Modelo de aprendizaje automatico para evaluación de deportistas de alto rendimiento en artes marciales mixtas

Machine learning model for evaluation of high performance athletes in mixed martial arts

Velasco J.

***Abstract*—Machine learning is a field of research, which has multiple applications in the sports sector, being a support for athletes and coaches, since trends can be established from the results obtained. In recent years there has been an increase in athletes who want to establish themselves in this field for economic reasons, being a professional route with good benefits. With this, it can be said that it is necessary to know if an athlete establishes the necessary characteristics to be part of a team that presents promising results, so that they can access an incentive to excel in sport. This leads us to use the KDDbased methodology, so that a machine learning model can be evaluated and improved through analysis with different tests. Therefore, a clean data entry was established as a result, in which an exhaustive analysis of the data set was previously established, so that inconsistent values or those with an inappropriate structure could be found, being treated by statistical criteria. As a result, it was obtained that the best prediction algorithm was Random Forest with an accuracy of 75.55%, noting that the model specializes in correctly predicting negative values, since it has a high specificity.**

***Index Terms*—Data mining, Knowledge Discovery, Sports performance evaluation, Sports data mining**

***Resumen*—El aprendizaje automático es un campo de investigación, que tiene múltiples aplicaciones en el sector deportivo, siendo un apoyo para los deportistas y entrenadores, ya que se puede establecer tendencias a partir de los resultados obtenidos. En los últimos años se ha establecido un incremento de los deportistas que quieren establecerse en este campo por razones económicas, siendo una ruta profesional con buenas prestaciones. Con esto se puede decir que es necesario el saber si un deportista establece las características necesarias para formar parte de un equipo que presente resultados prometedores, así pueda acceder a un incentivo por destacar deportivamente. Esto nos lleva a usar la metodología basada en KDD, de manera que se pueda evaluar un modelo de aprendizaje automático y mejorarlo a través del análisis con diferentes pruebas. Por lo tanto, se estableció como resultado una entrada de datos limpia, en el que con anterioridad se estableció un análisis exhaustivo al conjunto de datos, de manera que se pudo encontrar valores inconsistentes o con una estructura poco apropiada, siendo tratados mediante criterios estadísticos. Como resultado se obtuvo que el mejor algoritmo de predicción fue Random Forest con una exactitud del 75,55 %, acotando que el modelo se especializa en predecir de manera correcta los valores negativos, ya que tiene una alta especificidad.**

***Palabras Claves*—Minería de datos, Descubrimiento de conocimiento, Evaluación del rendimiento deportivo, Minería de datos deportivos**

I. Introducción

EL deporte y la ciencia han tenido una sinergia en los tiempos actuales, donde gracias al uso de dispositivos electrónicos, que calculan todo tipo de métricas, se han generado millones de datos que pueden aplicarse para predecir el rendimiento de un deportista mediante aprendizaje automático [1]. En Ecuador debido a la reciente aparición de deportistas destacados, que tuvieron premiaciones en los juegos olímpicos de Tokio 2021, ha surgido el plan de alto rendimiento, en el que dicta una ayuda a los deportistas que tengan resultados prometedores, por el cual se establece un crecimiento del 46.20 % aproximadamente para el 2025, con respecto al numero actual de beneficiados por el plan actual [2], por lo que se puede decir que a pesar de los pocos recursos existentes a los deportistas ecuatorianos sin tomar en cuenta futbolistas, estos han sabido destacarse por motivación e inversión propia, de tal manera que llegaron a la vista del gobierno, imponiendo la idea que hay más deportistas que pudieran destacarse en competencias internacionales.

Se ha establecido un incremento de personas que buscan involucrarse en el medio deportivo, uno de los referentes deportes es el MMA (Mixed Martial Arts), con la empresa líder UFC (Ultimate Fighting Championship), que ha establecido ciertas pautas que dictan el ingreso de un peleador a ligas profesionales con directos en televisión de paga, como lo es haber tenido peleas profesionales debidamente homologadas y tener un rango de edad de 21 a 34 años [3], por lo que los datos generales de entrada podrán servir a las empresas para estimar posibles talentos. En estas situaciones es beneficioso usar como apoyo el aprendizaje automático, para dar contexto a los datos, de manera que se pueda clasificar a los diferentes deportistas por su habilidad. Hay que acotar que hay ciertos motivos que involucran que un deportista obtenga una victoria, que pueden relacionarse al tipo de peleador que sea, pudiendo observarse en sus estadísticas de pelea, aspectos como el número de golpes por minuto, defensa por minuto, porcentaje de golpes conectados; al tipo de complexión física, en las que dicta su posición de combate habitual, altura, peso, alcance de brazo, entre otros. Las características físicas son importantes para determinar un enfoque diferente con respecto a otras investigaciones, en las que involucran únicamente datos que se

establecen en peleas anteriores, a pesar de que esta temática se la menciona de manera muy superficial, no se la toma en cuenta por falta de datos; siendo destacable que en los deportes de contacto, estos datos son imprescindibles por diferentes aspectos tal como: requisitos previos a una pelea o clasificación de peleadores por peso y experiencia.

Es deseable que haya un nivel de aceptación alta con respecto al rendimiento del deportista (peleador), siendo el objetivo realizar un análisis estadístico basado en los enfrentamientos históricos y realizar implementaciones de algoritmos de aprendizaje automático, así podremos realizar estimaciones acerca del resultado de una próxima pelea. Para predecir la victoria del peleador capturamos y preprocesamos datos relevantes mediante un análisis estadístico y los aplicamos a los algoritmos SVM (Máquinas de vectores de soporte), que segmenta los datos; DT (Arboles de decisión), que integra capas en cascada con reglas de encaminamiento; RF (Arboles aleatorios), que integra varios DT generados de manera aleatoria con una respuesta mediante trabajo en conjunto [4]. El algoritmo Random Forest fue elegido para la investigación propuesta, por las diferentes problemáticas que se establecieron con respecto a los datos obtenidos del dataset "UFC-Fight historical data from 1993 to 2021" en el sitio web de Kaggle, en el que se pudo observar que hay una muestra pequeña y múltiples características, siendo útil el uso del algoritmo para el propósito de la clasificación de victoria del peleador [5], ya que la principal ventaja es que puede determinar información oculta a través de estimación de datos faltantes. Concluyendo con los resultados de cada uno de los algoritmos podemos analizar de manera comparativa cada uno de ellos, para determinar la factibilidad de los modelos de forma individual.

[1]

II. Trabajos Relacionados

En el trabajo de [6], establece una predicción mediante algoritmos de aprendizaje automática el rendimiento de un deportista (maratonista), para que así pueda servir como una referencia al deportista para que pueda ejercer o no, un cambio con respecto a su manera de entrenar. Para el uso de los algoritmos de aprendizaje profundo se atribuyó a dos tipos de aplicación, una en la que los datos se encuentren limpios y otra que los datos se encuentren con valores atópicos; al aplicar los algoritmos se pudo deducir que, debido a la limitada cantidad de datos, el algoritmo GRU tuvo una mayor precisión, siendo del 95 %. Con esto podemos acotar que debido a que los datos fueron preprocesados y limpiados con la regla de tres sigmas, estos ofrecieron una mayor precisión que los datos que no se realizó estas tareas.

Los datos que se atribuyen al deporte son muy tediosos de conseguir debido a que no hay un marco de trabajo para extraer estos, por lo que en [4], se establece los datos como un valor para las universidades al realizar programas educativos deportivos, ya que pueden adherirse a los planes de manera que puedan competir con países con mayor trayectoria deportiva. Para dar valor, se realizó un preprocesamiento de los datos, categorizando variables en numéricas y estandarizando para normalizar la distribución de los datos, así poder aplicar los algoritmos de aprendizaje automático; destacando SVM que tuvo una precisión de 76.14 %, en cambio los otros modelos obtuvieron valores, tal como: regresión logística con 75.44 %, Random Forest con 71.57 %, Decision tree con 66.66 %, K Nearest Vector con 72.28 %, Naïve Bayes con 62.11 % y XG Boost con 71.92 %. Los resultados arrojados por el algoritmo son adaptados como tablas de seguimiento de un deportista, para ver si su tendencia es a mejorar o empeorar.

Ciertos datos suelen ser innecesarios para realizar una predicción con una alta precisión, por lo que existen algoritmos de aprendizaje automático que permiten la eliminación de variables como lo es LSSVM (SVM mínimos cuadrados) así lo establece [7], estableciendo que características son más relevante para la variable independiente. El objetivo de la investigación se centró en predecir el rendimiento de un corredor de una carrera de 1000m, lo cual, para la elección del algoritmo de aprendizaje automática, se basó en la idea de que puede haber diferentes métricas que sean irrelevantes para la predicción, contemplando que un corredor (talentoso) puede realizar un salto más largo en un ejercicio determinado. El algoritmo que utilizan para el experimento, destacando sobre otros fue LSSVM-New con un error cuadrático de 0.1945, por otro lado, SVM obtuvo 0.1945 y BPNN dispuso de 0.3923.

Según [5], un algoritmo de predicción de SVM con optimización por enjambre de partículas puede utilizarse para universitarios que no sean atletas, refiriendo a que los datos pueden usarse como un medio para mejorar el rendimiento físico con respecto a medios tradicionales como lo son los profesores de educación física, anteponiéndose ante algoritmos como las redes neuronales y la regresión lineal. Los algoritmos que se dispusieron en la investigación, se utilizaron principalmente en carreras de 100m; estableciendo una precisión de 95.23 % para SVM, 88.63 % redes neuronales, 80.196 % regresión lineal, por tanto, se puede aplicar a diferentes ejercicios, sirviendo como un análisis de progreso estudiantil.

Al predecir un resultado habrá que tomar en cuenta que nuestro conjunto de datos debe ser equilibrado, de tal manera que cada una de las variaciones de nuestra variable independiente posean números significativos, esto se lo puede observar en [8], al usar técnicas de aprendizaje para predecir la victoria de un equipo de futbol, que se toma en cuenta las características individuales de cada jugador y del equipo en conjunto, establece que habrá una posibilidad casi nula de que haya una predicción de empate, ya que en el conjunto de datos se encuentran un número de predicciones de empate muy pequeña. En la investigación se compara diferentes algoritmos, siendo uno de los destacados gradient boosting con una precisión de 80,42 %, SVM 56,62 %, logistic regression 58,73 %, decision tree 51,85 % y redes neuronales 53,83 %.

Hay ciertas situaciones que involucran que una predicción pueda conllevar un mayor a error, según [9], un algoritmo de predicción puede fallar debido a la exigencia la que se enfrenta un deportista, evidenciando qué en enfrentamientos

[1]El aporte de este trabajo es detallar los resultados encontrados, de manera que se presenta el link para encontrar el proceso establecido: Repositorio de Github y dataset limpio. Las secciones de este documento son: *(ii)*. Trabajos relacionados, *(iii)*. Metodología, *(iv)*. Experimentos y Análisis, *(v)*. Resultados, *(vi)*. Discusión, *(vii)*. Conclusiones.

de campeonato hay un considerable retroceso en la precisión, por ello enfatiza su algoritmo a cambios constantes a partir de una actualización dinámica de Bayes. Con esto concluye que hay un mejor rendimiento con respecto a metodologías tradicionales estableciendo una precisión del 74 % utilizando regresión lineal, ya que hay datos empíricos que pueden cambiar la predicción durante un enfrentamiento deportivo.

En el trabajo de [10], establece una comparación entre algoritmos de clasificación y de clasificación basados en regresión, en el que se hace uso de redes neuronales, decision tree y SVM. En el estudio se realiza la recolección de datos en sitios web con información deportiva, así disponer de estadística de equipos de futbol americano (2011- 2012). Al realizar las respectivas predicciones a partir de datos de entrenamiento y prueba, se pudo observar que los algoritmos de clasificación eran superiores en predicción deportiva, específicamente el algoritmo que mayor precisión tenía era el árbol de decisión dando un margen de 86.86 %.

Según [11], se pueden realizar predicciones a partir de fórmulas de algebra lineal, que son habitualmente usados por las casas de apuestas, estableciendo que los algoritmos de aprendizaje automático, específicamente arboles de decisión, arboles aleatorios, redes neuronales; son superiores para la predicción, siendo poco eficientes los algoritmos habituales. Para establecer la comparación se utilizó matrices de confusión para aclarar que hay predicciones que no corresponden con la verdadera respuesta, siendo un tema de falla recurrente ”los empates”. Adicionalmente, se enfocó en aplicar los algoritmos en un entorno real (casa de apuestas), pudiendo las redes neuronales, generar una ganancia más alta con respecto a los demás algoritmos con un 450 % con respecto a la inversión inicial.

Con el tiempo se ha predispuesto usar algoritmos de aprendizaje automático para predicciones, pero es considerable el realizar un preprocesamiento de los datos, así lo explica Zhao et al. [12], siendo su objetivo que mediante predicciones estadísticas encontrar conocimiento a partir de los datos, aclarando que en sus resultados, dispuso de una mayor precisión con respecto a algoritmos de aprendizaje automático, en el que los datos no se encontraban predispuestos. El estudio fue realizado a partir de juegos universitarios de futbol americano y consistió en encontrar similitudes comparables entre equipos y enfrentamientos pasados, en el que se dicta por pesos a los diferentes equipos, distinguiendo como ganador al equipo con mayor peso, pudiendo discernir el error si estos no se hubieran enfrentado en ocasiones anteriores, obteniendo una precisión del 91.43 %.

En su trabajo [13], propone una metodología que dicta que se debe realizar una excavación y validación de los datos previa para poder determinar los factores de influencia que podrá determinar si un equipo o jugador generará una victoria y al generarse los nuevos datos, el modelo se reconstruirá anteponiendo reglas que permitan evaluar la predicción con los nuevos datos generados en el entorno real, con las predicciones del modelo. En el experimento se realizo dos pruebas utilizando la metodología, en la primera se aplico a carreras de autos obteniendo una precisión del 85 % y en la segunda prueba en proyectos de entretenimiento estipulando una precisión del 75 %.

En su investigación [14], utilizaron un enfoque de flujo de conocimiento con el software WEKA para poder predecir con técnicas de aprendizaje automático el equipo más efectivo para participar en competencias de Cricket. Para la predicción del modelo se lo aplico en la clasificación de los jugadores en sus respectivas posiciones, usando específicamente Decision tree, SVM, Random Forest; apreciando una mayor precisión el algoritmo Random Forest con un 95.78 %.

En su trabajo [15], aplicaron conceptos de aprendizaje automático para el diseño de diferentes métodos que podrán trabajar en conjunto para ser aplicados en entornos deportivos, extrayendo datos de un dispositivo (weareable). Se enfocaron en determinar la aplicación de algoritmos en situaciones específicas, agrupando por tipo de actividad, rendimiento, duración; clasificar por sus capacidades y predecir nuevos entrenamientos, con ello atribuir a un asistente personalizado para deportistas.

Zhang et al. [16], consideraron que en la predicción deportiva puede haber muchas variantes, por lo que propone utilizar un algoritmo no lineal, específicamente AS-LSTM. Especificaron que el algoritmo determinará mediante un enfoque de atención, las características que afectan a partir del tiempo y con un sliding time windows, dispondrá que los últimos registros tendrán una mayor influencia para la predicción, teniendo una precisión del 80 %.

Weissbock et al. [17], proponen una metodología basada en el meta aprendizaje, por el cual se dispone de dos capas en cascada para predecir la victoria en partidos de Hockey, en el cual se obtendrá los datos a partir de análisis estadístico y elementos textuales. Estipula que en la primera capa se usaron 3 clasificadores, en el primero se dispuso de la predicción mediante datos numéricos con una precisión 58.57 % aplicando redes neuronales; el segundo la predicción mediante ideas textuales con una precisión 57,32 % aplicando JRIT, un algoritmo basado en reglas; el tercero la predicción mediante emociones en frases textuales con una precisión 45,39 % aplicando Naive Bayes. Como enfoque establecieron metas clasificadores en la segunda capa de manera, que el primer clasificador se basará en disponer de la información de la primera capa en un modelo, específicamente SVM; el segundo en disponer de la predicción con mayor confianza de la primera capa y tercero mediante una votación mayoritaria, que tuvo una precisión 60.25 %, siendo superior al resto.

Konstatinos et al. [18], se enfoca mediante la elección de jugadores de futbol con una trayectoria lineal, la predicción de la posición del jugador; número de goles por temporada y tiros realizados en un partido. Usaron tres modelos de aprendizaje automático (SVM, RF, LR, SVC), lo cual el que tuvo una mayor precisión en todas las tareas fue Random Forest, lo cual para el experimento principal obtuvo una precisión del 81.5 %. Para la predicción tuvieron ciertos inconvenientes con datos inconsistentes, por el cuál segmentaron los datos en aspectos individuales con jugadores de un nivel de rendimiento alto.

III. Metodología

En esta sección se detalla la metodología utilizada, de manera que podamos percibir las diferentes tareas que se

llevarán a cabo. Esto se realizará con el uso de un enfoque, que permita establecer una estructura capaz de obtener un modelo que presenta una predicción de rendimiento deportivo. Por ello se presenta una serie de instrucciones que siguen la metodología KDD, de manera que nuestro modelo tenga una constante retrospectiva, que busca encontrar el máximo beneficio. Nuestro marco de trabajo como se lo ve en la Figura 1, se basará en que un modelo de aprendizaje automático resuelva la problemática de predecir un peleador para acceder a un beneficio deportivo, mediante un proceso de cinco pasos.

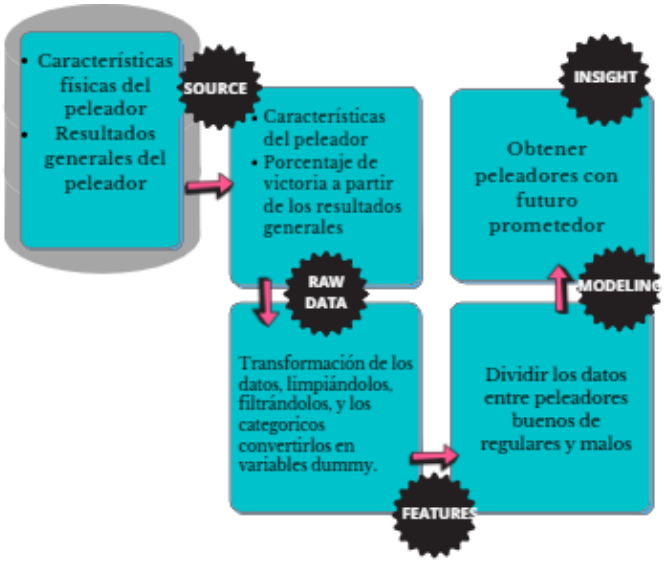

Fig. 1: Flujo de trabajo

A. *Entendimiento de los datos y del negocio*

En esta subsección se presentan dos de los pasos de nuestro flujo de trabajo, en que especifican determinar los datos e información relevante para nuestro modelo. En ellos no solamente se establece que obtengamos la información sino que también determinemos qué características son influyentes para arrojar un resultado apegado a nuestra solución.Por ello este artículo, se seleccionó un dataset de registros históricos de peleas específicamente de la UFC, para nuestro modelo propuesto. Considerando que el deporte de la MMA es un deporte de contacto, ya que trata de impactar partes del cuerpo en el contrincante. Con ello se manifiesta que no muchas personas poseen la habilidad o la moralidad para practicarlo, ya que es catalogado como un deporte temerario. A pesar de ello es un deporte que ha tenido un auge en estos tiempos, debido a la cantidad de dinero que maneja, pudiendo atraer a personajes talentosos. Esto se lo recalca debido a que se necesita de información sobre el rendimiento histórico como peleador y no como un deportista, siendo válido el dataset, ya que presenta información del desempeño que han tenido los peleadores profesionales.

Nuestro dataset, se recolecto en la página Kaggle, específicamente (https://www.kaggle.com/rajeevw/ufcdata), una comunidad que realiza participaciones públicas sobre aprendizaje automático, importando diferentes dataset de diferentes temáticas. El nombre del dataset escogido es “UFC-Fight historical data from 1993 to 2021”, proporcionando datos históricos de peleadores de la UFC, que se expresa en su documentación oficial que se obtuvo partir de valores estadísticos y resultados de las páginas oficiales de la UFC. Hay que especificar que el dataset consta de dos secciones: los detalles del peleador y resultados de peleas. El acceder a los datos directos de la UFC puede involucrar que tengamos una mayor precisión, ya que se posee varias características y se da la certeza de que los valores no sean erróneos o poco satisfactorios.

Los datos recopilados en la sección de resultados de peleas establecen 144 columnas que especifican cada uno de los aspectos técnicos establecidos al final de pelea. Al poseer tantas columnas debemos optar por tomar las características que nos ayuden en la metodología, ya que nuestro proyecto involucra características históricas y no de un resultado en particular. Las características tomadas son: El nombre del peleador, numero de victorias del peleador, numero de empates y numero de derrotas, acotando que se tomará los datos de las dos esquinas (roja y azul). En cambio, en la sección de detalles del peleador tomaremos las 14 columnas, que especifican un promedio de las habilidades obtenidas a través del resultado de múltiples peleas. Las características tomadas se describen en la Tabla 1, donde se describe el desempeño de un peleador a lo largo de su carrera profesional. Con ello podemos percibir que al ser un dataset de peleadores de artes marciales mixtas, estos tendrán tres capacidades desarrolladas, como lo son: la pelea de pie, especificados en STANCE, SLpM, Str. Acc; defensa contra golpes, proporcionados en SApM, Str. Def y derribos, determinados por TD Avg, TD Acc, TD Def, Sub. Avg. Hay que evidenciar que hay más características proporcionadas por el derribo y esto se debe a que para llevar a un oponente al piso primero se lo deberá derribar y luego aplicar alguna técnica de sumisión, por ello se lo toma como dos pasos, siendo en realidad una misma clasificación.

TABLA I: Características de detalles del peleador

Caracteristica	*Detalle*	*Dato*
HEIGHT	Altura del peleador	Object
WEIGHT	Peso del peleador	Object
REACH	Alcance de brazo de peleador	Object
STANCE	Tipo de parada del peledaor	Object
DOB	Fecha de nacimiento	Object
SLpM	Golpes significativos aterrizados por minuto	float64
Str. Acc	Precisión de golpe significativa	Object
SApM	Golpes significativos absorbidos por minuto	float64
Str. Def	Defensa contra golpes significativos	Object
TD Avg	Promedio de derribos aterrizados por 15 minutos	float64
TD Acc	Precisión de derribo	Object
TD Def	Defensa contra derribos	Object
Sub. Avg	Promedio de sumisión intentados por 15 minutos	float64

B. *Preparación de los datos y selección de características*

En esta subsección se va a detallar cada uno de los pasos que conllevo a la transformación de los datos, en información que pueda servir de entrada para un modelo de aprendizaje automático. Esto se realiza debido a que un modelo de aprendizaje automático como tal, no presenta predicciones acertadas de manera automática, ya que puede haber valores

incoherentes o atípicos, que atribuyen a un modelo defectuoso. Esto significa, que se debe realizar un preprocesamiento previo para exhibir las debilidades de nuestro dataset y proveer adecuaciones a las características, así generar una estructura más limpia. Una de sus debilidades más obvias es que el dataset como tal no posee la característica que permita resolver la problemática de nuestro modelo. Por ello se planea realizar un proceso para encontrar un porcentaje de victoria, de manera que nos proporcione la información que requiere nuestro modelo. El porcentaje de victoria se obtendrá a partir de las columnas de victoria, empate y derrota. Por consiguiente, vamos a dividir el número de victorias con la suma de las columnas. El realizar el proceso de conseguir el porcentaje de victoria puede conllevar a obtener datos repetidos, siendo poco beneficiosos para el modelo, ya que no existen datos anteriores de la referencia de detalles del peleador. Siendo necesario eliminar los datos repetidos mediante la agrupación de ellos y exhibiendo solamente el último registro que existe.

El dataset como tal es extraído de páginas oficiales de la UFC, pudiendo referenciar que contienen información del mundo real. Por lo tanto, es necesario verificar que el dataset no posea datos faltantes o con una estructura errónea, siendo evidente que puede haber errores del sistema o humanos en su proceso de recolección. Como primer paso, se observará los datos nulos que hay en cada una de las características del dataset. Con ello se pudo observar que hay ciertas características que contienen muchos datos nulos, siendo poco beneficiosas para el modelo, optando por eliminarlas y hacer desuso de ellas. Por otro lado, se encontró características con menos del 5 % en datos nulos, siendo el motivo de usar el concepto de CCA para eliminar los datos correspondientes. Así mismo se contempló datos nulos categóricos, siendo la opción para nuestro modelo establecer el concepto de moda para rellenar los datos nulos. Una vez que se resolvió la problemática de valores nulos, es necesario realizar una conversión de las características que sean categóricas, determinando un valor de existencia, para ello usaremos el concepto de valor dummy, así adjuntar columnas a nuestro dataset, pero resolviendo el problema de ser una característica categórica.

Un modelo de aprendizaje automático puede tener ciertas problemáticas con los datos, siendo necesario el realizar un escalado a los datos. Esto puede conllevar a que el modelo no disponga mayoritariamente de una variable, solo por poseer características numéricas significativas. El algoritmo que utilizaremos para nuestra problemática será el escalado robusto, ya que nuestros datos se manejan con valores atípicos, siendo una ventaja del algoritmo. Una vez que nuestros datos sean correctamente funcionales, se puede adherirlos a un modelo de aprendizaje automático, pero se presenta la problemática, de no saber si adherir todas las características o proporcionar sólo una proporción de ellos. Para ello vamos a utilizar una matriz que calcula cada uno de los coeficientes de correlación como se lo puede visualizar en la Figura 2, estableciendo cada una de las características que posean una alta correlación con respecto a nuestra variable de salida. Como otra opción se usará los datos antes establecidos en un modelo de aprendizaje automático, específicamente Random Forest, para que nos arroje las variables más utilizadas en el modelo. Como resultado obtendremos las características adecuadas para no establecer un modelo que utiliza muchos recursos.

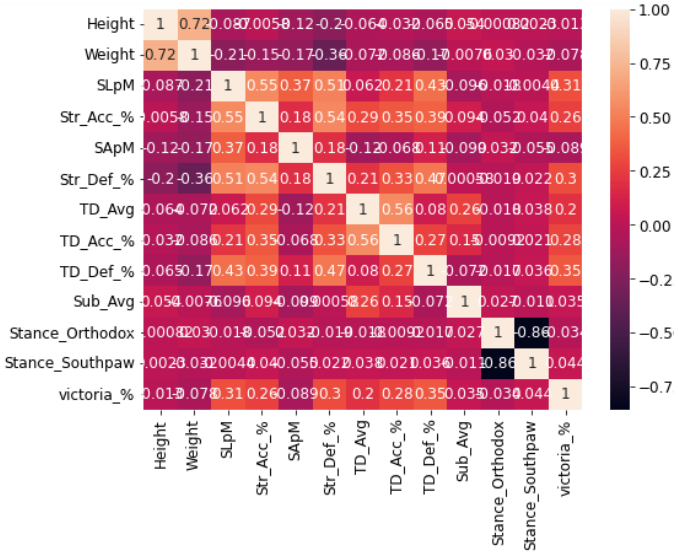

Fig. 2: Matriz de correlación

C. *Selección de modelos*

En esta subsección se analizará diferentes algoritmos de aprendizaje automático pudiendo comparar diferentes resultados así tener una mayor perspectiva, de manera que se pueda observar la factibilidad de usar Random Forest en predicciones deportivas. A través de ello, se establece una manera, que una empresa, ya sea, gubernamental o privada puede evaluar a un deportista, si tiene las características necesarias para acceder a grupo de apoyo deportivo. La predicción poco precisa puede conllevar pérdidas económicas, que no representan un beneficio, además de que ofrece una perspectiva a un peleador acerca del nivel que tiene actualmente. Por lo que, a través de nuestro modelo, se dicta a las empresas una predicción de posibles candidatos a un beneficio económico.

En esta investigación, se establece y compara diferentes modelos de aprendizaje automático, así poder seleccionar la perspectiva más efectiva, de tal manera que proporcione la mayor precisión al momento de predecir la victoria de un peleador o un posible beneficiario. En este documento se plantea utilizar diferentes modelos de aprendizaje automático, lo cuál incluye árboles de decisión, Random Forest y máquinas de vectores de soporte (SVM).

Para entender Random Forest debemos entender sus inicios, por lo que es factible utilizar un árbol de decisión ya que es una estructura con decisiones en cascada, de manera que permitirá segmentar los datos según una determinada condición. Por tanto, al entrar un dato en un árbol de decisión, este será dirigido de la hoja padre, que se le menciona como nodo inicio, a cada una de las diferentes etapas o alturas que tiene el árbol, siendo la última hoja que se posicione el dato, la decisión final. Todo este proceso trata de subdividir los datos en diferentes regiones con una única clase [19]. Siendo resaltable que para que un árbol de decisión haga un trabajo correcto tiene que subdividir los datos de manera que no exista

un direccionamiento directo a una clase, con ello decimos que los datos deben ser distribuidos por las diferentes condiciones según la ganancia de información.

Random Forest es un algoritmo que permite la utilización de varios árboles de decisión que trabajan en conjunto para llegar a una decisión lo cuál podemos observar su funcionamiento en la Figura 3. La elección de qué estructura tendrá cada uno de los árboles creados, se determinará a partir de una separación aleatoria de datos. Cada separación será usada para crear un árbol de decisión, disponiendo de estructuras diferentes. Al realizar el proceso de creación de las estructuras, todos los datos se usarán como entrada para todos los árboles, de manera que sigan el proceso normal de un árbol de decisión, que involucra disponer de una clase, mediante la subdivisión de los datos. Una vez obtenida la respuesta de cada árbol, se contempla la clase con mayor incidencia, de manera que sea la respuesta definitiva del modelo [20]. Para concluir pudimos contemplar que gracias a que todos los árboles de decisión en el modelo presentan datos faltantes, el modelo puede hacer caso omiso de datos faltantes en la entrada general, por tanto, ayudarnos a mejorar nuestra predicción. Todo lo mencionado se puede observar a través del pseudocódigo que detalla el funcionamiento de Random Forest, que en el caso del proyecto será nuestro algoritmo principal.

Algorithm 1 Algoritmo Random Forest

```
procedure RANDON FOREST(data,number_stimators)
    repeat
        for numero de estimadores do
            Tomar características aleatorias
            Construye un arbol de decisión
            Arboles.add(Arbol de decisión)
        end for
for each: arbol in Arboles
        Almacenar respuesta
        return respuesta con más apariciones
```

Por último como comparación se va a utilizar a SVM, que se lo especifica como un algoritmo que permite segmentar los datos de manera lineal mediante un vector de soporte. Para segmentar los datos se calcula el mayor margen posible entre los datos, de manera que pueda haber una segmentación correcta, a esto se le llama el vector de soporte. Hay ocasiones en las que es necesario tener en cuenta que el algoritmo no presenta una segmentación idónea, ya que muchas veces una separación lineal no es la mejor opción. Por lo tanto, en esos casos, lo obligatorio es aplicar un kernel diferente, estableciendo los valores de manera que se cree una nueva dimensión, así después regresarlos al espacio de dos dimensiones [21]. Para terminar, podemos determinar que SVM es un clasificador robusto, ya que posee la capacidad de trabajar con gran cantidad de variables, que mediante sus diferentes kernels, que agregan nuevas dimensiones, permite encontrar soluciones óptimas.

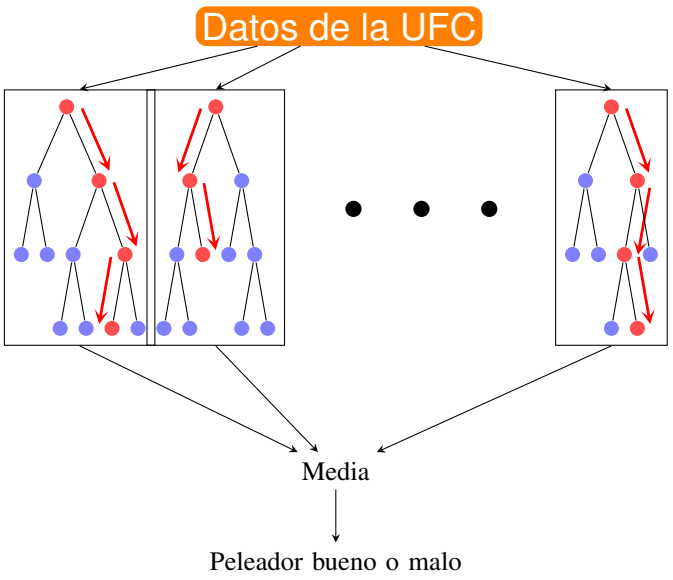

Fig. 3: Algoritmo de Random Forest

D. *Evaluación de modelos*

En esta subsección se explica como existen diversos criterios que pueden ayudar a determinar la calidad de un modelo de aprendizaje automático. El criterio a utilizar para especificar la importancia de una métrica será el objetivo de un proyecto, ya que no todas las veces se establecerá un modelo totalmente perfecto. Por ello es necesario aplicar diversas métricas y entender cómo se encuentra estructurado el modelo. En el caso de nuestro modelo, que es un algoritmo de clasificación se utilizará la matriz de confusión para determinar varias métricas. Los valores que se muestran en una matriz de confusión son: TP (true positive), que indica un valor positivo predicho de manera correcta; TN (true negative), que establece un valor negativo predicho de manera correcta; FP (false positive), que indica un valor positivo predicho de manera incorrecta; FN (false negative) que determina un valor negativo predicho de manera incorrecta. Con ello se establecen diferentes fórmulas para resumir de manera conceptual y como se encuentra construido un modelo, esto se lo puede observar en las ecuaciones (1)(2)(3)(4)(5).Con ellas podemos determinar la calidad del modelo y como maneja los diferentes parámetros. Se puede por ejemplo en especificidad, observar la proporción de valores negativos predichos correctamente; sensibilidad, determinar la proporción de valores positivos predichos correctamente; precision, establecer la proporción de predicciones positivas correctas, negative predictive value, la proporción de predicciones negativas correctas.

$$\textbf{Sensivity: } \frac{TP}{TP+FN} \tag{1}$$

$$\textbf{Specificity: } \frac{TN}{TN+FP} \tag{2}$$

$$\textbf{Precision: } \frac{TP}{TP+FP} \tag{3}$$

$$\textbf{Negative Predictive Value: } \frac{TN}{TN+FN} \tag{4}$$

$$\textbf{Accuracy: } \frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

IV. Experimentos y análisis

A. segmentación de los datos

Un modelo de aprendizaje automático requiere de valores de entrada y valores de prueba que nos ayude a establecer nuestros criterios de evaluación. Es por ello que es necesario establecer conceptos que nos ayude a establecer esos valores de prueba, de manera que, mediante los mismos datos, podamos establecer los datos de prueba. Uno de los conceptos que se utiliza, es el Train/Test, que consiste en establecer un porcentaje al azar de los datos a los datos de prueba. Otro concepto aplicado es el algoritmo K-fold, que nos indica que nuestros datos van a iterar a través de un split, en el cuál se va a hacer uso de todos los datos sin dejar alguno. Esto puede conllevar a que haya una mejora en el modelo, ya que usa todos los datos y toma toda la información posible [22].

B. Elección de parametros mediante GridSearch

Los diferentes conceptos antes manejados pueden conllevar a una mejora del modelo de aprendizaje automático, pero deberá ser el modelo quien deba manejar todos estos datos y establecer una predicción que se anteponga de conceptos como el underfitting y overfitting. Es por ello que es necesario el usar diferentes parámetros, ya que todos los algoritmos de aprendizaje automático, manejan una base matemática. Estas bases pueden conllevar a manejar diferentes casos en los datos, ya sea manejando información oculta o discernir de usar ciertas características. Para esto se utilizo Grid Search, así iterar entre diferentes parámetros de cada modelo, así determinar los mejores valores para ellos [23].

En el modelo de SVM se usará C, gamma, kernel; estableciendo que C, especificará la penalidad de error; gamma, determina la importancia de los puntos lejanos y cercanos; kernel, especificará el tipo de Kernel para disponer de un plano tridimensional. En el modelo decision tree se utilizará criterion, determinando la ganancia de información; max_depth, la altura del árbol. Para el modelo de random forest se usará criterion y max_depth, establecidos anteriormente en el árbol de decisión, adicionalmente se especificará max_features, estableciendo el número máximo de características para llegar a una solución; n_estimators, que nos permite establecer el número de sub-arboles. Mediante esta selección se pudo determinar los siguientes parámetros:

- Random Forest: 'criterion': 'entropy', 'max_depth': 7, 'max_features': 'auto', 'n_estimators': 200
- SVM: 'C': 10, 'gamma': 'scale', 'kernel': 'rbf'
- Decision tree: 'criterion': 'entropy', 'max_depth': 8

C. Escalado de los datos

Hay que establecer que al tratarse de datos que se presentan en la vida real, pueden contener valores muy variados. Esto puede suponer un problema, ya que puede el modelo considerar con una mayor disposición a los datos con una cifra grande con respecto a cifras pequeñas.Esto supone realizar un escalado de los datos. Para resolver esta problemática se procedió a usar RobustScaler, por el motivo de que es un algoritmo que no es influenciado por valores atípicos. Esto se puede observar en la Figura 4, donde se muestran nuestros datos graficados mediante una distribución de frecuencias. Con ello se logró un resultado, en el que nuestros datos seguían una distribución normal, sin tener una gran influencia de valores atípicos, a pesar de ello todavía existen pero en una menor cantidad.

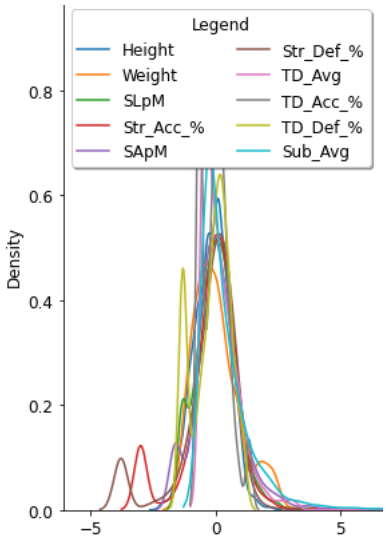

Fig. 4: Escalado robusto

V. Resultados

El determinar la calidad de un modelo de aprendizaje automático puede conllevar utilizar una matriz de confusión, ya que nuestro modelo se establece como un algoritmo de clasificación. Una vez establecido, la mejor manera de entender nuestro modelo es utilizar las métricas asociadas a la matriz de confusión. Obtenidos las diferentes métricas, podemos darnos cuenta que los mejores modelos de nuestra investigación son Random Forest y SVM con una exactitud del 75,55 % y 74,76 % respectivamente, por ello se opto por hacer uso de las diferentes métricas, así dar contexto a las tendencias de predicción de nuestro modelo, como se puede observar en la Tabla II.

Por último se graficó cada uno de los algoritmos mediante las métricas principales, por el cuál se puede observar en la Figura 5, cómo es superior Random Forest, a los demás algoritmos. Es evidente que al visualizar se puede decir que no hay una diferencia notoria con respecto a SVM, siendo casi similar al medir sus métricas. Por lo que podemos decir que también es una opción válida para predicciones deportivas.

VI. Discusión

Con las diferentes métricas, se pudo evidenciar que nuestro modelo tiene una alta precisión con respecto a predecir un valor negativo, por lo que se considera, que es más importante para el modelo no obtener valores positivos falsos, que

TABLA II: Métricas de los modelos utilizados

Metrica \ Modelo	Decision tree	Random Forest	SVM
Train Accuracy	80,91 %	82,19 %	77,15 %
Test Accuracy	69,28 %	75,55 %	74,76 %
Precision	52,28 %	63,93 %	62,11 %
Recall	60,87 %	56,52 %	57,00 %
Negative predictive value	79,69 %	80,22 %	80,13 %
Specificity	73,32 %	84,69 %	83,29 %

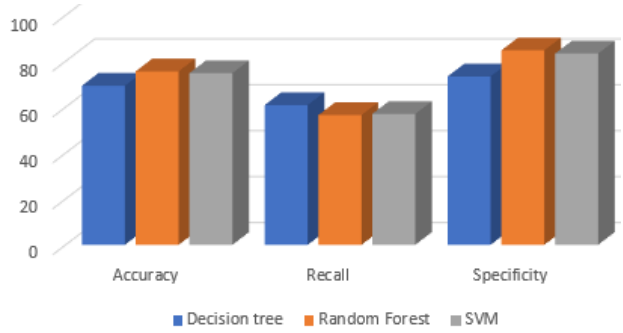

Fig. 5: Gráficas de métricas principales

predecir de manera precisa un valor positivo. Esto se lo puede deducir en que haya menos valores positivos verdaderos. La calidad del modelo se puede observar en la Figura 6, que se puede observar que en ningún momento choca con la recta, que indica una predicción aleatoria. Con ella podemos reafirmar que nuestro modelo no actúa de manera agresiva, de manera que no perdamos una gran cantidad de positivos verdaderos.

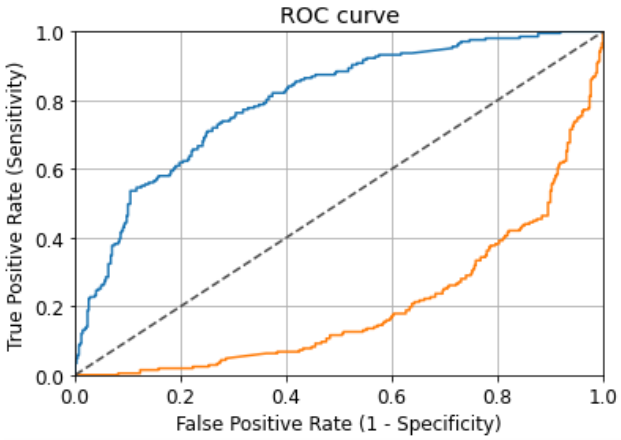

Fig. 6: Curva ROC

VII. Conclusiones

En este estudio, se utilizaron diferentes algoritmos de aprendizaje automático, de manera que se logró el objetivo de predecir si un deportista puede acceder a un beneficio económico deportivo. Esto se realizó utilizando los registros históricos que poseen de manera individual cada deportista, ya que con esto podremos especificar cada una de sus habilidades obtenidas y cuales son las más importantes para ser un próximo deportista destacado. La clasificación se la realizó utilizando tres algoritmos basados en clasificación, Random forest, SVM y decision tree. Con los resultados se pudo establecer que nuestro modelo se puede aplicar al campo real, ya que se presenta como un algoritmo capaz de comprender cuando un peleador no posee las habilidades correctas, siendo beneficioso para las organizaciones gubernamentales o privadas que deseen respaldar a un deportista en concreto. La exactitud del modelo se estableció con 75,55 % de parte de Random Forest, acotando que obtuvo una gran predicción con respecto a la evaluación de un deportista que no se presenta como posible beneficiario. El trabajo futuro consistirá en poder recolectar más información a través de las diferentes empresas que se emplean en este medio, así obtener una mayor cantidad de datos, de manera que podamos indicar el resultado de una pelea en concreto, siendo de referencia para mejorar la calidad de entrenamiento.

Referencias

[1] P. Sri Harsha Vardhan Goud, Y. Mohana Roopa, and B. Padmaja, *Player Performance Analysis in Sports: with Fusion of Machine Learning and Wearable Technology*.

[2] *Plan de alto rendimiento 2022*. Ministerio del deporte, 2021. [Online]. Available: https://www.deporte.gob.ec/wp-content/uploads/2022/01/MINDEP_PlanAltoRendimiento_2022.pdf

[3] ""TUF 28 Aplicación ufc"." [Online]. Available: https://www.ufcespanol.com/news/tuf-28-application

[4] E. M. Torralba, "Sports ed 3.5: Establishing the value of data-driven sports development programs for universities through machine learning models," in *PervasiveHealth: Pervasive Computing Technologies for Healthcare*, vol. PartF168341. Association for Computing Machinery, Dec 2020, p. 51–57.

[5] F. Zhang, Q. Jiang, and B. Zhou, "Prediction and analysis of college students' sports achievements based on support vector machine and particle swarm optimization," in *PervasiveHealth: Pervasive Computing Technologies for Healthcare*. ICST, Jun 2019, p. 128–132.

[6] H. T. El-Kassabi, K. Khalil, and M. A. Serhani, "Deep learning approach for forecasting athletes' performance in sports tournaments," in *PervasiveHealth: Pervasive Computing Technologies for Healthcare*. ICST, Sep 2020, p. 203–208.

[7] Z. Fan, "Modeling of sports performance based on nonlinear screening factors and weighting to improve prediction accuracy," in *ACM International Conference Proceeding Series*. Association for Computing Machinery, Oct 2018.

[8] Y. Cho, J. Yoon, and S. Lee, "Using social network analysis and gradient boosting to develop a soccer win–lose prediction model," *Engineering Applications of Artificial Intelligence*, vol. 72, p. 228–240, Jun 2018.

[9] S. Kovalchik and M. Reid, "A calibration method with dynamic updates for within-match forecasting of wins in tennis," *International Journal of Forecasting*, vol. 35, no. 2, p. 756–766, Apr 2019.

[10] D. Delen, D. Cogdell, and N. Kasap, "A comparative analysis of data mining methods in predicting ncaa bowl outcomes," *International Journal of Forecasting*, vol. 28, no. 2, p. 543–552, Apr 2012.

[11] G. Kyriakides, K. Talattinis, and S. George, "Rating systems vs machine learning on the context of sports," in *ACM International Conference Proceeding Series*, vol. 02-04-October-2014. Association for Computing Machinery, Oct 2014.

[12] C. K. Leung and K. W. Joseph, "Sports data mining: Predicting results for the college football games," in *Procedia Computer Science*, vol. 35, no. C. Elsevier B.V., 2014, p. 710–719.

[13] B. Zhao and L. Chen, "Prediction model of sports results base on knowledge discovery in data-base," in *Proceedings - 2016 International Conference on Smart Grid and Electrical Automation, ICSGEA 2016*. Institute of Electrical and Electronics Engineers Inc., Nov 2016, p. 288–291.

[14] A. Balasundaram, D. Jayashree, S. Ashokkumar, and S. Magesh Kumar, *Data mining based Classification of Players in Game of Cricket*.

[15] I. Fister, D. Fister, and S. Fong, "Data mining in sporting activities created by sports trackers," in *Proceedings - 2013 International Symposium on Computational and Business Intelligence, ISCBI 2013*. IEEE Computer Society, 2013, p. 88–91.

[16] Q. Zhang, X. Zhang, H. Hu, C. Li, Y. Lin, and R. Ma, "Sports match prediction model for training and exercise using attention-based lstm network," *Digital Communications and Networks*, 2021.

[17] J. Weissbock and D. Inkpen, *Combining Textual Pre-game Reports and Statistical Data for Predicting Success in the National Hockey League*. [Online]. Available: http://puckprediction.

[18] A. Konstantinos and C. Tjortjis, *Sports Analytics algorithms for performance prediction*.

[19] M. Ahmad, V. Tundjungsari, D. Widianti, P. Amalia, and U. A. Rachmawati, "Diagnostic decision support system of chronic kidney disease using support vector machine," *2017 Second International Conference on Informatics and Computing (ICIC)*, 2017.

[20] F.-J. Yang, "An extended idea about decision trees," *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, 2019.

[21] A. T. Prihatno, H. Nurcahyanto, and Y. M. Jang, "Predictive maintenance of relative humidity using random forest method," *2021 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, 2021.

[22] J. C. Prieto, A. Fernandez-Isabel, and F. Ortega, *A Supervised Learning Approach to Detect Copyright Infrigments*. IEEE, 2018.

[23] R. P. Bunker and F. Thabtah, "A machine learning framework for sport result prediction," *Applied Computing and Informatics*, vol. 15, no. 1, p. 27–33, Jan 2019.