Predicción de supervivencia en pacientes con cáncer de mama utilizando modelos de clasificación

(Reto Kaggle)

Katherine Restrepo Gutiérrez

Monografía para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor

Dr. Sergio Sanes Negrete

Universidad de Antioquia

Facultad de ingeniería, Departamento de ingeniería de sistemas

Especialización en Analítica y Ciencia de Datos

Medellín, Antioquia Colombia

2022

Cita	(Restrepo Gutiérrez, 2022)
Referencia **Estilo APA 7 (2020)**	Restrepo Gutiérrez, K., (2022). *supervivencia del cáncer de mama utilizando modelos de clasificación –Reto Kaggle.* [Trabajo de grado especialización]. Universidad de Antioquia, Medellín.

Especialización en Analítica y Ciencia de Datos, Cohorte III.

UNIVERSIDAD DE ANTIOQUIA
Vicerrectoría de Docencia

Sistema de Bibliotecas

Centro de Documentación Ingeniería (CENDOI)

Repositorio Institucional: http://bibliotecadigital.udea.edu.co

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes
Decano/director: Jesús Francisco Vargas Bonilla
Jefe departamento: Diego José Luis Botia Valderrama

Contenido

1. Resumen ejecutivo

El objetivo de este trabajo es predecir la supervivencia del cáncer de mama aplicando modelos de clasificación, para el desarrollo de este objetivo se utiliza la base de datos del Consorcio Internacional de Taxonomía Molecular del Cáncer de Mama (METABRIC), que contiene datos de 1.904 pacientes con atributos clínicos y 331 genes con niveles de puntuación z de ARNm y mutación de 175 genes, la base de datos está disponible en la plataforma de Kaggle publicado por RAGHAD ALHARBI. Para este trabajo se modela la supervivencia del cáncer de mama como un problema de clasificación binaria utilizando modelos de aprendizaje estadístico supervisado como es la regresión logística, árbol de clasificación, random forest y Gradient boosting. En un primer experimento, se utilizaron únicamente datos clínicos como variables explicativas. Como primer resultado se obtiene que la regresión logística es el mejor modelo.

En un segundo experimento, se realiza una modelación incluyendo datos clínicos y parte de las variables de expresión genética, lo que aumenta la dimensión de variables explicativas a 6.271, debido a esto, se aplica una técnica de reducción de dimensionalidad por análisis de componentes principales. El mejor modelo nuevamente es la regresión logística, pero el resultado se encuentra por debajo del primer modelo.

Finalmente se realiza un tercer experimento o iteración que busca mejorar el resultado de la segunda modelación, en esta última se incluyen las mismas variables clínicas y algunas variables genéticas que por estudios las han clasificado como principales factores de riesgo en el desarrollo de este tipo de cáncer. El mejor modelo continúa siendo la regresión logística y el resultado mejora respecto a los modelos anteriores.

Como conclusión, se puede evidenciar que el mejor desempeño se logra en el tercer experimento con el modelo de R*egresión Logística*, Accuracy de 84%, f1-score de 82% cuando se predice la muerte y f1-score del 85% en la predicción de la supervivencia del paciente.

Palabras Claves: Aprendizaje Estadístico, Supervivencia, Regresión, Predicción, Clasificación

2. Descripción del problema

El cáncer de mama es una de las principales enfermedades a nivel mundial, es una enfermedad que, si bien puede aparecer en edades tempranas, los riesgos de padecerla son más altos después de alcanzar una edad mayor de 40 años, el cáncer afecta en mayor porcentaje a las mujeres, se dice que en hombres ocurre aproximadamente en el 1% de esta población.

Como lo define la organización mundial de la salud el cáncer es una proliferación anormal y desordenada de células lo que indica que, la enfermedad está relacionada por alteraciones genéticas solo que algunos casos son de tipo hereditario y otros no hereditarios, pero que en ambas situaciones es producto de la alteración del ADN lo que lleva a que alguna célula mute y se desarrolle el cáncer, una vez esto ocurre tiende a multiplicarse y disiparse por todo el tejido mamario y en muchas ocasiones a otros órganos(metástasis). Solo en el 2020 se diagnosticaron a nivel mundial 2.3 millones de personas de las cuales el 30% fallecieron, el cáncer de mama es considerada la enfermedad más prevalente a diferencia de otro tipo de cáncer. (Organización Mundial de la Salud, 2021)

En Colombia actualmente el cáncer de mama se encuentra entre las enfermedades priorizadas, siendo esta la que más muertes registra, también se ha identificado que tan sólo el 7% de la población reportada se encontraba en etapa temprana (in situ) un resultado que refleja la falta de prevención y oportunidad en la atención de los pacientes. (Organización Mundial de la Salud, 2021)

En Colombia, durante el periodo comprendido entre el 2 de enero de 2019 y el 1° de enero de 2020, el cáncer de mama fue el más frecuente entre los casos nuevos de los 11 tipos de cáncer priorizados por el Ministerio de Salud y Protección Social. También fue el responsable de la mayor cantidad de muertes notificadas a la cuenta de alto costo (CAC) en el mismo período y continúa siendo el más común en las mujeres con el 27,99% del total de casos nuevos. (Fondo Colombiano de Enfermedades de Alto costo, 2021)

Se sabe que un diagnóstico temprano aumenta la probabilidad de supervivencia y evita los tratamientos prolongados e invasivos, debido a lo anterior el acceso oportuno al tratamiento es fundamental para mejorar el pronóstico de la enfermedad y la calidad de vida de las pacientes, según datos del Fondo Colombiano de enfermedades de alto costo, a nivel nacional, la mediana de días de espera para la confirmación del diagnóstico de cáncer de mama es de 36 días y hasta el primer tratamiento es de 60 días. Lo anterior indica que como país debemos continuar trabajando en el acceso al tratamiento, logrando que se inicie antes de los 30 días. (Fondo Colombiano de Enfermedades de Alto costo, 2021)

El tratamiento del cáncer suele ser eficaz si se detecta en etapas tempranas, el tratamiento consiste en una combinación de extirpación quirúrgica, radioterapia y medicación (terapia hormonal, quimioterapia) por consiguiente un tratamiento a tiempo puede mejorar la calidad de vida, evitar la progresión del cáncer o erradicarla por completo.

2.1 Problema de negocio

En Colombia las EPS[1] y las IPS[2] juegan un papel fundamental en el tratamiento oportuno de los pacientes que han sido diagnosticados tanto en etapas tempranas como avanzadas de la enfermedad y es aquí donde como sistema de salud aún no hay grandes avances en tecnología que permita hacer uso de la información histórica y actual del paciente, sin embargo, una buena iniciativa nace en el 2019 cuando se crea el sistema de información SISCAC[3], un desarrollo creado por la cuenta de alto costo y el ministerio de salud de Colombia donde se aplican técnicas de inteligencia artificial, que permite recopilar la información del paciente de diferentes fuentes y devolver al prestador y demás entidades la información más completa de estado actual del paciente y de esta manera poder llevar a cabo seguimientos en tiempo real y lo más importante realizar una gestión del riesgo, sin embargo su enfoque principal ha estado obstaculizado, y esto se debe a que deben intervenir diferentes actores en el proceso lo que hace que no se tenga información completa y a tiempo que permita tomar decisiones.

De acuerdo con lo anterior desde las EPS o IPS se debe empezar a desarrollar una cultura y estrategia analítica donde se dé valor a los datos. El problema general radica en que como entidades prestadoras de servicios de salud no se está aportando información suficiente al paciente, información que es determinante en el estado de salud e incluso en su supervivencia. En estas entidades se tiene acceso al historial médico del paciente tal como: la historia clínica, resultados del diagnóstico, resultados de los exámenes de laboratorio, procedimientos quirúrgicos, tratamientos aplicados entre otras atenciones, sin embargo, aun teniendo esta información no se ha logrado extraer su potencial valor, y esto se debe principalmente por la falta de calidad de los datos donde simplemente se han convertido en recolectores de datos sin que nadie intervenga y haga uso correcto de estos, como instituciones de salud la práctica del análisis predictivo aún no está muy desarrollada en las áreas de la salud en Colombia, especialmente en el campo de la oncología donde hasta hace relativamente poco se están haciendo acercamientos en la construcción de modelos que permitan identificar factores de riesgo futuros en pacientes ya dados de alta, predicciones que permiten identificar si un paciente es propenso a desarrollar metástasis, modelos de clasificación de patologías, análisis predictivos que permitan determinar un mejor curso del tratamiento entre otros.

[1] EPS: Entidad promotora de salud

[2] IPS: Institución prestadora de servicios de salud

[3] SISCAC: sistema de información en salud dispuesta por la cuenta de alto costo

2.2 Aproximación desde la analítica de datos

La parte más importante en el proceso de toma de decisiones clínicas en pacientes con cáncer, es lograr la estimación más precisa del estado de salud y el pronóstico de supervivencia. El uso de técnicas de aprendizaje automático utilizando información clínica y principalmente genética tiene el potencial de proporcionar una estimación con más certeza del tiempo de supervivencia, evitar procedimientos quirúrgicos y tratamientos innecesarios como también el optimizar recursos que son de alto costo y garantizar una mejor calidad de vida en los pacientes; además en Colombia pocas instituciones están aplicando técnicas de analítica avanzada en los procesos de diagnóstico y tratamiento del cáncer, se hace analítica pero más desde el punto de vista descriptivo.

Lo que se busca con este trabajo es empezar a hacer un ejercicio académico que permita tener un primer acercamiento de la estimación del cálculo de la supervivencia usando algoritmos de aprendizaje automático de clasificación y de esta manera poder adquirir conocimiento analítico donde se empiezan a desarrollar y aplicar modelos en las IPS.

Los códigos con los cuales se realizan los análisis y se estiman resultados en este trabajo se encuentran disponibles en un repositorio[4], en el anexo 1 se describe el contenido de cada uno de los archivos allí almacenados.

2.3 Origen de los datos

La base de datos utilizada es del Consorcio Internacional de Taxonomía Molecular del Cáncer de Mama (METABRIC), es un proyecto entre Canadá y el Reino Unido que contiene datos de secuenciación específica de 1.980 muestras primarias de cáncer de mama. Se encuentra disponible en la plataforma de Kaggle llamada perfiles de expresión genética del cáncer de mama (METABRIC), este dataset contiene en total de 1904 registros y 693 variables.

2.4 Métricas de desempeño

Las métricas tradicionales usadas para evaluar el rendimiento de modelos de clasificación binaria de aprendizaje automático son el Accuracy, *precisión*, *Recall (TPR), true negative rate (TNR)*, *F1-score*.

De acuerdo con el problema de clasificación se expone que el Accuracy se refiere a la proporción de resultados verdaderos (tanto verdaderos positivos, como falsos positivos) dividido entre el número total de predicciones realizadas, la precisión es la proporción de verdaderos positivos dividido entre todos los resultados positivos (tanto verdaderos positivos, como falsos positivos), el Recall es la proporción de verdaderos positivos que son

[4] El repositorio con el código y la base de datos se encuentran en: https://github.com/katherine2022Udea/TesisEspecializacion

correctamente observados por el clasificador y la especificidad está dada por la proporción de verdaderos negativos que son correctamente identificados (Barrios Arce, 2019)

3. Datos

3.1 Datos originales

Para predecir la supervivencia del cáncer de mama se utiliza la base de datos del Consorcio Internacional de Taxonomía Molecular del Cáncer de Mama (METABRIC) que contiene datos de 1.904 pacientes con sus atributos clínicos y 331 genes con niveles de puntuación z[5] de ARNm[6] y 175 genes mutados (Mukherjee, A., Russell, R., Chin, SF, 2018). La base de datos se descargó desde la plataforma de Kaggle publicada por RAGHAD ALHARBI.[7]

Los datos usados contienen la información de 1.904 pacientes con 693 variables entre atributos clínicos y variables genéticas. Entre los atributos clínicos se tienen 35 variables descritas en la Tabla 1.

Tabla 1. Diccionario de variables

Variable	Descripción
patient_id	ID del paciente
ageatdiagnosis	Edad de la paciente al momento del diagnóstico.
typeofbreast_surgery	Tipo de cirugía de cáncer de mama: 1- MASTECTOMÍA, que se refiere a una cirugía para extirpar todo el tejido mamario de una mama como una forma de tratar o prevenir el cáncer de mama. 2- CONSERVADORA DE MAMA, que se refiere a una urgencia donde solo la parte de la mama que tiene cáncer se remueve
cancer_type	tipos de cáncer de mama: 1- Cáncer de mama o 2- Sarcoma de mama
cancertypedetailed	Tipos detallados de cáncer de mama: 1- Carcinoma ductal invasivo de mama 2- Carcinoma ductal y lobulillar mixto de mama 3- Carcinoma lobular invasivo de mama 4- Carcinoma mucinoso mixto invasivo de mama 5- Cáncer de mama metaplásico

[5] Puntuación z indica a cuántas desviaciones estándar por encima o por debajo de la media se ubica un valor

[6] El ARNm es un tipo de ARN (ácido ribonucleico) que tiene información genética que se necesita para elaborar las proteínas, esta información viaja desde el ADN hasta el citoplasma donde allí se elaboran

[7] Reto Kaggle: https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric

Variable	Descripción
cellularity	Celularidad del cáncer después de la quimioterapia, que se refiere a la cantidad de células tumorales en la muestra y su disposición en grupos.
chemotherapy	El paciente recibió quimioterapia como tratamiento (sí / no)
pam50+claudin-low_subtype	Es una prueba de perfil tumoral que ayuda a mostrar si es probable que algunos cánceres de mama con receptores de estrógeno positivos (ER positivos) y HER2 negativos hagan metástasis (cuando el cáncer de mama se disemina a otros órganos). El subtipo de cáncer de mama con claudina baja se define por las características de expresión génica, principalmente: baja expresión de genes de adhesión célula-célula, alta expresión de genes de transición epitelial-mesenquimatosa (EMT) y patrones de expresión génica similares a células madre / menos diferenciados
cohort	La cohorte es un grupo de sujetos que comparten una característica definitoria (toma un valor de 1 a 5)
erstatusmeasuredbyihc	Para evaluar si los receptores de estrógeno se expresan en las células cancerosas mediante el uso de inmunohistoquímica (un tinte utilizado en patología que se dirige a un antígeno específico, si está allí, dará un color, no está allí, el tejido del portaobjetos se coloreará) (positivo negativo)
user_status	Las células cancerosas son positivas o negativas para los receptores de estrógeno
neoplasmhistologicgrade	Determinado por patología al observar la naturaleza de las células, se ven agresivas o no (toma un valor de 1 a 3)
her2statusmeasuredbysnp6	Evaluar si el cáncer es positivo para HER2 o no mediante el uso de técnicas moleculares avanzadas (tipo de secuenciación de próxima generación)
her2_status	Si el cáncer es positivo o negativo para HER2
tumorotherhistologic_subtype	Tipo de cáncer basado en el examen microscópico del tejido canceroso (toma un valor de 'Ductal / NST', 'Mixto', 'Lobular', 'Tubular / cribiforme', 'Mucinoso', 'Medular', 'Otro', 'Metaplástico')
hormone_therapy	Si la paciente tuvo tratamiento hormonal o no (sí / no)
inferredmenopausalstate	Si la paciente es posmenopáusica o no (post / pre)
integrative_cluster	Subtipo molecular del cáncer basado en alguna expresión genética (toma un valor de '4ER +', '3', '9', '7', '4ER-', '5', '8', '10', ' 1 ',' 2 ',' 6 '
primarytumorlaterality	Ya sea que involucre el seno derecho o el izquierdo
lymphnodesexamined_positive	Para tomar muestras del ganglio linfático durante la cirugía y ver si estaba involucrado por el cáncer

Variable	Descripción
mutation_count	Número de gen que tiene mutaciones relevantes
nottinghamprognosticindex	Se utiliza para determinar el pronóstico después de una cirugía por cáncer de mama. Su valor se calcula utilizando tres criterios patológicos: el tamaño del tumor; el número de ganglios linfáticos afectados; y el grado del tumor
oncotree_code	OncoTree es una ontología de código abierto que se desarrolló en el Memorial Sloan Kettering Cáncer Center (MSK) para estandarizar el diagnóstico de tipos de cáncer desde una perspectiva clínica al asignar a cada diagnóstico un código OncoTree único.
overallsurvivalmonths	Duración desde el momento de la intervención hasta la muerte
overall_survival	Variable objetivo si el paciente está vivo o muerto
pr_status	Las células cancerosas son positivas o negativas para los receptores de progesterona
radio_therapy	Si el paciente recibió radio como tratamiento (sí / no)
3-geneclassifiersubtype	Subtipo de clasificador de tres genes Toma un valor de 'ER- / HER2-', 'ER + / HER2- High Prolif', nan, 'ER + / HER2- Low Prolif', 'HER2+'
tumor_size	Tamaño del tumor medido por técnicas de imagen.
Etapa del cáncer	Etapa del cáncer según la participación de las estructuras circundantes, los ganglios linfáticos y la diseminación a distancia
deathfromcancer	si la muerte del paciente se debió a cáncer o no (sí / no)

3.2 Dataset

La base de datos para realizar la modelación se obtuvo de 1.904 pacientes, está compuesta de 693 variables: 35 atributos clínicos, 331 genes y 175 genes mutados. Dentro de las variables se tiene ***overall_survival*** que describe si el paciente está vivo o fallecido, por lo tanto, será la variable objetivo a modelar y es de naturaleza categórica y binaria.

Lo primero que se hace es analizar la variable ***death_from_cancer***, para identificar si la paciente murió por causa del cáncer o por otras causas, en este trabajo sólo nos interesa analizar los pacientes que fallecen y sobreviven por causa del cáncer.

Tabla 2. Distribución proporcional de muertes por cáncer

Death_from_cancer	N°	%
Living	801	42,1%
Died of disease	622	32,7%
Died of other causes	480	25,2%
Total	**1.903**	**100%**

En la Tabla 2 se observa que 42,1% de las pacientes sobreviven y el 57,9 % mueren, pero debemos excluir las muertes derivadas por otras causas diferentes al cáncer. Por tanto, la muestra que se usa para el análisis es la representada en la Tabla 3.

Tabla 3. Distribución proporcional variable objetivo

Death_from_cancer	N°	%
Living	801	56,3%
Died of disease	622	43,7%
Total	**1.423**	**100%**

Una vez se define la población estudio, se identifican los datos nulos de la base de datos, en la tabla 4 se observa el resultado.

Tabla 4. Proporción de valores nulos por variable

Variable	Registros	Tipo variable	% Nulos
tumor_stage	343	float64	24,1%
3-gene_classifier_subtype	149	object	10,5%
primary_tumor_laterality	80	object	5,6%
cellularity	42	object	2,9%
mutation_count	39	float64	2,7%
neoplasm_histologic_grade	38	float64	2,6%
er_status_measured_by_ihc	20	object	1,4%
type_of_breast_surgery	16	object	1,2%
tumor_size	12	float64	0,8%
cancer_type_detailed	10	object	0,7%
tumor_other_histologic_subtype	10	object	0,7%
oncotree_code	10	object	0,7%
death_from_cancer	1	object	0,1%

De acuerdo con la Tabla 4, vemos la proporción de datos nulos de la cual se decide realizar una imputación para todas las variables excepto para *death_from_cancer* que presenta el 0.1% de datos nulos, se toma la decisión de eliminar el registro *tumor_stage* de la base de datos por el porcentaje alto de datos nulos que contiene de 24,1%, para la imputación de las

variables numéricas se aplicó imputación por mediana debido a la robustez que presenta ante valores atípicos y para las variables categóricas se usa la moda. Se realiza un procedimiento de recodificación de variables categóricas y numéricas. Posteriormente se divide la base de datos en dos conjuntos uno de entrenamiento y otro de prueba.

3.3 Descriptiva

Antes de iniciar la etapa de entrenamiento de modelos de clasificación, para predecir y evaluar el rendimiento de estos con respecto a la supervivencia del cáncer, es necesario realizar un análisis exploratorio y descriptivo de los datos, que permita comprender los datos que se utilizaran en el entrenamiento de los modelos, mediante técnicas simples de resumen de datos, estadística descriptiva y pruebas de hipótesis.

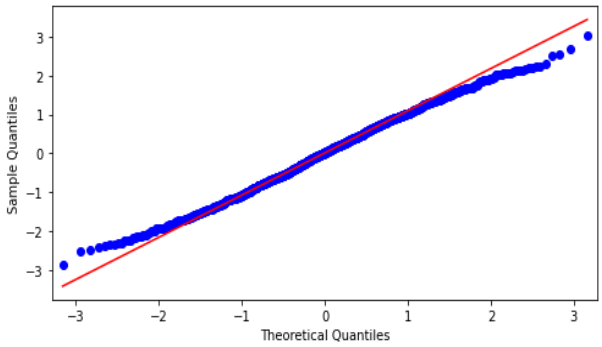


Gráfico 1. Distribución edad del paciente

Como parte inicial del análisis se busca saber si los datos siguen una distribución normal, para esto se utilizó la variable edad, al realizar el gráfico Q-Q[8] parecía indicar que sí, sin embargo, se realizan algunas pruebas de hipótesis que permite tener más certeza de que exista normalidad, para esto se aplicó la prueba de Shapiro-Wilk y según lo que indica la interpretación de la prueba nos muestra que no existe normalidad en la variable edad al ser el valor p menor a 0,05, sin embargo, como en ocasiones se sugiere no sólo aplicar un test, es por esta razón que se realiza el Test K^2 de D'Agostino que nos arroja nuevamente como resultado que no existe normalidad en los datos.

[8] Los gráficos Q-Q (cuantil-cuantil) comparan dos distribuciones de probabilidad mediante el trazado de sus cuantiles uno contra el otro.

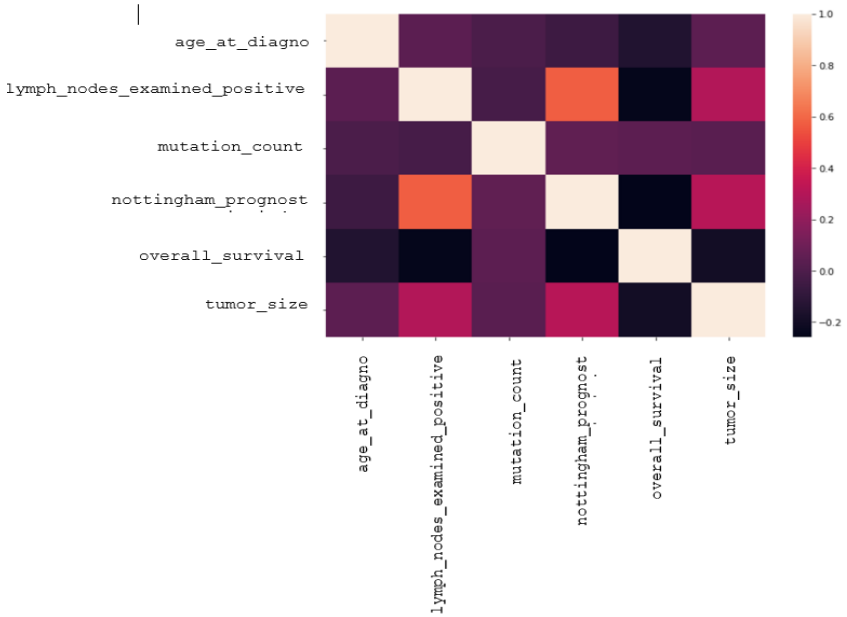


Gráfico 2. Matrix de correlación

Al analizar la matriz de correlación se evidencia que no existe casi ninguna relación entre las variables, sin embargo, entre las variables **nottingham_prognostic_index** y **lymph_nodes_examined_positive** existe una relación positiva, aunque no es una relación fuerte.

3.3.1 Impacto de los tratamientos y el uso de la terapia hormonal sobre la supervivencia del cáncer

Como primer acercamiento a los datos relacionados con la supervivencia, se analiza el efecto que tiene los diferentes tratamientos en la prolongación de la vida de los pacientes, lo cual está dado por medio de la variable ***overallsurvivalmonths*** que representa la duración o el tiempo que ha transcurrido desde el momento de la intervención hasta la muerte.

A continuación, en la Tabla 5 se muestra la relación entre la supervivencia y el tratamiento con quimioterapia, vemos que los pacientes que sobreviven y les aplicaron este tratamiento son el 10% de la muestra, mientras que los pacientes que sobreviven y no les aplicaron quimioterapia es del 32% y finalmente los pacientes que si se les aplicó quimioterapia y no sobrevivieron fue del 11%.

Tabla 5. Relación de supervivencia y tratamiento quimioterapia

Quimioterapia	Supervivencia					
	Frecuencia	No	Si	%	No	Si
	No	891	617	NO	47%	32%
	Si	212	184	SI	11%	10%

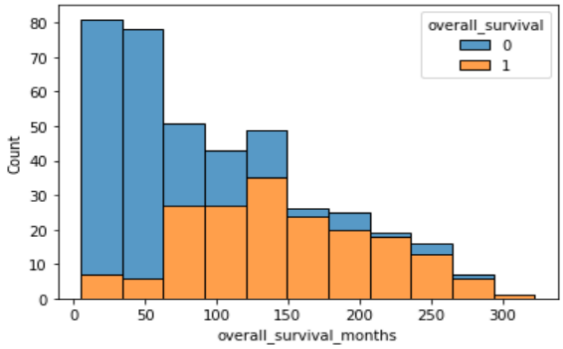
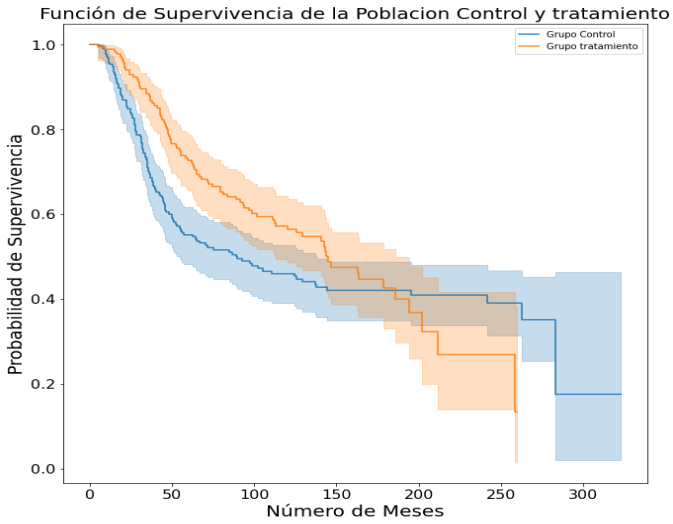


Gráfico 3. Tiempo de supervivencia en paciente con tratamiento de quimioterapia

Si evaluamos la independencia de las dos variables (quimioterapia y meses de supervivencia) a través de una prueba chi2 de tablas de contingencia, encontramos que el p valor es 0.053 por lo tanto no hay suficiente evidencia para rechazar la hipótesis nula: las variables Quimioterapia y Supervivencia son independientes. Parece indicar que la supervivencia es irrelevante si se aplica el tratamiento de quimioterapia en un paciente, esto es un poco contradictorio, ya que la quimioterapia ha demostrado ser efectiva en el tratamiento del cáncer cuando se detecta en etapas tempranas, este resultado podría presentarse, ya que no se está teniendo en cuenta la etapa del cáncer en la que se encuentra en paciente al momento de iniciar la quimioterapia, de esta manera entonces, se seleccionan los pacientes a los que se les aplicó quimioterapia independiente de la etapa en la que se encuentre. Para analizar la supervivencia se utiliza la función Kaplan-Meier muy utilizada para estimar la supervivencia.

En la Gráfica 4 se observa la función de Kaplan-Meier, vemos que la probabilidad de supervivencia se reduce considerablemente cuando han transcurrido 150 meses en promedio desde la intervención (línea azul).



Grafica 4. Supervivencia en pacientes con quimioterapia y terapia hormonal

Cuando se compara con el impacto que puede tener el tratamiento hormonal en los pacientes que se les aplicó quimioterapia, se evidencia como la estimación de la curva de los pacientes que se les aplicó tratamiento hormonal (línea naranja) se expande más a la derecha el tiempo de supervivencia, la mediana para el grupo de control es de 88 meses mientras que la mediana del tiempo de supervivencia es de 143 meses para el grupo tratamiento. Parece indicar que el tratamiento hormonal aplicado a los pacientes que recibieron quimioterapia prolonga su vida.

Para validar la premisa anterior, se realiza la prueba LogranK, donde se plantea la siguiente hipótesis:

H0: *No hay diferencias en las dos poblaciones incluyendo el tratamiento hormonal*
Ha: *Hay diferencias en las dos poblaciones incluyendo el tratamiento hormonal*

Con un **α = 0.05** se aplica la prueba brindando como resultado un p valor de 0.08. Se puede concluir que las probabilidades de supervivencia de los dos grupos de las muestras de pacientes que se les aplicó quimioterapia son diferentes.

Tabla 6. Relación de supervivencia y tratamiento de radioterapia

Radioterapia	Supervivencia					
	Frecuencia	No	Si	%	No	Si
	No	496	271	No	26%	14%
	Si	607	530	Si	32%	28%

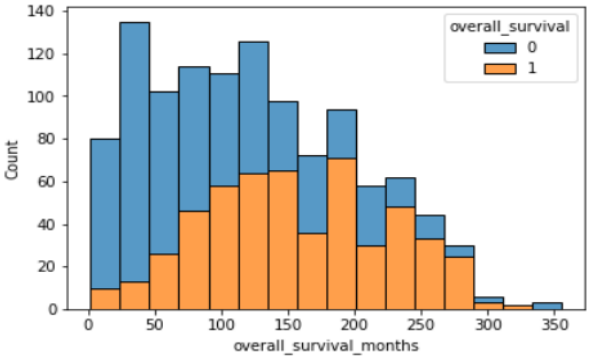
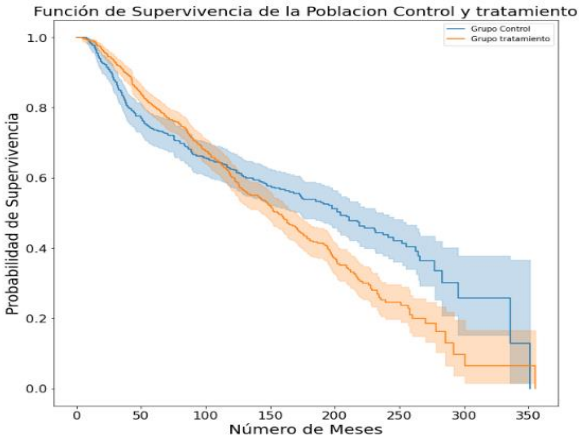


Gráfico 5. Tiempo de supervivencia en paciente con tratamiento de quimioterapia

Como se observa en la tabla 6 se analiza la relación entre la supervivencia y el tratamiento con radioterapia, los pacientes que sobreviven y recibieron el tratamiento es del 28%, mientras que los que sobrevivieron y no se les aplicó radioterapia fue del 14%, para los que sobrevivieron y no recibieron fue del 32%.



Grafica 6. Supervivencia en pacientes con tratamiento de radioterapia y terapia hormonal

En la Gráfica 6 se compara los pacientes que recibieron radioterapia (línea azul) y los que recibieron radioterapia y tratamiento hormonal (línea naranja), se observa cómo se expande a la derecha ligeramente el grupo de tratamiento en los primeros meses, sin embargo, este disminuye considerablemente después del mes 125 con una mediana para el grupo de control de 204 meses, mientras que la mediana del tiempo de supervivencia es de 143 meses para el grupo tratamiento. Parece indicar que el tratamiento hormonal aplicado a los pacientes que recibieron sólo radioterapia no mejora la probabilidad de supervivencia.

Para validar la premisa anterior, se realiza la prueba LogranK, planteando la misma prueba de hipótesis anterior, con un **α = 0.05** muestra como resultado un p valor de 0.01, se puede concluir que las probabilidades de supervivencia de los dos grupos de las muestras de pacientes que se les aplicó radioterapia son iguales.

Distribución de la Variable Supervivencia

Ahora se analiza la distribución de la variable objetivo **overall_survival**, para revisar si existe desbalanceo de clases en los datos, debido a que los modelos de clasificación binaria por lo general en una clase mayoritaria provocan desbalanceo en los datos, en los casos relacionados con la salud, por ejemplo en el cáncer de mama, la población mayoritaria serían las mujeres que no padecen de cáncer, si tenemos demasiadas muestras de pacientes que sobreviven al cáncer y pocas que fallecen, el modelo va a sesgar las predicciones hacia la clase mayoritaria mostrando un buen desempeño, mientras que en la clase minoritaria el resultado sería una baja predicción.

La distribución de la variable objetivo, para la clase 1 (supervivencia) el total de muestras es de 56.29% (801) mientras que para la clase 0 (muerte) tiene un 43.71% (622) vemos que la diferencia es de 12.50 puntos, por lo tanto, el desbalanceo no es tan grande como puede suceder en otros casos. Para efectos de este trabajo no se realiza ningún método de balanceo de datos.

4. Proceso de analítica

4.1 Pipeline principal

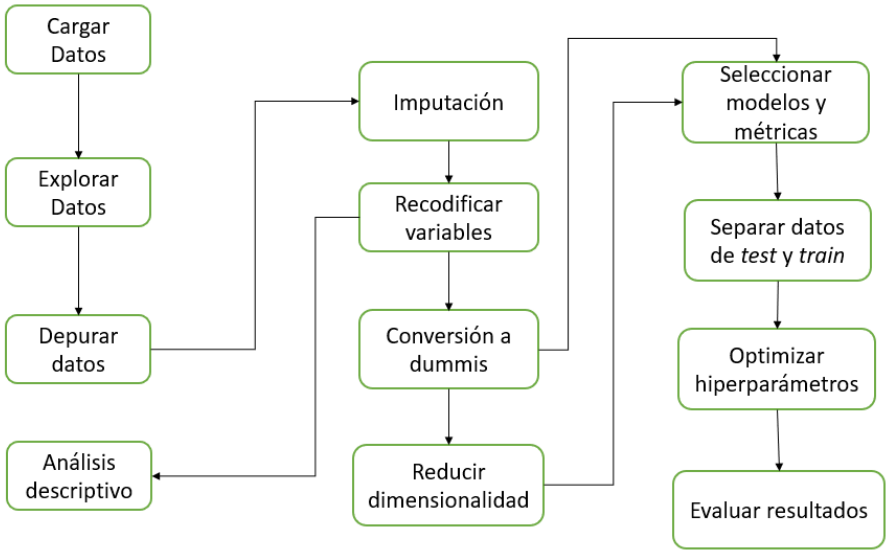


Gráfico 7. Flujo de trabajo general de los datos

En la Gráfica 7 se describe las fases o pasos utilizados para el procesamiento y resultado final de los modelos, en primer lugar se inicia con la carga de los datos en Google Colab[9], lugar donde se lleva a cabo todo el procesamiento, posterior a esto se realiza una exploración donde la finalidad era conocer cada una de las variables que hacen parte del dataset, como parte importante del proceso también es identificar los valores nulos y cuál es el porcentaje en cada variable de datos faltantes, con el fin de determinar si es viable incluir o no en el modelo, para este ejercicio es necesario aplicar técnicas como la imputación para tratar los datos faltantes. Posterior a esto, se codifican nuevamente las variables categóricas y numéricas. Con esta primera parte procesada se crea un análisis descriptivo tomando algunas variables que permita encontrar información significativa al problema de interés, al igual que determinar si existe normalidad en el conjunto de los datos, una vez se conoce el estado de los datos se convierten las variables en *"dummies",* es necesario aplicar reducción de dimensionalidad para el segundo análisis donde se incluyen variables genéticas que no han sido consideradas en el análisis inicial, en ambas iteraciones se utilizaron los mismo modelos de clasificación (árbol de decisión, random forest, boosting y regresión logística), se dividieron los datos en entrenamiento y test, se aplica validación cruzada en cada modelo y finalmente se evalúa los resultados.

[9] Permite programar y ejecutar código Python desde el navegador
https://colab.research.google.com/?hl=es#scrollTo=5fCEDCU_qrC0

4.2 Preprocesamiento

Antes de iniciar la modelación, se hace un procesamiento de las variables explicativas tipo factor, se transforma en variables dummies en k-1 dimensión cada variable categórica con atributos clínicos y variables genéticas. Para la modelación se plantean tres escenarios:

- **Experimento I: Modelación únicamente con atributos clínicos**

Se utiliza del conjunto de datos sólo las variables predictoras con atributos clínicos, se transforma de 30 variables a 65 variables.

- **Experimento II: Modelación con atributos clínicos y todas las variables genéticas**

Se analiza el conjunto de datos de las variables predictoras con los atributos clínicos y las variables genéticas, se transforma de 495 variables a 6.271 variables. Debido a la alta dimensionalidad de esta base de datos se aplica análisis de componentes principales (PCA).

- **Experimento III: Modelación con atributos clínicos y algunas variables genéticas consultadas en la literatura relacionadas con el desarrollo del cáncer de mama**

Esta tercera modelación se crea con el fin de mejorar el resultado de la segunda modelación, ya que al incluir todas las variables genéticas el resultado del modelo no mejoró, incluso estuvo por debajo del primero modelo. Las variables genéticas incluidas en esta modelación se seleccionaron a partir de la literatura que las han relacionado como los principales factores de riesgo en el desarrollo de este tipo de cáncer, unos de tipo hereditario y otros no hereditario. Los genes considerados de alta influencia son: BRCA1, BRCA2, TP53, PTEN, SKT11, CDH1 y MMR. Los denominados de influencia intermedia son: CHEK2, ATM, PALB2, BRIP1, RAD51C, RAD51D, BARD1, MRE11, RAD50, NBS1 y FANCM, en muchos casos responsables en parte del cáncer mamario familiar.[10] (Miguel-Soca, 2016)

De estos 19 genes mencionados se seleccionaron 13 variables, ya que las otras no se encontraban disponibles en la base de datos. Las variables seleccionadas fueron: BRCA1, BRCA2, TP53, PTEN, SKT11, CDH1, RAD51C, RAD51D, BARD1, CHEK2, ATM, PALB2, RAD50.

4.3 Modelos

Para el problema planteado inicialmente se utilizaron los algoritmos de clasificación como random forest, árbol de clasificación, boosting y regresión logística.

[10] Diccionario del instituto nacional de cáncer https://www.cancer.gov/espanol/publicaciones/diccionarios/diccionario-cancer

4.3.1 Árbol de clasificación

Los árboles de decisión son uno de los algoritmos más utilizados en machine learning gracias a su representación gráfica y sus reglas binarias (si/no) lo hace intuitivo y fácil de entender, no requiere de mucha limpieza y preprocesamiento y los *outliers no influyen mucho en su resultado*, con el tiempo se ha convertido en un referente para solucionar diferente tipo de problemas, tanto que hoy en día es una herramienta utilizada en el diagnóstico médico.

Amat Rodrigo (2020) expone que la desventaja de aplicar este método puede verse reflejada a la hora de tener una mejor capacidad predictiva, ya que puede existir *overfitting* y alta varianza, los afecta los datos desbalanceados. Para evitar que el modelo tuviera baja capacidad predictiva se utilizó una combinación de múltiples árboles como lo es random forest y boosting.

4.3.2 Random forest

Este modelo está basado en el árbol de decisión, tal y como lo menciona Amat Rodrigo (2020) está formado por un conjunto de árboles de decisión, que obtienen datos de la parte de entrenamiento (*bootstrapping*) de manera aleatoria, este modelo genera buenos resultados predictivos, al igual que el anterior modelo no se ve afectado por *outliers* ni requiere de un gran preprocesamiento ni limpieza de datos, utiliza *Out-of-Bag Error*[11], lo que evita que deba realizar estimación cruzada. Como desventaja identificada está que el tener tantos árboles puede llegar a ser complejo interpretarse.

4.3.3 Gradient Boosting

Como en el caso del anterior este modelo está formado por un conjunto de árboles de decisión, pero con la diferencia que está entrenado de manera secuencial lo que significa que cada árbol que se va creando trata de mejorar los errores del anterior y su predicción se obtiene del resultado de los árboles individuales. Como ventajas tiene las ya mencionadas en los modelos anteriores además de que tiene buena escalabilidad y puede aplicarse en un número alto de observaciones. A este modelo le ocurre lo mismo que al Random Forest que al tener tantos árboles puede perderse la interpretabilidad. (Amat Rodrigo, 2020).

4.3.4 Modelo lineal generalizado (logit)

Asimismo, en lo que respecta al modelo lineal, Amat Rodrigo (2020) expone que este modelo corresponde a un método estadístico de clasificación para variables cualitativas binarias,

[11] El error fuera de la bolsa (OOB) es el error promedio para cada zi calculado usando predicciones de los árboles que no contienen zi en su respectiva muestra bootstrap. Esto permite RandomForestClassifierestar en forma y validado mientras se entrena. Anexo página en bibliografía

haciendo que esta sea la principal razón de utilizarla, se convierte en una regresión logística al incluir varias variables independientes.

4.4 Métricas

Las métricas para evaluar los modelos son las derivadas de la matriz de confusión, como el **accuracy** que mide el porcentaje de casos que el modelo ha acertado, esta métrica no es recomendada para clases desbalanceadas, pero en este caso estamos trabajando con una base de datos balanceada, por lo tanto, es pertinente aplicarla. Se recomienda evaluar las demás métricas derivadas de las tradicionales, una de ellas es el **f1-score**, se utiliza para combinar las medidas de **precisión** y **recall** en un sólo valor. Esto es práctico porque hace más fácil el poder comparar el rendimiento combinado de la precisión y el recall. Sí lo que se busca es mantenerse lejos de falsos positivos y falsos negativos, f1-score es una buena opción.

Teniendo el balance entre las dos métricas anteriores (accuracy y recall), podemos optimizar los clasificadores, sin verse éste afectado por una distribución desbalanceada de clases. También hay que señalar que f1-score suele ser más útil que la accuracy, especialmente si tiene una distribución de clases desigual; el accuracy funciona mejor si los falsos positivos y los falsos negativos tienen un costo similar, si el costo de los falsos positivos y los falsos negativos es muy diferente, es mejor mirar tanto precisión como Recall y en este caso podemos usar f1-score como medida de robustez de los modelos, ya que por ejemplo puede ser muy costoso para las entidades de salud clasificar un paciente en un estado positivo de supervivencia, cuando realmente tiene probabilidad de supervivencia negativa, ya que los tratamientos que se aplicarán a este paciente serán los erróneos y esto podría afectar su estado de salud, si sucede el caso contrario que un paciente que realmente tiene alta probabilidad de supervivencia, y es clasificado como una persona que no va a sobrevivir el costo puede ser menor debido a que en un segundo diagnóstico o seguimiento se puede evaluar que su supervivencia es positiva y no se pone en riesgo la salud del paciente.

Por lo tanto, en este trabajo se evaluarán los modelos con el f1-score buscando reducir la tasa de falsos negativos. Para efectos del desarrollo de este trabajo, se utiliza el módulo metrics de la librería sklearn.[12]

[12] La información de la librería sklearn se encuentra disponible en: estahttps://scikit-learn.org/stable/

5. Metodología

5.1 Baseline

Experimento I: Modelación con todos los atributos clínicos

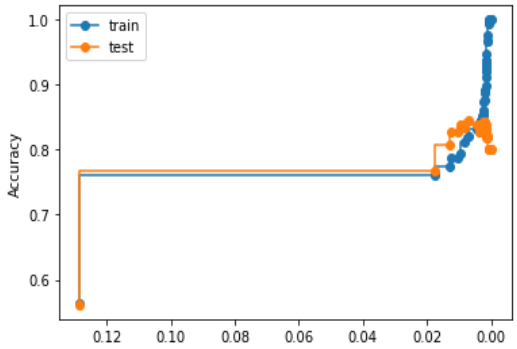
La primera iteración de modelado se realiza con todos los atributos clínicos (65 variables), como se cuenta con muchas variables explicativas se inicia con el proceso de entrenamiento utilizando un árbol de decisión, por ser un modelo simple no paramétrico que no hace suposiciones acerca de la distribución y estructura del modelo, tiene una ventaja importante y es que nos permite identificar la importancia de variables a partir de un índice de información como Gini o entropía, lo cual nos puede servir para reducir la dimensionalidad.

Árbol de clasificación

En esta iteración no se realiza el proceso de optimización de hiperparámetros sobre el nivel de complejidad del árbol, lo cual está relacionada con la profundidad, al no ser un árbol optimizado, el resultado obtenido es un árbol complejo con un nivel de profundidad de 7 mientras el número de nodos terminales es de 64. Teóricamente los árboles presentan problemas de sobreajuste u overffiting, como predicciones muy buenas en la base de entrenamiento y predicciones más bajas en la base de prueba (Amat Rodrigo,2020), si se revisa el accuracy de entrenamiento vemos que es de 89,2% mientras que para accuracy de prueba es de 75,0% aproximadamente, se evidencia problemas de sobreajuste, por tal motivo, se inicia la fase 2 de la primera iteración y es el proceso de podado del árbol.

Poda del Árbol

Con el fin de encontrar el balance entre la profundidad y complejidad del árbol con respecto a la capacidad predictiva del modelo en datos de prueba, normalmente se hace crecer el árbol de decisión hasta su mayor extensión y luego se ejecuta el proceso de poda para identificar el subárbol óptimo. Se ejecuta el modelo con diferentes niveles de complejidad alfa, para revisar hasta qué punto podemos tener una complejidad óptima.



Gráfica 8. Tasa de árbol optimizado

En la Gráfica 8 se observa la tasa óptima para alfa alrededor de un valor cercano a 0 donde el accuracy del entrenamiento es muy cercano al accuracy de prueba. Una vez se establece el alfa óptimo cercano a cero, se realiza un proceso de validación cruzada para obtener los parámetros óptimos para **`max_depth`** y el **`min_samples_leaf,`** los cuales son 5 y 3 respectivamente. Finalmente, el árbol podado tiene como profundidad 5 y un número de nodos terminales de 28.[13]

Del árbol anterior se obtiene un accuracy de 0.84% para el entrenamiento y un accuracy 85% para la base de prueba, logrando reducir el overfitting del modelo. Aunque los resultados son aceptables se conoce de antemano que los árboles son modelos inestables, por lo que no se recomienda este modelo como un modelo final para ser llevado a producción, el ejercicio se hace más para revisar la importancia de las variables predictoras y reducir la dimensionalidad de la base para modelos más robustos.

Adicionalmente se ejecuta un segundo modelo que permite acabar con el problema de la inestabilidad de un árbol de decisión. El proceso a continuación es la iteración del Random Forest y luego comparamos la importancia de variables de ambos modelos para seleccionar las variables predictoras más importantes.

Random Forest

El modelo Random Forest es un método de ensamble, el término ensamble significa grupo. Los métodos tipo ensamblador están formados de un grupo de modelos predictivos que permiten alcanzar una mejor precisión y estabilidad del modelo. Estos proveen una mejora significativa a los modelos de árboles de decisión. No se hace ningún proceso de optimización de Hiperparámetros.

Los resultados de esta iteración es un accuracy del 99% para los datos de entrenamiento y un accuracy del 78% para los datos de prueba, se evidencia un problema de overfitting, también vemos que este modelo ajusta o arroja mejores resultados que el modelo simple de árbol de clasificación. Para solucionar el problema del overfitting se realiza un proceso de optimización de hiperparámetros por validación cruzada, un parámetro a optimizar es el número de árboles que se usan como estimadores para construir el bosque aleatorio, el resultado nos arroja que el número óptimo de árboles es de 35.

Random Forest óptimo

Del resultado anterior, tenemos que el árbol óptimo es RandomForestClassifier con un resultado de 3 en el número mínimo de observaciones, 35 árboles incluidos en el modelo y 77 semillas para un resultado reproducible. El modelo presenta un accuracy de 99% con los datos de entrenamiento y un accuracy de prueba de 78% con los datos de prueba, un

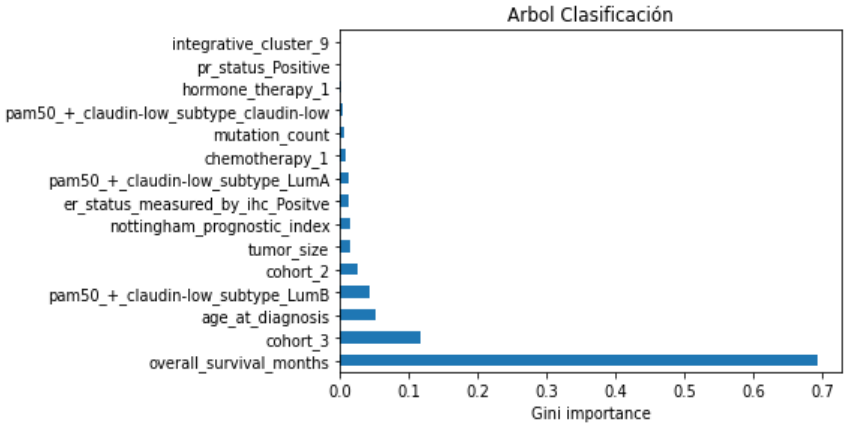
[13] El gráfico de árbol de decisión podado se encuentra en los notebooks del repositorio

resultado similar a la primera iteración del Random Forest sin realizar validación cruzada. No se logró reducir el overffiting, sin embargo, vamos a revisar la importancia de las variables y comparar las variables con la iteración del árbol de decisión.

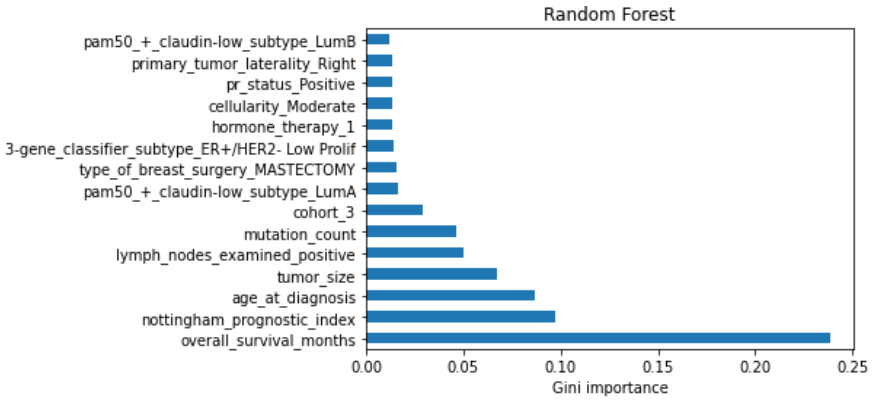
Importancia de Variables

Con el indicador de importancia de Gini se realiza un ranking de las primeras 15 variables más importantes tanto para el modelo de Árbol simple optimizado y el Random Forest Óptimo. En las Gráficas 9 y 10 podemos observar la importancia relativa de las variables según los dos modelos, en el árbol de decisión vemos que las variables meses de supervivencia tienen una participación cerca del 70%, el factor 3 de la variable cohorte con una importancia del 11% y la edad de la paciente con una importancia del 5%.

Al sumar estas tres variables encontramos una importancia del 85%. Mientras que el Random Forest arroja datos un poco más balanceados, la variable meses de supervivencia sigue siendo la más importante, aunque con un valor más discreto del 23%, la segunda variable más importante es nottingham_prognostic_index con una participación del 9%. Podemos evidenciar que según el árbol tranquilamente podemos tomar las primeras 5 variables del ranking, mientras que Random Forest sugiere tomar todas las 15 variables.



Gráfica 9. Importancia de variables en el árbol de clasificación



Gráfica 10. Importancia de variables en random forest

Finalmente, las variables seleccionadas son las 15 del ranking de Random Forest, ya que presenta más coherencia con la realidad del diagnóstico clínico, por ejemplo, la variable nottingham_prognostic_index es la segunda variable más importante, en la medicina se utiliza para determinar el pronóstico después de una cirugía por cáncer de mama. Su valor se calcula utilizando tres criterios patológicos: el tamaño del tumor; el número de ganglios linfáticos afectados y el grado del tumor y tiene sentido que sea una variable importante, en cambio el árbol de decisión muestra esta variable de importancia baja. Las siguientes iteraciones se realizan con las 15 variables seleccionadas para un modelo Gradient Boosting y un modelo paramétrico clásico de regresión logística.

5.2 Validación

Los datos de entrenamiento hace referencia a la base de datos con la que se entrenan los modelos de clasificación, en donde el modelo debe aprender los patrones y características de los pacientes con respecto a la variable objetivo (supervivencia del cáncer de mama) esta base de datos corresponde al 80% de las pacientes que sobreviven y las que mueren por la enfermedad, la muestra se selecciona a partir de un muestreo aleatorio simple estratificado, con el fin de mantener la proporción de pacientes fallecidas y vivas. En la tabla 8 vemos que la base de entrenamiento está representada por un 56.3% de pacientes que sobreviven y un 43.7% de pacientes que fallecen.

Tabla 7. Frecuencia de datos utilizados para entrenamiento y prueba

Base	Sobrevive	Fallece	Total
Entrenamiento	641	497	1.138
Prueba	160	125	285
Total	801	622	1.423

Tabla 8. Proporción de datos utilizados para entrenamiento y prueba

Base	Sobrevive	Fallece	Total
Entrenamiento	56,3%	43,7%	80%
Prueba	56,1%	43,9%	20%
Total	56,3%	43,7%	100%

El 20% de la muestra restante corresponde a la base de datos de prueba, que será utilizada para evaluar la capacidad predictiva de los modelos a entrenar, el objetivo es comprobar cómo se aproxima las predicciones de cada modelo a los verdaderos valores de la variable respuesta y cuantificar el error de predicción para comparar los modelos entrenados y seleccionar el mejor según las métricas de evaluación, para llevar a cabo esta tarea se necesita disponer de un conjunto de observaciones, de las que se conozca la variable respuesta, pero que el modelo no haya "visto", es decir que no haya participado en su entrenamiento.

Es importante verificar que la distribución de la variable respuesta es similar en el conjunto de entrenamiento y en el de prueba. Con la función train test split () de scikit-learn permite en problemas de clasificación, identificar con el argumento stratify la variable en la que se aplicará la división. Este tipo de repartición estratificada asegura que el conjunto de entrenamiento y el de prueba sean similares en cuanto a la variable respuesta como se puede observar en la tabla 7.

Fase de Validación Cruzada

Para solucionar el problema anterior se implementa el método *Leave One Out Cross-Validation (LOOCV),* un método iterativo que se inicia empleando como conjunto de entrenamiento todas las observaciones disponibles excepto una, que se excluye para emplearla como validación. La principal desventaja de este método es su costo computacional. El proceso requiere que el modelo sea reajustado y validado tantas veces como las observaciones disponibles.

Debido al alto costo computacional del método *LOOCV* en este trabajo se emplea el ***método K-Fold Cross-Validation*** (Amat Rodrigo, 2020), es un proceso iterativo que divide los datos de forma aleatoria en k grupos de aproximadamente el mismo tamaño, k-1 grupos se emplean para entrenar el modelo y uno de los grupos se emplea como validación, este emplea menos observaciones de entrenamiento que LOOCV, pero un número suficiente como para no tener un sesgo excesivo, por lo que el método K-fold CV con valores de k= [5, 10] consigue un mejor balance final.

5.3 Iteraciones y evolución

En la primera iteración se entrenaron 4 modelos únicamente con los atributos clínicos, un árbol de clasificación sencillo y el mismo árbol realizando una optimización de parámetros con variación cruzada, luego se itera un Random Forest sin optimización, y finalmente se optimiza el Random Forest mediante validación cruzada.

El resultado de los modelos en términos de predicción se analiza en la sección 6, pero *lo importante de esta primera iteración es la selección de 15 variables importantes* para reducir la dimensionalidad de la base de datos, para entrenar modelos posteriores. Por tanto, la siguiente iteración que se hace es la del modelo Gradient Boosting y posteriormente se realiza la iteración del modelo de Regresión Logística.

Una vez se completa la primera iteración con los seis modelos propuestos, se realiza una segunda iteración o experimento con los mismos modelos, pero utilizando todas las variables tanto clínicas como genéticas y finalmente se realiza una última iteración con los mismos modelos, pero tomando como predictoras las variables genéticas recomendadas en la literatura más las variables clínicas de base. En la sección 6, se exponen los resultados de las tres iteraciones.

5.4 Herramientas

La herramienta principal utilizada para el desarrollo y procesamiento de los datos fue Google Colaboratory, herramienta muy utilizada para escribir y ejecutar código. En esta herramienta se utiliza el lenguaje de programación de Python y las librerías numpy, pandas, matplotlib, seaborn, scipy, lifelines y sklearn.

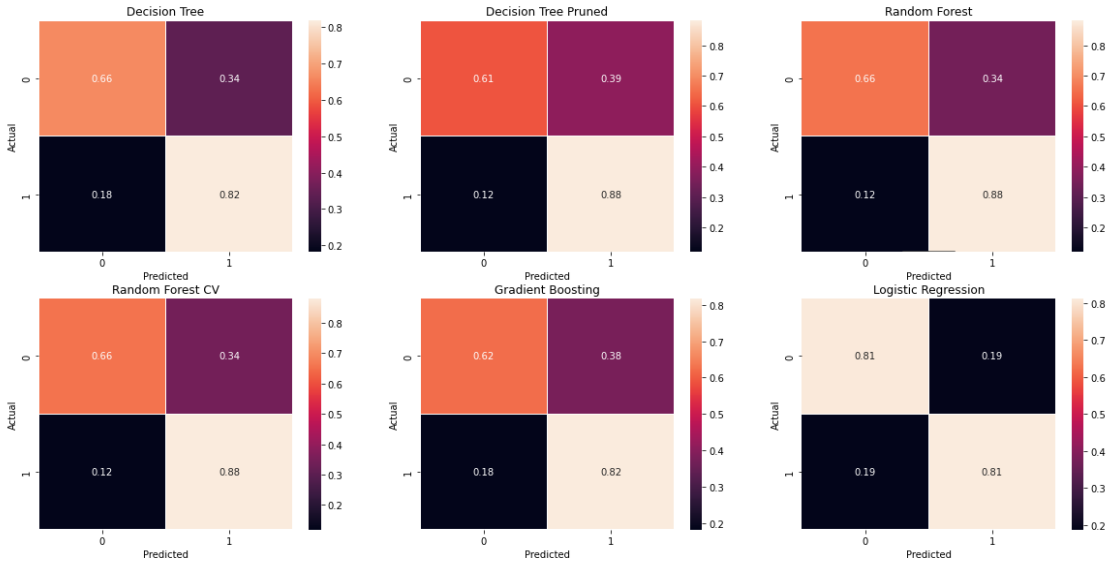
6. Resultados

A continuación, se exponen los resultados de tres iteraciones de modelos, la primera consiste en exponer los resultados de los modelos únicamente tomando como variables predictoras los atributos clínicos; los modelos entrenados son: Árbol de clasificación simple, Árbol de clasificación Optimizado, Random Forest Simple, Random Forest Óptimo, el Clasificador Gradient Boosting y Regresión Logística.

La segunda iteración es el resultado de los mismos modelos usados en la primera iteración, pero teniendo en cuenta tanto variables clínicas como genéticas y por último se exponen los resultados de la tercera iteración donde se incluye algunas de las variables genéticas recomendadas en la literatura buscando mejorar el resultado del segundo modelo.

6.1 Métricas

Experimento I – Resultados

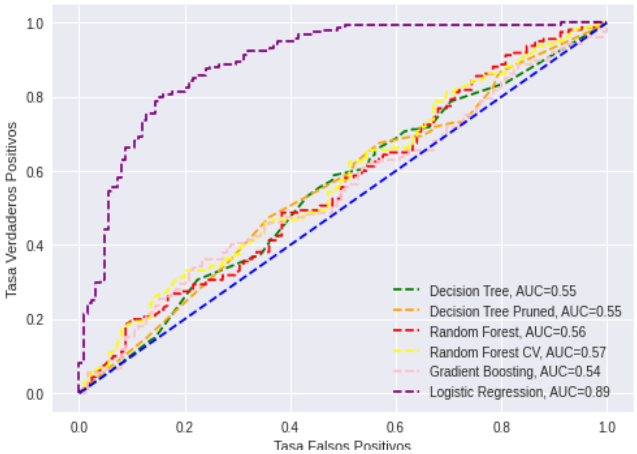


Gráfica 11. Matriz de confusión de cada modelo

Los resultados de la primera iteración (datos clínicos) se presentan en la Tabla 9, estas métricas se hallan cuando se evalúan los modelos en los datos de entrenamiento.

Tabla 9. Resultado de las métricas de los modelos en el primer

Modelos	Accuracy	f1_0	f1_1
RegresionLogistica	0.81	0.79	0.83
RandomForest	0.78	0.73	0.82
RandomForestOptimo	0.78	0.73	0.82
Árbol de decisión	0.76	0.69	0.81
Árbol de Optimizado	0.75	0.70	0.79
GradientBoosting	0.73	0.67	0.78



Gráfica 12. Curva ROC en modelos de clasificación

En la Tabla 9 se muestran las métricas de cada modelo, donde los modelos se ordenan teniendo en cuenta el f1_score de la clase 1, debido a que el objetivo es seleccionar un modelo que se mantenga lejos de falsos positivos y falsos negativos, además el balance entre las dos métricas (precisión y recall), permite que f1_score sea más útil que la precisión, especialmente si tiene una distribución de clases desigual. El accuracy funciona mejor si los falsos positivos y los falsos negativos tienen un costo similar. Pero el costo de los falsos positivos y los falsos negativos es muy diferente, cuando se quiere predecir la supervivencia del cáncer, ya que puede ser muy costoso para la IPS, clasificar una paciente en un estado positivo de supervivencia cuando este realmente tiene probabilidades de supervivencia en estado negativo.

Estas matrices las podemos ver en la Gráfica 9, donde se evidencia que el modelo con mejor rendimiento es la Regresión Logística, presenta una tasa de verdaderos positivos del 81% y una tasa de verdaderos negativos del 81%, mientras que la proporción de falsos positivos y falsos negativos es del 19%.

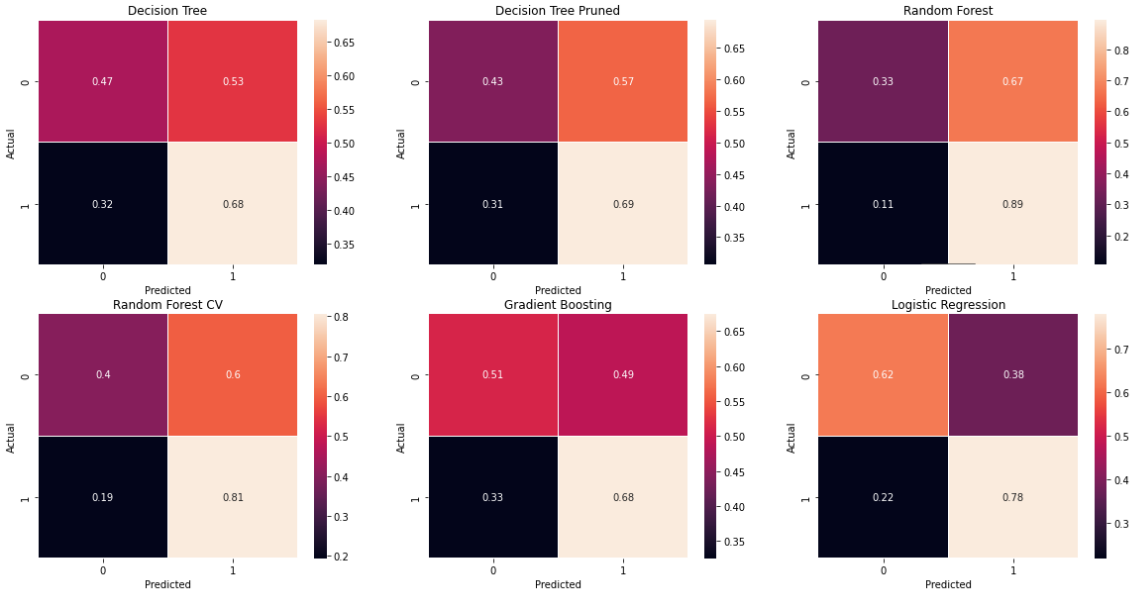
De la Tabla 9 se evidencia que la regresión logística es el modelo con mejor desempeño en las tres métricas, Accuracy de 81%, f1 score de la clase 0 (Muerte) de 79% y f1_score para clase 1 (Supervivencia) de 83%. El f1_score para la supervivencia de una paciente es superior al 80%, lo que se considera como un modelo aceptable.

Tabla 10. Clasificación Regresión Logística

	precision	recall	f1-score	support
0	0.77	0.81	0.79	125
1	0.84	0.81	0.83	160
accuracy			0.81	285
macro avg	0.81	0.81	0.81	285
weighted avg	0.81	0.81	0.81	285

En el reporte de clasificación de la Tabla 10 en la regresión logística se detalla las predicciones, presenta un recall de 81% y una precisión de 77%, se puede considerar que el modelo tiene un alto recall y una aceptable precisión, por tanto, el modelo escogido maneja bien la predicción de supervivencia.

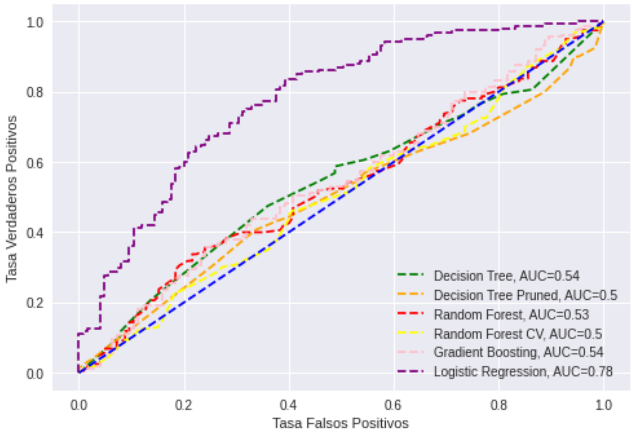
Experimento II – Resultados



Gráfica 13. Matriz de confusión de cada modelo

Tabla 11. Segunda Iteración incorporando atributos genéticos

Modelos	Accuracy	f1_0	f1_1
RegresionLogistica	0.71	0.66	0.75
RandomForest	0.65	0.45	0.74
RandomForestOptimo	0.63	0.49	0.71
Árbol de decisión	0.60	0.53	0.66
Árbol de Optimizado	0.59	0.50	0.65
GradientBoosting	0.58	0.47	0.65



Gráfica 14. Curva ROC en modelos de clasificación

Para la realización del experimento o iteración con todas la variables clínicas y genéticas, primero se hace una depuración de las variables genéticas, se seleccionaron únicamente las variables numéricas, debido a que las de tipo factor, presentan muchos valores con 0 y al incluirlas la dimensión de la base de datos aumentan alrededor de 6.271 variables, aplicado reducción de dimensionalidad por análisis de componentes principales (PCA), se realiza una reducción con las primeras 650 componentes que explican la varianza de los datos en un 72% aproximadamente, sin embargo, al entrenar los modelos las métricas no fueron satisfactorias.

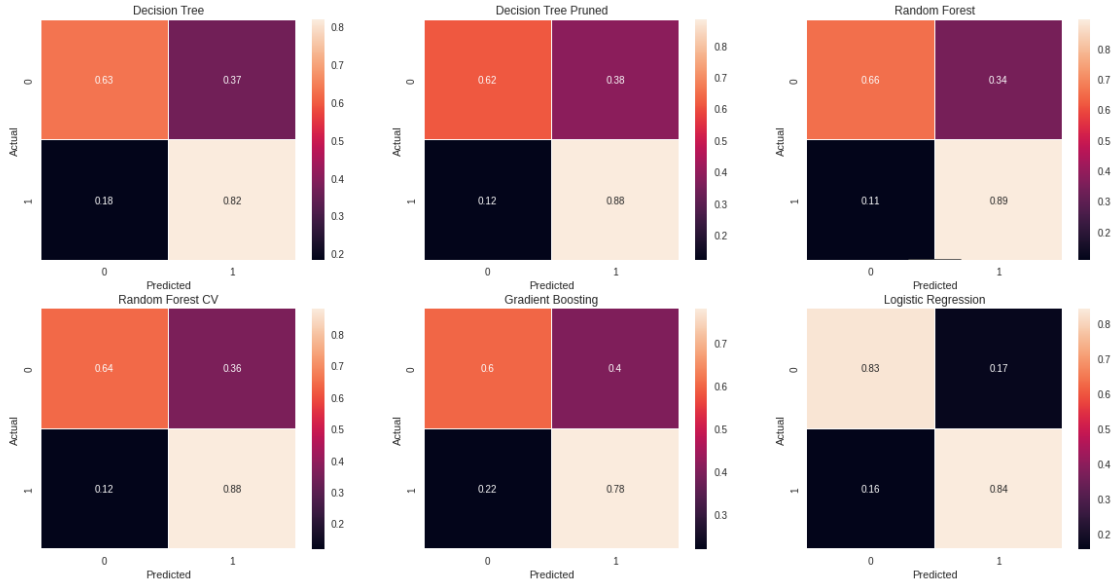
Se vuelve a analizar y se identifica que las variables genéticas de tipo factor continúan presentando muchos valores en cero, razón por la cual son excluidos nuevamente de la base de datos, como resultado se tiene 554 variables explicativas, se realiza reducción de dimensionalidad por análisis de componentes principales, se seleccionan las primeras 200 componentes que explican la varianza de los datos en un 84.3%. Una vez se hace esta depuración se realiza el entrenamiento y la evaluación de los modelos.

Tabla 12. Clasificación Regresión Logística

	precision	recall	f1-score	support
0	0.69	0.62	0.66	125
1	0.73	0.78	0.75	160
accuracy			0.71	285
macro avg	0.71	0.70	0.70	285
weighted avg	0.71	0.71	0.71	285

En términos generales no logra superar los resultados de la primera iteración (con solo atributos clínicos), es decir no logra conseguir un buen desempeño, como se observa en la Tabla 12 las métricas de la matriz de confusión f1 score para la clase 1 es del 75% mientras que la f1 score de la clase 0 es del 66%, con un accuracy de 71%.

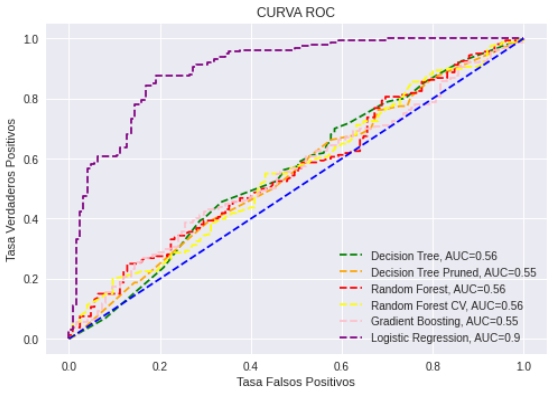
Experimento III – Resultados



Gráfica 15. Matriz de confusión de cada modelo

Tabla 13. Segundo experimento incorporando 20 atributos genéticos

Modelos	Accuracy	f1_0	f1_1
RegresionLogistica	0.84	0.82	0.85
RandomForest	0.79	0.73	0.83
RandomForestOptimo	0.78	0.71	0.82
Árbol de decisión	0.76	0.70	0.81
Árbol de Optimizado	0.74	0.68	0.78
GradientBoosting	0.70	0.64	0.75



Gráfica 16. Curva ROC en modelos de clasificación

En este último experimento se seleccionaron nuevamente todas las variables clínicas que se habían utilizado inicialmente en los dos ejercicios pasados más 13 variables genéticas que influyen en el desarrollo del cáncer (Miguel-Soca, 2016), esto se hizo con la finalidad de mejorar el resultado del segundo experimento, ya que al incluir la mayoría de las variables genéticas el modelo obtuvo un resultado por debajo del primer experimento. Con esta última modelación se logra obtener el mejor rendimiento respecto a los dos experimentos anteriores. Cómo se observa en la Tabla 13 las métricas de la matriz de confusión f1 score para la clase 1 es del 85% mientras que la f1 score de la clase 0 es del 82%, y un accuracy de 84%.

Tabla 14. Clasificación Regresión Logística

	precision	recall	f1-score	support
0	0.81	0.83	0.82	125
1	0.87	0.84	0.85	160
accuracy			0.84	285
macro avg	0.84	0.84	0.84	285
weighted avg	0.84	0.84	0.84	285

Si un AUC = 1 es un clasificador perfecto. Cuando se usa este modelo de predicción, hay al menos un valor umbral para obtener una predicción perfecta. En la mayoría de las situaciones de predicción, no existe un clasificador perfecto. En este caso la Regresión Logística tiene una buena medida de AUC de separación y es el mejor modelo de todos.

6.2 Evaluación cualitativa

Para evaluar el sobreajuste de los modelos seleccionados se calculan las métricas de error de la matriz de confusión en la base de entrenamiento y se compara con el error de prueba para verificar si existe un sobreajuste, es decir el error de prueba es mayor al error de entrenamiento.

En términos de accuracy, el modelo de regresión logística del experimento III presenta un buen balance entre las dos métricas, el accuracy de entrenamiento es del 83.0% mientras que el accuracy de la prueba es del 84.0 % en ese sentido podemos ver que no existe una diferencia significativa y que el modelo generaliza bien.

Ahora si comparamos el f score de la clase 1, vemos que el f1 score de la clase 1 para la base de entrenamiento es del 85% mientras que para la base de prueba es del 85% y

con respecto al f1 score para la clase 0 en la base de entrenamiento es del 80% mientras que para la base de prueba es del 82%.

7. Conclusiones

De los tres experimentos realizados se logra obtener un desempeño aceptable en general, resultado que se logra en la última iteración realizada donde se combina variables clínicas y 13 variables genéticas que por estudios han identificado como posibles determinantes en el desarrollo del cáncer de mama, el mejor modelo fue la regresión logística en los tres experimentos, del primer experimento se obtiene un accuracy de 81% al realizar el segundo ejercicio no se logra mejorar el desempeño del modelo y cae el accuracy a 71%, es decir disminuye en un 10% y finalmente se realiza el último experimento donde se obtiene un resultado de 84% superando el primer modelo en 3% y al segundo en 13%.

Es importante resaltar de la base de datos el conjunto de las *variables clínicas* que explican en gran parte el resultado de predecir la supervivencia, ya que algunas de estas variables se clasificaron dentro de las más relevante como lo son el índice de pronóstico de Nottingham, la edad del paciente en el momento del diagnóstico y el promedio mensual de supervivencia.

La principal limitación que se encuentra durante el desarrollo es cómo manejar los datos de las variables genéticas, ya que en algunos casos contenían resultados numéricos muy extensos y no se lograba determinar si era error en la calidad del dato o un resultado genético real.

Es posible que el resultado mejore al incluir otras variables de tipo genético en el modelo que no fueron contempladas en el último experimento y que quizás sean determinantes en el resultado, se podría considerar de igual manera también explorar otros modelos de clasificación bajo otros parámetros. Sin embargo, el resultado obtenido fue superior en comparación a los ejercicios desarrollados por algunos participantes en kaggle. Por tratarse de un ejercicio tipo académico, el modelo de predicción de supervivencia en pacientes con cáncer de mama no será llevado a producción en una IPS.

8. Bibliografía

Amat Rodrigo, J. (2020). Árboles de decisión con Python: regresión y clasificación. Ciencia de datos Net. https://www.cienciadedatos.net/documentos/py07_arboles_decision_python.html

Amat Rodrigo, J. (2020). Árboles de decisión con Python: regresión y clasificación. Ciencia de datos Net. https://www.cienciadedatos.net/documentos/30_cross-validation_oneleaveout_bootstrap#:~:text=K%2DFold%20Cross%2DValidation,-El%20m%C3%A9todo%20K&text=Consiste%20en%20dividir%20los%20datos,como%20validaci%C3%B3n%20en%20cada%20iteraci%C3%B3n

Barrios Arce, J. (26 de 07 de 2019). *Health Big Data.* Obtenido de Health Big Data: https://www.juanbarrios.com/la-matriz-de-confusion-y-sus-metricas/#:~:text=Falso%20positivo%3A%20El%20valor%20real,conoce%20como%20error%20tipo%20I

Errores OOB para bosques aleatorios. (s.f.). Scikit-learn. Recuperado de : https://scikit-learn.org/stable/auto_examples/ensemble/plot_ensemble_oob.html

Fondo Colombiano de Enfermedades de Alto costo. (14 de 10 de 2021). *Fondo Colombiano de Enfermedades de Alto costo.* Obtenido de Fondo Colombiano de Enfermedades de Alto costo. https://cuentadealtocosto.org/site/cancer/dia-mundial-contra-el-cancer-de-mama-2021/

Miguel-Soca, Pedro, Argüelles González, Ivis, & Peña González, Marisol. (2016). Factores genéticos en la carcinogénesis mamaria. Revista Finlay, 6(4), 299-316. Recuperado en 21 de mayo de 2022, de http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S2221-24342016000400007&lng=es&tlng=es.

Ministerio de Salud. (s.f.). Cáncer de mama. Ministerio de Salud. Recuperado de: https://www.minsalud.gov.co/salud/publica/ssr/Paginas/Cancer-de-

mama.aspx#:~:text=%E2%80%8BEl%20c%C3%A1ncer%20de%20mama,c%C3%A1ncer%20entre%20las%20mujeres%20colombianas.

Mukherjee, A., Russell, R., Chin, SF. *et al*. Associations between genomic stratification of breast cancer and centrally reviewed tumour pathology in the METABRIC cohort. *npj Breast Cancer* 4, 5 (2018). https://doi.org/10.1038/s41523-018-0056-8

Organización Mundial de la Salud. (2021). Cáncer de mama. *Organización Mundial de la Salud.* Recuperado de: https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer

9. Anexos

Tabla 15. Descripción de los archivos almacenados en el repositorio de GitHub

Archivo Repositorio GitHub	Descripción
1. Depuración Base Datos	Este archivo de colab contiene el tratamiento de los datos como limpieza, imputación, recodificación de variables, conversión de las variables a dummies, análisis de reducción de dimensionalidad y la selección de variables genéticas estudiadas por la literatura
2)Análisis Tiempo Supervivencia (Quimioterapia)	En estos dos colab se incluye un análisis de la supervivencia en meses de los pacientes que han recibido el tratamiento de quimioterapia o radioterapia combinado cada uno con el tratamiento de terapia hormonal y como este ha influido en la supervivencia
2)Análisis Tiempo Supervivencia (Radioterapia)	
3. Experimento I	Aplicación de los 6 modelos de clasificación utilizando sólo variables clínicas
3.1 Experimento II	Aplicación de los 6 modelos utilizando toda la base de datos, aplicando reducción dimensionalidad
3.2 Experimento III	Aplicación de los 6 modelos de clasificación utilizando variables clínicas y algunas variables genéticas propuestas por la literatura
BD Experimento _I	Corresponde a la base de datos de tipo csv que contiene sólo información clínica del paciente, es decir en este archivo no se incluye variables de genéticas
BD Experimento_II	Corresponde a la base de datos de tipo csv que contiene toda la información de la base de datos
BD Experimento_III	Corresponde a la base de datos de tipo csv que contiene la información clínica del paciente y algunas variables genéticas