

Modelo de predicción de falla cardíaca (HF prediction UdeA): Machine Learning en la prevención de enfermedades cardiovasculares.

Carlos Alfredo Pinto Hernández, Omar Alberto Torres, Leidy Castaño Castaño, Yuly Alvear Romo

Departamento de Ingeniería de sistemas, Universidad de Antioquia, Medellín, Colombia

Email: carlos.pinto@udea.edu.co, omara.torres@udea.edu.co, leidy.castanoc@udea.edu.co, yuly.alvear@udea.edu.co

Abstract—La falla cardíaca, como parte de las enfermedades cardiovasculares tiene un gran impacto en la salud pública de las comunidades dado el con gran participación en la mortalidad de los adultos de edad media. El desarrollo de herramientas o modelos predictivos podría ser una estrategia para ayudar a identificar de manera precoz esta entidad con el fin de detener la progresión o aparición de las complicaciones fatales. Como fuente de datos para el desarrollo del modelo predictivo HF prediction UdeA se utilizó el set de datos BD Heart Failure obtenida de Kaggle con 746 registros. Se diseñaron modelos utilizando 5 modelos implementando de la librería sklearn, los métodos relacionados Funciones Discriminantes Gaussianas (No GMMs), KNN vecinos más cercanos, Feed-Forward Neural Networks, Random Forest y Máquinas de Vectores de Soporte. Posteriormente se realizó proceso de selección y extracción de características utilizando la librería mlxtend.feature_selection. El modelo final fue desarrollado con el modelo de KNN vecinos más cercanos con todas las características de la base de datos original obteniendo una sensibilidad del 92.5% y una especificidad del 92.9%. Este es un acercamiento académico al machine learning que tiene un gran potencial en el área de la salud y que podría presentar grandes herramientas para la toma de decisiones clínicas y administrativas.

Index Terms—Machine learning, falla cardíaca, enfermedades cardiovasculares, KNN vecinos más cercanos, Random forest, Máquina de soporte vectorial, funciones discriminantes gaussianas (No GMMs), Redes neuronales artificiales

I. INTRODUCCIÓN

Las enfermedades cardiovasculares (ECV) son la principal causa de muerte a nivel mundial, cobrando un estimado de 17,9 millones de vidas cada año, lo que representa el 31% de todas las muertes en todo el mundo. Cuatro de cada 5 muertes por ECV se deben a ataques cardíacos y accidentes cerebrovasculares [1].

La enfermedad isquémica del corazón o infarto es la responsable de mayor número de casos fatales en Colombia, de acuerdo con información obtenida del Ministerio de Salud y Protección Social. Además, la trombosis, la hipertensión y la diabetes están en la lista de las 10 principales causas de mortalidad, lo cual pone al país frente a una epidemia de las enfermedades cardiovasculares (ECV). Según la Organización Mundial de la Salud, para el año 2011, la tasa de mortalidad de Colombia por cada 100 mil habitantes para ECV y diabetes era de 166,7 para mujeres y de 205,9 para hombres [2].

La falla cardíaca es un problema de salud pública mundial que afecta a más de 23 millones de personas. Las cifras en los Estados Unidos de Norteamérica estiman 5.1 millones de afectados según el registro NHANES (National Health and Nutrition Examination Survey), y se espera que su prevalencia se incremente un 46% del año 2012 al año 2030. En contraste con otras enfermedades cardiovasculares, la prevalencia, la incidencia y la mortalidad por falla cardíaca están en aumento [3].

Las herramientas de simulación y análisis computacional actualmente son fundamentales para predecir escenarios en una gran diversidad de problemas tanto físicos, químicos, biológicos y/o económicos que afectan de manera directa e indirecta la calidad de vida de las personas. En el contexto de salud tiene gran aplicabilidad como herramientas de apoyo para los profesionales de la salud, principalmente en el contexto de salud pública y políticas de promoción de la salud y prevención de la enfermedad. Dado el impacto de la falla cardíaca a nivel mundial y local la generación de un modelo de predicción de esta enfermedad permitiría que el personal médico pudiera tomar decisiones de manera precoz con el fin de impactar en población en riesgo y disminuir el progreso de la entidad o reducir sus complicaciones asociadas.

Para el desarrollo del modelo de clasificación (HF Prediction UdeA), desarrollo o no de falla cardíaca, se toma información de bases de datos (BD) “Heart Failure Prediction Dataset” que relaciona características cuantitativas y cualitativas de 1.190 pacientes. Las variables de entrada de la BD son 11: edad, sexo, tipo de dolor torácico, presión arterial sistólica en reposo, nivel sérico de colesterol total, alteración de glucosa en sangre en ayunas, frecuencia cardíaca máxima, presencia de angina con el ejercicio y dos variables de hallazgos en el electrocardiograma. Como variable única de salida la presencia o no de falla cardíaca.

II. ESTADO DEL ARTE

Múltiples grupos a nivel mundial han realizado estudios para generar modelos de predicción de patología cardiovascular usando modelo de aprendizaje de máquina, obteniendo diversos resultados en sus trabajos. En estos estudios han aplicado diferentes técnicas y enfoques

que buscan obtener niveles de precisión cada vez más altos.

En el departamento de bioestadística de la universidad Trakya en Turquía, Ture y colaboradores [4], realizaron un estudio evaluando el rendimiento de diferentes técnicas de clasificación para predecir la hipertensión esencial. Evaluaron el rendimiento de tres árboles de decisión, cuatro algoritmos estadísticos y dos redes neuronales. Se evidenció que los modelos con redes neuronales realizaron mejores predicciones comparado con las otras técnicas.

Resul Das y colaboradores [5], a través del uso del software SAS base 9.1.3 plantearon diagnosticar falla cardíaca usando una red neuronal en el centro del modelo propuesto. Posterior a la generación de múltiples modelos que sirvieron de base para el modelo final se obtuvo una precisión de clasificación de 89.1% tomando datos de la cohorte de Cleveland. Además, este modelo presentó una sensibilidad y especificidad del 80.95% y del 95.91% respectivamente.

En el año 2017, Gonzales Aranda desarrolló como un trabajo con el fin de validar modelos de predicción de riesgo para enfermedades cardiovasculares en pacientes con problemas de sueño [6]. Para esto utilizó diversas herramientas de validación y selección de variables, algoritmos de ML como Support Vector Machine (SVM), Red neuronal y Naives Bayes. De acuerdo con el autor cada algoritmo arrojaba unas métricas valiosas, pero era difícil decidir cuál es la mejor, por lo tanto, con la ayuda de la curva ROC compararon el AUC de cada modelo y se concluyó que el mejor algoritmo era el Logit (Regresión logística) para el dominio del problema que ellos estudiaron.

Recientemente en la 12a Conferencia Iberoamericana de Complejidad, Informática y Cibernética -CICIC, realizada en 2022, se presentó un trabajo realizado por Olivera y colaboradores donde se propuso el diseño de una aplicación móvil cuya funcionalidad era la predicción de pertenecer o no al grupo de riesgo cardiovascular para personas mayores de 50 años [7]. Utilizaron técnicas de machine learning (Naive Bayes, Random Forest, Logistic Regression y Support Vector Machine) y después de evaluar los modelos concluyeron que el algoritmo de Naive Bayes era el de mejor precisión y estabilidad combinada.

III. EXPERIMENTOS

A. Fuente de los datos

Para el desarrollo del modelo de predicción HF prediction UdeA se obtuvo información del sitio web Kaggle, Base de datos nombrada "Heart Failure Prediction Dataset" actualizada en el año 2021 (2). Este conjunto de datos se creó mediante la combinación de diferentes conjuntos de datos que ya estaban disponibles de forma independiente, pero que no se habían combinado antes. En este conjunto de datos, se combinan 5 cohortes (Cleveland, Húngaro, Suiza, Long Beach y Stalog) sobre 11 características comunes, lo que lo convierte en el

conjunto de datos de enfermedades cardíacas más grande disponible hasta ahora para fines de investigación [8].

B. Características de Base de datos

La base de datos original tenía 1.190 muestras, luego de aplicar proceso de ingeniería de características se determinó trabajar con 746 muestras, dado que algunas muestras presentaban datos null o cero en variables como constantes vitales o niveles séricos de colesterol total, los cuales no son datos válidos en este dominio; 356 muestras tenían salida (1) y los restantes salida (0) lo cual representa un 47.72% y 52.28% respectivamente, por lo cual consideramos que la BD estaba balanceada.

Las variables cualitativas como sexo, tipo de dolor torácico, alteración de glucosa en ayunas, angina con el ejercicio y alteración (ST Slope) en el electrocardiograma se modificaron a variables numéricas para realizar la aplicación de los algoritmos de machine learning. Al final se obtuvieron 20 variables de entrada y una variable de salida.

C. Modelos de predicción

Para la creación de los modelos se utilizaron, de la librería sklearn, los métodos o algoritmos relacionados con cada modelo: Funciones Discriminantes Gaussianas (No GMMs), K vecinos más cercanos, Feed-Forward Neural Networks, Random Forest y Máquinas de Vectores de Soporte. En el proceso de generación de los modelos se realizó la evaluación de la precisión de estos a través de metodología crossvalidation con KFold (n_splits=10, shuffle=True).

Para todos los modelos, excepto la red neuronal, se ejecutaron los algoritmos en 5 ocasiones con el fin de verificar su variabilidad. Se analizaron los resultados de la matriz de confusión con el fin de determinar la sensibilidad y especificidad, además se revisaron los valores de precisión, recall y f1-score.

1) Funciones Discriminantes Gaussianas (No GMMs):

Para este modelo de predicción se utilizó el método Gaussian-ProcessClassifier y a través de la metodología de validación se obtuvieron los resultados descritos en la tabla 1, los cuales son los promedios de los modelos obtenidos. En promedio se obtuvo una variabilidad muy baja entre modelos de esta categoría (0.8%).

Métricas FDG (No GMMs)	
Recall (S)	84.0%
Especificidad	83.6%
Precisión	84.1%
Accuracy	84.4%
F1-score	83.9%

Tabla 1. Métricas Modelo Funciones Discriminantes Gaussianas

2) *Vecinos más cercanos - KNN*: Para este modelo se determinó utilizar 5 vecinos más cercanos, dado que presentaba mejores resultados y además el tiempo de ejecución era menor a 5 segundos. Se utilizó el método `KNeighborsClassifier` de la librería `sklearn.neighbors`. En promedio, se obtuvo una variabilidad entre resultados de modelos generados del 1.2%. En la siguiente tabla se presentan las métricas relacionadas con este modelo.

Métricas KNN	
Recall (S)	92.5%
Especificidad	92.9%
Precisión	91.8%
Accuracy	92.3%
F1-score	92.3%

Tabla 2. Métricas Vecinos más cercanos KNN

3) *Random Forest*: El número de árboles utilizado para este modelo fue de 5, medida de pureza por default (gini) y `n_jobs` de 4, estos parámetros fueron establecidos en el método `RandomForestClassifier` de la librería `sklearn.ensemble` para la generación del modelo. La variabilidad de los resultados obtenidos fue del 1.3%, en la Tabla 3 se describen las métricas en promedio de los modelos generados

Métricas Radom Forest	
Recall (S)	86.6%
Especificidad	86.0%
Precisión	86.3%
Accuracy	86.7%
F1-score	86.5%

Tabla 3. Métricas Random Forest

4) *Máquina de soporte vectorial - SVM*: Los modelos se generaron a partir del método `SVC` de la librería `sklearn.svm` con parámetro `gamma` auto, `probability` en true y `cache_size` de 400. En estos modelos se presentó una variabilidad mayor comparada con los anteriores, en este caso se presentó un porcentaje del 3.5. Las métricas para este modelo se presentan en la siguiente tabla.

Métricas SVM	
Recall (S)	89.3%
Especificidad	90.0%
Precisión	89.2%
Accuracy	89.2%
F1-score	89.1%

Tabla 4. Métricas Máquina de soporte vectorial

5) *Red Neuronal Artificial - RNA*: En este modelo se generó una red con 30 neurona en la primera capa con kernel de iniciación "uniform", dos capas internas con 10 y 5 neuronas con función de activación ReLu y al final una neurona de salida con función de activación softmax. En el proceso de validación se usó como parámetro 10 épocas. Se importaron las librerías `keras` y `tensorflow` y para la generación de la red neuronal se utilizó el método `Sequential` de la librería `keras.models`.

Métricas RNA	
Recall (S)	64.5%
Especificidad	-
Precisión	64.5%
Accuracy	-
F1-score	-

Tabla 5. Métricas Red Neuronal Artificial

IV. SELECCIÓN DE CARACTERÍSTICAS

Para este proceso se eligieron los 3 mejores modelos de predicción de Falla cardíaca (Ver Tabla 6), esto fueron KNN vecinos más cercanos, Random forest y máquinas de soporte vectorial. Para estos modelos se realizó un proceso de selección de característica con el fin de verificar si se podían optimizar estos, para esto determinó realizar el método de búsqueda secuencial ascendente a través de la implementación del método `SequentialFeatureSelector` de la librería `mlxtend.feature_selection`. Se realizó comparación de selección de 5, 10 y 15 características para cada modelo.

Métricas	Modelos				
	FDG	KNN	RF	SVM	RNA
Recall (S)	84.0%	92.5%	86.6%	89.3%	64.5%
Especificidad	83.6%	92.9%	86.0%	90.0%	-
Precisión	84.1%	91.8%	86.3%	89.2%	64.5%
Accuracy	84.4%	92.3%	86.7%	89.2%	-
F1-score	83.9%	92.3%	86.5%	89.1%	-

Tabla 6. Métricas de todos los Modelos

A. Vecinos más cercanos - KNN

Para el modelo vecinos más cercanos los resultados se presentan en la tabla 7, con un menor error para 15 características (Age, RestingBp, Cholesterol, FastingBs, MaxHR, Oldpeak, Sex_F, Sex_M, ChesPaint_Type_ATA, ChesPaint_Type_NAP, ChesPaint_Type_ASY, ChesPaint_Type_TA, RestingECG_Normal, RestingECG_ST, RestingECG_LVH)

No. Características	5	10	15
Media Error	0.1435	0.15	0.1407
SD Error	0.0365	0.057	0.0321
t ejecución (s)	27.7	20.4	11.1

Tabla 7. Error y tiempo de ejecución SFS del Modelo KNN

Al evaluar estas características nuevamente en el algoritmo de `KNeighborsClassifier` se evidenció una disminución del poder predictivo del modelo, con un promedio de recall y accuracy del 78.04% y 78.11% respectivamente, lo cual representa una disminución de los parámetros comparados con el modelo con todas las características.

B. Radom Forest

En el caso del modelo de Random Forest el menor error también estuvo asociado a la selección de 15 características, pero estas fueron diferente al modelo anterior. Las características seleccionadas para este modelo fueron: Age, RestingBp, Cholesterol, FastingBs, MaxHR, Oldpeak, Sex_M, ChesPaint_Type_ATA, ChesPaint_Type_ASY, ChesPaint_Type_TA, RestingECG_Normal, RestingECG_LVH,

ST_Slope_Up, ST_Slope_Flat, ExerciseAngina_Y. Los resultados se presentan en la siguiente tabla.

No. Características	5	10	15
Media Error	0.1729	0.1527	0.1475
SD Error	0.0472	0.312	0.0338
t ejecución (s)	97.17	99.1	15.6

Tabla 8. Error y tiempo de ejecución SFS del Modelo Random Forest

Con estas 15 características se evaluaron 5 modelos con el método RandomForestClassifier y se evidenció en promedio un recall del 83.9% y un accuracy de 84.05%, una leve disminución de estos parámetros comparados con modelo con las características completas.

C. Máquina de Soporte Vectorial

El último modelo evaluado fue el Máquina de soporte vectorial e igual a los modelos anteriores la selección de 15 características demostró tener le menor error asociado. Las variables seleccionadas fueron Age, RestingBp, FastingBs, MaxHR, Oldpeak, Sex_F, Sex_M, ChesPaint_Type_ATA, ChesPaint_Type_ASY, ChesPaint_Type_TA, RestingECG_ST, RestingECG_LVH, ST_Slope_Flat, ST_Slope_Down, ExerciseAngina_N. En la tabla 9 se muestran los resultados de validación de 5, 10 y 15 características.

No. Características	5	10	15
Media Error	0.1422	0.1412	0.1381
SD Error	0.0482	0.358	0.1389
t ejecución (s)	36.21	76.06	118

Tabla 9. Error y tiempo de ejecución SFS del Modelo SVM

Evaluando las características seleccionadas se evidencia una reducción del 9.2% de las métricas evaluadas en los modelos con las características iniciales. Se calculó en promedio un recall y un accuracy del 89.3% y 89.3% respectivamente.

V. EXTRACCIÓN DE CARACTERÍSTICAS

Al igual para la selección de características en el caso de la extracción de estas se eligieron los mejores modelos de predicción (KNN vecinos más cercanos, Random forest y máquinas de soporte vectorial) para realizar la comparación de las métricas obtenidas. Se utilizó el método PrincipalComponentAnalysis de la librería mlxtend.feature_extraction. Para todos los modelos se realizó la extracción de 5, 10 y 15 características.

A. Vecinos más cercanos

En la extracción de características para el modelo KNN vecinos se evidenció que no hubo mejoría del rendimiento del modelo comparado con el modelo con las 20 características. De los 3 n se evidenció mejores métricas con 15 variables, pero aun así es mejor el modelo completo. En la siguiente tabla se resumen los resultados obtenidos y la comparación con el modelo original con todas las características.

Parámetros	Características			
	5	10	15	Todas
Recall (S)	82.0%	87.9%	91.8%	92.5%
Precisión	82.6%	87.8%	89.4%	91.8%
Accuracy	82.4%	87.8%	90.5%	92.3%
F1-score	82.2%	87.8%	90.1%	92.3%

Tabla 10. Métricas con ejecución PCA en Modelo KNN

B. Radom Forest

Para el modelo de Random Forest, la extracción de 15 características presenta una mejoría de las métricas del modelo al compáralo con todas las características iniciales. Se evidencia un incremento en promedio del 1.3% en todas las métricas. La tabla 11 presenta los resultados obtenidos para este modelo.

Parámetros	Características			
	5	10	15	Todas
Recall (S)	83.4%	84.8%	87.7%	86.6%
Precisión	82.9%	86.8%	87.9%	86.3%
Accuracy	82.4%	85.1%	87.8%	86.7%
F1-score	82.4%	83.2%	87.8%	86.5%

Tabla 11. Métricas con ejecución PCA en Modelo Random Forest

C. Máquina de Soporte Vectorial

En el caso del modelo de máquina de soporte vectorial se evidencia un cambio positivo del 0.1% en las métricas analizadas en un modelo con 10 características extraídas. En la siguiente tabla se resumen los resultados obtenidos.

Parámetros	Características			
	5	10	15	Todas
Recall (S)	86.9%	89.3%	84.7%	89.3%
Precisión	81.2%	89.6%	85.9%	89.2%
Accuracy	86.5%	89.2%	85.1%	89.2%
F1-score	86.5%	89.2%	84.9%	89.1%

Tabla 12. Métricas con ejecución PCA en Modelo SVM

VI. DISCUSIÓN FINAL

En la evaluación de un modelo de predicción de falla cardíaca se evidencia que los modelos de KNN vecinos más cercanos, Random forest y máquinas de soporte vectorial presentan mejores métricas, siendo KNN vecinos mas cercanos el mejor de todos con una sensibilidad y especificidad del 92.5% y 92.9% respectivamente.

El modelo de RNA fue el que presentó el peor desempeño para la base de datos analizada, siendo resultados contrarios a los publicado por el grupo de Resul Das y colaboradores. En su trabajo toman la cohorte de Cleveland la cual está incluida en la base de datos que utilizamos para el desarrollo del modelo. Los resultados publicados para este modelo en cuanto a precisión, sensibilidad y especificidad fueron 89.1%, 80.95% y 95.91% respectivamente. Comparando los resultados de su modelo con nuestro mejor modelo (KNN vecinos más cercanos con 20 variables) se evidencia una sensibilidad y especificidad mayores del 92% con una precisión del 91.8%, es decir, se mejora la precisión y la sensibilidad respecto al del Resul Das [5] y colaboradores.

Al realizar selección de características no se evidencian ninguna mejoría al evaluar 5, 10 o 15 características para ninguno de los 3 mejores modelos. En el caso de extracción se evidenció que el modelo de maquina vectorial con 10 característico presentó en promedio 0.1% adicional en las métricas con un mejor comportamiento en la precisión pasando de 89.2% con todas las variables a 89.6% con la extracción de solo 10. En el dominio clínico, aunque el modelo de 10 característica extraído presenta mejores resultados no sería un modelo para tener en cuenta dado que este no es interpretable.

El Machine Learning aplicado al área de salud proporciona herramientas con gran utilidad para el apoyo diagnóstico y pronóstico de la población general. Desarrollar mejores modelos predictivos es el reto para las comunidades académicas y científicas con el fin de contribuir al bienestar de la sociedad.

REFERENCES

- [1] O. M. de la Salud OMS, “Enfermedades cardiovasculares,” 5 2017.
- [2] M. de Salud y protección social, “Colombia enfrenta epidemia de enfermedades cardiovasculares y diabetes,” 10 2014.
- [3] J. E. Gómez-Mesa, C. I. Saldarriaga, L. E. Echeverría, and P. Luna, “Registro colombiano de falla cardiaca (recolfaca): metodología y datos preliminares,” *Rev Colomb Cardiol*, vol. 28, no. 3, pp. 217 – 230, 2021.
- [4] M. Turea, I. Kurta, A. T. Kurumb, and K. Ozdamarc, “Comparing classification techniques for predicting essential hypertension,” *Expert Systems with Applications*, vol. 29, p. 583 – 588, 2005.
- [5] R. Das, I. Turkoglu, and A. Sengur, “Effective diagnosis of heart disease through neural networks ensembles,” *Expert Systems with Applications*, vol. 36, p. 7675 – 7680, 2009.
- [6] M. J. H. Aranda, “Selección, implementación y validación de modelos de predicción de riesgo para enfermedades cardiovasculares en pacientes con problemas de sueño,” Ph.D. dissertation, Universidad de Chile, 2017.
- [7] O. C. Olivera, L. G. Honores, A. B. Padilla, and M. C. Galvez, “Aplicación móvil para predecir la probabilidad de pertenecer al grupo de riesgo cardiovascular utilizando machine learning,” *Memorias de la Décima Segunda Conferencia Iberoamericana de Complejidad, Informática y Cibernética*, vol. 1, pp. 158 – 163, 2022.
- [8] F. S. Palacios, “Heart failure prediction dataset,” *Kaggle*, 2021.