Análisis de Supervivencia en Enfermedades Cardíacas

Oscar Hernández - Erik Becerra - Camila Patiño

2025-03-15

Contents

INTRODUCCIÓN

El presente estudio tiene como objetivo usar el data set de ecocardiogramas disponible en el UCI Machine Learning Repository para analizar la relación entre diferentes variables clínicas y la probabilidad de supervivencia de los pacientes. Se emplearán técnicas de minería de datos, ciencia de datos y algoritmos de aprendizaje automático, como los bosques aleatorios, para generar modelos predictivos que permitan anticipar correctamente el pronóstico de los pacientes basados en sus datos ecocardiográficos.

DESCRIPCION DEL PROBLEMA

Las enfermedades cardiovasculares (ECV) son la principal causa de muerte a nivel mundial. Hallar estos problemas en el corazón, como la disfunción del ventrículo izquierdo, es de vital importancia para mejorar la calidad de vida y reducir la mortalidad. Partiendo de allí, los ecocardiogramas proporcionan información elemental sobre la estructura y el funcionamiento del corazón, permitiendo identificar posibles anomalías.

JUSTIFICACIÓN

Las enfermedades del corazón representan un problema de salud pública global. Según la Organización Mundial de la Salud (OMS) , las ECV son responsables de aproximadamente 17.9 millones de todas las muertes a nivel mundial por año[https://www.who.int/es/health-topics/cardiovascular-diseases#tab=tab_1]. La posibilidad de desarrollar modelos predictivos basados en datos médicos puede facilitar el diagnóstico temprano y la toma de decisiones clínicas, mejorando así los tratamientos y reduciendo la mortalidad.

El uso de técnicas de minería de datos en el ámbito de la salud permite encontrar patrones ocultos en los datos clínicos que podrían pasar desapercibidos para los especialistas. Un modelo predictivo efectivo podría ser utilizado como una herramienta complementaria en la evaluación de pacientes con enfermedades cardiacas, optimizando los recursos hospitalarios y mejorando la eficiencia en los tratamientos.4

OBJETIVOS

Objetivo General

Desarrollar un modelo predictivo basado en técnicas de aprendizaje automático para estimar la probabilidad de supervivencia de pacientes con enfermedades cardiovasculares, utilizando el dataset de ecocardiogramas del UCI Machine Learning Repository, con el fin de mejorar la toma de decisiones clínicas y optimizar el diagnóstico temprano.

Objetivos Específicos

- Realizar un análisis exploratorio de datos para identificar patrones, distribuciones, valores atípicos y posibles problemas en el dataset.
- Preprocesar los datos mediante la limpieza, imputación de valores faltantes y conversión de tipos de datos para garantizar la calidad del análisis.
- Evaluar la correlación entre variables clínicas y ecocardiográficas para determinar su influencia en la supervivencia de los pacientes.
- Desarrollar y entrenar modelos de aprendizaje automático, incluyendo bosques aleatorios, para predecir la probabilidad de supervivencia.
- Comparar el desempeño de distintos modelos predictivos mediante métricas de evaluación como precisión, sensibilidad, especificidad y área bajo la curva ROC.
- Validar y ajustar el modelo final para maximizar su precisión y minimizar sesgos en la predicción de la supervivencia de los pacientes.
- Proponer recomendaciones basadas en los hallazgos obtenidos para su posible aplicación en entornos clínicos y hospitalarios.

PREGUNTAS DE INVESTIGACIÓN

Para guiar el análisis y la construcción de modelos predictivos, se plantearon las siguientes preguntas de investigación:

1. ¿Cuáles son los factores más influyentes en la predicción de la supervivencia de los pacientes con problemas cardiacos según el dataset de ecocardiogramas?

2. ¿Qué tan preciso puede ser un modelo basado en bosques aleatorios para predecir la probabilidad de supervivencia de los pacientes?

3. ¿Existe una relación significativa entre las mediciones del ecocardiograma y la probabilidad de mortalidad en los pacientes analizados?

4. ¿Cómo se comparan los modelos de aprendizaje automático con los diagnósticos tradicionales en la predicción de la supervivencia de los pacientes?

METAS, PLANTEAMIENTOS E HIPÓTESIS

Metas

- Desarrollar un modelo predictivo de alta precisión para estimar la probabilidad de supervivencia de pacientes con enfermedades cardiovasculares.
- Identificar los factores clínicos y ecocardiográficos más relevantes en la predicción de la supervivencia.
- Evaluar diferentes técnicas de aprendizaje automático para seleccionar la más efectiva.
- Mejorar la calidad de los datos mediante técnicas de preprocesamiento. Presentar hallazgos que puedan ser útiles en entornos clínicos y hospitalarios.

Planteamientos

- El uso de modelos de aprendizaje automático en el análisis de ecocardiogramas puede mejorar la predicción de la supervivencia de los pacientes en comparación con métodos tradicionales.
- Existen variables clave en los ecocardiogramas que tienen mayor impacto en la probabilidad de supervivencia.
- La calidad de los datos afecta significativamente la precisión de los modelos predictivos.
- La combinación de técnicas estadísticas y algoritmos de aprendizaje automático puede generar mejores resultados en el diagnóstico clínico.

Hipótesis

Hipótesis Principal

Un modelo de aprendizaje automático bien entrenado puede predecir con alta precisión la probabilidad de supervivencia de los pacientes basándose en datos de ecocardiogramas.

Hipotesis Secundarias

- Algunas variables clínicas y ecocardiográficas tienen una mayor correlación con la supervivencia, por lo que la selección de características mejorará la precisión del modelo.
- La implementación de técnicas de preprocesamiento de datos, como la imputación de valores faltantes y la normalización, mejorará el rendimiento del modelo predictivo.
- Los modelos de aprendizaje automático, como los bosques aleatorios, proporcionarán un mejor desempeño predictivo en comparación con modelos más simples como la regresión logística.

DESCRIPCIÓN DE LA DATA

Nombre y Fuente del Dataset

- Nombre: Echocardiogram
- DatasetFuente: UCI Machine Learning Repository
- Enlace: https://archive.ics.uci.edu/dataset/38/echocardiogram

El data set de ecocardiogramas contiene información clínica y resultados de exámenes ecocardiográficos de pacientes con enfermedades cardiovasculares. Su objetivo principal es predecir la probabilidad de supervivencia de los pacientes con base en diferentes factores clínicos y ecocardiográficos.

Variables Disponibles y su Significado

El dataset contiene varias variables relacionadas con el estado de salud de los pacientes y los resultados de sus ecocardiogramas. Algunas de las variables clave incluyen:

- Survival: Días de supervivencia después del ecocardiograma.
- Still_alive: Indicador binario (1: el paciente sigue vivo, 0: el paciente falleció).
- Age_at_heart_attack: Edad del paciente en el momento del infarto.
- pericardial_effusion: Indica la presencia de derrame pericárdico (1: sí, 0: no).
- fractional_shortening: Indicador de la función del ventrículo izquierdo.
- epss (E-point septal separation): Medición de la separación del septo y la válvula mitral.
- lvdd (Left Ventricular End-Diastolic Dimension): Diámetro del ventrículo izquierdo en diástole.
- wall_motion_score: Puntaje de la movilidad de la pared del corazón.
- wall_motion_index: Índice calculado a partir del puntaje de movilidad de la pared.
- mult (Multiplying factor for wall motion index): Factor de ajuste del índice de movilidad.
- group: Clasificación del paciente en un grupo específico.
- alive_at_1_year: Indica si el paciente sigue vivo un año después del infarto (1: sí, 0: no).

Identificación de Posibles Problemas Iniciales en los Datos

Al tratarse de un data set médico, pueden presentarse diversos problemas en los datos:

- Valores faltantes: Algunas variables contienen datos ausentes, lo que puede afectar el análisis y el modelado.
- Ruido en los datos: Algunas mediciones pueden contener errores debido a la variabilidad en la toma de datos médicos.
- Distribución desigual de las clases: Si la cantidad de pacientes que sobrevivieron es mucho mayor o menor que la de aquellos que no lo hicieron, podría generar sesgo en el modelo predictivo.
- Valores atípicos: Puede haber registros con valores extremos que distorsionen el análisis.
- Conversión de tipos de datos: Algunas variables numéricas pueden estar almacenadas como texto y requerir conversión.

RESUMEN ESTADÍSTICO

Carga y Visualización de Datos

Importación del Dataset

```
library(readr)
library(dplyr)


# Cargar librerías necesarias library(readr) library(dplyr)
datos <- read_delim("C:\\Users\\12114\\OneDrive\\Documentos\\ITQ\\Semestre 8 (Bogotá)\\Minería de Datos\

# Mostrar las primeras filas del dataset
head(datos)
```

```
## # A tibble: 6 x 12
##   survival still_alive age_at_heart_attack pericardial_effusion
##   <chr>          <dbl> <chr>                              <dbl>
## 1 11                 0 71                                     0
## 2 19                 0 72                                     0
## 3 16                 0 55                                     0
## 4 57                 0 60                                     0
## 5 19                 1 57                                     0
## 6 26                 0 68                                     0
## # i 8 more variables: fractional_shortening <chr>, epss <chr>, lvdd <chr>,
## #   wall_motion_score <chr>, wall_motion_index <chr>, mult <chr>, group <chr>,
## #   alive_at_1_year <chr>
```

```
# Identificar la cantidad de filas y columnas
dim(datos) # Devuelve el número de filas y columnas
```

```
## [1] 131  12
```

Calcular la Media, Mediana y Desviación Estándar de las Variables Numéricas Principales

Carga de Librerías

```
library(readr)
```

Carga del Dataset

```
df <- read_delim("C:\\Users\\12114\\OneDrive\\Documentos\\ITQ\\Semestre 8 (Bogotá)\\Minería de Datos\\e
```

Verificación de Columnas

```
print(colnames(df))
```

```
##  [1] "survival"              "still_alive"           "age_at_heart_attack"
##  [4] "pericardial_effusion"  "fractional_shortening" "epss"
##  [7] "lvdd"                  "wall_motion_score"     "wall_motion_index"
## [10] "mult"                  "group"                 "alive_at_1_year"
```

Limpieza de Nombres de Columnas

```
colnames(df) <- trimws(colnames(df))
```

Conversión de Columnas Numéricas

```
numeric_cols <- c("survival", "still_alive", "age_at_heart_attack", "fractional_shortening",
                  "epss", "lvdd", "wall_motion_score", "wall_motion_index", "mult", "alive_at_1_year")
numeric_cols <- intersect(numeric_cols, colnames(df))  # Filtrar solo las que existen en df
df[numeric_cols] <- lapply(df[numeric_cols], function(x) as.numeric(gsub(",", ".", x)))
```

```
## Warning in FUN(X[[i]], ...): NAs introducidos por coerción
## Warning in FUN(X[[i]], ...): NAs introducidos por coerción
## Warning in FUN(X[[i]], ...): NAs introducidos por coerción
## Warning in FUN(X[[i]], ...): NAs introducidos por coerción
## Warning in FUN(X[[i]], ...): NAs introducidos por coerción
## Warning in FUN(X[[i]], ...): NAs introducidos por coerción
## Warning in FUN(X[[i]], ...): NAs introducidos por coerción
## Warning in FUN(X[[i]], ...): NAs introducidos por coerción
## Warning in FUN(X[[i]], ...): NAs introducidos por coerción
```

Cálculo de Estadísticas Descriptivas

```
stats <- data.frame(
  Variable = numeric_cols,
  Media = sapply(df[numeric_cols], mean, na.rm = TRUE),
  Mediana = sapply(df[numeric_cols], median, na.rm = TRUE),
  DesviacionEstandar = sapply(df[numeric_cols], sd, na.rm = TRUE)
)

# Mostrar resultados
print(stats)
```

```
##                                   Variable      Media Mediana
## survival                           survival 22.1829231  23.500
## still_alive                     still_alive  0.3282443   0.000
## age_at_heart_attack     age_at_heart_attack 62.8137222  62.000
## fractional_shortening fractional_shortening  0.2167339   0.205
## epss                                   epss 12.1647692  11.000
## lvdd                                   lvdd  4.7631570   4.650
## wall_motion_score         wall_motion_score 14.4381250  14.000
## wall_motion_index         wall_motion_index  1.3780000   1.216
## mult                                   mult  0.7767188   0.786
## alive_at_1_year             alive_at_1_year  0.3243243   0.000
##                       DesviacionEstandar
## survival                      15.8582672
## still_alive                    0.4713768
```

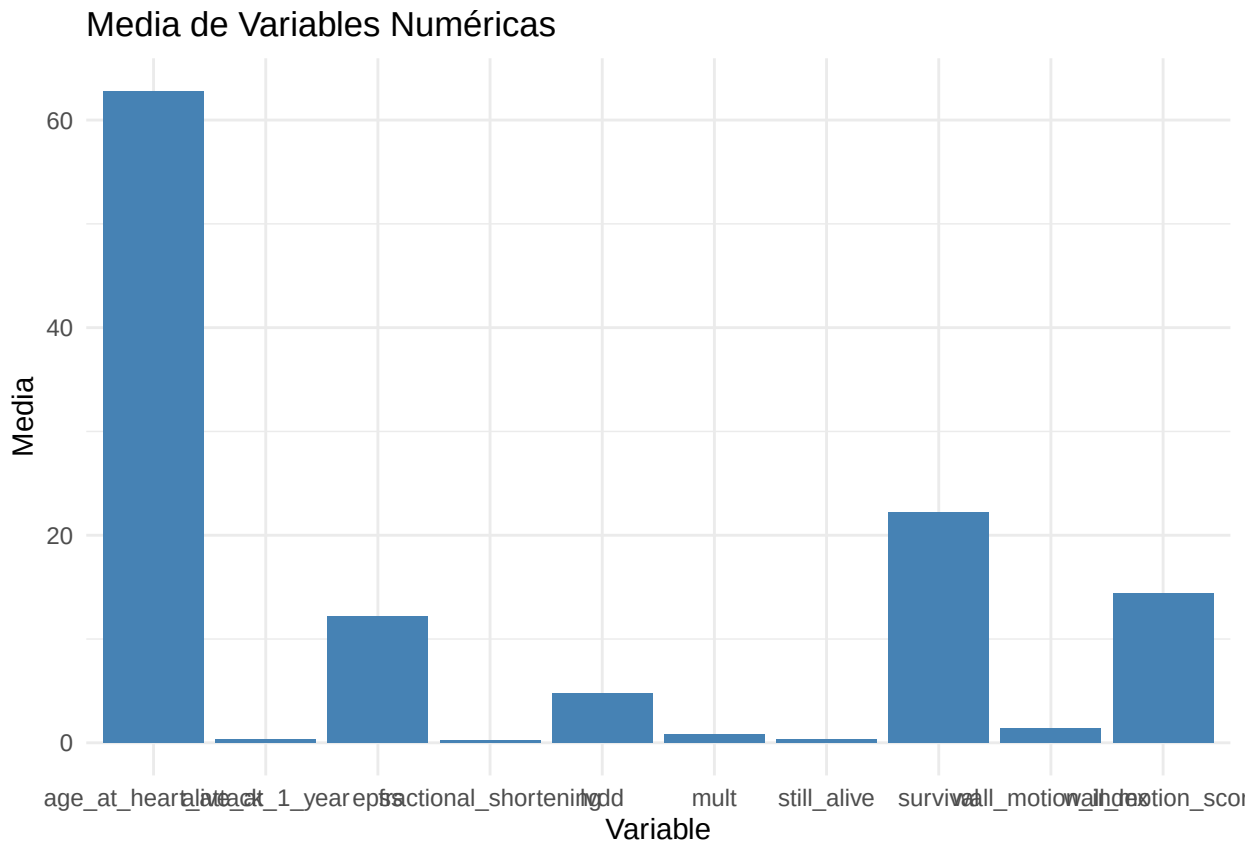
```
## age_at_heart_attack            8.3421099
## fractional_shortening          0.1075128
## epss                           7.3701595
## lvdd                           0.8100130
## wall_motion_score              5.0185664
## wall_motion_index              0.4518500
## mult                           0.1990783
## alive_at_1_year                0.4713172
```

Visualización de Resultados

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.3
```

```
ggplot(stats, aes(x = Variable, y = Media)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  theme_minimal() +
  labs(title = "Media de Variables Numéricas", x = "Variable", y = "Media")
```




```
# Cargar paquete necesario
library(dplyr)

# Seleccionar solo las variables categóricas
```

```r
categoricas <- df %>% select(where(is.character))

# Calcular frecuencias de cada variable categórica
frecuencias <- lapply(categoricas, table)

# Mostrar resultados
print(frecuencias)
```

```
## $pericardial_effusion
##
##   0   1
## 107  24
##
## $group
##
##  ?  1  2
## 22 24 85
```

CONCLUSIÓN

Los resultados obtenidos refuerzan la importancia de la recopilación y el procesamiento adecuado de los datos clínicos, ya que estos influyen directamente en la precisión y efectividad de los modelos. Además, el uso de enfoques estadísticos y de aprendizaje automático puede optimizar el diagnóstico y reducir la carga de trabajo de los profesionales de la salud.

En futuras investigaciones, sería recomendable explorar modelos más avanzados y ampliar la base de datos con información actualizada, lo que permitiría mejorar la precisión de las predicciones y contribuir al desarrollo de soluciones más robustas en el ámbito de la salud digital.

REFERENCIAS

- [1] Organización Mundial de la Salud (OMS), "Enfermedades cardiovasculares", World Health Organization, 2023. [En línea]. Disponible en: https://www.who.int/es/health-topics/cardiovascular-diseases#tab=tab_1. [Accedido: 15-Mar-2025].
- [2] UCI Machine Learning Repository, "Echocardiogram Data Set," UCI Machine Learning Repository, 1989. [En línea]. Disponible en: https://archive.ics.uci.edu/dataset/38/echocardiogram. [Accedido: 15-Mar-2025].