas	83	97	103	102	88	131	127	125	132	162	150	169	188	173	192	192	201	209	193	188
Testículo	122	129	136	153	149	142	165	163	156	145	133	143	152	130	140	152	165	109	150	174
Cuerpo uterino	81	78	93	101	105	101	124	116	123	136	144	171	141	175	167	173	178	160	156	180

Vejiga	88	91	93	94	105	113	108	114	90	108	107	116	141	134	164	130	150	135	129	134
Huesos y cartílago	68	70	92	80	86	100	107	101	90	107	93	97	114	96	126	99	112	107	133	82
Ojo	53	62	81	51	79	69	69	80	83	72	81	100	78	94	99	122	127	114	127	107
Mieloma	48	51	45	40	72	56	66	57	56	61	63	67	78	76	110	94	112	96	102	125
Linfoma de hodgkin	72	61	59	68	76	67	82	68	72	69	60	56	68	79	83	64	81	72	59	79
Esófago	45	40	45	52	36	49	59	74	74	80	66	71	71	73	72	74	83	73	72	70
Ano	52	35	50	45	44	56	57	62	42	71	60	59	73	72	77	94	64	71	90	93
Laringe	69	61	68	45	57	52	59	67	51	68	57	71	58	77	64	59	71	66	61	56
Vías biliares	36	30	27	30	51	33	48	61	57	69	57	68	77	79	77	68	93	96	74	75
Pene	24	36	27	40	42	40	40	41	48	40	46	61	46	49	42	45	64	60	67	54
Otras leucemias	5	8	14	26	29	40	46	46	52	44	47	58	56	69	71	59	58	60	36	29
Vulva	33	29	40	22	38	29	39	29	39	44	44	42	50	49	41	47	49	42	42	47
Senos paranasales	34	32	31	33	31	47	45	36	44	36	33	41	45	46	40	39	39	44	35	30
Fosa nasal	33	26	27	31	27	31	25	30	28	33	32	26	30	34	42	38	37	36	39	39
Coriocarcinoma	37	29	27	45	34	27	36	38	27	30	20	20	23	18	16	23	23	17	16	13
Otros	198	175	166	183	187	198	239	213	208	215	186	209	227	232	253	287	265	246	263	257
Total	7676	7701	8153	8095	8658	8747	9626	9607	9789	10232	10121	10601	11291	11431	11612	12423	12770	11547	11865	11557

Nota. Tabla obtenida del Instituto Nacional de Enfermedades Neoplásicas - https://portal.inen.sld.pe/wp-content/uploads/2022/08/Casos-nuevos-registrados-en-el-INEN-2000-2019.pdf

Métodos de diagnóstico de cáncer

El diagnóstico de cáncer tiene como objetivo identificar el sitio original de las células cancerosas y el tipo de células anormales presentes en él. El cáncer puede desarrollarse en cualquier parte del cuerpo humano excepto en las uñas, el cabello y los dientes. Aquí, el sitio se refiere a la ubicación del cáncer dentro del cuerpo. El órgano del cuerpo en el que se desarrolla inicialmente el cáncer se conoce como sitio primario. Estos sitios brindan información adecuada, como el comportamiento del tumor, hacia dónde y en qué dirección se puede propagar y qué síntomas puede causar el tumor. Los sitios más comunes en el cuerpo humano son la piel, los pulmones, los senos femeninos, la próstata, el colon, el recto y el cuerpo uterino. El sitio secundario se refiere a la parte del cuerpo donde crecen las células cancerosas. El cáncer siempre se describe por el sitio primario, incluso si se ha diseminado a otra parte del cuerpo.

Con el avance de la ciencia médica, los síntomas suelen indicar la presencia de cáncer, y estos pueden observarse directamente o a través de diversas tecnologías de imagen, como la tomografía computarizada (TC), la resonancia magnética nuclear (RMN), etc., o confirmarse mediante diversas pruebas. en un laboratorio, por ejemplo, la orina rosada o rojiza puede ser causada por una infección en el riñón o por cáncer. Un análisis de sangre puede confirmar un síntoma.

Se prefiere una biopsia para diagnosticar el cáncer, lo que implica la extirpación del tejido afectado y el examen a través de un microscopio. El muestreo de tejido es otro método que puede recuperar fácilmente un tumor de la superficie del cuerpo. Si el tumor es inaccesible, las tecnologías de imágenes son efectivas para localizar visualmente un tumor antes de realizar la biopsia. Un tipo histológico de cáncer se puede diagnosticar

fácilmente mediante un examen microscópico del tumor. Las biopsias con tecnologías de imágenes se utilizan ampliamente para la confirmación del cáncer en sus ubicaciones primarias y, posiblemente, de la ubicación en la que se puede propagar.

Es primordial identificar qué tipos de células están presentes en un tumor; los diferentes tipos de cáncer tienen diferentes tasas de aumento que siguen siendo de naturaleza diferente. Múltiples tipos de células pueden estar presentes en el mismo tumor. Así, una vez confirmado el cáncer, la identificación celular es obligatoria para conocer su efecto sobre las células sanas.

(Shaikh & Krishnan, Artificial Intelligence in Breast Cancer Early Detection and Diagnosis, 2021, págs. 8-9)

Cáncer de Mama

Cuando las células sanguíneas de la mama se vuelven recalcitrantes, la afección se conoce como cáncer de mama. Existen varios tipos de cáncer de mama, según las células afectadas; las tres partes principales de los senos son los lóbulos, los conductos y el tejido conectivo. Los lóbulos son donde se produce la leche, mientras que los conductos son canales que llevan la leche al pezón. El tejido conectivo rodea y mantiene todo unido. La mayoría de los cánceres ocurren en los lobulillos o los conductos y se pueden propagar a otros órganos del cuerpo a través de los vasos sanguíneos. Los tipos más comunes de cáncer de mama son los siguientes:

Carcinoma ductal invasivo: en este tipo de cáncer, las células cancerosas se producen fuera de los conductos.

Carcinoma lobulillar invasivo: en este tipo de cáncer, las células cancerosas se producen fuera de los lobulillos.

(Shaikh & Krishnan, Artificial Intelligence in Breast Cancer Early Detection and Diagnosis, 2021, págs. 12-13)

Síntomas del cáncer de mama.

Hay otros tipos de cáncer de mama como el cáncer de mama medular, mucinoso e inflamatorio. Varios pacientes tienen diferentes síntomas de cáncer de mama. Algunos pacientes pueden no tener todos los síntomas o ningún síntoma. Sin embargo, hay algunos síntomas o signos de cáncer:

- Nuevo bulto en el seno o la axila
- Engrosamiento o hinchazón de cualquier parte del seno
- Irritación o formación de hoyuelos en la piel del seno
- Enrojecimiento o piel escamosa en el área del pezón o el seno
- Tirar del pezón o dolor en el área del pezón
- Secreción del pezón distinta de la leche materna, incluida la sangre
- Cualquier cambio en el tamaño o la forma del seno
- Dolor en cualquier área del seno

(Shaikh & Krishnan, Artificial Intelligence in Breast Cancer Early Detection and Diagnosis, 2021, pág. 13)

Factores de riesgo del cáncer de mama.

Un factor que aumenta las posibilidades de contraer una enfermedad se denomina factor de riesgo. Pero tener un factor de riesgo no significa que sea probable que contraiga la enfermedad.

Los factores que están asociados con un mayor riesgo de cáncer de mama incluyen:

- **Ser mujer:** Las mujeres son mucho más propensas que los hombres a desarrollar cáncer de mama.
- **Edad creciente:** Su riesgo de cáncer de mama aumenta a medida que envejece.
- Antecedentes personales de afecciones mamarias. Si se sometió a una biopsia de seno que encontró carcinoma lobulillar in situ (LCIS) o hiperplasia atípica del seno, tiene un mayor riesgo de cáncer de seno.
- **Antecedentes personales de cáncer de mama:** Si ha tenido cáncer de mama en un seno, tiene un mayor riesgo de desarrollar cáncer en el otro seno.
- **Antecedentes familiares de cáncer de mama:** Si a su madre, hermana o hija le diagnosticaron cáncer de mama, especialmente a una edad temprana, su riesgo de cáncer de mama aumenta. Aun así, la mayoría de las personas diagnosticadas con cáncer de mama no tienen antecedentes familiares de la enfermedad.
- **Genes heredados que aumentan el riesgo de cáncer:** Ciertas mutaciones genéticas que aumentan el riesgo de cáncer de mama pueden transmitirse de padres a hijos. Las mutaciones genéticas más conocidas se denominan BRCA1 y BRCA2. Estos genes pueden aumentar en gran medida el riesgo de cáncer de mama y otros tipos de cáncer, pero no hacen que el cáncer sea inevitable.

- **Exposición a la radiación:** Si recibió tratamientos de radiación en el pecho cuando era niño o adulto joven, su riesgo de cáncer de mama aumenta.
- **Obesidad:** Ser obeso aumenta el riesgo de cáncer de mama.
- **Comenzar su período a una edad más temprana:** Comenzar su período antes de los 12 años aumenta su riesgo de cáncer de mama.
- **Comienzo de la menopausia a una edad más avanzada:** Si comenzó la menopausia a una edad mayor, es más probable que desarrolle cáncer de mama.
- **Tener su primer hijo a una edad mayor:** Las mujeres que dan a luz a su primer hijo después de los 30 años pueden tener un mayor riesgo de cáncer de mama.
- **Nunca haber estado embarazada:** Las mujeres que nunca han estado embarazadas tienen un mayor riesgo de cáncer de mama que las mujeres que han tenido uno o más embarazos.
- **Terapia hormonal posmenopáusica:** Las mujeres que toman medicamentos de terapia hormonal que combinan estrógeno y progesterona para tratar los signos y síntomas de la menopausia tienen un mayor riesgo de cáncer de mama. El riesgo de cáncer de mama disminuye cuando las mujeres dejan de tomar estos medicamentos.
- **Bebiendo alcohol:** El consumo de alcohol aumenta el riesgo de cáncer de mama.

(Mayo Clinic, Breast cancer, 2022)

ADN

Las instrucciones que determinan todas las características y funciones de un organismo se encuentran en su material genético: el ADN (ácido desoxirribonucleico).

El conocimiento del ADN, su estructura y función, fue determinante para el desarrollo de la biotecnología moderna.

La estructura de doble hélice del ADN, que los investigadores James Watson y Francis Crick propusieran en el año 1953 proporcionó respuestas a muchas preguntas que se tenían sobre la herencia. Predijo la autorreplicación del material genético y la idea de que la información genética estaba contenida en la secuencia de las bases que conforman el ADN. Más aún, con el correr de los años y de las investigaciones, se pudo determinar que todos los seres vivos contienen un ADN similar, formado a partir de las mismas unidades: los nucleótidos. Este código genético mediante el cual se "escriben" las instrucciones celulares es común a todos los organismos. Es decir que el ADN de un ser humano puede ser "leído" dentro de una bacteria, y una planta puede interpretar la información genética de otra planta diferente. A esta propiedad de la información genética se la conoce como "universalidad del código genético".

El código genético universal es uno de los conceptos básicos para comprender los procesos de la biotecnología moderna. Por ejemplo, la posibilidad de generar organismos transgénicos, y que las instrucciones del ADN de un organismo puedan determinar nuevas características en organismos totalmente diferentes. (Chile BIO, s.f.).

La función del ADN.

El ADN tiene la función de "guardar información". Es decir, contiene las instrucciones que determinan la forma y características de un organismo y sus funciones.

Además, a través del ADN se transmiten esas características a los descendientes durante la reproducción, tanto sexual como asexual. Todas las células, procariotas y eucariotas, contienen ADN en sus células. En las células eucariotas el ADN está contenido dentro del núcleo celular, mientras que, en las células procariotas, que no tienen un núcleo definido, el material genético está disperso en el citoplasma celular. (Chile BIO, s.f.).

La estructura del ADN.

El ADN está organizado en cromosomas. En las células eucariotas los cromosomas son lineales, mientras que los organismos procariotas, como las bacterias, presentan cromosomas circulares. Para cada especie, el número de cromosomas es fijo. Por ejemplo, los seres humanos tienen 46 cromosomas en cada célula somática (no sexual), agrupados en 23 pares, de los cuales 22 son autosomas y un par es sexual. Una mujer tendrá un par de cromosomas sexuales XX y un varón tendrá un par XY.

Cada cromosoma tiene dos brazos, ubicados por arriba y por debajo del centrómero. Cuando los cromosomas se duplican, previo a la división celular, cada cromosoma está formado por dos moléculas de ADN unidas por el centrómero, conocidas como cromátidas hermanas.

El ADN se compone de dos cadenas, cada una formada por nucleótidos. Cada nucleótido, a su vez, está compuesto por un azúcar (desoxirribosa), un grupo fosfato y una base nitrogenada. Las bases nitrogenadas son cuatro: adenina (A), timina (T), citosina (C), y guanina (G), y siempre una A se enfrenta a una T y una C se enfrenta a una G en la doble cadena. Las bases enfrentadas se dice que son complementarias. El ADN adopta una forma de doble hélice, como una escalera caracol donde los lados son cadenas de azúcares y fosfatos conectadas por “escalones”, que son las bases

nitrogenadas. La molécula de ADN se asocia a proteínas, llamadas histonas, y se encuentra muy enrollada y compactada para formar el cromosoma.

La doble hélice de ADN con las bases nitrogenadas complementarias que se ubican hacia dentro y establecen uniones no covalentes (o fuerzas de atracción) entre sí que mantienen la estructura de la molécula. Las desoxirribosas (azúcares) y los grupos fosfato constituyen las columnas de la molécula.

Cuando la célula se divide, cada nueva célula que se forma debe portar toda la información genética, que determine sus características y funciones. Para eso, antes de dividirse, el ADN debe replicarse, es decir generar una copia de sí mismo. Durante la replicación, la molécula de ADN se desenrolla, separando sus cadenas. Cada una de éstas servirá como molde para la síntesis de nuevas hebras de ADN. Para eso, la enzima ADN-polimerasa coloca nucleótidos siguiendo la regla de apareamiento A-T y C-G. El proceso de replicación del ADN es semiconservativo, ya que, al finalizar la duplicación, cada nueva molécula de ADN estará conformada por una hebra "vieja" (original) y una nueva. (Chile BIO, s.f.).

Genes.

Los genes son las unidades básicas de la herencia en todos los organismos vivos. Un gen es un segmento de ADN, las moléculas que contienen instrucciones para el desarrollo y funcionamiento de los organismos vivos. Se estima que los humanos tenemos aproximadamente 30.000 genes que componen nuestro genoma. El ADN se encuentra dentro del núcleo de cada célula y está organizado en estructuras llamadas cromosomas. Los seres humanos tienen 46 cromosomas: dos conjuntos de 23, uno de los cuales proviene de cada padre. Se puede pensar en el genoma humano como un conjunto

de enciclopedias con 23 volúmenes, donde cada cromosoma representa un volumen. El código de ADN es como las letras que se utilizan para construir las palabras, párrafos y páginas de texto en esos volúmenes. Debido a que los genes varían en tamaño, se pueden considerar como un solo párrafo o un capítulo completo dentro de cada volumen. (American Civil Liberties Union, 2009).

Un gen no es una estructura que se vea, sino que se define a nivel funcional. Es una secuencia que va a empezar en algún lugar del ADN y va a terminar en otro. Para conocer un gen se secuencia, se determina la cantidad de los nucleótidos que lo forman y el orden en que se ubican.

Todas las células de un organismo tienen el mismo genoma, o conjunto de genes. Pero, en cada célula se expresan los genes que se usan. Por ejemplo, aunque una célula de la piel tiene toda la información genética al igual que la célula del hígado, en la piel solo se expresarán aquellos genes que den características de piel, mientras que los genes que dan características de hígado estarán allí "apagados". Por el contrario, los genes que dan rasgos de "hígado" estarán activos en el hígado e inactivos en la piel. Lo que no se usa se encuentra mayormente compactado. Este empaquetamiento puede ser temporal o definitivo. (Chile BIO, s.f.).

Proteínas.

Las proteínas son macromoléculas que cumplen funciones variadas. Hay proteínas estructurales, otras son enzimas, otras transportan oxígeno como la hemoglobina, hay proteínas involucradas en la defensa inmunitaria, como los anticuerpos, otras cumplen funciones de hormonas como la insulina, etc.

Así como el ADN está compuesto a partir de nucleótidos, las proteínas están compuestas a partir de aminoácidos. Hay 20 aminoácidos diferentes, y cada proteína tiene una secuencia de aminoácidos particular.

El proceso de síntesis de proteínas consta básicamente de dos etapas: la transcripción y la traducción. En la primera etapa, las "palabras" (genes) escritas en el ADN en el lenguaje de los nucleótidos se copian o transcriben a otra molécula, el ARN mensajero (ARNm). Luego, en la etapa siguiente, el ARNm se traduce al idioma de las proteínas, el de los aminoácidos. Este flujo de información se conoce como el "dogma central de la biología". (Chile BIO, s.f.).

Código genético.

El código genético se refiere a las instrucciones que contiene un gen y que le indican a una célula cómo producir una proteína específica. El código de cada gen usa las cuatro bases nitrogenadas del ADN — adenina (A), citosina (C), guanina (G) y timina (T) — de diversas maneras para deletrear los "codones" de tres letras que especifican qué aminoácido se necesita en cada posición dentro de una proteína.

El código genético es el término que usamos para nombrar la forma en que las cuatro bases del ADN - A, C, G y T - se encadenan de forma que la maquinaria celular, el ribosoma, pueda leerlos y convertirlos en una proteína. En el código genético, cada tres nucleótidos consecutivos actúan como un triplete que codifica un aminoácido. De este modo cada tres nucleótidos codifican para un aminoácido. Las proteínas se componen a veces de cientos de aminoácidos. Así que el código de una proteína podría contener cientos, a veces incluso miles, de tripletes. (National Human Genome Research Institute, 2022).

Genes BRCA

Los genes BRCA1 y BRCA2 (Breast cancer 1 and 2, DNA repair associated) son genes humanos que codifican proteínas supresoras de tumores. Las proteínas codificadas desempeñan un papel en la reparación del ADN dañado, y por tanto tienen una función importante en asegurar la estabilidad del material genético.

Los genes BRCA1 y BRCA2 se heredan mediante un patrón autosómico dominante, pues están situados en los cromosomas 17 y 13, respectivamente, y es suficiente heredar un único alelo para tener un riesgo aumentado de cáncer. La penetrancia del rasgo "cáncer hereditario", aunque incompleta, es elevada, pues un alto porcentaje de los portadores del gen mutado desarrollarán un cáncer de mama u ovario.

Además, en la mujer, las mutaciones en BRCA1 se han relacionado con un aumento de riesgo de cáncer de las trompas de Falopio y de cáncer primario del peritoneo.

En el varón, las mutaciones en BRCA2 y, en menor medida, en BRCA1, se han relacionado con un aumento de riesgo de cáncer de mama. Las mutaciones dañinas en BRCA1 y BRCA2 se han asociado también a un riesgo aumentado de cáncer de próstata.

En ambos sexos, las mutaciones en BRCA1 y BRCA2 se han revelado como causantes de un aumento en el riesgo de cáncer de páncreas. Las mutaciones en BRCA2 (conocidas como FANCD1), cuando se heredan de ambos padres, pueden provocar un subtipo de anemia de Fanconi, un síndrome que se asocia a tumores sólidos en los niños y a Leucemia Mieloide Aguda. Por su parte, las mutaciones en BRCA1 (también denominadas FANCS), cuando se heredan también de ambos padres, están relacionadas con otro subtipo de anemia de Fanconi. (Álvarez Gama, 2016, pág. 3).

BRCA1.

El gen BRCA1 está situado en el brazo largo del cromosoma 17 (17q21.31), a lo largo de una secuencia de unos 79 kb de longitud, y consta de 24 exones (figura 1), destacando sobre el resto la longitud del exón 1110.

La traducción de este gen consiste en un producto de 1863 aminoácidos de longitud y unos 220 kDa de masa (figura 2). En esta proteína destacan los siguientes dominios:

- Un dominio con un dedo de zinc (RING). Este dominio es crucial en la unión con la proteína BARD1, que forma un heterodímero BRCA1/BARD1 que parece ser esencial en la actividad antitumoral de BRCA1. Por otro lado, este dominio desempeña un papel en la ubiquitinación de proteínas, es decir, el marcaje de estas para su destrucción, lo que a su vez es importante para la regulación del ciclo celular. En esta función intervienen también las proteínas BARD1 y BRCA2
- Un dominio rico en serina (SCD, Serine Cluster Domain). Parte de este dominio corresponde a los exones 11 a 13, una zona de alta frecuencia de mutaciones. La fosforilación de estos residuos de serina por quinasas ATM/ATR, que a su vez son activadas por el daño al ADN, permite a la proteína BRCA1 y los polímeros de que forma parte situarse junto al ADN dañado y proceder a su reparación. Por ello, mutaciones en estos residuos de serina producen una disminución de la actividad antitumoral de la proteína.

- Dos dominios carboxiterminales de BRCA1 (BRCT). El extremo carboxiterminal de la proteína BRCA1 contiene dos dominios BRCT, sujetos a diversas variaciones. Estos dominios son esenciales en la reparación del ADN, la regulación de la transcripción y la función de supresión tumoral. Mutaciones de cambio de sentido en estas regiones afectan de forma devastadora la función de la proteína, incrementando significativamente el riesgo de cáncer.

(Álvarez Gama, 2016, págs. 5-7).

BRCA2.

El gen BRCA2 está situado en el brazo largo del cromosoma 13 (13q13.1), a lo largo de una secuencia de unos 84 kb de longitud, y consta de 27 exones (figura 3), destacando sobre el resto la longitud del exón 11, y en menor medida, la del exón 10.

La traducción de este gen consiste en un producto de 3418 aminoácidos de longitud y unos 348 kDa de masa (figura 4). En esta proteína destacan los siguientes dominios:

- Un extremo amino-terminal relacionado con la unión a PALB2, que es otra proteína supresora tumoral
- Un dominio que contiene 8 repeticiones BRC (unas repeticiones de unos 35 aminoácidos), crítico para la unión a RAD51, que es una proteína cuya función está relacionada con la reparación del ADN de doble cadena

- Un dominio helicoidal (H), que se une al polipéptido DSS1, la deleción de cuyo gen está relacionado con un síndrome que se presenta con un conjunto de malformaciones congénitas
- Un dominio OB1, con estructura de barril beta, relacionado también con la unión a DSS1, así como con la unión a ADN de cadena sencilla.
- Un dominio OB3, también con estructura de barril beta y relacionado asimismo con la unión a ADN de cadena sencilla.
- Un dominio Tower (T), relacionado con la unión a ADN de doble cadena, esencial en la función antitumoral de la proteína BRCA2.
- Un extremo carboxiterminal que contiene una NLS (una señal de localización nuclear, esencial para el transporte de la proteína al núcleo celular, donde desempeña su función), y un sitio de fosforilación por la quinasa dependiente de ciclinas CDK2, que también es un lugar de unión a RAD51.

(Álvarez Gama, 2016, págs. 8-9).

Alteraciones Genéticas

¿Qué es una mutación?

A los cambios estables en la cadena de ADN que son capaces de ser heredados, se les conoce como mutaciones.

Las mutaciones realmente transcendentes para la descendencia son las que están presentes u ocurren en las células germinales (óvulos y espermatozoides).

Las mutaciones que se producen entonces pueden dar lugar a pequeños cambios, grandes cambios (causando enfermedad: mutaciones patógenas) o ser silentes.

A la mutación que heredamos de nuestros padres se le llama mutación heredada, a la que se da en el individuo sin que haya un progenitor con la misma mutación, se le conoce como mutación de Novo.

Tipos de mutaciones

Las mutaciones pueden darse en tres niveles diferentes:

- Molecular (génicas o puntuales): Son mutaciones a nivel molecular y afectan la constitución química de los genes, es decir a las bases o "letras" del ADN.
- Cromosómico: El cambio afecta a un segmento de cromosoma (de mayor tamaño que un gen), por tanto, a su estructura. Estas mutaciones pueden ocurrir porque grandes fragmentos se pierden (deleción), se duplican, cambian de lugar dentro del cromosoma.
- Genómico: Afecta al conjunto del genoma, aumentando el número de juegos cromosómicos (poliploidía) o reduciéndolo a una sola serie (haploidía o mono ploidía) o bien afecta al número de cromosomas individualmente (por defecto o por exceso), como la trisomía 21 o Síndrome de Down.

Podemos clasificar los diferentes tipos de mutaciones en:

1. Mutaciones silenciosas

En este tipo de mutación hay un cambio en una de las bases del ADN de forma que el triplete de nucleótidos se modifica, pero sigue codificando para el mismo aminoácido. Esto es así porque el código genético tiene cierto margen de

seguridad y para cada aminoácido hay varias combinaciones de tripletes que lo determinan.

Por ejemplo, lo tripletes CCA y CCC determinan que en esta posición de la proteína se sitúe una prolina. Así, si se produce por error este cambio, será un cambio silente, porque el aminoácido codificado por ambos tripletes es el mismo, la prolina.

2. Polimorfismos

En este tipo de mutaciones hay un cambio de una de las bases de ADN, de tal manera que el triplete de nucleótidos que es una parte se cambia, pero incluso si se necesita un cambio de aminoácido, el aminoácido que entra en el lugar en cuestión resulta tener poco o ningún impacto en la función de la proteína.

Los polimorfismos pueden incluso conducir a una reducción de la función de la proteína en cuestión, pero por sí sola no es suficiente para causar la enfermedad (de lo contrario no serían llamados polimorfismos, pero mutaciones patógenas). Ellos pueden, sin embargo, ser factores de riesgo cuando más de una junta.

3. Missense mutation

En este tipo de mutación hay un cambio en una de las bases del ADN de forma que el triplete codifica para un aminoácido diferente del que debería, es decir, en esa posición de la proteína habrá un aminoácido incorrecto, lo que puede alterar más o menos la función de la proteína dependiendo de su localización e importancia.

4. Nonsense mutation

En este tipo de mutación hay un cambio en una de las bases del ADN de forma que el nuevo triplete que se forma determina la señal de fin de la cadena de aminoácidos. Esto es, se trunca la proteína, no se continúa formando a partir de ahí. Según dónde quede truncada la proteína será capaz de preservar algo de función o no.

5. Inserción

En este tipo de mutación se añade una o más bases al ADN original. De esta forma se puede alterar el marco de lectura para formar la proteína o insertar aminoácidos extra que son inadecuados.

6. Deleción

En este tipo de mutación se pierden una o más bases, es decir, se pierde un trozo de ADN alterando la cadena proteica que debería formarse y su función. De esta forma se puede alterar el marco de lectura para formar la proteína o eliminar aminoácidos que son propios de la cadena proteica. En ocasiones las deleciones son tan largas que pueden comprometer un gen entero o varios genes contiguos.

7. Duplicación

En este tipo de mutación hay un fragmento de ADN que está copiado una o varias veces, lo que altera la formación de la cadena de aminoácidos y la función de la proteína. De esta forma se puede alterar el marco de lectura (ver punto 8) para formar la proteína o insertar aminoácidos extra que son inadecuados.

8. Cambio de marco de lectura (Frameshift mutation)

Este tipo de mutación se da cuando por inserción o pérdida de pares de bases se cambia el marco de lectura. Para la decodificación, las bases se leen de tres en tres, esto es, cada tres bases determinan un aminoácido.

Si se cambia el marco de lectura, cambia la forma de agrupar esas tres bases y se colocaran aminoácidos erróneos habiendo la posibilidad de un triplete STOP prematuro. Las inserciones, duplicaciones y deleciones pueden dar lugar a este tipo de mutaciones.

9. Expansión por repetición

Muchas veces no son consideradas mutaciones puntuales. Se trata de repeticiones de tripletes o cuatripletes de nucleótidos, pequeñas secuencias de ADN de 3 o 4 pares de bases que se repiten en serie.

Una mutación por expansión es una mutación en la que el número de repeticiones ha aumentado, lo que puede hacer que la proteína final no funcione correctamente.

Enfermedades paradigmáticas en este tipo de mutaciones son el Síndrome de X Frágil o las Ataxias Espinocerebelosas (SCA). En este último caso se repite el triplete de nucleótidos CAG de forma que determina una gran cadena de glutaminas (poli glutamina).

(DIVULGACIÓN MÉDICA, 2013).

Clasificación de variantes genéticas.

La secuencia de ADN de un gen se puede alterar de varias formas. Las variantes genéticas (también conocidas como mutaciones) pueden tener diversos efectos sobre la

salud, según dónde se produzcan y si alteran la función de las proteínas esenciales. Los tipos de mutaciones incluyen:

Sustitución

Este tipo de variante reemplaza un bloque de construcción de ADN (nucleótido) por otro. Además, las variantes de sustitución pueden clasificarse por el efecto que tienen sobre la producción de proteína a partir del gen alterado.

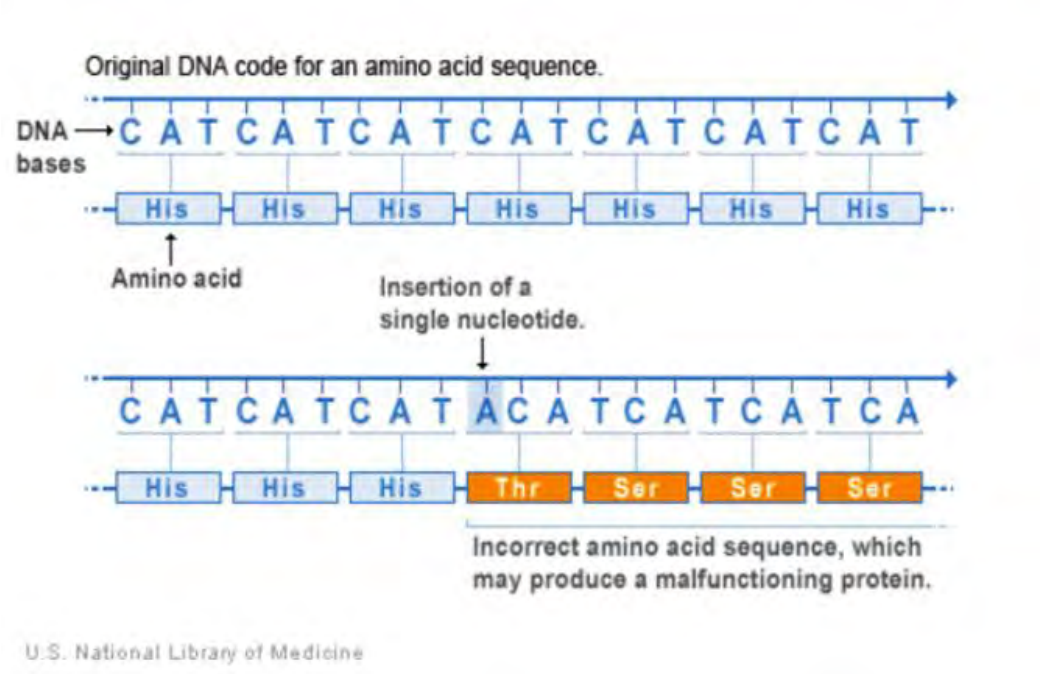
- **Cambio de sentido:** Una variante con cambio de sentido es un tipo de sustitución en la que el cambio de nucleótido resulta en el reemplazo de un bloque de construcción de proteína (aminoácido) por otro en la proteína hecha a partir del gen. El cambio de aminoácidos puede alterar la función de la proteína.
- **Sin sentido:** Una variante sin sentido también es un tipo de sustitución. En vez de causar un cambio en un aminoácido, la secuencia de ADN alterada resulta en una señal de detención que prematuramente indica a la célula que deje de fabricar una proteína. Este tipo de variante da como resultado una proteína acortada que puede funcionar mal, no funcionar o romperse.

Inserción

Una inserción cambia la secuencia de ADN en un gen al agregar uno o más nucleótidos al gen. Como resultado, la proteína producida del gen puede no funcionar correctamente.

Figura 5

Variante deiInserción



Nota. Imagen obtenida de la siguiente web:

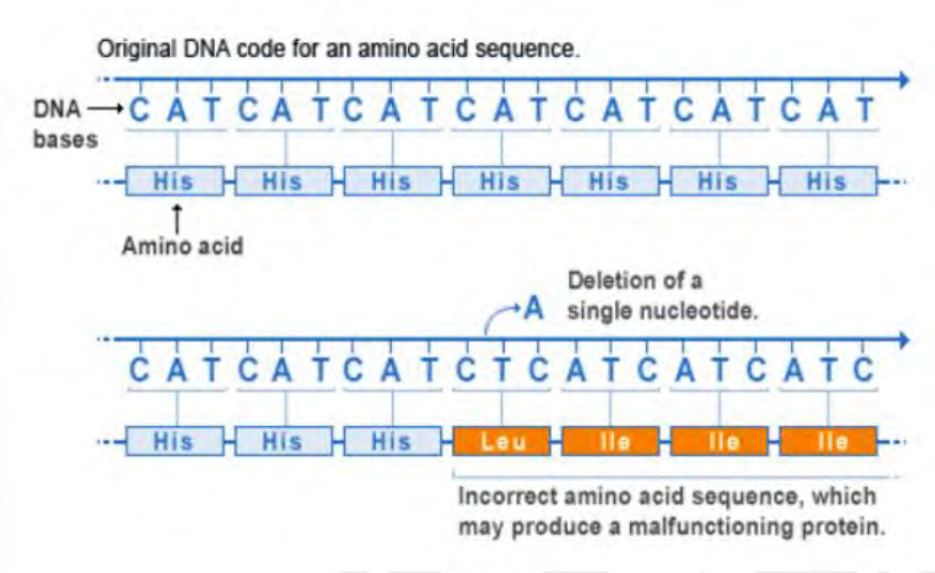
https://medlineplus.gov/spanish/genetica/entender/variantesytrastornos/posiblesvariantes/

Deleción

Una deleción cambia la secuencia de ADN al eliminar al menos un nucleótido en un gen. Las deleciones pequeñas pueden eliminar uno o algunos pares de bases dentro de un gen, mientras que las deleciones más grandes pueden eliminar un gen completo o varios genes vecinos. El ADN eliminado puede alterar la función de las proteínas resultantes.

Figura 6

Variante de deleción



Nota. Imagen obtenida de la siguiente web:

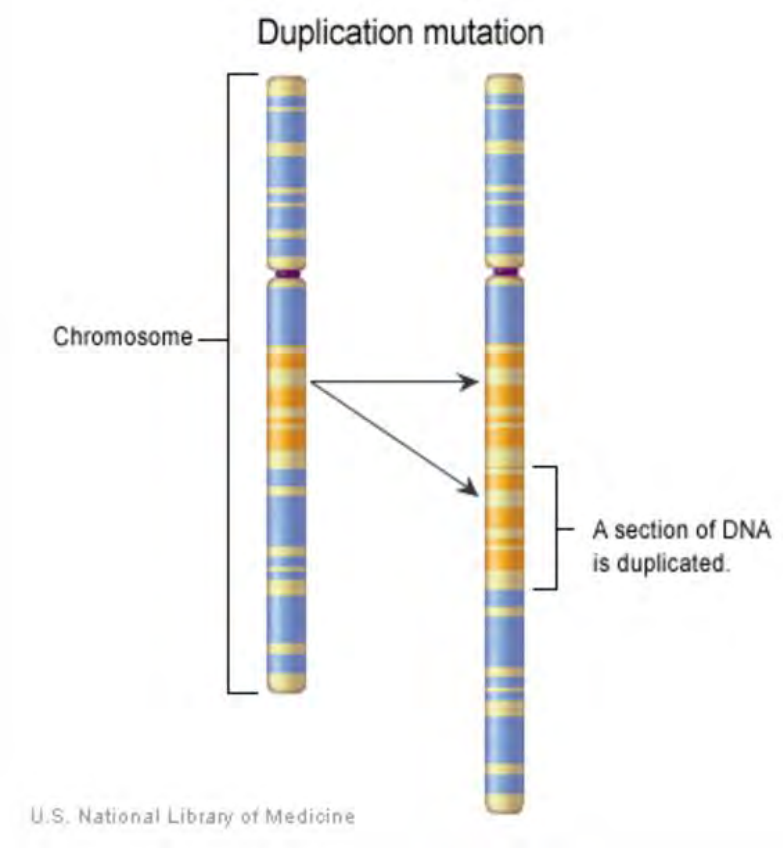
https://medlineplus.gov/spanish/genetica/entender/variantesytrastornos/posiblesvariantes/

Duplicación

Una duplicación ocurre cuando un tramo de uno o más nucleótidos en un gen se copia y se repite junto a la secuencia de ADN original. Este tipo de variante puede alterar la función de la proteína elaborada a partir del gen.

Figura 7

Variante de duplicación



Nota. Imagen obtenida de la siguiente web:

https://medlineplus.gov/spanish/genetica/entender/variantesytrastornos/posiblesvariantes/

Inversión

Una inversión cambia más de un nucleótido en un gen al reemplazar la secuencia original con la misma secuencia en orden inverso.

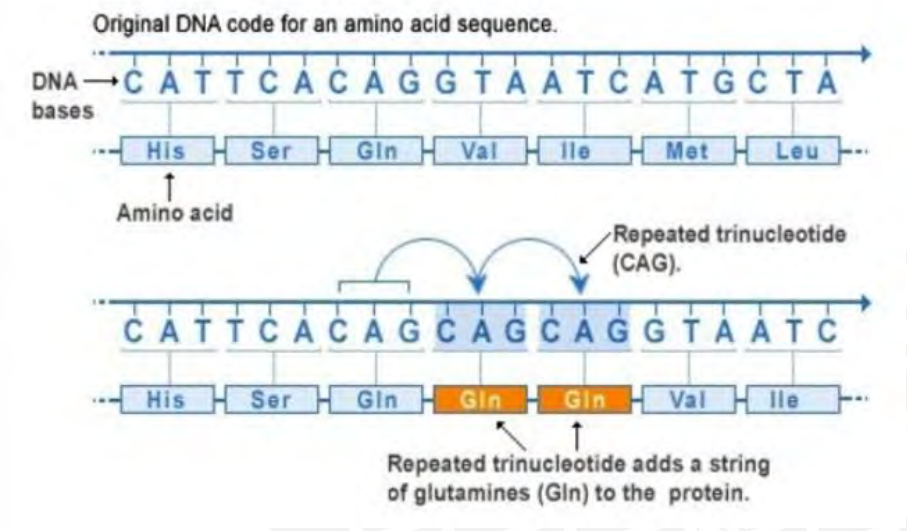
Expansión repetida

Algunas áreas de ADN contienen secuencias cortas de nucleótidos que se repiten varias veces seguidas. Por ejemplo, una repetición de trinucleótidos está formada por secuencias de tres nucleótidos y una repetición de tetranucleótidos está formada por

secuencias de cuatro nucleótidos. Una expansión repetida es una variante que aumenta el número de veces que se repite la secuencia corta de ADN. Este tipo de variante puede hacer que la proteína resultante funcione en forma incorrecta.

Figura 8

Variante de expansión repetida



Nota. Imagen obtenida de la siguiente web: https://medlineplus.gov/spanish/genetica/entender/variantesytrastornos/posiblesvariantes/ (Instituto Nacional de Investigación del Genoma Hum, 2021)

Inhibidor PARP.

Los inhibidores de la PARP son una nueva clase de medicamentos que bloquean la reparación del ADN en las células tumorales y, por tanto, provocan la muerte celular.

En las personas con cáncer de mama localmente avanzado o metastásico HER2-negativo, con mutación en la línea germinal de los genes BRCA, los inhibidores de la PARP ofrecen una mejora de la supervivencia sin progresión, y probablemente mejoran la supervivencia general y las tasas de respuesta tumoral. Esta revisión sistemática proporciona evidencia que apoya la administración de los inhibidores de la PARP como

parte de la estrategia terapéutica para las pacientes con cáncer de mama en este subgrupo. El perfil de toxicidad de los inhibidores de la PARP probablemente no sea peor que el de la quimioterapia, pero se necesita más información sobre los desenlaces de la calidad de vida, lo que pone de relieve la importancia de recopilar estos datos en estudios futuros. Los estudios futuros también deberían tener el poder estadístico necesario para detectar diferencias clínicamente importantes en la supervivencia general y podrían centrarse en el papel de los inhibidores de la PARP en otras poblaciones relevantes de cáncer de mama, como las HER2-positivas, BRCA-negativas/deficientes en la reparación por recombinación homóloga y positivas al ligando de muerte programada 1 (PDL1). (AM, y otros, 2021).

Deep Learning

Un algoritmo de aprendizaje profundo o automático es un proceso computacional que utiliza datos de entrada para lograr una tarea deseada sin estar literalmente programado (es decir, "codificado de forma rígida") para producir un resultado particular. Estos algoritmos son, en cierto sentido, "codificados por software" en el sentido de que alteran o adaptan automáticamente su arquitectura a través de la repetición (es decir, la experiencia) para que sean cada vez mejores en el logro de la tarea deseada. El proceso de adaptación se denomina entrenamiento, en el que se proporcionan muestras de datos de entrada junto con los resultados deseados. Luego, el algoritmo se configura a sí mismo de manera óptima para que no solo produzca el resultado deseado cuando se le presenten las entradas de entrenamiento, sino que pueda generalizar para producir el resultado deseado a partir de datos nuevos, nunca vistos. Este entrenamiento es la parte de "aprendizaje" de los procesos de aprendizaje automático y profundo. El entrenamiento no tiene por qué

limitarse a una adaptación inicial durante un intervalo finito. Al igual que con los humanos, un buen algoritmo puede practicar el aprendizaje "permanente" a medida que procesa nuevos datos y aprende de sus errores. (El Naqa & Murphy, Machine and Deep Learning in Oncology, Medical Physics and Radiology, 2022, p. 3)

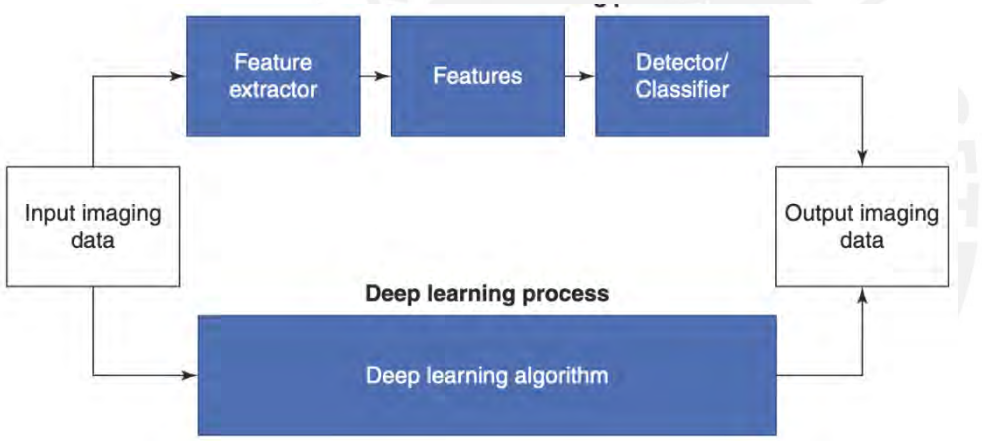
El aprendizaje profundo (DL), como se señaló anteriormente, comprende una subcategoría de aprendizaje automático que se ocupa del aprendizaje de representación, donde la información o los datos sin procesar se alimentan directamente al algoritmo, que luego puede descubrir automáticamente los patrones subyacentes (características) necesarios para la detección o tarea de clasificación. Conceptualmente, se puede aplicar a cualquier tecnología de aprendizaje automático como se muestra en la **Figura 9**, pero se ha demostrado en la práctica que actualmente es más eficaz con métodos de redes neuronales profundas. (El Naqa & Murphy, Machine and Deep Learning in Oncology, Medical Physics and Radiology, 2022, p. 7)

La capacidad de aprender a través de la entrada del entorno circundante ya sea jugando damas o juegos de ajedrez, o reconociendo patrones escritos, o resolviendo los abrumadores problemas en física médica, oncología o radiología, es la clave para una aplicación exitosa de aprendizaje automático. El aprendizaje se define en este contexto como la estimación de dependencias a partir de datos. Los campos de la minería de datos y el aprendizaje automático están entrelazados. La minería de datos utiliza algoritmos de aprendizaje automático para consultar grandes bases de datos y descubrir conocimiento oculto en los datos, mientras que muchos algoritmos de aprendizaje automático emplean datos

El aprendizaje automático/profundo tiene aspectos de la ciencia de la ingeniería, como estructuras de datos, algoritmos, probabilidad y estadísticas, y aspectos de la teoría de la información y el control y las ciencias sociales que se basan en ideas de la psicología y la filosofía. (El Naqa & Murphy, Machine and Deep Learning in Oncology, Medical Physics and Radiology, 2022, p. 8)

Figura 9

Aprendizaje automático "superficial" convencional (arriba) versus algoritmos de aprendizaje profundo, donde la representación y clasificación de datos de imagen se manejan dentro del mismo marco



Red neuronal.

Durante décadas, hemos soñado con construir máquinas inteligentes con cerebros como el nuestro: asistentes robóticos para limpiar nuestros hogares, autos que se manejan solos, microscopios que detectan enfermedades automáticamente. Pero construir estas máquinas con inteligencia artificial requiere que resolvamos algunos de los problemas computacionales más complejos con los que nos hemos enfrentado; problemas que nuestro cerebro ya puede resolver en microsegundos. Para abordar estos problemas, tendremos que desarrollar una forma radicalmente diferente de programar una

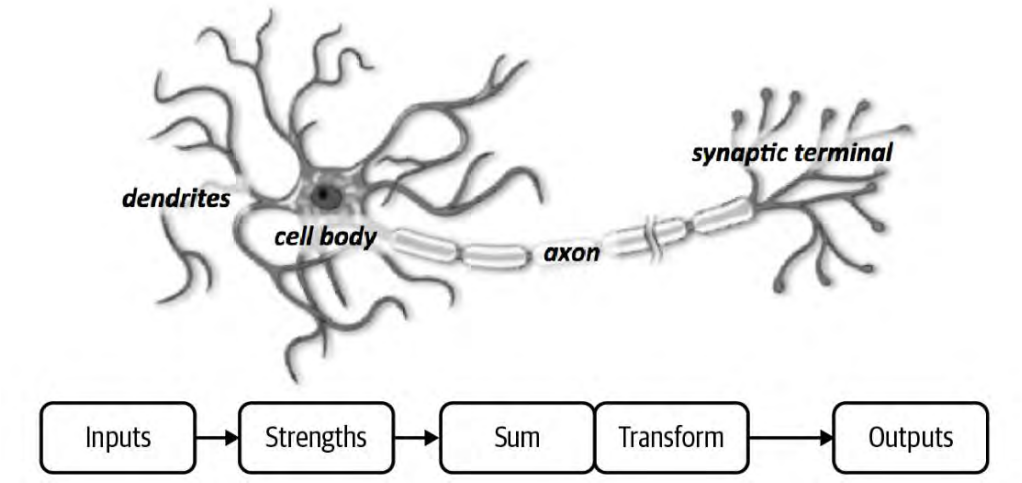
computadora utilizando técnicas desarrolladas en gran medida durante la última década. Este es un campo extremadamente activo de inteligencia artificial informática, a menudo denominado aprendizaje profundo. (Buduma, Buduma, & Papa, 2022, p. 39)

La neurona.

La neurona está optimizada para recibir información de otras neuronas, procesar esta información de una manera única y enviar su resultado a otras células. Este proceso se resume en la **Figura 10**. La neurona recibe sus entradas a lo largo de estructuras parecidas a antenas llamadas dendritas. Cada una de estas conexiones entrantes se fortalece o debilita dinámicamente según la frecuencia con la que se usa (así es como aprendemos nuevos conceptos), y es la fuerza de cada conexión lo que determina la contribución de la entrada a la salida de la neurona. Después de ser ponderados por la fuerza de sus respectivas conexiones, las entradas se suman en el cuerpo de la celda. Esta suma luego se transforma en una nueva señal que se propaga a lo largo del axón de la célula y se envía a otras neuronas. (Buduma, Buduma, & Papa, 2022, p. 46)

Figura 10

Una descripción funcional de la estructura de una neurona biológica

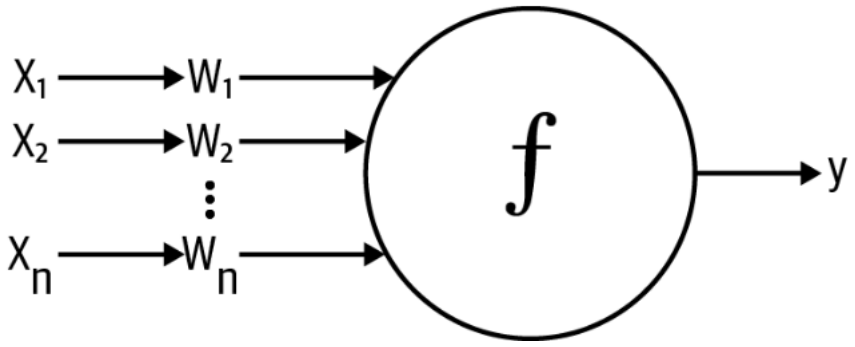


Podemos traducir esta comprensión funcional de las neuronas de nuestro cerebro en un modelo artificial que podemos representar en nuestra computadora. Dicho modelo se describe en la **Figura 11**, aprovechando el enfoque iniciado por primera vez en 1943 por Warren S. McCulloch y Walter H. Pitts. Al igual que en las neuronas biológicas, nuestra neurona artificial recibe cierto número de entradas, x1, x2..., xn, cada uno de los cuales se multiplica por un peso específico, w1, w2, ..., wn. Estas entradas ponderadas se suman, como antes, para producir el logit de la neurona, z = ∑n w x. En muchos casos, el logit también incluye un sesgo, que es i=0ii

constante (no se muestra en la figura). Luego, el logit se pasa a través de una función f para producir la salida y = f z. Esta salida se puede transmitir a otras neuronas. (Buduma, Buduma, & Papa, 2022, p. 46)

Figura 11

Esquema de una neurona en una red neuronal artificial



Reformulemos las entradas como un vector x = [x1 x2... xn] y los pesos de la neurona como w = [w1 w2... wn]. Entonces podemos volver a expresar la salida de la neurona como y = f x·w+b , donde b es el término de sesgo. Podemos calcular la salida realizando el producto escalar de los vectores de entrada y peso, sumando el término de

sesgo para producir el logit y luego aplicando la función de transformación. (Buduma, Buduma, & Papa, 2022, p. 47)

Categorías.

El aprendizaje automático o profundo se puede dividir según la naturaleza del etiquetado de datos en aprendizaje supervisado, no supervisado, semi supervisado y de refuerzo, como se muestra en la **Figura 12**. El aprendizaje supervisado se utiliza para estimar un mapeo de entrada-salida desconocido a partir de muestras de entrada-salida conocidas, donde la salida está etiquetada (por ejemplo, clasificación y regresión). En el aprendizaje no supervisado, solo se dan muestras de entrada al sistema de aprendizaje (por ejemplo, agrupación y estimación de la función de densidad de probabilidad). El aprendizaje semi supervisado es una combinación de ambos, supervisado y no supervisado, donde parte de los datos está parcialmente etiquetada y la parte etiquetada se usa para inferir la parte no etiquetada (p. ej., sistemas de recuperación de texto/imágenes). En el aprendizaje por refuerzo, el algoritmo de aprendizaje automático tiene como objetivo controlar el aprendizaje al acomodar un sistema de retroalimentación, en el que un agente intenta realizar una secuencia de acciones que pueden maximizar una recompensa acumulativa, como ganar un juego de damas, por ejemplo. Este tipo de enfoque es particularmente útil para aplicaciones de toma de decisiones adaptativas o secuenciales.

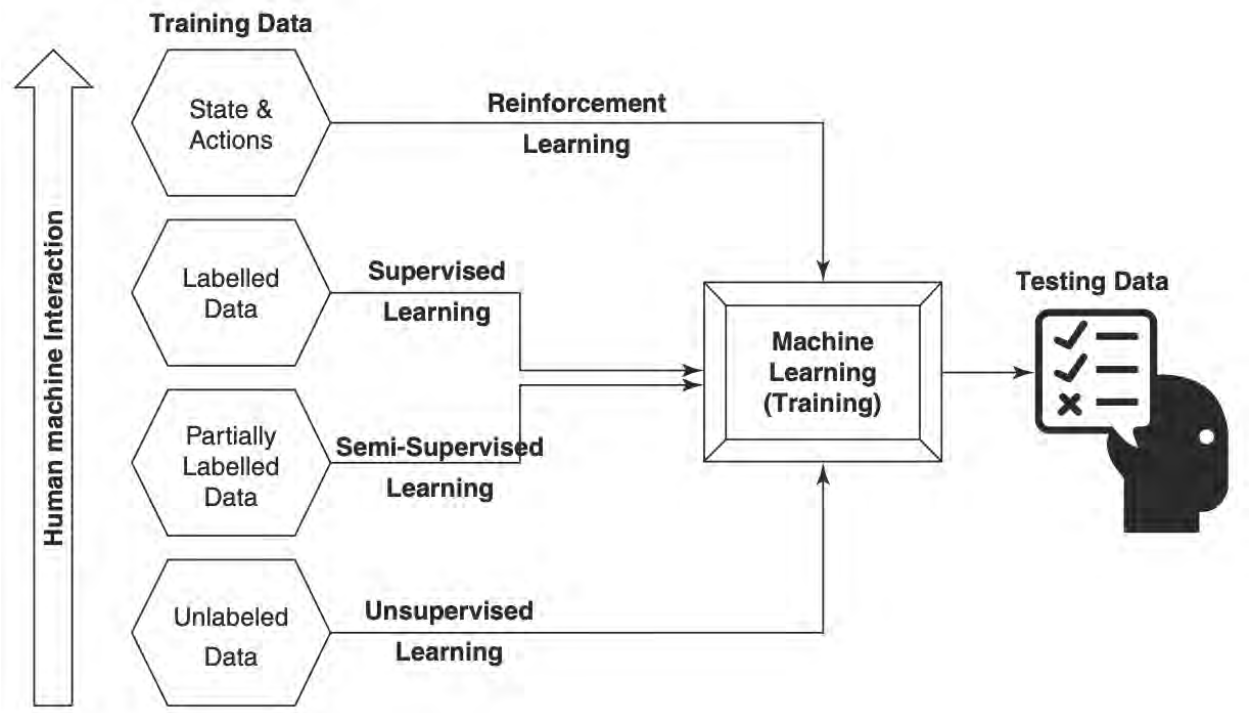
Desde una perspectiva de aprendizaje de conceptos, el aprendizaje automático se puede clasificar en aprendizaje transductivo e inductivo. El aprendizaje transductivo involucra la inferencia de casos de entrenamiento específicos a casos de prueba específicos usando etiquetas discretas como en el agrupamiento o usando etiquetas

continuas como en el aprendizaje múltiple. Por otro lado, el aprendizaje inductivo tiene como objetivo predecir salidas de entradas que no se ha encontrado antes. En este sentido, Mitchell argumenta la necesidad de un sesgo inductivo en el proceso de entrenamiento para permitir que un algoritmo de aprendizaje automático se generalice más allá de la observación invisible.

Desde una perspectiva probabilística, los algoritmos de aprendizaje automático se pueden dividir en modelos discriminantes o generativos. Un modelo discriminante mide la probabilidad condicional de una salida dadas entradas típicamente deterministas, como redes neuronales o una máquina de vectores de soporte. Un modelo generativo es totalmente probabilístico tanto si utiliza una técnica de modelado gráfico como las redes bayesianas como si no, como en el caso de Naïve Bayes. (El Naqa & Murphy, Machine and Deep Learning in Oncology, Medical Physics and Radiology, 2022, p. 8)

Figura 12

Categorías de algoritmos de aprendizaje automático según la naturaleza de los datos de entrenamiento



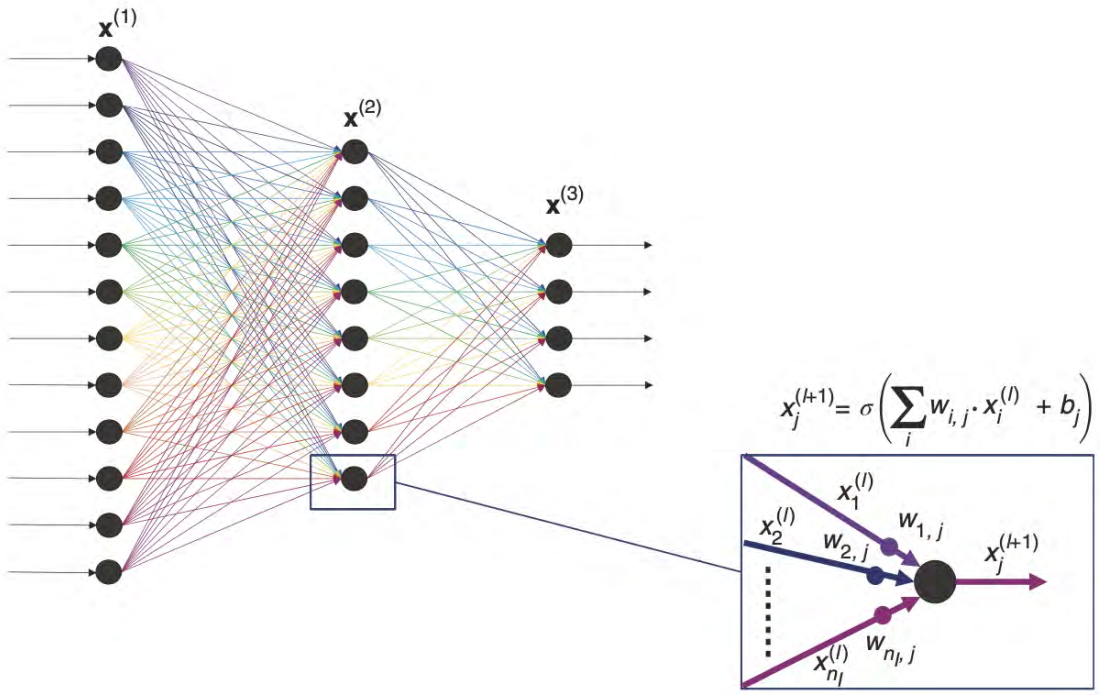
La red neuronal vainilla.

Una red neuronal estándar “totalmente conectada” o “vainilla” consta de capas de neuronas (también denominadas “unidades” o “nodos”). En una capa dada, cada nodo tiene conexiones ponderadas con cada nodo de la capa anterior y cada nodo de la capa siguiente. Los nodos no comparten conexiones dentro de la misma capa. Este flujo de información dirigido hacia adelante es la razón por la que las redes neuronales estándar también se conocen como "redes neuronales de avance". En la literatura, también se puede ver el término "perceptrón multicapa" utilizado para referirse a redes neuronales feed forward totalmente conectadas.

La estructura de una red neuronal vainilla de tres capas se visualiza en la **Figura 13**. La primera capa de la red neuronal es la capa de entrada, x(1). La capa de entrada toma las entradas de datos sin procesar y las propaga a la siguiente capa. La capa final, x(3), es la capa de salida. Los resultados de la capa de salida representan la salida de la red y se utilizan para definir una función de pérdida. El ancho (es decir, la cantidad de nodos) para las capas de entrada y salida suele estar determinado por las características inherentes de los datos y la tarea que realiza la red. Por ejemplo, una red diseñada para usar imágenes en escala de grises con 128 × 128 píxeles tendrán 16 384 nodos en su capa de entrada, donde cada nodo representa la intensidad de un píxel en la imagen. Si la tarea de la red es clasificar dígitos escritos a mano, entonces la capa de salida tendrá un ancho de 10, con cada nodo generando la puntuación de predicción sin procesar (que luego se normalizará en una probabilidad) de un dígito dado. Las capas ocultas (x(2)) consisten en todas las capas entre la entrada y la salida. El ancho y el número total de capas ocultas son hiper parámetros, lo que significa que sus valores los elige el usuario y no se actualizan durante el entrenamiento. El acto de seleccionar/sintonizar hiper parámetros para un modelo dado es un área activa de investigación y se considera un arte en sí mismo.

Figura 13

Una red neuronal de tres capas (capa de entrada, capa oculta y capa de salida)



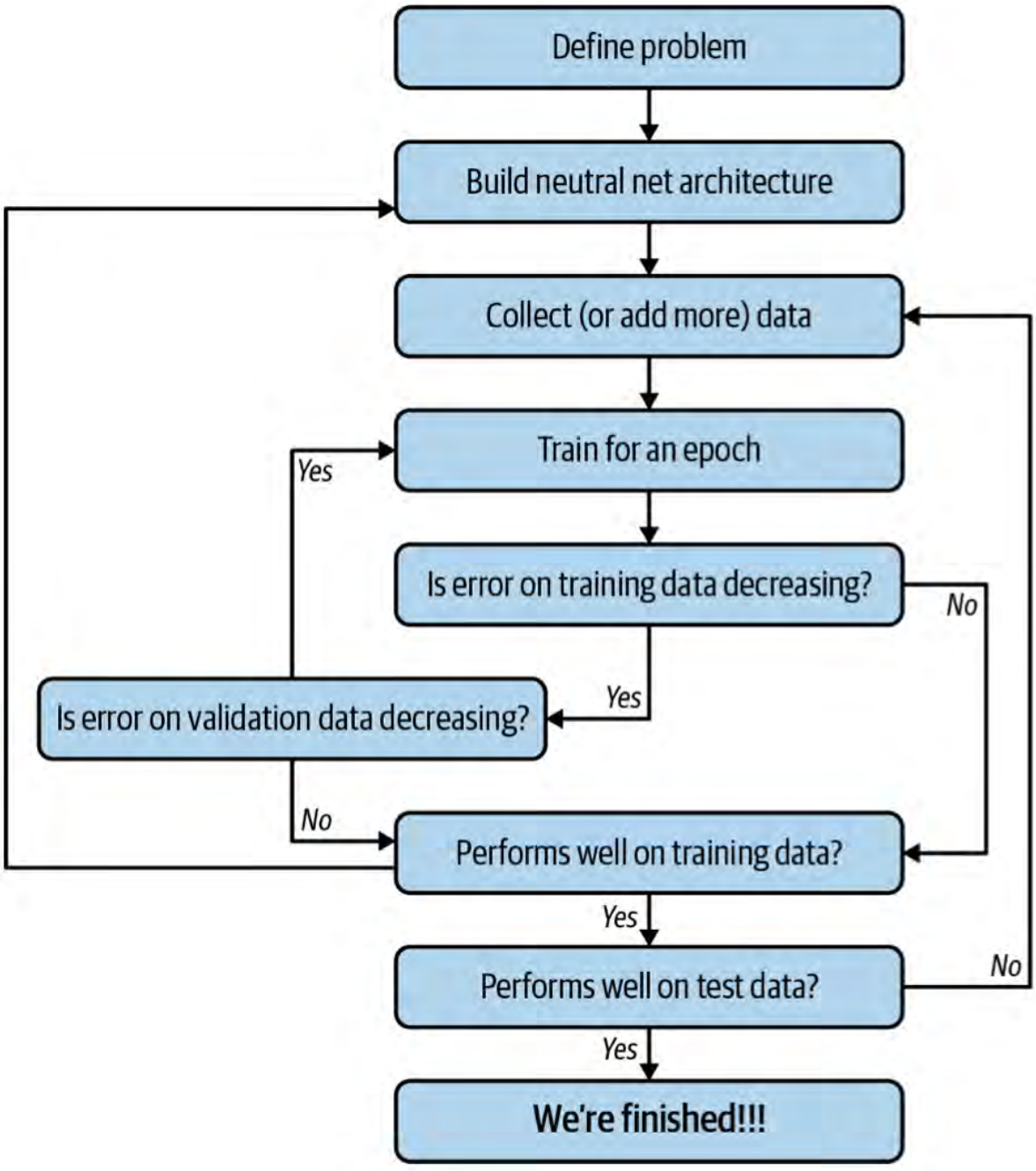
La salida de cada nodo se determina realizando una transformación no lineal (conocida como función de activación) en la suma de las entradas ponderadas más un término de sesgo adicional que el usuario elige y no actualiza durante el entrenamiento. (El Naqa & Murphy, Machine and Deep Learning in Oncology, Medical Physics and Radiology, 2022, p. 54)

Entrenamiento de una red neuronal

Describamos el flujo de trabajo que usamos cuando construimos y entrenamos modelos de aprendizaje profundo. El flujo de trabajo se describe en detalle en la **Figura 14.** Es un poco complicado, pero es fundamental comprender la canalización para garantizar que estamos entrenando adecuadamente nuestras redes neuronales. (Buduma, Buduma, & Papa, 2022, p. 70)

Figura 14

Flujo de trabajo detallado para entrenar y evaluar un modelo de aprendizaje profundo



Hiper parámetros asociados al entrenamiento.

Cuatro hiper parámetros significativos asociados con el proceso de entrenamiento son el **tamaño del lote**, el **número de épocas, la tasa de aprendizaje y la tasa de abandono**. El tamaño del lote es el número de muestras de datos alimentadas a la red antes de que se actualicen los pesos y sesgos. El tamaño del lote puede variar desde 1 (la red se actualiza después de una sola muestra) hasta el tamaño del conjunto de datos de entrenamiento (la red solo se actualiza después de haber visto todas las muestras de datos posibles). Los esquemas de entrenamiento con un tamaño de lote de 1 se denominan descenso de gradiente estocástico, mientras que los esquemas de entrenamiento con un tamaño de lote igual al número de muestras de entrenamiento se denominan descenso de gradiente o descenso de gradiente por lotes. Si el tamaño del lote es un número entre 1 y el tamaño de la muestra de entrenamiento, se dice que el algoritmo sufre un descenso de gradiente de mini lote. De estos tres esquemas de entrenamiento, el descenso de gradiente por lotes es el menos ruidoso porque utiliza cada muestra de datos al calcular el gradiente, lo que significa que no se verá afectado por las variaciones dentro del conjunto de datos. Sin embargo, para conjuntos de datos muy grandes (del orden de millones de ejemplos de entrenamiento) es computacionalmente costoso pasar todos los datos a través de la red antes de cada actualización, lo que aumenta el tiempo total de entrenamiento. El descenso de gradiente estocástico conduce a un proceso de aprendizaje mucho más ruidoso porque la red se actualiza después de una sola muestra de datos, que puede no ser representativa del conjunto de datos como un todo. Un beneficio del descenso de gradiente estocástico es que, al actualizar después de una sola muestra, la función de pérdida puede acercarse a un valor cercano al mínimo global a un ritmo más rápido. Sin

embargo, esto también significa que la red debe actualizarse con mayor frecuencia, lo que también es computacionalmente costoso. El descenso de gradiente de mini lote sirve como un compromiso entre estos dos extremos mediante el uso de un tamaño de lote lo suficientemente grande como para ser algo representativo de todo el conjunto de datos, lo que minimiza el ruido de entrenamiento, pero lo suficientemente pequeño como para que la red pueda actualizarse con más frecuencia, lo que lleva a una mayor rapidez. tiempos generales de entrenamiento. Durante el entrenamiento, es estándar normalizar las entradas de la red lote por lote, un procedimiento llamado normalización por lotes, para mejorar la velocidad de entrenamiento y la estabilidad del modelo.

Una época se refiere a una instancia en la que todo el conjunto de datos de entrenamiento ha pasado a través de la red. El número de épocas es, por lo tanto, un hiper parámetro que describe cuántas veces se pasa el conjunto de datos completo a través de la red durante el entrenamiento. Por lo general, una red requiere que todo el conjunto de datos de entrenamiento pase a través de ella varias veces antes de que la función de pérdida converja al mínimo.

La tasa de aprendizaje es un coeficiente fraccionario que se aplica al gradiente de la función de pérdida y puede considerarse como el tamaño de paso utilizado al actualizar los pesos y sesgos para minimizar la función de pérdida. Una opción popular para la optimización de redes neuronales es utilizar una tasa de aprendizaje adaptativa definida mediante el algoritmo de optimización estocástica de Adam.

El abandono es una práctica estándar utilizada durante el entrenamiento de redes neuronales para evitar el sobreajuste del modelo. Consiste en seleccionar aleatoriamente un número determinado de nodos en la red y establecer su salida en 0, descartándolos

efectivamente (y sus conexiones) de la red. La fracción de neuronas en una capa dada que se eliminan durante el entrenamiento se denomina tasa de deserción. (El Naqa & Murphy, Machine and Deep Learning in Oncology, Medical Physics and Radiology, 2022, p. 57)

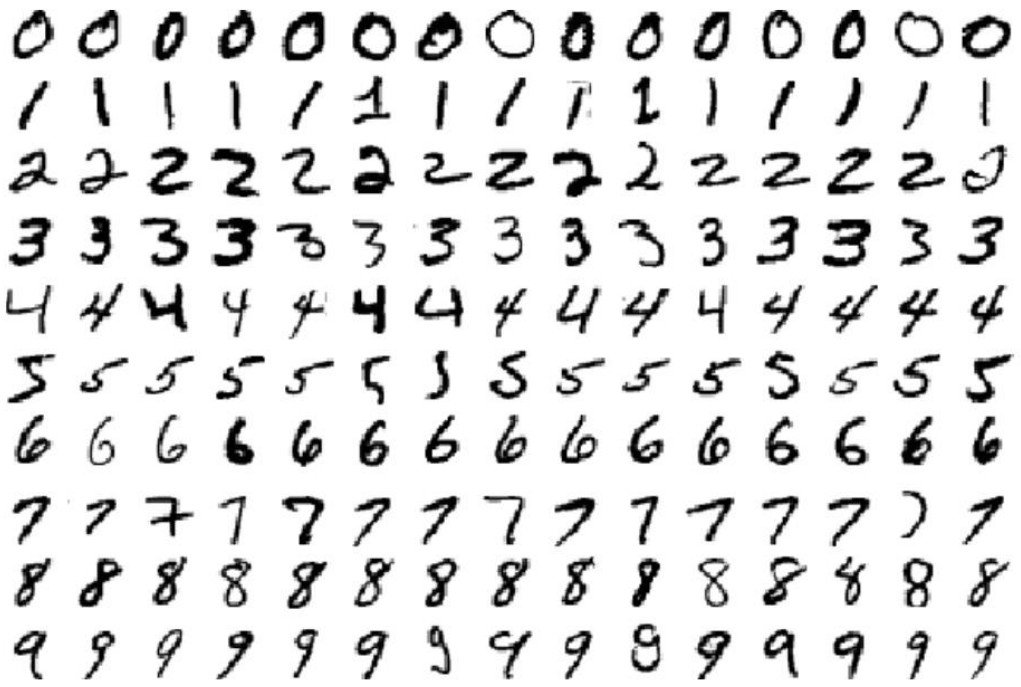
Autocodificadores.

Los límites de los programas informáticos tradicionales

¿Por qué exactamente ciertos problemas son tan difíciles de resolver para las computadoras? Bueno, resulta que los programas de computadora tradicionales están diseñados para ser muy buenos en dos cosas: (1) realizar operaciones aritméticas realmente rápido y (2) seguir explícitamente una lista de instrucciones. Entonces, si desea hacer un gran crujido de números financieros, está de suerte. Los programas de computadora tradicionales pueden hacer el truco. Pero digamos que queremos hacer algo un poco más interesante, como escribir un programa para leer automáticamente la escritura a mano de alguien. La **Figura 15** servirá como punto de partida.

Figura 15

Imagen del conjunto de datos de dígitos escritos a mano del MNIST2



Aunque cada dígito de la **Figura 6** está escrito de una manera ligeramente diferente, podemos reconocer fácilmente cada dígito de la primera fila como un cero, cada dígito de la segunda fila como uno, etc. Tratemos de escribir un programa de computadora para romper esta tarea. ¿Qué reglas podríamos usar para distinguir un dígito de otro?

Bueno, ¡podemos empezar de manera simple! Por ejemplo, podríamos afirmar que tenemos un cero si nuestra imagen tiene solo un bucle cerrado único. Todos los ejemplos de la figura 3-1 parecen cumplir este requisito, pero en realidad no es una condición suficiente. ¿Qué pasa si alguien no cierra perfectamente el ciclo en su cero? Y, como en la **Figura 16**, ¿cómo distingue un cero desordenado de un seis?

Figura 16

Un cero que es algorítmicamente difícil de distinguir de un seis

Potencialmente, podría establecer algún tipo de límite para la distancia entre el punto inicial del bucle y el punto final, pero no está exactamente claro dónde deberíamos dibujar la línea. Pero este dilema es solo el comienzo de nuestras preocupaciones. ¿Cómo distinguimos entre tres y cinco? ¿O entre cuatro y nueve? Podemos agregar más y más reglas o funciones a través de una observación cuidadosa y meses de prueba y error, pero está bastante claro que este no será un proceso fácil.

Muchas otras clases de problemas entran en esta misma categoría: reconocimiento de objetos, comprensión del habla, traducción automática, etc. No sabemos qué programa escribir porque no sabemos cómo lo hace nuestro cerebro. E incluso si supiéramos cómo hacerlo, el programa podría ser terriblemente complicado. (Buduma, Buduma, & Papa, 2022, p. 40)

Redes neuronales convolucionales.

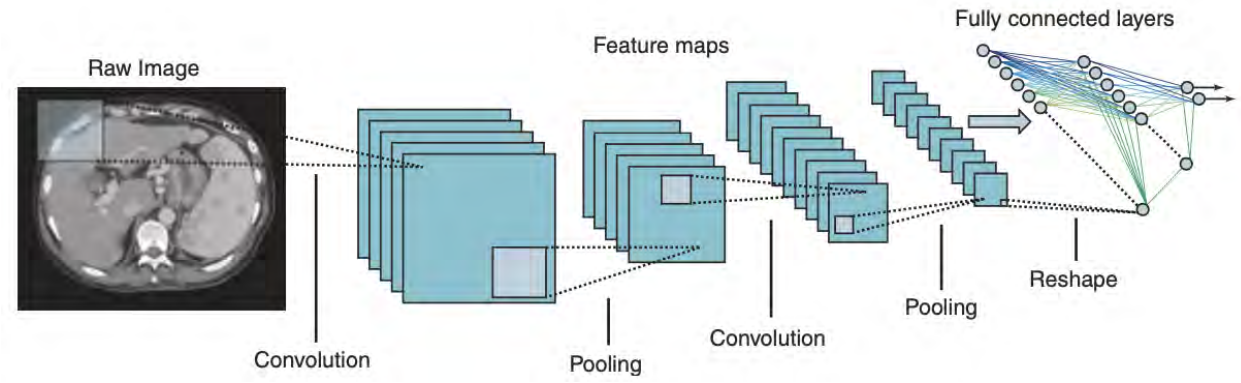
Entre los algoritmos de aprendizaje profundo más conocidos se encuentran las redes neuronales convolucionales (CNN). Inspiradas biológicamente en la corteza visual de los animales, las CNN se propusieron por primera vez en 1980 y desde entonces han revolucionado el campo de la visión artificial. Una característica importante de las CNN

son sus propiedades de traducción invariante, que es una de las principales atracciones de su popularidad en aplicaciones de imagen y visión por computadora. Las CNN se diferencian de las redes neuronales *feedforward* estándar en que son capaces de aprender características que dependen de las relaciones estructurales locales dentro de los datos, lo que las hace particularmente hábiles en tareas que involucran datos de imágenes.

Los bloques de construcción estándar de una CNN son capas convolucionales, que son responsables de crear mapas de características de los datos de entrada; capas de agrupación, que reducen el número de parámetros en el modelo y ayudan a evitar el sobreajuste; y una o más capas completamente conectadas al final de la red. La **Figura 17** muestra una CNN simple con dos capas convolucionales, dos capas de agrupación y tres completamente con capas conectadas. (El Naqa & Murphy, Machine and Deep Learning in Oncology, Medical Physics and Radiology, 2022, pp. 63-64)

Figura 17

Esquema de una CNN con 2 capas convolucionales y 2 capas completamente conectadas



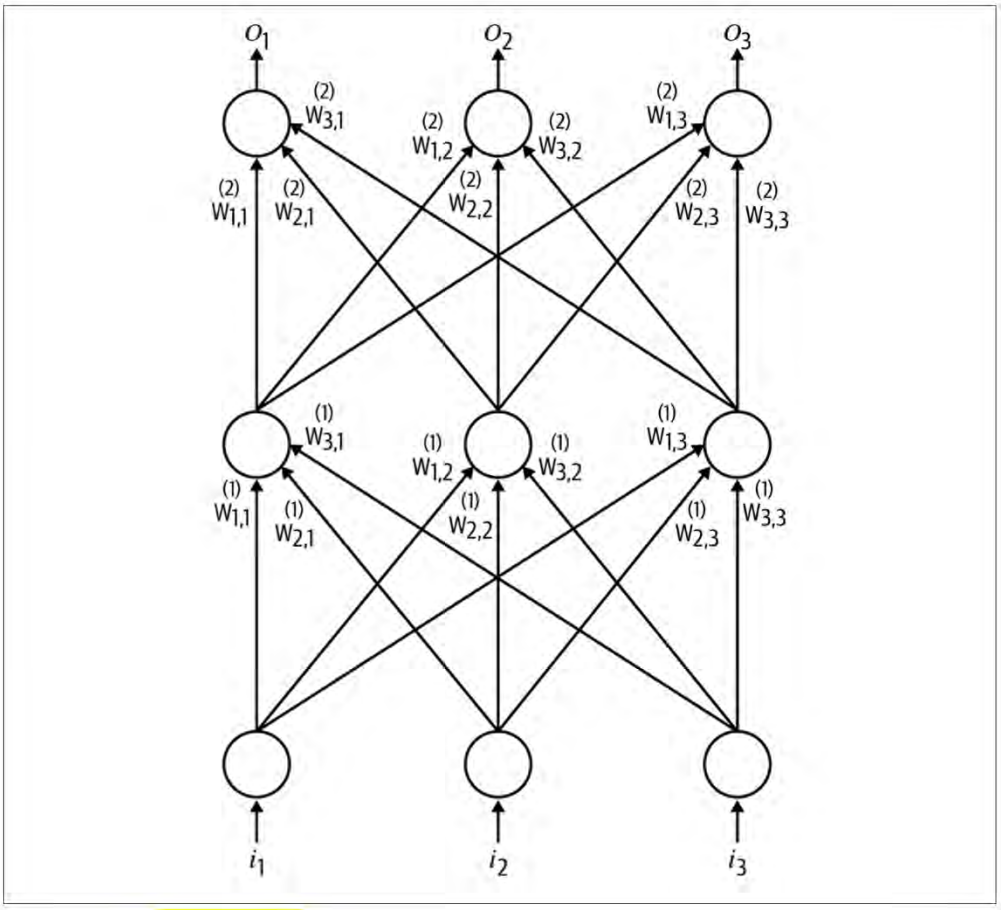
Red neuronal del tipo feedforward.

La red neuronal de tipo feedforward es el tipo más simple de red neuronal artificial. En una red de tipo feedforward, la información se desplaza solo en una dirección: desde la capa de entrada a la de salida. Las redes neuronales de tipo

feedforward transforman una entrada pasándola por una serie de capas ocultas. Cada capa consta de un conjunto de neuronas, donde cada capa está totalmente conectada a todas las neuronas de la capa anterior. Por último, hay una última capa totalmente conectada (la capa de salida) que representa las predicciones generadas. (Buduma, Buduma, & Papa, 2022, p. 48)

Figura 18

Una red neuronal feed-forward con tres capas (entrada, una oculta y salida) y tres neuronas por capa



Red neuronal concurrente.

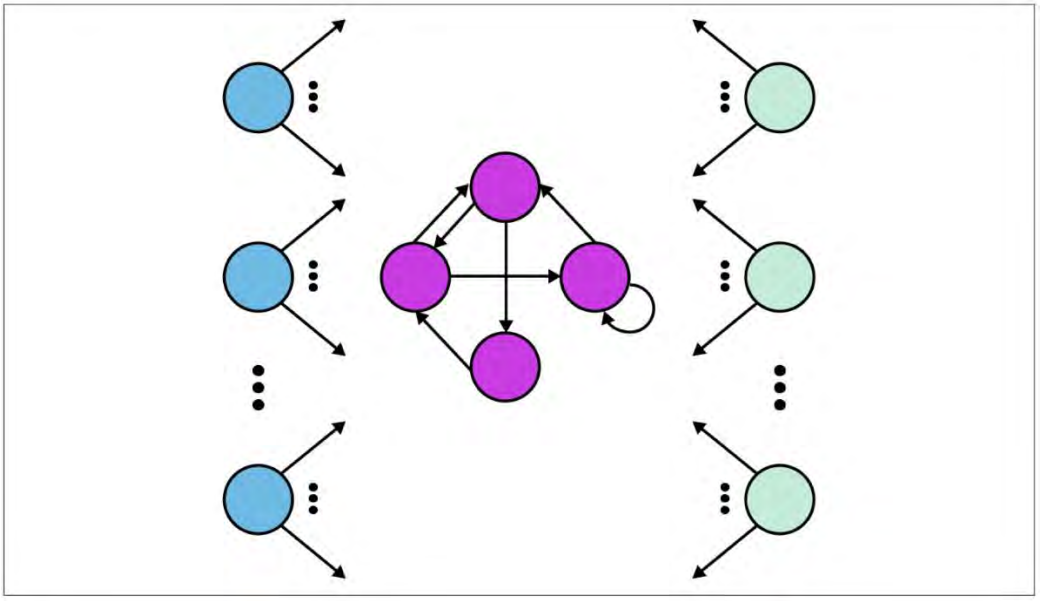
Las redes neuronales recurrentes son una red neuronal artificial que se usa ampliamente. Estas redes guardan la salida de una capa y la reenvían a la capa de entrada

para poder predecir el resultado de esa capa. Las redes neuronales recurrentes tienen grandes capacidades de aprendizaje. Suelen utilizarse en tareas complejas, como la predicción de series temporales, el aprendizaje de escritura a mano y el reconocimiento de idiomas

Las RNN son diferentes de las redes feed-forward porque aprovechan un tipo especial de capa neuronal, conocida como capas recurrentes, que permiten que la red mantenga el estado entre los usos de la red. (Buduma, Buduma, & Papa, 2022, pp. 207-208)

Figura 19

Conexiones entre neuronas que se encuentran en la misma capa



Nota. Ilustración obtenida del libro Fundamentals of Deep Learning_ Designing Next-Generation Machine Intelligence – Pág. 208

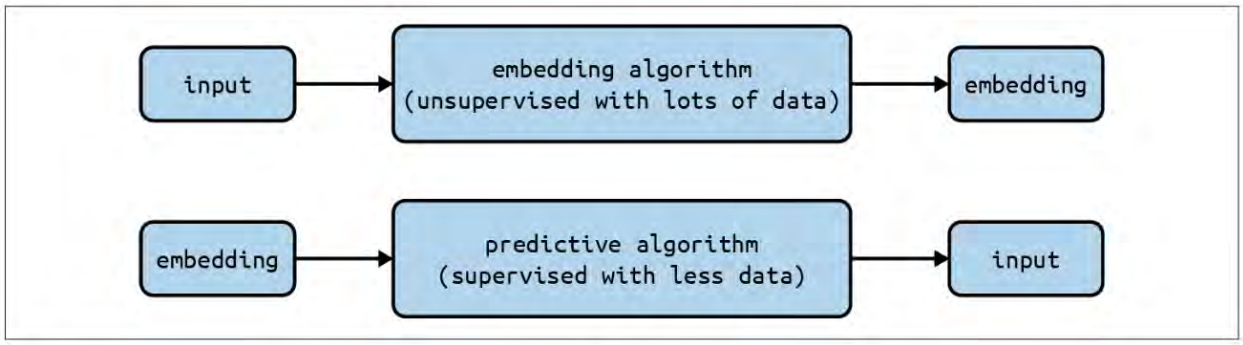
Aprendizaje de incrustación y representación.

Para muchos problemas, los datos etiquetados son escasos y costosos de generar, el aprendizaje de incrustación permite desarrollar modelos de aprendizaje efectivos en

situaciones donde los datos etiquetados son escasos, pero los datos salvajes y sin etiquetar son abundantes, podemos usar las incrustaciones generadas para resolver problemas de aprendizaje usando modelos más pequeños que requieren menos datos. Este proceso se resume en la **Figura 20.** (Buduma, Buduma, & Papa, 2022, pp. 157-158)

Figura 20

Uso de incrustaciones para automatizar la selección de funciones ante la escasez de datos etiquetados



Nota. Ilustración obtenida del libro Fundamentals of Deep Learning Designing Next-Generation Machine Intelligence – Pág. 157

Modelos para análisis de secuencias.

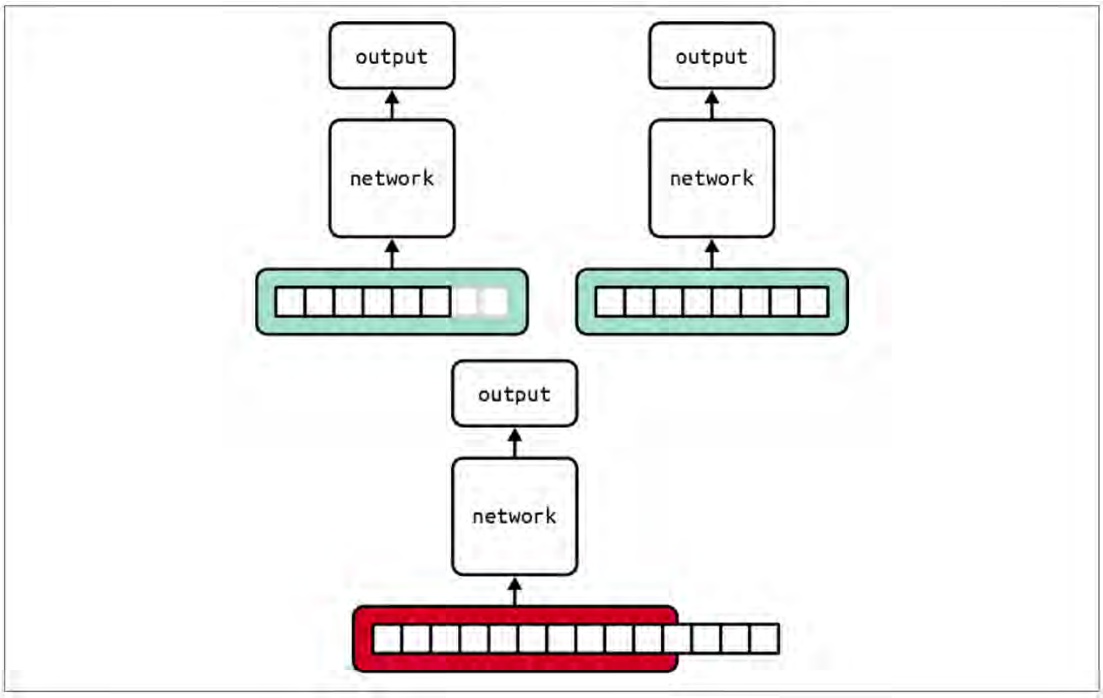
La **Figura 21** ilustra cómo se rompen nuestras redes neuronales de alimentación hacia adelante al analizar secuencias. Si la secuencia tiene el mismo tamaño que la capa de entrada, el modelo puede funcionar como se esperaba. Incluso es posible manejar entradas más pequeñas al agregar ceros al final de la entrada hasta que tenga la longitud adecuada. Sin embargo, en el momento en que la entrada supera el tamaño de la capa de entrada, el uso ingenuo de la red feed-forward ya no funciona.

Las redes feed-forward prosperan con problemas de tamaño de entrada fijo. El relleno cero puede abordar el manejo de entradas más pequeñas, pero cuando se utilizan

de forma ingenua, estos modelos se rompen cuando las entradas superan el tamaño de entrada fijo. (Buduma, Buduma, & Papa, 2022, p. 189)

Figura 21

Red de retroalimentación rota



Métodos en Interpretabilidad.

La interpretabilidad define la capacidad de un modelo para "explicar" su toma de decisiones a un tercero. Hay muchas arquitecturas modernas que no tienen esta capacidad solo por construcción. Una red neuronal, por ejemplo, es un excelente ejemplo de una de estas arquitecturas modernas. El término "opaco" se usa a menudo para describir redes neuronales, tanto en los medios como en la literatura. Esto se debe a que, sin técnicas post hoc para explicar la clasificación final o el resultado de regresión de una red neuronal, las transformaciones de datos que ocurren dentro del modelo entrenado no son claras y difíciles de interpretar para el usuario final. Todo lo que sabemos es que ingresamos un ejemplo y obtuvimos un resultado.

Todo esto plantea la pregunta: ¿por qué nos preocupamos por la interpretabilidad en primer lugar? En un mundo cada vez más dominado por la tecnología, los algoritmos complejos y el aprendizaje automático, la capacidad de explicar la toma de decisiones es imperativa. Especialmente en campos como la medicina, donde la vida de los pacientes está en juego, o en las finanzas, donde está en juego el sustento financiero de las personas, la capacidad de explicar la toma de decisiones de un modelo es un paso clave hacia la adopción generalizada. En la siguiente sección, cubriremos algunos modelos clásicos que tienen fuertes nociones de interpretabilidad integradas en su diseño. (Buduma, Buduma, & Papa, 2022, p. 276)

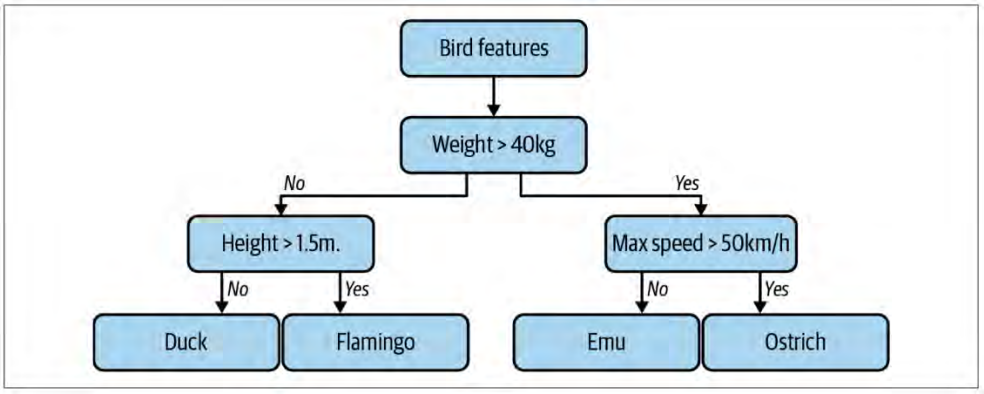
Árboles de Decisión y Algoritmos Basados en Árboles.

Los árboles de decisión están diseñados para clasificar una entrada en función de una serie de declaraciones condicionales, donde cada nodo del árbol está asociado con una declaración condicional. Para comprender cómo toma una decisión un modelo basado en un árbol entrenado, todo lo que debemos hacer para cualquier entrada dada es seguir la rama correcta en cada nodo del árbol **(Figura 22).**

Figura 22

Un árbol de decisión entrenado para clasificar especies de aves. Dado un conjunto de

características de aves, siga la rama derecha "Sí" o "No" en cada nodo para llegar a una clasificación final



Nota. Ilustración obtenida del libro Fundamentals of Deep Learning_ Designing Next-Generation Machine Intelligence – Pág. 276

También se pueden interpretar algoritmos basados en árboles más complejos, como el algoritmo de bosque aleatorio, que se compone de un conjunto de grandes árboles de decisión. Por ejemplo, en el caso de la clasificación, los algoritmos de bosque aleatorio funcionan ejecutando una entrada dada a través de cada árbol de decisión y luego tomando la clase de salida mayoritaria entre los árboles de decisión como la salida final (o un promedio en el caso de regresión). Por la construcción del algoritmo, sabemos exactamente cómo Random Forest llegó a una conclusión con respecto a la entrada.

Además de la interpretabilidad a nivel de ejemplo individual, los árboles de decisión y sus conjuntos más complejos tienen métricas integradas para la importancia de las características a nivel global. Por ejemplo, cuando se entrena un árbol de decisión, debe determinar en qué característica se dividirá y los umbrales de esa característica en los que se dividirá. En el régimen de clasificación, una metodología para hacer esto es calcular la ganancia de información dividiendo una característica propuesta en un umbral propuesto. Para enmarcar nuestro pensamiento, pensemos en las posibles etiquetas de

entrenamiento como el dominio de una distribución de probabilidad discreta, donde la probabilidad de cada etiqueta es la frecuencia con la que esa etiqueta aparece en el conjunto de datos de entrenamiento. (Buduma, Buduma, & Papa, 2022, pp. 276-277)

Regresión Lineal.

Una breve introducción a la regresión lineal: dado un conjunto de características y una variable de destino, nuestro objetivo es encontrar la "mejor" combinación lineal de características que se aproxime a la variable de destino. Implícita en este modelo está la suposición de que las características de entrada están linealmente relacionadas con la variable de destino. Definimos "mejor" como el conjunto de coeficientes que da como resultado la combinación lineal con el error cuadrático medio más bajo cuando se compara con la realidad básica: $y=\beta \cdot x+\varepsilon, \varepsilon \sim N\ 0, \sigma 2$

Donde β representa el vector de coeficientes. Nuestra noción integrada y global de la importancia de las características se deriva directamente de esto. Las características que se corresponden con los coeficientes de mayor magnitud son, globalmente, las características más importantes de la regresión. (Buduma, Buduma, & Papa, 2022, p. 280)

Técnicas del aprendizaje profundo y del aprendizaje automático.

En el aprendizaje automático, se debe indicar al algoritmo cómo realizar una predicción precisa; para ello debe conseguir más información (por ejemplo, realizando la extracción de características). En el aprendizaje profundo, en cambio, el algoritmo puede obtener información sobre cómo hacer una predicción precisa a través de su propio procesamiento de datos, gracias a la estructura de red neuronal artificial. (Microsoft Ignite, 2022)

En la tabla siguiente se comparan las dos técnicas con más detalle:

Tabla 3

Tabla comparativa de aprendizaje automático vs aprendizaje profundo

	Todo el aprendizaje automático	**Solo aprendizaje profundo**
Número de puntos de datos	Puede usar pequeñas cantidades de datos para hacer predicciones.	Necesita usar grandes cantidades de datos de entrenamiento para hacer predicciones.
Dependencias del hardware	Puede trabajar en equipos lentos. No necesita una gran cantidad de potencia de cálculo.	Depende de máquinas rápidas. Realiza intrínsecamente un gran número de operaciones de multiplicación de matrices. Una GPU puede optimizar eficazmente estas operaciones.
Proceso de características	Requiere que los usuarios creen e identifiquen con precisión las características.	Aprende las características de alto nivel de los datos y crea nuevas características automáticamente.
Enfoque del aprendizaje	Divide el proceso de aprendizaje en pasos más pequeños. Luego, combina los resultados de cada paso en una salida.	Pasa por el proceso de aprendizaje mediante la resolución del problema de un extremo a otro.
Tiempo de ejecución	Comparativamente, tarda menos tiempo en entrenarse; puede tardar unos segundos o unas pocas horas.	Normalmente, tarda demasiado tiempo en entrenarse, porque los algoritmos de aprendizaje profundo tienen muchas capas.
Salida	La salida suele ser un valor numérico, como una puntuación o una clasificación.	La salida puede tener varios formatos, como texto, una puntuación o un sonido.

Nota. Tabla obtenida del siguiente enlace web https://learn.microsoft.com/es-es/azure/machine-learning/concept-deep-learning-vs-machine-learning

Máquinas de Vectores de Soporte

Es un algoritmo que está entrenado para aprender reglas de clasificación y regresión a partir de los datos adquiridos. Opera creando una línea o un hiperplano y, por lo tanto, divide grupos de datos en clases. Su función es separar los datos hasta una distancia mínima elevada. SVM se puede utilizar para resolver problemas tanto lineales como no lineales. Se considera uno de los mejores clasificadores de uso general para la detección del cáncer de mama.

Redes Neuronales Artificiales

Es una técnica utilizada para replicar el sistema neuronal del cerebro humano y darle a la computadora un cerebro propio. En esta técnica, al igual que el cerebro humano, el algoritmo consta de varios nodos, que tienen dos valores, es decir, 0 o 1. Cero aquí significa que el nodo está activo, mientras que 1 significa que está inactivo. Estos nodos también tienen un peso que nuevamente tiene sus tipos: positivo y negativo, que se utilizan para ajustar la fuerza del nodo. Se supone que la máquina entrenada busca patrones ocultos en los datos proporcionados y se utiliza para buscar entre los registros de los pacientes para resaltar dichos patrones y ayudar a identificar tumores.

K Vecinos más próximos

Es un método de aprendizaje supervisado que se utiliza para diagnosticar y clasificar cáncer En este método, se identifican puntos de datos, que se conocen como K. El número de puntos de datos puede ser dado previamente o ser establecido por la propia computadora. El método funciona comprobando la distancia de cada punto con todos los puntos de datos proporcionados y asignando cada punto al punto de datos más cercano, también conocido como vecino. En este método, es adecuado seleccionar un gran conjunto de datos y establecer el valor de K como un número impar

Árbol de Decisión

Es una técnica de minería de datos que es útil en el diagnóstico precoz del cáncer de mama. En esta técnica, los datos se dividen en nodos. La representación de estos nodos cuando se juntan parece un árbol con sus ramas, de ahí su nombre. A través de varios algoritmos al escanear los datos disponibles, el árbol decide por sí mismo formar sub nodos que se puede suponer que son las ramas del árbol. Después de varias iteraciones, se logra un resultado al final del árbol

Evaluación del Modelo

Hay dos tipos diferentes de evaluación de modelos según el tipo de datos: evaluación de modelos en línea, que usa datos en vivo, y evaluación de modelos fuera de línea, que usa datos históricos. La evaluación del modelo en términos de métricas de rendimiento generalmente está relacionada con los modelos fuera de línea, y la evaluación del modelo después de la producción utilizando datos en vivo es para garantizar que los modelos sean sólidos, justos, calibrados y que funcionen. Para la evaluación del modelo, idealmente, los datos de entrada que se usaron para construir el modelo deberían ser como los datos de entrada que el modelo tendrá que usar en la producción, pero esto no siempre es posible porque existe una gran variabilidad en los datos del mundo real que el modelo ve en producción. Como resultado, los datos de entrada en producción son mucho más ruidosos en comparación con los datos de entrada durante el desarrollo del modelo. Para abordar esto, podemos hacer pequeños cambios en el proceso de desarrollo del modelo. Podemos agregar algo de ruido aleatorio a los datos de entrenamiento durante el proceso de entrenamiento del modelo para que el algoritmo pueda aprender del ruido aleatorio y luego pueda usar esos aprendizajes durante las predicciones sobre datos ruidosos del mundo real.

Las tareas de clasificación son tareas relacionadas con la predicción de la etiqueta de clase de una observación. Las tareas de clasificación binaria involucran dos etiquetas de clase, mientras que las tareas de clasificación multiclase involucran más de dos etiquetas de clase. Hay varias métricas de clasificación relacionadas con ambos tipos de clasificación, como se muestra a continuación. (Devisetty, 2022, p. 169)

Precisión:

Esta es la métrica de clasificación más utilizada para medir el rendimiento de un modelo. Es simplemente una medida de la frecuencia con la que el clasificador hace las predicciones correctas. Se define como la relación entre las predicciones correctas y el número total de todas las predicciones, como se muestra aquí:

$$Precisión = \frac{\#prediccionescorrectas}{\#total\ predicciones}$$

Cuanto mayor sea la precisión, mejor será el modelo, y viceversa. Alta precisión significa que el modelo se desempeñó bien con los datos de prueba.

Aunque la precisión es fácil de entender, no le brinda una imagen completa de las predicciones. Si hay dos clases, por ejemplo, factor de transcripción (TF) o sin TF, la precisión da la misma preferencia a ambas clases, lo que a veces no es suficiente. Es posible que le interese saber cuántas de muchas observaciones no caen en ningún TF versus TF porque es importante saber si el costo de la clasificación errónea puede diferir para las dos clases, o si hay un problema de clase desequilibrada en los datos. Aquí es cuando necesita algo llamado matriz de confusión que muestre un desglose más detallado de la clasificación (clase correcta versus clase no correcta). En una matriz de confusión típica, como la siguiente tabla, las filas corresponden a la verdad fundamental (o etiquetas originales) y las columnas corresponden a observaciones (predicciones), como se ilustra en el siguiente ejemplo: (Devisetty, 2022, p. 170)

Tabla 4

Matriz de confusión típica

	Predicción Positiva	**Predicción Negativa**
Etiqueta Positiva	Verdaderos Positivos (TP)	Falsos Negativos (FN)
Etiqueta Negativa	Falsos Positivos (FP)	Verdaderos Negativos (TN)

Nota. Tabla obtenida del libro "DeepLearning for Genomics" página 189.

Ahora la precisión será representada mediante:

$$Precisión = \frac{TP + TN}{TP + TN + FP + FN}$$

Precisión y recuperación

Estas son métricas de clasificación importantes que ayudarían a comprender los detalles más finos del rendimiento de la clasificación. La precisión informa de todas las decisiones tomadas por el clasificador para ser relevantes, cuantas de ellas son verdaderamente relevantes. Por el contrario, recordar informa de todos los elementos relevantes en los datos, cuántos de ellos son hechos por el clasificador. (Devisetty, 2022, p. 171)

Matemáticamente, la precisión y la recuperación se calculan de la siguiente manera:

$$Precisión = \frac{\#\ predicciones\ correctas}{\#\ decisiones\ hechas\ por\ clasificador}$$

$$Recuperación = \frac{\#\ predicciones\ correctas}{\#total\ elementos\ relevantes}$$

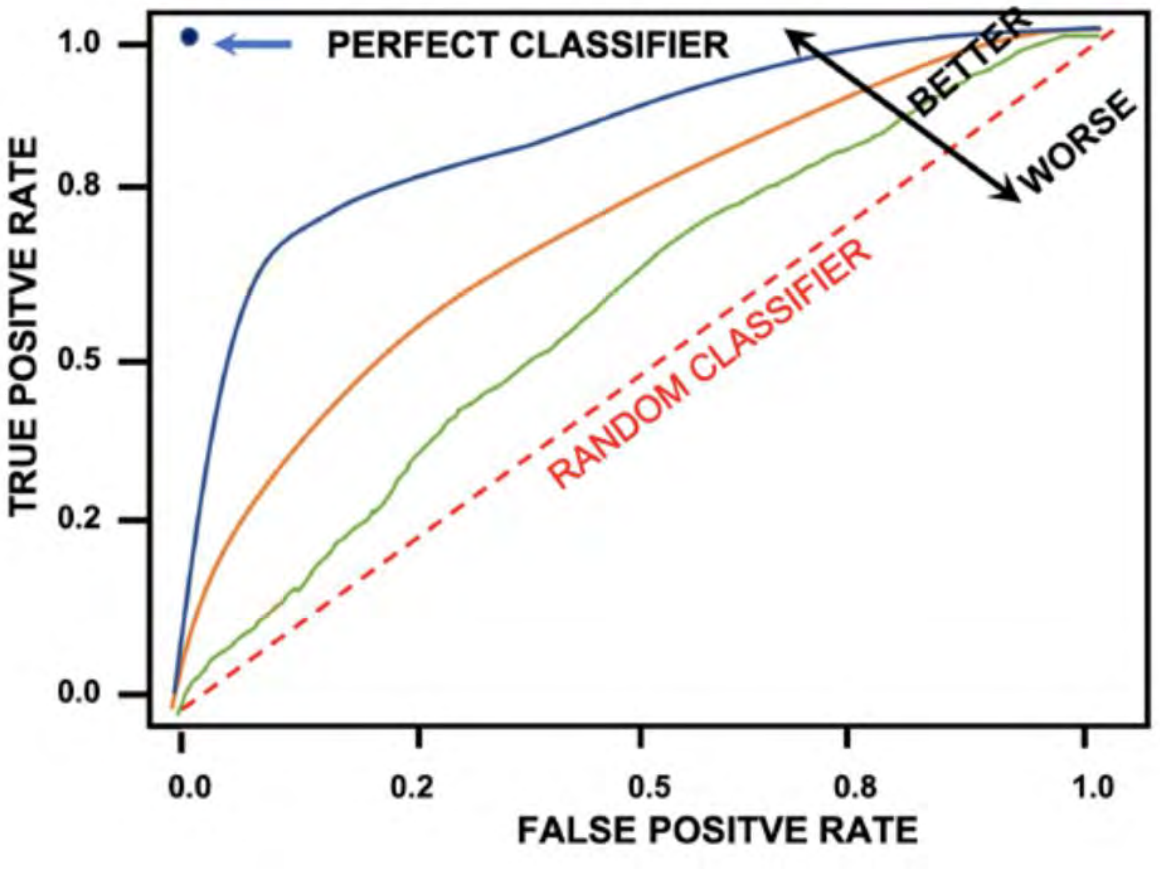
ROC

Esto representa la curva característica operativa del receptor (ROC), que es una medida del rendimiento de un modelo. Generalmente se usa para comparar múltiples modelos para identificar el modelo con mejor rendimiento. Esta curva muestra la sensibilidad del clasificador al trazar la Tasa de verdaderos positivos (TPR) como una función de la Tasa de falsos positivos

(FPR). La medición del área bajo la curva (AUC) indica el rendimiento del clasificador. Aunque AUC está representado por un solo número (entre 0 y 1), el número resume el ROC y proporciona información sobre el clasificador. Un modelo de alto rendimiento tendrá una medida de AUC de 1, lo que nunca ocurre en el mundo real, y un modelo de bajo rendimiento tendrá una medida de AUC de 0,5, lo que representa que la predicción no es mejor que aleatoria, como se muestra a continuación. diagrama. El ROC-AUC no depende de la distribución de clases en el conjunto de datos, por lo que es una métrica preferida para datos desequilibrados: (Devisetty, 2022, p. 172)

Figura 23

Un ejemplo de un gráfico ROC-AUC



Nota. Tabla obtenida del libro “DeepLearning for Genomics” página 191.

Definición de Términos del Estudio

Limitaciones

Para la construcción del modelo no se cuenta con información de datos demográficos ni genéticos de personas en el Perú. Se van a usar bases de datos públicas, que contienen información de datos genéticos de personas de diferentes partes del mundo.

Esta data se está extrayendo del repositorio público Google Cloud Life Sciences, la cual cuenta con más de un millón de datos registrados, esta data es anónima, por lo cual no es posible rastrear a los colaboradores, ya que se obtiene de diferentes regiones del mundo, protegiendo así la protección de datos personales de los pacientes, recalcando también que esta data fue extraída bajo su consentimiento de cada paciente.

Delimitaciones

- Diseño del modelo para las alteraciones de los genes BRCA1 y BRCA2 enfocado a la detección del cáncer de mama.

El trabajo no contemplará:

- Implementación del modelo predictivo.

Resumen

El presente capitulo tiene como finalidad exponer la problemática y mortalidad del cáncer de mama sobre la población, exponiendo las causas y métodos actuales para combatir esta enfermedad, enfocando nuestra tesis en un método de detección temprana mediante un modelo de predicción de cáncer de mama.

Capítulo II: Revisión de la Literatura

Los genes BRCA 1 y BRCA 2 son genes que inhiben los tumores malignos en los seres humanos, al presentarse una alteración en estos genes, los tumores no se inhiben como se espera, con lo cual el riesgo de sufrir de algún tipo de cáncer aumenta, dicho esto, para poder identificar estos genes es necesario un examen genético utilizando una muestra de sangre, una muestra de saliva o un hisopo.

El antecedente más antiguo que se tiene sobre estudios para la detección preventiva del cáncer utilizando los genes BRCA1 y BRCA2 se remonta a una investigación realizada en el año 2014 en Israel por la bióloga molecular Galit Yahalom, la fue líder del equipo de los laboratorios EventusDX, la cual usando una muestra de sangre permite determinar si existe la mutación en los genes BRCA1 y BRCA2.

Este examen permite que el diagnóstico se realice en un lapso máximo de 3 horas. En palabras de la investigadora, la detección temprana del cáncer o un diagnóstico basado en probabilidad permite tener algún tratamiento preventivo que ayude a no desarrollar la enfermedad.

Octava Pink ha demostrado una efectividad de 95% en la detección temprana del cáncer en mujeres sanas y un 75% en el caso de mujeres que padecen del mismo.

Otro método que se utiliza para la detección temprana del cáncer de mama es el conocido como *bluebox*, el cual tiene como principal ventaja el poder ser utilizado de manera doméstica, ya que se utiliza un método no invasivo, el cual consiste en una muestra de orina dentro de una caja. Este método nació como un prototipo desarrollado por Judit Giró en su tesis de grado de Ingeniería Biomédica por la Universidad de Barcelona en el año 2018.

Lo destacable de este producto es que utiliza algoritmos de inteligencia artificial para poder determinar de manera exitosa la probabilidad de padecer la enfermedad, aunque es muy importante detallar que la inventora del producto indica que su invento no reemplaza para nada la consulta que se puede hacer con un profesional de salud, sino que esto sirva como una actividad complementaria durante el diagnóstico preventivo de la enfermedad.

Según se menciona en el libro "Artificial Intelligence in Breast Cancer Early Detection and Diagnosis", antes de estos exámenes no había un diagnóstico 100% preventivo, en la edad moderna los diagnósticos se realizaban cuando la enfermedad ya estaba presente, es decir no es un diagnóstico como tal, sino que es un método para determinar qué tipo de cáncer ya padece una persona.

El termino cáncer se remonta a la antigua Grecia donde Hipócrates de Cos uso por primera vez dicha palabra la cual viene derivada de los términos *Carcinos* y *Carcinoma* los cuales describen una no formada úlcera y un tumor en una forma de úlcera. En el dialecto griego, el cáncer fu descrito e imaginado como un cangrejo llamado *Subphylum Crustacea* debido al hecho que en los cadáveres se veía con la forma de un cangrejo dentro del cuerpo humano. Auleus Cornelius Celsus, generalizó estos términos en el latín, mientras que Galeno Nicón de Pérgamo usó el término *oncos* para describir los tumores. En términos generales, la analogía del cangrejo aún persiste para la descripción de tumores malignos, mientras que *oncos* describe una categoría de cáncer dentro de la oncología.

En el libro "BRCA Variations Risk Assessment in Breast Cancers Using Different Artificial Intelligence Models" se indica que, desde estos primeros estudios a la actualidad, los médicos se han encargado de observar los diferentes órganos del hombre, para determinar el avance del cáncer e identificar las nuevas posibles variantes.

En este estudio se analizaron los datos de un total de 268 pacientes con cáncer de mama para 16 factores de riesgo diferentes, incluidas las clasificaciones de variantes genéticas. En total, se usaron 61 genes BRCA1, 128 BRCA2 y 11 genes BRCA1 y BRCA2 asociados con datos de pacientes con cáncer de mama para entrenar el sistema utilizando el método de inferencia difusa de Mamdani y el método de red neuronal Feed-Forward como software modelo en MATLAB. Se realizaron dieciséis pruebas diferentes en doce sujetos diferentes que no habían sido introducidos al sistema antes. Las tasas de red neuronal fueron del 99,9 % para el éxito del entrenamiento, del 99,6 % para el éxito de la validación y del 99,7 % para el éxito de la prueba. (Yazici, y otros, 2020)

En este estudio se desarrolló un algoritmo considerando todas las características clínicas, demográficas y genéticas de los pacientes para identificar la negatividad de BRCA1/2. Se creó un conjunto de datos experimental con la recopilación de todas las características clínicas, demográficas y genéticas de pacientes con cáncer de mama durante 20 años. Este conjunto de datos constaba de 125 características de 2070 pacientes con cáncer de mama de alto riesgo. Todos los datos se numeraron y normalizaron para la detección de la negatividad de BRCA1/2 en el algoritmo de Deep Learning. Se encontró que las tasas de precisión nearest neighbours (KNN) y del árbol de decisiones (DT) de 9 características que involucran el conjunto de datos 2 son las más efectivas. Se demostró que la eliminación de los datos innecesarios en el conjunto de datos al reducir la cantidad de características aumenta la tasa de precisión del algoritmo en comparación con el DT. La negatividad de BRCA1/2 se identificó sin realizar la prueba del gen BRCA1/2 con una precisión del 92,88%.

A continuación, en la siguiente tabla se detalla una comparación computacional entre métodos de aprendizaje para la detección del cáncer de mamá.

Tabla 5

Comparación computacional de varios métodos de aprendizaje automático para la detección del cáncer de mama

Referencia	Metodología	Base de datos	Rendimiento	Dataset	Comentarios
Gc S., R. Kasaudhan, T. K. Heo, and H. D. Choi, "Variability Measurement for Breast Cancer Classification of Mammographic Masses," in Proceedings of the 2015 Conference on Research in Adaptive and Convergent Systems (RACS), Prague, Czech Republic, pp. 177–182, 2015.	SVM	Mamografía	Variance: 95%, Range: 94%, Compactness: 86%	Digital Database for Screening Mammography (DDSM)	SVM es uno de los mejores contendientes por ser un clasificador ideal.
Wang C., W. Wang, S. Shin, and S. I. Jeon, "Comparative Study of Microwave Tomography Segmentation Techniques Based on GMM and KNN in Breast Cancer Detection," in Proceedings of the 2014 Conference on Research in Adaptive and Convergent Systems (RACS '14), Towson, Maryland, 2014, pp. 303–308.	GMM, KNN	Imagen de tomografía	Sensitivity: KNN – 87%, GMM – 67%, Accuracy: KNN – 85%,GMM – 75%, MCC: KNN – 67%, GMM – 48%	Electronics and Telecommunications Research Institute (ETRI)	Tanto KNN como GMM son candidatos adecuados, pero KNN sigue siendo el mejor de los dos.

Chowdhary C. L., and D. P. Acharjya, "Breast Cancer Detection using Intuitionistic Fuzzy Histogram Hyperbolization and Possibilitic Fuzzy c-Mean Clustering Algorithms with Texture Feature-based Classification on Mammography Images," in Proceedings of the International Conference on Advances in Information Communication Technology & Computing, Bikaner, India, 2016, pp. 1–6	SVM, KNN, Rough Set Data Analysis (RSDA)	Mamografía	Accuracy: Normal – 100%, Benign – 96.7%, Malignant – 94%	Mammographic Image Analysis Society (MIAS)	Este método mostró 94% precisión para detectar lesiones mamarias malignas
Aminikhanghahi S., S. Shin, W. Wang, S. I. Jeon, S. H. Son, and C. Pack, "Study of Wireless Mammography Image Transmission Impacts on Robust Cyber-aided Diagnosis Systems," Proc. 30th Annu. ACM Symp. Appl. Comput. - SAC ' 15, pp. 2252–2256, 2015.	SVM, GMM	Mamografía	MCC: SVM – 78.8%, GMM – 72.06%, Sensitivity: SVM –82%, GMM –-84%, Specificity: SVM – 96%, GMM – 86%	DDSM University of South Florida	SVM tiene una mejor precisión, pero GMM es más seguro.
Durai S. G., S. H. Ganesh, and A. J. Christy, "Novel Linear Regressive Classifier for the Diagnosis of Breast Cancer," In Computing and Communication Technologies (WCCCT), 2017 World Congress on 2017.	Logical Regression Classifier (LRC)	Datos estándar	Accuracy: LRC – 99.25, B-Flow Imaging (BFI) – 95.46, Iterative Dichotomiser (ID) – 3-92.99, J48 – 98.14, SVM – 96.40	University of California, Irvine (UCI)	LRC es el más preciso de todo.

Wang H., and S. W. Yoon, “Breast Cancer Prediction Using Data Mining Method,” IIE Annu. Conf. Expo 2015, pp. 818–828, 2015.	SVM, ANN, NB, AdaBoost tree, PCA	Datos estándar	Highest accuracy: WBC – 97.47% (PCs-SVM), WDBC – 99.63% (PCj-ANN)	WBC/WDBC	PCA puede ser un factor crítico para mejorar el rendimiento.
S. Hafizah, S. Ahmad, R. Sallehuddin, and N. Azizah, “Cancer Detection Using Artificial Neural Network and Support Vector Machine: A Comparative Study,” J. Teknol, vol. 65, pp. 73–81, 2013.	ANN, SVM	Datos estándar	Accuracy: SVM – 99.51%, ANN– 98.54%, Sensitivity: SVM –99.25%, ANN – 99.25%, Specificity: SVM – 100%,	WDBC	SVM es mejor que ANN, aunque ambos clasificadores tuvieron un alto rendimiento.
Azar A. T., and S. A. El-Said, “Performance Analysis of Support Vector Machines Classifiers in Breast Cancer Mammography Recognition,” Neural Comput. Appl., vol. 24, no. 5, pp. 1163–1177, 2014.	ST-SVM, PSVM, LSVM, NSVM, LPSVM, SSVM	Mamografía	Highest accuracy: LPSVM – 97.1429, Highest sensitivity: LPSVM – 98.2456, Highest specificity: SSVM, NSVM – 96.5517, Highest ROC: LPSVM-99.38	WDBC	LPSVM tiene el más alto rendimiento.
Deng C., and M. Perkowski, “A Novel Weighted Hierarchical Adaptive Voting Ensemble Machine Learning Method for Breast Cancer Detection,” Proc. Int. Symp. Mult. Log., vol. 2015, pp. 115–120, 2015.	WHAVED Neuro Fuzzy (NF) rule-based method, DT, Naive Baiyes (NB), SVM	Mamografía	Accuracy: Disjunctive Normal Form (DNF) – 65.72, DT – 94.74, NB – 84.5, SVM – 99.54, Hybrid – 99.54, KNN – 97.14, Quadratic classifier – 97.14, WHAVE – 99.8	WDBC	WHAVE ha demostrado lograr el más alto rendimiento valor del 99,8%.

Rehman A. U., N. Chouhan, and A. Khan, "Diverse and Discriminative Features Based Breast Cancer Detection Using Digital Mammography," 2015 13th Int. Conf. Front. Inf. Technol., pp. 234–239, 2015.	SVM RBF kernel	Mamografía	Highest accuracy: Model I – 76%, Model II – 68%, Model III – 80%, Highest specificity: Model I – 76%, Model II – 64%, Model III – 76%	Mammographic Image Analysis Society (MIAS)	El modelo III mostró la mejor resultados de los tres
Mejia T. M., M. G. Perez, V. H. Andaluz, and A. Conci, "Automatic Segmentation and Analysis of Thermograms Using Texture Descriptors for Breast Cancer Detection," 2015 Asia-Pacific Conf. Comput. Aided Syst. Eng., pp. 24–29, 2015	KNN	Termograma	Accuracy: Normal – 94.44%, Abnormal – 88.88%	Federal Fluminense University Hospital	Implementación de KNN mejorado la precisión: 88,88% para anormal, mientras que 94.445 para normal.
Ayeldeen H., M. A. Elfattah, O. Shaker, A. E. Hassanien, and T.-H. Kim, "Case-Based Retrieval Approach of Clinical Breast Cancer Patients," 2015 3rd Int. Conf. Comput. Inf. Appl., pp. 38–41, 2015	Bayes Net (BN), Multiclass classifier, DT, RBF, RF	Transfusión de sangre	Highest Accuracy: 99% (RF algorithm)	Department of Biochemistry and Molecular Biology of Kasr Alainy	El algoritmo de RF mostró la resultado más alto, con un 99% de rendimiento
Avramov T. K., and D. Si, "Comparison of Feature Reduction Methods and Machine Learning Models for Breast Cancer Diagnosis," Proc. Int. Conf. Comput. Data Anal. - ICCDA ' 17, pp. 69–74, 2017	Logistic regression (LR), DT.KNN, Cubic SVM (CSVM)	Imagen de microscopio digital	Highest accuracy: 98.56% (SVM, CSVM)	UCI	SVM y CSVM dieron la mejor precisión, con una precisión mejorada del 98,56 %.

Ngadi M., A. Amine, and B. Nassih, “A Robust Approach for Mammographic Image Classification Using NSVC Algorithm,” Proc. Mediterr. Conf. Pattern Recognit. Artif. Intell. – MedPRAI 2016, pp. 44–49, 2016	NSVC	Mamografía	Accuracy – 99%	UCI	NSVC fue mejor que los otros métodos.
Jiang Z., and W. Xu, “Classification of Benign and Malignant Breast Cancer Based on DWI Texture Features,” ICBCI 2017 Proceedings of the International Conference on Bioinformatics and Computational Intelligence 2017	RF-Recursive Feature Elimination (RF-RFE) method	Imagen de resonancia magnética ponderada por difusión	Highest accuracy: RF-RFE and RF, Histogram + GLCM 77.05, Highest sensitivity: RFRFE and RF, Histogram + GLCM 84.21, Highest specificity: RF-RFE and RF, Histogram + GLCM 65.21, Highest AUC: RF-RFE and RF, Histogram + GLCM 0.76	Zhejiang Cancer Hospital	La textura basada en características es muy importante en rendimiento estabilizador y mejorar la detección.
Salma M. U., “Fast Modular Artificial Neural Network for the Classification of Breast Cancer Data,” Proc. Third Int. Symp. Women Comput. Informatics - WCI ’ 15, pp. 66–72, 2015.	Fast modular artificial neural network (FM-ANN)	Radiografía	Highest accuracy – 99.96% (KDD)	WBCD, KDD Cup 2008	Comparando los resultados, FM ANN demostró ser más preciso.

Bevilacqua V., A. Brunetti, M. Triggiani, D. Magaletti, M. Telegrafo, and M. Moschetta, "An Optimized Feed-forward Artificial Neural Network Topology to Support Radiologists in Breast Lesions Classification," Proc. 2016 Genet. Evol. Comput. Conf. Companion - GECCO ' 16 Companion, pp. 1385–1392, 2016.	Optimized ANN	Imagen de resonancia magnética	Highest accuracy – 100%, Average accuracy – 89.7%, Sensitivity – 89.08%, Specificity – 90.46%	Radiologists of the University of Bari Aldo Moro	Los resultados fueron significativamente mejorado mediante el uso de la algoritmo

Nota. Tabla obtenida del libro "Cancer Prediction for Industrial IoT 4.0" página 76.

Según la información mostrada, en el libro “Cancer Prediction for Industrial IoT 4.0” se concluye lo siguiente “El principal candidato para la detección del cáncer de mama es SVM, para poder seleccionar una alternativa a SVM, es necesario realizar un balance entre precisión y facilidad de implementación, el que obtiene mejores resultados con respecto a precisión es ANN, pero los resultados más rápidos y con facilidad KNN. Dependiendo del problema, se puede tomar una mejor decisión.” (Gupta, Jain, Solanki, & Al-Turjman, 2022)

Resumen

En este capítulo se presentan los trabajos o estudios realizados para la detección temprana de cáncer de mamá utilizando una herramienta relacionada a la inteligencia artificial.

Capitulo III: Metodología

Diseño de la Investigación

La presente investigación ha sido diseñada como una investigación tanto cualitativa como cuantitativa. Es cualitativa debido a que los datos de entrada del modelo es un valor no numérico (secuencia de ADN), mientras que la parte cuantitativa es debido a que se hace uso de un algoritmo que hace uso de patrones, correlaciones y una relación de causa efecto entre la variable de entrada y el resultado.

Justificación del Diseño

La presente investigación tiene por finalidad buscar una alternativa más rápida y económica para la detección temprana del cáncer de mama tomando como referencia la secuencia genética de un determinado paciente, la cual se obtiene luego de analizar una gota de sangre o saliva con la máquina conocida como secuenciador genético.

Población

La presente investigación usa como población toda información documentada en la plataforma Google Cloud Life Sciences. En este repositorio está disponible la información sobre las distintas enfermedades genéticas, la secuencia genética que origina la enfermedad (mutación) y el gen en el cual se encuentra dicha mutación, dentro de la cual se cuenta con más de un millón de datos registrados en dicho repositorio.

Muestra

De la información disponible en Google Cloud Life Science estamos tomando como muestra toda la información sobre mutaciones en los genes BRCA 1 y BRCA 2, en los cuales se encuentra la mutación que origina el cáncer de mama, esta muestra la hemos complementado con

la información disponible en el repositorio BRCA Exchange, teniendo un total de 68962 datos que serán utilizados para alimentar el modelo de Deep Learning.

Consentimiento Informado

La legislación peruana para el tratamiento de la información sensible de las personas publicó la ley 29733, la cual regula el uso de los datos sensibles de las personas, la transferencia de la información hacía un tercero dentro del territorio nacional y la transferencia de la información hacía una ubicación fuera del país.

La ley de protección indica que el consentimiento a otorgar debe ser inequívoco, exacto y explícito. También se debe tener en cuenta que la ley tiene 4 aspectos que se protegen con este consentimiento los cuales son Protección en el uso de los datos personales, transferencia de los datos personales hacía un tercero dentro del territorio nacional, transferencia de los datos personales hacía un tercero fuera del país (transfronterizo) u el uso de la imagen y la información sensible de la persona.

Para la presente investigación, dado el alcance que se tiene, es importante que nosotros solicitemos el consentimiento para la transferencia de datos a nivel transfronterizo, dado que vamos la herramienta Google Colab para el procesamiento de estos datos. Dado que los resultados que se obtengan de las personas serán para fines académicos y serán anónimos, no se va a ser necesario solicitar un consentimiento para el uso de sus datos personales, imagen e información sensible.

Procedimiento de Recolección de Datos

La recolección de los datos a estudiar ha sido sencilla ya que al usar la información que existe en el repositorio BRCA Exchange, sólo hemos necesitado exportar dicha información y

realizar el proceso de Transformación de los Datos el cual se detalla en la sección del mismo nombre dentro del Capítulo IV.

Instrumentos de Medición

Para la medición de los resultados, hemos usado un método de comparación entre los datos que están disponibles en el BRCA Exchange y los datos que obtiene el modelo luego de su ejecución. Luego estos resultados serán presentados al médico genetista el cual nos brindará su juicio de experto a los resultados obtenidos del modelo.

Análisis e Interpretación de Datos

Todo lo referente al análisis e interpretación de los datos está descrito en el capítulo IV, en la sección Verificación de los Resultados del Modelo. En este capítulo se muestran los gráficos que permiten verificar a nivel estadístico la fiabilidad del modelo propuesto y su % de éxito en las predicciones.

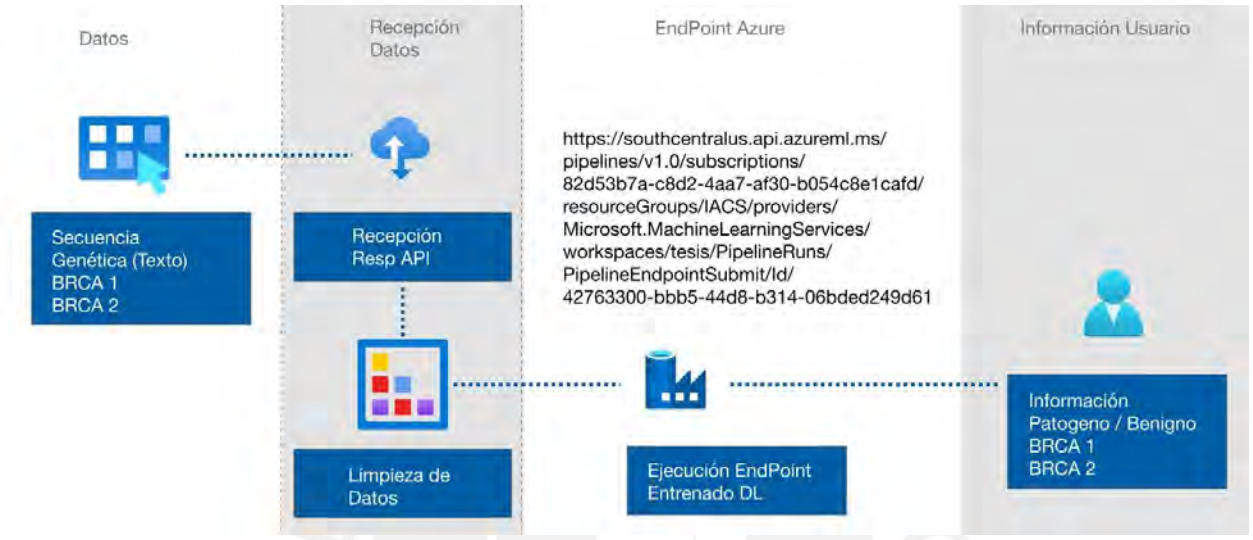
Validez y Confiabilidad

Los resultados del modelo serán sometidos a juicio de experto por parte del médico genetista que nos ha asesorado durante el desarrollo de la presente investigación, con lo cual los resultados obtenidos tendrán el respaldo de las investigaciones realizadas por el mismo, cabe resaltar que, por la naturaleza de la presente investigación, se tiene una población de expertos en el tema muy reducida, siendo un total de 28 profesionales en el Perú.

Para el respectivo juicio de experto se está proporcionando la ruta de un API al médico genetista, con el fin de que pueda validar la confiabilidad del modelo presentado, comparando así, de manera interna la información respectiva para dar su juicio final, para esto se presenta la siguiente arquitectura con la que trabajará el API; finalmente se menciona que dicho enlace estará únicamente habilitado el tiempo necesario para la validación del experto.

Figura 24

Publicación de servicio de deep learning para juicio de experto



Se detalla en los anexos los documentos respectivos sobre el presente.

Resumen

En el presente capítulo se explican las bases de la investigación realizada, así como desde donde hemos obtenido la información y algunas limitaciones que tenemos desde el punto de vista legal.

Capitulo IV: Presentación y Análisis de Resultados

A continuación, en cumplimiento del primer objetivo de la tesis, se contactó con un médico genetista, el doctor Yasser Sullcahuaman, con más de 15 años de experiencia en genética clínica y bioinformática, cofundador de la Maestría en Genética Humana y Residencia en Genética Medica de la UPCH, Exjefe del Servicio de Genética y Biología Molecular del INEN. Actual director médico de IGENOMICA; obteniendo así un total de 3 reuniones de apoyo el cual nos brindó, obteniendo así el correcto rumbo y entendimiento sobre la presente tesis, estas reuniones se detallan a continuación.

Sesión 1 (03/06/2022): en esta reunión el doctor Sullcahuaman realizó una presentación introductoria sobre la genética, en donde recordamos muchos temas como son la herencia, los genes, la replicación celular, el funcionamiento de los aminoácidos y proteínas. Luego el doctor nos explicó temas relacionados a su día a día laboral como son las alteraciones genéticas y determinación probabilística de ciertas enfermedades como es el cáncer, sobre las alteraciones genéticas aprendimos que éstas se heredan de los padres hacia los hijos y la importancia de determinarlas tempranamente para poder estar preparados para poder realizar un tratamiento o terapia personalizado. Un primer indicio que usa el doctor para determinar alguna alteración genética es a partir de un rasgo físico de la persona, por ejemplo, una mancha en la parte inferior en los labios de una niña de 8 años puede causar un cáncer de mamá cuando ésta tenga 40 años, para estos casos se recomienda realizar una prueba de sangre para extraer la información genética, esto puede durar 1 o 2 días, luego sigue un proceso de interpretación de la variación que puede tomar semanas o meses donde se realiza búsquedas a bases de datos de alteraciones genéticas para determinar si hay una relación de una alteración con una enfermedad. A partir de eso se determina la probabilidad de contraer la enfermedad y se le brinda recomendaciones al

cliente dependiendo de la posible enfermedad, por ejemplo, si está supera los 70% para un cáncer de mamá se puede recomendar un tratamiento como resonancia o retirar el seno u ovario antes de empezar una quimioterapia.

Un punto importante para tener en cuenta es que no se puede compartir la información genética de los pacientes debido a la ley de protección de los datos personales, también es importante entender sobre las barreras que existen para que las personas puedan realizarse una prueba genética y acceder a su información genética. por ejemplo, todas las personas que son detectadas con cáncer deberían pasar por esta evaluación, pero solo pasan menos del 1%, debido al costo de las pruebas y a que los seguros no lo cubren. Un estudio para detectar cáncer cuesta 800 dólares, un estudio para detectar 10000 enfermedades cuesta 1400 dólares y un estudio para realizar el genoma complejo cuesta 4000 dólares.

Figura 25

Paciente Mujer con Diagnostico de Cáncer de Mama a los 34 años.

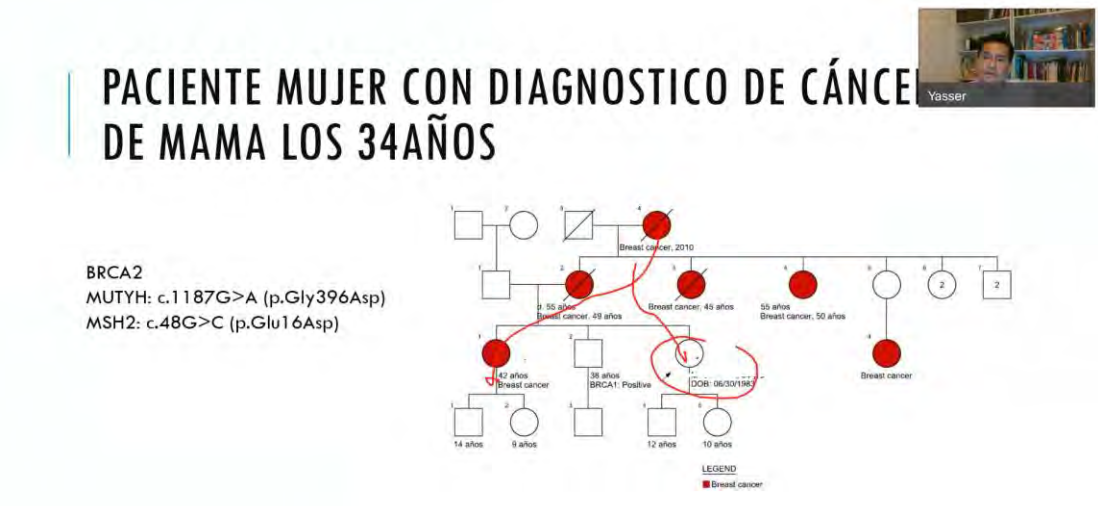


Figura 26

Resultados de Genes con Variación Patogénica en Población Peruana.

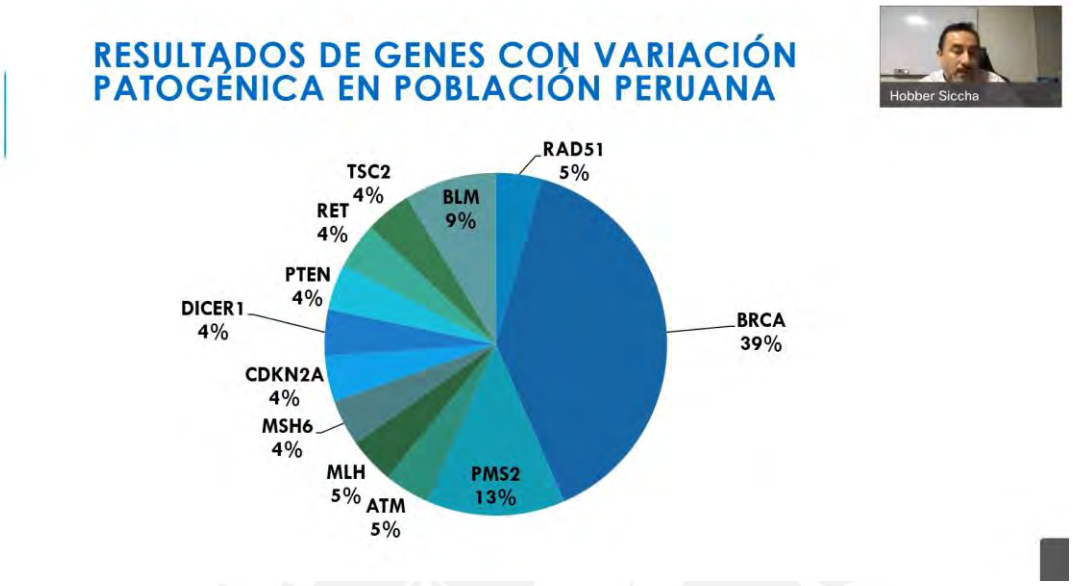


Figura 27

Cambios del color de piel, Hipo Pigmentación e Hiperpigmentación en tres pacientes.

A partir de esta reunión de entendimiento se identificó la necesidad y se plantearon los siguientes beneficios:

Beneficios para el paciente:

- Reducción en los tiempos de espera y de dinero en el diagnóstico preventivo del cáncer.
- Predecir el riesgo personal y familiar de presentar cáncer.
- Tomar decisiones anticipadas para realizar un tratamiento personalizado con el fin de evitar la muerte por cáncer o tener una mejor calidad de vida.

Beneficios para la sociedad:

- Diagnóstico preventivo rápido y a menores costos.
- Posibilidad de identificación de nuevos tratamientos de cáncer con orígenes en variaciones genéticas, la quimioterapia no sería la única alternativa.
- Escalamiento para su uso en otros genes para identificar enfermedades con origen en alteraciones genéticas.
- Expandir el conocimiento de enfermedades con origen en alteraciones genéticas

Beneficios para un gerente de TI en el sector salud:

- Contar con una propuesta innovadora basada en IA, única en el Perú, que puede ser implementada y desplegada por un equipo de científicos de datos.
- Al implementar la solución se tendrá una herramienta de apoyo para tomar decisiones más certeras, en menos tiempo y más confiables.

Sesión 2 (06/08/2022): Para esta reunión llegamos con una propuesta de solución a la necesidad expuesta por el doctor Sullcahuaman, para esto previamente le proporcionamos un cuestionario de 10 preguntas para obtener un mayor entendimiento sobre las mutaciones en los genes y sobre los datos que necesitamos para poder ejecutar nuestra solución propuesta. A continuación, se detalla las preguntas y respuestas:

1. **¿Qué tipos de mutaciones de los genes BRCA1 /BRCA2 prevalecen más en la población peruana?**

   Todos los tipos de mutación (SNP, CNV, Re-arreglos, etc.) No hay muchos estudios, les envió las dos publicaciones que realice, una es la del 2015 y otra hace un par de semanas Julio 2022

   - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4374018/
   - https://ascopubs.org/doi/full/10.1200/GO.22.00104

2. **¿Posee información de las estadísticas de cáncer de mama en el Perú? (departamento, etnia, edad, etc.)**

   https://portal.inen.sld.pe/indicadores-anuales-de-gestion-produccion-hospitalaria/

3. **¿Qué tipo de análisis se puede realizar con los datos arrojados por el secuenciador para determinar el cáncer de mama? ¿Cuánto demora realizar ese análisis?**

   Se puede diagnosticar si es hereditario o no, determinar el tipo de mutación, las variantes del gen, la frecuencia de mutaciones en una población. Demora 4 a 6 semanas

4. **Explicar proceso de reconocimiento de cáncer de mama según la mutación del BRCA 1 y BRCA 2 (Como diferencia entre el cáncer de Ovario y Cáncer de mama)**

   La persona con cáncer de mama puede o no tener mutación de BRCA1 o BRCa2 u otros genes. La diferencia de cáncer de mama u ovario se basa en la ubicación del cáncer en un determinado órgano, hay casos en los cuales se presenta cáncer de páncreas, estomago o melanoma en personas con mutación patogénica del gen BRCA.

5. **¿Qué variables o factores de riesgos se utilizan para poder detectar el cáncer de mama?**

   Edad, sexo, estilos de vida, obesidad, anticonceptivos, etc.

6. **¿El resultado del análisis genético brindado al paciente es una probabilidad de ocurrencia del cáncer?**

   Si

7. **¿Qué tanto pueden influir los antecedentes familiares en el cáncer de mama?**

   Depende de la alteración o variante genética presente o ausente en la familia

8. **¿Qué es el Score Grantham y cómo nos ayuda en la detección del cáncer?**

   No lo sé, revisare cual es el uso de este score

9. **¿Posee información arrojada por el secuenciador para poder usada en nuestra tesis?**

   Si no lo tuviera que base de datos públicas podríamos usar.

   La información está protegida de acuerdo con la ley, se debe pedir autorización por escrito a cada paciente mediante un consentimiento informado que debe tener la firma del paciente, la firma del médico tratante, la firma del investigador, de un testigo y ser avalado por el comité de ética de la universidad y aprobado por el Dpto. de investigación del INEN

10. **Nos podría brindar una explicación de cáncer de mama enfocado en la genética**

    El cáncer de mama es producto de la acumulación aleatoria de mutación en la mama, estas mutaciones pueden ser de origen hereditario o esporádico.

    - https://www.cancer.gov/espanol/cancer/causas-prevencion/genetica

En la reunión entramos a detalle a cada una de las respuestas y le mostramos un bosquejo de la solución a la problemática detectada en la primera reunión, esta solución se basa en un modelo de redes neuronales que predice si el cáncer de mamá es patógeno o benigno con base solo en información genética. El doctor sugirió complementar la información genética con información relacionados propios de la persona y a factores medioambientales como edad, si es

fumador, el género y su estilo de vida, sin embargo, nos indicó que eso sería complicado porque no habría mucha información disponible comparado con la información genética que puede ser accedida de forma gratuita a través de bases de datos con Ensembl (https://www.ensembl.org/). En dicha página se nos explicó como es el proceso de análisis y entendimiento del código de secuencia del gen, como son:

- Entendimiento de la lectura del código genético
- Identificación de las proteínas
- Entender la posición del gen, región codificante y posición del aminoácido
- Entendimiento del código de colores de la secuencia genética
- Como es una secuencia normal de un código genético
- Como identificar la mutación en el código genético

Figura 28

Análisis del código genético del repositorio Ensembl

- **Sesión 3 (05/11/2022):** Para esta reunión se tuvo como objetivo validar los datos de entrada que usaríamos en nuestro modelo predictivo, para eso el doctor Sullcahuaman explicó a más detalle los 4 criterios o categorías para determinar si una alteración es patógena o benigna.
- Frecuencia en la población: si se encuentra que la alteración se encuentra en más del 1% de la población es una variante benigna, es algo propio de los seres humanos.
- Tipo de alteración genética: existen variantes genéticas como la Missense, Nonsense y Frameshift. En el caso de alteraciones Nonsense las proteínas dejan de sintetizar lo que podría determinar que la variante en el gen sea patógena. En este tipo de alteraciones también se modifican los aminoácidos lo que afectaría a las proteínas.
- Predicción computacional: en este caso se nos mostró el ejemplo en donde diversos animales pueden tener en una determinada posición un nucleótido de Adenina y el ser humano Timina, esto podría ser un indicador de una variación patógena. Las alteraciones de tipo Splicing también se encuentran en esta categoría ya que afectan el marco de lectura genética.
- Criterio clínico: son alteraciones genéticas hereditarias y evidentes que se pueden visualizar de generación en generación.
- Luego de la explicación de estos criterios pudimos identificar las variables candidatas para poder ejecutar nuestro modelo de predicción propuesto.

Fase 1: Comprensión de los datos

Recopilación inicial de los datos

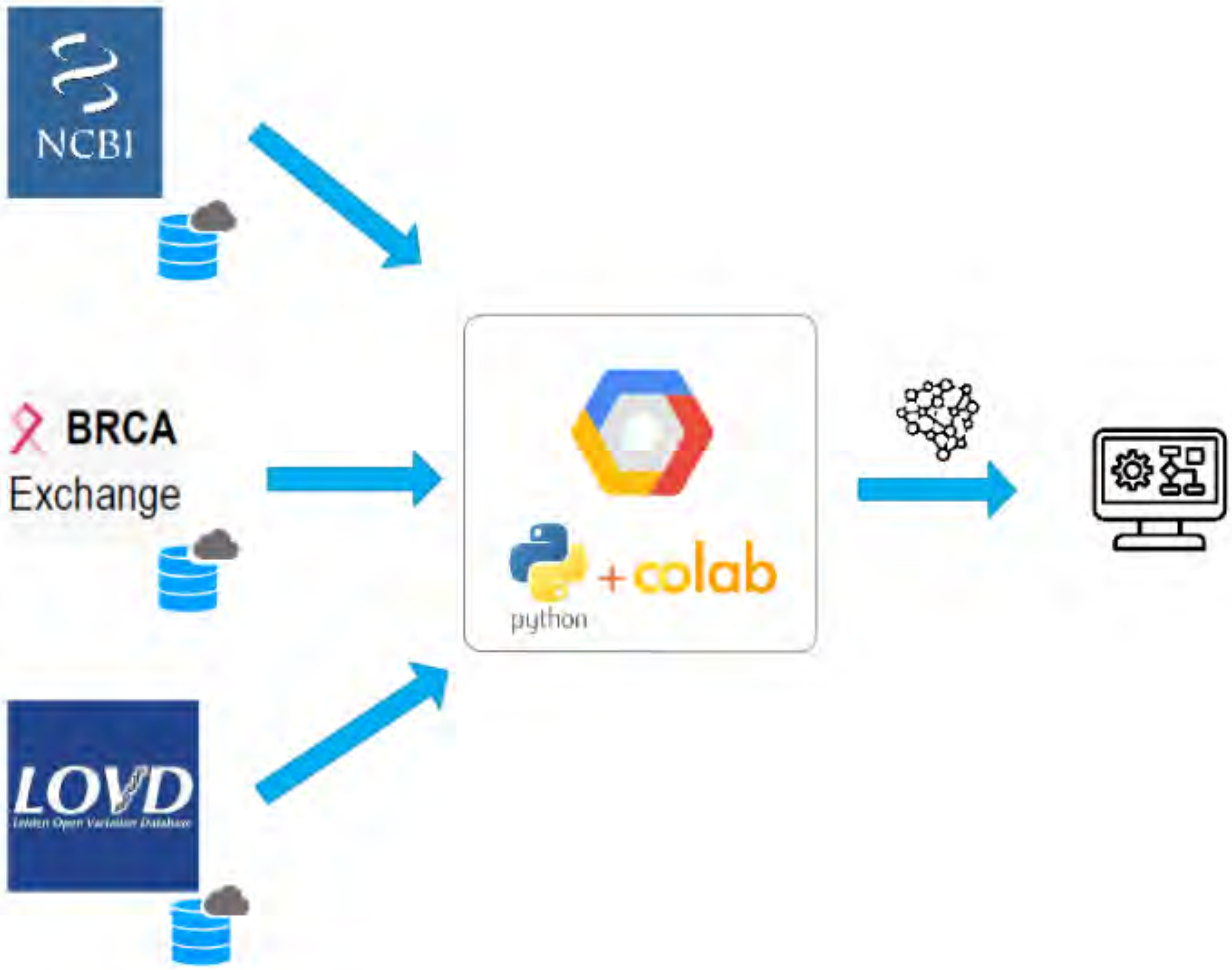
Para el presente trabajo se está trabajando con información pública que se encuentra disponible en la plataforma NCBI (National Center of Biotechnology Information). Esta es una

biblioteca pública de Estados Unidos la cual se encarga de almacenar y actualizar información referente a secuencias genómicas en el repositorio Genbank (Fuente: Wikipedia). Al ser esta información pública, parte del trabajo que realizaremos será exportar esta información desde esta biblioteca y llevarla hacía la plataforma Google Colab, donde se procederá a realizar el análisis y modelo predictivo.

Dicho esto, la arquitectura que se propone para la obtención y análisis de los datos sería el siguiente:

Figura 29

Arquitectura de la solución



Descripción de los datos

Una vez identificados los datos, hemos procedido a fusionar la información de los diferentes repositorios, utilizando como clave primaria el campo Pos y Chr, los cuales hacen referencia al gen en el cual está identificada la mutación y la posición que posee el gen en el cromosoma con lo cual el archivo que utilizaremos para construir nuestro modelo tendría la siguiente estructura:

Figura 30

Estructura de la tabla obtenida luego de cruzar la información de NCBI, BRCA Exchange y LOVD

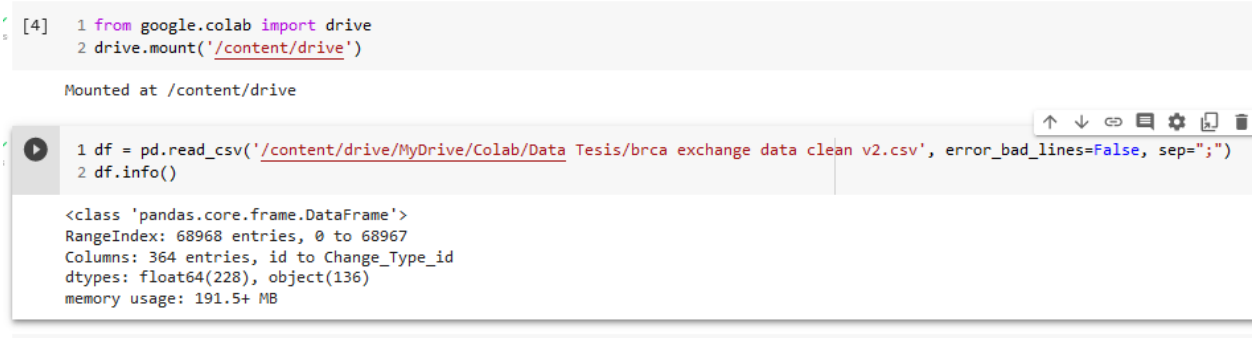
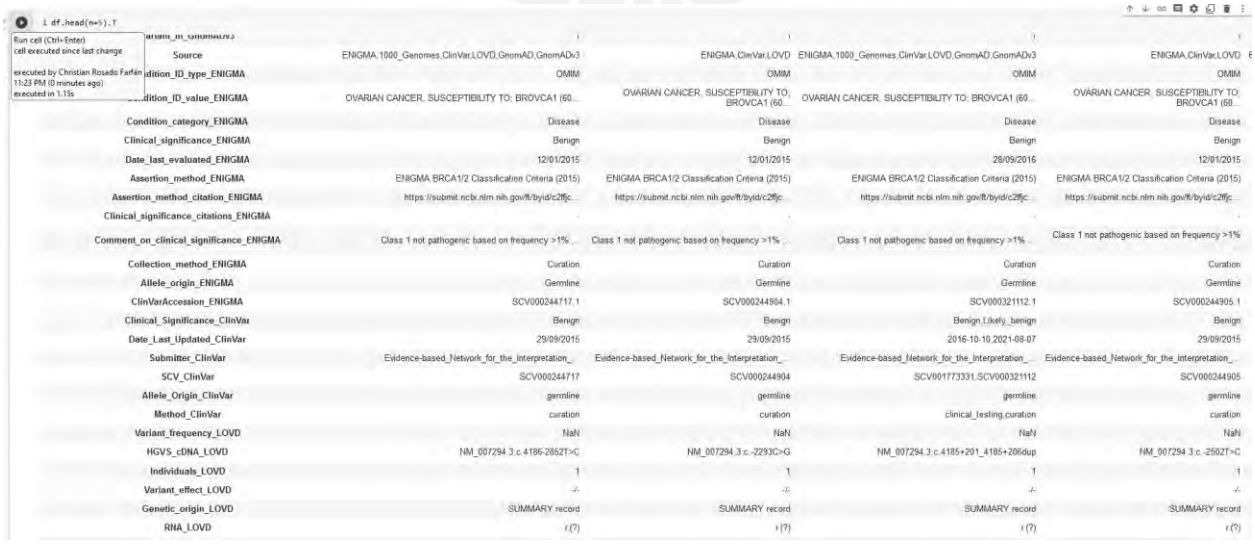


Figura 31

Extracto de los datos encontrados en la tabla que cruza la información



Las variables obtenidas para este análisis son de 364 en total, de los cuales se han identificado 226 datos numéricos que pertenecen a datos estadísticos sobre la identificación de las mutaciones en las distintas regiones del mundo, 133 datos adicionales que no se van a considera para el modelo corresponden a registros que documentan como se ha obtenido la información, es decir información sobre donde se obtuvo el dato, la historia clínica que documenta lo sucedido con el paciente, los datos de las alteraciones genéticas ocurridas, la zona dentro del cromosoma (exon) donde se encuentra la mutación, la secuencia de ADN que posee la alteración y demás datos informativos.

Para poder hacer el símil con lo que esperamos obtener del secuenciador genético, hemos tomado la variable Alt, la cual posee la secuencia proteínica del ADN y convertirla en valores numéricos, el artificio utilizado es el siguiente:

- Adenina (A) = 1
- Citocina (C) = 2
- Guanina (G) = 3
- Tiamina (T) = 4

Exploración de los datos

Dado que se ha realizado una personalización de la columna Alt, y a su vez esta se ha dividido en 4 partes para poder procesar los números sin ningún inconveniente, a continuación, se muestra un detalle sobre el contenido de los datos a analizar, esto incluye una visualización rápida de los 5 primeros registros

Figura 32

Resultados la exploración de los datos cargados en Google Colab

```
[7]  1 df.head(n=5)
```

	Pathogenic_Class	Chr	Pos	Ref	Alt	Alt-1	Alt-2	Alt-3	Alt-4
0	0	17	43085427	A	G	3	0	0	0
1	0	17	43127544	G	C	2	0	0	0
2	0	17	43090737	T	TGTGCGC	434	32	32	0
3	0	17	43127753	A	G	3	0	0	0
4	0	17	43079681	G	C	2	0	0	0

De esta tabla resultado, se pueden extraer los máximos, mínimos y promedios para los datos numéricos.

Figura 33

Exploración de datos numéricos

Feature	Count	Unique Value Count	Missing Value Count	Min	Max	Mean	Mean Deviation	1st Quartile	Median	3rd Quartile	Mode	Range	Sample
Pathogenic_Class	68962	2	0	0	1	0.071068	0.132035	0	0	0	0	1	0.06601
Chr	68962	2	0	13	17	14.99797	1.999998	13	13	17	13	4	4.00005
Pos	68962	47431	0	32314514	43127866	37714379.102723	5365847.221588	32349191	32399563.5	43091756	32395960	10813352	287932
Alt-1	68962	624	0										
Alt-2	68962	426	0										
Alt-3	68962	440	0										
Alt-4	68962	430	0										

A continuación, procedemos a detallar el significado de cada uno de los datos numéricos que serán utilizados en el modelo:

- **Pathogenic_class:** Hace referencia a si la variación que se está analizando es patogénica (1), o es benigna (0)
- **Chr:** En este caso, se hace referencia al cromosoma en el cual se encuentra el gen en que encuentra la mutación (Chr 17 para BRCA 1 y Chr 13 para BRCA 2).
- **Pos:** En esta columna se almacena la ubicación de la mutación dentro del gen.
- **Alt:** Posee la secuencia genética (ATGC) con la mutación documentada. Para nuestro caso hicimos el artificio para poder convertir esta secuencia en números, utilizando la equivalencia ya antes descrita (Adenina (A) = 1, Citocina (C) = 2, Guanina (G) = 3, Tiamina (T) = 4)
- **Alt-1, Alt-2, Alt-3, Alt-4:** porciones de la secuencia genética para un mejor análisis de los datos de entrada, se divide en grupos de 1000 valores.

Verificación de la calidad de los datos

Como la información obtenida es resultado de cruzar las tres bases de datos anteriormente detallada se ha consolidado en un archivo Google Sheet, se van a utilizar sentencias en Visual Basic para la verificación de Calidad de la Información.

Estos comandos se utilizaron para comprobar que los tipos de datos de las columnas a utilizar tengan los valores correctos y no tener problemas en el procesamiento del modelo.

Figura 34

Código VBA utilizado para verificar la calidad de los datos

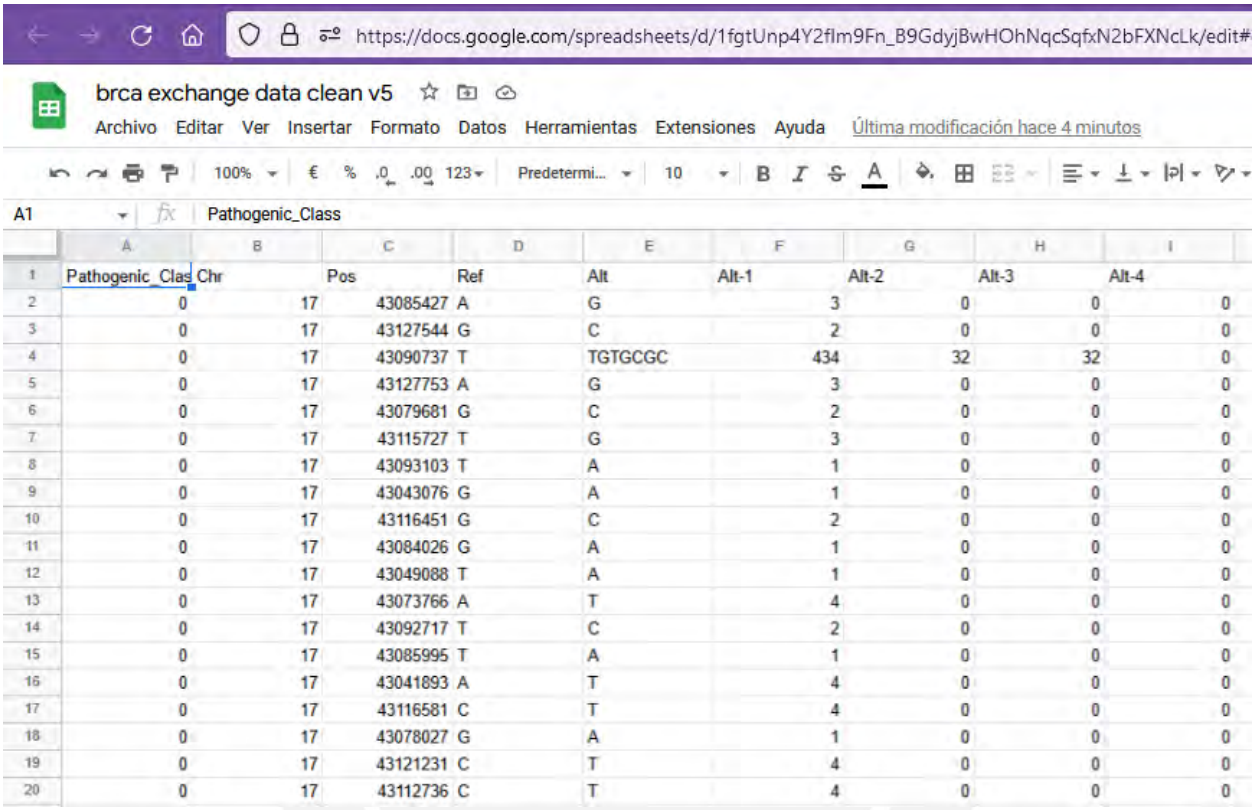
```
Sub VerificaDatosXLS()
    Dim nroColumnas As Integer
    Dim ColumnLetter As String
    Dim nroFilas As Integer

    nroColumnas = 273
    nroFilas = 476

    For C = 1 To nroColumnas Step 1
        ColumnLetter = Split(Cells(1, C + 1).Address, "$")(1)

        columns(ColumnLetter & ":" & ColumnLetter).Select
        Selection.Insert Shift:=xlToRight, CopyOrigin:=xlFormatFromLeftOrAbove
        range(ColumnLetter & "2").Select
        ActiveCell.FormulaR1C1 = "=IFERROR(VALUE(RC[-1]),RC[-1])"

        ColumnLetter = Split(Cells(1, C + 1).Address, "$")(1)

        Selection.AutoFill Destination:=range(ColumnLetter & "2:" & ColumnLetter & "476")
        range(ColumnLetter & "2:" & ColumnLetter & "476").Select
        Selection.Copy

        ColumnLetter = Split(Cells(1, C).Address, "$")(1)

        range(ColumnLetter & "2").Select
        Selection.PasteSpecial Paste:=xlPasteValues, Operation:=xlNone, SkipBlanks:=False, Transpose:=False

        ColumnLetter = Split(Cells(1, C + 1).Address, "$")(1)

        columns(ColumnLetter & ":" & ColumnLetter).Select
        Application.CutCopyMode = False
        Selection.Delete Shift:=xlToLeft

        ColumnLetter = Split(Cells(1, C).Address, "$")(1)

        For r = 1 To nroFilas Step 1
            If range(ColumnLetter & r).Value = "-" Then
                range(ColumnLetter & r).Value = Null
            End If
        Next
    Next
End Sub
```

Fase 2: Preparación de los datos

Para empezar la preparación de los datos, lo primero que realizaremos es la importación de la información hacia la plataforma Google Drive, luego utilizando Google Colab, haremos la conexión para así poder visualizar y manipular los datos con python. Adicional a esto, este mismo archivo será cargado en la plataforma Microsoft Azure Machine Learning Factory con la finalidad de tener dos algoritmos en diferentes plataformas y comparar los resultados.

Figura 35

Carga de información desde Google Drive a Google Colab



https://docs.google.com/spreadsheets/d/1fgtUnp4Y2flm9Fn_B9GdyjBwHOhNqcSqfxN2bFXNcLk/edit#

brca exchange data clean v5

Archivo Editar Ver Insertar Formato Datos Herramientas Extensiones Ayuda Última modificación hace 4 minutos

A1 Pathogenic_Class

	A	B	C	D	E	F	G	H	I
1	Pathogenic_Clas	Chr	Pos	Ref	Alt	Alt-1	Alt-2	Alt-3	Alt-4
2	0	17	43085427	A	G	3	0	0	0
3	0	17	43127544	G	C	2	0	0	0
4	0	17	43090737	T	TGTGCGC	434	32	32	0
5	0	17	43127753	A	G	3	0	0	0
6	0	17	43079681	G	C	2	0	0	0
7	0	17	43115727	T	G	3	0	0	0
8	0	17	43093103	T	A	1	0	0	0
9	0	17	43043076	G	A	1	0	0	0
10	0	17	43116451	G	C	2	0	0	0
11	0	17	43084026	G	A	1	0	0	0
12	0	17	43049088	T	A	1	0	0	0
13	0	17	43073766	A	T	4	0	0	0
14	0	17	43092717	T	C	2	0	0	0
15	0	17	43085995	T	A	1	0	0	0
16	0	17	43041893	A	T	4	0	0	0
17	0	17	43116581	C	T	4	0	0	0
18	0	17	43078027	G	A	1	0	0	0
19	0	17	43121231	C	T	4	0	0	0
20	0	17	43112736	C	T	4	0	0	0

Figura 36

Carga de data desde Google Drive a Google Colab

```
[4]  1 from google.colab import drive
     2 drive.mount('/content/drive')
```

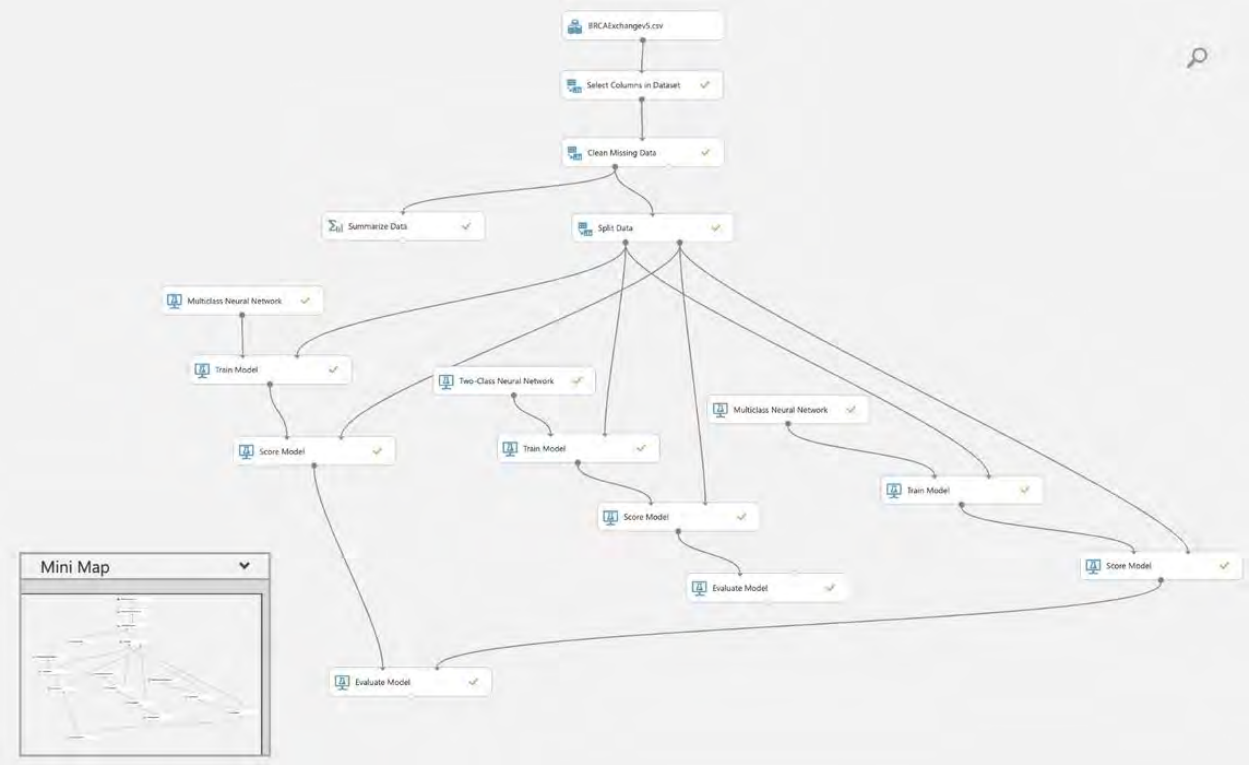
```
Mounted at /content/drive
```

```
[6]  1 df = pd.read_csv('/content/drive/MyDrive/Colab/Data Tesis/brca exchange data clean v5.csv', error_bad_lines=False, sep=";")
     2 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 68962 entries, 0 to 68961
Data columns (total 9 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Pathogenic_Class  68962 non-null  int64 
 1   Chr              68962 non-null  int64 
 2   Pos              68962 non-null  int64 
 3   Ref              68962 non-null  object
 4   Alt              68962 non-null  object
 5   Alt-1            68962 non-null  object
 6   Alt-2            68962 non-null  object
 7   Alt-3            68962 non-null  object
 8   Alt-4            68962 non-null  object
dtypes: int64(3), object(6)
memory usage: 4.7+ MB
```

Figura 37

Carga de datos a Microsoft Azure Machine Learning Factory



A diferencia de Google, Microsoft nos ayuda en la construcción de un workflow que se encarga de la carga, transformación y análisis de los datos, por lo cual en las siguientes fases sólo se mencionará las tareas que se realizarán en Google.

Fase 3: Preparación de los datos

Limpieza de los datos

Para realizar la limpieza de los datos cargados, ya que estos se encuentran en Google Colab, se usará la siguiente sentencia en python que permitirá realizar esta tarea:

Figura 38

Limpieza de datos

```
[28]  1 # Reemplazando comas por puntos:
      2 for i in df.columns:
      3     df[i] = df[i].astype(str).str.replace(',','.') # forzando las cadenas para que reemplace la data
      4
      5 # Dando formato numérico:
      6
      7 # assert not df.any(df.isnan(x))
```

```
[29]  1 for i in df.columns:
      2     #print(isinstance(i,(int, float)))
      3     # if df[i].astype(String) != String: # se desea que toda columna menos 'date' tenga formato numérico
      4     try:
      5       if isinstance(i,(int, float, object)):
      6         df[i] = df[i].astype(float) # continuo
      7     except ValueError as e:
      8       print(e)
      9
```

Construcción de datos

Luego de todo el análisis realizado, junto con el doctor Sullcahuaman, el campo que vamos a utilizar para el análisis será *Pathogenic_Class*, el cual tendrá la información que queremos simular, indicando si una alteración genética se considera benigna (valor = 0) o patogénica (valor = 1).

Utilizando Google Colab, vamos a ejecutar las siguientes instrucciones con lo cual vamos a dividir la información en una proporción de 80-20, utilizando el 80% de los datos para predecir el valor del campo Pathogenic_Class y el 20% para corroborar si la predicción fue exitosa.

Figura 39

Instrucciones python para iniciar la construcción del modelo

```python
from sklearn.model_selection import train_test_split # función para partir el conjunto de datos a entrenaiento y prueba

# Es buena práctica definir una lista con los nombres de las variables explicativas y la variable objetivo a predecir

# División de los datos en train y test
# ==============================================================================
predictores = ['chr', 'pos', 'ref', 'alt', 'alt-1', 'alt-2', 'alt-3', 'alt-4']

objetivo = ['Pathogenic Class']

# División train-test!
X_entrenamiento, X_prueba, Y_entrenamiento, Y_prueba= train_test_split(
    df[predictores], # aquí se introduce el conjunto de variables predictoras
    df[objetivo], # aquí se introduce el objetivo a predecir
    test_size = 0.2, # indicamos que queremos entrenar sobre el 80% de datos y evaluar sobre el 20% restante
    random_state = 1234
)
```

Integración de los datos

Al tener toda la información ya cargada en Google Colab y habiendo ejecutado sentencias previas en VBA, ya tenemos toda la información integrada y no necesitamos un paso adicional. El detalle de las tareas realizadas en este punto se puede consultar en la sección *Recopilación Inicial de los Datos.*

Formateo de los datos

Siguiendo con la metodología CRISP-DM, lo que se va a mostrar a continuación es el modelo planteado, el cual se basa en una *Neural Network*, la cual a su vez utilizará como función de activación *Relu* en las primeras capas y en la última usará la función *Softplus* dado que lo que buscamos es la simulación de valores 1 y 0.

Adicional a esto, hemos aplicado un artificio a la columna *Alt*, la cual posee la secuencia genética que ha producido la mutación, esto se ha cambiado para que los aminoácidos

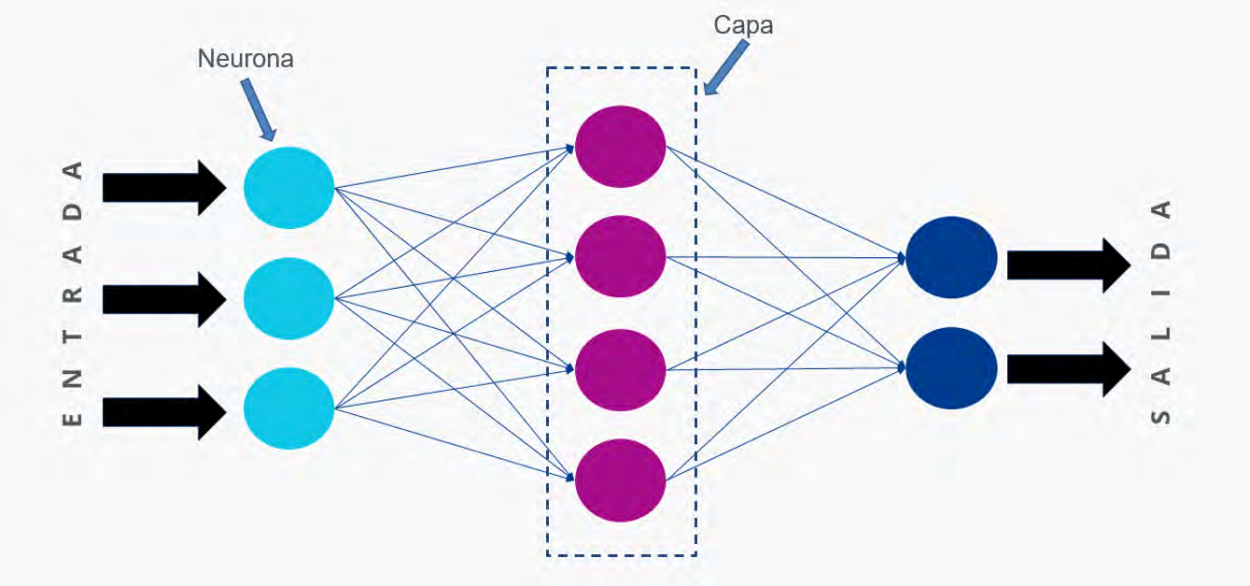
Fase 4: Modelado

Selección de la técnica de modelado

Como se indicó en el punto anterior, la técnica seleccionada será Neural Network, la cual será ejecutada desde Google Colab y nos permitirá obtener mejores resultados.

Figura 40

Vista gráfica de una neural network



Diseño de la evaluación

Para la óptima evaluación del modelo, usaremos las mejores prácticas para el entrenamiento, haciendo que el modelo posea tantas neuronas como variables hayamos identificado para la simulación, pasando a reducir su cantidad en la mitad hasta llegar al valor 1, es decir nosotros en las etapas de análisis hemos identificado 8 variables, con lo cual la neural network poseerá 8, 4, 2, 1 neuronas distribuidas en 4 capas respectivamente.

Figura 41

Construcción de la red neuronal

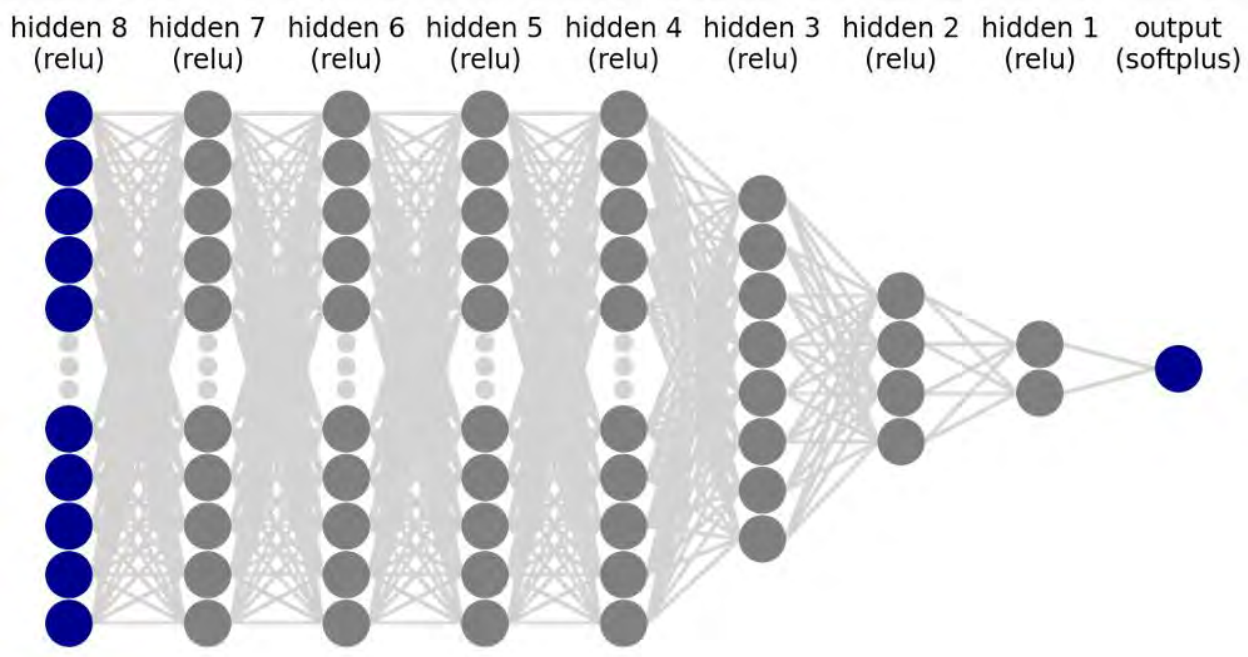
```
from sklearn.model_selection import train_test_split # función para partir el conjunto de datos a entrenaiento y prueba

# Es buena práctica definir una lista con los nombres de las variables explicativas y la variable objetivo a predecir

# División de los datos en train y test
# ==============================================================================
predictores = ['Exon', 'Variant_effect_number', 'transcrip_id', 'protdesc', 'GEN',
               'Minor_allele_frequency_percent_substr', 'Allele_frequency_ExAC', 'type_id']

objetivo = ['Pathogenic Class']

# División train-test!
X_entrenamiento, X_prueba, Y_entrenamiento, Y_prueba= train_test_split(
    df[predictores], # aquí se introduce el conjunto de variables predictoras
    df[objetivo], # aquí se introduce el objetivo a predecir
    test_size = 0.2, # indicamos que queremos entrenar sobre el 80% de datos y evaluar sobre el 20% restante
    random_state = 1234
)
```

Construcción del modelo

A continuación, se muestra de manera gráfica la neural network producida para el modelo propuesto

Figura 42

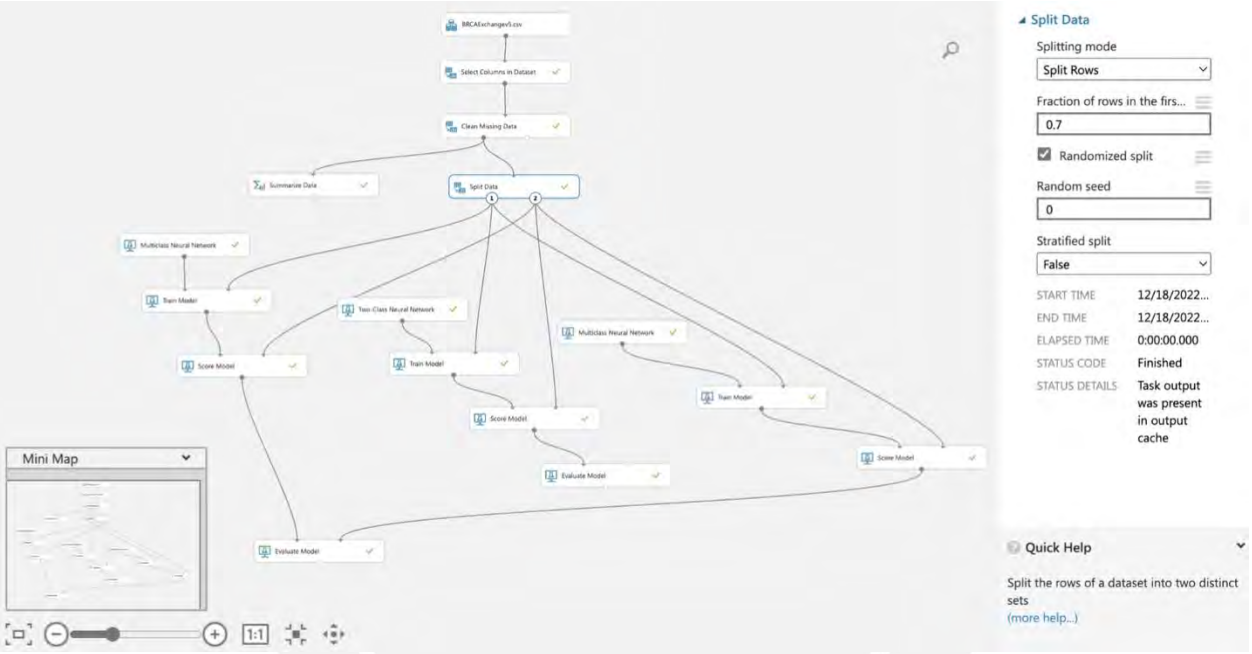
Representación de la neural network construida



Ahora dentro de Machine Learning Factory realizaremos el trabajo de configuración del modelo para poder ejecutar la simulación y obtener los resultados:

Figura 43

Configurando el modelo en machine learning factory (Microsoft Azure)



Verificación de las Predicciones Del Modelo.

Luego de varias ejecuciones del modelo, revisamos los resultados y obtuvimos un nivel de precisión de casi 88% en la plataforma Azure, mientras que en Google sólo pudimos llegar al 84%. Para mejorar esta ratio, las sentencias generadas en la herramienta Google Colab fueron replicadas en Microsoft Azure, con lo cual los resultados del modelo mejoraron y llegamos a un 91% de éxito.

Corroboramos los resultados con la ayuda del doctor sulcahuaman, y nos indicó que eran resultados muy atractivos y positivos.

El doctor si nos comentó que los resultados son muy buenos para los primeros resultados obtenidos y que las siguientes tareas serán la de incluir nuevos datos al modelo y nuevas variables para poder afinarlo.

Figura 44

Estadísticas del modelo ejecutado en machine learning factory mostrando los resultados

Figura 45

ROC del modelo

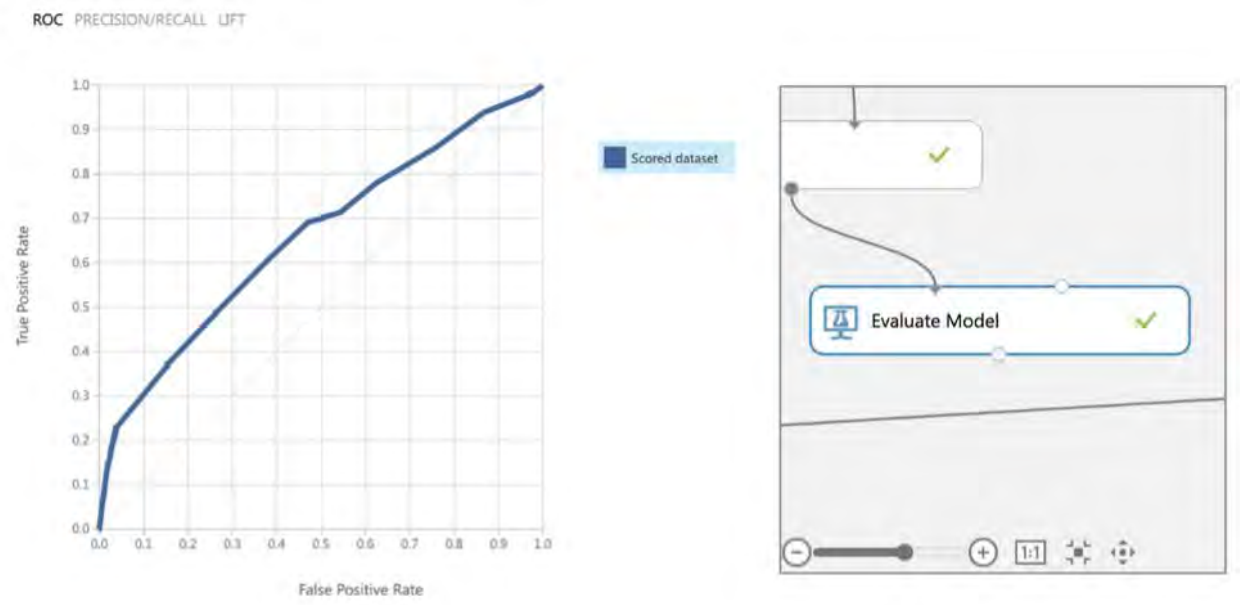
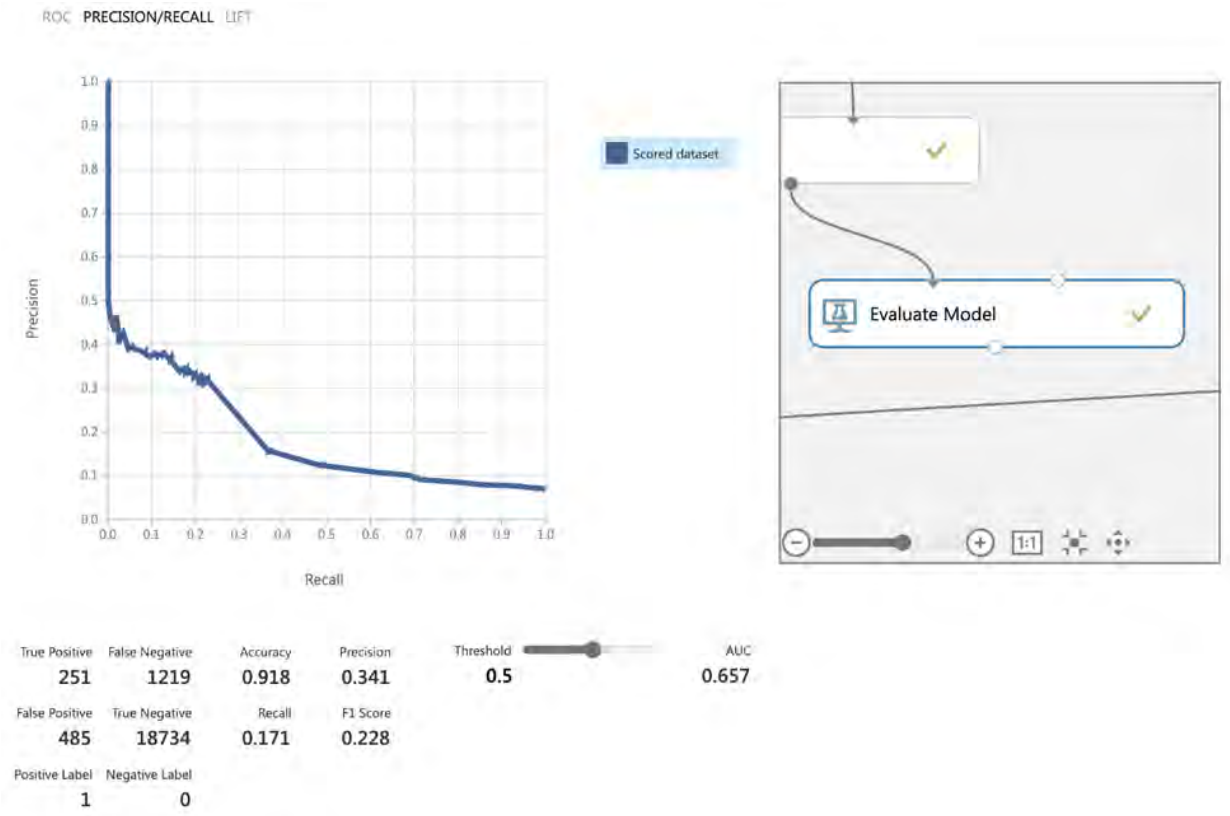


Figura 46

Precisión del modelo construido



Lecciones Aprendidas

Durante el presente trabajo se han identificado varias oportunidades de mejora para hacer madurar el modelo y acercarnos a mejores predicciones. A continuación, enumeramos algunas de ellas:

- **Tener acceso más información**, dado que no tenemos un repositorio público y anónimo en Perú, es importante poder contar con más datos para hacer madurar el modelo con información regional.
- **Consulta con otros expertos,** Si bien el especialista consultado nos ha ayudado entender la necesidad de este tipo de implementaciones, es necesario contar en el equipo con alguien de tecnología que conozca el sector salud, agilizando así la revisión de los datos y el entendimiento de la información obtenida.

- **La importancia de la asesoría de expertos,** el modelo no busca reemplazar la labor que hacen investigadores como el especialista consultado, sino que busca que el trabajo que este realice sea muchísimo más rápido, permitiendo así que estos exámenes sean de costo menor y de menor tiempo de espera.

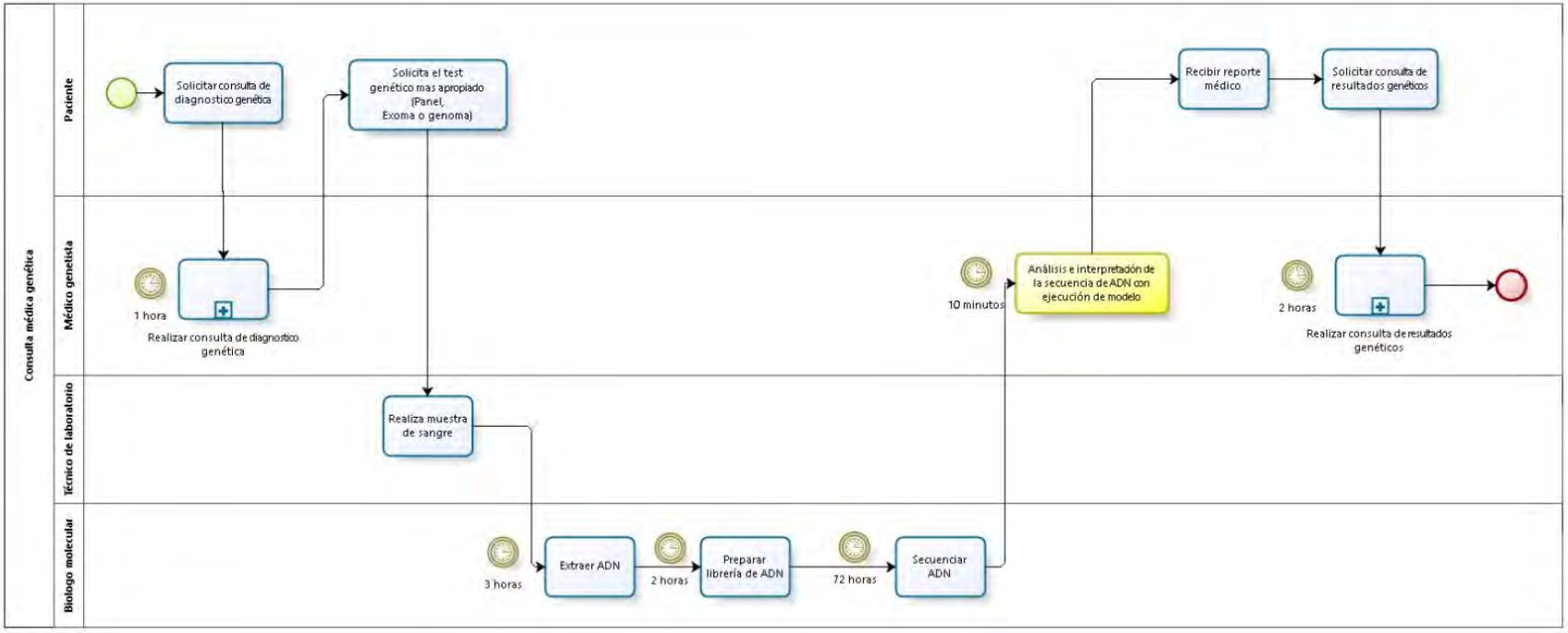
Análisis y Resultados

Analizar e interpretar la secuencia de ADN de una persona implica una parte de trabajo de laboratorio y otra, habitualmente más larga, de análisis de la información y filtrado de las variantes genéticas de potencial interés, que hay que integrar e interpretar en el contexto clínico del paciente, lo cual puede tomar días y hasta semanas. Sin embargo, algunos pacientes no pueden esperar semanas para recibir los resultados de la secuenciación de su genoma, debido a que conocer la patología de su enfermedad puede ser una cuestión de vida o muerte.

Lo que consigue el presente trabajo es reducir el trabajo de análisis e interpretación de secuencia de ADN de días o semanas a minutos.

Figura 47

Flujo TO BE consulta médica genética



Próximos Pasos

Al terminar el presente trabajo, hemos visto que estos deberían ser los próximos pasos para obtener un modelo maduro.

- **Adaptar el modelo para la lectura de archivos BAM,** los archivos de este formato son aquellos que se generan desde el secuenciador genético, al permitirle al modelo entender este tipo de archivos, estaríamos ahorrando el tiempo de conversión de archivo BAM a archivo CSV.
- **Actualizar el modelo según nuevos datos obtenidos por investigadores,** tenemos que continuar con la investigación, tomando siempre como insumo la nueva información que investigadores genéticos obtienen. Esto ayudará a que el modelo pueda seguir y seguir aprendiendo de nuevas variaciones vigentes.
- **Desplegar el modelo en plataformas tipo AWS o Microsoft Azure.** Si bien es cierto, hemos trabajo realizando la construcción del modelo en las plataformas Microsoft Azure y Google, se debe tener en cuenta que para obtener tiempos de respuesta aceptables es necesario realizar el despliegue en la nube debido a su gran capacidad de cómputo.

Resumen

En el presente capítulo se explica cómo se ha utilizado la metodología CRISP-DM como marco metodológico para el análisis y ejecución de la presente tesis, explicando los resultados de la investigación realizada, mostrando los valores obtenidos luego de las simulaciones del modelo y validando estos resultados en compañía de profesionales del sector para corroborar el éxito de estos.

Capítulo V: Conclusiones y Recomendaciones

Conclusiones

- La definición de los requerimientos funcionales fueron realizadas gracias a entrevistas que se tuvieron con un médico genetista con más de 15 años de experiencia, en éstas descubrimos la necesidad de contar con una herramienta de apoyo que permita una primera manera de detectar el cáncer mamá sin realizar el test genético de manera que esta información sea más accesible para los peruanos para que puedan descubrir la causa hereditaria del cáncer y predecir el riesgo personal y familiar de presentar cáncer. En estas entrevistas también descubrimos los factores de riesgo asociados al cáncer mamá, incluidas las variaciones genéticas, las cuales sirvieron para determinar los datos de entrada candidatos para el diseño del modelo predictivo.
- Los modelos basados en Deep Learning son tan buenos como los son sus datos de entrada, por lo que la calidad de los datos es un componente crítico de la construcción de modelos óptimos, la clave para la construcción de nuestro modelo fue determinar cómo se iban a obtener los datos y si éstos serían de la calidad correctos. La recopilación de estos datos fue la etapa que nos demandó más tiempo, para eso tuvimos que revisar diferente base de datos disponibles públicamente como son NCBI o BRCA Exchange. Posteriormente tuvimos que trabajar los datos en procesos de limpieza y transformación. Dentro del proceso de limpieza de datos eliminamos los valores atípicos y duplicados, filtramos datos inexactos o irrelevantes y corregimos valores faltantes. Luego, para el proceso de transformación usamos la normalización de datos para garantizar que todos los datos de entrada se encuentren dentro del rango entre 0 y 1, esto ayudó durante los siguientes pasos del modelado.

- Cuando el modelo empiece a funcionar, se debe considerar de la ley de protección de datos los consentimientos para el uso de los datos a nivel transfronterizo y el uso de datos sensibles.
- La evaluación del modelo es una fase crítica del ciclo de vida de Deep Learning porque la precisión de las predicciones una vez el modelo se despliegue en producción dependerán de esta fase. Hay varias métricas relacionadas a modelos de clasificación binaria, para nuestro modelo utilizamos las métricas de precisión y de recall, apoyándonos también en la matriz de confusión y la curva ROC. Nuestro modelo superó la precisión de 75% que se había definido en los objetivos especificó al lograr una precisión de 92%.

Recomendaciones

- El modelo predictivo puede probarse y evaluarse para validar su precisión con otro tipo de enfermedades que tengan como base una alteración genética, por ejemplo, otros canceres, epilepsia, defectos de nacimiento, etc.
- El modelo no busca reemplazar el trabajo de los investigadores, sino busca complementar su trabajo brindándoles una herramienta que les facilite la rápida detección de nuevas alteraciones y su tipificación entre benigno y patógeno.
- Se recomienda seguir entrenando el modelo de manera que logremos un rendimiento de por lo menos un 90% durante los siguientes 3 meses y no caer en problemas de sobreajuste(overfitting), el cual ocurre cuando el modelo se desempeña bien en los datos de entrenamiento, pero no generaliza bien en datos no vistos, esto lo podemos lograr afinando el modelo al modificar alguno de los hiperparámetros.
- El modelo está utilizando los datos recopilados desde las fuentes NCBI, BRCA Exchange y LOVD. Al momento de ir al ambiente de implementación se deben utilizar sólo los

valores de las secuencias genéticas y de los exones donde aparecen las variaciones a fin que el modelo pueda predecir de los resultados utilizando la información del secuenciador

Referencias

Álvarez Gama, D. (2016). *Análisis de Mutaciones en BRCA1 y BRCA2 Asociadas al Cancer de Mama.* Universidad de Cantabria - Facultad de Medicina. Santander: Universidad de Cantabria. Obtenido de https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjf1dC7xKn7AhW2LrkGHd_RAMcQFnoECA4QAQ&url=https%3A%2F%2Fdigital.csic.es%2Fbitstream%2F10261%2F164963%2F1%2FBRCAmuta.pdf&usg=AOvVaw00-xSY9IsnLytXoNco14KK

AM, T., Hang, C. D., M, T., SM, P., TA, T., ME, R., & M, K. (22 de 4 de 2021). *Inhibidores de la PARP para el cáncer de mama localmente avanzado o metastásico*. Obtenido de https://www.cochrane.org/es/evidence: https://www.cochrane.org/es/CD011395/BREASTCA_inhibidores-de-la-parp-para-el-cancer-de-mama-localmente-avanzado-o-metastasico

American Civil Liberties Union. (27 de Mayo de 2009). *Legal Challenge to Human Gene Patents*. Obtenido de https://www.aclu.org/other/brca-faqs#01: https://www.aclu.org/other/brca-faqs#01

Bitgenia. (19 de 12 de 2018). *Recomendaciones del ACMG/AMP para la clasificación de variantes*. Obtenido de https://www.diagnosticsnews.com: https://www.diagnosticsnews.com/noticias/31600-una-mirada-desde-la-bioquimica-molecular-recomendaciones-del-acmg-amp-para-la-clasificacion-de-variantes

Brazier, Y., & Rush, T. (21 de Junio de 2022). *What are the different types of tumor?* Obtenido de https://www.medicalnewstoday.com: https://www.medicalnewstoday.com/articles/249141

Buduma, N., Buduma, N., & Papa, J. (2022). *Fundamentals of Deep Learning Designing Next-Generation Machine Intelligence Algorithms.* O'Reilly Media, Inc. .

Chile BIO. (s.f.). *La Estructura del ADN, los genes y el código genético*. Obtenido de https://www.chilebio.cl: https://www.chilebio.cl/el-adn-los-genes-y-el-codigo-genetico/

Cleveland Clinic medical. (09 de Agosto de 2021). *What is cancer?* Obtenido de https://my.clevelandclinic.org/health/diseases/12194-cancer: https://my.clevelandclinic.org/health/diseases/12194-cancer

Devisetty, U. K. (2022). *Deep Learning for Genomics.* BIRMINGHAM: Packt Publishing Ltd.

DIVULGACIÓN MÉDICA. (23 de 7 de 2013). *Tipos de mutaciones*. Obtenido de https://metabolicas.sjdhospitalbarcelona.org: https://metabolicas.sjdhospitalbarcelona.org/noticia/tipos-mutaciones

El Naqa, I., & Murphy, M. (2022). *Machine and Deep Learning in Oncology, Medical Physics and Radiology.*

El Naqa, I., & Murphy, M. (2022). *Machine and Deep Learning in Oncology, Medical Physics and Radiology.* Springer Cham.

El Peruano. (05 de Febrero de 2022). *Más de 17,000 nuevos casos de cáncer se registraron en el 2021*. Obtenido de https://elperuano.pe: https://elperuano.pe/noticia/138721-mas-de-17000-nuevos-casos-de-cancer-se-registraron-en-el-2021#:~:text=04%2F02%2F2022%20El%20Instituto,son%20del%20interior%20del%20país

Gupta, M., Jain, R., Solanki, A., & Al-Turjman, F. (2022). *Cancer Prediction for Industrial IoT 4.0: A Machine Learning Perspective.* Londres: Chapman & Hall.

Instituto Nacional de Investigación del Genoma Hum. (4 de Noviembre de 2021). *¿Qué tipos de variantes de genes son posibles?* Obtenido de https://medlineplus.gov: https://medlineplus.gov/spanish/genetica/entender/variantesytrastornos/posiblesvariantes/

Instituto Nacional del Cancer. (10 de Noviembre de 2020). *Cáncer metastásico*. Obtenido de https://www.cancer.gov: https://www.cancer.gov/espanol/tipos/cancer-metastatico#:~:text=Las%20células%20cancerosas%20se%20diseminan%20por%20el%20cuerpo%20en%20varios,a%20otras%20partes%20del%20cuerpo.

Instituto Nacional del Cáncer de EE.UU. (5 de Mayo de 2021). *¿Qué es el cáncer?* Obtenido de https://www.cancer.gov/espanol/cancer/naturaleza/que-es: https://www.cancer.gov/espanol/cancer/naturaleza/que-es

Mayo Clinic. (17 de Septiembre de 2021). *Salud de la mujer*. Obtenido de https://www.mayoclinic.org: https://www.mayoclinic.org/es-es/healthy-lifestyle/womens-health/in-depth/breast-cancer-prevention/art-20044676

Mayo Clinic. (22 de Abril de 2022). *Breast cancer*. Obtenido de https://www.mayoclinic.org: https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470

Microsoft Ignite. (12 de Octubre de 2022). *Aprendizaje profundo frente a aprendizaje automático en Azure Machine Learning*. Obtenido de https://learn.microsoft.com: https://learn.microsoft.com/es-es/azure/machine-learning/concept-deep-learning-vs-machine-learning

Ministerio de Salud. (19 de Octubre de 2020). *https://www.gob.pe*. Obtenido de El cáncer de mama tiene un 90% de probabilidades de curación si se detecta a tiempo:

https://www.gob.pe/institucion/minsa/noticias/308976-el-cancer-de-mama-tiene-un-90-de-probabilidades-de-curacion-si-se-detecta-a-tiempo

National Human Genome Research Institute. (27 de Septiembre de 2022). *CÓDIGO GENÉTICO*. Obtenido de https://www.genome.gov: https://www.genome.gov/es/genetics-glossary/Codigo-genetico#:~:text=El%20código%20genético%20es%20el,y%20convertirlos%20en%20una%20proteína

Patel, A. (Julio de 2020). Benign vs Malignant Tumors. *JAMA Oncol*. doi:10.1001/jamaoncol.2020.2592

Puente , J., & de Velasco, G. (16 de Diciembre de 2019). *¿Qué es el cáncer y cómo se desarrolla?* Obtenido de https://seom.org/informacion-sobre-el-cancer/que-es-el-cancer-y-como-se-desarrolla: https://seom.org/informacion-sobre-el-cancer/que-es-el-cancer-y-como-se-desarrolla

República, C. d. (2011). *Ley N° 29733 Ley de Protección de Datos Personales.* Lima: Editora Perú. Obtenido de https://www.gob.pe/institucion/congreso-de-la-republica/normas-legales/243470-29733

Senturk, N., Tuncel, G., Dogan, B., Aliyeva, L., Dundar, M. S., Sag, S. O., . . . Dundar, M. (9 de Noviembre de 2021). BRCA Variations Risk Assessment in Breast Cancers Using Different Artificial Intelligence Models. *MDPI*. doi:https://doi.org/10.3390/genes12111774

Shaikh, K., & Krishnan, S. (2021). *Artificial Intelligence in Breast Cancer Early Detection and Diagnosis.* RESEARCHGATE. doi:10.1007/978-3-030-59208-0

Shaikh, K., Krishnan, S., & Thanki, R. (2021). *Artifcial Intelligence in Breast Cancer Early Detection and Diagnosis.* Junio: Springer. doi:10.1007/978-3-030-59208-0

T. A. (20 de Noviembre de 2020). *¿Qué es el cáncer?* (S. A. Cáncer, Ed.) doi:cancer.org | 1.800.227.2345

The American Cancer Society. (20 de Septiembre de 2019). *Tipos de cáncer de seno*. Obtenido de https://www.cancer.org: https://www.cancer.org/es/cancer/cancer-de-seno/acerca/tipos-de-cancer-de-seno.html#:~:text=Los%20tipos%20más%20comunes%20son,todos%20los%20cánceres%20de%20seno

The American Cancer Society medical. (1 de Febrero de 2020). American Cancer Society - Fever. *American Cancer Society*, 4. Obtenido de https://www.cancer.org/content/dam/CRC/PDF/Public/8893.00.pdf

The American Society Cancer. (20 de Septiembre de 2019). *Comprensión de un diagnóstico de cáncer de seno*. doi:cancer.org | 1.800.227.2345

Weinberg, R. A. (Septiembre de 1996). How Cancer Arises, An explosion of research is uncovering the long-hidden molecular underpinnings of cancer—and suggesting new therapies. *Scientific American, Inc.* doi:10.1038/scientificamerican0996-62

Yazici, H., Odemis, D. A., Aksu, D., Erdogan, O. S., Tuncer, S. B., Avsar, M., . . . Aydin, M. A. (2020). New Approach for Risk Estimation Algorithms of BRCA1/2 Negativeness Detection with Modelling Supervised Machine Learning Techniques. *HINDAWI*, 7. doi:8594090

Apéndice A: Carta de Presentación Solicitud de Validación de Experto

ANEXO A: VALIDÉZ DE LOS INSTRUMENTOS

CARTA DE PRESENTACIÓN

Dr. YASSER SULLCAHUAMAN ALLENDE

Fecha: 22 de febrero del 2023

Presente

Asunto: Validación de instrumentos a través de juicio de experto

Nos es muy grato comunicarme con usted para expresarle mi saludo y así mismo, hacer de su conocimiento que, siendo estudiante del Programa Académico de Maestría en Gerencia de Tecnología de la Información correspondiente a la escuela de negocios CENTRUM PUCP BUSINESS SCHOOL de la PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ (PUCP), en la sede Lima Este "Santiago de Surco", requiero validar el instrumento con el cual recogeré la información necesaria para poder desarrollar mi trabajo de investigación.

El título nombre del proyecto de investigación es: **Propuesta de un Modelo de Predicción de Cáncer de Mama Utilizando Deep Learning** y siendo imprescindible contar con la aprobación de docentes especializados para poder aplicar los instrumentos en mención, he considerado conveniente recurrir a usted, ante su connotada experiencia en temas educativos y/o investigación educativa.

Para la validación respectiva le compartimos el enlace del API con las instrucciones de uso.
Expresándole mis sentimientos de respeto y consideración me despido de usted, no sin antes agradecerle por la atención que dispense a la presente.

Atentamente

Firma:

Nombre completo y DNI:

Jorge Antonio Páez Cumpa
DNI 43461755

Henry Edward Palomino Delgado
DNI 71479229

Christian Rosado Farfán
DNI 43071779

Elmer Ronald Salazar Huamanjulca
DNI 45352290