UNIVERSIDAD DE SANTIAGO DE CHILE

FACULTAD DE INGENIERÍA

Departamento de Ingeniería Informática

PREDICCIÓN DE LA DIFICULTAD PERCIBIDA ANTE UNA TAREA DE BÚSQUEDA EN ESTUDIANTES DE ENSEÑANZA BÁSICA A PARTIR DEL ANÁLISIS DE DINÁMICAS DE TIPEO

Álvaro Ignacio Valenzuela Alfaro

Profesor guía: Roberto González Ibañez

Profesor co-guía: Daniel Gacitúa Vásquez

Tesis para optar al título de Ingeniero Civil en Informática

Santiago – Chile

2025

Proyecto FONDECYT Regular - N°1201610 - TUTELAGE.

RESUMEN

Recientes investigaciones evidencian que los estudiantes escolares en Chile presentan dificultades al momento de realizar búsquedas de información en la web. Entre los factores asociados a estas dificultades se encuentra la percepción subjetiva del estudiante respecto al nivel de complejidad de una tarea de búsqueda. Sin embargo, existe un déficit de mecanismos automáticos para detectar y anticipar estas dificultades durante las sesiones de búsqueda de información. Por esta razón, el estudio abordará la problemática de la estimación de la dificultad percibida de un desafío de información sin que el estudiante deba explicitarlo, a través de la detección de la dificultad percibida mediante una técnica no intrusiva (sin sensores externos) llamada Keystroke Dynamics, la cual permite analizar la forma en que un usuario escribe en un computador, monitorizando las pulsaciones de teclado e intentando identificarlas basándose en patrones rítmicos habituales en su forma de escribir (Monrose & Rubin, 2000).

Con el fin de dar respuesta a esta problemática de manera sistemática y reproducible, la investigación se estructura siguiendo la metodología de trabajo OSEMN (Obtain, Scrub, Explore, Model, iNterpret), en donde se abordan dos grandes objetivos; el primero, capturar y transformar los datos utilizando los flujos de trabajo desarrollados por Brevis (2022) y Latapiatt (2022). Y el segundo, utilizar los aprendizajes y resultados empíricos obtenidos por Gacitúa y González (2022) en su estudio sobre la predicción de la dificultad percibida en sesiones de búsqueda realizadas por adultos.

Se espera que este estudio permita identificar patrones en el comportamiento de escritura (Keystroke Dynamics) asociados a niveles de dificultad percibida durante sesiones de búsqueda de información en estudiantes escolares. En concreto, se espera obtener al menos un modelo de clasificación binaria con resultados similares a los alcanzados por Gacitúa (Gacitúa & González, 2022) en tareas y grupos demográficos distintos a los empleados en este estudio.

Palabras Claves: Alfabetización Informacional, Dinámica de Pulsaciones, Dificultad Percibida, Estudiantes

La educación no es algo que uno puede terminar – Isaac Asimov

AGRADECIMIENTOS

Quiero agradecer a todas las personas que me apoyaron durante este largo proceso de aprendizaje, quienes han sido un pilar fundamental desde el primer día en que inicie mis estudios.

En primer lugar, deseo agradecer a Tamara por su compañía y apoyo constante. En los momentos más difíciles, cuando todo parecía salir mal, estuvo siempre presente, brindándome su amor y respaldo incondicional.

Agradezco a mis padres, Carola y Sergio, por haberme transmitido valores esenciales y mostrarme que la vida está para descubrirla. Aprecio de manera especial el privilegio que me otorgaron al permitirme estudiar y enseñarme que, con esfuerzo y dedicación, todos los sueños y anhelos son alcanzables.

También agradezco de una manera muy especial a mi abuela Olga, quien pudo acompañarme durante gran parte de mi vida. Sé que estarías muy orgullosa de todo lo que he logrado.

Además, quisiera agradecer profundamente a cada uno de los profesores del Departamento de Ingeniería en Informática que formaron parte de mi proceso de aprendizaje durante estos años. En especial, a los profesores Dra. Luz Chourio Acevedo, Dr. Roberto Gonzalez-Ibáñez y Dr. Daniel Gacitúa Vázquez, por su excelente disposición y profesionalismo durante la realización de mi trabajo. Su apoyo fue clave para poder llegar hasta aquí.

Por último, agradezco al proyecto TUTELAGE, FONDECYT Regular N° 1201610, por el acceso a los datos utilizados a lo largo de este trabajo.

ÍNDICE GENERAL

1 Introducción **1**
- 1.1 Antecedentes y motivación . . . 1
- 1.2 Descripción del problema . . . 2
- 1.3 Solución propuesta . . . 4
- 1.4 Pregunta de investigación . . . 5
- 1.5 Objetivos y alcances del proyecto . . . 5
  - 1.5.1 Objetivo general . . . 5
  - 1.5.2 Objetivos específicos . . . 6
  - 1.5.3 Alcances . . . 6
- 1.6 Organización del documento . . . 7

2 Marco teórico **8**
- 2.1 Marco conceptual . . . 8
  - 2.1.1 Alfabetización Informacional . . . 8
  - 2.1.2 Carga cognitiva . . . 9
  - 2.1.3 Variables proxy . . . 11
  - 2.1.4 Aprendizaje Automático . . . 11
    - 2.1.4.1 Algoritmos bayesianos . . . 12
    - 2.1.4.2 Árboles de decisión . . . 12
    - 2.1.4.3 Regresión logística . . . 14
    - 2.1.4.4 Support Vector Machine . . . 15
    - 2.1.4.5 Modelo de Mezcla Gaussiana (GMM) . . . 16
  - 2.1.5 Ecosistema NEURONE . . . 18
- 2.2 Estado del arte . . . 18
  - 2.2.1 Keystroke Dynamics . . . 18
  - 2.2.2 Comportamiento y habilidades de búsqueda de información . . . 20
  - 2.2.3 Dificultad percibida en una tarea de búsqueda . . . 22
  - 2.2.4 Keystroke Dynamics y carga cognitiva . . . 23
  - 2.2.5 Habilidades de escritura en teclado en estudiantes de enseñanza básica . . . 25
- 2.3 Resumen . . . 25

3 Metodología **27**
- 3.1 Origen de los datos . . . 27
  - 3.1.1 Diagnóstico . . . 28
  - 3.1.2 Intervención . . . 29
  - 3.1.3 Descripción de los datos . . . 31
- 3.2 Pipeline de trabajo . . . 33
- 3.3 Construcción de tabla de características . . . 36
- 3.4 Técnicas de minería de datos . . . 39
- 3.5 Ambiente experimental y hardware requerido . . . 42
- 3.6 Resumen . . . 43

4 Análisis exploratorio **44**
4.1 Exploración preliminar de los datos . . . 45
4.1.1 Eliminación de valores atípicos . . . 50
4.1.2 Distribución de segmentos de búsqueda . . . 53
4.1.3 Valores faltantes . . . 53
4.1.4 Desbalances . . . 54
4.1.5 Resumen de observaciones eliminadas . . . 55
4.2 Caracterización de patrones de tipeo mediante técnicas no supervisadas . 55
4.2.1 Reducción de dimensionalidad (PCA) . . . 56
4.2.2 Identificación de perfiles . . . 60
4.2.2.1 Caracterización de los perfiles de tipeo identificados . . 60
4.2.2.2 Dinámica de transición entre perfiles . . . 62
4.3 Relación entre velocidad tipeo y dificultad percibida . . . 64
4.3.1 Comparación estadística entre niveles de dificultad . . . 65
4.3.2 Relaciones entre métricas de tipeo . . . 67
4.3.3 Regresión logística para la predicción de dificultad percibida . . . 68
4.4 Factores contextuales y experimentales . . . 71
4.4.1 Análisis global de la muestra . . . 71
4.4.2 Análisis diferenciado por grupo experimental . . . 73
4.5 Relación entre tipo de tarea y dificultad percibida . . . 74
4.6 Resumen . . . 75
4.6.1 Existencia de perfiles de comportamiento . . . 75
4.6.2 Validación de las métricas como proxy de dificultad . . . 75
4.6.3 Sensibilidad a la intervención pedagógica . . . 76

5 Modelamiento y resultados **77**
5.1 Selección de características . . . 77
5.2 Evaluación de modelos en pre y post test . . . 78
5.2.1 Comparativas de desempeño en pre test . . . 79
5.2.2 Comparativas de desempeño en post test . . . 80
5.3 Pruebas en diagnóstico . . . 83
5.4 Importancia de características . . . 86
5.5 Resumen . . . 88
5.5.1 Viabilidad técnica . . . 88
5.5.2 Dependencia del nivel de competencia . . . 89
5.5.3 Prioridad en la detección . . . 89

6 Discusión **90**
6.1 Dinámicas de tipeo como proxy de la dificultad percibida . . . 90
6.2 El efecto de la intervención . . . 91
6.3 Resultados cercanos al azar en diagnóstico . . . 91
6.4 Comparación con el estado del arte . . . 93
6.5 Resumen . . . 94

7 Conclusiones **95**

7.1 Objetivos específicos 95
7.2 Objetivo general 97
7.3 Implicancias 98
7.3.1 Implicancias teóricas 98
7.3.2 Implicancias prácticas 98
7.4 Alcances y limitaciones 99
7.5 Trabajo futuro 99
7.6 Reflexiones finales 100

Referencias bibliográficas **102**

Anexos **107**

A Documentos administrativos **107**
A.0.1 Licencia 107
A.0.2 Carta de compromiso 108

B Definiciones técnicas **109**
B.0.1 Diccionario de datos 109

C Resultados Diagnóstico **113**

D Test pareado pre test y post test en Intevención **121**

E Análisis estadístico Diagnóstico **125**

F Análisis por nivel socioeconómico **130**

G Análisis de valores atípicos **134**

ÍNDICE DE TABLAS

Tabla 3.1 Tabla de características 39
Tabla 3.2 Modelos ordenados según AUC obtenido 41
Tabla 3.3 Configuración de grillas de hiperparámetros por modelo 42

Tabla 4.1 Estadísticas descriptivas de variables de 1er Segmento de consulta, Intervención 46
Tabla 4.2 Resultados de la prueba de normalidad Shapiro–Wilk 51
Tabla 4.3 Proceso de filtrado y limpieza del conjunto de datos 55
Tabla 4.4 Resultados de pruebas de contraste por variable – Wilcoxon no pareado (pre y post test) 66
Tabla 4.5 Resultados de pruebas pareadas Pre–Post y validación de normalidad ($N = 77$) 72

Tabla 5.1 Variables del conjunto pre-test seleccionadas mediante Boruta y RFE (ordenadas por importancia) 78
Tabla 5.2 Variables del conjunto post-test seleccionadas mediante Boruta y RFE (ordenadas por importancia) 78
Tabla 5.3 Comparación de clasificadores 80
Tabla 5.4 Desempeño de clasificadores (validación cruzada) 82
Tabla 5.5 Métricas de clasificación para el dataset diag-q1 (CV 10-fold, 3 repeticiones) 85
Tabla 5.6 Distribución de clases de dificultad percibida por conjunto de datos . 86

Tabla B.1 Estructura de Queries 109
Tabla B.2 Estructura de Questionnaires 109
Tabla B.3 Estructura de Studies 110
Tabla B.4 Estructura de UserChallenge 110
Tabla B.5 Estructura de UserDatas 110
Tabla B.6 Estructura de UserQuestionnaires 111
Tabla B.7 Estructura de Users 111
Tabla B.8 Estructura de UserStudies 111
Tabla B.9 Estructura de VisitedLinks 112

Tabla C.1 Estadísticas descriptivas de variables del conjunto completo de tipeo en diagnóstico. 114
Tabla C.2 Conteo de valores nulos por variable 115
Tabla C.3 Pruebas de significancia entre grupos de dificultad percibida (Fácil vs Difícil). 120

Tabla D.1 Resultados de pruebas pareadas Pre–Post (Grupo Control, $N = 25$) . 121
Tabla D.2 Resultados de pruebas pareadas Pre–Post (Grupo Tradicional, $N = 19$) 122
Tabla D.3 Resultados de pruebas pareadas Pre–Post (Grupo Asistente Virtual, $N = 28$) 123

Tabla D.4 Resultados de pruebas pareadas Pre–Post (Grupo Ludificado, $N = 30$) 124

Tabla E.1 Resultados de pruebas (4° Básico) 125
Tabla E.2 Resultados de pruebas no pareadas Pre–Post (5° Básico) 126
Tabla E.3 Resultados de pruebas no pareadas Pre–Post (6° Básico) 127
Tabla E.4 Resultados de pruebas no pareadas Pre–Post (7° Básico) 128
Tabla E.5 Resultados de pruebas no pareadas Pre–Post (8° Básico) 129

Tabla F.1 Caracterización socioeconómica de los establecimientos participantes 130
Tabla F.2 Resultados de la prueba de Kruskal-Wallis por variable en GSE . . . 131
Tabla F.3 Resultados completos del análisis post hoc mediante la prueba de Dunn con corrección de Bonferroni 131

Tabla G.1 Resultados de la prueba de suma de rangos de Wilcoxon: Comparativa entre Dataset Filtrado y Outliers. 135

ÍNDICE DE FIGURAS

Figura 2.1 Estructura de la memoria humana según la teoría de la carga cognitiva. 10
Figura 2.2 Representación gráfica de Support Vector Machine 15
Figura 2.3 Agrupamiento utilizando GMM en distribuciones de distintas covarianzas 17
Figura 2.4 Métricas para Keystroke Dynamics en dígrafos 20
Figura 2.5 Representación visual de los acercamientos para la agregación de los segmentos de consulta. 23

Figura 3.1 Demografía de los estudiantes en el estudio de diagnóstico 28
Figura 3.2 Fases del estudio de intervención 30
Figura 3.3 Diseño del estudio de intervención 30
Figura 3.4 Diagrama de la metodología OSEMN 34
Figura 3.5 Diagrama de flujo del pipeline implementado 36
Figura 3.6 Tabla visitedLinks filtrada 37
Figura 3.7 Ejemplo ventana de tiempo utilizada en la agregación 38

Figura 4.1 Gráfico de caja, intervención, primer segmento de consulta 48
Figura 4.2 Distribuciones de características, intervención, primer segmento . 49
Figura 4.3 Eliminación de valores atípicos, 1er segmento de consulta, intervención. 52
Figura 4.4 Distribución de segmentos de consulta, intervención, primer segmento de consulta 53
Figura 4.5 Desbalance en conjunto de datos de Intervención 54
Figura 4.6 Gráfico de codo, 1er segmento, intervención 57
Figura 4.7 Análisis de componentes principales, 1er segmento, intervención 58
Figura 4.8 Gráfico de contribución por componente 59
Figura 4.9 Gráfico de caja para grupos generados por GMM 61
Figura 4.10 Grupos generados por GMM 62
Figura 4.11 Gráfico aluvial para perfiles de pre test y post test 64
Figura 4.12 Matriz de correlaciones, 1er segmento, intervención 68
Figura 4.13 Regresión logistica para PC1 69
Figura 4.14 Regresión logistica para PC2 70
Figura 4.15 Representación gráfica test chi cuadrado 74

Figura 5.1 Área bajo la curva (ROC) en intervención pre test 79
Figura 5.2 Área bajo la curva (ROC) en intervención post test 81
Figura 5.3 Matriz de confusión de modelo XGBoost 83
Figura 5.4 Área bajo la curva (ROC) en diagnóstico 84
Figura 5.5 Importancia de características en pre test 87
Figura 5.6 Importancia de características en post test 88

Figura 6.1 Variabilidad en el intervalo entre pulsaciones de teclado (IKI, Inter-Key Interval, en milisegundos) versus la edad 93

Figura A.1 Carta de compromiso firmada 108

Figura C.1 Eliminación de valores atípicos, 1er segmento de consulta, diagnóstico. 113
Figura C.2 Distribución de variables, primer segmento de consulta. 116
Figura C.3 Boxplot variables luego de eliminación de valores atípicos, primer segmento de consulta. 117
Figura C.4 Correlacion de spearman para características, primer segmento de consulta. 118
Figura C.5 Análisis componentes principales, primer segmento de consulta . 119

Figura G.1 Distribución de valores con valores atípicos 136
Figura G.2 Distribución de datos atípicos versus datos retenidos 136
Figura G.3 Porcentaje de observaciones de copy-paste 137

CAPÍTULO 1. INTRODUCCIÓN

1.1 ANTECEDENTES Y MOTIVACIÓN

Estudios de la Asociación Internacional para la Evaluación del Logro Educativo (IEA) evidencian que, en Chile, un 38 % de los estudiantes de educación primaria tienen problemas para manejar, conocer y seleccionar la información relevante proveniente de las Tecnologías de la Información (TIC) (Agencia de Calidad de la Educación, 2020). Dada la importancia de estas habilidades en el mundo actual, las ciencias de la computación deben contribuir a este campo con herramientas para apoyar el desarrollo de estas competencias en las personas (Chourio-Acevedo et al., 2024). En primer lugar, el desarrollo de la alfabetización informacional es de suma importancia para los niños, ya que les habilita a desarrollar habilidades para investigar de forma autónoma más allá del aula (American Library Association (ALA) & Libraries (ACRL), 2000). En segundo lugar, en cuanto a las nuevas generaciones, estas demuestran facilidad para utilizar e integrarse a las tecnologías de la información; sin embargo, esto no significa que la discriminación de información y el uso de herramientas digitales sean uno de sus puntos fuertes (Agencia de Calidad de la Educación, 2020). En tercer lugar, en Chile se ha demostrado un notorio avance en materias de infraestructura tecnológica a través de los años; sin embargo, el último informe de la Agencia de Calidad de la Educación, 2020 dejó en evidencia que los estudiantes experimentan dificultades al formular consultas en sistemas de recuperación de información.

Estas dificultades en la formulación de buenas consultas no son solo un obstáculo puntual dentro del aula, sino un desafío a más largo plazo: es evidente que un estudiante no podrá aprender todos los contenidos durante su etapa escolar debido principalmente a la explosión de información que caracteriza al siglo XXI, a menudo denominado la "era de la información". La cantidad de datos y fuentes de información crece a una velocidad sorprendentemente rápida en la sociedad actual, lo que hace imposible que cualquier persona asimile todo el conocimiento necesario en unos pocos años de formación académica (Naik & Padmini, 2014). Por ello, las competencias de búsqueda

y selección de información son de suma importancia para los niños, ya que les habilitan a desarrollar habilidades de búsqueda autónoma mas allá del aula de clases (American Library Association (ALA) & Libraries (ACRL), 2000).

Ahora bien, para que esta autonomía sea posible en la práctica, el manejo eficiente de los sistemas de información es indispensable para desarrollar habilidades de búsqueda. Si bien las nuevas generaciones han demostrado ventajas en el uso de estos sistemas, esto no implica necesariamente que posean habilidades efectivas de búsqueda de información (Çoklar et al., 2017).

Este problema no es únicamente conceptual, sino que también se observa empíricamente en el contexto chileno. Chile presenta un 87,9 % de logro en desarrollo digital escolar (Agencia de Calidad de la Educación, 2020). Si bien esta cifra refleja principalmente la adquisición de infraestructura, no incorpora el desarrollo de competencias en los estudiantes. De hecho, solo un 29 % posee competencias básicas en alfabetización digital, como por ejemplo: buscar, localizar y utilizar información para tareas básicas, mientras que el promedio internacional supera el 48 %, lo que evidencia que muchos estudiantes aún requieren apoyo para realizar búsquedas simples y evaluar información digital con criterio (Agencia de Calidad de la Educación, 2020).

Frente a este escenario —donde existe acceso a tecnología pero persisten brechas en las competencias de búsqueda— se vuelve fundamental poder identificar cuándo un niño está experimentando dificultades reales durante una sesión de búsqueda. Si bien el auto-reporte puede ser una alternativa, esta no siempre refleja la realidad, pudiendo existir sobreestimaciones o subestimaciones (Mayer et al., 2007). Debido a esto, resulta clave poder contar con herramientas para hacer una estimación certera o con bajo nivel de error.

1.2 DESCRIPCIÓN DEL PROBLEMA

Tal como se discutió en la sección anterior, existe una brecha entre el acceso a tecnologías digitales y la capacidad real de los estudiantes para buscar, evaluar y utilizar

información de manera autónoma. El Informe Nacional del Estudio de Alfabetización Computacional y Manejo de Información (Agencia de Calidad de la Educación, 2020) demostró que en Chile existe una proporción significativa de estudiantes que carecen de competencias informacionales, especialmente en las etapas iniciales de la educación formal. En consecuencia, se evidenció que los niños presentan dificultades al momento de realizar búsquedas de información (Agencia de Calidad de la Educación, 2020; Chourio-Acevedo & González-Ibañez, 2024). Frente a esta realidad, surge la necesidad de crear herramientas y estrategias educativas que permitan identificar y abordar estas dificultades a tiempo.

Para un mejor entendimiento de la problemática señalada, se debe considerar el concepto de dificultad percibida, la que se ha definido como:

> La valoración subjetiva e individual que una persona hace sobre el esfuerzo, complejidad o carga mental/física requerida para realizar una tarea o enfrentar una situación. Esta no depende objetivamente de la tarea, sino de cómo el individuo interpreta sus propios recursos para superarla.

Actualmente, el diagnóstico de este tipo dificultad reside en el auto-reporte del estudiante, la cual es una técnica que interrumpe la sesión de búsqueda y que puede estar sesgada por sobreestimaciones o subestimaciones del propio usuario.

Para efectos de comprender y detectar las dificultades reales de los estudiantes durante la búsqueda de información, bajo el proyecto TUTELAGE Chourio-Acevedo y González-Ibañez (2024) generaron dos bases de datos en torno a las sesiones de búsqueda de información de alumnos entre cuarto y octavo básico en cuatro escuelas de la Región Metropolitana. Estas bases de datos incluyeron, entre otras características, información sobre la dificultad percibida auto-reportada por los estudiantes, utilizando una escala Likert de seis puntos, antes y después de cada una de las tareas de búsqueda, así como también información sobre cada una de las teclas que fueron presionadas en estas sesiones. Sin perjuicio de la existencia de los datos, a la fecha estos no han sido investigados en relación a las dinámicas de tipeo y su incidencia en la predicción del concepto de dificultad percibida.

Complementariamente a lo mencionado, la revisión de la literatura evidenció

la existencia de investigaciones en cuanto a las dinámicas de tipeo de niños en un contexto médico, dificultad percibida por niños en contexto pedagógico y dinámicas de tipeo para la predicción de dificultad percibida en adultos, pero persiste una brecha en cuanto a la relación entre los conceptos de dinámicas de tipeo y dificultad percibida aplicado en niños en un contexto educacional.

1.3 SOLUCIÓN PROPUESTA

Dada la ausencia de herramientas que permitan anticipar estas dificultad de manera automática y anticipada, se propone el diseño y entrenamiento de modelos de clasificación supervisada basado en técnicas de Aprendizaje Automático (Machine Learning) utilizando como base teórica los resultados obtenidos por Gacitúa y González (2022), quienes demostraron la viabilidad de predecir estados cognitivos mediante variables de dinámicas de tipeo en usuarios adultos.

El problema de clasificación se define como la necesidad de discriminar, de forma temprana, si una sesión de búsqueda de un alumno de educación básica, pertenece a una de las siguientes categorías:

- Clase “Fácil”: Tareas en las que el estudiante reportó una baja percepción de dificultad, correspondiente a los niveles 1, 2 y 3 de una escala Likert de seis puntos.
- Clase “Difícil”: Tareas en las que el estudiante reportó una alta percepción de dificultad, correspondiente a los niveles 4, 5 y 6 de una escala Likert de seis puntos.

Para desarrollar y resolver este desafío, la solución utilizó las dinámicas de tipeo (Keystroke Dynamics). Esta es una técnica biométrica no intrusiva, que consiste en el análisis de los patrones rítmicos de escritura de un usuario, tales como el tiempo de presión de teclas y las latencias entre ellas, mediante la monitorización de las pulsaciones de teclado. De esta manera, las variables fueron utilizadas como variable Proxy de la dificultad percibida.

Para la ejecución de esta propuesta, se utilizó el flujo de ingeniería de datos desarrollado por Brevis (2022) y Latapiatt (2022). Este flujo permitió agilizar las etapas de extracción, limpieza y transformación de los datos contenidos en las bases de datos antes referidas.

Finalmente, con la tabla de características obtenida se desarrollaron nueve modelos predictivos de clasificación supervisada priorizando el rendimiento del indicador de sensibilidad como criterio preponderante para la comparación de los resultados de los modelos entrenados.

1.4 PREGUNTA DE INVESTIGACIÓN

A partir del problema mencionado, se detallará la pregunta central de investigación.

¿Cómo puede estimarse la dificultad percibida por estudiantes chilenos de cuarto a octavo básico durante una tarea de búsqueda de información a través de motores de búsqueda web?

1.5 OBJETIVOS Y ALCANCES DEL PROYECTO

1.5.1 Objetivo general

Construir un modelo de clasificación de la dificultad percibida por escolares de 4° a 8° básico durante tareas de búsqueda de información, mediante el análisis de sus dinámicas de tipeo (keystroke dynamics) al formular consultas, con el fin de contar con indicadores que permitan brindar asistencia oportuna en el proceso de enseñanza-aprendizaje de competencias de investigación en línea.

1.5.2 Objetivos específicos

1. Aplicar los pipelines de procesamiento desarrollados por Brevis (2022) y Latapiatt (2022) para llevar a cabo el preprocesamiento de datos vinculados a dinámicas de tipeo de estudiantes considerando procesos de limpieza, transformación y normalización.
2. Caracterizar las dinámicas de tipeo de estudiantes durante la formulación de consultas en tareas de búsqueda de información factual y evaluación crítica, identificando patrones asociados a distintos niveles de dificultad percibida.
3. Construir modelos de clasificación de dificultad percibida a partir de las dinámicas de tipeo usando técnicas de aprendizaje automático del estado del arte.
4. Comparar los modelos a partir de métricas tales como AUC, F1-Score y sensibilidad.

1.5.3 Alcances

Sobre los alcances de este estudio, se deben considerar los siguientes:

1. **Privacidad de los datos**: La muestra no contiene información sobre los nombres de los niños que fueron sometidos al experimento; por lo tanto, se mantendrá el anonimato de los estudiantes.
2. **Ritmo y Patrones sin registro de teclas**: Los análisis solamente se centrarán en las dinámicas de tipeo, como, por ejemplo, el ritmo de escritura o los patrones de escritura. No se registrarán las teclas que se presionaron en particular; es decir, no se almacenará el texto que el usuario escribió.
3. **Bases de datos**: Se utilizarán únicamente las bases de datos de los estudios de diagnostico e intervención desarrollados por Chourio-Acevedo y González-Ibañez (2024).

1.6 ORGANIZACIÓN DEL DOCUMENTO

Luego del presente capitulo, se encuentra el capitulo dos que aborda el marco conceptual de este trabajo. En este se describen los conceptos y estudios más relevantes que fueron utilizados como base teórica para el presente trabajo. El tercer capítulo detalla el estado del arte al momento de abordar el presente trabajo. Se describen los trabajos más relevantes dentro del área de dinámicas de tipeo y sobre las habilidades de escritura en teclado. El cuarto capitulo detalla la metodología utilizada en el presente trabajo. Se muestran los procedimientos técnicos para consolidar la tabla de características a utilizar por los modelos predictivos. El quinto capitulo muestra los resultados obtenidos durante el análisis exploratorio de los datos, en donde se realizan distintas comparaciones en torno a las distribuciones de los datos. El sexto capitulo muestra un análisis de caracterizaciones de los estudiantes, en donde se realizan técnicas estadísticas y se generan técnicas de análisis no supervisado. El séptimo capitulo detalla los resultados obtenidos después de realizar la etapa de entrenamiento de múltiples modelos, en la que se realizan comparaciones utilizando métricas tales como ROC y sensibilidad. Finalmente, el octavo capitulo resume todo lo realizado en el presente trabajo y también propone posibles trabajos futuros.

CAPÍTULO 2. MARCO TEÓRICO

En el presente capítulo se abordará el marco teórico pertinente a este trabajo. Este se dividirá en dos secciones: Marco conceptual, en donde se revisaron definiciones y conceptos como, por ejemplo, la alfabetización informacional, la carga cognitiva, entre otros; y el estado del arte, en donde se revisaron los trabajos más recientes y/o relevantes sobre las dinámicas de tipeo y comportamientos de búsqueda web.

2.1 MARCO CONCEPTUAL

2.1.1 Alfabetización Informacional

El concepto de alfabetización informacional ha pasado por numerosas re-definiciones a lo largo de los años. Según Johnston y Webber (2003), el concepto surge a raíz de las influencias de las ciencias de la información y el área bibliográfica/bibliotecaria. Un primer acercamiento a una definición formal surgió de Paul Zurkowski, presidente de la Asociación de la Industria de la Información, en 1974, donde mencionó lo siguiente:

> Las personas entrenadas en la aplicación de recursos de información en sus trabajos pueden ser llamados «alfabetizados informacionalmente». Ellos han aprendido las técnicas y habilidades para utilizar un gran abanico de herramientas de información así como fuentes primarias con el fin de dar forma a soluciones de información para sus problemas. (Zurkowski, 1974)

Si bien esta definición se gesta diferenciando la perspectiva tecnológica del concepto de información como tal, no se aprecian explícitamente las diferencias entre ambas definiciones. Es por esto que Horton Jr (1983) separa el concepto de alfabetización informacional del de alfabetización computacional. Específicamente, Horton declara que la alfabetización computacional tiene que ver con incrementar nuestro conocimiento sobre lo que las computadoras pueden y no pueden hacer, mientras que la definición de Paul

Zurkowski aborda el concepto desde una perspectiva de las competencias en la gestión de la información.

Según Carbo (1997), numerosas definiciones y estrategias para enseñar estas competencias en ambientes educativos surgieron en los años ochenta. Debido a esto, la exagencia del gobierno de los Estados Unidos llamada "Comisión Nacional de Bibliotecas y Ciencias de la Información" (NCLIS) convocó, a principios de los años 90, un foro en el que se reunieron 46 organizaciones a lo largo del país. Como resultado, se definió el concepto de alfabetización informacional como: "la habilidad para acceder, evaluar y usar la información desde una variedad de fuentes" (Carbo, 1997).

2.1.2 Carga cognitiva

La carga cognitiva se utiliza como medida para evaluar, modelar y predecir el rendimiento humano en distintas disciplinas, como, por ejemplo, la psicología y la psicología educacional (Gwizdka, 2008). Dada la complejidad de esta materia, existe más de una definición, pero la que mayor consenso ha tenido es la presentada por Moray (1979), que estipula que la carga cognitiva se entiende como la relación entre la demanda de recursos mentales impuestos por una tarea y la habilidad de una persona para suplir estos recursos, entendiéndose que los recursos mentales son limitados.

El primer modelo que se utilizó para representar los estados mentales demandados por una tarea fue desarrollado por Atkinson y Shiffrin (1968), en el que se definió el proceso en tres etapas:

1. **Memoria sensorial**: Es el punto de entrada de la información. Esta memoria tiene la virtud de procesar grandes cantidades de información, tanto visual como auditiva, pero presenta una baja retención.

2. **Memoria de trabajo**: La información se transfiere desde la memoria sensorial y se eleva a la conciencia. Esta memoria posee una capacidad de procesamiento muy limitada, de no más de siete elementos de información a la vez; además, la

información puede ser retenida solamente por un máximo de aproximadamente 30 segundos. La Figura 2.1 muestra un esquema en donde se puede apreciar la limitación generada por la memoria de trabajo. Según Sweller (1988), dadas las limitadas características de esta memoria, provoca un cuello de botella al momento de generar el aprendizaje en una persona.

3. **Memoria a largo plazo**: La información de la memoria de trabajo es transferida como esquemas cognitivos; Es decir, constructos representacionales del objeto de información a almacenar. A diferencia de la memoria de trabajo, esta supone una cantidad ilimitada de almacenamiento de información.

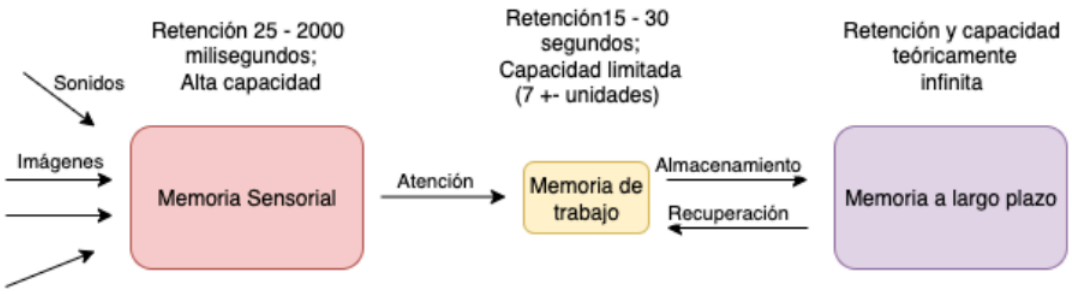

Figura 2.1: Estructura de la memoria humana según la teoría de la carga cognitiva.

Fuente: Elaboración propia

Posteriormente, Sweller utilizó como base teórica el trabajo de Atkinson y Shiffrin (1968) para formular la teoría de la carga cognitiva. En su estudio, reconoció tres tipos, los cuales definió de la siguiente manera:

1. **Carga intrínseca**: Es inherente a los aspectos esenciales de la tarea, así como a la pericia de la persona para abordarla.

2. **Carga extrinseca**: Carga asociada a los elementos no esenciales de la tarea, como por ejemplo: la interfaz web o distracciones externas de la persona al momento de resolver un problema. Además, tanto la carga intrínseca como la extrinseca son aditivas.

3. **Carga relacional**: Carga asociada a los métodos de aprendizaje de la persona. Es importante mencionar que si la carga intrínseca y extrinseca son muy grandes, no habrá espacio para que la carga relacional se desarrolle; por lo tanto, supone una limitante para el proceso de aprendizaje de la persona.

2.1.3 Variables proxy

Una variable proxy es aquella que se utiliza para medir una cantidad de interés no observable (American Psychological Association, 2018). Normalmente, estos escenarios ocurren cuando la variable de interés es costosa o difícil de medir; es por esto que un camino podría ser omitir dicha variable, pero esto podría inducir sesgos en el análisis o en la interpretación de los resultados. Es aquí donde se recurre a una variable sustituta que se asume relacionada con el fenómeno que se requiere estudiar. Por ejemplo, una variable compleja de medir es la calidad de vida, que puede ser aproximada por el ingreso per-cápita o la habilidad cognitiva con los años de escolaridad.

En el contexto de este estudio, la variable de interés a medir es la dificultad percibida ante la tarea de búsqueda web. Si bien este fenómeno no es directamente observable, se utilizará la variable proxy generada a través de las dinámicas de tipeo, en donde métricas como, por ejemplo, la velocidad de escritura, el tiempo de permanencia en una tecla o los tiempos de vuelo podrían describir el fenómeno de interés.

2.1.4 Aprendizaje Automático

El aprendizaje automático (Machine Learning) constituye una subdisciplina de la inteligencia artificial que integra la teoría estadística, las ciencias de la computación y la inteligencia artificial (Alnuaimi & Albaldawi, 2024). El concepto central es el autoaprendizaje, el cual es un proceso en el que los sistemas descubren patrones en los datos para poder predecir resultados.

Según Alnuaimi y Albaldawi (2024), la literatura clasifica los algoritmos en cuatro categorías, dependiendo de los datos utilizados

1. **Aprendizaje supervisado**: Utiliza datos etiquetados para entrenar al sistema y predecir los datos de entrada. Imita el aprendizaje bajo la supervisión de un tutor y se divide en dos categorías dependiendo de la variable de salida: Clasificación, para resultados categóricos, y regresión, para resultados continuos.

2. **Aprendizaje no supervisado**: No utiliza etiquetas para predecir resultados, sino que detecta patrones intrínsecos en los datos para descubrir su estructura interna. De esta manera, se pueden categorizar utilizando los grupos formados por esta metodología.

3. **Aprendizaje semi-supervisado**: Técnica que combina datos etiquetados con datos no etiquetados cuando los datos son costosos o limitados.

4. **Aprendizaje por refuerzo**: Se basa en la experimentación de prueba y error, donde un agente aprende de su entorno mediante un sistema de recompensa y castigo.

Existen distintos tipos de modelos en el aprendizaje supervisado, dependiendo de la problemática a enfrentar:

2.1.4.1 Algoritmos bayesianos

La incertidumbre puede ser cuantificada y administrada utilizando técnicas estadísticas, y los algoritmos bayesianos son un ejemplo en el que se utilizan técnicas basadas en la teoría de la probabilidad.

2.1.4.2 Árboles de decisión

Los árboles de decisión son una metodología utilizada en estadística y en minería de datos (Rokach & Maimon, 2005). La fácil interpretabilidad del modelo

generado hace que esta técnica sea una de las más populares.

Un árbol de decisión es un árbol con raíz y se compone de distintos tipos de nodos: Raíz, que es el nodo inicial del árbol y no tiene aristas entrantes; Nodos internos que se encargan de dividir el espacio de instancias en dos o mas subespacios y las hojas (o nodos terminales), que son aquellos que no tienen aristas salientes. A cada arista se le asigna una clase que representa el valor objetivo mas apropiado.

También existen diversos tipos de criterios para subdividir los nodos. Los mas comunes son los criterios de división univariados, que son aquellos que dividen el espacio de instancias en dos. Algunos de estos criterios son:

- **Information Gain**: Es un criterio basado en impureza (entropía), el cual tiene su origen en la teoría de la información, donde se mide el desorden o la impureza en una base de datos. En otras palabras, si un subconjunto del conjunto general de datos es homogéneo, es decir, todas las observaciones pertenecen a una sola clase, entonces no existe impureza o desorden en el conjunto (Tangirala, 2020).

$$\begin{aligned} \text{IG}(a_i, S) = \text{Entropy}(y, S) \\ - \sum_{v_{i,j} \in \text{dom}(a_i)} \frac{|a_i = v_{i,j}, S|}{|S|} \cdot \text{Entropy}(y, a_i = v_{i,j}, S) \end{aligned} \tag{2.1}$$

donde:

$$Entropy(y, S) = \sum_{c_j \in dom(y)} -\frac{|y = c_j, S|}{|S|} \cdot \log_2 \frac{|y = c_j, S|}{|S|} \tag{2.2}$$

Fuente: Rokach y Maimon (2005)

Cuando esta metodología se implementa como función evaluadora en un nodo de decisión, se realiza una evaluación de la diferencia entre la entropía de la clase y la entropía condicional de la clase y de la característica puntual. En este caso, si la ganancia de información aumenta, significa que la característica f es mas útil para la clasificación (Tangirala, 2020).

$$IG(L, f) = Entropy(L) - \sum_{v=1}^{V} \frac{|L^v|}{|L|} Entropy(L^v) \tag{2.3}$$

Fuente: Tangirala (2020)

- **Gain Ratio**: Normaliza la ganancia de información de la siguiente manera:

$$GainRatio(a_i, S) = \frac{InformationGain(a_i, S)}{Entropy(A_i, S)}$$

Según Quinlan (1988), este método tiene mejores resultados que el cálculo de entropía normal, tanto en calidad como en complejidad de clasificación.

2.1.4.3 Regresión logística

Esta metodología está basada en métodos estadísticos, los cuales pueden ser considerados como una extensión de los modelos lineales. sin embargo, en vez de producir una salida continua, produce una salida dicotómica que representa la ocurrencia (o no ocurrencia) de un evento. En síntesis, este método estima la probabilidad de pertenencia de una instancia a una clase conocida utilizando una función logística llamada sigmoide, la cual se define de la siguiente manera:

$$g(z) = \frac{1}{1 + e^{-z}} \tag{2.4}$$

Los casos de uso más comunes de este método son aquellos en los que se requiere encontrar la relación de mejor ajuste entre la variable dependiente (o variable objetivo) y las variables independientes (o predictores).

2.1.4.4 Support Vector Machine

Este método explica la relación entre las variables independientes y la variable dependiente al encontrar un hiperplano (margen) que maximiza la distancia entre las dos clases de datos. La figura 2.2 ilustra con una linea continua el hiperplano que separa las clases y con una línea discontinua el margen el cual se denomina vector de soporte. Este vector está relacionado con la distancia máxima entre dos hiperplanos paralelos que pueden separar los dos conjuntos de puntos.

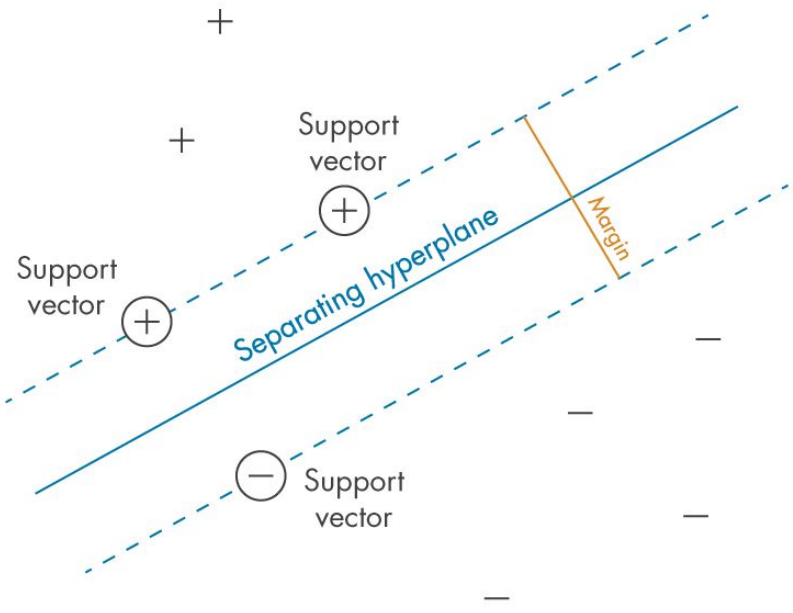

Figura 2.2: Representación gráfica de Support Vector Machine

Fuente: MathWorks (2025)

Existen casos en que la estructura intrínseca de los datos no es linealmente separable. En ese caso, se hace uso de funciones matemáticas complejas llamadas *kernels* para lograr esta separabilidad de clases.

2.1.4.5 Modelo de Mezcla Gaussiana (GMM)

Un Modelo de Mezclas Gaussianas (GMM) es una categoría de modelo probabilístico que establece que todas las muestras de datos generadas se derivan de una mezcla de un número finito de densidades gaussianas (Wan et al., 2019). A diferencia de K-Means, que asume que los clusters son esféricos, GMM puede adaptarse mejor a formas más complejas, como elipses, gracias a su enfoque basado en la probabilidad. Cada una de estas densidades se denomina componente y está definida por dos parámetros principales:

- La media (μ): El centro de la densidad gaussiana.
- La matriz de covarianza (Σ): Que define la forma y dispersión de la densidad.

El modelo en su totalidad se denomina como una suma ponderada de estas componentes, donde cada una tiene un peso de mezcla (w_k) que indica su importancia relativa en la población total.

Dado que los parámetros (pesos, medias y covarianzas) suelen ser desconocidos, se utiliza el algoritmo de Expectación-Maximización (EM) para estimarlos mediante un proceso iterativo de dos pasos:

1. Paso de Expectación (E-step): El algoritmo calcula la probabilidad (o verosimilitud) de que cada punto de datos pertenezca a cada una de las componentes, basándose en los parámetros actuales de las distribuciones gaussianas.
2. Paso de Maximización (M-step): Se actualizan los parámetros de cada componente (media, covarianza y pesos) utilizando las probabilidades calculadas en el paso anterior para maximizar la verosimilitud del modelo. Básicamente, el algoritmo intenta encontrar la mejor configuración de las gaussianas para que representen los datos de manera más precisa.

Estas iteraciones continúan hasta que se alcanza una condición de parada o un número máximo de repeticiones.

La Figura 2.3 ilustra las capacidades de agrupamiento del Modelo de Mezcla Gaussiana bajo distintos tipos de matrices de covarianza. Esta flexibilidad permite modelar distribuciones de datos complejas presentes en las dinámicas de tipeo, superando las limitaciones de esfericidad de algoritmos como K-Means.

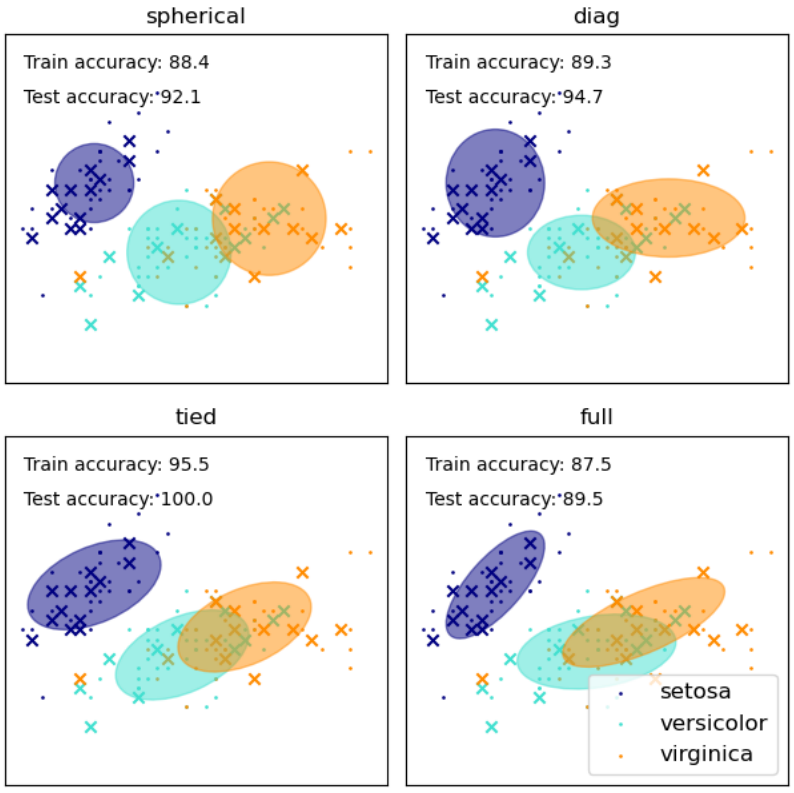

Figura 2.3: Agrupamiento utilizando GMM en distribuciones de distintas covarianzas

Fuente: www.scikit-learn.com

Finalmente, cuando se emplea como clasificador, el GMM utiliza la teoría de decisión de Bayes. El modelo estima la densidad de probabilidad de cada clase utilizando varias distribuciones gaussianas y luego asigna una nueva muestra a la clase que produce la probabilidad a posteriori máxima.

2.1.5 Ecosistema NEURONE

NEURONE es una plataforma de software web diseñada para apoyar investigaciones y evaluaciones de competencias de investigación en línea (Online Inquiry Competences, OIC) (Gonzalez-Ibañez, 2016). Esta plataforma fue concebida para medir y evaluar habilidades de búsqueda de información en internet, como por ejemplo: evaluación critica de la información, búsqueda y selección de fuentes, entre otras funcionalidades.

Con el paso del tiempo, NEURONE evolucionó desde un entorno principalmente investigativo hacia un ecosistema más amplio. En esa línea, Trivia fue implementado dentro del ecosistema NEURONE, que es un juego del tipo preguntas-respuestas cuyo propósito es poner a prueba las competencias de investigación en línea de los estudiantes, en relación con la búsqueda de información, la evaluación crítica de fuentes de información y el uso de la información (Chourio-Acevedo et al., 2024).

Y es así como esta evolución derivó en NEURONE Trainer, que es una plataforma que incorpora trivia como parte de sus experiencias, orientada al entrenamiento de competencias de investigación en linea. En esta plataforma se pueden generar experiencias que incluyen videos instructivos y actividades guiadas diseñadas para mejorar habilidades de búsqueda y selección crítica de información.

Este ecosistema fue utilizado durante dos etapas, las cuales serán detalladas más adelante, para la obtención de los datos utilizados en este estudio.

2.2 ESTADO DEL ARTE

2.2.1 Keystroke Dynamics

El campo de detección de patrones a través de pulsaciones por teclado, también conocido como Keystroke Dynamics (KD), es un área de estudio que se ha popularizado en las últimas dos décadas a raíz de fortalecer las barreras de protección en materia de

ciberseguridad (Moskovitch et al., 2009). Esta técnica consiste en recuperar la información del tiempo en que un usuario presiona las teclas de un teclado.

Esta área es aplicable a dos tipos de escritura en teclado: texto libre y texto fijo. Según Crawford (2010), el texto fijo consiste en que el usuario debe escribir el texto predefinido, como, por ejemplo, el nombre de usuario y la contraseña. Por otro lado, el texto libre permite al usuario ingresar el texto que desee.

A partir de estas acciones en el teclado se desprenden métricas temporales de acciones, dentro de las cuales se encuentran:

- Intervalo (Interval): Tiempo transcurrido entre la pulsación de la tecla A y la tecla B.
- Latencia (Latency): Tiempo transcurrido entre la pulsación de teclas consecutivas.
- Tiempo de vuelo (Flight time): Tiempo entre la pulsación de la tecla A y el inicio de la siguiente pulsación de la tecla B.
- De liberación a liberación[1] (Up to Up): Tiempo entre el momento en que se suelta la tecla A y el momento en que se suelta la tecla B.
- Tiempo de permanencia (Dwell time): tiempo transcurrido en la presión y liberación de una tecla.

Complementariamente, la figura 2.4 ilustra estas métricas aplicadas a dígrafos.

[1] Traducción libre

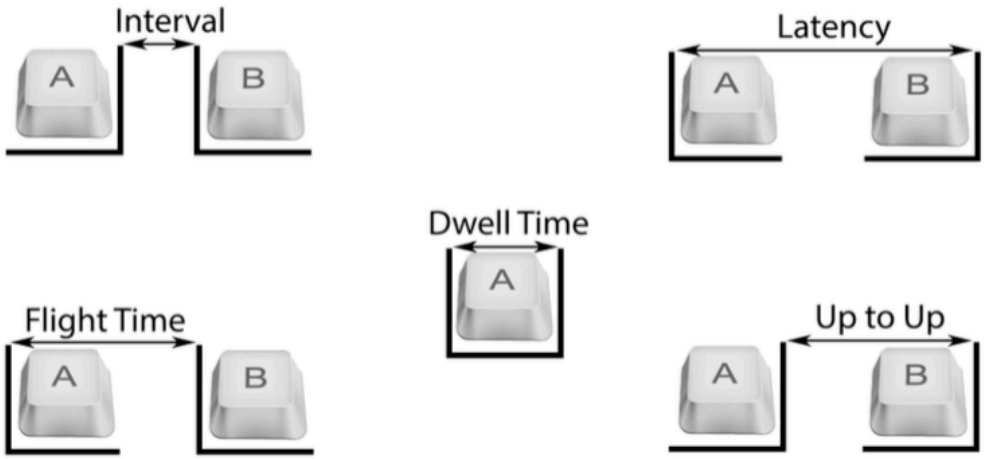

Figura 2.4: Métricas para Keystroke Dynamics en dígrafos

Fuente: Moskovitch et al. (2009)

Para analizar esta información, las teclas se agrupan en dígrafos (o n-grafos). Un dígrafo es una secuencia de dos teclas consecutivas (Iapa & Cretu, 2021) en la que se pueden integrar las métricas descritas anteriormente

Finalmente, una de las ventajas mas importantes de KD es su naturaleza no intrusiva, ya que no requiere de sensores externos y solamente de un teclado de computadora tradicional (Moskovitch et al., 2009)

2.2.2 Comportamiento y habilidades de búsqueda de información

Bilal (2000) desarrolló uno de los estudios seminales en el área del comportamiento de búsqueda de información enfocado en niños. En particular, analizó el comportamiento cognitivo, físico y afectivo de los niños al utilizar un motor de búsqueda web llamado Yahooligans!. Una de las particularidades de este motor de búsqueda es que sus resultados estaban orientados netamente al contenido infantil. Durante el estudio, se examinó el comportamiento de búsqueda de 22 alumnos de séptimo grado y se capturó información cualitativa de los registros de búsqueda, así como cuantitativa proveniente de los logs del sistema y de una cámara web. Entre sus resultados, se destaca que se observaron ciclos

repetitivos, retrocesos y patrones de navegación ineficientes. Adicionalmente, se observó que los conocimientos previos de los niños y la interfaz diseñada para este grupo de edad no implicaron una mejor navegación.

Bilal y Gwizdka (2018) durante su estudio sobre el comportamiento de búsqueda de información en niños, observó 72 sesiones de 24 niños entre las edades de 11 y 13 años utilizando tecnologías de rastreo ocular y rastreo de interacciones en el sistema de recuperación de información. Entre los resultados obtenidos, y al igual que los hallazgos de Druin et al. (2010), Duarte Torres et al. (2010), Kammerer y Bohnacker (2012) y Vanderschantz y Hinze (2017), se encontró que los estudiantes de sexto grado construyeron el mayor número de consultas de tipo pregunta, mientras que los estudiantes de octavo grado formularon mayoritariamente consultas de tipo palabra clave (keywords). También se evidenció que los niños raramente repetían palabras exactas (el autor propone que esto puede ser provocado por el bajo manejo del vocabulario), así como una baja relación entre la longitud de la consulta y el éxito de esta.

Vanderschantz y Hinze (2017) durante su estudio sobre entender el proceso de búsqueda de los niños en un ambiente educacional, utilizando Google como herramienta, también evidenciaron patrones similares en los que la longitud de la consulta era mayor cuando se utilizaba lenguaje natural. Además, un aspecto destacable que se evidenció en este estudio es que, si bien existe una correlación positiva entre las dificultades al momento de formular una consulta y la cantidad de palabras clave, los niños, a pesar de entender que el uso de palabras clave generaría mejores resultados, tendían a abandonar esta estrategia a favor del uso de lenguaje natural. Lo cual se alinea con lo expuesto por Bilal (2000) en torno a que los niños, a pesar de tener conocimiento previo en materias de búsqueda o, en este caso, sobre cómo utilizar una herramienta de recuperación de información, no mejoran sus habilidades de búsqueda.

2.2.3 Dificultad percibida en una tarea de búsqueda

Huang et al. (2020) durante su estudio, buscó explorar cómo influye la dificultad de una tarea de búsqueda en el comportamiento de un usuario, específicamente contrastando la dificultad objetiva y subjetiva. La dificultad percibida fue definida como "la percepción individual del usuario sobre la dificultad de la tarea". Esta fue medida utilizando una escala Likert de 5 puntos, considerando como difíciles las tareas con puntuaciones superiores a 3. Por otro lado, la dificultad objetiva fue definida mediante la proporción de la muestra que puede completar los requisitos de una tarea. Esta se midió considerando difíciles las tareas con puntuaciones superiores a la puntuación promedio de los 30 participantes. Los resultados observados condujeron a la conclusión de que es mejor desarrollar modelos de predicción orientados a la dificultad subjetiva, porque estos modelos son más estables.

Gacitúa y González (2022) en su estudio, proponen un modelo de clasificación que identifica cuando una persona requiere asistencia durante una sesión de búsqueda de información. A lo largo del estudio, se desarrollaron diversas tareas, tales como el reclutamiento de los usuarios a estudiar, la adaptación de la plataforma en la cual los usuarios llevaron a cabo el experimento, y la definición de las variables a observar, entre otras. Es importante destacar que se optó por utilizar la dificultad previa a la tarea, capturada a través de una encuesta, ya que esta permite una apreciación inicial y temprana de la dificultad percibida. Además, debido a la importancia de predecir en los momentos iniciales de una búsqueda de información en la web, las consultas se clasificaron en tres tipos: Primer segmento (1qs), que corresponde a la primera consulta de la sesión; Segmentos acumulados, que corresponden al segmento 1qs+n de la búsqueda, y sesión completa, que corresponde a todas las consultas realizadas en la sesión (revisar Figura 2.5). Sobre los resultados del análisis de datos, se observó que las consultas de una tarea difícil tienen significativamente más caracteres, reformulaciones, tiempo de escritura y pausas que una tarea definida como fácil. Finalmente, sobre los resultados de los modelos de clasificación, se observó que XGBoost (AUC = 0.780) y SVM-RadialKernel (AUC = 0.790) obtuvieron los mejores resultados en las consultas del primer segmento.

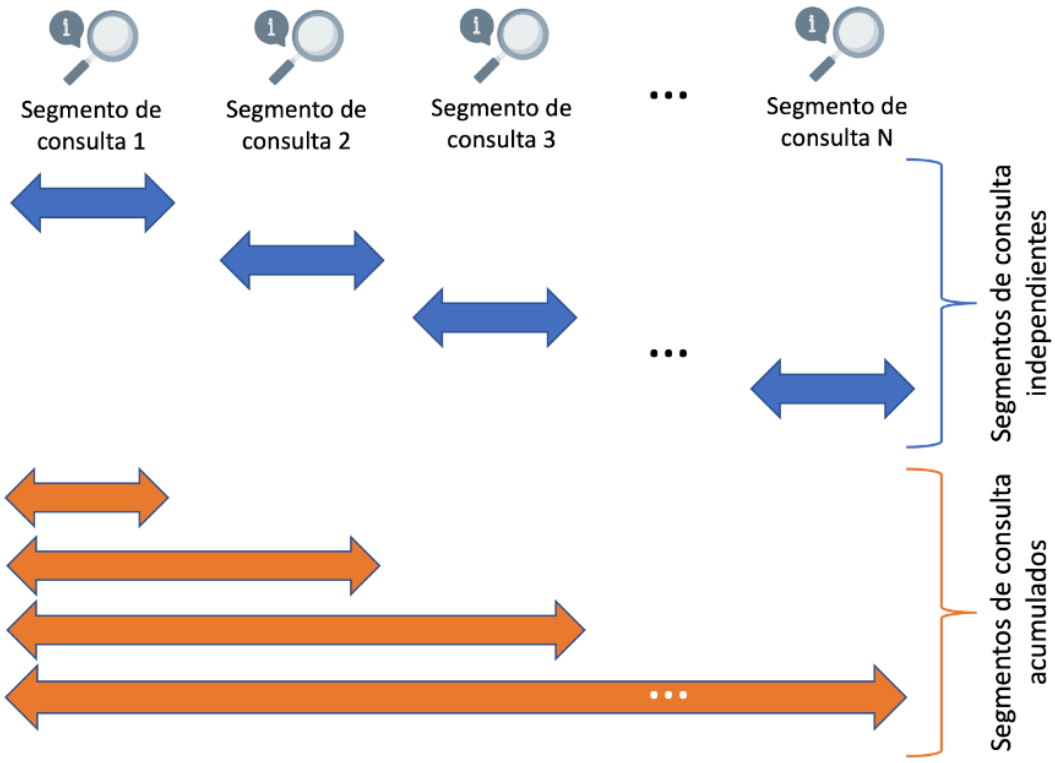

Figura 2.5: Representación visual de los acercamientos para la agregación de los segmentos de consulta.

Fuente: Gacitúa y González (2022)

2.2.4 Keystroke Dynamics y carga cognitiva

La relación entre la dinámica de tipeo y la carga cognitiva ha sido objeto de estudio en áreas como la medicina y la interacción humano-computador. Los estudios demuestran que la carga cognitiva fomenta cambios en la dinámica de tipeo, tales como una mayor cantidad de pausas durante la escritura y el tiempo empleado. Entre las características más relevantes, se puede observar que las pausas entre palabras, el número de revisiones y el tiempo total de escritura son los más sensibles ante un cambio en la complejidad de la tarea, mientras que los intervalos entre pulsaciones tienden a ser menos sensibles.

Révész et al. (2017) en su estudio, investigó cómo la complejidad de la

tarea afecta el proceso de escritura en una segunda lengua (L2), particularmente en sus comportamientos de escritura y en la complejidad lingüística de su producción escrita, utilizando herramientas de grabación de pulsaciones de teclas. Setenta y tres estudiantes avanzados en L2 fueron considerados para el estudio, los cuales fueron asignados a uno de dos grupos; el primero, de tarea simple, en el que a los estudiantes se les proporcionaron ideas y subtemas específicos para incluir en el ensayo, reduciendo así la carga cognitiva en los procesos de planificación; mientras que en el segundo grupo (de tarea compleja), los estudiantes debían escribir un ensayo argumentativo basado en una consigna estándar de un examen de Trinity College London (ISE III) sin ayuda adicional para generar ideas. El estudio evidenció que el grupo de tarea simple invirtió menos esfuerzo mental al desarrollar el ensayo, así como también en la cantidad de palabras. Por el contrario, en el grupo de tarea compleja se evidenció un mayor uso de palabras.

Conijn et al. (2019) en su estudio investigó la variabilidad de las características del registro de pulsaciones de teclas a partir de las diferentes demandas cognitivas de la tarea. También definió dos categorías para agrupar las características de los registros de teclas: En primer lugar, el grupo relacionado con la duración de la pulsación, donde se encuentran las características de duración de la pulsación, latencia de la tecla, tiempo entre palabras, entre otras; y en segundo lugar, el grupo relacionado con el contenido, donde se encuentran las características de cantidad de palabras eliminadas y cantidad de palabras escritas. La metodología del estudio consistió en aplicar un enfoque experimental a un grupo de usuarios, quienes tuvieron que transcribir un correo electrónico en comparación con otro grupo que tuvo que redactarlo. Los resultados evidenciaron que el tiempo total, las palabras tecleadas, la cantidad de revisiones y el tiempo total entre palabras resultaron ser las características más sensibles a variaciones en la carga cognitiva.

Urry y Edwards (2024) en su estudio buscó comprender la carga cognitiva en cursos introductorios de programación en Python en la Universidad de Utah. Para ello, el estudio abordó la carga cognitiva desde la perspectiva de la duración de las pausas durante las pulsaciones de teclas en actividades de búsqueda de información, asociando una mayor duración de estas a una alta carga cognitiva. Cuarenta y tres estudiantes fueron considerados para el estudio, en el que se registraron alrededor de 900,000 registros de

teclas. Los resultados del estudio evidencian que la duración de las pausas correlaciona positivamente con la carga cognitiva de los estudiantes; sin embargo, es importante destacar que la duración total de la tarea de búsqueda no necesariamente indica que un estudiante esté empleando una alta carga cognitiva, ya que este podría estar utilizando una mayor cantidad de pausas, pero de menor duración.

2.2.5 Habilidades de escritura en teclado en estudiantes de enseñanza básica

Sobre las habilidades de escritura en teclado de los estudiantes de enseñanza básica, un estudio realizado por Gahshan y Weintraub (2024) compara el rendimiento (velocidad y precisión) así como el papel de las funciones subyacentes en la predicción del rendimiento con el teclado de 57 alumnos de cuarto grado de dos escuelas distintas. El estudio evidenció que los estudiantes escribían más rápido en tareas de dictado que en tareas de copia y también que los estudiantes adaptaban la velocidad de escritura dependiendo de los requerimientos de las tareas.

Sobre la poca capacitación de los niños en torno a las habilidades de escritura en teclado en las escuelas de los Estados Unidos, Daniels (2020) realizó un estudio para comprobar si una intervención de 10 semanas de duración tiene (o no) efectos en el rendimiento de la escritura. La muestra estuvo compuesta por 134 estudiantes que cursan cuarto y quinto grado. La intervención demostró mejoras en la calidad, como la velocidad de escritura, después de la intervención.

2.3 RESUMEN

El marco teórico desarrollado en este capítulo aborda las definiciones básicas para llevar a cabo el desarrollo de este trabajo, y en el estado del arte se evidencian los últimos avances significativos en torno a la pregunta de investigación. Principalmente, se define el concepto de alfabetización informacional como la habilidad para acceder,

evaluar y usar la información desde una variedad de fuentes (Carbo, 1997). Sobre la carga cognitiva, esta se explica y se definen los conceptos relevantes orientados al marco de estudio, así como el significado de una variable proxy. Finalmente, se explican los métodos de clasificación, enfatizando en XGBoost y SVM, ya que son algunos métodos utilizados durante el desarrollo de este trabajo.

CAPÍTULO 3. METODOLOGÍA

La investigación se abordó bajo el paradigma de investigación experimental, puesto que durante la revisión de la literatura se evidenció que es un fenómeno poco estudiado. Tal como se mencionó en la introducción, la finalidad del proyecto consiste en desarrollar un modelo de predicción binaria; para ello, se deberán limpiar y procesar los datos generados por Chourio-Acevedo y González-Ibañez (2024), que contienen información sobre la dificultad percibida y las teclas presionadas durante la sesión de búsqueda. Para acelerar el desarrollo de las etapas iniciales de esta investigación, se utilizaron los proyectos de ingeniería de datos de Brevis (2022) y Latapiatt (2022). Estos proyectos constituyen una metodología estandarizada para soportar proyectos derivados de la plataforma NEURONE.

Luego, en la etapa de análisis de datos, se caracterizaron las dinámicas de tipeo de los estudiantes durante la formulación de consultas para identificar patrones asociados a distintos niveles de dificultad percibida, como por ejemplo, generar análisis univariados y bivariados para entender distribuciones y asociaciones entre variables, así como también análisis de componentes principales.

Finalmente, se construirán los modelos de clasificación de la dificultad percibida y sus resultados se compararán con los obtenidos por Gacitúa y González (2022) en su estudio, que consideró un rango etario distinto al presentado. En esta etapa, se considerará como variable objetivo la variable binaria de dificultad percibida (fácil o difícil).

3.1 ORIGEN DE LOS DATOS

Los datos utilizados en este estudio provienen de un mismo sistema empleado en dos estudios diferentes. A continuación, se detalla el origen y los objetivos de ambos estudios.

3.1.1 Diagnóstico

El objetivo de este estudio fue caracterizar el nivel de desarrollo de competencias de búsqueda de información en línea en estudiantes de educación básica. El estudio se llevó a cabo en una única sesión de 45 minutos con 4 desafíos. Estos desafíos consistieron en respuestas de texto y en la justificación de las fuentes confiables utilizadas. El estudio se desarrollo utilizando la plataforma NEURONE - Trivia, que permitió registrar consultas, selección de resultados, tiempos de navegación, registro de tipeos en teclado, percepción de la dificultad ante los distintos desafíos, entre otros. La participación fue de 347 estudiantes de 9 a 15 años (4° a 8° grado) de cuatro establecimientos educacionales de la Región Metropolitana, con 1055 observaciones válidas. La Figura 3.1 muestra las comunas pertenecientes a los establecimientos educacionales que participaron del estudio.

Uno de los hallazgos principales fue que los estudiantes realizaron consultas poco elaboradas, es decir, la mayoría realizó solo una consulta por desafío, sin reformular ni utilizar operadores de búsqueda. También se observó que la cantidad de consultas copiadas incrementaba positivamente en función de la edad de los participantes.

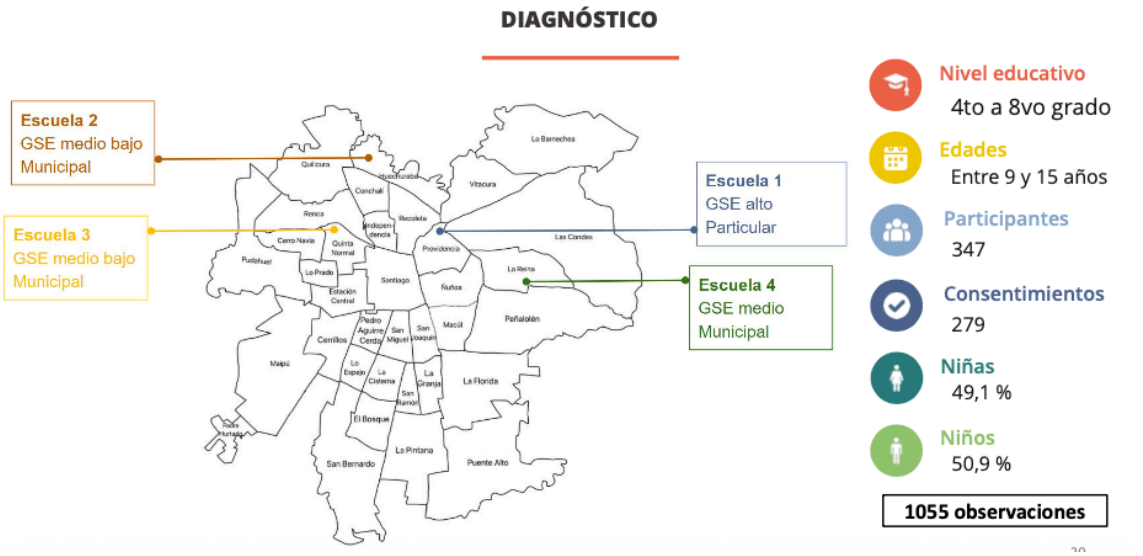

Figura 3.1: Demografía de los estudiantes en el estudio de diagnóstico

Fuente: Chourio-Acevedo y González-Ibañez (2024)

3.1.2 Intervención

El objetivo de este estudio fue desarrollar competencias de búsqueda mediante un modelo tecnológico y procedimental apoyado por un asistente virtual y estrategias de gamificación. El estudio tuvo un diseño experimental que involucró cuatro grupos: Control, Tradicional, Ludificado y Asistente Virtual, en los cuales el comportamiento de los alumnos fue analizado en etapas previas y posteriores al experimento. El nivel educativo de los participantes fue de 6° básico, donde la participación fue de 381 estudiantes. La duración del estudio fue de dos sesiones de 45 minutos cada una. Las Figuras 3.2 y 3.3 ilustran las fases desarrolladas durante el experimento. Similar al estudio de Diagnóstico, el estudio se realizó a través de la plataforma NEURONE-Trainer. Esta plataforma toma la base de NEURONE-Trivia, pero le añade módulos donde se incorpora un asistente virtual, entre otras mejoras. Las actividades realizadas fueron formular consultas ante un desafío, refinar las consultas con operadores, identificar partes de los resultados y validar fuentes.

Algunos hallazgos relevantes de este estudio fueron que los grupos intervenidos (Ludificado y Asistente Virtual) mostraron mejoras significativas en el post-test respecto al grupo de control, y además, el Asistente Virtual presentó un efecto moderado en ambas clases y diferencias significativas en la primera.

Figura 3.2: Fases del estudio de intervención

Figura: Chourio-Acevedo y González-Ibañez (2024)

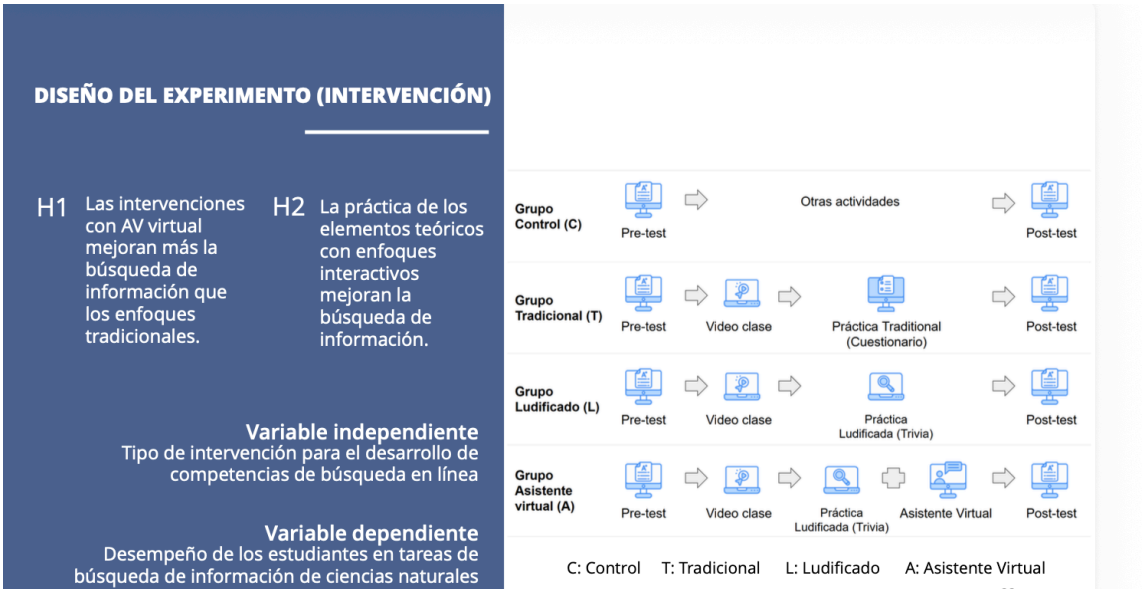

Figura 3.3: Diseño del estudio de intervención

Figura: Chourio-Acevedo y González-Ibañez (2024)

Un aspecto importante a destacar es que ambos estudios capturaron cada presión de tecla realizada durante el transcurso de los desafíos, lo cual será la información que se agregará y procesará para generar la tabla de características.

3.1.3 Descripción de los datos

Como se pudo adelantar en la sección anterior, los datos utilizados consistieron en dos bases de datos generadas en el marco del proyecto TUTELAGE Chourio-Acevedo y González-Ibañez, 2024. Ambos conjuntos de datos, al provenir de un mismo sistema común, comparten las mismas cantidades de tablas y su estructura. A continuación, se detallará una breve descripción de las tablas utilizadas en este estudio:

Study Tabla que define el tipo de estudio que se llevará a cabo. Cada instancia de esta tabla posee múltiples desafíos (challenges) que se efectúan durante la sesión de investigación. La tabla posee un ID único, el nombre del estudio, entre otros.

UserStudy Tabla que contiene cada uno de los estudios que llevaron a cabo los estudiantes. Cada registro es una instancia en la que un alumno tuvo que responder una serie de desafíos. La tabla contiene un ID único, el estudiante y los ID de cada uno de los desafíos iniciados.

Challenges Desafíos definidos al momento de realizar el estudio. Cada desafío posee un ID único, un ID que corresponde al estudio al cual pertenece y una columna question, la cual contiene la pregunta del desafío. Cada registro se interpreta como un desafío a realizar en un estudio dado.

UserChallenges Desafíos realizados por los estudiantes. La tabla contiene un ID único, un ID de un desafío, un ID de un estudiante y una columna Answer en la que se registra la respuesta del estudiante. Cada registro se interpreta como un desafío realizado por un alumno en un estudio dado.

Questionnaires Listado de cuestionarios a realizar durante el pre y post test. La tabla contiene un ID único, el título de la pregunta y las posibles respuestas. Estas preguntas están orientadas a conocer si un estudiante estuvo (o no) motivado al momento de realizar el estudio, o si estos encontraban que los desafíos serían fáciles o difíciles. Cada una de las

posibles respuestas está registrada bajo una escala del 1 al 6, siendo 6 una dificultad alta y 1 una dificultad baja.

UserQuestionnaires Las respuestas de los alumnos durante la realización del cuestionario. La tabla contiene un ID único, un ID de un cuestionario, un ID de estudio, un ID de un estudiante y un ID de un desafió en particular. La tabla, además de la respuesta del usuario, contiene si esta fue realizada durante el pre o post test. Cada registro de esta tabla se interpreta como una respuesta efectuada por un usuario durante el pre o post test.

keystrokes Tabla con los registros de tipeo de teclas. La tabla contiene un ID único, un ID de usuario, un código de tecla presionada, el carácter presionado y si esta acción fue de tipo key-up o key-down. Cada registro de la tabla se interpreta como una acción efectuada sobre una tecla. Se espera visualizar dos registros por tecla, es decir, un key-down y luego un key-up por cada una de ellas.

Queries Tabla con todas las consultas realizadas durante los desafíos. La tabla contiene un ID único, una columna query en donde se registra la búsqueda realizada y una columna URL en donde se registra la página en donde se efectuó la consulta. Cada registro se interpreta como una búsqueda puntual realizada por el estudiante.

Consentimiento usuarios Tabla en la que se registraron todos los estudiantes que contaban con el consentimiento de sus padres y, por lo tanto, accedieron a compartir los datos generados durante el estudio.

Visited links Tabla en la que se registran todos los enlaces (o páginas) visitadas durante el desafío. La tabla contiene un ID único y un estado que puede ser PageEnter (si un estudiante ingresó a una página) o PageExit (si un estudiante salió de una página). También existe una columna llamada *url* , la cual contiene la dirección web en la que se efectuó dicha acción. Cada registro se interpreta como una página visitada o el cierre de una página en un momento dado.

Es importante mencionar que cada una de estas tablas contiene una columna timestamp en la que se registra la hora y la fecha en que tuvo lugar el registro. Finalmente, un mayor detalle de las tablas puede encontrarse en el Anexo B.0.1

3.2 PIPELINE DE TRABAJO

Para el desarrollo de este trabajo se optó por utilizar la metodología *OSEMN* debido a que es la estrategia implementada en el trabajo realizado por Brevis (2022) y Latapiatt (2022) ya que, como se mencionó anteriormente, constituyen los flujos referenciales de este estudio.

La metodología *OSEMN* propuesta por Mason y Wiggins (2010), ilustrada en la Figura 3.4, consta de los siguientes pasos:

1. **O(Obtain)**: Recopilar los datos brutos de las fuentes de interés.
2. **S(scrub)**: Limpiar los datos a un formato que sea entendible para la máquina.
3. **E(explore)**: Análisis exploratorio de los datos (*EDA*).
4. **M(model)**: Construir el/los modelos predictivo(s).
5. **N(interpret)**: Mostrar los resultados en un formato comprensible.

De este modo, esta metodología no solo garantiza coherencia con el trabajo de Brevis (2022) y de Latapiatt (2022), sino que también ofrece un flujo de trabajo claro y adaptable a las necesidades de este estudio.

Figura 3.4: Diagrama de la metodología OSEMN

Fuente: www.datascience-pm.com

La adopción del pipeline de Brevis (2022) comenzó con un análisis para dimensionar la cantidad de cambios a realizar, ya que, si bien uno de los objetivos del trabajo de Brevis fue estandarizar el procesamiento de datos de proyectos derivados de NEURONE, los análisis implementados en este tuvieron foco en los objetivos propuestos por el autor. El flujo de este pipeline puede visualizarse en la Figura 3.5 que describe cada una de las etapas por las cuales pasa los datos hasta ser procesados por los modelos de aprendizaje automático. Una vez realizada la inspección inicial del pipeline, se propuso la siguiente estructura de ejecución:

1. Carga del dataset crudo (Obtain)

    a) **Importar datos crudos**: En este paso se realizó la carga inicial de archivos.

    b) **Limpieza de datos**: En este paso se realizó una limpieza preliminar de los datos, asegurando que cada observación corresponda a una en donde hubo consentimiento en el uso de los datos por parte del usuario.

2. Limpieza de datos (Scrub)

    a) **Convertir variable de tiempo**: Se capturaron los instantes de tiempo y se transforman a Epoch para facilitar los cómputos.

    b) **Filtrado de datos**: Se almacenaron unicamente los datos de interés al estudio. Además se realizaron cruces de datos con tablas de interés para el posterior

cálculo de las variables de dinámicas de tipeo.

c) **Construcción de tabla de características**: En esta etapa se realizó el procesamiento de los datos para construir la tabla de características de dinámicas de tipeo.

d) **Unir datos adicionales**: Una vez construida la tabla de características, se unió información adicional, como por ejemplo: edad, tipo de pregunta, entre otros.

3. Exploración de datos (Explore)

a) **Análisis exploratorio**: En esta etapa se realizaron diversos análisis, incluyendo exploración de distribuciones, escalado y análisis de componentes principales, entre otros.

b) **Caracterizaciones**: Se realizaron un análisis de caracterizaciones con el fin de perfilar a los estudiantes en función de la dificultad percibida.

4. Modelamiento (Model)

a) **Preprocesamiento**: En esta etapa se transformaron las variables binarias a valores 0 o 1 y se realiza one-hot encoding a variables categóricas.

b) **Selección de características**: Se realizó la selección automática de características utilizando Boruta y recursive feature elimination (RFE).

c) **Entrenamiento de modelos**: En esta etapa se realizó el entrenamiento utilizando una grilla de modelos para la selección de hiperparámetros. Se utilizó el cluster XI NETRAIDER del Departamento de Ingeniería Informática de la Universidad de Santiago.

d) **Evaluación de modelos**: Se realizaron comparaciones de los modelos entrenados, en donde se evalúan los rendimientos en función de AUC, F1-score y sensibilidad.

e) **Importancia de características**: En esta etapa se realizó un análisis de importancia de características para explicar los resultados obtenidos.

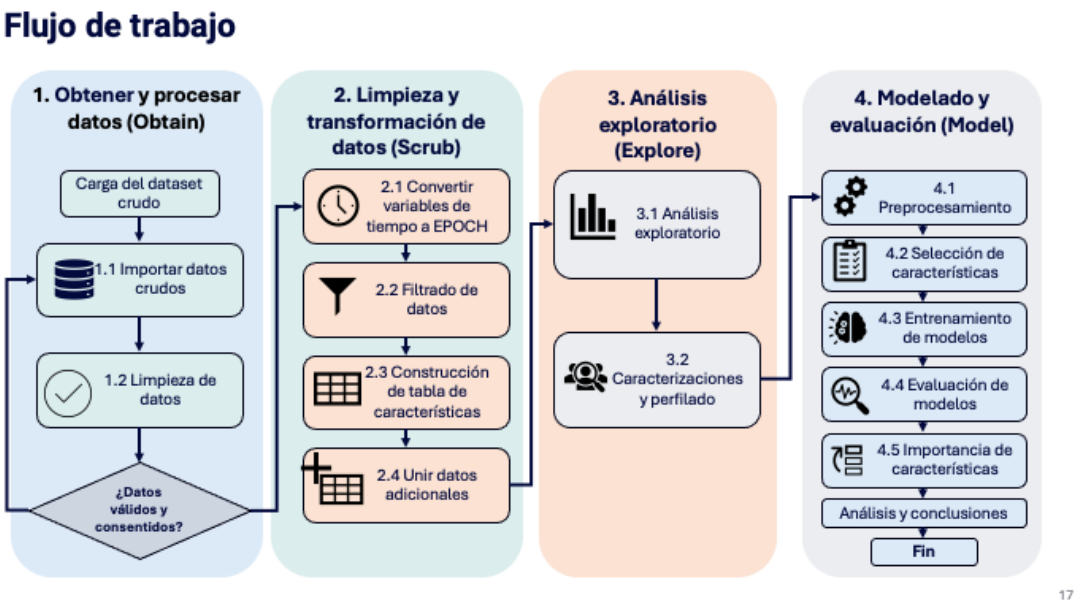

Figura 3.5: Diagrama de flujo del pipeline implementado

Fuente: Elaboración propia

3.3 CONSTRUCCIÓN DE TABLA DE CARACTERÍSTICAS

Para la construcción de la tabla de características, se considero utilizar las variables con mejores resultados durante el estudio de Gacitúa y González (2022) puesto que durante este estudio se realizó una comparación exhaustiva entre el rendimiento de todas las variables de Keystroke Dynamics y los distintos niveles de dificultad percibida. A continuación se detallarán cada uno de los pasos para construir esta tabla, correspondiente al paso 2.c del flujo del pipeline descrito en la sección anterior. Por último se muestra un resumen de la tabla de características con la descripción de cada una de estas. Los pasos, en donde los títulos hacen referencia a los módulos desarrollados en el pipeline, son los siguientes:

Convert to epoch Se cargaron las tablas desde el modulo anterior (Obtain) en formato .csv. Luego, para cada una de las tablas cargadas, se realiza una conversión de timestamp

a epoch en la columna de timestamp local. Epoch es una representación del tiempo como el número de segundos o mili-segundos transcurridos desde un punto de referencia fijo conocido como la «época», que es el 1 de enero de 1970 a las 00:00:00 UTC. Esto facilita las conversiones y cálculos utilizando información temporal. Es importante destacar que se utiliza únicamente la columna de timestamp local, puesto que estos registros corresponden a los tiempos obtenidos en el computador en donde se efectuaron los desafíos y no los tiempos de servidor para que los cálculos posteriores sean más precisos. Posteriormente se exportaron estas en formato .csv.

Set time and corrects Se ajustó la tabla userQuestionnaires, que es aquella que que contiene los valores reportados de dificultad pre/post test por los alumnos. Luego se filtró y procesó la tabla visitedLinks para capturar únicamente las páginas de búsquedas que contengan urls similares a '/session/search' y también los instantes de tiempos correspondientes al acceso a estos sitios. Luego, se crearon ventanas temporales (enter_time y exit_time) utilizando los instantes de tiempo de visitedLinks en donde se pueda apreciar que el usuario ingresó y salió del sitio. Esta ventana de tiempo, ilustradas en la Figura 3.6, fueron utilizada para agrupar cada una de las presiones de teclas y construir la consulta.

/session/search	PageEnter	2023-10-17 18:43:58.551
/session/search	PageQuery	2023-10-17 18:44:24.928
/session/search-result/deporte%20corazon%20salud...	PageEnter	2023-10-17 18:44:24.930
/session/search	PageExit	2023-10-17 18:44:24.937

Figura 3.6: Tabla visitedLinks filtrada

Fuente: Elaboración propia

Aggregated Keystrokes Utilizando la tabla visitedLinks procesada, se procedió a hacer la agregación de la tabla keystrokes. Para ello, se utilizaron las ventanas de tiempo en donde un estudiante entra a la página de búsqueda hasta que se emite una consulta. La Figura 3.7 muestra un ejemplo.

/session/search	PageEnter	2023-10-17 18:43:58.551
/session/search	PageQuery	2023-10-17 18:44:24.928
/session/search-result/deporte%20corazon%20salud...	PageEnter	2023-10-17 18:44:24.930
/session/search	PageExit	2023-10-17 18:44:24.937

Figura 3.7: Ejemplo ventana de tiempo utilizada en la agregación

Fuente: Elaboración propia

Entre esos dos instantes de tiempo, se buscó cada una de las teclas presionadas, lo cual conformaría la consulta realizada. Con este subconjunto de registros de pulsaciones de teclas, se realizaron los cálculos y agrupaciones para generar las variables de dinámicas de tipeo. La tabla resultante tiene las variables descritas en la Tabla 3.1

Get user keystrokes Finalmente, la tabla de características, calculada en la función anterior, fue complementada con valores demográficos de los usuarios, tales como sexo, edad, entre otros. También se calculan las categorías de la variable objetivo; en este caso, se toman los valores obtenidos del auto reporte de dificultad percibida ingresada por los usuarios, los cuales se encuentran en una escala de 1 a 6. Se considera como fácil si el valor está entre 1 y 3 y difícil si este se encuentra entre 4 y 6.

Finalmente, los datos fueron exportados en formato .csv, para ser utilizados por el módulo Explore en donde se realizó el análisis exploratorio de datos. La tabla de características resultante de este flujo es la presentada en la Tabla 3.1.

Tabla 3.1: Tabla de características

Variable	Descripción
effectiveCharacterCount	Número total de caracteres efectivos escritos en la consulta.
charactersPerMinute	Velocidad de tipeo medida en caracteres por minuto.
pauseCount	Cantidad total de pausas realizadas durante la escritura.
reformulations	Número de re-formulaciones o modificaciones de la consulta inicial.
dwellTimeAvg	Tiempo promedio de permanencia en cada tecla presionada (ms).
edad	Edad del participante en años.
queryTime	Tiempo total empleado en la construcción de la consulta (ms).
dwellTimeStd	Desviación estándar del tiempo de permanencia en teclas.
sexo	Sexo declarado del participante (categórica).
writingTime	Tiempo neto de escritura, excluyendo pausas (ms).
latencyAvg	Latencia promedio entre dos teclas consecutivas (ms).
pauseTime	Tiempo total acumulado en pausas (ms).
upToUpStd	Variabilidad en los intervalos de presión a presión (Up-to-Up).
copypaste[1]	Distancia de Levenshtein entre la consulta formulada y la consulta enviada.

[1] Esta variable de comportamiento se incorpora con la finalidad de caracterizar patrones de interacción y explorar relaciones entre las dinámicas de tipeo registradas.

3.4 TÉCNICAS DE MINERÍA DE DATOS

Durante la etapa de modelamiento, se diseñó un protocolo experimental orientado a la robustez y comparabilidad de los resultados.

- **División del conjunto de datos**: Para la división del conjunto de datos se utilizó validación cruzada repetida de 10 pliegues. Este método divide un conjunto inicial de 10 particiones en 9 para el entrenamiento y 1 para las pruebas. Este proceso se repite 10 veces alternando las particiones a lo largo del conjunto de datos.

- **Selección automática de características**: Para la selección automática de características, se utilizaron los siguientes métodos:

- **Eliminación Recursiva de Características (RFE)**: Es un método de selección de características que elimina iterativamente las variables con menor peso predictivo hasta alcanzar un subconjunto óptimo. Esto es especialmente útil para la problemática actual puesto que se evidenció alta colinealidad observadas en las métricas de dinámicas de tipeo. Este método permite reducir la redundancia del conjunto de datos, y la capacidad de generalización de los modelos.
- **Boruta**: A diferencia de RFE, que busca el subconjunto mínimo de características, este es un algoritmo de relevancia total. Este identifica las variables más relevantes de un conjunto de datos al compararlas con variables «sombras» generadas aleatoriamente. Su uso se justifica por la naturaleza del comportamiento humano en el tipeo, en donde algunas interacciones entre variables correlacionadas podrían indicar señales sutiles de dificultad frente a una tarea de búsqueda.

- **Técnicas de remuestreo**: Ante el escenario de que los datos no estén balanceados, se consideró realizar submuestreo (undersampling) al conjunto de entrenamiento.

- **Instantes de medición**: Corresponde a la metodología de calculo y agrupación de las variables de dinámicas de tipeo.

  - **Primer segmento de consulta (1qs)**: Corresponde a todas las acciones realizadas por un usuario desde el momento en que ingresa al buscador y envía la primera consulta.
  - **Sesión completa (all)**: Corresponde a todas las acciones realizadas por un usuario durante la sesión completa de búsqueda.

- **Algoritmos de clasificación**: La estrategia de modelado se estructuró a partir de los antecedentes de Gacitúa y González (2022), seleccionando inicialmente aquellos algoritmos que demostraron un desempeño superior en términos de AUC (ver Tabla 3.2). No obstante, con el objetivo de capturar patrones de mayor complejidad y asegurar la robustez de los resultados, el espectro de evaluación

se amplió mediante la integración de arquitecturas de aprendizaje adicionales. Esta expansión metodológica permitió contrastar los referentes bibliográficos con modelos de ensamble avanzado y técnicas de regularización (detallados en la Tabla 3.3), garantizando una exploración exhaustiva del espacio de soluciones y una optimización más precisa frente a la naturaleza específica de los datos utilizados en este estudio.:

Tabla 3.2: Modelos ordenados según AUC obtenido

Fuente: Gacitúa y González (2022)

Modelo	**AUC**
Random Forest	0.775
SVM Radial Kernel	0.759
Neural Network	0.710
Naive Bayes	0.662
RPart	0.633
C5.0	0.617
SVM Linear Kernel	0.546
Regularized Logistic Reg.	0.491
GLM	0.479

- **Hiperparámetros**: Para los modelos previamente seleccionados, se consideró la siguiente grilla de hiperparámetros utilizando 3 repeticiones.

Tabla 3.3: Configuración de grillas de hiperparámetros por modelo

Fuente: Elaboración propia

Modelo	**Hiperparámetros evaluados**	**Total de combinaciones (1 rep)**
K-Nearest Neighbors (KNN)	$k = \{3, 5, 7, \ldots, 99\}$	30
SVM (Kernel Lineal)	$C = 2^{[-10,10]}$ con pasos de 0,5	41
SVM (Kernel Radial)	$\sigma = 2^{[-12,12]}$ y $C = 2^{[-12,12]}$	$25 \times 25 = 625$
SVM (Kernel Sigmoid)	$C = 2^{[-10,10]}$ (paso 0,5), *degree* = 1, *scale* = $\{1, 1{,}5, 2, 2{,}5, 3\}$	$41 \times 5 = 205$
Random Forest	$mtry = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 18, 21, 25, 30\}$	15
XGBoost	$nrounds = \{100, 250, 500, 750, 1000\}$, *max_depth* = {2, 3, 4, 6}, $\eta = \{0{,}01, 0{,}025, 0{,}05, 0{,}1\}$, $\gamma = \{0, 0{,}5, 1\}$, *colsample_bytree* = {0.3, 0.4, 0.6, 0.8}, *min_child_weight* = {1, 2, 3, 5}, *subsample* = {0.25, 0.5, 0.75, 1}	$5 \times 4 \times 4 \times 3 \times 4 \times 4 \times 4 = 15{,}360$
Elastic Net (glmnet)	$\alpha = \{0, 0{,}1, \ldots, 1\}$, $\lambda = 2^{[-10,2]}$ (paso 0,5)	$11 \times 25 = 275$
Gradient Boosting Machine (GBM)	$n.trees = \{500, 1000, 1500, 2000\}$, *interaction.depth* = {2, 3, 5, 7}, *shrinkage* = {0.001, 0.01, 0.05, 0.1}, *n.minobsinnode* = {5, 10, 20}	$4 \times 4 \times 4 \times 3 = 192$
Decision Tree (RPart)	$cp = 10^{[-4,-0{,}5]}$ (30 valores espaciados logarítmicamente)	30

Nota: Cada combinación de hiperparámetros fue evaluada mediante validación cruzada 10-fold repetida 3 veces (30 ejecuciones por combinación).

- **Métricas de evaluación de modelos**: Para la evaluación de modelos se utilizaron las métricas de Accuracy (exactitud), AUC (área bajo la curva ROC) y F1-Score.

3.5 AMBIENTE EXPERIMENTAL Y HARDWARE REQUERIDO

La ejecución del flujo, a excepción de la etapa de entrenamiento de modelos, tuvo lugar en una computadora personal. Esto se debe a los largos tiempos de cómputo en la etapa de búsqueda de hiperparámetros (fine tuning). La computadora personal utilizada tiene las siguientes características: Macbook Pro de 16 pulgadas (2019). Procesador de

2,6 GHz Intel Core i7 de seis núcleos. Tarjeta gráfica AMD Radeon Pro 5300M de 4 GB, Memoria de 16 GB a 2667 MHz DDR4 y disco duro de estado sólido de 256 GB de almacenamiento. Adicionalmente, el software a utilizado en la máquina personal fue el entorno de desarrollo RStudio en la versión 2025.05.0+496 y en un entorno que garantice la conexión continua a internet para la descarga de paquetes dentro del entorno R.

Por otro lado, para la etapa de entrenamiento, se utilizó el cluster llamado Xi NetRAIDER 64 LT del Departamento de Ingeniería en Informática de la Universidad de Santiago de Chile. El cluster combina 2 nodos CPU de alto rendimiento (EPYC 7513) con 2 nodos GPU A30, totalizando 4 nodos optimizados para tareas de procesamiento paralelo y aprendizaje profundo

3.6 RESUMEN

En este capitulo se formuló la arquitectura metodológica del estudio, describiendo el origen de los datos provenientes de los estudios de diagnóstico e intervención del proyecto TUTELAGE. Se definió el pipeline de procesamiento de los datos que comienza con la obtención de estos hasta, finalmente, el entrenamiento y evaluación de modelos predictivos. Este procesamiento está alineado con la metodología OSEMN, la cual consta de cinco etapas en las que comienza con la obtención de los datos y finaliza con el análisis de los resultados de los modelos.

Durante los siguientes capítulos se mostrarán y analizaran los resultados obtenidos al ejecutar este pipeline con las bases de datos del proyecto TUTELAGE.

CAPÍTULO 4. ANÁLISIS EXPLORATORIO

Tras haber definido el pipeline de procesamiento de datos y las variables de estudio en el capitulo anterior, este capitulo tiene como objetivo analizar los datos en la base de datos de intervención observando únicamente el primer segmento de consulta. El objetivo principal de este capitulo es caracterizar comportamientos de tipeo que puedan (o no) asociarse a la dificultad percibida durante la búsqueda de información. En paralelo al análisis principal, se examinó la base de datos correspondiente al experimento de diagnóstico, con el objetivo de complementar los resultados obtenidos. Dicho análisis se presenta en el Anexo B.

Preparación y calidad de los datos (Sección 4.1) Se realizó un análisis preliminar del conjunto completo de los datos para conocer las distribuciones, y luego se realizó una limpieza de valores atípicos.

Caracterización de Patrones (Sección 4.2) Se aplicaron técnicas de aprendizaje no supervisado (Análisis de Componentes Principales y Modelo de Mezcla Gaussiana) para descubrir perfiles de tipeo intrínseco en los estudiantes, independientes de lo que estos hayan reportado. Esto permite identificar patrones generales de los datos basados en la fluidez y producción de escritura.

Validación con la dificultad percibida (Sección 4.3) Se realizó un contraste entre estos perfiles y las métricas de tipeo, así como las variables auto reportadas por los estudiantes, utilizando pruebas estadísticas para validar si la fluidez de escritura es una variable proxy de la dificultad percibida.

Análisis del Contexto Experimental (Secciones 4.4 y 4.5) Finalmente, se evaluó como factores externos, tales como el tipo de intervención pedagógica (Pre vs post test y grupos ludificados, tradicionales y asistente virtual), influyen en las dinámicas de tipeo. Esto proporciona una visión general del fenómeno antes de iniciar el proceso de modelado predictivo.

En resumen, este enfoque permite entender no solo las estrategias de escritura de un estudiante, sino también cómo se relaciona esto con los distintos niveles de dificultad percibida ante una tarea de búsqueda de información.

4.1 EXPLORACIÓN PRELIMINAR DE LOS DATOS

Sobre los datos utilizados para este estudio, la base de datos de intervención cuenta con 1776 registros. A continuación, se detallarán los hallazgos encontrados en esta etapa. En primera instancia, se realiza un análisis sobre los datos sin procesar para conocer sus características preliminares. Los resultados se presentan en la Tabla 4.1.

Tabla 4.1: Estadísticas descriptivas de variables de 1er Segmento de consulta, Intervención

Tipo	Variable (unidad)	Mínimo	1er. Cuartil	Mediana	Media	3er. Cuartil	Máximo
Segmento de query	qs (conteo)	1	1	1	1.75	2	13
Comportamiento de búsqueda	effectiveCharacterCount (caracteres)	0	43	60	67.8	88	267
	totalCharacterCount (caracteres)	0	2	30	40.1	64	434
	charactersPerMinute (caract./min)	0	44.8	86.2	119	142	7059
	reformulations (conteo)	0	0	1	1.49	2	20
	copypaste (binaria)	0	0	1	0.572	1	1
	queryTime (ms)	0	5307	18626	27615	37210	291583
	writingTime (ms)	0	5074	18495	27451	36902	291510
	pauseTime (ms)	0	2802	10461	19647	25063	283761
	pauseCount (conteo)	0	1	5	10.4	15	106
Dinámica de tipeo	dwellTimeAvg (ms)	5	99	125	159	174	963
	dwellTimeStd (ms)	0	29.5	46.5	97.4	116	1362
	flightTimeAvg (ms)	12	335	510	1218	815	102928
	flightTimeStd (ms)	0.707	391	774	1617	1499	71964
	latencyAvg (ms)	20.3	336	598	1744	1458	102920
	latencyStd (ms)	7.07	407	840	2159	1915	99224
	intervalAvg (ms)	20.3	336	598	1744	1458	102920
	intervalStd (ms)	7.07	407	840	2159	1915	99224
	upToUpAvg (ms)	2	431	646	1595	1065	103002
	upToUpStd (ms)	1.41	481	965	2068	1922	100771
Similitud de consulta	copypaste (0–1)	0	0.216	0.486	0.503	0.771	1
Dificultad	dificulty_pre-test (Likert 1–6)	1	2	3	3.17	4	6
	dificulty_post-test (Likert 1–6)	1	1	3	2.91	4	6
Demografía	edad (años)	11	12	12	11.9	12	14

Fuente: Elaboración propia

Lo primero que se puede observar de la Tabla 4.1 son las distancias entre los valores mínimos y máximos en las características de dinámicas de tipeo lo cual expone la potencial existencia de datos atípicos en el dataset (o de distribuciones cargadas hacia la derecha). Además no se observa la presencia de valores negativos.

Para revisar más a fondo este comportamiento, la Figura 4.1 muestra un gráfico de caja en donde se representan todos los valores de las características numéricas de la base de datos en torno a la dificultad percibida. Este tipo de gráficos muestra la distribución de los datos a través de cuartiles, identificando la mediana, el primer y tercer cuartil (la caja). Adicionalmente, los bigotes de la caja representan los datos que van desde el mínimo hasta el primer cuartil y desde el tercer cuartil hasta el máximo. Cualquier valor que no se encuentre dentro de estos rangos se consideró como un valor atípico y, en la Figura, es marcado individualmente con un punto.

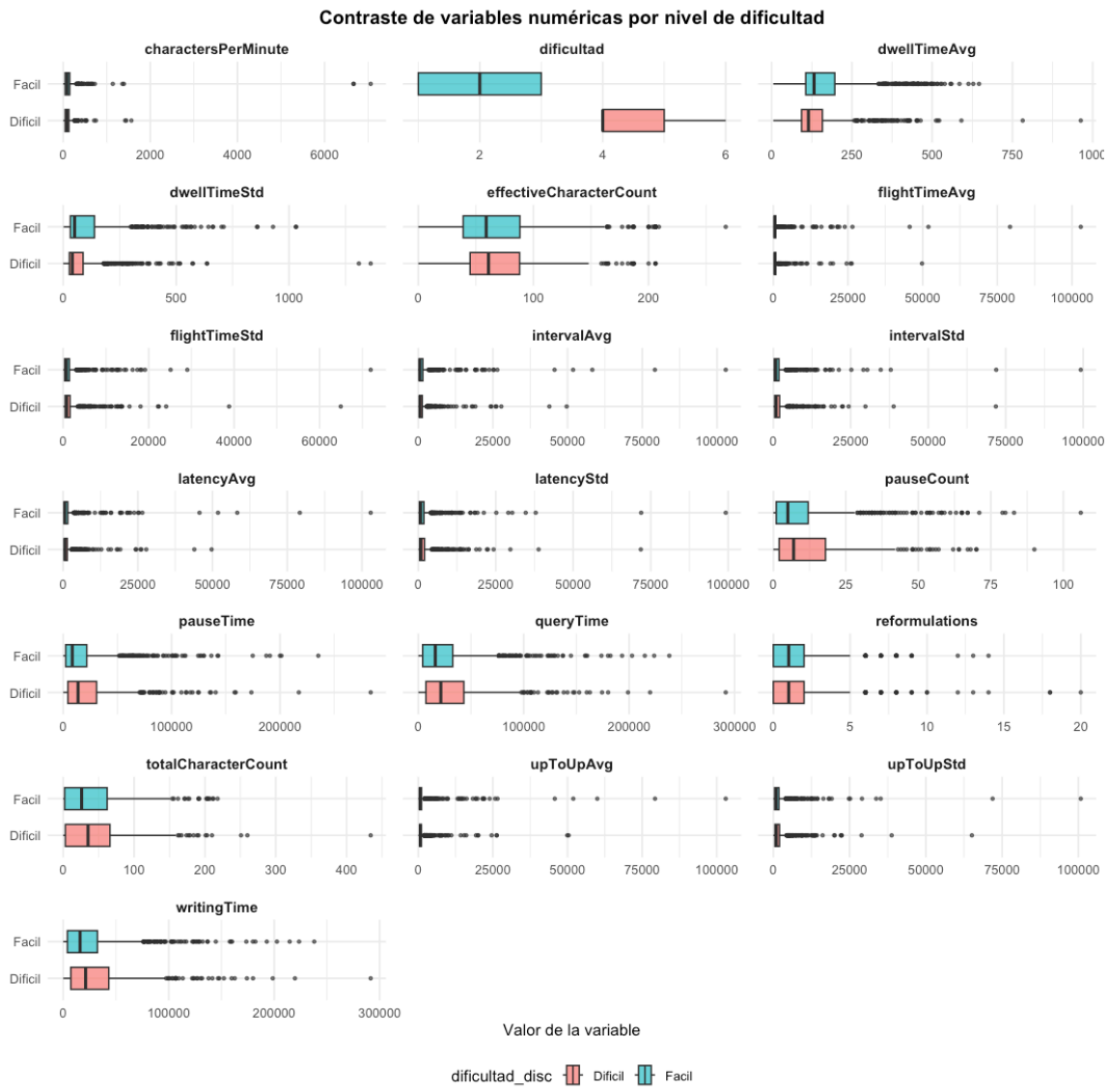

Figura 4.1: Gráfico de caja, intervención, primer segmento de consulta

Fuente: Elaboración propia

En la Figura 4.1 se puede apreciar una gran cantidad de datos atípicos y distribuciones desbalanceadas en la mayoría de las características. Otra mirada se ilustra en la Figura 4.2, donde los gráficos de distribución univariados muestran que la mayoría de las distribuciones están sesgadas hacia la derecha. Además, como se pudo revisar en las figuras anteriores, se observa la posible presencia de valores atípicos. Por ejemplo, los tiempos de escritura (writingTime) superiores a 20 minutos, entendiendo que en ese tiempo el estudiante ya debería haber terminado el desafío.

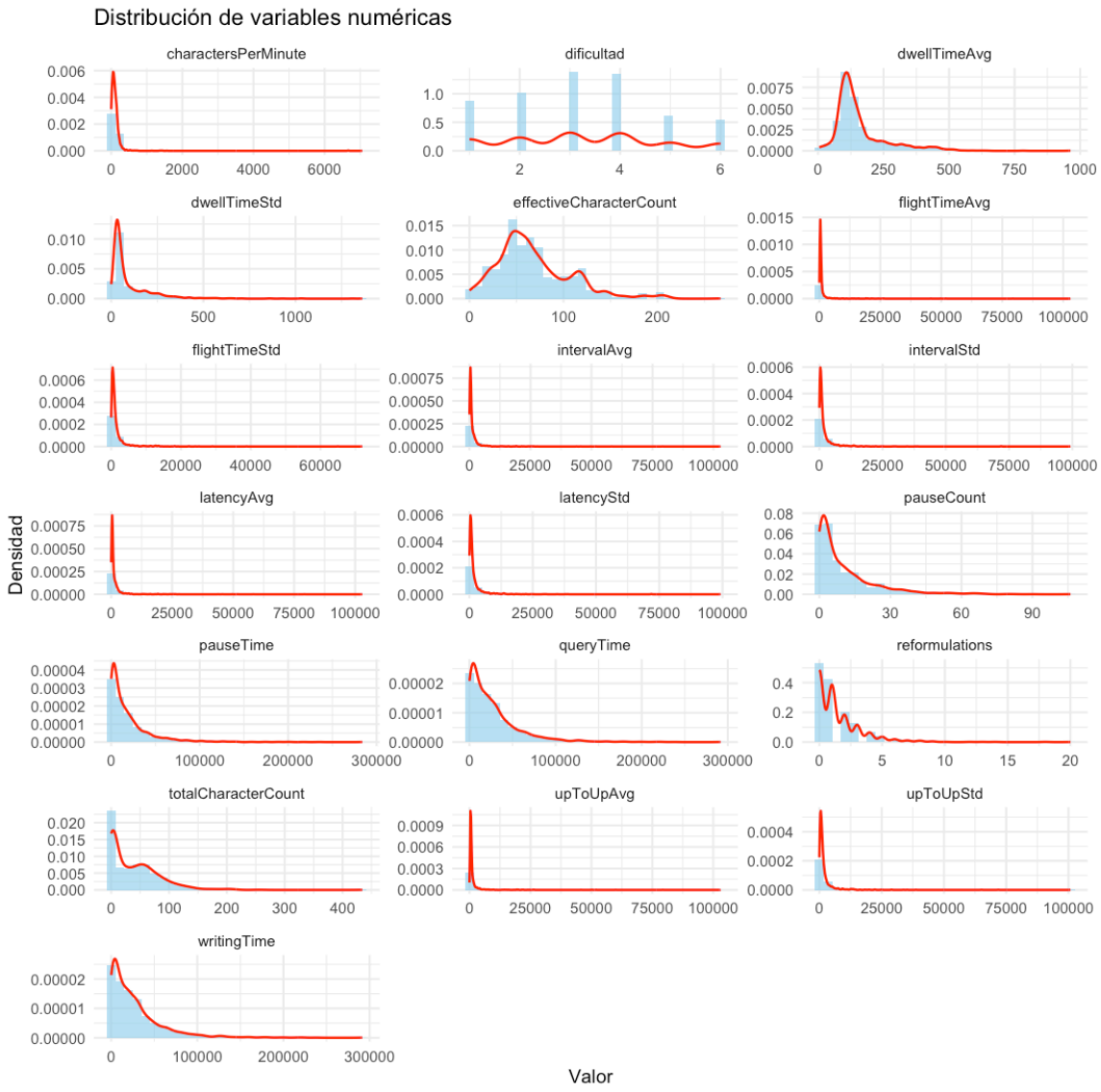

Figura 4.2: Distribuciones de características, intervención, primer segmento

Fuente: Elaboración propia

Estos resultados conducen a realizar técnicas de eliminación de valores atípicos ya que estos valores podrían afectar la estabilidad y el desempeño de los modelos de aprendizaje automático.

4.1.1 Eliminación de valores atípicos

Para definir la naturaleza de los análisis estadísticos subsiguientes, como por ejemplo la elección entre pruebas paramétricas o no paramétricas, y además, determinar la estrategia más adecuada para el tratamiento de valores atípicos, se realizó un test de bondad de ajuste de Shapiro-Wilk (ya que el tamaño del dataset es menor a 5000 observaciones) con un nivel de significancia α = 0.05. Este test plantea las siguientes hipótesis:

1. Hipótesis Nula (H_0): La variable sigue una distribución normal.
2. Hipótesis Alternativa (H_1): La variable no sigue una distribución normal.

Como se observa en la tabla 4.2 (y gráficamente en la Figura 4.2) ninguna de las características describe una distribución normal ($p < 0{,}05$) y, además, la mayoría están sesgadas hacia la derecha. En este escenario, es conveniente utilizar variantes no paramétricas para la detección de outliers, ya que en estas distribuciones, que se asemejan más a una distribución log-normal, es común que la media sea mucho mayor que la mediana debido a los valores extremos de la cola derecha que inflan el promedio (Seo, 2006).

Tabla 4.2: Resultados de la prueba de normalidad Shapiro–Wilk

Fuente: Elaboración propia

Variable	**W**	**p-value**
effectiveCharacterCount	0.964	$8{,}35 \times 10^{-11}$
totalCharacterCount	0.982	$1{,}52 \times 10^{-6}$
queryTime	0.945	$7{,}50 \times 10^{-14}$
writingTime	0.945	$7{,}29 \times 10^{-14}$
pauseTime	0.900	$< 1 \times 10^{-16}$
pauseCount	0.919	$< 1 \times 10^{-16}$
charactersPerMinute	0.965	$1{,}65 \times 10^{-10}$
dwellTimeAvg	0.979	$1{,}61 \times 10^{-7}$
dwellTimeStd	0.863	$< 1 \times 10^{-16}$
latencyAvg	0.944	$4{,}24 \times 10^{-14}$
flightTimeAvg	0.953	$1{,}10 \times 10^{-12}$
intervalAvg	0.944	$4{,}24 \times 10^{-14}$
intervalStd	0.893	$< 1 \times 10^{-16}$
latencyStd	0.893	$< 1 \times 10^{-16}$
upToUpAvg	0.943	$3{,}34 \times 10^{-14}$
copypaste	0.454	$< 1 \times 10^{-16}$

Dada esta información, se optó por escoger una metodología de eliminación de valores atípicos robusta que no dependa de la distribución de las variables. Es por esto que se optó por utilizar el método de rango intercuartilico (IQR), que consiste en detectar valores que se encuentran fuera de este rango.

El proceso de limpieza implicó la eliminación de 333 datos para diagnóstico y 525 valores eliminados para el estudio de intervención. Cabe destacar que la metodología utilizada, aparte de detectar los valores atípicos, elimina la observación completa si contiene al menos un valor atípico. En la Figura 4.3 se muestran en color rojo los valores atípicos capturados mediante este método, mientras que los negros son los valores que permanecerán en la base de datos.

Figura 4.3: Eliminación de valores atípicos, 1er segmento de consulta, intervención.

Fuente: Elaboración propia

Al observar los resultados de esta etapa, se puede identificar que gran parte de los valores fueron eliminados. Es importante destacar que existen casos en los que una observación puede tener mas de un valor atípico. Independientemente de esto, como se mencionó anteriormente, la observación completa fue descartada. Para mayor detalle sobre las características de este subconjunto eliminado, visitar el Anexo G.

4.1.2 Distribución de segmentos de búsqueda

Adicionalmente al proceso de eliminación de valores atípicos, se investigó cuantos segmentos de búsqueda existen en la base de datos. Como se mencionó anteriormente, un usuario puede realizar una o más consultas en la caja de búsqueda durante el transcurso de un desafío (Figura 2.5). En particular se puede observar, a través de la Figura 4.4, que en la base de datos de intervención existe una tendencia a realizar solamente una consulta por desafío.

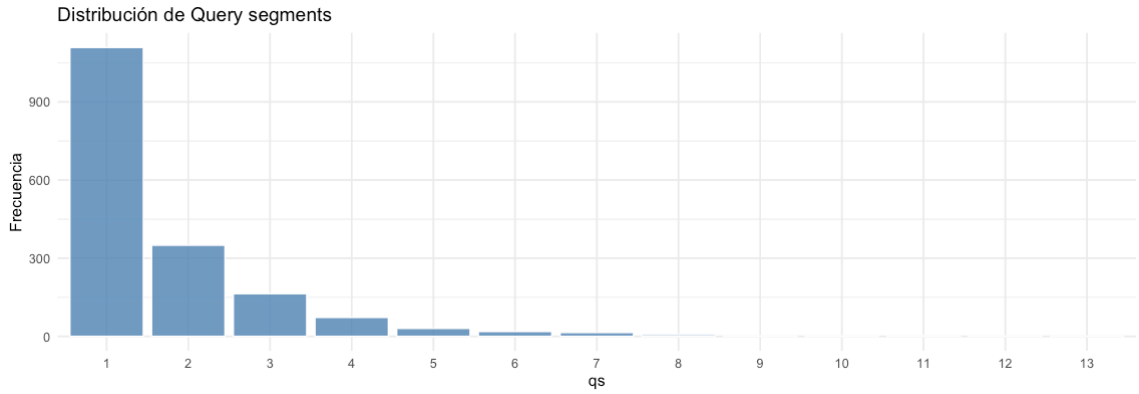

Figura 4.4: Distribución de segmentos de consulta, intervención, primer segmento de consulta

Fuente: Elaboración propia

4.1.3 Valores faltantes

Para distinguir los valores faltantes de la base de datos, se realizó un análisis exploratorio para identificar cuantos valores faltantes existen en cada una de las bases de datos. De manera similar a como se hizo para la búsqueda y eliminación de valores atípicos, se eligió el criterio de eliminar la observación completa en caso de encontrar al menos un valor faltante. Los resultados obtenidos muestran que en los datos de intervención, alrededor de 234 observaciones serán descartadas.

4.1.4 Desbalances

Al visualizar el desbalance de ambos datasets en la Figura 4.5, se puede observar que existe alrededor de un 13 % de desbalance en la variable dependiente. Si bien esto no supone un sesgo para metodologías mas robustas de modelamiento, como por ejemplo XGBoost, dada la característica exploratoria de este trabajo, se optará por aplicar metodologías de sub-muestreo para garantizar la coherencia entre modelos que si son sensibles a esta diferencia de muestras.

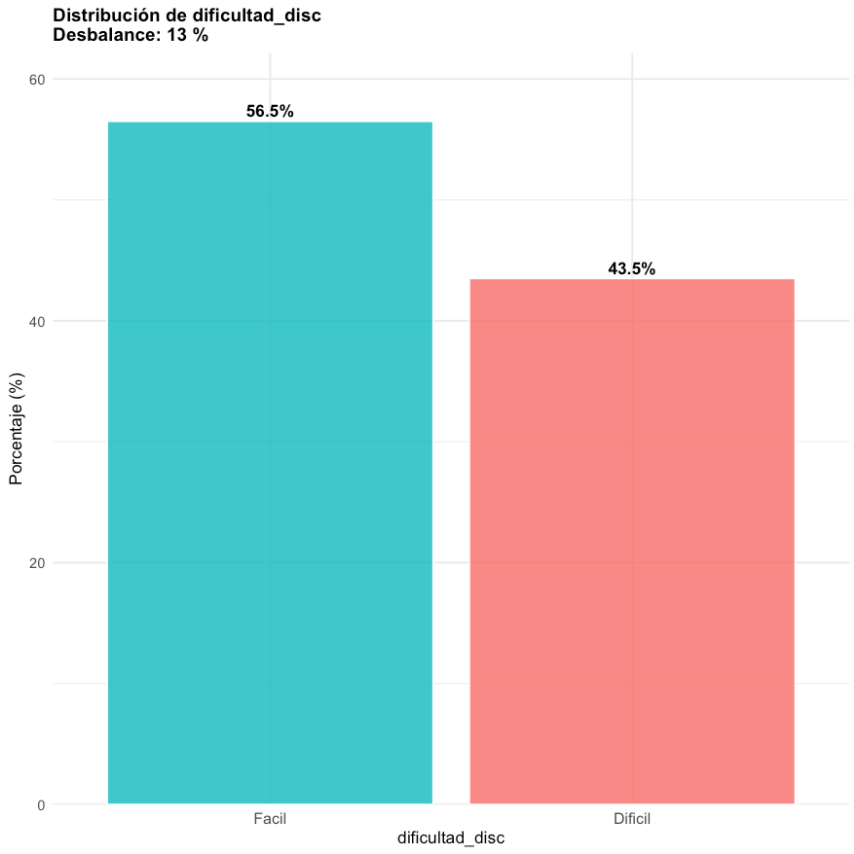

Figura 4.5: Desbalance en conjunto de datos de Intervención

Fuente: Elaboración propia

4.1.5 Resumen de observaciones eliminadas

La Tabla 4.3 resume el proceso de limpieza del conjunto de datos. A partir de 1776 registros iniciales, se aplicó un filtrado por segmento de consulta (QS = 1), seguido de la eliminación de observaciones con valores faltantes y la detección de valores atípicos. Como resultado, se obtuvo un conjunto final de 584 observaciones válidas para el análisis.

Tabla 4.3: Proceso de filtrado y limpieza del conjunto de datos

Fuente: Elaboración propia

Etapa del proceso	**Registros eliminados**	**Registros restantes**
Registros iniciales	–	1 776
Filtrado por segmento de consulta (QS = 1)	667	1 109
Eliminación de valores faltantes (NA)	234	875
Eliminación de valores atípicos (outliers)	291	584
Total final de registros	**1 192**	**584**

4.2 CARACTERIZACIÓN DE PATRONES DE TIPEO MEDIANTE TÉCNICAS NO SUPERVISADAS

Con el objetivo de descubrir patrones en la estructura interna de los datos, independientemente de las etiquetas de dificultad percibida o del grupo experimental asignado, se aplicaron técnicas de aprendizaje no supervisado. Este enfoque permite identificar la estructura natural del comportamiento de búsqueda y perfilar a los estudiantes según sus métricas de dinámica de tipeo.

4.2.1 Reducción de dimensionalidad (PCA)

Dada la elevada correlación teórica entre las variables de dinámica de tipeo —por ejemplo, se podría suponer que un estudiante con una alta velocidad de escritura tendería a presentar valores bajos en las variables de latencia e intervalo— se decidió aplicar un análisis de componentes principales (PCA).

Esta técnica de reducción de la dimensionalidad se basa en un algoritmo que transforma un conjunto inicial de variables entre sí correlacionadas en un nuevo conjunto de variables ortogonales, denominadas componentes principales. De este modo, se reduce la dimensionalidad del espacio de características y se facilita tanto el análisis estadístico como la representación y visualización de los datos.

Para encontrar el número de componentes, se realizó el método del codo para visualizar la varianza explicada por el conjunto de datos. Este análisis, ilustrado en la Figura 4.6, evidencia que entre las primeras dos componentes se explica el 70,1 % de la varianza. Aunque una tercera componente elevaría la varianza explicada al 79,3 %, se optó por trabajar con las dos primeras para facilitar la interpretación de los perfiles de fluidez y producción textual. Posteriormente, se realizó el análisis de componentes principales. Este análisis, el cual está representado en las Figuras 4.7 y 4.8, evidenció que las variables de dinámicas de tipeo están representadas en la primera componente principal y los de producción textual en la segunda componente principal. A continuación se definirán estas componentes en más detalle:

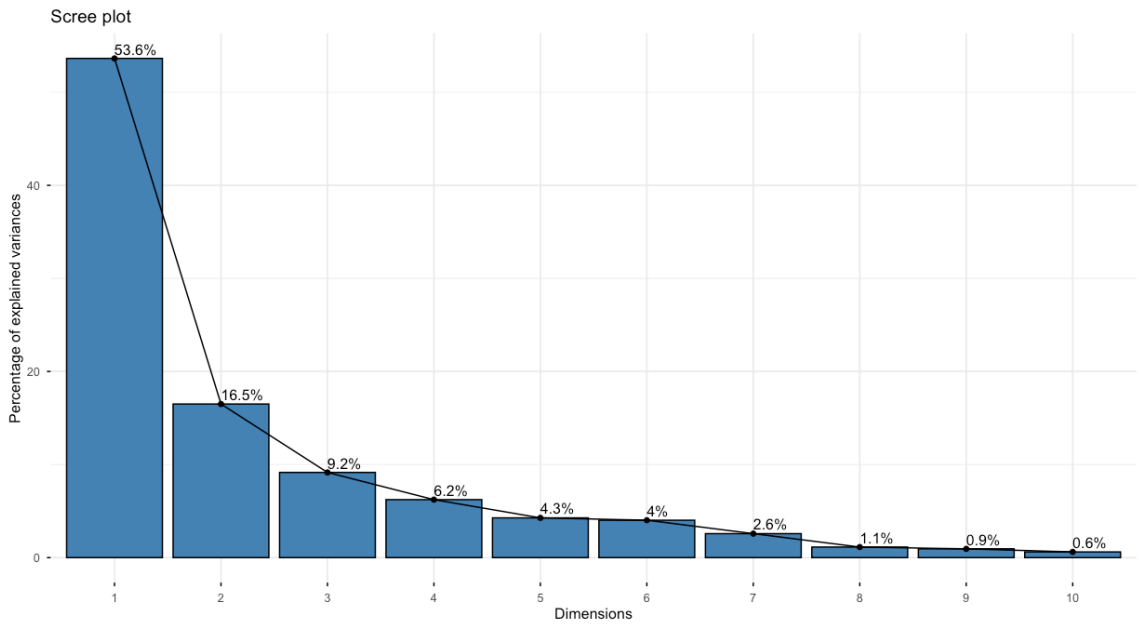

Figura 4.6: Gráfico de codo, 1er segmento, intervención

Fuente: Elaboración propia

- **PC1 - Componentes de fluidez de escritura (53.6 % de la varianza)**: Esta dimensión está influenciada por la mayoría de las características de las dinámicas de tipeo, como, por ejemplo, los promedios de las latencias, tiempos de vuelo, tiempos de intervalos y también las variables de sus respectivas desviaciones estándar. Además, estas variables correlacionan negativamente con la cantidad de caracteres escritos, indicando que, a mayor lentitud en la fluidez de escritura, menor es la producción textual. Este comportamiento muestra que esta dimensión explica la fluidez de tipeo.

- **PC2 - Componente de producción textual (16.5 % de la varianza)**: Esta dimensión está influenciada por las características de producción efectiva de texto, como, por ejemplo, reformulaciones y cantidad de caracteres efectivos (effectiveCharacterCount). Esto sugiere que PC2 explica el volumen de texto generado durante la formulación de la consulta.

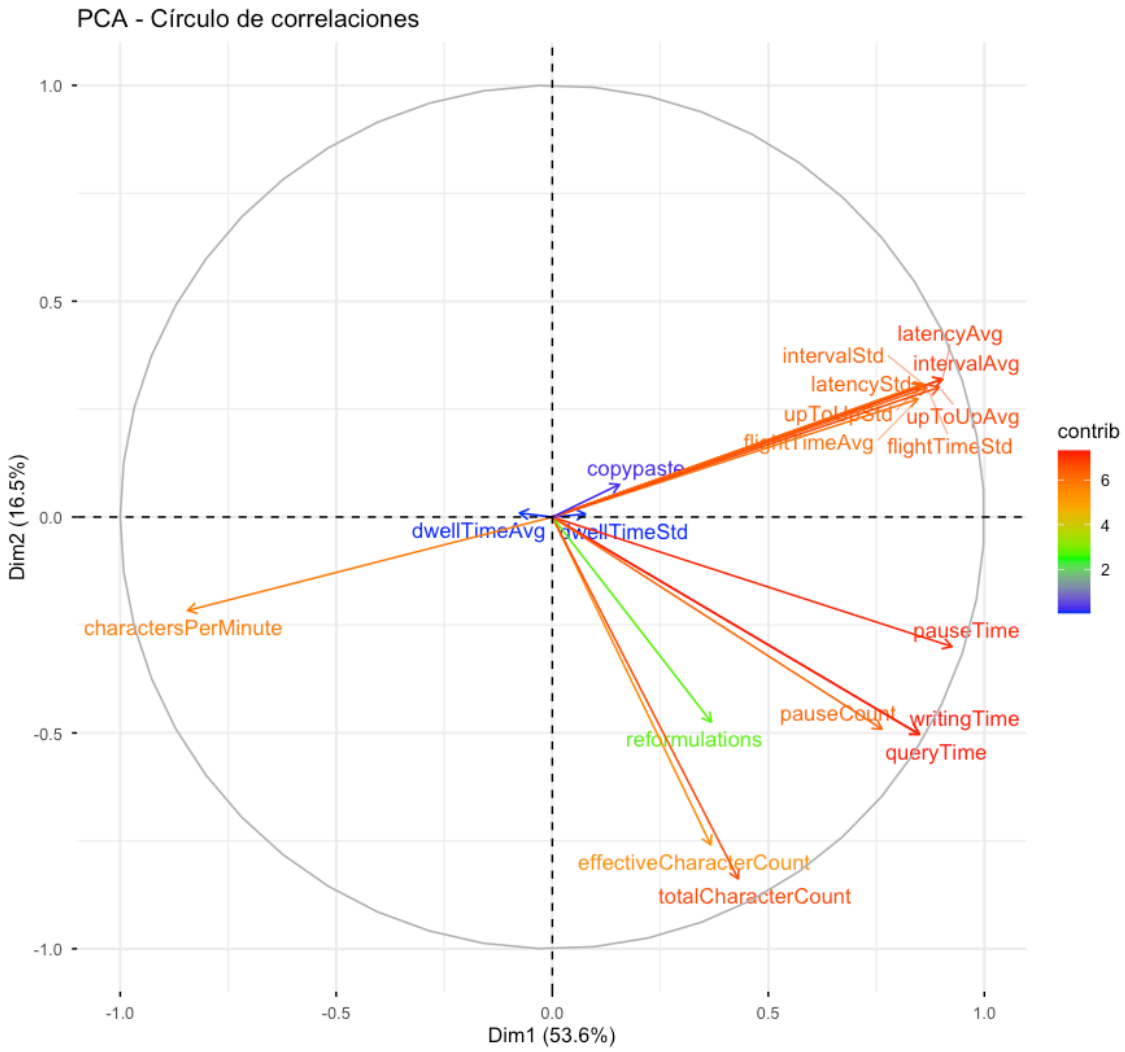

Figura 4.7: Análisis de componentes principales, 1er segmento, intervención

Fuente: Elaboración propia

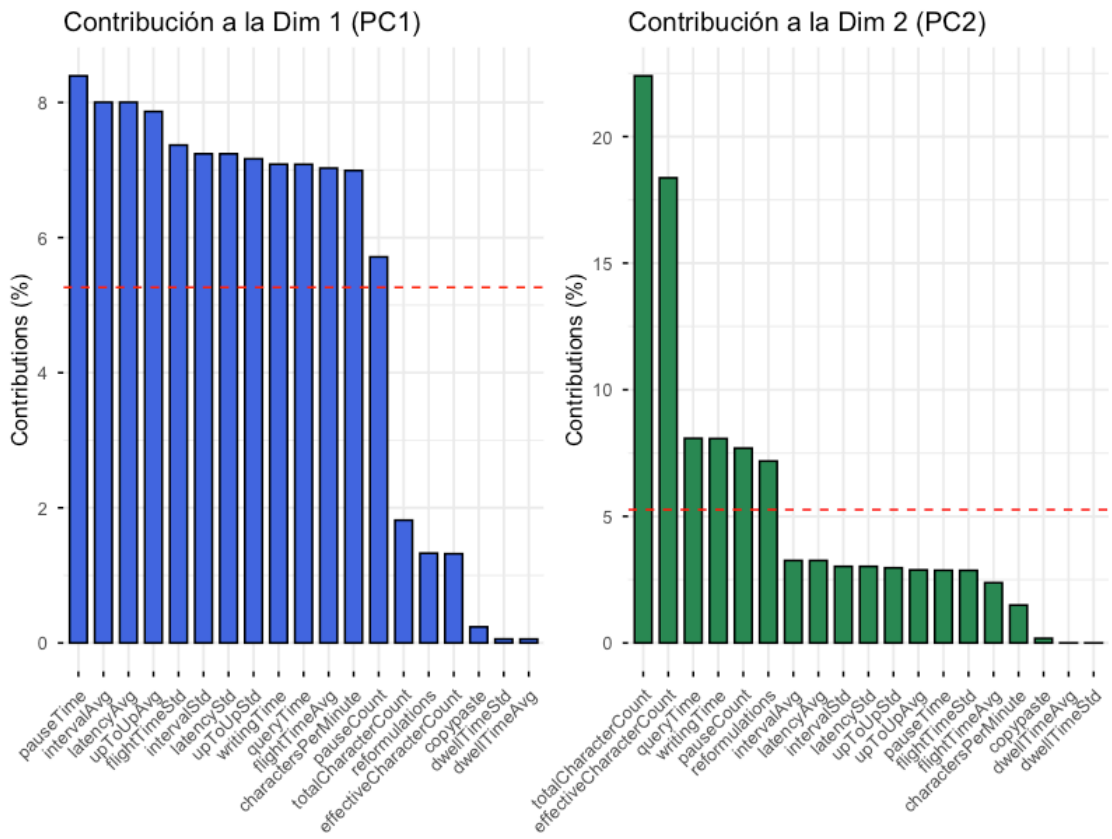

Figura 4.8: Gráfico de contribución por componente

Fuente: Elaboración propia

Los resultados obtenidos en PCA perfilan distintos tipos de comportamiento frente a la producción de consultas.

1. Cuadrante superior derecho: Estudiantes con baja fluidez de escritura y bajo volumen de producción textual: Estudiantes que formularon consultas breves y a un ritmo lento de escritura.

2. Cuadrante inferior derecho: Estudiantes con baja fluidez de escritura y alto volumen de producción textual: Estudiantes que, a pesar de escribir a un menor ritmo, elaboraron consultas extensas o bien revisadas.

3. Cuadrante superior izquierdo: Estudiantes con alto ritmo y volumen (más caracteres efectivos, más reformulaciones) de producción textual: Estudiantes que formularon consultas largas, ricas en contenido.

4. Cuadrante superior izquierdo: Estudiantes con alta fluidez de escritura y bajo volumen de producción textual: Estudiantes que formularon consultas rápidamente y de manera concisa.

4.2.2 Identificación de perfiles

Si bien el análisis de componentes principales (PCA) permitió reducir la dimensionalidad del conjunto de datos para entender la estructura latente de estos, esta técnica no segmenta explícitamente a los sujetos para generar un análisis cualitativo de la información. Es por esto que se realiza un análisis de agrupamiento para identificar los perfiles de estudiantes y relacionarlos con las estrategias de búsqueda.

Es importante mencionar que, para realizar este análisis, se utilizaron las componentes principales generadas en la sección anterior, dada la existencia de multicolinealidad entre las características.

4.2.2.1 Caracterización de los perfiles de tipeo identificados

A partir de los grupos generados mediante el Modelo de Mezcla Gaussiana (GMM), los cuales se pueden visualizar en las Figuras 4.9 y 4.10, se definieron tres perfiles de comportamiento de búsqueda.

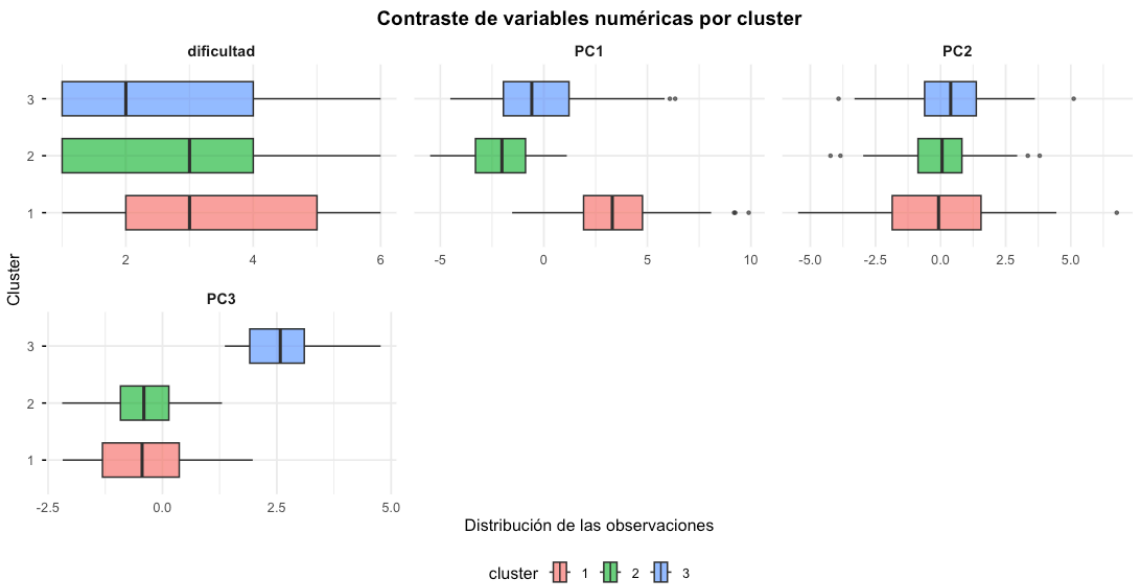

Figura 4.9: Gráfico de caja para grupos generados por GMM

Fuente: Elaboración propia

- **Perfil de baja fluidez de escritura** (Grupo 1 - Rojo): Corresponde a los estudiantes asociados a los valores más altos en la primera componente principal; es decir, cuando existe una baja fluidez de escritura y, además, se observa que predomina una alta dificultad percibida. Este grupo podría asociarse a un alto esfuerzo mental.

- **Perfil de alta fluidez de escritura** (Grupo 2 - Verde): Corresponde al grupo con los valores mas bajos en la primera componente principal; es decir, cuando existe una alta fluidez de escritura. Este comportamiento sugiere un dominio avanzado en la búsqueda de información.

- **Perfil Intermedio** (Grupo 3 - Azul): Corresponde a los estudiantes que se encuentran en un grupo intermedio. Esto sugiere que, si bien el estudiante no está en los niveles de un buscador veloz, logra formular las consultas a un ritmo constante y sin interrupciones, como los observados en el grupo de alta fluidez de escritura.

Figura 4.10: Grupos generados por GMM

Fuente: Elaboración propia

4.2.2.2 Dinámica de transición entre perfiles

Finalmente, para visualizar el impacto de las intervenciones realizadas en la modificación de los patrones de búsqueda, se analizó la movilidad de los estudiantes entre el pretest y el postest. Los resultados pueden ser visualizados en la Figura 4.11.

Primero, se puede observar que el perfil «Alta fluidez» presenta una migración de algunos estudiantes hacia los perfiles «Intermedio» y «Baja fluidez».

En segundo lugar, se observa un flujo considerable de alumnos que migraron del perfil «Baja fluidez» hacia el perfil «Alta fluidez». Al observar las proporciones de los grupos de intervención que migraron, se puede notar que una cantidad moderada de

estudiantes del grupo Control migró. Esto podría indicar que el estudio de intervención no fue la causa raíz de este cambio de comportamiento, sino más bien elementos externos, como la familiarización progresiva con la tarea, el efecto de práctica, el desarrollo natural de habilidades de escritura y búsqueda de información, o bien variables contextuales no controladas durante el período de observación.

Finalmente, se observa una migración importante desde el perfil «Intermedio» hacia «Alta fluidez». Este fenómeno también podría explicarse por la interferencia de componentes externos, como, por ejemplo, la falta de tiempo para terminar los desafíos. Esto podría provocar un ritmo mayor de escritura, no necesariamente consciente, para finalizar en los tiempos establecidos.

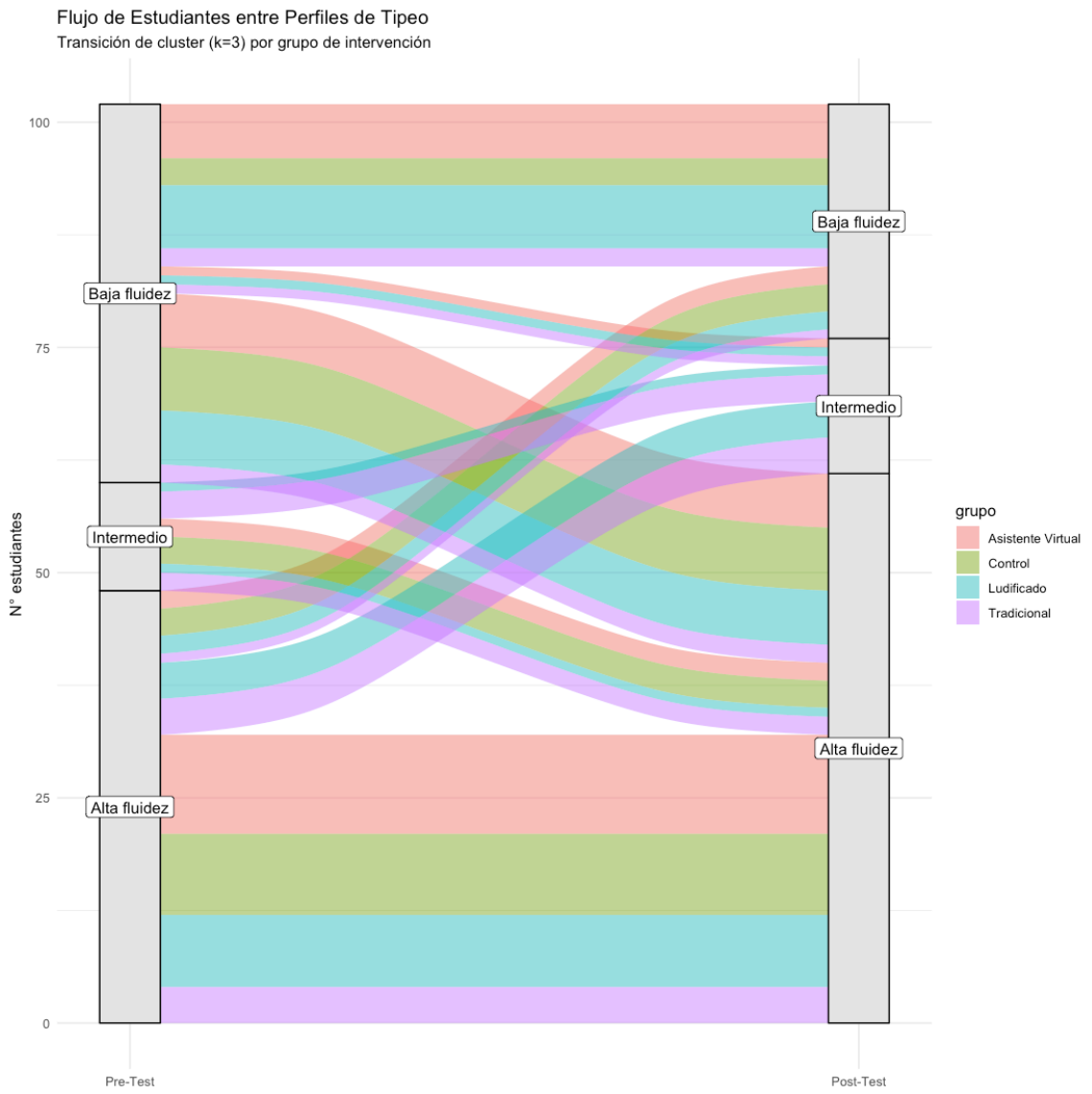

Figura 4.11: Gráfico aluvial para perfiles de pre test y post test

Fuente: Elaboración propia

4.3 RELACIÓN ENTRE VELOCIDAD TIPEO Y DIFICULTAD PERCIBIDA

En la presente sección se mostrarán los resultados obtenidos al comparar las variables de dinámicas de tipeo con la dificultad percibida reportada por los estudiantes.

Para lograr este objetivo, se utilizaron técnicas estadísticas, técnicas de correlación entre variables y análisis de agrupamiento no supervisado.

4.3.1 Comparación estadística entre niveles de dificultad

Con el objetivo de descubrir si las métricas de dinámicas de tipeo difieren estadisticamente entre tareas percibidas como fáciles o difíciles, se aplicaron pruebas no paramétricas (Wilcoxon-mann-whitney) de dos colas. Este análisis permite identificar que aspectos caracterizan ambas percepciones de una tarea, como por ejemplo la latencia de tipeo, la velocidad de tipeo, entre otros. Las hipótesis planteada en este análisis fue la siguiente:

1. Hipótesis nula (H_0): El ritmo de producción textual y las latencias de tipeo son independientes del nivel de dificultad percibida en una tarea de búsqueda.
2. Hipótesis Alternativa (H_A): El nivel de dificultad percibida de una tarea de búsqueda ejerce un efecto significativo sobre la fluidez y el ritmo de la producción textual.

Tabla 4.4: Resultados de pruebas de contraste por variable – Wilcoxon no pareado (pre y post test)

Fuente: Elaboración propia

Variable	Estadístico	p-valor	Método	Tamaño del efecto
latencyAvg	29599	$\mathbf{1{,}13 \times 10^{-9}}$	Wilcoxon	$-0{,}294$
intervalAvg	29599	$\mathbf{1{,}13 \times 10^{-9}}$	Wilcoxon	$-0{,}294$
upToUpAvg	30410	$\mathbf{1{,}28 \times 10^{-8}}$	Wilcoxon	$-0{,}274$
flightTimeAvg	30588	$\mathbf{2{,}13 \times 10^{-8}}$	Wilcoxon	$-0{,}270$
charactersPerMinute	52848	$\mathbf{6{,}28 \times 10^{-8}}$	Wilcoxon	$0{,}261$
pauseTime	33130	$\mathbf{1{,}40 \times 10^{-5}}$	Wilcoxon	$-0{,}210$
dwellTimeAvg	50481	$\mathbf{2{,}23 \times 10^{-5}}$	Wilcoxon	$0{,}205$
pauseCount	33564	$\mathbf{3{,}60 \times 10^{-5}}$	Wilcoxon	$-0{,}199$
intervalStd	33578	$\mathbf{3{,}76 \times 10^{-5}}$	Wilcoxon	$-0{,}199$
latencyStd	33578	$\mathbf{3{,}76 \times 10^{-5}}$	Wilcoxon	$-0{,}199$
flightTimeStd	33788	$\mathbf{5{,}88 \times 10^{-5}}$	Wilcoxon	$-0{,}194$
upToUpStd	33860	$\mathbf{6{,}83 \times 10^{-5}}$	Wilcoxon	$-0{,}192$
writingTime	35426	**0,00134**	Wilcoxon	$-0{,}155$
queryTime	35435	**0,00136**	Wilcoxon	$-0{,}154$
effectiveCharacterCount	37195	**0,0197**	Wilcoxon	$-0{,}113$
dwellTimeStd	44797	0.153	Wilcoxon	$0{,}0689$
totalCharacterCount	40261	0.415	Wilcoxon	$-0{,}0393$
copypaste	42237	0.848	Wilcoxon	$0{,}0078$

Nota: En negrita se indican las diferencias significativas al nivel $\alpha = 0{,}05$. El tamaño del efecto corresponde a Cliff's Delta.

Los resultados presentados en la tabla 4.4. En términos generales, se evidenciaron diferencias significativas en la mayoría de las metricas de dinámica de tipeo (p <0.05). Las variables con mayor evidencia son: promedio de latencia (latencyAvg), promedio de intérvalos (intervalAvg) y promedio de liberaciones (upToUpAvg). Estos resultados evidencian una relación entre una mayor fluidez en la escritura y menores niveles de dificultad percibida en la realización de una tarea.

En conjunto, estos resultados indicarían que la percepción de dificultad, está relacionada con aspectos de fluidez en la escritura, donde se ve degradada en tareas consideradas difíciles. A partir de estos resultados, se llevó a cabo un análisis de correlaciones para entender las relaciones entre variables.

4.3.2 Relaciones entre métricas de tipeo

En esta sección se realizó un análisis de correlaciones para identificar y caracterizar las relaciones entre las características de las dinámicas de tipeo, así como visualizar que variables están fuertemente relacionadas entre si.

Considerando los antecedentes previamente revisados en el análisis exploratorio, donde se mostró que ninguna de las variables numéricas describe una distribución normal (o cercana a esta), se optó por utilizar la correlación de Spearman, puesto que es la alternativa no paramétrica de la correlación de Pearson. La Figura 4.12 es una representación gráfica de la matriz de correlaciones. Los valores positivos se muestran en rojo y los negativos en azul. La intensidad del color está representada proporcionalmente en los coeficientes de Spearman. En la leyenda se encuentra una guía visual de la dinámica descrita. Los principales resultados son los siguientes:

- A medida que aumenta la producción textual (charactersPerMinute), disminuyen las variables de tiempo (writingTime, queryTime, pauseTime) y la cantidad de pausas.
- Variables de dinámicas de tipeo altamente correlacionadas ([0.8; 1])

Sobre los segmentos de consulta, estos presentaron una leve correlación positiva con las variables de dinámicas de tipeo y una leve correlación negativa con las variables temporales. Esto indica que los primeros segmentos de consulta se relacionan con consultas más rápidas y una mayor fluidez en la escritura.

Por otro lado, la correlación negativa entre la edad y los tiempos de escritura podría interpretarse como un indicador indirecto de la experiencia en la producción textual. Los estudiantes de mayor edad tienden a formular consultas en menor tiempo, lo cual podría estar relacionado con alumnos que poseen una mejor comprensión lectora o experiencia en la producción textual.

También existe una alta correlación entre todas las variables de las dinámicas de tipeo. Esto confirma la validez interna del conjunto de datos, puesto que, en conjunto, describen la fluidez de la escritura.

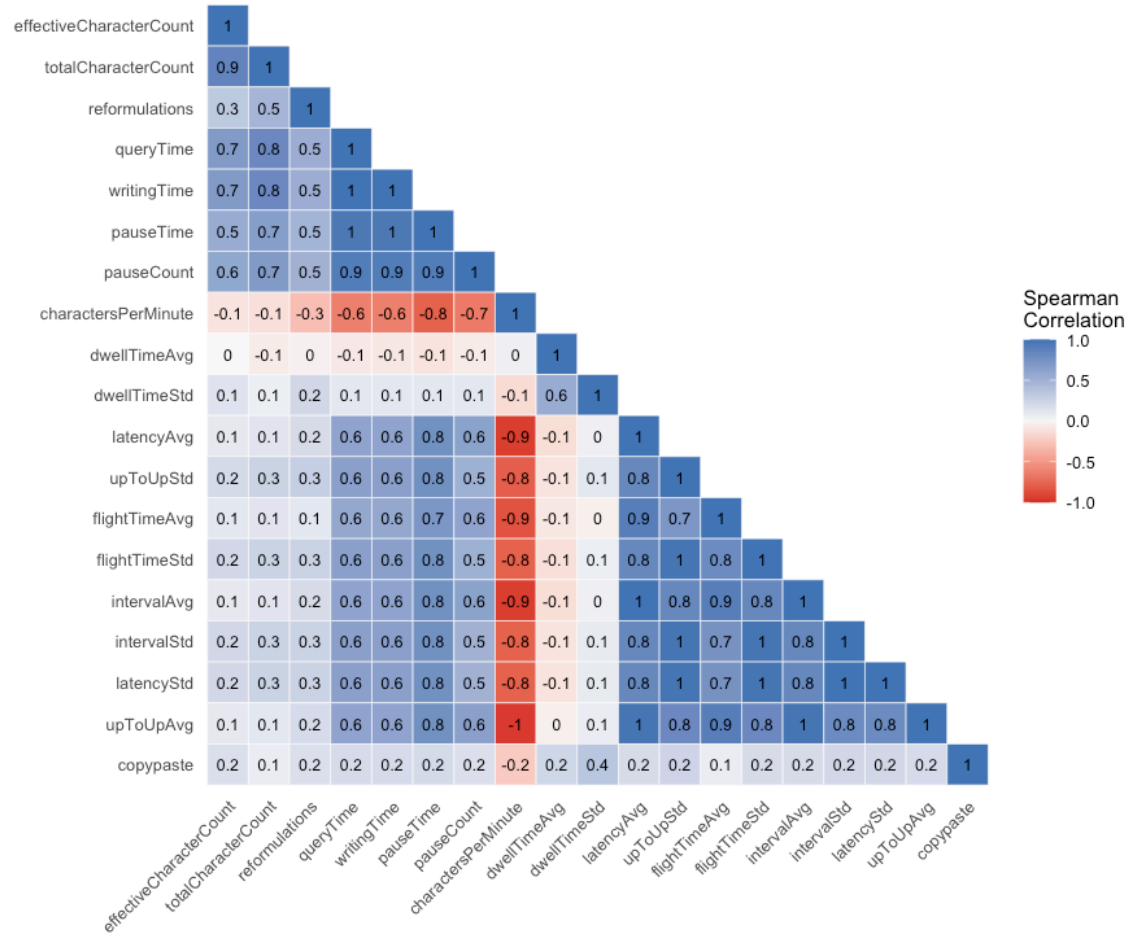

Figura 4.12: Matriz de correlaciones, 1er segmento, intervención

Fuente: Elaboración propia

4.3.3 Regresión logística para la predicción de dificultad percibida

A partir de las dimensiones generadas en el análisis de componentes principales, surge la pregunta de qué tan bien estas componentes modelan la probabilidad de que una observación pertenezca a la clase «Fácil» o «Difícil». Debido a esto, se realizó una regresión logística utilizando las dos primeras componentes que mejor explican la varianza.

La Figura 4.13 muestra los resultados del análisis en la primera componente principal. Los puntos en la parte superior indican los valores en los que la tarea fue indicada como «Difícil», mientras que los puntos inferiores muestran donde la tarea fue indicada como «Fácil». La línea azul es la probabilidad predicha por el modelo de regresión logística con un intervalo de confianza del 95 % (banda gris). Esta línea explica como cambia la probabilidad de que una observación sea «Difícil» en función de PC1.

Al observar los resultados obtenidos en la Figura 4.13, esta linea describe una pendiente positiva, lo que explica que, mientras menor sea el valor de PC1, menor será la probabilidad de encontrar una tarea indicada como «Facil». En caso contrario, mientras más grande sea el valor de PC1, existirá una mayor probabilidad de que la tarea esté indicada como «Dificil».

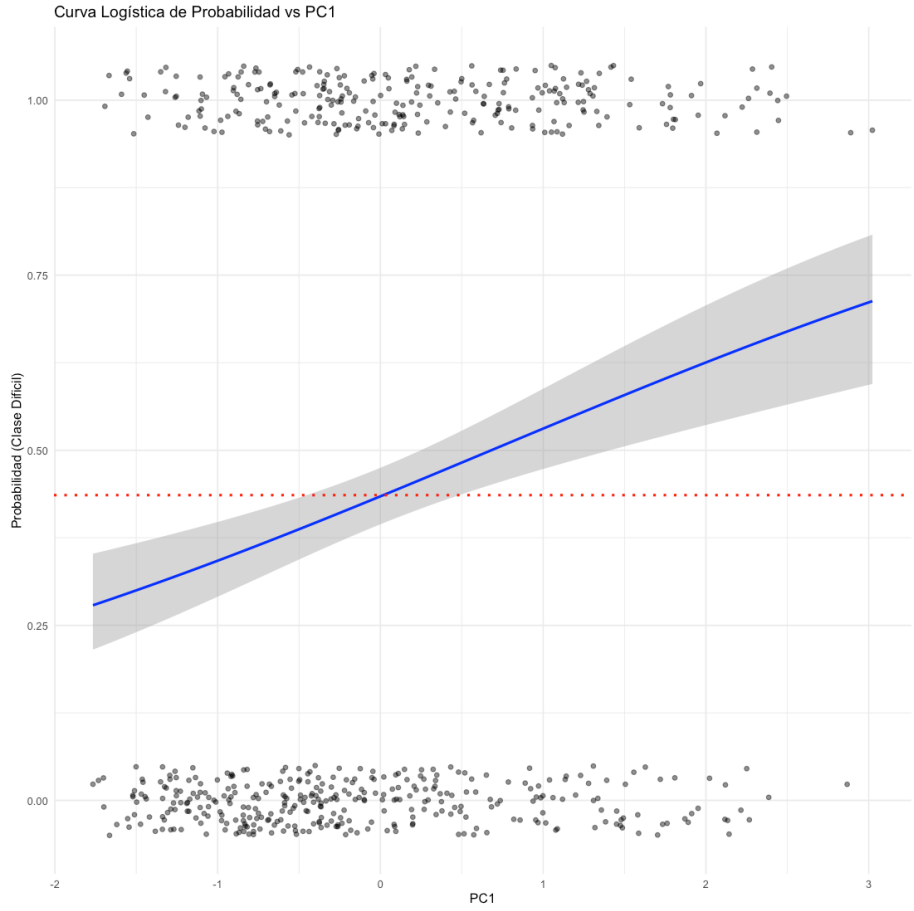

Figura 4.13: Regresión logistica para PC1

Fuente: Elaboración propia

Por otro lado, se realizó el mismo análisis para la segunda componente principal, el cual su resultado se ilustra en la Figura 4.14; sin embargo, a diferencia de la primera componente, el modelo no es capaz de clasificar los datos debido a la alta homogeneidad que existe entre estos.

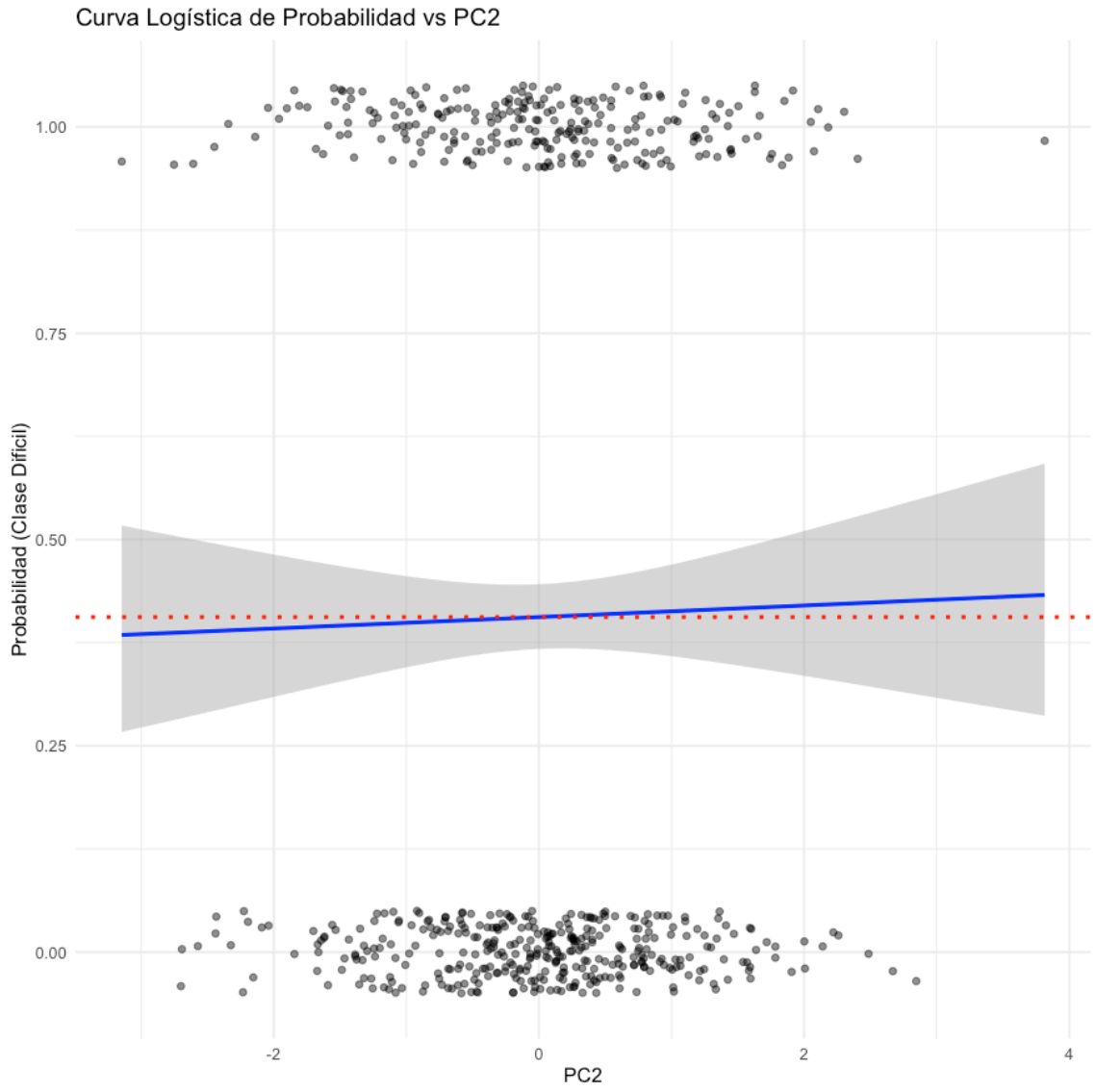

Figura 4.14: Regresión logistica para PC2

Fuente: Elaboración propia

Al observar los resultados previos, la primera componente principal (PC1) describe el ritmo de tipeo en un estudiante, donde valores más altos indican que el ritmo de tipeo fue más lento. Complementando esos resultados con este análisis, se puede concluir

que un bajo ritmo de tipeo está relacionado con una mayor probabilidad de encontrar tareas indicadas como «Difícil».

4.4 FACTORES CONTEXTUALES Y EXPERIMENTALES

Tras haber caracterizado los patrones de tipeo asociados a la dificultad percibida, resulta de interés identificar que variables de dinámicas de tipeo presentaron más variabilidad entre el pre y post test. Para ello, primero se realizó un test de bondad de ajuste (Shapiro-Wilk) en cada una de las características para determinar si el test corresponde a uno paramétrico o no paramétrico. Luego, dependiendo del resultado del test, se realizó un t-test pareado o un Wilcoxon pareado por cada una de las características.

4.4.1 Análisis global de la muestra

El test de bondad de ajuste se realizó utilizando un valor de significancia de $p < 0{,}05$. Las hipótesis planteadas son las siguientes:

1. Hipótesis Nula (H_0): Los datos siguien una distribución normal.
2. Hipótesis Alternativa (H_A): Los datos no siguen una distribución normal.

Para las variables que cumplieron con el test de normalidad, cuyos resultados se pueden observar en la columna «p-norm» de la tabla 4.5, se aplicó un t-test de Student de muestras pareadas. En caso contrario, se aplicó el test Wilcoxon de signos para muestras pareadas. Las hipótesis planteadas para este test estadístico fueron las siguientes:

- Hipótesis Nula (H_0): La media (o mediana dependiendo del test estadístico aplicado) de las diferencias entre el pre-test y el post-test es igual a cero. No existe un cambio real en el promedio de la variable observada.

- Hipótesis Alternativa (H_A): La media (o mediana dependiendo del test estadístico aplicado) de las diferencias es distinta a cero. Existe un cambio significativo en el promedio de la variable observada.

Al analizar la muestra, omitiendo los valores de control, los resultados evidenciaron cambios significativos en la conducta de escritura de los estudiantes tras la intervención. Al observar la tabla 4.5, se puede evidenciar que existe una reducción significativa en la mayoría de las variables de dinámicas de tipeo, como, por ejemplo, latencyAvg (p = 0.00306) e intervalAvg (p = 0.00306). Esto sugiere que la intervención modificó la fluidez de la escritura, sin embargo, como se observó en la Figura 4.11, estos cambios reflejarían el aprendizaje o dominio de la herramientas utilizadas y no un efecto atribuible a la intervención.

Tabla 4.5: Resultados de pruebas pareadas Pre–Post y validación de normalidad ($N = 77$)

Fuente: Elaboración propia

Variable	p**-norm**	**Método**	**Est.**	p**-value**	**Tam. Efecto**
latencyAvg	0.1580	t-test pareado	−3,06	**0,00306**	−0,300
intervalAvg	0.1580	t-test pareado	−3,06	**0,00306**	−0,300
dwellTimeStd	0.0002	Wilcoxon pareado	949	**0,00507**	−0,368
upToUpAvg	0.1430	t-test pareado	−2,75	**0,00745**	−0,262
intervalStd	0.3820	t-test pareado	−2,30	**0,02430**	−0,314
latencyStd	0.3820	t-test pareado	−2,30	**0,02430**	−0,314
upToUpStd	0.4980	t-test pareado	−2,29	**0,02450**	−0,310
flightTimeStd	0.4960	t-test pareado	−1,90	0.06180	−0,258
pauseTime	0.0329	Wilcoxon pareado	1141	0.06760	−0,240
dwellTimeAvg	0.0169	Wilcoxon pareado	1054	0.07320	−0,240
charactersPerMinute	0.0829	t-test pareado	1.78	0.07930	0.187
queryTime	0.0667	t-test pareado	−1,73	0.08790	−0,185
writingTime	0.0672	t-test pareado	−1,72	0.08980	−0,184
effectiveCharacterCount	0.3260	t-test pareado	1.37	0.17400	0.175
pauseCount	0.0396	Wilcoxon pareado	1233	0.23500	−0,157
flightTimeAvg	0.1100	t-test pareado	−1,17	0.24700	−0,122
dificultad	NA	Wilcoxon pareado	886	0.40300	−0,122
totalCharacterCount	0.6230	t-test pareado	−0,61	0.54600	−0,072
copypaste	<0.001	Wilcoxon pareado	730	0.91400	−0,017

Nota: p-norm corresponde al test de Shapiro-Wilk sobre las diferencias. En negrita se indican diferencias significativas ($\alpha = 0{,}05$). El tamaño del efecto es Cohen's d para t-test y Rank-Biserial para Wilcoxon.

4.4.2 Análisis diferenciado por grupo experimental

Para tener más detalle de los resultados expuesto, se realizó un test segregado por grupo experimental (Control, Tradicional, Ludificado y Asistente Virtual) en el dataset de Intervención. Con esto se buscó explorar qué metodologías impactaron más en la fluidez de escritura (ver Anexo D).

Grupo control Se observó una disminución significativa en las métricas temporales, como pauseTime ($p = 0{,}017$) y latencyAvg ($p = 0{,}041$), indicando que los estudiantes escribieron más rápido y con menos pausas. Sin embargo, es importante destacar que no hubo diferencias significativas en los niveles de dificultad percibida. Estos resultados se podrían interpretar como que los estudiantes se volvieron más familiarizados con la interfaz, pero no experimentaron que la tarea fuera más fácil.

Grupo tradicional Se observó en este grupo un comportamiento contrario al del grupo de control. No se observó ninguna reducción significativa en la fluidez y los tiempos de escritura, lo cual indicaría que la intervención no tuvo efecto en términos de dinámicas de tipeo y comportamientos de búsqueda en este grupo experimental.

Grupo ludificado Se observó una mejoría en la fluidez de la escritura. Se observaron reducciones significativas en upToUp ($p = 0{,}0157$), intervalStd ($p = 0{,}0186$) y la variabilidad del tipeo (dwellTimeStd, $p = 0{,}029$). A diferencia del grupo tradicional, donde solamente se visualizó un video previo al post test, el ludificado mejoró la estabilidad y el ritmo de la escritura. Estos resultados indican que la estrategia gamificada promueve una mayor fluidez en la escritura, probablemente inducida por un aumento en la confianza del estudiante al enfrentarse con un buscador en fases previas al desafío.

Grupo Asistente Virtual Finalmente, se observó que el grupo Asistente Virtual no presentó diferencias significativas entre los grupos para ninguna de las variables observadas. Es posible que los alumnos fueran más conscientes del proceso de formulación de consultas en el post test; por tanto, no hubo un incentivo para escribir más rápido.

4.5 RELACIÓN ENTRE TIPO DE TAREA Y DIFICULTAD PERCIBIDA

Tras haber caracterizado a los estudiantes y asociado los distintos tipos de estudio con la dificultad percibida, este capitulo finaliza con un análisis entre los dos tipos de preguntas existentes durante la intervención y su asociación con la dificultad percibida. Estas preguntas pueden ser de dos tipos: justificación de la fuente confiable y respuesta de texto. Para determinar esta relación, se realizó un test de chi cuadrado de independencia entre ambas variables.

Figura 4.15: Representación gráfica test chi cuadrado

Fuente: Elaboración propia

Los resultados mostrados en la imagen 4.15 indican que no existen relaciones significativas entre el tipo de pregunta y la dificultad percibida.

4.6 RESUMEN

El análisis exploratorio realizado en este capítulo ha permitido establecer las bases teóricas y prácticas para la construcción de los modelos predictivos. A través de técnicas estadísticas y de modelado no supervisado, se ha podido triangular el ritmo de tipeo con los efectos propios de la intervención. Esto conduce a tres conclusiones fundamentales que argumentan la relación entre las dinámicas de tipeo y la dificultad percibida en los estudiantes al momento de enfrentarse a una tarea de búsqueda de información.

4.6.1 Existencia de perfiles de comportamiento

Mediante técnicas de reducción de dimensionalidad (PCA) y de Mezcla Gaussiana (GMM) se identificaron estructuras latentes compuestas por tres perfiles de estudiantes: «Grupo de alta fluidez de escritura», «Grupo intermedio» y «Grupo de baja fluidez de escritura». Estos perfiles reflejan patrones distintivos, evidenciando que los estudiantes no abordan las tareas de búsqueda de manera homogénea.

4.6.2 Validación de las métricas como proxy de dificultad

Los análisis bivariados confirmaron que las métricas de tipeo son sensibles a la dificultad percibida. En particular, se observó que los estudiantes que reportaron alta dificultad experimentaron comportamientos significativamente menos fluidos, validando el supuesto de que esta metodología describe la dificultad percibida por un estudiante. Por el contrario, no se observaron relaciones entre el tipo de tarea (justificación y respuesta) y la dificultad percibida, indicando que esta variable no es relevante para el modelo.

4.6.3 Sensibilidad a la intervención pedagógica

El análisis entre pre y post test reveló que las dinámicas de tipeo no son comportamientos estáticos. Si bien se observó una mejora global en la fluidez, no todas las intervenciónes aportaron el mismo efecto. En particular, se observó que la estrategia ludificada logró mejoras significativas en la fluidez de escritura.

En síntesis, este capitulo demuestra la existencia de una señal diferenciable entre el estado cognitivo del estudiante. Estos resultados justifican el paso a la siguiente fase de esta investigación, en donde se evaluó la capacidad de los algoritmos de aprendizaje supervisado en la predicción de estos estados de dificultad percibida.

CAPÍTULO 5. MODELAMIENTO Y RESULTADOS

Tras la caracterización de los patrones de tipeo, este capitulo aborda la construcción y evaluación de modelos predictivos. El objetivo principal es determinar si es posible clasificar la dificultad percibida de un estudiante basándose exclusivamente en sus dinámicas de tipeo. Para ello, se evaluaron nueve algoritmos de clasificación utilizando validación cruzada de 10 pliegues (10-fold CV) con tres repeticiones. Los modelos seleccionados provienen de distintas bases matemáticas: Basados en instancias (KNN), basados en árboles (Random Forest, XGBoost y GBM) y basados en hiperplanos (SVM con distintos kernels)

Es importante destacar que la selección del mejor modelo se realizó bajo un criterio de maximización de la sensibilidad, dado que el costo pedagógico de cometer falsos negativos, es decir, no ofrecer ayuda cuando si la requiere, es mayor que ofrecer ayuda innecesaria.

Finalmente, el área bajo la curva ROC (AUC) se utilizó como métrica principal de comparación, ya que evalúa la capacidad del modelo para distinguir entre clases, independientemente del umbral de decisión.

5.1 SELECCIÓN DE CARACTERÍSTICAS

Previo al entrenamiento, se utilizaron los algoritmos Boruta y Recursive Feature Elimination (RFE) sobre ambos conjuntos de datos con la finalidad de reducir la dimensionalidad, identificar los predictores más robustos y eliminar el ruido. La razón de la elección de estos métodos de selección de características se origina en

Se obtuvo que ambos métodos, en la base de datos de intervención, convergieron a un conjunto optimo de variables en donde prevalecen las de dinámicas de tipeo (ver tabla 5.1 y 5.2). En particular, es llamativo que los algoritmos, además de seleccionar las variables de interes, seleccionaron variables conductuales, como por ejemplo la distancia de Levenshtein, lo cual puede sugerir que la transcripción de texto podría tener un efecto

en discriminar la dificultad de una tarea.

En resumen, los resultados obtenidos a traves de la selección de características refuerzan el planteamiento de que las variables de dinámicas de tipeo pueden modelar la dificultad percibida por un estudiante.

Tabla 5.1: Variables del conjunto pre-test seleccionadas mediante Boruta y RFE (ordenadas por importancia)

Fuente: Elaboración propia

Boruta	**RFE**
dwellTimeAvg	dwellTimeAvg
dwellTimeStd	dwellTimeStd
latencyAvg	flightTimeAvg
flightTimeAvg	lev_similarity_qs
intervalAvg	effectiveCharacterCount

Tabla 5.2: Variables del conjunto post-test seleccionadas mediante Boruta y RFE (ordenadas por importancia)

Fuente: Elaboración propia

Boruta	**RFE**
dwellTimeStd	dwellTimeAvg
dwellTimeAvg	dwellTimeStd
dwellTimeStd	flightTimeAvg
flightTimeAvg	lev_similarity_qs
latencyAvg	effectiveCharacterCount

5.2 EVALUACIÓN DE MODELOS EN PRE Y POST TEST

Tras la selección de variables relevantes, se realizaron los entrenamientos con los subconjuntos pre y post de la base de datos de intervención (alumnos de sexto básico).

5.2.1 Comparativas de desempeño en pre test

En la fase pre test del estudio, los estudiantes se enfrentaron a la tarea sin haber recibido la intervención pedagógica. Los resultados obtenidos sobre este conjunto de datos arrojaron resultados que, si bien superan el azar, la poca separabilidad de los datos interfiere en lograr mejores resultados.

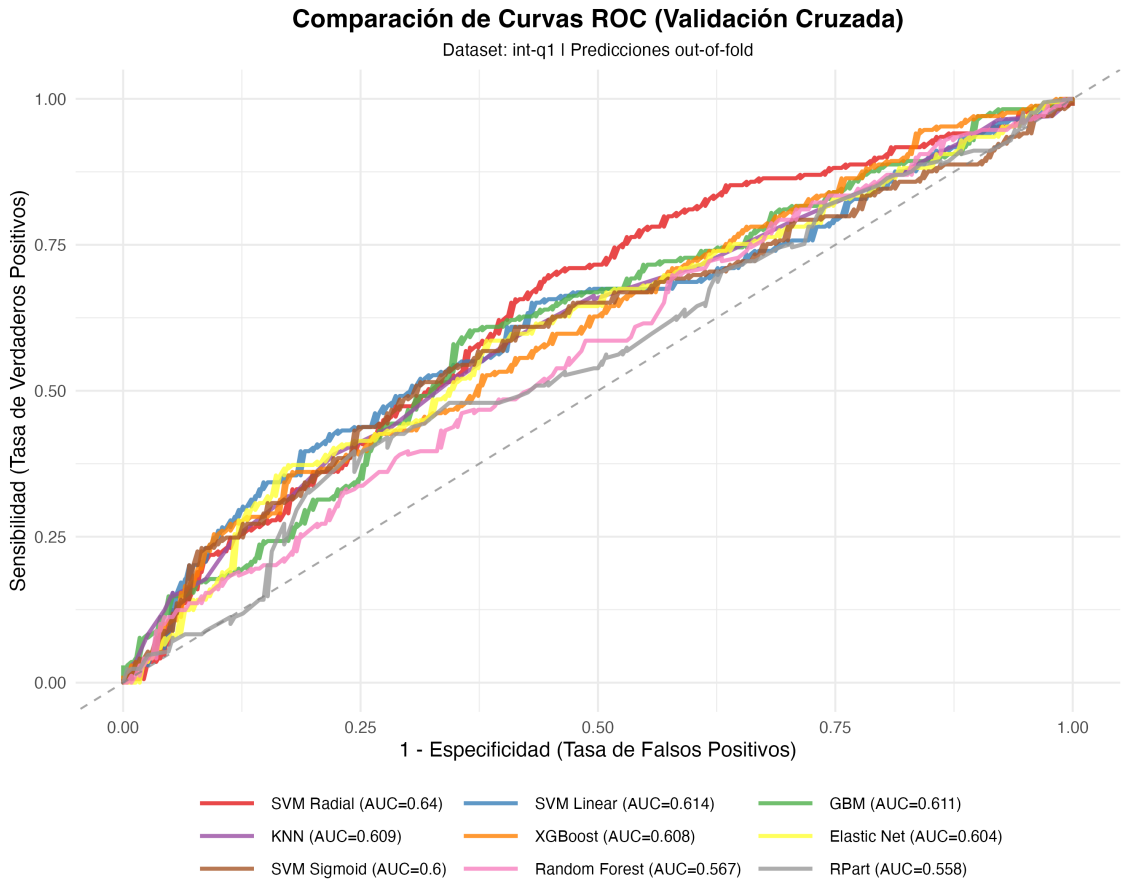

Figura 5.1: Área bajo la curva (ROC) en intervención pre test

Fuente: Elaboración propia

La Tabla 5.3 muestra que el mejor modelo obtuvo 0.64 en AUC y una sensibilidad de 0.65. Además, al hacer una comparación visual entre los modelos, utilizando la Figura 5.1, se puede observar que no existe una convergencia entre las métricas de los demás modelos. Este comportamiento sugiere que, en esta etapa inicial

del estudio, la señal de dificultad está mezclada con el ruido conductual de los alumnos. Por ejemplo, durante el pre test un estudiante puede mostrar latencias altas (flightTimeAvg) o pausas (pauseCount) por distintas razones indistinguibles para el modelo:

1. Dificultad Intrínseca: La tarea realmente representa un desafío para el estudiante
2. Dificultad Instrumental: El estudiante en etapas tempranas no domina la interfaz o carece de estratégias básicas de búsqueda.

Tabla 5.3: Comparación de clasificadores

Fuente: Elaboración propia

Clasificador	Accuracy	Kappa	AUC	Sensibilidad	Especificidad	Precisión	F1
SVM Radial	**0.617**	**0.237**	**0.640**	**0.657**	0.587	0.539	**0.592**
SVM Linear	0.586	0.167	0.614	0.574	0.596	0.511	0.540
GBM	0.604	0.206	0.611	0.615	0.596	0.528	0.568
KNN	0.594	0.169	0.609	0.521	**0.648**	0.521	0.521
XGBoost	0.561	0.124	0.608	0.580	0.548	0.485	0.528
Elastic Net	0.581	0.163	0.604	0.598	0.570	0.505	0.547
SVM Sigmoid	0.591	0.177	0.600	0.580	0.600	0.516	0.546
Random Forest	0.551	0.085	0.567	0.485	0.600	0.471	0.478
RPart	0.531	0.060	0.558	0.527	0.535	0.454	0.488

Finalmente, es notable destacar el bajo rendimiento de modelos simples, como, por ejemplo, Random Forest y RPart. Esto indica que no existen reglas de decisión simples para separar el conjunto de datos.

5.2.2 Comparativas de desempeño en post test

Luego, se realizó este mismo ejercicio con los datos posteriores a la intervención. Con estos datos, el rendimiento de la mayoría de los modelos alcanzó valores aproximados al 71 % en AUC lo cual puede observarse en la Figura 5.2. Este incremento debe

interpretarse no como una mejora en los algoritmos, como por ejemplo al extender la grilla de hiperparámetros (ya que se utilizó la misma para ambos conjuntos), sino como una mejora en la calidad de los datos.

En la Figura 4.11 de la sección anterior, podemos notar que existió un cambio conductual en los estudiantes luego de someterse a la intervención. Esto podría haber generado un efecto en como el estudiante expresa la carga cognitiva en sus dinámicas de tipeo. Por ejemplo, si un estudiante vacila con el ritmo de escritura en el post test, es mucho mas probable que se deba a un efecto de la dificultad intrínseca de la tarea y no por motivos de experiencia o desconocimiento de estratégias de búsqueda.

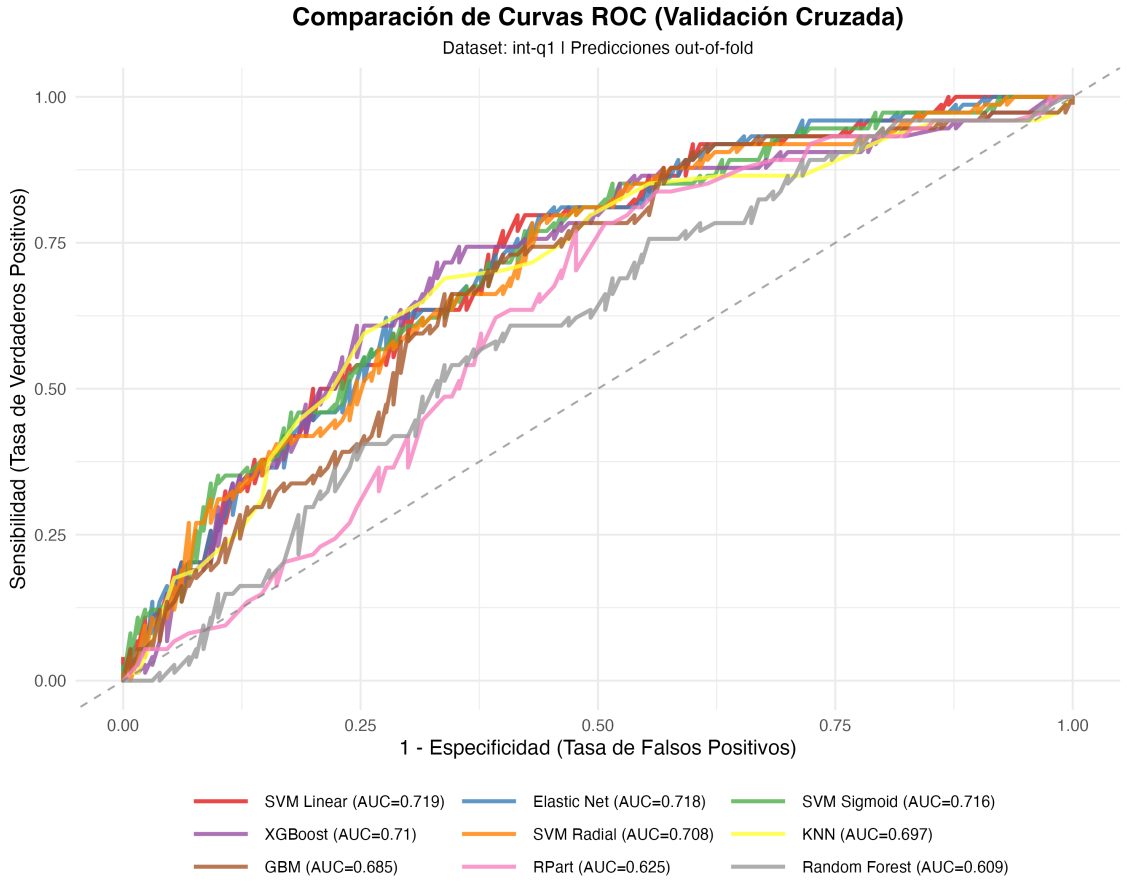

Figura 5.2: Área bajo la curva (ROC) en intervención post test

Fuente: Elaboración propia

Tabla 5.4: Desempeño de clasificadores (validación cruzada)

Fuente: Elaboración propia

Clasificador	Accuracy	Kappa	AUC	Sens.	Espec.	Precisión	F1
SVM Linear	0.662	0.299	**0.719**	0.635	0.677	0.528	0.577
Elastic Net	**0.686**	0.337	0.718	0.622	**0.723**	**0.561**	0.590
SVM Sigmoid	0.652	0.279	0.716	0.622	0.669	0.517	0.564
XGBoost	**0.686**	**0.362**	0.710	**0.716**	0.669	0.552	**0.624**
SVM Radial	0.657	0.287	0.708	0.622	0.677	0.523	0.568
KNN	0.672	0.319	0.697	0.649	0.685	0.539	0.589
GBM	0.642	0.290	0.685	**0.716**	0.600	0.505	0.592
RPart	0.593	0.188	0.625	0.635	0.569	0.456	0.531
Random Forest	0.608	0.185	0.609	0.554	0.638	0.466	0.506

Puesto que la estructura interna de este conjunto de datos permite modelar la dificultad percibida de la tarea en una manera más estable, se analizaron los modelos en función de las distintas métricas obtenidas durante el proceso de entrenamiento. La tabla 5.4 permite seleccionar y descartar los modelos más robustos para modelar el fenómeno.

- **Modelos seleccionados**: Los modelos basados en hiperplanos (SVM) mostraron consistentemente los mejores resultados. El mejor de estos (SVM Lineal) alcanzó un AUC de 0.72, superando a los modelos basados en árboles, tales como Random Forest (AUC $\approx$ 0.57) y árboles de decisión simples (AUC $\approx$ 0.56). Sin embargo, se puede observar que XGBoost es igual de competitivo en términos de AUC y tiene un mayor nivel de sensibilidad (AUC $\approx$ 0.71). Este factor hace que este modelo predomine sobre aquellos con mayor AUC, ya que el coste de cometer falsos positivos, es decir, clasificar una tarea difícil como fácil, es superior a indicar una tarea fácil como difícil. Se puede tener una mirada más detallada al observar la Figura 5.3, en donde el modelo clasificó correctamente un 71.6 % de las observaciones en las que un estudiante indicó a priori que consideraba que la tarea era difícil. Mientras que un 28.4 % del total fue donde el modelo clasificó erroneamente esta tarea como fácil.

- **Modelos descartados**: A partir de los resultados obtenidos, se puede inferir que los

modelos descartados son aquellos que obtuvieron un AUC menor al 70 %. Se hace especial énfasis en Random Forest, puesto que este presentó valores bajos en AUC (AUC $\approx$ 0.60) y en sensibilidad (AUC $\approx$ 0.55). Esto sugiere que los arboles de decisión, en esta base de datos, tienden a sobreajustarse al ruido.

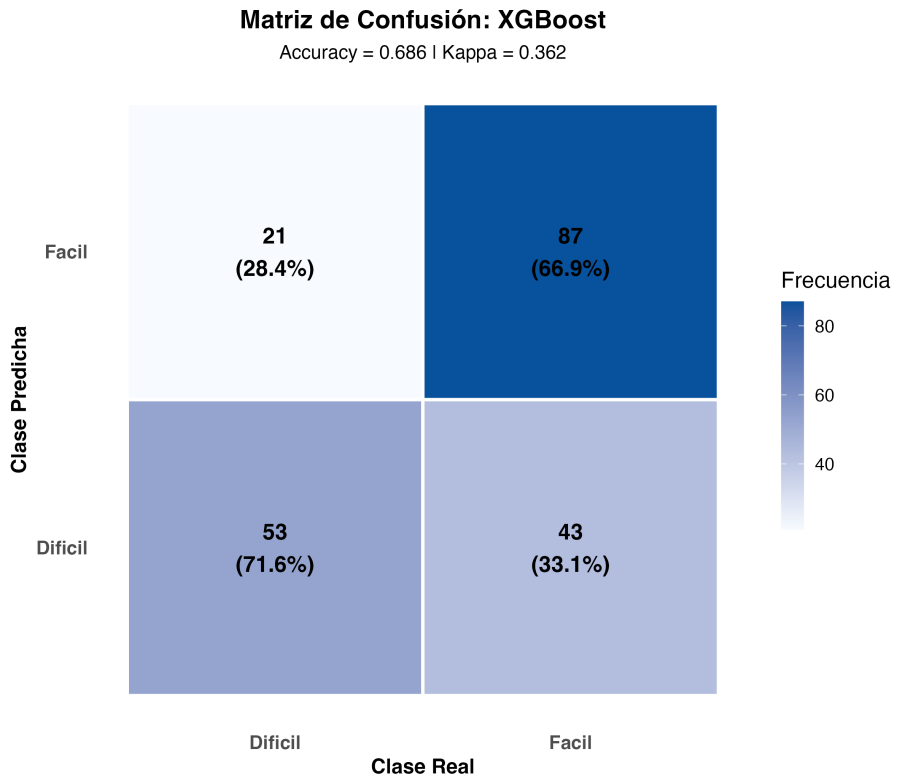

Figura 5.3: Matriz de confusión de modelo XGBoost

Fuente: Elaboración propia

5.3 PRUEBAS EN DIAGNÓSTICO

Análogamente, se realizó el mismo proceso de modelado para la base de datos de diagnóstico. Un aspecto a recordar es que este estudio se realizó en distintos niveles escolares (niños desde cuarto básico hasta octavo básico); por tanto, las técnicas de búsqueda, así como la fluidez de escritura, podrían variar entre los distintos grupos etarios.

Al realizar el entrenamiento de modelos, se evidenció en las Figuras 5.4 y en la tabla 5.5 que KNN obtuvo el mejor resultado con AUC en 0.58. Estos resultados muestran que la señal de la dificultad percibida podría estar contaminada por otros factores, como, por ejemplo, las distintas estrategias de búsqueda según el grupo etario y/o los distintos niveles de destreza al momento de utilizar el teclado al abordar la tarea. Esta cualidad dificultaría la separabilidad de los datos en la base de datos de diagnóstico

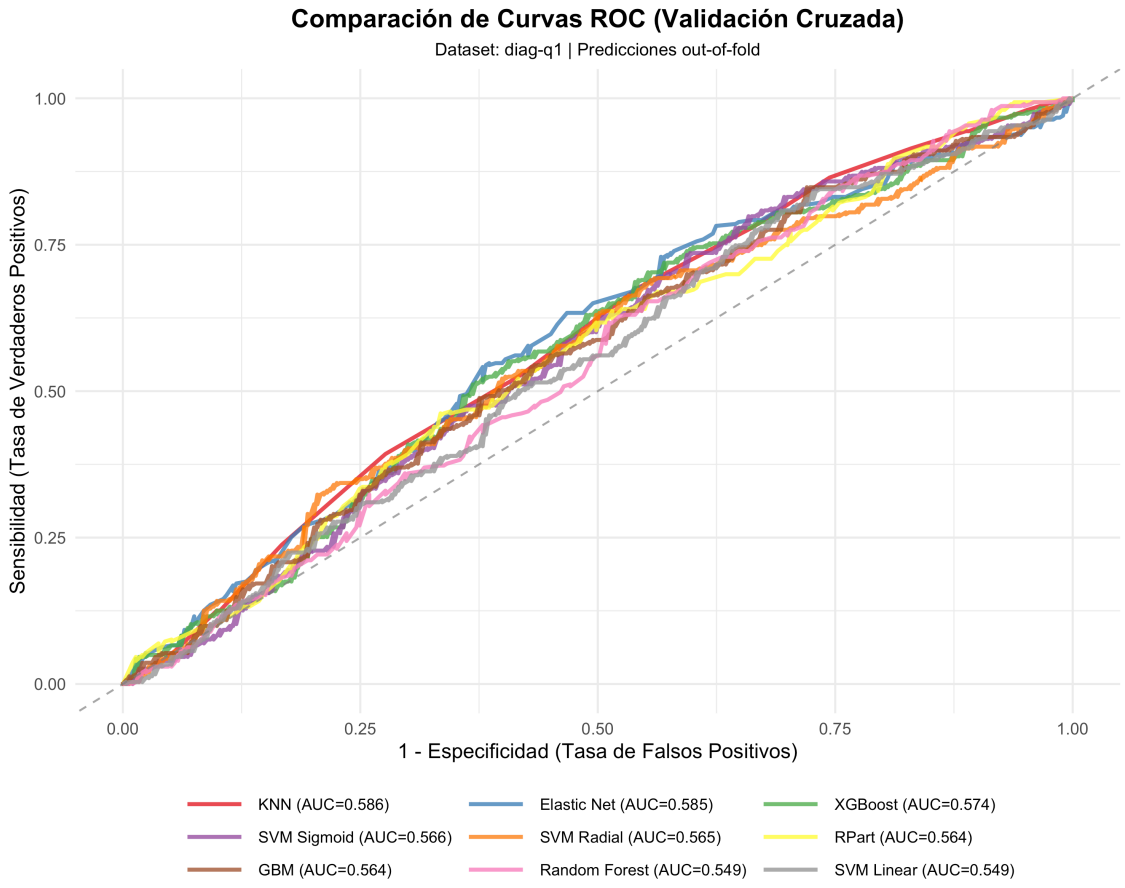

Figura 5.4: Área bajo la curva (ROC) en diagnóstico

Fuente: Elaboración propia

Tabla 5.5: Métricas de clasificación para el dataset diag-q1 (CV 10-fold, 3 repeticiones)

Clasificador	Accuracy	Kappa	AUC	Sensibilidad	Especificidad	Precisión	F1
KNN	0.554	0.108	0.586	0.525	0.584	0.566	0.545
Elastic Net	**0.579**	**0.157**	**0.585**	0.614	0.543	**0.581**	0.597
XGBoost	0.569	0.135	0.574	**0.640**	0.495	0.567	**0.602**
SVM Sigmoid	0.542	0.085	0.566	0.521	0.563	0.552	0.537
SVM Radial	0.552	0.104	0.565	0.551	0.553	0.560	0.556
GBM	0.554	0.107	0.564	0.558	0.549	0.561	0.560
RPart	0.547	0.095	0.564	0.521	**0.573**	0.558	0.539
SVM Linear	0.540	0.082	0.549	0.508	**0.573**	0.552	0.529
Random Forest	0.545	0.089	0.549	0.597	0.491	0.548	0.572

*Se presenta en color amarillo el modelo seleccionado

El bajo rendimiento observado en los modelos predictivos evidencia la presencia de ruido estructural en el conjunto de datos, lo que dificulta una clasificación binaria efectiva. Para descubrir el origen de esta variabilidad, se realizó un análisis estadístico estratificado por grupo etario para determinar si estos distintos grupos son un factor en términos de rendimiento para el modelo.

Los resultados publicados en el Anexo E evidencian la ausencia de variables comunes entre los distintos grupos etarios. Esto podría indicar la existencia de distintas estrategias de búsqueda dependiendo de la edad del estudiante; lo cual, al agrupar estos comportamientos en un único conjunto de datos de entrenamiento, hace que las señales especificas de cada grupo formen estructuras homogéneas indistinguibles para el modelo.

Otro camino para comprobar dicho efecto es el re-entrenamiento de los modelos estratificado por grupo etario; no obstante, al particionar el conjunto de datos original, los grupos estratificados quedan con alrededor de 30 observaciones cada uno, considerando el submuestreo aplicado debido al desbalance existente (a excepción de las observaciones de 4to básico, que son alrededor de 200, Tabla 5.6). Esta baja cantidad de observaciones puede provocar sobreajuste en los modelos predictivos (Ying, 2019); por tanto, se descarta esta estrategia.

Tabla 5.6: Distribución de clases de dificultad percibida por conjunto de datos

Fuente: Elaboración propia

Grupo Escolar	Difícil	Fácil	Total	Desbalance (%)
4° Básico	**141**	129	270	4.4
5° Básico	**33**	23	56	17.9
6° Básico	31	**62**	93	33.4
7° Básico	**50**	35	85	17.6
8° Básico	**28**	26	54	3.8

Nota: La clase mayoritaria en cada grupo se indica en negrita, correspondiente a la categoría susceptible de ser nivelada mediante técnicas de subsampleo.

5.4 IMPORTANCIA DE CARACTERÍSTICAS

Una vez entrenados los modelos y analizados sus rendimientos, es importante identificar que características contribuyeron a los resultados obtenidos. Es por esto que se utilizó la metodología de importancia de las características mediante permutación. Este método mezcla los valores de una característica con los valores de la variable dependiente con el objetivo de que esta pierda la correlación. Luego, se entrena el modelo con esta variable barajada y se compara el resultado con las características originales.

Las Figuras 5.5 y 5.6 muestran los resultados obtenidos al aplicar esta metodología al modelo con mejor rendimiento (XGBoost). Se indetificó como importante únicamente la variable flightTimeAvg. Este resultado indica que esta variable se consolida como el predictor dominante. El tiempo de vuelo, en particular, refleja el tiempo transcurrido entre la presión de dos teclas distintas; por tanto, este tiempo reflejaría la pausa mental necesaria para procesar información antes de ejecutar la siguiente acción motora.

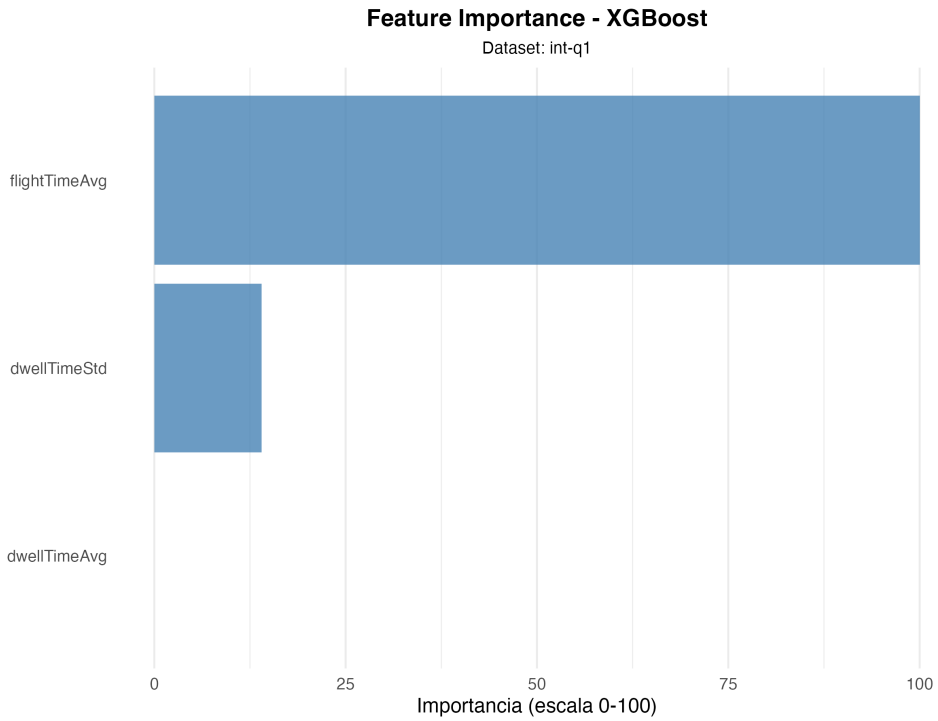

Figura 5.5: Importancia de características en pre test

Fuente: Elaboración propia

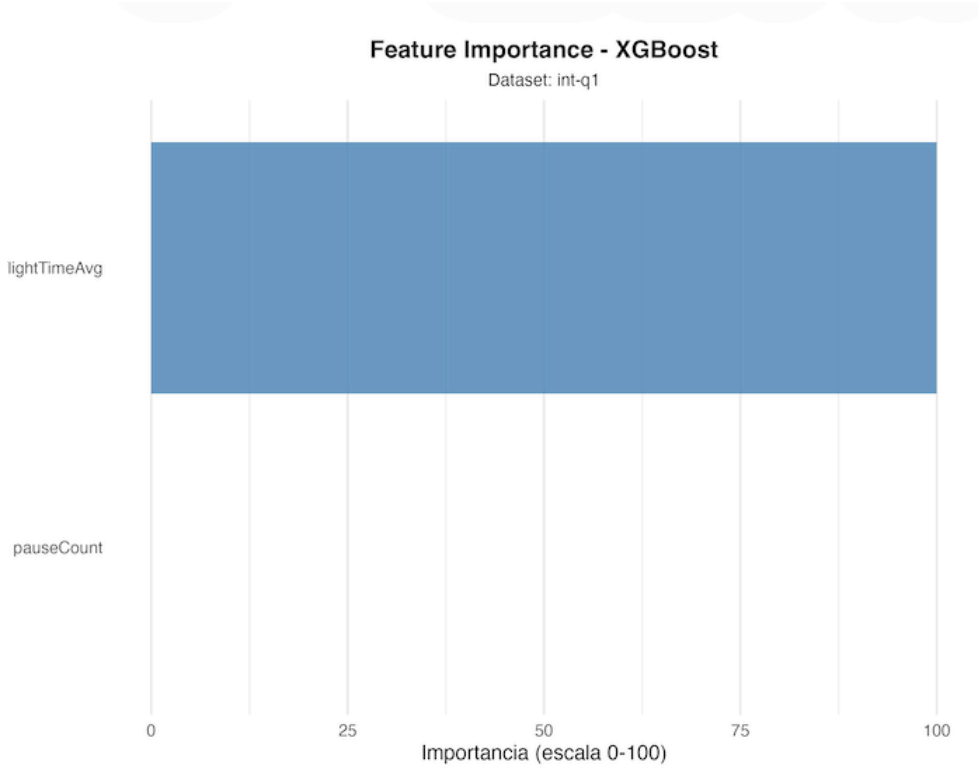

Figura 5.6: Importancia de características en post test

Fuente: Elaboración propia

5.5 RESUMEN

El análisis comparativo entre los modelos generados permite establecer tres conclusiones fundamentales:

5.5.1 Viabilidad técnica

Es posible predecir la dificultad percibida ante una tarea de búsqueda superando el azar. Dentro de los modelos con mejor rendimiento se encuentran las máquinas de soporte vectorial (SVM) y XGboost. Se identificó que estos modelos ofrecen un mejor equilibrio entre sensibilidad y especificidad, especialmente en la fase post-test

($AUC \approx 0{,}72$). Además, se observa que la variable de tiempo de vuelo (flightTimeAvg) y la variabilidad los tiempos de permanencia (dwellTimeStd) son los indicadores mas fuertes para la predicción.

5.5.2 Dependencia del nivel de competencia

La capacidad predictiva de los modelos esta estrechamente relacionada con la barrera instrumental inicial del experimento. La falta de familiaridad con la plataforma NEURONE o una experiencia previa limitada en herramientas de búsqueda digital pudieron haber generado patrones de escritura con ruido sistemático, lo cual pudo haber afectado el rendimiento de los modelos predictivos.

5.5.3 Prioridad en la detección

A raíz de la importancia de detectar verdaderos positivos, el modelo seleccionado debe estar ligado a su rendimiento, utilizando la sensibilidad como métrica a comparar, incluso si esto implica una menor precisión global.

CAPÍTULO 6. DISCUSIÓN

Durante el desarrollo de este trabajo, se pudo observar cuantitativamente el comportamiento de los estudiantes al momento de abordar una tarea de búsqueda de información. El presente capítulo tiene como propósito interpretar y contextualizar estos hallazgos, referenciando a los autores presentados en el marco teórico.

6.1 DINÁMICAS DE TIPEO COMO PROXY DE LA DIFICULTAD PERCIBIDA

El objetivo teórico de este estudio estuvo orientado a validar si un fenómeno no observable, en particular la dificultad percibida ante una tarea de búsqueda web, podría inferirse a través de un comportamiento observable, como en este caso las dinámicas de tipeo. Los resultados obtenidos validan esta premisa, puesto que la selección de características realizada identificó constantemente el tiempo de vuelo (flightTimeAvg) y la variabilidad de la precisión (dwellTimeStd) como los predictores mas fuertes de la dificultad percibida en los modelos en los que se obtuvieron los mejores resultados.

Esto podría asociarse directamente con la teoría de la carga cognitiva propuesta por Sweller, mencionada en el marco teórico. Cuando un estudiante percibe como dificil una tarea, esta satura la memoria de trabajo y compite por los recursos mentales disponibles para la planificación motora. Este efecto puede observarse en los retrasos capturados por el teclado, los cuales no estarían relacionados meramente con un retraso mecánico en la escritura, sino que reflejarían un cuello de botella cognitivo.

6.2 EL EFECTO DE LA INTERVENCIÓN

Uno de los hallazgos mas importantes de este estudio fue descubrir un aumento en las métricas de predicción de los modelos desde las fases de pre test (AUC $\approx 0{,}64$) hacia post test (AUC $\approx 0{,}72$).

En el pre test, los estudiantes, teóricamente, pudieron verse enfrentados a una alta carga cognitiva extrínseca, puesto que no solamente debían resolver las preguntas de investigación, sino que también se enfrentaban a una interfaz y a instrucciones propias de la metodología del experimento. Esto podría generar ruido en las métricas de dinámicas de tipeo.

La intervención posiblemente redujo esta carga extrinseca al enseñar nuevas competencias de alfabetización informacional (que son el conjunto de habilidades para acceder y evaluar la información); de esta manera, la barrera técnica se redujo, acentuando patrones mas cercanos a la dificultad percibida de la tarea.

Luego, en el post test, los alumnos ya tuvieron una experiencia inicial de búsqueda utilizando la plataforma; es por esto que las variables de dinámicas de tipeo comenzaron a reflejar la verdadera carga intrínseca de la tarea. En otras palabras, la señal se limpió, permitiendo que variables, como por ejemplo el tiempo de vuelo, pudieran representar de manera mas directa la dificultad percibida por el estudiante. Esto sugiere que la precisión técnica de las dinámicas de tipeo aumenta a medida que las habilidades digitales se estabilizan.

6.3 RESULTADOS CERCANOS AL AZAR EN DIAGNÓSTICO

Para predecir la dificultad percibida en estudiantes de cuarto a octavo básico a partir de dinámicas de tipeo, el rendimiento de los modelos puede verse afectado por múltiples factores que no siempre están directamente relacionados con la habilidad cognitiva o informacional que se desea modelar. A continuación, se describen los

principales aspectos que pueden interferir en la predicción

Por un lado, la habilidad mecanográfica es variable entre distintos grupos etarios. En particular Van Waes et al. (2021) demostró que la velocidad de tipeo sigue un desarrollo no lineal con la edad de la persona. La Figura 6.1 representa en el eje *X* la edad de un individuo, mientras que en el eje *Y* se muestra el intervalo entre pulsaciones de teclado (IKI) en mili-segundos. Cada punto es una observación individual y los colores representan los tipos de tareas ejecutadas en un estudio de mecanografía. Se puede observar que el efecto de la edad influye en la velocidad de las pulsaciones, donde disminuye desde la niñez hasta la adolescencia y la adultez temprana. Esta variabilidad podría ser erróneamente interpretada por el modelo como una señal de dificultad o de alta dificultad percibida, cuando en realidad es una limitación de destreza motora

Por otro lado, existen diferencias relacionadas con la atención visual entre los niños con menor destreza mecanográfica. Se ha demostrado que los usuarios con mayor experiencia mecanográfica, durante la producción textual, priorizan su atención en la pantalla para ver el texto escrito y miran el teclado solamente cuando es necesario. Por el contrario, los usuarios con menor experiencia mecanográfica centran su atención en las teclas a presionar para coordinar los movimientos motores de sus manos y ocasionalmente ven la pantalla para ver lo que han escrito (Feit et al., 2016; Gahshan-Haddad & Weintraub, 2023; Johansson et al., 2010; Logan & Crump, 2011).

En definitiva, los resultados obtenidos en la base de datos de diagnóstico sugieren que la predicción de la dificultad percibida a través de dinámicas de tipeo requiere de una estratificación por grupo etario debido a la variabilidad motora observada entre estos distintos niveles.

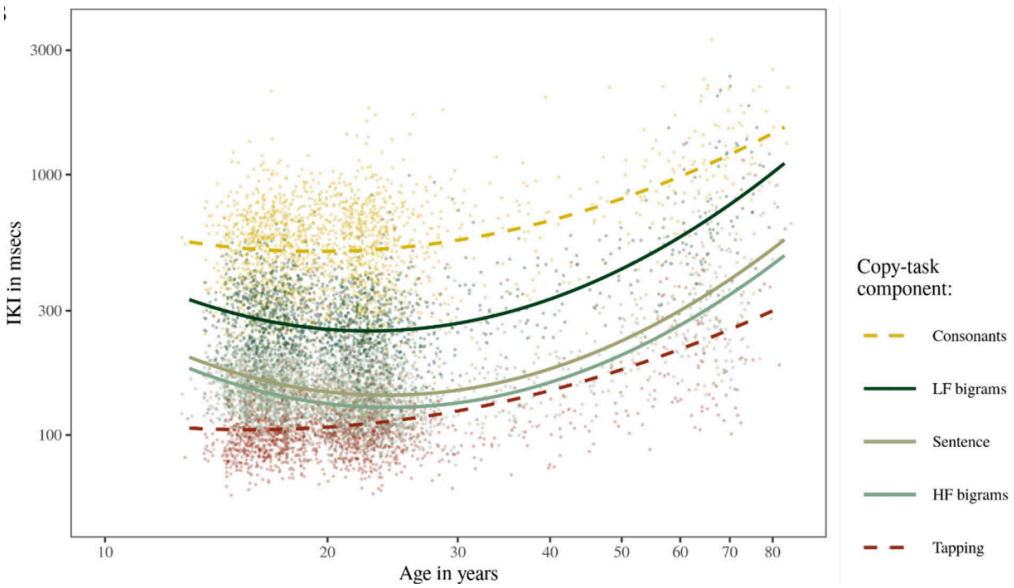

Figura 6.1: Variabilidad en el intervalo entre pulsaciones de teclado (IKI, Inter-Key Interval, en milisegundos) versus la edad

Fuente: Van Waes et al. (2021)

6.4 COMPARACIÓN CON EL ESTADO DEL ARTE

Finalmente, los resultados obtenidos son consistentes con los reportados en la literatura. Considerando que los niños poseen un desarrollo psicomotor en proceso y una menor capacidad de retención que los adultos, era esperable obtener resultados inferiores a los de Gacitúa y González (2022), donde se obtuvo 0.78 en AUC. No obstante, los resultados obtenidos en el post test fueron bastante cercanos (0.72 en AUC) a los reportados.

Es importante destacar que si bien estos resultados se sitúan en el rango de discriminación aceptable según los criterios de Hosmer Jr et al. (2013), no existen aplicaciones de este tipo de metodologías a problemáticas similares al de este estudio, por tanto, se sugiere cautela en la interpretación. Ante la falta de literatura comparativa específica, estos resultados no solo deben evaluarse por su suficiencia estadística, sino como una línea de base prometedora que demuestra la viabilidad técnica de predecir

estados cognitivos en niños a través de dinámicas de tipeo.

Por otro lado, se destaca el bajo rendimiento de los modelos simples, como, por ejemplo, los árboles de decisión. Esto sugiere que la relación entre la dificultad percibida y el movimiento motor no es lineal ni jerárquica.

6.5 RESUMEN

La conjugación entre la teoría, el análisis exploratorio y el modelado predictivo confirma que la técnica de dinámicas de tipeo puede ser utilizada como herramienta para predecir si un estudiante está teniendo problemas al abordar una tarea de búsqueda en la web.

CAPÍTULO 7. CONCLUSIONES

En el presente estudio se planteó el objetivo de predecir la dificultad percibida ante una tarea de búsqueda web utilizando técnicas no intrusivas basadas en las dinámicas de tipeo. En el presente capitulo se presentan las conclusiones pertinentes a los resultados obtenidos y expuestos en los capítulos anteriores de este escrito. A continuación se presentan las conclusiones respecto a los objetivos del proyecto, además de presentar una mirada hacia trabajos futuros, implicancias y limitaciones técnicas.

7.1 OBJETIVOS ESPECÍFICOS

O1: Aplicar los pipelines de procesamiento desarrollados por Brevis (2022) y Latapiatt (2022) para llevar a cabo el preprocesamiento de datos vinculados a dinámicas de tipeo de estudiantes considerando procesos de limpieza, transformación y normalización.

Este objetivo fue concluido en su totalidad. La primera fase del desarrollo de este escrito consistió en adaptar el flujo inicial de Brevis (2022), puesto que este ya utilizaba los datos de Diagnóstico. Si bien esto ayudo a disminuir notoriamente los tiempos iniciales de desarrollo. Además, en las etapas iniciales de procesamiento de datos se incorporó lógica para utilizar la base de datos de intervención, ya que, si bien esta fue generada a partir de un sistema similar al de NEURONE, existían ciertas diferencias en las tablas de los desafíos, lo que provocaba errores en la ejecución.

O2: Caracterizar las dinámicas de tipeo de estudiantes durante la formulación de consultas en tareas de búsqueda de información factual y evaluación crítica, identificando patrones asociados a distintos niveles de dificultad percibida.

La caracterización se llevó a cabo utilizando múltiples métodos analíticos, incluidos métodos estadísticos, regresión logística y métodos no supervisados. Los resultados permitieron identificar patrones consistentes en las dinámicas de tipeo.

- Se pudo observar la existencia de tres perfiles de estudiantes basados en sus dinámi-

cas de tipeo: «Alta fluidez de escritura», «Intermedio» y «Baja fluidez de escritura». Estos perfiles muestran que los alumnos poseen diferentes comportamientos al momento de abordar una tarea de búsqueda. En particular, se observó un aumento en la fluidez de escritura entre el pre y el post test y, además, según la teoría de la carga cognitiva de Sweller (1988), una aparente reducción en la carga cognitiva luego de que los estudiantes abordaran la intervención, debido a la adaptación con la herramienta; sin embargo, no es posible afirmar concretamente un aumento o disminución de la carga cognitiva, dado que la investigación de donde provienen los datos no fue diseñada con este propósito. Ahora bien, si se asume que teóricamente se observó una reducción de la carga cognitiva, este efecto podría incluso no estar relacionado con el efecto experimental de la intervención, sino más bien asociarse a que, en algunos casos, los estudiantes, al incorporar nuevas herramientas de búsqueda, así como el efecto de práctica y familiarización con la herramienta, pudieron verse reflejados en el aumento de la fluidez de escritura. Por lo tanto, se sugiere que, en estudios futuros, se analice este fenómeno con mayor profundidad.

- Se observó en la base de datos de diagnóstico que existen diferencias significativas entre los distintos grupos de edad y las variables de dinámica de tipeo en los estudiantes con baja y alta percepción de la dificultad de la tarea. Esto indicaría que la edad es un factor relevante al momento de generalizar las estrategias de búsqueda de los estudiantes. Esta baja convergencia resultó en un bajo poder predictivo en los modelos creados para el conjunto de datos de diagnóstico.

Estos hallazgos argumentan a favor de la evidencia sobre la utilidad de las métricas de dinámicas de tipeo para predecir la dificultad percibida en tareas de búsqueda de información.

O3: Construir modelos de clasificación de dificultad percibida a partir de las dinámicas de tipeo usando técnicas de aprendizaje automático del estado del arte.

Se generaron diversos modelos de aprendizaje automático. Particularmente, se utilizaron los modelos seleccionados en el estudio de Gacitúa y González (2022) sobre la predicción de la dificultad percibida en adultos, considerando buenas prácticas de

programación, como, por ejemplo, el control de fuga de información y el balance de clases para la base de datos de diagnóstico.

O4: Comparar los modelos a partir de métricas tales como AUC, F1-Score y sensibilidad.

Se compararon todos los modelos generados utilizando métricas de AUC, F1-Score y sensibilidad sobre los datos de intervención y diagnóstico. El modelo con mejor rendimiento en los datos de intervención fue Support Vector Machine con kernel sigmoide (AUC = 0.712), mientras que para el diagnóstico fue XGBoost (AUC = 0.57).

7.2 OBJETIVO GENERAL

El objetivo general planteado para este trabajo es el siguiente:

O1: Construir un modelo de clasificación de la dificultad percibida por escolares de 4° a 8° básico durante tareas de búsqueda de información, mediante el análisis de sus dinámicas de tipeo (keystroke dynamics) al formular consultas, con el fin de contar con indicadores que permitan brindar asistencia oportuna en el proceso de enseñanza aprendizaje de competencias de investigación en línea.

Para abordar este objetivo, se utilizó la metodología OSEMN, en la que se modificó un flujo de trabajo ya existente para acelerar las etapas tempranas de obtención de datos. Luego, estos datos fueron procesados y se realizaron diversos análisis exploratorios y estadísticos para caracterizar el conjunto de datos. Finalmente, se realizó la etapa de generación de modelos de clasificación utilizando el cluster de la Universidad de Santiago de Chile.

Si bien Los resultados obtenidos fueron inferiores a los obtenidos por el trabajo de Gacitúa y González (2022), ello puede explicarse debido a diversos factores, tales como la variabilidad entre las técnicas de búsqueda entre los estudiantes, así como la poca experiencia ante un buscador de contenido web, lo cual introduce variabilidad adicional a los resultados obtenidos. Sin embargo, los resultados fueron coherentes y siguen siendo útiles para futuros desarrollos de aplicaciones enfocadas en la ayuda oportuna en tareas de

búsqueda de información.

7.3 IMPLICANCIAS

A raíz de los resultados obtenidos, se plantean las implicancias teóricas y practicas para este trabajo:

7.3.1 Implicancias teóricas

- Este trabajo amplía el estado del arte sobre la predicción de la dificultad percibida ante una tarea de búsqueda en el contexto escolar, ya que durante la revisión de la literatura no se encontraron artículos relacionados.
- Se demuestra que es posible caracterizar a los estudiantes utilizando métricas de dinámicas de tipeo, donde se observaron distintos patrones durante la formulación de una consulta.
- El presente estudio puede reutilizarse como base teórica para futuras investigaciones en el contexto educativo.
- Bajo desempeño al entrenar con muestras que abarcan diferentes rangos de edad. Esto podría indicar que el desarrollo psicomotor de los estudiantes aún está en curso, lo que dificulta la capacidad de generalización del modelo y, en consecuencia, conduce a resultados deficientes.

7.3.2 Implicancias prácticas

- Este trabajo habilita la posibilidad de implementar herramientas de monitoreo en tiempo real que alerten cuando un estudiante requiera ayuda durante una sesión de

búsqueda de información.

- Dada la naturaleza poco intrusiva de la obtención de las métricas de dinámicas de tipeo, se habilita que esta técnica pueda ser implementada sin una inversión elevada de dinero en los establecimientos educacionales.

7.4 ALCANCES Y LIMITACIONES

- El presente estudio utiliza bases de datos provenientes de sistemas que no fueron explícitamente diseñados para la obtención de las métricas de dinámicas de tipeo, lo cual podría implicar cierto margen de error al momento de comparar los resultados con otros estudios.
- Los modelos de aprendizaje supervisado utilizaron datos obtenidos exclusivamente de dinámicas de tipeo, excluyendo información potencialmente relevante, como, por ejemplo, el historial de búsqueda y el contenido de la consulta.
- La baja cantidad de observaciones al estratificar por edad en la base de datos de diagnóstico imposibilitó realizar nuevos entrenamientos con la finalidad de corroborar el supuesto de que los estudiantes poseen distintas estrategias de búsqueda y que, por tanto, esto conduce a malas predicciones al agrupar todas las observaciones.

7.5 TRABAJO FUTURO

A partir de los resultados obtenidos durante la realización de este trabajo, se proponen las siguientes líneas de investigación:

- Modelar la problemática utilizando redes neuronales y comparar los resultados con los obtenidos en este estudio.

- Agregar información sobre comportamientos de búsqueda, como, por ejemplo, los movimientos del ratón dentro de una página para explorar distintos perfiles de usuarios.
- Explorar conjuntos de datos relacionados con otros niveles educativos y utilizar los resultados presentados en este estudio como linea base.
- Realizar estudios similares a Bilal (2000) en los que se capture el movimiento ocular de los alumnos.
- Desarrollar modelos de predicción de la dificultad percibida, estratificados por grupos de edad.
- Explorar técnicas avanzadas para el tratamiento del desbalance de la variable dependiente, dado que las pruebas iniciales mostraron que la ausencia de balanceo reduce el AUC en aproximadamente 3 puntos respecto al sub-muestreo. Con el fin de evitar la pérdida de información valiosa de la clase mayoritaria, se sugiere implementar la generación de muestras sintéticas mediante algoritmos como SMOTE o ADASYN.

7.6 REFLEXIONES FINALES

El presente trabajo constituye un camino hacia la predicción anticipada de la dificultad percibida por un estudiante en una tarea de búsqueda de información. Si bien los resultados pueden catalogarse como débiles en términos predictivos, estos muestran consistencia al compararse con los obtenidos por Gacitúa y González (2022), lo cual abre un camino hacia la investigación de herramientas predictivas de baja intrusión dentro de contextos escolares.

En este sentido, el presente trabajo aporta directamente a la problemática indicada en el marco teórico, habilitando nuevos caminos de investigación en contextos

escolares para mitigar los bajos niveles de alfabetización informacional reportados en el país.

REFERENCIAS BIBLIOGRÁFICAS

Agencia de Calidad de la Educación. (2020). *Informe nacional del estudio de alfabetización computacional y manejo de información* (inf. téc.). Asociación Internacional para la Evaluación del Logro Educativo. https://www.agenciaeducacion.cl/estudios-internacionales/icils/

Alnuaimi, A. F., & Albaldawi, T. H. (2024). An overview of machine learning classification techniques. *BIO Web of Conferences*, *97*, 133. https://www.bio-conferences.org/articles/bioconf/pdf/2024/16/bioconf_iscku2024_00133.pdf

American Library Association (ALA), A. f. C., & Libraries (ACRL), R. (2000). Information Literacy Competency Standards for Higher Education. http://hdl.handle.net/10150/105645

American Psychological Association. (2018). APA Dictionary of Psychology. https://dictionary.apa.org/

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. En *Psychology of learning and motivation* (pp. 89-195, Vol. 2). Elsevier.

Bilal, D. (2000). Children's use of the Yahooligans! Web search engine: I. Cognitive, physical, and affective behaviors on fact-based search tasks. *Journal of the American Society for Information Science*, *51*(7), 646-665. https://doi.org/10.1002/(sici)1097-4571(2000)51:7<646::aid-asi7>3.0.co;2-a

Bilal, D., & Gwizdka, J. (2018). Children's query types and reformulations in Google search. *Information Processing & Management*, *54*(6), 1022-1041. https://doi.org/10.1016/j.ipm.2018.06.008

Brevis, V. (2022). *Flujo de procesamiento para la caracterización de estudiantes según comportamientos de búsqueda y tipos de tarea* [Trabajo de titulación de pregrado]. Universidad de Santiago de Chile. https://usach.primo.exlibrisgroup.com/discovery/fulldisplay?docid=alma992071740806116&context=L&vid=56USACH_INST:REPOSITORIO&lang=es&adaptor=Local%20Search%20Engine&query=any,contains,brevis

Carbo, T. (1997). Mediacy: Knowledge and Skills to Navigate the Information Highway. *International Information & Library Review*, *29*(3-4), 393-401. https://doi.org/10.1080/10572317.1997.10762447

Chourio-Acevedo, L., & González-Ibañez, R. (2024). Evaluation of Online Inquiry Competencies of Chilean Elementary School Students: A Dataset. *Data*, *9*(7), 85. https://doi.org/10.3390/data9070085

Chourio-Acevedo, L., Köhler, J., Coscarelli, C., Gacitúa, D., Proaño-Ríos, V., & González-Ibáñez, R. (2024). Information literacy development and assessment at school level: a systematic review of the literature [Version Number: 1]. https://doi.org/10.48550/ARXIV.2404.19020

Çoklar, A. N., Yaman, N. D., & Yurdakul, I. K. (2017). Information literacy and digital nativity as determinants of online information search strategies. *Computers in Human Behavior*, *70*, 1-9. https://doi.org/10.1016/j.chb.2016.12.050

Conijn, R., Roeser, J., & Zaanen, M. v. (2019). Understanding the keystroke log: the effect of writing task on keystroke features. *Reading and Writing*, *32*, 2353-2374. https://doi.org/10.1007/s11145-019-09953-8

Crawford, H. (2010). Keystroke dynamics: Characteristics and opportunities. *2010 Eighth International Conference on Privacy, Security and Trust*, 205-212. https://doi.org/10.1109/PST.2010.5593258

Daniels, S. (2020). *Understanding Students' Typing Skills: Evaluating the Effects and Efficiency of a Typing Intervention for Elementary School Students* [Tesis doctoral, University of Tennessee]. https : / / trace . tennessee . edu / entities / publication / 6b578115-5488-4cfb-bce5-028aa53f7bdc

Druin, A., Foss, E., Hutchinson, H., Golub, E., & Hatley, L. (2010). Children's roles using keyword search interfaces at home. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 413-422. https://doi.org/10.1145/1753326.1753388

Duarte Torres, S., Hiemstra, D., & Serdyukov, P. (2010). An analysis of queries intended to search information for children. *Proceedings of the third symposium on Information interaction in context*, 235-244. https://doi.org/10.1145/1840784.1840819

Feit, A. M., Weir, D., & Oulasvirta, A. (2016). How we type: Movement strategies and performance in everyday typing. *Proceedings of the 2016 chi conference on human factors in computing systems*, 4262-4273. https://dl.acm.org/doi/10.1145/2858036.2858233

Gacitúa, D., & González, R. (2022). Predicting Search Task Difficulty through a Discrete-Time Action Log Representation on Spectrum Kernel. *Proceedings of*

the Association for Information Science and Technology, *59*(1), 690-692. https://doi.org/10.1002/pra2.692

Gahshan, N., & Weintraub, N. (2024). Effect of an Instructional Program for Word Processing and efficient typing on 'Year 4 students' composition. *Reading and Writing*. https://doi.org/10.1007/s11145-024-10613-9

Gahshan-Haddad, N., & Weintraub, N. (2023). Underlying functions associated with keyboarding performance of elementary-school students. *Scandinavian Journal of Occupational Therapy*, *30*(8), 1415-1423. https://doi.org/10.1080/11038128.2023.2188254

Gonzalez-Ibañez, R. (2016). NEURONE (oNlinE inqUiRy evaluatiON ecosystEm). https://informatica.usach.cl/proyecto/neurone-online-inquiry-evaluation-ecosystem/

Gwizdka, J. (2008). Revisiting search task difficulty: Behavioral and individual difference measures. *Proceedings of the American Society for Information Science and Technology*, *45*(1), 1-12. https://doi.org/10.1002/meet.2008.1450450249

Horton Jr, F. W. (1983). Information literacy vs. computer literacy. *Bulletin of the American Society for Information Science*, *9*(4), 14-16. https://eric.ed.gov/?id=EJ284227

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons. https://onlinelibrary.wiley.com/doi/book/10.1002/9781118548387

Huang, K., Chen, J., Liu, C., & Zhang, L. (2020). A Comparative Study of the Relationship between the Subjective Difficulty, Objective Difficulty of Search Tasks and Search Behaviors. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 421-424. https://doi.org/10.1145/3383583.3398614

Iapa, A.-C., & Cretu, V.-I. (2021). Modified Distance Metric That Generates Better Performance For The Authentication Algorithm Based On Free-Text Keystroke Dynamics. *2021 IEEE 15th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, 000455-000460. https://doi.org/10.1109/SACI51354.2021.9465601

Johansson, R., Wengelin, Å., Johansson, V., & Holmqvist, K. (2010). Looking at the keyboard or the monitor: relationship with text production processes. *Reading and writing*, *23*(7), 835-851. https://link.springer.com/article/10.1007/s11145-009-9189-3

Johnston, B., & Webber, S. (2003). Information Literacy in Higher Education: A review and case study. *Studies in Higher Education*, *28*(3), 335-352. https://doi.org/10.1080/03075070309295

Kammerer, Y., & Bohnacker, M. (2012). Children's web search with Google: the effectiveness of natural language queries. *Proceedings of the 11th International Conference on Interaction Design and Children*, 184-187. https://doi.org/10.1145/2307096.2307121

Latapiatt, J. (2022). *Pipeline de procesamiento para el análisis de comportamientos del usuario y su relación con la percepción de confianza en fuentes de información en las búsquedas en línea* [Trabajo de titulación de pregrado]. Universidad de Santiago de Chile. https://usach.primo.exlibrisgroup.com/discovery/collectionDiscovery?vid=56USACH_INST:REPOSITORIO&collectionId=8126208340006116

Logan, G. D., & Crump, M. J. (2011). Hierarchical control of cognitive processes: The case for skilled typewriting. En *Psychology of learning and motivation* (pp. 1-27, Vol. 54). Elsevier.

Mason, H., & Wiggins, C. (2010). A taxonomy of data science. *dataists. com*. https://introdatasci.dlilab.com/pdf/A_Taxonomy_of_Data_Science.pdf

MathWorks. (2025). Introducción a Support Vector Machines (SVM) - MATLAB & Simulink. https://la.mathworks.com/discovery/support-vector-machine.html

Mayer, R. E., Stull, A. T., Campbell, J., Almeroth, K., Bimber, B., Chun, D., & Knight, A. (2007). Overestimation Bias in Self-reported SAT Scores. *Educational Psychology Review*, *19*(4), 443-454. https://doi.org/10.1007/s10648-006-9034-z

Monrose, F., & Rubin, A. D. (2000). Keystroke dynamics as a biometric for authentication. *Future Generation Computer Systems*, *16*(4), 351-359. https://doi.org/https://doi.org/10.1016/S0167-739X(99)00059-X

Moskovitch, R., Feher, C., Messerman, A., Kirschnick, N., Mustafic, T., Camtepe, A., Lohlein, B., Heister, U., Moller, S., Rokach, L., & Elovici, Y. (2009). Identity theft, computers and behavioral biometrics. *2009 IEEE International Conference on Intelligence and Security Informatics*, 155-160. https://doi.org/10.1109/ISI.2009.5137288

Naik, M. M., & Padmini, I. (2014). Importance of information literacy. *International Journal of Digital Library Services*, *4*(3), 92-100. https://www.ijodls.in/uploads/3/6/0/3/3603729/9434.pdf

Révész, A., Kourtali, N., & Mazgutova, D. (2017). Effects of Task Complexity on L2 Writing Behaviors and Linguistic Complexity. *Language Learning*, *67*, 208-241. https://doi.org/10.1111/LANG.12205

Rokach, L., & Maimon, O. (2005). Decision trees. En *Data mining and knowledge discovery handbook* (pp. 165-192). https://link.springer.com/book/10.1007/978-0-387-09823-4

Seo, S. (2006). *A review and comparison of methods for detecting outliers in univariate data sets* [PhD Thesis]. University of Pittsburgh. https://d-scholarship.pitt.edu/concern/etds/3ffb48fc-81a3-44fd-9bfe-2fdcf844867f

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive science*, *12*(2), 257-285. https://onlinelibrary.wiley.com/doi/10.1207/s15516709cog1202_4

Tangirala, S. (2020). Evaluating the impact of GINI index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, *11*(2), 612-619.

Urry, J., & Edwards, J. (2024). A framework that explores the cognitive load of cs1 assignments using pausing behavior. *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, 1347-1353.

Van Waes, L., Leijten, M., Roeser, J., Olive, T., & Grabowski, J. (2021). Measuring and assessing typing skills in writing research. *Journal of Writing Research*, *13*(1), 107-153.

Vanderschantz, N., & Hinze, A. (2017). A Study of Children's Search Query Formulation Habits. *Electronic Workshops in Computing*. https://doi.org/10.14236/ewic/hci2017.7

Wan, H., Wang, H., Scotney, B., & Liu, J. (2019). A Novel Gaussian Mixture Model for Classification. *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 3298-3303. https://doi.org/10.1109/SMC.2019.8914215

Ying, X. (2019). An overview of overfitting and its solutions. *Journal of physics: Conference series*, *1168*, 022022. https://iopscience.iop.org/article/10.1088/1742-6596/1168/2/022022/meta

Zurkowski, P. G. (1974). The Information Service Environment Relationships and Priorities. Related Paper No. 5. https://eric.ed.gov/?id=eD100391

ANEXO A. DOCUMENTOS ADMINISTRATIVOS

A.0.1 Licencia

© **Álvaro Ignacio Valenzuela Alfaro** , **Roberto González Ibañez** 2025
● Algunos derechos reservados. Esta obra está bajo una Licencia Creative Commons Atribución–NoComercial–CompartirIgual 4.0 Internacional (CC BY-NC-SA 4.0). Sus condiciones de uso pueden ser revisadas en: https://creativecommons.org/licenses/by-nc-sa/4.0/.

A.0.2 Carta de compromiso

USACH

PATROCINIO TEMA DE TITULACIÓN
VESPERTINO

Estudiante (RUN, nombre, carrera):

19.111.889-8. Álvaro Ignacio Valenzuela Alfaro

Ingeniería Civil en Informática

Título tentativo:

PREDICCIÓN DE LA DIFICULTAD PERCIBIDA ANTE UNA TAREA DE BÚSQUEDA EN ESTUDIANTES DE ENSEÑANZA BÁSICA A PARTIR DEL ANÁLISIS DE DINÁMICAS DE TIPEO

Profesor patrocinador:

Roberto González Ibáñez

Tema propuesto por profesor guía

Declaro entender que el estudiante, para acceder a presentar su Propuesta de Tema de Titulación ante una Comisión Revisora, requiere mostrar avances parciales adecuados durante el semestre, y que la decisión del cumplimiento de este requisito recae en el(los) profesor(es) del curso de Seminario.

Fecha: 12 de Junio de 2025

Firma Estudiante

Firma Profesor

#SOMOSUSACH

Figura A.1: Carta de compromiso firmada

ANEXO B. DEFINICIONES TÉCNICAS

B.0.1 Diccionario de datos

Tabla B.1: Estructura de Queries

Atributo	Descripción	Formato
Query	Texto de consulta	Character
URL	URL donde ocurrió la búsqueda	Character
LocalTime	Tiempo local	TimeStamp
UserId	Identificador de usuario	Character
StudyId	Identificador de estudio	Character
ChallengeId	Identificador de desafío	Character
CreatedAt	Fecha de creación	TimeStamp

Fuente: Elaboración propia

Tabla B.2: Estructura de Questionnaires

Atributo	Descripción	Formato
Name	Nombre del cuestionario	Character
Type	Tipo de cuestionario	Character
questions.n._id	ID de la pregunta	Character
questions.n.question	Texto de la pregunta	Character
questions.n.number	Número de la pregunta	Integer
questions.n.options.n.option	Respuesta seleccionada	Character
createdAt	Fecha de creación	TimeStamp

Fuente: Elaboración propia

Tabla B.3: Estructura de Studies

Atributo	Descripción	Formato
Name	Nombre estudio	Character
Cooldown	Tiempo espera	Integer
Max_per_interval	Máx. desafíos	Integer
CreatedAt	Fecha creación	TimeStamp
Description	Descripción	Character
Image_url	Imagen portada	Character
Image_id	ID estudio	Character
Gm_code	Código NEURONE-GM	Character

Fuente: Elaboración propia

Tabla B.4: Estructura de UserChallenge

Atributo	Descripción	Formato
HintUsed	Uso pista	Boolean
UserId	Usuario	Character
ChallengeId	Desafío	Character
answers.n.answer	Respuesta	Character
answers.n.urls.n	URLs	Character
timeLeft	Tiempo restante	Integer
distance	Diferencia respuesta	Integer
createdAt	Fecha	TimeStamp
pointsObtained	Puntaje	Integer

Fuente: Elaboración propia

Tabla B.5: Estructura de UserDatas

Atributo	Descripción	Formato
Birthday	Fecha nacimiento	TimeStamp
Course	Curso	Integer
Institution	Establecimiento	Character
Institution_commune	Comuna	Character
Institution_region	Región	Character
CreatedAt	Registro	TimeStamp
Sex	Sexo	Character
Registered_via	Vía registro	Character

Fuente: Elaboración propia

Tabla B.6: Estructura de UserQuestionnaires

Atributo	Descripción	Formato
ChallengeId	Desafío	Character
UserId	Usuario	Character
Questionnaire	Cuestionario	Character
Type	Tipo	Character
Answers.n	Respuesta	Character
CreatedAt	Fecha	TimeStamp

Fuente: Elaboración propia

Tabla B.7: Estructura de Users

Atributo	Descripción	Formato
Confirmed	Correo confirmado	Character
CreatedAt	Registro	Character
Has_played	Vio playback	Boolean
StudyId	Estudio	Character
Gm_code	Código GM	Character
UserDataId	ID demográfico	Integer
Cooldown_start	Inicio intervalo	TimeStamp
Interval_answers	Máx respuestas	Character
Registered_via	Vía registro	Character

Fuente: Elaboración propia

Tabla B.8: Estructura de UserStudies

Atributo	Descripción	Formato
Assent	Aceptación consentimiento	Boolean
Finished	Finalizó	Boolean
UserId	Usuario	Character
StudyId	Estudio	Character
Challenges.n.*	Estados por desafío	Boolean / Character / TimeStamp
CreatedAt	Inicio estudio	TimeStamp
FinishedAt	Término estudio	TimeStamp

Fuente: Elaboración propia (2022)

Tabla B.9: Estructura de VisitedLinks

Atributo	Descripción	Formato
ChallengeId	Desafío	Character
Url	Página visitada	Character
State	Entrada o salida	Character
Title	Título página	Character
LocalTimeStamp	Hora cliente	Character
UserId	Usuario	Character
CreatedAt	Hora servidor	TimeStamp

Fuente: Elaboración propia

ANEXO C. RESULTADOS DIAGNÓSTICO

En este anexo se presentarán todos los resultados obtenidos en la ejecución del pipeline para la base de datos Diagnóstico en el primer segmento de consulta.

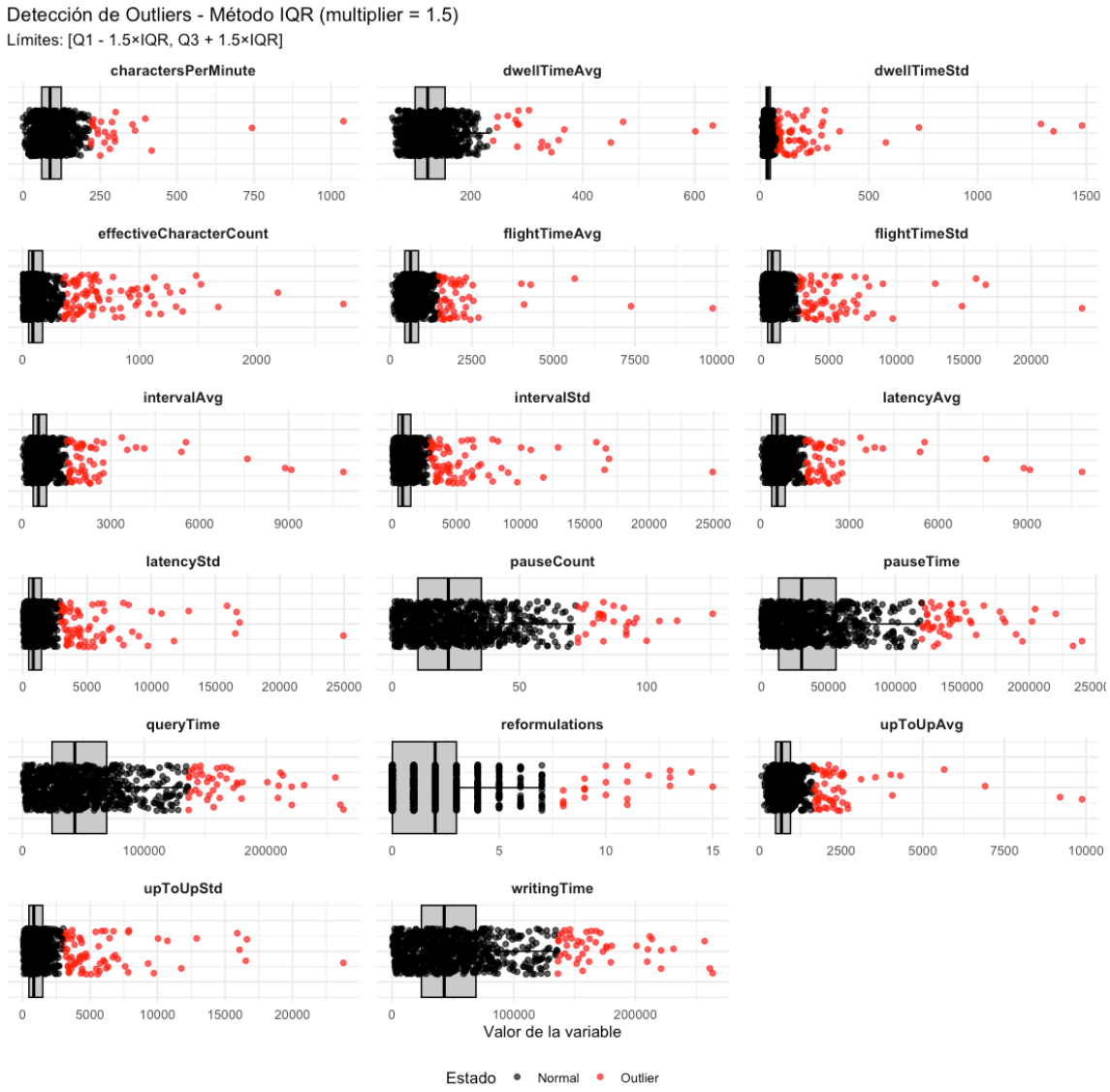

Figura C.1: Eliminación de valores atípicos, 1er segmento de consulta, diagnóstico.

Fuente: Elaboración propia

Tabla C.1: Estadísticas descriptivas de variables del conjunto completo de tipeo en diagnóstico.

Primer segmento de consulta

Tipo	Variable	Mínimo	1er. Cuartil	Mediana	Media	3er. Cuartil	Máximo
Segmento de query	qs	1.00	1.00	1.00	1.00	1.00	1.00
Comportamiento de búsqueda	effectiveCharacterCount	0.00	46.00	83.50	162.00	164.00	2742.00
	reformulations	0.00	0.00	1.00	1.97	3.00	15.00
	queryTime	4.00	18682.00	40014.00	49034.00	66640.00	263544.00
	writingTime	0.00	18567.00	39926.00	48888.00	66422.00	263438.00
	pauseTime	0.00	9682.00	26914.00	37244.00	52276.00	239619.00
	pauseCount	0.00	8.00	19.00	23.30	34.00	126.00
	charactersPerMinute	0.00	58.40	86.90	111.00	124.00	8571.00
	copypaste	0.00	0.00	0.00	0.23	0.00	1.00
Dinámica de tipeo	dwellTimeAvg	3.00	101.00	126.00	143.00	161.00	632.00
	dwellTimeStd	2.83	28.40	36.80	60.80	50.30	1480.00
	flightTimeAvg	10.50	431.00	611.00	741.00	846.00	9882.00
	flightTimeStd	2.12	465.00	786.00	1269.00	1380.00	23766.00
	latencyAvg	36.50	384.00	573.00	811.00	879.00	11495.00
	latencyStd	21.70	470.00	815.00	1360.00	1450.00	24945.00
	intervalAvg	36.50	384.00	573.00	811.00	879.00	11495.00
	intervalStd	21.70	470.00	815.00	1360.00	1450.00	24945.00
	upToUpAvg	2.00	480.00	665.00	848.00	949.00	11830.00
	upToUpStd	15.60	489.00	860.00	1387.00	1509.00	23773.00
Dificultad	dificulty_pre-test	1.00	3.00	4.00	3.57	5.00	6.00
	dificulty_post-test	1.00	2.00	3.00	3.36	5.00	6.00
Demografía	edad	9.00	10.00	10.00	10.70	12.00	16.00

Fuente: Elaboración propia

Tabla C.2: Conteo de valores nulos por variable

Variable	**Primer segmento de consulta**
effectiveCharacterCount	0
reformulations	0
queryTime	0
writingTime	0
pauseTime	0
pauseCount	0
charactersPerMinute	31
dwellTimeAvg	0
dwellTimeStd	34
latencyAvg	53
upToUpStd	67
edad	3
flightTimeAvg	31
flightTimeStd	58
intervalAvg	53
intervalStd	110
latencyStd	110
upToUpAvg	33
dificulty_pre-test	0
dificulty_post-test	0
copypaste	0
qs	0

Nota: "—" indica que la variable no está presente en ese dataset.

Fuente: Elaboración propia.

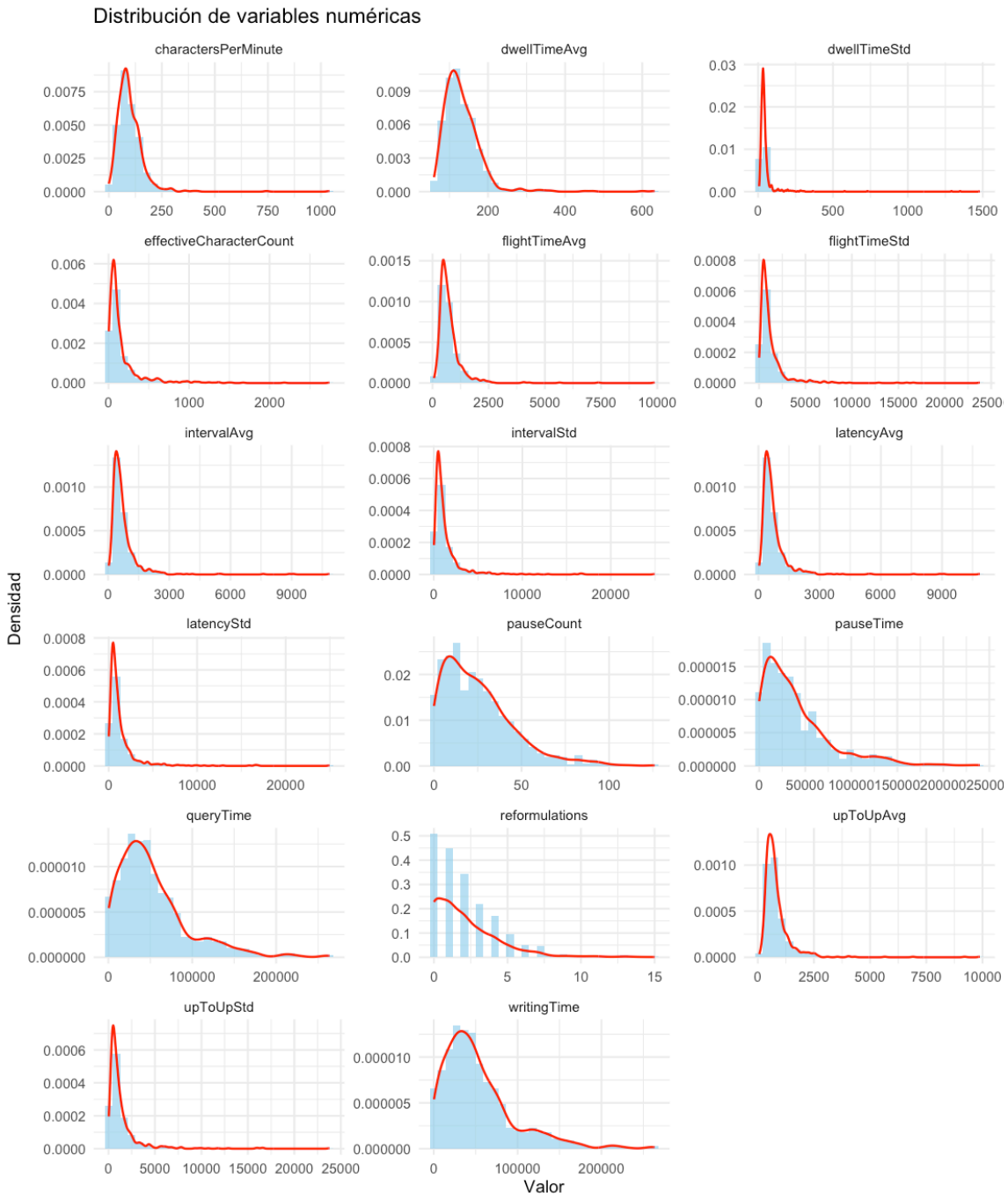

Figura C.2: Distribución de variables, primer segmento de consulta.

Fuente: Elaboración propia

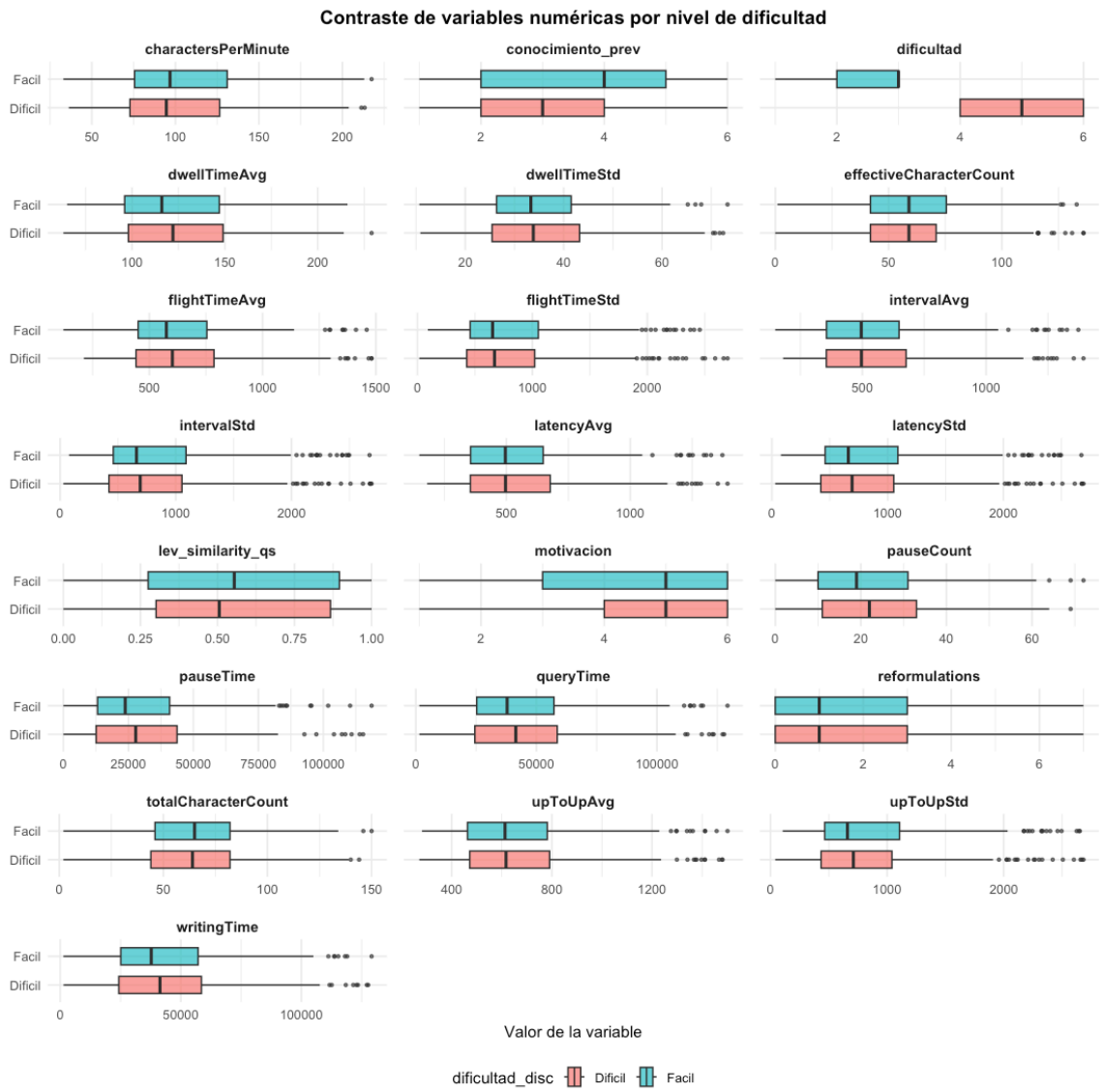

Figura C.3: Boxplot variables luego de eliminación de valores atípicos, primer segmento de consulta.

Fuente: Elaboración propia

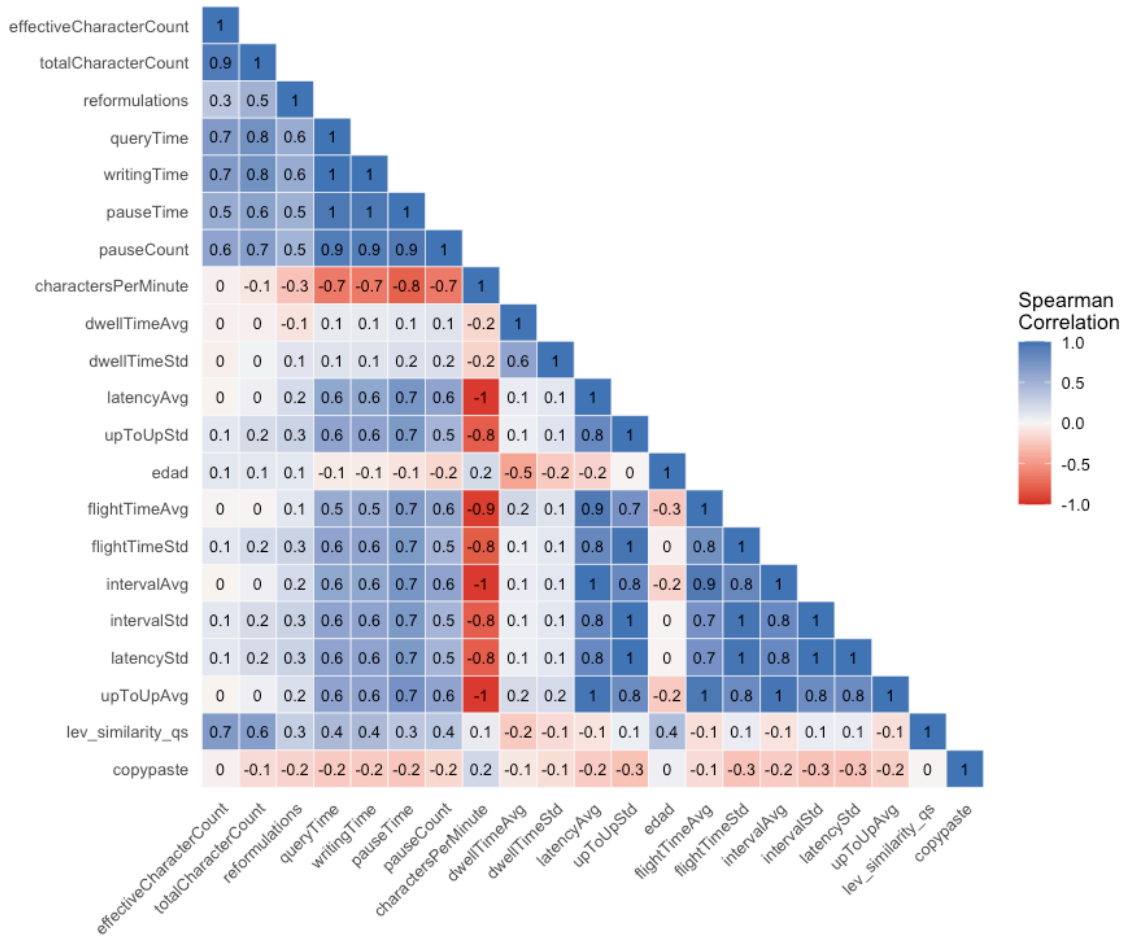

Figura C.4: Correlacion de spearman para características, primer segmento de consulta.

Fuente: Elaboración propia

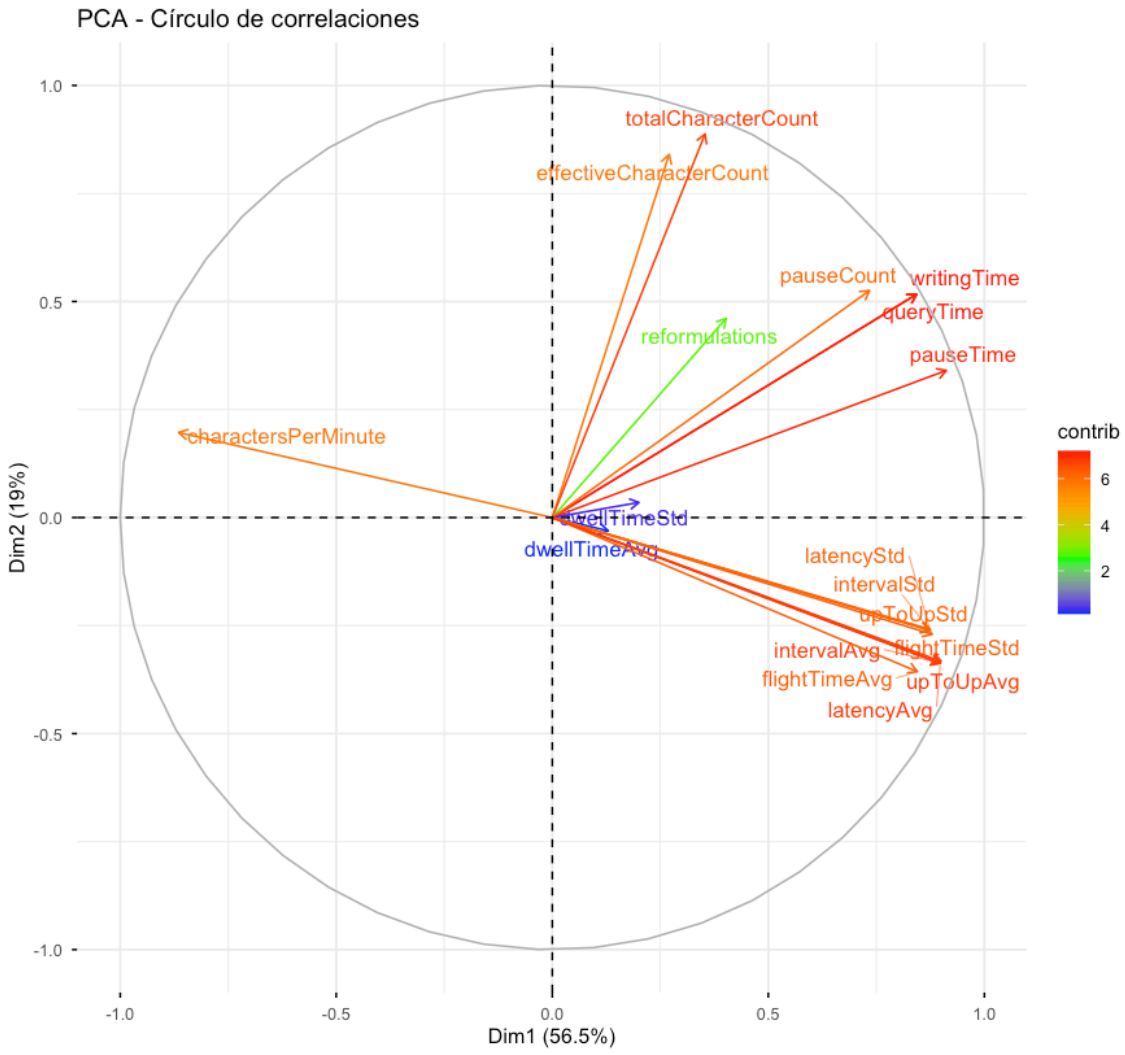

Figura C.5: Análisis componentes principales, primer segmento de consulta

Fuente: Elaboración propia

Tabla C.3: Pruebas de significancia entre grupos de dificultad percibida (Fácil vs Difícil).

Primer segmento de consulta.

Variable	W	p_value	método	Tamaño efecto
pauseCount	42564	0.202	wilcox	-0.060
dwellTimeAvg	42718	0.229	wilcox	-0.056
pauseTime	43324	0.359	wilcox	-0.043
flightTimeAvg	43554	0.418	wilcox	-0.038
writingTime	43784	0.483	wilcox	-0.033
queryTime	43792	0.485	wilcox	-0.033
charactersPerMinute	46401	0.600	wilcox	0.025
upToUpAvg	44290	0.642	wilcox	-0.022
dwellTimeStd	44320	0.652	wilcox	-0.021
latencyAvg	44326	0.654	wilcox	-0.021
intervalAvg	44326	0.654	wilcox	-0.021
intervalStd	45986	0.742	wilcox	0.016
latencyStd	45986	0.742	wilcox	0.016
upToUpStd	45934	0.760	wilcox	0.014
flightTimeStd	45709	0.842	wilcox	0.009
totalCharacterCount	44926	0.867	wilcox	-0.008
effectiveCharacterCount	45544	0.903	wilcox	0.006
lev_similarity_qs	45516	0.913	wilcox	0.005

Fuente: Elaboración propia

ANEXO D. TEST PAREADO PRE TEST Y POST TEST EN INTEVENCIÓN

A continuación se presentarán todos los resultados obtenidos en el test pareado bilateral entre el pre test y el post. Los análisis realizados se establecieron con un nivel de significancia de 0.05:

Tabla D.1: Resultados de pruebas pareadas Pre–Post (Grupo Control, $N = 25$)

Fuente: Elaboración propia

Variable	Método	N	Est.	p-value	Tamaño de efecto
pauseTime	t-test pareado	25	-2.73	**0,0116**	−0,508
queryTime	t-test pareado	25	-2.63	**0,0145**	−0,542
writingTime	t-test pareado	25	-2.63	**0,0146**	−0,541
charactersPerMinute	Wilcoxon pareado	25	253	**0,0155**	0,557
upToUpAvg	t-test pareado	25	-2.39	**0,0250**	−0,343
latencyAvg	Wilcoxon pareado	25	79	**0,0255**	−0,514
intervalAvg	Wilcoxon pareado	25	79	**0,0255**	−0,514
pauseCount	t-test pareado	25	-2.35	**0,0276**	−0,458
dwellTimeAvg	Wilcoxon pareado	25	81	**0,0293**	−0,502
dwellTimeStd	Wilcoxon pareado	25	84	**0,0358**	−0,483
intervalStd	Wilcoxon pareado	25	92	0.0596	−0,434
latencyStd	Wilcoxon pareado	25	92	0.0596	−0,434
upToUpStd	t-test pareado	25	-1.74	0.0953	−0,374
flightTimeAvg	t-test pareado	25	-1.67	0.107	−0,248
totalCharacterCount	Wilcoxon pareado	25	112	0.179	−0,311
flightTimeStd	t-test pareado	25	-1.37	0.183	−0,313
dificultad	Wilcoxon pareado	25	154.	0.388	−0,213
effectiveCharacterCount	t-test pareado	25	-0.814	0.424	−0,229

Nota: En negrita se indican las diferencias significativas al nivel $\alpha = 0{,}05$. El tamaño del efecto corresponde a Cliff's Delta para pruebas no paramétricas y a Cohen's d para pruebas paramétricas.

Tabla D.2: Resultados de pruebas pareadas Pre–Post (Grupo Tradicional, $N = 19$)

Fuente: Elaboración propia

Variable	**Método**	**N**	**Est.**	**p-value**	**Tamaño de efecto**
effectiveCharacterCount	t-test pareado	19	1.96	0.0661	0.510
dwellTimeAvg	t-test pareado	19	-1.63	0.120	-0.249
pauseCount	t-test pareado	19	-0.840	0.412	-0.143
flightTimeStd	t-test pareado	19	0.720	0.481	0.186
upToUpStd	t-test pareado	19	0.624	0.540	0.158
intervalStd	t-test pareado	19	0.596	0.558	0.153
latencyStd	t-test pareado	19	0.596	0.558	0.153
pauseTime	t-test pareado	19	-0.516	0.612	-0.0880
dwellTimeStd	Wilcoxon pareado	19	82.0	0.615	-0.137
latencyAvg	t-test pareado	19	-0.407	0.689	-0.0703
intervalAvg	t-test pareado	19	-0.407	0.689	-0.0703
upToUpAvg	t-test pareado	19	-0.320	0.753	-0.0511
queryTime	t-test pareado	19	-0.223	0.826	-0.0430
writingTime	t-test pareado	19	-0.220	0.828	-0.0423
charactersPerMinute	t-test pareado	19	-0.104	0.919	-0.0236
dificultad	Wilcoxon pareado	19	50.5	0.925	-0.0381
flightTimeAvg	t-test pareado	19	0.0592	0.953	0.0118
totalCharacterCount	t-test pareado	19	0.000754	0.999	0.000204

Nota: No se observaron diferencias estadísticamente significativas al nivel $\alpha = 0{,}05$. El tamaño del efecto corresponde a Cliff's Delta para pruebas no paramétricas y a Cohen's d para pruebas paramétricas.

Tabla D.3: Resultados de pruebas pareadas Pre–Post (Grupo Asistente Virtual, $N = 28$)

Fuente: Elaboración propia

Variable	**Método**	**N**	**Est.**	**p-value**	**Tamaño de efecto**
dwellTimeStd	Wilcoxon pareado	28	102	**0,0221**	−0,498
latencyAvg	t-test pareado	28	−2,10	**0,0448**	−0,308
intervalAvg	t-test pareado	28	−2,10	**0,0448**	−0,308
upToUpAvg	t-test pareado	28	−1,87	0.0717	−0,274
dwellTimeAvg	t-test pareado	28	−1,71	0.0995	−0,183
intervalStd	t-test pareado	28	−1,67	0.106	−0,378
latencyStd	t-test pareado	28	−1,67	0.106	−0,378
upToUpStd	t-test pareado	28	−1,63	0.116	−0,360
charactersPerMinute	t-test pareado	28	1.57	0.127	0,257
flightTimeStd	t-test pareado	28	−1,42	0.166	−0,308
dificultad	Wilcoxon pareado	28	139	0.210	−0,324
flightTimeAvg	t-test pareado	28	−1,24	0.224	−0,193
pauseTime	t-test pareado	28	−1,11	0.277	−0,199
queryTime	t-test pareado	28	−0,772	0.447	−0,140
writingTime	t-test pareado	28	−0,765	0.451	−0,139
pauseCount	t-test pareado	28	−0,572	0.572	−0,0985
totalCharacterCount	t-test pareado	28	−0,322	0.750	−0,0638
effectiveCharacterCount	t-test pareado	28	0.238	0.814	0,0516

Nota: En negrita se indican las diferencias significativas al nivel $\alpha = 0{,}05$. El tamaño del efecto corresponde a Cliff's Delta para pruebas no paramétricas y a Cohen's d para pruebas paramétricas.

Tabla D.4: Resultados de pruebas pareadas Pre–Post (Grupo Ludificado, $N = 30$)

Fuente: Elaboración propia

Variable	**Método**	**N**	**Est.**	**p-value**	**Tamaño de efecto**
upToUpStd	t-test pareado	30	$-2{,}57$	**0,0157**	$-0{,}555$
intervalStd	t-test pareado	30	$-2{,}49$	**0,0186**	$-0{,}546$
latencyStd	t-test pareado	30	$-2{,}49$	**0,0186**	$-0{,}546$
latencyAvg	t-test pareado	30	$-2{,}46$	**0,0201**	$-0{,}448$
intervalAvg	t-test pareado	30	$-2{,}46$	**0,0201**	$-0{,}448$
dwellTimeStd	t-test pareado	30	$-2{,}30$	**0,0286**	$-0{,}398$
upToUpAvg	t-test pareado	30	$-2{,}22$	**0,0341**	$-0{,}392$
flightTimeStd	t-test pareado	30	$-2{,}15$	**0,0396**	$-0{,}482$
pauseTime	t-test pareado	30	$-1{,}81$	0.0802	$-0{,}322$
queryTime	t-test pareado	30	$-1{,}74$	0.0926	$-0{,}303$
writingTime	t-test pareado	30	$-1{,}73$	0.0939	$-0{,}302$
charactersPerMinute	t-test pareado	30	1.68	0.103	$0{,}270$
pauseCount	Wilcoxon pareado	30	191	0.399	$-0{,}178$
flightTimeAvg	t-test pareado	30	$-0{,}723$	0.475	$-0{,}146$
dificultad	Wilcoxon pareado	30	147	0.476	$-0{,}162$
copypaste	Wilcoxon pareado	30	95	0.487	$-0{,}177$
totalCharacterCount	t-test pareado	30	$-0{,}614$	0.544	$-0{,}114$
dwellTimeAvg	t-test pareado	30	0.588	0.561	$0{,}0730$
effectiveCharacterCount	t-test pareado	30	0.466	0.645	$0{,}0888$

Nota: En negrita se indican las diferencias significativas al nivel $\alpha = 0{,}05$. El tamaño del efecto corresponde a Cohen's d para pruebas paramétricas y a Cliff's Delta para pruebas no paramétricas.

ANEXO E. ANÁLISIS ESTADÍSTICO DIAGNÓSTICO

A continuación se presentarán los resultados obtenidos al realizar un análisis estadístico estratificado entre los grupos etarios de la base de datos de diagnóstico.

Tabla E.1: Resultados de pruebas (4° Básico)

Fuente: Elaboración propia

Variable	**Método**	**Est.**	**p-value**	**Tamaño de efecto**
lev_similarity_qs	Wilcoxon	9691	0.191	-0.0886
pauseTime	Wilcoxon	9975	0.361	-0.0619
queryTime	Wilcoxon	10059	0.426	-0.0540
writingTime	Wilcoxon	10061	0.427	-0.0538
pauseCount	Wilcoxon	10080	0.442	-0.0521
charactersPerMinute	Wilcoxon	11134	0.488	0.0471
flightTimeAvg	Wilcoxon	10168	0.518	-0.0438
flightTimeStd	Wilcoxon	10278	0.622	-0.0334
latencyAvg	Wilcoxon	10298	0.642	-0.0316
intervalAvg	Wilcoxon	10298	0.642	-0.0316
upToUpAvg	Wilcoxon	10333	0.677	-0.0283
intervalStd	Wilcoxon	10336	0.680	-0.0280
latencyStd	Wilcoxon	10336	0.680	-0.0280
upToUpStd	Wilcoxon	10344	0.688	-0.0272
effectiveCharacterCount	Wilcoxon	10392	0.739	-0.0227
dwellTimeAvg	Wilcoxon	10446	0.795	-0.0177
totalCharacterCount	Wilcoxon	10482	0.835	-0.0142
dwellTimeStd	t-test	0.097	0.923	0.0113

Tabla E.2: Resultados de pruebas no pareadas Pre–Post (5° Básico)

Fuente: Elaboración propia

Variable	Método	Est.	p-value	Tamaño de efecto
pauseCount	Wilcoxon	317	**0.0224**	-0.341
effectiveCharacterCount	Wilcoxon	333	**0.0392**	-0.308
pauseTime	Wilcoxon	336	**0.0437**	-0.301
queryTime	Wilcoxon	339	**0.0482**	-0.295
writingTime	Wilcoxon	339	**0.0482**	-0.295
lev_similarity_qs	Wilcoxon	374	0.135	-0.223
charactersPerMinute	Wilcoxon	580	0.169	0.206
flightTimeAvg	t-test	-1.12	0.270	-0.288
latencyAvg	Wilcoxon	402	0.273	-0.164
intervalAvg	Wilcoxon	402	0.273	-0.164
upToUpAvg	Wilcoxon	405	0.292	-0.158
totalCharacterCount	t-test	-0.89	0.379	-0.229
dwellTimeStd	Wilcoxon	499	0.807	0.0374
flightTimeStd	Wilcoxon	464	0.818	-0.0353
intervalStd	Wilcoxon	465	0.829	-0.0333
latencyStd	Wilcoxon	465	0.829	-0.0333
dwellTimeAvg	Wilcoxon	496	0.845	0.0301
upToUpStd	Wilcoxon	470	0.883	-0.0229

Tabla E.3: Resultados de pruebas no pareadas Pre–Post (6° Básico)

Fuente: Elaboración propia

Variable	**Método**	**Est.**	**p-value**	**Tamaño de efecto**
lev_similarity_qs	Wilcoxon	1290	0.321	0.120
totalCharacterCount	Wilcoxon	1036	0.409	-0.100
dwellTimeStd	t-test	-0.699	0.488	-0.162
pauseCount	Wilcoxon	1084	0.625	-0.0595
writingTime	Wilcoxon	1091	0.664	-0.0530
pauseTime	Wilcoxon	1092	0.669	-0.0521
effectiveCharacterCount	Wilcoxon	1096	0.689	-0.0486
queryTime	Wilcoxon	1096	0.690	-0.0486
upToUpAvg	Wilcoxon	1208	0.690	0.0486
dwellTimeAvg	Wilcoxon	1190	0.790	0.0326
charactersPerMinute	t-test	-0.230	0.819	-0.0484
latencyAvg	Wilcoxon	1179	0.849	0.0234
intervalAvg	Wilcoxon	1179	0.849	0.0234
flightTimeStd	Wilcoxon	1133	0.894	-0.0165
intervalStd	Wilcoxon	1140	0.934	-0.0104
latencyStd	Wilcoxon	1140	0.934	-0.0104
flightTimeAvg	Wilcoxon	1159	0.963	0.0061
upToUpStd	Wilcoxon	1151	0.997	-0.0009

Tabla E.4: Resultados de pruebas no pareadas Pre–Post (7° Básico)

Fuente: Elaboración propia

Variable	Método	Est.	p-value	Tamaño de efecto
dwellTimeAvg	Wilcoxon	722	**0.0192**	-0.288
effectiveCharacterCount	Wilcoxon	1292	**0.0256**	0.274
totalCharacterCount	Wilcoxon	1222	0.0959	0.205
lev_similarity_qs	Wilcoxon	1208	0.120	0.191
dwellTimeStd	Wilcoxon	820	0.121	-0.191
writingTime	Wilcoxon	1152	0.270	0.136
queryTime	Wilcoxon	1151	0.274	0.135
flightTimeAvg	Wilcoxon	908	0.398	-0.105
upToUpAvg	Wilcoxon	916	0.434	-0.097
latencyAvg	Wilcoxon	919	0.449	-0.094
intervalAvg	Wilcoxon	919	0.449	-0.094
pauseCount	Wilcoxon	1106	0.463	0.091
pauseTime	Wilcoxon	1104	0.473	0.089
charactersPerMinute	t-test	0.621	0.536	0.136
upToUpStd	Wilcoxon	961	0.674	-0.052
intervalStd	Wilcoxon	967	0.709	-0.046
latencyStd	Wilcoxon	967	0.709	-0.046
flightTimeStd	Wilcoxon	973	0.745	-0.040

Tabla E.5: Resultados de pruebas no pareadas Pre–Post (8° Básico)

Fuente: Elaboración propia

Variable	Método	Est.	p-value	Tamaño de efecto
flightTimeStd	Wilcoxon	527	**0.0269**	0.346
intervalStd	Wilcoxon	524	**0.0304**	0.338
latencyStd	Wilcoxon	524	**0.0304**	0.338
upToUpStd	Wilcoxon	519	**0.0373**	0.326
charactersPerMinute	t-test	-1.84	0.0712	-0.491
latencyAvg	Wilcoxon	497	0.0851	0.269
intervalAvg	Wilcoxon	497	0.0851	0.269
upToUpAvg	Wilcoxon	497	0.0851	0.269
flightTimeAvg	Wilcoxon	494	0.0944	0.262
dwellTimeStd	Wilcoxon	458	0.279	0.170
pauseTime	Wilcoxon	430	0.533	0.098
totalCharacterCount	Wilcoxon	358	0.582	-0.087
queryTime	Wilcoxon	422	0.623	0.078
writingTime	Wilcoxon	422	0.623	0.078
dwellTimeAvg	Wilcoxon	420	0.652	0.072
effectiveCharacterCount	Wilcoxon	404	0.836	0.033
pauseCount	Wilcoxon	383	0.896	-0.022
lev_similarity_qs	Wilcoxon	392	1.000	0.001

ANEXO F. ANÁLISIS POR NIVEL SOCIOECONÓMICO

Se realizó un análisis estadístico (Kruskal-Wallis) para observar si existen diferencias significativas entre las variables de dinámicas de tipeo y los distintos niveles socioeconómicos de los establecimientos educacionales (GSE). Cabe recordar que la base de datos de intervención cuenta con observaciones de cuatro establecimientos educacionales, y estos están categorizados mediante un modelo de clasificación desarrollado por la Asociación de Investigadores de Mercado (AIM) con la finalidad de segmentar la población chilena por factores sociales y económicos (ingresos, educación, acceso a internet).

Tabla F.1: Caracterización socioeconómica de los establecimientos participantes

Fuente: Elaboración propia

Establecimiento	**GSE**
Establecimiento 1	Medio
Establecimiento 2	Medio bajo
Establecimiento 3	Medio bajo
Establecimiento 4	Alto

Los resultados del análisis están presentados en la Tabla F.2. Se observa la existencia de diferencias significativas en todas las variables de dinámicas de tipeo, lo que sugiere que el nivel socioeconómico podría ser un factor clave al explicar la variabilidad en las métricas de las dinámicas de tipeo entre estudiantes de diferentes establecimientos educacionales.

Además, la Tabla F.3 muestra los resultados de un análisis post-hoc utilizando el método estadístico Bonferroni sobre los resultados previamente obtenidos para identificar los grupos con estas diferencias significativas. Se aprecia que las mayores diferencias significativas (según el valor p) se dan entre los grupos alto y medio bajo.

Tabla F.2: Resultados de la prueba de Kruskal-Wallis por variable en GSE

Variable	**Método**	**p-value**
dwellTimeAvg	Kruskal-Wallis	$3{,}60 \times 10^{-22}$
flightTimeAvg	Kruskal-Wallis	$5{,}29 \times 10^{-16}$
latencyAvg	Kruskal-Wallis	$6{,}78 \times 10^{-15}$
intervalAvg	Kruskal-Wallis	$6{,}78 \times 10^{-15}$
upToUpAvg	Kruskal-Wallis	$3{,}92 \times 10^{-12}$
pauseTime	Kruskal-Wallis	$4{,}83 \times 10^{-12}$
charactersPerMinute	Kruskal-Wallis	$2{,}68 \times 10^{-10}$
pauseCount	Kruskal-Wallis	$3{,}96 \times 10^{-10}$
writingTime	Kruskal-Wallis	$9{,}66 \times 10^{-10}$
queryTime	Kruskal-Wallis	$1{,}04 \times 10^{-9}$
flightTimeStd	Kruskal-Wallis	$1{,}68 \times 10^{-8}$
intervalStd	Kruskal-Wallis	$2{,}10 \times 10^{-7}$
latencyStd	Kruskal-Wallis	$2{,}10 \times 10^{-7}$
dwellTimeStd	Kruskal-Wallis	$3{,}46 \times 10^{-7}$
upToUpStd	Kruskal-Wallis	$4{,}40 \times 10^{-7}$
copypaste	Kruskal-Wallis	$5{,}74 \times 10^{-7}$
dificultad	Kruskal-Wallis	$2{,}16 \times 10^{-4}$
estado_animo	Kruskal-Wallis	$2{,}16 \times 10^{-4}$
copypaste_bin	Kruskal-Wallis	0,0038
totalCharacterCount	Kruskal-Wallis	0,0083
motivacion	Kruskal-Wallis	0,0456
reformulations	Kruskal-Wallis	0,0500
conocimiento_prev	Kruskal-Wallis	0,0931
effectiveCharacterCount	Kruskal-Wallis	0,2512
qs	Kruskal-Wallis	NaN

Nota: Los valores extremadamente pequeños se presentan en notación científica.

Tabla F.3: Resultados completos del análisis post hoc mediante la prueba de Dunn con corrección de Bonferroni

Variable	**Grupo 1**	**Grupo 2**	n_1	n_2	**Estadístico**	p_{adj}
dwellTimeAvg	alto	medio_bajo	336	150	-9.754	$< 10^{-20}$
flightTimeAvg	alto	medio_bajo	336	150	8.304	$< 10^{-16}$
latencyAvg	alto	medio_bajo	336	150	8.030	$< 10^{-15}$
intervalAvg	alto	medio_bajo	336	150	8.030	$< 10^{-15}$
edad	alto	medio_bajo	336	150	-8.029	$< 10^{-15}$
edad	alto	medio	336	98	-7.541	$< 10^{-13}$
upToUpAvg	alto	medio_bajo	336	150	7.195	$< 10^{-12}$
pauseTime	alto	medio_bajo	336	150	7.147	$< 10^{-12}$

Continúa en la página siguiente

Variable	**Grupo 1**	**Grupo 2**	n_1	n_2	**Estadístico**	p_{adj}
dwellTimeAvg	medio	medio_bajo	98	150	-6.720	$< 10^{-10}$
charactersPerMinute	alto	medio_bajo	336	150	-6.595	$< 10^{-10}$
pauseCount	alto	medio_bajo	336	150	6.398	$< 10^{-9}$
writingTime	alto	medio_bajo	336	150	6.340	$< 10^{-9}$
queryTime	alto	medio_bajo	336	150	6.329	$< 10^{-9}$
flightTimeStd	alto	medio_bajo	336	150	5.983	$< 10^{-6}$
intervalStd	alto	medio_bajo	336	150	5.545	$< 10^{-6}$
latencyStd	alto	medio_bajo	336	150	5.545	$< 10^{-6}$
upToUpStd	alto	medio_bajo	336	150	5.405	$< 10^{-6}$
dwellTimeStd	alto	medio_bajo	336	150	-5.178	$< 10^{-6}$
copypaste	alto	medio_bajo	336	150	-5.032	$< 10^{-6}$
dwellTimeStd	medio	medio_bajo	98	150	-4.168	$9{,}22 \times 10^{-5}$
dificultad	alto	medio_bajo	336	150	3.714	$6{,}12 \times 10^{-4}$
estado_animo	alto	medio_bajo	336	150	3.714	$6{,}12 \times 10^{-4}$
latencyAvg	medio	medio_bajo	98	150	3.453	$1{,}66 \times 10^{-3}$
intervalAvg	medio	medio_bajo	98	150	3.453	$1{,}66 \times 10^{-3}$
dificultad	medio	medio_bajo	98	150	3.439	$1{,}75 \times 10^{-3}$
estado_animo	medio	medio_bajo	98	150	3.439	$1{,}75 \times 10^{-3}$
flightTimeAvg	medio	medio_bajo	98	150	3.333	$2{,}57 \times 10^{-3}$
flightTimeAvg	alto	medio	336	98	3.332	$2{,}59 \times 10^{-3}$
copypaste_bin	alto	medio_bajo	336	150	-3.212	$3{,}96 \times 10^{-3}$
pauseCount	alto	medio	336	98	3.172	$4{,}54 \times 10^{-3}$
copypaste	alto	medio	336	98	-3.111	$5{,}59 \times 10^{-3}$
flightTimeStd	medio	medio_bajo	98	150	3.097	$5{,}87 \times 10^{-3}$
totalCharacterCount	alto	medio_bajo	336	150	3.056	$6{,}73 \times 10^{-3}$
upToUpStd	medio	medio_bajo	98	150	3.038	$7{,}14 \times 10^{-3}$
upToUpAvg	medio	medio_bajo	98	150	3.019	$7{,}62 \times 10^{-3}$
intervalStd	medio	medio_bajo	98	150	2.971	$8{,}92 \times 10^{-3}$
latencyStd	medio	medio_bajo	98	150	2.971	$8{,}92 \times 10^{-3}$
latencyAvg	alto	medio	336	98	2.962	$9{,}17 \times 10^{-3}$
intervalAvg	alto	medio	336	98	2.962	$9{,}17 \times 10^{-3}$
pauseTime	alto	medio	336	98	2.869	$1{,}23 \times 10^{-2}$
pauseTime	medio	medio_bajo	98	150	2.867	$1{,}24 \times 10^{-2}$
charactersPerMinute	medio	medio_bajo	98	150	-2.797	$1{,}55 \times 10^{-2}$
writingTime	alto	medio	336	98	2.782	$1{,}62 \times 10^{-2}$
queryTime	alto	medio	336	98	2.775	$1{,}66 \times 10^{-2}$
upToUpAvg	alto	medio	336	98	2.739	$1{,}85 \times 10^{-2}$
charactersPerMinute	alto	medio	336	98	-2.477	$3{,}98 \times 10^{-2}$
writingTime	medio	medio_bajo	98	150	2.334	$5{,}88 \times 10^{-2}$
queryTime	medio	medio_bajo	98	150	2.333	$5{,}90 \times 10^{-2}$
motivacion	alto	medio_bajo	336	150	2.318	$6{,}14 \times 10^{-2}$
pauseCount	medio	medio_bajo	98	150	2.033	$1{,}26 \times 10^{-1}$

Continúa en la página siguiente

Variable	**Grupo 1**	**Grupo 2**	n_1	n_2	**Estadístico**	p_{adj}
motivacion	medio	medio_bajo	98	150	1.975	$1{,}45 \times 10^{-1}$
copypaste_bin	alto	medio	336	98	-1.719	$2{,}57 \times 10^{-1}$
flightTimeStd	alto	medio	336	98	1.614	$3{,}19 \times 10^{-1}$
intervalStd	alto	medio	336	98	1.382	$5{,}01 \times 10^{-1}$
latencyStd	alto	medio	336	98	1.382	$5{,}01 \times 10^{-1}$
totalCharacterCount	alto	medio	336	98	1.281	$6{,}00 \times 10^{-1}$
upToUpStd	alto	medio	336	98	1.186	$7{,}07 \times 10^{-1}$
totalCharacterCount	medio	medio_bajo	98	150	1.178	$7{,}17 \times 10^{-1}$
copypaste	medio	medio_bajo	98	150	-1.055	$8{,}74 \times 10^{-1}$
copypaste_bin	medio	medio_bajo	98	150	-0.909	1,00
dwellTimeAvg	alto	medio	336	98	-0.740	1,00
dificultad	alto	medio	336	98	-0.714	1,00
estado_animo	alto	medio	336	98	-0.714	1,00
edad	medio	medio_bajo	98	150	0.595	1,00
dwellTimeStd	alto	medio	336	98	0.286	1,00
motivacion	alto	medio	336	98	-0.252	1,00

ANEXO G. ANÁLISIS DE VALORES ATÍPICOS

Los resultados de los valores atípicos expuestos en la tabla 4.3 muestran que 291 observaciones fueron descartadas del dataset original. Dada la magnitud de esta reducción muestral, resulta de interés caracterizar este subconjunto segregado. Los siguientes análisis descriptivos buscan comprender si estos datos representan errores de medición, o por el contrario, manifiestan patrones de comportamiento extremos pero legítimos.

Inicialmente, el gráfico de caja ilustrado en la Figura G.1 evidencia que la existencia de valores atípicos se concentra principalmente en los rangos superiores de cada variable. Los tiempos de latencia, de vuelo y de liberación muestran mayor cantidad de datos anómalos. Complementariamente, el gráfico G.2 ilustra una comparativa entre las observaciones retenidas (en rojo) y las descartadas (en azul). Se puede observar que las variables descartadas se caracterizan por tiempos de consulta, tiempos de pausa y niveles de re formulación significativamente superiores a la masa de datos retenidos. Para validar este supuesto, se realizó un test estadístico de suma de rangos de Wilcoxon de dos colas. Las hipótesis planteadas son las siguientes:

- Hipótesis Nula (H_0): No existen diferencias significativas en la distribución de las variables temporales y de comportamiento entre el grupo de observaciones retenidas y el grupo de observaciones descartadas.
- Hipótesis Alternativa (H_A): Existen diferencias significativas en la distribución de dichas variables entre ambos grupos, lo que sugiere que pertenecen a poblaciones con dinámicas de comportamiento distintas.

Los resultados del test ($p < 0{,}05$) permitieron rechazar la hipótesis nula, validando formalmente el supuesto de que las observaciones atípicas provienen de distribuciones distintas y no representan fluctuaciones aleatorias de la muestra principal.

Adicionalmente, con el objetivo de identificar el origen de estas anomalías, se generó la métrica de distancia levenshtein para detectar casos en donde se haya hecho uso de copiar el enunciado de la pregunta en el buscador. Este escenario es problemático para el calculo de variables de dinámicas de tipeo, puesto que el estudiante estaría omitiendo el uso del teclado para generar la consulta. La figura G.3 evidencia que un 66.3 %, de los valores indicados como atípicos, corresponden a escenarios donde el alumno hizo uso de esta estrategia de búsqueda.

En conclusión, la caracterización de las observaciones segregadas permite arrojar luz sobre la naturaleza de las observaciones extremas en este estudio. Por un lado, se ha evidenciado de manera cuantitativa que una gran proporción de los datos filtrados ($66{,}5\,\%$) corresponden a acciones de copiar y pegar el enunciado en el buscador. Esta estrategia de búsqueda invalida uno de los supuestos necesarios para el análisis de dinámicas de tecleo al suprimir el uso del teclado. Por otro lado, la existencia de pausas excesivamente prolongadas o anomalías extremas podría sugerir posibles fallas instrumentales o ruidos puntuales en el proceso de recolección automatizada de los datos.

Sin embargo, esta última suposición no es concluyente. Las altas latencias en las variables de dinámicas de tipeo también podrían reflejar dificultades reales de los

estudiantes al momento de formular una consulta, como por ejemplo, bloqueos cognitivos, dudas sobre la ortografía o problemas de la comprensión de la tarea. Dado este escenario, la elección de la estrategia de filtrado de datos mediante el rango intercuartil (IQR) responde a una postura metodológica conservadora para garantizar la homogeneidad de la muestra. Por lo tanto, se deja abierta la vía para que futuros trabajos utilicen otras estratégias de tratamiento de valores atípicos.

Tabla G.1: Resultados de la prueba de suma de rangos de Wilcoxon: Comparativa entre Dataset Filtrado y Outliers.

Variable	**Estadístico (W)**	**p-value**	**Método**	**Tamaño Efecto (r)**
pauseCount	115,370	<0.001	Wilcoxon	-0.277
latencyAvg	115,352	<0.001	Wilcoxon	-0.277
intervalAvg	115,352	<0.001	Wilcoxon	-0.277
upToUpAvg	116,658	<0.001	Wilcoxon	-0.269
pauseTime	117,570	<0.001	Wilcoxon	-0.263
charactersPerMinute	198,604	<0.001	Wilcoxon	0.245
flightTimeAvg	120,596	<0.001	Wilcoxon	-0.244
queryTime	122,532	<0.001	Wilcoxon	-0.232
00000 writingTime	122,538	<0.001	Wilcoxon	-0.232
dwellTimeAvg	185,854	2.44e-06	Wilcoxon	0.165
upToUpStd	133,500	2.92e-06	Wilcoxon	-0.163
intervalStd	133,512	2.95e-06	Wilcoxon	-0.163
latencyStd	133,512	2.95e-06	Wilcoxon	-0.163
flightTimeStd	134,176	5.23e-06	Wilcoxon	-0.159
effectiveCharacterCount	142,494	0.002	Wilcoxon	-0.107
totalCharacterCount	143,140	0.003	Wilcoxon	-0.103
dwellTimeStd	166,544	0.211	Wilcoxon	0.043
copypaste	162,852	0.486	Wilcoxon	0.020

Fuente: Elaboración propia

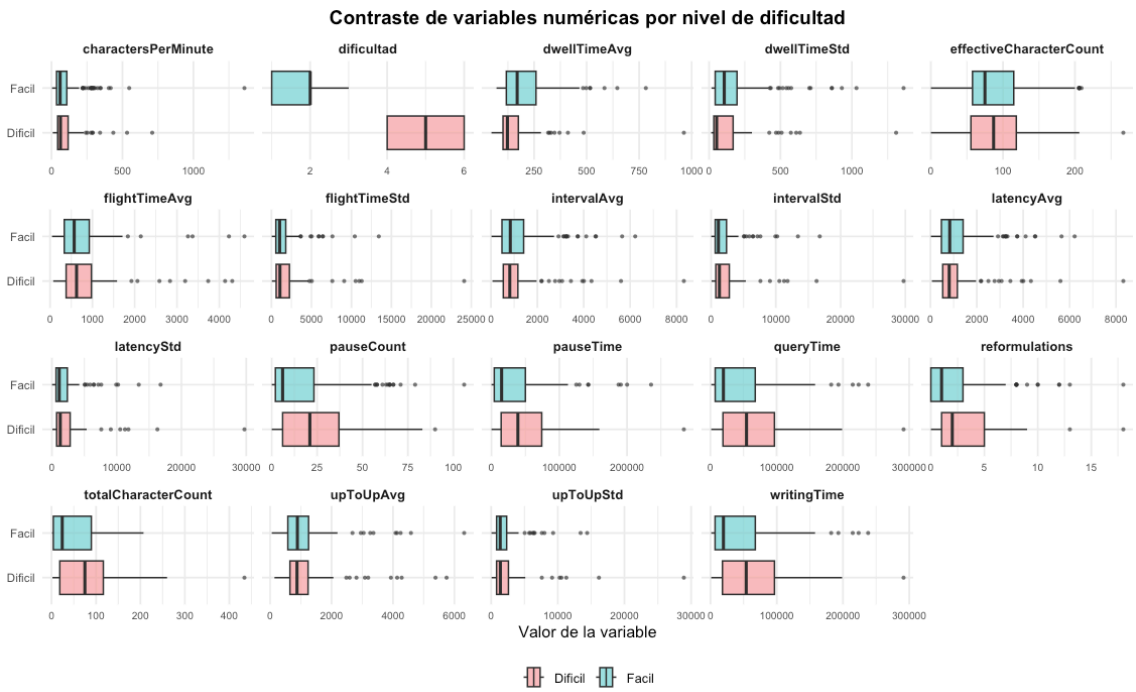

Figura G.1: Distribución de valores con valores atípicos

Fuente: Elaboración propia

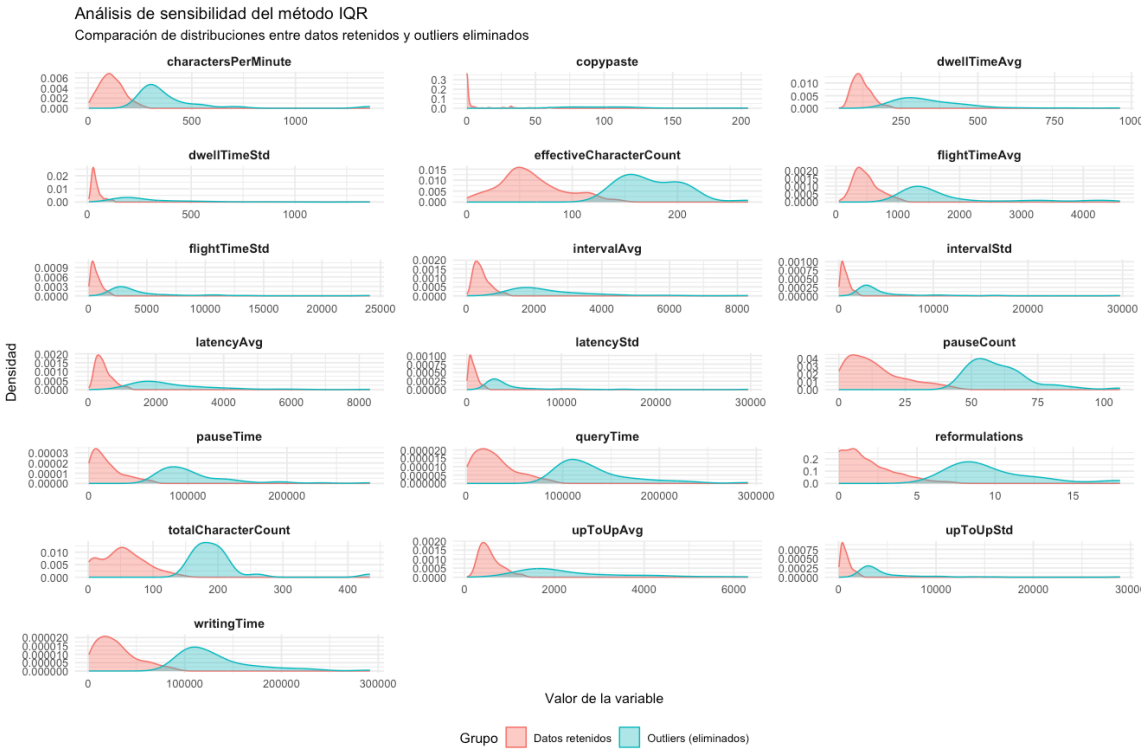

Figura G.2: Distribución de datos atípicos versus datos retenidos

Fuente: Elaboración propia

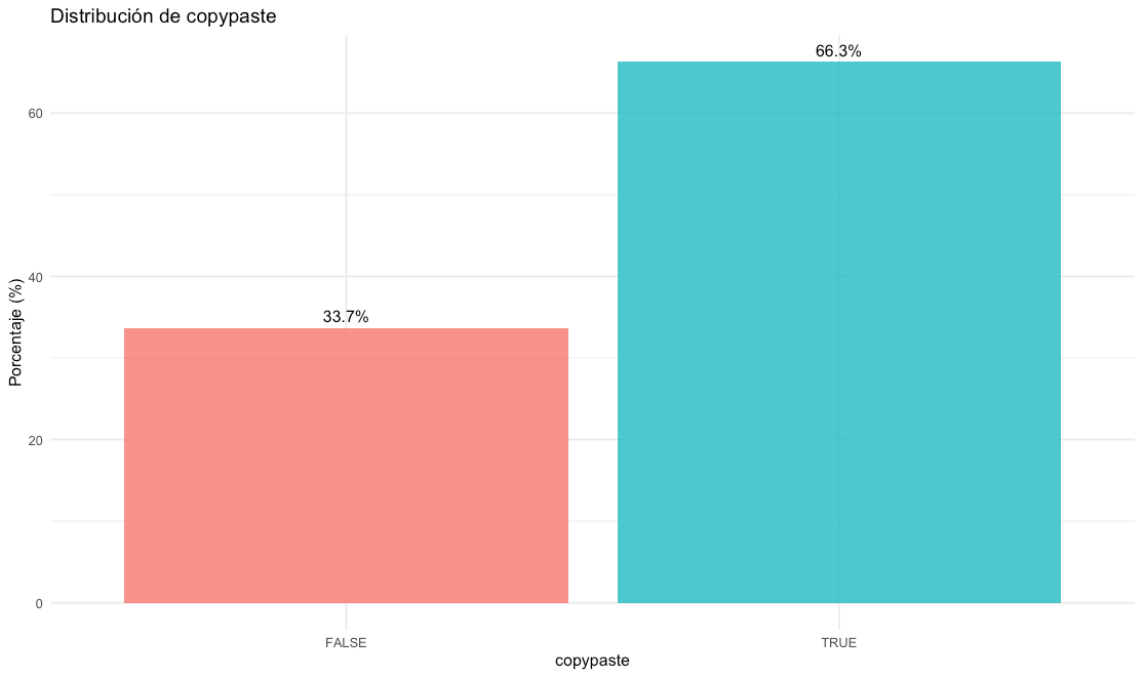

Figura G.3: Porcentaje de observaciones de copy-paste

Fuente: Elaboración propia