Data Analysis and Intro to ML
Taller

Juan Felipe Acevedo Pérez

Diciembre 2022

Parámetros de entrega

I. El taller debe ser entregado de forma individual o en parejas, **vía correo electrónico**.

II. En caso de ser entregado de forma individual, se ha de enviar con el siguiente nombre: **T1_ 1erNombre_1erApellido**

III. En caso de ser entregado en parejas, se debe respetar la siguiente marcación: **T1_ 1erNombreEst1_1erApellidoEst1_Y_ 1erNombreEst2_1erApellidoEst2**

IV. El formato del taller debe ser **PDF** (es decir, **NO** se corrigen archivos en Word, Excel, etc.).

V. La ortografía y la redacción serán tenidas en cuenta al momento de la evaluación. Debe ser breve y puntual con las respuestas, **no se extienda innecesariamente**.

VI. **Todo ejercicio debe tener el desarrollo del procedimiento**, de lo contrario el punto será anulado. **El script elaborado debe ser adjuntado en formato .ipynb**.

VII. La entrega del taller se permite hasta las **11:59 pm** del **sábado 10 de diciembre** del 2022. Tras pasar este límite temporal, ningún taller será aceptado bajo ninguna circunstancia.

El incumplimiento de cualquiera de los parámetros de entrega, tendrán una afectación negativa en la nota del taller.

1. Análisis de datos (80 %)

Se realiza la selección de dos bases de datos. Para cada una de ellas encontrará ejercicios enfocados a preguntas concretas y, un punto, en el cuál deberá utilizar su creatividad para presentar información de forma visual que, a criterio propio, considere relevante e interesante; la información que utilice para ello, no debe ser exactamente la misma que empleó para responder a las preguntas (puede utilizarla, pero debe añadir más información o nuevos cálculos). *Estas preguntas tendrán un mayor peso que las preguntas concretas, no serán valoradas de igual manera.*

1.1. Base de datos 2017 NIS-Child (50 %)

Para este punto, verá los datos de vacunas de 2017 de los Centros para el Control y Prevención de Enfermedades (CDC) de Estados Unidos. Su archivo de datos se encuentra en el moodle Taller/NISPUF17.csv. Para realizar el ejercicio satisfactoriamente, se suministra una guía de usuarios de los datos, disponible en Taller/NIS-PUF17-DUG. Deberá mapear las variables en los datos para resolver las preguntas que se hacen.

I) Escriba una función que devuelva la proporción de niños en el conjunto de datos que: tenían una madre con niveles de educación menores a secundaria (< 12), secundaria (12), más de secundaria pero no graduado universitario (> 12) y título universitario.

II) Exploremos la relación entre ser alimentado con leche materna cuando era niño y recibir una vacuna contra la influenza estacional de un proveedor de atención médica. Devuelva una tupla del número promedio de vacunas contra la influenza para aquellos niños que sabemos que recibieron leche materna cuando eran niños y aquellos que sabemos que no.

III) Sería interesante ver si hay evidencia de un vínculo entre la efectividad de la vacuna y el sexo del niño. Calcular la razón del número de niños que contrajeron varicela pero fueron vacunados contra ella (al menos una dosis de varicela) versus los que fueron vacunados pero no contrajeron varicela. Devolver resultados por sexo.

IV) A partir de la información disponible, realice un análisis de su interés. Para dar respuesta a este punto, puede emplear métodos de agrupación, reformar, pivotear, etc. Se espera al menos dos gráficas y una tabla (si va a trabajar en latex consulte la función .to_latex() de pandas) **con su respectiva explicación y análisis** (de lo contrario, el punto no será válido).

1.2. De todas las universidades del mundo, ¿cuáles son las mejores? (30 %)

Clasificar universidades es una práctica difícil, política y controvertida. Hay cientos de diferentes sistemas de clasificación de universidades nacionales e internacionales, muchos de los cuales no están de acuerdo entre sí. Este conjunto de datos contiene tres rankings universitarios globales de lugares muy diferentes. Debe emplear para dar respuesta a las preguntas concretas el dataset del Center for World University Rankings (Taller/archive/cwurData.csv); sin embargo, para el último punto de esta sección, podrá emplear los datos suplementarios. Toda la información necesaria se encuentra disponible AQUÍ.

I) ¿Cuántos países de América Latina se encuentran en el ranking? ¿Cuántas universidades de Colombia?

II) Queremos crear una nueva columna llamada Rank_Level, donde las instituciones con clasificación mundial 1-100 se categorizan como primer nivel, aquellas con clasificación 101-200 son segundo nivel, clasificación 201-300 son tercer nivel y después de 301 es otras universidades del top. Tras realizar esta clasificación, compare las universidades por Rank_Level versus el país de las universidades en términos de puntaje general (promedio). Por facilidad, puede mostrar únicamente las 5 primeras filas en el entregable (utilizar df.head() para ello).

III) Encuentre el país que tiene el puntaje promedio máximo en el nivel universitario

superior de primer nivel (resultante de la clasificación realizada en el punto anterior). Para ello, les recomiendo consultar la función idxmax() de pandas).

IV) A partir de la información disponible, realice un análisis de su interés. Para dar respuesta a este punto, puede emplear métodos de agrupación, reformar, pivotear, etc. Se espera al menos dos gráficas y una tabla (si va a trabajar en latex consulte la función .to_latex() de pandas) **con su respectiva explicación y análisis** (de lo contrario, el punto no será válido).

Machine Learning (20 %)

Realice un modelo de Random Forest Classifier utilizando la información disponible AQUÍ. Evalúe su nivel de ajuste empleando el oob_score y explique el resultado.

Para facilidad de ustedes, pueden importar los datos con estas líneas de código:

```
# Sugerencia de código para almacenar los datos
names = ["classes","cap-shape","cap-surface","cap-color","bruises","odor",
"gill-attachment", "gill-spacing","gill-size","gill-color","stalk-shape",
"stalk-root","stalk-surface-above-ring","stalk-surface-below-ring",
"stalk-color-above-ring","stalk-color-below-ring","veil-type","veil-color",
"ring-number","ring-type","spore-print-color","population","habitat"]

data = read_csv('http://archive.ics.uci.edu/ml/machine-learning-databases/
mushroom/agaricus-lepiota.data', names=names)
```