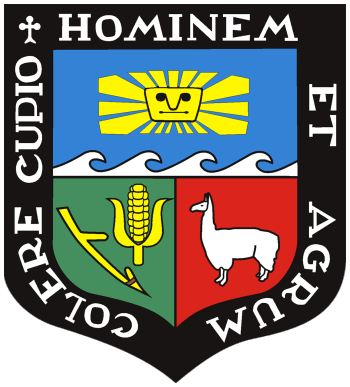
UNIVERSIDAD NACIONAL AGRARIA
LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN



"CLASIFICACIÓN DE DATOS TEXTUALES PROVENIENTES DE UN STREAMING APLICANDO EL MÉTODO DE REPRESENTACIÓN DE TEXTO TF-IDF EN UNA REGRESIÓN LOGÍSTICA"

TESIS PARA OPTAR TÍTULO DE
INGENIERA ESTADÍSTICA INFORMÁTICA

FIORELLA ALEXANDRA FLORES CÁCERES

LIMA – PERÚ

2024

TESIS_final.docx

INFORME DE ORIGINALIDAD

16%	15%	5%	9%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	towardsdatascience.com Fuente de Internet	1%
2	ichi.pro Fuente de Internet	1%
3	repositorio.ucv.edu.pe Fuente de Internet	1%
4	www.lamolina.edu.pe Fuente de Internet	1%
5	repository.usta.edu.co Fuente de Internet	1%
6	rstudio-pubs-static.s3.amazonaws.com Fuente de Internet	<1%
7	github.com Fuente de Internet	<1%
8	hdl.handle.net Fuente de Internet	<1%
9	Submitted to Universidad Internacional de la Rioja	<1%

UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

FACULTAD DE ECONOMÍA Y PLANIFICACIÓN

“CLASIFICACIÓN DE DATOS TEXTUALES PROVENIENTES DE UN STREAMING APLICANDO EL MÉTODO DE REPRESENTACIÓN DE TEXTO TF-IDF EN UNA REGRESIÓN LOGÍSTICA”

TESIS PARA OPTAR EL TÍTULO DE
INGENIERA ESTADÍSTICA INFORMÁTICA

PRESENTADA POR:
FIORELLA ALEXANDRA FLORES CÁCERES

SUSTENTADA Y APROBADA ANTE EL SIGUIENTE JURADO:

Dr. Raphael Félix Valencia Chacón PRESIDENTE	Dr. Jaime Carlos Porras Cerrón ASESOR
Dr. Iván Dennys Soto Rodríguez MIEMBRO	Dr. Cesar Higinio Menacho Chiok MIEMBRO

LIMA – PERÚ

2024

DEDICATORIA

A mi madre y a mi abuela materna, mis pilares eternos.

A mi prometido, mi pilar emergente.

AGRADECIMIENTOS

Agradezco al Vicerrectorado de Investigación de la UNALM,

por el financiamiento otorgado en el 11° Concurso de Subvención de Tesis de Pregrado.
Asimismo, al Mg.Sc. Duber Chinguel por su orientación en todo el proceso del concurso.

A mi asesor, el profesor Jaime Porras,

por su disposición de orientación desde el primer momento, continua guía y compromiso con la investigación.

A mis amigos Katja, Lita, José Luis, Evelyn, Vilma e Ivonne,

por estar siempre, sobre todo en los momentos cruciales.

A mi prometido Peter y a su abuelo Don Hilario García,

por la excesiva paciencia y enorme apoyo durante el desarrollo del presente trabajo de investigación.

RESUMEN

El presente trabajo de investigación tuvo como finalidad implementar un modelo de regresión logística utilizando datos textuales transformados mediante el método de representación de texto TF-IDF, con el objetivo de clasificar comentarios de docentes en *streamings* de orientación sobre la estrategia Aprendo en Casa realizados por el Ministerio de Educación. El procedimiento de análisis se dividió en pre-procesamiento de los datos, análisis exploratorio de los datos, aplicación del método de representación de texto TF-IDF, estimación y evaluación del modelo; y clasificación de nuevos comentarios. Para la etapa de pre-procesamiento se realizó la limpieza y estandarización de los datos textuales de los comentarios; mientras que en el análisis exploratorio se obtuvieron indicadores descriptivos de los comentarios de cada categoría utilizando n-gramas. En la aplicación del método de representación de texto TF-IDF se elaboró la matriz documento-término a partir de la muestra de entrenamiento y se utilizó la prueba Chi Cuadrado para la selección de variables. En la estimación del modelo de clasificación se obtuvo el modelo final ajustado con los datos de entrenamiento provenientes de la matriz documento-término. Para la evaluación del modelo se aplicó el método TF-IDF a la muestra de prueba, a fin de obtener su matriz documento-término para realizar la clasificación y hallar los resultados de las métricas de evaluación, donde se consiguió una exactitud de 0.81. Posteriormente, se evaluó el modelo de clasificación mediante el método K-Fold de Validación Cruzada y se clasificaron nuevos comentarios. En base a los resultados de la presente investigación se concluye que la implementación del modelo desarrollado es adecuada.

Palabras clave: Tokenización, N-gramas, Matriz dispersa, Validación Cruzada.

ABSTRACT

The purpose of this research work was to implement a logistic regression model using transformed textual data using the TF-IDF text representation method, with the aim of classifying teacher comments in guidance streamings on the "Aprendo en Casa" strategy carried out by the Ministry of Education. The analysis procedure was divided into data pre-processing, exploratory data analysis, application of the TF-IDF text representation method, model estimation and evaluation; and classification of new comments. For the pre-processing stage, the textual data of the comments were cleaned and standardized; while in the exploratory analysis, descriptive indicators of the comments of each category were obtained using n-grams. In the application of the TF-IDF text representation method, the document-term matrix was created from the training sample and the Chi-Square test was used for variable selection. In the estimation of the classification model, the final model adjusted with the training data from the document-term matrix was obtained. To evaluate the model, the TF-IDF method was applied to the test sample, in order to obtain its document-term matrix to perform the classification and find the results of the evaluation metrics, where an accuracy of 0.81 was achieved. Subsequently, the classification model was evaluated using the K-Fold Cross-Validation method and new comments were classified. Based on the results of this research, it is concluded that the implementation of the developed model is adequate.

Keywords: Tokenization, N-grams, Sparse matrix, Cross-Validation.

ÍNDICE GENERAL

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

ÍNDICE DE ANEXOS

I. INTRODUCCIÓN

Durante las últimas décadas, la cantidad de documentos de texto en formato digital ha crecido exponencialmente (Cull, 2011). A raíz de ello, surgieron diferentes métodos destinados a clasificar de manera automatizada y eficiente los documentos en categorías según el contenido del texto (Trstenjak et al., 2014).

En general, los documentos en formato digital pueden contener texto, imágenes, música, entre otros. Asimismo, cada tipo de contenido requiere métodos de clasificación especiales (Trstenjak et al., 2014). Por lo tanto, la capacidad de realizar con exactitud la labor de clasificación de textos depende de la representación del texto en los documentos a clasificar; es decir, la transformación del texto en vectores numéricos. A diferencia de la minería de datos donde se analizan los datos bien estructurados, la minería de texto trata con una colección de documentos semiestructurados, incluso no estructurados. Por tal motivo, uno de los principales temas de estudio en la minería de texto consiste en la representación del texto (Zhang et al., 2011).

La Frecuencia del término-Frecuencia Inversa del documento (TF-IDF por sus siglas en inglés) es uno de los métodos que se utiliza a menudo en el procesamiento del lenguaje natural y la minería de texto. En efecto, este método determina el peso; esto es la medida que evalúa la importancia de los términos (o palabras) en la colección de documentos (Trstenjak et al., 2014).

La regresión logística es uno de los modelos estadísticos utilizados para la clasificación de textos, dado que es adecuada para los datos textuales que se caracterizan por su gran dimensión y dispersión. En muchos trabajos de investigación, la aplicación del método de representación de texto TF-IDF en un modelo de regresión logística para la clasificación de datos textuales tuvo buenos indicadores de desempeño (Gebre et al., 2013).

Por otro lado, la pandemia del COVID-19 condujo al incremento del uso de las herramientas digitales en el mundo. En ese sentido, muchas personas optaron por utilizarlas con el objeto de comunicarse y seguir desempeñando sus actividades cotidianas, como el trabajo y el estudio. De la misma manera, gobiernos e instituciones desarrollaron nuevas estrategias para sostener su normal funcionamiento (Iivari et al., 2020).

El Ministerio de Educación del Perú, en el marco de la emergencia sanitaria, propuso "Aprendo en casa", una estrategia nacional de educación a distancia de libre acceso y sin costo que propuso experiencias de aprendizaje alineadas al currículo nacional, con la finalidad de garantizar la continuidad del servicio educativo. Esta se implementó en televisión, radio y web (Ministerio de Educación [MINEDU], 2020a, 2020b).

Sin embargo, la brecha digital de maestros y familias con niños en edad escolar obstaculizó la efectividad de la estrategia (RPP, 2020; UNESCO, 2020). Ante múltiples consultas y reclamos de la comunidad educativa sobre la estrategia, el MINEDU optó por ofrecer respuestas y orientaciones a través de las plataformas digitales. Una de ellas es su página de Facebook "Portal PerúEduca", donde realizaban *streamings* (transmisiones en vivo) en las que especialistas del ministerio aclaraban las dudas y explicaban en mayor detalle el funcionamiento de la estrategia.

En esa medida, al ser visto en vivo y en directo por miles de personas (en su mayoría docentes de educación básica), cada *streaming* generaba cientos de comentarios, entre los cuales se observaban desde saludos hasta consultas muy importantes. Cada *streaming* duraba aproximadamente una hora y se daba una vez cada dos semanas en promedio. Esto último se debió al tiempo que demoraba la preparación del contenido. De igual modo, en cada *streaming* se daban nuevas orientaciones, se reforzaban algunas explicadas anteriormente y se respondían las inquietudes más frecuentes e importantes de los espectadores en el *streaming* pasado, puesto que no se podían responder preguntas en vivo para evitar desordenar las pautas.

Por lo tanto, con el fin de responder las consultas y los reclamos más importantes en cada *streaming*, se necesitó hallar con rapidez aquellos comentarios más importantes entre los cientos que los espectadores han dejado. En ese sentido, debido a la gran cantidad de comentarios, revisar cada uno ellos manualmente tomaría bastante tiempo, por lo que una posible solución consistió en implementar un modelo de regresión logística aplicando el

método de representación de texto TF-IDF para clasificar de manera automatizada los comentarios.

Teniendo en cuenta este contexto, el presente trabajo de investigación tuvo como objetivo implementar un modelo de clasificación de datos textuales al aplicar el método de representación de texto TF-IDF en un modelo de regresión logística para clasificar de manera automatizada los comentarios de los *streamings* de orientación a docentes sobre la estrategia "Aprendo en casa".

1.1 Justificación de la investigación

La pandemia del COVID-19 afectó el desenvolvimiento habitual en la vida cotidiana, lo que obligó a las personas a adoptar nuevas maneras de continuar en el trabajo y el estudio (Iivari et al., 2020). En marzo de 2020, se interrumpió el inicio del año escolar en Perú al declararse la emergencia sanitaria a nivel nacional. En ese sentido, el Ministerio de Educación dispuso la suspensión de las clases presenciales y, a partir del 6 de abril, se dio inicio al año escolar a través de la estrategia "Aprendo en Casa" (ANDINA, 2020a, 2020b, 2020c; IPE, 2020). Sin embargo, la brecha digital es un problema presente en el país, puesto que muchos docentes y estudiantes no contaban con acceso a internet, radio o televisión, o no estaban capacitados para el uso de la tecnología (IPE, 2020; RPP, 2020; UNESCO, 2020).

Por consiguiente, el MINEDU llevó a cabo diferentes acciones para afrontar la brecha digital (RPP, 2020); una de estas consistió en los *streamings* (transmisiones en vivo) con orientaciones sobre la estrategia "Aprendo en casa". Estos *streamings* generaron comentarios, como consultas por parte de los docentes. El equipo encargado de revisarlos realizó la depuración de datos manualmente a partir de palabras clave y técnicas básicas de minería de texto para hallar las palabras y las frases más frecuentes, con el objeto de identificar los focos de las temáticas de consulta. No obstante, la clasificación final se hacía a través de revisión manual y tomaba mucho tiempo, debido a la cantidad a revisar.

En el presente trabajo de investigación se propuso la automatización del análisis de los datos mediante la implementación de un modelo de clasificación de datos textuales, por medio del método de representación de texto TF-IDF en un modelo de regresión logística con el objeto de clasificar los comentarios de los nuevos *streamings* de "Aprendo en casa". De este modo, se posibilitó un mejor análisis sobre estos y se agilizó el trabajo que realizó el equipo encargado de la revisión de consultas en los próximos *streamings*. En esa medida, el modelo permitió obtener los resultados en menor tiempo, lo que propició una mejor preparación del

contenido en el siguiente *streaming* para absolver dudas y brindar mejores orientaciones a los maestros, a fin de facilitar su labor en esta situación de emergencia sanitaria. Finalmente, este trabajo de investigación permitió difundir el conocimiento de herramientas disponibles de los campos de la minería de texto y el procesamiento del lenguaje natural.

1.2 Objetivos de la investigación

1.2.1 Objetivo general

Implementar un modelo de clasificación de regresión logística utilizando datos textuales transformados mediante el método de representación de texto TF-IDF para clasificar de manera automatizada los comentarios de los *streamings* de orientación a docentes sobre la estrategia "Aprendo en casa", a fin de mejorar la eficiencia en el equipo de la DIFODS.

1.2.2 Objetivos específicos

- Obtener indicadores descriptivos de los comentarios utilizando n-gramas para examinar la temática de las categorías "Relevante" y "No Relevante".
- Realizar la transformación de datos no estructurados (texto) a datos estructurados mediante el método de representación de texto TF-IDF para poder implementar los datos obtenidos de forma estructurada.
- Evaluar el modelo de clasificación utilizando las métricas: Exactitud, Precisión, Sensibilidad, Especificidad, Medida F1, AUC ROC (área bajo la curva ROC) y AUC PR (área bajo la curva de Precisión-Recall) para estimar el rendimiento del clasificador.
- Evaluar el modelo de clasificación a través de la Validación Cruzada para comprobar la validez de los resultados de las métricas de evaluación del modelo.

II. REVISIÓN DE LITERATURA

2.1 Antecedentes

2.1.1 Trabajos de investigación relacionados

En primer lugar, Luhn (1957) desarrolló un sistema basado en un enfoque estadístico para la lectura y búsqueda automática de texto. En este, el autor propuso la Frecuencia del Término (TF, por sus siglas en inglés) como método para hallar la ponderación de un término en un documento de texto. El sistema se probó en una colección de documentos de 1200 informes técnicos y se logró una codificación completamente automática de documentos para una futura recuperación de información.

Por su parte, Spärck (1972) planteó que la especificidad de un término (nivel de detalle con el que se representa un concepto) se debe interpretar estadísticamente, además de su aspecto semántico, como una función del uso del término, que se denominó Frecuencia inversa del documento (IDF, por sus siglas en inglés). En ese sentido, la autora argumentó que los términos deben ponderarse según la frecuencia de recopilación en documentos de modo que los menos frecuentes y más específicos sean de mayor valor que los frecuentes. El sistema de ponderación de términos resultante, conocido como TF-IDF, se probó en tres diferentes colecciones de documentos, obteniendo mejoras considerables de rendimiento.

Asimismo, Li et al. (2007) propusieron un método para la extracción de palabras clave basado en TF-IDF con un conjunto de estrategias especiales para documentos de texto chino. Dicho esto, este método se probó en una muestra aleatoria de 400 documentos de noticias chinos, donde se obtuvo una precisión, sensibilidad y medida F1 de 74.16 %, 74.19 % y 74.18 %, respectivamente; en efecto, estos resultados superaron al método de referencia en un 25 %, aproximadamente.

Por otro lado, Van Zaanen y Kanters (2010) implementaron un sistema de clasificación automática de estados de ánimo basado en letras de canciones. Este sistema obtuvo un 75 % de exactitud en los resultados; para ello, se utilizó el método de representación de texto TF-IDF con el objeto de medir la relevancia de las palabras para las diferentes clases de estados de ánimo, empleando el clasificador de aprendizaje automático "K-vecinos más cercanos".

De manera similar, Gebre et al. (2013) presentaron un sistema que permitió la identificación de la lengua nativa presente en 12100 ensayos de la Prueba de Inglés como Idioma Extranjero (TOEFL, por sus siglas en inglés) escritos en inglés por hablantes nativos de 11 idiomas distintos. Para su desarrollo, el sistema se basó en el método de representación de texto TF-IDF al utilizar tres clasificadores lineales: Máquinas de Soporte Vectorial, Regresión Logística y Perceptrones. En efecto, se obtuvo una exactitud general de 81.4 % para el conjunto de idiomas. Sin embargo, en las pruebas de validación cruzada se alcanzó una exactitud promedio de 84.55 % en la categorización de uno de los idiomas al emplear unigramas y bigramas (datos textuales compuestos por una y dos palabras, respectivamente) en el método TF-IDF. Como resultado, se halló que la Regresión Logística y las Máquinas de Soporte Vectorial tuvieron un desempeño similar en este estudio.

Aunado a esto, Pranckevičius y Marcinkevičius (2017) presentaron una comparación al emplear la frecuencia del término (TF) entre los modelos de Naïve Bayes, Random Forest, Árboles de decisión, Máquinas de Soporte Vectorial y Regresión Logística para clasificar textos breves sobre la revisión de productos de Amazon. En ese sentido, los modelos se implementaron en la plataforma informática Apache Spark; asimismo, se evaluaron según la exactitud de clasificación, el tamaño de los conjuntos de datos de entrenamiento y el número de n-gramas. En efecto, los hallazgos indicaron que la Regresión Logística logró la exactitud de clasificación más alta (mínimo 32.43 % y máximo 58.50 % en los experimentos) en comparación con los demás modelos.

En el trabajo realizado por Gaydhani et al. (2018) se planteó una solución para la detección de mensajes de odio y lenguaje ofensivo en Twitter al emplear el método de representación de texto TF-IDF en modelos de clasificación. Por tal motivo, se llevó a cabo un análisis comparativo entre los modelos de Regresión Logística, Naïve Bayes y Máquinas de Soporte Vectorial para clasificar automáticamente los tuits de Twitter en tres clases: odiosos, ofensivos y limpios. En esa medida, los resultados mostraron que la Regresión Logística tuvo el mejor desempeño, alcanzando una exactitud en clasificación del 95.6 %.

Finalmente, Li et al. (2019) desarrollaron un sistema de detección automática de propaganda en 500 artículos de noticias proporcionados por Da San Martino, Barron-Cedeno et al. (2019). En ese sentido, el sistema se basó en la Regresión Logística y, en esa medida, los autores utilizaron el método de representación de texto TF-IDF, el modelo de procesamiento de lenguaje natural BERT, la longitud de la oración, el grado de legibilidad, el léxico de emociones LIWC y el contenido enfático en el texto para clasificar automáticamente si una oración era propagandística o no. Como resultado, se obtuvo una medida F1 del 66.16 %, lo que superó el resultado del estudio base (Da San Martino, Yu et al., 2019).

2.2 Marco teórico

2.2.1 Representación de texto

En primera instancia, Chen et al. (2016) menciona que el incremento de documentos de texto en formato digital ha generado que el procesamiento de datos basado en texto sea considerado más importante. Según Zhang et al. (2011), uno de los principales temas de estudio en la minería de texto es la representación de texto, es decir, la transformación del texto en vectores numéricos, dado que es fundamental e indispensable para el procesamiento de los datos ya que la minería de texto utiliza datos semiestructurados o no estructurados.

Generalmente, la representación de texto consiste en la indexación y ponderación de términos del documento, donde se le asigna la ponderación a cada término para medir su importancia en el documento (Zhang et al., 2011). En la clasificación automática de texto, los documentos textuales suelen ser representados por vectores numéricos y luego ser asignados a categorías predefinidas a través de técnicas de aprendizaje automático supervisado, por lo que la ponderación de términos afecta directamente la exactitud de la clasificación (Chen et al., 2016).

2.2.2 Tokenización

Para representar el texto en vectores numéricos, Feldman y Sanger (2007), Hvitfeldt y Silge (2020), Silge y Robinson (2017) y Žižka et al. (2019) indican que la tokenización es el enfoque utilizado con mayor frecuencia en los sistemas de minería de texto. Consiste en dividir el texto en componentes significativos denominados tokens. Estos pueden ser caracteres individuales, secuencias de caracteres (incluyendo dígitos, puntuación, caracteres

especiales, etc.), palabras, oraciones, párrafos y estructuras más complejas contenidas en el documento de texto. La tokenización más utilizada suele ser a nivel de palabra.

Entre los tipos de tokens que se pueden obtener, también se encuentran los n-gramas. En lingüística, un n-grama es un término utilizado para referirse a una secuencia contigua de n elementos proveniente de una secuencia de texto o habla. Los elementos pueden ser fonemas, sílabas, letras o palabras. Sin embargo, los n-gramas generalmente son utilizados para denotar un grupo de n palabras (Hvitfeldt y Silge, 2020; Silge y Robinson, 2017). Los tipos más comunes de n-gramas de palabras son los siguientes:

- Unigrama: Contiene solo un elemento.
- Bigrama: Contiene dos elementos contiguos.
- Trigrama: Contiene tres elementos contiguos.

Cada tipo de n-grama permite extraer diferente nivel de detalle de los datos textuales. Los unigramas pueden mostrar las palabras individuales que se han utilizado con mayor frecuencia, pero no pueden capturar la información sobre el orden de las palabras. En contraste, los bigramas y trigramas sí capturan esa información, pero solo pueden detectar las palabras que aparecen junto a otras determinadas palabras con frecuencia. Es por ello que se suele utilizar más de un tipo de n-grama en un análisis de datos textuales (Hvitfeldt y Silge, 2020).

Al realizar la tokenización, es necesario conocer el procedimiento que se está utilizando para identificar los tokens. En la mayoría de los idiomas, es difícil establecer con claridad las reglas que definen algunos tokens específicos. Por ejemplo, las palabras en algunos idiomas, como el español o inglés, están delimitadas por espacios en blanco, por lo que el procedimiento para extraer las palabras consiste en dividir el texto donde haya espacios en blanco. Sin embargo, en otros idiomas, como el chino, no hay espacios entre palabras, lo que implicaría seguir una diferente manera de poder identificar las palabras en el texto (Hvitfeldt y Silge, 2020; Žižka et al., 2019).

En el idioma español, los signos de puntuación y la codificación de texto pueden influenciar en el proceso de tokenización. Generalmente se entiende que los signos de puntuación (coma, punto y coma, punto, comillas, signos de interrogación, signos de exclamación, apóstrofos, corchetes, etc.) separan los tokens (palabras, oraciones, párrafos). Sin embargo, pueden tener otros propósitos, como el uso del punto en abreviaciones (Dr., Srta., vol., Gral.,

etc.). Por lo que no sería adecuado considerar los signos de puntuación en los criterios para tokenizar un texto en español.

Por otro lado, los caracteres especiales generados por la codificación de texto utilizada pueden referirse a nombres de correos electrónicos, enlaces web, formatos de páginas HTML, emoticones, entre otros. Estos caracteres pueden variar debido a las diferentes posiciones que ocupan en diferentes tablas de caracteres. Probablemente el enfoque de tokenización más simple consista en dividir el texto donde haya espacios en blanco, luego de remover los signos de puntuación y los caracteres especiales; sin embargo, puede ocurrir una pérdida de información si la intención es analizar la interacción social en textos informales o coloquiales, caracterizados por presentar un uso excesivo de espacios en blanco, signos de interrogación y exclamación, emoticones, entre otros (Feldman y Sanger, 2007; Hvitfeldt y Silge, 2020; Žižka et al., 2019).

Por ejemplo, dado el siguiente texto: “Buenas noches. Excelente ponencia, profesor.”, luego de remover los signos de puntuación, algunos de los n-gramas que se pueden hallar se presentan a continuación:

Tabla 1: Unigramas, bigramas y trigramas del texto “Buenas noches. Excelente ponencia, profesor”

Unigrama	Bigrama	Trigrama
Buenas	Buenas noches	Buenas noches Excelente
noches	noches Excelente	noches Excelente ponencia
Excelente	Excelente ponencia	Excelente ponencia profesor
ponencia	ponencia profesor	
profesor		

Generalmente, n-gramas como "noches Excelente" y "ponencia profesor" no suelen ser tomados en cuenta en un análisis de datos textuales porque no encapsulan completamente una idea o mensaje.

Otro aspecto a considerar es el grado de compresión de la tokenización. Si se elige obtener pocos elementos en cada token, como es el caso de los unigramas, las tareas computacionales posteriores se realizarían de manera más rápida, pero se perdería información sobre el orden de las palabras, lo cual se puede mantener si se emplean tokens con mayor cantidad de elementos, como los bigramas y trigramas. Sin embargo, el espacio vectorial de los tokens aumentaría drásticamente y el costo computacional sería mayor (Hvitfeldt y Silge, 2020).

En trabajos de investigación, como el realizado por Gebre et al. (2013), se hallaron mejores resultados utilizando unigramas y bigramas de manera conjunta que por separado.

2.2.3 Frecuencia del término (TF)

Propuesto por Hans Peter Luhn en 1957 para la recuperación de información. La frecuencia del término, también conocida como TF por sus siglas en inglés, se refiere al número de veces que aparece un determinado término (token) en un documento. Se mide cuán importante es el término en el documento en base a su frecuencia (Chen et al., 2016; Gebre et al., 2013; Silge y Robinson, 2017).

Según Gebre et al. (2013), los términos que aparecen con mayor frecuencia que otros, pueden identificar o especificar mejor el documento. La frecuencia del término, como ponderación de un término en un documento, es más preciso y razonable que el valor binario (1 ó 0) que representa la presencia o ausencia del término en el documento, porque generalmente los términos relevantes o clave aparecen frecuentemente en el documento y corresponde asignarles mayor ponderación que a los términos raramente mencionados (Chen et al., 2016).

2.2.4 Frecuencia inversa del documento (IDF)

Zhang et al. (2011) señala que un término que aparece en muchos documentos no es un buen discriminador. Silge y Robinson (2017) explican que hay términos en un documento que aparecen frecuentemente pero que pueden no ser importantes, por ejemplo, los artículos (el, la, los, las, ...). Determinar una lista de términos frecuentes no importantes que debieran excluirse del procesamiento no es una solución adecuada porque es posible que algunos de estos términos sean más importantes en algunos documentos que en otros.

Por otra parte, Gebre et al. (2013) menciona que la importancia efectiva de un término también depende de cuán infrecuente sea el término en otros documentos. La frecuencia inversa del documento (IDF por sus siglas en inglés) disminuye la ponderación de los términos frecuentes y aumenta la ponderación de los términos no frecuentes en una colección de documentos (Silge y Robinson, 2017). Fue propuesta por Karen Spärck Jones en 1972, y se define, para cualquier término dado, de la siguiente manera:

$$idf(término) = ln\left(\frac{n_{documentos}}{n_{\text{documentos que contienen el término}}}\right) \qquad (1)$$

2.2.5 Representación de texto TF-IDF

La Frecuencia del término - Frecuencia Inversa del documento (TF-IDF por sus siglas en inglés) es un método de ponderación de términos ampliamente utilizado en los campos del procesamiento del lenguaje natural, la recuperación de información y la minería de texto (Chen et al., 2016; Gebre et al., 2013; Trstenjak et al., 2014).

Fue propuesto también por Karen Spärck Jones en 1972. El método combina las ponderaciones de TF e IDF multiplicando ambos elementos para medir la importancia de un término en una colección de documentos. TF da más peso a un término frecuente en un documento e IDF reduce el peso si el término aparece en muchos documentos, de esta manera se controla el problema de la presencia de los términos frecuentes no importantes (Bafna et al., 2016; Gebre et al., 2013; Silge y Robinson, 2017; Zhang et al., 2011).

Para realizar esta representación, se construye la matriz documento-término, la cual describe una colección de documentos, donde cada fila representa un documento, cada columna representa un término y cada valor es la ponderación TF-IDF de un término en un determinado documento. Se trata de una matriz dispersa, es decir, la mayoría de los valores de la matriz es cero, debido a que la mayoría de los documentos no contienen la mayoría de los términos (Hvitfeldt y Silge, 2020; Silge y Robinson, 2017).

Dado que los documentos no suelen presentar la misma longitud (cantidad de términos), si un término aparece muchas más veces en documentos largos que en documentos más cortos, tendría un impacto en la ponderación de términos. Por lo tanto, es necesario estandarizar la frecuencia de cada término (TF) dividiéndola por la cantidad total de términos en el documento (Gebre et al., 2013; Hvitfeldt y Silge, 2020). De manera ilustrativa, se tiene un ejemplo en el Anexo 1.

2.2.6 Otros métodos de representación de texto

Word Embeddings

Propuesto en Mikolov et al. (2013), es un conjunto de técnicas de representación de palabras como vectores numéricos reales en un espacio vectorial para el aprendizaje de características y modelado de lenguaje natural. En Word Embeddings, cada palabra ocupa un vector dentro del espacio. Los principales métodos para generar estos espacios vectoriales son: Redes neuronales, Reducción de dimensionalidad en la matriz de co-ocurrencia de palabras y Modelos probabilísticos.

Estas técnicas son utilizadas en los algoritmos de aprendizaje para lograr mejor rendimiento en tareas de procesamiento de lenguaje natural en tareas como análisis sintáctico y análisis de sentimientos.

Uno de los principales beneficios de la aplicación de estas técnicas, es que, debido a que las palabras se encuentran en espacios vectoriales, se pueden realizar operaciones vectoriales y, por ende, se pueden realizar operaciones semánticas como cercanía de palabras. Esto permite medir la similitud lingüística y semántica de palabras.

Las principales limitaciones de Word Embeddings son la homonimia (coincidencia en la escritura de dos palabras que tienen distinto significado) y la polisemia (cuando una misma palabra tiene varios significados), pues con esta técnica las palabras son clasificadas en una sola representación (un solo vector en el espacio semántico), ocasionando que solo puedan tener un significado. Actualmente, entre las principales técnicas de obtención de Word Embeddings se encuentran Word2Vec y GloVe.

- **Word2vec**

  Es una técnica para el procesamiento del lenguaje natural. La técnica de word2vec utiliza un modelo de red neuronal para aprender asociaciones de palabras de un corpus de texto (colección grande de textos). Una vez entrenado el corpus, dicho modelo puede detectar palabras sinónimas o sugerir palabras adicionales para una oración parcial. Como su nombre lo indica, word2vec representa cada palabra distinta con una lista particular de números denominada vector dentro de un espacio. Estos vectores se eligen cuidadosamente de modo que una función matemática simple (la similitud del coseno entre los vectores) indica el nivel de similitud semántica entre las palabras representadas por esos vectores (Mikolov et al., 2013). Esta técnica es utilizada para producir word embeddings en un menor tiempo y comprende redes neuronales superficiales de dos capas que están capacitadas para reconstruir contextos lingüísticos de palabras. Word2vec toma como entrada un gran corpus de texto y produce un espacio vectorial, típicamente de varios cientos de dimensiones, y a cada palabra única del corpus se le asigna un vector correspondiente en el espacio. Los vectores de palabras se colocan en el espacio vectorial de manera que las palabras que comparten contextos comunes en el corpus se ubican cerca unas de otras en el espacio.

- **GloVe**

  Es un algoritmo de aprendizaje no supervisado para obtener representaciones vectoriales de palabras (Word embeddings). El entrenamiento se realiza en espacios vectoriales globales de coocurrencia palabra-palabra agregadas de un corpus, y las representaciones resultantes muestran interesantes sub-estructuras lineales del espacio vectorial de palabras (Pennington et al., 2014).

ELMo

Es un tipo de representación profunda de palabras contextualizadas que modela las características complejas del uso de palabras y la manera en cómo se usan en diferentes contextos lingüísticos. Esto se logra a través de vectores con funciones de aprendizaje que contienen estados internos (celdas LSTM) y modelos de lenguaje bidireccionales profundos (biLM). Estas representaciones de palabras se pueden agregar fácilmente a modelos existentes y mejorar significativamente los resultados para la respuesta a preguntas, análisis de sentimientos y clasificación de textos (Peters et al., 2018).

Si bien este método no es una extensión de las técnicas de word embeddings, usa celdas LSTM bidireccionales pre-entrenadas por un corpus para observar una frase en su totalidad y asignar un embedding a cada una de las palabras.

Actualmente ELMo es usado para representaciones de texto grandes debido a que soluciona problemas de polisemia y homonimia, dando más de un significado a una palabra en el espacio vectorial.

2.2.7 Selección de Variables

Generalmente, la cantidad de términos diferentes entre documentos es grande, llegando a ser cientos de miles en grandes colecciones de documentos, demandando de esta forma un gran costo computacional debido al cálculo de la ponderación de todos los términos en la matriz documento-término. Sin embargo, la mayoría de estos términos no contribuyen en la clasificación de datos textuales. Por lo tanto, pueden eliminarse sin afectar la clasificación e incluso mejorar el rendimiento debido a la reducción de ruido (Feldman y Sanger, 2007).

El procedimiento que consiste en eliminar los términos irrelevantes se denomina Selección de Características o Variables. Muchos sistemas de clasificación de datos textuales eliminan los términos comunes no relevantes, llamados *stopwords*, como paso inicial. Posteriormente, se puede considerar utilizar métodos más sofisticados, como los métodos de filtro, los cuales

son adecuados para problemas a gran escala debido a sus bajos costos computacionales y a la capacidad de reducir la dimensionalidad en un factor de 100 sin perjudicar el desempeño de la clasificación, o incluso mejorándola (Feldman y Sanger, 2007; Žižka et al., 2019). Un método popular para la Selección de Variables es la Prueba Chi Cuadrado. En Estadística, este método es utilizado para medir la independencia entre dos variables categóricas en una muestra aleatoria a través de una tabla de contingencia con frecuencias. En la Selección de Variables para la Minería de texto, este método se utiliza para medir la dependencia entre un término y una categoría, ambas variables con valores "sí" o "no" para representar su ocurrencia, siendo la hipótesis nula que las dos variables son completamente independientes entre sí (Kumar y Paul, 2016; Žižka et al., 2019). El valor del estadístico χ^2 se puede calcular de la siguiente manera:

Si:

- x es un término
- y es una categoría
- A_{xy} es el número de veces que x e y ocurren simultáneamente
- B_x es el número de veces que x ocurre sin y
- C_y es el número de veces que y ocurre sin x
- D_0 es el número de veces que x e y no ocurren
- n es el número de documentos

Entonces:

$$\chi^2 = \frac{n * (A_{xy} * D_0 - C_y * B_x)^2}{(A_{xy} + C_y) * (B_x + D_0) * (A_{xy} + B_x) * (C_y + D_0)} \qquad (2)$$

Por ejemplo, si se tuviese 14 comentarios, la variable x fuese la importancia del término “competencia” con valores TF-IDF y la variable y sea “categoría” con valores y = 0 (categoría no relevante) e y = 1 (categoría relevante), el cálculo del valor de χ^2 sería el siguiente:

Tabla 2: Cálculo de los componentes de la fórmula de la obtención de χ^2 para el término "competencia" (Ejemplo)

Nro. de comentario	x "competencia"	y "categoría"	Ocurre A_{xy} x = Sí y = Sí	Ocurre B_x x = Sí y = No	Ocurre C_y x = No y = Sí	Ocurre D_0 x = No y = No
1	0.23 (Sí)	1 (Sí)	✓			
2	0.42 (Sí)	1 (Sí)	✓			
3	0.056 (Sí)	1 (Sí)	✓			
4	0 (No)	0 (No)				✓
5	0 (No)	0 (No)				✓
6	0 (No)	0 (No)				✓
7	0.78 (Sí)	1 (Sí)	✓			
8	0 (No)	1 (Sí)			✓	
9	0.015 (Sí)	0 (No)		✓		
10	0.46 (Sí)	1 (Sí)	✓			
11	0.52 (Sí)	1 (Sí)	✓			
12	0 (No)	0 (No)				✓
13	0 (No)	1 (Sí)			✓	
14	0 (No)	1 (Sí)			✓	

En la Tabla 2, los valores de ambas variables se reemplazan con valores "Sí" o "No" para representar su ocurrencia en cada observación (comentario). Por ejemplo, en el comentario 9 el término "competencia" tiene un valor TF-IDF mayor que cero, es decir, sí aparece textualmente en el comentario, por consiguiente, corresponde asignarle el valor "Sí"; mientras que en la variable "categoría" se indica que este comentario se clasificó como No relevante, es decir, no ocurrió el evento de interés, por lo tanto, se le asigna el valor "No".

Posteriormente, se realiza el conteo de las observaciones que coincidan con lo que representa cada componente de la fórmula, resultando en $A_{xy} = 6$, $B_x = 1$, $C_y = 3$ y $D_0 = 4$. Por lo tanto:

$$\chi^2 = \frac{14 * (6 * 4 - 3 * 1)^2}{(6+3) * (1+4) * (6+1) * (3+4)} = \frac{6174}{2205} = 2.8 \qquad (3)$$

Los valores de χ^2 permiten identificar los términos útiles para la clasificación de texto, como se describe a continuación:

El valor de χ^2 igual a cero indica que el término y la categoría son independientes, mientras que valores de χ^2 altos indican la existencia de alguna dependencia entre la categoría y el término, cuanto más alto sea el valor de χ^2, más estrecha será la relación entre las dos variables. Los términos que tienen valores altos de χ^2 son útiles para la clasificación de texto

porque tienen la capacidad de distinguir las categorías, por lo tanto, la ocurrencia de estos términos puede determinar que una categoría sea más o menos probable (Kumar y Paul, 2016; Žižka et al., 2019).

En la clasificación de texto no se necesita hacer afirmaciones sobre la independencia estadística; de efectuarlas, tener en cuenta que el valor calculado de χ^2 se distribuye como una χ^2 con un grado de libertad, por lo tanto, se debe utilizar la corrección de Yates, la cual no se encuentra disponible en muchos *softwares*. A pesar de ello, en la clasificación de texto no se considera tan problemático incluir o eliminar por error algunos términos adicionales en el conjunto de las variables seleccionadas. Es por tal motivo que varios expertos recomiendan enfocarse en la utilidad de los términos (Manning et al., 2009; Žižka et al., 2019).

Otros métodos de filtro comúnmente utilizados para la Selección de Variables en la Minería de texto son: Información Mutua, Ganancia de Información, Separación Bi-Normal, Entropía Cruzada Esperada, Índice de Gini, Fuerza del Término, Clasificación Basada en la Entropía y Contribución del Término (Žižka et al., 2019).

2.2.8 Regresión Logística

Según Cramer (2002), la función de regresión logística fue desarrollada como un modelo de crecimiento de población llamado “Logístico” por Pierre-François Verhulst entre 1830 y 1840. Posteriormente, fue desarrollada independientemente en 1883 como un modelo de autocatálisis y finalmente fue redescubierta en 1920 por Raymond Pearl y Lowell Reed.

Según Acosta Zúñiga (2020), este modelo analiza la relación entre múltiples variables independientes que pueden ser categóricas o numéricas, y una variable dependiente categórica. Para establecer esta relación, se utiliza la función logística para modelar una variable dependiente, siendo esta binaria o de una extensión más compleja con más categorías. Asimismo, según la cantidad de categorías de la variable dependiente, el modelo se considerará de regresión logística binaria, donde la variable dependiente tiene únicamente dos categorías, o regresión logística multinomial, donde la variable dependiente tiene más de dos categorías. En los casos de regresión logística multinomial, esta variable puede ser nominal u ordinal.

El modelo logístico

Sea Z una variable aleatoria respuesta binaria o dicotómica, siendo Z = 1 si el resultado es un éxito y Z = 0 si el resultado es un fracaso, donde "éxito" y "fracaso" se refieren a los términos genéricos de las dos categorías, con probabilidades P(Z = 1) = π y P(Z = 0) = 1 – π, el valor de la variable respuesta dada la variable explicativa X puede expresarse como Z = π + ε, donde ε puede asumir uno de dos posibles valores: Si Z = 1 entonces ε = 1 – π, y si Z = 0 entonces ε = -π.

Si existen n variables aleatorias independientes $Z_1, \ldots, Z_n$ con $P(Z_j = 1) = \pi_j$, entonces su probabilidad conjunta es

$$\prod_{j=1}^{n} \pi_j^{Z_j} (1-\pi_j)^{1-Z_j} = exp\left[\sum_{j=1}^{n} Z_j \log\left(\frac{\pi_j}{1-\pi_j}\right) + \sum_{j=1}^{n} \log(1-\pi_j)\right] \qquad (4)$$

por lo que pertenece a la familia exponencial. Si todos los π_j son iguales y se define que

$$Y = \sum_{j=1}^{n} Z_j \qquad (5)$$

entonces la variable aleatoria Y es el número de éxitos en n 'ensayos', que presenta una distribución Binomial (n, π), de modo que:

$$P(Y = y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \qquad y = 0, 1, \ldots, n \qquad (6)$$

La distribución Binomial es la distribución estadística sobre la cual se basa el análisis de regresión logística.

El modelo de regresión logística permite estimar la probabilidad de ocurrencia de un evento específico en función de un conjunto de variables. Es un modelo de clasificación que forma parte de los Modelos Lineales Generalizados al pertenecer a la familia exponencial, y es ampliamente utilizado para analizar datos que comprenden respuestas binarias o binomiales y varias variables explicativas.

Sea p el número de variables independientes explicativas, la forma específica del modelo de regresión logística es:

$$\pi = \frac{exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)}{1 + exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p)} \qquad (7)$$

Equivalente a:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \qquad (8)$$

Donde la combinación lineal $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$ representa al predictor lineal conformado por los parámetros $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ y las variables explicativas (que pueden ser covariables o factores) $X_1, X_2, X_3, \ldots, X_p$.

El término $\log\left(\frac{\pi}{1-\pi}\right)$ generalmente recibe el nombre de función de enlace logit o logística, que permite relacionar π con el predictor lineal y garantizar que π se encuentre limitado en el intervalo de cero a uno.

Para ajustar el modelo de regresión logística a un conjunto de datos es necesario estimar los valores de los parámetros desconocidos β. El método de Máxima Verosimilitud estima estos valores a través de la función de verosimilitud, la cual expresa la probabilidad de obtener el conjunto de datos observado como una función de los parámetros desconocidos, donde los estimadores resultantes serán aquellos valores que maximicen esta función, es decir, aquellos que concuerden más con los datos observados. El procedimiento de estimación se desarrolla utilizando el logaritmo de la función de verosimilitud, también llamado función de log-verosimilitud. Para el caso general de N variables aleatorias independientes $Y_1, Y_2, \ldots, Y_N$ correspondientes a los números de éxitos en N subgrupos o estratos diferentes, si $Y_i \sim$ Binomial (n_i, π_i), la función de log-verosimilitud es

$$l(\pi_1, \ldots, \pi_N; y_1, \ldots, y_N) = \sum_{i=1}^{N} \left[y_i \, log \, \pi_i + (n_i - y_i) \log(1 - \pi_i) + \log \binom{n_i}{y_i} \right] \qquad (9)$$

Esta función se deriva matemáticamente con respecto a los p + 1 parámetros y cada expresión resultante se iguala a cero, de modo que se tienen p + 1 ecuaciones, las cuales son resueltas

utilizando métodos iterativos especiales debido a que las expresiones en las ecuaciones no son lineales.

Interpretación del modelo

Si se denomina *odds* a la probabilidad que el evento de interés ocurra respecto a que no ocurra, denotado por:

$$\text{Odds} = \frac{P(Y=1)}{P(Y=0)} = \frac{P(Y=1)}{1-P(Y=1)} = \frac{\pi}{1-\pi} = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p) \quad (10)$$

Entonces, Hosmer y Lemeshow (2000) indican que para el caso de una variable explicativa dicotómica, se denomina *odds ratio* a la relación o razón entre los *odds* de $x = 1$ y los *odds* de $x = 0$, dada por la ecuación

$$\text{OR} = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]}$$

$$\text{OR} = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)/\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)/\left(\frac{1}{1+e^{\beta_0}}\right)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1} \quad (11)$$

El *odds ratio* es una medida de asociación que indica cuánto más probable (o improbable) es que el evento de interés esté presente cuando $x = 1$ con respecto a que $x = 0$. Por ejemplo, si Y representa la presencia o ausencia de enfermedad cardíaca, y si X representa si la persona es fumadora, entonces OR = 2 estima que la enfermedad cardíaca tiene el doble de probabilidades de ocurrir entre los fumadores que entre los no fumadores en la población de estudio.

Para el caso de una variable explicativa continua, Myers et al. (2010) explica que, si el predictor lineal contiene solo este regresor, el valor ajustado del predictor lineal en un valor particular de X, X_i, es $b_0 + b_1X_i$. Mientras que el valor ajustado para $X_i + 1$ es $b_0 + b_1(X_i+1)$. La diferencia entre estos valores es b_1, es decir,

$$b_0 + b_1(X_i+1) - (b_0 + b_1X_i) = \ln(\text{odds}_{X_{i+1}}) - \ln(\text{odds}_{X_i})$$

$$= ln\left(\frac{odds_{Xi+1}}{odds_{Xi}}\right) = b_1 \quad (12)$$

Utilizando el antilogaritmo se obtiene el *odds ratio* estimado:

$$\widehat{OR} = \frac{odds_{Xi+1}}{odds_{Xi}} = e^{b_1} \qquad (13)$$

El *odds ratio* estimado puede ser interpretado como el aumento estimado en las probabilidades de éxito asociado con un cambio de una unidad en el valor de la variable explicativa. En general, el aumento estimado en el *odds ratio* asociado con un cambio de d unidades en la variable explicativa es $\exp(db_1)$.

La interpretación de los coeficientes en el modelo de regresión logística múltiple es similar a la del caso en el que el predictor lineal contiene solo un regresor, dado que el *odds ratio* para el regresor X_j es $\exp(\beta_j)$, suponiendo que todas las demás variables explicativas son constantes.

Problemas que pueden ocurrir en un modelo

El sobreajuste y sub-ajuste son problemas comunes en la estimación de un modelo. El primero ocurre cuando el modelo se ajusta tan bien al conjunto de datos de entrenamiento que no generaliza y se vuelve sensible a las particularidades de los datos de entrenamiento. Se evidencia cuando el desempeño del modelo disminuye al ser evaluado en un conjunto de datos que no ha sido previamente visto. Este modelo posee un exceso de parámetros, por lo que los conjuntos pequeños de datos tienden a experimentar sobreajuste con mayor frecuencia en comparación a conjuntos grandes de datos.

Por el contrario, el sub-ajuste ocurre cuando se emplea un modelo demasiado simple para representar un conjunto específico de datos, resultando en un modelo sin capacidad de capturar la variabilidad de los datos (Khalaf y Zaman, 2015).

2.2.9 Otros modelos de clasificación de datos textuales

Naïve Bayes

Se trata de un clasificador probabilístico de aprendizaje automático supervisado, basado en el Teorema de Bayes. Asigna una clase a un documento calculando el producto de las probabilidades a posteriori de los atributos observados en los datos de entrenamiento por la

probabilidad a priori de que ocurra una clase por elemento textual clasificado, bajo el supuesto que todas las variables son importantes e independientes entre sí, de modo que el cálculo sea más fácil y aplicable en la práctica (Jurafsky y Martin, 2021; Kumar y Paul, 2016; Žižka et al., 2019).

Naïve Bayes es muy popular en la clasificación de datos textuales debido a que no necesita un gran conjunto de datos de entrenamiento, ni tantos ciclos computacionales, a diferencia de otros clasificadores avanzados y sofisticados; además, puede trabajar con una gran cantidad de variables y cuando la dimensionalidad de los datos es alta (Kumar y Paul, 2016; Žižka et al., 2019).

Máquinas de Soporte Vectorial

Las Máquinas de Soporte Vectorial (SVM por sus siglas en inglés) son algoritmos de aprendizaje funcional que construyen un modelo de clasificación mediante entrenamiento o aprendizaje supervisado (Chen et al., 2016).

Fue introducido por Vapnik en 1995 para resolver problemas de reconocimiento de patrones de dos clases. El método originalmente se define sobre un espacio vectorial donde un clasificador SVM encuentra un hiperplano (superficie de decisión) que separa los datos en dos clases con un margen máximo. Actualmente se emplea el enfoque "Uno contra el resto" para la categorización multi-clase de los datos (Gebre et al., 2013; Zhang et al., 2011).

Las Máquinas de Soporte Vectorial pueden utilizarse en tareas de regresión y clasificación. Son adecuadas para la clasificación de texto porque pueden trabajar con datos dispersos y de gran dimensionalidad, como lo son los datos textuales; y pueden brindar un buen rendimiento (Chen et al., 2016; Gebre et al., 2013; Hvitfeldt y Silge, 2020).

K-vecinos más cercanos

Es un algoritmo de aprendizaje que consiste en clasificar directamente un documento comparando su similitud con los documentos de las clases predefinidas, a través de la distancia euclidiana, definida como la distancia entre dos puntos en el espacio euclidiano. Este algoritmo plantea que es posible clasificar documentos en el espacio euclidiano como puntos. La menor distancia euclidiana entre los documentos indica su mayor similitud, por lo tanto, los documentos completamente iguales tendrán una distancia igual a cero (Chen et al., 2016; Trstenjak et al., 2014).

No requiere de datos de entrenamiento para realizar la clasificación. Sin embargo, éstos pueden ser usados durante la fase de prueba. El algoritmo determina qué documentos de las clases predefinidas se comparan con cada nuevo documento. El número de "vecinos más cercanos", K, indica el número de documentos requeridos de la colección más cercanos al documento seleccionado. Se optimiza mediante la evaluación del rendimiento en un conjunto de datos específico (Chen et al., 2016; Trstenjak et al., 2014).

2.2.10 Métricas de evaluación del modelo

Chicco y Jurman (2020), Tharwat (2021) y Žižka et al. (2019) indican que existen varias formas para evaluar la calidad de un clasificador. La mayoría son métricas escalares y algunas son métodos gráficos. A continuación, se describen las más utilizadas.

Matriz de confusión

Describe el rendimiento de un modelo de clasificación mediante una tabla cruzada de las clases observadas y predichas para los datos. Las celdas diagonales representan las predicciones correctas, mientras que el resto de las celdas muestran las predicciones incorrectas (Chicco y Jurman, 2020; Kuhn y Johnson, 2013; Tharwat, 2021; Žižka et al., 2019). En el caso de una clasificación binaria, la matriz de confusión está conformada por cuatro entradas, tal como se observa en la siguiente tabla.

Tabla 3: Matriz de confusión para una clasificación binaria

	Clase A predicha	Clase B predicha
Clase A observada	Verdadero positivo	Falso negativo
Clase B observada	Falso positivo	Verdadero negativo

- Verdadero Positivo: Indica el número de observaciones clasificadas correctamente a la clase de interés.
- Falso Positivo: También conocido como Error Tipo I, indica la cantidad de observaciones incorrectamente asignadas a la clase de interés.
- Verdadero Negativo: Se refiere al número de observaciones correctamente clasificadas como no pertenecientes a la clase de interés.
- Falso Negativo: También conocido como Error Tipo II, indica el número de observaciones incorrectamente clasificadas como no pertenecientes a la clase de interés.

Exactitud

Analizar cada entrada de la matriz de confusión puede tomar mucho tiempo, por lo tanto, algunas métricas fueron definidas utilizando la información que proporciona la matriz de confusión con el objetivo de describir rápidamente la calidad de una predicción (Chicco y Jurman, 2020).

Muchos investigadores consideran que la métrica más razonable de usar es la exactitud (llamada *accuracy* en inglés) debido a que emplea la información de todas las entradas de la matriz de confusión. Está definida como la relación entre el número de predicciones correctas y el número total de predicciones (Chicco y Jurman, 2020; Tharwat, 2021; Žižka et al., 2019), representada por la siguiente fórmula:

$$\text{Exactitud} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{VN} + \text{FP} + \text{FN}} \qquad (14)$$

donde:

VP = Verdadero Positivo
VN = Verdadero Negativo
FP = Falso Positivo
FN = Falso Negativo

De esta forma, el rango del valor de la exactitud se encuentra en el intervalo [0, 1], donde el valor de 0 representa una clasificación completamente errónea, y el valor de 1 representa una clasificación completamente correcta (Chicco y Jurman, 2020).

La exactitud es una de las métricas más utilizadas para evaluar el desempeño de una clasificación. Sin embargo, no es adecuada cuando el conjunto de datos es desbalanceado (la cantidad de observaciones en una clase es mucho mayor que la cantidad en otras clases) porque proporciona una estimación demasiado optimista de la capacidad del clasificador en la clase mayoritaria (Chicco y Jurman, 2020; Tharwat, 2021; Wu et al., 2018).

Precisión

Es la proporción de las observaciones asignadas a la clase de interés que fueron clasificadas correctamente (Wu et al., 2018; Žižka et al., 2019), dada por la fórmula:

$$\text{Precisión} = \frac{\text{Verdadero positivo}}{\text{Verdadero positivo} + \text{Falso positivo}} \qquad (15)$$

Sensibilidad

También llamada tasa de verdaderos positivos, tasa de aciertos o recall. Es la proporción de las observaciones pertenecientes a la clase de interés que fueron clasificadas correctamente (Kuhn y Johnson, 2013; Tharwat, 2021; Žižka et al., 2019), representada por la fórmula:

$$\text{Sensibilidad} = \frac{\text{Verdadero positivo}}{\text{Verdadero positivo} + \text{Falso negativo}} \qquad (16)$$

Especificidad

También conocida como la tasa de verdaderos negativos o el recall inverso. Es la proporción de las observaciones no pertenecientes a la clase de interés que fueron clasificadas correctamente (Kuhn y Johnson, 2013; Tharwat, 2021), expresada como:

$$\text{Especificidad} = \frac{\text{Verdadero negativo}}{\text{Verdadero negativo} + \text{Falso positivo}} \qquad (17)$$

De esta manera, puede considerarse a la sensibilidad como la exactitud para las observaciones correspondientes a la clase de interés, y a la especificidad como la exactitud para las observaciones que no forman parte de la clase de interés (Kuhn y Johnson, 2013; Tharwat, 2021).

Medida F1

También conocida como Puntuación F1, es el promedio armónico de la precisión y la sensibilidad (Chicco y Jurman, 2020; Wu et al., 2018; Žižka et al., 2019); proveniente de la Medida F, definida como el promedio armónico ponderado de la precisión y la sensibilidad, cuya fórmula es:

$$\text{Medida F} = \frac{(\beta^2 + 1) * \text{precisión} * \text{sensibilidad}}{\beta^2 * \text{precisión} + \text{sensibilidad}} \qquad (18)$$

donde β es el factor de ponderación de la precisión con respecto a la sensibilidad

Se puede observar que, si β aumenta, la precisión tendría mayor importancia (ponderación) que la sensibilidad. De esta manera, la Medida F es denominada Medida o Puntuación F1

cuando la precisión y la sensibilidad tienen la misma importancia, es decir, β = 1 (Žižka et al., 2019). Por lo tanto, la fórmula de la Medida o Puntuación F1 es la siguiente:

$$\text{Medida F1} = \frac{2 * \text{precisión} * \text{sensibilidad}}{\text{precisión} + \text{sensibilidad}} = \frac{2 * \text{VP}}{2 * \text{VP} + \text{FP} + \text{FN}} \qquad (19)$$

El valor de la medida F1 oscila en el intervalo [0,1], donde el valor de 0 indica que todas las observaciones pertenecientes a la clase de interés han sido clasificadas incorrectamente, es decir, VP = 0; mientras que el valor de 1 representa una clasificación perfecta, es decir, FP = FN = 0. Por lo tanto, un valor cercano a 1 indicaría un alto rendimiento de clasificación (Chicco y Jurman, 2020; Tharwat, 2021).

Curva ROC (*Receiver Operating Characteristic*)

La sensibilidad y la especificidad dependen del punto de corte de probabilidad que se haya decidido utilizar para la clasificación de las observaciones (por defecto se utiliza el valor de 0.5). La curva ROC representa la tasa de verdaderos positivos (sensibilidad) y la tasa de falsos positivos (1 - especificidad) en un gráfico bidimensional para diferentes puntos de corte de probabilidad posibles, en el que se puede hallar el punto de corte óptimo que maximice adecuadamente el equilibrio entre sensibilidad y especificidad, es decir, un balance entre beneficio y costo (Aldas y Jimenez, 2017; Kuhn y Johnson, 2013; Myers et al., 2010; Tharwat, 2021).

En el gráfico, el punto en la coordenada (0,1) representa una clasificación perfecta, de manera que, mientras más se aproxime una curva ROC a ese punto, se tratará de un mejor clasificador. Mientras que, si se aproxima a la línea diagonal, representará a un clasificador realizando predicciones aleatorias, similar a lanzar una moneda al aire para determinar la clase de cada observación. Una curva ROC que se encuentre por debajo de la línea diagonal representaría a un clasificador completamente deficiente (Aldas y Jimenez, 2017; Kuhn y Johnson, 2013; Tharwat, 2021).

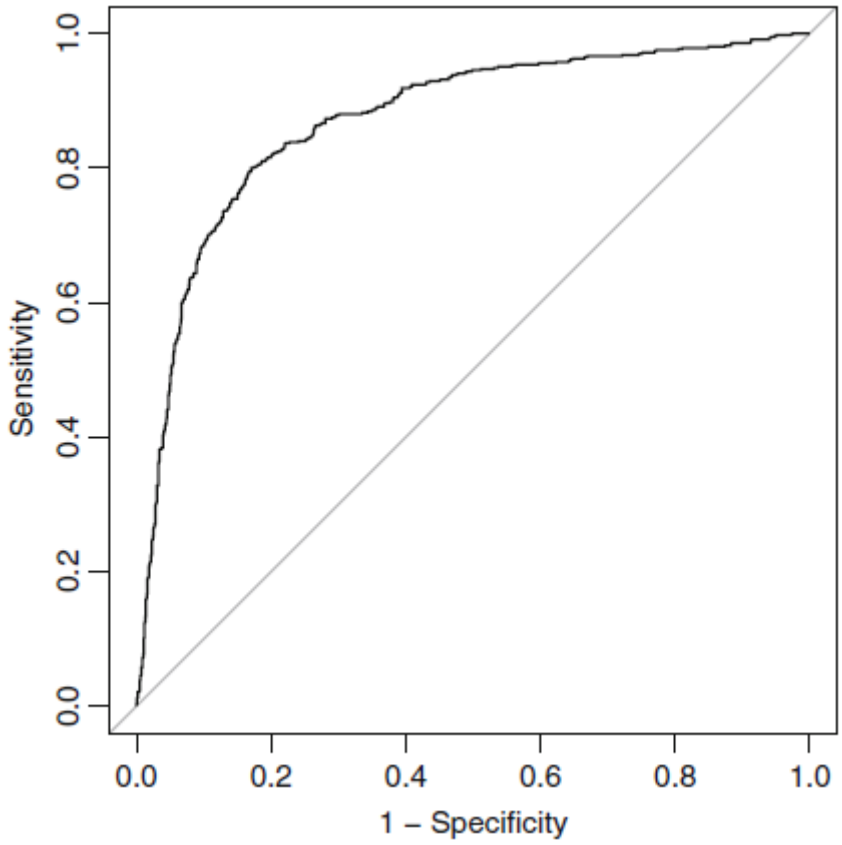


Figura 1: Curva ROC

Nota. Tomado de *Applied Predictive Modeling,* por M. Kuhn y K. Johnson, 2013, Springer.

El área bajo la curva (AUC por sus siglas en inglés) mide la capacidad del modelo para discriminar entre las observaciones pertenecientes a la clase de interés y las que no. El AUC puede calcularse sumando el área de todos los trapezoides que conforman el área bajo la curva ROC. El valor del AUC varía entre 0 y 1, sin embargo, no existe ningún clasificador realista que presente un AUC inferior a 0.5, ya que este valor hace referencia a un modelo deficiente, mientras que un valor igual a 1 haría referencia a un modelo perfecto que siempre asigna mayores probabilidades de ocurrencia a las observaciones donde el evento de interés realmente ocurre que cuando no es así (Aldas y Jimenez, 2017; Kuhn y Johnson, 2013; Myers et al., 2010; Tharwat, 2021). Hosmer y Lemeshow (2000) sugieren seguir los siguientes criterios como regla general:

- Si ROC = 0.5, se considera que no hay discriminación (similar a lanzar una moneda).
- Si $0.7 \leq ROC < 0.8$, se considera discriminación aceptable.
- Si $0.8 \leq ROC < 0.9$, se considera discriminación excelente.
- Si $ROC \geq 0.9$, se considera discriminación sobresaliente.

La sensibilidad y la especificidad pueden evaluar el rendimiento de una clasificación con datos desbalanceados, por lo que la curva ROC también tiene la cualidad de no ser sensible a los datos desbalanceados (Kuhn y Johnson, 2013; Tharwat, 2021; Wu et al., 2018).

Curva de Precisión - Recall

Representa la relación entre la precisión y la sensibilidad (recall) en un gráfico bidimensional para diferentes puntos de corte de probabilidad. La curva PR suele ser una curva en zigzag. En el gráfico, el número de observaciones correspondientes a cada clase (n_A y n_B) define la línea de base del nivel de rendimiento del clasificador de la siguiente manera: la línea horizontal que pasa por la ordenada $n_A/(n_A+n_B)$ representa a un clasificador que realiza predicciones aleatorias.

Una curva PR que se halle por debajo de esa línea significa que presenta un bajo desempeño, mientras que, si se encuentra por encima de la línea representará un buen desempeño. El punto en la coordenada (1,1) representa una clasificación perfecta, esto quiere decir que, cuanto más se aproxime una curva PR a ese punto, mejor será el rendimiento de la clasificación (Tharwat, 2021).

El área bajo la curva PR (AUC PR) puede calcularse de la misma manera que el AUC ROC, esto es, sumando el área de todos los trapezoides bajo la curva. Asimismo, el valor del AUC PR para un clasificador perfecto es igual a uno (Tharwat, 2021). La precisión es una métrica sensible a los datos desbalanceados, por lo tanto, la curva de Precisión - Recall también comparte esta propiedad debido al eje de precisión que la compone. A pesar de ello, muchos investigadores consideran que la curva de Precisión - Recall es más informativa que la curva ROC (Chicco y Jurman, 2020; Tharwat, 2021).

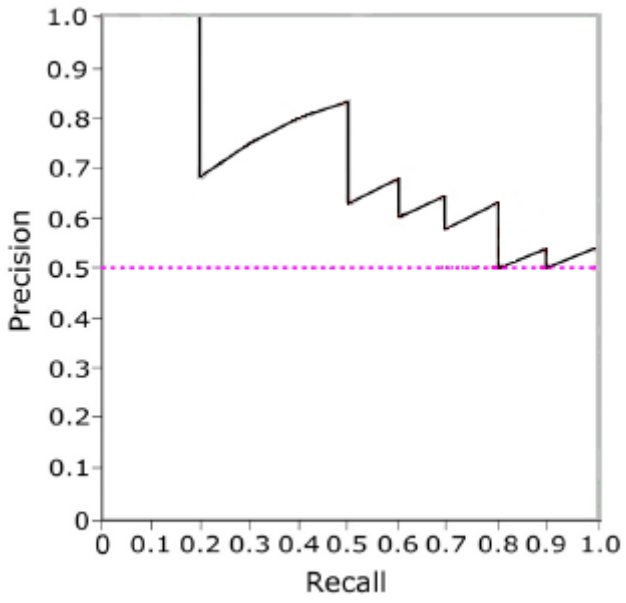


Figura 2: Curva de Precisión – Recall

Nota. Tomado de "Classification assessment methods", por A. Tharwat, 2021, en *Applied Computing and Informatics, 17*(1), 168-192.

Las métricas mencionadas anteriormente pueden calcularse para cada clase para evaluar los resultados en un problema de clasificación. Sin embargo, Chen et al. (2016) y Kuhn y Johnson (2013) explican que existen dos métricas que pueden representar el rendimiento general de la clasificación, las cuales se denominan macro-promedio (o promedio simple) y micro-promedio (o promedio ponderado). El primero se basa en la media aritmética, es decir, promedia los resultados obtenidos de una métrica para cada clase, otorgando el mismo peso a cada resultado de la métrica. Mientras que la segunda métrica tiene en cuenta cada elemento clasificado, por lo tanto, cada resultado de la métrica tendrá un peso que dependerá del tamaño de la clase.

Por ejemplo, si la precisión para la clase A es 0.5 y para la clase B es 0.9, donde $n_A = 20$ y $n_B = 100$, entonces el macro-promedio o promedio simple de la precisión es igual a 0.7, mientras que el micro-promedio o promedio ponderado de la precisión es igual a

$$0.5 * \frac{20}{(20 + 100)} + 0.9 * \frac{100}{(20 + 100)} = 0.83 \qquad (20)$$

2.2.11 Validación Cruzada

Un buen clasificador debe ser capaz de extraer y aprender los patrones o relaciones subyacentes entre los atributos dependientes e independientes. La validación cruzada es una técnica para validar este aspecto en el rendimiento de un modelo (Kumar y Paul, 2016). Se explican dos métodos de validación cruzada:

Método "K-Fold"

El conjunto de datos se divide aleatoriamente en k partes o sub-muestras (*folds*) de tamaño similar y se ejecuta el proceso de entrenamiento y prueba del modelo k veces, utilizando en cada iteración una sub-muestra diferente como muestra de prueba y las (k-1) restantes como muestra de entrenamiento. Las métricas de evaluación del modelo se calculan para cada iteración y los resultados se promedian al final (Feldman y Sanger, 2007; Kumar y Paul, 2016; Žižka et al., 2019).

Método "Dejar uno afuera"

Siendo n el tamaño del conjunto de los datos, el proceso de entrenamiento y prueba se repite n veces, en cada repetición se deja una única observación fuera de la muestra de entrenamiento; posteriormente, el modelo se prueba solamente en esa observación omitida, de manera que todas las observaciones del conjunto de datos hayan conformado la muestra de prueba en alguna iteración. Finalmente, se promedian los resultados de las métricas de evaluación de cada iteración. Este método es computacionalmente costoso porque el número de veces que se ajusta el modelo es igual a la cantidad total de datos disponibles; asimismo, en cada iteración se utiliza una muestra de entrenamiento casi del mismo tamaño que el conjunto total de datos (Kuhn y Johnson, 2013; Kumar y Paul, 2016).

Investigadores han encontrado resultados similares entre los métodos de "Dejar uno afuera" y "K-Fold", indicando que este último aplicado con k = 10 resulta más atractivo debido a su eficiencia computacional. Métodos "K-Fold" con valores pequeños de k, como k = 2 o k = 3, presentan un alto sesgo, aunque computacionalmente son muy eficientes. Se puede recurrir a la repetición de todo el procedimiento de la validación cruzada para aumentar la precisión de las estimaciones de manera efectiva, manteniendo un pequeño sesgo al mismo tiempo. Lo común es repetir el procedimiento k veces, por lo tanto, el procedimiento de entrenamiento y prueba se realizarían k^2 veces en total (Kuhn y Johnson, 2013; Kumar y Paul, 2016).

Si los resultados de la evaluación del modelo en el procedimiento de Validación Cruzada son inferiores significativamente a los obtenidos previamente a la aplicación de la técnica, se estaría presentando un típico caso de sobreajuste en el modelo. De lo contrario, esto indicaría estabilidad en el modelo, por lo tanto, un buen ajuste del modelo a los datos (Kumar y Paul, 2016; Žižka et al., 2019).

III. METODOLOGÍA

3.1 Formulación de hipótesis

La implementación de un modelo de regresión logística utilizando datos textuales transformados mediante el método de representación de texto TF-IDF es adecuada para la clasificación de comentarios en los *streamings* de orientación sobre la estrategia "Aprendo en casa", lo cual a su vez es sustentado por una adecuada tasa de correcta clasificación (exactitud).

3.2 Población

La aplicación de la investigación se realizó sobre datos textuales correspondientes a comentarios por parte de la comunidad educativa (conformada en su mayoría por docentes) que recibió orientaciones de la estrategia "Aprendo en casa" a través de los tres primeros *streamings* (transmisiones en vivo) a cargo del equipo de la Dirección de Formación Docente en Servicio (DIFODS) del Ministerio de Educación en la página de Facebook "Portal PerúEduca" durante los meses de mayo y junio de 2020. Los comentarios de cada *streaming* se recolectaron horas después de concluir la transmisión. Posteriormente, fueron clasificados por el área pedagógica del equipo de la DIFODS.

Tabla 4: Descripción general de los tres *streamings*

Streaming	Fecha	Cantidad de comentarios
1	8 de mayo de 2020	1455
2	25 de mayo de 2020	2085
3	5 de junio de 2020	6082

El cálculo de la obtención del tamaño de la muestra se detalla a continuación:

$N = 9622$

$\alpha = 0.01$

$E = 0.03$

$$n_0 = \frac{Z_{tab}^2(\pi)(1-\pi)}{E^2} = \frac{2.58^2(0.5)(0.5)}{0.03^2} = 1849$$

$$n = \frac{n_0\, N}{N + n_0 - 1} = \frac{1849(9622)}{9622 + 1849 - 1} = 1551.0966$$

n = 1552 (redondeado por exceso)

La muestra se distribuyó de la siguiente manera:

Tabla 5: Distribución de la muestra

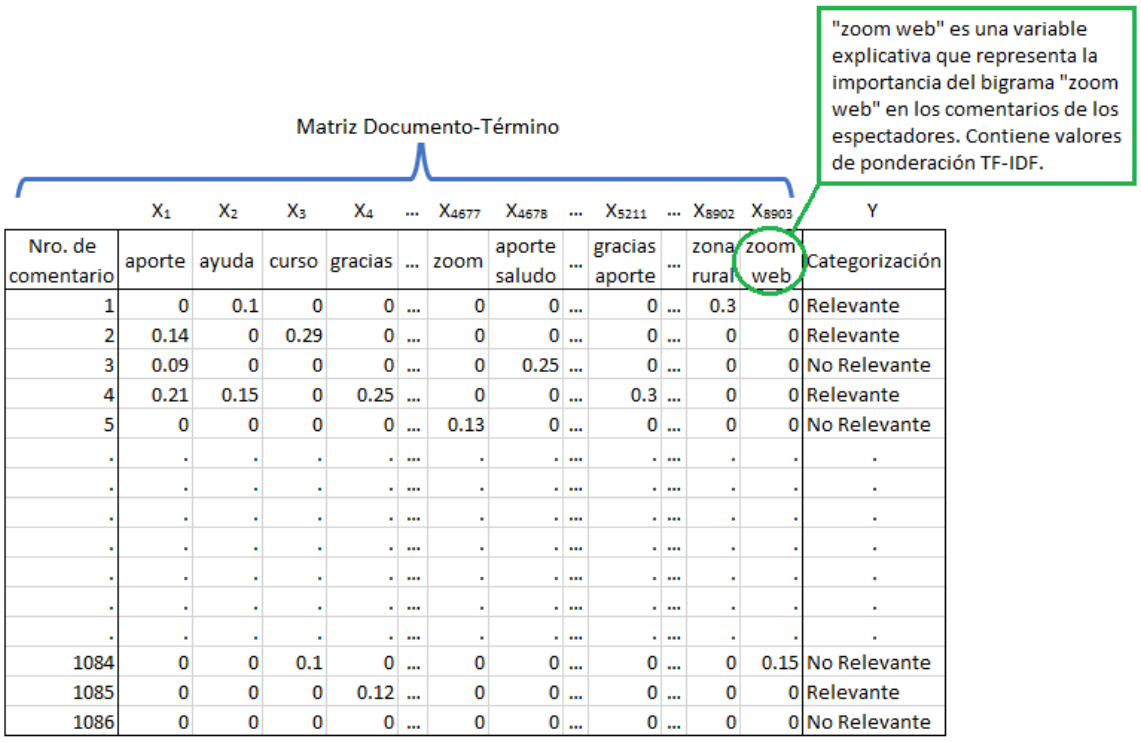
Tipo de comentario	Comentarios
Relevantes	769
No relevantes	783

3.3 Variables

Las variables se definieron de la siguiente manera:

Tabla 6: Variables

ID	Variable	Tipo de variable	Descripción
X_i	$Término_i$	Cuantitativa Continua	Importancia del término i en los comentarios de los espectadores
Y	Categorización	Cualitativa Nominal	Categoría del comentario (Relevante o No relevante)



	X_1	X_2	X_3	X_4	...	X_{4677}	X_{4678}	...	X_{5211}	...	X_{8902}	X_{8903}	Y
Nro. de comentario	aporte	ayuda	curso	gracias	...	zoom	aporte saludo	...	gracias aporte	...	zona rural	zoom web	Categorización
1	0	0.1	0	0	...	0	0	...	0	...	0.3	0	Relevante
2	0.14	0	0.29	0	...	0	0	...	0	...	0	0	Relevante
3	0.09	0	0	0	...	0	0.25	...	0	...	0	0	No Relevante
4	0.21	0.15	0	0.25	...	0	0	...	0.3	...	0	0	Relevante
5	0	0	0	0	...	0.13	0	...	0	...	0	0	No Relevante
.	.	.	.	.	...	.	.	...	.	...	.	.	.
.	.	.	.	.	...	.	.	...	.	...	.	.	.
.	.	.	.	.	...	.	.	...	.	...	.	.	.
.	.	.	.	.	...	.	.	...	.	...	.	.	.
.	.	.	.	.	...	.	.	...	.	...	.	.	.
.	.	.	.	.	...	.	.	...	.	...	.	.	.
.	.	.	.	.	...	.	.	...	.	...	.	.	.
1084	0	0	0.1	0	...	0	0	...	0	...	0	0.15	No Relevante
1085	0	0	0	0.12	...	0	0	...	0	...	0	0	Relevante
1086	0	0	0	0	...	0	0	...	0	...	0	0	No Relevante

Figura 3: Estructura de datos

3.4 Tipo de investigación

El tipo de investigación fue de carácter explicativo, se predijo la categoría correspondiente ("Relevante" o "No relevante") a cada comentario de los siguientes *streamings* de "Aprendo en casa" a través de la implementación del método de representación de texto TF-IDF en un modelo de regresión logística.

3.5 Diseño de investigación

El diseño de esta investigación fue de carácter no experimental y transversal, debido a que no existió manipulación de las variables en estudio y la recolección de los datos de cada grupo de espectadores de los *streamings* se realizó una sola vez.

3.6 Procedimiento de análisis

1. Pre-procesamiento de datos. Consistió en la limpieza de los datos textuales, es decir, la corrección de escritura, la supresión de signos de puntuación, enlaces web, caracteres especiales y palabras comunes no relevantes (por ejemplo, las preposiciones), y la conversión de todas las palabras en minúscula. El procedimiento se realizó en el programa Python con las librerías pandas, numpy, re y nltk.

2. Análisis exploratorio de datos. Se obtuvieron indicadores descriptivos de los comentarios por categoría utilizando unigramas, bigramas y trigramas, realizados en el programa R con los paquetes readxl, dplyr, tidytext, tidyr, tm, RColorBrewer, ggplot2, reshape2 y scales.

3. Aplicación del método de representación de texto TF-IDF. Se hizo una partición de datos para definir los conjuntos de datos de entrenamiento y prueba. Se aplicó el método TF-IDF al conjunto de datos de entrenamiento para obtener la matriz documento-término que contiene las ponderaciones de la importancia de cada término (unigramas y bigramas). Para reducir la dimensionalidad de la matriz, se seleccionaron los términos más importantes a través de la prueba Chi-Cuadrado. El procedimiento se realizó utilizando las librerías re, nltk, sklearn, seaborn y matplotlib.pyplot del programa Python.

4. Estimación del modelo de clasificación. La matriz documento-término conformó el conjunto de variables explicativas del modelo de Regresión Logística, es decir, cada variable explicativa consistió en la ponderación de la importancia de un término y fue representada en una columna de la matriz. Se descartaron las variables no significativas del modelo, además de comprobarse de que no exista una fuerte correlación entre las variables del modelo final de clasificación.

La matriz de correlación entre las variables fue hallada utilizando el programa R con el paquete GGally, mientras que la estimación del modelo se realizó en el programa Python con la librería sklearn y en el programa R con el paquete stats.

5. Evaluación del modelo de clasificación. Se obtuvieron los resultados de la Matriz de confusión, Exactitud, Precisión, Sensibilidad, Especificidad, Medida F1, AUC ROC, AUC PR y Validación cruzada para evaluar el modelo. El procedimiento se realizó en el programa Python con la librería sklearn.

6. Clasificación de nuevos comentarios. Se clasificaron nuevos comentarios en categorías Relevante o No Relevante, realizando previamente el pre-procesamiento de los datos y obteniendo la matriz documento-término para hallar los valores TF-IDF de las variables explicativas.

A continuación, se presenta el diagrama de flujo del procedimiento de análisis con imágenes referenciales de cada etapa para su mejor comprensión.

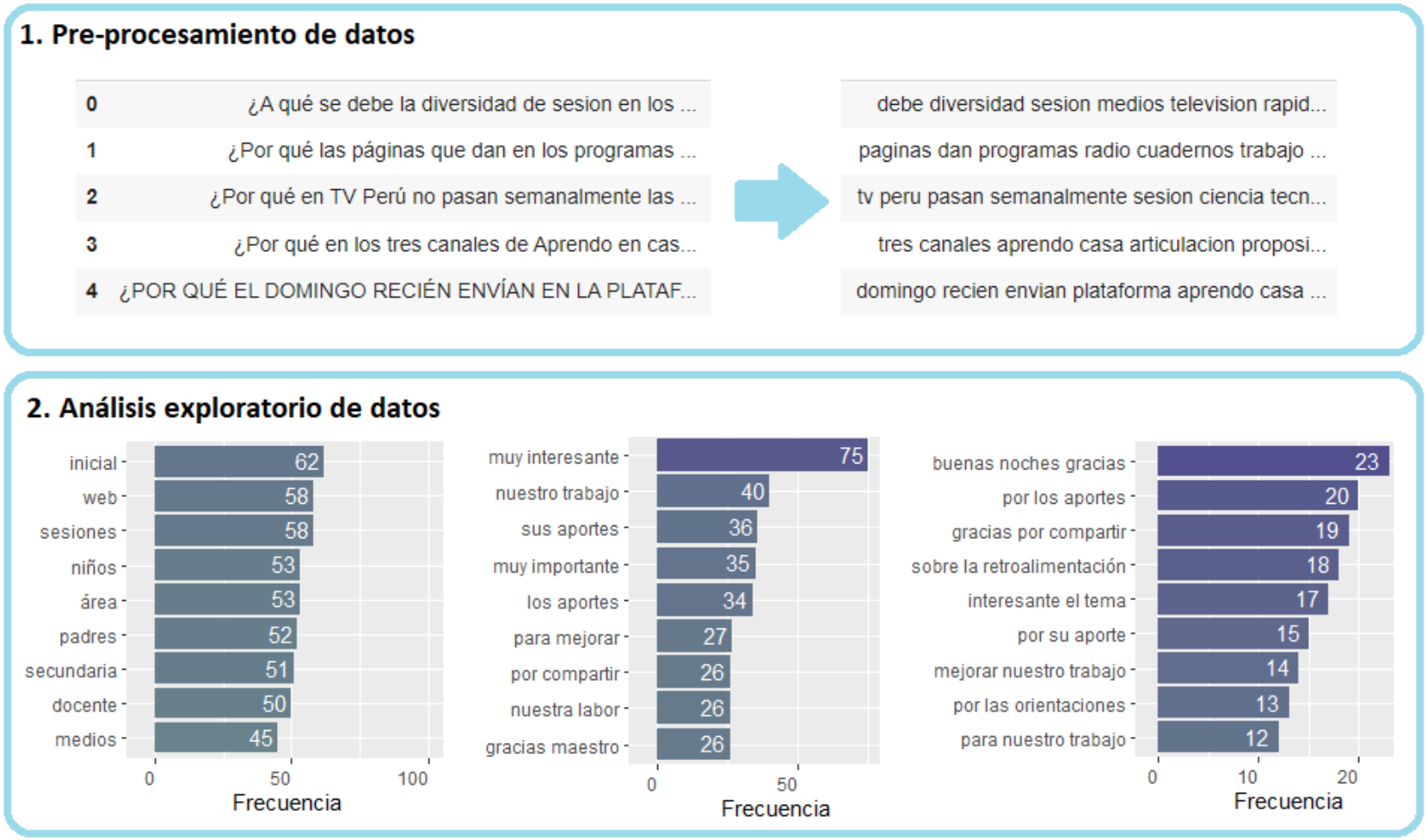
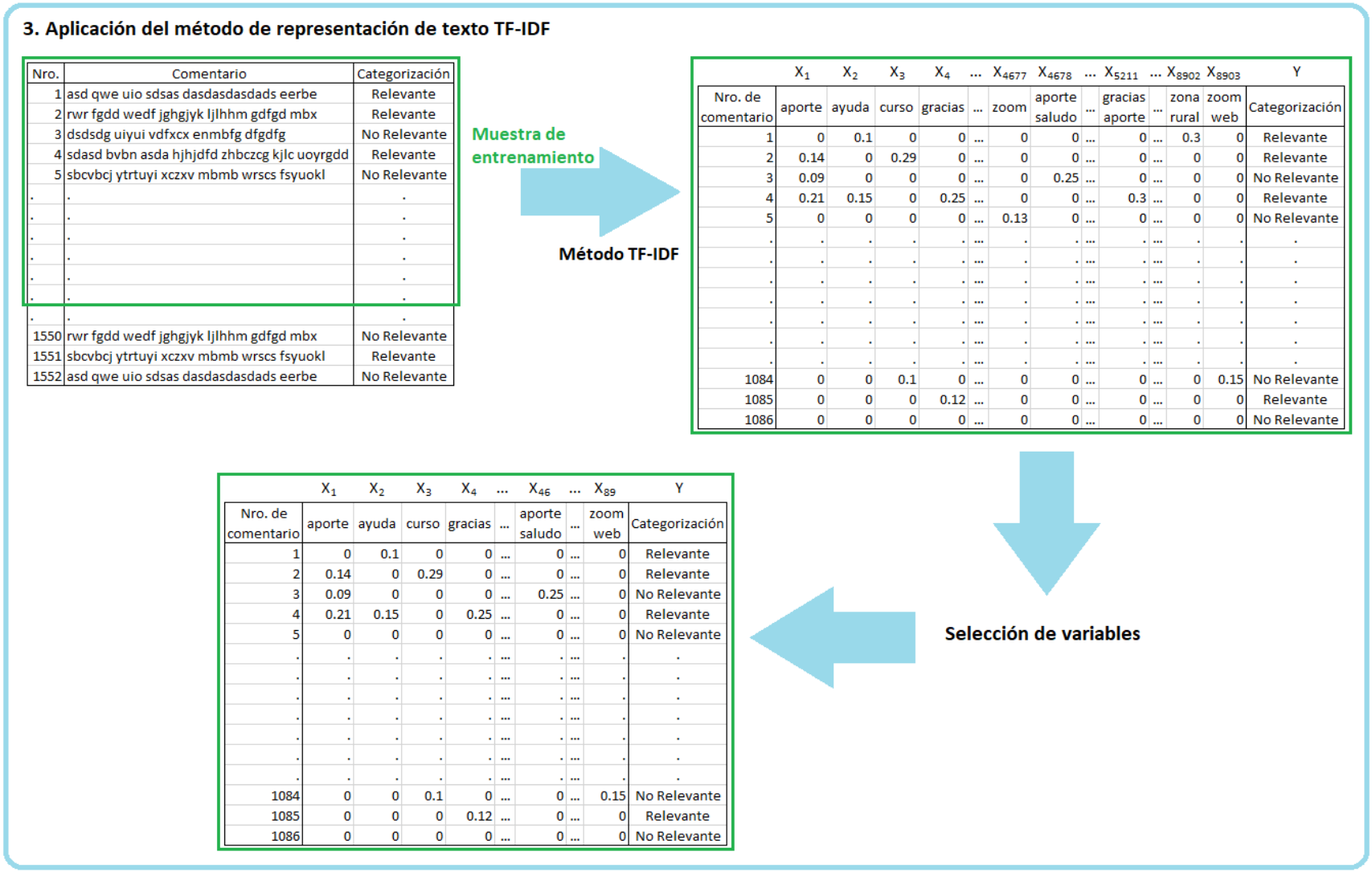


Figura 4: Diagrama de flujo del procedimiento de análisis (Etapas 1 y 2)

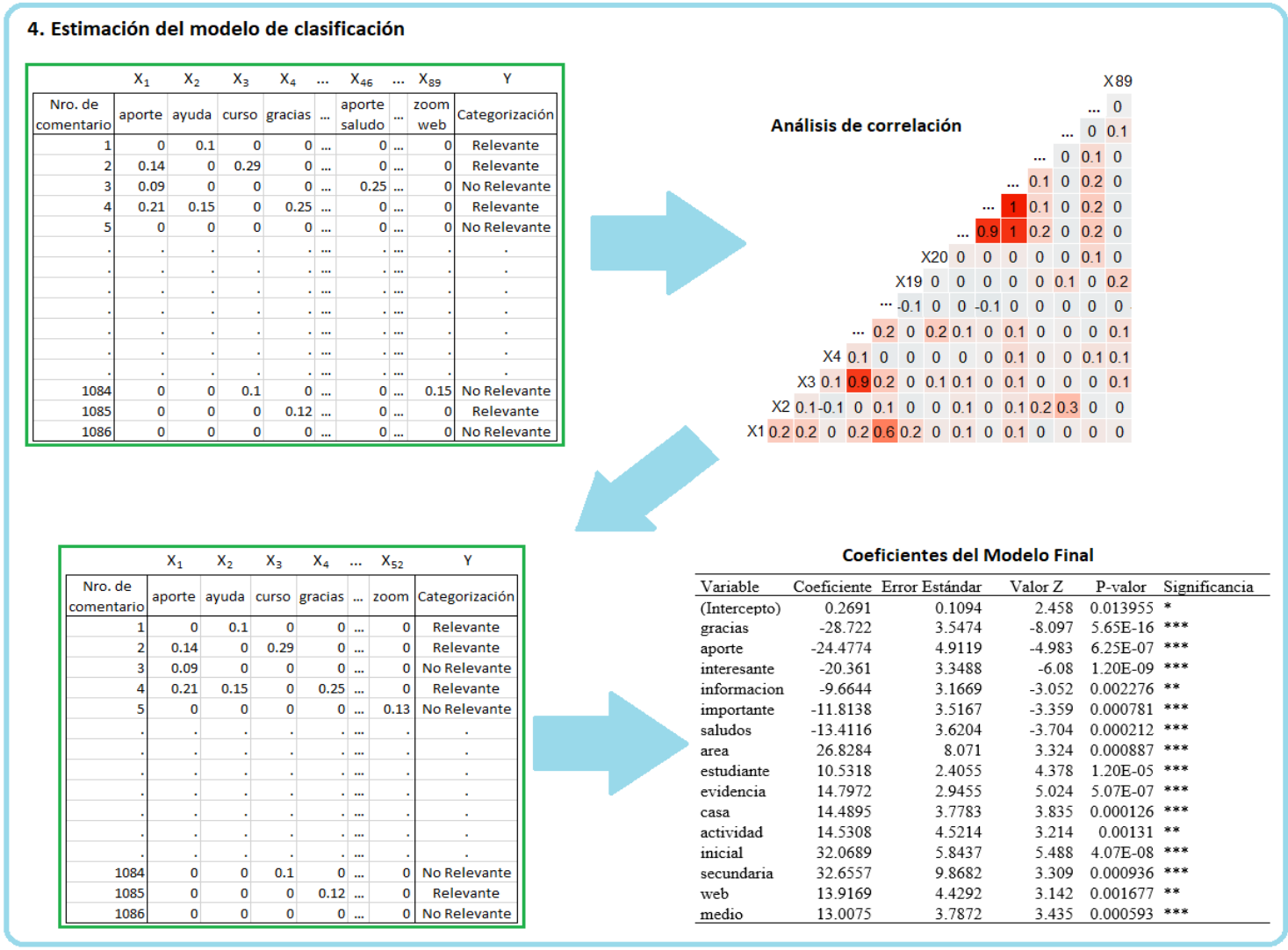


Nro.	Comentario	Categorización
1	asd qwe uio sdsas dasdasdasdads eerbe	Relevante
2	rwr fgdd wedf jghgjyk ljlhhm gdfgd mbx	Relevante
3	dsdsdg uiyui vdfxcx enmbfg dfgdfg	No Relevante
4	sdasd bvbn asda hjhjdfd zhbczcg kjlc uoyrgdd	Relevante
5	sbcvbcj ytrtuyi xczxv mbmb wrscs fsyuokl	No Relevante
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
.	.	.
1550	rwr fgdd wedf jghgjyk ljlhhm gdfgd mbx	No Relevante
1551	sbcvbcj ytrtuyi xczxv mbmb wrscs fsyuokl	Relevante
1552	asd qwe uio sdsas dasdasdasdads eerbe	No Relevante

	X_1	X_2	X_3	X_4	...	X_{4677}	X_{4678}	...	X_{5211}	...	X_{8902}	X_{8903}	Y
Nro. de comentario	aporte	ayuda	curso	gracias	...	zoom	aporte saludo	...	gracias aporte	...	zona rural	zoom web	Categorización
1	0	0.1	0	0	...	0	0	...	0	...	0.3	0	Relevante
2	0.14	0	0.29	0	...	0	0	...	0	...	0	0	Relevante
3	0.09	0	0	0	...	0	0.25	...	0	...	0	0	No Relevante
4	0.21	0.15	0	0.25	...	0	0	...	0.3	...	0	0	Relevante
5	0	0	0	0	...	0.13	0	...	0	...	0	0	No Relevante
.	.	.	.	.	...	.	.	...	.	...	.	.	.
.	.	.	.	.	...	.	.	...	.	...	.	.	.
.	.	.	.	.	...	.	.	...	.	...	.	.	.
.	.	.	.	.	...	.	.	...	.	...	.	.	.
.	.	.	.	.	...	.	.	...	.	...	.	.	.
.	.	.	.	.	...	.	.	...	.	...	.	.	.
1084	0	0	0.1	0	...	0	0	...	0	...	0	0.15	No Relevante
1085	0	0	0	0.12	...	0	0	...	0	...	0	0	Relevante
1086	0	0	0	0	...	0	0	...	0	...	0	0	No Relevante

	X_1	X_2	X_3	X_4	...	X_{46}	...	X_{89}	Y
Nro. de comentario	aporte	ayuda	curso	gracias	...	aporte saludo	...	zoom web	Categorización
1	0	0.1	0	0	...	0	...	0	Relevante
2	0.14	0	0.29	0	...	0	...	0	Relevante
3	0.09	0	0	0	...	0.25	...	0	No Relevante
4	0.21	0.15	0	0.25	...	0	...	0	Relevante
5	0	0	0	0	...	0	...	0	No Relevante
.	.	.	.	.	...	.	...	.	.
.	.	.	.	.	...	.	...	.	.
.	.	.	.	.	...	.	...	.	.
.	.	.	.	.	...	.	...	.	.
.	.	.	.	.	...	.	...	.	.
.	.	.	.	.	...	.	...	.	.
.	.	.	.	.	...	.	...	.	.
1084	0	0	0.1	0	...	0	...	0.15	No Relevante
1085	0	0	0	0.12	...	0	...	0	Relevante
1086	0	0	0	0	...	0	...	0	No Relevante

Figura 5: Diagrama de flujo del procedimiento de análisis (Etapa 3)

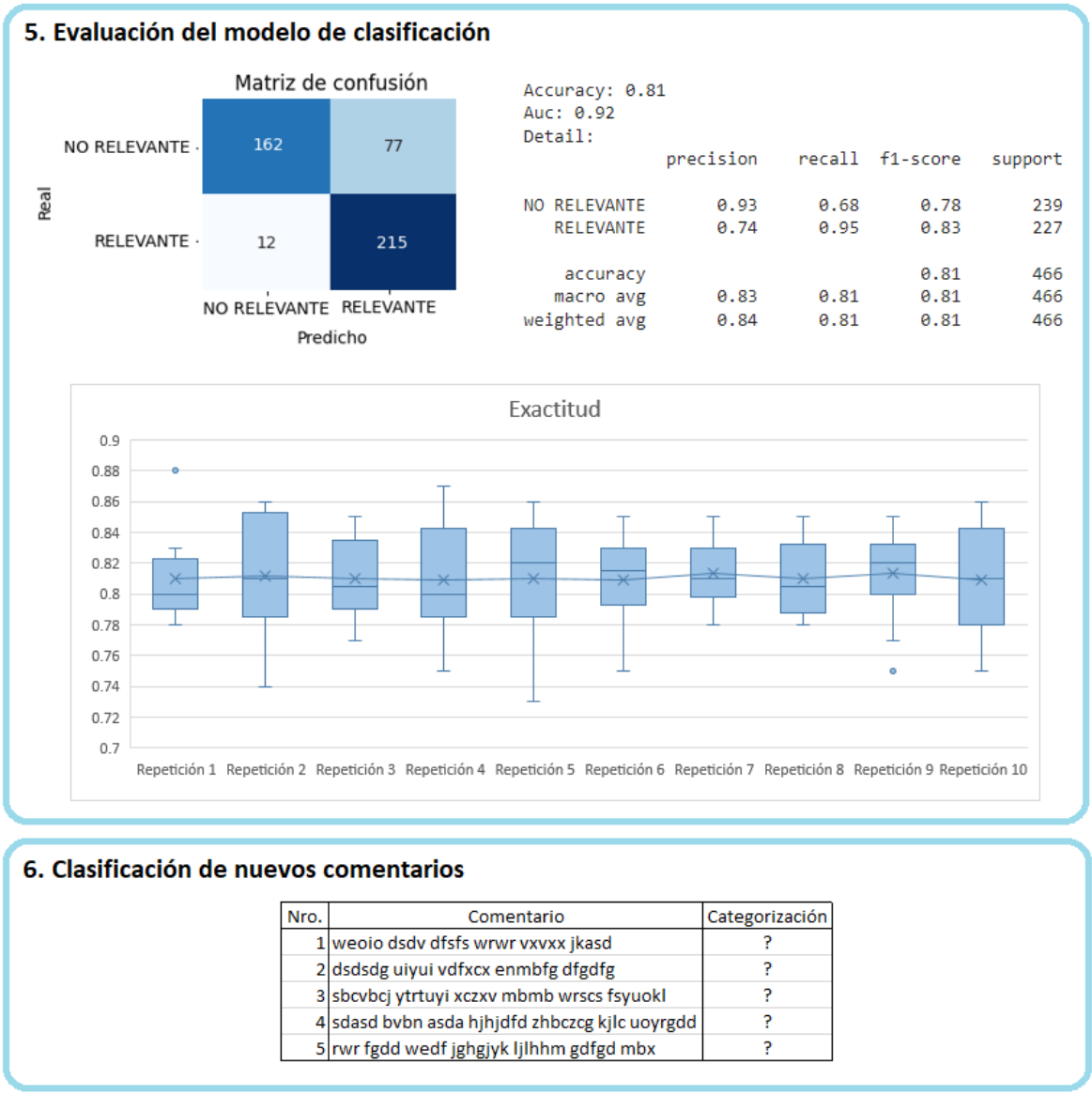


	X_1	X_2	X_3	X_4	...	X_{46}	...	X_{89}	Y
Nro. de comentario	aporte	ayuda	curso	gracias	...	aporte saludo	...	zoom web	Categorización
1	0	0.1	0	0	...	0	...	0	Relevante
2	0.14	0	0.29	0	...	0	...	0	Relevante
3	0.09	0	0	0	...	0.25	...	0	No Relevante
4	0.21	0.15	0	0.25	...	0	...	0	Relevante
5	0	0	0	0	...	0	...	0	No Relevante
.	.	.	.	.	...	.	...	.	.
.	.	.	.	.	...	.	...	.	.
.	.	.	.	.	...	.	...	.	.
.	.	.	.	.	...	.	...	.	.
.	.	.	.	.	...	.	...	.	.
.	.	.	.	.	...	.	...	.	.
.	.	.	.	.	...	.	...	.	.
1084	0	0	0.1	0	...	0	...	0.15	No Relevante
1085	0	0	0	0.12	...	0	...	0	Relevante
1086	0	0	0	0	...	0	...	0	No Relevante

	X_1	X_2	X_3	X_4	...	X_{52}	Y
Nro. de comentario	aporte	ayuda	curso	gracias	...	zoom	Categorización
1	0	0.1	0	0	...	0	Relevante
2	0.14	0	0.29	0	...	0	Relevante
3	0.09	0	0	0	...	0	No Relevante
4	0.21	0.15	0	0.25	...	0	Relevante
5	0	0	0	0	...	0.13	No Relevante
.	.	.	.	.	...	.	.
.	.	.	.	.	...	.	.
.	.	.	.	.	...	.	.
.	.	.	.	.	...	.	.
.	.	.	.	.	...	.	.
.	.	.	.	.	...	.	.
.	.	.	.	.	...	.	.
1084	0	0	0.1	0	...	0	No Relevante
1085	0	0	0	0.12	...	0	Relevante
1086	0	0	0	0	...	0	No Relevante

Variable	Coeficiente	Error Estándar	Valor Z	P-valor	Significancia
(Intercepto)	0.2691	0.1094	2.458	0.013955	*
gracias	-28.722	3.5474	-8.097	5.65E-16	***
aporte	-24.4774	4.9119	-4.983	6.25E-07	***
interesante	-20.361	3.3488	-6.08	1.20E-09	***
informacion	-9.6644	3.1669	-3.052	0.002276	**
importante	-11.8138	3.5167	-3.359	0.000781	***
saludos	-13.4116	3.6204	-3.704	0.000212	***
area	26.8284	8.071	3.324	0.000887	***
estudiante	10.5318	2.4055	4.378	1.20E-05	***
evidencia	14.7972	2.9455	5.024	5.07E-07	***
casa	14.4895	3.7783	3.835	0.000126	***
actividad	14.5308	4.5214	3.214	0.00131	**
inicial	32.0689	5.8437	5.488	4.07E-08	***
secundaria	32.6557	9.8682	3.309	0.000936	***
web	13.9169	4.4292	3.142	0.001677	**
medio	13.0075	3.7872	3.435	0.000593	***

Figura 6: Diagrama de flujo del procedimiento de análisis (Etapa 4)



Nro.	Comentario	Categorización
1	weoio dsdv dfsfs wrwr vxvxx jkasd	?
2	dsdsdg uiyui vdfxcx enmbfg dfgdfg	?
3	sbcvbcj ytrtuyi xczxv mbmb wrscs fsyuokl	?
4	sdasd bvbn asda hjhjdfd zhbczcg kjlc uoyrgdd	?
5	rwr fgdd wedf jghgjyk ljlhhm gdfgd mbx	?

Figura 7: Diagrama de flujo del procedimiento de análisis (Etapas 5 y 6)

IV. RESULTADOS Y DISCUSIÓN

4.1 Pre-procesamiento de datos

En esta primera etapa se realizó la limpieza y estandarización de los datos textuales de los comentarios, la cual consistió en:

- Corrección de errores de escritura.
- Supresión de signos de puntuación, enlaces web y caracteres especiales.
- Conversión de todas las palabras en minúscula.
- Conversión de palabras de plural a singular, a excepción de las palabras que naturalmente son expresadas en plural, como “gracias”.

Para fines descriptivos, en el Análisis exploratorio de datos no se consideraron los procedimientos de limpieza de conversión de palabras de plural a singular ni la supresión de tildes. En las siguientes etapas sí se incluyeron estos procedimientos de limpieza como pasos previos.

4.2 Análisis exploratorio de datos (AED)

En esta etapa se obtuvieron indicadores descriptivos de los comentarios de cada categoría utilizando unigramas, bigramas y trigramas.

Las categorías de los comentarios son las siguientes:

Relevantes: Comentarios que contienen un mensaje importante acerca de la estrategia Aprendo en casa, tales como consultas, sugerencias y reclamos.

Tabla 7: Ejemplo de comentarios pertenecientes a la categoría "Relevante"

Nro.	Comentarios Relevantes
1	Buenas noches, profesor Julio ¿a qué se refiere con los canales de comunicación comunitaria? Si hablamos de que están sin conectividad.
2	¿Por qué las páginas que dan en los programas de radio no son de los cuadernos de trabajo de este año (Resolvemos Problemas)?
3	¿Cómo atender a más de 150 estudiantes en cuanto a la retroalimentación formativa?
4	Buenas tardes, en Vida Activa Secundaria ¿qué trabajos se puede dejar responsablemente?
5	¿Por qué no brindan los planificadores de Ciencias Sociales de 5°? No están en la plataforma.
6	¿Cómo convencer a aquellos padres de familia que no apoyan a sus niños?, porque piensan que esa es nuestra labor.
7	¿A qué se debe la diversidad de sesiones en los medios y en la televisión la rapidez de su desarrollo?
8	En el caso de los docentes de educación técnico-productiva ¿cómo podemos pedir sus evidencias?
9	¿Qué hacer con el estudiante que se encuentra en el campo y no tiene ningún medio de comunicación para trabajar Aprendo en casa?
10	Las clases por tv deben ser de acuerdo al cuaderno de trabajo del estudiante, se facilita y aprenden mejor.

No Relevantes: Comentarios que no contienen un mensaje importante acerca de la estrategia Aprendo en casa, tales como saludos, felicitaciones, agradecimientos por la capacitación, apreciación del *streaming*, entre otros.

Tabla 8: Ejemplo de comentarios pertenecientes a la categoría "No Relevante"

Nro.	Comentarios No Relevantes
1	Bastante claridad que nos llevamos a mejorar nuestra evaluación formativa.
2	Muy buenas precisiones. Gracias por compartir y fortalecer nuestras competencias pedagógicas como maestras.
3	Amigo Julio interesantes tus aportes para la construcción de la identidad docente.
4	Excelente las orientaciones de la retroalimentación. Bendiciones.
5	Muy buena información de las dudas que tenía y que me permitirá mejorar en la enseñanza aprendizaje de los estudiantes en la estrategia Aprendo en Casa.
6	Así es, estamos día a día fortaleciendo las innovaciones de la tecnología.
7	Buena explicación para aclarar muchas dudas. Gracias Maestro.
8	Ahora sí me quedó claro muy buena explicación maestro.
9	San Martín, qué bueno, hay muchos maestros interesados, decididos al cambio.
10	Muy buenos aportes, ojalá quede grabado el programa, saludos desde Ica.

En la obtención de los unigramas se omitieron las palabras comunes no relevantes (por ejemplo, las preposiciones) llamadas *stopwords*.

En las Figuras 8 y 9 se observan los unigramas más frecuentes de cada categoría.

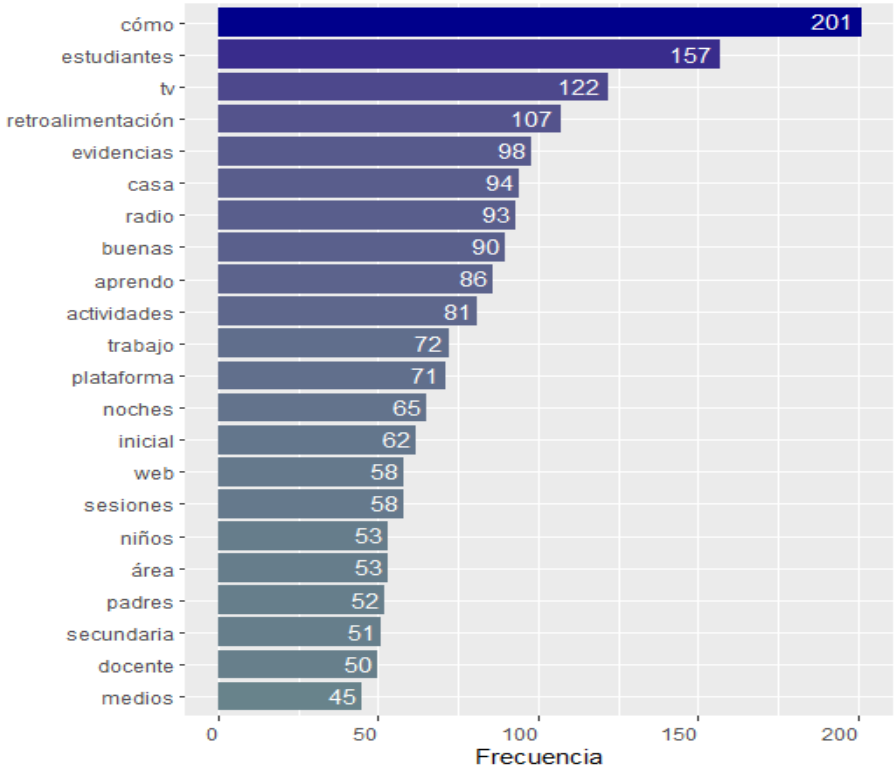


Figura 8: Unigramas más frecuentes de la categoría “Relevante”

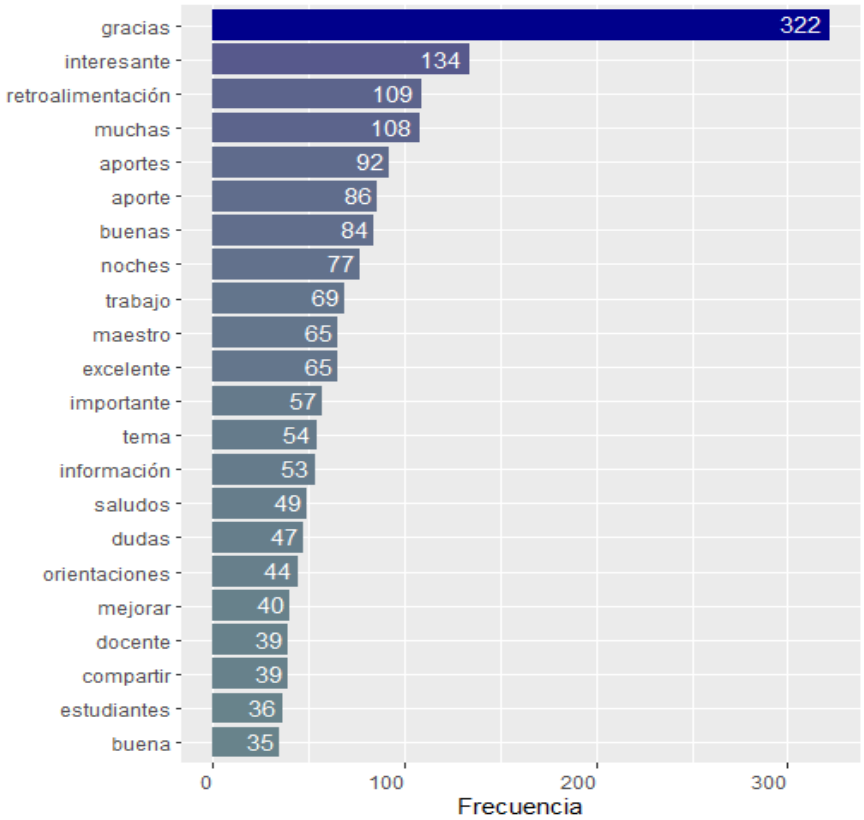


Figura 9: Unigramas más frecuentes de la categoría “No Relevante”

En la Figura 8 se observó que los unigramas más frecuentes de la categoría Relevante fueron "cómo", "estudiantes", "tv" y "retroalimentación", mientras que en la Figura 9 los unigramas más frecuentes de la categoría No Relevante fueron "gracias", "interesante", "retroalimentación" y "muchas". Además, la lista de unigramas de cada categoría en las Figuras 8 y 9 denota la temática de los comentarios: Educación y la estrategia Aprendo en casa. Sin embargo, las categorías difieren en qué aspectos se enfocan: La categoría Relevante se enfoca en consultas, sugerencias y reclamos (evidenciado principalmente por el unigrama "cómo"), mientras que la categoría No Relevante representa los saludos, felicitaciones, agradecimientos y apreciación del *streaming*, representados por unigramas como "gracias", "interesante", "excelente", "importante" y "saludos".

En las Figuras 10 y 11 se observan los bigramas y trigramas más frecuentes de la categoría Relevante, respectivamente. Asimismo, en las Figuras 12 y 13 se muestran los bigramas y trigramas más frecuentes de la categoría No Relevante, respectivamente.

Los bigramas y trigramas complementan la información proporcionada por los unigramas sobre las categorías; a través de ellos se observan con mayor detalle las diferencias entre las categorías. A diferencia de los unigramas, para fines descriptivos, los bigramas y trigramas requieren de los *stopwords* para obtener ideas completas o entendibles; por lo tanto, no se excluyeron los *stopwords* en los bigramas y trigramas de las Figuras 10, 11, 12 y 13.

Para mayor detalle, se puede visualizar la lista completa de *stopwords* y los 100 unigramas, bigramas y trigramas sin *stopwords* más frecuentes de cada categoría en los Anexos 2 y 3, respectivamente.

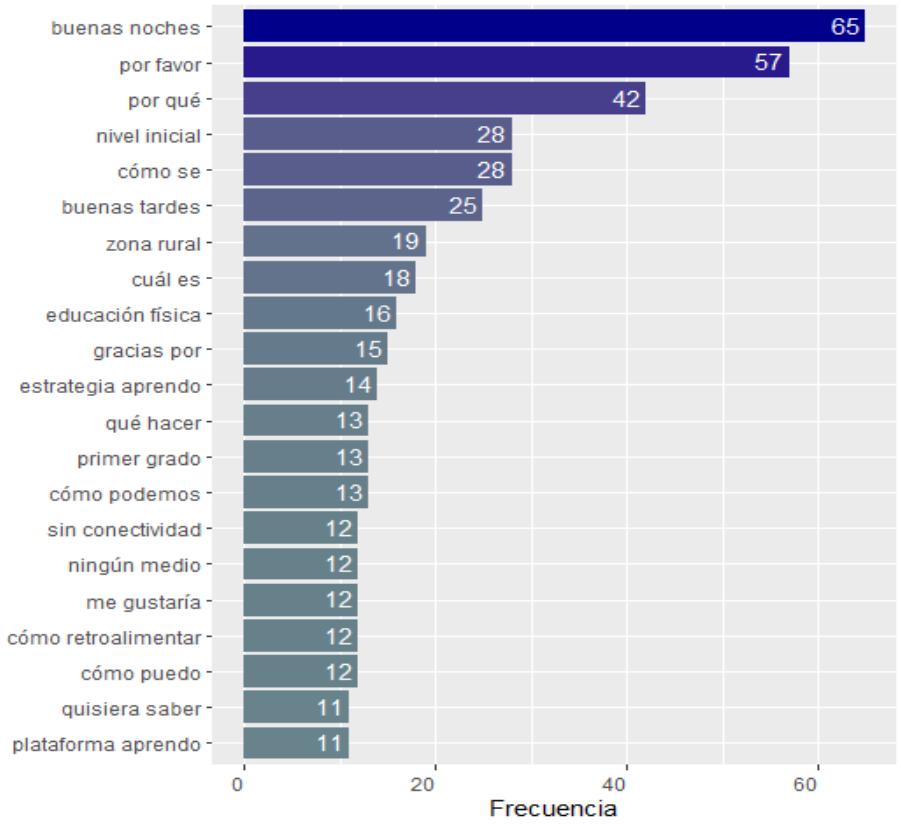


Figura 10: Bigramas más frecuentes de la categoría “Relevante”

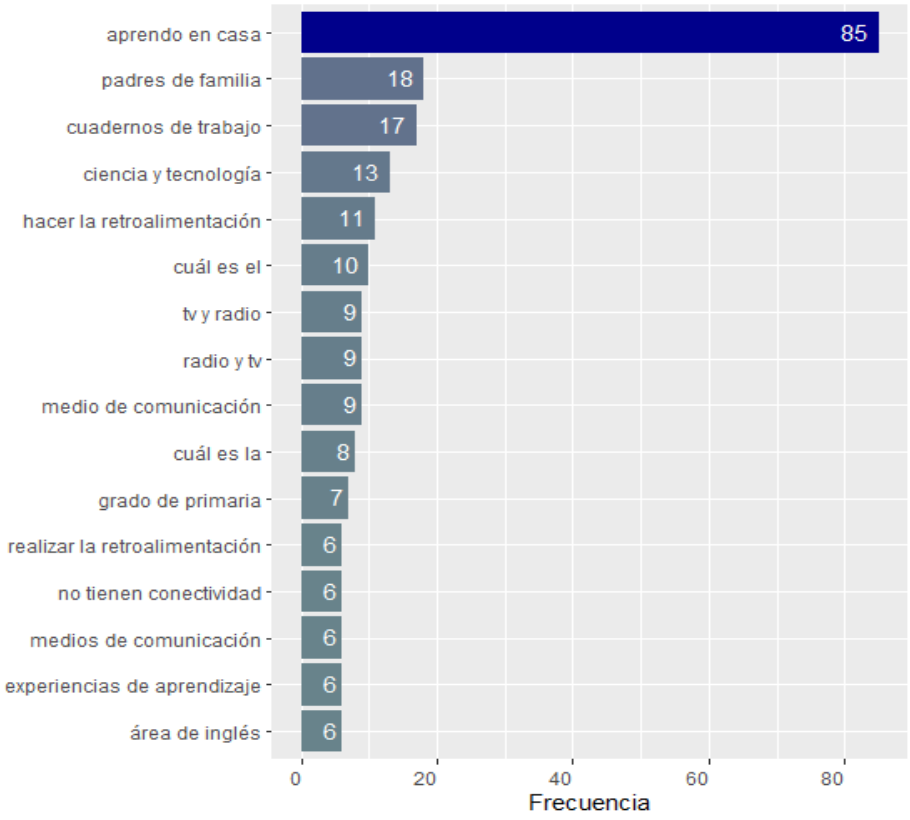


Figura 11: Trigramas más frecuentes de la categoría “Relevante”

En la Figura 10 se puede observar lo siguiente:

- Los bigramas “buenas noches” y “buenas tardes” pueden denotar un saludo cortés de parte de los docentes al momento de realizar su consulta.
- Los bigramas “por favor”, “por qué”, “cómo se”, “cuál es”, “qué hacer”, “cómo podemos”, “me gustaría”, “cómo puedo” y “quisiera saber” denotan la inquietud y la intención de los docentes de consultar.
- Los bigramas “nivel inicial”, “zona rural”, “educación física”, “estrategia aprendo”, “primer grado”, “sin conectividad”, “ningún medio”, “cómo retroalimentar” y “plataforma aprendo” indican los temas frecuentes de los que se han hecho consultas.

En la Figura 11 se observa con mayor detalle algunos temas frecuentes que se observaron en la Figura 10 y también se muestran otros temas de consulta de los docentes, tales como “padres de familia”, “cuadernos de trabajo”, “ciencia y tecnología”, “tv y radio”, “experiencias de aprendizaje” y “área de inglés”.

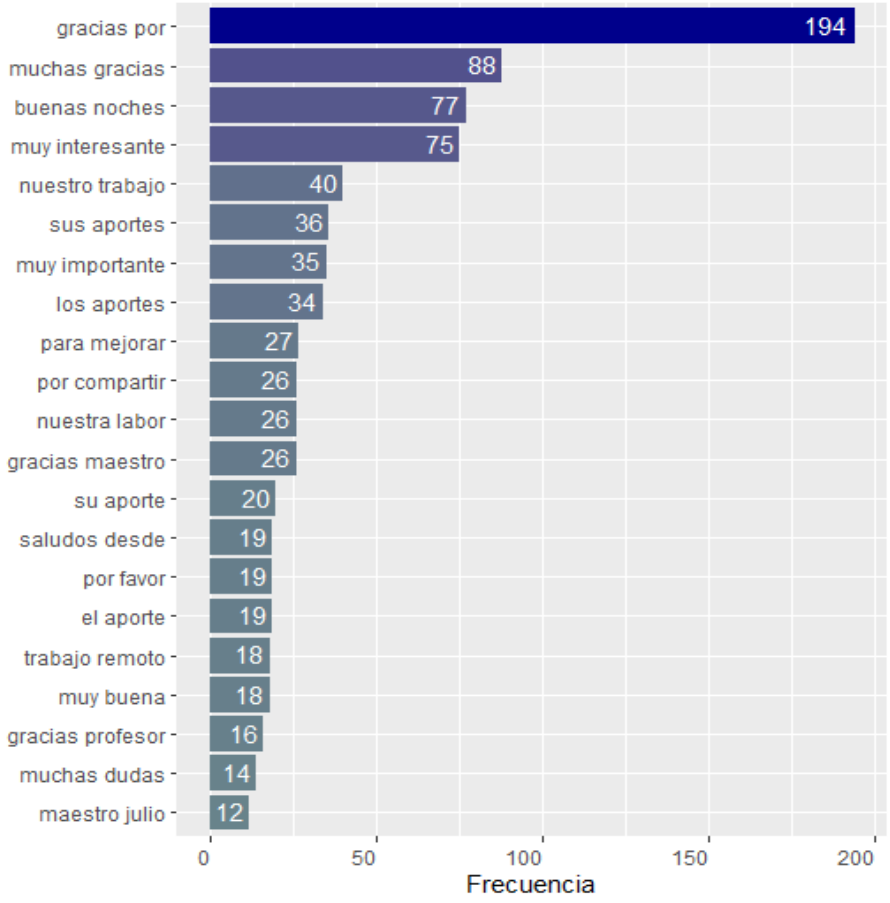


Figura 12: Bigramas más frecuentes de la categoría “No Relevante”

En la Figura 12 se puede observar lo siguiente:

- Los bigramas “buenas noches” y “saludos desde” denotan un saludo cortés por parte de los espectadores.
- Los bigramas “gracias por”, “muchas gracias”, “gracias maestro” y “gracias profesor” reflejan los agradecimientos por la capacitación de la estrategia “Aprendo en casa”.
- Los bigramas “muy interesante”, “sus aportes”, “muy importante”, “los aportes”, “para mejorar”, “por compartir”, “muy buena”, “nuestro trabajo” y “nuestra labor” pueden denotar la apreciación de la importancia del contenido del *streaming*.

En la Figura 13 se puede observar de una manera más completa las ideas mostradas anteriormente en la Figura 12.

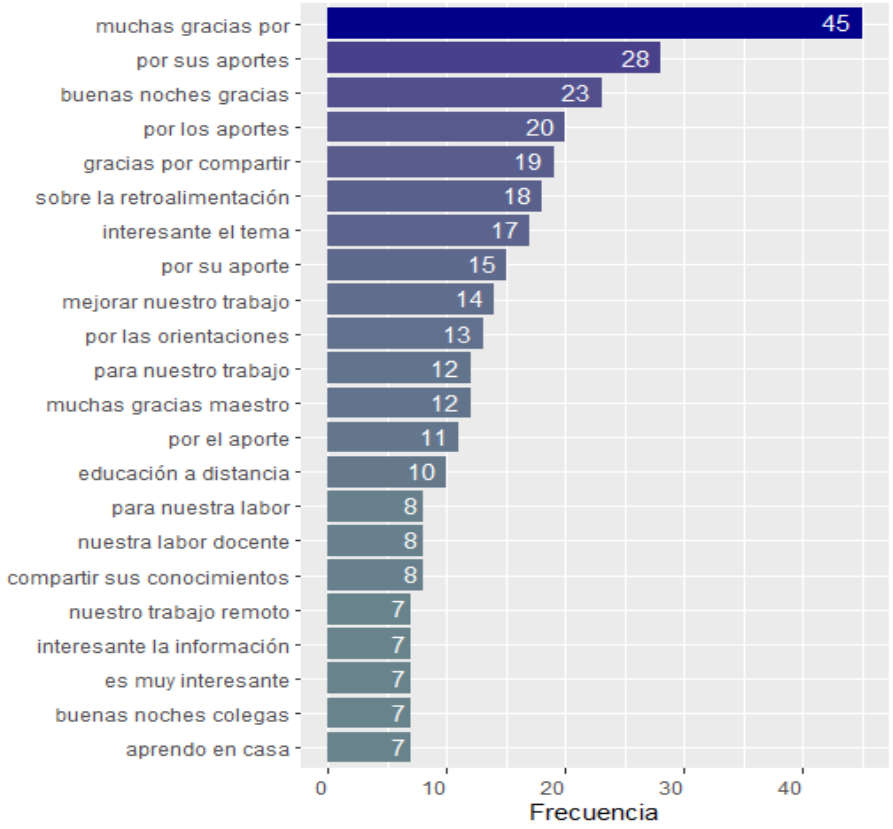


Figura 13: Trigramas más frecuentes de la categoría “No Relevante”

La frecuencia de los n-gramas disminuye conforme se añada un término, debido a que solamente se deben considerar las ideas contenidas únicamente en n términos (solo dos términos para los bigramas y solo tres términos para los trigramas), y que además se coincida en el orden de los términos. Es por ello que la frecuencia de los bigramas es menor a la frecuencia de los unigramas, así como la frecuencia de los trigramas es menor a la frecuencia de los bigramas.

4.3 Aplicación del método de representación de texto TF-IDF

En esta etapa se realizó la partición de los datos para obtener la muestra de entrenamiento y la muestra de prueba, las cuales consistieron en el 70 % y 30 % de la muestra original, respectivamente.

A continuación, se aplicó el método de representación de texto TF-IDF a la muestra de entrenamiento y se obtuvo la matriz documento-término, donde cada fila representa un comentario, cada columna corresponde a un término (considerándose únicamente unigramas y bigramas para este estudio) y cada valor representa la ponderación TF-IDF de un término en un determinado comentario.

La matriz documento-término obtenida está conformada por 1086 filas y 8903 columnas, es decir, 1086 comentarios y 8903 unigramas y bigramas hallados en los 1086 comentarios. Dado que es una matriz dispersa, los valores TF-IDF diferentes de cero son escasos, menos del 1 % de los valores en la matriz.

En este procedimiento de análisis, cada columna de la matriz documento-término, que representa la ponderación TF-IDF de un término, consiste en una variable explicativa para la estimación del modelo de clasificación; por lo tanto, en esta etapa se tienen 8903 variables. Para disminuir esta cantidad, se realizó una selección de variables utilizando la prueba Chi Cuadrado a un nivel de significancia de 0.05. Siendo "i" el término i-ésimo en la matriz documento-término, el planteamiento de las hipótesis fue el siguiente:

H_0: La importancia del término "i" y la categoría No Relevante son independientes.
H_1: La importancia del término "i" y la categoría No Relevante no son independientes.

H_0: La importancia del término "i" y la categoría Relevante son independientes.
H_1: La importancia del término "i" y la categoría Relevante no son independientes.

A un nivel de significancia de 0.05, se tuvo suficiente evidencia estadística para afirmar que las variables que no son independientes de las categorías Relevante y No Relevante fueron las siguientes:

Tabla 9: Variables no independientes pertenecientes a la categoría No Relevante

Variable	P-valor	χ^2
gracias	1.93179E-14	58.5983689
aporte	8.6996E-12	46.6014395
interesante	8.34661E-10	37.677397
muchas gracias	2.04035E-07	26.9945025
muchas	3.85193E-07	25.7671279
gracias aporte	1.79605E-06	22.8017142
excelente	2.3248E-06	22.3059904
informacion	1.64278E-05	18.5641888
importante	6.12237E-05	16.0644162
duda	0.000100425	15.1286951
orientacion	0.000157309	14.2826352
maestro	0.000182302	14.0052371
mejorar	0.000182903	13.9990534
compartir	0.000815324	11.2060199
gracias maestro	0.000995857	10.8352538
tema	0.001063375	10.7138333
gracias orientacion	0.001484875	10.0972906
aporte retroalimentacion	0.001641875	9.91219139
mejorar trabajo	0.001643916	9.90990538
interesante aporte	0.001701752	9.84628364
saludos	0.002083878	9.47414649
buena	0.00211031	9.45102643
labor	0.002723795	8.98382759
felicitaciones	0.002832845	8.91211823
curso	0.003134674	8.72737516
julio	0.003456004	8.54958821
alcance	0.00419685	8.19662031
exposicion	0.004775548	7.96253522
pedagogica	0.00493919	7.9015705
noches gracias	0.005301572	7.77358525
ayuda	0.00701419	7.26933151
gran	0.008582897	6.90762393
profesor	0.009301349	6.7640295
interesante tema	0.009465844	6.73275211
ponencia	0.009860015	6.66001384
gracias compartir	0.010086387	6.61957545
retroalimentacion	0.011557019	6.37759105
aclarar	0.012180311	6.28443201
gracias informacion	0.013805282	6.06287421
practica	0.014637015	5.9596417
claro	0.015738172	5.83189387
excelente aporte	0.017850254	5.61078604
gracias profesor	0.018869334	5.51358479

tema retroalimentacion	0.019080747	5.49409999
tv	0.019679145	5.44013529
buenas noches	0.019691533	5.43903611
noches	0.019691533	5.43903611
ayudara	0.021989068	5.24664979
servira	0.025807503	4.96887
maestro julio	0.028265099	4.81182481
agradecida	0.028819163	4.77838341
trabajo	0.030629854	4.67358904
gracias alcance	0.030907203	4.65810952
valioso	0.030910207	4.6579427
area	0.031421561	4.62978172
agradece	0.036649711	4.36661574
perueduca	0.037509458	4.32712923
radio	0.039821985	4.22545312
mejorar labor	0.040586264	4.19320647
gracias interesante	0.043874517	4.06138436
acuna	0.043990119	4.056941
mejora	0.044660132	4.0314277
plataforma	0.045525929	3.99904955
oportunidad	0.046070574	3.97901218
excelente exposicion	0.046409189	3.96668003
bendiciones	0.04661588	3.95919901
diapositiva	0.048181631	3.90363926
labor docente	0.048221241	3.90225853

Tabla 10: Variables no independientes pertenecientes a la categoría Relevante

Variable	P-valor	χ^2
tv	3.21624E-06	21.6830274
area	9.855E-06	19.5393221
si	1.75571E-05	18.4374757
estudiante	3.55298E-05	17.0963894
radio	4.06326E-05	16.8416061
evidencia	4.87135E-05	16.4975314
plataforma	6.54093E-05	15.9392177
casa	7.66633E-05	15.638848
sesion	0.0001018	15.1030189
aprendo	0.000121793	14.7647122
aprendo casa	0.00013261	14.6043173
actividad	0.000137252	14.5394956
conectividad	0.000368151	12.6872842
inicial	0.000386849	12.5946653
hacer	0.000492526	12.1437503
gracias	0.000752871	11.3539646
secundaria	0.000832784	11.1667045

aporte	0.000932612	10.9567904
web	0.001060907	10.7181323
padre	0.001280236	10.3708059
solo	0.001871728	9.67126055
nivel	0.00237839	9.23186737
medio	0.002484174	9.15221327
deben	0.002839922	8.90756146
programacion	0.003065406	8.76812364
zona	0.003356046	8.60302795
interesante	0.00378198	8.38563525
caso	0.00387709	8.34050122
comunicacion	0.004154809	8.21488742
puede	0.004985765	7.88459462
trabajar	0.006171649	7.49948949
estan	0.009563352	6.71447353
alumno	0.010087453	6.61938731
semana	0.010389387	6.56688104
nivel inicial	0.011365253	6.40729332
rural	0.011701547	6.35553782
ejemplo	0.012172178	6.28561578
cuaderno	0.012200031	6.28156513
tv radio	0.012706693	6.20949422
muchas gracias	0.013668064	6.08051862
ningun	0.014602167	5.96384452
debe	0.014816662	5.93813823
ser	0.015416944	5.86818448
deberian	0.015506697	5.85796677
competencia	0.016482596	5.75064585
diferente	0.017226471	5.67315258
quisiera	0.01797151	5.59892421
retroalimentar	0.018787382	5.52119864
grado	0.01879038	5.52091946
gracias aporte	0.019333292	5.47111459
acceso	0.019568444	5.44998953
excelente	0.020957519	5.33032151
plataforma aprendo	0.021742562	5.26626928
ningun medio	0.022111736	5.23697134
educacion fisica	0.022248824	5.22622077
fisica	0.022248824	5.22622077
cuaderno trabajo	0.023005935	5.16805693
zona rural	0.023345813	5.1425886
primaria	0.02339543	5.13890277
tiempo	0.024777692	5.03935652
tambien	0.024951274	5.02726476
tema	0.025266827	5.00550526
nino	0.025658097	4.9789132

favor	0.02750252	4.85898444
saber	0.028879974	4.77475388
envian	0.029214357	4.75493733
mismo	0.030373399	4.68803296
familia	0.031094255	4.64775129
realizar	0.032856455	4.55324321
ciencia	0.033226662	4.53406198
clase	0.033471277	4.52151006
radio tv	0.03467986	4.46086852
muchas	0.035103449	4.44013609
cuentan	0.037595264	4.32324045
podria	0.038933641	4.26376061
planificacion	0.040070573	4.21489388
si estudiante	0.040375898	4.20201826
puedo	0.040534959	4.19535109
matematica	0.041318532	4.16290291
pueden	0.041876527	4.14018764
padre familia	0.044038835	4.05507222
ciencia tecnologia	0.044163985	4.05028131
evaluar	0.045451091	4.00182255
whatsapp	0.048132373	3.90535796

También se observa que estas variables presentaron altos valores del estadístico χ^2.

Es necesario recalcar que, debido al proceso de limpieza de datos, los términos de las variables no llevan tilde.

Se descartaron las variables que representaban términos como conectores lógicos, adverbios y verbos, entre ellos: "si", "solo" y "hacer", respectivamente. Asimismo, se observó multicolinealidad, es decir, fuerte correlación entre variables, especialmente en unigramas y bigramas relacionados por sus términos, como se ilustra en la Figura 14.

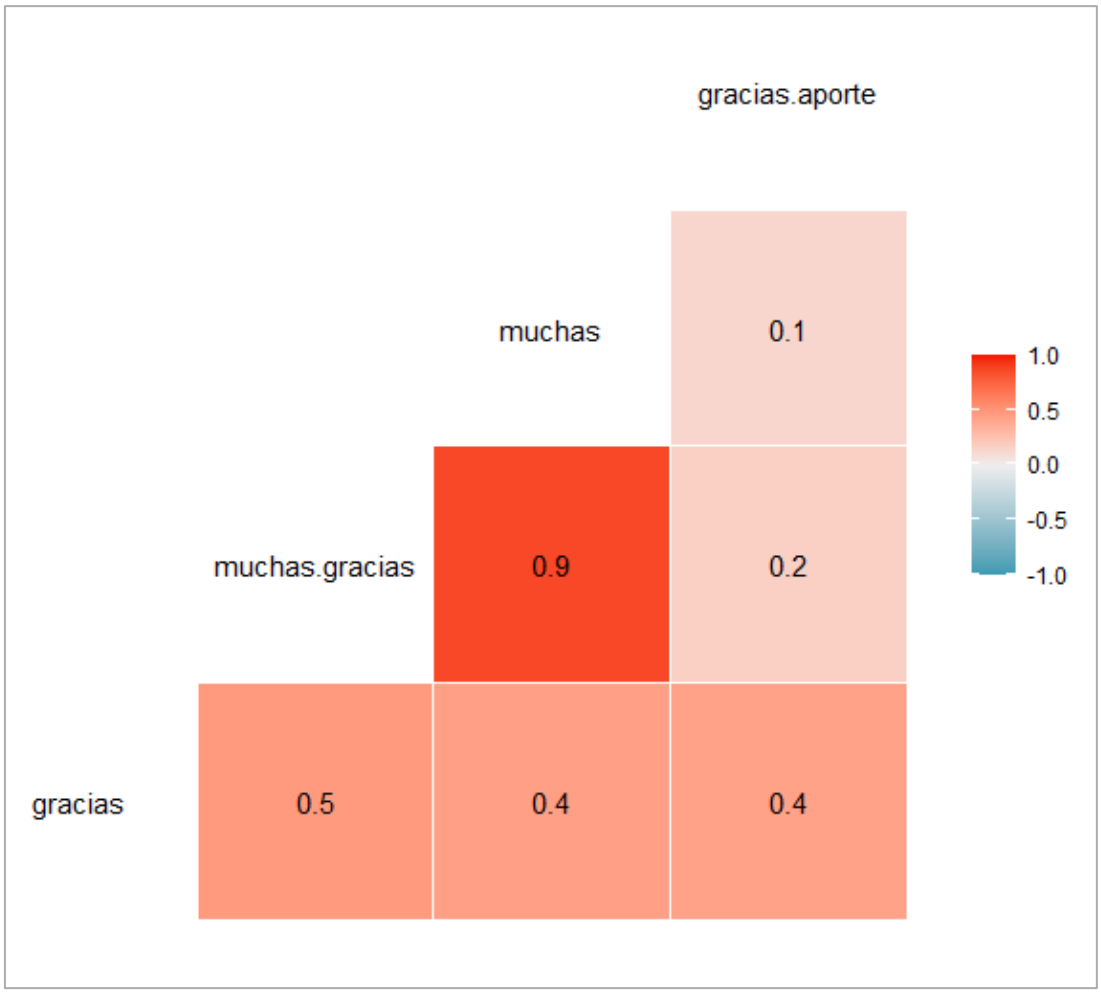


Figura 14: Correlación entre la variable del unigrama “gracias” y variables relacionadas a ella por sus términos

La solución ante la presencia de alta correlación entre esas variables fue escoger solo una variable que represente a las demás. En el caso de las variables de la Figura 14, "gracias" permaneció en el proceso.

4.4 Estimación del modelo de clasificación

En esta cuarta etapa se obtuvo el modelo final de clasificación de Regresión Logística ajustado con los datos de entrenamiento provenientes de la matriz documento-término; donde previamente se descartaron las variables no significativas del modelo.

A continuación, en la Figura 15 se presenta la correlación entre las variables explicativas del modelo final, donde se observa que no existe una correlación fuerte entre las variables, dado que el más alto valor de correlación ha sido de 0.4; por lo tanto, se comprueba la ausencia de multicolinealidad.

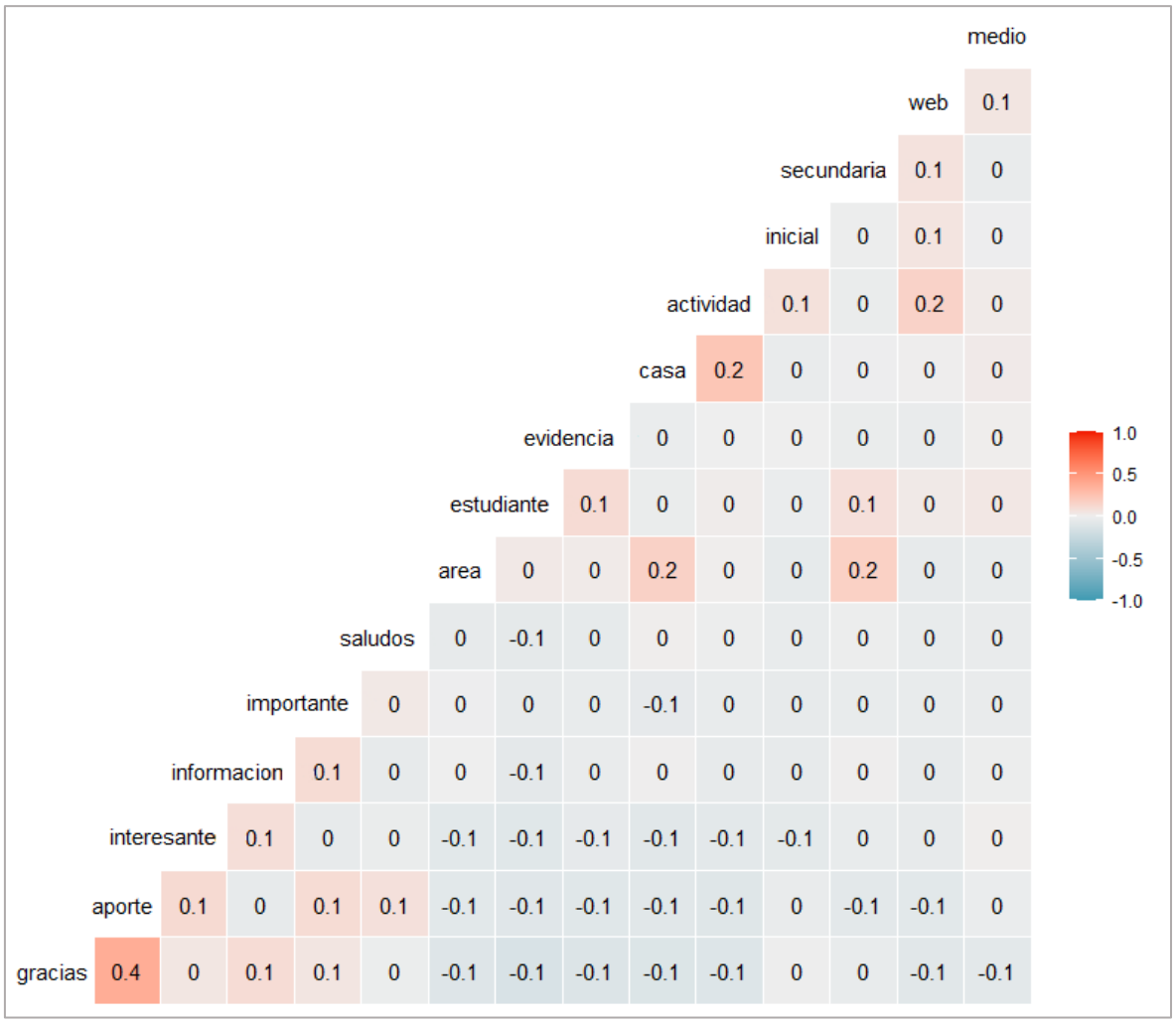


Figura 15: Correlación entre las variables del modelo de clasificación

En la Tabla 11 se muestran los coeficientes estimados de las variables significativas del modelo final de clasificación. Se observa que las variables con mayor significancia fueron la importancia de los términos “gracias”, “aporte”, “interesante”, “importante”, “saludos”, “area”, “estudiante”, “evidencia”, “casa”, “inicial”, “secundaria” y “medio”.

Asimismo, el signo de los coeficientes indica que el sentido de la relación entre la variable dependiente y las variables de los términos “gracias”, “aporte”, “interesante”, “información”, “importante” y “saludos” es negativa; por lo tanto, la probabilidad de que un comentario sea Relevante disminuye al incrementarse la importancia de estos términos en el texto; según ponderación TF-IDF. Por el contrario, las variables de los términos “area”, “estudiante”, “evidencia”, “casa”, “actividad”, “inicial”, “secundaria”, “web” y “medio” presentan una relación positiva con respecto a la variable dependiente, es decir, la probabilidad de que un comentario sea Relevante aumenta al incrementarse la importancia de estos términos.

No se consideró conveniente interpretar la magnitud de los coeficientes, los cuales son más altos de los que resultan de análisis de regresión de aplicaciones diferentes a la clasificación de datos textuales, generando *odds ratios* en rangos muy extremos. Esto se debe a la naturaleza de los datos: la matriz dispersa contiene menos del 1 % de valores diferentes de cero; a su vez, estos valores son ponderaciones del método TF-IDF, los cuales están limitados en un rango de 0 a 1.

Tabla 11: Coeficientes del modelo de clasificación

Variable	Coeficiente	Error Estándar	Valor Z	P-valor	Significancia
(Intercepto)	0.2691	0.1094	2.458	0.013955	*
gracias	-28.722	3.5474	-8.097	5.65E-16	***
aporte	-24.4774	4.9119	-4.983	6.25E-07	***
interesante	-20.361	3.3488	-6.08	1.20E-09	***
informacion	-9.6644	3.1669	-3.052	0.002276	**
importante	-11.8138	3.5167	-3.359	0.000781	***
saludos	-13.4116	3.6204	-3.704	0.000212	***
area	26.8284	8.071	3.324	0.000887	***
estudiante	10.5318	2.4055	4.378	1.20E-05	***
evidencia	14.7972	2.9455	5.024	5.07E-07	***
casa	14.4895	3.7783	3.835	0.000126	***
actividad	14.5308	4.5214	3.214	0.00131	**
inicial	32.0689	5.8437	5.488	4.07E-08	***
secundaria	32.6557	9.8682	3.309	0.000936	***
web	13.9169	4.4292	3.142	0.001677	**
medio	13.0075	3.7872	3.435	0.000593	***

Significancia de las variables:

* si el p-valor está en el rango de 0.01 – 0.05

** si el p-valor está en el rango de 0.001 – 0.01

*** si el p-valor está en el rango de 0 – 0.001

4.5 Evaluación del modelo de clasificación

En la quinta etapa del procedimiento, se aplicó el método de representación de texto TF-IDF a la muestra de prueba, con el objetivo de obtener su matriz documento-término para realizar la clasificación y hallar los resultados de las métricas de evaluación del modelo.

La muestra de prueba estuvo conformada por 239 observaciones de la clase No Relevante y 227 observaciones de la clase Relevante, por lo tanto, se trata de una muestra relativamente balanceada.

En la Figura 16 se tiene la matriz de confusión, donde se puede apreciar la cantidad de observaciones correctamente clasificadas de cada clase, esto es, 162 comentarios correctamente asignados a la clase No Relevante (Verdaderos negativos) y 215 comentarios de la clase Relevante correctamente clasificados (Verdaderos positivos); por lo tanto, existen 12 Falsos negativos y 77 Falsos positivos. En consecuencia, la exactitud (*accuracy*) indica que la proporción de comentarios correctamente clasificados, entre Relevantes y No Relevantes, ha sido 0.81.

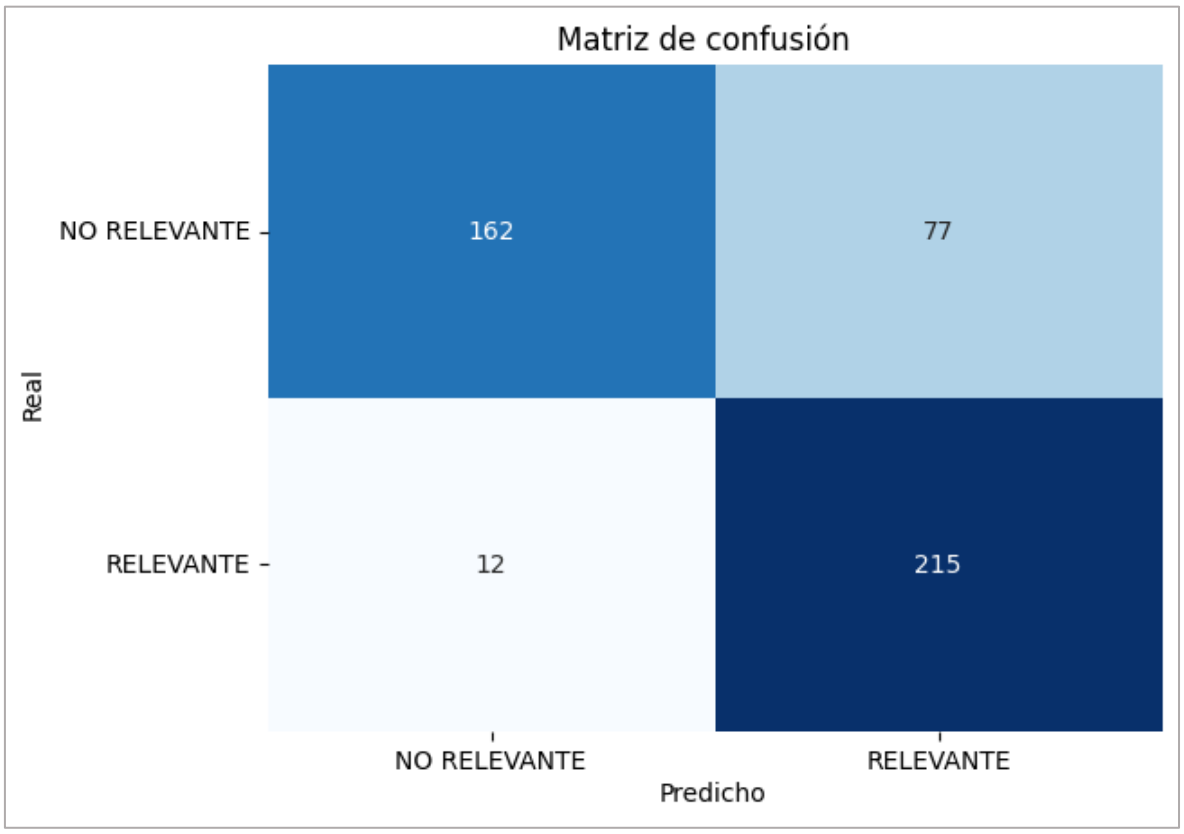


Figura 16: Matriz de confusión

Dado que la categoría No Relevante resulta de interés para los análisis posteriores, en los resultados de las métricas de evaluación también se contempló el caso donde la categoría No Relevante es la clase de interés, además de los resultados correspondientes al promedio simple y ponderado entre las dos categorías.

En la Tabla 12 se muestran los resultados de las métricas de evaluación del modelo para ambas clases. Tal como se observó en la Figura 16, el modelo clasificó como Relevante a 292 comentarios, siendo errónea la clasificación de 77 de ellos; por consiguiente, la precisión fue 0.74, mientras que la precisión con respecto a la clase No Relevante fue 0.93.

Tabla 12: Métricas de evaluación del modelo de clasificación

Métricas	Relevante	No Relevante	Promedio simple	Promedio ponderado
Precisión	0.74	0.93	0.83	0.84
Sensibilidad	0.95	0.68	0.81	0.81
Especificidad	0.68	0.95	0.81	0.81
Medida F1	0.83	0.78	0.81	0.81

La sensibilidad indica que se acertó en un 95 % la clasificación de los comentarios que realmente eran Relevantes.

La especificidad indica que el 68 % de los comentarios que realmente eran No Relevantes fueron clasificados correctamente.

La medida F1, la cual es el promedio armónico de la precisión y la sensibilidad, resultó en 0.78 para la categoría No Relevante y 0.83 para la categoría Relevante; lo cual indica un buen rendimiento de clasificación.

Las Curvas ROC para ambas clases se presentan en la Figura 17, donde se observa que ambas curvas ROC se encuentran por encima de la línea diagonal y se aproximan al punto en la coordenada (0,1). Asimismo, el área bajo la curva (AUC ROC) fue 0.92 para ambas categorías. Por lo tanto, se puede considerar que el modelo tiene un rendimiento sobresaliente de clasificación.

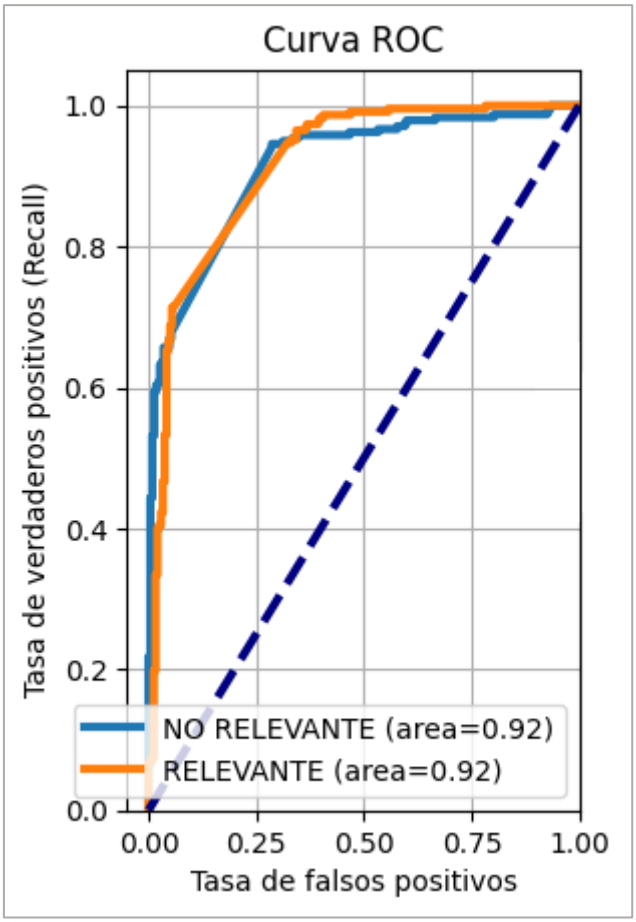


Figura 17: Curvas ROC para las categorías Relevante y No Relevante

Las Curvas de Precisión-Recall para ambas clases se presentan en la Figura 18, donde se puede observar que las curvas de PR se acercan a la coordenada (1,1) y se encuentran por encima de la línea de base del nivel de rendimiento del clasificador, representada por la línea horizontal que pasa por la ordenada 0.487, cuyo valor depende de la proporción de observaciones de la clase Relevante en la muestra de prueba. Además, el área bajo la curva (AUC PR) fue 0.92 para la clase Relevante y 0.94 para la clase No Relevante; estos resultados demuestran que el modelo de clasificación ha logrado tener un excelente desempeño.

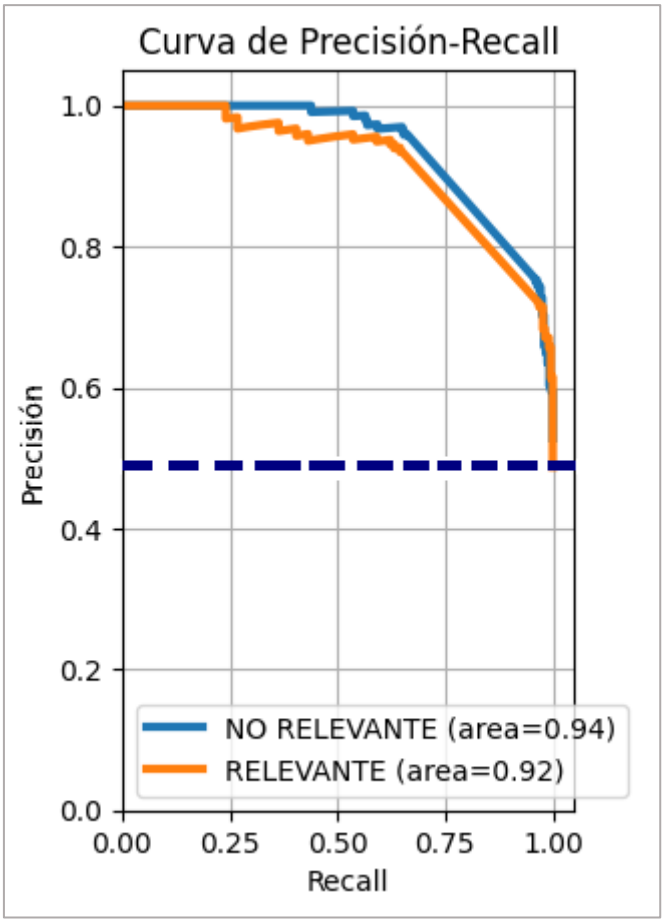


Figura 18: Curvas de Precisión-Recall para las categorías Relevante y No Relevante

Posteriormente, se evaluó el modelo de clasificación mediante el método *K-Fold* de Validación Cruzada. La muestra total (1552 observaciones) se dividió aleatoriamente en 10 sub-muestras (K=10) con 155 comentarios en cada una, a excepción de la última sub-muestra con 157 comentarios. Se seleccionó como muestra de prueba a una sub-muestra diferente en cada ejecución del modelo y se obtuvieron los resultados de las métricas de evaluación (exactitud, precisión, sensibilidad, especificidad, medida F1, AUC ROC, AUC PR). Este procedimiento se repitió nueve veces más, por lo tanto, se obtuvo 100 conjuntos de resultados para las métricas de evaluación, hallándose un promedio total de 0.8105 para la exactitud. En la Tabla 13 se presentan los resultados de las métricas para cada clase.

Tabla 13: Promedio de las métricas de evaluación del modelo para cada categoría en el proceso de Validación Cruzada

Métricas	Promedio de la categoría Relevante	Promedio de la categoría No Relevante	Promedio simple entre categorías	Promedio ponderado entre categorías
Precisión	0.7408	0.9322	0.8363	0.8388
Sensibilidad	0.9491	0.6747	0.8126	0.8105
Especificidad	0.6747	0.9491	0.8126	0.8105
Medida F1	0.8320	0.7819	0.8068	0.8072
AUC ROC	0.9156	0.9156	0.9156	0.9156
AUC PR	0.9158	0.9305	0.9232	0.9239

Se observa mucha similitud entre los resultados de la Validación Cruzada (Tabla 13) y los obtenidos anteriormente (Tabla 12 y Figuras 17 y 18), lo cual demuestra que no hay sobreajuste en el modelo. Asimismo, la ligera diferencia percibida en milésimas del promedio simple y ponderado de cada métrica demuestra que los datos son balanceados.

Los resultados detallados de las métricas en cada repetición se encuentran en el Anexo 4.

A continuación, se muestran los resultados a través de diagramas de cajas para cada métrica, donde cada caja representa los 10 resultados de una repetición en el proceso de Validación Cruzada.

En la Figura 19 se observa estabilidad en el promedio de la Exactitud, debido a que permaneció en el rango de 0.8 a 0.82 en las 10 repeticiones.

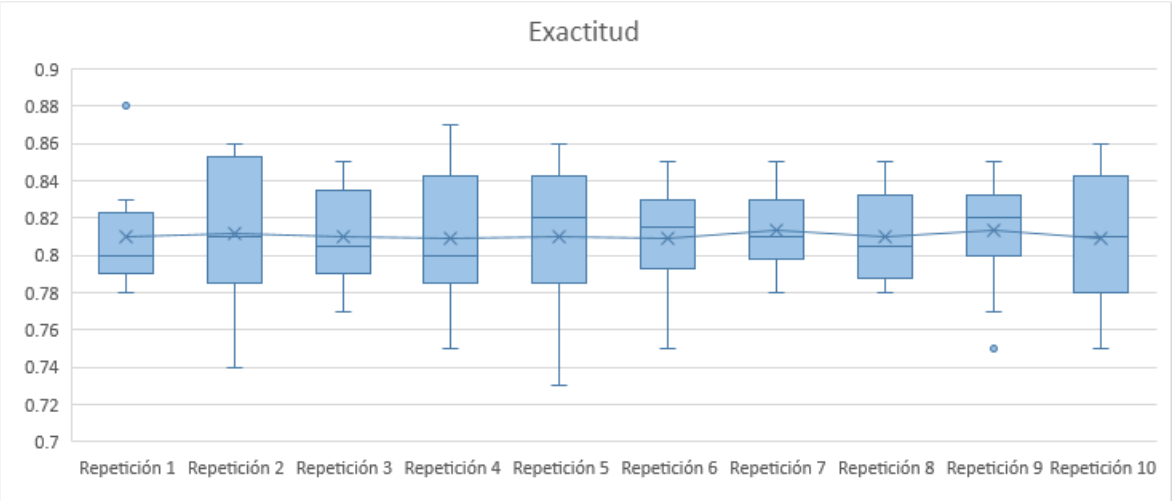


Figura 19: Diagrama de cajas de la Exactitud en las 10 repeticiones de Validación Cruzada

En la Figura 20 también se observa muy poca variabilidad en el promedio del AUC ROC, el cual se encontró en el rango de 0.9 a 0.92.

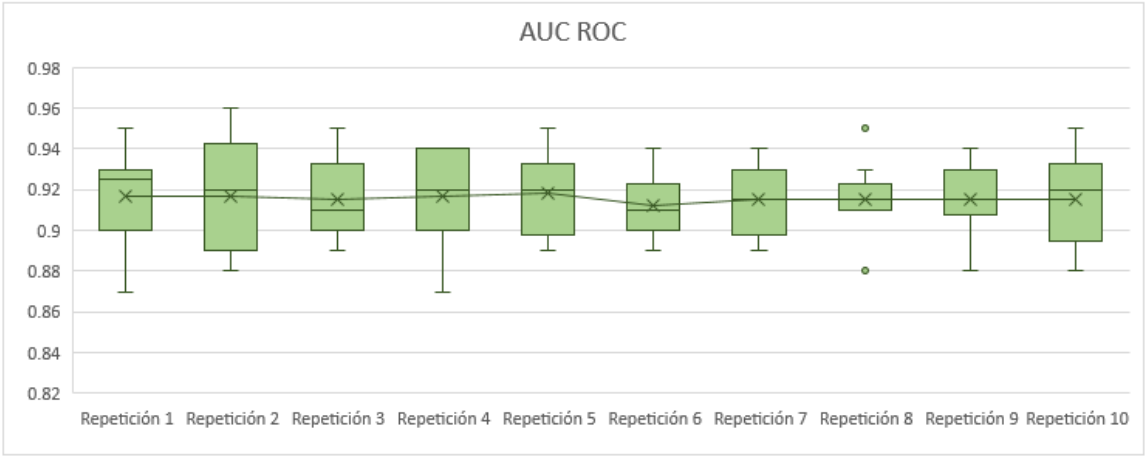


Figura 20: Diagrama de cajas del AUC ROC en las 10 repeticiones de Validación Cruzada

En las Figuras 21, 22, 23, 24 y 25 se muestran dos series de cajas para cada métrica, donde la serie que contiene cajas azules representa el caso en el que la categoría No Relevante es la clase de interés, mientras que la serie que contiene cajas anaranjadas representa los resultados en el que la categoría Relevante es la clase de interés.

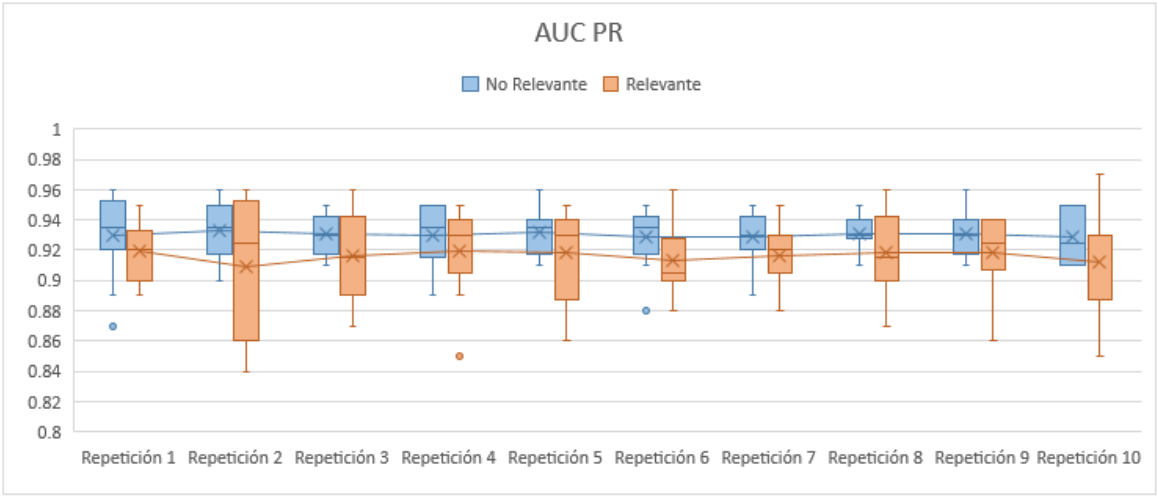


Figura 21: Diagrama de cajas del AUC PR en las 10 repeticiones de Validación Cruzada, según categoría

En la Figura 21 se observa estabilidad en el promedio del AUC PR de ambas series: el rango del promedio de la serie de la categoría No Relevante fue de 0.9 a 0.92, y el rango del promedio de la serie de la categoría Relevante fue de 0.92 a 0.94. Por lo tanto, los rangos de ambas categorías se encuentran muy cercanos entre sí.

En la Figura 22 se aprecia poca variabilidad en el promedio de la Precisión en ambas series de cajas. Sin embargo, los rangos del promedio de ambas categorías se encuentran distanciados: El rango del promedio para la serie de la categoría No Relevante fue de 0.9 a 0.95, mientras que para la categoría Relevante fue de 0.7 a 0.75.

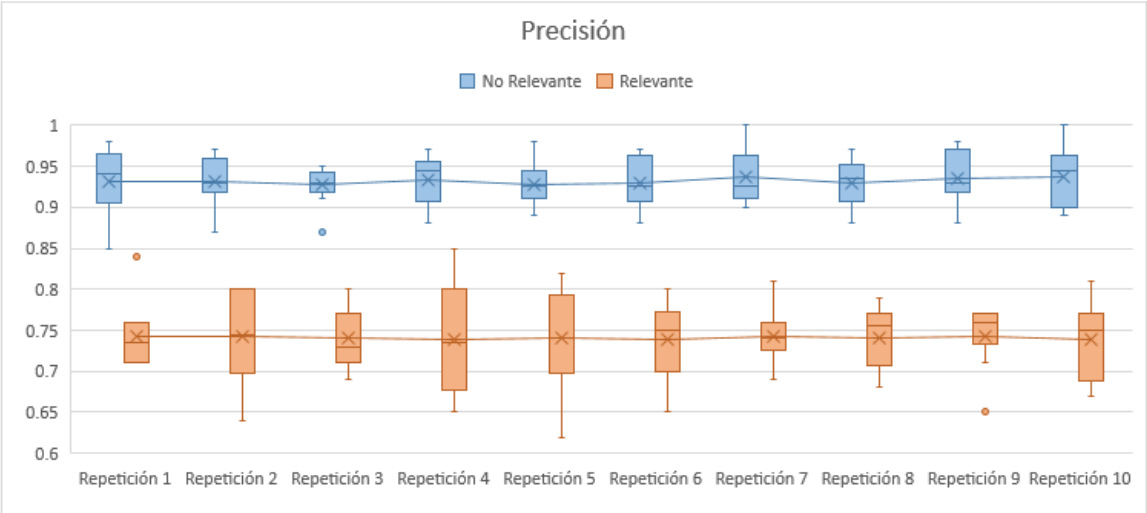


Figura 22: Diagrama de cajas de la Precisión en las 10 repeticiones de Validación Cruzada, según categoría

En la Figura 23, el promedio de la Sensibilidad en la serie de la clase Relevante es muy estable, se encuentra alrededor de 0.95. Sin embargo, está distanciado del promedio de la serie de la clase No Relevante, el cual tiene un rango entre 0.65 y 0.7.

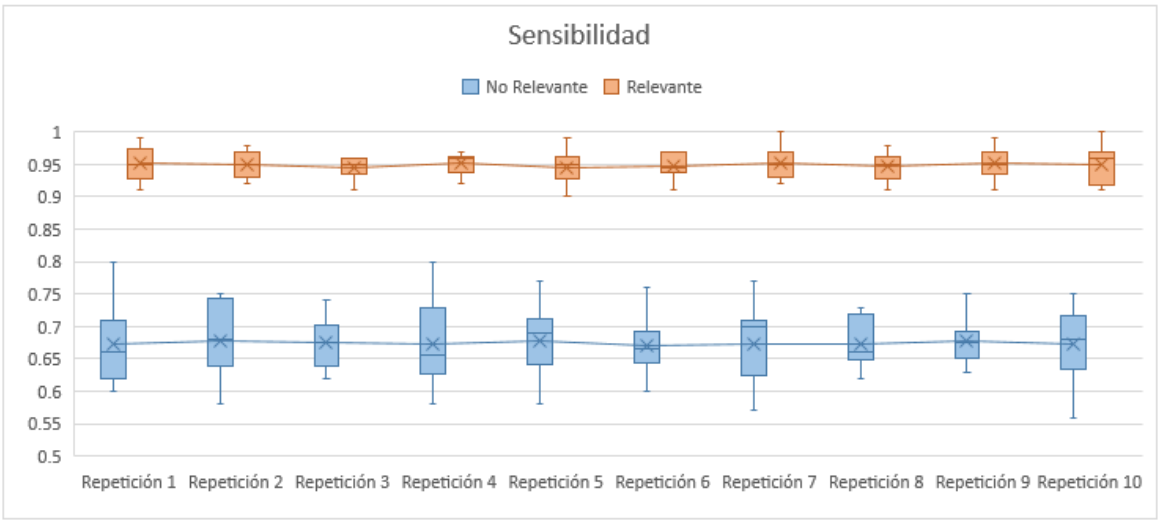


Figura 23: Diagrama de cajas de la Sensibilidad en las 10 repeticiones de Validación Cruzada, según categoría

En la Figura 24 se observa lo opuesto a la Figura 18, debido a que se muestran los resultados de la Especificidad y Sensibilidad, respectivamente. Por lo tanto, el promedio de la Especificidad en la serie de la clase No Relevante se encuentra muy estable, oscila alrededor de 0.95, y está distanciado del promedio de la serie de la clase Relevante, el cual presenta un rango entre 0.65 y 0.7.

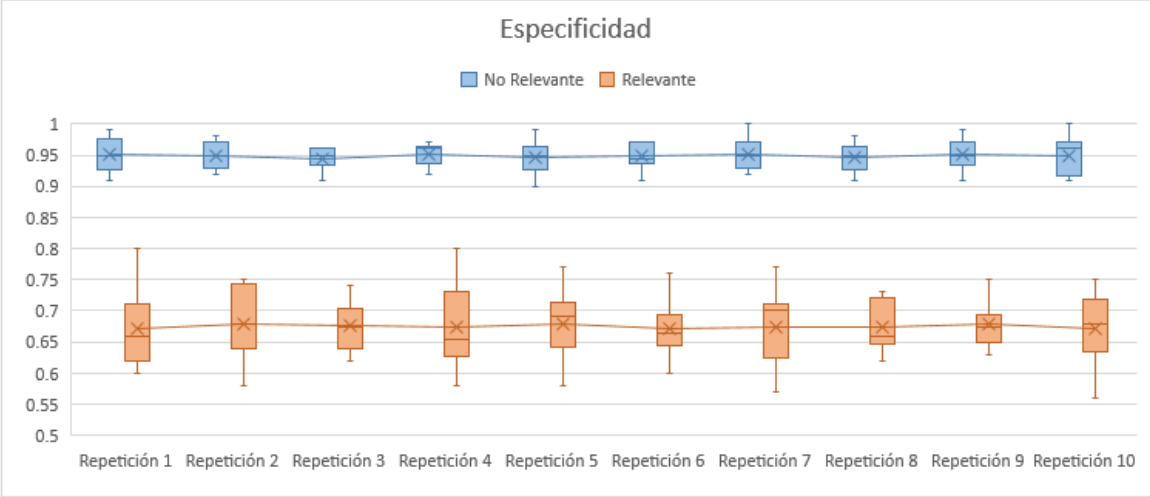


Figura 24: Diagrama de cajas de la Especificidad en las 10 repeticiones de Validación Cruzada, según categoría

En la Figura 25 se observa estabilidad en el promedio de la Medida F1 de ambas series de cajas. El promedio de la serie de la categoría Relevante se encuentra en el rango de 0.8 a 0.85, y el rango del promedio de la serie de la categoría No Relevante se encuentra entre 0.75 y 0.8.

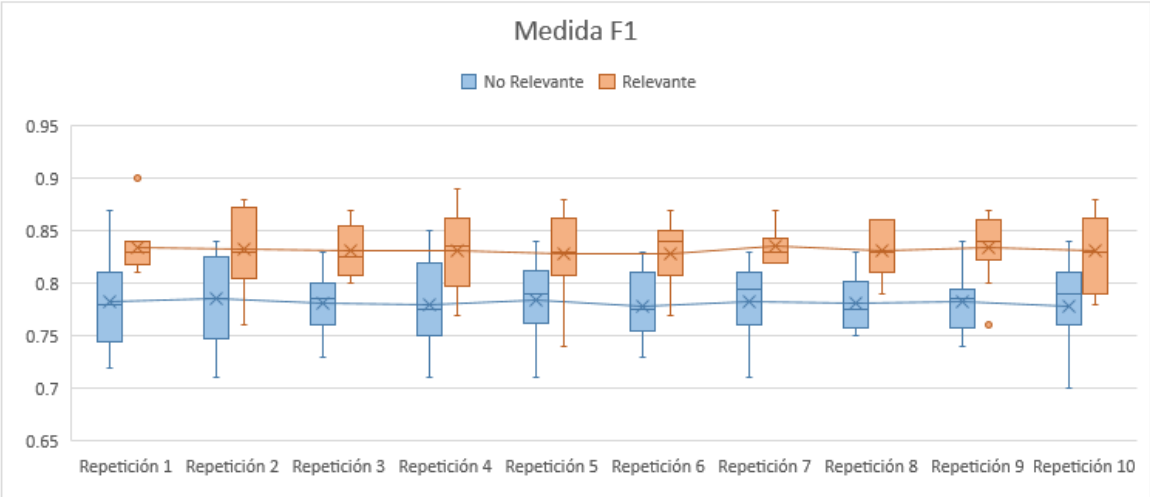


Figura 25: Diagrama de cajas de la Medida F1 en las 10 repeticiones de Validación Cruzada, según categoría

Las Figuras 26, 27, 28, 29 y 30 presentan dos series de cajas para cada métrica, donde la serie que contiene cajas azules representa los resultados del promedio simple entre las dos categorías de cada ejecución del modelo, mientras que la serie que contiene cajas anaranjadas representa los resultados correspondientes al promedio ponderado entre las dos categorías de cada ejecución del modelo.

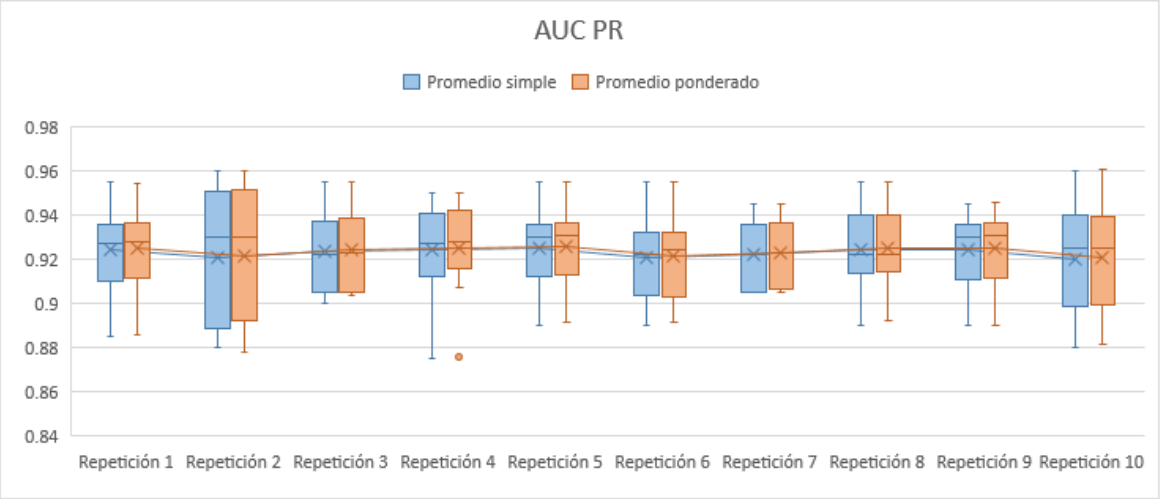


Figura 26: Diagrama de cajas del AUC PR en las 10 repeticiones de Validación Cruzada, según promedio simple y ponderado de las categorías

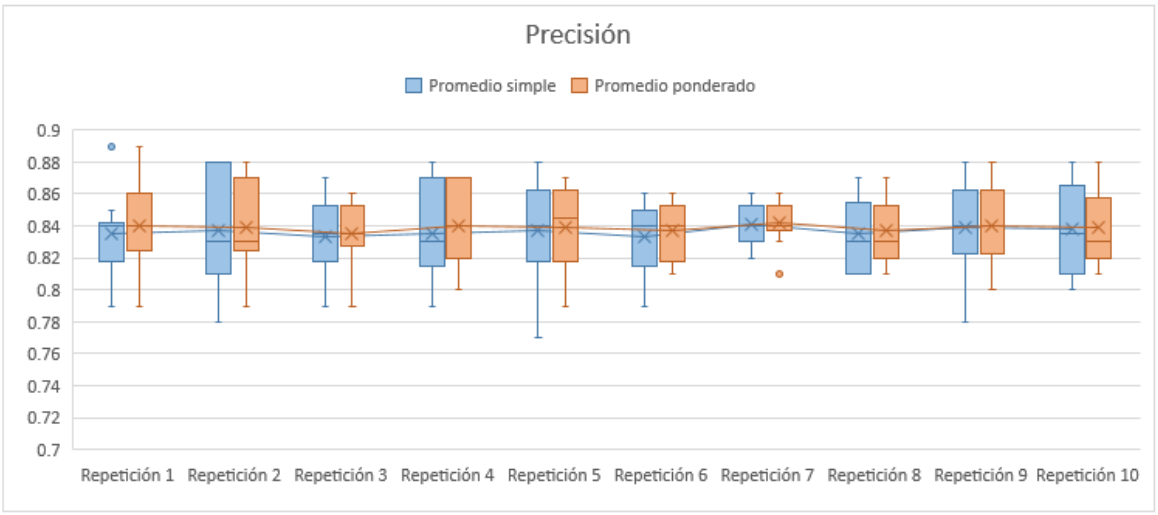


Figura 27: Diagrama de cajas de la Precisión en las 10 repeticiones de Validación Cruzada, según promedio simple y ponderado de las categorías

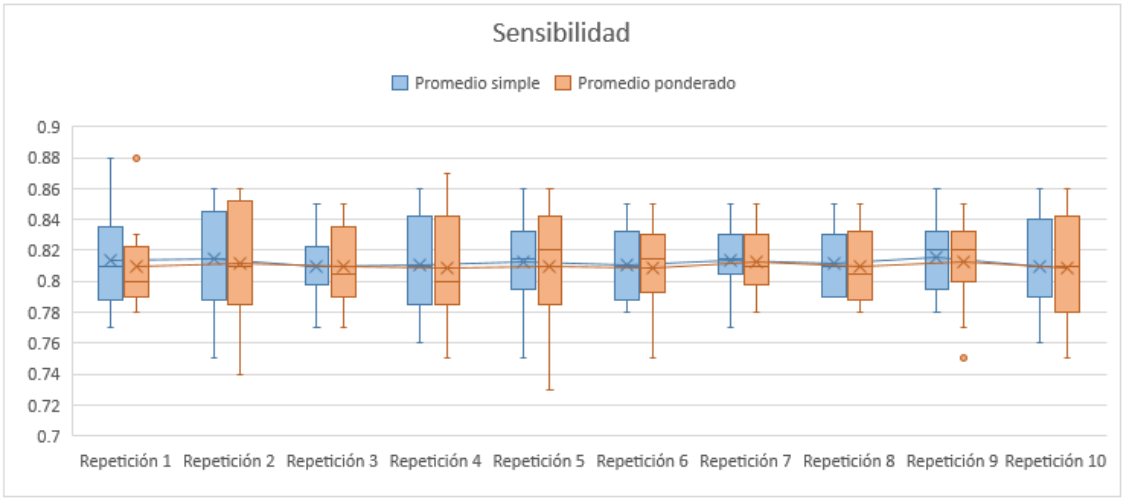


Figura 28: Diagrama de cajas de la Sensibilidad en las 10 repeticiones de Validación Cruzada, según promedio simple y ponderado de las categorías

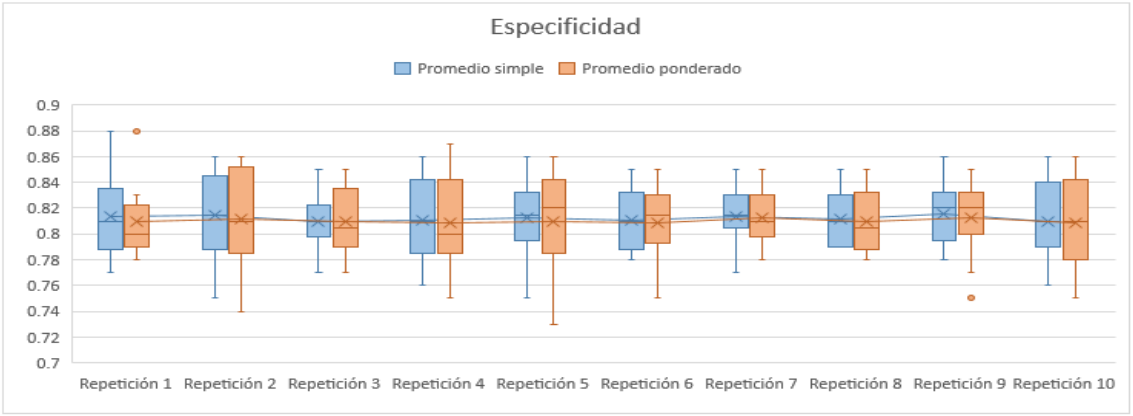


Figura 29: Diagrama de cajas de la Especificidad en las 10 repeticiones de Validación Cruzada, según promedio simple y ponderado de las categorías

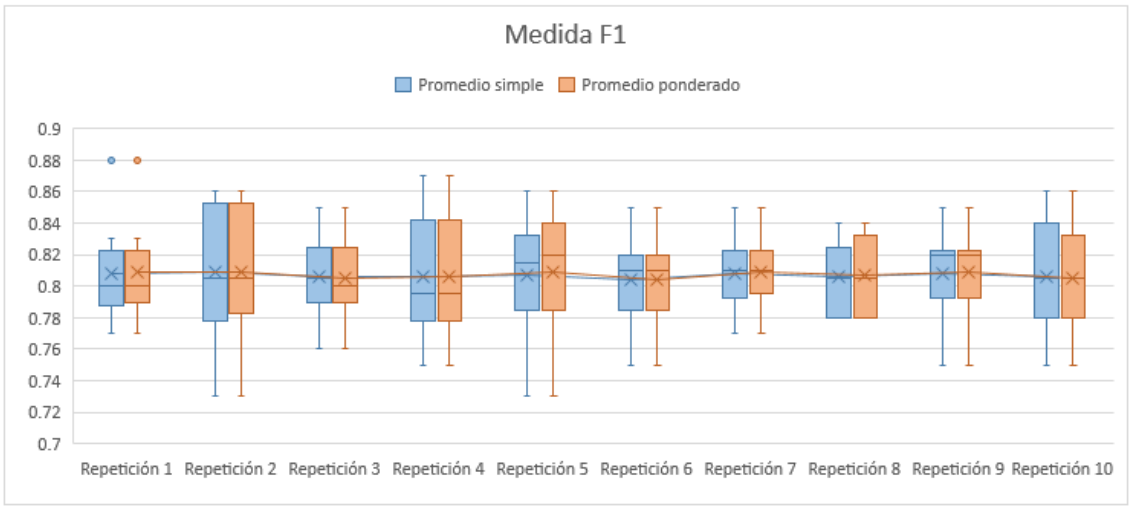


Figura 30: Diagrama de cajas de la Medida F1 en las 10 repeticiones de Validación Cruzada, según promedio simple y ponderado de las categorías

En las Figuras 26, 27, 28, 29 y 30 se puede observar que las dos series son muy similares, tanto en el promedio total de cada caja, como en la distribución de sus datos internos. En consecuencia, el promedio total de cada repetición a lo largo de ambas series es muy similar; asimismo, se encuentran estables. Estas características indican que los datos de las sub-muestras en las 10 repeticiones fueron balanceados.

El promedio total del AUC PR en cada repetición, tanto para la serie del promedio simple, como para la serie del promedio ponderado, oscila alrededor de 0.92. Para el caso de la Precisión, los promedios totales oscilan alrededor de 0.84; mientras que para la Sensibilidad, Especificidad y Medida F1, el promedio total de cada serie se encuentra en el rango de 0.8 a 0.82.

4.6 Clasificación de nuevos comentarios

En esta sexta etapa se clasificaron cinco nuevos comentarios. Previamente, se realizó el pre-procesamiento de los datos textuales y se obtuvo la matriz documento-término para hallar los valores TF-IDF de las variables explicativas. La clasificación de los nuevos comentarios se observa en detalle en la Tabla 14.

Tabla 14: Clasificación de nuevos comentarios

Nro.	Comentario	X_{aporte}	X_{area}	$X_{estudiante}$	$X_{gracias}$	$X_{interesante}$	$X_{saludos}$	$\hat{\pi}$	Clasificación
1	Muy interesantes los aportes profesor. Saludos desde Jauja.	0.33333	0	0	0	0.33333	0.33333	4.8E-09	No relevante
2	¿Cómo trabajar los enfoques transversales?	0	0	0	0	0	0	0.56687	Relevante
3	¿Se implementará más áreas en la plataforma?	0	0.37797	0	0	0	0	0.99997	Relevante
4	Excelente exposición maestro, gracias.	0	0	0	0.37797	0	0	2.5E-05	No relevante
5	¿Qué hacer con los estudiantes que aún no han sido contactados?	0	0	0.33333	0	0	0	0.97768	Relevante

V. CONCLUSIONES

1. Se obtuvieron indicadores descriptivos de los comentarios utilizando n-gramas para examinar la temática de las categorías “Relevante” y “No Relevante”, la cual fue consultas, sugerencias y reclamos de la estrategia “Aprendo en casa” para la categoría Relevante, y saludos, felicitaciones, agradecimientos y apreciación del *streaming* para la categoría No Relevante.
2. Se realizó la transformación de datos no estructurados (texto) a datos estructurados mediante el método de representación de texto TF-IDF para poder implementar los datos obtenidos de forma estructurada. De esta manera, se logró implementar un modelo de regresión logística binaria para la clasificación automatizada de los comentarios de los *streamings* de orientación a docentes sobre la estrategia “Aprendo en casa”.
3. Se evaluó el modelo de clasificación mediante las métricas de Exactitud, Precisión, Sensibilidad, Especificidad, Medida F1, AUC ROC (área bajo la curva ROC) y AUC PR (área bajo la curva de Precisión-Recall) para estimar el rendimiento del clasificador, obteniendo resultados en un rango de muy bueno a excelente.
4. Se evaluó también el modelo de clasificación a través de la Validación Cruzada para comprobar la validez de los resultados de las métricas de evaluación del modelo, donde se obtuvo resultados muy similares a los mostrados en los datos de prueba, concluyendo que no existe sobreajuste en el modelo y los resultados de las métricas de evaluación son confiables.
5. La implementación de un modelo de regresión logística utilizando datos textuales transformados mediante el método de representación de texto TF-IDF es adecuada para la clasificación de comentarios de los *streamings* de orientación sobre la estrategia “Aprendo en casa”; esto posibilita la reducción de los tiempos de análisis de data textual y, de esta manera, mejora la eficiencia en el equipo de la DIFODS, propiciando una mejor preparación del contenido en los nuevos *streamings* para absolver dudas y brindar mejores orientaciones a los maestros sobre la estrategia “Aprendo en casa”.

VI. RECOMENDACIONES

1. Probar otras estrategias de pre-procesamiento de datos textuales, dado que el nivel de minuciosidad en la limpieza y estandarización afecta directamente la calidad de la clasificación.
2. Comparar el método de representación de texto TF-IDF con otros métodos para representar numéricamente los datos textuales en modelos de clasificación, por ejemplo: Word2Vec, Glove y ELMo.
3. Comparar la Prueba Chi Cuadrado con otros métodos de selección de variables para mejorar los resultados de la clasificación; estos métodos pueden ser: Información Mutua, Ganancia de Información, Separación Bi-Normal, Entropía Cruzada Esperada, Índice de Gini, Fuerza del Término, Clasificación Basada en la Entropía y Contribución del Término.
4. Comparar el modelo de regresión Logística con otros modelos de clasificación con el objetivo de obtener mejores resultados, como Naïve Bayes, Máquinas de Soporte Vectorial y K-vecinos más cercanos.
5. Probar la regularización Ridge (L2) en el modelo de clasificación como alternativa de solución ante el problema de multicolinealidad que se puede presentar en las variables.
6. Los datos textuales del presente estudio se pueden utilizar para otros análisis de texto, por ejemplo, los datos de la categoría No Relevante se pueden usar para el Análisis de Sentimientos.

VII. BIBLIOGRAFÍA

Acosta Zúñiga, E. R. (2020). *Estudio de técnicas de ultrasonido cuantitativo aplicadas a la detección de inflamación o IFTA en riñones trasplantados* (tesis de pregrado, Pontificia Universidad Católica del Perú). Recuperada de Repositorio PUCP. https://tesis.pucp.edu.pe/repositorio/handle/20.500.12404/18239

Aldas, J., & Jimenez, E. (2017). *Análisis multivariante aplicado con R.* Ediciones Paraninfo.

ANDINA. (2020a). *Coronavirus: El Gobierno declara la emergencia sanitaria a escala nacional por 90 días*. https://andina.pe/agencia/noticia-coronavirus-gobierno-declara-emergencia-sanitaria-a-escala-nacional-90-dias-787949.aspx

ANDINA. (2020b). *Cronología del coronavirus en el Perú*. https://andina.pe/agencia/interactivo-cronologia-del-coronavirus-el-peru-488.aspx

ANDINA. (2020c). *Presidente Vizcarra anuncia postergación del inicio del año escolar ante coronavirus*. https://andina.pe/agencia/noticia-presidente-vizcarra-anuncia-postergacion-del-inicio-del-ano-escolar-ante-coronavirus-787851.aspx

Bafna, P., Pramod, D., & Vaidya, A. (2016). Document clustering: TF-IDF approach. *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 61–66.

Chen, K., Zhang, Z., Long, J., & Zhang, H. (2016). Turning from TF-IDF to TF-IGM for term weighting in text classification. *Expert Systems with Applications*, *66*, 245–260. https://doi.org/10.1016/j.eswa.2016.09.009

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics, 21*(6). https://doi.org/10.1186/s12864-019-6413-7

Cramer, J. S. (2002). *The origins of logistic regression*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=360300

Cull, B. W. (2011). *Reading revolutions: Online digital text and implications for reading in academe*. First Monday.

Da San Martino, G., Barron-Cedeno, A., & Nakov, P. (2019). Findings of the nlp4if-2019 shared task on fine-grained propaganda detection. *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, 162–170.

Da San Martino, G., Yu, S., Barrón-Cedeno, A., Petrov, R., & Nakov, P. (2019). Fine-grained analysis of propaganda in news article. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 5640–5650.

Feldman, R., & Sanger, J. (2007). *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press.

Gaydhani, A., Doma, V., Kendre, S., & Bhagwat, L. (2018). *Detecting hate speech and offensive language on twitter using machine learning: An n-gram and tfidf based approach*. ArXiv Preprint ArXiv:1809.08651: https://arxiv.org/abs/1809.08651

Gebre, B. G., Zampieri, M., Wittenburg, P., & Heskes, T. (2013). Improving native language identification with tf-idf weighting. *The 8th NAACL Workshop on Innovative Use of NLP for Building Educational Applications (BEA8)* (pp. 216–223). Association for Computational Linguistics.

Hosmer, D., & Lemeshow, S. (2000). *Applied Logistic Regression.* John Wiley & Sons, Inc.

Hvitfeldt, E., & Silge, J. (2020). *Supervised Machine Learning for Text Analysis in R*. https://smltar.com/

Iivari, N., Sharma, S., & Ventä-Olkkonen, L. (2020). Digital transformation of everyday life–How COVID-19 pandemic transformed the basic education of the young generation and why information management research should care? *International Journal of Information Management*, *55*, 102183. https://doi.org/10.1016/j.ijinfomgt.2020.102183

IPE. (2020). *EDUCACIÓN EN LOS TIEMPOS DEL COVID-19*. 1. https://www.ipe.org.pe/portal/educacion-en-los-tiempos-del-covid-19-aprendo-en-casa/

Jurafsky, D., & Martin, J. (2021). *Speech and Language Processing.* https://web.stanford.edu/~jurafsky/slp3/ed3book.pdf

Khalaf, H., & Zaman, R. (2015). Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). *Computer Science, Communication & Instrumentation Devices*, 163-172. http://dx.doi.org/10.3850/978-981-09-5247-1_017

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling.* Springer.

Kumar, A., & Paul, A. (2016). *Mastering Text Mining with R.* Packt Publishing.

Li, J., Qi'na, F., & Kou, Z. (2007). Keyword extraction based on tf/idf for Chinese news document. *Wuhan University Journal of Natural Sciences, 12*, 917–921 .

Li, J., Ye, Z., & Xiao, L. (2019). Detection of propaganda using logistic regression. *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, 119–124.

Luhn, H. (1957). A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development, 1*(4), 309 - 317. https://doi.org/10.1147/rd.14.0309

Manning, C., Raghavan, P., & Schütze, H. (2009). *An introduction to information retrieval.* Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space.* ArXiv Preprint ArXiv:1301.3781: https://arxiv.org/abs/1301.3781

Ministerio de Educación [MINEDU]. (2020a). *¿Qué es Aprendo en casa y cómo funciona?* Aprendo en Casa, 3: https://resources.aprendoencasa.pe/perueduca/orientaciones/familia/familia-orientaciones-que-es-aprendo-en-casa.pdf

Ministerio de Educación [MINEDU]. (2020b). *Orientaciones para implementar la estrategia Aprendo en Casa en el nivel de educación inicial.* Aprendo En Casa, 12: http://www.dreapurimac.gob.pe/inicio/images/archiv-2020/com/Orientaciones-inicial.pdf

Myers, R., Montgomery, D., Vining, G., & Robinson, T. (2010). *Generalized Linear Models: With Applications in Engineering and the Sciences.* John Wiley & Sons.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations.* ArXiv Preprint ArXiv:1802.05365: https://aclanthology.org/N18-1202

Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, *5*(2), 221-232. http://dx.doi.org/10.22364/bjmc.2017.5.2.05

RPP. (2020). *¿Cómo afecta la brecha digital a la educación remota?* https://rpp.pe/peru/actualidad/como-afecta-la-brecha-digital-a-la-educacion-remota-noticia-1267377?ref=rpp

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. " O'Reilly Media, Inc."

Sparck, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation, 28*(1), 11-21. https://doi.org/10.1108/eb026526

Tharwat, A. (2021). "Classification assessment methods. *Applied Computing and Informatics, 17*(1), 168-192. https://doi.org/10.1016/j.aci.2018.08.003

Trstenjak, B., Mikac, S., & Donko, D. (2014). KNN with TF-IDF based framework for text categorization. *Procedia Engineering*, *69*, 1356–1364. https://doi.org/10.1016/j.proeng.2014.03.129

UNESCO. (2020). *La educación en tiempos de la pandemia de COVID-19*. http://www.grade.org.pe/creer/recurso/la-educacion-en-tiempos-de-la-pandemia-de-covid-19/

van Zaanen, M., & Kanters, P. (2010). *Automatic mood classification using tf*idf based on lyrics.* https://pure.uvt.nl/ws/files/1299515/ismir_Zaanen.pdf

Wu, Z., Lin, W., & Ji, Y. (2018). An integrated ensemble learning model for imbalanced fault diagnostics and prognostics. *IEEE Access, 6*, 8394-8402. https://doi.org/10.1109/ACCESS.2018.2807121

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert Systems with Applications*, *38*(3), 2758–2765. https://doi.org/10.1016/j.eswa.2010.08.066

Žižka, J., Dařena, F., & Svoboda, A. (2019). *Text Mining with Machine Learning: Principles and Techniques*. CRC Press.

VIII. ANEXOS

Anexo 1: Ejemplo de aplicación del método de representación de texto TF-IDF

Se tienen los siguientes comentarios:

Nro.	Comentarios
1	Buenas noches. Excelente ponencia, profesor.
2	Muy agradecido por el soporte pedagógico. Muy valioso.
3	Importante aporte para nuestro trabajo.
4	Gracias por su gran aporte a nuestro trabajo docente.
5	Excelente aporte sobre retroalimentación, muy práctico y útil.
6	Muy importante la capacitación, claro y preciso.

En la aplicación del método TF-IDF, un documento de texto puede ser un conjunto de párrafos, un solo párrafo, un conjunto de oraciones, o incluso una sola oración. En este ejemplo se considerará a un comentario como un documento, de manera que se tienen seis documentos.

En primer lugar, se realiza una limpieza de datos, la cual consiste en la supresión de signos de puntuación (puntos y comas) y stopwords hallados en los seis documentos.

Nro.	Documentos
1	Buenas noches Excelente ponencia profesor
2	agradecido soporte pedagógico valioso
3	Importante aporte trabajo
4	Gracias gran aporte trabajo docente
5	Excelente aporte retroalimentación práctico útil
6	importante capacitación claro preciso

Luego, se tokenizan los documentos en unigramas, bigramas y trigramas:

Unigramas						
Buenas	profesor	valioso	Gracias	docente	práctico	claro
noches	agradecido	Importante	gran	Excelente	útil	preciso
Excelente	soporte	aporte	aporte	aporte	importante	
ponencia	pedagógico	trabajo	trabajo	retroalimentación	capacitación	

Bigramas		
Buenas noches	Importante aporte	aporte retroalimentación
noches Excelente	aporte trabajo	retroalimentación práctico
Excelente ponencia	Gracias gran	práctico útil
ponencia profesor	gran aporte	importante capacitación
agradecido soporte	aporte trabajo	capacitación claro
soporte pedagógico	trabajo docente	claro preciso
pedagógico valioso	Excelente aporte	

Trigramas	
Buenas noches Excelente	gran aporte trabajo
noches Excelente ponencia	aporte trabajo docente
Excelente ponencia profesor	Excelente aporte retroalimentación
agradecido soporte pedagógico	aporte retroalimentación práctico
soporte pedagógico valioso	retroalimentación práctico útil
Importante aporte trabajo	importante capacitación claro
Gracias gran aporte	capacitación claro preciso

A continuación, se procede a calcular los valores TF-IDF para cada n-grama utilizando las siguientes fórmulas:

$$\text{Valor TF} = \frac{\text{Frecuencia del token en el documento}}{\text{Número total de tokens en el documento}}$$

$$\text{Valor IDF} = \ln\left(\frac{\text{Número total de documentos}}{\text{Número de documentos que contienen el token}}\right)$$

$$\text{Valor TF-IDF} = \text{Valor TF} * \text{Valor IDF}$$

Algunos n-gramas se repiten (resaltados en amarillo), este detalle será tomado en cuenta en el cálculo del componente IDF.

Token (N-grama)	Frecuencia del token en el documento	Nro. total de tokens en el documento	Valor TF	Nro. de docs. que contienen el token	Valor IDF	Valor TF-IDF
Buenas	1	5	0.2	1	1.791759	0.358352
noches	1	5	0.2	1	1.791759	0.358352
Excelente	1	5	0.2	2	1.098612	0.219722
ponencia	1	5	0.2	1	1.791759	0.358352
profesor	1	5	0.2	1	1.791759	0.358352
agradecido	1	4	0.25	1	1.791759	0.447940
soporte	1	4	0.25	1	1.791759	0.447940
pedagógico	1	4	0.25	1	1.791759	0.447940
valioso	1	4	0.25	1	1.791759	0.447940
Importante	1	3	0.33333	2	1.098612	0.366204
aporte	1	3	0.33333	3	0.693147	0.231049
trabajo	1	3	0.33333	2	1.098612	0.366204
Gracias	1	5	0.2	1	1.791759	0.358352
gran	1	5	0.2	1	1.791759	0.358352
aporte	1	5	0.2	3	0.693147	0.138629
trabajo	1	5	0.2	2	1.098612	0.219722
docente	1	5	0.2	1	1.791759	0.358352
Excelente	1	5	0.2	2	1.098612	0.219722
aporte	1	5	0.2	3	0.693147	0.138629
retroalimentación	1	5	0.2	1	1.791759	0.358352
práctico	1	5	0.2	1	1.791759	0.358352
útil	1	5	0.2	1	1.791759	0.358352
importante	1	4	0.25	2	1.098612	0.274653
capacitación	1	4	0.25	1	1.791759	0.447940
claro	1	4	0.25	1	1.791759	0.447940
preciso	1	4	0.25	1	1.791759	0.447940
Buenas noches	1	4	0.25	1	1.791759	0.447940
noches Excelente	1	4	0.25	1	1.791759	0.447940
Excelente ponencia	1	4	0.25	1	1.791759	0.447940
ponencia profesor	1	4	0.25	1	1.791759	0.447940
agradecido soporte	1	3	0.33333	1	1.791759	0.597253
soporte pedagógico	1	3	0.33333	1	1.791759	0.597253
pedagógico valioso	1	3	0.33333	1	1.791759	0.597253
Importante aporte	1	2	0.5	1	1.791759	0.895880
aporte trabajo	1	2	0.5	2	1.098612	0.549306
Gracias gran	1	4	0.25	1	1.791759	0.447940
gran aporte	1	4	0.25	1	1.791759	0.447940
aporte trabajo	1	4	0.25	2	1.098612	0.274653
trabajo docente	1	4	0.25	1	1.791759	0.447940
Excelente aporte	1	4	0.25	1	1.791759	0.447940
aporte retroalimentación	1	4	0.25	1	1.791759	0.447940

retroalimentación práctico	1	4	0.25	1	1.791759	0.447940
práctico útil	1	4	0.25	1	1.791759	0.447940
importante capacitación	1	3	0.33333	1	1.791759	0.597253
capacitación claro	1	3	0.33333	1	1.791759	0.597253
claro preciso	1	3	0.33333	1	1.791759	0.597253
Buenas noches Excelente	1	3	0.33333	1	1.791759	0.597253
noches Excelente ponencia	1	3	0.33333	1	1.791759	0.597253
Excelente ponencia profesor	1	3	0.33333	1	1.791759	0.597253
agradecido soporte pedagógico	1	2	0.5	1	1.791759	0.895880
soporte pedagógico valioso	1	2	0.5	1	1.791759	0.895880
Importante aporte trabajo	1	1	1	1	1.791759	1.791759
Gracias gran aporte	1	3	0.33333	1	1.791759	0.597253
gran aporte trabajo	1	3	0.33333	1	1.791759	0.597253
aporte trabajo docente	1	3	0.33333	1	1.791759	0.597253
Excelente aporte retroalimentación	1	3	0.33333	1	1.791759	0.597253
aporte retroalimentación práctico	1	3	0.33333	1	1.791759	0.597253
retroalimentación práctico útil	1	3	0.33333	1	1.791759	0.597253
importante capacitación claro	1	2	0.5	1	1.791759	0.895880
capacitación claro preciso	1	2	0.5	1	1.791759	0.895880

Finalmente, se construye la matriz documento - término, donde cada columna corresponde a un token (n-grama) ordenado alfabéticamente, y cada fila corresponde a un documento. Para una mejor visualización de la matriz documento - término en el presente trabajo de investigación, la matriz elaborada se encuentra transpuesta. Se aprecia que es una matriz dispersa: solo el 18.5 % de la matriz contiene datos diferentes de cero. Este porcentaje puede reducirse a 1 % en matrices que representen mayor cantidad de documentos.

Token	Doc. 1	Doc. 2	Doc. 3	Doc. 4	Doc. 5	Doc. 6
agradecido	0	0.4479399	0	0	0	0
agradecido soporte	0	0.5972532	0	0	0	0
agradecido soporte pedagógico	0	0.8958797	0	0	0	0
aporte	0	0	0.2310491	0.1386294	0.1386294	0

aporte retroalimentación	0	0	0	0	0.4479399	0
aporte retroalimentación práctico	0	0	0	0	0.5972532	0
aporte trabajo	0	0	0.5493061	0.2746531	0	0
aporte trabajo docente	0	0	0	0.5972532	0	0
Buenas	0.3583519	0	0	0	0	0
Buenas noches	0.4479399	0	0	0	0	0
Buenas noches Excelente	0.5972532	0	0	0	0	0
capacitación	0	0	0	0	0	0.4479399
capacitación claro	0	0	0	0	0	0.5972532
capacitación claro preciso	0	0	0	0	0	0.8958797
claro	0	0	0	0	0	0.4479399
claro preciso	0	0	0	0	0	0.5972532
docente	0	0	0	0.3583519	0	0
Excelente	0.2197225	0	0	0	0.2197225	0
Excelente aporte	0	0	0	0	0.4479399	0
Excelente aporte retroalimentación	0	0	0	0	0.5972532	0
Excelente ponencia	0.4479399	0	0	0	0	0
Excelente ponencia profesor	0.5972532	0	0	0	0	0
Gracias	0	0	0	0.3583519	0	0
Gracias gran	0	0	0	0.4479399	0	0
Gracias gran aporte	0	0	0	0.5972532	0	0
gran	0	0	0	0.3583519	0	0
gran aporte	0	0	0	0.4479399	0	0
gran aporte trabajo	0	0	0	0.5972532	0	0
Importante	0	0	0.3662041	0	0	0.2746531
Importante aporte	0	0	0.8958797	0	0	0
Importante aporte trabajo	0	0	1.7917595	0	0	0
importante capacitación	0	0	0	0	0	0.5972532
importante capacitación claro	0	0	0	0	0	0.8958797
noches	0.3583519	0	0	0	0	0
noches Excelente	0.4479399	0	0	0	0	0
noches Excelente ponencia	0.5972532	0	0	0	0	0
pedagógico	0	0.4479399	0	0	0	0
pedagógico valioso	0	0.5972532	0	0	0	0
ponencia	0.3583519	0	0	0	0	0

ponencia profesor	0.4479399	0	0	0	0	0
práctico	0	0	0	0	0.3583519	0
práctico útil	0	0	0	0	0.4479399	0
preciso	0	0	0	0	0	0.4479399
profesor	0.3583519	0	0	0	0	0
retroalimentación	0	0	0	0	0.3583519	0
retroalimentación práctico	0	0	0	0	0.4479399	0
retroalimentación práctico útil	0	0	0	0	0.5972532	0
soporte	0	0.4479399	0	0	0	0
soporte pedagógico	0	0.5972532	0	0	0	0
soporte pedagógico valioso	0	0.8958797	0	0	0	0
trabajo	0	0	0.3662041	0.2197225	0	0
trabajo docente	0	0	0	0.4479399	0	0
útil	0	0	0	0	0.3583519	0
valioso	0	0.4479399	0	0	0	0

Anexo 2: Lista completa de Stopwords

Nro.	Stopword	Nro.	Stopword	Nro.	Stopword	Nro.	Stopword	Nro.	Stopword
1	de	21	como	41	donde	61	algunos	81	algunas
2	la	22	más	42	quien	62	qué	82	algo
3	que	23	pero	43	desde	63	unos	83	nosotros
4	el	24	sus	44	todo	64	yo	84	mi
5	en	25	le	45	nos	65	otro	85	mis
6	y	26	ya	46	durante	66	otras	86	tú
7	a	27	o	47	todos	67	otra	87	te
8	los	28	este	48	uno	68	él	88	ti
9	del	29	sí	49	les	69	tanto	89	tu
10	se	30	porque	50	ni	70	esa	90	tus
11	las	31	esta	51	contra	71	estos	91	ellas
12	por	32	entre	52	otros	72	mucho	92	nosotras
13	un	33	cuando	53	ese	73	quienes	93	vosotros
14	para	34	muy	54	eso	74	nada	94	vosotras
15	con	35	sin	55	ante	75	muchos	95	os
16	no	36	sobre	56	ellos	76	cual	96	mío
17	una	37	también	57	e	77	poco	97	mía
18	su	38	me	58	esto	78	ella	98	míos
19	al	39	hasta	59	mí	79	estar	99	mías
20	lo	40	hay	60	antes	80	estas	100	tuyo

Nro.	Stopword	Nro.	Stopword	Nro.	Stopword	Nro.	Stopword	Nro.	Stopword
101	tuya	143	estabais	185	habrías	227	serás	269	tengáis
102	tuyos	144	estaban	186	habríamos	228	será	270	tengan
103	tuyas	145	estuve	187	habríais	229	seremos	271	tendré
104	suyo	146	estuviste	188	habrían	230	seréis	272	tendrás
105	suya	147	estuvo	189	había	231	serán	273	tendrá
106	suyos	148	estuvimos	190	habías	232	sería	274	tendremos
107	suyas	149	estuvisteis	191	habíamos	233	serías	275	tendréis
108	nuestro	150	estuvieron	192	habíais	234	seríamos	276	tendrán
109	nuestra	151	estuviera	193	habían	235	seríais	277	tendría
110	nuestros	152	estuvieras	194	hube	236	serían	278	tendrías
111	nuestras	153	estuviéramos	195	hubiste	237	era	279	tendríamos
112	vuestro	154	estuvierais	196	hubo	238	eras	280	tendríais
113	vuestra	155	estuvieran	197	hubimos	239	éramos	281	tendrían
114	vuestros	156	estuviese	198	hubisteis	240	erais	282	tenía
115	vuestras	157	estuvieses	199	hubieron	241	eran	283	tenías
116	esos	158	estuviésemos	200	hubiera	242	fui	284	teníamos
117	esas	159	estuvieseis	201	hubieras	243	fuiste	285	teníais
118	estoy	160	estuviesen	202	hubiéramos	244	fue	286	tenían
119	estás	161	estando	203	hubierais	245	fuimos	287	tuve
120	está	162	estado	204	hubieran	246	fuisteis	288	tuviste
121	estamos	163	estada	205	hubiese	247	fueron	289	tuvo
122	estáis	164	estados	206	hubieses	248	fuera	290	tuvimos
123	están	165	estadas	207	hubiésemos	249	fueras	291	tuvisteis
124	esté	166	estad	208	hubieseis	250	fuéramos	292	tuvieron
125	estés	167	he	209	hubiesen	251	fuerais	293	tuviera
126	estemos	168	has	210	habiendo	252	fueran	294	tuvieras
127	estéis	169	ha	211	habido	253	fuese	295	tuviéramos
128	estén	170	hemos	212	habida	254	fueses	296	tuvierais
129	estaré	171	habéis	213	habidos	255	fuésemos	297	tuvieran
130	estarás	172	han	214	habidas	256	fueseis	298	tuviese
131	estará	173	haya	215	soy	257	fuesen	299	tuvieses
132	estaremos	174	hayas	216	eres	258	siendo	300	tuviésemos
133	estaréis	175	hayamos	217	es	259	sido	301	tuvieseis
134	estarán	176	hayáis	218	somos	260	tengo	302	tuviesen
135	estaría	177	hayan	219	sois	261	tienes	303	teniendo
136	estarías	178	habré	220	son	262	tiene	304	tenido
137	estaríamos	179	habrás	221	sea	263	tenemos	305	tenida
138	estaríais	180	habrá	222	seas	264	tenéis	306	tenidos
139	estarían	181	habremos	223	seamos	265	tienen	307	tenidas
140	estaba	182	habréis	224	seáis	266	tenga	308	tened
141	estabas	183	habrán	225	sean	267	tengas		
142	estábamos	184	habría	226	seré	268	tengamos		

Anexo 3: N-gramas sin stopwords más frecuentes de cada categoría

- Categoría Relevante: Unigramas

Nro.	Unigrama	Frecuencia	Nro.	Unigrama	Frecuencia
1	cómo	201	41	caso	36
2	estudiantes	157	42	aprendizaje	33
3	tv	122	43	día	33
4	si	110	44	semana	33
5	retroalimentación	107	45	estrategia	32
6	evidencias	98	46	medio	32
7	casa	94	47	programación	32
8	hacer	94	48	temas	32
9	radio	93	49	competencias	30
10	buenas	90	50	grado	29
11	aprendo	86	51	realizar	29
12	actividades	81	52	alumnos	27
13	trabajo	72	53	ser	27
14	plataforma	71	54	mismo	26
15	noches	65	55	podemos	26
16	inicial	62	56	diferentes	25
17	puede	58	57	familia	25
18	sesiones	58	58	rural	25
19	web	58	59	tardes	25
20	favor	57	60	ejemplo	24
21	solo	56	61	internet	24
22	área	53	62	cada	23
23	niños	53	63	cuál	23
24	padres	52	64	poder	23
25	secundaria	51	65	zona	23
26	docente	50	66	sesión	22
27	medios	45	67	celular	21
28	deben	44	68	cuadernos	21
29	docentes	44	69	deberían	21
30	educación	44	70	enviar	21
31	estudiante	44	71	primaria	21
32	áreas	43	72	evaluación	20
33	debe	42	73	evidencia	20
34	gracias	42	74	muchas	20
35	nivel	40	75	acceso	19
36	clases	39	76	envían	19
37	conectividad	38	77	minedu	19
38	trabajar	38	78	podría	19
39	comunicación	37	79	puedo	19
40	retroalimentar	37	80	zonas	19

Nro.	Unigrama	Frecuencia	Nro.	Unigrama	Frecuencia
81	ejemplos	18	91	días	16
82	física	18	92	dice	16
83	hacemos	18	93	maestro	16
84	manera	18	94	mejor	16
85	pueden	18	95	planificación	16
86	quisiera	18	96	saber	16
87	tiempo	18	97	señal	16
88	tres	18	98	cuentan	15
89	whatsapp	18	99	dan	15
90	ningún	17	100	elaborar	15

- Categoría Relevante: Bigramas

Nro.	Bigrama	Frecuencia	Nro.	Bigrama	Frecuencia
1	buenas noches	65	26	programa aprendo	7
2	nivel inicial	28	27	puedo hacer	7
3	buenas tardes	25	28	retroalimentación si	7
4	zona rural	19	29	cómo hago	6
5	educación física	16	30	educación inicial	6
6	estrategia aprendo	14	31	mismas competencias	6
7	puede hacer	14	32	muchas gracias	6
8	cómo podemos	13	33	noches cómo	6
9	primer grado	13	34	radio tv	6
10	cómo puedo	12	35	aula virtual	5
11	cómo retroalimentar	12	36	año resolvemos	5
12	ningún medio	12	37	debe ser	5
13	plataforma aprendo	11	38	debemos hacer	5
14	quisiera saber	11	39	estudiantes cómo	5
15	zonas rurales	11	40	evidencias cómo	5
16	cómo hacemos	10	41	mismo grado	5
17	cómo trabajar	9	42	noches quisiera	5
18	deben ser	9	43	podemos hacer	5
19	trabajo remoto	9	44	resolvemos problemas	5
20	tv radio	9	45	actividades complementarias	4
21	ciencias sociales	8	46	cada semana	4
22	cómo hacer	8	47	canales comunitarios	4
23	cómo realizar	7	48	demás áreas	4
24	muchas veces	7	49	diferentes medios	4
25	plataforma web	7	50	educación religiosa	4

Nro.	Bigrama	Frecuencia	Nro.	Bigrama	Frecuencia
51	envían evidencias	4	76	cómo evaluar	3
52	estudiantes si	4	77	cómo podría	3
53	evaluación formativa	4	78	cómo podríamos	3
54	hacer sesiones	4	79	cómo retroalimento	3
55	hacer si	4	80	dar retroalimentación	3
56	inicial cómo	4	81	debe realizar	3
57	mismo día	4	82	educación primaria	3
58	mismos temas	4	83	educación secundaria	3
59	noches profesor	4	84	estudiantes solo	3
60	plataforma virtual	4	85	evidencia si	3
61	realizar retroalimentación	4	86	favor podría	3
62	saber cómo	4	87	guía docente	3
63	temas deben	4	88	instituciones educativas	3
64	tv perú	4	89	necesidades especiales	3
65	web cómo	4	90	nivel secundaria	3
66	web radio	4	91	noches maestro	3
67	aquellos estudiantes	3	92	norma dice	3
68	buen día	3	93	pasa si	3
69	buenos días	3	94	programación semanal	3
70	casa cómo	3	95	puede elaborar	3
71	clase virtual	3	96	puede ser	3
72	clases si	3	97	pueden enviar	3
73	cuarto grado	3	98	pueden hacer	3
74	cuánto tiempo	3	99	página web	3
75	cómo debemos	3	100	radio solo	3

- Categoría Relevante: Trigramas

Nro.	Trigrama	Frecuencia
1	buenas noches cómo	6
2	año resolvemos problemas	5
3	buenas noches quisiera	5
4	cómo puedo hacer	5
5	buenas noches profesor	4
6	quisiera saber cómo	4
7	temas deben ser	4
8	buenas noches maestro	3
9	cómo realizar retroalimentación	3
10	radio tv web	3
11	buenas noches si	2
12	cómo podemos hacer	2
13	cómo podemos tener	2
14	cómo podemos trabajar	2

Nro.	Trigrama	Frecuencia
15	cómo puedo acceder	2
16	cómo recojo evidencias	2
17	deben ser dialogadas	2
18	deben ser publicados	2
19	educación básica especial	2
20	estudiantes deben ser	2
21	estudiantes solo indican	2
22	nivel inicial cómo	2
23	noches quisiera saber	2
24	profesor buenas noches	2
25	puede hacer si	2
26	puede planificarse trabajar	2
27	quisiera saber si	2
28	saber cómo vamos	2
29	tv radio web	2
30	acciones realizadas gracias	1
31	acompañantes pedagógicos pregunto	1
32	actividad menos compleja	1
33	actividades deben compartir	1
34	actividades deben hacerlo	1
35	actividades diarias cómo	1
36	actividades semanales deberían	1
37	actividades significativas iniciales	1
38	además muchas veces	1
39	agregar razonamiento verbal	1
40	agreguen información adicional	1
41	aguilar carlos ie	1
42	alguien podría proponer	1
43	alguien puede publicar	1
44	alguna línea telefónica	1
45	alguna sugerencia respecto	1
46	algún producto realizado	1
47	almirante miguel grau	1
48	alumno cuánto tiempo	1
49	alumnos quiero decir	1
50	alumnos solo cuentan	1
51	ambos guardan relación	1
52	analizarlos cada vez	1
53	angulo buenas noches	1
54	antelación necesitamos adecuar	1
55	aprendiendo cómo cuidarnos	1

Nro.	Trigrama	Frecuencia
56	aprendizaje deben ser	1
57	aprendizaje integrando áreas	1
58	aprendizajes haciendo uso	1
59	areas cómo deben	1
60	arequipa quisiera saber	1
61	arte ccss dpcc	1
62	atrás ugel puno	1
63	atte roger tintaya	1
64	audio video videollamada	1
65	aula virtual aclare	1
66	autorizado hacer clases	1
67	año ningún estudiante	1
68	aún puedo desarrollar	1
69	borja cómo puedo	1
70	brindan espero puedan	1
71	brindar retroalimentación si	1
72	brinde ejemplos acerca	1
73	buen día agradecida	1
74	buen día sugiero	1
75	buen día suplico	1
76	buena coordinación cómo	1
77	buenas noches cierto	1
78	buenas noches colegas	1
79	buenas noches cuál	1
80	buenas noches debieran	1
81	buenas noches desearía	1
82	buenas noches felicitaciones	1
83	buenas noches gracias	1
84	buenas noches julio	1
85	buenas noches podría	1
86	buenas noches profesora	1
87	buenas noches quiero	1
88	buenas tardes ciencia	1
89	buenas tardes cómo	1
90	buenas tardes escribo	1
91	buenas tardes maestra	1
92	buenas tardes pregunta	1
93	buenas tardes quisiera	1
94	buenas tardes si	1
95	buenos días gracias	1
96	cada medio remoto	1
97	cada sábado envío	1
98	calificación certificadora podrían	1
99	campo real resulta	1

Nro.	Trigrama	Frecuencia
100	canales comunitarios si	1

- Categoría No Relevante: Unigramas

Nro.	Unigrama	Frecuencia	Nro.	Unigrama	Frecuencia
1	gracias	322	36	gran	19
2	interesante	134	37	remoto	19
3	retroalimentación	109	38	distancia	18
4	muchas	108	39	orientación	18
5	aportes	92	40	pedagógica	18
6	aporte	86	41	alcances	16
7	buenas	84	42	claro	16
8	noches	77	43	práctica	16
9	trabajo	69	44	buen	15
10	excelente	65	45	colegas	15
11	maestro	65	46	evidencias	15
12	importante	57	47	maestros	15
13	tema	54	48	ponencia	15
14	información	53	49	seguir	15
15	saludos	49	50	acuña	14
16	dudas	47	51	día	14
17	orientaciones	44	52	importantes	14
18	mejorar	40	53	poder	14
19	compartir	39	54	agradecida	13
20	docente	39	55	cómo	13
21	estudiantes	36	56	conocimientos	13
22	buena	35	57	espero	13
23	julio	35	58	ie	13
24	profesor	35	59	interesantes	13
25	labor	33	60	video	13
26	curso	28	61	apoyo	12
27	felicitaciones	27	62	diapositivas	12
28	docentes	25	63	estudiante	12
29	educación	23	64	bendiciones	11
30	ayuda	22	65	bueno	11
31	aclarar	21	66	fortalecer	11
32	aprendizaje	21	67	grabado	11
33	favor	20	68	oportunidad	11
34	bien	19	69	aclaraciones	10
35	exposición	19	70	ayudará	10

Nro.	Unigrama	Frecuencia	Nro.	Unigrama	Frecuencia
71	capacitación	10	86	cada	8
72	estrategia	10	87	casa	8
73	niños	10	88	claridad	8
74	pedagógico	10	89	educativa	8
75	puede	10	90	educativo	8
76	reto	10	91	excelentes	8
77	servirá	10	92	formativa	8
78	siempre	10	93	hacer	8
79	valioso	10	94	internet	8
80	explicación	9	95	mejora	8
81	nuevo	9	96	precisiones	8
82	si	9	97	puedo	8
83	tener	9	98	realizar	8
84	aportaciones	8	99	va	8
85	buenos	8	100	virtual	8

- Categoría No Relevante: Bigramas

Nro.	Bigrama	Frecuencia	Nro.	Bigrama	Frecuencia
1	muchas gracias	88	25	valioso aporte	5
2	buenas noches	77	26	aclarar muchas	4
3	gracias maestro	26	27	buena iniciativa	4
4	noches gracias	23	28	buena oportunidad	4
5	trabajo remoto	18	29	buenos aportes	4
6	gracias profesor	16	30	excelente gracias	4
7	muchas dudas	14	31	excelente información	4
8	maestro julio	12	32	felicitaciones maestro	4
9	profesor julio	12	33	feliz día	4
10	julio acuña	9	34	gran reto	4
11	labor docente	9	35	importante aporte	4
12	buen aporte	8	36	interesante tema	4
13	excelente aporte	8	37	ir mejorando	4
14	práctica pedagógica	8	38	noches maestro	4
15	interesante gracias	7	39	buena explicación	3
16	noches colegas	7	40	buenas tardes	3
17	labor pedagógica	6	41	buenos días	3
18	buen tema	5	42	clases virtuales	3
19	evaluación formativa	5	43	cómo hago	3
20	excelente exposición	5	44	educación remota	3
21	excelentes aportes	5	45	enseñanza aprendizaje	3
22	gran aporte	5	46	estrategia aprendo	3
23	interesante aporte	5	47	excelente ponencia	3
24	trabajo pedagógico	5	48	gracias bendiciones	3

Nro.	Bigrama	Frecuencia	Nro.	Bigrama	Frecuencia
49	gran ayuda	3	75	aportes brindados	2
50	labor educativa	3	76	aportes saludos	2
51	maestro excelente	3	77	bien claro	2
52	muchísimas gracias	3	78	bonita exposición	2
53	noches interesante	3	79	buena capacitación	2
54	poder mejorar	3	80	buena información	2
55	quede grabado	3	81	buena orientación	2
56	quehacer educativo	3	82	cada día	2
57	retroalimentación gracias	3	83	ciertas dudas	2
58	retroalimentación oportuna	3	84	colegas interesante	2
59	retroalimentación saludos	3	85	curso rol	2
60	saludos cordiales	3	86	cómo puedo	2
61	saludos maestro	3	87	cómo realizar	2
62	suma importancia	3	88	dejar grabado	2
63	tan importante	3	89	dr julio	2
64	tema gracias	3	90	dudas gracias	2
65	tema tan	3	91	dudas muchas	2
66	trabajo diario	3	92	estimado julio	2
67	abrazo maestro	2	93	excelente apoyo	2
68	aclarado muchas	2	94	excelente maestro	2
69	agradecida saludos	2	95	excelente tema	2
70	alfabetización digital	2	96	excelente trabajo	2
71	aportaciones maestro	2	97	excelentes alcances	2
72	aporte interesante	2	98	exposición ojalá	2
73	aporte muchas	2	99	felicitaciones profesor	2
74	aporte profesor	2	100	gracias colega	2

- Categoría No Relevante: Trigramas

Nro.	Trigrama	Frecuencia
1	buenas noches gracias	23
2	muchas gracias maestro	12
3	buenas noches colegas	7
4	gracias profesor julio	5
5	buenas noches maestro	4
6	aclarar muchas dudas	3
7	buenas noches interesante	3
8	maestro julio acuña	3
9	muchas gracias profesor	3
10	tema tan importante	3

Nro.	Trigrama	Frecuencia
11	abrazo maestro julio	2
12	aclarado muchas dudas	2
13	aporte muchas gracias	2
14	bonita exposición ojalá	2
15	dudas muchas gracias	2
16	excelente ponencia gracias	2
17	gracias maestro excelente	2
18	gracias maestro julio	2
19	hola buenas noches	2
20	interesante gracias maestro	2
21	julio buenas noches	2
22	maestro excelente exposición	2
23	maestro muchas gracias	2
24	muchas dudas gracias	2
25	muchas dudas muchas	2
26	profesor julio acuña	2
27	pueda dejar grabado	2
28	saludos maestro julio	2
29	abrazo colegas valemos	1
30	aclaraciones nuevas experiencias	1
31	aclaraciones tan puntuales	1
32	aclarado interesantes puntos	1
33	aclaran muchas dudas	1
34	aclarando muchas dudas	1
35	aclarar ciertas dudas	1
36	aclaró dudas seguiremos	1
37	aclaró muchas dudas	1
38	acompañamiento pedagógico agradece	1
39	agradecido maestro acuña	1
40	alcántara saniz importante	1
41	alfabetización digital espero	1
42	alto mirador rioja	1
43	amazonas ugel san	1
44	amigo julio interesantes	1
45	aportaciones maestro muchas	1
46	aportación felicitaciones maestro	1
47	aporte excelente apoyo	1
48	aporte gladys yparraguirre	1
49	aporte profesor julio	1
50	aporte quedé satisfecha	1
51	aportes brindados enriquece	1
52	aportes brindados mediante	1

Nro.	Trigrama	Frecuencia
53	aportes estimado profesor	1
54	aportes maestro fortalece	1
55	aportes muchas gracias	1
56	aportes ojalá quede	1
57	aportes profesor julio	1
58	aportes saludos cordiales	1
59	aportes tan necesarios	1
60	arequipa retroalimentando saberes	1
61	así darnos conocimientos	1
62	así podemos guiar	1
63	ayudado aclarar muchas	1
64	bastante creatividad sé	1
65	bendiciones miss yani	1
66	bertilde periche buen	1
67	bien amigo gracias	1
68	bien estimado prof	1
69	bien paty importante	1
70	brindan soportes útiles	1
71	brisas pertenezco ugel	1
72	buen aporte saludos	1
73	buen aporte sres	1
74	buen día saludos	1
75	buen tema ayudo	1
76	buen tema felicitaciones	1
77	buen tema retroalimentación	1
78	buena coordinación saludos	1
79	buena disertación compartan	1
80	buena explicación maestro	1
81	buena explicación muchas	1
82	buena noche grizela	1
83	buenas noches buen	1
84	buenas noches cecilia	1
85	buenas noches estimados	1
86	buenas noches excelente	1
87	buenas noches felicito	1
88	buenas noches gran	1
89	buenas noches gusto	1
90	buenas noches maestros	1
91	buenas noches muchas	1
92	buenas noches mándenos	1
93	buenas noches profesor	1
94	buenas noches profesores	1
95	buenas noches saludos	1
96	buenas noches solicitamos	1

Nro.	Trigrama	Frecuencia
97	buenas orientaciones despejaron	1
98	buenas precisiones gracias	1
99	buenas tardes gracias	1
100	buenas tardes interesante	1

Anexo 4: Resultados de las métricas de evaluación en cada repetición del proceso de Validación Cruzada

- Repetición 1

Repetición 1		
Sub-muestra de prueba	Métricas	
	Exactitud	AUC ROC
Sub-muestra 1	0.88	0.95
Sub-muestra 2	0.82	0.93
Sub-muestra 3	0.79	0.91
Sub-muestra 4	0.79	0.93
Sub-muestra 5	0.79	0.9
Sub-muestra 6	0.83	0.93
Sub-muestra 7	0.82	0.92
Sub-muestra 8	0.81	0.93
Sub-muestra 9	0.79	0.9
Sub-muestra 10	0.78	0.87

Repetición 1					
Sub-muestra de prueba	Métricas (Categoría Relevante)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.84	0.96	0.8	0.9	0.95
Sub-muestra 2	0.73	0.99	0.67	0.84	0.92
Sub-muestra 3	0.71	0.97	0.62	0.82	0.92
Sub-muestra 4	0.71	0.99	0.6	0.83	0.94
Sub-muestra 5	0.71	0.95	0.65	0.81	0.89
Sub-muestra 6	0.73	0.95	0.74	0.83	0.9
Sub-muestra 7	0.76	0.95	0.68	0.84	0.92
Sub-muestra 8	0.74	0.93	0.7	0.83	0.92
Sub-muestra 9	0.76	0.92	0.64	0.83	0.93
Sub-muestra 10	0.74	0.91	0.62	0.81	0.9

Repetición 1					
Sub-muestra de prueba	Métricas (Categoría No Relevante)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.95	0.8	0.96	0.87	0.96
Sub-muestra 2	0.98	0.67	0.99	0.8	0.95
Sub-muestra 3	0.96	0.62	0.97	0.76	0.93
Sub-muestra 4	0.98	0.6	0.99	0.75	0.94
Sub-muestra 5	0.93	0.65	0.95	0.77	0.93
Sub-muestra 6	0.96	0.74	0.95	0.84	0.96
Sub-muestra 7	0.93	0.68	0.95	0.79	0.93
Sub-muestra 8	0.92	0.7	0.93	0.8	0.94
Sub-muestra 9	0.86	0.64	0.92	0.73	0.89
Sub-muestra 10	0.85	0.62	0.91	0.72	0.87

Repetición 1					
Sub-muestra de prueba	Métricas (Promedio Simple entre las 2 categorías)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.89	0.88	0.88	0.88	0.955
Sub-muestra 2	0.85	0.83	0.83	0.82	0.935
Sub-muestra 3	0.84	0.8	0.8	0.79	0.925
Sub-muestra 4	0.84	0.79	0.79	0.79	0.94
Sub-muestra 5	0.82	0.8	0.8	0.79	0.91
Sub-muestra 6	0.84	0.85	0.85	0.83	0.93
Sub-muestra 7	0.84	0.82	0.82	0.82	0.925
Sub-muestra 8	0.83	0.82	0.82	0.81	0.93
Sub-muestra 9	0.81	0.78	0.78	0.78	0.91
Sub-muestra 10	0.79	0.77	0.77	0.77	0.885

Repetición 1					
Sub-muestra de prueba	Métricas (Promedio Ponderado entre las 2 categorías)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.89	0.88	0.88	0.88	0.9548
Sub-muestra 2	0.86	0.82	0.82	0.82	0.9359
Sub-muestra 3	0.84	0.79	0.79	0.79	0.9252
Sub-muestra 4	0.85	0.79	0.79	0.79	0.9400
Sub-muestra 5	0.83	0.79	0.79	0.79	0.9109
Sub-muestra 6	0.86	0.83	0.83	0.83	0.9345
Sub-muestra 7	0.84	0.82	0.82	0.82	0.9249
Sub-muestra 8	0.83	0.81	0.81	0.81	0.9305
Sub-muestra 9	0.81	0.79	0.79	0.79	0.9122
Sub-muestra 10	0.79	0.78	0.78	0.77	0.8862

- Repetición 2

Repetición 2		
Sub-muestra de prueba	Métricas	
	Exactitud	AUC ROC
Sub-muestra 1	0.86	0.96
Sub-muestra 2	0.74	0.89
Sub-muestra 3	0.86	0.94
Sub-muestra 4	0.81	0.92
Sub-muestra 5	0.79	0.93
Sub-muestra 6	0.77	0.88
Sub-muestra 7	0.84	0.92
Sub-muestra 8	0.81	0.89
Sub-muestra 9	0.79	0.89
Sub-muestra 10	0.85	0.95

Repetición 2					
Sub-muestra de prueba	Métricas (Categoría Relevante)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.8	0.97	0.75	0.88	0.96
Sub-muestra 2	0.64	0.93	0.58	0.76	0.86
Sub-muestra 3	0.8	0.96	0.75	0.87	0.95
Sub-muestra 4	0.74	0.93	0.7	0.83	0.93
Sub-muestra 5	0.71	0.95	0.65	0.81	0.92
Sub-muestra 6	0.66	0.97	0.61	0.79	0.84
Sub-muestra 7	0.78	0.94	0.74	0.85	0.93
Sub-muestra 8	0.74	0.95	0.66	0.83	0.88
Sub-muestra 9	0.75	0.92	0.65	0.83	0.86
Sub-muestra 10	0.8	0.98	0.7	0.88	0.96

Repetición 2					
Sub-muestra de prueba	Métricas (Categoría No Relevante)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.97	0.75	0.97	0.84	0.96
Sub-muestra 2	0.91	0.58	0.93	0.71	0.92
Sub-muestra 3	0.95	0.75	0.96	0.84	0.95
Sub-muestra 4	0.92	0.7	0.93	0.8	0.94
Sub-muestra 5	0.93	0.65	0.95	0.77	0.94
Sub-muestra 6	0.96	0.61	0.97	0.75	0.93
Sub-muestra 7	0.92	0.74	0.94	0.82	0.93
Sub-muestra 8	0.93	0.66	0.95	0.77	0.91
Sub-muestra 9	0.87	0.65	0.92	0.74	0.9
Sub-muestra 10	0.96	0.7	0.98	0.81	0.95

Repetición 2					
Sub-muestra de prueba	Métricas (Promedio Simple entre las 2 categorías)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.88	0.86	0.86	0.86	0.96
Sub-muestra 2	0.78	0.75	0.75	0.73	0.89
Sub-muestra 3	0.88	0.86	0.86	0.86	0.95
Sub-muestra 4	0.83	0.82	0.82	0.81	0.935
Sub-muestra 5	0.82	0.8	0.8	0.79	0.93
Sub-muestra 6	0.81	0.79	0.79	0.77	0.885
Sub-muestra 7	0.85	0.84	0.84	0.84	0.93
Sub-muestra 8	0.83	0.81	0.81	0.8	0.895
Sub-muestra 9	0.81	0.78	0.78	0.78	0.88
Sub-muestra 10	0.88	0.84	0.84	0.85	0.955

Repetición 2					
Sub-muestra de prueba	Métricas (Promedio Ponderado entre las 2 categorías)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.88	0.86	0.86	0.86	0.9600
Sub-muestra 2	0.79	0.74	0.74	0.73	0.8929
Sub-muestra 3	0.87	0.86	0.86	0.86	0.9500
Sub-muestra 4	0.83	0.81	0.81	0.81	0.9352
Sub-muestra 5	0.83	0.79	0.79	0.79	0.9305
Sub-muestra 6	0.83	0.77	0.77	0.76	0.8905
Sub-muestra 7	0.85	0.84	0.84	0.84	0.9300
Sub-muestra 8	0.83	0.81	0.81	0.8	0.8949
Sub-muestra 9	0.81	0.79	0.79	0.79	0.8783
Sub-muestra 10	0.87	0.85	0.85	0.85	0.9555

- Repetición 3

Repetición 3		
Sub-muestra de prueba	Métricas	
	Exactitud	AUC ROC
Sub-muestra 1	0.81	0.93
Sub-muestra 2	0.85	0.94
Sub-muestra 3	0.85	0.95
Sub-muestra 4	0.79	0.89
Sub-muestra 5	0.77	0.9
Sub-muestra 6	0.8	0.9
Sub-muestra 7	0.83	0.92
Sub-muestra 8	0.79	0.92
Sub-muestra 9	0.81	0.9
Sub-muestra 10	0.8	0.9

Repetición 3					
Sub-muestra de prueba	Métricas (Categoría Relevante)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.71	0.96	0.68	0.82	0.92
Sub-muestra 2	0.8	0.95	0.71	0.87	0.95
Sub-muestra 3	0.8	0.95	0.74	0.87	0.96
Sub-muestra 4	0.69	0.94	0.67	0.8	0.87
Sub-muestra 5	0.71	0.91	0.62	0.8	0.91
Sub-muestra 6	0.73	0.96	0.64	0.83	0.9
Sub-muestra 7	0.76	0.96	0.69	0.85	0.93
Sub-muestra 8	0.72	0.95	0.64	0.82	0.94
Sub-muestra 9	0.75	0.95	0.67	0.84	0.89
Sub-muestra 10	0.73	0.92	0.7	0.81	0.89

Repetición 3					
Sub-muestra de prueba	Métricas (Categoría No Relevante)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.95	0.68	0.96	0.79	0.95
Sub-muestra 2	0.92	0.71	0.95	0.8	0.94
Sub-muestra 3	0.93	0.74	0.95	0.83	0.95
Sub-muestra 4	0.94	0.67	0.94	0.78	0.93
Sub-muestra 5	0.87	0.62	0.91	0.73	0.91
Sub-muestra 6	0.94	0.64	0.96	0.76	0.91
Sub-muestra 7	0.95	0.69	0.96	0.8	0.94
Sub-muestra 8	0.93	0.64	0.95	0.76	0.93
Sub-muestra 9	0.93	0.67	0.95	0.78	0.92
Sub-muestra 10	0.91	0.7	0.92	0.79	0.93

Repetición 3					
Sub-muestra de prueba	Métricas (Promedio Simple entre las 2 categorías)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.83	0.82	0.82	0.81	0.935
Sub-muestra 2	0.86	0.83	0.83	0.84	0.945
Sub-muestra 3	0.87	0.85	0.85	0.85	0.955
Sub-muestra 4	0.81	0.8	0.8	0.79	0.9
Sub-muestra 5	0.79	0.77	0.77	0.76	0.91
Sub-muestra 6	0.84	0.8	0.8	0.79	0.905
Sub-muestra 7	0.85	0.82	0.82	0.82	0.935
Sub-muestra 8	0.82	0.79	0.79	0.79	0.935
Sub-muestra 9	0.84	0.81	0.81	0.81	0.905
Sub-muestra 10	0.82	0.81	0.81	0.8	0.91

Repetición 3					
Sub-muestra de prueba	Métricas (Promedio Ponderado entre las 2 categorías)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.84	0.81	0.81	0.8	0.9365
Sub-muestra 2	0.86	0.85	0.85	0.84	0.9455
Sub-muestra 3	0.86	0.85	0.85	0.85	0.9552
Sub-muestra 4	0.83	0.79	0.79	0.79	0.9037
Sub-muestra 5	0.79	0.77	0.77	0.76	0.9100
Sub-muestra 6	0.83	0.8	0.8	0.79	0.9050
Sub-muestra 7	0.85	0.83	0.83	0.82	0.9350
Sub-muestra 8	0.83	0.79	0.79	0.79	0.9350
Sub-muestra 9	0.84	0.81	0.81	0.81	0.9047
Sub-muestra 10	0.82	0.8	0.8	0.8	0.9111

- Repetición 4

Repetición 4		
Sub-muestra de prueba	Métricas	
	Exactitud	AUC ROC
Sub-muestra 1	0.77	0.9
Sub-muestra 2	0.75	0.87
Sub-muestra 3	0.79	0.93
Sub-muestra 4	0.87	0.94
Sub-muestra 5	0.83	0.92
Sub-muestra 6	0.85	0.92
Sub-muestra 7	0.79	0.91
Sub-muestra 8	0.8	0.9
Sub-muestra 9	0.8	0.94
Sub-muestra 10	0.84	0.94

Repetición 4					
Sub-muestra de prueba	Métricas (Categoría Relevante)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.65	0.94	0.65	0.77	0.89
Sub-muestra 2	0.67	0.95	0.58	0.79	0.85
Sub-muestra 3	0.68	0.97	0.64	0.8	0.93
Sub-muestra 4	0.85	0.93	0.8	0.89	0.93
Sub-muestra 5	0.8	0.92	0.73	0.86	0.93
Sub-muestra 6	0.8	0.96	0.73	0.87	0.92
Sub-muestra 7	0.7	0.96	0.64	0.81	0.91
Sub-muestra 8	0.76	0.96	0.59	0.85	0.94
Sub-muestra 9	0.71	0.96	0.66	0.82	0.94
Sub-muestra 10	0.77	0.97	0.71	0.86	0.95

Repetición 4					
Sub-muestra de prueba	Métricas (Categoría No Relevante)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.94	0.65	0.94	0.77	0.92
Sub-muestra 2	0.92	0.58	0.95	0.71	0.9
Sub-muestra 3	0.97	0.64	0.97	0.77	0.95
Sub-muestra 4	0.9	0.8	0.93	0.85	0.94
Sub-muestra 5	0.88	0.73	0.92	0.8	0.93
Sub-muestra 6	0.95	0.73	0.96	0.82	0.93
Sub-muestra 7	0.95	0.64	0.96	0.76	0.94
Sub-muestra 8	0.91	0.59	0.96	0.72	0.89
Sub-muestra 9	0.95	0.66	0.96	0.78	0.95
Sub-muestra 10	0.97	0.71	0.97	0.82	0.95

Repetición 4					
Sub-muestra de prueba	Métricas (Promedio Simple entre las 2 categorías)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.79	0.79	0.79	0.77	0.905
Sub-muestra 2	0.8	0.76	0.76	0.75	0.875
Sub-muestra 3	0.82	0.81	0.81	0.79	0.94
Sub-muestra 4	0.88	0.86	0.86	0.87	0.935
Sub-muestra 5	0.84	0.82	0.82	0.83	0.93
Sub-muestra 6	0.87	0.85	0.85	0.85	0.925
Sub-muestra 7	0.82	0.8	0.8	0.78	0.925
Sub-muestra 8	0.83	0.77	0.77	0.78	0.915
Sub-muestra 9	0.83	0.81	0.81	0.8	0.945
Sub-muestra 10	0.87	0.84	0.84	0.84	0.95

Repetición 4					
Sub-muestra de prueba	Métricas (Promedio Ponderado entre las 2 categorías)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.82	0.77	0.77	0.77	0.9076
Sub-muestra 2	0.8	0.75	0.75	0.75	0.8761
Sub-muestra 3	0.84	0.79	0.79	0.78	0.9412
Sub-muestra 4	0.87	0.87	0.87	0.87	0.9345
Sub-muestra 5	0.84	0.83	0.83	0.83	0.9300
Sub-muestra 6	0.87	0.85	0.85	0.85	0.9248
Sub-muestra 7	0.83	0.79	0.79	0.78	0.9261
Sub-muestra 8	0.82	0.8	0.8	0.79	0.9187
Sub-muestra 9	0.84	0.8	0.8	0.8	0.9454
Sub-muestra 10	0.87	0.84	0.84	0.84	0.9500

- Repetición 5

Repetición 5		
Sub-muestra de prueba	Métricas	
	Exactitud	AUC ROC
Sub-muestra 1	0.81	0.92
Sub-muestra 2	0.84	0.93
Sub-muestra 3	0.73	0.89
Sub-muestra 4	0.79	0.89
Sub-muestra 5	0.83	0.92
Sub-muestra 6	0.85	0.93
Sub-muestra 7	0.77	0.9
Sub-muestra 8	0.79	0.91
Sub-muestra 9	0.86	0.94
Sub-muestra 10	0.83	0.95

Repetición 5					
Sub-muestra de prueba	Métricas (Categoría Relevante)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.74	0.9	0.72	0.81	0.93
Sub-muestra 2	0.8	0.94	0.71	0.86	0.94
Sub-muestra 3	0.62	0.92	0.58	0.74	0.88
Sub-muestra 4	0.71	0.95	0.65	0.81	0.86
Sub-muestra 5	0.79	0.95	0.68	0.86	0.94
Sub-muestra 6	0.79	0.96	0.7	0.87	0.93
Sub-muestra 7	0.7	0.93	0.62	0.8	0.92
Sub-muestra 8	0.69	0.97	0.65	0.81	0.89
Sub-muestra 9	0.82	0.95	0.77	0.88	0.94
Sub-muestra 10	0.74	0.99	0.7	0.85	0.95

Repetición 5					
Sub-muestra de prueba	Métricas (Categoría No Relevante)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.89	0.72	0.9	0.8	0.94
Sub-muestra 2	0.91	0.71	0.94	0.8	0.93
Sub-muestra 3	0.91	0.58	0.92	0.71	0.93
Sub-muestra 4	0.93	0.65	0.95	0.77	0.92
Sub-muestra 5	0.92	0.68	0.95	0.78	0.91
Sub-muestra 6	0.94	0.7	0.96	0.81	0.94
Sub-muestra 7	0.91	0.62	0.93	0.74	0.91
Sub-muestra 8	0.96	0.65	0.97	0.77	0.94
Sub-muestra 9	0.93	0.77	0.95	0.84	0.94
Sub-muestra 10	0.98	0.7	0.99	0.82	0.96

Repetición 5					
Sub-muestra de prueba	Métricas (Promedio Simple entre las 2 categorías)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.82	0.81	0.81	0.81	0.935
Sub-muestra 2	0.85	0.83	0.83	0.83	0.935
Sub-muestra 3	0.77	0.75	0.75	0.73	0.905
Sub-muestra 4	0.82	0.8	0.8	0.79	0.89
Sub-muestra 5	0.86	0.82	0.82	0.82	0.925
Sub-muestra 6	0.87	0.83	0.83	0.84	0.935
Sub-muestra 7	0.81	0.78	0.78	0.77	0.915
Sub-muestra 8	0.83	0.81	0.81	0.79	0.915
Sub-muestra 9	0.88	0.86	0.86	0.86	0.94
Sub-muestra 10	0.86	0.84	0.84	0.83	0.955

Repetición 5					
Sub-muestra de prueba	Métricas (Promedio Ponderado entre las 2 categorías)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.82	0.81	0.81	0.81	0.9353
Sub-muestra 2	0.85	0.84	0.84	0.84	0.9355
Sub-muestra 3	0.79	0.73	0.73	0.73	0.9087
Sub-muestra 4	0.83	0.79	0.79	0.79	0.8914
Sub-muestra 5	0.85	0.83	0.83	0.83	0.9268
Sub-muestra 6	0.86	0.85	0.85	0.84	0.9346
Sub-muestra 7	0.81	0.77	0.77	0.77	0.9149
Sub-muestra 8	0.84	0.79	0.79	0.79	0.9174
Sub-muestra 9	0.87	0.86	0.86	0.86	0.9400
Sub-muestra 10	0.87	0.83	0.83	0.83	0.9554

- Repetición 6

Repetición 6		
Sub-muestra de prueba	Métricas	
	Exactitud	AUC ROC
Sub-muestra 1	0.83	0.91
Sub-muestra 2	0.81	0.92
Sub-muestra 3	0.75	0.9
Sub-muestra 4	0.77	0.91
Sub-muestra 5	0.83	0.9
Sub-muestra 6	0.81	0.9
Sub-muestra 7	0.8	0.89
Sub-muestra 8	0.82	0.93
Sub-muestra 9	0.85	0.92
Sub-muestra 10	0.82	0.94

Repetición 6					
Sub-muestra de prueba	Métricas (Categoría Relevante)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.74	0.97	0.7	0.84	0.91
Sub-muestra 2	0.71	0.97	0.69	0.82	0.9
Sub-muestra 3	0.65	0.96	0.6	0.77	0.91
Sub-muestra 4	0.67	0.91	0.66	0.77	0.9
Sub-muestra 5	0.77	0.95	0.69	0.85	0.9
Sub-muestra 6	0.76	0.94	0.67	0.84	0.88
Sub-muestra 7	0.77	0.93	0.63	0.84	0.9
Sub-muestra 8	0.78	0.94	0.66	0.85	0.95
Sub-muestra 9	0.8	0.94	0.76	0.87	0.92
Sub-muestra 10	0.74	0.97	0.65	0.84	0.96

Repetición 6					
Sub-muestra de prueba	Métricas (Categoría No Relevante)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.97	0.7	0.97	0.81	0.94
Sub-muestra 2	0.97	0.69	0.97	0.81	0.95
Sub-muestra 3	0.95	0.6	0.96	0.74	0.93
Sub-muestra 4	0.91	0.66	0.91	0.77	0.94
Sub-muestra 5	0.93	0.69	0.95	0.79	0.91
Sub-muestra 6	0.91	0.67	0.94	0.77	0.92
Sub-muestra 7	0.88	0.63	0.93	0.73	0.88
Sub-muestra 8	0.9	0.66	0.94	0.76	0.93
Sub-muestra 9	0.92	0.76	0.94	0.83	0.94
Sub-muestra 10	0.96	0.65	0.97	0.78	0.95

Repetición 6					
Sub-muestra de prueba	Métricas (Promedio Simple entre las 2 categorías)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.85	0.84	0.84	0.82	0.925
Sub-muestra 2	0.84	0.83	0.83	0.81	0.925
Sub-muestra 3	0.8	0.78	0.78	0.75	0.92
Sub-muestra 4	0.79	0.79	0.79	0.77	0.92
Sub-muestra 5	0.85	0.82	0.82	0.82	0.905
Sub-muestra 6	0.83	0.81	0.81	0.81	0.9
Sub-muestra 7	0.82	0.78	0.78	0.79	0.89
Sub-muestra 8	0.84	0.8	0.8	0.81	0.94
Sub-muestra 9	0.86	0.85	0.85	0.85	0.93
Sub-muestra 10	0.85	0.81	0.81	0.81	0.955

Repetición 6					
Sub-muestra de prueba	Métricas (Promedio Ponderado entre las 2 categorías)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.86	0.83	0.83	0.82	0.9261
Sub-muestra 2	0.85	0.81	0.81	0.81	0.9281
Sub-muestra 3	0.82	0.75	0.75	0.75	0.9214
Sub-muestra 4	0.81	0.77	0.77	0.77	0.9230
Sub-muestra 5	0.85	0.83	0.83	0.82	0.9048
Sub-muestra 6	0.83	0.81	0.81	0.81	0.8988
Sub-muestra 7	0.81	0.8	0.8	0.79	0.8914
Sub-muestra 8	0.83	0.82	0.82	0.81	0.9412
Sub-muestra 9	0.86	0.85	0.85	0.85	0.9298
Sub-muestra 10	0.85	0.82	0.82	0.81	0.9551

- Repetición 7

Repetición 7		
Sub-muestra de prueba	Métricas	
	Exactitud	AUC ROC
Sub-muestra 1	0.81	0.92
Sub-muestra 2	0.83	0.94
Sub-muestra 3	0.85	0.93
Sub-muestra 4	0.81	0.9
Sub-muestra 5	0.8	0.93
Sub-muestra 6	0.79	0.89
Sub-muestra 7	0.81	0.91
Sub-muestra 8	0.82	0.91
Sub-muestra 9	0.78	0.89
Sub-muestra 10	0.83	0.93

Repetición 7					
Sub-muestra de prueba	Métricas (Categoría Relevante)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.73	0.93	0.71	0.82	0.91
Sub-muestra 2	0.76	0.95	0.7	0.85	0.95
Sub-muestra 3	0.81	0.93	0.77	0.87	0.93
Sub-muestra 4	0.76	0.92	0.7	0.83	0.92
Sub-muestra 5	0.71	0.97	0.64	0.82	0.93
Sub-muestra 6	0.74	0.95	0.58	0.83	0.92
Sub-muestra 7	0.73	0.96	0.66	0.83	0.91
Sub-muestra 8	0.76	0.93	0.71	0.84	0.88
Sub-muestra 9	0.69	1	0.57	0.82	0.89
Sub-muestra 10	0.74	0.97	0.7	0.84	0.92

Repetición 7					
Sub-muestra de prueba	Métricas (Categoría No Relevante)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.92	0.71	0.93	0.81	0.94
Sub-muestra 2	0.93	0.7	0.95	0.8	0.94
Sub-muestra 3	0.91	0.77	0.93	0.83	0.92
Sub-muestra 4	0.9	0.7	0.92	0.79	0.92
Sub-muestra 5	0.96	0.64	0.97	0.77	0.95
Sub-muestra 6	0.91	0.58	0.95	0.71	0.89
Sub-muestra 7	0.95	0.66	0.96	0.78	0.93
Sub-muestra 8	0.92	0.71	0.93	0.8	0.93
Sub-muestra 9	1	0.57	1	0.73	0.92
Sub-muestra 10	0.97	0.7	0.97	0.81	0.95

Repetición 7					
Sub-muestra de prueba	Métricas (Promedio Simple entre las 2 categorías)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.83	0.82	0.82	0.81	0.925
Sub-muestra 2	0.85	0.83	0.83	0.82	0.945
Sub-muestra 3	0.86	0.85	0.85	0.85	0.925
Sub-muestra 4	0.83	0.81	0.81	0.81	0.92
Sub-muestra 5	0.84	0.81	0.81	0.8	0.94
Sub-muestra 6	0.82	0.77	0.77	0.77	0.905
Sub-muestra 7	0.84	0.81	0.81	0.8	0.92
Sub-muestra 8	0.84	0.82	0.82	0.82	0.905
Sub-muestra 9	0.84	0.79	0.79	0.77	0.905
Sub-muestra 10	0.86	0.83	0.83	0.83	0.935

Repetición 7					
Sub-muestra de prueba	Métricas (Promedio Ponderado entre las 2 categorías)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.84	0.81	0.81	0.81	0.9263
Sub-muestra 2	0.85	0.83	0.83	0.82	0.9450
Sub-muestra 3	0.86	0.85	0.85	0.85	0.9252
Sub-muestra 4	0.83	0.81	0.81	0.81	0.9200
Sub-muestra 5	0.84	0.8	0.8	0.8	0.9405
Sub-muestra 6	0.81	0.79	0.79	0.78	0.9066
Sub-muestra 7	0.84	0.81	0.81	0.8	0.9203
Sub-muestra 8	0.84	0.82	0.82	0.82	0.9055
Sub-muestra 9	0.85	0.78	0.78	0.77	0.9055
Sub-muestra 10	0.86	0.83	0.83	0.83	0.9357

- Repetición 8

Repetición 8		
Sub-muestra de prueba	Métricas	
	Exactitud	AUC ROC
Sub-muestra 1	0.84	0.95
Sub-muestra 2	0.82	0.91
Sub-muestra 3	0.79	0.91
Sub-muestra 4	0.85	0.92
Sub-muestra 5	0.79	0.92
Sub-muestra 6	0.78	0.91
Sub-muestra 7	0.79	0.91
Sub-muestra 8	0.83	0.93
Sub-muestra 9	0.78	0.88
Sub-muestra 10	0.83	0.91

Repetición 8					
Sub-muestra de prueba	Métricas (Categoría Relevante)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.79	0.94	0.73	0.86	0.96
Sub-muestra 2	0.76	0.92	0.72	0.84	0.91
Sub-muestra 3	0.71	0.96	0.64	0.82	0.91
Sub-muestra 4	0.77	0.97	0.72	0.86	0.92
Sub-muestra 5	0.71	0.93	0.65	0.81	0.94
Sub-muestra 6	0.7	0.96	0.62	0.81	0.9
Sub-muestra 7	0.75	0.91	0.66	0.82	0.9
Sub-muestra 8	0.77	0.98	0.66	0.86	0.95
Sub-muestra 9	0.68	0.94	0.65	0.79	0.87
Sub-muestra 10	0.76	0.96	0.69	0.85	0.92

Repetición 8					
Sub-muestra de prueba	Métricas (Categoría No Relevante)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.92	0.73	0.94	0.81	0.95
Sub-muestra 2	0.9	0.72	0.92	0.8	0.93
Sub-muestra 3	0.94	0.64	0.96	0.76	0.93
Sub-muestra 4	0.97	0.72	0.97	0.83	0.94
Sub-muestra 5	0.91	0.65	0.93	0.76	0.94
Sub-muestra 6	0.94	0.62	0.96	0.75	0.93
Sub-muestra 7	0.88	0.66	0.91	0.75	0.92
Sub-muestra 8	0.96	0.66	0.98	0.78	0.93
Sub-muestra 9	0.93	0.65	0.94	0.77	0.91
Sub-muestra 10	0.95	0.69	0.96	0.8	0.93

Repetición 8					
Sub-muestra de prueba	Métricas (Promedio Simple entre las 2 categorías)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.85	0.83	0.83	0.84	0.955
Sub-muestra 2	0.83	0.82	0.82	0.82	0.92
Sub-muestra 3	0.83	0.8	0.8	0.79	0.92
Sub-muestra 4	0.87	0.85	0.85	0.84	0.93
Sub-muestra 5	0.81	0.79	0.79	0.78	0.94
Sub-muestra 6	0.82	0.79	0.79	0.78	0.915
Sub-muestra 7	0.81	0.79	0.79	0.79	0.91
Sub-muestra 8	0.87	0.82	0.82	0.82	0.94
Sub-muestra 9	0.81	0.8	0.8	0.78	0.89
Sub-muestra 10	0.85	0.83	0.83	0.82	0.925

Repetición 8					
Sub-muestra de prueba	Métricas (Promedio Ponderado entre las 2 categorías)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.85	0.84	0.84	0.84	0.9552
Sub-muestra 2	0.83	0.82	0.82	0.82	0.9201
Sub-muestra 3	0.83	0.79	0.79	0.79	0.9203
Sub-muestra 4	0.87	0.85	0.85	0.84	0.9303
Sub-muestra 5	0.82	0.79	0.79	0.78	0.9400
Sub-muestra 6	0.83	0.78	0.78	0.78	0.9157
Sub-muestra 7	0.81	0.79	0.79	0.79	0.9095
Sub-muestra 8	0.86	0.83	0.83	0.83	0.9408
Sub-muestra 9	0.82	0.78	0.78	0.78	0.8922
Sub-muestra 10	0.85	0.83	0.83	0.82	0.9250

- Repetición 9

Repetición 9		
Sub-muestra de prueba	Métricas	
	Exactitud	AUC ROC
Sub-muestra 1	0.77	0.9
Sub-muestra 2	0.81	0.88
Sub-muestra 3	0.83	0.91
Sub-muestra 4	0.82	0.92
Sub-muestra 5	0.81	0.92
Sub-muestra 6	0.82	0.93
Sub-muestra 7	0.84	0.91
Sub-muestra 8	0.75	0.91
Sub-muestra 9	0.85	0.94
Sub-muestra 10	0.83	0.93

Repetición 9					
Sub-muestra de prueba	Métricas (Categoría Relevante)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.71	0.91	0.65	0.8	0.92
Sub-muestra 2	0.74	0.94	0.68	0.83	0.87
Sub-muestra 3	0.74	0.97	0.7	0.84	0.86
Sub-muestra 4	0.76	0.95	0.69	0.84	0.94
Sub-muestra 5	0.76	0.95	0.65	0.84	0.94
Sub-muestra 6	0.76	0.95	0.69	0.84	0.94
Sub-muestra 7	0.77	0.99	0.67	0.87	0.92
Sub-muestra 8	0.65	0.92	0.63	0.76	0.92
Sub-muestra 9	0.77	0.97	0.75	0.86	0.93
Sub-muestra 10	0.77	0.96	0.67	0.86	0.94

Repetición 9					
Sub-muestra de prueba	Métricas (Categoría No Relevante)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.88	0.65	0.91	0.74	0.91
Sub-muestra 2	0.91	0.68	0.94	0.78	0.91
Sub-muestra 3	0.97	0.7	0.97	0.81	0.94
Sub-muestra 4	0.93	0.69	0.95	0.79	0.93
Sub-muestra 5	0.92	0.65	0.95	0.76	0.92
Sub-muestra 6	0.93	0.69	0.95	0.79	0.94
Sub-muestra 7	0.98	0.67	0.99	0.79	0.93
Sub-muestra 8	0.92	0.63	0.92	0.75	0.94
Sub-muestra 9	0.97	0.75	0.97	0.84	0.96
Sub-muestra 10	0.94	0.67	0.96	0.78	0.93

Repetición 9					
Sub-muestra de prueba	Métricas (Promedio Simple entre las 2 categorías)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.8	0.78	0.78	0.77	0.915
Sub-muestra 2	0.83	0.81	0.81	0.8	0.89
Sub-muestra 3	0.85	0.84	0.84	0.82	0.9
Sub-muestra 4	0.84	0.82	0.82	0.82	0.935
Sub-muestra 5	0.84	0.8	0.8	0.8	0.93
Sub-muestra 6	0.84	0.82	0.82	0.82	0.94
Sub-muestra 7	0.88	0.83	0.83	0.83	0.925
Sub-muestra 8	0.78	0.78	0.78	0.75	0.93
Sub-muestra 9	0.87	0.86	0.86	0.85	0.945
Sub-muestra 10	0.86	0.82	0.82	0.82	0.935

Repetición 9					
Sub-muestra de prueba	Métricas (Promedio Ponderado entre las 2 categorías)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.8	0.77	0.77	0.77	0.9149
Sub-muestra 2	0.83	0.81	0.81	0.8	0.8901
Sub-muestra 3	0.86	0.83	0.83	0.82	0.9028
Sub-muestra 4	0.84	0.82	0.82	0.82	0.9350
Sub-muestra 5	0.83	0.81	0.81	0.81	0.9307
Sub-muestra 6	0.84	0.82	0.82	0.82	0.9400
Sub-muestra 7	0.87	0.84	0.84	0.83	0.9246
Sub-muestra 8	0.8	0.75	0.75	0.75	0.9316
Sub-muestra 9	0.88	0.85	0.85	0.85	0.9461
Sub-muestra 10	0.85	0.83	0.83	0.82	0.9354

- Repetición 10

Repetición 10		
Sub-muestra de prueba	Métricas	
	Exactitud	AUC ROC
Sub-muestra 1	0.85	0.95
Sub-muestra 2	0.86	0.94
Sub-muestra 3	0.75	0.88
Sub-muestra 4	0.84	0.93
Sub-muestra 5	0.75	0.91
Sub-muestra 6	0.79	0.88
Sub-muestra 7	0.81	0.92
Sub-muestra 8	0.81	0.92
Sub-muestra 9	0.81	0.9
Sub-muestra 10	0.82	0.92

Repetición 10					
Sub-muestra de prueba	Métricas (Categoría Relevante)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.81	0.93	0.74	0.87	0.97
Sub-muestra 2	0.8	0.97	0.75	0.88	0.93
Sub-muestra 3	0.67	0.97	0.56	0.79	0.85
Sub-muestra 4	0.75	1	0.68	0.86	0.93
Sub-muestra 5	0.68	0.96	0.56	0.79	0.93
Sub-muestra 6	0.69	0.91	0.7	0.78	0.85
Sub-muestra 7	0.73	0.96	0.66	0.83	0.9
Sub-muestra 8	0.76	0.91	0.71	0.83	0.93
Sub-muestra 9	0.75	0.92	0.68	0.83	0.9
Sub-muestra 10	0.75	0.96	0.68	0.85	0.93

Repetición 10					
Sub-muestra de prueba	Métricas (Categoría No Relevante)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.9	0.74	0.93	0.81	0.95
Sub-muestra 2	0.97	0.75	0.97	0.84	0.95
Sub-muestra 3	0.96	0.56	0.97	0.7	0.91
Sub-muestra 4	1	0.68	1	0.81	0.95
Sub-muestra 5	0.94	0.56	0.96	0.7	0.92
Sub-muestra 6	0.91	0.7	0.91	0.79	0.91
Sub-muestra 7	0.95	0.66	0.96	0.78	0.94
Sub-muestra 8	0.89	0.71	0.91	0.79	0.92
Sub-muestra 9	0.9	0.68	0.92	0.78	0.91
Sub-muestra 10	0.95	0.68	0.96	0.79	0.93

Repetición 10					
Sub-muestra de prueba	Métricas (Promedio Simple entre las 2 categorías)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.86	0.84	0.84	0.84	0.96
Sub-muestra 2	0.88	0.86	0.86	0.86	0.94
Sub-muestra 3	0.81	0.76	0.76	0.75	0.88
Sub-muestra 4	0.88	0.84	0.84	0.84	0.94
Sub-muestra 5	0.81	0.76	0.76	0.75	0.925
Sub-muestra 6	0.8	0.8	0.8	0.79	0.88
Sub-muestra 7	0.84	0.81	0.81	0.8	0.92
Sub-muestra 8	0.83	0.81	0.81	0.81	0.925
Sub-muestra 9	0.82	0.8	0.8	0.8	0.905
Sub-muestra 10	0.85	0.82	0.82	0.82	0.93

Repetición 10					
Sub-muestra de prueba	Métricas (Promedio Ponderado entre las 2 categorías)				
	Precisión	Sensibilidad	Especificidad	Medida F1	AUC PR
Sub-muestra 1	0.85	0.85	0.85	0.84	0.9610
Sub-muestra 2	0.88	0.86	0.86	0.86	0.9397
Sub-muestra 3	0.82	0.75	0.75	0.75	0.8814
Sub-muestra 4	0.88	0.84	0.84	0.83	0.9402
Sub-muestra 5	0.81	0.75	0.75	0.75	0.9249
Sub-muestra 6	0.82	0.79	0.79	0.79	0.8845
Sub-muestra 7	0.84	0.81	0.81	0.8	0.9206
Sub-muestra 8	0.82	0.81	0.81	0.81	0.9250
Sub-muestra 9	0.82	0.81	0.81	0.8	0.9049
Sub-muestra 10	0.85	0.82	0.82	0.82	0.9300

Anexo 5: Códigos utilizados en el procesamiento de los datos

En R:

Cargando paquetes

Paquete para la lectura de datos

library(readxl)

Paquetes para el pre-procesamiento de datos y la minería de texto

library(dplyr)

library(tidytext)

library(SnowballC)

```
library(tidyr)
library(tm)

# Paquetes para visualización de datos
library(RColorBrewer)
library(ggplot2)
library(reshape2)
library(scales)

## Lectura de datos
dat<- read_excel(file.choose(), sheet = 1)
dat1<-dat[1:769,]
dat2<-dat[770:1552,]
dat1<-dat1$COMENTARIO
dat2<-dat2$COMENTARIO

## Limpieza de datos
cleanfun<-function(htmlstring){
  return(gsub("<.*?>", "", htmlstring))
}

dat1<-cleanfun(dat1)
dat2<-cleanfun(dat2)

## N-gramas

# Almacenando stopwords en un objeto dataframe
stopw1<-as.data.frame(stopwords("spanish"))

# Unigramas

text_df1 <- tibble(text = dat1)
unigr<-text_df1 %>%
  unnest_tokens(unigram, text,token="ngrams",n=1) %>%
  filter(!unigram %in% stopw1$`stopwords("spanish")`) %>%
  count(unigram, sort = TRUE)

gr01h<-as.data.frame(unigr[1:110,])
gr01 <-head(gr01h,22)
```

```
gr01 <-gr01 %>% rename(Frecuencia = n)
gr01$unigram<-factor(gr01$unigram,levels=gr01$unigram[order(gr01$Frecuencia)])

graf1 <- ggplot(gr01, aes(unigram, Frecuencia,fill=Frecuencia))+
  geom_bar(stat="identity")+
  theme(axis.text.x=element_text(hjust=1))+
  xlab(NULL) +
  coord_flip() +
  geom_text(aes(label=Frecuencia),hjust=1.2,vjust=0.4,colour="white")+
  scale_fill_gradient(low="lightblue4", high = "blue4")
graf1

text_df15 <- tibble(text = dat2)

unigr<-text_df15 %>%
  unnest_tokens(unigram, text,token="ngrams",n=1) %>%
  filter(!unigram %in% stopw1$`stopwords("spanish")`) %>%
  count(unigram, sort = TRUE)

gr015h<-as.data.frame(unigr[1:110,])
gr015 <-head(gr015h,22)
gr015 <-gr015 %>% rename(Frecuencia = n)
gr015$unigram<-factor(gr015$unigram,levels=gr015$unigram[order(gr015$Frecuencia)])

graf15 <- ggplot(gr015, aes(unigram, Frecuencia,fill=Frecuencia))+
  geom_bar(stat="identity")+
  theme(axis.text.x=element_text(hjust=1))+
  xlab(NULL) +
  coord_flip() +
  geom_text(aes(label=Frecuencia),hjust=1.2,vjust=0.4,colour="white")+
  scale_fill_gradient(low="lightblue4", high = "blue4")
graf15

# Bigramas

text_df2 <- tibble(text = dat1)

bigr<-text_df2 %>%
  unnest_tokens(bigram, text,token="ngrams",n=2) %>%
```

```
  separate(bigram, c("word1", "word2"), sep = " ") %>%
  filter(!word1 %in% stopw1$`stopwords("spanish")`) %>%
  filter(!word2 %in% stopw1$`stopwords("spanish")`) %>%
  count(word1, word2, sort = TRUE) %>%
  unite(bigram, word1, word2, sep = " ")

gr2h <- as.data.frame(bigr[1:210,])
gr2h <- gr2h %>% rename(Palabras = bigram, Frecuencia = n)
grupo2 <-head(gr2h,22)
grupo2$Palabras<-
factor(grupo2$Palabras,levels=grupo2$Palabras[order(grupo2$Frecuencia)])

graf2 <- ggplot(grupo2, aes(Palabras, Frecuencia,fill=Frecuencia))+
  geom_bar(stat="identity")+
  theme(axis.text.x=element_text(hjust=1))+
  coord_flip() +
  xlab(NULL) +
  geom_text(aes(label=Frecuencia),hjust=1.2,vjust=0.4,colour="white")+
  scale_fill_gradient(low="lightblue4", high = "blue4")
graf2

text_df25 <- tibble(text = dat2)

bigr<-text_df25 %>%
  unnest_tokens(bigram, text,token="ngrams",n=2) %>%
  separate(bigram, c("word1", "word2"), sep = " ") %>%
  filter(!word1 %in% stopw1$`stopwords("spanish")`) %>%
  filter(!word2 %in% stopw1$`stopwords("spanish")`) %>%
  count(word1, word2, sort = TRUE) %>%
  unite(bigram, word1, word2, sep = " ")

gr25h <- as.data.frame(bigr[1:210,])
gr25h <- gr25h %>% rename(Palabras = bigram, Frecuencia = n)
grupo25 <-head(gr25h,22)
grupo25$Palabras<-
factor(grupo25$Palabras,levels=grupo25$Palabras[order(grupo25$Frecuencia)])
```

```
graf25 <- ggplot(grupo25, aes(Palabras, Frecuencia,fill=Frecuencia))+
  geom_bar(stat="identity")+
  theme(axis.text.x=element_text(hjust=1))+
  coord_flip() +
  xlab(NULL) +
  geom_text(aes(label=Frecuencia),hjust=1.2,vjust=0.4,colour="white")+
  scale_fill_gradient(low="lightblue4", high = "blue4")
graf25

# Trigramas

text_df3 <- tibble(text = dat1)

trig<-text_df3 %>%
  unnest_tokens(trigram, text,token="ngrams",n=3) %>%
  separate(trigram, c("word1", "word2", "word3"), sep = " ") %>%
  filter(!word1 %in% stopw1$`stopwords("spanish")`) %>%
  filter(!word2 %in% stopw1$`stopwords("spanish")`) %>%
  filter(!word3 %in% stopw1$`stopwords("spanish")`) %>%
  count(word1, word2, word3, sort = TRUE) %>%
  unite(trigram, word1, word2, word3, sep = " ")

gr3h <- as.data.frame(trig[1:210,])
gr3h <- gr3h %>% rename(Palabras = trigram, Frecuencia = n)
grupo3 <-head(gr3h,22)
grupo3$Palabras<-
factor(grupo3$Palabras,levels=grupo3$Palabras[order(grupo3$Frecuencia)])

graf3 <- ggplot(grupo3, aes(Palabras, Frecuencia,fill=Frecuencia))+
  geom_bar(stat="identity")+
  theme(axis.text.x=element_text(hjust=1))+
  coord_flip() +
  xlab(NULL) +
  geom_text(aes(label=Frecuencia),hjust=1.4,vjust=0.4,colour="white")+
  scale_fill_gradient(low="lightblue4", high = "blue4")
graf3

text_df35 <- tibble(text = dat2)
```

```
trig<-text_df35 %>%
  unnest_tokens(trigram, text,token="ngrams",n=3) %>%
  separate(trigram, c("word1", "word2", "word3"), sep = " ") %>%
  filter(!word1 %in% stopw1$`stopwords("spanish")`) %>%
  filter(!word2 %in% stopw1$`stopwords("spanish")`) %>%
  filter(!word3 %in% stopw1$`stopwords("spanish")`) %>%
  count(word1, word2, word3, sort = TRUE) %>%
  unite(trigram, word1, word2, word3, sep = " ")

gr35h <- as.data.frame(trig[1:210,])
gr35h <- gr35h %>% rename(Palabras = trigram, Frecuencia = n)
grupo35 <-head(gr35h,22)
grupo35$Palabras<
factor(grupo35$Palabras,levels=grupo35$Palabras[order(grupo35$Frecuencia)])

graf35 <- ggplot(grupo35, aes(Palabras, Frecuencia,fill=Frecuencia))+
  geom_bar(stat="identity")+
  theme(axis.text.x=element_text(hjust=1))+
  coord_flip() +
  xlab(NULL) +
  geom_text(aes(label=Frecuencia),hjust=1.4,vjust=0.4,colour="white")+
  scale_fill_gradient(low="lightblue4", high = "blue4")
graf35
```

En Python:

```
## Instalando y cargando librerías

# librería para manejo de dataframes
import pandas as pd

# librería para operaciones matemáticas
import numpy as np

# librería para mostrar gráficos
import matplotlib.pyplot as plt

# librería para la visualización de datos
import seaborn as sns
```

```
# librería para caracteres de texto especiales
!pip install unidecode
import unidecode

# librería para cadenas de texto y expresiones regulares (regular expressions)
import re

# librería para procesamiento de lenguaje natural (NLP)
import nltk
nltk.download('omw-1.4')

# librería para el procesamiento estadístico (pruebas, modelos, métricas de evaluación)
from sklearn import feature_extraction, feature_selection, model_selection, pipeline,
manifold, preprocessing, metrics

## Lectura de datos

df = pd.read_excel('/content/datos_.xlsx')
df
df.describe()
df.CATEGORÍA.unique()
dtf = df
dtf = dtf.rename(columns={"CATEGORÍA":"y", "COMENTARIO":"text"})
dtf.sample(5)

fig, ax = plt.subplots()
fig.suptitle("y", fontsize=12)
dtf["y"].reset_index().groupby("y").count().sort_values(by=
      "index").plot(kind="barh", legend=False,
       ax=ax).grid(axis='x')
plt.show()

## Pre-procesamiento de datos

# Función para el pre-procesamiento de datos

def utils_preprocess_text(text, lst_stopwords=None):

   # Limpieza de datos
```

```
    text = unidecode.unidecode(text)
    text = re.sub(r'[^\w\s]', ' ', str(text).lower().strip())
    text = re.sub(r'[.,;]', ' ', str(text).lower().strip())

    # Tokenización
    lst_text = text.split()

    # Eliminar stopwords
    if lst_stopwords is not None:
        lst_text = [word for word in lst_text if word not in lst_stopwords]

    # Devolver datos listos del pre-procesamiento
    text = " ".join(lst_text)
    return text

# Configurando y verificando stopwords al idioma español

nltk.download('stopwords')
lst_stopwords = nltk.corpus.stopwords.words("spanish")
lst_stopwords

# Aplicando y verificando la función en los datos

nltk.download('wordnet')
dtf["text_clean"] = dtf["text"].apply(lambda x:
        utils_preprocess_text(x, lst_stopwords=lst_stopwords))
dtf.head()

# Dividiendo los datos en muestra de entrenamiento y de prueba

dtf_train, dtf_test = model_selection.train_test_split(dtf, test_size=0.3)
y_train = dtf_train["y"].values
y_test = dtf_test["y"].values

## Aplicación de TF-IDF en la muestra de entrenamiento

# Configurando hiperparámetros de TF-IDF
vectorizer = feature_extraction.text.TfidfVectorizer(max_features=20000,
ngram_range=(1,2))

# Aplicación de TF-IDF y almacenamiento del resultado en nuevos objetos
X = dtf
```

```
corpus = dtf_train["text_clean"]
vectorizer.fit(corpus)
X_train = vectorizer.transform(corpus)
dic_vocabulary = vectorizer.vocabulary_

# Inspección de los resultados en los nuevos objetos
X_train
len(dic_vocabulary.keys())
print(dic_vocabulary)
word = "aprendo"
dic_vocabulary[word]
word = "gracias"
dic_vocabulary[word]
list(dic_vocabulary.keys())[list(dic_vocabulary.values()).index(5338)]

## Selección de variables

y = dtf_train["y"]
X_names = vectorizer.get_feature_names_out()
p_value_limit = 0.95
dtf_features = pd.DataFrame()

for cat in np.unique(y):
    chi2, p = feature_selection.chi2(X_train, y==cat)
    dtf_features = dtf_features.append(pd.DataFrame(
                {"feature":X_names, "score":1-p, "y":cat, "chi2":chi2}))

dtf_features = dtf_features.sort_values(["y","score"], ascending=[True,False])
dtf_features = dtf_features[dtf_features["score"]>p_value_limit]
dtf_features.to_excel('Features.xlsx')

for cat in np.unique(y):
   print("# {}:".format(cat))
   print("  . selected features:",
         len(dtf_features[dtf_features["y"]==cat]))
   print("  . top features:", ",".join(
dtf_features[dtf_features["y"]==cat]["feature"].values[:10]))
   print(" ")
```

```
dtf_features = pd.read_excel('/content/Features_final.xlsx')
X_names = dtf_features["feature"].unique().tolist()
X_names
len(X_names)
X_train_df = pd.DataFrame(X_train.toarray())
X_train_df
column_indices = [dic_vocabulary[word] for word in X_names]
X_train_new = X_train[:, column_indices]
print(X_train_new)

# Creando un objeto dataframe para leer la matriz documento-término en R.

X_train_nuevorden = pd.DataFrame(X_train_new.toarray())
X_train_nuevorden
X_train_nuevorden.to_excel('X_train_new_para_R.xlsx')
```

En R:

```
## Hallando las correlaciones de las variables
library(GGally)
ggcorr(datos, label = TRUE, method = c("pairwise", "pearson"))

## Selección del mejor modelo de clasificación
datos<-read.delim("clipboard", header=TRUE)
modelo<-glm(y~., data=datos, family = binomial())
summary(modelo)
```

En Python:

```
## Estimación del modelo final de clasificación

# Configuración de los hiperparámetros del modelo
classifier = sklearn.linear_model.LogisticRegression(penalty=None, solver='lbfgs')
model = pipeline.Pipeline([("classifier", classifier)])

# Entrenando al clasificador
model["classifier"].fit(X = X_train_new, y = y_train)

# Aplicación de TF-IDF en los datos de prueba
vectorizer = feature_extraction.text.TfidfVectorizer(max_features=20000,
ngram_range=(1,2))
```

```
corpus1 = dtf_test["text_clean"].values
vectorizer.fit(corpus1)
X_test = vectorizer.transform(corpus1)
dic_vocabulary1 = vectorizer.vocabulary_
column_indices1 = [dic_vocabulary1[word] for word in X_names]
X_test_New = X_test[:, column_indices1]
X_test_New

# Clasificación del modelo
predicted = model.predict(X_test_New)
predicted_prob = model.predict_proba(X_test_New)
print(classifier.coef_)
print(classifier.intercept_)
classifier.coef_.shape

## Evaluación del clasificador

classes = np.unique(y_test)
y_test_array = pd.get_dummies(y_test, drop_first=False).values

# Métricas de evaluación

accuracy = metrics.accuracy_score(y_test, predicted)
auc = metrics.roc_auc_score(y_test, predicted_prob[:, 1])
print("Accuracy:",  round(accuracy,2))
print("Auc:", round(auc,2))
print("Detail:")
print(metrics.classification_report(y_test, predicted))

# Gráfico de Matriz de confusión

cm = metrics.confusion_matrix(y_test, predicted)
fig, ax = plt.subplots()
sns.heatmap(cm, annot=True, fmt='d', ax=ax, cmap=plt.cm.Blues,
            cbar=False)
ax.set(xlabel="Pred", ylabel="True", xticklabels=classes,
       yticklabels=classes, title="Confusion matrix")
plt.yticks(rotation=0)
fig, ax = plt.subplots(nrows=1, ncols=2)
```

```
# Curva ROC

for i in range(len(classes)):
    fpr, tpr, thresholds = metrics.roc_curve(y_test_array[:,i],
                           predicted_prob[:,i])
    ax[0].plot(fpr, tpr, lw=3,
               label='{0} (area={1:0.2f})'.format(classes[i],
                          metrics.auc(fpr, tpr)))

ax[0].plot([0,1], [0,1], color='navy', lw=3, linestyle='--')
ax[0].set(xlim=[-0.05,1.0], ylim=[0.0,1.05],
          xlabel='False Positive Rate',
          ylabel="True Positive Rate (Recall)",
          title="Curva ROC")
ax[0].legend(loc="lower right")
ax[0].grid(True)

# Curva de Precisión-Recall

for i in range(len(classes)):
    precision, recall, thresholds = metrics.precision_recall_curve(
                y_test_array[:,i], predicted_prob[:,i])
    ax[1].plot(recall, precision, lw=3,
               label='{0} (area={1:0.2f})'.format(classes[i],
                           metrics.auc(recall, precision))
              )
ax[1].set(xlim=[0.0,1.05], ylim=[0.0,1.05], xlabel='Recall',
          ylabel="Precisión", title="Curva de Precisión-Recall")
ax[1].legend(loc="best")
ax[1].grid(True)
plt.show()

## Validación Cruzada

# División de la muestra total en 10 sub-muestras
dtf = dtf.sample(frac=1).reset_index(drop=True)
```

```
# Sub-muestra 1 (K1)
test = dtf.iloc[:155]
train = dtf.iloc[155:]

# Aplicación de TF-IDF en la muestra de entrenamiento
vectorizer = feature_extraction.text.TfidfVectorizer(max_features=20000,
ngram_range=(1,2))
X = dtf
corpus = dtf_train["text_clean"]
vectorizer.fit(corpus)
X_train = vectorizer.transform(corpus)
dic_vocabulary = vectorizer.vocabulary_
len(dic_vocabulary.keys())

# Estimación del modelo
classifier = sklearn.linear_model.LogisticRegression(penalty=None, solver='lbfgs')
model = pipeline.Pipeline([("classifier", classifier)])
model["classifier"].fit(X = X_train_new, y = y_train)

# Aplicación de TF-IDF en los datos de prueba
vectorizer = feature_extraction.text.TfidfVectorizer(max_features=20000,
ngram_range=(1,2))
corpus1 = dtf_test["text_clean"].values
vectorizer.fit(corpus1)
X_test = vectorizer.transform(corpus1)
dic_vocabulary1 = vectorizer.vocabulary_
column_indices1 = [dic_vocabulary1[word] for word in X_names]
X_test_New = X_test[:, column_indices1]
X_test_New

# Clasificación del modelo
predicted = model.predict(X_test_New)
predicted_prob = model.predict_proba(X_test_New)
print(classifier.coef_)
print(classifier.intercept_)
classifier.coef_.shape
```

```
# Evaluación del clasificador

y_test = test["y"].values
classes = np.unique(y_test)
y_test_array = pd.get_dummies(y_test, drop_first=False).values

accuracy = metrics.accuracy_score(y_test, predicted)
auc = metrics.roc_auc_score(y_test, predicted_prob[:, 1])
print("Accuracy:",  round(accuracy,2))
print("Auc:", round(auc,2))
print("Detail:")
print(metrics.classification_report(y_test, predicted))

for i in range(len(classes)):
    fpr, tpr, thresholds = metrics.roc_curve(y_test_array[:,i],
                           predicted_prob[:,i])
    print("AUC ROC -",  classes[i],":", round(metrics.auc(fpr, tpr),2))

for i in range(len(classes)):
    precision, recall, thresholds = metrics.precision_recall_curve(
                y_test_array[:,i], predicted_prob[:,i])
    print("AUC PR  -",  classes[i],":", round(metrics.auc(recall, precision),2))

# Sub-muestra 2 (K2)
test = dtf.iloc[155:310]
train1 = dtf.iloc[:155]
train2 = dtf.iloc[310:]
train = pd.concat([train1, train2])
# Repetir los pasos de sub-muestra 1

# Sub-muestra 3 (K3)
test = dtf.iloc[310:465]
train1 = dtf.iloc[:310]
train2 = dtf.iloc[465:]
train = pd.concat([train1, train2])
# Repetir los pasos de sub-muestra 1

# Sub-muestra 4 (K4)
test = dtf.iloc[465:620]
```

```
train1 = dtf.iloc[:465]
train2 = dtf.iloc[620:]
train = pd.concat([train1, train2])
# Repetir los pasos de sub-muestra 1

# Sub-muestra 5 (K5)
test = dtf.iloc[620:775]
train1 = dtf.iloc[:620]
train2 = dtf.iloc[775:]
train = pd.concat([train1, train2])
# Repetir los pasos de sub-muestra 1

# Sub-muestra 6 (K6)
test = dtf.iloc[775:930]
train1 = dtf.iloc[:775]
train2 = dtf.iloc[930:]
train = pd.concat([train1, train2])
# Repetir los pasos de sub-muestra 1

# Sub-muestra 7 (K7)
test = dtf.iloc[930:1085]
train1 = dtf.iloc[:930]
train2 = dtf.iloc[1085:]
train = pd.concat([train1, train2])
# Repetir los pasos de sub-muestra 1

# Sub-muestra 8 (K8)
test = dtf.iloc[1085:1240]
train1 = dtf.iloc[:1085]
train2 = dtf.iloc[1240:]
train = pd.concat([train1, train2])
# Repetir los pasos de sub-muestra 1

# Sub-muestra 9 (K9)
test = dtf.iloc[1240:1395]
train1 = dtf.iloc[:1240]
train2 = dtf.iloc[1395:]
```

```
train = pd.concat([train1, train2])
# Repetir los pasos de sub-muestra 1
# Sub-muestra 10 (K10)
test = dtf.iloc[1395:]
train = dtf.iloc[:1395]
# Repetir los pasos de sub-muestra 1
# Repetir todo el procedimiento de Validación Cruzada 9 veces más.
```