Laboratorio 3.
Clasificación

INSTRUCCIONES:

Una cooperativa de productores de manzana ha recolectado, durante varias cosechas, múltiples observaciones con las siguientes características por cada manzana:

ID
Tamaño
Peso
Dulzura
Textura
Humedad
Madurez
Acidez
Calidad

La cooperativa está interesada en dtener un modelo, basado en estos datos, que le permita predecir la calidad de nuevas manzanas al recibir las características de las mismas. Inicialmente, solicitaron al estudiante extranjero, Siul Nalruf, que les hiciera el modelo pero, por falta de tiempo, lo único que logró fue hacer una limpieza básica de los datos y normalizarlos.

El objetivo de este laboratorio es que ustedes continúen con el desarrollo del modelo. Los datos se adjuntan a esta guía.

Deben hacer un análisis exploratorio para entender mejor los datos, sabiendo que el objetivo final es predecir si la calidad de la manzana es buena o mala. Recuerden explicar bien cada uno de los hallazgos que encuentren. La forma más organizada de hacer un análisis exploratorio es generando ciertas preguntas de las líneas que les parezca interesante investigar. Generen un informe en pdf con las explicaciones de los pasos que llevaron a cabo y los resultados obtenidos. Recuerden que la investigación debe ser reproducible, por lo que deben guardar el código que han utilizado para resolver los ejercicios y/o cada uno de los pasos llevados a cabo si utilizan una herramienta visual.
NOTA: Si trabajan con un notebook (.ipyjnb), no es necesario entregar un informe por aparte...pueden dejar toda la documentación requerida dentro del notebook.

Este laboratorio debe realizarse en **PAREJAS.** Para recibir una calificación, los miembros de la pareja deben estar inscritos en algún grupo de CANVAS.

DESCRIPCIÓN DEL CONJUNTO DE DATOS

Al estar normalizados los datos, no es mucho lo que pueden decir sobre los datos (ie. rangos de valores, etc).

EJERCICIOS

1. **Parte 1 – Análisis exploratorio de datos**
   1.1. Realicen una exploración rápida de sus datos. Para eso hagan un resumen de su conjunto de datos.
   1.2. Enuncien el tipo de cada una de las variables del conjunto de datos (cualitativa o categórica, cuantitativa continua, cuantitativa discreta)
   1.3. Incluyan los gráficos exploratorios, siendo consecuentes con el tipo de variable que están representando.
   1.4. Aíslen las variables numéricas de las categóricas, hagan un análisis de correlación entre las mismas.
   1.5. Utilicen las variables categóricas, haga tablas de frecuencia, proporción, gráficas de barras o cualquier otra técnica que le permita explorar los datos
   1.6. Realicen la limpieza de variables utilizando las técnicas vistas en clase, u otras que piense pueden ser de utilidad
2. **Parte II – Desarrollo de un modelo de Clasificación para determinar si la calidad de una manzana es buena o mala.**
   2.1. Prueben todos los modelos que se vieron en clase. Sí hay alguno que no consideran pertinente, expliquen porqué.
   2.2. ¿Cuál es el modelo con mejor rendimiento? Utilice las métricas vistas en clase para dar respaldo a su respuesta. Recuerden que también pueden afinar los hiperparámetros.

EVALUACIÓN

(30 puntos) **Análisis Exploratorio**:
- Identifican el tipo de archivo .data y la forma de leerlo
- Estudian las variables cuantitativas mediante técnicas de estadística descriptiva
- Hacen gráficos exploratorios como histogramas, diagramas de cajas y bigotes, gráficos de dispersión que ayudan a explicar los datos
- Analizan las correlaciones entre las variables, tratan de explicar los datos atípicos (outliers) y toman decisiones acertadas ante la presencia de valores faltantes.
- Estudian las variables categóricas
- Elaboran gráficos de barra, tablas de frecuencia y de proporciones
- Elaboran gráficos adecuados según el tipo de dato que representan
- Explican muy bien todos los procedimientos y los hallazgos que van haciendo.
- Realizan la limpieza de datos para que se tenga un conjunto de datos que permita usar modelos

(56 puntos) **Clasificación**
- Realizan modelos de todos los algoritmos vistos en clase, analizan los resultados para identificar el de mejor rendimiento. Si hay algún modelo que no se considere pertinente, dan explicaciones de porqué no.

(14 puntos) **Hallazgos y conclusiones**.

- Hacen un resumen de los hallazgos en el análisis exploratorio y los modelos de clasificación

MATERIAL A ENTREGAR

- Archivo .pdf con el informe de análisis exploratorio. **Si el código lo desarrollan dentro de un notebook .ipyjnb, el informe puede ser parte de este notebook.**
- Link de Google drive donde trabajó el grupo.
- Script de R (.r o .rmd) o de Python que utilizó para responder las preguntas con el código utilizado o archivo de flujo de trabajo de KNime
- Link de github o el versionador que se utilizó.