Descripción General: el proyecto consiste en resolver un problema sobre un conjunto de datos. Se debe analizar el contexto, hacer un análisis exploratorio de datos, y seleccionar el mejor modelo al evaluar diferentes técnicas de clasificación cómo: modelos lineales generalizados (GLM) o modelos aditivos generalizados (GAM). Así como proponer un modelo de agrupación usando modelo mixtos Gaussianos que ayuden a complementar el análisis. También deben involucrar técnicas de reducción de la dimensión como PCA, KPCA, ICA y t-SNE, con el fin de mejorar los modelos o ayudar a la visualización de la información.

Nota: El proyecto de RNAseq no involucra métodos supervizados, es decir, que sólo se aplicarán técnicas de reducción de la dimensión y clustering.

Directrices: el proyecto consistirá en

1. Estudiar el contexto del problema a resolver.
2. Una primera parte del proyecto, consiste en realizar un análisis exploratorio de los datos, en donde se realicen conclusiones, mediante gráficas y cálculos estadísticos básicos (aquí pueden involucrar técnicas de reducción de la dimensión para visualización).
3. Realizar un Jupyter Notebook en el que se describan claramente el problema, el desarrollo, análisis, metodología y los procedimientos desarrollados y las pertinentes conclusiones que obtiene en el contexto del problema considerado. El código del mismo debe estar bien documentado y organizado.
4. Se realizarán unas preguntas de defensa del proyecto, programadas individualmente por cada grupo.

Requerimientos: el proyecto debe:

1. Desarrollarse en Python.
2. Depurar y preprocesar el conjunto de datos.
3. Realizar una análisis exploratorio de datos.
4. Involucrar técnicas de reducción de la dimensión (PCA, KPCA, ICA, t-SNE) ya sea para mejorar el pipeline completo del proceso o para ayudar a visualizar la información de los datos.
5. Realizar modelos de regresión logística (con y sin regularización L1 y L2), análisis discriminante lineal, Naive Bayes, KNN, modelos GLM y GAM (se pueden usar árboles de decisión, random forest, boosted machines, etc.)
6. Evaluar y seleccionar los modelos usando validación cruzada. Y posteriormente, evaluar el modelo sobre un conjunto independiente de prueba.
7. Desarrollar un modelo de clustering usando modelos mixtos Gaussianos, justificando el número de clusters seleccionados y comparando los resultados con las etiquetas verdaderas (si las hay).
8. Concluir con las observaciones generadas a través del análisis, creación y evaluación de los modelos construidos.

9. El trabajo deber estar bien escrito, sin errores gramaticales y ortográficos.
10. El trabajo puede entregarse en un notebook de Jupyter organizado, como un informe que posee ilustraciones y código, pero que se pueda leer de forma amena.

Entregables y Fechas:

Entrega: Notebook de Jupyter con la descripción del problema considerado, el análisis exploratorio de datos, el conjunto de técnicas utilizadas y los procedimientos desarrollados, una discusión y análisis de los resultados y las conclusiones obtenidas.

Fecha de entrega: a más tardar el **16 de Octubre de 2021**.

Puntaje: 70 puntos

Descripción del problema, análisis exploratorio y calidad de las conclusiones: 20 puntos

Descripción del procedimiento, análisis y discusión realizados: 20 puntos.

Claridad argumentativa, ortografía, presentación: 10 puntos.

Código Python **en notebook y documentado** que respalde todos los análisis/procedimientos y permita replicar todos los resultados: 10 puntos

Defensa: Sustentación del trabajo desarrollado mediante preguntas.

Fecha: Se acordará una sesión del grupo con el profesor, posterior a la entrega del documento.

Puntaje: 30 puntos

Otros:

Grupos: el proyecto se realizará en los grupos designados al inicio del semestre.

Caso especial: si se detecta que un estudiante no participó en la realización de su proyecto, su nota será cero.