Universidad del Valle de Guatemala

Facultad de Ingeniería

Departamento de Ciencias de la Computación

Proyecto 2: Análisis Exploratorio de Datos

Tema 12: Reconocimiento de deletreo manual en Lengua de Señas Americana (ASL)

Curso: CC3084 – Data Science

Integrantes del grupo:

- Diederich Solis 22952
- Juan Cordón
- Sebas Juarez
- Sara Guzman

Semestre II – 2025

Índice

1. Introducción

La lengua de señas constituye el medio principal de comunicación para más de 70 millones de personas sordas en el mundo, y se estima que más de 1500 millones de personas padecen algún grado de pérdida auditiva [1]. El *American Sign Language* (ASL) es el sistema más comúnmente utilizado en los Estados Unidos, y dentro de este, el **deletreo manual** cumple un papel fundamental para expresar nombres propios, direcciones, números y otros términos que no cuentan con un signo específico.

El reto tecnológico consiste en que, a diferencia del reconocimiento automático de voz o la traducción de texto, los sistemas de inteligencia artificial para el reconocimiento del ASL aún se encuentran en desarrollo. La principal dificultad radica en la complejidad de los gestos, las variaciones individuales entre usuarios y la falta de conjuntos de datos suficientemente grandes y diversos.

Este informe busca analizar de manera exploratoria un conjunto de datos de deletreo manual en ASL, con el fin de establecer las bases para futuros modelos de reconocimiento automático basados en visión por computadora.

2. Planteamiento inicial del problema

2.1. Situación problemática

Los avances en inteligencia artificial han revolucionado el reconocimiento automático de voz y la traducción de texto, facilitando la interacción entre humanos y dispositivos. Sin embargo, estos desarrollos aún no se he extendido plenamente al ámbito de la lengua de señas, lo que genera una brecha tecnológica para millones de personas sordas.

El deletreo manual del ASL permite formar palabras letra por letra mediante configuraciones específicas de la mano. En la vida cotidiana, se utiliza para introducir información esencial como nombres propios, direcciones de correo electrónico, números de teléfono o acrónimos. Muchos usuarios sordos de teléfonos inteligentes pueden deletrear palabras con sus manos de manera más rápida que escribir en teclados virtuales, alcanzando hasta 57 palabras por minuto frente a un promedio de 36 en usuarios oyentes [2].

No obstante, los sistemas automáticos de reconocimiento del deletreo manual aún presentan desafíos:

- Variabilidad entre usuarios (tamaño de manos, estilo personal, destreza).
- Diferencias en condiciones de captura (iluminación, ángulos de cámara, accesorios).
- Falta de representatividad en los datos (pocos tonos de piel, predominio de usuarios de ciertas regiones).

2.2. Problema científico

Derivado de lo anterior surge la siguiente pregunta de investigación:

> **¿Es posible desarrollar un modelo de visión por computadora que reconozca con precisión el deletreo manual del lenguaje de señas estadounidense (ASL), a partir de imágenes y metadatos recolectados en condiciones reales de uso, considerando la variabilidad entre usuarios y los posibles sesgos del dataset?**

2.3. Objetivos

Objetivo general

Analizar de manera exploratoria el conjunto de datos de deletreo manual en ASL con el fin de sentar las bases para un modelo de reconocimiento automático que contribuya a la accesibilidad tecnológica para personas sordas y con pérdida auditiva.

Objetivos específicos

1. Describir y caracterizar el dataset de ASL Fingerspelling, identificando variables, estructura y calidad de los datos.
2. Implementar procesos de limpieza y preprocesamiento que aseguren la consistencia de la información.

3. Realizar un análisis estadístico y visual de los datos para detectar patrones, desbalances y posibles sesgos.

4. Establecer implicaciones y lineamientos para la construcción de modelos de clasificación basados en visión artificial.

3. Descripción de los datos

3.1. Origen de los datos

El conjunto de datos utilizado en este proyecto fue recolectado con la participación de más de **100 personas usuarias de la Lengua de Señas Americana (ASL)** en distintas regiones de Estados Unidos. Todas las personas participantes fueron reclutadas a través de la Red de Artes Profesionales para Sordos, garantizando que el ASL fuera su lengua materna y principal medio de comunicación. Esta diversidad aporta variabilidad en características físicas, tonos de piel y estilos de deletreo, lo que enriquece la representatividad del dataset.

3.2. Proceso de recopilación

La recopilación de datos se realizó mediante **smartphones** que incluían una aplicación especialmente diseñada para registrar secuencias de video. Cada participante debía presionar un botón para iniciar y finalizar la grabación, durante la cual **deletreaban en ASL** el texto que aparecía en la pantalla. Las condiciones de captura fueron intencionalmente diversas para reflejar escenarios reales de uso:

- **Variación en las manos:** algunos participantes utilizaron la mano izquierda, otros la derecha, e incluso alternaron entre ambas en diferentes secuencias.

- **Variación postural y de encuadre:** diferencias en distancia a cámara, zoom, pose corporal, iluminación y presencia de accesorios.

- **Frases objetivo:** palabras, nombres y direcciones que normalmente se introducen en dispositivos móviles, reflejando un uso práctico del deletreo manual.

3.3. Contenido del dataset

Para efectos de este proyecto únicamente se utilizará el archivo `supplemental_metadata.csv`, ya que contiene la información necesaria para realizar el análisis exploratorio solicitado. Este archivo incluye variables descriptivas de las secuencias capturadas, sin necesidad de procesar los archivos de landmarks (`.parquet`) debido a su gran tamaño y complejidad, lo cual excede el alcance de esta entrega.

Las variables principales contenidas en el archivo de metadatos son:

- `path`: ruta o identificador del archivo de referencia.
- `file_id`: identificador único del archivo.
- `participant_id`: identificador de la persona firmante.
- `sequence_id`: identificador de la secuencia de video.
- `phrase`: palabra o frase deletreada en la secuencia.

3.4. Variables categóricas

Existen varias variables categóricas de interés para el análisis:

- `participant_id`: identifica a la persona que firma. Permite estudiar variaciones entre usuarios.
- `phrase`: corresponde a la palabra deletreada. Puede descomponerse en letras individuales para análisis de balance de clases.
- `letter`: variable derivada que representa cada carácter del alfabeto manual.

3.5. Estructura general

En la Tabla 1 se resumen las principales variables del archivo `supplemental_metadata.csv`.

Cuadro 1: Resumen de variables principales del archivo de metadatos

Variable	Tipo	Descripción
`participant_id`	Categórica	ID anónimo del participante.
`sequence_id`	Entera	ID único de secuencia.
`file_id`	Entera	Identificador único del archivo.
`path`	Texto	Ruta del archivo de referencia.
`phrase`	Categórica	Palabra o frase deletreada.

3.6. Estructura y dimensiones

El archivo `supplemental_metadata.csv` contiene un total de **43,998 secuencias** y **509 frases distintas**, asociadas a los participantes mediante identificadores únicos. No se detectaron valores nulos en las variables principales (`path`, `file_id`, `sequence_id`, `participant_id`, `phrase`), lo que indica una buena calidad de datos a nivel de metadatos.

3.7. Distribución de frases

El análisis inicial muestra que las frases más frecuentes en el dataset incluyen expresiones como *"find a nearby parking spot"* y *"most judges are very honest"*, cada una con más de 100 repeticiones. Estas frases reflejan ejemplos de lenguaje cotidiano que suelen escribirse en dispositivos móviles, lo que confirma la intención práctica del conjunto de datos.

3.8. Distribución por letras

A partir de las frases se derivó la variable `letter`, correspondiente a la primera letra de cada palabra o frase deletreada.

La Figura 1 muestra la distribución de secuencias por letra. Se observa un **fuerte desbalance de clases**:

- La letra **t** concentra más de 10,000 ejemplos, representando casi una cuarta parte del total.
- Letras como **z** (87 ejemplos), **k** (180 ejemplos) y **q** (338 ejemplos) aparecen con muy baja frecuencia.

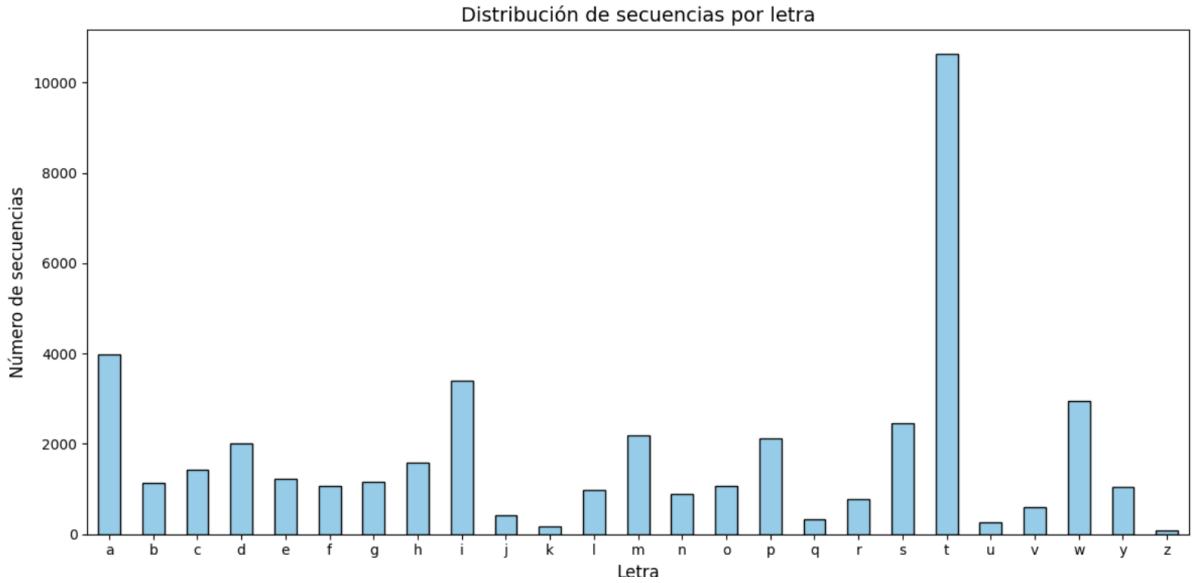

Figura 1: Distribución de secuencias por letra derivada de las frases.

- Otras letras como **a** (3,980), **i** (3,398) y **w** (2,959) están mejor representadas.

Este desbalance tendrá implicaciones importantes en fases posteriores de modelado, donde será necesario aplicar técnicas de balanceo o ponderación de clases.

3.9. Distribución por participantes

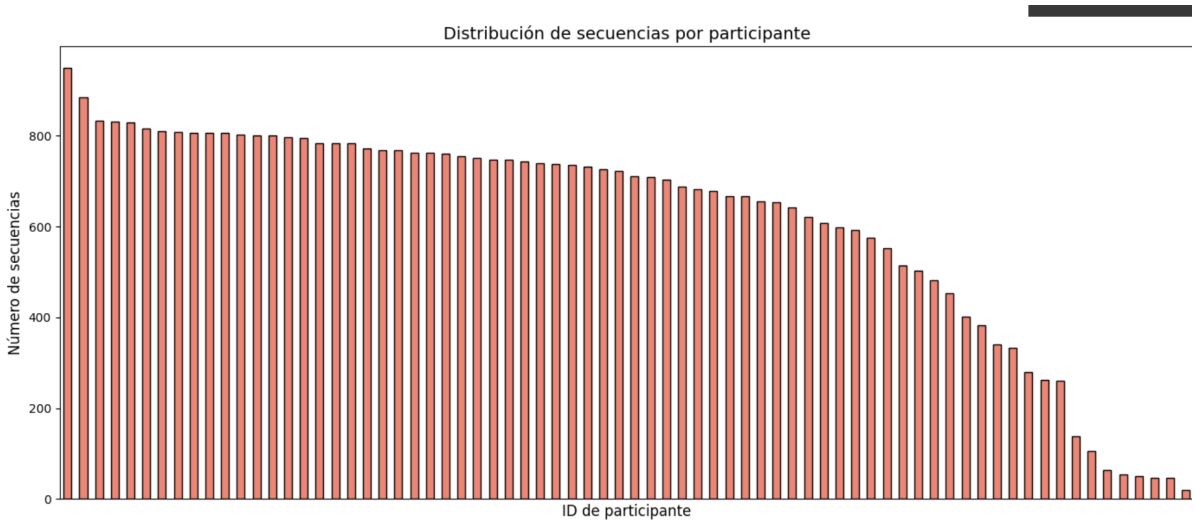

Figura 2: Distribución de secuencias por participante.

La Figura 2 muestra la distribución de secuencias por participante. Se observa que:

- La participación no es homogénea: algunos usuarios aportaron más de 900 secuencias, mientras que otros contribuyeron con menos de 200.
- Esto introduce un **sesgo potencial**, ya que un pequeño grupo de participantes podría dominar el entrenamiento de futuros modelos.
- Aun así, la variedad de 72 personas firmantes constituye un avance frente a datasets más pequeños o menos diversos.

3.10. Notas preliminares

Los hallazgos de esta sección resaltan dos características clave del dataset:

1. Existe un **desbalance significativo entre letras**, lo cual podría sesgar modelos de clasificación hacia las clases mayoritarias.
2. La **contribución desigual de los participantes** puede limitar la generalización si no se controla durante el entrenamiento.

Estos aspectos serán abordados en la fase de limpieza y preprocesamiento, donde se definirán estrategias de normalización y balanceo de datos.

4. Limpieza y Preprocesamiento de los datos

4.1. Estrategia general

El conjunto de datos analizado presenta buena calidad en cuanto a completitud, ya que no se detectaron valores nulos en las variables principales. Sin embargo, el análisis exploratorio reveló dos aspectos críticos que deben atenderse antes de su uso en fases posteriores de modelado:

1. **Desbalance de clases:** se observó que algunas letras (t, a, i, w) concentran un número muy elevado de ejemplos, mientras que otras (z, k, q) tienen muy pocas instancias. Esto puede sesgar los modelos hacia clases mayoritarias.

2. **Contribución desigual de participantes:** algunos usuarios aportaron más de 900 secuencias, mientras que otros menos de 200. Esto podría introducir un sesgo hacia ciertos estilos de señado.

4.2. Acciones de limpieza propuestas

Para garantizar consistencia y robustez en la información, se definieron las siguientes acciones de limpieza y preprocesamiento:

- **Normalización de texto:** transformar todas las frases a minúsculas y eliminar espacios en blanco al inicio y final para evitar duplicidades.
- **Derivación de variable categórica `letter`:** generar la columna correspondiente a la primera letra de cada frase, con el fin de estudiar la distribución de clases y su equilibrio.
- **Validación de claves:** comprobar que no existan duplicados en los identificadores `sequence_id` y `file_id`, asegurando unicidad de cada observación.
- **Chequeo de valores faltantes:** aunque no se detectaron inicialmente, se verificará sistemáticamente la presencia de celdas vacías o inconsistencias en las columnas principales.
- **Análisis de balance de clases:** calcular la frecuencia relativa de cada letra para identificar el nivel de desbalance y documentar sus implicaciones.
- **Análisis de distribución por participante:** revisar si existen participantes con muy baja contribución (outliers) que podrían distorsionar el entrenamiento de modelos.

4.3. Resultados de la limpieza y preprocesamiento

La Figura 3 Tras aplicar las acciones de limpieza propuestas, se obtuvieron los siguientes hallazgos clave:

- **Duplicados:** No se encontraron duplicados en los identificadores `sequence_id`, lo que confirma que cada secuencia es única. En el caso de `file_id`, se observaron coincidencias debido a que varios fragmentos pueden provenir del mismo archivo de video, lo cual es consistente con el diseño del dataset.
- **Valores nulos:** Ninguna de las variables analizadas (`path`, `file_id`, `sequence_id`, `participant_id`, `phrase`, `letter`) presentó valores faltantes, lo que refuerza la completitud y calidad de los metadatos.
- **Dimensiones finales:** El conjunto limpio cuenta con **52,958 secuencias**, **509 frases distintas** y **72 participantes**.

Distribución por letras

El análisis de la variable `letter` revela un desbalance significativo:

- La letra **t** concentra el 24.1 % de todas las secuencias (12,790 instancias).
- Otras letras como **a** (9.0 %), **i** (7.8 %) y **w** (6.6 %) también aparecen con alta frecuencia.
- En contraste, letras como **z** (104 ejemplos), **k** (216) y **u** (321) están muy poco representadas, con proporciones inferiores al 1 %.

El **índice de Gini calculado para la distribución de letras fue 0.487**, lo que refleja un nivel de desbalance medio-alto (donde 0 indica perfecta equidad y 1 desbalance extremo). Esto implica que los modelos entrenados con este dataset tenderían a sesgarse hacia letras mayoritarias si no se aplican técnicas de balanceo.

Distribución por participantes

La contribución de los 72 participantes también muestra cierta desigualdad:

- En promedio, cada persona aportó unas **735 secuencias**, con una desviación estándar de 295.
- El rango va desde un mínimo de solo **20 secuencias** hasta un máximo de **1,151 secuencias**.

- Este comportamiento sugiere que algunos estilos de señado están más representados que otros, lo cual puede introducir un sesgo hacia participantes con mayor volumen de datos.

4.4. Implicaciones para el modelado

El análisis de limpieza y preprocesamiento permite concluir que, aunque el dataset presenta buena calidad estructural y no requiere correcciones mayores, los **problemas de desbalance entre letras y participantes** deben abordarse en etapas posteriores.

Referencias

[1] World Health Organization (2021). *World Report on Hearing*. Disponible en: `https://www.who.int/publications/i/item/world-report-on-hearing`

[2] Goldin-Meadow, S. (2010). *The Resilience of Language: What Gesture Creation in Deaf Children Can Tell Us About How All Children Learn Language*. Psychology Press.

```
Distribución por letra (primeras filas):
        count   ratio
letter
a        4788  0.0904
b        1384  0.0261
c        1680  0.0317
d        2410  0.0455
e        1500  0.0283
f        1280  0.0242
g        1400  0.0264
h        1904  0.0360
i        4126  0.0779
j         524  0.0099

Top 5 letras más frecuentes:
        count   ratio
letter
t       12790  0.2415
a        4788  0.0904
i        4126  0.0779
w        3543  0.0669
s        2952  0.0557

Bottom 5 letras menos frecuentes (sin 'otra'):
        count   ratio
letter
z         104  0.0020
k         216  0.0041
u         321  0.0061
q         417  0.0079
j         524  0.0099

Índice de Gini (letras): 0.4867  -> 0=perfectamente balanceado, 1=desbalance extremo

Participantes únicos: 72
Estadísticos #secuencias por participante:
count      72.000000
mean      735.527778
std       295.356837
min        20.000000
25%       613.750000
50%       865.500000
75%       937.500000
max      1151.000000
Name: count, dtype: float64

Dataset limpio guardado en: /content/eda_outputs/metadata_clean.csv

Resumen limpieza/preprocesamiento:
rows_iniciales          52958.000000
rows_limpios            52958.000000
cols                        6.000000
n_participantes            72.000000
n_secuencias            52958.000000
n_frases                  508.000000
n_duplicados_sequence       0.000000
n_duplicados_file       52958.000000
gini_letras                 0.486737
```

Figura 3: Resultado limpieza.