**Proyecto Final - Modelos Lineales**

**Maestría en Ciencia de Datos**

**Objetivo del Proyecto**

Aplicar modelos lineales y sus extensiones al análisis de un conjunto de datos real, con el fin de modelar, interpretar y comunicar relaciones entre variables predictoras y una variable de interés.

**Equipos: de 2 a 3 personas**

**Fases del Proyecto**

**1. Elección del conjunto de datos**

* Debe ser real y público (INEGI, Kaggle, UCI, OpenML, etc.)
* Mínimo 200 observaciones y 4 variables predictoras
* Variable respuesta: continua, categórica o de conteo

**2. Análisis exploratorio y limpieza**

* Estadísticos descriptivos
* Gráficos (distribuciones, correlaciones, boxplots)
* Manejo de NA, atípicos y transformaciones

**3. Modelado**

Aplicar **al menos 2 enfoques** distintos entre los siguientes:

* Regresión lineal múltiple
* GLM: logístico, Poisson u otra familia
* GAM con al menos una función suave
* Regresión no paramétrica (loess, splines)
* Análisis de varianza (si aplica diseño experimental)

**4. Evaluación y comparación de modelos**

* Validación cruzada, AIC, BIC, R^2, AUC, etc.
* Gráficos de residuos, predicciones vs reales, curvas ROC

**5. Informe final**

Debe presentarse en formato tipo artículo científico:

* Introducción y objetivos
* Descripción del conjunto de datos
* Metodología y modelos ajustados
* Resultados y discusión
* Conclusiones
* Anexos: código, figuras y tablas

Entrega: archivo PDF + código fuente (Rmd, Quarto o script R limpio)

**Recursos sugeridos**

* Librerías: ggplot2, dplyr, mgcv, MASS, car, performance, caret
* Conjuntos de datos: Kaggle, UCI, TidyTuesday, INEGI, datos.gob.mx, World Bank

**Fecha de entrega sugerida:** 29 de julio presentación del proyecto por equipos y entrega del archivo. En caso de haber cambios, se podrá entregar el archivo final el día 1 de agosto a más tardar.

**Entrega en formato PDF + código reproducible.**

**Rúbrica de Evaluación**

| **Criterio** | **Puntaje** | **Descripción** |
| --- | --- | --- |
| Selección y descripción del conjunto de datos | 10 pts | Datos adecuados, bien contextualizados. No ejemplos típicos. |
| Limpieza y análisis exploratorio | 15 pts | Buen manejo de valores faltantes, gráficos y estadísticas |
| Aplicación de múltiples modelos | 25 pts | Incluye al menos 2 enfoques apropiados |
| Interpretación de resultados | 20 pts | Coherente, con foco en inferencia y relación con la pregunta de interés |
| Evaluación y comparación de modelos | 15 pts | Uso adecuado de métricas y validación cruzada |
| Presentación del informe | 10 pts | Redacción clara, bien estructurado, visualmente organizado |
| Claridad y reproducibilidad del código | 5 pts | Código limpio, comentado y funcional |
| **Total** | **100 pts** |  |

**Nota: Para evitar uso de las mismas bases de datos, en el chat del grupo, cada equipo subirá la base de datos con la que trabajará. Las bases de datos que sean declaradas a utilizar no podrán usarla el resto de los equipos**