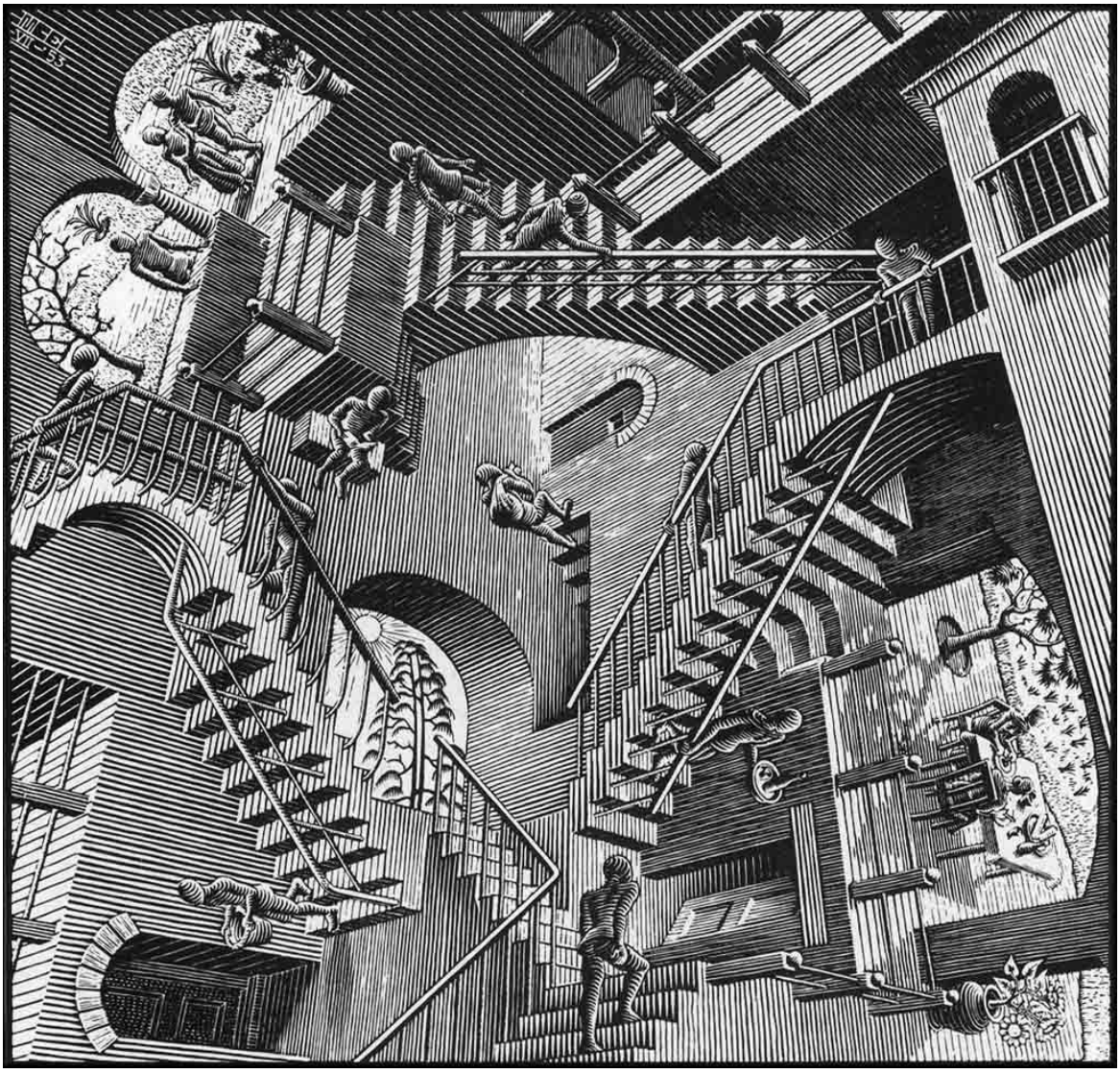
Tarea 2: Preparación proyecto final

Taller de Modelado de Datos



"Relativity" (1953), M. C. Escher. Litografía: esta composición refleja la interconexión y estructura de múltiples sistemas coexistiendo bajo reglas distintas, una metáfora visual del modelado de datos: distintos planos que se ordenan en un todo coherente pese a su complejidad aparente.

Cristopher Corona Velasco
Ingeniería en Ciencia de Datos

ITESO, Universidad Jesuita de Guadalajara

I. Descripción de las temáticas del proyecto

A. Tema 1: Clasificación: Reconocimiento de caracteres manuscritos con EMNIST

Tomaremos imágenes pequeñas de letras y números escritos a mano y enseñaremos a un sistema a reconocer qué carácter es en cada imagen. Probaremos primero una opción sencilla (regresión logística/k-NN) y luego otra un poco más capaz (CNN básica), comparando sus aciertos. El trabajo consiste en preparar ejemplos, entrenar, evaluar y elegir la versión más práctica y estable para la clasificación.

B. Tema 2: Regresión: Predicción de la nota final (Student Performance, UCI)

Usaremos una tabla de datos de estudiantes para estimar su nota final a partir de información básica (hábitos, ausencias, etc.). Construiremos una predicción numérica y compararemos una solución simple contra otra más flexible. El foco es medir el error y entender qué factores pesan más en el resultado.

II. Motivación del proyecto

A. Interés en clasificación de EMNIST

Un modelo que lea caracteres manuscritos permite automatizar la captura de información en formularios, guías de envío y documentos escaneados, mejorando tiempos y reduciendo errores humanos. También habilita OCR básico, apoyo a accesibilidad (texto a voz) y digitalización histórica. Empezar con EMNIST nos da un terreno controlado para luego trasladar la idea a escenarios reales.

B. Interés en predicción de nota

Predecir la nota final ayuda a crear alertas tempranas para estudiantes en riesgo, personalizar apoyos (tutorías, material) y planear recursos en escuelas. Más que “adivinar”, buscamos una herramienta que oriente decisiones útiles: dónde intervenir, qué hábitos importan y cómo mejorar resultados de forma práctica y responsable.

III. Incógnitas abordadas

A. Incógnitas de la clasificación EMNIST

1. ¿Qué precisión logramos con una solución simple vs. una más capaz?
2. ¿En qué pares de caracteres se confunde más (p. ej., “O/0”, “1/1”)?
3. ¿Cuánto mejora con pequeños aumentos de datos (rotaciones/ruido leves)?
4. ¿Cuál es el mejor equilibrio entre acierto y simplicidad para uso práctico?
5. ¿Qué tiempo de predicción por imagen tenemos en una computadora común?

B. Incógnitas de la predicción de nota

1. ¿Cuál es el error promedio razonable de la predicción?
2. ¿Qué factores son más influyentes y, sobre todo, accionables?
3. Si quitamos columnas que ya anticipan la nota final, ¿cambia mucho el desempeño?
4. ¿Una solución simple basta o conviene una más flexible?
5. ¿Qué tan temprano en el curso la predicción es útil con un error aceptable?

IV. Conjuntos de datos

A. Conjunto 1: Clasificación

Nombre del conjunto de datos: EMNIST (Extended MNIST) — *split* “Balanced”.

Enlace al conjunto de datos: Página oficial de NIST (descarga y resumen).

Variable objetivo: Clase del carácter (letra o dígito) por imagen 28×28.

Por qué es clasificación: la salida es categórica (una etiqueta entre varias clases definidas en el *split*).

B. Conjunto 2: Regresión

Nombre del conjunto de datos: Student Performance (UCI Machine Learning Repository).

Enlace al conjunto de datos: Ficha oficial en UCI

Variable objetivo: G3 (calificación final numérica de 0–20).

Por qué es regresión: la salida a predecir es continua (un valor numérico), y el dataset está planteado para regresión además de clasificación.

V. Conclusión

En esta tarea, vamos a trabajar en dos áreas de aprendizaje supervisado: clasificación con EMNIST, que es sobre reconocer letras y números en imágenes de 28×28, y regresión con Student Performance, donde vamos a estimar la calificación final G3 usando variables de contexto. Así, cubrimos las dos salidas típicas del aprendizaje automático: etiquetas y números. Nuestros objetivos son medir la precisión y las confusiones en EMNIST, y el error (MAE/RMSE) y la importancia de las variables en el rendimiento académico, asegurándonos de evitar la fuga de datos.

En las próximas semanas, vamos a preparar los datos, trabajarlos y entrometernos en un proceso de prueba y error con la única intención de concluir con un proyecto de alto valor y lo suficientemente eficaz para obtener los resultados buscados.

VI. Referencias

- National Institute of Standards and Technology. (2017, April 4). *The EMNIST dataset*. *https://www.nist.gov/itl/products-and-services/emnist-dataset NIST*
- Cohen, G., Afshar, S., Tapson, J., & van Schaik, A. (2017, February 17). *EMNIST: An extension of MNIST to handwritten letters*. arXiv. *https://arxiv.org/abs/1702.05373 arXiv*
- *UCI Machine Learning Repository. (2014, November 26). Student Performance. https://archive.ics.uci.edu/dataset/320/student%2Bperformance archive.ics.uci.edu*