CLASIFICACIÓN MULTIVARIANTE DE ADULTOS EN EDAD DE RETIRO EN COSTA RICA SEGÚN SU ESTADO DE PRESIÓN ARTERIAL

Steven Alemán Enríquez[1], Adriana Monge Zúñiga[1], María Ortiz Ureña[1]
steven.aleman@ucr.ac.cr, adriana.mongezuniga@ucr.ac.cr ,

mariaortizzz1997@gmail.com

RESUMEN

En Costa Rica la hipertensión arterial se ha convertido en un problema de salud que afecta cada día más a sus habitantes, frente al creciente número de víctimas producto de enfermedades cardiovasculares es importante la buena clasificación de personas según su estado de presión arterial ya que esto puede ser información valiosa para la detección temprana de enfermedades de este tipo. Por lo que el objetivo de este estudio es determinar el método de clasificación más adecuado para la identificación del estado de presión arterial (Tiene HTA y no tiene HTA) mediante las características sociodemográficas, ambientales y estilo de vida y de acceso de los individuos. Se ponen a prueba 6 métodos de clasificación: regresión logística binomial, k vecinos más cercanos, árboles de decisión, bagging, bosques aleatorios y boosting. Para el análisis se calibran los métodos y se realiza validación cruzada para comparar y contrastar los indicadores de desempeño de cada uno de los métodos. El método que presenta mejor desempeño es bosques aleatorios debido a que muestra menor error, falsos positivos y negativos y a su vez mayor precisión negativa, asertividad negativo, AUC y KS.

PALABRAS CLAVE: Métodos de clasificación, regresión logística multinomial, calibración, validación cruzada, indicadores de desempeño

INTRODUCCIÓN

La presión arterial es un indicador vital de la salud cardiovascular y un factor clave en la detección temprana de enfermedades relacionadas con el corazón. Según Osorio y Amariles (2017), cuando se presentan niveles elevados de la presión sanguínea en las arterias se está frente a una condición llamada hipertensión arterial; la cual es una enfermedad crónica que se caracteriza por constituir un factor de riesgo para otros problemas cardiovasculares.

Según Berenguer (2016), las causas específicas que desencadenan esta condición no se han identificado, sin embargo, las personas que presentan esta enfermedad suelen coincidir en algunos factores, entre los cuales se puede mencionar, la dieta, la actividad física, la edad, el sobrepeso y la obesidad, además del consumo excesivo de tabaco o alcohol.

Por otra parte, Ortiz, Torres, Peña, Alcántara, Supliguicha, Vasquez, Añez, Rojas y Bermúdez en un estudio sobre factores de riesgo asociados a la hipertensión arterial realizado en el 2017, encontraron que a nivel de características sociodemográficas la edad presenta una asociación significativa con la hipertensión arterial presentando una mayor prevalencia conforme aumenta la edad de las personas, por otra parte, también encontraron que la

[1]Estudiantes de Estadística de la Universidad de Costa Rica

prevalencia de la hipertensión se daba en mayor proporción en personas divorciadas y viudas en relación con los solteros. Además, analizaron algunos factores de estilo de vida, donde encontraron que la actividad física y el consumo de alcohol se encuentran asociados a esta enfermedad, donde una mayor actividad física y un menor consumo de alcohol se consideran como un factor protector ante dicha enfermedad.

Las enfermedades cardiovasculares cada año dejan como saldo cerca de 17 millones de muertes a nivel mundial. La hipertensión arterial como parte de este grupo de enfermedades, causa por sí sola aproximadamente 9,4 millones de defunciones anualmente según Berenguer (2016).

En Costa Rica la hipertensión arterial se ha convertido en un problema de salud que afecta cada día más a sus habitantes. Según Morera y Cortés (2021), la prevalencia de hipertensión arterial aumentó de 25% a 36 % en la población de 20 años y más durante el decenio 2004-2014. Además, según datos del Ministerio de Salud durante el 2020 en el país se reportó un total de 336 defunciones asociadas a la hipertensión arterial, aumentando en un 26% con respecto al año anterior. Además, para el año 2021 se presentó un promedio de 53 personas diagnosticadas diariamente con hipertensión arterial.

La clasificación de las personas según su estado de presión arterial puede proporcionar información valiosa para la identificación temprana de riesgos cardiovasculares además de contribuir con el desarrollo de políticas de salud para mejorar la calidad de vida de las personas que padecen hipertensión. Asimismo, es necesario concientizar a la sociedad en general sobre este tema de salud.

Por lo tanto, el objetivo es determinar el método de clasificación más adecuado para la identificación del estado de presión arterial (Tiene HTA y no tiene HTA) mediante las características sociodemográficas, ambientales y estilo de vida y de acceso de los individuos.

METODOLOGÍA

Para efectuar el análisis de este estudio se utilizan datos del Centro Centroamericano de Población, específicamente del proyecto “Estudio costarricense de Longevidad y Envejecimiento Saludable, cohorte de jubilación (CRELES-RC)”. Estos datos contienen información sobre salud física auto informada, salud psicológica, condiciones de vida y comportamientos de salud de personas en edad de retiro nacidos entre 1945 y 1955 en Costa Rica.

A estos datos se les realiza un proceso de depuración y revisión de calidad, se eliminan los valores faltantes y se elimina de la variable respuesta la categoría de hipertensión desconocida debido a que esta no es importante para el objetivo del estudio, posteriormente se crea una nueva subbase solo con aquellas variables de interés para el presente estudio. De esta manera se obtiene una muestra total de 2387 adultos costarricenses y 16 variables.

Descripción de las variables:

Como variable respuesta se tiene la hipertensión (HTA) la cual tiene dos niveles, tiene HTA y no tiene HTA. Se cuenta con una variable independiente cuantitativa la cual es la edad y las siguientes variables independientes categóricas (cuadro 1).

Cuadro 1.

Variables independientes categóricas

Variables independientes categóricas

	Variable	Categorias
Sociodemográficas	Sexo	Hombre = 1 ; Mujer = 0
	Estado Conyugal	En unión = 1 ; Otro = 0
	Nivel educativo	Sin educación = 0
		Educación básica = 1
		Colegio o más = 2
	Nivel socioeconómico percibido	Bueno = 0
		Regular = 1
		Malo = 2
	Vive dentro de la GAM	Sí = 1 , No = 0
Ambientales y de estilo de vida	Actividad física	Baja = 1
		Moderada = 2
		Alta = 3
	Fumado	Fumador = 1
		Antes fumaba = 2
		Nunca ha fumado = 3
	Ingesta de alcohol	Tomador diario = 1
		Antes consumia = 2
		Nunca ha tomado = 3
	Índice de masa corporal	Bajo peso u peso normal = 1
		Sobre peso u obesidad = 2
	Presencia de alguna discapacidad	Sí = 1 , No = 2
De acceso	Cercanía al centro médico	Cerca (Menos de 25 min) = 1
		Medio (Entre 25 y 45 min) = 2
		Lejos (Más de 45 min) = 3
	Visita de ATAP	Sí = 1 , No = 2
	Percepción de la salud	Mala = 1 , Buena = 0
	Tiene pensión	Sí = 1 , No = 2

Con los datos bien definidos, se procede a realizar una comparación de técnicas de clasificación multivariante para observar cuál de las técnicas logra predecir de mejor manera el estado de presión arterial de una persona. Los métodos clasificadores aplicados en la presente investigación son regresión logística binomial, k vecinos más cercanos (KNN), árboles de decisión, bagging, bosques aleatorios y boosting.

Cada uno de los métodos es sometido a una etapa de calibración, para determinar los valores más adecuados de los parámetros de cada modelo. Para esto se realizó una validación cruzada para cada uno de los posibles valores de los parámetros a probar y de esta se obtuvieron las medias de los indicadores de desempeño por cada pliegue (se utilizaron 5 pliegues) para compararlas entre sí y así escoger el mejor valor para ese parámetro.

La primera clasificación corresponde a la construcción de un modelo de regresión logístico binomial, el cual según Dominguez (2021), establece las clasificaciones mediante

probabilidades de pertenencia a cada categoría de la variable dependiente. A esta técnica se le aplica un proceso de selección de variables y se construye un modelo de clasificación con las variables que fueron seleccionadas.

Como segunda técnica se utiliza k vecinos más cercanos (KNN), según Berástegui y Galar (2018), esta técnica consiste en tomar la distancia de una nueva observación a cada una de las observaciones existentes, dichas distancias se ordenan de menor a mayor para ir seleccionando la categoría a la que se va a asignar esa observación, finalmente la categoría a la que se asigna es aquella que tenga una mayor frecuencia entre los vecinos con las k menores distancias. La calibración de esta técnica consiste en ir cambiando el número de vecinos, utilizando valores entre 1 y 11 y como distancia entre individuos la distancia de Gower ya que se tienen variables tanto numéricas como categóricas, se elige el valor de k que tenga la mayor área bajo la curva (AUC).

Seguidamente se utilizan árboles de decisión, esta técnica consiste en realizar una división jerárquica y secuencial del problema, cada una de las divisiones o nodos describen gráficamente las decisiones posibles y por lo tanto los resultados de las distintas combinaciones de decisiones y eventos. Según Trujillano, Sarria, Esquerda, Badia, Palma y March (2008), la clasificación de patrones se realiza según una serie de preguntas sobre las variables predictoras, empezando por el nodo padre y siguiendo el camino por las variables independientes elegidas, a la derecha o a la izquierda según sea la característica, este proceso se detiene hasta llegar a un nodo hoja donde se determina la clasificación asignada. A esta técnica se le debe fijar los parámetros de minisplit y minbucket que corresponden al 1% y 0.5% del total de datos de la base de entrenamiento, seguidamente se deben calibrar el parámetro de complejidad (cp) y la profundidad máxima (maxdepth), para esto, se inicia calibrando el cp en un rango de 0.0005 a 0.01 y cuando se elige un valor apto de cp se procede a calibrar la profundidad del árbol (maxdepth) probando valores de 1 a 15. Para elegir ambos parámetros se utilizan los indicadores de desempeño: error de clasificación, área bajo la curva y Kolmogorov-Smirnov.

La cuarta técnica utilizada es bagging con árboles de decisión, en esta técnica se generan *m* muestras tipo bootstrap de tamaño *n*, luego se ajustan los árboles de decisión con las *m* muestras y finalmente se resume la información mediante la mayoría de los votos. Según Giraldo (2018), la idea básica de la técnica bagging es remuestrear los datos originales y calcular las predicciones sobre el conjunto de datos remuestreados. Para calibrar esta técnica se fijan los parámetros del árbol que se obtuvieron en la técnica anterior y se prueban diferentes números de árboles, en este caso de 1 a 19, pero tomando en cuenta sólo números impares para evitar el inconveniente de que suceda un empate. La calibración de esta técnica se elige utilizando todos los indicadores de desempeño.

La quinta técnica elegida para clasificar es la denominada bosques aleatorios, la cual utiliza la agregación Bootstrap para combinar diferentes árboles; cada árbol es construido con observaciones y variables aleatorias en cada nodo. Medina y Ñique (2017) afirman que esta técnica se puede ver como la ponderación de un conjunto de árboles de decisión. Para calibrar esta técnica se comienza calibrando la cantidad de variables aleatorias, las cuales van desde 1 hasta la raíz cuadrada del número total de variables disponibles y fijando el número de árboles en 500, después de decidir el número de variables, se fija en el parámetro mtry y se procede a

cambiar el número de árboles tomando como valores posibles 10, 50, 100, 200 y 500. Ambos parámetros se calibran utilizando como criterio el área bajo la curva.

Finalmente se utiliza el método de potencialización o boosting; según James, Witten, Hastie y Tibshirani (2021), en esta técnica los modelos son construidos secuencialmente, es decir, cada modelo es construido para aprender de los errores de mala clasificación de los anteriormente construidos, así, cada vez que se construya un nuevo modelo este se enfocará más en los individuos mal clasificados, mejorando su clasificación. La calibración de este método consiste en ir cambiando el número de árboles o iteraciones, en este caso se prueba un número de árboles que va desde 1 hasta 15. Para decidir el número de árboles se utilizan todos los indicadores de desempeño.

Los indicadores de desempeño utilizados son el error de clasificación (E) que es la proporción del número total de predicciones que son incorrectas respecto al total, los falsos negativos (FN) y falsos positivos (FP) que son es la proporción de casos negativos o positivos que fueron clasificados incorrectamente como positivos o negativos según corresponda, la presión positiva (PP) y precisión negativa (PN) los cuales son la proporción de casos positivos o negativos que fueron predichos correctamente, la asertividad positiva (AP) y asertividad negativa (AN) los cuales indican la proporción de buena predicción para los positivos o negativos. Se busca un E, FN, FP bajos y una PP, PN, AN, AP altos.

Por otro lado, se tienen los indicadores de área bajo la curva ROC (AUC) y Kolmogrov-Smirnov (KS). El AUC representa la probabilidad de que el modelo clasifique un positivo aleatorio más alto que un negativo aleatorio, se espera un AUC entre 60% y 90% y el KS representa el ajuste del modelo con la distribución de los datos, se espera que tenga valores entre 20% y 70%. Estos parámetros fueron tomados en cuenta para la escogencia del método que tenga un mejor desempeño en estos indicadores.

Por último, estos seis métodos son comparados entre sí, mediante una validación cruzada, para escoger cuál de ellos produce la mejor clasificación de los individuos.

El análisis estadístico se realiza mediante el software estadístico R Studio en su versión 4.3.0, se hace uso de los paquetes y librerías *haven* (Wickham H, Miller E y Smith D, 2022), *dplyr* (Wickham H, François R, Henry L, Müller K y Vaughan D, 2023) *caret* (Kuhn, 2022), *rocr* (Sing T, Sander O, Beerenwinkel N y Lengauer T, 2005). Para k vecinos más cercanos se utilizó las librerías *kknn* (Schliep K y Hechenbichler K, 2016), *cluster* (Maechler M, Rousseeuw P, Struyf A, Hubert M y Hornik K, 2022), para realizar los árboles de decisión se utilizó la librería *rpart* (Therneau y Atkinson, 2022), para bagging *rsample* (Frick H, Chow F, Kuhn M, Mahoney M, Silge J y Wickham H, 2022), para bosques aleatorios *randomForest* (Liaw y Wiener, 2002), y, por último, para boosting *adabag* (Alfaro E, Gámez M y García N, 2013).

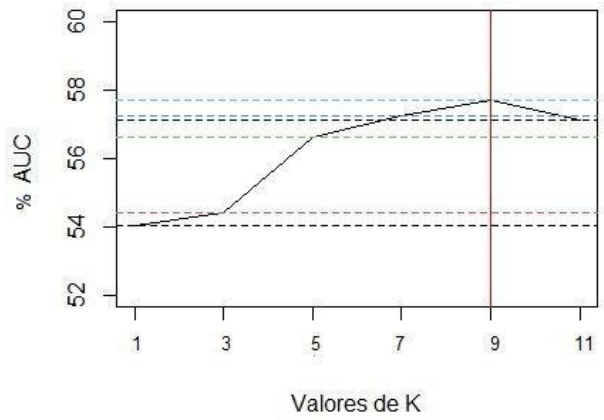
RESULTADOS

En primer lugar, se realiza la construcción de las bases de entrenamiento y validación de 80% y 20% respectivamente. Seguidamente, se comienzan a realizar los métodos clasificadores. La regresión logística binomial da como resultado un modelo con las variables predictoras: visita de ATAP, incapacidad, vive en la GAM, edad, actividad física, sexo, situación económica, índice de masa corporal, cercanía con el centro médico, percepción salud.

Para el método de K vecinos más cercanos (KNN) para la calibración del número de vecinos, se obtiene que la mejor cantidad es de 9 vecinos debido a que este presenta el mejor desempeño en el indicador de AUC (Figura 1).

Figura 1.

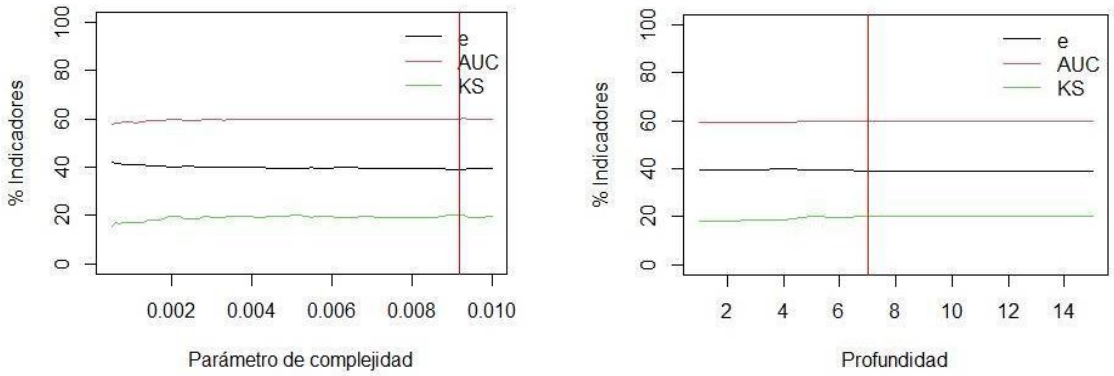
Calibración del número de k para el método de k vecinos más cercanos (KNN)



Para la técnica de árboles de decisión la calibración indica que el cp más adecuado a utilizar dado los indicadores de desempeño de AUC, KS y error de clasificación es de 0.0092 y una profundidad máxima (maxdepth) de 7 (Figura 2).

Figura 2.

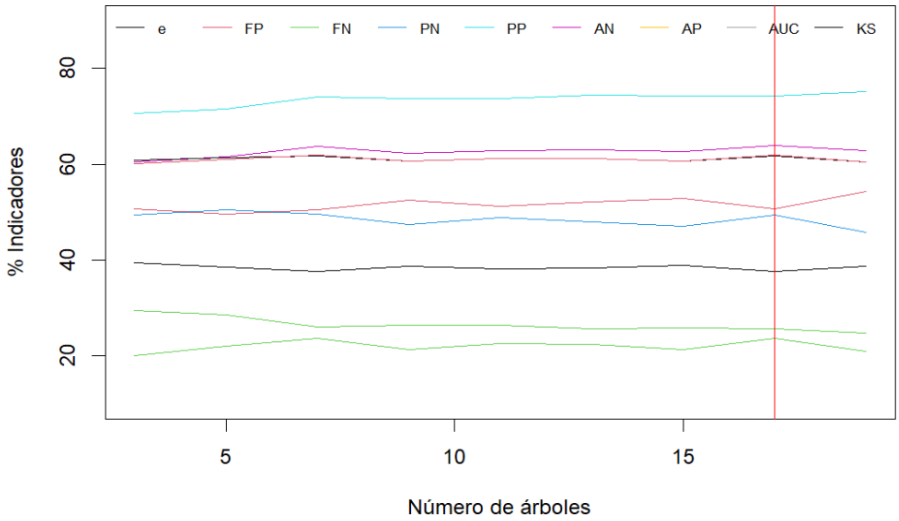
Calibración del parámetro de complejidad y profundidad máxima para el método de árboles de decisión.



Por otro lado, al calibrar el número de árboles (ntrees) para bagging esta indica que la mejor cantidad a utilizar es de 17 árboles ya que en comparación con otros valores este presenta porcentajes bajos en los indicadores de error, FN, FP y altos en el PP, PN, AUC y KS (Figura 3).

Figura 3.

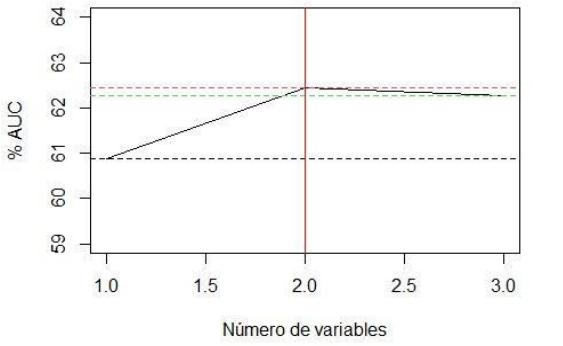
Calibración del número de árboles para el método de bagging.

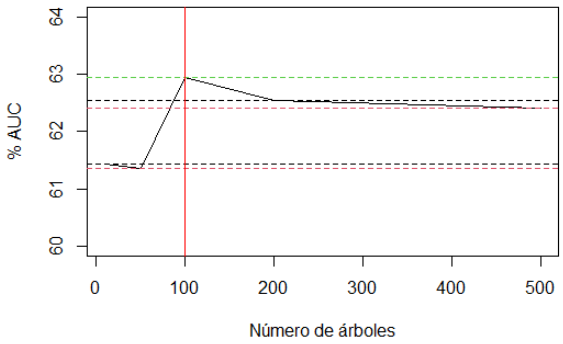


Con el método de árboles aleatorios (Random Forest) la calibración del número de variables indica que lo ideal es utilizar dos variables y una cantidad de 100 árboles ya que estos presentan el AUC más alto en comparación con los otros valores probados (Figura 4).

Figura 4.

Calibración del número de variables y número de árboles para el método de bosques aleatorios.

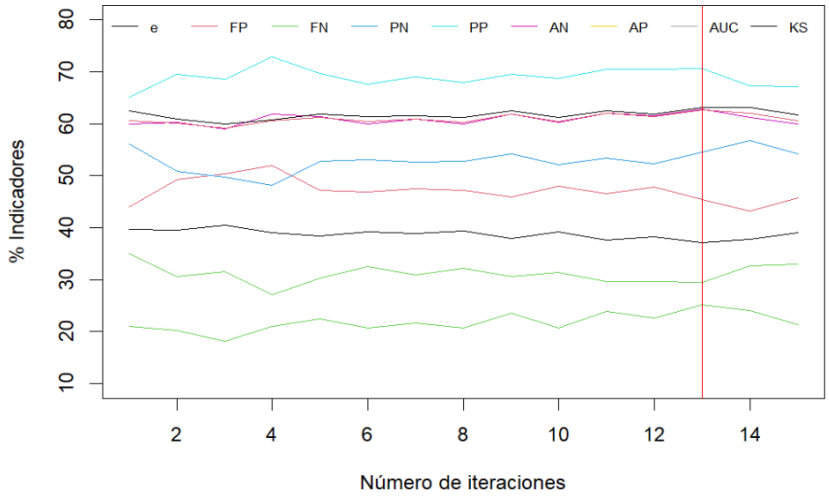




Para el método de boosting se realiza la calibración de la cantidad de iteraciones y esta indica que el número adecuado a utilizar es de 13 árboles o iteraciones, debido a que presenta porcentajes bajos en error, falsos positivos y negativos y altos en precisión, asertividad, área bajo la curva y kolmogorov-smirnov (Figura 5).

Figura 5.

Calibración del número de iteraciones para el método de boosting.



Por último, se realiza la validación cruzada de los diferentes métodos ya calibrados utilizando los distintos indicadores de desempeño, con el fin de escoger el método ideal para clasificar a los individuos. Se muestra que la técnica de bosques aleatorios tiene mejores resultados en el error de clasificación (36.57%), falsos positivos (47.41%), precisión negativa (52.59%), asertividad positivo (62.95%), AUC (62.95%) y KS (25.90%) comparado con las demás técnicas. La siguiente clasificación que muestra mejores resultados es árboles de decisión ya que presenta buenos resultados en falsos negativos (17.19%), precisión positiva (82.81%) y asertividad negativo (66.47%). Las demás técnicas no resultan adecuadas al observar sus indicadores de desempeño (cuadro 2).

Cuadro 2.

Indicadores de desempeño según técnica de clasificación.

Técnica	E	FP	FN	PN	PP	AN	AP	AUC	KS
Logística Binomial	37.53	46.77	29.07	53.23	70.94	62.54	62.48	62.08	24.16
Arboles de decisión	39.63	63.94	17.19	36.06	82.81	66.47	58.74	59.44	18.87
Bosques aleatorios	36.57	47.41	26.69	52.59	73.31	64.26	62.95	62.95	25.90
Bagging	38.75	52.53	26.26	47.47	73.74	62.33	60.71	60.60	21.20
Boosting	37.12	45.42	29.46	54.58	70.54	62.73	63.09	62.56	25.12
Knn	43.27	30.57	54.90	69.43	45.10	53.46	62.04	57.27	14.53

Nota: Los valores son el % promedio de los pliegues de cada indicador de desempeño

El método con mejor desempeño según los indicadores anteriores es bosques aleatorios; al utilizar el modelo de esta técnica se obtiene como resultado la siguiente clasificación de los individuos (cuadro 3):

Cuadro 3.

Clasificación de individuos con el método de bosques aleatorios.

	No tiene HTA	Tiene HTA
No tiene HTA	679	458
Tiene HTA	290	960

CONCLUSIONES

En conclusión, al comparar los diferentes métodos de clasificación por medio de los indicadores de desempeño (E, FN, FP, PP, PN, AN, AP, AUC, KS) se muestra que, entre las técnicas menos adecuadas para realizar la clasificación de individuos, se encuentran la regresión logística binomial, bagging y boosting ya que su desempeño en comparación con el de los demás no es tan bueno. Por otro lado, la técnica de los vecinos más cercanos se desempeña mejor que los mencionados anteriormente, sin embargo, no es la técnica ideal.

Quienes muestran el mejor desempeño son los árboles de decisiones y bosques aleatorios ya que presentan la mayor cantidad favorable de indicadores, sin embargo, al compararlos entre sí, quien obtiene un óptimo desempeño es la técnica de bosques aleatorios, por lo que se considera como el mejor método de clasificación. Por lo tanto, en términos de clasificaciones futuras el método con mejor desempeño para clasificar los individuos según su estado de presión arterial son los bosques aleatorios.

Un punto importante de notar es que todos los métodos, incluso el elegido como clasificador final presentan altos errores de clasificación lo que podría significar que ninguno de los modelos es suficiente para aprender o entrenar la relación entre los datos de entrada y salida y podría darse un subajuste de los mismos, resultando en un rendimiento pobre del modelo.

Finalmente, cabe destacar como limitante a este estudio, la falta de investigaciones previas sobre la clasificación de personas que tienen o no hipertensión arterial, ya que los estudios previos son un apoyo para entender mejor el problema de investigación.

BIBLIOGRAFIA

Berástegui, G., Galar, M. (2018). Implementación del algoritmo de los k vecinos más cercanos (k-NN) y estimación del mejor valor local de k para su cálculo. https://academica-e.unavarra.es/bitstream/handle/2454/29112/Memoria.pdf?sequence=2

Berenguer, L .(2016).Algunas consideraciones sobre la hipertensión arterial. *MEDISAN, 20(11).* http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1029-30192016001100015

Domínguez, M. (2021). Regresión Logística y métodos de Aprendizaje. Aplicaciones. https://zaguan.unizar.es/record/110300/files/TAZ-TFG-2021-3094.pdf

Giraldo, A. (2018). DESARROLLO Y APLICACIÓN DE LA METODOLOGÍA BAGGING Y ADABOOST PARA LA DETECCIÓN DE PÉRDIDAS NO TÉCNICAS EN EL SISTEMA DE DISTRIBUCIÓN DE LA EMPRESA DE ENERGÍA DE PEREIRA S.A. ESP. https://repositorio.utp.edu.co/server/api/core/bitstreams/77fbeabd-e0dc-4fd4-80dd-21b01162ee17/content

James, G., Witten, D., Hastie, T., Tibshirani, R. (2021). An Introduction to Statistical Learning with Applications in R. Springer, *Second Edition.*

Medina, R., Ñique, I. (2017). Bosques aleatorios como extensión de los árboles de clasificación con los programas R y Python. ISSN 1993-4912. :10.26439/interfases2017.n10.1775

Ministerio de Salud. (2022). 53 personas son diagnosticadas diariamente con hipertensión arterial.https://www.ministeriodesalud.go.cr/index.php/prensa/52-noticias-2022/1311-53-personas-son-diagnosticadas-diariamente-con-hipertension-arterial

Morera, M., Cortés, A. (2021). Análisis multinivel del control óptimo de pacientes hipertensos en la atención primaria en Costa Rica. *Gestión en Salud y Seguridad Social*, *1(1), ISSN: 2215-6216.*
https://www.binasss.sa.cr/ojssalud/index.php/gestion/article/download/176/315/

Ortiz, R., Torres, M., Peña, S., Alcántara, V., Supliguicha, M., Vasquez, X., Añez, R., Rojas, J., Bermúdez, V. (2017) .Factores de riesgo asociados a hipertensión arterial en la población rural de Quingeo Ecuador. *Revista Latinoamericana de Hipertensión, 12(3),. 95-103*. https://www.redalyc.org/pdf/1702/170252187004.pdf

Osorio, E., Amariles, P. (2017).Hipertensión arterial en pacientes de edad avanzada: una revisión estructurada. *Revista Colombiana de Cardiología, 25(3), 209-221* https://doi.org/10.1016/j.rccar.2017.10.006

Trujillano, J., Sarria, A., Esquerda, A., Badia, M., Palma, M., March, J. (2008). Aproximación a la metodología basada en árboles de decisión (CART). Mortalidad hospitalaria del infarto agudo de miocardio. Approach to the methodology of classification and regression trees. *Gaceta Sanitaria, 22(1), 65-72.* https://www.sciencedirect.com/science/article/pii/S0213911108712044