Laboratorio 2.
Análisis Exploratorio y Regresiones

INSTRUCCIONES:

Esta hoja de trabajo utiliza un conjunto de datos extraidos de la página https://www.baseball-reference.com/.

Este conjunto de datos no está limpio, por lo que se deben realizar algunos pasos de limpieza y preprocesamiento antes de poder analizarlo:

Debe hacer un análisis exploratorio para entender mejor los datos, sabiendo que el objetivo final es predecir la asistencia de aficionados al estadio. Recuerde explicar bien cada uno de los hallazgos que haga. La forma más organizada de hacer un análisis exploratorio es generando ciertas preguntas de las líneas que le parece interesante investigar. Genere un informe con las explicaciones de los pasos que llevó a cabo y los resultados obtenidos. Recuerde que la investigación debe ser reproducible por lo que debe guardar el código que ha utilizado para resolver los ejercicios y/o cada uno de los pasos llevados a cabo si utiliza una herramienta visual.

Este laboratorio debe realizarse en **PAREJAS.** Para que se pueda calificar su laboratorio debe estar inscrito en algún grupo de Canvas.

DESCRIPCIÓN DEL DATASET

El dataset contiene datos de 2463 partidos de baseball y 17 variables que describen diferentes aspectos de estos.

EJERCICIOS

1. **Parte 1** – Análisis exploratorio de datos
   - 1.1. Haga una exploración rápida de sus datos para ello, haga un resumen del conjunto de datos.
   - 1.2. Determine el tipo de cada una de las variables.
   - 1.3. Incluya los gráficos exploratorios siendo consecuentes con el tipo de variable que están representando.
   - 1.4. Aísle las variables numéricas de las categóricas, haga un análisis de correlación entre las mismas.
   - 1.5. Utilice las variables categóricas, haga tablas de frecuencia, proporción, gráficas de barras o cualquier otra técnica que le permita explorar los datos
   - 1.6. Realice la limpieza de variables utilizando las técnicas vistas en clase, u otras que piense pueden ser de utilidad

2. **Parte 2** – Pruebe todos los modelos de Regresión vistos en clase y encuentre el mejor modelo de ellos para predecir el número de asistentes a un partido. Para cada modelo:
   2.1. Siga los procedimientos vistos en clase para realizar una regresión con los datos dados
   2.2. ¿Cuál es el rendimiento de su modelo? Calcule el parámetro R^2 para dar respaldo a su respuesta
   2.3. Si es uno de los modelos lineales, obtenga las constantes del modelo y exprese la ecuación que representan
   2.4. Esta interesado en predecir cuál será la asistencia a un partido en el que se enfrenten X y Y equipos (Ud decide cuáles), así como el día de la semana, la hora y el estadio (también los decide Ud) y otras variables que exija su modelo. Para estos valores, ¿cuál es la predicción de la asistencia?

EVALUACIÓN

(45 puntos) Análisis Exploratorio:

- Estudia las variables cuantitativas mediante técnicas de estadística descriptiva
- Hace gráficos exploratorios como histogramas, diagramas de cajas y bigotes, gráficos de dispersión que ayudan a explicar los datos
- Analiza las correlaciones entre las variables, trata de explicar los outliers (puntos atípicos) y toma decisiones acertadas ante la presencia de valores faltantes.
- Estudia las variables categóricas
- Elabora gráficos de barra, tablas de frecuencia y de proporciones
- Elabora gráficos adecuados según el tipo de dato que representan
- Explica muy bien todos los procedimientos y los hallazgos que va haciendo.
- Realiza la limpieza de datos para que se tenga un conjunto de datos que permita usar modelos

(41 puntos) Regresiones

- Realiza los procedimientos necesarios para asegurar que los modelos realizados dén resultados aceptables
- Evalúa los modelos para ver su rendimiento
- Si aplica, interpreta los coeficientes de la ecuación resultante.

(14 puntos) Hallazgos y conclusiones.

- Hace un resumen de los hallazgos en el análisis exploratorio y el mejor modelo de regresión para este conjunto de datos

MATERIAL A ENTREGAR

- Informe de análisis exploratorio.
- 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.