MODELOS ESTADÍSTICOS LINEALES - IIND 4100
DEPARTAMENTO DE INGENIERÍA INDUSTRIAL
Periodo 2022-20. Profesor: Carlos Valencia

PROYECTO FINAL

- π El resultado de su análisis será evaluado a partir de una presentación.
- π Las presentaciones serán programadas para el martes 6 de diciembre a partir de las 4pm.
- π El proyecto puede ser realizado en grupos de máximo 3 personas.
- π Use procedimientos claros y rigurosos.

En el proyecto, cada grupo debe seleccionar una base de datos de su preferencia en donde el análisis se pueda realizar mediante un modelo de regresión lineal o un modelo lineal generalizado. El análsis debe responder a una pregunta de interés y se debe enmarcar dentro de un contexto, para poder interpretar los resultados. Pueden usar datos propios de algunos integrantes del grupo (o sus organizaciones), o datos públicos. Idealmente, el problema a resolver debe estar fundamentado en probar algún tipo de hipótesis (pueden ser varias pruebas) sobre el contexto de interés, o sobre la necesidad de explicar y/o predecir la variable de respuesta.

Los datos pueden ser públicos o privados y pueden ser obetenidos de cualquier fuente, siempre y cuando se respeten los derechos de autor y se den los créditos necesarios. Algunas fuentes para bases de datos libres son:

- Datos mundiales sobre problemas considerados de interés general para la humanidad (*Our World in Data*) recolectados y administrados por la Universidad de Oxford: `https://ourworldindata.org/`. En pariticular, hay muchos datos disponibles sobre el comportamiento del Covid-19, por ejemplo, `https://ourworldindata.org/grapher/covid-stringency-index`
- Estadísticas del DANE (Colombia): `https://www.dane.gov.co/index.php/estadisticas-por-tema`
- Repositorio para Machine Learnig de UC Irvine: `https://archive.ics.uci.edu/ml/index.php`
- Estadísticas oficiales USA: `https://www.data.gov/`
- Datos del Banco Mundial: `https://data.worldbank.org/`
- Datos de la Organización Mundial de la Salud: `http://apps.who.int/gho/data/node.home`

El análisis de los datos mediante modelos lineales debe tener en cuenta:

1. Análisis del contexto del problema y relevancia del análsis: Esto debe responder a la pregunta de por qué es interesante o importante estudiar los datos que seleccionan.
2. Entendimiento de los datos: análisis descriptivo de las variables, selección de variable de respuesta, selección de variables independientes, análsis gráfico.
3. Aplicación de modelo lineales: correcto uso de los modelos vistos en clase para estimación de parámetros de interés o respuesta a hipótesis de investigación.
4. Validación de los modelos propuestos: análisis riguroso de para saber si los resultados obtenidos son válidos.

El proyecto se evaluará mediante una presentación oral, en la cual deben interactuar con los evaluadores para responder preguntas que tienen que ver con el entendimiento de los datos y de los modelos usados.

Adicional a las fuentes suministradas, y otros que econsideren de interés, se propone un temas con su respectiva base de datos:

- Explicación y predicción de bancarrota: En este problema se busca explicar si una empresa (*reatil*) se va a bancarrota en siguiente año o no, dependiendo de la información contenida en los estados financieros. Corresponde a un modelo generalizado con respuesta binaria. Los datos de este proyecto serán suministrados bajo para los equipos interesados. Información sobre este análisis la puede consultar el artículo:
  Carlos Valencia, Sergio Cabrales, Laura Garcia, Juan Ramirez & Diego Calderona (2019). Generalized additive model with embedded variable selection for bankruptcy prediction: Prediction versus interpretation, Cogent Economics & Finance, 7:1, DOI: 10.1080/23322039.2019.1597956.

Un ejemplo de un estudio con objetivos reales e interesantes lo pueden ver encontrar en: Liang, LL., Tseng, CH., Ho, H.J. et al. Covid-19 mortality is negatively associated with test number and government effectiveness. Sci Rep 10, 12567 (2020). https://doi.org/10.1038/s41598-020-68862-x.