Universidad de Costa Rica
Facultad de Ciencias
Escuela de Matemática
Departamento de Matemática Pura y Ciencias Actuariales

HERRAMIENTAS DE CIENCIA DE DATOS I

CA0204

Proyecto Individual

Estudiantes:

Zúñiga Mora Andrés

C387733

Diciembre de 2025

Reporte ejecutivo del proyecto individual

1. Contexto y motivación

El curso CA-0204 *Herramientas de Ciencia de Datos I* propone, como proyecto individual, el desarrollo de un trabajo aplicado en R que integre manipulación de datos, modelado estadístico y comunicación de resultados mediante un reporte ejecutivo. En este proyecto se construyó, a modo de caso de estudio, un modelo de riesgo a 10 años de mortalidad basado en datos de salud poblacional, con el objetivo de ilustrar un flujo de trabajo completo en R: desde la obtención de datos crudos hasta la generación de gráficos y tablas interpretables para una audiencia no necesariamente técnica.

Para ello se utilizaron los datos de la Encuesta Nacional de Salud y Nutrición de Estados Unidos (NHANES), enlazados con los archivos públicos de mortalidad (*Linked Mortality Files*), que permiten seguir a las personas encuestadas y determinar si fallecieron durante el periodo de seguimiento y en qué momento ocurrió el evento. El resultado de interés es la mortalidad por cualquier causa (*all-cause mortality*) a 10 años, medida como un tiempo de supervivencia en días y un indicador de evento (1 = falleció durante el seguimiento, 0 = censurado).

2. Datos y construcción del conjunto analítico

2.1. Fuente de datos

NHANES es una encuesta por conglomerados, representativa de la población adulta de Estados Unidos, que combina entrevistas, mediciones físicas y exámenes de laboratorio. En este proyecto se emplearon los ciclos 1999–2014, enlazados con archivos de mortalidad que reportan tiempo de seguimiento y estado vital (vivo/fallecido) hasta una fecha de corte definida por los Centros para el Control y la Prevención de Enfermedades (CDC).

2.2. Variables y predictores clínicos

El script `01_build_predictors_ALL.R` construye dos objetos centrales:

- **Núcleo de supervivencia** (`nhanes_surv_core_9914.rds`): combina la mortalidad enlazada con un núcleo demográfico mínimo por persona (`seqn`, sexo, edad), filtrando observaciones con sexo, edad, tiempo de seguimiento y estado vital no faltantes.
- **Predictores "ricos"** (`nhanes_pred_all_full_9914.rds`): para cada ciclo de NHANES se extraen y armonizan variables clínicas relevantes:
  - Sexo biológico (codificado como 1 = hombres, 0 = mujeres).
  - Edad (años).
  - Presión arterial sistólica media (promedio de tres mediciones, `sbp`).
  - Colesterol total (`chol`).

- Colesterol HDL (`hdl`).
- Índice de masa corporal (`bmi`).
- Indicadores binarios: tabaquismo actual (`smoker`), diagnóstico de diabetes (`diabetes`) y tratamiento para hipertensión (`htn_treated`).

Dado que NHANES se organiza por ciclos bienales con estructuras de tabla ligeramente distintas, el código implementa una estrategia robusta de lectura:

- Uso de `nhanesA::nhanes()` vía la función auxiliar `nh_get()` con manejo de errores; si una tabla no existe en un ciclo, se retorna `NULL`.
- Búsqueda de las variables clave en diferentes tablas candidatas (por ejemplo, colesterol total en `L13`, `TCHOL` o `LAB13`), deteniéndose cuando se encuentra una columna compatible.
- Normalización de variables categóricas (sexo, tabaquismo, diabetes) a codificaciones numéricas consistentes en todos los ciclos.

Finalmente, se construyen dos capas de información: un conjunto "core" con las variables mínimas de supervivencia y un conjunto ampliado de predictores clínicos, que luego se unen mediante el identificador `seqn` en los scripts posteriores.

3. Metodología de modelado

3.1. Modelo de riesgos proporcionales de Cox

Para modelar el tiempo hasta la muerte se empleó el modelo semiparamétrico de riesgos proporcionales de Cox, ampliamente utilizado en análisis de supervivencia. En este modelo, la tasa de riesgo instantánea para una persona con covariables X se expresa como

$$\lambda(t \mid X) = \lambda_0(t) \exp(\beta^\top X),$$

donde $\lambda_0(t)$ es la función de riesgo basal común a todos los individuos y β es el vector de coeficientes que mide el efecto multiplicativo de cada covariable sobre el riesgo relativo.

Este enfoque es especialmente apropiado para modelos de riesgo cardiovascular de 10 años, como los derivados del estudio de Framingham, que combinan edad, presión arterial, lípidos, tabaquismo y diabetes en una función de riesgo absoluto.

3.2. Ajuste por sexo y modelo combinado

El script `02_fit_cox_ALL.R` realiza los siguientes pasos:

1. Convierte el tiempo de seguimiento de meses a días (`time_days`) a partir de la variable `time_mo` del núcleo de supervivencia.

2. Une el núcleo con los predictores clínicos para construir tres conjuntos:
   - Hombres (sexo = 1),
   - Mujeres (sexo = 0),
   - Muestra combinada (ambos sexos).
3. Implementa un filtro de calidad de variables opcionales (`sbp`, `chol`, `hdl`, `bmi`, `smoker`, `diabetes`, `htn_treated`) mediante la función `pick_vars()`:
   - Exige al menos un 40% de datos no faltantes en la muestra correspondiente.
   - Descarta covariables con un único valor distinto (sin variabilidad).

   De este modo, el modelo de Cox evita incluir predictores muy incompletos o degenerados.
4. Ajusta un modelo de Cox para hombres, otro para mujeres y un modelo combinado que incluye el indicador de sexo como covariable adicional.
5. Calcula, para cada modelo:
   - La función de riesgo basal acumulada mediante `basehaz()`.
   - La supervivencia basal a 10 años $S_0(10\ años)$ evaluada en 3652.5 días.
   - El índice de concordancia (C-index) como medida de discriminación, a partir de `survival::concordance()`.
6. Exporta los coeficientes, razones de riesgo (*hazard ratios*) e intervalos de confianza a archivos `CSV`, junto con el valor de $S_0(10\ años)$ de cada modelo.

El C-index resume la capacidad del modelo para asignar un mayor riesgo a quienes efectivamente fallecen antes; valores cercanos a 0.5 indican discriminación nula (equivalente a azar), mientras que valores más altos reflejan mejor capacidad predictiva.

3.3. Función de riesgo a 10 años y `scoring`

El script `03_score.R` empaqueta los modelos ajustados en una función de riesgo a 10 años:

- `risk_10y(sex, age, sbp, chol, hdl, bmi, smoker, diabetes, htn_treated)` devuelve la probabilidad estimada de fallecer en los próximos 10 años para un perfil clínico dado.
- Internamente, se corrige cada covariable restando la media usada en el ajuste (*centrado*) y se aplica la fórmula estándar de riesgo absoluto:

$$Risk(10\ años) = 1 - S_0(10\ años)^{\exp(L)},$$

donde L es el predictor lineal del modelo de Cox.

- Se dispone de un modelo específico por sexo (si está disponible) y de un modelo combinado de respaldo que incluye el indicador de sexo cuando uno de los modelos específicos no se pudo ajustar.

La función `predict_risk_df()` extiende el cálculo de riesgo a un `data.frame` completo, lo que permite generar columnas de riesgo individual (`risk10`) para todos los participantes elegibles.

4. Exploración de datos y visualización

4.1. Análisis exploratorio de biomarcadores

El script `08_eda_relations.R` se centra en la relación entre edad, sexo y los principales biomarcadores (presión arterial, colesterol total y HDL, IMC). Los pasos clave son:

- Unir el núcleo de supervivencia y los predictores, manejando de forma cuidadosa las columnas duplicadas (`.x`, `.y`) y los posibles cambios de tipo (numérico, entero, factor).
- Construir un `data.frame` "canónico" con variables continuas (`age`, `sbp`, `chol`, `hdl`, `bmi`) y binarias (`sex`, `smoker`, `diabetes`, `htn_treated`).
- Generar un conjunto de figuras en `ggplot2`, entre ellas:
  - Distribución de edad por sexo (violín + caja).
  - Dispersión de colesterol versus edad, con suavizado por sexo.
  - Gráficos de densidad y diagramas de caja de los biomarcadores por sexo.
  - Matriz de correlaciones entre variables numéricas.

Estas figuras permiten evaluar visualmente patrones epidemiológicos plausibles, como el aumento de presión arterial con la edad o diferencias sistemáticas por sexo en colesterol o IMC, así como detectar posibles problemas de calidad de datos (valores extremos, colas pesadas, etc.).

4.2. Curvas de riesgo y supervivencia

El script `09_risk_plots.R` conecta directamente el modelo de Cox con gráficos interpretables:

- Barras de riesgo promedio a 10 años por bandas de edad (en intervalos de 10 años) y sexo, a partir de la columna `risk10`.
- Curvas de riesgo a 10 años según la edad y distintos niveles de presión arterial sistólica (por ejemplo, 120, 140 y 160 mmHg), estratificadas por sexo. Estas curvas ilustran cómo el incremento de la presión arterial desplaza hacia arriba el riesgo estimado en todo el rango etario.

- Curvas de supervivencia basal $S_0(t)$ por sexo, obtenidas de los modelos de Cox, y curvas de Kaplan–Meier empíricas. La comparación entre ambas curvas es útil para detectar desviaciones importantes respecto al supuesto de riesgos proporcionales o mala especificación del modelo.

Las figuras se guardan como archivos PNG en la carpeta `output/`, lo que facilita su inclusión posterior en el reporte o en la presentación del proyecto.

4.3. Calibración por deciles de riesgo

Finalmente, el script `10_risk_deciles_plots.R` evalúa la calibración del modelo a través de deciles de riesgo:

- Se calcula, para cada persona, el riesgo a 10 años (`risk10`) y el indicador de evento a 10 años (`y10`), definido como muerte ocurrida dentro de 3652.5 días desde la entrevista.
- Se forman deciles globales de riesgo y se computan, para cada decil, la tasa observada de eventos y el riesgo promedio predicho.
- Se producen tablas y gráficos de barras y líneas, tanto globales como estratificados por sexo, comparando *observado vs. predicho* en cada decil.

Este tipo de análisis, recomendado en la literatura de modelos de riesgo, permite verificar si el modelo tiende a sobrestimar o subestimar sistemáticamente el riesgo en distintos estratos de la población.

5. Discusión y conclusiones

Desde la perspectiva del curso, el proyecto logra integrar varios componentes fundamentales de la ciencia de datos aplicada:

- **Obtención y limpieza de datos:** se trabajan datos reales, con problemas típicos de archivos de salud pública (múltiples fuentes por ciclo, nombres de variables que cambian, valores faltantes y codificaciones heterogéneas). La construcción del núcleo de supervivencia y de los predictores clínicos ilustra el uso intensivo de `dplyr`, `purrr` y funciones auxiliares robustas.
- **Modelado estadístico en R:** el ajuste de modelos de Cox por sexo y combinado muestra cómo traducir un problema clínico (estimar riesgo a 10 años) a una formulación de supervivencia, obteniendo tanto parámetros interpretables (razones de riesgo) como funciones de riesgo absoluto.
- **Validación y comunicación de resultados:** el uso de C-index, curvas de supervivencia, gráficos de riesgo por edad y análisis de calibración por deciles ejemplifica buenas prácticas para evaluar modelos predictivos, más allá de reportar únicamente coeficientes o p-valores.

- **Automatización y reproducibilidad:** la organización en scripts numerados (`01_...` a `10_...`), el uso sistemático de lectura/escritura de archivos `RDS` y `CSV`, y la generación automática de figuras en `output/` favorecen un flujo de trabajo reproducible y fácil de extender.

Como trabajo futuro, el mismo flujo podría ampliarse en varias direcciones:

- Incorporar pesos muestrales y diseño complejo de NHANES para obtener estimaciones más representativas de la población.
- Comparar el desempeño del modelo con funciones de riesgo existentes, como los scores derivados del estudio de Framingham u otros modelos recientes de riesgo cardiovascular.
- Explorar modelos alternativos (por ejemplo, árboles de supervivencia, *random survival forests* o modelos de riesgo flexibles) y evaluar si mejoran la discriminación o la calibración.
- Desplegar una aplicación interactiva en Shiny que permita a una persona usuaria ingresar sus variables clínicas y obtener una estimación personalizada del riesgo a 10 años, reforzando la conexión con la parte de reportería interactiva del curso.

En resumen, el proyecto demuestra que, con un conjunto relativamente pequeño de variables clínicas y una implementación cuidadosa en R, es posible construir un modelo de riesgo de mortalidad a 10 años sobre datos reales y complejos, reforzando las competencias de programación, análisis y comunicación que persigue el curso.

6. Referencias

References

[1] National Center for Health Statistics. (2019). *Continuous NHANES public-use linked mortality files*. Hyattsville, MD: Centers for Disease Control and Prevention.

[2] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, 34(2), 187–220.

[3] D'Agostino, R. B., Sr., Vasan, R. S., Pencina, M. J., et al. (2008). General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*, 117(6), 743–753.

[4] Crowson, C. S., Atkinson, E. J., & Therneau, T. M. (2013). Assessing calibration of prognostic risk scores. *Statistical Methods in Medical Research*, 22(2), 169–187.