Instituto Tecnológico y Estudio Superiores de Monterrey

Modelación del aprendizaje con inteligencia artificial
TC2034.101

E2 Proyecto Aprendizaje no supervisado:
Reporte

Profesor/a:
Dra. María Valentina Narváez Terán

Autoría:
David Fernando Armendáriz Torres
A01570813

14 de marzo de 2023

INTRODUCCIÓN

Las tecnologías de la información en la época moderna se han cimentado como un aspecto vital de la sociedad humana. El acelerado crecimiento de la industria tecnológica ha facilitado, y requerido, el desarrollo e implementación de diversos métodos capaces de procesar las cantidades masivas de datos que se generan día a día. Estos métodos buscan producir valor a raíz de datos brutos.

Una rama de las ciencias computacionales la cual se ha posicionado a la vanguardia del creciente movimiento tecnológico es la Inteligencia Artificial. Encapsulado dentro de la Inteligencia Artificial se encuentra el aprendizaje automatizado el cual será el foco de este reporte.

El aprendizaje automatizado, busca simular el aprendizaje humano a raíz del ajuste de diversos algoritmos a un set de datos proporcionado. A su vez, éste se subdivide en dos ramas, aprendizaje supervisado y aprendizaje no supervisado. El aprendizaje supervisado se da cuando se utilizan datos etiquetado para el entrenamiento de los algoritmos con fines de clasificación y predicción. En cambio, el aprendizaje no supervisado busca analizar sets de datos carentes de etiquetas de clase en busca de patrones y similitudes con base en las cuales generar agrupaciones de datos (IBM, s.f.a).

A lo largo de este reporte se implementan los algoritmos de Bosques Aleatorios, Regresión Logística, K-Means Clustering y DBSCAN Clustering a un set de datos pertinente a la problemática de salud de cáncer de mama. Se profundiza mas acerca de estos temas más adelante.

PROBLEMÁTICA

La Inteligencia Artificial, y en especifico el Aprendizaje Automatizado, tiene diversas aplicaciones como el reconocimiento de voz, detección de fraudes y manejo de inversiones, entre otras. No obstante, aquella pertinente a este escrito es la aplicación en el área de la medicina, donde se utiliza como una herramienta para facilitar el proceso de investigación, detección y diagnóstico.

La problemática sobre la cual se centra este reporte corresponde al cáncer de mama. Este cáncer es el más prevalente en el mundo; según un artículo de la Organización Mundial de la Salud, *cerca de una de cada 12 mujeres enfermarán de cáncer de mama a lo largo de su vida*. La tasa de supervivencia a mas de 5 años de pacientes con este padecimiento es superior al 90% en países primermundistas, sin embargo, decrece considerablemente en países tercermundistas y en desarrollo (WHO, 2021). Adicionalmente el tratamiento suele ser sumamente eficaz si se detecta de forma temprana, aquí es donde yace la contribución de valor; se busca demostrar la eficacia y utilidad de aplicar modelos de aprendizaje automatizado para realizar diagnósticos preliminares de tumores malignos (cancerígenos) o benignos (no cancerígenos).

DATASET

El data set utilizado, *Breast Cancer Wisconsin Dataset*, posee información correspondiente a características físicas de tumores de pecho detectados en 569 mujeres. Éste cuenta con 32 variables; 30 numéricas correspondientes a dichas características; una categórica correspondiente a la etiqueta de clase, *diagnosis*, la cual indica si el diagnóstico del paciente es un tumor benigno (B) o maligno (M); y una llave primaria, *id*.

Previo al entrenamiento y manejo de los modelos, se realizó un proceso de limpieza y preprocesamiento de los datos con el objetivo de facilitar su ajuste y acatar de mejor manera las limitaciones de los modelos.

Variables del dataset:

- Llave primaria: id
- Etiqueta de clase: diagnosis
- Numéricas:
    - radius_mean
    - texture_mean
    - perimeter_mean
    - area_mean
    - smoothness_mean
    - compactness_mean
    - concavity_mean
    - concave points_mean
    - symmetry_mean
    - fractal_dimension_mean
    - radius_se
    - texture_se
    - perimeter_se
    - area_se
    - smoothness_se
    - compactness_se
    - concavity_se
    - concave points_se
    - symmetry_se
    - fractal_dimension_se
    - radius_worst
    - texture_worst
    - perimeter_worst
    - area_worst
    - smoothness_worst
    - compactness_worst
    - concavity_worst
    - concave points_worst
    - symmetry_worst
    - fractal_dimension_worst

Limpieza de Datos

Dado que el set de datos no contenía registros nulos, la eliminación de estos no fue necesaria. Ya que la columna, *id*, no provee información relevante para los modelos, esta fue eliminada.

Adicionalmente para evitar un sesgo en dirección de aquellas variables con ordenes de magnitud mayores o menores, se normalizaron los datos numéricos.

Preprocesamiento

Aprendizaje Supervisado

Exclusivo a la implementación de los datos en los modelos de aprendizaje supervisado, se transformó la variable correspondiente a las etiquetas de clase, *diagnosis*, en una de tipo dummy al asignar los valores binarios 0 (benigno) y 1 (maligno).

Adicionalmente, se realizó una segmentación de los datos, dividiéndolos en la variable dependiente o target, *diagnosis,* y las independientes, es decir, todas menos *diagnosis* e *id.*

Aprendizaje No Supervisado

Exclusivo a la implementación de los datos en los modelos de aprendizaje no supervisado, se eliminaron las etiquetas de clase, *diagnosis*, para propiciar una naturaleza amena a modelos de clustering.

Adicionalmente, clusters generados utilizando las 30 variables numéricas como criterios de entrenamiento se verían afectados por la Maldición de la Dimensión. Los problemas de alta dimensionalidad al generar clusters surgen como resultado del hecho de que una cantidad fija de puntos de datos se vuelven mas dispersos conforme se incrementan las dimensiones (Steinbach et álii, 2004). Para lidiar con este obstáculo, se utilizó la técnica de reducción de dimensionalidad de Análisis de Componentes Principales.

El análisis de componentes principales permite simplificar un set de datos al reducir su dimensionalidad a modo que se conserve la mayor parte de la información. Esto se logra al generar un conjunto de componentes principales, es decir, combinaciones lineales de las variables iniciales sin correlación entre sí que maximizan la varianza y por ende son representativos de una parte significativa de la información del dataset original.

Utilizando esta técnica se redujo el dataset a dos componentes principales con base en los cuales se generaron y graficaron los clusters.

	PC1	PC2
0	0.199095	-0.016644
1	0.036075	-0.062749
2	0.114292	-0.023569
3	0.120728	0.115834
4	0.093005	-0.027815
...	...	...
564	0.151367	-0.068353
565	0.075628	-0.053767
566	0.028349	-0.001966
567	0.202568	0.029699
568	-0.097152	-0.015005

Ilustración 1 Primero y ultimos elementos de los componentes princpales generado

MODELOS DE APRENDIZAJE SUPERVISADO

Se implementaron los siguientes modelos de aprendizaje supervisado con el fin de, una vez ajustados al set de datos, facilitar la clasificación de tumores en benignos o malignos.

Bosque Aleatorio

Un algoritmo de Bosque Aleatorio es un ensamble de múltiples árboles de decisión los cuales trabajan en conjunto para generar un solo resultado. A su vez los árboles de decisión son algoritmos de aprendizaje supervisado que buscan encontrar la mejor división para particionar los datos en subconjuntos, esto se ve reflejado en los nodos decisión los cuales clasifican con base en el criterio de la partición. Métricas como la impureza de Gini y el error cuadrático medio pueden utilizarse para evaluar y seleccionar la calidad de la partición (IBM, s.f.b).

La implementación del modelo se llevó a cabo con las siguientes consideraciones e híperparámetros:

- Holdout Cross Validation: División aleatoria del set de datos en una proporción de 70% para el entrenamiento y 30% para la evaluación.
- Cantidad de estimadores (árboles): 10
- Criterio de calidad de partición: Impureza de Gini
- Cantidad Máxima de niveles: 5
- Función Python: sklearn.ensemble.RandomForestClassifier()

Regresión Logística

Los modelos de Regresión Logística estiman la probabilidad de suceso de un evento y son principalmente empleados en la clasificación y predicción. En un contexto de aprendizaje automatizado, dicho evento corresponde a la variable objetivo, la cual debe ser

dicotómica y se clasifica con base en la probabilidad, menor o mayor a 0.5 que predice o clasifica en 0 o 1 respectivamente (IBM, s.f.c).

Las variables independientes manejadas por el modelo deben carecer de multicolinealidad. Con esto en mente se utilizó RFE, *Recurssive Feature Extraction*, para seleccionar las variables más representativas. Una vez hecho se aplicó el modelo de Regresión Logística a estas mismas, tomando como variable target *diagnosis*. A partir de esto se eliminaron las variables con alta correlación entre sí; se utilizó la regla de decisión de eliminar aquellas con a α mayor a 0.05, es decir, si p-value<0.05 entonces la variable es significativa.

La implementación del modelo se llevó a cabo con las siguientes consideraciones e híper parámetros:

- Holdout Cross Validation: División aleatoria del set de datos en una proporción de 70% para el entrenamiento y 30% para la evaluación.
- Variables seleccionadas con RFE: 10
    - Variables independientes utilizadas: 'area_mean', 'concave points_mean', 'radius_se', 'area_se' & 'area_worst'.
- Función Python: sklearn.linear_model.LogisticRegression()

MODELOS DE APRENDIZAJE NO SUPERVISADO

Se implementaron los siguientes modelos de aprendizaje no supervisado con el fin de, una vez ajustados al set de datos, facilitar la clasificación de tumores en benignos o malignos.

K-Means Clustering

K-Means es un algoritmo basado en centroides a los cuales se les asignan los objetos mas cercanos generando clusters los cuales se van ajustando mediane un proceso iterativo utilizando las medias de los datos intra-cluster. Se busca minimizar la suma de las distancias entre los datos y el centroide del cluster identificando así el grupo correcto al cual pertenecen (Sharma, 2023).

Ya que el modelo de K-Means requiere una cantidad de k predeterminada de clusters los cuales generar, se selecciona la cantidad ideal utilizando el método del codo. Este método consiste en graficar la suma de errores cuadráticos con respecto al número de clusters, aquella cantidad ideal será donde se ubique el punto de inflexión mas significativo, aquel que se asemeje a un “codo”.

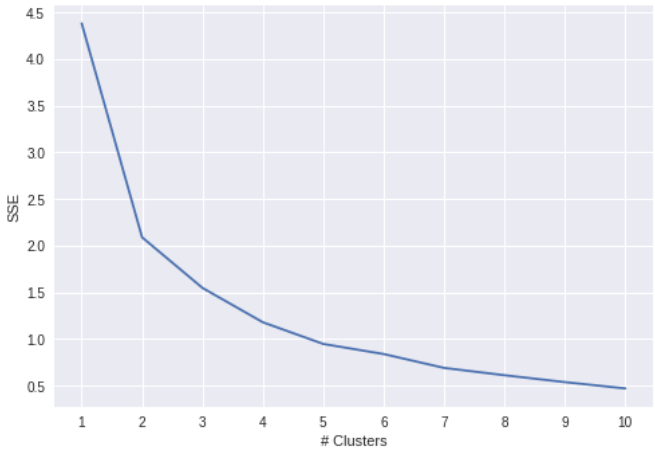


Ilustración 2 Gráfica de codo (SSE) con punto de inflexión en 2 clusters.

La implementación del modelo se llevó a cabo con las siguientes consideraciones e híper parámetros:

- Número k de clusters: 2, este resultado es observable en la ilustración 2.
- Función Python: sklearn.cluster.KMeans()

DBSCAN Clustering

DBSCAN, *Density-based spatial clustering of applications with noise*, funciona en base a la densidad de los datos. El algoritmo busca identificar regiones, clusters, con alta concentración de datos separadas por áreas vacías. Aquellos elementos que no pertenecen a ningún cluster son etiquetados como ruido (ArcGIS Pro 3.0, s.f.).

El algoritmo DBSCAN requiere la especificación de dos híper parámetros épsilon y puntos mínimos. Épsilon especifica la vecindad dentro de la cual deben recaer dos puntos para ser considerados miembro de un cluster, esto se determina mediante el método del codo graficando la distancia mínima promedio de un punto a sus vecinos. La cantidad de puntos mínimos simplemente especifica la cantidad mínima de puntos requeridos para formar un cluster, esta en cambio se seleccionó arbitrariamente mediante un proceso de prueba y error buscando aquel que diera los mejores resultados.

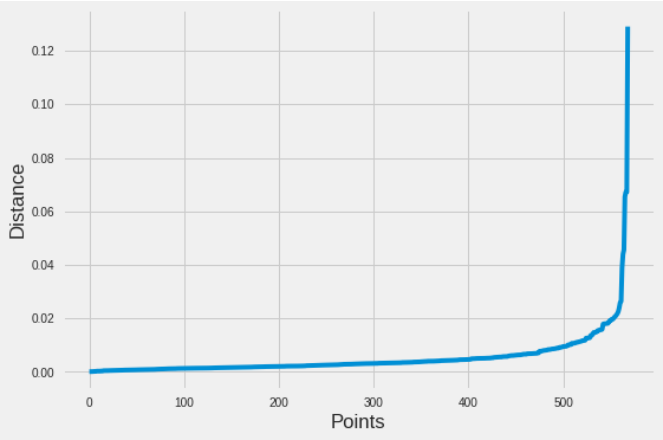


Ilustración 3 Gráfica de codo (Distancia mínima entre puntos vecinos) con punto de inflexión en distancia de 0.02.

La implementación del modelo se llevó a cabo con las siguientes consideraciones e híper parámetros:

- Épsilon: 0.02, este resultado es observable en la ilustración 3.
- Puntos mínimos: 15
- Función Python: sklearn.cluster.DBSCAN()

RESULTADOS

A continuación, se despliegan y analizan los resultados generados como resultado de aplicar los 4 modelos de aprendizaje automatizado.

Bosque Aleatorio

Utilizando como métrica la *accuracy* promedio, es decir la proporción de predicciones correctas con respecto a las predicciones totales, se obtuvo un resultado de 0.959. Asimismo, el modelo cuenta con una precisión de 0.924 y sensibilidad de 0.968 resultados propios de un modelo con un manejo sobresaliente de la clase, es decir puede detectar con facilidad y confianza tumores malignos.

Adicionalmente se generó una matriz de confusión para observar el comportamiento del modelo y de tal forma evaluar su precisión de forma porcentual. En el primer cuadrante se observa que el modelo acertó en un 95% al clasificar los verdaderos negativos; es decir los pacientes que presentaban tumores benignos; de la misma manera se puede observar en el cuarto cuadrante que el 97% de las ocasiones se clasificaron correctamente los verdaderos positivos, es decir pacientes que poseían tumores malignos. En cambio hay un bajo porcentaje de falsos positivos y falsos negativos lo que indica que pocas veces se generaron predicciones erróneas.

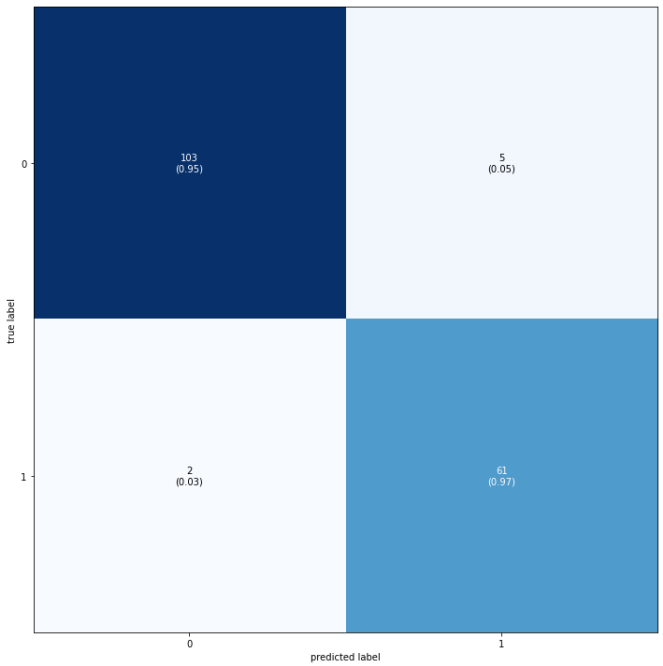


Ilustración 4 Matriz de Confusión de predicciones de Bosque Aleatorio

Estos resultados indica que el modelo generalizó bastante bien y resulta bastante efectivo en el proceso de predicción de si un tumor es benigno o maligno.

Regresión Logística

Manejando los mismos criterios de evaluación que el bosque aleatorio, las predicciones generadas con regresión logística generaron un *accuracy* del 0.690, una precisión de 1 y sensibilidad de 0.159.

A su vez la matriz de confusión indica que el modelo acertó el 100% de las veces al clasificar los verdaderos negativos, sin embargo, solamente el 16% de las veces clasificó los datos correctamente como verdaderos positivos. Más importante aún el modelo clasifico falsos negativos en un 84% de las ocasiones.

Esto indico que el modelo fue inefectivo en la predicción de la clasificación de tumores. Específicamente, el modelo posee dificultades para detectar tumores malignos, no obstante, cuando lo hace es altamente confiable.

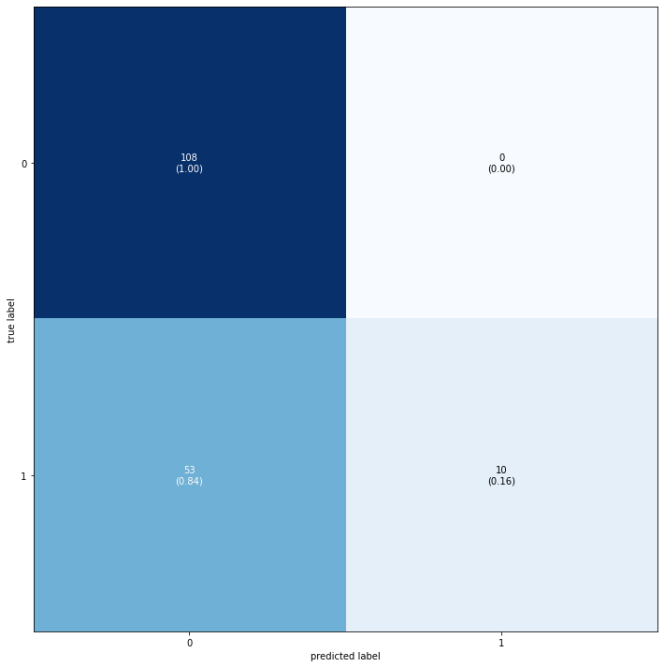


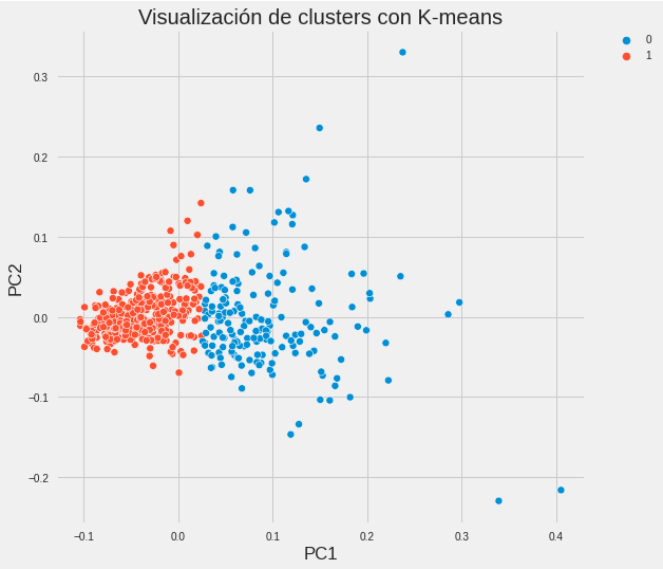
Ilustración 5 Matriz de Confusión de predicciones de Regresión Logística

K-Means Clustering

El proceso de la evaluación de los resultados de agrupamiento generado por los modelos de aprendizaje no supervisado se limitó a el coeficiente de silueta promedio de los datos y la visualización e interpretación de los clusters.

El coeficiente de silueta, una métrica que indica que tan similar es un dato a aquellos de su mismo cluster genera resultados en un rango de -1 a 1. Un valor cercano a -1 indica que los datos fueron posicionados en clusters erróneos, un valor cercano a 0 indica que fueron asignados aleatoriamente y uno cercano 1 indica que se asignaron correctamente. Este coeficiente fue de 0.546 para este método lo cual es un resultado no favorable.

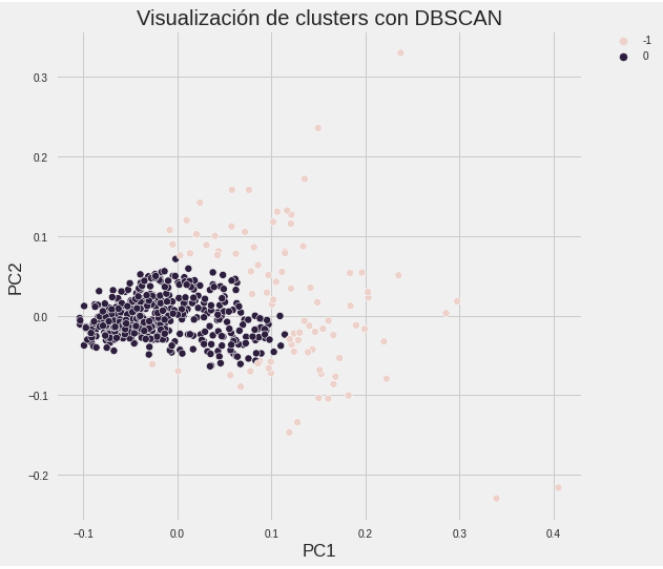
En el algoritmo de K-Means implementado se manejan dos grupos como la cantidad ideal lo cual concuerda con el comportamiento real de los datos, los cuales originalmente contaban con 2 etiquetas de clase. A su vez, el agrupamiento tiende a una forma esférica la cual parece seccionar los datos, esto debido a las suposiciones que el modelo hace con respecto a la densidad de los datos.



DBSCAN Clustering

Nuevamente, bajo los mismos criterios de evaluación de K-Means, el modelo generado con DBSCAN resulta en un coeficiente de silueta de 0.521, también un resultado deficiente en el agrupamiento de los tumores.

De manera similar, con base en los híper parámetros, el algoritmo generó automáticamente 2 clusters para agrupar los datos. Esto en cambio, no tienden hacia una forma esférica como su contraparte.



CONCLUSIONES

Tras el previo análisis de resultados se es fácil determinar que el modelo que mejor se ajusta a los datos y generaliza las características de los tumores es el modelo de Bosques Aleatorio con una sobresaliente accuracy de 0.959, precisión de 0.924 y sensibilidad de 0.968. Esto indica, como se mencionó previamente que el modelo genera predicciones acerca de si un tumor es benigno o maligno con facilidad y un alto nivel de confianza.

El resto de los modelos no resultaron en niveles de desempeño aceptables. El modelo de Regresión Logística difícilmente predice falsos positivos respecto a la presencia de tumores, sin embargo, genera demasiados falsos negativos, lo cual en un ámbito médico conlleva un riesgo inaceptable.

De manera general, el desempeño de los modelos de clustering, aunque no completamente equiparable, fue inferior al de aquellos de clasificación, con coeficientes de silueta oscilando en 0.50.

Como explicación para estos resultados se presentan los siguientes puntos:

- La implementación del método de reducción de dimensionalidad PCA, resultó en perdida de información la cual perjudicó el desempeño de los modelos no supervisados.
- K-Means asume una forma de cluster esférica y DBSCAN asume la misma densidad para todos los clusters, ambos aspectos, al no ser representativos de la distribución de los datos, presentan un detrimento para el agrupamiento.
- Regresión Logística se desempeña mal cuando se poseen pocos registros o las variables independientes están correlacionadas entre sí. Ambas siendo características de la base de datos al esta poseer 569 registros y variables correlacionadas al estas ser características físicas de un mismo tumor.
- El modelo de Bosques Aleatorios, comúnmente considerado como la Panacea de los modelos de clasificación, es relativamente inmune a la presencia de variables correlacionadas y a la presencia de outliers y al ser un modelo de ensamble reduce el sesgo. Características que propiciaron el buen desempeño a pesar de los ‘defectos’ del dataset.

Adicionalmente se presentan las siguientes recomendaciones:

- Incrementar la cantidad de registros para fomentar un mejor generalización y desempeño.
- Utilizar estrategias de optimización de híper parámetros como GridSearchCV y RandomizedSearchCV.
- Implementar otros métodos de reducción de dimensionalidad que mejor se ajusten al dataset y reduzcan la cantidad de información perdida.

Finalmente, con un enfoque mas centrado en la problemática, se puede apreciar la utilidad de métodos de aprendizaje automatizado en el área de salud. Aunque estos algoritmos no posean un nivel de efectividad que permita remplazar el rol de un médico profesional, estos pueden ser implementados para facilitar y agilizar las labores de estos. Dichos métodos pueden ser implementados para generar clasificaciones preliminares con

base en las cuales priorizar la atención a ciertos pacientes o generar predicciones preventivas del estado de salud del paciente.

Referencias

ArcGIS Pro 3.0. (s.f.). Cómo funciona el clustering basado en densidad. Recuperado en Marzo 12, 2023, de https://pro.arcgis.com/es/pro-app/latest/tool-reference/spatial-statistics/how-density-based-clustering-works.htm

Cáncer de Mama. (s.f.). Recuperado en Marzo 1, 2023, de https://www.who.int/es/news-room/fact-sheets/detail/breast-cancer#:~:text=Cerca%20de%20una%20de%20cada,como%20consecuencia%20de%20esa%20enfermedad.

IBM. (s.f.a). What is machine learning? Recuperado en Marzo 12, 2023, de https://www.ibm.com/topics/machine-learning?mhsrc=ibmsearch_a&mhq=what+is+machine+learning

IBM. (s.f.b). What is Random Forest? Recuperado en Marzo 12, 2023, de https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.

IBM. (s.f.c). What is logistic regression? Recuperado en Marzo 13, 2023, de https://www.ibm.com/topics/logistic-regression#:~:text=Resources-,What%20is%20logistic%20regression%3F,given%20dataset%20of%20independent%20variables.

Sharma, N. (2023, Febrero 20). K-means clustering explained. Recuperado en Marzo 14, 2023, de https://neptune.ai/blog/k-means-clustering

Steinbach, M., Ertöz, L., & Kumar, V. (2004). The Challenges of Clustering High Dimensional Data. In L. T. Wille (Ed.), *New Directions in Statistical Physics: Econophysics, Bioinformatics, and Pattern Recognition* (pp. 273–309). doi:10.1007/978-3-662-08968-2_16

Bibliografía

Análisis de Componentes Principales (principal component analysis, PCA ... (s.f.). Recuperado en Septiembre 5, 2022, de https://www.cienciadedatos.net/documentos/35_principal_component_analysis

A step-by-step explanation of principal component analysis (PCA). Built In. (s.f.). Recuperado en Septiembre 5, 2022, de https://builtin.com/data-science/step-stepexplanation-principal-component-analysis

IBM. (s.f.). Artificial Intelligence in medicine. Recuperado en Marzo 14, 2023, de https://www.ibm.com/topics/artificial-intelligence-medicine#:~:text=How%20is%20artificial%20intelligence%20used,health%20outcomes%20and%20patient%20experiences.

Sala Pulido, J. C., Márquez Campos, D., Carsellé Aldac, S., Armendariz Torres, D. F., & Lugo Gutiérrez, V. E. (2023, Marzo 2). Entregable 1: Situación problema: Proyecto de aprendizaje supervisado [Digital image]. Recuperado en Marzo 12, 2023, de https://docs.google.com/document/d/1UH4W526Tu_gJMZitj9FJPjHjVYac0-2n44msF-xZB5Y/edit?usp=sharing

González, L. (2021, Septiembre 08). DBSCAN Teoría. Recuperado en Marzo 13, 2023, de https://aprendeia.com/dbscan-teoria/

Lee, W.-M. (2022, Febrero 2). Using principal component analysis (PCA) for Machine Learning. Medium. Recuperado en Septiembre 5, 2022, de https://towardsdatascience.com/using-principal-component-analysis-pca-formachine-learning-b6e803f5bf1e

Gutiérrez López, R. I., M.C. (2023, Marzo). *4. Exploración de métricas de desempeño de un modelo* [PDF]. Monterrey: ITESM | Departamento Acad´emico: Computaci´on.