Universidad Austral
Maestría en Ciencia de Datos
Laboratorio de Implementación I , 2024 Virtual

Fecha de liberación oficial: martes 19 de marzo 23:01

Tarea para el Hogar UNO

No intente hacer esta Tarea para el Hogar UNO antes de la fecha de liberación oficial, ya que este documento y los que dependen de él se irán actualizando hasta ese momento.

La asignatura *Laboratorio de Implementación I* es la más intensa de toda la Maestría en Ciencia de Datos de Universidad Austral, lo ha sido en forma ininterrumpida desde el inicio de la maestría en el año 2006.

Las Tareas para el Hogar tienen actividades que debe realizar ya que serán discutidas en la siguiente clase sincrónica de Zoom y serán el punto de partida a mejores métodos.

Iniciar la asignatura implica instalar todas las herramientas que utilizaremos en el laboratorio.

En general, en las Tareas para el Hogar hay tres secciones :

- Pasado
- Deseable
- Complementaria.

Primero está la *Sección Pasado*, tareas que usted *ya debería haber hecho* para estar al día, por si quizas faltó a alguna de las clases.
Luego está la *Sección Deseable*, ejercicios y problemas que TODOS los alumnos deberían hacer para estar al día, entender lo que se discute la siguiente clase y finalmente aprobar la asignatura.
Por último está la *Sección Complementaria*, que es para alumnos que dispongan principalmente de tiempo y motivación.

Usted debe venir a la segunda clase el martes 26 de marzo a las 18:00, con la tarea realizada, en particular la instalación de Google Cloud la cual es realmente desafiente para personas con bajo *computer literacy*.

Sección Pasado

1. Del Libro de la Asignatura realice por única vez todas las altas en plataformas e instalaciones de aplicaciones solicitadas en el
    - capítulo 3 Arranque en Frío

    solicite ayuda en Zulip en el stream `#v-Arranque en Frio` si se le presenta algun problema, especialmente con Jupyter Labs.

    Notas de color
    Jupyter se pronuncia en inglés de la siguiente forma https://www.youtube.com/watch?v=PyDYfrp9_ys donde las letras "py" se pronuncian distinto a Python
    https://www.youtube.com/watch?v=e_PMuMQ9F2U

    Kaggle se pronuncia de la siguiente forma https://www.youtube.com/watch?v=pNn4XTe6T9s

2. Del Libro de la Asignatura leer por única vez
    (tiempo estimado 10 minutos, dificultad muy baja)
    - capítulo 1 La Asignatura y la forma de evaluación

    preste atención al cronograma, entienda la posibilidades que posee de personalizar la forma en la que será evaluado

3. Si aún no lo ha hecho, ver los *dos primeros* videos correspondientes a la Clase1 que están en https://campusvirtual.austral.edu.ar/course/view.php?id=14206&section=1#tabs-tree-start
    - Primeros Pasos
    - Presentación del Problema

    (tiempo estimado 20 minutos a 1.5x , dificultad baja)

Los videos están alojados en la plataforma Vimeo, que en la configuración de UAustral no permite alterar la velocidad del video. Ver esos videos a velocidad normal es uno de los peores castigos a los que puede ser sometido un estudiante.
Por lo cual, instale esta extensión para Google Chome que presionando la tecla "d" aumenta la velocidad y con la tecla "s" la disminuye. https://chrome.google.com/webstore/detail/video-speed-controller/nffaoalbilbmmfgbnbgppjihopabppdk

Sección Deseable

4. Propuesta de Matrix de Ganancia
   Complete en las celdas con su nombre, en la Google Sheet Colaborativa de la asignatura la hoja "Matrix Ganancia" , los seis valores que a su mejor entender representan la verdadera ganancia de la campaña de marketing en el banco.

5. Instalación Google Cloud
   Prerequisito: *NA*
   En esta asignatura entrenaremos complejos modelos predictivos que demandarán de decenas de horas de procesamiento en grandes servidores en la nube.
   Primero debe tener una cuenta de gmail, y luego crear una cuenta de Google Cloud en donde le van a regalar USD 300 por 3 meses https://cloud.google.com/
   Allí es donde Google le va a pedir una tarjeta de crédito, le va a debitar un dólar y se los vuelven a acreditar, este paso lo realiza Google para evitar abusos.

   Seguir este instructivo https://storage.googleapis.com/open-courses/austral2024-fc72/GoogleCloud_labo2024v.pdff
   La instalación le demandará unas dos horas, de las cuales 60 minutos serán de forma desatendida.

   El profesor está a disposición para ayudar via Zulip, incluso por Jitsi Meet compartiendo pantalla a quienes sea sobrepasados por la situación.
   (tiempo estimado 120 minutos totales, dificultad alta)

6. Leer detenidamente el archivo `DiccionarioDatos` que se encuentra en
   https://storage.googleapis.com/open-courses/austral2024-fc72/DiccionarioDatos_2024.ods
   Esto le permitirá conocer los campos que posee el dataset y comenzar a ganar intuición sobre las variables más importantes que afectan la predicción y así poder crear variables derivadas, tarea llamada Feature Engineering en la Ciencia de Datos y que permite aumentar el poder predictivo de los modelos, en nuestro caso, aumentar la ganancia.
   Plantear dudas y observaciones en Zulip.
   (tiempo estimado 10 minutos, dificultad baja)

7. Si aún no lo ha hecho, ver el *tercer* video correspondientes a la Clase1 que está en
   https://campusvirtual.austral.edu.ar/course/view.php?id=14206&section=1#tabs-tree-start
   - Hiperparámetros de un Árbol de Decisión

   (tiempo estimado 10 minutos a 1.5x , dificultad baja)

8. Experimente con el script `src/rpart/z101_PrimerModelo.R` modificando los hiperparámetros de rpart `<minsplit, minbucket, maxdepth>` e intente quedar primero en Kaggle.
   Recuerde que Kaggle limita a 20 submits diarios, el corte es a las 21:00 hora Buenos Aires
   (tiempo estimado 20 minutos, dificultad baja)

Sección Complementaria

9. Lea https://en.wikipedia.org/wiki/Stratified_sampling para entender que es dividir el dataset en <Training, Testing> de forma estratificada en la clase.

   Lea con detalle el script `src/rpart/z111_traintest_estratificado.r` , córralo linea a linea desde RStudio, entienda en profundidad lo que hace.
   En el Check In que usted realizó a la asignatura anterior en Zulip, usted cargó 5 numeros primos, esas son sus semillas.
   Pruebe correrlo con cada una de sus cinco semillas aleatorias, cambiándolas en la linea 8
   Carge en la Google Sheet Colaborativa, hoja `z111` , *solo el resultado de la corrida con su primer semilla.*

   ¿Esperaba la variabilidad que observó? ¿Cómo compara con respecto a sus compañeros?
   (tiempo estimado 15 minutos, dificultad media)

10. Ver el video correspondiente a la Clase2 que están en
    https://campusvirtual.austral.edu.ar/course/view.php?id=14206&section=2#tabs-tree-start
    1. **Optimización** de Hiperparámetros en un Arbol de Decisión
    (tiempo estimado 9 minutos a 1.5x, dificultad media)

11. En el lenguaje R tradicional se utiliza el objeto `dataframe` que posee severas deficiencias en cuanto a performance y sintaxis complicada. En la materia utilizamos la librería `data.table` que es ampliamente superadora y que permite manejar grandes volúmenes de datos, tiene una sintaxis más simple PERO muy distinta a la de dataframes, por lo que es necesario aprenderla.
    Leer los siguientes artículos
    https://towardsdatascience.com/data-table-rs-best-data-object-c95b7d5f0104
    https://towardsdatascience.com/blazing-fast-data-wrangling-with-r-data-table-de5045cc4b4d

    Si realiza alguna prueba por su cuenta, haga los comentarios en Zulip, de igual forma si es un férreo defensor de tidyverse , dplyr, p Pandas en Python ¡ No le tenemos miedo !

    (tiempo estimado 15 minutos, dificultad baja)