Tarea 4 – Introducción a Data Science
Inferencia Estadística

Profesor: Brian Keith Norambuena, brian.keith@ucn.cl.
Ayudante: Matías Salas Villanueva, matias.salas@alumnos.ucn.cl.

Objetivos

1. Reconocer predictores estadísticamente significativos.
2. Diagnosticar un modelo de regresión lineal mediante técnicas

Entregables

- Un cuaderno interactivo *Jupyter Notebook* con todo el código fuente, resultados, y análisis.
- El cuaderno interactivo debe utilizar apropiadamente celdas de *markdown* y de código.

Fecha de Entrega

3 de octubre a las 23.59 por Campus Virtual con el formato indicado en la sección de Consideraciones de Evaluación y Administrativas.

Requerimientos Mínimos

Se describen a continuación los requerimientos mínimos para esta tarea.

1. Ajuste del Modelo [1.5 puntos]:
   - Cargue y divida el conjunto en conjuntos de entrenamiento y prueba en un 80%/20% respectivamente. **(+0.5 puntos)**
   - Ajuste un modelo de regresión lineal múltiple utilizando todas las variables predictoras disponibles para explicar la variable objetivo. **(+0.5 puntos)**
   - Calcule los valores ajustados (fitted values) y los residuos del modelo. **(+0.5 puntos)**
2. Inferencia Estadística [4.5 puntos]:
   - Diagnóstico de residuos [3 puntos].
     - Verifique las 4 suposiciones del modelo (Independencia de residuos, Homocedasticidad, Normalidad de residuos y Multicolinealidad) mediante gráficos o pruebas estadísticas. **(+0.5 c/u)**
     - Con base en los resultados de las verificaciones anteriores, responda: ¿Son confiables los resultados de inferencia del modelo? Explique por qué. **(+1 punto)**
   - Significancia de variables [1.5 puntos].
     - **En base a una significancia de 0.05** (es decir con una confianza del 95%), determine los predictores que no son estadísticamente significativos y elimínelos. **(+0.5 puntos)**
     - Ajuste nuevamente el modelo con las variables significativas y compare los coeficientes y errores estándar con el modelo original. **(+0.5 puntos)**
     - **Discuta** cómo la eliminación de variables no significativas afecta la interpretación del modelo y la confiabilidad de los valores-p restantes. **(+0.5 puntos)**

Descripción del conjunto de datos

Se debe hacer uso del conjunto de datos proporcionado wine_Tarea4.csv para la ejecución de esta tarea, a continuación, una breve descripción de cada columna:

- **fixed acidity**: Cantidad de acidez fija en la muestra, generalmente medida en gramos por litro.
- **volatile acidity**: Cantidad de acidez volátil, que se refiere a los ácidos que pueden evaporarse, como el ácido acético. Un nivel alto puede dar lugar a un aroma avinagrado.
- **citric acid**: Cantidad de ácido cítrico presente en la muestra, que puede aportar frescura y mejorar la estabilidad del vino.
- **residual sugar**: Cantidad de azúcar residual en gramos por litro después de la fermentación. Afecta el dulzor del vino.

- chlorides: Cantidad de cloruros (sales, principalmente cloruro de sodio) en la muestra, que pueden influir en el sabor y la estabilidad del vino.
- free sulfur dioxide: Cantidad de dióxido de azufre libre en la muestra, un compuesto utilizado como conservante para evitar la oxidación y el crecimiento de microorganismos.
- total sulfur dioxide: Cantidad total de dióxido de azufre en la muestra, incluyendo el SO_2 libre y el combinado con otros compuestos del vino.
- density: Densidad del vino, que puede indicar la concentración de azúcar, alcohol y otros compuestos.
- pH: Medida del nivel de acidez o alcalinidad del vino en una escala de 0 a 14. Valores más bajos indican mayor acidez.
- sulphates: Concentración de sulfatos en la muestra, que pueden influir en la percepción del sabor y en la conservación del vino.
- quality: Puntuación de calidad del vino basada en evaluaciones sensoriales y químicas.
- alcohol: Porcentaje de alcohol en volumen en la muestra, clave en el cuerpo y sabor del vino. (variable objetivo)

Consideraciones de Evaluación y Administrativas

- El formato de entrega será el siguiente "NombreApellido1_NombreApellido2.ipynb" (sin tildes) mediante plataforma de campus virtual.
  Ejemplo: PedroRojas_JosePerez.ipynb
- A partir de las 00:00 hrs del día siguiente del día de entrega comenzará un descuento de 10 décimas por cada hora de retraso, Ejemplo: Si el límite de entrega es el 29 de agosto a las 23:59 hrs y la tarea es entregada el 30 de agosto a las 01:00 hrs, tiene derecho a nota máxima 6.0.
- La escala de evaluación será al 60%.
- Para recibir el puntaje completo debe ser asegurarse de que el cuaderno pueda ser ejecutado de corrido, de lo contrario se aplicará la mitad del puntaje.
- Los análisis de resultados y correspondientes conclusiones serán evaluados según la claridad de sus ideas y su presentación (escritura y formato).
- Si utiliza Inteligencia Artificial generativa (e.g., ChatGPT) durante su trabajo, por favor incluya todas sus *prompts* al final del cuaderno interactivo como documentación o entregue un anexo con esta información. No habrá penalización por el uso de estas herramientas como apoyo, pero no deberían ser utilizadas para resolver la tarea por ustedes.
- Los Créditos Extras serán puntos extras que se sumarán al final del semestre a tareas que no tengan el puntaje completo, comenzando por la tarea con menor puntaje. Los créditos extra sobrantes se anularán si el estudiante tiene puntaje máximo en todas las tareas.