**ISPC - TSCD**

**MÓDULO CIENTÍFICO DE DATOS -**

**PROCESAMIENTO DEL HABLA**

**COHORTE 2022**



Profesor MOISÉS TINTE

Grupo 4 - Integrantes

* Cristian Falco
* Eduardo Figueroa
* Sol Figueroa
* Jorge Flores
* Carlos Giménez
* Cinthia Yael Gómez
* Monserrat Gutiérrez
* Walter Nieto
* Jorgelina Tissera

# **Evidencia 2 - informe de avance del proyecto**

## Repositorio

<https://github.com/monsegutier/Grupo4/tree/main/Procesamiento%20del%20Habla>

## Tema / objetivo

**PROYECTO DE DEEP LEARNING EN EL ÁREA DE LA GENÓMICA**

**Clasificación de secuencias de ADN codificantes y no codificantes**

El proyecto tiene como finalidad desarrollar un modelo de clasificación, mediante redes neuronales y basado en procesamiento de lenguaje natural, que pueda procesar secuencias genómicas, en este caso, de la bacteria E.coli, e identificar las regiones codificantes y no codificantes, dentro de su secuencia de ADN.

## Dataset

Se selecciona la E.coli por ser un organismo altamente estudiado y que posee una cantidad similar de secuencias codificantes y no codificantes, lo cual sera necesario para posteriormente establecer las clases (codificante [0] y no codificante [1])

(en un principio se iba a utilizar el genoma del bacteriófago lambda el cual posee una cantidad de bases mucho menor (348k el bacteriófago y 4kk la E.coli) lo cual haría mucho mas fácil el manejo de la información, pero al tener las clases muy desbalanceadas, no seria posible posteriormente entrenar el modelo de manera eficaz.)

## Grado de Avance

**1.Incorporación del Dataset:**

Recolección y adquisición de datos de diversas fuentes.

Almacenamiento en bases de datos o sistemas de archivos.

**Estado: Completado.**

**2. Exploración y Análisis de Datos (EDA):**

Análisis inicial de los datos para entender su estructura y características.

Identificación de patrones, anomalías y relaciones entre variables.

Visualización de datos para obtener insights preliminares.

**Estado: Completado.**

**3. Preprocesamiento de Datos:**

Limpieza de datos (manejo de valores faltantes, corrección de errores).

Transformación de datos (escalado, normalización, codificación de variables categóricas).

División del dataset en conjuntos de entrenamiento y prueba.

**Estado: Completado**

**4. Ingeniería de Características:**

Creación de nuevas variables a partir de las existentes.

Selección de características más relevantes para el modelo.

**Estado: Completado**

**5. Selección y Entrenamiento de Modelos:**

Evaluación de diferentes algoritmos de machine learning.

Entrenamiento de modelos iniciales y ajuste de hiperparámetros.

**Estado: Completado**

**6. Implementación de Modelos:**

Implementación de modelos de regresión, clasificación o clustering.

Implementación de Redes Neuronales Convolucionales (CNNs) si aplica.

**Estado: Completado**

**7. Evaluación de Modelos:**

Evaluación de la precisión y desempeño de los modelos utilizando métricas adecuadas.

Validación cruzada para asegurar la robustez del modelo.

**Estado: Completado**

**8. Obtención y Comunicación de Resultados:**

Interpretación de resultados y generación de informes.

Visualización de resultados y elaboración de conclusiones.

Informe Técnico

**Estado: Completado**

9. **Actualización Repositorio en Github con el material del trabajo**

**Estado: En progreso**, 80% Completado.

10. **Presentación / Exposición Final:**

**Estado: En progreso**, aproximadamente 80 % completado.