**APLICACIONES DE LA ANALÍTICA: ANALITICA EN SALUD**

****

**POR:**

**DANIELA LÓPEZ ANAYA**

**VALENTINA MÚNERA PULGARÍN**

**DOCENTE:**

**JUAN CAMILO ESPAÑA**

**INGENIERÍA INDUSTRIAL**

**FACULTAD DE INGENIERÍA**

**UNIVERSIDAD DE ANTIOQUIA**

**2023**

# DISEÑO DE LA SOLUCIÓN

Una Clínica de Medellín de régimen privado, maneja 4 indicadores para llevar un seguimiento adecuado de su gestión mensualmente. Estos indicadores son:

* Mortalidad.
* Evaluación de la productividad de la clínica.
* Satisfacción del paciente.
* Factores de riesgo

Actualmente, la tasa mortalidad es uno de los indicadores más importantes que tiene la clínica, el cual se mantiene en constante observación y seguimiento. En los últimos meses se han generado alertas por su constante incremento, convirtiéndolo en el indicador más crítico. Debido a lo anterior se realizó una investigación de la causa de muerte de los pacientes, en donde se determinó que la mayoría de los pacientes que han muerto han sido pacientes con Insuficiencia Cardiaca (IC).

Por otro lado, la alta tasa de mortalidad está ocasionando pérdidas en contrataciones importantes con entidades promotoras de la salud, ya que al resultado de este indicador genera perdida en la confiabilidad de los servicios prestados por la clínica causando problemas financieros. Por esta razón, el gerente de la clínica decide contratar un grupo de analistas de datos para que indaguen la situación y así mismo brinden una solución.

El problema analítico consta de crear un modelo que prediga si un paciente con insuficiencia cardiaca al ingresar a la clínica va a sobrevivir o morir, este modelo se entrenará utilizando datos históricos brindados por la clínica.

Por otro lado, se tiene como diseño de solución lo siguiente:

1. Al ingresar un paciente con deficiencia cardiaca, se ingresan los datos a una base datos.
2. Cada que se ingrese los datos de un paciente se programa la ejecución automática del modelo, que predice si el paciente ingresado puede sobrevivir o no.
3. La situación de los pacientes identificada por el modelo es enviada al personal ejecutante presente, para que tengan un panorama de la predisposición del futuro estado del paciente y tome las debidas acciones al respecto.
4. Todos los días a las 3 am se toma la información recolectada de los pacientes con insuficiencia cardiaca y se entrena el modelo.

# LIMPIEZA Y TRANSFORMACIÓN

Inicialmente, fue suministrado una base de datos para examinar, la cual contiene registros de 1177 pacientes y 51 variables, donde se encuentra información de las características demográficas, signos vitales, comorbilidades y variables de laboratorio, de los pacientes con insuficiencia cardiaca (IC) de un centro hospitalario. Adicionalmente, se tiene la variable "outcome", la cual indica si el paciente sobrevivió (0) o falleció (1).

De archivo mencionado anteriormente, se realizó un proceso de limpieza y transformación, de la siguiente forma:

1. Análisis general de la cantidad de variables y registros.
2. Análisis de valores únicos por cada variable.
3. Estudio de valores nulos por variable y cálculo del porcentaje de nulos por archivo.
4. En caso de presencial de nulos, selección y aplicación de métodos para tratamiento de estos.
5. Observación del tipo de variable y transformación según conveniencia.

Al realizar los análisis descriptos anteriormente se decide eliminar las variables BMI, Neutrophils, Basophils, Lymphocyte, Creatine kinase, PH, Lactic acid y PCO2, ya que tienen un porcentaje de nulos mayor al 10%. Por otro lado, en la variable Outcome se encuentra un valor nulo, por esto se decide eliminar este registro, quedando de esta manera 1776 registros en total y 41 variables.

# ANÁLISIS EXPLORATORIO

Seguidamente, tras lograr obtener un archivo consolidado, y teniendo como objetivo la construcción de un modelo de aprendizaje supervisado de tipo clasificación, se procedió a realizar el análisis exploratorio de los datos, encontrando lo siguiente.

En la Ilustración # 1, observamos que, de los 1777 registros de pacientes, el 13.5% de los pacientes que ingresan a UCI con IC mueren y el 86.5% de los pacientes viven.

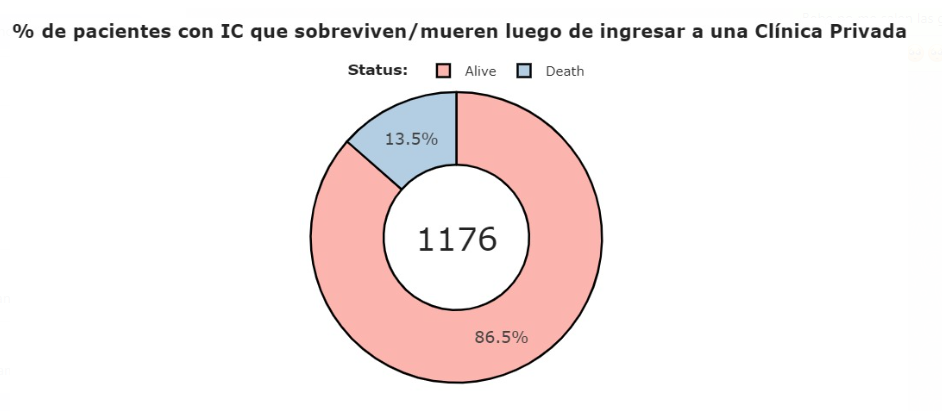


Ilustración 1. Porcentaje de pacientes con IC que sobreviven o mueren.

En cuanto a la frecuencia cardiaca promedio segundo el género del paciente, se encontró que tanto los pacientes hombres y mujeres que fallecen tienen un promedio de frecuencia cardiaca alto comparado con los pacientes que sobreviven.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Ilustración 2. Frecuencia cardiaca promedio según el género.

Por otro lado, se exploró ciertas variables de laboratorio, la frecuencia respiratoria, creatinina, nitrógeno y potasio en sangre para todos los pacientes, observándose que, para todas las variables, aunque fuera leve su diferencia, para los pacientes que fallecieron presentaron un promedio superior respecto a los que sobrevivieron.

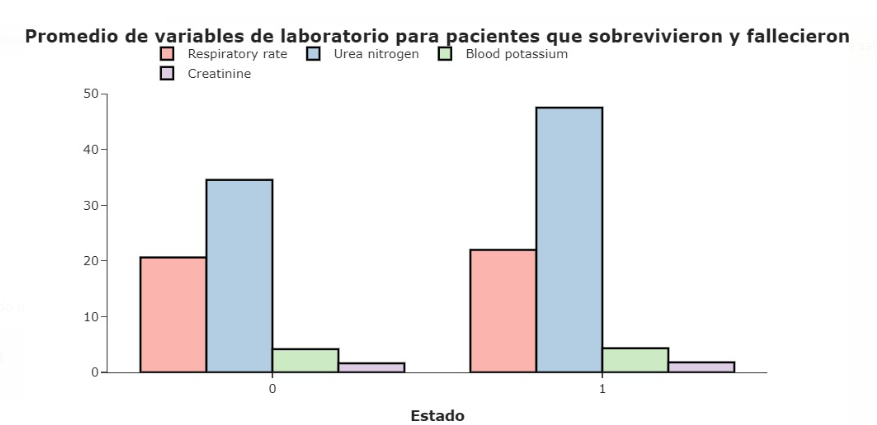


Ilustración 4. Promedio de variables de laboratorio para pacientes que sobrevivieron y fallecieron.

Referente a fibrilación auricular, observar Ilustración 5, se encontró que el 42.1% de los pacientes que sobreviven y el 57.9% de los pacientes fallecen.

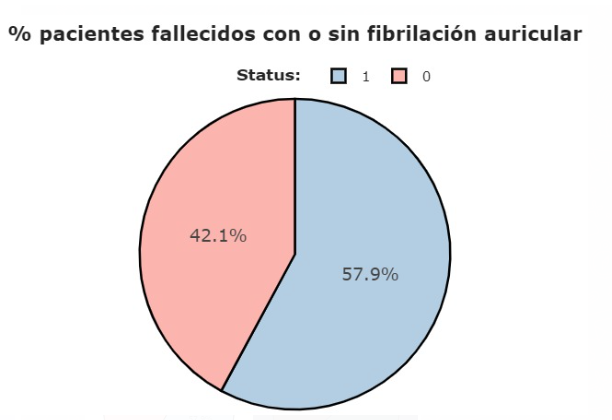


Ilustración 5. Porcentaje de pacientes fallecidos con o sin fibrilación auricular.

Finalmente, se realizó una exploración acerca de los pacientes fallecidos con o sin hipertensión arterial, observar Ilustración 6, se encontró que el 36.5% de los pacientes fallecidos no tenían hipertensión arterial y el 63.5% si tenían.

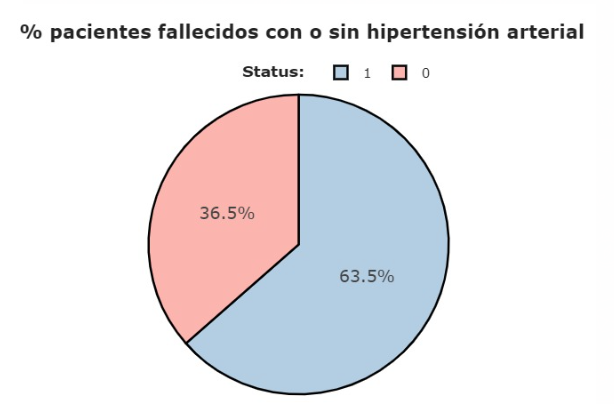


Ilustración 6. Porcentaje de pacientes fallecidos con o sin hipertensión arterial.

# SELECCIÓN DE VARIABLES

Tras realizar la limpieza de los datos, 10 variables fueron eliminadas, quedando en total 40 variables de entrada y 1 de salida. Sin embargo, se observó la necesidad de aplicar métodos de selección de variables para así determinar cuales tenían mayor influencia con la variable de salida.

Los métodos utilizados fueron KBest y Recursive Feature Elimination, seleccionando en cada uno 15 variables, para luego realizar una comparación de resultados y seleccionar las variables en común, las cuales fueron:

* Anion gap
* Bicarbonate
* Leucocyte
* Urea nitrogen
* Blood calcium
* Urine output
* RDW
* INR
* Blood potassium
* heart rate
* Renal failure

# DEFINICIÓN DE ARQUITECTURA Y DE LOS HIPERPARÁMETROS DE LA RED NEURONAL

Para la definición de arquitectura y de los hiperparámetros de una red neuronal, es vital entender el contexto del problema que se va a tratar, en este caso puntual es importante tener presente que se cuenta con pocas observaciones (1176 registros) para 11 variables y una variable de respuesta binaria donde 0 significa que el paciente sobrevivió y 1 que falleció. Con la información anterior un modelo de red neuronal adecuado sería una red neuronal multicapa (Multilayer Perceptron, MLP) ya que ésta, es una red neuronal artificial comúnmente utilizada para la clasificación binaria, por lo que es fácil de implementar, es capaz de aprender relaciones no lineales entre las variables de entrada y la variable de salida, y se puede entrenar eficientemente con un conjunto de datos de tamaño moderado, la cual contara con una capa de entrada, una o más capas ocultas y una capa de salida. El número de neuronas en la capa de salida debería ser igual a 1, ya que la variable de salida es binaria (0 o 1).

Ahora bien, para determinar la cantidad de neuronas para la capa de entrada y oculta, se utilizará la técnica de búsqueda aleatoria, con el fin de ajustar el número de neuronas en cada capa mediante pruebas y evaluación del rendimiento del modelo.

En cuanto a la capa de salida, teniendo en cuenta que la variable de salida es binaria, se utilizó una función de activación sigmoidea, la cual es una función matemática en forma de “S” que toma un valor de entrada y lo transforma a un valor entre 0 y 1.

# EVALUACIÓN Y ANÁLISIS DEL MODELO

Para la evaluación y análisis del modelo, se utilizó la técnica aleatoria para la capa de entrada y oculta como se mencionó anteriormente y se modifico el hiperparametro optimizer, en la Tabla1 se puede observar el accuracy obtenido para cada uno de los modelos implementados.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Capa de entrada | Capa oculta | Capa de salida | Optimizer | Accuracy |
| 10 | 1 | 1 | “Adam” | 86.44 |
| 30 | 3 | 1 | “rmsprop” | 88.14 |
| 20 | 4 | 1 | “Adam” | 87.57 |
| 6 | 3 | 1 | “rmsprop” | 88.42 |
| 4 | 1 | 1 | “Adam” | 87.57 |
| 6 | 1 | 1 | “Adam” | 86.44 |

Tabla 1. Comparación de modelos

Dado los resultados obtenidos para el accuracy de cada modelo, se eligió el mayor de ellos correspondiente al modelo # 4.

Seguidamente, se utilizó la métrica de desempeño de Matriz de Confusión (Ilustración # 7), de la cual se puede observar que el modelo predijo certeramente que 3000 pacientes sobrevivirán cuando realmente lo harán, sin embargo, el modelo también predijo que 35 pacientes sobrevivirán cuando realmente fallecerán. Adicionalmente, el modelo también predijo que 13 personas fallecerán cuando realmente lo harán y 6 personas fallecerán cuando realmente no lo harán.

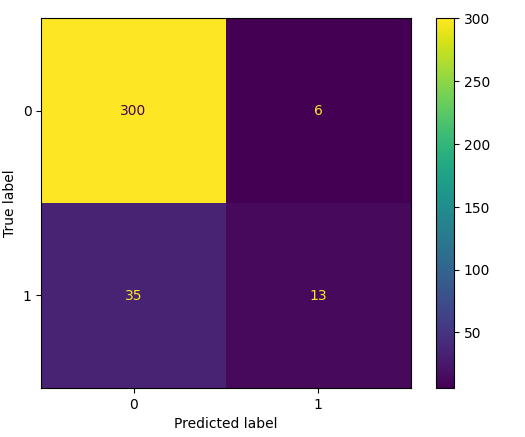


Ilustración 7. Matriz de confusión.

# DESPLIEGUE DEL MODELO

Para un buen despliegue del modelo se debe tener en cuenta lo siguiente:

1. Al ingresar un paciente con deficiencia cardiaca, se ingresan los datos a una base datos.
2. Luego, las predicciones, se realizarán en el mismo instante que sean ingresados los datos. Estas predicciones estarán configuradas para realizarse de forma automática.
3. Las predicciones generadas serán almacenadas en formato CSV y serán cargadas a un repositorio en la nube, donde serán tomadas por el personal ejecutante presente, para que tengan un panorama de la predisposición del futuro estado del paciente y tome las debidas acciones al respecto.

# REFERENCIAS

*In Hospital Mortality Prediction*. (2021, 3 septiembre). Kaggle. https://www.kaggle.com/datasets/saurabhshahane/in-hospital-mortality-prediction