



**TECNOLÓGICO NACIONAL DE MÉXICO**

**INSTITUTO TECNOLÓGICO DE TIJUANA**

**SUBDIRECCIÓN ACADÉMICA**

Departamento de Sistemas y Computación

**EXAMEN**

Carrera: Ingeniería En Sistemas Computacionales/ Tecnologías de la información/ Informática

Materia: Datos Masivos

Unidad (es) a evaluar: Unidad 2

Catedrático: Jose Christian Romero Hernandez

Grupo:

Tipo de examen: Práctico

Firma del maestro:

Período: **Agosto-Diciembre 2021**

Salón:

Fecha:

Calificación:

Alumno: \_\_\_\_\_ No. Control: \_\_\_\_\_

**Instrucciones**

**Desarrollar las siguientes instrucciones en Spark con el lenguaje de programación Scala, utilizando solo la documentacion de la librería de Machine Learning Mllib de Spark y Google.**

1. Cargar en un dataframe **Iris.csv** que se encuentra en <https://github.com/jcromerohdz/iris>, elaborar la limpieza de datos necesaria para ser procesado por el siguiente algoritmo (**Importante, esta limpieza debe ser por medio de un script de Scala en Spark**) .
  - a. Utilice la librería Mllib de Spark el algoritmo de Machine Learning **multilayer perceptron**
2. ¿Cuáles son los nombres de las columnas?
3. ¿Cómo es el esquema?
4. Imprime las primeras 5 columnas.
5. Usa el metodo describe () para aprender mas sobre los datos del DataFrame.
6. Haga la transformación pertinente para los datos categoricos los cuales serán nuestras etiquetas a clasificar.
7. Construya el modelo de clasificación y explique su arquitectura.
8. Imprima los resultados del modelo

**Instrucciones de evaluación**

- Tiempo de entrega 4 días
- Al terminar poner el código y la explicación en la rama (branch) correspondiente de su github así mismo realizar su explicación de la solución en su google drive.
- Finalmente defender su desarrollo en un video de 8-10 min el cual servirá para dar su calificación, este video debe subirse a youtube para ser compartido por un link público.