Aplicación de Data Mining para el diagnóstico tardío del cáncer de mama metastásico

Marietha Kristeen Alexandra Córdova Delgado
Facultad de Ingeniería
Universidad del Pacífico
Lima, Perú
m.cordovad@alum.up.edu.pe

María del Carmen Mendoza Mechan
Facultad de Ingeniería
Universidad del Pacífico
Lima, Perú
m.mendozame@alum.up.edu.pe

Sebastian Antonio Valentino Guevara Peralta
Facultad de Ingeniería
Universidad del Pacífico
Lima, Perú
sa.guevarap@alum.up.edu.pe

Fiorella Ariana Tamariz Pantoja
Facultad de Ingeniería
Universidad del Pacífico
Lima, Perú
fa.tamarizp@alum.up.edu.pe

Santiago Wiesse Paredes
Facultad de Ingeniería
Universidad del Pacífico
Lima, Perú
s.wiessep@alum.up.edu.pe

***Abstract*—El cáncer de mama triple negativo es uno de los tipos más agresivos y difíciles de detectar en etapas tempranas, lo cual repercute negativamente en las tasas de supervivencia. Este trabajo aplica técnicas de minería de datos y aprendizaje automático para analizar información clínica, socioeconómica, demográfica y ambiental con el objetivo de identificar factores asociados al diagnóstico tardío de este tipo de cáncer en mujeres estadounidenses. Se utilizaron algoritmos como K-Means, PCA y KNN sobre un conjunto de datos obtenido de Kaggle, agrupando variables por código postal (ZIP) para entender su comportamiento a nivel territorial. El análisis exploratorio reveló patrones relevantes relacionados con desigualdades raciales, cobertura de seguro y exposición ambiental. Los resultados permiten detectar zonas de mayor vulnerabilidad, proponiendo un enfoque predictivo que prioriza variables no clínicas y resalta la utilidad de modelos integradores para la detección temprana y la formulación de políticas públicas más equitativas.**

***Index Terms*—Minería de datos, cáncer de mama, salud, ubicación, diagnóstico, K-Means, KNN, PCA.**

I. Introducción

El cáncer de mama representa un problema crítico en la salud pública, no solo por su frecuencia, si no también debido a su alta morbimortalidad [1]. De todos los diagnósticos de cáncer en mujeres al año, el 30% de ellos resulta ser de mama [2]. Asimismo, casi un 25% de los cánceres de mama presentan metástasis durante su evolución, y alrededor de un 5% son metastásicos de entrada [3].

En específico, el cáncer de mama triple negativo representa aproximadamente entre el 10 y el 15% de todos los cánceres de mama. Al ser más agresivo, es más probable que se haya propagado al momento de su detección y que reaparezca después del tratamiento. Es decir, que los diagnósticos de esta enfermedad suelen ocurrir en etapas muy avanzadas, lo que reduce la tasa de supervivencia, a 5 años de enfermedad, de 99% a un 77% [4].

Sin embargo, las posibilidades de diagnóstico oportuno y supervivencia son desiguales debido determinantes sociales [5], acceso a seguro médico [6], acceso a atención médica en zonas rurales [7] y factores de riesgo físico [5]. Asímismo, en términos económicos, un diagnóstico tardío del cáncer de mama significa mayores costos hospitalarios y deficiencia en la gestión de recursos [8], además de mayor carga económica para los pacientes [7], es decir que a medida que avanza la enfermedad los costos son mayores y los ingresos que generan los pacientes se reducen por su incapacidad de trabajar a causa del cáncer y su tratamiento.

Por lo mencionado, esta investigación se enfocará en el uso de técnicas de Minería de Datos y Machine Learning aplicado a información sobre el cáncer de mama triple negativo en mujeres estadounidenses desde una perspectiva poblacional (asociado a los pacientes ubicados en una misma región geográfica, condado).

Sin embargo, para entender el trabajo se explicarán algunos conceptos clave como son:

1) **Cáncer de mama triple negativo (CMTN):** Subtipo de cáncer de mama que no expresa los receptores de estrógeno, progesterona ni HER2. Se caracteriza por ser más agresivo y de peor pronóstico.
2) **ZIP code:** Sistema de códigos postales en EE.UU. utilizado para organizar geográficamente a los pacientes y contextualizar variables por zonas.
3) **KDD (Knowledge Discovery in Databases):** Proceso que incluye la selección, transformación, limpieza, minería de datos y evaluación de patrones en grandes volúmenes de información.
4) **K-Means:** Algoritmo no supervisado que agrupa datos en clústeres según su similitud, asignando cada punto al centro más cercano.
5) **PCA:** Técnica de reducción de dimensionalidad que permite simplificar los datos conservando las variables más informativas.
6) **KNN:** Algoritmo supervisado que predice la clase de una observación basándose en la cercanía con otras

observaciones previamente clasificadas.

La aplicación de Data Mining (DM) y Machine Learning (ML) se centrará en descubrir patrones ocultos entre variables que pueden estar asociadas al diagnóstico tardío del cáncer de mama triple negativo, como son:

- Variables clínicas agregadas desde la base de pacientes (por ZIP).
- Variables ambientales temporales (calidad del aire por mes).
- Información demográfica y social complementaria, como población por distrito.

Además, se aplicará el proceso de KDD que incluirá:

- Limpieza e imputación de datos individuales y su transformación en indicadores por ZIP.
- Análisis exploratorio: heatmaps de correlación, gráficos de series temporales y mapas geográficos interactivos.
- Técnicas de DM: agrupamiento de zonas según factores de riesgo comunes, y reglas de asociación para identificar relaciones significativas entre variables condales.

De esta manera, el análisis ayudará a identificar cuáles son las áreas con mayor vulnerabilidad frente al cáncer de mama triple negativo y los factores contextuales que puedan influir en su diagnóstico tardío.

En ese sentido, el presente proyecto se traza como objetivo principal:

> Identificar patrones territoriales y factores contextuales asociados al cáncer de mama triple negativo mediante técnicas de minería de datos, para aportar a una comprensión más integral de su distribución geográfica y desigualdades en el diagnóstico.

Y los siguientes objetivos específicos:

1) Realizar limpieza e imputación de valores faltantes en la base de datos de entrenamiento.
2) Separar las bases de datos a nivel de pacientes y a nivel de condados (código postal).
3) Visualizar la evolución temporal de variables ambientales mediante gráficos de series.
4) Explorar correlaciones y asociaciones relevantes.
5) Identificar zonas geográficas con mayor riesgo potencial para enfoques de intervención.

Con este fin, se expondrán primero el Estado del Arte sobre investigaciones relevantes que hayan abordado este mismo problema o similares. Posteriormente, se presentará el diseño del experimento, el cuál comprende una descripción del conjunto de datos a usar y la metodología empleada. Finalmente, se explicarán las tareas realizadas en la etapa de experimentación y sus correspondientes resultados.

II. Estado del Arte

Anteriormente ya se han realizado trabajos para poder pronosticar la presencia de cáncer en pacientes basados en técnicas de data mining y machine learning. Pese a que los estudios acerca del cáncer son generalmente biológicos, estas técnicas pueden asistir a un mejor modelo de diagnóstico [9]. Esto mediante enlazar atributos relacionados al cáncer del paciente con su resultado de supervivencia.

En el análisis de características para el diagnóstico de tumores de mama, se busca mejorar la velocidad y precisión del diagnóstico del método de aspiración con aguja fina mediante el uso de procesamiento de imágenes y técnicas de machine learning [10].

Más recientemente, se desarrollaron modelos predictivos de riesgo y supervivencia en cáncer utilizando únicamente determinantes sociales de la salud como ingresos, educación, empleo, seguro de salud y condiciones de vivienda. Mediante algoritmos como Random Forest y XGBoost aplicados a datos del NHANES, lograron alta precisión predictiva, destacando el valor de factores no médicos en el análisis del cáncer [11].

De forma complementaria, en otro paper se examinó cómo variables demográficas y ambientales influyen en la percepción del impacto del cáncer en sobrevivientes infantiles y adolescentes. Encontró que el sexo, el tipo de seguro y la residencia rural estaban significativamente asociados con percepciones más negativas de la experiencia oncológica, reforzando la importancia de considerar el contexto social en estudios de cáncer [12].

A diferencia de los trabajos previos que se enfocan en datos clínicos o sociales de manera aislada, el presente estudio propone un enfoque integrador que combina datos clínicos, socioeconómicos, demográficos y ambientales agregados por código postal (ZIP) para identificar factores asociados al diagnóstico tardío del cáncer de mama triple negativo. Además, se aplica un análisis territorial mediante técnicas no supervisadas como K-Means y PCA, lo cual no ha sido ampliamente explorado en la literatura revisada. Este enfoque geoespacial y multidimensional permite identificar zonas de vulnerabilidad y desigualdad en el diagnóstico, aportando así una nueva perspectiva para la formulación de políticas públicas con base en datos no clínicos. Por tanto, el presente trabajo no solo complementa los estudios existentes, sino que también introduce una metodología novedosa de análisis poblacional y territorial para un cáncer altamente agresivo.

III. Diseño del experimento

A. Descripción del conjunto de datos

Para el proyecto se identificaron dos bases de datos que fueron obtenidas mediante Kaggle [13]. La primera base de datos escogida es un conjunto de datos de 12906 observaciones y 83 atributos. Esta corresponde a un set de datos de entrenamiento con información que se divide en cuatro ramas:

- La primera, brindada por Health Verity, ecosistema de datos de atención médica de los Estados Unidos, es sobre datos demográficos, opciones de diagnóstico y tratamiento, y seguro proporcionado sobre pacientes que fueron diagnosticadas con cáncer de mama triple negativo metastásico entre 2015 y 2018. Cuenta con atributos como edad, raza o IMC a nivel de pacientes. Cabe resaltar que atributos como nombres y apellidos no se encuentran a forma de proteger la identidad de las pacientes.

- La segunda rama corresponde a datos geodemográficos y socioeconómicos a nivel de código postal como ingresos, educación o alquiler, obtenidos a través de las páginas del Servicio Postal de EE. UU., la Oficina del Censo de EE. UU., el Servicio Meteorológico Nacional, la Encuesta de la Comunidad Estadounidense y el IRS.
- La tercera, también utilizando el nivel de código postal, se agregaron datos toxicológicos del aire de la NASA y la Universidad de Columbia.
- Finalmente, una única columna que indica si el cáncer se diagnosticó dentro de los 90 días.

Para simplificar las dimensiones de las ramas y la identificación de los atributos métricos y categóricos se presenta la Tabla 1:

TABLE I: Distribución de variables por rama de datos

Rama	**Número de columnas**	**Número de variables métricas**	**Número de variables categóricas**
Pacientes	15	2	13
Geodemográficos y socioeconómicos	64	64	-
Toxicológicos	3	3	-
Diagnóstico	1	-	1

El segundo conjunto de datos corresponde a datos de validación, por lo que contiene la misma cantidad de atributos del set anterior a excepción la cuarta rama (variable de salida "Diagnóstico"), ya que es la que se va predecir; pero 5792 observaciones.

Finalmente, se tiene un conjunto de datos con información de los códigos zip con 41483 observaciones y 9 atributos, entre los cuales se encuentra la latitud y longitud de los puntos de código postal [14].

B. Metodología

Para poder analizar correctamente el conjunto de datos es fundamental optar por estrategias y algoritmos que nos garanticen un tratamiento riguroso de los datos. En el presente estudio se utilizó Python y librerías como Pandas, Numpy y Scikit-Learn para aplicar estrategias que permitan capturar información valiosa a partir del conjunto de datos.

1) Estrategia de preprocesamiento: Para el preprocesamiento de los datos se aplicaron enfoques de la metodología KDD para su transformación; de esta manera, se realizaron las siguientes tareas:

- Tratamiento de campos con datos faltantes: Se calculó la cantidad de valores nulos de cada campo y se eliminaron aquellos con una suma de valores nulos mayor al 50%, ya que consideramos que no aportarán valor real al estudio.
- Imputar datos faltantes: Se realizó en 2 partes. En la primera, se consideraron las observaciones correspondientes a un de un mismo código postal (ZIP). Si la variable era categórica, se imputó utilizando la moda dentro del mismo código postal; si era numérica, se usó la media de las observaciones del conjunto asociado al mismo ZIP. En la segunda parte, para aquellos casos que no pudieron ser imputados por este criterio, se aplicó una imputación general: la moda global para las variables categóricas y la media global para las variables numéricas.
- Inspección y corrección de errores: se exploraron los campos del conjunto de datos para detectar errores comunes como valores inconsistentes.
- Tratamiento de outliers: Se generaron visualizaciones mediante diagramas de caja (boxplots) con el fin de analizar la distribución de cada variable del conjunto de datos. Dado que el conjunto incluye zonas con diferentes densidades poblacionales, no se consideró apropiado realizar una imputación basada en cuantiles, pues esto podría distorcionar la variablidad real de los datos. Por ello, no se aplicó tratamiento de valores atípicos en la medida que se consideró que estos aportaban valor al estudio.
- Transformación de variables: Con el objetivo de consolidar la información de variables redundantes y facilitar su uso, se realizaron 2 transformaciones. Para la variable de edad, se creó una nueva variable denominada age_range, que agrupa a los pacientes según rangos etarios definidos (menores de 10, 10 a 19, 20 a 29, sucesivamente hasta mayores de 80 años). De manera similar, para la variable raza, se generó la variable race_range, que consolida las columnas asociadas a la raza del paciente (race_white, race_black, etc.).
- Eliminación de variables irrelevantes y redundantes: Una vez creadas las variables age_range y race_range, se eliminaron las columnas usadas para su construcción, ya que no aportar información adicional al modelo y con ello se logra reducir la dimensionalidad del conjunto de datos. Asimismo, se eliminaron variables que no se usan en el análisis: patient_id y patient_zip3.
- Disión del conjunto de datos: El conjunto de datos fue dividido en tres subconjuntos con fines de entrenamiento y evaluación del modelo. En una primera etapa, se separó el 70% de los datos para entrenamiento (train_clean) y el resto para validación y prueba interna. Posteriormente, este 30% se dividió en partes iguales para obtener un 15% destinado a validación (valid_clean) y un 15% para prueba (test_clean).

2) Algoritmos a emplear: Se optó por una combinación de enfoques complementarios para la selección y evaluación de variables predictoras. En primer lugar, se aplicaron cuatro métodos para la selección de características: tres de ellos corresponden a técnicas propias de minería de datos (Random Forest, Mutual Information y Chi-cuadrado), y uno asociado a algoritmos de aprendizaje automático supervisado (XGBoost). El objetivo de emplear múltiples enfoques fue comparar sus rankings y determinar qué variables muestran una importancia consistente en distintos modelos. Posteriormente, se utilizó un método adicional de aprendizaje supervisado para construir modelos predictivos utilizando subconjuntos de variables seleccionadas:

- KDD: Metodología estructurada para la exploración sis-

temática y extracción de patrones relevantes en los datos.

- StratifiedGroupKFold: Esta técnica permite entrenar y evaluar modelos respetando el balance de clases en cada grupo y que los datos del mismo grupo no se crucen en el entrenamiento y en la validación.
- Random Forest: algoritmo de aprendizaje supervisado que construye múltiples árboles de decisión durante el entrenamiento y devuelve la clase que representa la mayoría de los votos entre los árboles individuales. En el contexto de selección de variables, Random Forest permite identificar aquellas variables que tienen mayor influencia en la predicción del resultado, asignándoles un puntaje de importancia basado en la reducción del índice de impureza (Gini) o en el descenso en la precisión del modelo al permutar aleatoriamente una variable. Este enfoque tiene la ventaja de manejar tanto variables categóricas como numéricas y de ser robusto frente a valores atípicos y datos faltantes.
- Mutual Information (Información Mutua): La información mutua es una medida estadística que cuantifica la dependencia entre dos variables. En la selección de características, este método evalúa cuánta información aporta una variable independiente sobre la variable dependiente, sin asumir una relación lineal entre ellas. A diferencia de otras métricas que se centran en correlaciones lineales, la información mutua permite capturar relaciones más complejas. Por ello, es especialmente útil en conjuntos de datos heterogéneos donde las asociaciones entre variables pueden ser no lineales o no triviales.
- Chi-cuadrado: es una técnica estadística tradicionalmente utilizada para analizar la relación entre dos variables categóricas. En selección de variables, permite evaluar si existe una dependencia significativa entre una variable independiente categórica y la variable objetivo. Cuanto mayor es el valor de la estadística Chi-cuadrado, mayor es la asociación entre ambas variables. Este método es útil especialmente cuando se trabaja con variables codificadas, como las que resultan del uso de etiquetas o clases discretas, aunque requiere que los datos sean no negativos y que se cumplan ciertos supuestos de frecuencia esperada.
- XG-Boost: XGBoost (Extreme Gradient Boosting) es un algoritmo de aprendizaje automático supervisado basado en árboles de decisión que utiliza la técnica de gradient boosting para optimizar el rendimiento del modelo. A diferencia de Random Forest, que construye árboles de manera paralela, XGBoost construye árboles de forma secuencial, corrigiendo los errores de predicción de los árboles anteriores. Para la selección de variables, XGBoost calcula una medida de importancia basada en cuántas veces y con qué ganancia una variable se utiliza para dividir nodos en los árboles. Su capacidad para capturar relaciones no lineales, manejar interacciones entre variables y optimizar el modelo de forma eficiente lo convierte en una herramienta poderosa para la priorización de variables relevantes.

3) Selección de métricas de calidad de la tarea: Para evaluar la calidad de los modelos de clasificación, se seleccionaron métricas adecuadas a la naturaleza de la tarea, que corresponde a un problema de clasificación binaria (diagnóstico probable dentro o fuera del periodo de 90 días). Las métricas fueron calculadas tanto en la fase de validación cruzada como en el conjunto de prueba final.

- Precisión (accuracy): Medir la proporción de predicciones correctas. Es útil para evitar falsos positivos.
- Recall (sensibilidad): Medir la capacidad del modelo para identificar correctamente los casos positivos (pacientes con CMTN) y minimizar los falsos positivos.
- F1-score: Métrica principal de referencia al balancear precisión y recall. Es especialmente útil en conjuntos de datos desbalanceados como los relacionados a enfermedades raras o agresivas.
- AUC-ROC (Área bajo la curva ROC): Evaluar la capacidad general del modelo para discriminar entre clases.

4) Estrategia para la optimización de hiperparámetros: En esta etapa, se optó por una estrategia controlada para la selección de variables. Como punto de partida, se tomó como referencia una solución publicada en la plataforma Kaggle, en el marco de la competencia WiDS 2024. En dicha propuesta [14], se emplearon técnicas de feature importance aplicadas a modelos como LightGBM para identificar las nueve variables más relevantes en la predicción del desenlace oncológico.

Si bien esta aproximación fue desarrollada sobre un conjunto de datos distinto, proporcionó una guía práctica para la reducción de dimensionalidad y la priorización de variables con mayor poder predictivo. En base a esta referencia, se adaptó el proceso de selección de atributos a las características específicas del conjunto de datos utilizado en este estudio, el cual incorpora variables demográficas, socioeconómicas y ambientales.

IV. EXPERIMENTACIÓN Y RESULTADOS

A. Tarea 1. Análisis exploratorio

El análisis comienza con una serie de boxplots (véase Anexo 1) que permiten explorar la distribución de variables clave en el conjunto de datos. Por ejemplo, la edad de los pacientes muestra una mediana cercana a los 50 años, con una dispersión moderada y varios valores extremos hacia edades avanzadas, lo que indica la presencia de pacientes de edad avanzada. En contraste, la variable age20s muestra valores concentrados en cero o muy bajos, lo cual sugiere una baja representación de personas jóvenes en el grupo de estudio. La variable female, probablemente codificada como binaria (0/1), presenta una mediana en 0.5, lo que podría indicar un equilibrio entre géneros o un sesgo en la codificación. Variables socioeconómicas como incomehousehold35to50 tienen una mediana intermedia, pero revelan mucha variabilidad, con algunos outliers altos que podrían corresponder a zonas con mayores ingresos. En cuanto a las variables raciales, se observa que la población blanca predomina con valores más altos en su distribución, mientras que la población negra

está subrepresentada. También se examinan variables sociales como la pobreza, el porcentaje de personas sin seguro médico (healthuninsured) y los niveles de contaminación por PM2.5, los cuales muestran una distribución dispersa, con valores altos en ciertos sectores, lo cual indica desigualdades marcadas tanto en acceso a salud como en condiciones ambientales.

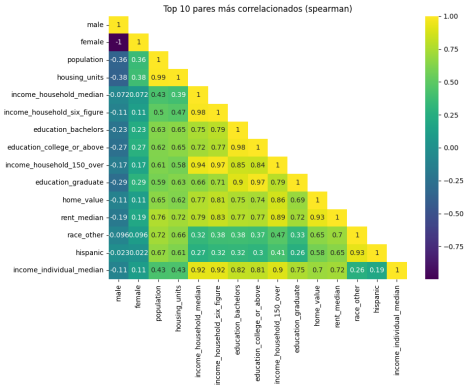


Fig. 1: Heatmap de correlación: top 10 de variables correlacionadas

A continuación, la matriz de correlación (véase figura 1) permite identificar asociaciones relevantes entre variables. Para el gráfico se utilizaron los 10 pares de variables más correlacionados en búsqueda de una mejor visualización del heatmap. Se destacan fuertes correlaciones entre factores socioeconómicos como ingreso, nivel educativo y pobreza, lo cual sugiere colinealidad y posibles redundancias que deben considerarse al construir modelos predictivos. Este patrón se confirma en las visualizaciones exploratorias: la distribución de pacientes por raza muestra una mayoría de pacientes blancos (más del 50

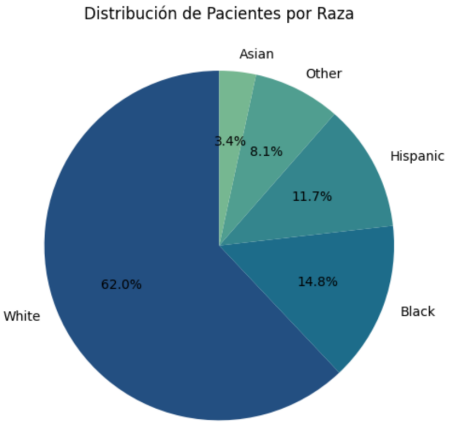


Fig. 2: Gráfico pie: Distribución de pacientes por raza

Posterior a ello, se realizó un gráfico pie (Véase Fig.2), el cual muestra una clara predominancia de pacientes de raza blanca, quienes representan el 62.0% de la población estudiada. Le siguen los pacientes afroamericanos (14.8%), hispanos (11.7%), otros grupos raciales (8.1%) y asiáticos (3.4%). Esta distribución desigual refleja potenciales disparidades en la representación poblacional dentro del estudio. Es importante considerar que la raza ha sido documentada como un factor que puede influir en los resultados clínicos del cáncer de mama, incluyendo el tiempo hasta la metástasis, debido a diferencias en factores biológicos, acceso a atención médica y adherencia a tratamientos.

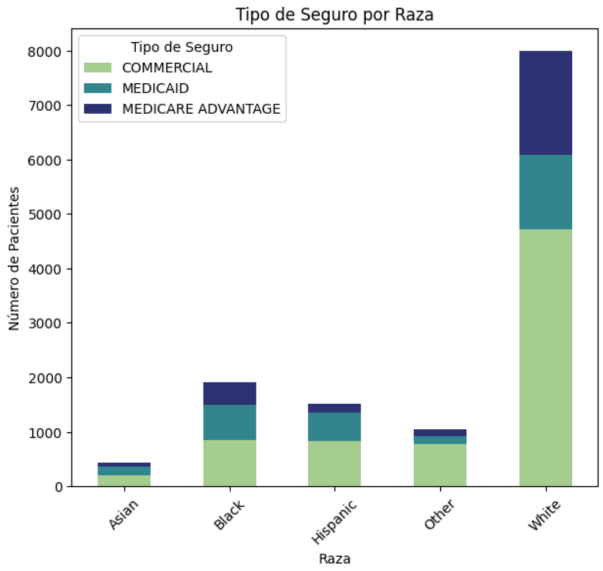


Fig. 3: Grafico de barras apiladas: Tipo de Seguro por Raza

Luego, se realizó un gráfico de barras (Véase Fig.3) que revela patrones preocupantes sobre las desigualdades en cobertura de salud entre grupos raciales. Los pacientes blancos muestran un acceso predominante a seguros Commercial y Medicare Advantage, indicando potencialmente mejor cobertura, mejor tratamiento y acceso a especialistas. En contraste, los pacientes afroamericanos e hispanos presentan una mayor proporción de cobertura Medicaid, lo que puede asociarse con limitaciones en el acceso a ciertos tratamientos o centros especializados. Estas disparidades en seguros podrían ser un predictor crítico del intervalo entre diagnóstico inicial y metástasis, ya que la calidad y continuidad de la atención médica, influenciadas por el tipo de seguro, son fundamentales para la detección temprana de recurrencias y manejo de la enfermedad.

B. Tarea 2. Selección de variables

En la etapa de selección de variables, se utilizaron cuatro métodos distintos: Random Forest, Mutual Information, Chi-cuadrado y XGBoost. Todos coinciden en destacar como la variable más importante al código de diagnóstico de cáncer de mama (breastcancerdiagnosiscode), lo cual es evidente, ya que la variable diágnostico hace referencia a la evolución de un cancer de mama inicial a uno metástasico. También se repiten variables como la edad del paciente, tamaño familiar, nivel educativo, contaminación ambiental (PM2.5) y pobreza, lo que resalta la influencia conjunta de factores clínicos, sociales y ambientales.

El análisis de la sección "Variables seleccionadas" permite identificar qué variables fueron destacadas de manera consistente. Solo una variable (breastcancerdiagnosiscode) fue seleccionada por los cuatro métodos, confirmando su alto valor predictivo. Otras catorce variables fueron seleccionadas por al menos dos métodos, entre ellas: ozone, commutetime, educationbachelors, incomehousehold10to15, limitedenglish, metastaticcancerdiagnosiscode, y payertype. Estas variables reflejan tanto condiciones de salud como factores sociales y demográficos, lo cual valida una perspectiva integral para el análisis.

En cuanto a las métricas de desempeño, los modelos presentan una tasa de acuerdo entre métodos del 81.4

C. Tarea 3. Regresión Logística y LDA

El modelo de Regresión Logística confirma estas observaciones. Su matriz de confusión muestra un buen desempeño general, aunque hay un número considerable de falsos positivos (664 casos). La curva ROC indica un AUC de 0.76, lo cual representa un modelo aceptable. El análisis de coeficientes muestra que el diagnóstico de cáncer de mama tiene el mayor peso positivo en la predicción, seguido por variables como ozone, metastaticcancerdiagnosiscode, commutetime, y poverty. En cambio, algunas variables como educationbachelors o incomehouseholdmedian tienen un impacto menor o incluso negativo, lo cual permite matizar su interpretación.

Finalmente, el análisis mediante LDA (Análisis Discriminante Lineal) presenta una matriz de confusión similar y un AUC de 0.76, mostrando un rendimiento comparable al de la regresión logística. La variable más influyente vuelve a ser breastcancerdiagnosiscode, seguida por familysize, payertype y incomehousehold75to100. Otras variables como commutetime, limitedenglish y poverty también tienen peso, y algunas como educationcollegeorabove tienen coeficientes negativos, lo que sugiere una relación inversa con la clase objetivo. Esto refuerza la idea de que los determinantes sociales tienen un rol importante y deben ser cuidadosamente considerados en intervenciones o políticas.

D. Tarea 4. Machine Learning - RandomForest

Se utilizó Random Forest como modelo de referencia dado que fue empleado en el trabajo de una competencia de Kagle como algoritmo base para la comparación de métodos predictivos en [área/dominio específico].

Para evaluar el desempeño del modelo, se implementaron tres configuraciones: el modelo base con parámetros por defecto, y dos variantes optimizadas mediante búsqueda de hiperparámetros, una optimizada para el conjunto de entrenamiento (T) y otra para el conjunto de validación (Tv).

Los resultados obtenidos revelan limitaciones significativas en la capacidad predictiva de todos los modelos evaluados. El modelo base, aunque superior a las variantes optimizadas, alcanzó un RMSE de 0.414 y un R^2 de apenas 0.269, indicando que explica únicamente el 26.9

Los gráficos de dispersión del RMSE se evidencian la baja capacidad predictiva de los modelos, observándose una considerable dispersión de los puntos respecto a la línea de predicción perfecta en los tres casos. Esta alta dispersión, particularmente notable en los valores extremos, sugiere que Random Forest presenta limitaciones inherentes para capturar los patrones subyacentes en este conjunto de datos específico, justificando la exploración de metodologías alternativas más sofisticadas.

V. DISCUSIÓN

¿Cómo podría ser mejorada la tarea?

Las mejoras se pueden presentar en dos ejes: calidad del dataset y enfoque geoespacial. Si bien los datos obtenidos de Kaggle son bastante completos, las variables son considerablemente redundantes y poco variables, lo que limita la capacidad de algunos modelos predictivos. Sería útil contar con variables con enfoque temporal para analizar la evolución de los pacientes. También sería interesante profundizar el análisis geoespacial, al integrar información interactiva mediante mapas que relacione el riesgo con variables sociales y ambientales.

¿Qué elementos del proceso le causaron más dificultad? ¿Cómo podrían superarlos?

La principal dificultad se presentó al desarrollar el preprocesamiento de datos, por la complejidad y dispersión de variables. Muchas columnas se componían por variables dummy que debían transformarse para evitar excesiva dimensionalidad. Esto requiere manejar con mucho cuidado el encoding, la imputación y la validación de integridad entre variables. Otra dificultad importante fue la imposibilidad de aplicar un modelo de PCA, ya que los valores de los datos eran muy lineales y los componentes no representaban adecuadamente la variabilidad.

VI. CONCLUSIÓN Y TRABAJOS FUTUROS

Este estudio ha evidenciado que el diagnóstico tardío del cáncer de mama triple negativo no depende únicamente de factores clínicos, sino que está profundamente condicionado por determinantes sociales, económicos y territoriales. El uso de técnicas de minería de datos permitió identificar patrones relevantes en variables como raza, tipo de seguro, nivel de ingresos y exposición ambiental, revelando focos de vulnerabilidad en ciertas poblaciones y zonas geográficas.

Los hallazgos permiten concluir que:

Existen desigualdades marcadas en el acceso y oportunidad de diagnóstico, asociadas a raza, seguro de salud y condiciones socioeconómicas.

Factores ambientales, como la contaminación del aire (PM2.5), y el nivel de pobreza en el código postal de residencia tienen un impacto potencial en el riesgo de diagnóstico tardío.

La edad al diagnóstico presenta una distribución no homogénea, lo cual sugiere la conveniencia de adaptar los modelos predictivos a distintos grupos etarios.

Se plantea como siguiente etapa del proyecto:

Implementar modelos más robustos de clasificación como XGBoost o redes neuronales, incluyendo ajuste de hiperparámetros.

Explorar modelos secuenciales o temporales que consideren la evolución de los pacientes en el tiempo.

Integrar mapas interactivos para la visualización de zonas de alto riesgo y su vinculación con políticas públicas.

Expandir la base de datos con fuentes adicionales (censales o clínicas) que enriquezcan el análisis contextual y territorial.

REFERENCES

[1] Dvir, K., Giordano, S., & Leone, J. P. (2024). Immunotherapy in Breast Cancer. International Journal of Molecular Sciences, 25(14), 7517. https://doi.org/10.3390/ijms25147517

[2] American Cancer Society (2025). Key Statistics for Breast Cancer. https://www.cancer.org/cancer/types/breast-cancer/about/how-common-is-breast-cancer.html

[3] Beuzeboc, P. (2015). Cáncer de mama metastásico. EMC - Ginecología-Obstetricia, 51(1), 1-14. https://doi.org/10.1016/s1283-081x(15)70034-2

[4] American Cancer Society (2023). Triple-negative Breast Cancer. https://www.cancer.org/cancer/types/breast-cancer/about/types-of-breast-cancer/triple-negative.html

[5] Rodríguez-González, N., Ramos-Monserrat, M. J., & De Arriba-Fernández, A. (2022). ¿Cómo influyen los determinantes sociales de la salud en el cáncer de mama? Revista de Senología y Patología Mamaria, 36(3), 100467. https://doi.org/10.1016/j.senol.2022.100467

[6] American Cancer Society (s.f.) Acceso a la cobertura de salud. (s/f). https://www.fightcancer.org/es/what-we-do/access-health-insurance

[7] Unger, J. M., McAneny, B. L., & Osarogiagbon, R. U. (2025). Cancer in rural America: Improving access to clinical trials and quality of oncologic care. CA A Cancer Journal For Clinicians. https://doi.org/10.3322/caac.70006

[8] National Cancer Institute [NCI] (2025). Cancer Disparities. https://www.cancer.gov/about-cancer/understanding/disparities

[9] Kaur, I., Doja, M., Ahmad, T. (2022). Data mining and machine learning in cancer survival research: An overview and future recommendations. Journal Of Biomedical Informatics, 128, 104026. https://doi.org/10.1016/j.jbi.2022.104026

[10] W. Nick Street, W. H. Wolberg, O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," Proc. SPIE 1905, Biomedical Image Processing and Biomedical Visualization, (29 July 1993); https://doi.org/10.1117/12.148698

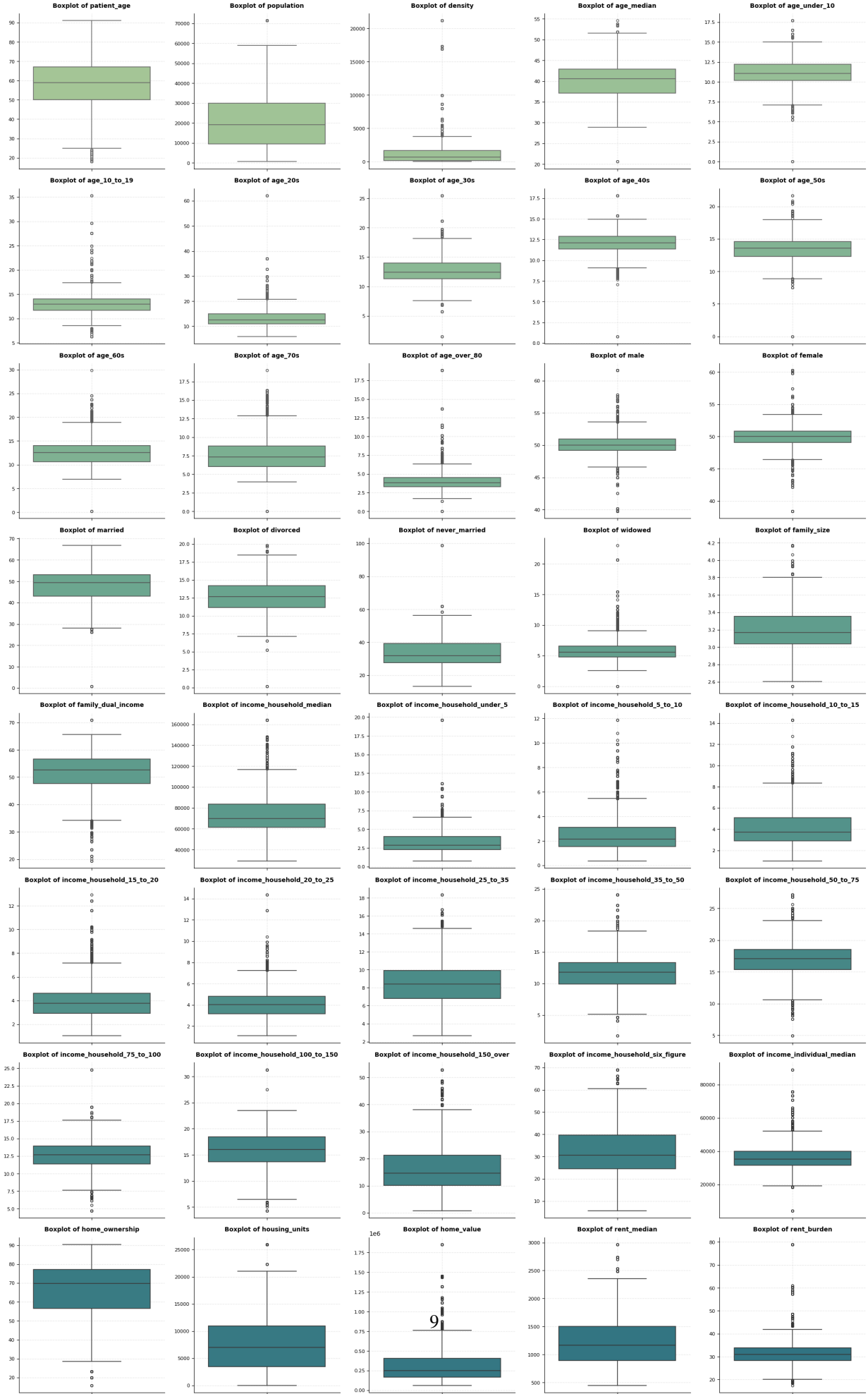
[11] Zhang, S., Jin, J., Zheng, Q., & Wang, Z. (2025). Building a cancer risk and survival prediction model based on social determinants of health combined with machine learning: A NHANES 1999 to 2018 retrospective cohort study. Medicine, 104(6), e41370. https://doi.org/10.1097/md.0000000000041370

[12] Cetin, Nazan, "Examining Demographic and Environmental Factors in Predicting the Perceived Impact of Cancer on Childhood and Adolescent Cancer Survivors" (2022). Dissertations and Theses. Paper 6114. https://doi.org/10.15760/etd.7974

[13] Ramakrishnan R (2024). WIDS2024Challenge1 - Baseline v1 [Notebook]. Kaggle. https://www.kaggle.com/code/ravi20076/wids2024challenge1-baseline-v1

[14] Syah I (2024). WiDS 2024 (1/2) - 2nd place solution [Notebook]. Kaggle. https://www.kaggle.com/code/iqbalsyahakbar/wids-2024-1-2nd-place-solution/notebook

VII. ANEXOS


Boxplot of patient_age
Boxplot of population
Boxplot of density
Boxplot of age_median
Boxplot of age_under_10
Boxplot of age_10_to_19
Boxplot of age_20s
Boxplot of age_30s
Boxplot of age_40s
Boxplot of age_50s
Boxplot of age_60s
Boxplot of age_70s
Boxplot of age_over_80
Boxplot of male
Boxplot of female
Boxplot of married
Boxplot of divorced
Boxplot of never_married
Boxplot of widowed
Boxplot of family_size
Boxplot of family_dual_income
Boxplot of income_household_median
Boxplot of income_household_under_5
Boxplot of income_household_5_to_10
Boxplot of income_household_10_to_15
Boxplot of income_household_15_to_20
Boxplot of income_household_20_to_25
Boxplot of income_household_25_to_35
Boxplot of income_household_35_to_50
Boxplot of income_household_50_to_75
Boxplot of income_household_75_to_100
Boxplot of income_household_100_to_150
Boxplot of income_household_150_over
Boxplot of income_household_six_figure
Boxplot of income_individual_median
Boxplot of home_ownership
Boxplot of housing_units
Boxplot of home_value
1e6
Boxplot of rent_median
Boxplot of rent_burden

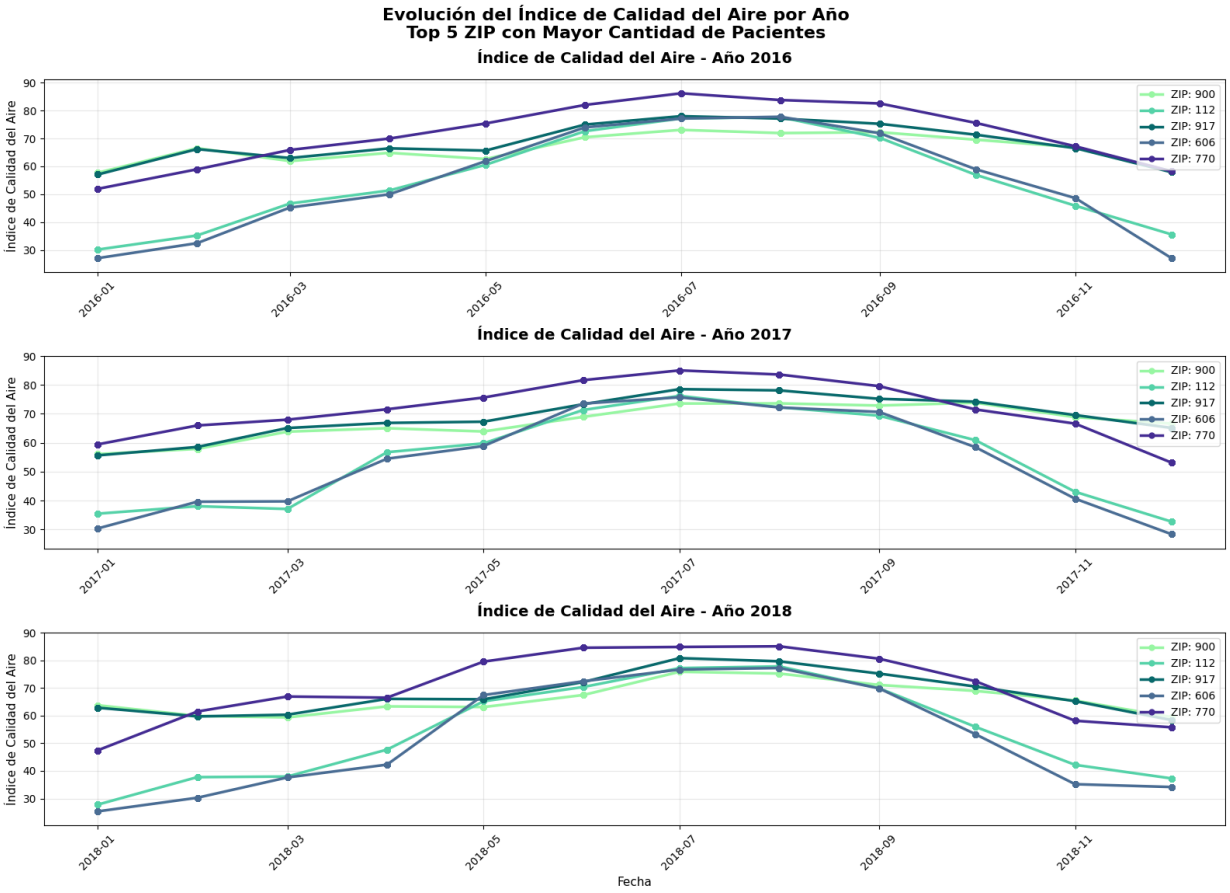


Fig. 5: Anexo 2: Series temporales del indice de calidad del aire