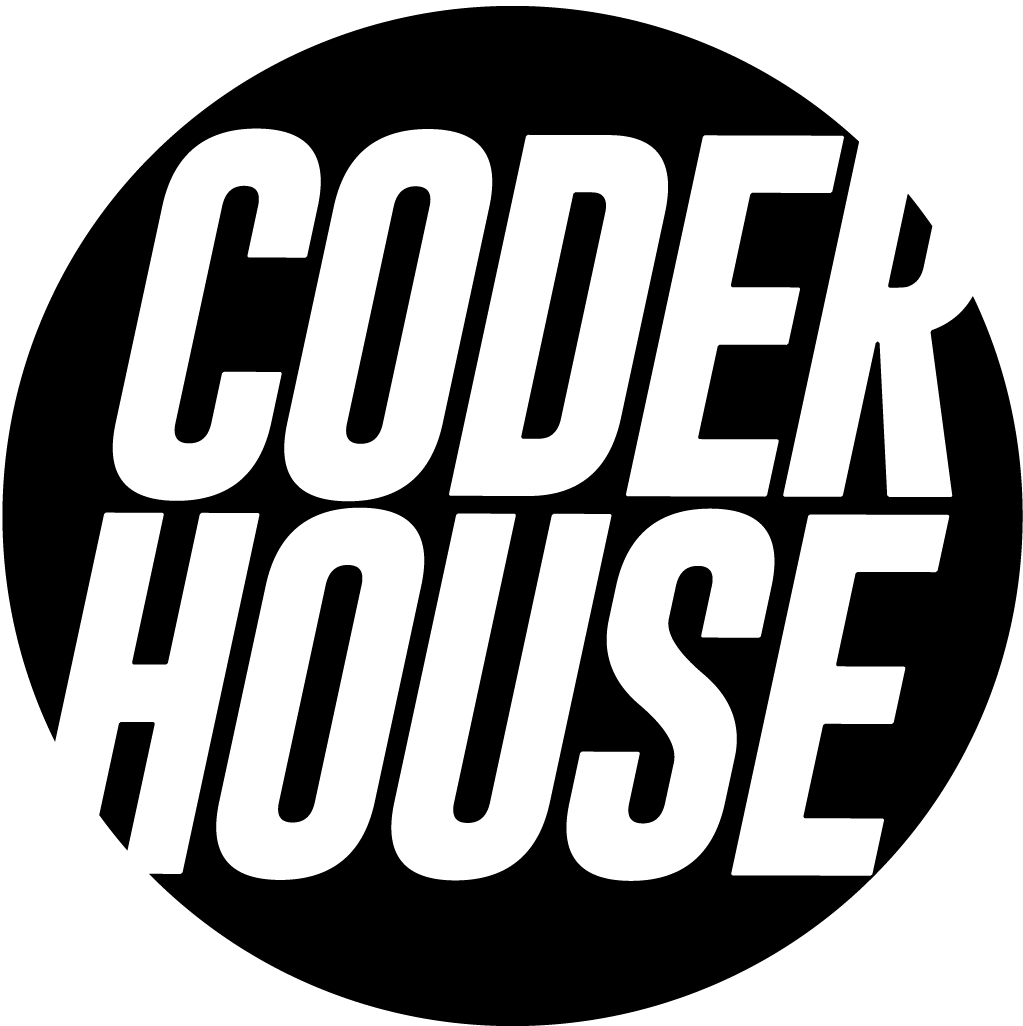
**Curso Data Science: Proyecto Final**

# **Modelo analítico de Machine Learning Aplicado a la NBA**

*Rodrigo Espósito, Horacio Torregrosa, Romila Lorente, Sebastián Kuperman*



**NBA players Season 2020-2021**

7 de diciembre del 2021

**

## Tabla de contenidos

[**Tabla de contenidos**](#_bxxr75819dw7) **1**

[**Introducción**](#_49v2sy2sp6y5) **2**

[**Descripción del caso de negocio**](#_ogvd022evtg9) **3**

[**El problema elegido**](#_hly1g3dlm5t) **3**

[**Descripción de los datos**](#_ds28vo9674i3)[**4**](https://docs.google.com/document/d/1ci9haqsRdmtqYBC0BLsgv600MbArdllh4O3rRC78BmE/edit#heading=h.e5om8k41pxgr)

**Proceso EDA** [**4**](https://docs.google.com/document/d/1ci9haqsRdmtqYBC0BLsgv600MbArdllh4O3rRC78BmE/edit#heading=h.str0z923729b)

[**Algoritmos elegidos y métricas de desempeño**](#_uns3o6vgxc1b)[**4**](https://docs.google.com/document/d/1ci9haqsRdmtqYBC0BLsgv600MbArdllh4O3rRC78BmE/edit#heading=h.2q74rdaf2pei)

[**Iteraciones de Optimización y métricas finales del Modelo Optimizado**](#_ild1m7pffg9a)[**4**](https://docs.google.com/document/d/1ci9haqsRdmtqYBC0BLsgv600MbArdllh4O3rRC78BmE/edit#heading=h.ooy2sosgsoxz)

[**Futuras líneas**](#_qkhd854550qp) **4**

[**Conclusiones**](#_lfasnxpcw32s) **4**

## 

## Introducción

Como se ha demostrado, el análisis simultáneo de información, junto con el procesamiento estadístico, suele ser una de las tareas más demandadas hoy en día a nivel mundial. Por tal motivo uno de los objetivos de este trabajo es el conocer como el machine learning puede convertirse en un pilar fundamental para el tratamiento de datos a gran escala dentro del mundo de los deportes.

Sectores como el informático, medicina, retail y logística llevan años implementando soluciones utilizando el aprendizaje automático y sus predicciones en búsqueda de mejorar el análisis de sus datos, en pro de una predicción futura, ya sea por la implementación de nuevos sistemas o simplemente el mejoramiento de los ya existentes, mediante el uso de algoritmos basados en información antigua o reciente que permita el funcionamiento óptimo del sistema a trabajar.

## Descripción del caso de negocio

Somos una agencia de basketball dedicada al servicio de asesoramiento profesional a clubes profesionales y basketbolistas. Nos especializamos en representación, negociación de contratos deportivos (jugadores & técnicos), marketing deportivo, management y marketing deportivo. Desarrollamos carreras y exportamos talento alrededor del mundo.

Vinculamos jugadores con clubes de distintos clubes del mundo de acuerdo a su perfil. Ayudamos en el desarrollo y búsqueda de las mejores oportunidades para la carrera del jugador, como en la identificación de nuevas oportunidades previo a la expiración de su contrato.

Poseemos un conocimiento profundo del mercado basketbolístico que nos permite brindar un servicio dedicado exclusivamente al pleno desarrollo de la actividad deportiva de jugador. Nuestro objetivo es formar y fortalecer una relación de confianza con nuestros basketbolístas y maximizar sus transferencias y beneficios economicos.

## El problema elegido

Para entender realmente el conjunto de datos, vamos a estudiar el significado y la relevancia de cada variable con respecto al problema. Nuestro objetivo es predecir qué jugadores serán titulares o suplentes y describir sus características a través de modelos de machine learning. Para ello, analizaremos las cualidades de los jugadores que participaron de la temporada **2020-2021** de la NBA.

**Para ello, nos concentraremos en preguntarnos sobre:**

Qué variables consideraremos relevantes para nuestro estudio.

Cuáles son las transformaciones que consideraremos necesarias realizar para obtener nuestro dataset definitivo.

La relevancia de las distintas variables al momento de predecir o agrupar a los jugadores titulares/suplentes.

Qué consejos o recomendaciones podríamos compartir con nuestros clientes al momento de considerar la relevancia deportiva de un jugador o club en función de sus características.

## Descripción de los datos

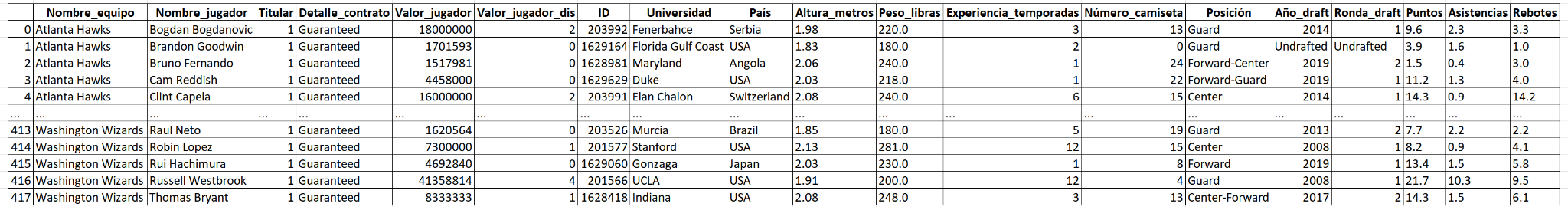
Para la construcción de nuestros modelos de Machine Learning, se utilizará una base de datos histórica extraída de la web kaggle, dado que es la única fuente con datos completos de cada área dentro del ecosistema de la NBA que ofrece información desagregada de jugadores, clubes y juegos lo suficientemente específico como para realizar un análisis de ciencia de datos que sea significativo.

Este conjunto de datos cuenta con tablas que consolidan datos históricos a través de diferentes tablas. Cada tabla contiene datos y estadísticas referente a los equipos, artículos periodísticos, resultados de juegos, rendimiento de jugadores, salarios, y datos de drafts.

Fuente: <https://www.kaggle.com/wyattowalsh/basketball>

Proceso EDA

Una vez importada la base de datos basketball.sqlite y habiendo realizado el merge de las tablas Player\_Salary y Player\_attributes, procedimos a implementar el proceso de exploración y limpieza de nuestro dataframe. Para ello, realizamos las tareas de limpieza, discretización y transformaciones necesarias para obtener nuestro dataset final:



Donde las variables son las siguientes: Nombre\_equipo, Nombre\_jugador, Titular, Detalle\_contrato, Valor\_jugador, Valor\_jugador\_dis, ID, Universidad, País, Altura\_metros, Peso\_libras, Experiencia\_temporadas, Número\_camiseta, Posición,Año\_draft, Ronda\_draft, Puntos Asistencias, Rebotes.

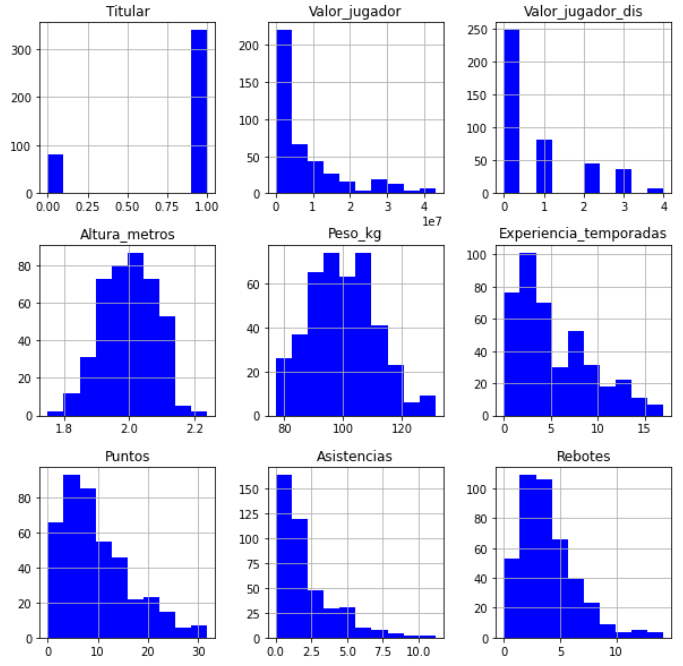
Total de registros: 418.

Total de columnas: 19.

Período temporal: temporada NBA 2020-2021.

Análisis univariado

En este análisis, visualizamos las variables numéricas del DataFrame final (en adelante **df\_final)** para ver las frecuencias y distribuciones: de los datos.

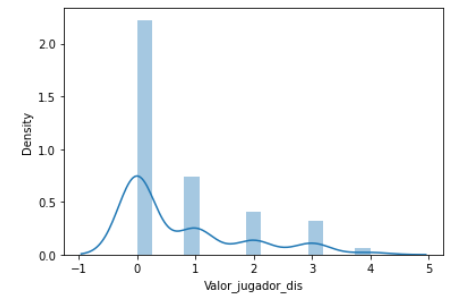


Asimismo, realizamos un gráfico de histograma de la variable discretizada Valor\_jugador\_dis. Notamos que, a simple vista se aprecia que la variable "Valor\_jugador\_dis" presenta:

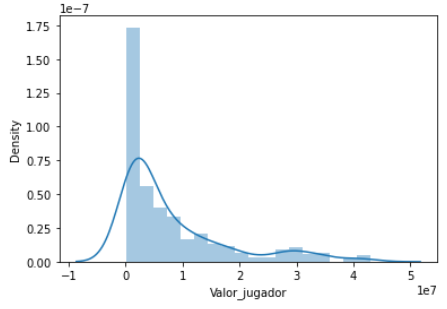
- Una distribución asimétrica positiva.

- Una distribución sesgada hacia la derecha.

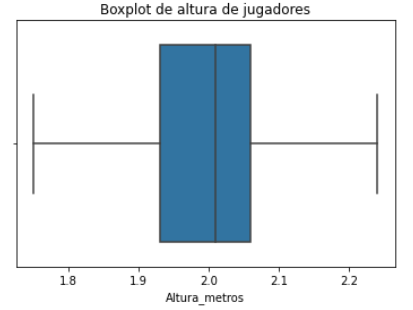
A su vez, notamos que existe una mayoría de jugadores con un valor de mercado clasificado en 0. Mientras que los jugadores que más ganan (3 y 4) representan a la minoría.



Siguiendo con el análisis de variables, observamos la cantidad de jugadores por categoría de valor de mercado obtenida gracias al cluster. Se puede observar la existencia de una mayoría de jugadores con categoría de valor de mercado 0. Para este caso específico, se trat de los jugadores con valores más bajos, mientras que los de categoría 4 son los jugadores con valor de mercado más alto.



Hacemos un boxplot de la altura de los jugadores. Podemos ver que el 50% de los jugadores mide 2 metros. Mientras que solo un 25% mide más de 2,10 metros

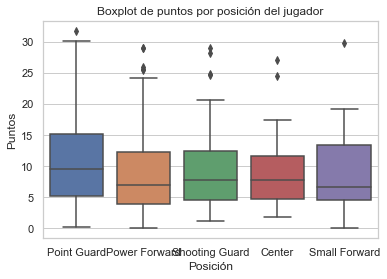


Análisis bivariado

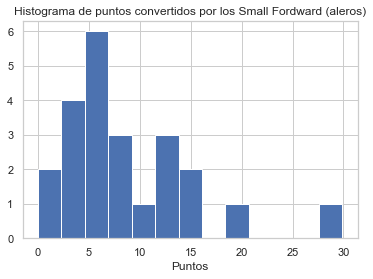
Podemos ver que los point guard (bases o armadores) convirtieron en promedio mayores anotaciones que los center.

Asimismo point guard tiene la mediana más alta, concentrándo a la mitad de los jugadores en ese rol con un promedio de entre casi 9 puntos y 15 puntos. Asimismo vemos que los small forward (aleros) tienen una distribución asimétrica de puntos.

Esto lo veremos en el siguiente gráfico:

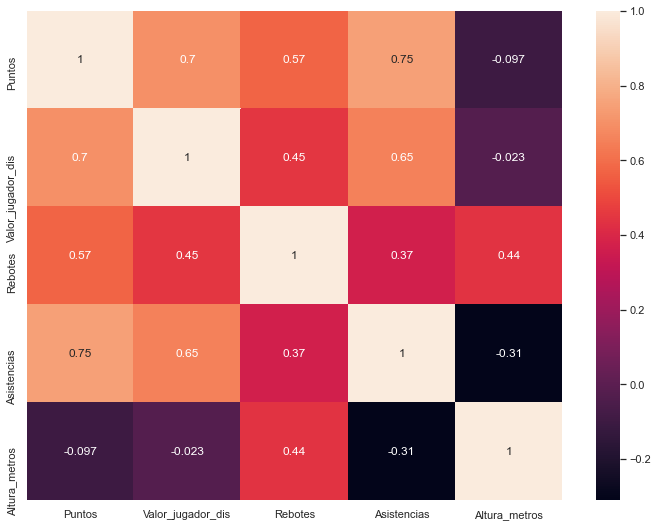


Analizamos la distribución de los puntos promedios por posición. Para ello, realizamos un histograma de puntos convertidos por los Small\_Fordward:



Para analizar a mayor detalle la correlación de valores entre las distintas variables, creamos un mapa de calor de las columnas de Puntos, Valor\_jugador\_dis, Rebotes, Asistencias y Altura\_metros con la intención de ver cuán positivas y/o negativas son las correlaciones.

Podemos ver que la variable Altura\_metros tiene una débil correlación con las otras variables, mientras que la variable Rebotes tiene una correlación media con las resantes variables.



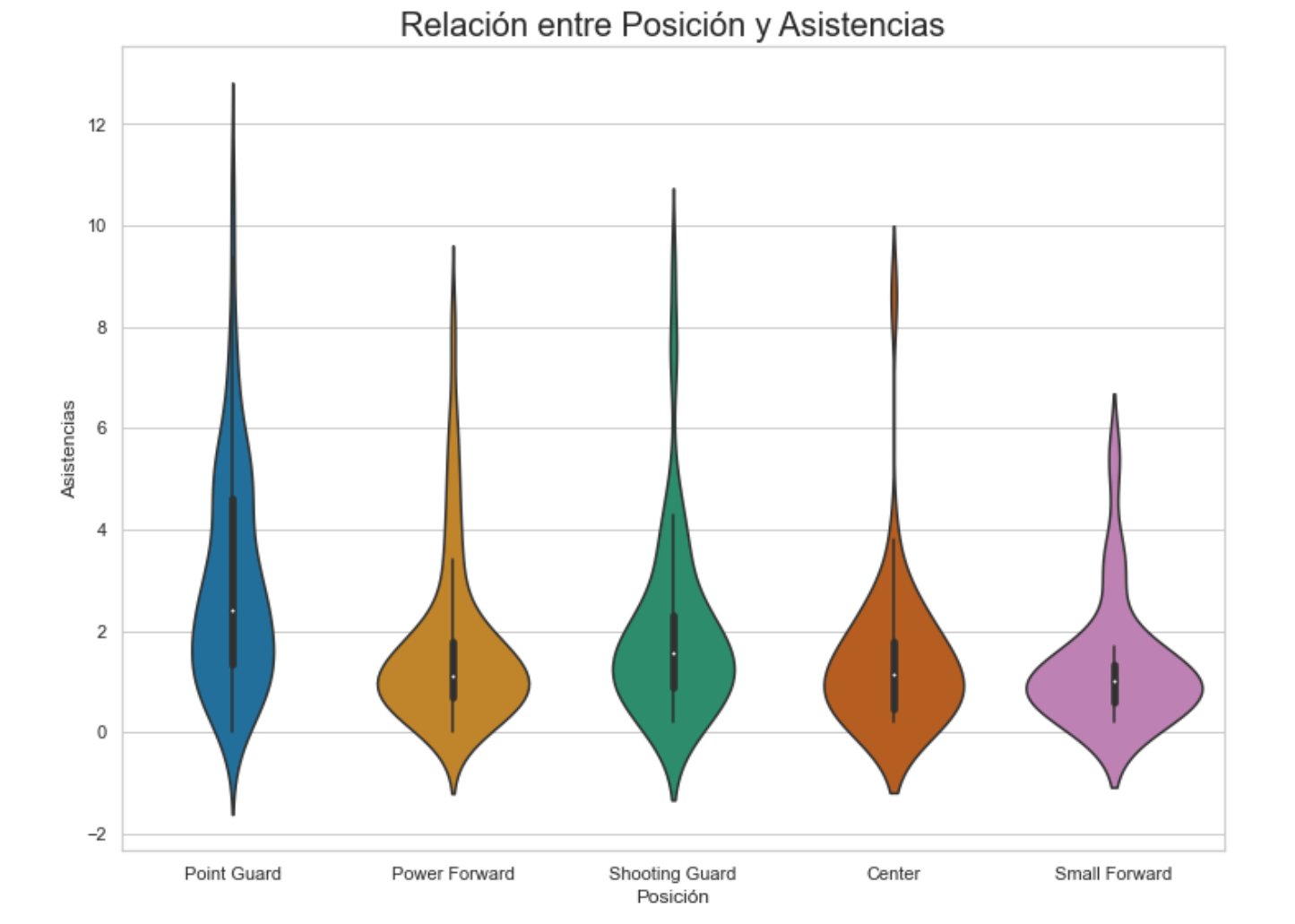
Calculamos cuáles son los 5 clubes que concentran la suma de los jugadores más caros del mercado:



Comparamos el promedio de asistencias en función de la posición de los jugadores.

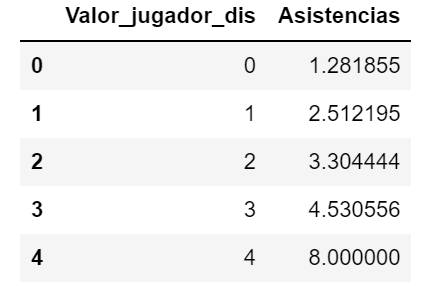
Vemos que los Point guard (bases o armadores) concentran la mayor cantidad de asistencias en promedio.

Por otro lado, vemos que los small forward (aleros) concentran en promedio la menor cantidad de asistencias.



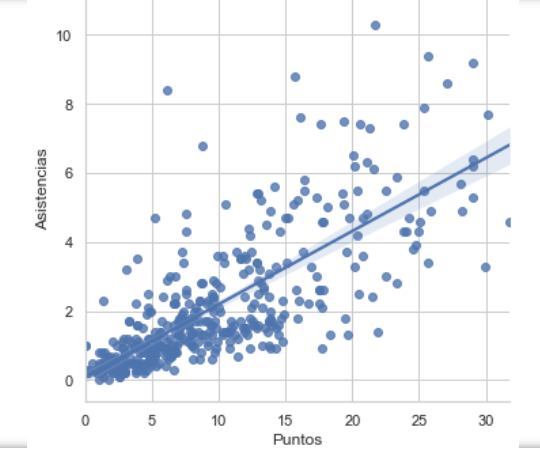
Calculamos el promedio de asistencias en función de la categoría del valor de mercado del jugador.

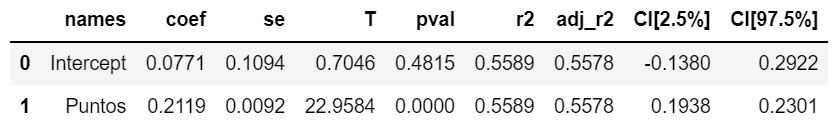
Podemos ver que los jugadores que se encuentran en las categorías 3 y 4 tienen en promedio más del doble de asistencias por partido que los restantes.



Realizamos un gráfico de correlación lineal entre las variables Puntos y Asistencias. Establecemos un intervalo de confianza del 90%. La zombra de la recta sirve para predecir el valor del eje Y en función de la variación de los valores del eje x.

En este caso, existe un 90% de confianza de que los valores del eje y se encuentren en la zona sombreada por arriba y por debajo de la línea.





Vemos los resultados de la regresión lineal entre las variables Puntos y Asistencias a partir del siguiente cuadro.

En este caso, el p-value es igual a 0 (menor a 0.05) y el valor de beta es 0,21 (pendiente positiva). Por convención estadística esto explica que existe una relación lineal ya positiva, es decir que no existe un porcentaje significativo como para suponer que el valor del coef. B sea igual 0 (lo que significaría que la pendiente de la recta de la regresión sería prácticamente nula).

Asimismo, miramos el intervalo de confianza. Entendemos que con una probabilidad del 95% el el valor del coeficiente B (la pendiente) se encuentra entre los valores 0.1938 y 0.2301. Esto quiere decir que, con un alto nivel de confianza, la pendiente (valor del coef. B) no estará comprendida en un intervalo de confianza que incluya al valor cero.

Por último, confirmamos esta suposición estadística al observar el valor del R2. Este coeficiente muestra qué porcentaje del total de las variaciones de la variable Y son explicadas por la variable X. Es decir, representa el porcentaje de variabilidad de los datos explicado por el modelo de regresión lineal. Se espera que R2 sea mas de 0,50, según el modelo.

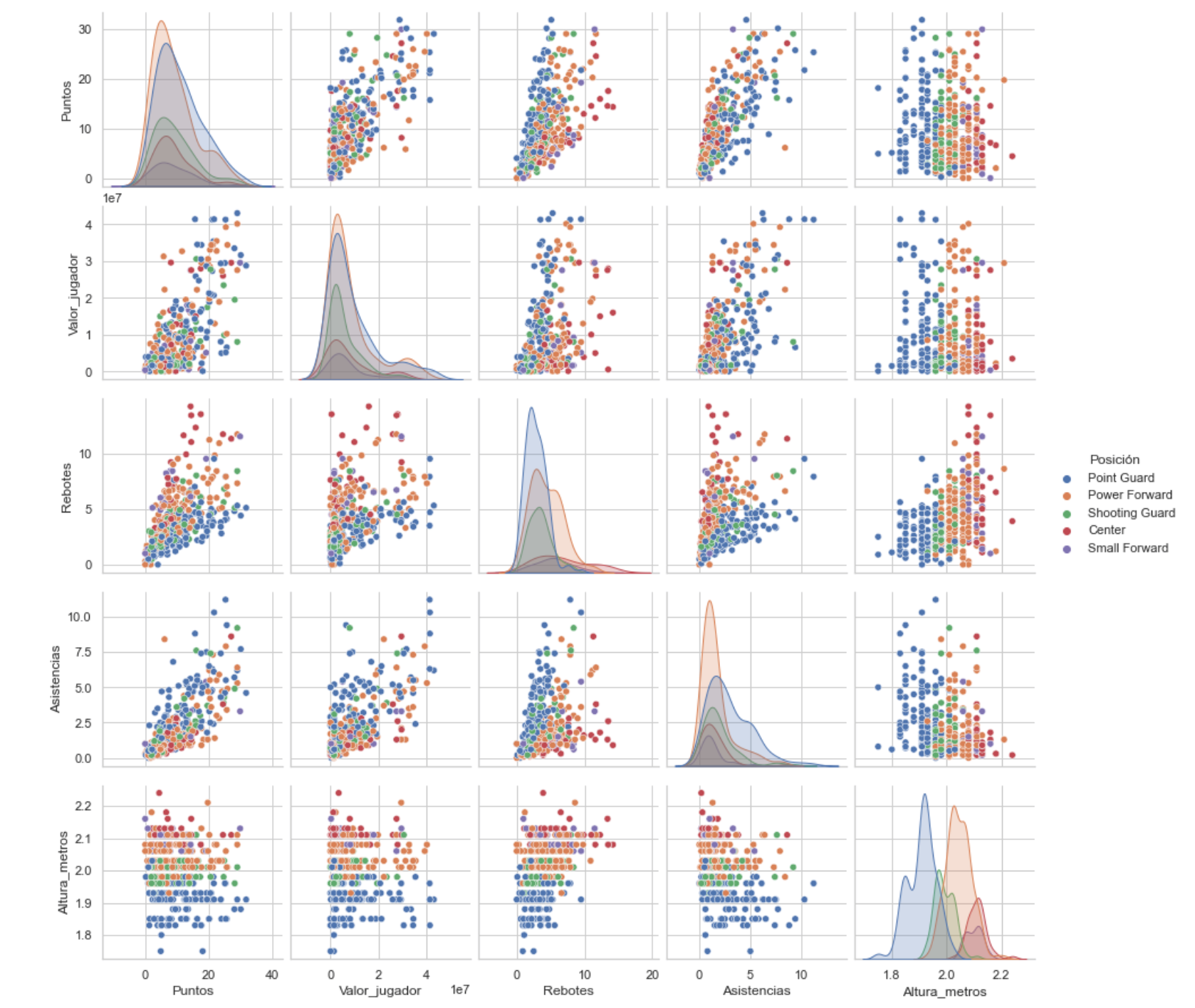
En este caso el R2 es 0,56, lo que significa que un 56% de la variación total de los valores de Y (asistencias) son explicados por el cambio en los valores de X (puntos).

Podríamos concluir que los jugadores que más se esfuerzan por anotar puntos, más pases de asistencia realizan a sus compañeros.

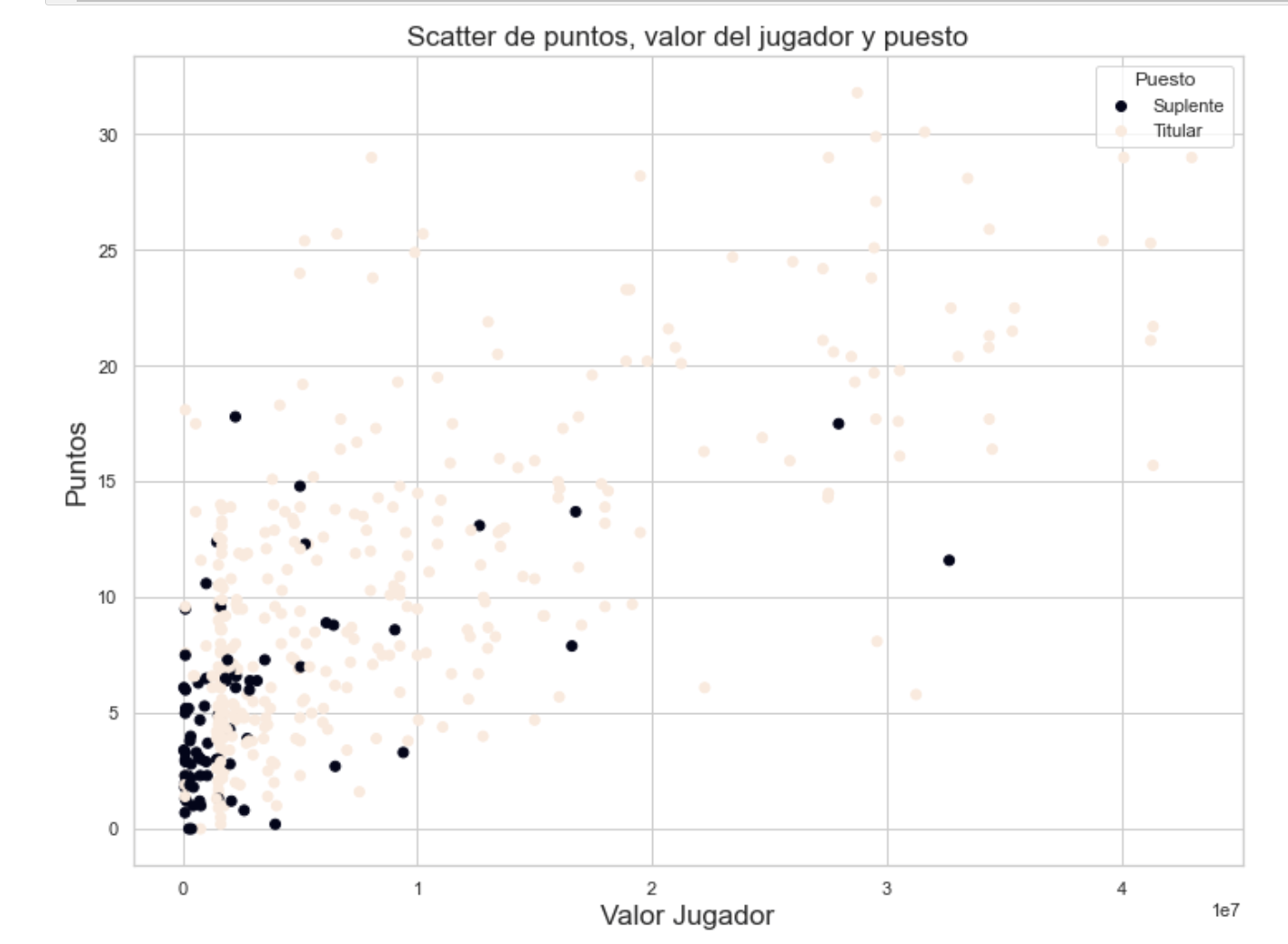
Esto se puede deber a que los jugadores estrella del equipo son los que más contacto tienen con el balón, teniendo la oportunidad de anotar puntos y asistir en muchas ocaciones a sus compañeros.

Análisis Multivariado

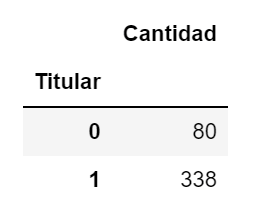
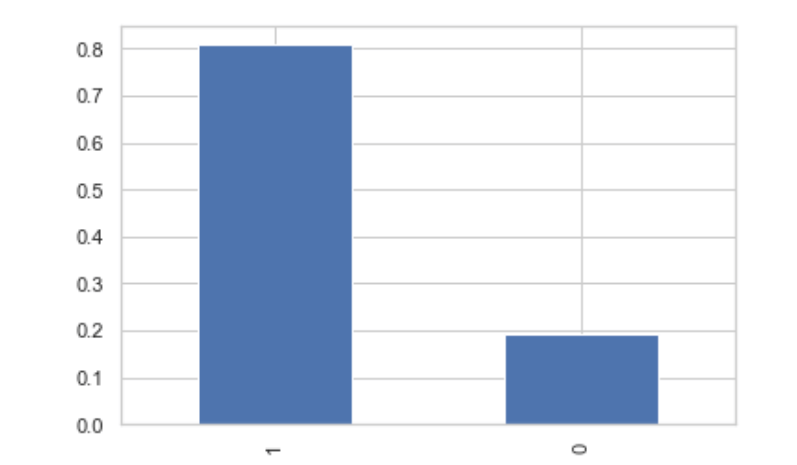
Luego de transitar la fase de data wrangling y EDA, exploramos las relaciones entre múltiples variables y descubrimos que existen distintas características de los jugadores que inducen a pensar que se pueden formar grupos peculiares en función de distintas características.



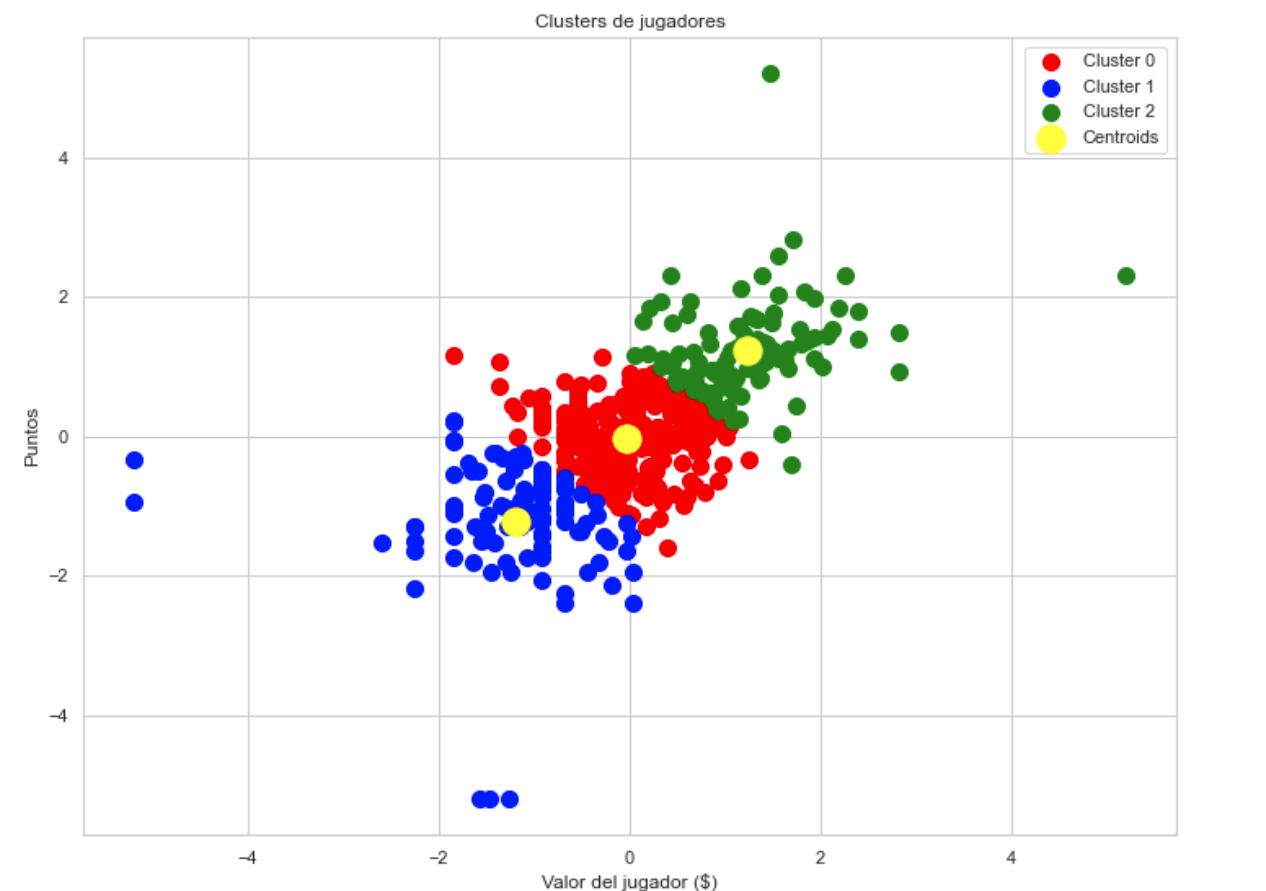
Como paso siguiente, implementamos un gráfico para visualizar la distribución de los jugadores suplentes y titulares en función de su valor de mercado y el promedio de puntos convertidos. Interpretamos que existen tendencias a explorar en las distribuciones del siguiente gráfico:



Asimismo, notamos que las clases relativas al puesto están desbalanceadas en nuestro dataset:



Como paso siguiente, procedimos a normalizar los valores de las variables valor jugador y puntos para obtener un gráfico de clusters k-means:

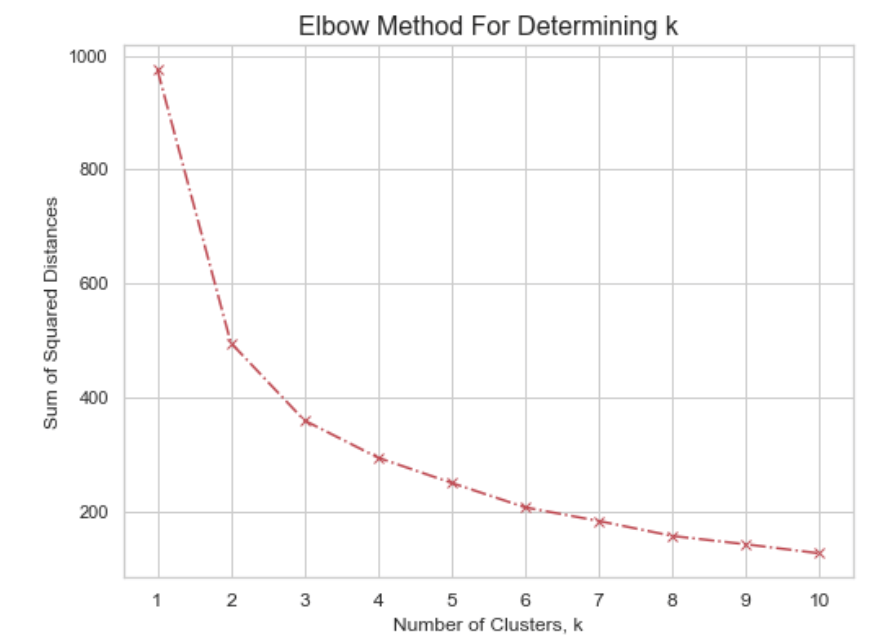


Podemos ver que existen tres grupos de jugadores. Estos son los azules, rojos y los verdes. Teniendo en cuenta que trabajamos con datos normalizados, podemos observar que los azules tienen un promedio de puntos muy disperso y un valor de mercado bajo (se deduce que muchos son los jugadores suplentes o recién iniciados de los equipos). Por otro lado, el grupo de los verdes también encuentra disperso en el plano, donde encontramos que muchos jugadores tienen similares promedios de puntos que los azules pero una mayor valoración monetaria en el mercado (deducimos que son las estrellas de la NBA). Por último tenemos a los jugadores promedio de la temporada (grupo de los rojos), los cuales tienen un buen promedio de puntos y una valoración de mercado promedio en relación al resto.

Como una primera reflexión de los grupos, entendemos que, además del promedio de puntos en la temporada, existen otros atributos que hacen que la valoración del jugador en el mercado sea más elevada. Estos factores pueden ser la condición del jugador (titular o suplente, rookie o experimentado, MVP de la temporada, la posición del jugador, entre otros).

Asimismo, la diferencia de puntos para los jugadores del grupo verde se puede deber a que pueden existir jugadores estrella que se lesionaron en la temporada, por lo que aportaron un menor promedio de puntos. Otra explicación que pensamos se puede relacionar al promedio de tiempo de juego del jugador.

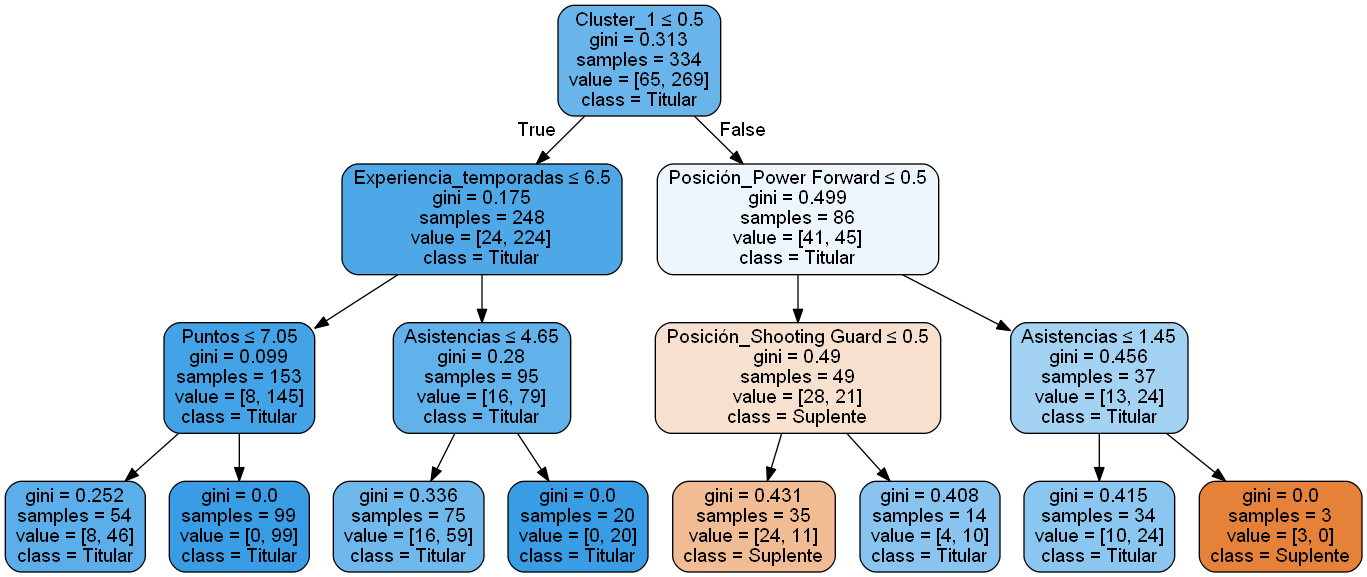
Seguidamente, implementaremos el esquema del método del codo (Elbow Method) para saber si acertamos al determinar un número óptimo de 𝑘. Entendemos que fue correcta nuestra suposición inicial al elegir un número de 𝑘=3.

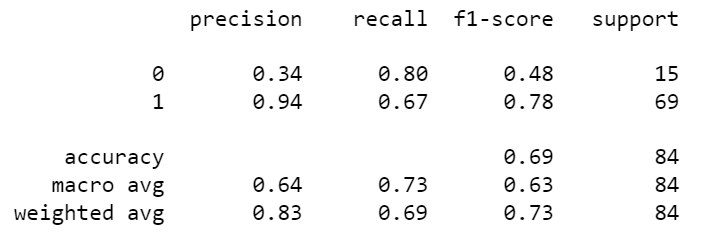


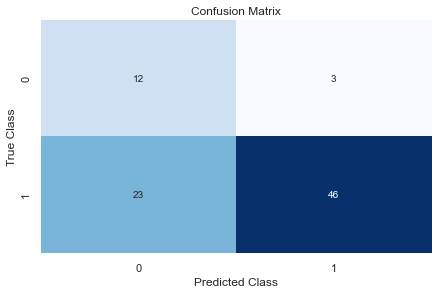
## Algoritmos elegidos y métricas de desempeño

Para enfrentar el desafío de descubrir cuáles son las características que más influyen al momento de predecir la condición de titular o suplente del player, implementamos los siguientes modelos de clasificación supervisada con técnicas de balanceo, oversampling, cross validation e hyperparameter tuning:

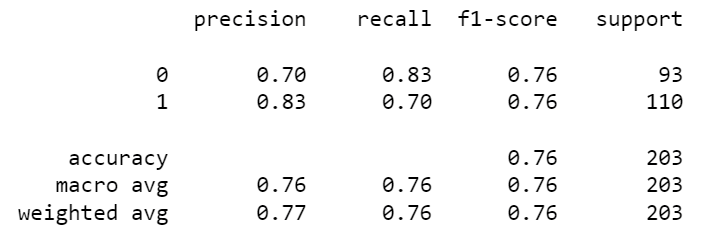
1. **Decision Tree con balanceo de clases**

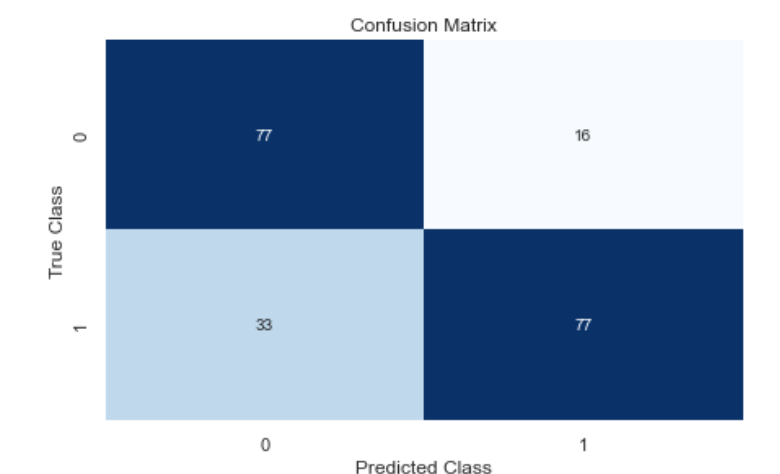
****

****

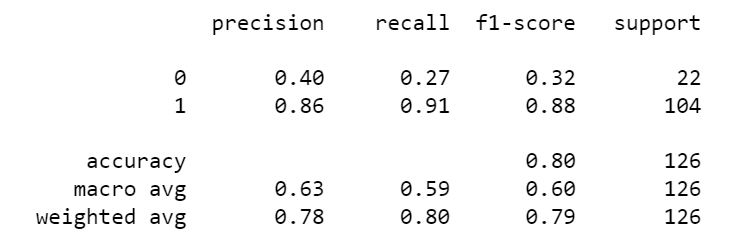
****

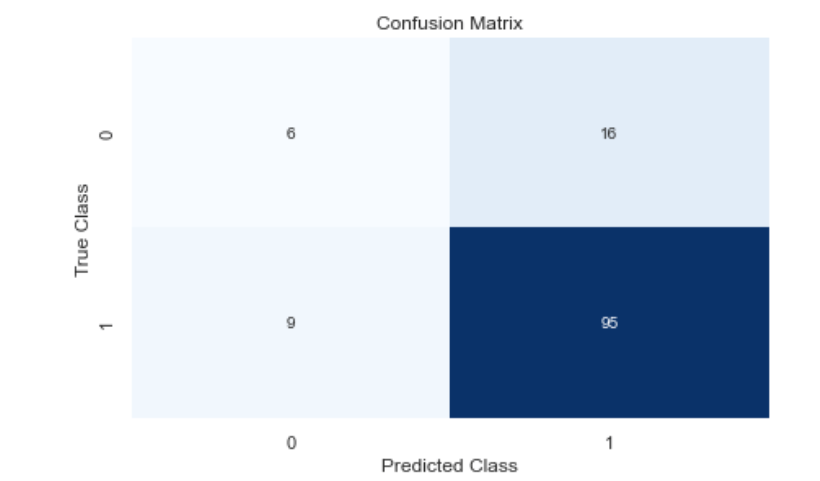
1. **Logistic Regression con balanceo de clases**

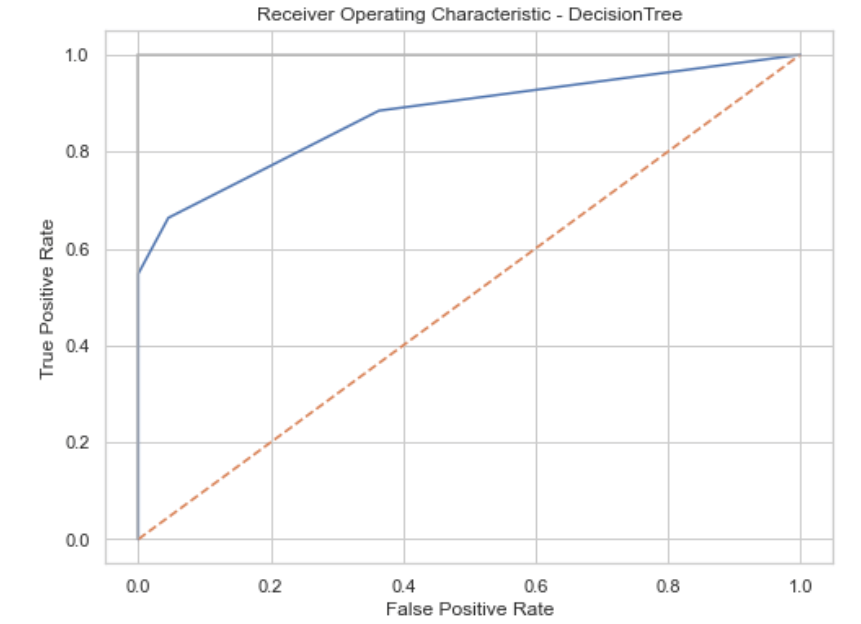
****

****

1. **Random Forest con balanceo de clases**

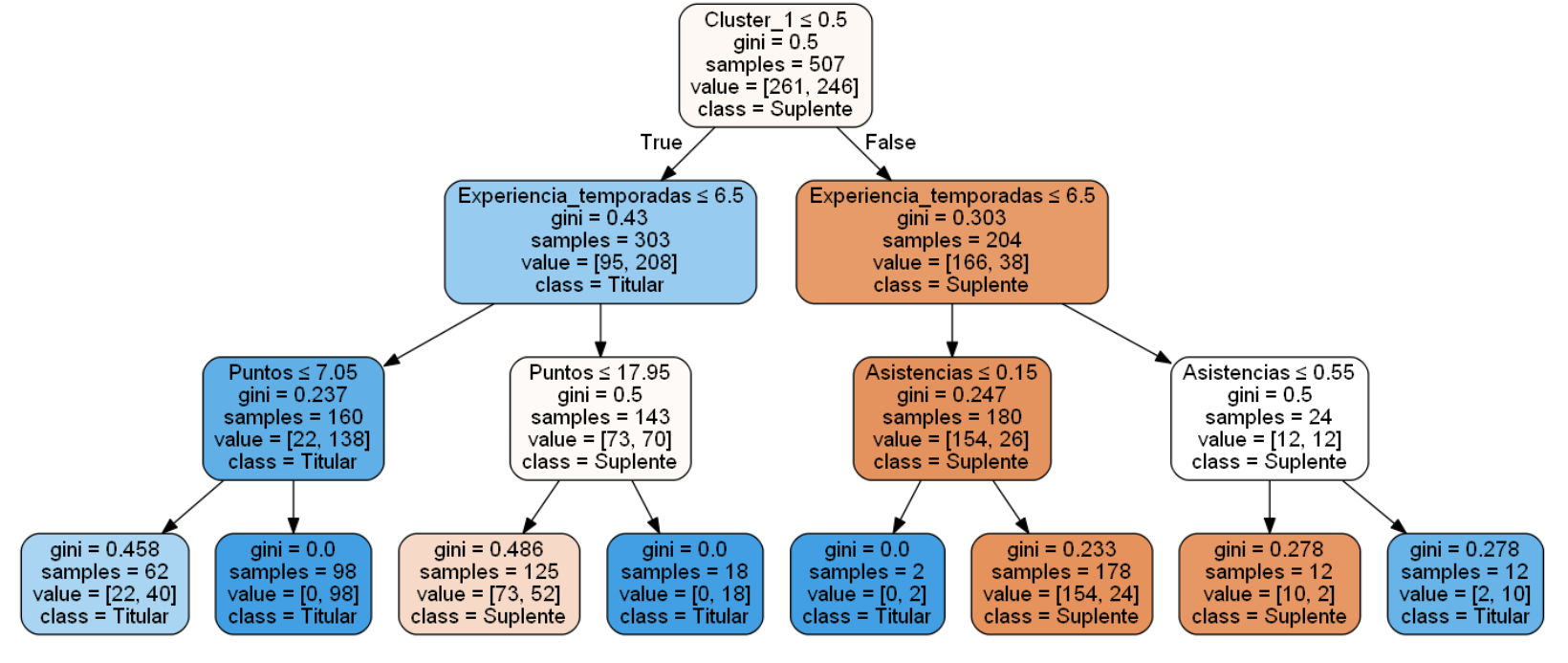
****

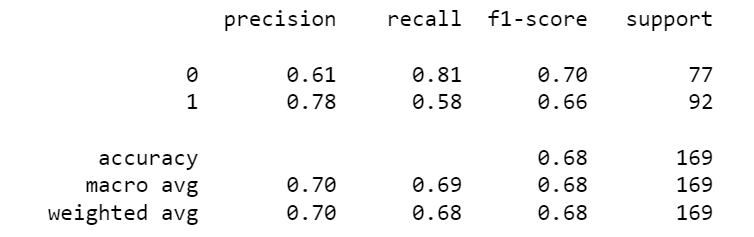
****

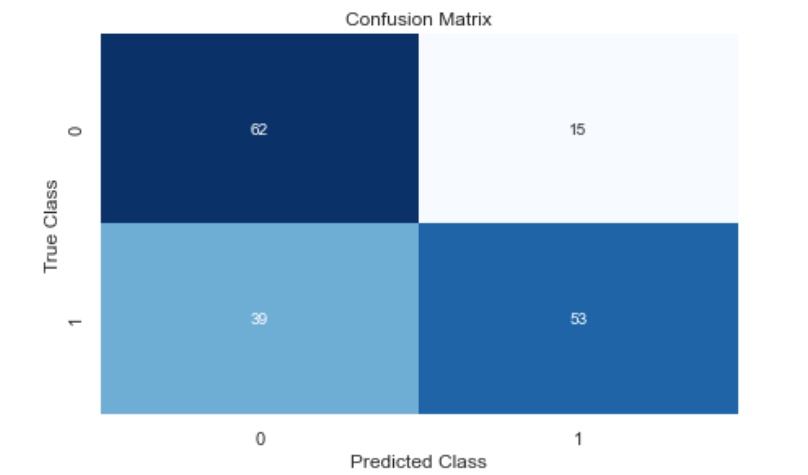


## Iteraciones de Optimización y métricas finales del Modelo Optimizado

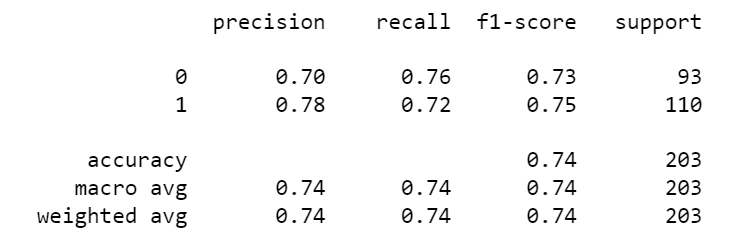
1. **Decision Tree whit oversampled**

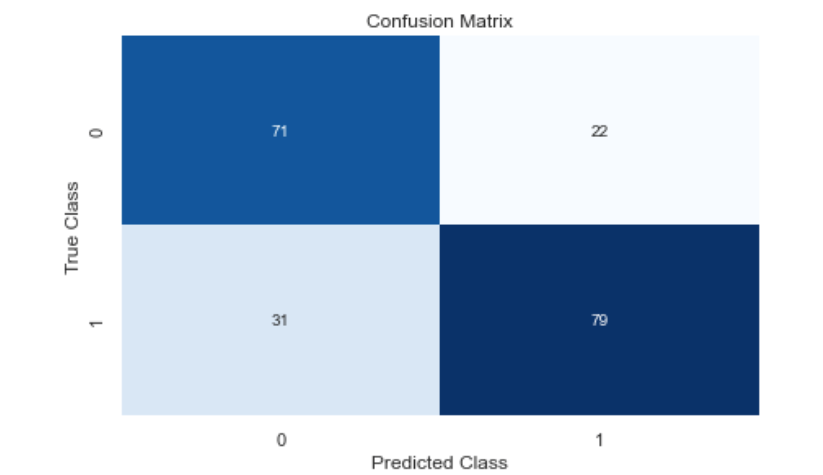
****

****

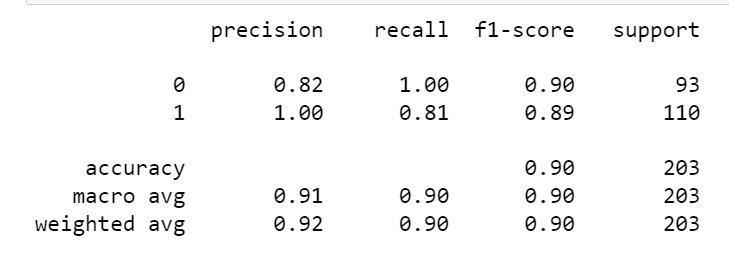
****

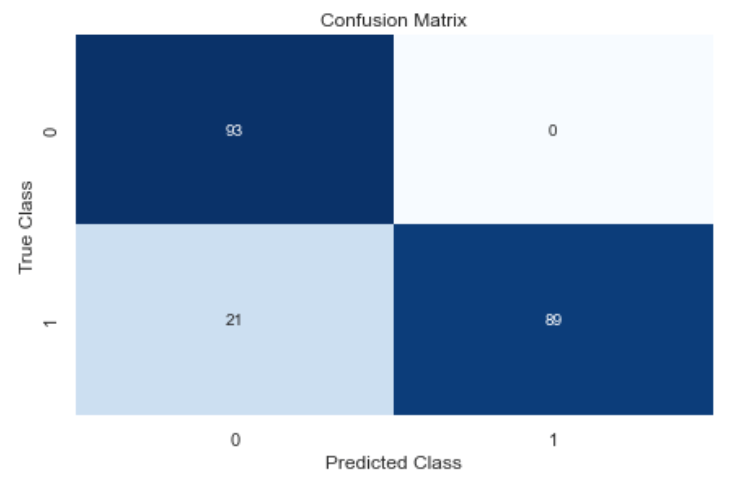
1. **Logistic Regression whit oversampled**

****

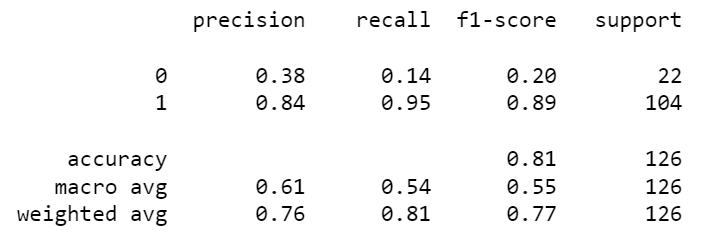
****

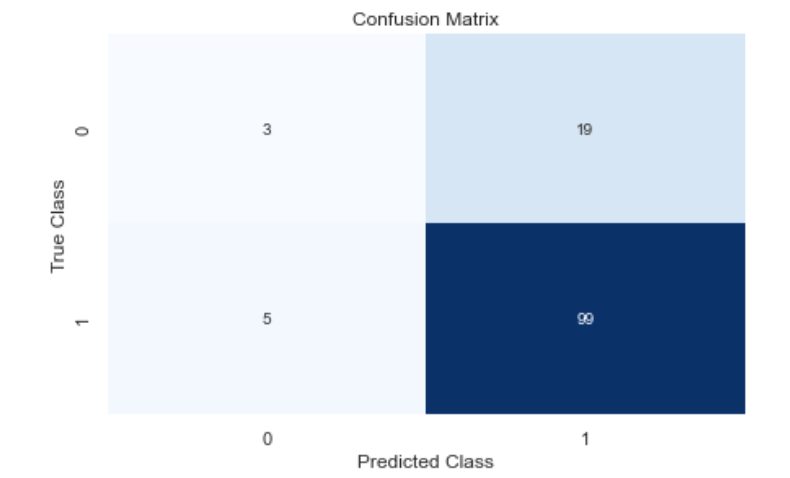
1. **Random Forest whit oversampled**

****

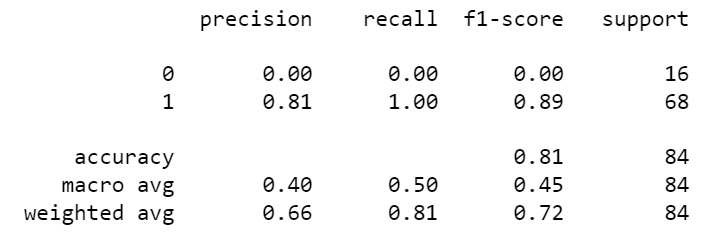
****

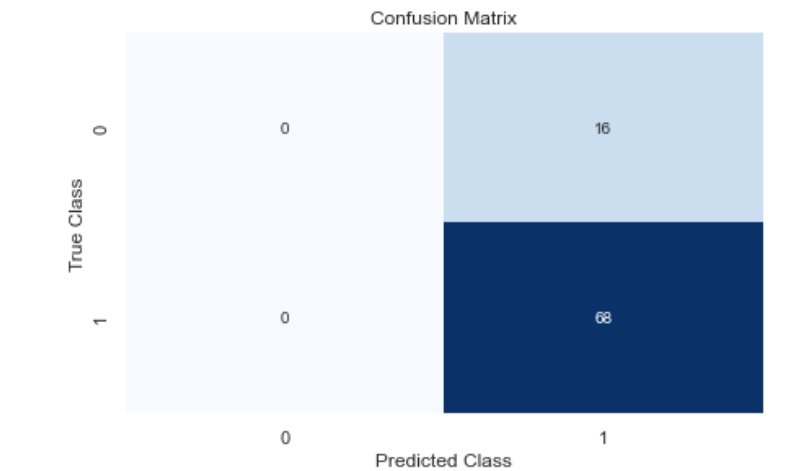
1. **Randomized Search CV aplicado a KNeighborsClassifier**

****

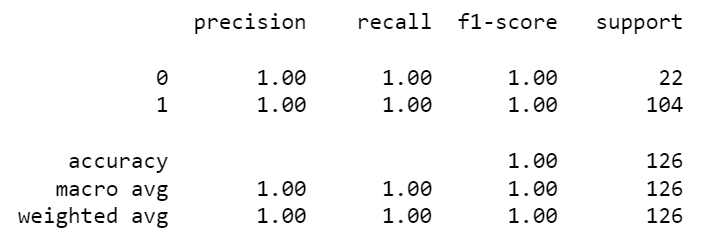
****

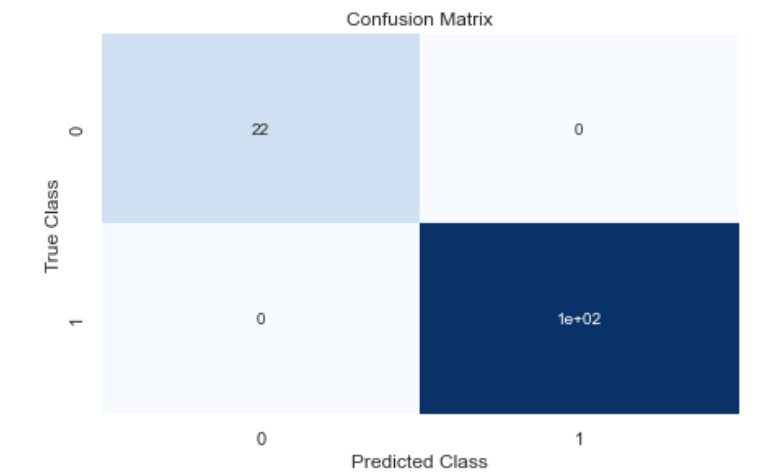
1. **Grid Search CV aplicado a Stratified K-Fold (SKF)**

****

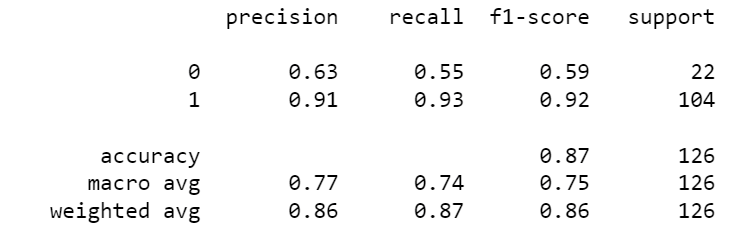
****

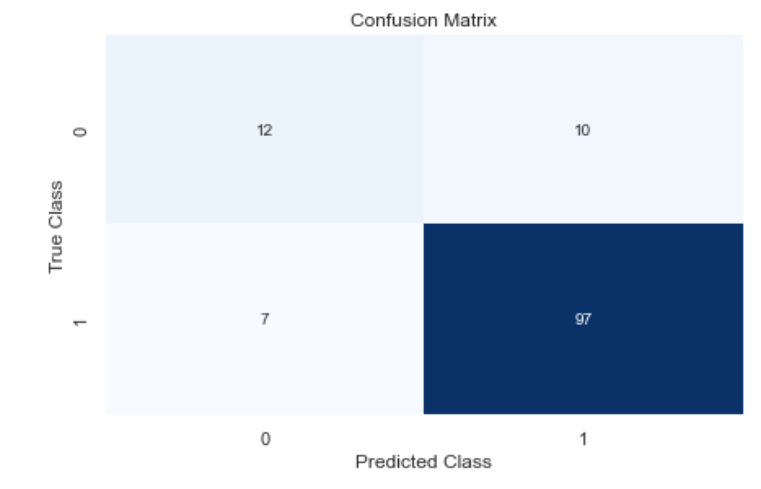
1. **XGBClassifier aplicado a Logistic Regressiony múltiples árboles.**

****

****

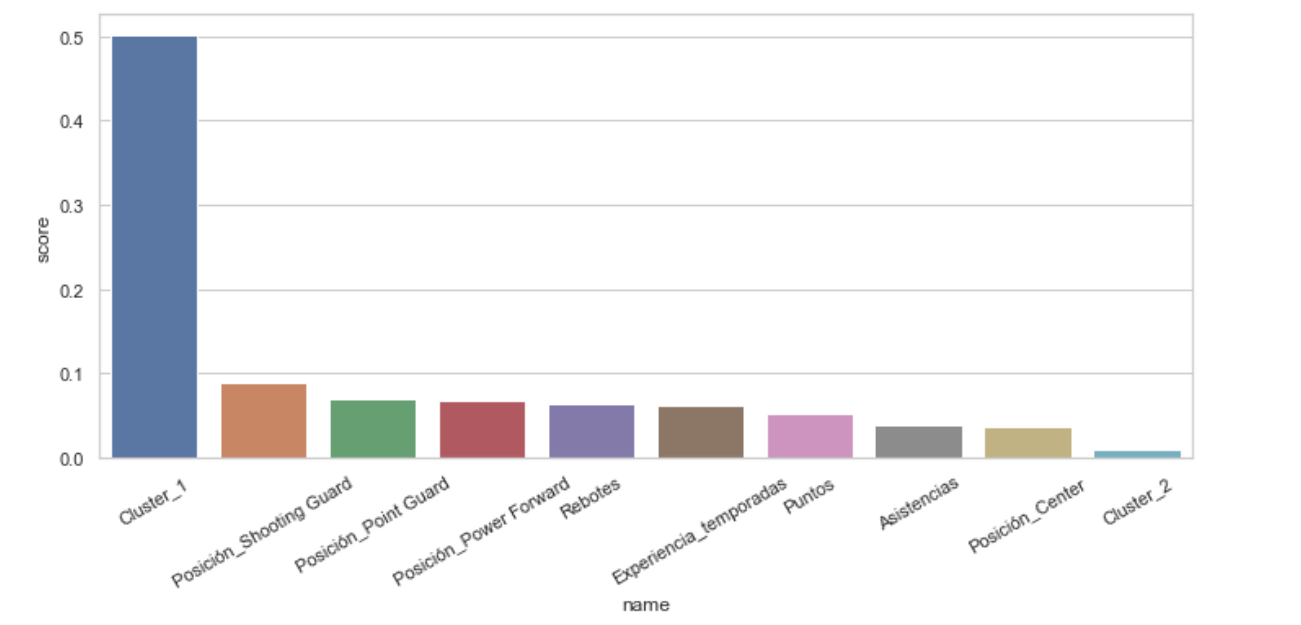
1. **XGBClassifier aplicado a múltiples árboles.**

****

****

**Variables con mayor peso en el modelo:**

****



## Futuras líneas

Como posible mejora, entendemos la necesidad de contar a futuro con un dataset con mayores observaciones al momento de implementar nuestros modelos de ML. Asimismo, esto nos permitirá obtener estadísticas más robustas y hacer comparaciones inter-temporadas.

## Conclusiones

El desafío con este proyecto fue implementar modelos de ML con un dataset limpio de 418 observaciones y con clases desbalanceadas para el caso de nuestra variable target. Entendíamos que los resultados podrían tender al overfitting y, por lo tanto, a no generalizar correctamente.

Por lo tanto, tomamos como principal criterio entender que puede ser muy probable que los mejores resultados, en términos de performance, provengan de modelos más simples o del ajuste de hiper parámetros.

Para ello, fue necesario balancear nuestro dataset y comparar los resultados obtenidos.

Finalmente nos quedamos

Podemos concluir que, si bien los modelos avanzados como el XGBOOST nos brindan métricas de accuracy más altas y una mejor sensibilidad (TPR), priorizaremos tener un modelo que no tienda al overfitting y que nos brinde una métrica de especificidad los más óptima posible.

Por lo tanto, podemos concluir que el modelo de Decision Tree con las clases de la variable target balanceadas es la solución más acorde a nuestro problema de clasificación.

El cluster 1 es la variable que mejor segmenta la condición de titular o suplente para el caso de los dos modelos. Esto confirma nuestra suposición respecto a que los jugadores con menor promedio de puntos y un valor de mercado bajo (posibles rookies) son más propensos a ser suplentes. Asimismo, la posición dentro del campo de juego, los rebotes, la experiencia y las asistencias también tienen un peso a considerar.