

## **Informe**

### **Trabajo # 4**

#### **Kelly Johana Collazo Quiroz (Ingeniería Biomédica)**

La parte principal de todo es la base de datos, en este caso se eligió una que posee la capacidad de predecir si el paciente sobrevivirá o no, sufriendo de insuficiencia cardíaca, esta posee esta capacidad gracias a que está cargada con un total de 5000 registros médicos de pacientes que ya padecieron esta enfermedad y así mismo contiene 13 características clínicas, tales como; la edad, padecimiento de anemia, niveles de creatinina, padecimiento de diabetes, fracción de eyección, niveles de la presión arterial, niveles de plaquetas, sexo, niveles de creatinina sérica, niveles de sodio, disidencias de paciente con fumar, días en el hospital, índice de muerte a lo largo dicho seguimiento.

Una vez se logró escoger la base de datos más adecuada y que lograr cumplir todos los parámetros ya establecidos para lograr el análisis de datos de forma exitosa, la metodología llevada a cabo abarca la carga, análisis, preprocesamiento, construcción de modelos y evaluación de un conjunto de datos vinculado a la predicción de insuficiencia cardíaca, que como fue mencionada con anterioridad está conformada como una serie de casos clínicos. El propósito es contrastar el rendimiento de dos modelos de clasificación: Support Vector Machine (SVM) y Random Forest (RF).

El conjunto de datos fue cargado desde un archivo CSV utilizando la biblioteca pandas. Se emplearon diversas gráficas para realizar un análisis exhaustivo de los datos, que incluyó la visualización como un aporte extra para el análisis exhaustivo de los datos, un resumen estadístico de las variables, la distribución de las variables objetivas y numéricas, y la evaluación de la correlación entre las variables. Durante el preprocesamiento, no se detectaron valores nulos.

El conjunto de datos fue segmentado en conjuntos de entrenamiento y prueba, realizando inicialmente una reducción del tamaño del conjunto para la fase de prueba. Se evaluaron ambos modelos utilizando el conjunto de prueba y se imprimieron los reportes de clasificación, matrices de Confusión y curvas de Aprendizaje.

Es muy importante tener en cuenta que para poder realizar este código se necesitaron una serie de herramientas tales como las bibliotecas de pandas, numpy, seaborn y matplotlib, para tener todos los recursos necesarios para el procesamiento y análisis de bases de datos de alto contenido.

Por otro lado, es muy importante tener en cuenta la importancia de la matriz de confusión, la cual es una herramienta utilizada en la programación, específicamente en el contexto de la clasificación y el aprendizaje automático, para evaluar el rendimiento de un modelo de clasificación. Proporciona una representación visual de los resultados de la clasificación al mostrar la relación entre las predicciones del modelo y los valores reales. Por las características de este durante el programa fue utilizado para evaluar el desempeño por

medio de la comparación de dos modelos de clasificación, Support Vector Machine (SVM) y Random Forest, aplicados a un conjunto de datos relacionado con la insuficiencia cardíaca.