

## Tarea 2

Predicción de hongos venenosos

**Profesor: Juan D. Velásquez**

Auxiliares: José Saffie V. y José Soza C.

Ayudantes: Mariela Palacios, Michelle Avendaño,  
Tomas Diaz, Claudia Navarro y Fernanda Saavedra

### Descripción del problema

El emperador de los túneles, al ver tan magnífico y espléndido trabajo que realizaron al modelar su sistema de Casinos, decide faltar a su promesa de dejarles libres y los obliga deliberadamente a ejecutar un nuevo estudio, esta vez ocupando análisis exploratorio de datos y Machine Learning.

Le comenta que, al ser una sociedad oculta bajo tierra, han tenido problemas significativos con distintas especies de hongos que brotan en las superficies de la cavernas. Su mentalidad empresarial e industrial, les lleva a aprovecharse de la situación y ocupar estos hongos para sus restaurantes, no obstante, algunos de ellos son venenosos y le piden a usted que pueda diseñar un modelo de predicción con tal de poder saber si pueden ocupar este preciado ingrediente o no.

Para ello, le proporcionan una base de datos con 11 columnas distintas, las cuales se encuentran detalladas en el diccionario anexado al final de este enunciado. De usted depende poder ayudar a esta sociedad secreta y poder volver a ver la luz del sol.

### Preguntas

**P1.-** Realice un análisis exploratorio de los datos (EDA), aplicando cada una de las herramientas básicas y avanzadas vistas en clases y clases auxiliares. Además, aplique cada uno de los pasos de la metodología de Feature Engineering, es decir, Feature Cleaning, Feature Generation y Feature Reduction, justificando las decisiones que toma en cada etapa. Luego, indique sus principales conclusiones obtenidas con cada herramienta. **(3 ptos.)**

**P2.-** Construya y evalúe dos modelos de clasificación para el problema de predecir si un hongo es o no venenoso. Indique claramente, en base al trabajo realizado en la pregunta anterior, cuáles serán las variables independientes y dependiente de sus modelos. Finalmente, indique cuál modelo es mejor y por qué. **(3 ptos.)**

## Reglas del juego

- Debe realizar esta tarea con su grupo de trabajo, publicado a través de U-Cursos.
- La resolución de dudas vía canales virtuales será realizada de forma exclusiva a través del foro del curso.
- Debe trabajar exclusivamente utilizando lenguaje Python en un documento *notebook* de *Google Colab*. Para ello, genere una copia de la plantilla del documento de Colab disponible en este [link](#). Puede agregar las celdas de código y texto que estime conveniente, siempre y cuando mantenga la estructura de secciones previamente definida en la plantilla.
- Las conclusiones preliminares y finales, así como su trabajo en general, deben ser congruentes con el objetivo del problema de construir un modelo de clasificación.
- El formato de su entrega debe ser el de un informe autocontenido, es decir, debe contener una introducción al trabajo a realizar, además de conclusiones sobre sus resultados. Cada etapa de su procedimiento debe contener celdas de texto explicando qué está realizando. Para ello, aproveche el paradigma de *programación literaria* de Google Colab que le permite intercalar bloques de texto y código. Al respecto, **código y procedimientos no explicados serán evaluados con puntaje cero**.
- Puede usar los ejemplos de código vistos en las clases auxiliares.
- Su entrega debe consistir de el documento notebook `.ipynb`, además de un archivo `.txt` que contenga el link hacia su notebook en Google Colab, **en modo lector**.
- Fecha de entrega: Lunes 3 de junio a las 23:59

## Anexo: Diccionario

1. cap-diameter: Diámetro del píleo.
2. cap-shape: Forma del píleo. (cada número representa una forma distinta).
3. gill-attachment: Morfología del himenio. (cada número representa una morfología distinta).
4. gill-color: Color del himenio. (cada número representa un color distinto).
5. stem-height: Altura del tallo.
6. stem-width: Ancho del tallo.
7. stem-color: Color del tallo.
8. season: Estación del año (cada número representa una estación distinta).
9. discovery: Fecha exacta de agregada la fila al dataset.
10. continent: Continente de origen.
11. class: 1 si es venenosa, 0 si no.