¿Se pueden predecir problemas cardiacos desde la ciencia de datos?

Can heart problems be predicted from data science?

Arnold Santiago Torres Anzola, Edison Fabian Tovar Castro, Nassir Santiago Cárdenas caldas

Pregrado Ciencia de Datos Universidad Externado, Bogotá – Colombia

Pregrado Ciencia de Datos Universidad Externado, Bogotá - Colombia

Pregrado Ciencia de Datos Universidad Externado, Bogotá – Colombia

Arnold.torres@est.uexternado.edu.co, edisson.tova1@est.uexternado.edu.co, Nassir.cardenas@est.uexternado.edu.co

RESUMEN

Para el campo de la medicina es importante recolectar datos médicos de pacientes, esto con el fin de analizar dichos datos y encontrar factores de riesgo, pero ¿es posible predecir si pacientes sufrirán problemas cardiacos? El fin del siguiente trabajo es desde datos médicos de pacientes con y sin problemas cardiacos analizar, validar, comparar variables de cuadros médicos e intentar observar si con datos médicos se pueden predecir posibles afectaciones cardiacas.

ABSTRACT

For the medical field it is important to collect medical data from patients to analyze such data and find risk factors, but is it possible to predict whether patients will suffer cardiac problems? The purpose of the following work is to analyze, validate, compare variables from medical charts and try to see if medical data can be used to predict possible cardiac problems.

1. INTRODUCCION

En el campo de la medicina la recopilación de datos de los pacientes posee usos prácticos. En un primer plano analizar la información sirve para validar el estado de salud de un paciente, los datos referentes a la enfermedad, sus causas o, en su defecto, sus consecuencias. A partir de modelos de predicción, los especialistas del área de la salud tendrán la posibilidad no solo de analizar el fenómeno que refiere a la sanidad de las personas, sino que se abre la probabilidad de pronosticar posibles complicaciones y consecuencias en la salubridad pública. Por este motivo se realiza la siguiente pregunta:

¿Es posible predecir problemas cardiacos con datos médicos?

Las causas de muerte por enfermedades cardiovasculares representan la causa número 1 de decesos a nivel global y, a partir de las nuevas tecnologías, es posible dar una predicción de que tan posible es que ocurra un suceso de riesgo cardiaco. Por ende, se abordará la siguiente pregunta:

A partir del análisis de datos ¿Es posible encontrar factores de riesgo que incidan en el desarrollo de problemas cardiacos?

Se realizará la exploración de datos médicos de pacientes para averiguar que factores perjudican la salud cardiovascular de las personas. Para lograrlo, se proponen los siguientes objetivos:

Objetivo general

Realizar, a partir de los datos médicos presentados, un estudio para encontrar los factores que producen problemas cardiacos.

Objetivos específicos

1) Realizar el estudio y la exploración de los datos seleccionados, escogiendo la base de datos y verificando las variables para observar una posible correlación entre datos.

2) Realizar la normalización de los datos para proceder a realizar la comparación entre los factores que inciden en el desarrollo de problemas cardiacos.

3) Concluir los fatores determinantes que inciden en el desarrollo de problemas cardiacos, a partir del análisis de variables y su comparación.

2. Análisis Exploratorio

Para el estudio se utilizará la siguiente base de datos:

https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data

La base de datos se recopilo en el año 2020 con el fin de observar patrones entre pacientes con problemas cardiacos y pacientes sanos, para así ver como se pueden relacionar posibles problemas cardiacos con el estado de salud y los diversos hábitos de los pacientes.

La base de datos cuenta con las siguientes variables:

- **Age**: edad de los pacientes. Variable de tipo numérica
- **Anemia**: Disminución de glóbulos rojos o hemoglobina. Variable de tipo dicotómica
- **creatinine_phosphokinase**: Nivel de la enzima CPK en la sangre (mcg/L). Variable de tipo numérico
- **diabetes**: Determina si el paciente tiene diabetes. Variable de tipo dicotómica
- **ejection_fraction**: Porcentaje de sangre que sale del corazón en cada contracción. Variable de tipo numérica
- **high_blood_pressure**: Determina si el paciente tiene hipertensión. Variable de tipo dicotómica
- **plateles**: Número de plaquetas en la sangre (kiloplatelets/mL). Variable de tipo numérica
- **serum_creatinine**: Nivel de creatinina sérica en la sangre (mg/dL). Variable de tipo numérica.
- **serum_sodium:** Nivel de sodio sérico en la sangre (mEq/L). Variable de tipo numérica.
- **sex**: Hombre o Mujer. Variable de tipo dicotómica.
- **smoking:** Determina si el paciente fuma. Variable de tipo dicotómica.
- **Time:** Determina el tiempo de seguimiento. Variable de tipo numérica.
- **DEATH_EVENT:** Determina si el paciente murió durante el seguimiento. Variable de tipo dicotómica.

2.1 Análisis Descriptivo

Para el análisis y la exploración de datos tendremos presentes alrededor de 299 datos clínicos de pacientes, caracterizando su estado de salud, edad, los hábitos que tienen los pacientes (si fuman), y los datos sobre su sistema circulatorio (ver figura 1).

Realizando el primer análisis, se observa que existen sesgos en la mayoría de las variables presentadas en la base de datos. No se observa varianza que afecte las variables.

2.1.1 Procesamiento de datos

Para el procesamiento de los datos se realizó la selección de las variables no dicotómicas para concretar su respecta normalización, la cual incluye el cambio de escala de miles a cientos para trabajar en las comparaciones de datos.

Entre las variables que fueron modificadas se pueden encontrar:

creatinine_phosphokinase
ejection_fraction
platelets
serum_creatinine
serum_sodium

2.2 Presentación de grafica

Se inicia la presentación de las variables contenidas en la base de datos:

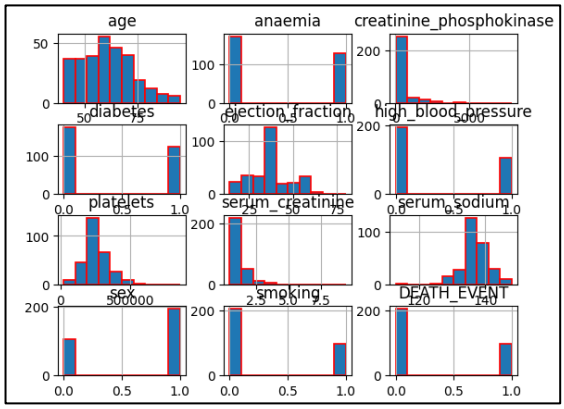

Fig. 1. Vista de las variables durante el análisis descriptivo.

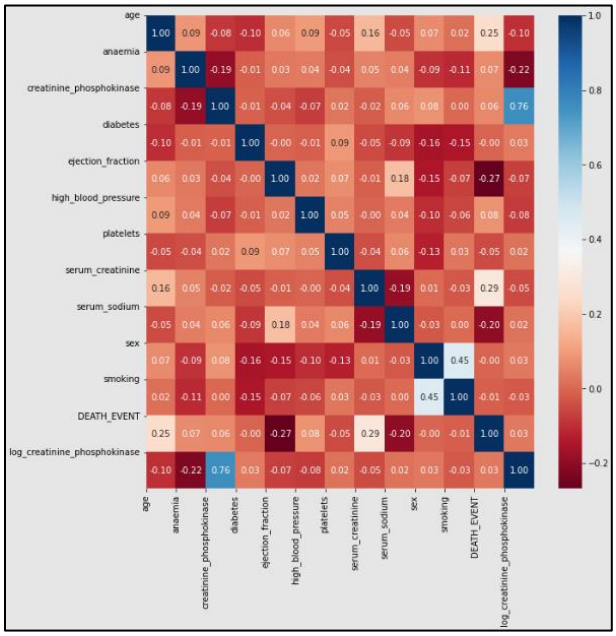

Fig. 2. Mapa de calor de las variables presentes en la base de datos. Se puede observar una buena distribución de dato.

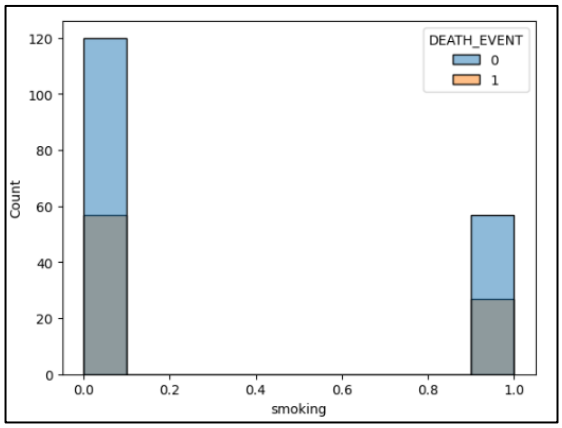

Fig. 3. En el eje x se sitúa la variable smoking (fumadores) donde 0 son los pacientes que fuman y 1 los que no fuman. Se observa que las muertes en pacientes fumadores rondan la mitad, ocurre algo similar en pacientes que no fuman.

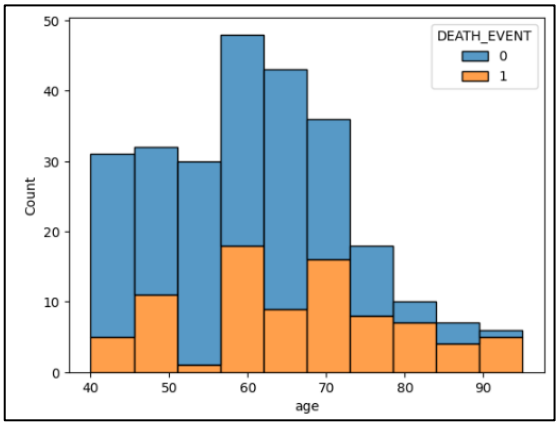

Fig. 4. En la gráfica se observa la cantidad de muertes dependiendo de la edad de los pacientes. Se concluye que cuando es mayor la edad de los pacientes, mayor es la cantidad de muertes es casi del 100%.

Datos cardiacos

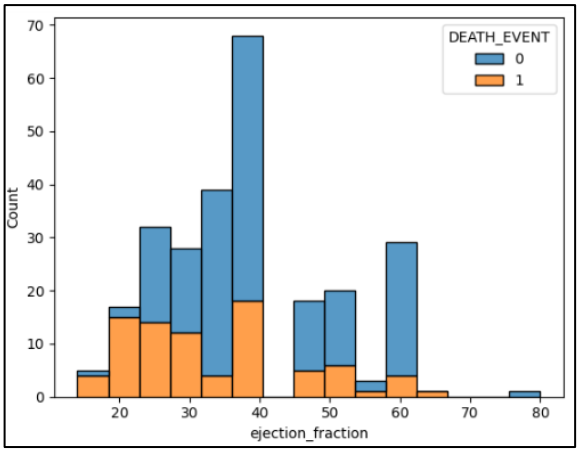

Fig. 5. La grafica muestra los eventos de muerte a partir de las Fracciones de eyección. Se encuentra que cuando las muertes se dan en personas con una fracción de eyección menor de 40, el riesgo de muerte es mayor.

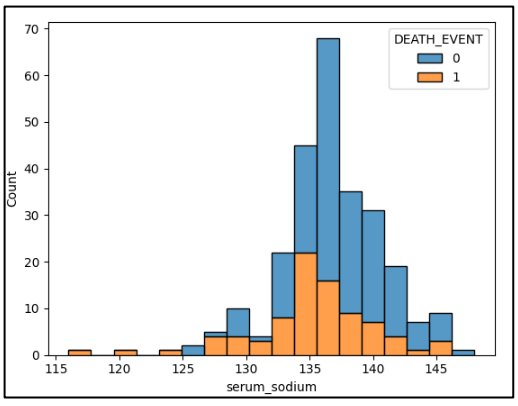

Fig. 6. Los altos niveles de sodio en la sangre también son causales de decesos por problemas cardiacos. Un nivel alto de sodio produce ritmos cardiacos acelerados. Las concentraciones normales de sodio en la sangre son de

135mEq/L. a 145mEq/L., por tal motivo la mayoría de los datos se encuentran en este rango.

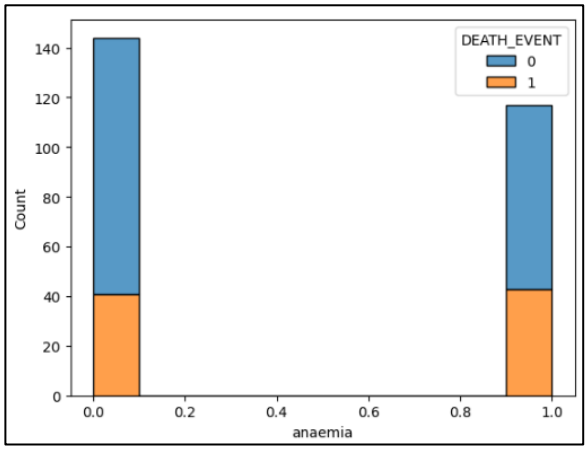

Fig. 7. Los datos sobre anemia presente en pacientes son relevantes, dado que los pacientes que padecen de esta enfermedad son propensos a sufrir arritmias cardiacas.

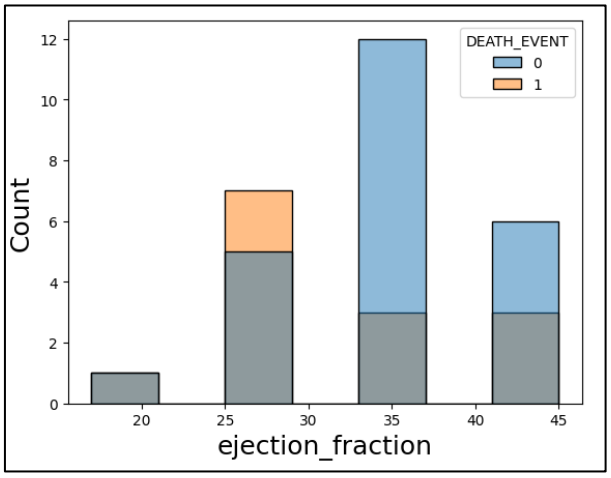

Fig. 8 En la figura se observa como las personas que poseen presión sanguínea menor al 50% son más propensas a tener problemas en la fracción de expulsión.

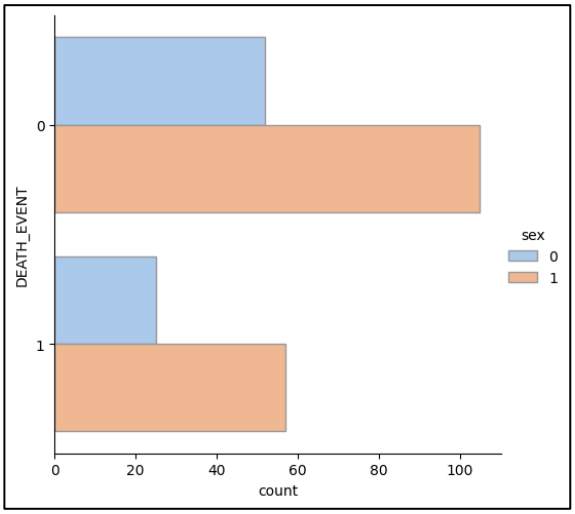

Fig.9 En la gráfica de barras se aprecia que los pacientes fallecidos tras la realización de la encuesta, los hombres, en correlación a su género, tuvieron una mortalidad más alta

2.3 Presentación de la Hipótesis

Después de realizar el análisis de datos, se plantea la siguiente pregunta: ¿Es posible predecir problemas cardiacos con datos médicos?

A partir de la pregunta, se plantean las siguientes hipótesis:

Desde la muestra de 299 pacientes, se concluye que la muerte por problemas cardiacos es mayor en personas que tienen más de 70 años.

Se observa que al obtener facciones de eyección menores a 30%, la mortalidad aumenta al 70% de las personas.

Comparando los datos de muertes de personas con y sin anemia, se puede concluir que la mortalidad en el grupo de pacientes con anemia supera la mortalidad de pacientes sin anemia.

2.4 Modelo de regresión

Para validar el modelo expuesto, se decidió validar a partir de las siguientes variables:

ejection_fraction
serum_sodium
sex
DEATH_EVENT
creatinine_phosphokinase

Las anteriores variables fueron seleccionadas puesto que, entre las encontradas en toda la base de datos, son las que tienen menos correlación entre sí, lo que implica más veracidad durante los resultados finales.

Se utilizará el siguiente modelo de regresión logística para verificar los resultados:

P (X) = probability (DEATH_EVENT | (variables escogidas)) = logistic(β0+β1X)

Donde se tomará como variable dependiente a DEATH_EVENT, ya que cuenta con relación frente a las otras variables escogidas.

2.5 Inferencia del modelo

Para la validez del modelo propuesto, se utilizó como mecanismo de validación una prueba por RMSE relacionando todas las variables del dataset. El modelo arroja un 0.15969053288022289% de error.

Con este porcentaje, se evidencia la precisión del modelo planteado, contando con una buena eficacia frente al problema presentado.

3. Resultados y Discusión

Una vez aplicados los ajustes requeridos en el modelo se obtuvieron los siguientes resultados:

Normalización de los datos.

Se redujo la escala de las variables en escala de 0.0 – 0.1, ya que la mayoría de las variables presentadas en el modelo no lograba relacionarse de manera adecuada.

Correlación de las variables.

Dentro de la correlación de las variables se podía observar poca correlación entre los datos.

Observaciones:

Para la crear la versión final del modelo, los datos seleccionados resaltan por su correlación y son los siguientes:

- ejection_fraction
- serum_sodium
- sex
- DEATH_EVENT
- creatinine_phosphokinase

Creación del modelo:

El modelo demostró pocos errores entre las variables utilizadas, esto nos da a entender que las variables seleccionadas van acorde a la necesidad.

RMSE

El RMSE nos indica como resultado 0.15969053288022289% de precisión frente a lo que se quiere obtener.

Predicción

La predicción del modelo nos dio como resultado la siguiente tabla:

DEATH_EVENT		
0	183	20
1	49	47

Donde 183 resultados fueron acertados y 20 fueron declarados como falsos o no correctos.

4. CONCLUSION

¿Se pueden predecir problemas cardiacos desde la ciencia de datos? A partir de dicha pregunta, se planteó el problema con el que inicio este estudio. Desde el análisis exploratorio de los datos, hasta la creación del modelo de regresión, se observo que la pregunta puede ser resuelta a partir de los datos suministrados por los pacientes. Siguiendo los pasos adecuados, se llega a las siguientes conclusiones:

- ¿Los datos clínicos son los indicados y suficientes para resolver un problema de esta magnitud?

Aunque los datos tomados en un inicio eran bastantes y parecían correctos para llegar a la finalidad del modelo, la mayoría fueron descartados para encontrar mayor veracidad en los resultados. Dado a la baja correlación que tenían los datos entre sí, el modelo se sobrestima y puede dar un resultado negativo. Para evitar resultados negativos en el modelo, se descartaron datos cruciales para tener una respuesta acertada, lo que implica que los datos utilizados no son suficientes para una respuesta certera de la problemática.

- ¿Los resultados obtenidos son acordes a la realidad?

Se puede observar que sí. Aunque se hallan excluido variables, los datos restantes eran acertados a la realidad, se pueden observar similitudes en los datos clínicos originales, tomados de pacientes en la base de datos sin modificar.

- ¿Se puede predecir un problema cardiaco?

Aunque el campo de la medicina cuenta con sus propios métodos para diagnosticar diferentes tipos de

padecimientos en el área de la salud, y la tecnología es bastante avanzada en el presente siglo, actualmente es necesario tener más información y conocimiento suficiente sobre un paciente para dar un veredicto frente a las predicciones de problemas cardiacos.

Aunque se mantuvo una estimación fiel a los datos suministrados por la base original, fueron excluidos datos importantes para el resultado final: lo que llevó a una afectación negativa del modelo. El modelo presenta una sobrestimación en sus resultados, y por tal motivo su precisión puede ser mayor o menor en comparación a la realidad.