Universidad de San Carlos de Guatemala

Facultad de ingeniería

Escuela de Estudios de Postgrado

Maestría en Ingeniería para la industria con Especialización

en Ciencias de la Computación

Introducción al análisis de datos

Carlos Luis Pablo Hernández Gramajo.

**Proyecto Documentación de procesos.**

[**https://github.com/HernandezGramajo/Proyecto1\_Aalisis\_de\_datos.git**](https://github.com/HernandezGramajo/Proyecto1_Aalisis_de_datos.git)

**Data Analisys**

Se realizo el análisis de los datos para poder validar la información que se tiene se verificaron los 6 archivos palyer para ver si contenían la misma información.

Y posterior mente se realizo la carga de los archivos y se unificaron

**Texto

Descripción generada automáticamente con confianza media**

**Data Cleaning**

Interfaz de usuario gráfica, Texto

Descripción generada automáticamenteSe empezó con la limpieza de los datos verificando primero los duplicados y eliminándolos, para esto se utilizo el campo “sofifa\_id” que es el identificador único de los jugadores.

Después de obtener el Dataset limpio se utilizo para verificar que columnas son las que contienen datos “NA”

Texto

Descripción generada automáticamente

En el cual se obtuvieron los siguientes resultados

Tabla

Descripción generada automáticamente

Tomando en cuenta las columnas que se tenían con datos “NA”, se tomaron diferentes decisiones :

Para las siguientes columnas se realizo un promedio de la columna y se coloco el valor sustituyéndolo por el “NA”, esto debido a que son clausulas y habilidades.

* Pace
* release\_clause\_eur
* passing
* dribbling
* defending
* physic

Código en R:

Texto

Descripción generada automáticamente

Para la coluna “mentality\_composure”, se tomo todos los datos que nos “NA” de esa columna y se agrego de forma aleatoria, ya que estos datos tenían dos formas de erce una con un número “48” y otro con un numero seguido de un signo y luego de otro número “65-2” esto indica que el jugador puede estar dentro de un rango.

Código en R:

Texto

Descripción generada automáticamente con confianza media

Para los siguientes datos se tuvo que identificar que jugadores juegan en posición de guardameta ya que las siguientes columnas son habilidades exclusivas de los porteros en la cual se incluyó aquellos jugadores que juegan en dos posiciones (incluyendo guardameta ) y se agregó el promedio sustituyendo el valor “NA”, para aquellos que no juegan en la posición de guardameta se les agrego “0”.

* gk\_diving
* gk\_handling
* gk\_kicking
* gk\_reflexes
* gk\_speed
* gk\_positioning

Código en R:

Texto

Descripción generada automáticamente

Para la parte de años de validación de contrado “contract\_valid\_until”, aquellos campos con “NA” se agrego “9999” para indicar año indefinido, es decir, que no tiene año de expiración de contrato.

Código en R:



Para las siguientes dos columnas que contiene “NA” al ser número de camisolas se tomo la decisión de que los números de camisolas no se pueden repetir por jugadores del mismo club es decir no existen un jugador al mismo tiempo con la misma camisola. El mismo principio se tomo para el número de camisola de su selección. Para esto se tomo los jugadores que si tuvieran número y que pertenecieran al club y se colocó un rango del 1 al 99 para agregar números de camisola y se excluye el número que ya se está utilizando por los jugadores del rango establecido.

* team\_jersey\_number
* nation\_jersey\_number

Código en R:

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Se verifica el formato de los datos

Texto

Descripción generada automáticamente

Se procede a cambiar el formato de la fecha de string a date

Interfaz de usuario gráfica, Texto, Aplicación

Descripción generada automáticamente con confianza mediaCódigo en R:

**Data Wrangling**

Se crearon las siguientes Columnas.

**Edad\_2024:** Esta columna se creo para obtener los años año día de hoy de los jugadores a partir de la fecha de nacimiento.

**Potencial\_crecimento:** Esta columna se crea a partir de “potential – overall”, para obtener el potencial de crecimiento de un jugador y ver donde esta ahora y que tan lejos puede llegar.

**valor\_relativo :**  Esta columna se crea a partir de “value\_eur / overall”, con esto se crea un índice que combina el valor del mercado y la calificación general par obtener una media del valor relativo.

**Promedio\_habilidad: :**  Esta columna se crea a partir de “value\_eur / age”, relaciona el valor del mercado con la edad del jugador para ver que tan eficiente es un jugador en relación a su valor y edad.

**valor\_mercado\_categoria:**  Se agrega una categoría al jugador dependiendo del valor del mercado en que se necuentre obedeciendo los siguiente:

value\_eur < 300000 = "Low",

value\_eur >= 300000 & value\_eur < 10000000 = "Medium",

value\_eur >= 10000000 = "High"

Se convierten las habilidades en un formato largo.

Imagen que contiene Gráfico de dispersión

Descripción generada automáticamente

**Data Transformation**

Seleccionamos la columnas númericas y se filtra el dataset, se normalizan los datos y se usara “preProcess()” del paquete “caret” que permite aplicar técnicas de preprocesamiento como normalización. Se transforman los datos en una rango de 0 a 1.

Texto

Descripción generada automáticamente con confianza media

**Datos normalizados**

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

**Interpretación de las columnas normalizadas**

1. **age (Edad):**

* Esto indica que los valores de edad han sido escalados a este rango (0 y 1). Si por ejemplo la edad original de un jugador fuera de 20 años y el rango de edad original fuera de 18 a 40 años, 20 años sería un valor cercano a 0.393.

1. **height\_cm (Altura en cm):**

* Los valores de altura (por ejemplo, 0.294 y 0.608) se normalizan entre 0 y 1. Esto sugiere que la altura de los jugadores se encuentra en un rango determinado, y la normalización ajusta esos valores al nuevo rango. Un valor de 0.294 es cercano al mínimo, y 0.608 es relativamente más alto, pero aún por debajo del máximo.

1. **weight\_kg (Peso en kg):**

* Similar a la altura, los valores de peso también están normalizados. Los valores 0.295 a 0.508 indican que el peso de los jugadores en el conjunto de datos se encuentra dentro de un rango determinado, con los valores más cercanos a 0 indicando los jugadores más ligeros y los más cercanos a 1 los más pesados.

1. **overall (Valoración general):**

* La columna overall ha sido normalizada entre 0 y 1, con los jugadores más valorados en la parte superior del rango. El valor 1 en la primera fila indica que este jugador tiene la valoración máxima posible dentro de este dataset. Un valor de 0.981 indica un jugador que está cerca del máximo, pero no lo alcanza.

1. **potential (Potencial):**

* La columna de potencial también muestra valores cercanos a 1 para los jugadores con mayor potencial. Un valor de 0.945 significa que este jugador tiene un potencial muy alto, pero no el máximo.

1. **value\_eur (Valor en euros):**

* Los valores de los jugadores en euros han sido normalizados entre 0 y 1. Un valor de 0 indica que el jugador no tiene valor o tiene un valor muy bajo, mientras que el valor 0 sugiere que estos jugadores tienen un valor de mercado muy.

1. **wage\_eur (Salario en euros):**

* Los salarios también están normalizados entre 0 y 1 y dependiendo del salario se puede categorizar al jugador.

1. **pace, gk\_diving, gk\_handling, gk\_kicking, gk\_reflexes, gk\_speed (Y otras habilidades específicas de los jugadores):**

* Estas columnas representan habilidades o atributos específicos de los jugadores. Los valores de estas columnas son valores normalizados entre 0 y 1, lo que significa que los jugadores que sobresalen en estas habilidades tendrán valores cercanos a 1, mientras que aquellos con menor habilidad tendrán valores cercanos a 0.

**Estandarización**

Texto

Descripción generada automáticamente

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

**Interpretación de columnas estandarizadas:**

1. **age (Edad)**:
   * El valor **0.763** en las primeras filas indica que estos jugadores tienen una edad superior al promedio de la muestra. Un valor de **1.20** en la fila 6 sugiere que este jugador es significativamente mayor que la media de la muestra en términos de edad.
2. **height\_cm (Altura en cm)**:
   * Los valores negativos (por ejemplo, **-1.78**) indican que estos jugadores tienen una altura por debajo de la media de la muestra. Esto significa que la mayoría de los jugadores en el conjunto de datos tienen una altura superior a los valores de estas primeras filas.
   * Un valor positivo (**0.624**) indica que este jugador es más alto que la media del conjunto de datos.
3. **weight\_kg (Peso en kg)**:
   * Al igual que con la altura, los valores negativos (como **-1.12**) indican que estos jugadores tienen un peso inferior a la media de la muestra. El valor positivo de **0.774** indica que el jugador de la fila 6 tiene un peso superior a la media.
4. **overall (Valoración general)**:
   * Un valor de **4.60** en varias filas indica que estos jugadores tienen una valoración general bastante alta, por encima de la media del conjunto de datos (suponiendo que el valor máximo sea superior a 4.60).
   * El jugador de la fila 6 tiene una valoración de **4.45**, que sigue siendo alta, pero ligeramente por debajo de los primeros jugadores.
5. **potential (Potencial)**:
   * Los jugadores con un **potencial de 4.21** en varias filas son valorados como teniendo un potencial alto (por encima de la media).
   * El jugador de la fila 6, con un valor de **3.73**, tiene un potencial ligeramente inferior al de los demás.
6. **value\_eur (Valor en euros)** y **wage\_eur (Salario en euros)**:
   * Los valores **-0.373** y **-0.386** indican que estos jugadores tienen valores por debajo de la media en cuanto a su valor de mercado y salario. Los valores negativos generalmente implican que estos jugadores tienen un valor de mercado o salario más bajo en comparación con el resto de la muestra.
7. **pace (Velocidad)**:
   * El valor **2.71** en todas las filas indica que la velocidad de estos jugadores es mucho más alta que la media del conjunto de datos (2.71 es un valor bastante alto en una distribución normalizada).
8. **gk\_diving, gk\_handling, gk\_kicking, gk\_reflexes, gk\_speed** **(Habilidades de portero):**
   * Los valores de **-0.351** en todas estas habilidades indican que estos jugadores tienen habilidades inferiores al promedio para estas características específicas. Este patrón sugiere que estos jugadores no son porteros de alto nivel, según sus habilidades estandarizadas.

**Datos estadísticos**

**Datos originales:**

Tabla

Descripción generada automáticamente

Tabla

Descripción generada automáticamente**Datos normalizados:**

**Tabla

Descripción generada automáticamente**

**Datos estandarizados:**

**Conclusiones**

* La edad de los jugadores oscila entre los 16 y 44 años, con una media de aproximadamente 23.5 años. Esto sugiere que la mayoría de los jugadores en los datos son relativamente jóvenes, con un fuerte enfoque en el talento emergente.
* La altura promedio es de 180.9 cm, con un rango que va desde 154 cm hasta 205 cm, lo que cubre una gran variedad de tipos físicos. El peso promedio es de 74.68 kg, lo que refleja una amplia gama de físicos, desde jugadores más ligeros hasta los más corpulentos.
* El valor promedio de los jugadores es de 325,908 EUR, mientras que el salario promedio es de 1,685 EUR. Sin embargo, el rango es extenso, con valores que varían desde 0 EUR hasta 32.5 millones de EUR, lo que indica la presencia de jugadores de élite que dominan las métricas.
* La habilidad promedio (overall) de los jugadores es de 62.07, lo que es consistente con una población de jugadores profesionales que incluyen tanto talentos emergentes como jugadores ya consolidados. La habilidad de los porteros (representada por las métricas de "gk") muestra valores bajos, ya que solo algunos jugadores son especializados en este rol.
* Al comparar los datos normalizados y estandarizados, se observa que ambos procesos afectan las métricas. El rango de valores en los datos normalizados va de 0 a 1, mientras que en los estandarizados va de valores negativos a positivos, con la media acercándose a 0. Esto hace que los datos estandarizados sean útiles para modelos de machine learning que asuman distribuciones estándar, mientras que los datos normalizados son más intuitivos para interpretar las proporciones relativas entre los jugadores.