Pipeline para la detección del trastorno específico del lenguaje (SLI) a partir de transcripciones de narrativas espontáneas

Santiago Arena[1][0009−0000−3032−1225] and
Antonio Quintero-Rincón[1][0000−0003−0186−4049]

Laboratorio de Ciencia de Datos e Inteligencia Artificial
Departamento de Ciencia de Datos
Universidad Católica Argentina, Buenos Aires, Argentina
santiagoarena@uca.edu.ar,antonioquintero@uca.edu.ar

Abstract. El Trastorno Específico del Lenguaje (SLI) es un trastorno que afecta la comunicación y puede afectar tanto la comprensión como la expresión. Este estudio se centra en la detección eficaz del SLI en niños, empleando transcripciones de narrativas espontáneas tomadas en 1063 entrevistas. Para dicho fin, proponemos un *pipeline* de tres etapas en cascada. En la primera etapa, se hace una extracción de características y una reducción de dimensionalidad de los datos usando en conjunto, los métodos de *Random Forest* (RF) y correlación de *Spearman*. En la segunda etapa se estiman las variables más predictivas de la primera etapa usando regresión logística, las cuales son usadas en la última etapa, para detectar el trastorno SLI en niños a partir de transcripciones de narrativas espontáneas usando un clasificador de vecinos más cercanos. Los resultados revelaron una precisión del 97,13% en la identificación del SLI, destacándose aspectos como el largo de las respuestas, la calidad de sus enunciados y la complejidad del lenguaje. Este nuevo enfoque enmarcado en procesamiento natural del lenguaje, ofrece beneficios significativos al campo de la detección de SLI, al evitar complejas variables subjetivas y centrarse en métricas cuantitativas directamente relacionadas con el desempeño del niño.

Keywords: Reducción de dimensionalidad · SLI · Random Forest · clasificación · k-NN · NLP

1 Introducción

El Trastorno Específico del Lenguaje (SLI por sus siglas en inglés, *Specific Language Impairment*), también llamado disfasia infantil, afecta entre el 2 y el 11 porciento de los niños menores de 13 años, siendo caracterizado por deficiencias en el lenguaje, sin discapacidad o irregularidades mentales o física evidente. En otras palabras, es un trastorno que afecta la comunicación y puede afectar tanto la comprensión como la expresión [1]. Es más común en niños que en niñas, con un fuerte vínculo genético, dado que entre el 50% y el 70% de los niños con SLI tienen un miembro de la familia con esta afección [2]. Un niño con SLI a menudo tiene antecedentes de retraso en el desarrollo del lenguaje

expresivo, presentando síntomas como dificultades en la construcción de oraciones, desarrollo del vocabulario, entablar conversaciones, comprender y seguir reglas gramaticales y/o comprender instrucciones habladas [3]. Un trastorno del lenguaje (SLI) ocurre cuando un niño muestra una incapacidad persistente para adquirir y usar habilidades lingüísticas, como se proyectaría según las expectativas normativas basadas en la edad [4]; este trastorno se considera *primario* o *específico* cuando no hay una explicación clara para estos retrasos en las habilidades lingüísticas. La mayoría de los niños identificados con trastorno primario del lenguaje al ingresar a la escuela seguirán teniendo habilidades lingüísticas significativamente deprimidas con el tiempo [5], mostrarán dificultades con la preparación para el jardín de infantes [6], y tendrán dificultades para aprender a leer [7], este último debido en parte a sus efectos en habilidades lingüísticas de nivel superior [8]. El trastorno del lenguaje en la primera infancia también está vinculado a un mayor riesgo de preocupaciones psiquiátricas, dificultades de atención, problemas socio-conductuales y discapacidades de aprendizaje en la adolescencia [9] [10]. Dada la incidencia relativamente alta de esta discapacidad infantil y sus efectos significativos en numerosas áreas de desarrollo, existe un gran interés en garantizar una identificación precisa y una intervención temprana para los niños afectados en edad temprana [11]. En la literatura se destaca la importancia de detectar trastornos específicos del lenguaje en edades tempranas, basándose en asociaciones de los pacientes con déficits en diversos aspectos del lenguaje [11]. Así mismo, se destaca la importancia de distinguir SLI de otros trastornos, como el autismo, para mejorar la precisión diagnóstica [12]. Múltiples enfoques se han desarrollado utilizando diversas herramientas de ML para la clasificación de SLI en niños. Mac Farlane et al. [13] utiliza un esquema de redes neuronales para clasificar niños a través de índices de desempeño de 10 indicadores de confección personal, estos demostraron que es posible clasificar niños bilingües en los lenguajes de español e inglés. En el trabajo de Sharma et al. [2], se propuso un modelo que aprovecha datos directos de las transcripciones sin ningún procesamiento, utilizando redes neuronales convolucionales y aprendizaje profundo para segregar los pacientes entre SLI y desarrollo típico (TD). Kaushik et al. [14] aplica un método de detección de SLI denominado SLINet, donde a partir de una red neuronal convolucional 2D obtuvieron en 98 sujetos (54 SLI y 44 controles), una precisión del 99.09% utilizando validación cruzada de diez pliegues. Por otra parte, Gray et al. [15] presentó un enfoque para la detección de SLI en niños donde utilizaron la fiabilidad de prueba/reprueba y el parámetro de precisión diagnóstica para evaluar el diagnóstico de SLI. Para este propósito, se realizó una evaluación en repetición de no-palabras, serie de dígitos, batería de evaluación *Kaufman* para Niños y el test de lenguaje expresivo fotográfico estructurado, los autores informaron una especificidad y sensibilidad del 100% y 95%, respectivamente con 44 niños en edad preescolar (22 SLI, 22 TD). Armon-Lotem y Meir [16] propusieron un método para la identificación de SLI en niños bilingües (hebreo, ruso), su estudio utilizó pruebas de repetición de dígitos, no-palabras y oraciones, logrando tasas de precisión, sensibilidad y especificidad del 94%, 80% y 97%, respectivamente, para ambos idiomas en

230 niños mono y bilingües (175 TD, 55 SLI). Slogrove y Haar [17] aplicaron coeficientes cepstrales de frecuencia *Mel* de señales de voz para la detección de SLI. Alcanzaron una tasa de precisión del 99%utilizando un clasificador *Random Forest* de forma aleatoria. Reddy et al. [18] utilizaron características de la fuente glotal, coeficientes cepstrales de frecuencia *Mel* y una red neuronal feed-forward para la detección de SLI en niños. Su estudio utilizó señales de habla de 54 pacientes con SLI y 44 niños con TD. Informaron una precisión del 98.82% con selección de características. Oliva et al. [19] propusieron un método de detección de SLI utilizando técnicas de aprendizaje automático. En su estudio, se utilizaron datos cómo la longitud media de las emisiones, oraciones no gramaticales, uso correcto de: artículos, verbos, *clíticos*, argumentos temáticos y proporción de estructuras ditransitivas, entre otras, de 24 niños con SLI y 24 niños con TD, informaron tasas de sensibilidad y especificidad del 97% y 100%, respectivamente. SLI actualmente es un campo de estudio de interés en la comunidad científica, como se ha expuesto en el estado-del-arte. La literatura resalta la importancia de la detección temprana de trastornos específicos del lenguaje, como el SLI, y la necesidad de distinguirlo de otros trastornos para un diagnóstico preciso. Varios estudios han explorado enfoques de aprendizaje automático, como redes neuronales y análisis de características de voz, con resultados prometedores en la clasificación de SLI versus desarrollo típico. Además, la detección de SLI es de sumo interés en la automatización del proceso mediante técnicas de Procesamiento Natural del Lenguaje (NLP) y *Machine Learning* (ML) y en especial, en el diseño de instrumentos de medición que se basen en aspectos cuantitativos del diagnóstico del paciente [2]. Precisamente, este estudio se centra en niños diagnosticados con SLI y busca identificar marcadores lingüísticos que permitan una detección temprana y eficiente de este trastorno. La presente investigación tiene como objetivo desarrollar un *pipeline* en cascada usando clásicas técnicas de ML, para detectar el trastorno SLI en niños a partir de transcripciones de narrativas espontáneas. Para ello, proponemos usar los métodos de *Random Forest* (RF) y correlación en conjunto, cómo selectores de características y así obtener una reducción de dimensionalidad de los datos. Luego con estos datos, se usa el modelo de regresión logística, con el objetivo de obtener solamente las variables más predictivas, para finalmente, usar un clasificador de vecinos más cercano para detectar SLI.

El artículo está organizado de la siguiente manera. La Sección 2 presenta la metodología propuesta, donde se introduce el esquema del *pipeline*, se explican conceptos clave como reducción de dimensionalidad, los modelos aplicados en el mismo y las métricas utilizadas para evaluar los resultados. Luego en la Sección 3, se detallan los resultados obtenidos en las distintas etapas del proceso. Se presentan los análisis y las interpretaciones correspondientes a cada paso del método propuesto, incluyendo la evaluación de la eficacia de las técnicas empleadas. Además, se discuten los hallazgos significativos y se comparan con resultados previos en la literatura, con el objetivo de validar y contextualizar los nuevos resultados obtenidos. Finalmente, en la Sección 4 se extraen conclusiones, se realizan comparaciones, se plantean limitaciones, fortalezas y se discute sobre trabajos futuros.

2 Metodología

El *pipeline* en cascada propuesto, se compone de tres etapas, ver Figura 1. En la primera etapa (letras de color azul), se hace una extracción de características y una reducción de dimensionalidad de los datos usando los métodos de *Random Forest* (RF) y correlación en conjunto, logrando reducir de 43 a 11 variables. En la segunda etapa (letras de color rojo), se busca hallar las variables más predictivas usando regresión logística, obteniéndose 6 variables finales. Estas variables son usadas en la última etapa (color negro), para detectar el trastorno SLI en niños a partir de transcripciones de narrativas espontáneas usando k-NN. Las métricas usadas para evaluar los resultaron, fueron: Precisión, Recall positivo, Recall negativo, F1-score, Curva ROC, Error de raíz cuadrática media (RMSE), Error medio absoluto (MAE), R^2 y Error OOB [32,33]. La implementación de los siguientes métodos fueron realizados usando el lenguaje de programación *RStudio 2023.09.1+494, Desert Sunflower Release*, el cual está desarrollado para computación estadística y visualización de datos. A continuación se introducen los métodos del *pipeline* propuesto, siguiendo la siguiente nomenclatura:

Sea X la matriz que contiene los datos de tamaño $N \times V$, donde N es cantidad de observaciones y V la cantidad de variables. Note que x corresponde a una observación de una variable específica. Recordar que el objetivo final, es la detección del trastorno específico del lenguaje (SLI) partir de transcripciones de narrativas espontáneas, este objetivo se enmarca en un problema de clasificación binaria, por ende es necesario considerar dos clases $C = 0$ para un desarrollo típico normal y $C = 1$ para SLI.

2.1 Base de Datos

La base de datos consiste en transcripciones de audio públicas de tres estudios diferentes, llamados: Conti-Ramsden 4, ENNI y GILLUM. Se puede accesar libremente en [20]. El conjunto de datos Conti-Ramsden 4 se recopiló para un estudio que evaluó la efectividad de las pruebas narrativas en adolescentes. Consiste en 99 muestras de desarrollo típico (TD) y 19 muestras de trastorno específico del lenguaje (SLI) de niños entre las edades de 13 y 16. Este contiene transcripciones de una tarea de narración basada en el libro de imágenes sin palabras. El conjunto de datos ENNI consta de 300 muestras de desarrollo típico (TD) y 77 muestras de trastorno específico del lenguaje (SLI) de niños entre 4 y 9 años. A cada niño se le presentaron dos historias de imágenes sin palabras, una más complicada que la otra. El conjunto de datos de Gillam se basa en otra herramienta para la evaluación narrativa conocida como *Test de Lenguaje Narrativo (TNL)*. Consiste en 250 niños con trastornos del lenguaje (SLI) y 520 TD de entre 5 y 12 años. Las bases de datos combinadas contiene 1163 observaciones y 62 variables. Dentro de estas variables se encuentra el diagnóstico correspondiente a cada niño, si este tiene un desarrollo típico o muestra SLI, esta variable corresponde a la variable objetivo.

A modo de simplificación se decidió retirar aquellas variables que no fueron una métrica numérica medible, cómo por ejemplo el conteo de palabras o sílabas. De esta manera se cuentan con variables que registran el desempeño de un niño

descomponiendo su narrativa en la calidad de sus enunciados, oraciones y palabras. Algunos ejemplos son la cantidad de palabras de relleno por oración, la cantidad de errores por palabra, el promedio de verbos en pasado sobre los verbos en infinitivo, el promedio de palabras,verbos o adjetivos por oración, entre otros. El objeto de estas observaciones consiste en evaluar la complejidad narrativa y en ella, poder predecir un desarrollo atípico en la misma. Escapa el objeto de la presente investigación indagar en profundidad sobre los patrones del lenguaje en niños y sus variantes; no obstante, se invita al lector a explorar [21, 22] para un análisis de narrativas y extracción de características.

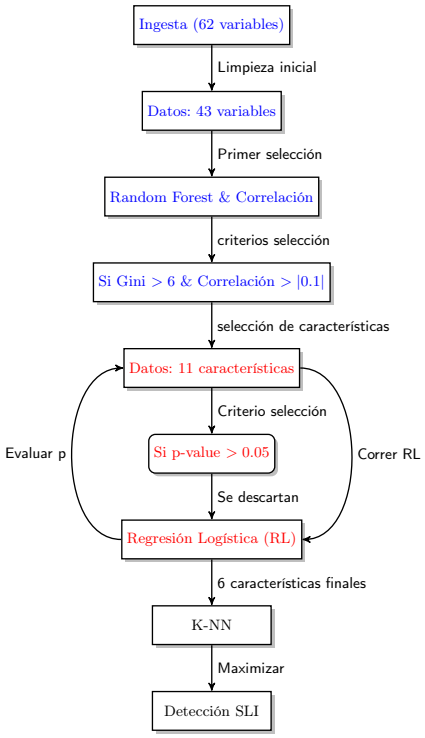


Fig. 1: *Pipeline* que ilustra la metodología propuesta.

2.2 Extracción de características

Estos métodos se utilizan para obtener el subconjunto de características más relevantes del conjunto principal, dicho subconjunto será aquel que maximice una función criterio determinada [23]. En esta etapa se consideró la relación entre la habilidad narrativa y el lenguaje. Específicamente se considera como un método válido la medición de la competencia comunicativa de un individuo [24]. Para lograr esto, solo se consideraron las variables cuantitativas relacionadas con la narrativa de los niños, dejando fuera aquellas que no representan un dato concreto y medible. Esta selección fue hecha manualmente obteniéndose 43 variables iniciales. El siguiente paso fue aplicar un esquema para determinar la importancia de estas variables, usando los métodos de *Random Forest* (RF) [25,26] y Correlación de *Spearman* (r_s) [28]; junto con su capacidad de predicción

de la variable objetivo ma través de la Regresión logística [29–31], dada por la variable grupo (ver Sección 2.1).

2.3 *k*-vecinos más cercanos (*k*-NN)

k-NN es un algoritmo de aprendizaje supervisado utilizado para la clasificación y regresión. En la clasificación k-NN, se asigna una etiqueta de clase al punto de datos desconocido, basándose en la mayoría de las etiquetas de clase de los k puntos de datos más cercanos en el conjunto de entrenamiento [34]. Usando el vector de características Θ_{Ct}, se considera una clasificación en dos posibles clases $C = 0$ (Desarrollo típico o normal) y $C = 1$ (SLI). La probabilidad de clasificar una muestra en una de las dos clases está dada por:

$$\rho(\Theta_{Ct}|c=0) = \frac{1}{N_0} \sum_{n \in \text{class } 0} \mathcal{N}(\Theta_{Ct}|\Theta_{nCt}, \sigma^2 I) = \frac{1}{N_0 (2\pi\sigma^2)^{D/2}} \sum_{n \in \text{class } 0} \exp\left(-\frac{(\Theta_{Ct} - \Theta_{nCt})^2}{2\sigma^2}\right)$$

$$\rho(\Theta_{Ct}|C=1) = \frac{1}{N_1} \sum_{n \in \text{class } 1} \mathcal{N}(\Theta_{Ct}|\Theta_{nCt}, \sigma^2 I) = \frac{1}{N_1 (2\pi\sigma^2)^{D/2}} \sum_{n \in \text{class } 1} \exp\left(-\frac{(\Theta_{Ct} - \Theta_{nCt})^2}{2\sigma^2}\right)$$

donde D es la dimensión de la muestra Θ_{Ct}, N_0 y N_1 son los números de muestras de entrenamiento de la clase 0 y clase 1, respectivamente, y σ^2 es la varianza. Usando la regla de Bayes para clasificar una nueva observación Θ^*_{Ct}, obtenemos la siguiente ecuación:

$$\rho(c=0|\Theta^*_{Ct}) = \frac{\rho(\Theta^*_{Ct}|c=0)\rho(c=0)}{\rho(\Theta^*_{Ct}|C=0)\rho(C=0) + \rho(\Theta^*_{Ct}|C=1)\rho(C=1)}, \tag{1}$$

donde la máxima verosimilitud es $\rho(C=0) = N_0/(N_0+N_1)$ y $\rho(C=1) = N_1/(N_0+N_1)$. Sustituyendo en la ecuación (1), se obtiene la probabilidad $\rho(C = 0|\Theta_{Ct})$. La expresión para $\rho(C = 1|\Theta_{Ct})$ se puede derivar de manera similar. Para determinar qué clase es más probable, se evalúa la proporción entre las dos expresiones: $\rho(C = 0|\Theta_{Ct})/\rho(C = 1|\Theta_{Ct}) = \rho(\Theta^*_{Ct}|C = 0)\rho(C = 0)/\rho(\Theta^*_{Ct}|C = 1)\rho(C = 1)$. Si la proporción es mayor que uno, Θ_{Ct} se clasifica como $C = 0$, de lo contrario se clasifica como $C = 1$. Es importante señalar que en el caso donde σ^2 es muy pequeño, tanto el numerador como el denominador estarán dominados por el término para el cual la muestra Θ_{n0Ct} en la clase-0 o Θ_{n1Ct} en la clase-1 están más cerca del punto Θ_{Ct}, tal que:

$$\frac{\rho(C=0|\Theta_{Ct})}{\rho(C=1|\Theta_{Ct})} = \frac{\exp\left(-\frac{(\Theta^*_{Ct}-\Theta_{n0Ct})^2}{2\sigma^2}\right)\rho(C=0)/N_0}{\exp\left(-\frac{(\Theta_{Ct}-\Theta_{n1Ct})^2}{2\sigma^2}\right)\rho(C=1)/N_1} = \frac{\exp\left(-\frac{(\Theta_{Ct}-\Theta_{n0Ct})^2}{2\sigma^2}\right)}{\exp\left(-\frac{(\Theta^*_{Ct}-\Theta_{n1Ct})^2}{2\sigma^2}\right)}.$$

En el límite $\sigma^2 \to 0$, Θ_{Ct} se clasifica como clase 0 si Θ_{Ct} tiene un punto en los datos de clase 0 que está más cerca que el punto más cercano en los datos de clase 1. El método k-NN se recupera así como el caso límite de un modelo generativo probabilístico. El parámetro k se elige basado en $N^{1/2}$, donde N es el número de muestras en el conjunto de datos de entrenamiento. Para un tratamiento completo de las propiedades matemáticas del clasificador de k-vecinos más cercanos, remitimos al lector a [35, 36].

2.4 Validación cruzada quíntuple

La validación cruzada es un método utilizado para evaluar el rendimiento predictivo de un modelo de aprendizaje automático [37]. Consiste en dividir el conjunto de datos en subconjuntos de entrenamiento y prueba repetidamente, ajustando y evaluando el modelo en cada iteración. Esta dada por: $\bar{p} = \frac{1}{k}\sum_{i=1}^{k} p_i$, donde k es el número de iteraciones de la validación cruzada, $\bar{p}$ es la precisión media, p_i es la precisión del modelo en la i-ésima iteración.

3 Resultados y discusión

En la presente sección se discuten los resultados hallados tras aplicar el *pipeline* propuesto siguiendo la Figura 1 de la Metodología, Sección 2. Todo sobre las 1063 observaciones y las 43 variables relacionadas con las narrativas espontáneas en niños, introducidas en la Sección 2.1.

La primera etapa consiste en reducir las cantidad de datos. Entonces, como primer instancia se necesitan estimar los parámetros del método RF, para así determinar, cuáles son las variables más importantes de los datos. Recordar que RF como clasificador binario, se centra en la votación mayoritaria de los árboles, por lo tanto, es necesario estimar la cantidad árboles óptimos del método. Para dicho fin, se busca el valor óptimo OOB, entre las 43 variables resultantes de limpiar los datos originales. La idea es encontrar el parámetro relacionado con el número de variables aleatorias, como candidatas en cada ramificación que registre el OOB óptimo. La Figura 2(a) muestra como varía el error OOB con el número de árboles y muestra cómo se comporta el modelo con respecto a cada clase de predicción. Observe los valores de los árboles son bastante constantes desde el valor 50, para el tipo 1 y para OOB, más sin embargo, se estabilizan en el valor 500. Teniendo el valor de la cantidad de árboles M, es posible entonces, estimar el método RF para las clases $C = 0 = normal$ y $C = 1 = SLI$. La meta en esta etapa, es encontrar la selección de características de mayor importancia y así hacer una primera reducción de dimensionaliad de los datos. Esta parte involucra el método RF con el *atributo importancia*, junto con el método del coeficiente de correlación con la variable *objetivo*. Este criterio de selección se compone de dos partes.

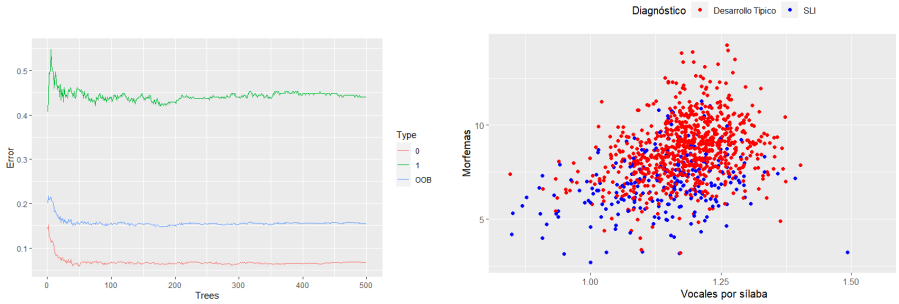
1. Criterio 1: Consiste en elegir aquellas variables que cumplan tener, una importancia superior al promedio aproximado entre la mediana y el tercer cuartil, cuyos valores son $4,5$ y $7,8$ respectivamente. La razón de esta decisión consiste en aprovechar no solo el mejor 25% de los datos (por delante del tercer cuartil), si no también, aquellas variables que se encontraban entre la mediana y el punto de corte del mencionado cuartil. Este enfoque permite aportar características de valor, que de otra forma hubiesen sido descartadas. En la Figura 3(a) es posible apreciar una importante aglomeración antes del punto de corte (linea roja-verde) entre la mediana y el tercer cuartil. Donde el eje x es el atributo importancia de RF y en el eje y, la correlación con la variable *objetivo*. Por otra parte en la figura 3(b) se aprecia la utilidad de la propuesta planteada, en tanto se identifican aquellas variables que no poseen una importancia significativa para el RF (eje x). De esta manera, se conservan variables de importancia sin arrastrar consigo aquellas de menor atributo de importancia, las cuales se encuentran alrededor de la mediana, por la izquierda de la linea verde.
2. Criterio 2: Consiste en corroborar que aquellas variables cuyo coeficiente de Gini sea mayor a 6 y que tengan una correlación no-nula con la variable *objetivo*, es decir, una $r_s > |0.1|$. Este enfoque detectó 13 variables como las más importantes, más, sin embargo, al analizar las correlaciones, se observó que todas las variables contaban con correlaciones válidas distintas de cero salvo

dos, identificadas como *Rellenos* y *Edad*, cuyas correlaciones eran de 0.015 y 0.016 respectivamente. De esta manera se separan 11 variables que pasan a ser las características finales. En La Figura 3(b) se puede observar la *importancia del atributo RF* contra la *correlación de la variable objetivo*. Note como se crean dos líneas de intersección que hacen que se puedan separar las variables más importantes (puntos de color negro). La línea de corte en la correlación (línea de color rojo) y el atributo importancia (línea de color verde) permiten seleccionarlas características más importantes. Finalmente, se obtienen las siguientes 11 características: Verbos sin declinar, Media de los morfemas por oración, Errores de palabras, Promedio de sílabas por palabra, Frecuencia de tipos de palabras, Uso regular del pasado, Media de las palabras por oración, Número de etiquetas y número de palabras.

La siguiente etapa del *pipeline* consiste en evaluar la predictibilidad de las 11 características seleccionadas en la etapa anterior, aplicando el modelo de regresión logística de manera consecutiva. Este enfoque, permite ir descartando aquellas variables que no consiguieran un p-value menor a 0.05. El experimento se repite cíclicamente, hasta ya no poder conseguir descartar más características. Recordar que la variable objetivo en este caso es la que esta etiquetada ya sea como un TD o SLI. Este proceso permitió seleccionar un conjunto final de 6 características relevantes para el análisis. En la Tabla 1 se detallan estas características: Relación de uso de verbos sin declinar ante los declinados, Longitud media de morfemas por oración, Errores de palabras identificados en las transcripciones, Promedio de sílabas por palabra, Frecuencia de tipos de palabras respecto al número total de palabras y uso regular del pasado. Además, en la tabla se presentan los estimados, errores estándar y valores z asociados a cada una de estas características, que son medidas fundamentales para comprender su contribución al modelo predictivo y su significancia estadística en la clasificación de niños con trastornos específicos del lenguaje (SLI).

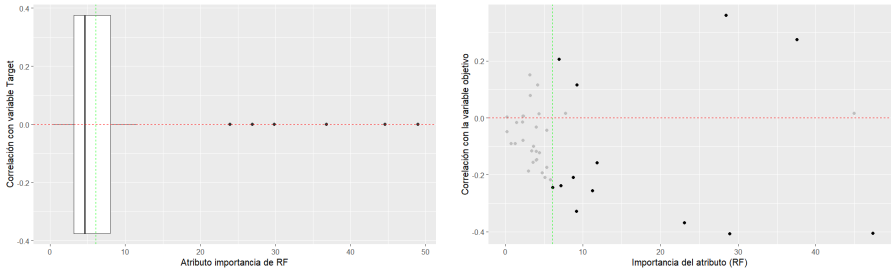
Finalmente, considérese dos posibles clases $C = 0$ (Desarrollo típico, TD) y $C = 1$ (trastorno específico del lenguaje, SLI) para el vector de características Θ_{Ct} dado por las 6 características dadas por al etapa anterior. Se propone usar 14-NN introducido en la Sección (2.3) como clasificador para detectar SLI a partir de transcripciones de narrativas espontáneas. A modo de ilustración, en la Figura 2(b), se puede apreciar una tendencia de los pacientes con SLI (Azul) con respecto a los pacientes con TD (Rojo), específicamente la relación entre las características *Vocales por sílabas* vs *Morfemas*. Se puede observar como la dispersión de los grupos están bastante superpuestos, lo que hace que el modelo 14-NN sea una buena opción de clasificación. El k seleccionado surge de aplicar la raíz cuadrada al numero de observaciones que componen el conjunto de entrenamiento, el cual es el 70% de las 1063 observaciones que forman los datos [35]. Los hiperparámetros de k-NN se eligieron usando validación cruzada de 5 pliegues, dentro de un rango posible de valores menor a 27 (El cual surge de calcular $\sqrt{N}$, donde N es igual al número de observaciones usados en el entrenamiento de modelo, el cual es de 730). De esta manera, el n que minimiza el promedio de error de MAE, RMSE y R^2 es igual a 14. Con este valor se procede

a dividir los datos con una selección aleatoria entre entrenamiento y prueba de 70% y 30% respectivamente. El modelo de clasificación k-NN, optimizado con validación cruzada quíntuple, demostró una mejora considerable en su precisión, a raíz de las selecciones de características propuestas en el *pipeline*.



(a) Modelo con respecto a cada clase (b) Relación Morfemas Vs Silabas

Fig. 2: (a) Comportamiento del modelo con respecto a cada clase de predicción. Observe como los árboles se empiezan a estabilizar a partir del valor 16 aproximadamente, para el tipo 1 (Clase 1) y OOB, mientras que para el tipo 0 (Clase 0), se empieza a estabilizar en 200 aproximadamente. Esta inspección visual permite establecer el valor de los árboles en 500. (b) Relación entre Cantidad de *Morfemas* por *Respuesta y promedio de silabas*



(a) Boxplot (b) Gráfico de dispersión

Fig. 3: Comportamiento para el atributo de importancia de RF y correlación usando (a) Boxplot y (b) Gráfico de dispersión. Se puede observar los criterios de selección en rojo y verde en ambos gráficos. Siendo el de correlación nula en rojo y de atributo importancia de *Random Forest* mayor a seis, en verde. Los puntos grises consisten variables descartados, en tanto los negros, características

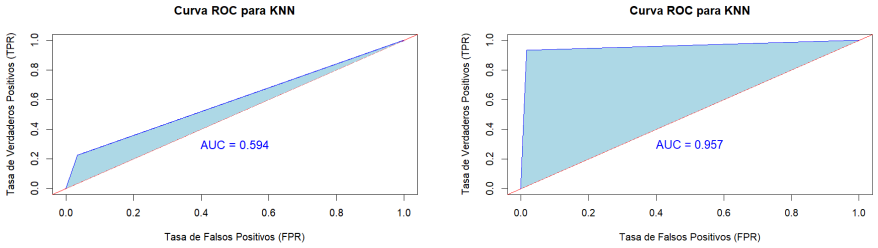
Para demostrar la potencialidad de este enfoque, considérese las 11 características dadas por RF y la correlación, contra las 6 características dadas por el modelo de regresión logística. Como se puede observar en las Figuras 4(a) y 4(b), se destaca la capacidad del modelo para distinguir entre clases positivas que mostró un incremento importante, en particular en el caso de las clases negativas en donde la especificidad escaló de 0.22 con 11 características a 0.95 con 6 características en la identificación de niños con dificultades en el lenguaje. Lo

cual sugiere que el *pipeline* diseñado es una buena herramienta en la detección de SLI.

Para obtener una visión integral de la eficacia del modelo, se utilizaron las métricas de precisión, F1-score, sensibilidad y especificidad. Los resultados, respaldan su utilidad clínica para la detección temprana de SLI, con suficiente evidencia estadística para afirmar que, los resultados no son casuales. Cómo pueden verse en la Tabla 3, la combinación de métodos de selección de características permitió, a través de las métricas de predicción evaluadas, determinar que el *pipeline* sugerido, puede ser una buena herramienta de análisis. Además, cabe notar, que la reducción de dimensionalidad fue bastante óptima, pues se pasaron de 43 variables iniciales a 6 finales. Es interesante notar que las métricas mejoraron ostensiblemente al usar solamente las 6 variables propuestas. una sensibilidad del 98% y una especificidad del 93%, indican su capacidad para identificar tanto los casos positivos como los negativos con alta precisión. Por otro lado, el F1-Score, que combina precisión y recall, alcanza un valor del 98% y 97% respectivamente.

Table 1: Resultados de aplicar regresión logística. El ciclo de Regresiones termina al encontrar un conjunto de características que cumpla la condición de p-value < 0.05. Se debe lograr ya no poder descartar elementos.

Características	Estimados	Std. error	z value	$p > \|z\|$
Verbos sin declinar	1.25727	0.28994	4.336	≈ 0
Morfemas por oración	-0.84605	0.08618	-9.817	≈ 0
Errores	0.85750	0.11493	7.461	≈ 0
Promedio silabas	-2.64640	1.10947	-2.385	0.01707
Frecuencia de tipos	-2.73838	0.91128	-3.005	0.00266
Pasado regular	-0.05929	0.01623	-3.652	0.00026



(a) 11 covariables (b) 6 covariables

Fig. 4: Resultados del modelo 14-NN para (a) 11 y (b) 6, covariables.

Estos resultados son respaldados por la matriz de confusión presentada en la Tabla 3. La matriz se compone de los resultados del modelo en las predicciones de los valores de prueba, que componen el 30% de los datos por un total de 314 observaciones. El modelo logró predecir correctamente 238 casos positivos y 67 casos negativos. Sin embargo, se observaron 5 falsos negativos y 4 falsos positivos en las predicciones. Estos resultados sugieren una mejora sustancial en

la precisión al reducir el número de variables, con un F1-score mucho más alto usando 6 características.

	P Predicho	N Predicho
P Real	238	5
N Real	4	67

Table 2: Matriz de confusión de 14-NN con 6 características.

características	**Resultados**			
	Prec	***F1sc***	***Sens***	***Espe***
11	82.8%	33.1%	96.4%	22%
6	97.13%	98.74%	98.71%	95.06%

Table 3: Resultados del Modelo 14-NN propuesto

Los resultados obtenidos sugieren que la combinación de técnicas de selección de características, como *Random Forest* y Regresión Logística, puede ser una herramienta potente para mejorar la eficacia del modelo de clasificación. La reducción del conjunto de variables a seís destacó la relevancia de características específicas, como *el uso de Verbos sin declinar, la media de morfemas por oración, los errores cometidos. el promedio de silabas por palabra, la frecuencia de tipos de palabras y el uso de verbos en pasado regular*. En este contexto, surge la importancia de la toma de muestras de narrativas espontáneas de niños en el ámbito de la psiquiatría infantil. Para ello cobran suma importancia que existan metodologías que utilicen variables cuantitativas directamente relacionadas con el desempeño del niño, como pueden ser, el número de morfemas o el promedio de vocales por palabra. Este tipo de técnicas estandarizadas podrían motivar la toma de muestras en nuevos pacientes, dado que son escasas las bases de datos y resulta complejo de acceder a material de audio o transcripciones de niños. La presente investigación se destaca por la simpleza del modelo, su precisión y su replicabilidad, en tanto no utiliza complejas variables que requieren del uso de indicadores abstractos de la psicología o herramientas de la fonoaudiologia, que lo vuelva al experimento complejo de replicar, o hasta incluso incompatible, con transcripciones en idioma inglés como español. Es interesante notar que el *pipeline* propuesto tiene la siguiente complejidad en términos de *Big* $\mathcal{O}$: El método *Random Forest* $\mathcal{O}(n\log(n)Vm)$, el coeficiente de correlación de *Spearman* $\mathcal{O}(n)$; regresión logística $\mathcal{O}(nV)$ y el clasificador de k vecinos más cercanos $\mathcal{O}(knV)$, donde V es el tamaño de la características y m es la profundidad de los árboles. Esto sugiere que el *pipeline* no presenta una alta complejidad computacional. La Tabla 4 muestra una comparación del *pipeline* propuesto, con otros trabajos del estado-de-arte, destacando su singularidad y eficacia en términos de sus métricas. Observe que los resultados son alentadores frente a modelos de alta complejidad como las redes neuronales o SVM.

4 Conclusiones

Este estudio se propuso un *pipeline* en cascada de 3 etapas, que permite, a través de un enfoque simple y eficiente, la detección de SLI en niños. La precisión del modelo de 97.13%, sugiere su viabilidad clínica como herramienta de detección temprana de SLI. En la primera etapa, se realizó una extracción de características y una reducción de dimensionalidad de los datos usando los métodos de *Random Forest* (RF) y correlación en conjunto, logrando reducir de 43 a 11 variables. En

la segunda etapa, se estimaron las variables más predictivas usando regresión logística, obteniéndose 6 variables finales de las 11 de la etapa anterior. Estas variables son usadas en la última etapa, para detectar el trastorno SLI en niños a partir de transcripciones de narrativas espontáneas. En resumen, el *pipeline* diseñado permitió reducir de 43 variables a 6 variables, lo que da una luz en la detección de SLI.

Table 4: Comparación con algunos trabajos del estado-del arte en clasificación de SLI. DDT: Datos directos de transcripciones procesadas con técnicas de NLP, RNN: Redes neuronales recurrentes, CNN: Redes neuronales convolucionales, ANN: Redes neuronales artificiales, SVM: Support Vector Machines.

Método	Características	Precisión	Ref
14-NN	DDT	97.77%	Este trabajo
CNN	DDT	99%	[2]
SLINet	CNN 2D	99.09%	[22]
likelihood ratios	Pruebas de repetición de dígitos, no-palabras y oraciones	94%	[16]
SVM, RF y RNN	Señales de voz	99.00%	[17]
SVM y feed-forward NN	Fuente glotal y coeficientes cepstrales de frecuencia *Mel*	98.82%	[18]
Naïve Bayes, SVM y ANN	Longitud media de emisiones y estructuras gramaticales	79%,80%,76%	[19]

El *pipeline* propuesto, presenta tres fortalezas notables. Es de baja complejidad computacional; presenta una reducción de dimensionalidad de los datos siguiendo criterios precisos a partir de los datos y permite un tratamiento de la información en varios estadios, permitiendo realizar una selección de las características que logran un gran nivel predictivo de la variable objetivo. La combinación de NLP y ML abre nuevas posibilidades para diagnósticos precisos y eficientes, con un potencial impacto en la identificación temprana y el diseño de intervenciones personalizadas.

En cuanto a las limitaciones, se considera lo experimental del proyecto, en tanto no fue probado con otros tipos de datos; también la estimación del valor k del clasificador, puede ser óptimo o no, en muchos casos es un valor empírico que busca minimizar el error en la etapa de clasificación, por ende puede tener un rango de posibles valores. Resulta importante destacar en cuanto a las iteraciones de las regresiones que no siempre es conveniente proceder hasta hallar el mínimo conjunto posible de características. Dado que pueden ocurrir dos escenario. Uno donde se tiene un numero importante de variables que afectan en gran medida la capacidad del modelo de predecir la variable objetivo, y otro donde se tiene una selección reducida de características. En el primer caso se recomienda continuar efectuando regresiones con el fin de reducir la dimensión de trabajo. No obstante en el segundo caso, se puede optar por no correr otra regresión, en tanto se perderían variables de interés por una mejoría despreciable en el modelo. Esto queda sujeto al caso de uso y la consulta de un profesional en el área.

Futuras investigaciones se centrarán en explorar la aplicación de este enfoque en poblaciones más amplias, evaluar su utilidad en entornos clínicos, adaptar el *pipeline* a varios tipos de datos, así como implementar esta metodología en otros campos.

5 Disponibilidad del software

En https://github.com/SantiagoarenaDS/Pipeline-SLI-JAIIO2024, accessed: 2024-24-03, esta disponible el software utilizado en este estudio.

References

1. Leonard, L.B.: Children with Specific Language Impairment. Bradford Books (2014)
2. Sharma, Y., Singh, B.K.: One-dimensional convolutional neural network and hybrid deep-learning paradigm for classification of specific language impaired children using their speech. Computer Methods and Programs in Biomedicine **213**, 106487 (2022)
3. Barua, P.D., Aydemir, E., Dogan, S., Erten5, M., Kaysi, F., Tuncer, T., Fujita, H., Palmer, E., Acharya1, U.R.: Novel favipiravir pattern-based learning model for automated detection of specific language impairment disorder using vowels. Neural Computing and Applications (35), 6065–6077 (2023)
4. Association, A.P.: Diagnostic and Statistical Manual of Mental Disorders, DSM-5^{TR}. Amer Psychiatric Pub Inc (2022)
5. Webster, R., Majnemer, A., Platt, R., Shevell, M.: The predictive value of a preschool diagnosis of developmental language impairment. Neurology **12**(63), 2327–2331 (2004)
6. Pentimonti, J.M., Murphy, K.A., Justice, L.M., Logan, J.A.R., Kaderavek, J.N.: School readiness of children with language impairment: predicting literacy skills from pre-literacy and social–behavioural dimensions. International Journal of Language & Communication Disorders **51**(2), 148–161 (2016)
7. Catts, H., Fey, M., Tomblin, J., Zhang, X.: A longitudinal investigation of reading outcomes in children with language impairments. Journal of Speech, Language, and Hearing Research **6**(45), 1142–1157 (2002)
8. Hogan, T.F., Bridges, M.S., Justice, L.M., Cain, K.: Increasing higher level language skills to improve reading comprehension. Focus on Exceptional Children **44**(3), 1–20 (2011)
9. Beitchman, J.H., Brownlie, E.B., Inglis, A., Wild, J., Ferguson, B., Schachter, D., Lancee, W., Wilson, B., Mathews, R.: Seven-year follow-up of speech/language impaired and control children: psychiatric outcome. Journal of Child Psychology and Psychiatry **37**(8), 961–970 (1996)
10. Stanton-Chapman, T., Justice, L., Grant, S.L.: Social and behavioral characteristics of preschoolers with specific language impairment. Linguistics, Education, Psychology **1** (2007)
11. Justice, L.M., Ahn, W.Y., Logan, J.A.R.: Identifying children with clinical language disorder: An application of machine-learning classification. Journal of Learning Disabilities **52**(5), 351–365 (2019)
12. Gabani, K., Solorio, T., Liu, Y., Hassanali, K., Dollaghan, C.A.: Exploring a corpus-based approach for detecting language impairment in monolingual English-speaking children. Artificial intelligence in medicine **53 3**, 161–70 (2011)
13. MacFarlane, H., Gorman, K., Ingham, R., Presmanes Hill, A., Papadakis, K., Kiss, G., van Santen, J.: Quantitative analysis of disfluency in children with autism spectrum disorder or language impairment. Plos One **12**(3), 1–20 (03 2017)
14. Kaushik, M., Baghel, N., Burget, R., Travieso, C.M., Dutta, M.K.: Slinet: Dysphasia detection in children using deep neural network. Biomedical Signal Processing and Control **68**, 102798 (2021)
15. Gray, S.: Diagnostic accuracy and test-retest reliability of nonword repetition and digit span tasks administered to preschool children with specific language impairment. J Commun Disord **36**(2), 129–151 (2003)

16. Armon-Lotem, S., Meir, N.: Diagnostic accuracy of repetition tasks for the identification of specific language impairment (SLI) in bilingual children: evidence from Russian and Hebrew. Int J Lang Commun Disord **51**(6), 715–731 (2016)
17. Slogrove, K.J., van der Haar, D.: Specific language impairment detection through voice analysis. In: Abramowicz, W., Klein, G. (eds.) Business Information Systems. Lecture Notes in Business Information Processing, vol. 389. Springer, Cham (2020)
18. Reddy, M.K., Alku, P., Rao, K.S.: Detection of specific language impairment in children using glottal source features. IEEE Access **8**, 15273–15279 (2020)
19. Oliva, J., Serrano, J.I., del Castillo, M.D., Ángel Iglesias: Computational cognitive modeling for the diagnosis of specific language impairment. In: et al., B.B. (ed.) Data and Knowledge for Medical Decision Support. IOS Press (2013)
20. Child language data exchange system. https://childes.talkbank.org/, accessed: 2024-24-06
21. Brown, R.: A first language: The early stages. George Allen & Unwin, London (1973)
22. Bowen, C.: Brown's stages of syntactic and morphological development. Retrieved from www. speech-language-therapy. com/index. php (1998)
23. Violini, M.L.: Selección de características. Su aplicación a clasificación de texturas. Bachelor's thesis, Universidad Nacional de La Plata (12 2014)
24. Botting, N., Conti-Ramsden, G.: The role of language, social cognition, and social skill in the functional social outcomes of young adolescents with and without a history of SLI. British Journal of Developmental Psychology **26**(2), 281–300 (2008)
25. Biau, G., Scornet, E.: A random forest guided tour. TEST (25), 197–227 (2016)
26. Quintero-Rincón, A., D'giano, C., Batatia, H.: A quadratic linear-parabolic model-based EEG classification to detect epileptic seizures. J Biomed Res. **34**(3), 205–212 (2019)
27. Breiman, L.: Random forests. Machine Learning **45**, 5–32 (2001)
28. Becker, R.A., Chambers, J.M., Wilks, A.R.: The New S Language. Wadsworth & Brooks/Cole (1988)
29. Cheng, Q., Varshney, P., Arora, M.: Logistic regression for feature selection and soft classification of remote sensing data. IEEE Geoscience and Remote Sensing Letters **3**(4), 491–494 (2006)
30. Zakharov, R., Dupont, P.: Ensemble logistic regression for feature selection. In: Loog, M., Wessels, L., Reinders, M.J.T., de Ridder, D. (eds.) Pattern Recognition in Bioinformatics. pp. 133–144. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)
31. Flach, P.: Machine Learning, The Art and Science of Algorithms that Make Sense of Data. Cambridge (2012)
32. James, G., Witten, D., Hastie, T., Tibshirani, R.: An Introduction to Statistical Learning: with Applications in R. Springer (2013)
33. Heumann, C., Schomaker, M., Shalabh: Introduction to Statistics and Data Analysis: With Exercises, Solutions and Applications in R. Springer (2022)
34. Quintero-Rincón, A., Muro, V., D'Giano, C., Prendes, J., Batatia, H.: Statistical model-based classification to detect patient-specific spike-and-wave in EEG signals. Computers **9**(4),1–14 (2020)
35. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006)
36. Barber, D.: Bayesian Reasoning and Machine Learning. Cambridge University Press (2012)
37. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International Joint Conference on Artificial Intelligence (1995)