Entregable 2

Integrantes

Sebastián Saldarriaga Cardona
Juan David Ríos Rodríguez
Wilmer Mario Leiva Esteban

Prof. Raul Ramos Pollan

Introducción a la Inteligencia Artificial

Universidad de Antioquia

2023-2

Introducción

Este informe tiene como propósito ofrecer una descripción detallada del proceso de implementación de un modelo de predicción basado en las características clave de un jugador de fútbol con el objetivo de predecir su valor en el mercado. A lo largo del informe, se detallan todas las etapas que conforman este proceso, junto con los resultados derivados de su ejecución. Además, tras alcanzar un progreso significativo en el proyecto, se presentan conclusiones extraídas de la implementación y los resultados obtenidos

exploración descriptiva del dataset

El dataset seleccionado es FIFA 19 complete player dataset de Kaggle que se compone 18.200 filas y 89 columnas, tras haber implementado un proceso riguroso de limpieza de datos descartamos de su aplicación un total de 18 columnas, lo que nos dejo 71 columnas que consideramos aportan peso a la decisión que puede tomar el modelo.

Descripción del progreso alcanzado

En el proceso implementado para poder desarrollar el modelo destacamos 3 fases importantes en su implementación:

1- Importación del Dataset

En primer lugar se debe obtener el token que kaggle suministra para poder importar un dataset directamente a colab, especificamos el dataset que queremos descomprimir, luego tras su implementación obtenemos un json, y con eso ya se pueden ejecutar las celdas.

A Partir de esto continuamos con el entendimiento del dataset que consiste en un análisis de las columnas para estudiar qué es lo que contienen, el tipo de dato de cada una y seleccionar aquellas que nos interesan. Esta etapa es importante ya que sin hacer un análisis minucioso nos permite descartar aquellos datos que son demasiado irrelevantes antes de empezar un proceso más sofisticado.

2- Limpieza del Dataset

El proceso de limpieza consta de 9 etapas, en cada una de ellas será descartado una fila o columna o se hará una conversión si consideramos que será relevante para el juicio sobre el precio del jugador, en cada una existe un criterio que justifica la eliminación de dicho valor o su conversión.

2.1 Eliminar columnas menos relevantes:

En el proceso de entendimiento del dataset a primera vista se pueden seleccionar columnas que no tienen el menor impacto dentro del proceso de selección ya sea por su categoría o por tipo de dato que contiene por lo que son suprimidas de inmediato del dataset.

2.2 Celdas con valores NaN:

En las columnas que contenían en su mayoría valores NaN fueron descartadas ya que no hay manera de implementar una conversión y carecen de peso para el modelo.

2.3 Jugadores Específicos:

Algunos jugadores carecen de información dentro de columnas con una alta relevancia como 'Overall' o 'Age', al tratarse de solo dos datos los consultamos y se los agregamos de forma manual.

2.4 Jugadores sin Equipo:

Los jugadores que no tienen ningún equipo presentan un patrón atípico ya que además no tiene información en todas las columnas relacionadas con valores monetarios por lo tanto al presentar dicho desbalance también fueron descartados.

2.5 Jugadores con muchas estadísticas desconocidas:

Existe un total de 48 jugadores que tiene campos vacíos en 10 columnas, por lo tanto determinamos que a la hora de entrenar el modelo podrían causar un desbalance, así que fueron descartados.

2.6 Eliminación de los Porteros:

Detectamos una irregularidad al revisar que los porteros tiene muchas de columnas con datos NaN, tras evaluar el dataset se llega a la conclusión que el total de los portero carecen de muchos datos por lo que son totalmente irrelevantes para entrenar el modelo, así que fueron descartados y las columnas que evaluaban específicamente este tipo de jugadores también fueron eliminadas.

2.7 Jugadores con Release Clause en NaN:

Se detectó otra columna que contenía valores NaN se trata de Release Clause, la razón es que en su momento estos se encontraban de préstamo por lo tanto el campo quedó nulo. Los jugadores que tuvieran este campo de esta manera fueron eliminados.

2.8 Conversión de datos numéricos:

Después de habernos asegurado que el dataset ya no tuviera columnas con datos nulos procedimos a hacer la conversión de múltiples columnas:

- **2.8.1 Conversión de valores monetarios:**

  Hay 3 columnas que hacen referencia a valores monetarios. Fue crucial convertir estas columnas de formato de texto a números.

  Las columnas en cuestión fueron: Value, Wage y Release Clause.

- **2.8.2 Conversión de calidad en diferentes posiciones:**

  Las columnas 'LS', 'ST', 'RS', 'LW', 'LF', 'CF', 'RF', 'RW', 'LAM', 'CAM', 'RAM', 'LM', 'LCM', 'CM', 'RCM', 'RM', 'LWB', 'LDM', 'CDM', 'RDM', 'RWB', 'LB', 'LCB', 'CB', 'RCB', 'RB'; hacen referencia a la valoración de un jugador para jugar en estas posiciones, las cuales, en su mayoría, no son su posición natural.

  Estas columnas presentan valores con formato X+Y por lo que reemplazamos este valor con el resultado de dicha suma.

- **2.8.3 Conversión de Height y Weight:**

  En estas columnas los valores se encontraban en pies,pulgadas y libras se hizo su respectiva conversión a centímetros y kilogramos y se suprimieron las unidades que los acompañaban.

2.9 Conversion de columnas categóricas:

Finalmente, el dataset presentaba 4 columnas categóricas: Nationality, Club, Preferred Foot y Position.

- Preferred Foot se pasó de right y left a 1 y 0.
- Nationality y Club consideramos que lo mejor era reemplazarlos por la media del valor de sus jugadores.
- Finalmente, para Position decidimos suprimir ya que con la calidad específica para cada posición es suficiente para esta categoría.

Tras haber realizado con éxito estos 9 procesos en la limpieza de los datos el Dataset terminó con un total de 71 columnas libres de valores NaN, unidades que los acompañan, datos irrelevantes y categóricos.

Se procedió a guardar el nuevo Dataset para continuar con la etapa de entrenamiento y Pruebas.

3- Pruebas en Modelos y métricas de desempeño

Una vez establecido el dataset final para general el modelo, hicimos pruebas en diferentes tipos de ellos para evaluar su precisión y desempeño, siguiendo estos pasos:

3.1 descargar el dataset procesado:

En este proceso, instalamos la biblioteca "gdown" para descargar archivos de Google Drive en Google Colab, importamos las bibliotecas "pandas" y "numpy" para el manejo de datos y operaciones matemáticas, definimos la URL y el nombre del archivo a descargar, lo descargamos con "gdown.download" y luego cargamos los datos en un DataFrame de Pandas llamado "df", ajustando la visualización para mostrar todas las columnas.

3.2 Preparar valores para los test:

Creamos dos conjuntos de datos, "X" (características) y "y" (variable objetivo) eliminando la columna 'Value'. Luego, dividimos el conjunto de datos en conjuntos de entrenamiento y prueba (80% para entrenamiento, 20% para prueba) utilizando "train_test_split" de scikit-learn, con "random_state" para reproducibilidad.

3.3 Modelos de Regresión:

Tras la creación de los conjuntos de datos y su división en conjuntos de entrenamiento y prueba, avanzamos hacia la aplicación de modelos de regresión en nuestra tarea de análisis. En este contexto, consideramos la **Regresión Lineal** para establecer relaciones lineales entre las características y la variable objetivo 'Value'. La **Regresión Polinómica** se vuelve relevante para capturar relaciones no lineales en los datos. Además, aplicamos técnicas de regularización, como **Ridge, Lasso y Elastic Net,** para combatir el sobreajuste y seleccionar las características más influyentes en el proceso de regresión.

3.4 Modelos de Árbol:

También evaluamos modelos basados en árboles, como **Decision Trees** y **Random Forest**, que nos permiten manejar relaciones no lineales de manera efectiva y capturar interacciones complejas en los datos.

3.5 Comparación de MAE (Mean Absolute Error):

Comparamos varios modelos de regresión para determinar cuál de ellos ofrece las predicciones más precisas. Se usó el Error Absoluto Medio (MAE) como métrica para evaluar el rendimiento de cada modelo. Se hizo con la finalidad de identificar cuál es el

más eficaz en la estimación de los datos, lo que es crucial para tomar decisiones informadas en tareas de análisis y predicción.

3.6 Porcentaje de las predicciones dentro del margen esperado:

De igual forma que hicimos la comparación de MAE, calculamos el porcentaje de predicciones que entran dentro de un rango del 10% del margen de error para cada uno de los modelos probados. Esto es importante ya que inicialmente habíamos establecido que al menos el 85% de las predicciones debía entrar en ese rango para que el modelo pudiera ser considerado válido.

Conclusiones:

La fase de pruebas en diferentes modelos de regresión y árboles demostró la importancia de elegir el enfoque adecuado para predecir el valor de mercado de los jugadores de fútbol. La comparación de métricas, como el Mean Absolute Error (MAE) y el porcentaje de predicciones dentro del margen esperado, permitió evaluar el desempeño de los modelos. Este análisis contribuyó a identificar el modelo más eficaz y validó la utilidad de la implementación.