CI-0115 Probabilidad y Estadística
Prof. Sebastián Ruiz Blais
I-2024
Grupos 2 y 3

Proyecto – Entrega 3 (10%)
Fecha de entrega: *2 de julio del 2024*

Objetivo general
Profundizar el análisis de datos utilizando métodos de modelado y aprendizaje de máquina.

Objetivos específicos
1. Utilizar distribuciones continuas para modelar datos
2. Utilizar modelos no lineales para generar predicciones
3. Utilizar técnicas de aprendizaje de máquina para construir modelos y estimar parámetros

Especificación
- **(20%) Para cada una de las variables de interés, realizar pruebas de normalidad**, utilizando ya sea el test Shapiro-Wilk o Kolmogorov-Smirnov (ver tutorial). **Explicar** el resultado obtenido. En caso de que los datos no estén distribuidos normalmente, **aplicar una transformación a los datos para obtener una distribución normal**. En caso de no obtener una distribución normal, probar por lo menos con dos tipos de transformaciones y documentar el proceso. Cuando realice su presentación, agregue este paso antes de realizar los ANOVA's (que asumen una distribución normal). En caso de haber tenido que transformar los datos, vuelva a realizar los ANOVA's para comprobar que no hubo cambios. Si no se pudo obtener una distribución normal utilizando una transformación de datos, omita realizar nuevos ANOVA's.
- **(30%)** Las relaciones entre variables pueden no ser lineales. Por ejemplo, la temperatura ideal para la vida humana posiblemente tenga un rango óptimo. La esperanza de vida no está relacionada linealmente con la temperatura promedio, ya que si la temperatura fuera muy baja (p.ej., -30º), la esperanza de vida sería muy baja, y si la temperatura fuera muy alta (p.ej., 60º), la esperanza de vida también sería muy baja. Otros procesos que siguen un patrón similar son la relación entre el desempeño y el nivel de estrés. Hay un nivel de estrés óptimo para que una persona se desempeñe de manera óptima en una tarea. **Identifique en su base de datos por lo menos dos pares de variables que no tengan una relación lineal**. Utilice este tutorial para **realizar una regresión no lineal a sus datos**, incluyendo un **ajuste inicial y luego un ajuste optimizado**. Investigue en caso de necesitar otro tipo de función y explique el procedimiento.
- **(50%)** Cuando analizamos datos, muchas veces queremos realizar clasificaciones. Por ejemplo, podemos entrenar un modelo que determine la raza de un perro a partir de sus características o predecir si una persona va a realizar una donación a partir de sus características psicológicas. **Identifique en su base de datos una o más variables que**

desee predecir que no sea numérica. Por ejemplo, la raza de un perro, el hecho de realizar donación o no, la presencia o ausencia de una enfermedad, o la región del mundo en la que se encuentra. También puede utilizar categorías similares a las de ANOVA de la entrega anterior. Utilizando este tutorial, **investigue sobre el aprendizaje de máquina** (machine learning) para **crear y entrenar un modelo que utilice los valores de varias variables para predecir la o las variables anteriormente escogidas.** Procure entrenar el modelo con un subconjunto de los datos (típicamente 90% de los datos, seleccionados aleatoriamente) y realizar pruebas con otro subconjunto (típicamente el 10% restante). **Evalúe el porcentaje de predicciones correctas con este modelo**. Para ello, obtenga el promedio en varias ejecuciones (donde se entrena el modelo con datos distintos debido a la muestra aleatoria). **Optimice el modelo realizando algunas modificaciones**. Explique todas las etapas realizadas y documente el ajuste de los modelos generados. Para los modelos utilizados, puede utilizar redes neuronales artificiales u otras aproximaciones.

Evaluación

- *Calidad del análisis*: logran implementar cada uno de los objetivos del proyecto con una base de datos lo suficientemente rica. Realizan cada paso correctamente. Seleccionan variables de interés de manera correcta, incluyendo las variables predictoras y las variables a predecir.
- *Calidad de los modelos*: realizan pasos de manera satisfactoria para construir, entrenar y optimizar los modelos.
- *Calidad de las explicaciones*: explican los pasos realizados y las preguntas realizadas. Contestan algunas de las preguntas basándose en el análisis de los datos. Presentan los pasos de manera ordenada y usando gráficas y visualizaciones para mejorar las explicaciones. El análisis es coherente y sigue una lógica clara. Se responde a las preguntas planteadas.

Referencias

Test de normalidad
Regresión no lineal
Clasificación usando Machine Learning