ARTIFICIAL INTELLIGENCE
Machine Learning Training:
Heart Classification

Diana Velásquez, Karen Zambrano

Abstract

En este artículo se estudió y aplicó el uso de tres algoritmos de supervised learning como: logistic regression (regresión logística), decision tree classification (clasificación del árbol de decisión) y random forest classification (clasificación aleatoria de bosques), para poder llevar a cabo el entrenamiento de machine learning y así determinar si una persona tiene o presentará problemas cardiacos mediante ciertos patrones y características de la base de datos seleccionada.

I. ASIGNACIÓN

El objetivo de esta investigación es realizar el entrenamiento de los modelos de machine learning y así evaluar el rendimiento de cada uno de los algoritmos de clasificación para seleccionar el que mayor grado de exactitud proporcione al momento de predecir si una persona es más propensa a padecer o no problemas cardiacos.

II. INTRODUCCIÓN

La inteligencia artificial (IA) es la simulación de la inteligencia humana en máquinas que están programadas para pensar y aprender. La IA se puede utilizar para realizar tareas que normalmente requieren inteligencia humana, como la percepción visual, el reconocimiento de voz, la toma de decisiones y la comprensión del lenguaje. El campo de la investigación de la IA se fundó en la creencia de que se puede hacer que una máquina piense como un ser humano si se utilizan los métodos correctos. Las aplicaciones de IA incluyen automóviles autónomos, asistentes personales virtuales y sistemas de diagnóstico médico.

Machine Learning es una rama de la inteligencia artificial que permite que las máquinas aprendan sin ser expresamente programadas para ello. Una habilidad indispensable para hacer sistemas capaces de identificar patrones entre los datos para hacer predicciones. Además, machine learning se clasifica en: supervised learning(datos de entrada y salida), unsupervised learning (no hay datos de salida pero si de entrada) y por último, reinforcement learning que es un aprendizaje en tiempo real, es decir, va aprendiendo de prueba error y asi es como se va autoajustando sus parámetros y se produce el modelo matemático.

Por otro lado, para poder realizar un determinado entrenamiento de un modelo de aprendizaje mediante la técnica de supervised learning se debe tener una base de datos. Además, para este tipo de entrenamiento supervisado hay dos tipos de problemas: clasificación y regresión. La base de datos a utilizar debe tener caracteristicas o parámetros de entrada y salida (respectiva etiquetas) para el problema de clasificación. En cambio, en el problema de regresión la diferencia radica en que la salida son valores continuos.

Desde luego, existen varios algoritmos que se pueden usar para problemas de clasificación como en el caso de estudio, algunos de los más populares incluyen: logistic regression, naive bayes, decision trees, random forest, support vector machines, neural networks, entre otros. Sin embargo, es importante tener en cuenta que el mejor algoritmo para un problema de clasificación en particular dependerá de factores como el tamaño y la estructura de los datos, la complejidad del problema y los recursos disponibles.

1) Logistic Regression

La regresión logística es útil cuando desea predecir la presencia o ausencia de una característica o resultado en función de un conjunto de valores predictores. Esto es similar a un modelo de regresión lineal, pero es adecuado para modelos con variables dependientes dicotómicas. Puede usar los coeficientes de regresión logística para estimar la razón de verosimilitud para cada variable independiente en su modelo. La regresión logística es aplicable a una gama más amplia de situaciones de investigación que el análisis discriminante. La regresión logística modela la probabilidad de que cada entradapertenezca a una categoría particular. Una función toma entradas y devuelve salidas. Para generar probabilidades, la regresión logística utiliza una función que da salidas entre 0 y 1 para todos los valores de X. Hay muchas funciones que cumplen con esta descripción, pero la utilizada en este caso es la función logística (sigmoide).

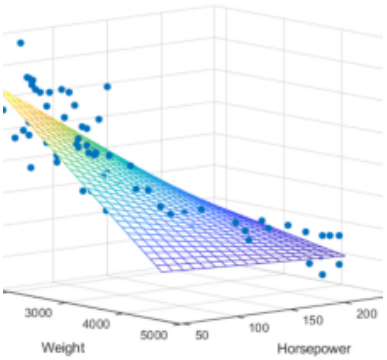

Fig. 1. An example of logistic regression. [1]

2) Decision Tree Classification

Es un tipo de algoritmo de aprendizaje supervisado utilizado principalmente en problemas de clasificación, pero funciona con variables de entrada y salida tanto categóricas como continuas. Los árboles de decisión identifican las variables más importantes y sus valores que producen el mejor conjunto homogéneo de población. Los métodos basados en árboles potencian modelos predictivos con alta precisión, estabilidad y facilidad de interpretación. A diferencia de los modelos lineales, mapean bastante bien las relaciones no lineales.

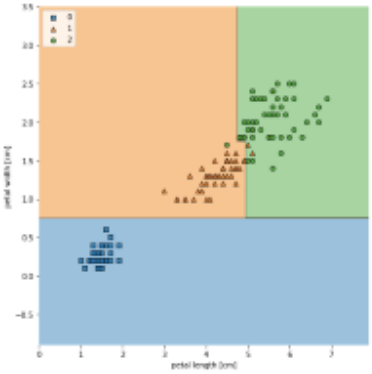

Fig. 2. An example of decision tree classification. [2]

3) Random Forest Classification

La clasificación aleatoria de bosques es una técnica de aprendizaje automático supervisada basada en árboles de decisión. Su principal ventaja es que obtiene un mejor rendimiento de generalización para un rendimiento durante entrenamiento similar. Esta mejora en la generalización la consigue compensando los errores de las predicciones de los distintos árboles de decisión. Para asegurarnos que los árboles sean distintos, lo que hacemos es que cada uno se entrena con una muestra aleatoria de los datos de entrenamiento. Esta estrategia se denomina bagging.

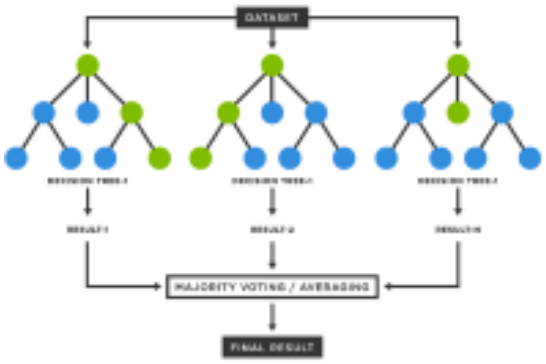

Fig. 3. An example of random forest classification. [3]

III. METODOLOGÍA

Todos los algoritmos implementados tienen sus parámetros cuyo valor se establece antes de que comience el proceso de aprendizaje, los valores de otros parámetros se obtienen a través del entrenamiento. Cuando se utiliza cualquier tipo de clasificador descrito anteriormente, se tienen en cuenta los parámetros predeterminados. Sin embargo, estos parámetros predeterminados muchas veces no son la elección correcta para la predicción óptima del algoritmo. De modo que nos vimos obligados a cambiar la mayoría de ellos, o al menos los más importantes, aquellos que pueden ser aptos para llevar a cabo el entrenamiento. Se analizan evaluaciones como la característica para comprender mejor cual está funcionando bien o no. Los algoritmos descritos en la parte de introducción fueron cuidadosamente seleccionados. Aun así, para tener un mejor y más amplio conocimiento del desempeño de cada uno de los modelos.

IV. EXPERIMENTACIÓN

En esta apartado, se evalúan los efectos de los tres algoritmos de clasificación en función de diferentes parámetros y también diferentes tamaños de características para encontrar el modelo óptimo para determinar si una determinada persona puede padecer o no de problemas cardiacos, es decir, al corazón. Posteriormente, el rendimiento de los modelos de entrenamiento de machine learning se mide en términos de precisión. Además, los algoritmos descritos en la parte de introducción, fueron cuidadosamente seleccionados ya que son parte de los principales algoritmos de clasificación.

Clasificador	**Puntuación de precisión**
Clasification logistic regression	84,29%
Decision tree clasification	75,74%
Random Forest	85,23%

Tabla I. Acurracy Score

La Tabla I muestra el desempeño de los algoritmos. El modelo más adecuado para este tipo de tareas es Random forest classification y Logistic Regression. Por el contrario, decision tree classification ocupó el último lugar para determinar con exactitud la precisión, es decir, no es el adecuado para poder predecir si los pacientes sufren o no de problemas cardiacos.

En la base de datos "heart"que se empleó para este entrenamiento se verficó que hay tipos de datos objetos y enteros; los parámetros con tipo de datos objetos se deben transformar a valores númericos para que el modelo pueda comprender el tipo de problema a resolver. Para comprender la base de datos se debe tener en cuenta las siguientes características (fixtures) más significativas:

- Las columnas de la tabla de la base de datos es lo mismo que un parámetro.
- **ChestpainType:** Tipo de dolor que sufre en el pecho.
- **MaxHR:** Ritmo cardiaco nivel máximo.
- **HeartDisease:** Parámetros de salida de clasificación que sirve para determinar si una persona sufre o no del corazón.

En el siguiente gráfico se puede observar el rango de valores de la edad va de los 30 a 70 años, el promedio o media se encuentra aproximadamente en los 50 a 60 años. La mayor parte de los datos de las personas que presentan problemas caradiacos o no están por la media de los 50 años. Con respecto al patrón del colesterol, se puede observar que el rango de valores aceptables va desde los 400 hasta los 150, y entre esos valores esta el diagrama o campana de distribución. Dentro de esos valores hay un sector en la cual hay valores medios que estan aproximadamente por los 200. Pero exiten ciertos valores fuera de los rangos aceptables tanto en la parte superior como inferior. Lo que significa que esa caracteristica presenta ruido y por ende son valores incorrectos o datos basuras. Mediante la gráfica se pudieron observar y analizar las columnas que no servían y para eliminarlas estas debian tener ciertas caracteristicas como poseer una media, limites superior e inferior, en la cual los datos dentro de esos límites son considerados válidos y los que se encuentran fuera son considerados basuras o ruidos. Es decir, de los 11 patrones se eliminaron tres columnas, quedando como resultado ochos patrones significativos para poder resolver este problema. Siendo asi más fexible y sencillo para entrenar el modelo.

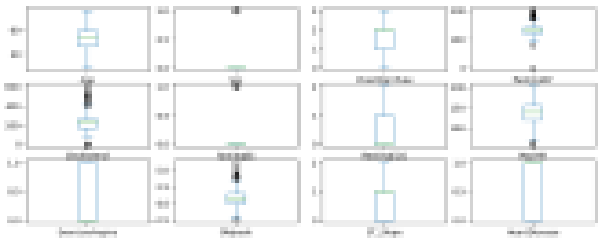

Fig. 4. Detección de valores anómalos o atípicos

VI. CONCLUSIÓN

En conclusion, se escogió la base de datos "heart" para realizar el entrenamiento del modelo de machine learning la cual determina si una persona tiene o va presentar problemas cardiacos de acuerdos a ciertos patrones o caracteristicas que se tomaron en cuenta. Ademas, para llevar acabo este entrenamiento se aplicaron los algoritmos o modelos como logistic regression, random forest y decision tree classification los cuales ayudan a identificar problemas de classificacion.

Gracias a las metricas de machine learning se puede evaluar el rendimiemto de los modelos en una base de datos y hay que terner en cuenta que existen diferentes métricas que son apropidas para diferentes tipos de problemas como la clasificación o regresión. Al momento de analizar la bases de datos "heart" se pudo establecer que era un problema de clasificación y como métrica de clasificación se trabajó con accuracy(exactitud), precision y recall. Las dos últimas métricas se encotraron embebidas dentro de la matriz de confusión.

Como resultado de este entrenamiento del modelo de machine learning se pudo determinar que el mejor algoritmo para utilizar en este problema de clasificación es el modelo random forest classification, con un puntaje de exactitud del 86% en la evaluación y el 100% en el entrenamiento. Ademas, se recomienda guardar estos modelos de machine learning para utlizarlos en trabajos futuros que sirvan como soportes para el diseño una aplicación tanto web como móvil.

REFERENCIAS

[1] https://www.tibco.com/es/reference-center/what-is-a-random-forest
[2] https://aprendeia.com/aprendizaje-supervisado-decision-tree-classification/
[3] https://www.ibm.com/docs/es/spss-statistics/saas?topic=regression-logistic