Exploración y Análisis Multivariado para el Modelado Predictivo en los datos Ultimate UFC

Ing. Nelson Ariza Morales, Lic. Juan Javier Monsivais Borjón
Centro de Investigación en Matemáticas. Unidad Monterrey
Email: nelson.ariza@cimat.mx, juan.monsivais@cimat.mx

***Abstract*— Este reporte detalla un análisis profundo de los datos de combates de la UFC, con el objetivo de predecir los ganadores de los enfrentamientos. Inicialmente, se aplicó el Escalamiento Multidimensional (MDS) para identificar posibles agrupaciones dentro de los datos. Aunque las primeras iteraciones no revelaron grupos distintos, ajustes siguientes permitieron obtener configuraciones que ofrecieron perspectivas útiles. Posteriormente, se implementaron modelos predictivos utilizando Análisis Discriminante Lineal (LDA), abordando retos como la colinealidad mediante la eliminación de variables altamente correlacionadas. Este enfoque no solo enriquece la comprensión del deporte, sino que también suministra herramientas predictivas esenciales para la formulación de estrategias competitivas y decisiones financieras como apuestas.**

I. Introducción

Ultimate UFC Dataset se encuentra disponible en Kaggle, representa una colección de datos sobre los combates en la Ultimate Fighting Championship (UFC). Este conjunto de datos se caracteriza por su riqueza en detalles, incluyendo perfiles exhaustivos de los peleadores, estadísticas detalladas de cada combate y cuotas de apuestas, lo que lo convierte en un recurso óptimo para el análisis y la modelización estadística multivariada.

La finalidad de este estudio es aplicar técnicas de estadística multivariada para predecir el ganador de los combates, utilizando como base los atributos únicos de los peleadores y las características específicas de cada enfrentamiento. Este análisis no solo apunta a mejorar las estrategias en entrenamiento y competición, sino también a proporcionar datos estadísticos a los seguidores y analistas del deporte.

En el marco de este análisis, iniciaremos utilizando la matriz de distancia de Gower para evaluar similitudes entre peleadores, seleccionando variables que se consideran críticas por su impacto potencial en los resultados de los combates. A continuación, se implementará el Escalamiento Multidimensional (MDS) para visualizar las agrupaciones y detectar patrones subyacentes que podrían no ser evidentes a primera vista.

Este enfoque preliminar es crucial para identificar grupos homogéneos dentro de los datos, lo que permitirá afinar los modelos predictivos que se desarrollarán más adelante. A partir de las agrupaciones identificadas mediante MDS, se llevará a cabo un análisis predictivo. El objetivo es mejorar la precisión de las predicciones y, por ende, el rendimiento de los modelos predictivos utilizados.

La integración de estos métodos no solo tiene la intención de ofrecer un enfoque más robusto para la predicción de resultados de combates, sino también de proporcionar un entendimiento valioso para el desarrollo de estrategias de entrenamiento avanzadas. Asimismo, estos análisis buscan enriquecer la narrativa y la comprensión analítica de los

entrenadores, peleadores, y aficionados, ofreciendo una perspectiva más amplia y fundamentada sobre las dinámicas complejas del combate en la UFC.

II. Objetivos

1. Realizar un análisis para entender las características y rendimientos de los peleadores, para aplicar MDS e identificar agrupaciones en los datos.
2. Desarrollar un modelo predictivo para pronosticar los ganadores de los combates, abordando el problema de la colinealidad entre variables.
3. Evaluar la precisión del modelo predictivo y discutir su potencial aplicación en estrategias de entrenamiento y análisis para mejorar las decisiones competitivas en la UFC.

III. Metodología

En esta sección, presentaremos los modelos teóricos que se usarán en este estudio, seguidos de su potencial para este tipo de análisis de los datos de la UFC. Cada método ha sido seleccionado para abordar diferentes aspectos de los datos y maximizar la precisión y la utilidad de los resultados obtenidos.

III-A. Análisis de factores

El análisis factorial es una técnica estadística multivariada utilizada para explicar la variabilidad observada en un conjunto de variables a través de un número menor de variables latentes no observadas, conocidas como factores. Estos factores representan la estructura subyacente en los datos y ayudan a reducir la dimensionalidad mientras se conserva la mayor cantidad de información posible.

El modelo básico de análisis factorial puede expresarse matemáticamente de la siguiente manera:

$$X_i = \mu_i + \lambda_{i1}F_1 + \lambda_{i2}F_2 + \cdots + \lambda_{im}F_m + \epsilon_i$$

Donde:

- X_i es la variable observada.
- μ_i es la media de la variable.
- λ_{ij} es la carga factorial, que representa la relación entre la variable observada X_i y el factor F_j.
- F_j es el factor común.
- ϵ_i es el error específico o factor específico asociado con la variable X_i.

En notación matricial, el modelo se puede representar como:

$$\mathbf{X} = \mu + \mathbf{\Lambda F} + \mathbf{E}$$

Donde:

- $\mathbf{X}$ es el vector de variables observadas.
- μ es el vector de medias.
- $\mathbf{\Lambda}$ es la matriz de cargas factoriales.
- $\mathbf{F}$ es el vector de factores comunes.
- $\mathbf{E}$ es el vector de errores específicos.

La varianza total de una variable observada X_i se puede descomponer en:

- **Varianza común (h_i^2)**: Proporción de la varianza que se explica por los factores comunes.
- **Varianza específica (ψ_i)**: Proporción de la varianza que es única para esa variable.

$$Var(X_i) = h_i^2 + \psi_i$$

Procedimiento del Análisis Factorial

1. **Recolección de Datos**: Obtener una matriz de covarianzas o correlaciones a partir de los datos observados.
2. **Extracción de Factores**: Determinar el número de factores y calcular las cargas factoriales iniciales. Los métodos comunes para la extracción de factores incluyen el método de componentes principales y el método de máxima verosimilitud.
3. **Rotación de Factores**: Rotar los factores para fa-

cilitar su interpretación. Las rotaciones pueden ser ortogonales (p. ej., Varimax) o oblicuas.

4. **Cálculo de Puntuaciones Factoriales**: Calcular las puntuaciones factoriales para cada observación, que pueden utilizarse para análisis adicionales.

Método del Factor Principal

El método del factor principal es una modificación del enfoque de componentes principales, enfocándose en identificar las variables latentes que contribuyen a la varianza común de las variables medidas. La matriz de correlaciones reducida $\mathbf{R_r}$ se puede factorizar aproximadamente como:

$$\mathbf{R_r} \approx \mathbf{L}^* \mathbf{L}^{*\prime}$$

Donde $\mathbf{L}^*$ son las cargas estimadas. Las comunalidades se estiman iterativamente y se ajustan hasta que la solución converge.

Rotación de Factores

La rotación de factores se realiza para obtener una estructura más interpretable. Los métodos de rotación comunes incluyen:

- **Varimax**: Maximiza la varianza de las cargas factoriales cuadradas.
- **Quartimax**: Minimiza el número de factores necesarios para explicar cada variable.
- **Equimax**: Combina criterios de Varimax y Quartimax.

Aplicación en el Análisis de UFC

En el contexto del análisis de datos de la UFC, el análisis factorial puede ayudar a identificar las variables latentes que influyen en los resultados de los combates. Al reducir la dimensionalidad de los datos, es posible obtener una comprensión más clara de los factores clave que determinan el desempeño de los peleadores y predecir los ganadores con mayor precisión.

*III-B. **Distancia de Gower***

La distancia de Gower se basa en su capacidad para manejar datos heterogéneos, adaptando la medida de similitud según el tipo de cada variable. Esta flexibilidad es crucial para análisis que involucran tanto datos numéricos como categóricos, como es común en conjuntos de datos extensos y variados de disciplinas como la Ultimate Fighting Championship (UFC).

Fórmula General

La distancia de Gower entre dos entidades i y j se calcula mediante la siguiente fórmula:

$$S(i,j) = \frac{\sum_{k=1}^{p} w_{ijk} s_{ijk}}{\sum_{k=1}^{p} w_{ijk}}$$

donde:

- p es el número total de variables.
- s_{ijk} representa la puntuación de similitud entre las entidades i y j para la variable k.
- w_{ijk} es el peso asignado a la similitud s_{ijk}, que puede ajustarse para manejar la importancia relativa de la variable o la presencia de datos faltantes.

Cálculo de Similitudes

- **Variables Numéricas:** La similitud se mide como $s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{R_k}$, donde R_k es el rango de la variable k, es decir, la diferencia entre su máximo y mínimo observados.
- **Variables Categóricas:** La similitud s_{ijk} se asigna un valor de 1 si $x_{ik} = x_{jk}$ y 0 en caso contrario.

En el análisis de datos de UFC, donde los datos abarcan desde resultados de combates hasta estadísticas detalladas de los peleadores, es decir, tanto datos numéricos como datos categóricos, la distancia de Gower permite una comparación integral de los competidores. Esta capacidad es esencial para desarrollar modelos predictivos que reflejen con precisión las capacidades y características de los peleadores.

Para el uso de este método, tenemos la opción de la librería de clúster en Python, que proporciona una implementación conveniente de la distancia de Gower a través de la librería ***Gower***.

III-C. ***Método de Escalamiento Multidimensional (MDS)***

Es una técnica de análisis multivariante utilizada para representar las proximidades entre un conjunto de objetos como distancias en un espacio de baja dimensión.
Las proximidades pueden ser de diversos tipos, tales como medidas de similitud o disimilitud, distancias, correlaciones, entre otras. En este estudio, utilizamos la distancia de Gower para calcular las disimilitudes entre los registros de combates en la base de datos de la UFC. La matriz de proximidades resultante se utilizó como entrada para el algoritmo MDS.

Modelo Clásico de MDS

El teorema del MDS clásico establece que si una matriz de distancias es euclidiana, se puede encontrar una configuración en un espacio de dimensión K tal que las distancias entre los puntos en este espacio correspondan a las distancias originales. Este modelo se basa en la descomposición espectral de matrices y utiliza los valores y vectores propios para encontrar la configuración óptima. El procedimiento para obtener las coordenadas principales a partir de una matriz de disimilaridades (distancias al cuadrado) es el siguiente:

1. Se construye la matriz $B = -\frac{1}{2}H\Delta H$, donde $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$.
2. Se obtienen los valores propios de B y se seleccionan los m valores propios positivos más grandes.
3. Las coordenadas principales de los puntos están dadas por los vectores propios asociados a estos valores propios.

El uso del MDS, complementado con la distancia de Gower, permite representar y analizar de manera efectiva las relaciones entre los datos de la UFC, facilitando la identificación de patrones y la predicción de resultados en combates. Este enfoque no solo enriquece el análisis deportivo, sino que también proporciona herramientas valiosas para la toma de decisiones estratégicas.

Minimos cuadrados

En el contexto del Modelo Clásico de MDS por mínimos cuadrados, el enfoque se centra en ajustar las distancias entre puntos en un espacio de baja dimensión a las disimilaridades observadas entre objetos. Aquí se utiliza un método de optimización que intenta minimizar la suma de los cuadrados de las diferencias entre las distancias teóricas y las representadas en el modelo. Este proceso se conoce como STRESS, el cual es una medida de qué tan bien el modelo de MDS se ajusta a las disimilaridades observadas.

III-D. ***Análisis de Discriminante Lineal (LDA)***

Una técnica estadística utilizada para la clasificación y reducción de dimensionalidad. En el contexto de este estudio, LDA se aplicó para predecir el ganador de combates de la UFC, considerando múltiples características de los peleadores y detalles específicos de cada encuentro.

LDA asume que las distintas clases (en este caso, los ganadores de las peleas) se distribuyen de manera normal multivariada y que todas las clases comparten una matriz de covarianzas común. Este método busca proyectar los datos en una nueva línea que maximiza la separación entre las clases mientras minimiza la varianza dentro de cada clase.

Definición LDA

Dado un vector de características $\mathbf{X} = (X_1, X_2, \ldots, X_p)'$, LDA se basa en las siguientes suposiciones:

- Las clases π_1 y π_2 tienen distribuciones normales multivariadas con medias μ_1 y μ_2 y una matriz de covarianzas común Σ.
- La función discriminante de Fisher para separar las dos clases se define como $Y = \mathbf{a}'\mathbf{X}$, donde el vector de pesos $\mathbf{a}$ maximiza la separación entre las medias de las clases, dado por $\mathbf{a} = \Sigma^{-1}(\mu_1 - \mu_2)$.

Cálculo

En la práctica, los parámetros poblacionales μ_1, μ_2 y Σ son desconocidos y deben ser estimados a partir de los datos de entrenamiento. Las medias de las clases se estiman como:

$$\hat{\mu}_1 = \frac{1}{n_1} \sum_{i \in \pi_1} \mathbf{x}_i, \quad \hat{\mu}_2 = \frac{1}{n_2} \sum_{i \in \pi_2} \mathbf{x}_i$$

Y la matriz de covarianzas común se estima como:

$$\hat{\Sigma} = \frac{1}{n_1 + n_2 - 2}$$

$$\left[\sum_{i \in \pi_1} (\mathbf{x}_i - \hat{\mu}_1)(\mathbf{x}_i - \hat{\mu}_1)' + \sum_{i \in \pi_2} (\mathbf{x}_i - \hat{\mu}_2)(\mathbf{x}_i - \hat{\mu}_2)' \right]$$

Clasificación

La regla de clasificación asigna una nueva observación $\mathbf{x}$ a la clase cuya función discriminante lineal sea mayor:

$$\delta_k(\mathbf{x}) = \mathbf{x}'\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k'\Sigma^{-1}\mu_k + \log(\pi_k)$$

Aplicación en Predicción de Combates de UFC

En nuestro estudio, se utilizó LDA para construir un modelo predictivo que determine el ganador de los combates de la UFC. Se implementaron pasos para tratar problemas de colinealidad y se evaluó la eficacia del modelo mediante técnicas de validación cruzada y matrices de confusión.

La implementación de LDA en este análisis se llevó a cabo utilizando la función lda del paquete MASS en R. Los pasos incluyeron la preparación de los datos, el ajuste del modelo y la evaluación de su rendimiento. Este proceso permitió construir un modelo capaz de clasificar y predecir el ganador de los combates de la UFC con una precisión significativa, demostrando la aplicabilidad del LDA en contextos deportivos y de análisis de datos complejos.

IV. Implementación y resultados

Los métodos descritos anteriormente fueron implementados con el fin de estudio de la base de datos 'Ultimate UFC Dataset' alojado en Kaggle y compilado por Matt Dabbert, contiene una extensa colección de registros de combates de la UFC.

Incluye datos de:

ufcstats.com: Estadísticas de peleas y peleadores.

bestfightodds.com: Cuotas de apuestas.

kaggle.com/martj42/ufc-rankings: Rankings de la UFC.

Variables Principales

El conjunto de datos abarca una amplia gama de variables, entre las cuales destacan:

1. Datos del Peleador:
   Físicas: Nombre, altura, peso, alcance, edad, estilo de combate.
2. Estadísticas de la Pelea:
   Resultados: Número de victorias, derrotas, empates, racha de victorias/derrotas.
   Acciones: Golpes significativos aterrizados, porcentaje de golpes significativos, intentos de sumisión, derribos.
3. Datos del Encuentro:
   Información de la pelea: Ubicación, clase de peso, si es pelea por el título, número de rondas.
   Detalles del resultado: Método de finalización de la pelea, detalles de finalización, tiempo total de pelea.

El propósito principal de este conjunto de datos es proporcionar una base para análisis y el desarrollo de

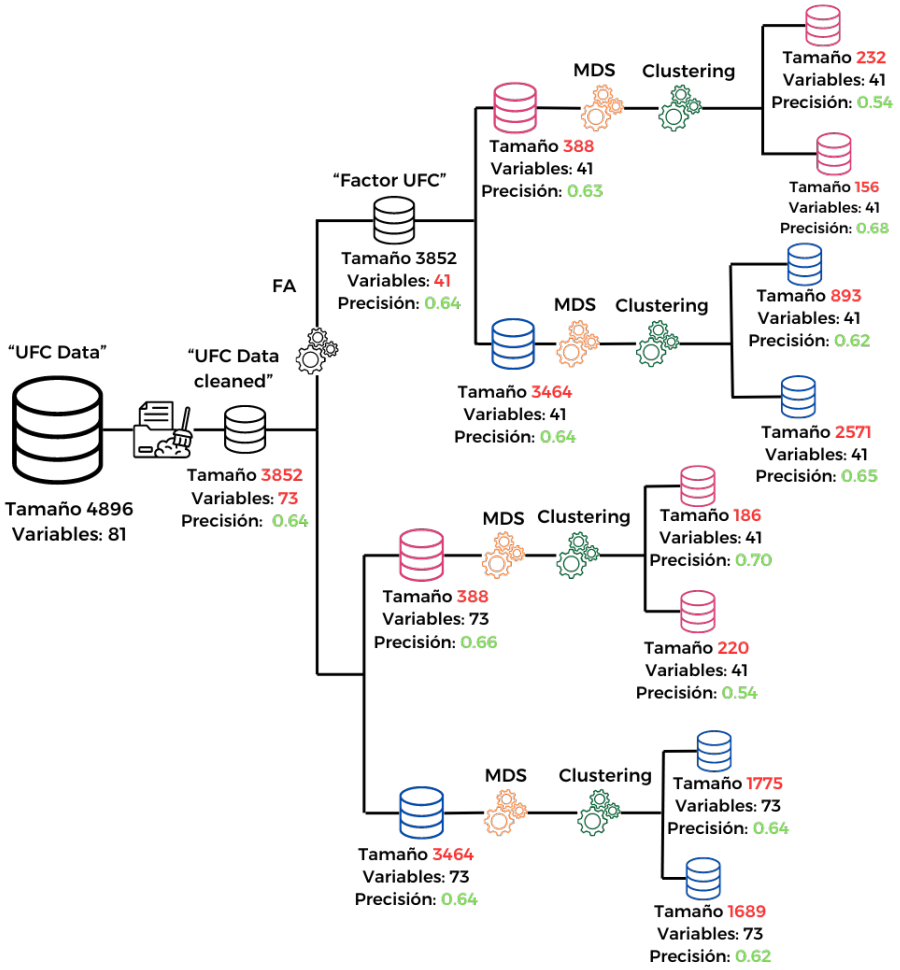

Fig. 1
FLUJO DE TRABAJO SOBRE LA BASE DE DATOS

modelos predictivos en el ámbito de las peleas de la UFC. Además, se busca fomentar la creación de modelos que puedan predecir resultados de peleas, identificar patrones y tendencias en los datos, y mejorar las estrategias de entrenamiento y competencia en el deporte.

Flujo de trabajo

En la figura 1 se muestra el flujo de trabajo sobre la base de datos, haciendo uso de las metodologías mencionadas anteriormente.

Entonces la estrategia es, a apartir de la base de datos con tamaño de 4896, limpiarla, reduciendo la base de datos un 20 %, eliminando las entradas nulas. A partir de aquí obtenemos otra base de datos aplicando análisis de factores, dejando solo aquellos que nos explique el 80 % de la varianza, reduciendo ahora el número de variables continuas en un 50 %, posterior a ello cada una de las 2 bases de datos (la original limpia y la reducida por factores) son subdivididas en de manera arbitraria en peleas de hombres y mujeres. Aa continuación aplicamos reducción multidimensional haciendo uso de las distancias gower, a estas reducciones tridimensionales aplicamos clusterización, en busca de nuevos subgrupos, finalmente a estos

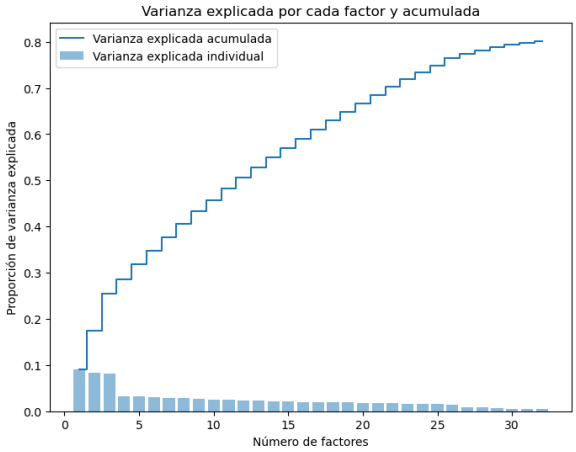

Fig. 2
VARIANZA ACUMULADA E INDIVIDUAL DE LOS FACTORES

subgrupos aplicamos LDA y comparamos el desempeño con nuestro baseline.

Resultados

Análisis de Factores

En este estudio, se aplicó el Análisis de Factores (FA) a los datos de la UFC para reducir la dimensionalidad del conjunto de datos y enfocarse en las características más relevantes. El gráfico de la figura 2 ilustra la varianza explicada por cada factor individualmente y la varianza acumulada a medida que se agregan más factores.

La selección de los primeros 32 factores se justifica porque:

1. Reducción de Dimensionalidad: Permite simplificar el modelo y hacerlo más manejable computacionalmente, sin perder una cantidad significativa de información.
2. Captura de Varianza Significativa: Con estos factores se captura una porción considerable de la variabilidad en los datos, asegurando que las características más importantes estén incluidas en los análisis posteriores.

Con la reducción de los factores, hemos definido cuatro enfoques clave para nuestro estudio:

1. Análisis basado exclusivamente en la base de datos de peleas de hombres.
2. Análisis de datos enfocado en la selección de las peleas de mujeres.
3. Análisis utilizando una base de datos con los 32 factores que explican la mayor varianza en las peleas de hombres.
4. Análisis utilizando una base de datos con los 32 factores que explican la mayor varianza en las peleas de mujeres.

Entonces se realizará el análisis siguiendo las particiones pasadas, con el fin de comparar, explorar y mejorar la precisión final del ganador de la pelea.

Ahora se presenta el nivel de importancia que le da cada factor a las variables implicadas, de esta manera podemos intentar darle cierta interpretación a los factores.

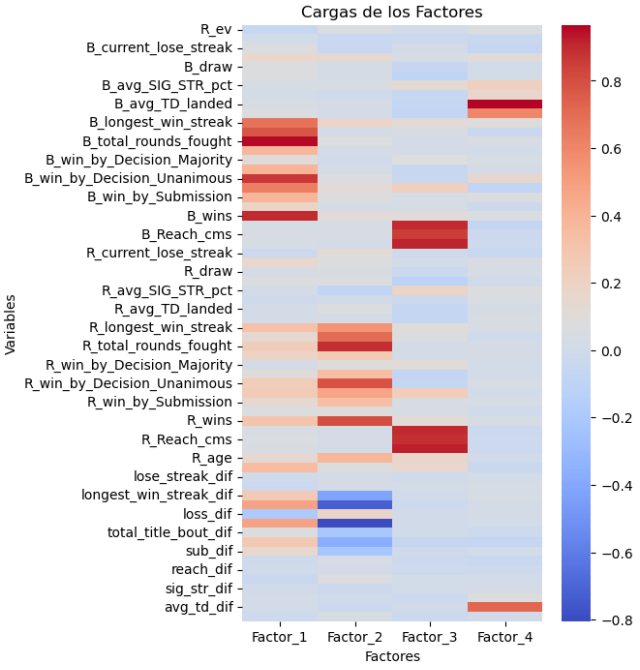

Fig. 3
MATRIZ DE CARGAS

Lo que observamos es que como ya vimos en la gráfica de las varianzas individuales de cada factor, los tres primeros factores son los predominantes sobre los demás, estos capturan información de los 2 peleadores (los primeros dos factores) y es las diferencias entre ambos peleadores, por lo que al factor se le relaciona al peleador de la esquiana azul y al segundo factor al peleador de la esquina roja. Y el finalmente el tercer factor es le asociado a las características físicas de ambos.

Balance de variable

Podemos observar que hay un desbalance en nuestra variable predictora tanto en ganadores y perdedores de mujeres como de hombres como se llega a mostrar en las figuras 4 y 5, respectivamente. En los datos de hombres, la proporción de ganadores es aproximadamente el 57.73 %, mientras que la de perdedores es del 42.27 %. En el caso de las mujeres, la proporción de ganadores es del 59.02 % y la de perdedores es del 40.98 %.

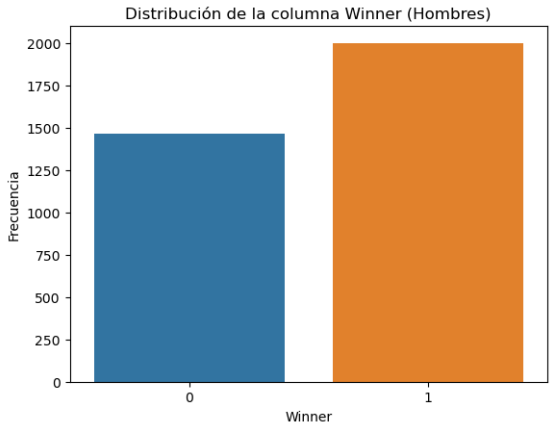

Fig. 4
BALANCE DE LA VARIABLE WINNER EN HOMBRES EN LOS DATOS DE LA UFC

Para abordar este desbalance en la variable predictora, se utilizó la técnica SMOTE (Synthetic Minority Oversampling Technique). SMOTE es una técnica de sobremuestreo que genera ejemplos sintéticos de la clase minoritaria para equilibrar el conjunto de datos y mejorar el rendimiento del modelo predictivo. Esta técnica es fundamental para garantizar que el modelo no esté sesgado hacia la clase mayoritaria y pueda hacer predicciones más precisas y equilibradas.

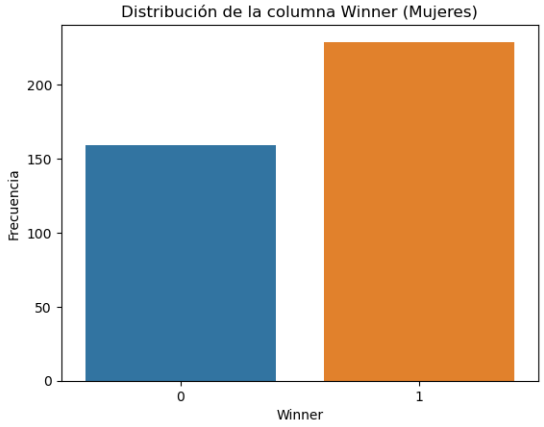

Fig. 5
BALANCE DE LA VARIABLE WINNER EN MUJERES EN LOS DATOS DE LA UFC

Escalamiento multidimensional

Continuando con el desarrollo del estudio se calculó la matriz de Gower y se realizó un análisis mediante MDS para las cuatro divisiones de la base de datos. Este proceso se orientó a descubrir agrupaciones significativas que facilitaran la diferenciación y el análisis predictivo más detallado por grupos homogéneos. La identificación de estos conglomerados con características compartidas en teoría debería permitir refinar la precisión de las predicciones.

En los análisis de MDS para las bases de datos de hombres y mujeres sin la reducción por factores, observamos patrones distintos en cada grupo.

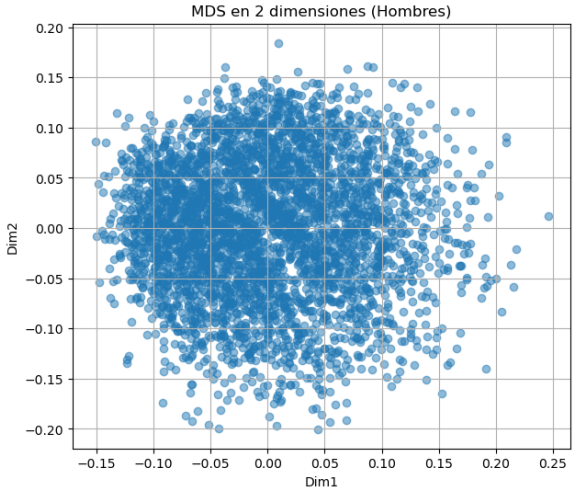

Fig. 6

PROYECCIÓN EN DOS DIMENSIONES APLICANDO MDS A LAS PELEAS DE LAS HOMBRES

Para los hombres, como se muestra en la figura 6 los puntos muestran una distribución que sugiere una posible división transversal, aunque no es completamente definida, sugiriendo que podría haber características que diferencian en dos grupos principales.

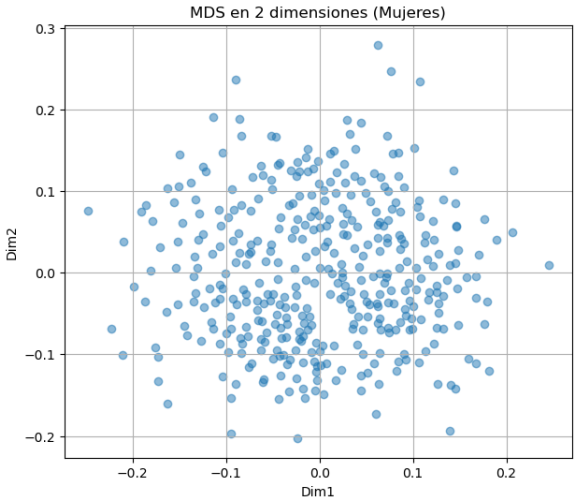

Fig. 7

PROYECCIÓN EN DOS DIMENSIONES APLICANDO MDS A LAS PELEAS DE LAS MUJERES

En mujeres, la dispersión de los puntos es más uniforme sin áreas claras de agrupación. Esto indica que, para este conjunto particular de características, las peleadoras no se agrupan en subconjuntos distintos basados en los datos analizados, lo cual puede ser indicativo de una menor variabilidad en las características o de que las dimensiones seleccionadas no capturan bien las diferencias entre las peleadoras. Este resultado podría motivar un análisis más detallado para identificar si otras variables o combinaciones de ellas podrían revelar patrones más claros como lo pueden ser las bases de datos de los factores.

En las representaciones de MDS de dos dimensiones para los datos de factores de hombres y mujeres, observamos distintas distribuciones. En la gráfica de hombres en la figura 8, se nota una posible distribución tipo dona que podría sugerir grupos diversos, especialmente concentrados en torno al centro del gráfico. Esto podría indicar diferencias en las características de los peleadores que son más o menos similares entre sí, claro que es necesario mejorar o tratar de abordar estos datos desde otra perspectiva para confirmarlo.

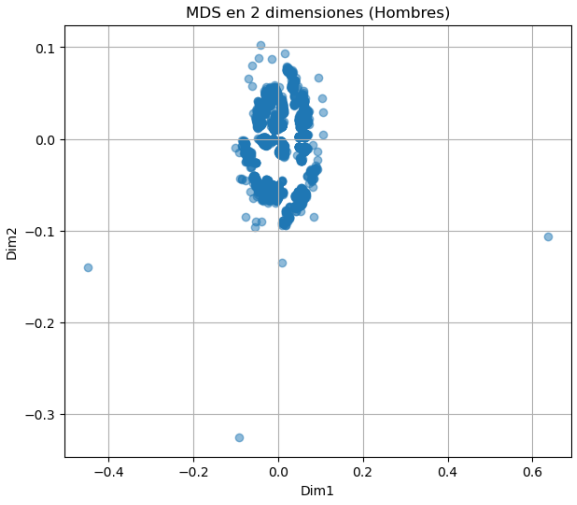

Fig. 8

PROYECCIÓN EN DOS DIMENSIONES APLICANDO MDS A LOS FACTORES DE LAS PELEAS DE HOMBRES

Por otro lado, la dispersión en la gráfica 9 de mujeres parece más homogénea y no muestra una separación clara,

lo que sugiere que las características de las peleadoras podrían no diferir tan significativamente o que los factores seleccionados no capturan completamente las diferencias relevantes.

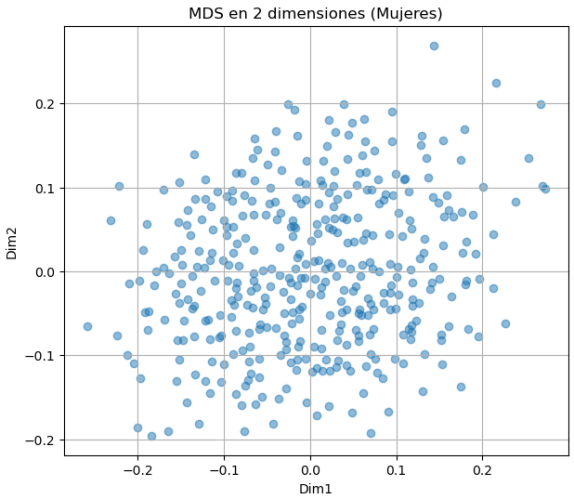

Fig. 9
PROYECCIÓN EN DOS DIMENSIONES APLICANDO MDS A LOS FACTORES DE LAS PELEAS DE MUJERES

Además de estos análisis en dos dimensiones, exploramos la opción de visualizar estas representaciones de los datos en tres dimensiones, lo cual podría revelar estructuras más complejas o agrupaciones no detectadas en las proyecciones bidimensionales. Analizar los datos en tres dimensiones nos permite una exploración más profunda y puede ofrecer nuevas perspectivas sobre las agrupaciones.

Para las peleas masculinas la proyección en 3D como se muestra en la figura 10, observamos una linea central con dos subdivisiones a las orillas como se mostraba en el caso bidimensional.

En el caso de las peleas femeninas, figura 11, la dispersión parece ser más clara a dos grupos.. Esto podría indicar diferencias menos marcadas entre las combatientes en términos de las variables consideradas.

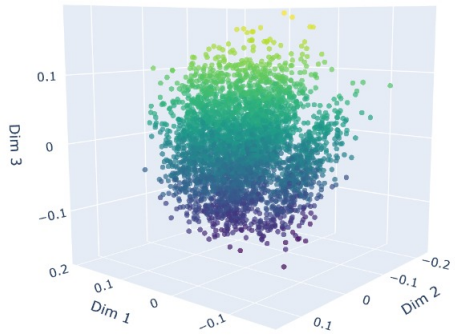

Fig. 10
PROYECCIÓN EN 3 DIMENSIONES DE LOS DATOS DE PELEAS DE HOMBRES

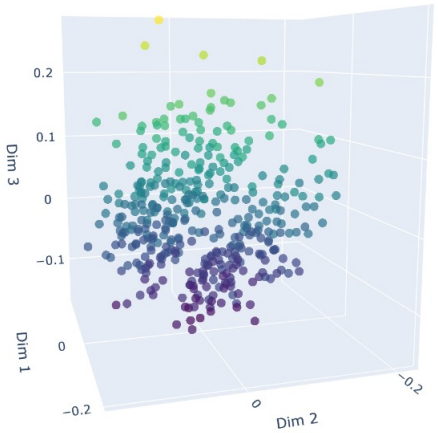

Fig. 11
PROYECCIÓN EN 3 DIMENSIONES DE LOS DATOS DE PELEAS DE MUJERES

Por otra parte, en los resultados del análisis de los factores para hombres y mujeres, observamos patrones distintos en la representación tridimensional.

Para los hombres, la distribución muestra una concentración densa con algunos puntos aislados que sugieren posibles outliers o factores con características únicas. Esta agrupación densa podría indicar similitudes en las características físicas y estadísticas de combate entre la mayoría de los peleadores tal como se demuestra en la figura 12, realmente no se encuentra un patrón importante en esta distribución

de los factores.

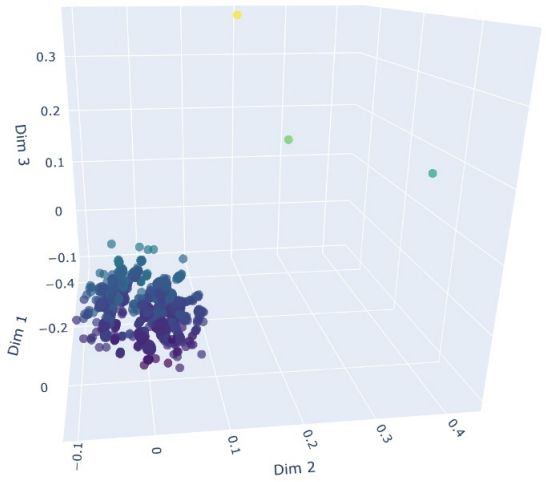

Fig. 12
PROYECCIÓN EN 3 DIMENSIONES DE LOS DATOS DE PELEAS DE HOMBRES EN LOS 32 FACTORES

En contraste, la distribución para las mujeres muestra una dispersión más clara como se demuestra en la figura 13, diferenciando bien dos grupos en 3D. Esto podría proporcionar oportunidades más claras para identificar patrones o tendencias que podrían predecir los resultados de las peleas.

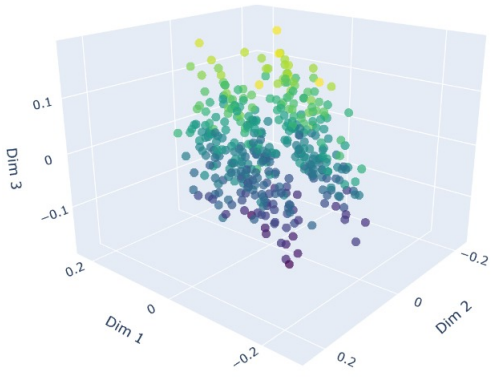

Fig. 13
PROYECCIÓN EN 3 DIMENSIONES DE LOS DATOS DE PELEAS DE MUJERES EN LOS 32 FACTORES

Clasificación

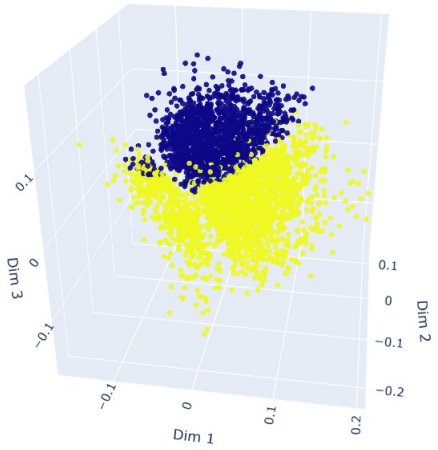

Fig. 14
AGRUPAMIENTO SEGÚN SPECTRAL CLUSTERING PARA LA BASE DE DATOS DE HOMBRES EN 3D (AMARILLO CLÚSTER 1, AZUL CLÚSTER 2)

Ahora que identificamos las particiones según el análisis de MDS, podemos realizar clasificaciones adecuadas para los datos por generos o factores.

Para la base de datos de hombres, que teniamos representada en 3D, se uso el algoritmo de clasificación Spectral Clustering ya que mejoraba bastante la clasificación en dos grupos representativos, logrando asi el resultado de la figura 14.

Siguiendo la misma metodología seguimos con las mujeres, donde identificabamos dos grupos de igual forma claros, en este sentido, el algoritmo de clasificación que mejor vista de la representación de los datos fue el de Kmeans, obteniendo los resultados de la figura 15. Donde representamos claramente los dos grupos con la separación esperada.

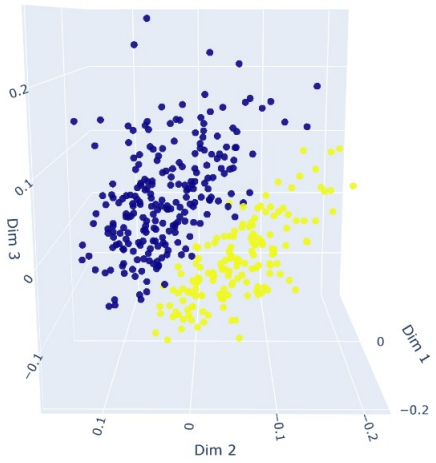

Fig. 15
AGRUPAMIENTO SEGÚN KMEANS PARA LA BASE DE DATOS DE MUJERES EN 3D (AMARILLO CLÚSTER 1, AZUL CLÚSTER 2)

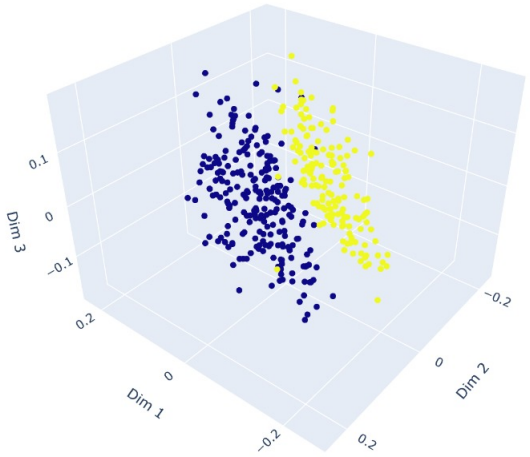

Fig. 16
AGRUPAMIENTO SEGÚN SPECTRAL CLUSTERING PARA LA BASE DE DATOS DE FACTORES DE MUJERES EN 3D (AMARILLO CLÚSTER 1, AZUL CLÚSTER 2)

En el caso de los factores se observó que el conjunto de datos de mujeres daba una mejor representación, lo que nos llevo a hacer la clusterización a este conjunto de factores con el algoritmo de Spectral Clustering, obteniendo como resultados los observados en la figura 16. Donde claramente representamos los dos grupos que se hacían mención al momento de obtener esta representación en la figura 11.

Ya obtenidas estas configuraciones podemos dar paso a lo siguiente que es encontrar la mejor configuración que nos ofrezca el modelo LDA para la predicción.

Predicción del ganador

Nuestro punto de partida es la clasificación sobre las variables continuas, sin realizar más que un proceso de estandarización, la metodología de clasificación usada en el presente es, como ya se mencionó al inicio, un discriminante lineal. Entonces los resultados para la clasificación con LDA se muestran a continuación

	precision	**recall**	**f1-score**
0	0.63	0.67	0.65
1	0.66	0.62	0.64
accuracy		0.64	
macro avg	0.65	0.65	0.64
weighted avg	0.65	0.64	0.64

TABLE I
BASE DE DATOS COMPLETA

Con una exactitud a partir de un Cross-validation de: $0{,}6210 \pm 0{,}0653$, veamos las distribuciones de las proyecciones dadas por LDA

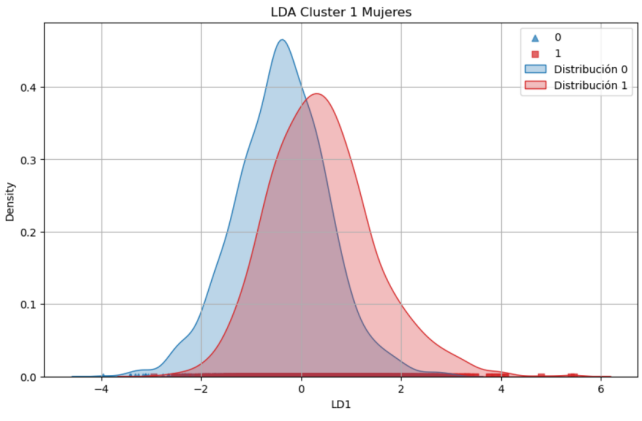

Fig. 17
DISTRIBUCIÓN LDA PARA LA BASE DE DATOS COMPLETA

La recuperación para la clase 0 es de 0.67, indicando

que el modelo fue capaz de identificar correctamente el 67 % de todos los casos reales de la clase 0.

Para la clase 1, la recuperación es de 0.62, mostrando que el modelo identificó correctamente el 62 % de todos los casos reales de esta clase.

Bien, ahora ya con estos datos de referencia, debemos saber que los análisis anteriores servirán para encontrar una configuración en la cual mejoremos nuestro rendimiento al momento de predecir los ganadores, es decir este es nuestro Baseline. Ahora analizamos cada una de estas configuraciones:

	precision	recall	f1-score
0	0.64	0.64	0.64
1	0.63	0.63	0.63
accuracy		0.64	
macro avg	0.63	0.63	0.63
weighted avg	0.63	0.64	0.63

TABLE II
LDA PARA LA SEPARACIÓN DE HOMBRES

	precision	recall	f1-score
0	0.69	0.72	0.70
1	0.61	0.58	0.60
accuracy		0.66	
macro avg	0.65	0.65	0.65
weighted avg	0.66	0.66	0.66

TABLE III
LDA PARA LA SEPARACIÓN DE MUJERES

Ahora para un análisis planteado por generos como se muestran en las tablas 2 y 3, obtenemos que los resultados sugieren que el modelo tiene un rendimiento ligeramente mejor al predecir los resultados de los combates para mujeres en comparación con los hombres. La mayor recuperación y precisión en la clase de perdedores en mujeres sugiere que el modelo es más efectivo en identificar los casos de derrotas en combates femeninos. Esto puede deberse a características inherentes en los datos de las mujeres que el modelo puede capturar más efectivamente.

Tras aplicar técnicas de clustering en tres dimensiones sobre los datos de hombres, identificamos dos grupos distintos. Aquí analizamos las métricas de predicción para cada uno de estos clústeres:

	precision	recall	f1-score
0	0.64	0.64	0.64
1	0.65	0.64	0.65
accuracy		0.64	
macro avg	0.64	0.64	0.64
weighted avg	0.64	0.64	0.64

TABLE IV
LDA HOMBRES (CLUSTER 1)

	precision	recall	f1-score
0	0.62	0.63	0.63
1	0.61	0.61	0.61
accuracy		0.62	
macro avg	0.62	0.62	0.62
weighted avg	0.62	0.62	0.62

TABLE V
LDA HOMBRES (CLUSTER 2)

Los resultados de ambos clústeres sugieren diferencias sutiles en el rendimiento del modelo, siendo el primer clúster ligeramente superior en términos de precisión y recall. Esto podría indicar diferencias en la homogeneidad o en las características de los datos dentro de cada clúster, que afectan la eficacia del modelo predictivo. Aunque si existe una pequeña diferencia entre estos dos reportes, no se esta logrando lo buscado que era que al dividir este conjunto de datos, podriamos rescatar más información haciendo mas eficiente el modelo.

Hacemos un análisis similar para el caso de las mujeres en donde el Cluster 1 muestra un rendimiento global superior,

con mejores métricas en todas las categorías con respecto al cluster 2, especialmente en la precisión y recuperación para la clase 0. El Cluster 2 presenta un rendimiento más equilibrado entre las clases, pero con métricas generalmente más bajas, indicando una homogeneidad menor y una capacidad de predicción más limitada.

	precision	recall	f1-score
0	0.71	0.67	0.69
1	0.69	0.73	0.71
accuracy		0.70	
macro avg	0.70	0.70	0.70
weighted avg	0.70	0.70	0.70

TABLE VI
LDA MUJERES (CLUSTER 1)

Como vemos además de tener un buen equilibrio en las clases, logramos tener una precisión del 70 %, por lo que debemos observar la distribución generada por la maximización de las distancias en las medias dado por LDA:

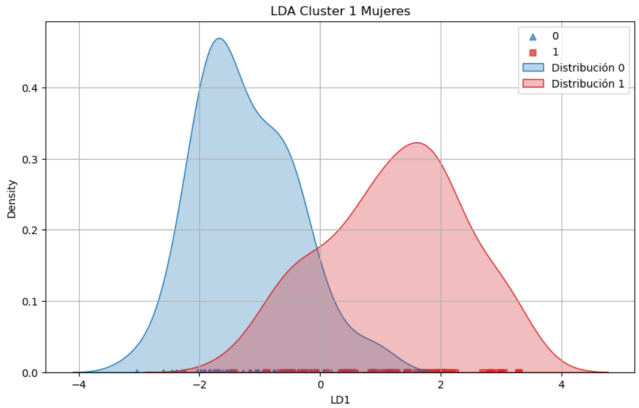

Fig. 18
DISTRIBUCIÓN DE LDA PARA EL CLUSTER 1 DE MUJERES

Si bien, es cierto que aún hay traslape entre las distribuciones, que es justo lo que está causando el error, también es cierto que las distribuciones ahora son identificables y mucho más claras que en el baseline, por lo que el corte lineal nos da resultados relativamente mejores.

	precision	recall	f1-score
0	0.52	0.61	0.56
1	0.61	0.52	0.56
accuracy		0.56	
macro avg	0.57	0.57	0.56
weighted avg	0.57	0.56	0.56

TABLE VII
MUJERES 2 (CLUSTER 2)

Clasificación con análisis de factores

Los resultados para los modelos predictivos desarrollados con los 32 factores obtenidos y la segmentación en dos clústeres para hombres y mujeres son presentados en las siguientes tablas:

	precision	recall	f1-score
0	0.61	0.59	0.60
1	0.63	0.64	0.63
accuracy		0.62	
macro avg	0.62	0.62	0.62
weighted avg	0.62	0.62	0.62

TABLE VIII
HOMBRES 1 FACTORES (CLUSTER 1)

El Cluster 1 para el caso de los hombres que se presenta en la tabla VIII muestra resultados bastante equilibrados entre las clases, pero con una precisión general moderada. Cluster 2 revela una mejora significativa en la precisión para la clase 0 y en la recuperación para la clase 1, indicando un modelo más efectivo para identificar correctamente a los perdedores y ganadores, respectivamente.

	precision	recall	f1-score
0	0.71	0.62	0.66
1	0.60	0.69	0.64
accuracy		0.65	
macro avg	0.65	0.65	0.65
weighted avg	0.66	0.65	0.65

TABLE IX
HOMBRES 2 FACTORES (CLUSTER 2)

	precision	recall	f1-score
0	0.55	0.55	0.55
1	0.54	0.54	0.54
accuracy		0.54	
macro avg	0.54	0.54	0.54
weighted avg	0.54	0.54	0.54

TABLE X
MUJERES 1 FACTORES (CLUSTER 1)

El Cluster 1 en mujeres muestra resultados bastante uniformes, pero con métricas relativamente bajas en todas las categorías, lo que sugiere una menor capacidad del modelo para diferenciar entre clases bajo este conjunto de factores. Por otro lado, el cluster 2 muestra una mejora notable en todas las métricas, especialmente en precisión y recuperación para ambas clases, indicando una mayor efectividad del modelo en identificar ganadores y perdedores basado en los factores seleccionados.

	precision	recall	f1-score
0	0.66	0.70	0.68
1	0.70	0.66	0.68
accuracy		0.68	
macro avg	0.68	0.68	0.68
weighted avg	0.68	0.68	0.68

TABLE XI
MUJERES 2 FACTORES (CLUSTER 2)

Consideraciones generales

El estudio aplicó técnicas avanzadas de análisis estadístico para predecir los ganadores en los combates de la UFC, destacando la eficacia de modelos simplificados en un deporte altamente regulado como la UFC. A continuación, se analizan resultados claves:

- Aplicar un modelo de Análisis Discriminante Lineal (LDA) a datos no segmentados resultó en un rendimiento comparativamente alto, demostrando que modelos simples pueden ser efectivos en contextos bien regulados.
- El uso combinado de análisis de factores y segmentación por clústeres mostró mejoras moderadas en la precisión de las predicciones, siendo más notables en los modelos segmentados por género.
- En deportes como la UFC, donde las peleas están diseñadas para ser equilibradas y competitivas, alcanzar una precisión del 64 % en la predicción de resultados es notablemente útil. Esto indica una capacidad significativa para anticipar los resultados basándose en análisis predictivos
- La aplicación de MDS y Gower para identificar conglomerados no reveló agrupaciones significativamente claras, lo que sugiere una gran homogeneidad en las características físicas y técnicas entre los peleadores, reflejando la efectividad de las políticas de igualación de la UFC.
- Para los apostadores y fanáticos, entender estas dinámicas no solo mejora la experiencia de visualización, sino que permite hacer apuestas informadas.
- Los resultados sugieren que las características que predicen los resultados en peleas de mujeres difieren notablemente de las de los hombres. Esta observación apoya la necesidad de estrategias de entrenamiento y preparación diferenciadas según género.
- El desbalance que notamos en terminos de generos en hombre y mujeres, si tiene consecuencia en los modelos de predicción al momento de la toma de desición.

V. Conclusiones

En conclusión, el estudio aplicó métodos estadísticos avanzados como el Análisis de Factores y el Escalamiento Multidimensional (MDS), complementados con técnicas de clustering, para explorar y predecir los resultados de combates en la UFC. A pesar de que la mejora en la precisión predictiva fue modesta, el uso de estos métodos ofreció insights valiosos sobre las dinámicas y características que influyen en los resultados de los combates.

El análisis reveló diferencias notables en los patrones de datos entre los géneros, subrayando la necesidad de enfoques diferenciados para hombres y mujeres en el entrenamiento y en la estrategia de combate. Además, los modelos predictivos basados en datos segmentados por género mostraron un desempeño diferencial, lo que destaca la importancia de personalizar los modelos predictivos según las características específicas de los peleadores.

Finalmente, los resultados del estudio enfatizan la importancia de continuar explorando y desarrollando modelos más sofisticados, como aquellos basados en aprendizaje profundo, que podrían capturar mejor la complejidad y la naturaleza multidimensional de los datos en deportes de combate.

References

- Dabbert, M. (2024). Ultimate UFC Dataset [Conjunto de datos]. Kaggle. Ultimate UFC Dataset en Kaggle
- Velicer, W.F., Eaton, C.A., Fava, J.L. (2000). Construct Explication through Factor or Component Analysis: A Review and Evaluation of Alternative Procedures for Determining the Number of Factors or Components. In Goffin, R. D., Helmes, E. (Eds.). Boston: Kluwer. (Pp. 41-71).
- Borg, I., Groenen, P. J. F. (2005). Modern Multidimensional Scaling: Theory and Applications (2nd ed.). Springer.
- Groenen, P. J. F., van de Velden, M. (2004). Multidimensional scaling. Econometric Institute Report EI 2004-15.
- Johnson, R.A., Wichern, D.W. (2007). Applied Multivariate Statistical Analysis.
- James, N., Mellalieu, S.D. (2005). Sports data mining: Predicting results for the college football games.