**Asignatura Aprendizaje Automático y Análisis de Datos**

**Guía de Trabajo**

**Febrero 24 de 2023**

**Juan José Restrepo**

**Isabella Ceballos**

A partir de lo explicado en el [video](https://youtu.be/IMfLXEOksGc) y en la lectura recomendada, responder las siguientes preguntas:

En máximo 60 minutos responder las preguntas 1-3

1. ¿Qué es el análisis discriminante lineal?

**El análisis discriminante lineal (LDA) es una técnica estadística utilizada en el análisis multivariante para encontrar una función lineal que pueda separar dos o más grupos de datos en función de un conjunto de variables predictoras, con el objetivo de encontrar la combinación lineal de variables predictoras que maximice la separación entre los grupos mientras minimiza la variación dentro de cada grupo.**

1. ¿Cómo se diferencian el análisis discriminante lineal y cuadrático?

**La principal diferencia entre LDA y QDA es la forma de la función discriminante.**

**En LDA, se asume que las varianzas en cada grupo son iguales y que siguen una distribución normal multivariante. Por lo tanto, la función discriminante que se encuentra es lineal y se utiliza para crear un hiperplano que maximice la separación entre los grupos.**

**LDA asume igualdad en las matrices de covarianza y se tiene un número grande de variables predictivas de los grupos, produciendo así una función de separación lineal**

**En el QDA, se permite que cada grupo tenga su propia matriz de covarianza y se asume que también sigue una distribución normal multivariante. En este caso, la función discriminante es cuadrática y se utiliza para crear una superficie que separa los grupos.**

**QDA permite matrices de covarianza diferentes para cada grupo y produce una función de separación cuadrática**

1. ¿Cómo se puede usar el análisis discriminante (lineal o cuadrático) para hacer clasificación?

Ahora, analizar el archivo “K-vecinos más cercanos” y a partir de este responder las preguntas 4-5. En este punto deben tardar máximo 30 minutos

1. ¿Cómo se diferencia el modelo de K-vecinos más cercanos de los demás estudiados hasta el momento?

**La diferencia del modelo KNN con respecto a los demás, es que KNN se utiliza para la clasificación y la regresión, mediante la manipulación los datos de entrenamiento y clasifica los nuevos datos de prueba en función de las métricas de distancia. KNN encuentra los k-vecinos más cercanos a los datos de prueba, y luego la mayoría de las etiquetas de clase realizan la clasificación de estos.**

1. ¿Cómo influye el valor de K en el modelo?

**K es el parámetro que puede ajustar la complejidad de KNN, por lo que, a un mayor k, más suave será el límite de clasificación. O también, se puede pensar que la complejidad de KNN es menor cuando k aumenta y se evitan tener límites de decisión inestables cuando se tiene un valor pequeño de K.**

Usar el resto de la sesión para realizar la parte práctica:

1. Utilizar el conjunto de datos que ya tienen preparado (adults.data) para aplicarle las técnicas vistas hasta el momento, esto implica las siguientes tareas:
   * 1. Separar el conjunto de datos en conjunto de entrenamiento y de prueba (\*\*)
     2. Usar los datos de entrenamiento para construir los siguientes modelos:
        1. Clasificador por regresión logística
        2. K-vecinos más cercanos (k=5)
        3. Análisis discriminante lineal
     3. Calcular la matriz de confusión sobre el conjunto de test para cada uno de los modelos desarrollados
     4. Calcular las métricas de desempeño precision, recall y F1-score
     5. Analizar los resultados e identificar: fortalezas y debilidades de cada técnica con respecto a este conjunto de datos. Comparar las técnicas y establecer cuál obtiene mejor desempeño