Facultade de Informática
UNIVERSIDADE DA CORUÑA

TRABALLO FIN DE GRAO
GRAO EN CIENCIA E ENXEÑARÍA DE DATOS

Aplicación de técnicas estatísticas avanzadas ao baloncesto

Estudante: Uxío Francisco Merino Currás

Dirección: Ricardo José Cao Abad

Francisco Camba Rodríguez

A Coruña, xullo de 2024.

Dedicatoria

Agradecementos

Agradecementos

Resumo

A idea deste traballo é mostrar a utilidade da aplicación de técnicas estatísticas no ámbito do baloncesto, usando diferentes ferramentas para dar solución ás situacións presentadas polo Obradoiro CAB, equipo profesional de baloncesto, a través do seu analista de datos. Todas as bases de datos utilizadas foron cedidas polo propio clube, e conteñen grandes cantidades de información da Liga ACB da tempada actual (2023/24) e a anterior (2022/23), tanto para xogadores (individualmente e agrupados por quintetos en xogo) como para equipos (variables de rendemento dos equipos en cada partido).

Partindo destas bases de datos, aplicamos distintas técnicas, cada unha con obxectivos distintos. A nivel individual dos xogadores, aplicaremos un ACP e técnicas de clustering, para resumir a variabilidade dos datos en menos compoñentes, facilitar a representación e etiquetar aos xogadores polo seu estilo de xogo de maneira obxectiva. Para poder optimizar o rendemento, construímos modelos de regresión (lineais e de Machine Learning) cos datos a nivel de equipo e de quintetos para buscar variables relevantes e configuracións de xogadores que dean mellores resultados

O deporte é unha rama máis na que o Big Data é decisivo para a toma de decisións estratéxicas e poder obter unha vantaxe competitiva con respecto aos rivais, o que se trata de mostrar neste traballo. O obxectivo global era poder obter unha ferramenta que levase á optimización de xogadores e do seu rendemento, e os resultados obtidos demostran que a aplicación destes métodos axudan en gran medida a lograr isto, transformando os datos en vantaxes.

Abstract

The idea of this work is to demonstrate the usefulness of applying statistical techniques in the field of basketball, using different tools to address the situations presented by Obradoiro CAB, a professional basketball team, through its data analyst. All databases used were provided by the club itself and contain large amounts of information from the ACB League for the current season (2023/24) and the previous one (2022/23), both for players (individually and grouped by playing lineups) and for teams (team performance variables in each game).

Starting with these databases, various techniques were applied, each with different objectives. At the individual player level, PCA and clustering techniques were used to summarize data variability into fewer components, facilitate representation, and objectively label players by their playing style. To optimize performance, regression models were built(linear and Machine Learning) with team and lineup level data to identify relevant variables and player configurations that yield better results.

Sports is another field where Big Data is crucial for strategic decision-making and gaining a competitive edge over rivals, which is what this work aims to demonstrate. The overall objective was to develop a tool that would lead to team optimization and performance improvement. The results obtained show that the application of these methods significantly helps achieve this, transforming data into advantages.

Palabras chave:

- Técnicas estatísticas
- Análise de datos
- Análise de Compoñentes Principais
- Clustering
- Modelos de regresión
- Random Forest
- Aprendizaxe Automático

Keywords:

- Statistical techniques
- Data analysis
- Principal Component Analysis
- Clustering
- Regression models
- Random Forest
- Machine Learning

Índice Xeral

1 Introdución **1**

1.1 Baloncesto, ACB e Datos . . . 1

1.2 Obxectivos . . . 3

2 Fundamentos Teóricos **4**

2.1 Análise de Compoñentes Principais (PCA) . . . 4

2.1.1 Obtención das compoñentes . . . 4

2.1.2 Utilidade . . . 5

2.2 Técnicas Clustering . . . 5

2.2.1 Clustering xerárquico . . . 6

2.2.2 Distancias . . . 6

2.2.3 Métodos de Encadeamento . . . 7

2.3 Regresión . . . 8

2.3.1 Regresión Lineal . . . 8

2.3.2 Random Forest . . . 9

2.3.3 Avaliación . . . 10

2.3.4 Selección de variables . . . 11

3 Análise realizado **13**

3.1 Preprocesado . . . 13

3.1.1 Xogadores individuais . . . 13

3.1.2 Partidos . . . 15

3.1.3 Agrupación por quintetos . . . 17

3.2 PCA . . . 18

3.2.1 Significado das compoñentes . . . 19

3.2.2 Utilidades . . . 23

3.3 Clustering realizado . . . 26

3.3.1 Distancia de Mahalanobis . . . 27

3.3.2 Distancia Euclídea 31
3.3.3 Significación dos clusters 34
3.4 Regresión 38
3.4.1 Modelo lineal 38
3.4.2 Random Forest 41

4 Conclusións **49**

A Material adicional **51**

Índice de Figuras

3.1 Histograma para a variable 'Puntos por partido' 14
3.2 Boxplot para a variable 'Puntos por partido' 14
3.3 Histograma para a variable 'Resultado' 17
3.4 Boxplot para a variable 'Resultado' 17
3.5 Porcentaxe de varianza explicada por cada compoñente 18
3.6 Coeficientes da Primeira Compoñente 19
3.7 Coeficientes da Segunda Compoñente 21
3.8 Coeficientes da Terceira Compoñente 22
3.9 Individuos representados sobre as dúas primeiras compoñentes 23
3.10 Individuos representados sobre as tres primeiras compoñentes 24
3.11 Mostra do gráfico interactivo 24
3.12 Mostra de resultado de cálculo de distancias 25
3.13 Resultado da procura de xogador máis próximo 25
3.14 Resultado da procura de xogador próximo con filtro de equipo 25
3.15 Matriz de distancias dos xogadores do Obradoiro CAB 26
3.16 Dendograma xerado pola distancia de Mahalanobis e encadeamento tipo centroide 27
3.17 Dendograma xerado pola distancia de Mahalanobis e encadeamento tipo Ward 28
3.18 Dendograma xerado pola distancia de Mahalanobis e encadeamento tipo Ward coloreado 28
3.19 Individuos segundo clúster nas 2 primeiras compoñentes principais 29
3.20 Individuos segundo clúster nas 2 primeiras compoñentes principais 30
3.21 Individuos segundo clúster nas 3 primeiras compoñentes principais 30
3.22 Dendograma xerado pola distancia euclídea e encadeamento tipo Ward.2 . . . 31
3.23 Dendograma xerado pola distancia euclídea e encadeamento tipo Ward.2 coloreado 32
3.24 Individuos segundo clústers nas 2 primeiras compoñentes principais 33

3.25 Individuos segundo clústers nas 2 primeiras compoñentes principais 33
3.26 Individuos segundo clústers nas 3 primeiras compoñentes principais 34
3.27 Área ocupada por cada clúster nas 2 primeiras compoñentes principais 35
3.28 Área ocupada por cada clúster na primeira e terceira compoñentes principais . 35
3.29 Coeficientes do modelo tras selección de variables 39
3.30 Variables con alta multicolinealidade 40
3.31 Modelo lineal final 40
3.32 Histograma da variable resposta 42
3.33 Histograma da variable resposta transformada 43
3.34 Observacións vs Predicións (diferenza total como resposta) 45
3.35 Observacións vs Predicións (diferenza por minuto como resposta) 46
3.36 Situación do quinteto con maior predición no histograma 47
3.37 Situación do quinteto con maior predición e configuración non probada no histograma 48

Índice de Táboas

3.1 Comparación da diferenza media de puntos cos resultados finais da Liga ACB 16
3.2 Xogadores con puntuación positiva e negativa na primeira compoñente 20
3.3 Xogadores con puntuación positiva e negativa na segunda compoñente 21
3.4 Xogadores con puntuación positiva e negativa na terceira compoñente 22

Capítulo 1

Introdución

No deporte profesional, a competitividade está á orde do día. Trátase dun mundo que move grandes cantidades de recursos económicos e unha masa social enorme, ao alcance de practicamente ningunha outra industria, onde as vitorias e derrotas resultan cruciais para afeccionados, contratos televisivos, patrocinadores...

Estando ademais nunha era onde a información e os datos son utilizados a diario en calquera empresa para optimizar os seus procesos e maximizar os beneficios, é lóxico pensar que estes dous mundos estaban destinados a atoparse. Neste contexto, o Big Data e a análise de datos emerxen como ferramentas fundamentais para obter un beneficio no rendemento deportivo. A través da análise detallada de datos, os equipos poden tomar decisións máis informadas e estratéxicas, tanto no terreo de xogo coma fóra del.

Un exemplo claro deste uso é o ámbito do baloncesto profesional. Neste traballo, realizado en colaboración co equipo Obradoiro CAB, buscaremos aplicar diversas técnicas estatísticas para mellorar o rendemento do equipo. A análise de datos non só permite avaliar o rendemento pasado, senón tamén predicir tendencias e formular estratexias futuras. A dispoñibilidade de datos detallados proporcionados polo equipo sobre os xogadores e os partidos ofrece unha base sólida para estas análises, que se poden traducir en melloras significativas no campo de xogo.

En resumo, este traballo pretende mostrar como a análise de datos pode transformar o xeito no que se toma decisións no baloncesto profesional, ofrecendo unha vantaxe aos equipos baseada en datos obxectivos e análises.

1.1 Baloncesto, ACB e Datos

O baloncesto é un deporte altamente popular no mundo enteiro, pero imos explicar uns conceptos básicos para favorecer a comprensión do que se está tratando de acadar neste traballo.

Trátase dun xogo onde 5 xogadores de cada equipo compiten durante 4 cuartos de 10 minutos para tratar de anotar máis puntos ca o rival. Non obstante, estes 5 xogadores non teñen por que ser os mesmos durante a duración total do partido. Un equipo pode presentar ata 12 xogadores para cada partido, e existen cambios ilimitados para favorecer o descanso e as diferentes estratexias que decida probar o adestrador. Incluso un equipo pode ter máis xogadores contratados dos 12 límite para un partido e facer unha selección chegado o momento do mesmo. Por todo isto, cobra especial importancia coñecer, de todos os xogadores dispoñibles nun equipo, qué combinación de 5 xogadores é a mellor.

Sabendo que hai 5 xogadores en todo momento na pista, no baloncesto tradicional fixouse unha posición para cada un deles, coas súas tarefas concretas asociadas. Estas posicións son as de base, escolta, alero, ala-pívot e pívot; e están moi relacionadas coa estatura dos xogadores. Así, tradicionalmente os xogadores máis baixos eran clasificados como bases ou escoltas, e debían saber botar, pasar e tirar dende lonxe; mentres que os xogadores máis altos eran clasificados como ala-pívots ou pívots e a súa responsabilidade estaba en xogar cerca do aro, poñer tapóns aos rivais e coller rebotes. Pero como todo, o baloncesto foise modernizando, e cada vez existen máis xogadores altos que xogan por fóra ou pequenos que lanzan máis que pasan. Estes novos tempos obrigan a desbotar (nunca por completo) as posicións tradicionais e tratar de buscar novas etiquetas para os xogadores.

Como se mencionou anteriormente, a base para gañar un partido é anotar máis puntos có teu rival. Pero, na práctica, para saír vitorioso entran en xogo moitas máis accións. Dende as máis sinxelas como rebotes (coller o balón despois dun tiro a canastra) ou roubos (recuperar o balón que estaba en posesión do rival) ata algunhas máis complexas coma os puntos tras rebote ofensivo (cantos puntos anota o teu rival tras capturar o rebote ofensivo, é dicir, do seu propio tiro). Á hora de preparar un partido, é primordial saber en cales de todas estas facetas centrarse e cales non son realmente tan importantes.

No noso país, a competición de baloncesto máis importante é a Liga ACB (Asociación de Clubs de Baloncesto), a primeira división e a única profesional en España. É considerada a liga nacional máis importante de Europa e a segunda do mundo, só por detrás da NBA. Nela compiten 18 equipos, dos cales 4 participan na Euroliga, a máxima competición europea de baloncesto e a máis importante do mundo (de novo só por detrás da NBA): Real Madrid, FC Barcelona, Baskonia e Valencia Basket.

Con todo este contexto, está claro que a Liga ACB é un referente mundial e un entorno propicio para innovar en aspectos como o uso dos datos. Precisamente, conscientes do novo mundo centrado en datos, é a propia Liga ACB a que proporciona todos os datos aos equipos a través dun acordo cunha empresa tecnolóxica de Reino Unido. Esta empresa nutre de datos exclusivos aos clubs a través dunha API, así como outros datos en directo durante os partidos a medios de comunicación e afeccionados.

Non obstante, o baloncesto sempre foi un deporte que prestou especial atención aos datos, e as estatísticas levan moitos anos no día a día deste mundo: o número de puntos, rebotes e asistencias por partido dun xogador é algo que leva instaurado moito tempo tanto na prensa coma entre os afeccionados á hora de falar de xogadores. A revolución chega, como en tantos outros campos, polo volume da información.

Os datos proporcionados pola liga inclúen 36 variables de rendemento por xogador, así como para cada combinación de 5 xogadores que probou cada equipo, que completan a información das estatísticas máis "tradicionais". Por exemplo, é moi importante ter en conta non só os rebotes por partido, senón a porcentaxe dos rebotes que colle ese xogador dos dispoñibles mentres está en pista. A información dispoñible é completa; o reto está en saber aproveitala ben.

1.2 Obxectivos

Esta grande cantidade de datos que ofrece cada partido de baloncesto fai que os equipos busquen neles solucións aos seus problemas e situacións concretas. Existen datos individualizados, por grupos de xogadores (en parellas, tríos, cuartetos e quintetos), por equipos, por partido... e este abano de datos abre un gran abano de posibilidades.

Neste caso, ao fixar o proxecto, o Obradoiro CAB a través do seu analista de datos mostrou o interese en centrar esta análise en atopar resposta ás preguntas propostas anteriormente: Cal é a mellor combinación de xogadores dispoñibles? Que variables son relevantes para lograr a vitoria?. Polo tanto, buscáronse as áreas de coñecemento da Ciencia de Datos que axuden a extraer esta información partindo das bases de datos das que dispoñemos, concluíndo que a regresión a través de modelos lineais e de Machine Learning era a mellor resposta.

É tamén altamente relevante obter conclusións claras e comprensibles destas análises, xa que se trata dunha aplicación a un campo real no que non todo o mundo ten que ter coñecementos de Estatística ou Machine Learning, sendo chave poder explicar a información de maneira sinxela e directa.

Capítulo 2

Fundamentos Teóricos

TODAS as ferramentas utilizadas están baseadas nunha serie de conceptos teóricos, que serán descritos e explicados un a un ao longo deste capítulo para poder comprender máis adiante todo o proceso realizado e as aplicacións destes.

2.1 Análise de Compoñentes Principais (PCA)

A idea básica do Análise de Compoñentes Principais consiste en obter combinacións lineais das variables aleatorias, de tal maneira que estas expliquen a maior cantidade de variabilidade posíbel dos datos sobre os que se aplica. É dicir, aplica unha redución da dimensión para resumir toda a información posíbel do vector aleatorio p-dimensional orixinal nun novo vector de menor número de compoñentes, todas incorreladas entre si.

2.1.1 Obtención das compoñentes

As compoñentes principais obtéñense en orde, segundo a cantidade de varianza que explican. De esta maneira, a primeira compoñente será a que maior información conteña e a que se obtén nunha primeira etapa; a continuación, obtense a segunda combinación lineal das variables con maior varianza, que será a segunda compoñente, e así de maneira iterativa.

A i-ésima compoñente principal obtense aplicando sobre o vector aleatorio orixinal de variables os coeficientes do autovector da matriz de varianzas-covarianzas (Σ) dos datos asociado ao autovalor i-ésimo (ordeados de maior a menor). De esta maneira, a primeira compoñente principal sería o resultado de aplicar o autovector de Σ asociado ao maior autovalor.

De feito, a porción de variabilidade total explicada pola compoñente i-ésima corresponde co autovalor i-ésimo partido do sumatorio total dos autovalores:

$$\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}$$

2.1.2 Utilidade

Como xa comentamos ao inicio da sección, unha das principais utilidades desta análise é a redución da dimensión dos datos iniciais, permitindo traballar cun número de compoñentes máis manexable.

A redución do número de compoñentes mantendo unha porcentaxe importante da información otorga a posibilidade de realizar visualizacións gráficas en dúas ou tres dimensións utilizando as primeiras compoñentes principais. Isto axuda a identificar patróns e poder ubicar os datos no espazo.

Ademáis, esta ferramenta identifica aquelas variables irrelevantes en canto á variabilidade, ás que outorga pouco peso nos seus coeficientes e poden chegar a quedar practicamente excluídas.

Ao transformar as variables orixinais en compoñentes principais, podemos identificar patróns ocultos e relacións entre as variables que non serían evidentes a partir dunha análise directa das variables orixinais, e podemos observar os efectos conxuntos destas que expliquen a variabilidade.

Finalmente, as compoñentes principais tamén poden servir de utilidade para outros algoritmos ou técnicas estatísticas, ao aplicalos directamente sobre os datos transformados. Isto pode mellorar o seu rendemento por moitos motivos, como a aceleración da converxencia debido á redución da dimensión ou menor risco de sobreaxuste pola eliminación de variables irrelevantes.

2.2 Técnicas Clustering

Coas de técnicas clustering estamos a falar de clasificación non supervisada. A clasificación non supervisada consiste na creación de grupos dentro dunha poboación de maneira que os individuos sexan similares entre si dentro de cada grupo pero heteroxéneos entre grupos.

Existen dous principais métodos de clasificación non supervisada: métodos xerárquicos e non xerárquicos.

Os métodos xerárquicos non crean unha única agrupación da poboación, senón que agrupa seguindo unha xerarquía de particións. Así, este tipo de métodos non fixan un número de grupos a crear, senón que mostran como se formarían os grupos segundo o número que decidamos crear.

Pola súa parte, os métodos non xerárquicos clasifican nun número fixo de grupos, K. Este valor pode ser escollido a priori, por exemplo por coñecemento sobre o significado real dos datos, ou elixido dentro do proceso do método non xerárquico realizado.

No noso caso, ao non coñecer de antemán o número de grupos a formar, decantámonos polos métodos xerárquicos. Ímonos centrar nestes para continuar cos fundamentos teóricos.

2.2.1 Clustering xerárquico

En todos os métodos xerárquicos, a partición inicial correspóndese cun grupo único que engloba a todos os individuos da poboación, e descende nas xerarquías ata alcanzar un grupo específico para cada individuo. Os métodos poden ser aglomerativos, se parten dun grupo por individuo e xuntan iterativamente grupos ata alcanzar o grupo único, ou divisivos, se comezan pola partición que engloba a todos os individuos e vai realizando divisións ata alcanzar tantos grupos coma individuos.

Para ambos tipos de métodos, o dendograma (tamén coñecido como árbore xerárquica) é a forma de representación de todas as particións posibles creadas polo método escollido. Consiste nunha árbore binaria, a cal se pode cortar a unha certa altura para obter os grupos formados. A altura representa a distancia á que están os grupos, polo que se unha rama inicialmente común bifurca a unha certa altura h significa que os dous novos grupos representados polas ramas bifurcadas están a esa distancia h.

Hai varias cousas a buscar nun dendograma. É importante nun dendograma un corte que de rango de movemento, é dicir, que cortando algo por enriba ou algo por abaixo desa altura as particións sexan as mesmas. Tamén é preferible que os cortes no dendograma devolvan grupos máis ou menos balanceados en número de individuos, e non algúns grupos practicamente baleiros fronte a outros moi numerosos.

2.2.2 Distancias

A base do problema é agrupar aos individuos de maneira que a distancia entre grupos sexa máxima e dentro deles mínima, polo que a maneira de calcular esa distancia é, claramente, unha decisión chave. As distancias máis utilizadas, e as que consideramos neste problema, son a euclídea e a de Mahalanobis.

A distancia euclídea correspóndese co concepto clásico de distancia como o entendería calquera persoa fora do ámbito matemático. É a percepción visual de distancia que observaríamos ao graficar os datos en dúas ou tres dimensións, pero de maneira xeneralizada para espazos de calquera tamaño. Ven dada pola fórmula:

$$d(\vec{x},\vec{y}) = \left(\sum_{i=1}^{p}(x_i - y_i)^2\right)^{1/2} = \left[(\vec{x}-\vec{y})^T(\vec{x}-\vec{y})\right]^{1/2}$$

A distancia de Mahalanobis inclúe na súa fórmula a matriz de varianzas-covarianzas, o que significa que ten en conta a estrutura de covarianzas dos datos cando calcula a distancia entre eles. A fórmula é a seguinte:

$$d_M(\vec{x},\vec{y}) = \left[(\vec{x}-\vec{y})^T\Sigma^{-1}(\vec{x}-\vec{y})\right]^{1/2}$$

O uso de distintas distancias leva a resultados totalmente diferentes, polo que a elección debe ser feita con cautela. Para iso, utilizaremos representacións visuais en varias dimensións (grazas ao PCA) dos datos por grupos e ao significado real dos mesmos, contando coa experiencia do titor empresarial e analista de datos do Obradoiro CAB, Francisco Camba.

2.2.3 Métodos de Encadeamento

Os métodos de encadeamento son os criterios para decidir que grupos fusionar (ou dividir, dependendo de se o enfoque é aglomerativo ou divisivo) en cada unha das etapas do algoritmo. Imos introducir unha breve explicación dos métodos probados:

- **Simple**: Este método fusiona grupos baseándose na menor distancia entre calquera par de puntos pertencentes a diferentes grupos. É dicir, mide a distancia entre os veciños máis próximos dos grupos, e fusiona os que presenten menor distancia.
- **Completo**: Como criterio da distancia entre grupos toma a distancia entre o par de observacións máis alonxadas (os veciños máis afastados).
- **Promedio**: Utiliza a distancia promedio entre as observacións de cada grupo.
- **Mediana**: Utiliza a mediana no lugar do promedio.
- **McQuitty**: Este método é recursivo, e baséase en que a distancia entre os dous clusters máis recentemente fusionados, A e B, e calquera outro cluster C é media das distancias de A a C e de B a C.
- **Centroide**: Fusiona grupos baseándose na distancia entre os centroides dos clusters. Normalmente estes centroides correspóndense coa media de cada cluster.
- **Ward**: Considera a unión de cada para de grupos, e fusiona aqueles que incrementen en menor medida a suma dos cadrados das desviacións entre cada punto e o centroide do cluster ao que pertence.

  Utilizamos tamén unha variación deste método de Ward, na que no lugar de utilizar o incremento na varianza utiliza o incremento na distancia euclídea ao cadrado dos puntos ao novo centroide resultante da fusión.

De igual maneira que coas distancias, o uso de distintos métodos de encadeamento leva a resultados distintos, polo que aplicaremos de novo visualizacións e a comprensión do seu significado real.

2.3 Regresión

A regresión é unha técnica estatística utilizada para analizar a relación entre unha ou máis variables independentes (tamén chamadas variables explicativas ou predictoras) e unha variable dependente (tamén chamada variable resposta). O obxectivo da regresión é comprender a natureza da relación entre estas variables e facer predicións ou inferencias sobre a variable dependente baseándose nas variables independentes.

Existen varios tipos de modelos de regresión, pero neste caso os utilizados foron os modelos de regresión lineal e modelos de Machine Learning para regresión.

2.3.1 Regresión Lineal

Os modelos de regresión lineal son aqueles que asumen unha relación lineal entre a resposta e as variables explicativas. O modelo matemático defínese como segue:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \tag{2.1}$$

Onde:

- β_0 é o coñecido como intercept, que corresponde co valor esperado da variable resposta se todas as explicativas tiveran valor cero.
- β_j é a taxa de cambio esperada en Y por incremento unitario en X_j cando X_r permanece constante para todo $r \neq j$.
- ε_i i.i.d. N(0, σ), i = 1, . . . , n; sendo σ a desviación estándar das respostas para un valor arbitrario de (X_1, X_2, . . . , X_k)

Por definición, os modelos de regresión lineal múltiple asumen as seguintes hipóteses estruturais:

- **Linealidade:** O valor esperado da variable resposta, dado o conxunto de variables explicativas, é unha combinación lineal desas variables.

$$E(\vec{Y}|\mathbf{X}) = m(\mathbf{X}) = \mathbf{X}\vec{\beta} \iff E(\vec{\varepsilon}) = \vec{0} \tag{2.2}$$

- **Homoscedasticidade:** A varianza das respostas, dado o conxunto de variables explicativas, é constante.

$$\mathrm{Var}(\mathbf{Y}|\mathbf{X}) = \sigma^2 \mathbf{I}_n \tag{2.3}$$

- **Independencia:** Os erros son independentes entre si.

$$\mathrm{Var}(\vec{\varepsilon}) = E[\vec{\varepsilon}\vec{\varepsilon}^T] = \sigma^2 \mathbf{I} \tag{2.4}$$

- **Normalidade:** As respostas, dado o conxunto de variables explicativas, seguen unha distribución normal multivariada.

$$\vec{Y}|\mathbf{X} \sim \mathcal{N}_n(\mathbf{X}\vec{\beta}, \sigma^2 \mathbf{I}_n) \iff \vec{\varepsilon} \sim \mathcal{N}_n(\vec{0}, \sigma^2 \mathbf{I}_n) \tag{2.5}$$

Multicolinealidade

A multicolinealidade é un problema que aparece cando algunhas das variables explicativas están altamente correlacionadas entre si, o que leva a que os efectos das variables explicativas sobre a resposta estén confundidos. Elevada multicolinealidade xera estimacións pouco precisas dos coeficientes β_j asociados a esas variables explicativas correladas. Aínda que realmente a multicolinealidade non afecta ao modelo á hora de predicir, o axuste resultante non é válido para unha análise estrutural do modelo.

Existen varias maneiras de detectar a multicolinealidade, pero a principalmente utilizada neste traballo foi a través dos Factores de Inflación da Varianza (FIV). O FIV_j mide, para cada variable, o factor de incremento na varianza de $\hat{\beta}_j$ na regresión múltiple con respecto á súa varianza nun modelo simple, con só esa variable X_j e a resposta. Os FIV calcúlanse da seguinte maneira:

$$\mathrm{FIV}_j = \frac{1}{1 - R^2_{(j)}} \tag{2.6}$$

, para j = 1, .., k sendo $R_(j)$ o coeficiente de correlación múltiple entre X_j e o resto de explicativas.

Un valor elevado suxire alta multicolinealidade para esa variable. É común comparalo co valor de $(1-R^2)^{-1}$, xa que un FIV maior ca este nivel quere dicir que existe maior correlación entre esta explicativa e o resto que entre a resposta e as explicativas.

Para solucionar este problema, é común levar a cabo un procedemento de selección de variables, vixiando non minguar demasiado a capacidade preditiva do axuste.

2.3.2 Random Forest

Un Random Forest é un algoritmo de Machine Learning utilizado para tarefas de regresión e clasificación, que combina múltiples árbores de decisión para mellorar a precisión e evitar o sobreaxuste.

Unha árbore de decisión é un modelo predictivo que utiliza unha estrutura en forma de árbore para representar decisións e as súas posibles consecuencias, incluíndo os resultados das decisións. En cada nodo da árbore, tomase unha decisión baseada nunha característica ou variable, que divide os datos en subconxuntos.

O Random Forest pertence á categoría dos métodos de ensemble, que combinan múltiples modelos base para crear un modelo máis robusto. O Random Forest usa un método chamado bagging (bootstrap aggregating), onde se crean múltiples subconxuntos de datos a partir do conxunto de datos orixinal mediante a técnica bootstrap (remostraxe con reemprazamento). Cada árbore de decisión é adestrada nun deses subconxuntos. Finalmente, as predicións das árbores son agregadas (por exemplo, mediante a media para regresión) para obter a predición final.

Este tipo de modelos ten varias vantaxes:

- Reduce o risco de sobreaxuste, ao combinar múltiples modelos.
- É robusto a outliers e ruído nos datos.
- Manexa automaticamente as interaccións entre as variables.

Validación cruzada

A validación cruzada é unha técnica utilizada para avaliar a xeneralización dun modelo, dividindo o conxunto de datos en subconxuntos. O procedemento xeral consiste en:

- Dividir o conxunto de datos en k subconxuntos ou folds.
- Adestrar o modelo en $k-1$ folds e avaliar co fold restante.
- Repetir o proceso k veces, cambiando o fold de validación cada vez.
- Promediar os resultados das k iteracións para obter unha medida de rendemento xeral.

LOOCV (Leave-One-Out Cross-Validation)

LOOCV é unha variante da validación cruzada onde o número de folds k é igual ao número de observacións no conxunto de datos. Isto significa que para cada observación no conxunto de datos, o modelo é adestrado coas $n-1$ observacións restantes e avaliado na observación deixada fóra. Este proceso repítese para cada observación, resultando en n modelos adestrados e n avaliacións. LOOCV proporciona unha estimación da precisión do modelo moi detallada, pero pode ser computacionalmente custosa para conxuntos de datos grandes. É o método de adestramento máis indicado cando o conxunto de datos dispoñíbel é pequeno.

2.3.3 Avaliación

Para avaliar a calidade dos modelos de regresión, úsanse varias métricas:

Coeficiente de determinación: R^2

O coeficiente de determinación R^2 mide a proporción da variabilidade na variable resposta que é explicada polas variables explicativas no modelo. A súa fórmula é:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2.7}$$

onde y_i son os valores observados, $\hat{y}_i$ son os valores preditos polo modelo, e $\bar{y}$ é a media dos valores observados.

O R^2 axustado é unha versión modificada do coeficiente de determinación que ten en conta o número de preditores no modelo, penalizando a inclusión de variables irrelevantes:

$$R^2_{\text{axustado}} = 1 - \left(\frac{1 - R^2}{n - p - 1}\right)(n - 1) \tag{2.8}$$

onde n é o número de observacións e p é o número de predictores. Un valor de R^2 de 1 implicaría un axuste do modelo perfecto, onde as variables explicativas capturan perfectamente a variabilidade da resposta.

Erro Cuadrático Medio (MSE)

O erro cuadrático medio mide a magnitude media dos erros ao cadrado das predicións do modelo:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.9}$$

Un valor máis baixo indica mellor axuste do modelo.

Erro Medio Absoluto (MAE)

O erro medio absoluto mide a magnitude media dos erros absolutos das predicións do modelo:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2.10}$$

De novo, un valor baixo indica mellor axuste.

2.3.4 Selección de variables

Un paso fundamental para obter un bo rendemento nos modelos de regresión é a selección de variables. Tanto para modelos lineais como para Random Forest, a presencia de variables irrelevantes pode levar a problemas coma o ruído, sobreaxuste ou multicolinealidade. Para cada un dos modelos construídos, aplicaremos procesos distintos para levar a cabo esta selección.

Para a regresión lineal, aplicamos un método de selección baseado no Criterio de Información de Akaike (AIC). Este estatístico baséase na verosimilitude do modelo axustado, incluíndo unha penalización polo número de parámetros do mesmo. Así, intenta seleccionar un modelo con bo axuste pero parsimonioso. Defínese da seguinte maneira:

$$AIC = 2p - 2\ln\left(L\left(\hat{\beta}\right)\right)$$

sendo p o número de parámetros e $L\left(\hat{\beta}\right)$ o valor da verosimilitude.

O algoritmo utilizado realiza unha selección por eliminación progresiva de variables, é dicir, parte do modelo completo e en cada paso elimina a variable que menor mellora ofrece no axuste. En máis detalle, o funcionamento é o seguinte:

1. Comeza co modelo completo M_k que inclúe todas as regresoras.

2. Para $j = k, k\text{–}1, \ldots, 1$:

   (a) Considera todos os j modelos que eliminan unha regresora de M_j.

   (b) Escolle o mellor dos novos j modelos segundo o que presente un R^2 máis elevado e denomínao M_{j-1}.

3. Selecciona o mellor entre $M_0, M_1, \ldots, M_k$ segundo o Criterio de Información de Akaike.

Este tipo de selección leva fácilmente a problemas de multicolinealidade se hai regresoras correladas, pero é un excelente método para evitar a eliminación de variables relevantes.

En canto ao modelo de aprendizaxe automático, o método de selección de variables utilizado é moi parecido. Tamén realizamos unha selección cara atrás (partimos do modelo completo e eliminamos paso a paso), pero neste caso o criterio para escoller o modelo será en base á importancia de cada variable preditora para o modelo. A importancia dunha variable para o modelo mídese segundo a diferenza positiva ou negativa na métrica escollida para a avaliación do modelo (MAE, MSE, R^2) ao ser esta variable incluida ou non.

Capítulo 3

Análise realizado

PARTINDO dos conceptos teóricos explicados no capítulo anterior, imos describir neste a aplicación que se lle deu a cada unha desas ferramentas, sobre que datos foron aplicadas e os resultados que obtivemos. Basicamente, explícase neste capítulo a análise realizada paso a paso e fundamentada nos conceptos xa explicados, para poder comprender as diversas aplicacións reais que teñen estes conceptos teóricos no campo no que estamos e a que resultados nos levan.

3.1 Preprocesado

O primeiro paso para poder realizar unha análise correcta da que extraer conclusións é realizar un preprocesado das bases de datos, comprendendo a súa natureza e comprobando a súa coherencia. Ao traballar con varias bases de datos en distintas fases (para os xogadores individuais, os partidos e os xogadores en quintetos), imos explicar os procesos que levamos a cabo para cada unha.

3.1.1 Xogadores individuais

Esta trátase da base de datos que contén os rexistros de todos os xogadores das últimas dúas tempadas na Liga ACB. Contén 38 rexistros (onde 35 son variables de rendemento e 3 son de información sobre o xogador) de 648 xogadores.

O primeiro paso vai ser realizar un filtrado destes xogadores por minutos e partidos xogados, para eliminar xogadores con participación residual e que vaian introducir máis ruído que información de calidade ás nosas análises. Consensuando co titor profesional, decidimos fixar un mínimo de 100 minutos e 7 partidos xogados ao longo de cada tempada individualmente. Tras este filtrado, os xogadores redúcense a 506.

A continuación, facemos unha inspección individual de cada variable. Todas elas son variables continuas (algunha, coma 'Puntos por partido', é realmente discreta, pero con suficientes valores como para tratala como continua). Observamos os valores mínimo e máximo de cada variable para comprobar a coherencia co seu significado real. Vendo isto, decatámonos de que existen valores negativos para variables que, por definición, non poden tomar valor menor ca 0. Estas variables son 'Puntos por posesión' e 'Porcentaxe tiros asistidos'. Como é lóxico, un xogador non pode anotar puntos negativos por cada posesión nin ter unha ratio negativa de tiros de campo acertados procedentes de pase. En ambos casos, o mínimo sería 0, por iso substituímos os 2 valores negativos de 'Porcentaxe tiros asistidos' e o único de 'Puntos por posesión' por 0.

A continuación, tratamos de comprender a distribución de cada variable a través de histogramas e diagramas de caixas. Por exemplo, para a variable 'Puntos Por Partido':

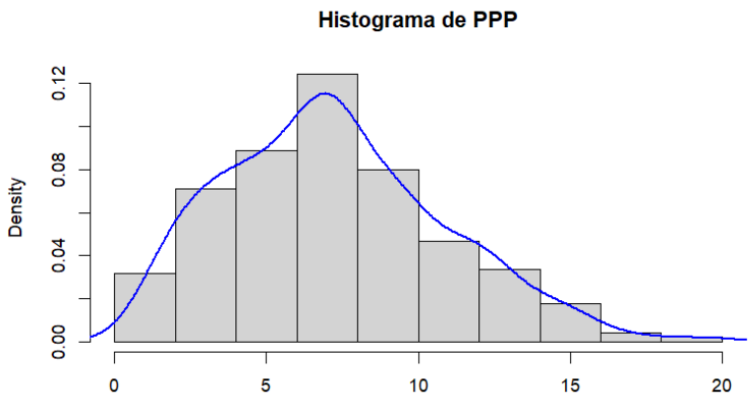

Figura 3.1: Histograma para a variable 'Puntos por partido'

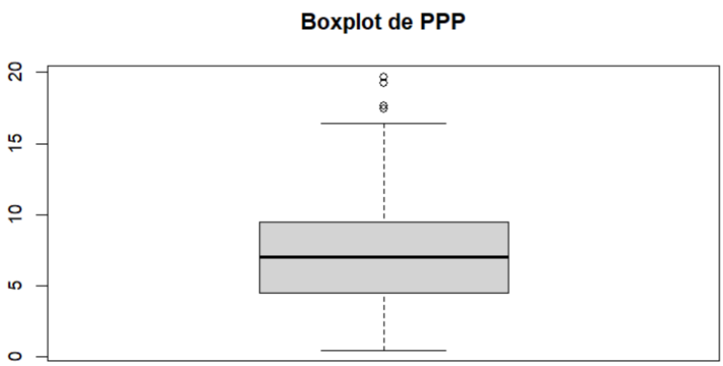

Figura 3.2: Boxplot para a variable 'Puntos por partido'

Así, observamos o número de datos atípicos que ten cada variable e a distribución dos propios datos. Podemos ver que os 4 datos atípicos se sitúan nos valores máis altos da variable (bastante razoable pensando no significado real desta variable) e que esta non parece seguir unha distribución normal. Comprobamos co test estatístico de Shapiro-Wilk para a normalidade, confirmando que esta variable non o é.

Realizamos estas análises para todas as variables, obtendo só 3 variables para as que non cabe rexeitar a normalidade co test utilizado e un alpha de 0.05: 'Puntos por minuto', 'Faltas recibidas' e 'Porcentaxe de tiros de campo intentados' (mide a ratio dos tiros do seu equipo que intenta o xogador).

É tamén interesante observar a matriz de correlacións dos datos, e comprobar se hai moitas variables altamente correladas. Fixando un limiar de, por exemplo, 0.7, comprobamos qué porcentaxe das variables superan este valor. O resultado foi que temos 41 pares correlados dun total de 595, dando unha porcentaxe de 6.89.

Por suposto, esta comparación fíxose en valor absoluto para ter en conta tamén as posíbeis correlacións negativas.

3.1.2 Partidos

A continuación, traballamos coa base de datos dos partidos desta tempada do Obradoiro CAB. Aínda que a base de datos contiña todos os partidos de todos os equipos, decidimos centrar a análise no Obradoiro CAB para poder ter unha ferramenta específica axustada cos datos do equipo. Filtramos así a base datos por equipos para obter só os do Obradoiro.

A continuación, interesaba obter unha métrica do nivel do rival, xa que parece lóxico pensar que non é o mesmo xogar un partido contra o primeiro clasificado que contra o último, polo que é unha variable que pode afectar ao resultado. A métrica escollida para ter en conta o rival foi a diferenza de puntos a favor e en contra nos seus partidos, polo que agrupamos por equipo e calculamos a media da diferenza. Esta medida aproxima bastante ben a clasificación real da liga:

Orde segundo diferencia media	**Clasificación final Liga ACB**
Unicaja	Unicaja
Real Madrid	Real Madrid
Barça	Barça
UCAM Murcia	Valencia Basket
Lenovo Tenerife	UCAM Murcia
Dreamland Gran Canaria	Lenovo Tenerife
Valencia Basket	Dreamland Gran Canaria
BAXI Manresa	BAXI Manresa
Baskonia	Baskonia
MoraBanc Andorra	Joventut Badalona
Surne Bilbao Basket	MoraBanc Andorra
Casademont Zaragoza	Casademont Zaragoza
Joventut Badalona	Surne Bilbao Basket
Monbus Obradoiro	Bàsquet Girona
Bàsquet Girona	Coviran Granada
Río Breogán	Río Breogán
Coviran Granada	Monbus Obradoiro
Zunder Palencia	Zunder Palencia

Táboa 3.1: Comparación da diferenza media de puntos cos resultados finais da Liga ACB

Queda así un data frame con tantas filas coma partidos e con 24 variables de rendemento.

Ademais, como a análise que imos realizar posteriormente é un modelo de regresión, analizamos a distribución da variable que será a resposta: Resultado, a diferenza de puntos a favor e en contra en cada partido.

No histograma, observamos que é unha variable desprazada cara a esquerda con respecto do 0, con máis valores negativos que positivos. Ademais é asimétrica, xa que a mediana é menor que a media (-6.5 e -4.346, respectivamente).

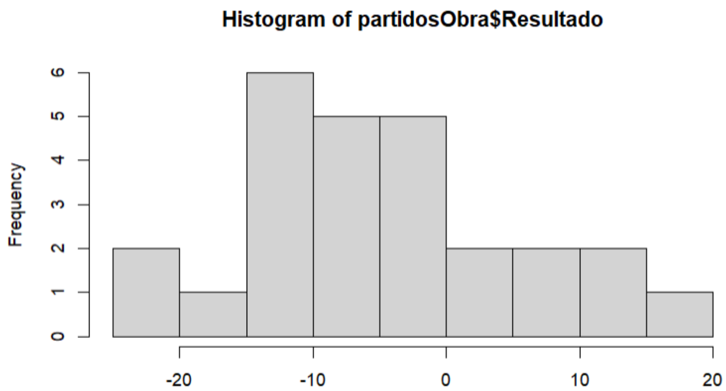

Figura 3.3: Histograma para a variable 'Resultado'

Vemos no diagrama de caixas que non existen datos atípicos, e confirmamos a asimetría positiva:

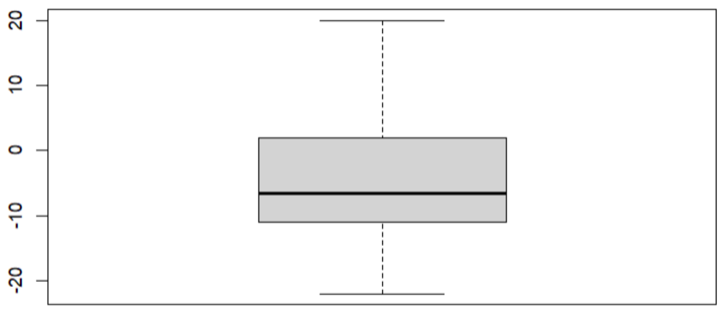

Figura 3.4: Boxplot para a variable 'Resultado'

3.1.3 Agrupación por quintetos

De novo, a base de datos contén datos de todos os quintetos de todos os equipos, pero volvemos decidir centrar a análise nos datos específicos do Obradoiro CAB para obter resultados personalizados do equipo.

Tamén imos realizar un corte segundo os minutos xogados por cada quinteto, pero isto se vai explicar nunha sección máis adiante xa que está relacionado coa análise específica a realizar.

Esta base de datos contén 20 variables de rendemento conseguidas por un quinteto de xogadores específico. Como o que se pretende é atopar a mellor configuración de xogadores tendo en conta o seu perfil, o que se fixo foi substituír os nomes dos xogadores polo seu clúster resultante na análise sobre a base de datos dos xogadores individuais. Non obstante, como a orde na combinación destes clusters é irrelevante (é o mesmo o quinteto de clusters A, A, A, B, B que a combinación B, B, A, A, A) o que se fixo foi transformar o data frame para obter cantos xogadores hai de cada cluster. Seguindo o exemplo anterior, e supoñendo tres cluster A, B e C, o data frame tería unha columna para cada cluster cos valores A: 3, B: 2 e C:0.

Cómpre destacar que o corte de minutos ten que ser, por lóxica, menor para os quintetos que para os xogadores individuais, polo tanto e ao tratarse de bases de datos distintas, nalgún quinteto que supere o corte de minutos pode estar contido un xogador que non superase o corte individual e non teña asignado un cluster. Como este era un problema que sucedía en poucas ocasións, decidimos directamente non ter en conta estes quintetos.

3.2 PCA

A primeira ferramenta aplicada foi a PCA, xa que logo servirá de axuda para análises posteriores. Obtemos as compoñentes principais coa base de datos normalizada, para evitar problemas coas escalas, e pasamos a escoller o número delas a utilizar.

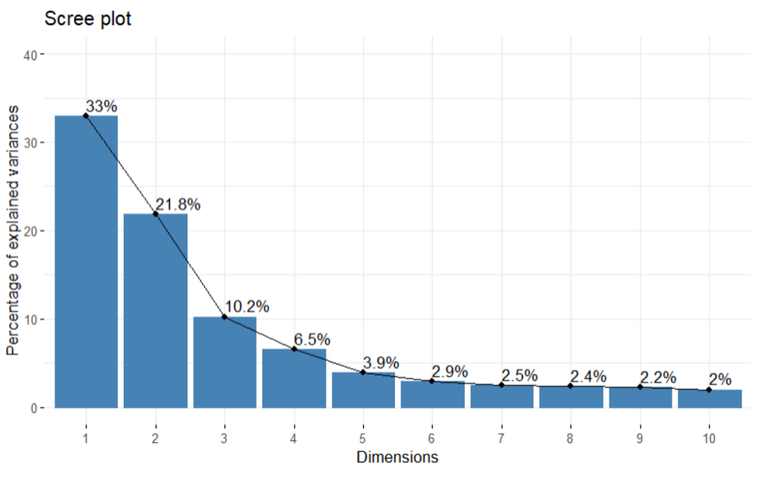

Figura 3.5: Porcentaxe de varianza explicada por cada compoñente

Como vemos no gráfico, o resultado é bastante bo. Con dúas compoñentes xa se explica máis dun 50% da varianza total. Con tres explica exactamente un 65%, pero sube ao 75.4% incluíndo cinco. Pasar das 35 variables orixinais a 5 mantendo un 75.4% da información total pareceu un resultado razoablemente bo, polo que imos traballar con 5 compoñentes.

3.2.1 Significado das compoñentes

Cada unha destas compoñentes ten unha significación real asociada, e a continuación imos mostrar a das primeiras tres compoñentes (que son as que explican a maior parte da variabilidade).

Primeira compoñente

Mostramos o peso de cada variable na primeira compoñente:

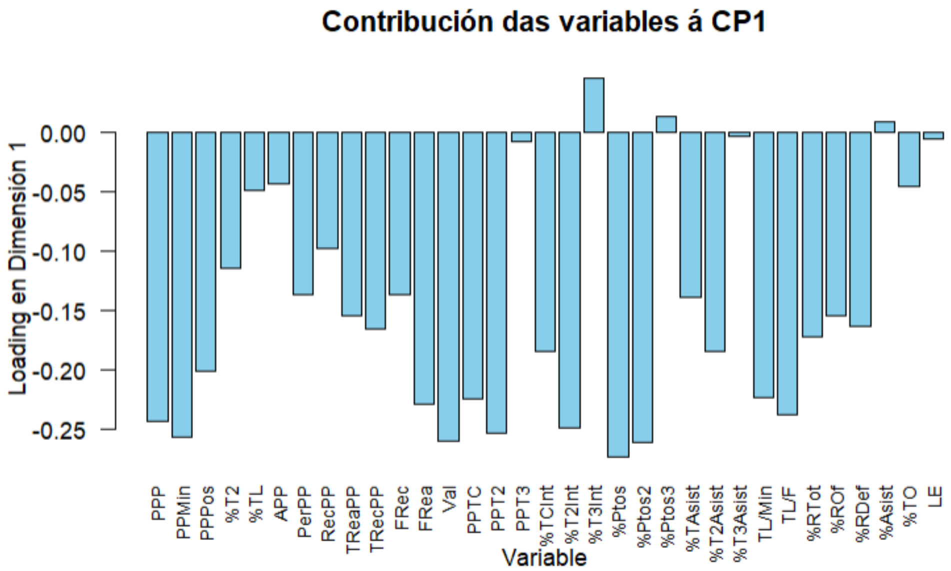

Figura 3.6: Coeficientes da Primeira Compoñente

Vemos aquí que a maioría das variables teñen un peso negativo, sendo as que máis influencia teñen as relacionadas coa anotación e lanzamentos (de dous puntos) e as porcentaxes de rebotes (porcentaxe total, ofensiva e defensiva). As únicas con peso positivo están relacionadas coa anotación de 3 puntos e a ratio de asistencias, pero todas con pouco peso.

Concluímos que os xogadores con puntuación negativa nesta compoñente son xogadores con moito peso na anotación e predominantemente interiores, que non lanzan de 3 puntos. Os xogadores con puntuación positiva serán xogadores exteriores sen moita relevancia ofensiva. Mostramos os xogadores con maior puntuación, tanto positiva coma negativa, na compoñente:

Puntuación positiva			Puntuación negativa		
Xogador	**Temporada**	**Puntuación**	**Xogador**	**Temporada**	**Puntuación**
POL FIGUERAS	2023/24	8.914533	DRAGAN BENDER	2022/23	-8.454069
ALBERT VENTURA	2022/23	8.213376	WILLY HERNANGOMEZ	2023/24	-8.068303
JORDI RODRIGUEZ	2023/24	8.086942	GIORGI SHERMADINI	2022/23	-7.957521
PABLO ALMAZAN	2022/23	7.801098	ETHAN HAPP	2023/24	-7.931423
JOVAN KLJAJIC	2022/23	7.783381	BRANDON DAVIES	2023/24	-7.735119

Táboa 3.2: Xogadores con puntuación positiva e negativa na primeira compoñente

Ao visualizar os xogadores, confirmamos que os perfís que intuíamos son correctos. Todos os xogadores con puntuación moi negativa xogan en posicións interiores e son dos anotadores máis destacados da competición, mentres que os xogadores con valores positivos grandes son xogadores principalmente exteriores, e con pouca relevancia no xogo ofensivo do seu equipo.

Segunda compoñente

No caso da segunda compoñente, destacan especialmente en positivo as variables relacionadas coa anotación exterior (puntos por tiro de tres intentados, porcentaxe de tiros de tres intentados, tiros de tres asistidos...), o volume de anotación (puntos por partido e por minuto) e as variables relacionadas coas asistencias (ratio de asistencias e asistencias por partido). Pola contra, en negativo destacan as variables reboteadoras, a eficiencia na anotación (puntos por posesión, en lugar de por partido ou minuto) e especialmente na anotación de dous puntos (porcentaxe en tiros de 2, tiros de 2 asistidos...) e os tapóns realizados.

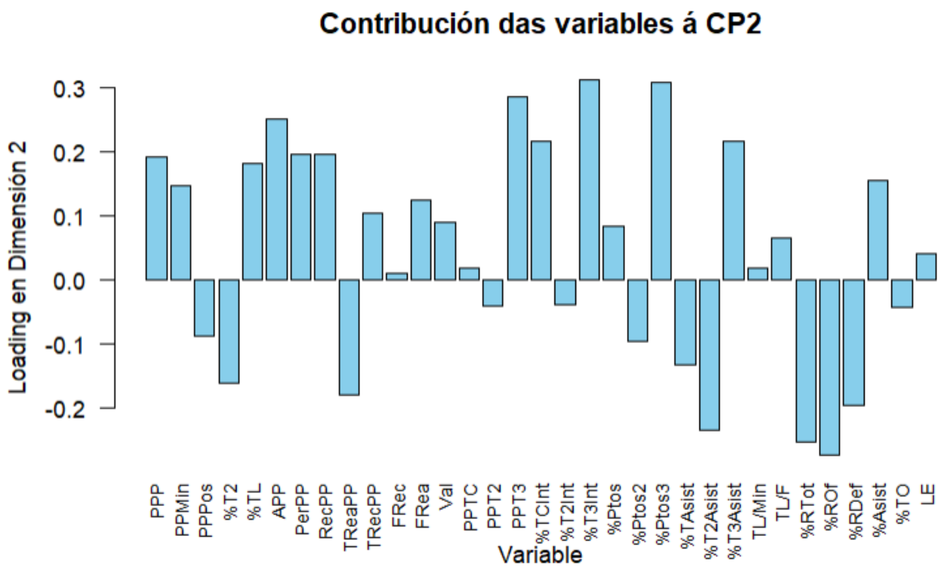

Figura 3.7: Coeficientes da Segunda Compoñente

Asociamos os valores positivos nesta dimensión a xogadores exteriores especialistas na anotación exterior, que toman moitos tiros pero que tamén asisten, é dicir, que teñen o balón moito tempo e con moita importancia no ataque do equipo. Os xogadores con valores negativos son xogadores interiores, con pouca influencia anotadora pero con moito peso reboteador e defensivo.

Mostramos os xogadores máis representativos desta compoñente:

Puntuación positiva			**Puntuación negativa**		
Xogador	**Temporada**	**Puntuación**	**Xogador**	**Temporada**	**Puntuación**
SHANNON EVANS	2022/23	6.940284	MARCUS LEE	2022/23	-6.249614
MARKUS HOWARD	2023/24	6.483216	BOUBACAR TOURE	2023/24	-5.642699
JEAN MONTERO	2022/23	5.487223	EDY TAVARES	2023/24	-5.583532
JEAN MONTERO	2023/24	5.229768	JAMES NNAJI	2023/24	-5.517565
MARKUS HOWARD	2022/23	5.215525	FELIPE DOS ANJOS	2023/24	-5.354127

Táboa 3.3: Xogadores con puntuación positiva e negativa na segunda compoñente

Efectivamente, todos os xogadores con puntuación alta son xogadores exteriores con volumes de anotación moi elevados, especialmente no lanzamento de 3 puntos. En canto aos xogadores destacadamente negativos, son en todos os casos xogadores interiores non especialmente dominantes en anotación, senón que destacan noutras facetas como a reboteadora.

Terceira compoñente

En canto á última compoñente, observamos como as variables máis influíntes son as porcentaxes de tiros asistidos (especialmente nos lanzamentos de 3) e as relacionadas coa eficiencia na anotación (puntos por tiro, puntos por posesión...). Tamén entran lixeiramente as porcentaxes de rebotes e os tapóns realizados. Cun peso negativo, temos as variables relacionadas con asistencias e pérdidas, a porcentaxe de tiros de 2 intentados e a porcentaxe de puntos de 2 que anota con respecto ao seu equipo.

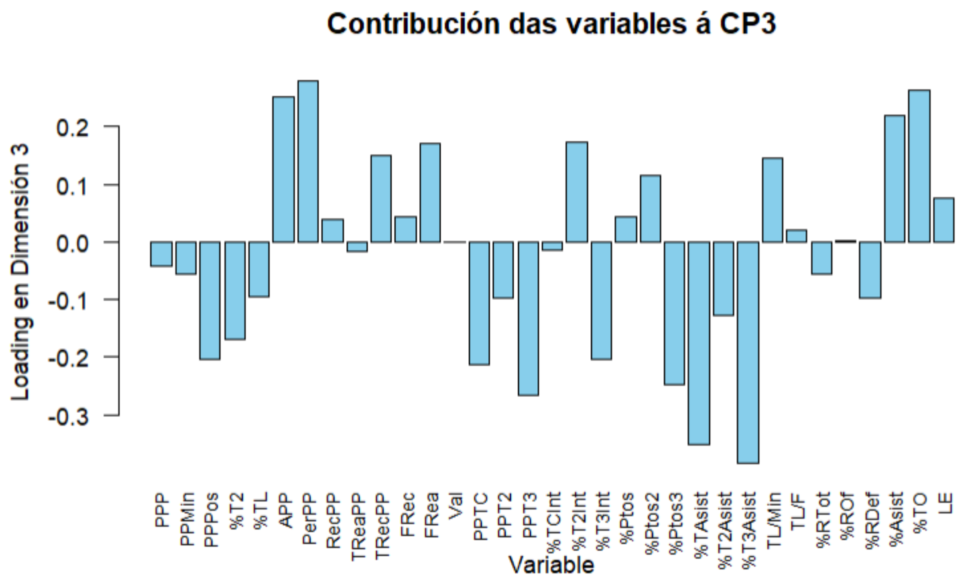

Figura 3.8: Coeficientes da Terceira Compoñente

Estes coeficientes fannos pensar que o xogador que teña unha puntuación positiva nesta variable será un manexador do balón (moitas asistencias e pérdidas) que non destaque pola anotación de 3 puntos pero poda anotar de 2. Os xogadores cunha puntuación negativa serán grandes especialistas no tiro de 3 puntos, sobre todo tras pase. A entrada das porcentaxes de rebote fannos pensar que ademais é un tirador de gran tamaño, non un xogador moi exterior.

Puntuación positiva			**Puntuación negativa**		
Xogador	**Tempada**	**Puntuación**	**Xogador**	**Tempada**	**Puntuación**
PIERRIA HENRY	2022/23	5.956300	AARON DOORNEKAMP	2022/23	-4.789964
YIFTACH ZIV	2023/24	4.379170	ALEX ABRINES	2023/24	-4.734096
SHANNON EVANS	2022/23	4.373322	ROKAS GIEDRAITIS	2022/23	-4.484615
JORDAN DAVIS	2022/23	4.063880	MIKE TOBEY	2022/23	-4.434831
MARCELINHO HUERTAS	2022/23	4.019464	TIM ABROMAITIS	2023/24	-4.211964

Táboa 3.4: Xogadores con puntuación positiva e negativa na terceira compoñente

Observando os xogadores destacados tanto en positivo coma en negativo, corroboramos os perfís formados anteriormente. Os xogadores con puntuación positiva son bases con bo manexo do balón e un gran dominio do pase e as asistencias, mentres que os xogadores que destacan por abaixo son tiradores especialistas de 3 puntos e, efectivamente, xogan en posicións non demasiado exteriores, sendo algún incluso un interior que sae a lanzar de fóra.

3.2.2 Utilidades

Como comentamos no contido teórico, a Análise de Compoñentes Principais ten moitas utilidades. Aquí imos mostrar algunhas das utilizadas neste traballo, ademais da significación propia de cada compoñente mencionada anteriormente.

Representación gráfica

Ao reducir máis dun 50% da varianza en tan só dúas compoñentes e máis dun 60% en tres, podemos utilizar estas para visualizar as posicións dos individuos sobre estas compoñentes e mellorar a comprensión visual dos datos de maneira directa. Por exemplo, en 2D:

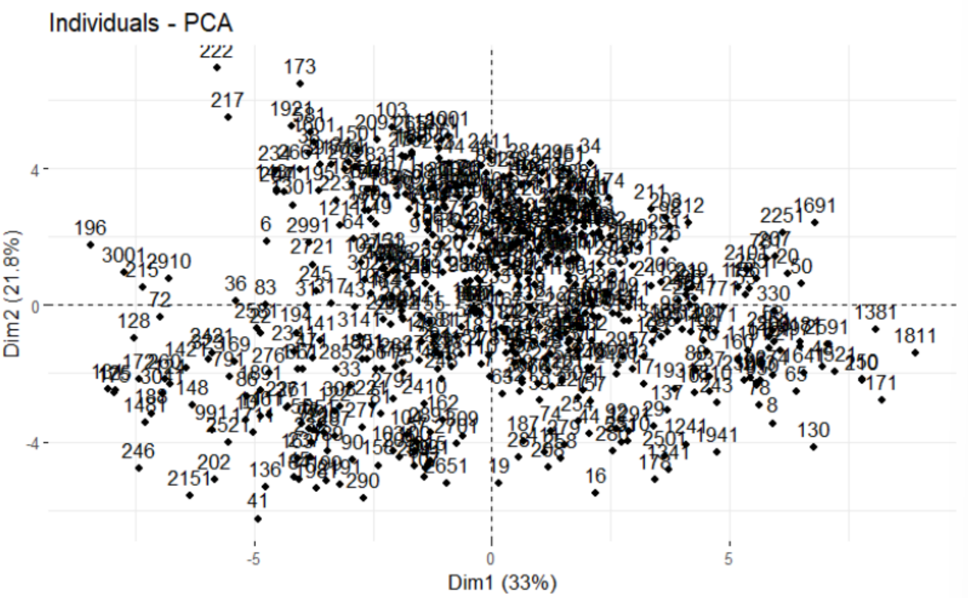

Figura 3.9: Individuos representados sobre as dúas primeiras compoñentes

Nestes gráficos xa podemos apreciar algún patrón, como por exemplo que non existen xogadores con valores altos tanto na dimensión 1 coma na dimensión 2 ao mesmo tempo, pero si con valores moi negativos en ambas.

Tamén podemos realizar gráficos en 3D collendo as tres primeiras compoñentes principais:

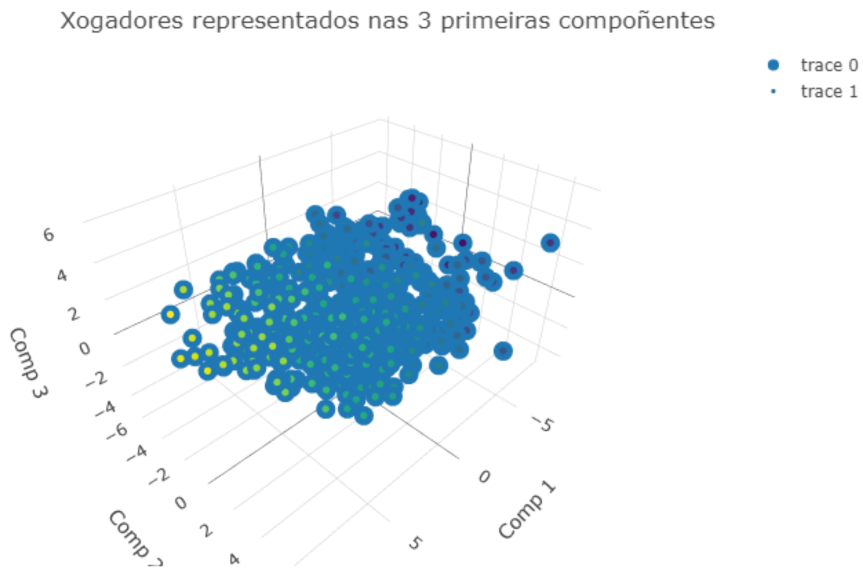

Figura 3.10: Individuos representados sobre as tres primeiras compoñentes

Este é un gráfico interactivo, que se pode rotar, afastar ou acercar como se desexe. Mostra o nome, tempada e puntuación sobre as tres primeiras compoñentes de cada individuo ao seleccionar o punto que lle corresponde sobre o espazo. Vemos aquí un exemplo disto:

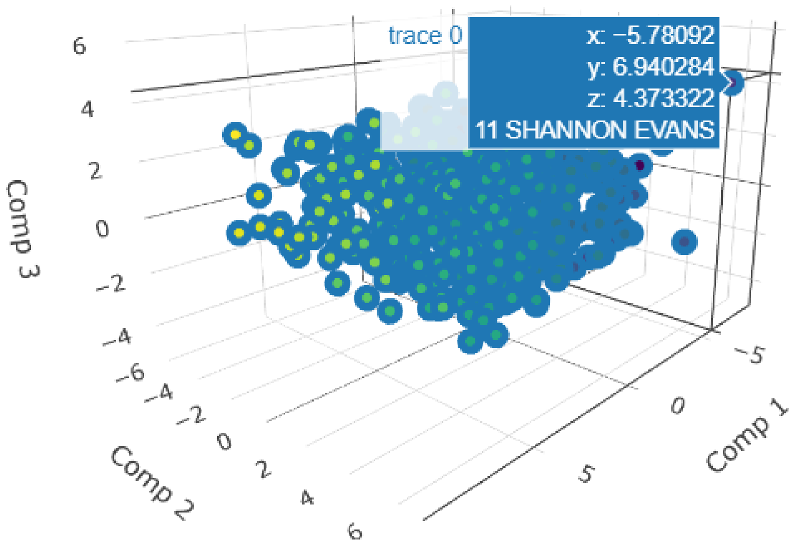

Figura 3.11: Mostra do gráfico interactivo

Isto facilita enormemente a comprensión visual dos datos. Sitúa aos xogadores nunhas compoñentes das cales coñecemos o significado, podendo comprender de maneira moi rápida en que destaca cada xogador.

Cálculo de distancias

Visualmente, a representación xa da unha aproximación das distancias entre os xogadores. Non obstante, estas poden ser calculadas de maneira moi sinxela utilizando os datos transformados nas cinco compoñentes principais, eliminando a influencia de variables irrelevantes, de ruído e das escalas das variables.

Así, podemos crear funcións para calcular a distancia entre dous xogadores concretos ou para buscar o xogador máis cercano a un dado:

```
> distancia(df_dist, '22 EDY TAVARES', "2023/24", '13 SERGIO RODRIGUEZ', "2023/24")
          203
2151 13.11286
> distancia(df_dist, '22 EDY TAVARES',"2023/24", '17 VINCENT POIRIER', "2023/24")
           202
2151 0.9514904
```

Figura 3.12: Mostra de resultado de cálculo de distancias

Como vemos no exemplo, o cálculo das distancias é moi intuitivo: a distancia entre Edy Tavares, pívot de 2.20 metros, e Sergio Rodríguez, base de 1.91 m, é moito maior que entre o propio Tavares e Vincent Poirer, co que comparte posición e estilo de xogo.

Precisamente, se buscamos o xogador máis próximo a Edy Tavares este ano, o resultado é o seu compañeiro Vincent Poirer. A esta función tamén se lle pode aplicar algún filtro, como pode ser o equipo.

```
> buscar_xogador(df_dist, "22 EDY TAVARES", año = "2023/24")
[1] "17 VINCENT POIRIER" "0.951490449633018"
```

Figura 3.13: Resultado da procura de xogador máis próximo

```
> buscar_xogador(df_dist, "22 EDY TAVARES", equipo = "Monbus Obradoiro", año = "2023/24")
[1] "11 MAREK  BLAZEVIC" "3.11048106022544"
```

Figura 3.14: Resultado da procura de xogador próximo con filtro de equipo

Seguindo de novo cos datos transformados, podemos calcular a matriz de distancias do equipo Obradoiro CAB, para obter as distancias entre todos os xogadores da plantilla:

	93 ALEX SUAREZ	5 FERNANDO ZURBRIGGEN	10 DEVON DOTSON	18 ROKO BADZIM	8 JANIS STRELNIEKS	50 OLEKSANDR KOVLIAR	11 MAREK BLAZEVIC	13 ARTEM PUSTOVYI	44 TRES TINKLE
93 ALEX SUAREZ	0.000000	7.976386	11.238408	6.701140	7.154800	8.023223	11.991555	14.317509	9.251161
5 FERNANDO ZURBRIGGEN	7.976386	0.000000	4.178792	4.369881	7.677732	1.740755	8.506633	9.578577	4.805866
10 DEVON DOTSON	11.238408	4.178792	0.000000	5.443728	9.365943	3.737667	8.542977	8.326448	5.090893
18 ROKO BADZIM	6.701140	4.369881	5.443728	0.000000	5.381518	3.369374	9.718423	10.796043	5.929656
8 JANIS STRELNIEKS	7.154800	7.677732	9.365943	5.381518	0.000000	6.400672	13.871832	15.224499	10.866720
50 OLEKSANDR KOVLIAR	8.023223	1.740755	3.737667	3.369374	6.400672	0.000000	9.052122	9.984892	5.507214
11 MAREK BLAZEVIC	11.991555	8.506633	8.542977	9.718423	13.871832	9.052122	0.000000	3.009649	6.108144
13 ARTEM PUSTOVYI	14.317509	9.578577	8.326448	10.796043	15.224499	9.984892	3.009649	0.000000	6.998177
44 TRES TINKLE	9.251161	4.805866	5.090893	5.929656	10.866720	5.507214	6.108144	6.998177	0.000000
1 THOMAS SCRUBB	9.256520	3.848066	5.380157	6.513250	11.008014	5.208461	6.435159	7.468253	2.538332
26 RIGOBERTO MENDOZA	7.951270	2.774770	4.082876	3.371151	8.120049	3.009545	7.212834	8.179652	3.714375
6 JANIS TIMMA	8.666470	3.959951	5.390276	5.918084	8.990259	4.363299	9.887849	10.895228	4.843923
32 RUBEN GUERRERO	5.091345	8.176312	11.025124	7.902980	9.288355	8.366170	8.601962	11.337005	8.708273
33 ALVARO MUNOZ	5.386725	3.271023	6.330548	4.022857	7.509413	3.775285	8.154194	9.957444	4.343870
7 POL FIGUERAS	5.175990	9.938922	13.059470	8.561927	6.034706	9.524581	14.372630	16.552302	12.724229
0 JORDAN HOWARD	10.576995	6.022836	5.496290	5.631151	10.488392	6.245770	9.439202	9.473882	5.497692

	1 THOMAS SCRUBB	26 RIGOBERTO MENDOZA	6 JANIS TIMMA	32 RUBEN GUERRERO	33 ALVARO MUNOZ	7 POL FIGUERAS	0 JORDAN HOWARD
93 ALEX SUAREZ	9.256520	7.951270	8.666470	5.091345	5.386725	5.175990	10.576995
5 FERNANDO ZURBRIGGEN	3.848066	2.774770	3.959951	8.176312	3.271023	9.938922	6.022836
10 DEVON DOTSON	5.380157	4.082876	5.390276	11.025124	6.330548	13.059470	5.496290
18 ROKO BADZIM	6.513250	3.371151	5.918084	7.902980	4.022857	8.561927	5.631151
8 JANIS STRELNIEKS	11.008014	8.120049	8.990259	9.288355	7.509413	6.034706	10.488392
50 OLEKSANDR KOVLIAR	5.208461	3.009545	4.363299	8.366170	3.775285	9.524581	6.245770
11 MAREK BLAZEVIC	6.435159	7.212834	9.887849	8.601962	8.154194	14.372630	9.439202
13 ARTEM PUSTOVYI	7.468253	8.179652	10.895228	11.337005	9.957444	16.552302	9.473882
44 TRES TINKLE	2.538332	3.714375	4.843923	8.708273	4.343870	12.724229	5.497692
1 THOMAS SCRUBB	0.000000	3.571920	5.062376	8.666062	4.317250	12.341038	5.535432
26 RIGOBERTO MENDOZA	3.571920	0.000000	5.435451	7.763506	3.483368	10.174101	4.153371
6 JANIS TIMMA	5.062376	5.435451	0.000000	9.627176	4.165950	11.799865	7.815020
32 RUBEN GUERRERO	8.666062	7.763506	9.627176	0.000000	5.867101	6.957414	11.244173
33 ALVARO MUNOZ	4.317250	3.483368	4.165950	5.867101	0.000000	8.666217	6.925737
7 POL FIGUERAS	12.341038	10.174101	11.799865	6.957414	8.666217	0.000000	12.911288
0 JORDAN HOWARD	5.535432	4.153371	7.815020	11.244173	6.925737	12.911288	0.000000

Figura 3.15: Matriz de distancias dos xogadores do Obradoiro CAB

Aquí, podemos observar que xogador é o máis afastado dentro do equipo, cales son os máis próximos, distancias concretas sen acudir á función anterior...

Cómpre destacar que para o cálculo de todas estas distancias a métrica utilizada foi a distancia euclídea. A razón para non empregar a distancia de Mahalanobis é que esta, ao utilizar a matriz de varianzas-covarianzas no cálculo, realiza unha estandarización das varianzas dos datos, o que daría o mesmo peso a cada dimensión. Ao aplicar a PCA, interesa conservar a importancia relativa de cada compoñente segundo a variabilidade que explican, e isto reflíctese mellor coa distancia euclídea.

3.3 Clustering realizado

Para levar a cabo o etiquetado dos individuos, procedemos a realizar un clustering xerárquico dos datos. A elección deste tipo de clustering ven dada por non coñecer de antemán o número de grupos existentes. Agora, imos comparar as métricas de distancia e os métodos de encadeamento para alcanzar un agrupamento óptimo dos datos.

Dentro de cada distancia, escollemos os métodos de encadeamento que xeran un dendograma máis correcto segundo as nocións comentadas no contido teórico. Un exemplo de dendograma non desexable sería o seguinte:

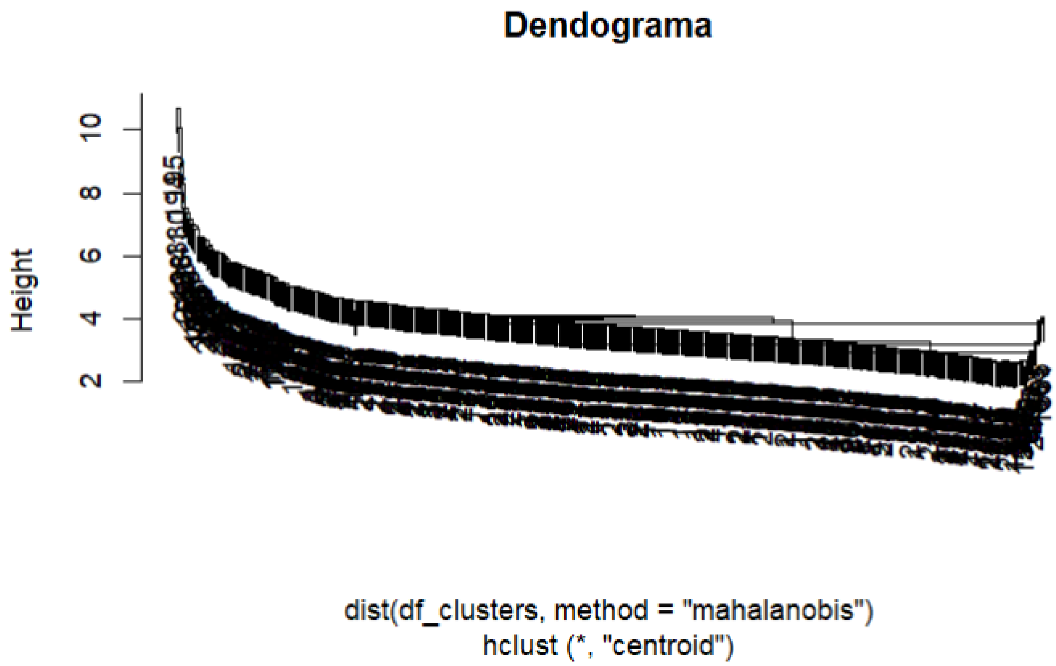

Figura 3.16: Dendograma xerado pola distancia de Mahalanobis e encadeamento tipo centroide

Aquí, un lixeiro cambio na altura do corte varía moito a partición formada, e algúns grupos estarían formados por un só individuo. Hai que tratar de evitar isto.

A continuación, compararemos os grupos resultantes tanto visualmente (comprobamos como se distribúen os grupos nas dimensións da Análise de Compoñentes Principais) como a través do coñecemento do titor profesional.

3.3.1 Distancia de Mahalanobis

Para a distancia de Mahalanobis, o mellor método de encadeamento foi o de Ward, xerando o seguinte dendograma:

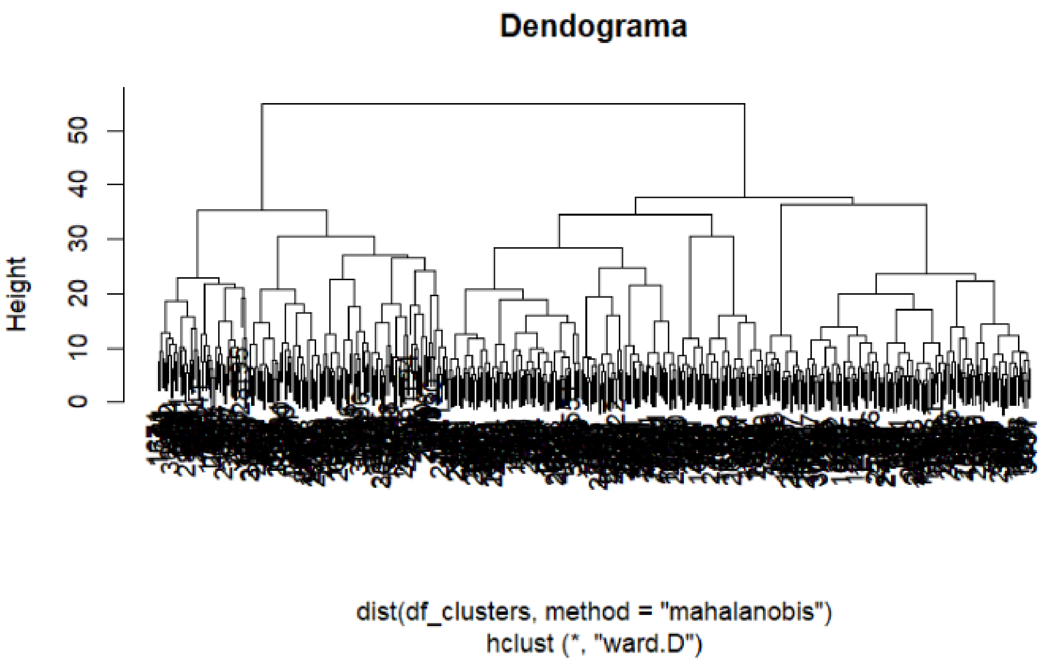

Figura 3.17: Dendograma xerado pola distancia de Mahalanobis e encadeamento tipo Ward

Ao realizar un corte por h = 30, obtemos as seguintes agrupacións:

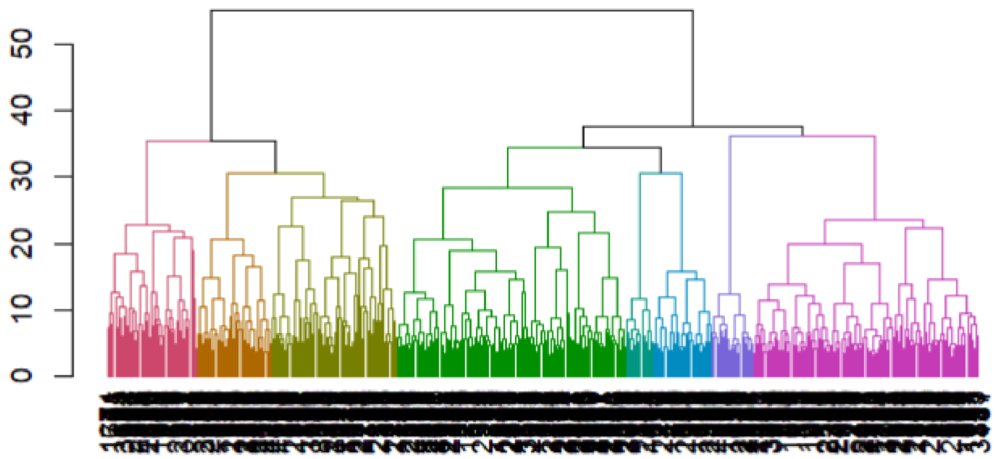

Figura 3.18: Dendograma xerado pola distancia de Mahalanobis e encadeamento tipo Ward coloreado

Vemos que hai clusters bastante desbalanceados, sendo os de cor verde e rosa moito máis poboados có de morado.

Mostramos agora os individuos coloreados por cluster en gráficos de dúas e tres dimensións:

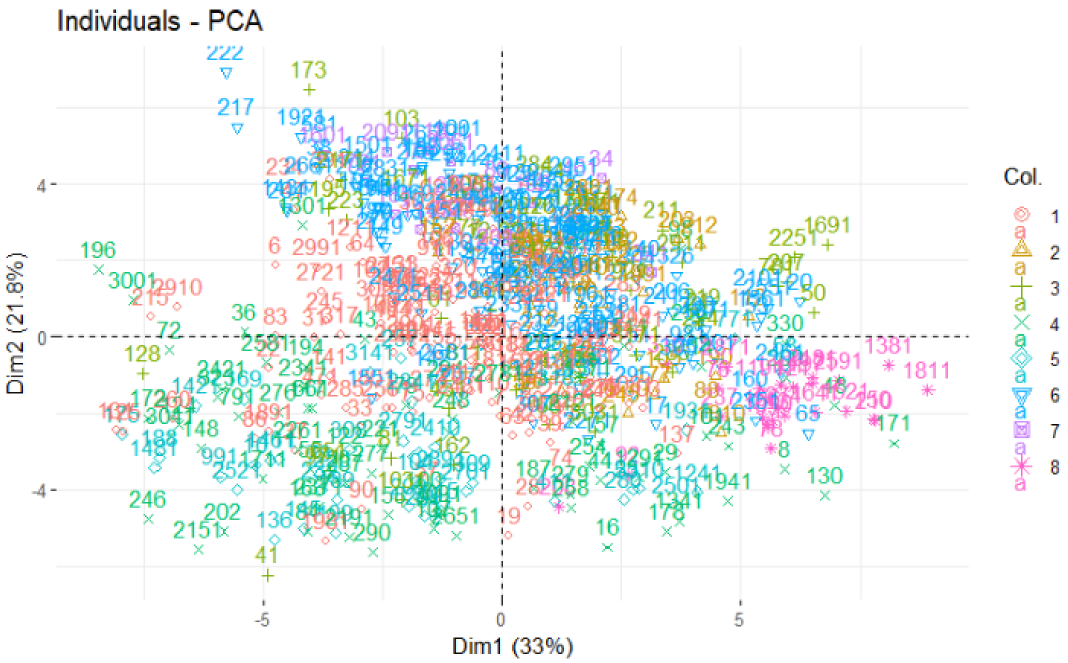

Figura 3.19: Individuos segundo clúster nas 2 primeiras compoñentes principais

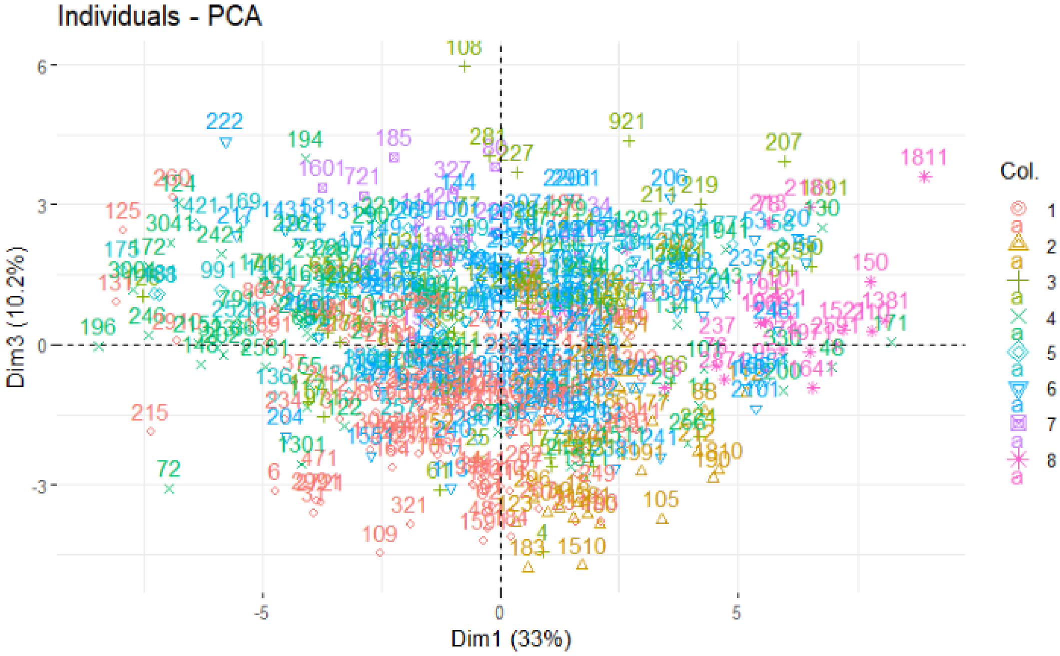

Figura 3.20: Individuos segundo clúster nas 2 primeiras compoñentes principais

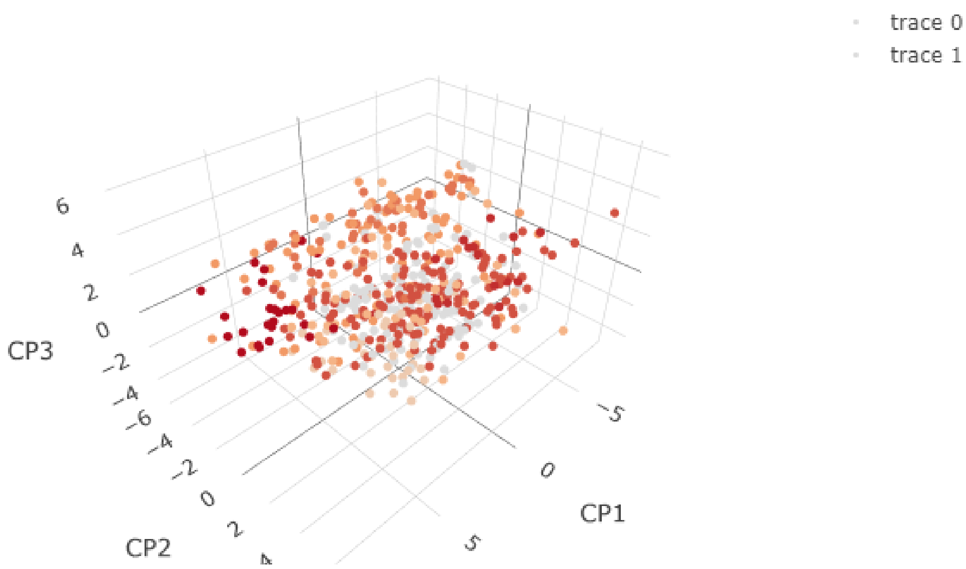

Figura 3.21: Individuos segundo clúster nas 3 primeiras compoñentes principais

Como podemos observar, non se aprecia ningún patrón na localización dos individuos dos mesmos clusters en ningunha das compoñentes. Os clusters repártense por calquera punto do espazo sen agruparse, o que nos leva a sospeitar que este agrupamento non se vai axustar á realidade.

Comprobando a significación real dos xogadores, o titor profesional non considerou que esta fose unha boa partición dos individuos que se axustase á realidade dos xogadores e os seus perfís.

3.3.2 Distancia Euclídea

Imos repetir o proceso anterior coa distancia euclídea, para comprobar se esta si que arroxa un agrupamento dos datos máis correcto.

O dendograma con mellor aspecto foi o devolto pola variante da función de Ward, que mostramos a continuación:

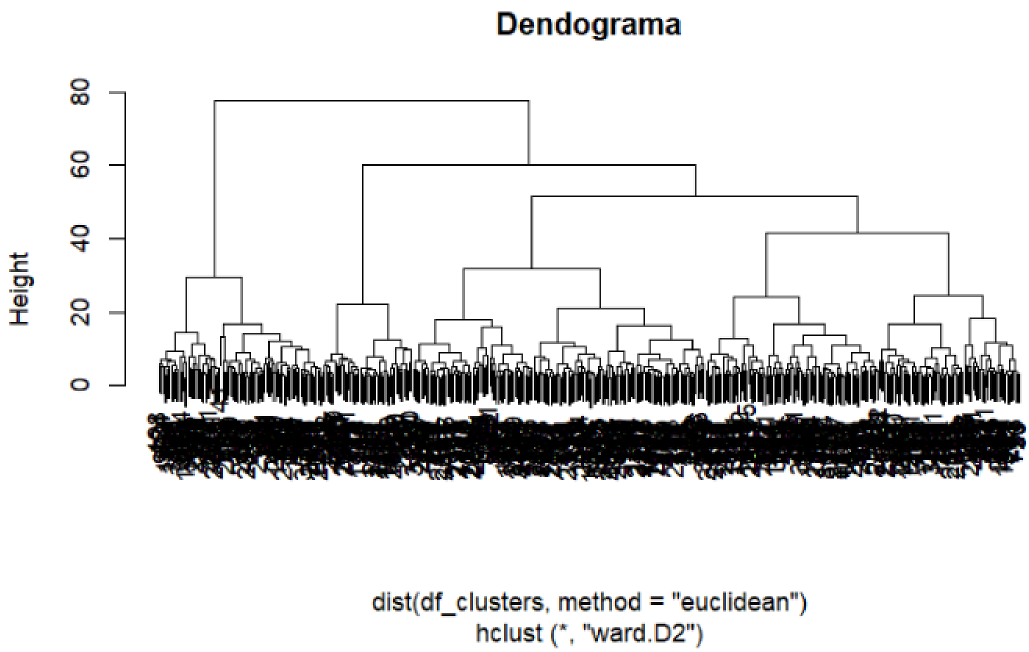

Figura 3.22: Dendograma xerado pola distancia euclídea e encadeamento tipo Ward.2

Tendo en conta o balanceo dos clusters e a marxe de maniobra en canto á altura, realizamos un corte por h = 30 que proporciona as seguintes agrupacións:

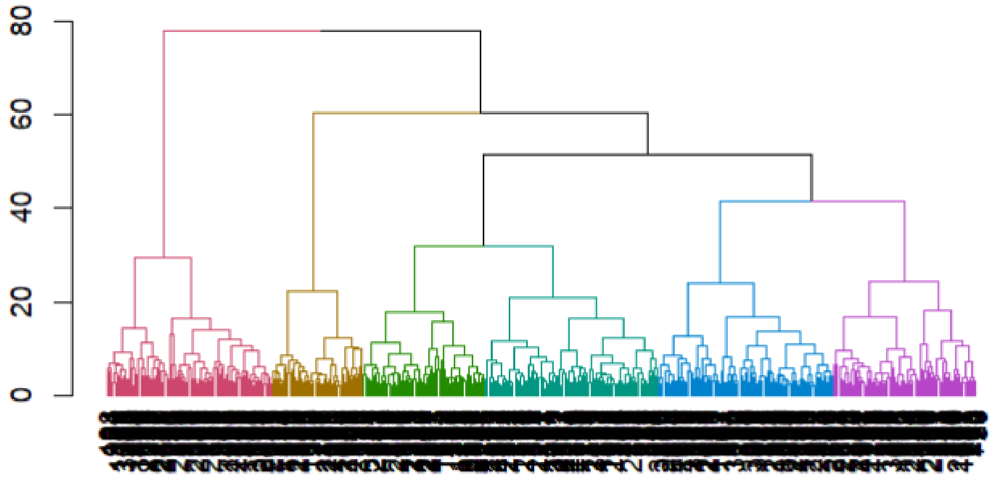

Figura 3.23: Dendograma xerado pola distancia euclídea e encadeamento tipo Ward.2 coloreado

Estes clusters parecen ter un tamaño aproximadamente similar, e xera un número de grupos acorde ao significado real (6 grupos, tendo en conta que existen 5 posicións tradicionais no baloncesto como se comentou na introdución).

Pasamos a comprobar gráficamente o comportamento destes clusters:

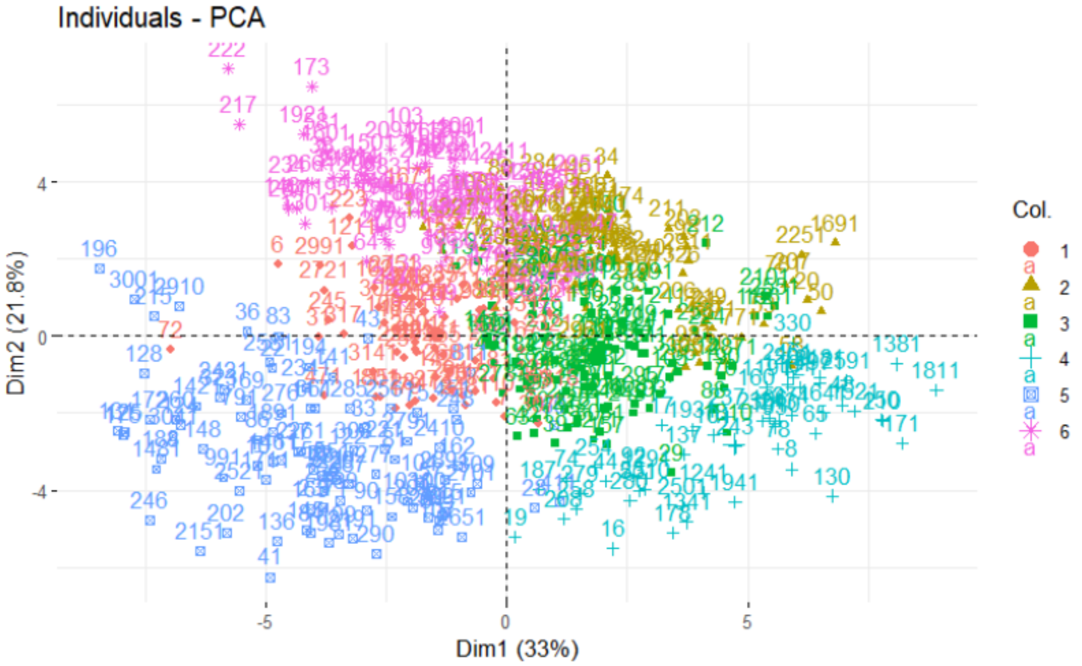

Figura 3.24: Individuos segundo clústers nas 2 primeiras compoñentes principais

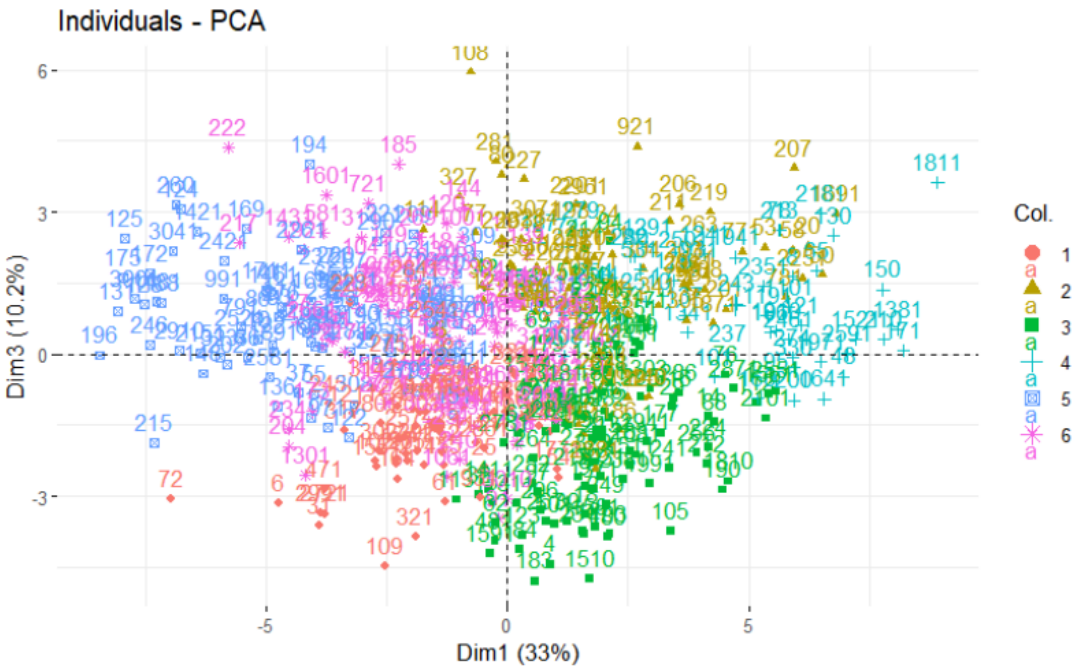

Figura 3.25: Individuos segundo clústers nas 2 primeiras compoñentes principais

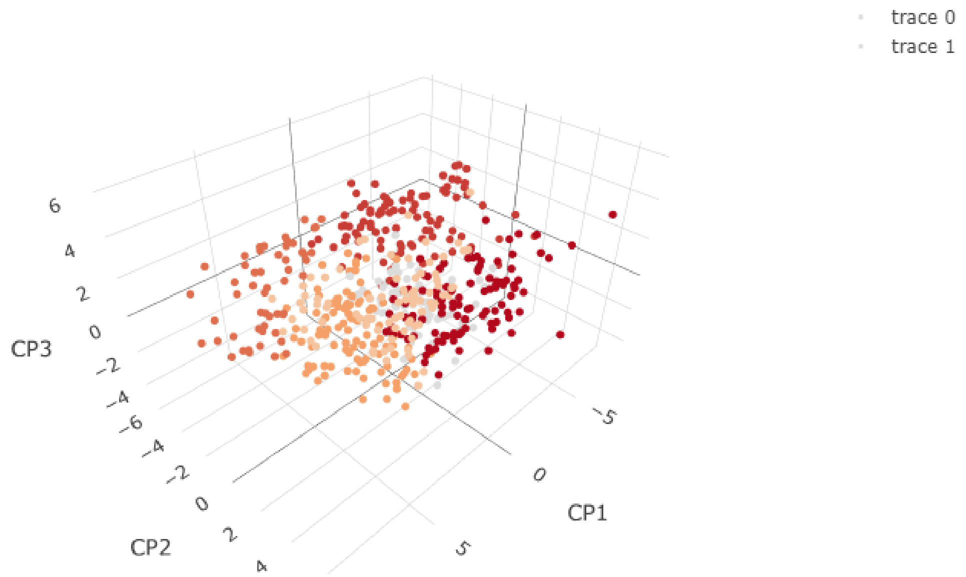

Figura 3.26: Individuos segundo clústers nas 3 primeiras compoñentes principais

Neste caso, observamos patróns claros de agrupamento por localización nas compoñentes principais. Por exemplo, os individuos do cluster 4 sitúanse maioritariamente no terceiro cuadrante na visualización 2D coa primeira e segunda compoñentes, é dicir, con valores negativos de ambas. O cluster 6, por exemplo, corresponde cos xogadores máis destacados na segunda compoñente.

Esta partición, na que podemos observar que os clusters se agrupan ben nas compoñentes principais, si que parece ser máis correcta. Esta tamén foi a valoración do titor profesional.

3.3.3 Significación dos clusters

Idealmente, cada un destes clusters engloba a un tipo de xogador homóxeno, cun significado na vida real. O obxectivo desta sección é extraer ese significado de cada un deles en base á localización dos individuos nas compoñentes principais, xa que poderemos intuír unha serie de patróns. Mostramos as áreas que forman os individuos agrupados sobre as compoñentes principais para xustificar as conclusións que vaiamos extraer:

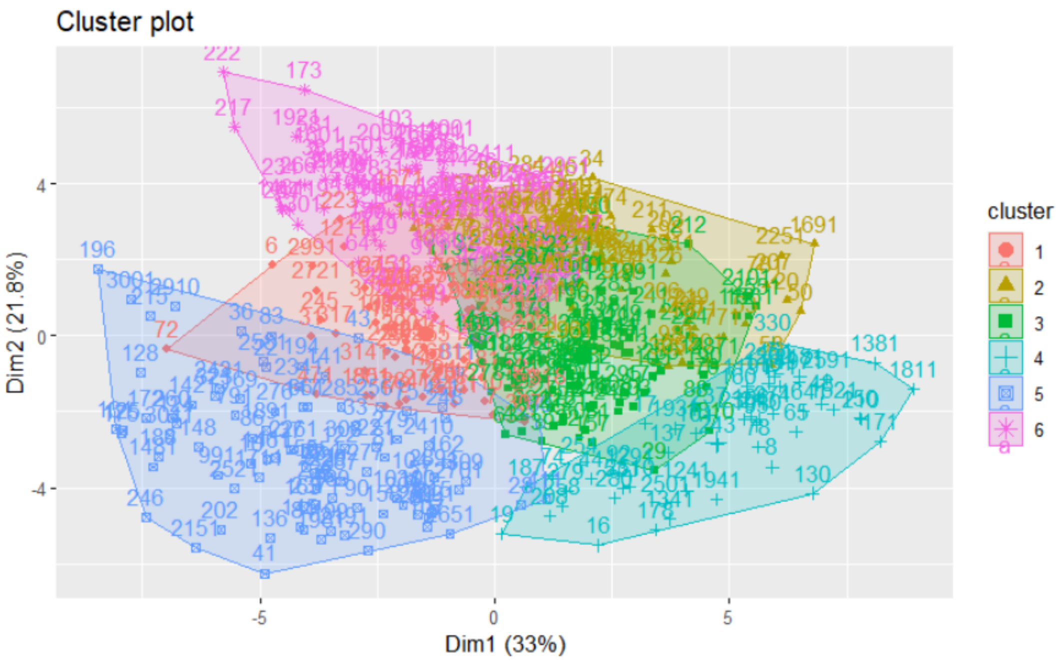

Figura 3.27: Área ocupada por cada clúster nas 2 primeiras compoñentes principais

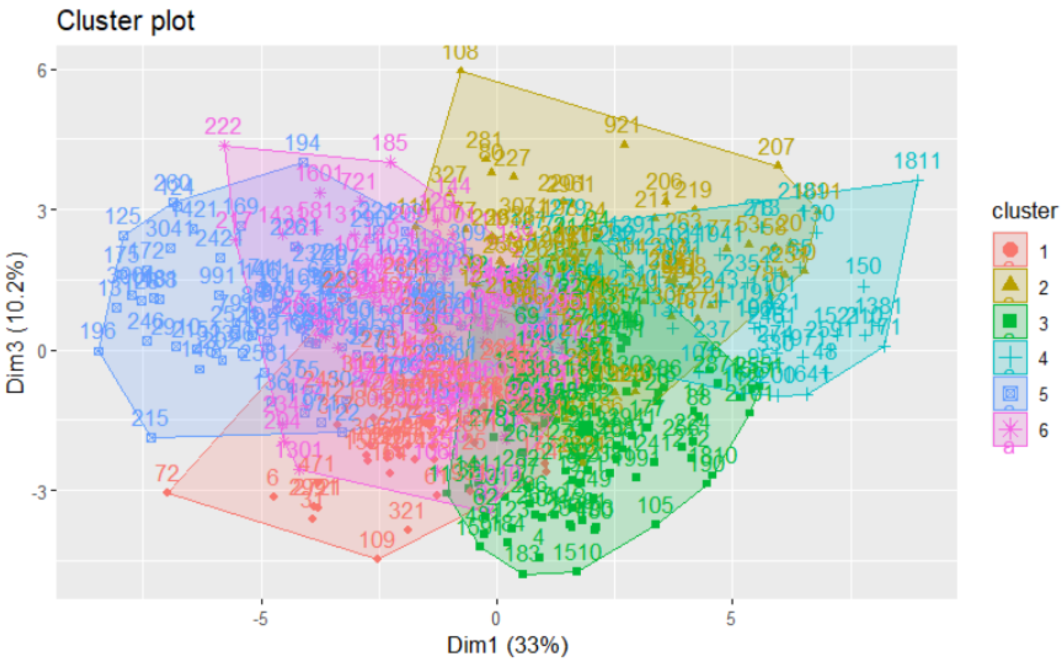

Figura 3.28: Área ocupada por cada clúster na primeira e terceira compoñentes principais

Clúster 1

Podemos observar na figura superior que os individuos do clúster 1 están relacionados maioritariamente cunha puntuación lixeiramente negativa na primeira compoñente principal. Esta, como comentamos anteriormente, asóciase con xogadores con moito peso na anotación e principalmente interiores.

Vemos ademais que tamén predominan os valores positivos da segunda compoñente, aínda que están bastante distribuídos e hai tanto valores positivos coma negativos. Nesta segunda compoñente, os xogadores con valores positivos son especialistas na anotación exterior, e os de puntuación negativa son predominantemente interiores.

En canto aos valores destes xogadores na terceira compoñente, vemos como son principalmente valores negativos, que se relacionan con xogadores interiores pero que poden lanzar dende o exterior.

Tendo en conta a influencia das primeiras compoñentes principais, concluímos que este clúster contén xogadores que poden xogar tanto por dentro como por fóra, con bo lanzamento exterior pero capacidade reboteadora e anotación interior. Algúns exemplos son Nikola Mirotic, ala-pívot do FC Barcelona que destaca polo seu tiro de tres, e Tres Tinkle ou Thomas Scrubb no Obradoiro, que son xogadores que alternan a posición de alero e ala-pívot e poden xogar abertos.

Clúster 2

En canto aos individuos do clúster 2, estes están asociados a valores positivos tanto na primeira como na segunda compoñente principal. Valores positivos da primeira compoñente principal indica que non teñen un excesivo peso na anotación e suxire que son xogadores exteriores. Os valores positivos da segunda compoñente indican que son xogadores exteriores e que teñen o balón unha gran parte do tempo, xerando tamén moitas asistencias.

Ademais, este clúster destaca claramente nos valores da terceira compoñente, os cales son moi positivos. Esta compoñente está asociada ás asistencias e as pérdidas, que son características dos xogadores que teñen moito tempo o balón nas mans e que dirixen o xogo do equipo.

En conclusión, os individuos do clúster 2 serán bases que dirixen o xogo do equipo. Pertencen a este clúster xogadores coma Ricky Rubio (FC Barcelona) e Sergio Rodríguez (Real Madrid), considerados dous dos mellores bases da historia de España. No Obradoiro, algún exemplo sería Janis Strelnieks ou Oleksandr Kovliar.

Clúster 3

O clúster 3 está marcado por valores positivos da primeira dimensión, indicando que os xogadores que estean englobados dentro deste cluster non terán demasiada influencia na anotación do equipo e que poden anotar dende fóra.

Ademais, os valores da segunda compoñente están repartidos sobre o 0, indicando que neste cluster existen tanto xogadores con valores lixeiramente positivos (poden anotar de 3 puntos) coma lixeiramente negativos (xogadores máis interiores, con capacidade reboteadora).

Na terceira compoñente, destacan principalmente os valores negativos, o que indica que son xogadores con bo lanzamento de 3 puntos e principalmente interiores.

Esta combinación de valores fainos pensar que os individuos do clúster 3 serán xogadores destacados polo seu tiro exterior, pero que non serán xogadores de posicións exteriores, senón aleros, ala-pívots ou pívots. Algúns xogadores representativos serían Mike Tobey, que no FC Barcelona xoga maioritariamente de pívot pero destaca por ser un gran lanzador de 3 puntos, ou Alex Suárez no Obradoiro, que é un alero / ala-pívot que lanza moito dende o exterior.

Clúster 4

En canto aos individuos que forman o 4º clúster, podemos observar como se sitúan todos no cuarto cuadrante do figura 3.27, o que quere dicir que teñen valor positivo na primeira compoñente pero negativo na segunda. Isto indica que son xogadores interiores, con pouco lanzamento de 3, e con moi pouca influencia na anotación do equipo.

Ademais, a puntuación lixeiramente positiva asociada á terceira compoñente fai ver que estes xogadores non destacan pola súa anotación de tres puntos, e poden ser bos pasadores do balón.

En resumo, este clúster asóciase cos xogadores interiores con pouca participación no ataque do equipo. Por exemplo, no Obradoiro CAB un exemplo sería Rubén Guerrero, que foi o terceiro pívot este ano e non gozou de moito protagonismo.

Clúster 5

O quinto clúster é quizáis o máis homoxéneo de todos. Sitúase claramente aillado no terceiro cuadrante da figura 3.27, mostrando valores negativos nas dúas primeiras compoñentes principais. Isto quere dicir que os individuos agrupados neste cluster serán xogadores interiores con moito peso ofensivo e gran capacidade anotadora.

De igual maneira có clúster anterior, tamén teñen unha puntuación lixeiramente positiva na terceira compoñente, o que indica que non destacan polo seu tiro exterior.

Vendo todo isto, concluímos que o clúster 5 é o asociado tradicionalmente aos pívots, sendo xogadores interiores destacados ofensivamente. Claros exemplos serían Edy Tavares (Real Madrid, considerado o mellor pívot de Europa) e Willy Hernangómez (FC Barcelona). No Obradoiro CAB, o mellor exemplo é Artem Pustovyi, pívot ucraíno de 2.18 metros.

Clúster 6

Finalmente, o clúster 6 destaca por valores fortes en positivo da segunda compoñente, o cal amosa que son grandes anotadores exteriores cun peso ofensivo moi elevado e gran lanzamento de 3 puntos. Ese peso ofensivo tamén se observa nos valores positivos da primeira compoñente.

En canto á terceira compoñente, este grupo adoita ter valores positivos, o que fai intuír que vai estar formado por bos manexadores exteriores do balón, que no adoitan tirar tras pase senón que crean os seus propios tiros.

Unindo todo isto, asociamos este clúster a xogadores exteriores predominantemente anotadores, que crean os seus tiros e manexan moi ben o balón. Os mellores exemplos serían Markus Howard (Baskonia, máximo anotador da pasada Euroliga dende a posición de escolta) e Kyle Guy (Lenovo Tenerife). Jordan Howard sería o representante deste clúster no Obradoiro CAB.

3.4 Regresión

Tendo por obxectivo a mellora da planificación deportiva, a ferramenta ideal para logralo era a regresión. Máis concretamente, buscaremos optimizar o rendemento global do equipo nos partidos e trataremos de atopar unha melloría na confección dos quintetos de xogadores que forman xuntos na pista.

No primeiro caso, interesa coñecer cales eran as variables relevantes para alcanzar unha maior diferencia de puntuación nos partidos. Isto axuda enormemente a saber como enfocar os partidos, en que facetas do xogo centrarse... No segundo caso, o relevante era obter, dentro dos quintetos que xogaran menos minutos, cales deberían ter desfrutado de máis tempo na pista. É dicir, distinguir quintetos que aínda que non alcanzaron un bo rendemento no seu escaso tempo de xogo, xa sexa por mala sorte ou por factores externos (o rival, algún xogador lesionado...), si o alcanzarían no longo prazo.

3.4.1 Modelo lineal

Para coñecer que variables levan a mellores diferencias de puntos, o modelo ideal para logralo é un modelo lineal que mostre ademais o peso que ten cada variable. Axustamos así un modelo de regresión lineal múltiple coas 24 variables das que dispoñemos para o equipo en cada partido (engadindo ademais a diferenza media do rival, como explicamos no preprocesado inicial) e como resposta a diferenza de puntos resultante en cada partido. Non obstante, como xa comentamos anteriormente, é relevante realizar unha selección desas variables para coñecer cales son realmente importantes e o seu peso. Realizamos esta selección como foi explicado no capítulo de fundamentos teóricos, a través do AIC, e o modelo resultante foi o seguinte:

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)   
(Intercept)   -168.0929    36.6962  -4.581  0.00377 **
PPPos          575.9571   168.7223   3.414  0.01425 * 
PPT           -902.8402   267.3668  -3.377  0.01492 * 
PPT3           334.1754    93.7261   3.565  0.01185 * 
TAsist        -137.9866    32.4044  -4.258  0.00533 **
T2Asist         33.7737    33.6596   1.003  0.35441   
TL_Min        -498.3360   164.6903  -3.026  0.02322 * 
TL_F           181.9475    66.0222   2.756  0.03303 * 
RebTot           5.9287     1.0808   5.485  0.00154 **
RebOf           -1.7153     0.5339  -3.213  0.01830 * 
RebDef          -3.9946     0.8470  -4.716  0.00327 **
`%TL`            0.5902     0.4157   1.420  0.20546   
TO              -1.3859     0.7039  -1.969  0.09651 . 
Rec             -5.1916     2.0886  -2.486  0.04744 * 
TRea            -1.2884     1.1158  -1.155  0.29211   
FRea             9.2005     2.7423   3.355  0.01532 * 
RitmoXogo        2.4353     0.5542   4.394  0.00460 **
PtosRebDef      -1.2923     0.3762  -3.435  0.01389 * 
PtosRebOf       -3.0601     0.6800  -4.500  0.00410 **
mediaDifRival   -2.3675     0.8014  -2.954  0.02548 * 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.597 on 6 degrees of freedom
Multiple R-squared:  0.9708,    Adjusted R-squared:  0.8784 
F-statistic: 10.51 on 19 and 6 DF,  p-value: 0.00394
```

Figura 3.29: Coeficientes do modelo tras selección de variables

Vemos que este modelo ten un gran R^2, de aproximadamente un 88%, pero hai claros problemas de multicolinealidade. Xa o intuímos polos coeficientes das variables (hai algúns que non teñen sentido, como que os puntos por tiro teñan un peso negativo na diferenza de puntos obtida), pero comprobamos cos Factores de Inflación da Varianza e comparándoos co limiar comentado na fundamentación teórica. As variables maiores que ese limiar e o seu correspondente FIV son:

```
    PPPos        PPT       PPT3     TL_Min       TL_F     RebTot      RebOf     RebDef      `%TL`        Rec       FRea mediaDifRival
2073.66960 3312.31386 1173.99313 1172.79623  958.62531   58.21841   37.48260   91.36547   37.28885   86.11981  185.24892      38.07914
```

Figura 3.30: Variables con alta multicolinealidade

Isto non sorprende, xa que sabíamos que este método de selección de variables podía levar a este problema. A pesar de que este modelo si sería válido para predicir, ademais cun coeficiente de determinación elevado, o obxectivo era coñecer o peso das variables relevantes, o cal non se podería facer con este problema da multicolinealidade. Así, realizamos unha segunda eliminación das variables en base a este problema da multicolinealidade, obtendo o seguinte modelo:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)   
(Intercept)   -66.1422    21.3553  -3.097  0.00622 **
PPT3           19.1512     6.4180   2.984  0.00796 **
RebTot          0.6885     0.2686   2.563  0.01955 * 
TO             -0.9204     0.4146  -2.220  0.03951 * 
Rec             1.7261     0.4415   3.909  0.00103 **
FRea            0.8809     0.3639   2.421  0.02626 * 
PtosRebOf      -0.8911     0.3339  -2.669  0.01565 * 
mediaDifRival  -0.6352     0.2321  -2.737  0.01354 * 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.034 on 18 degrees of freedom
Multiple R-squared:  0.7536,	Adjusted R-squared:  0.6578 
F-statistic: 7.867 on 7 and 18 DF,  p-value: 0.000203
```

Figura 3.31: Modelo lineal final

Vemos que o R^2 é menor, é dicir, sería peor para predicir, pero aquí si que podemos extraer unha análise das variables resultantes:

- **Intercept**: O seu coeficiente indica que, se o resto de variables tivesen valor 0, o Obradoiro CAB perdería o partido por 66.1422 puntos.
- **PPT3**: Esta variable reflexa os puntos que se obteñen por cada tripla. É dicir, se o equipo lanza 3 tiros de 3 puntos e só anota un, sería un valor de 1 punto por tiro de 3 (PPT3). O coeficiente desta variable é o maior de todos, e indica que un aumento unitario na variable PPT3 aumenta en 19.1512 a diferenza de puntos esperada para o equipo.
- **RebTot**: Neste caso, a variable indica a porcentaxe de rebotes totais capturados polo equipo. Un aumento dunha unidade nesta porcentaxe reporta un 0.6885 máis na diferenza de puntos do equipo.

- **TO**: Esta é a primeira das variables que ten un peso negativo, o cal ten sentido xa que reflexa o número de pérdidas de balón totais que cometeu o equipo. Por cada pérdida, a diferenza esperada para o Obradoiro CAB baixa en 0.9204 puntos.

- **Rec**: Esta variable representa as recuperacións totais logradas polo Obradoiro CAB, e volve a ter un peso positivo. Cada recuperación, retorna 1.7261 puntos á diferenza esperada.

- **FRea**: As faltas realizadas, que é o que representa esta variable, tamén ten un peso positivo. Cada falta realizada aumenta a diferenza esperada en 0.8809 puntos.

- **PtosRebOf**: Esta variable indica o número de puntos que encaixa un equipo tras un rebote ofensivo do rival, é dicir, cando non colles o rebote despois dun tiro do rival e te anotan. Esta medida específica de puntos recibidos é significativa cun peso negativo, e cada punto reporta unha diferenza esperada menor en 0.8911 puntos.

- **mediaDifRival**: Esta é a métrica que mide o nivel do rival a través da súa diferenza media. Loxicamente, todos sabemos que a maior nivel do rival, menor diferenza se espera obter a favor. Formalmente, cada punto máis na diferenza media do rival fai que a diferenza esperada decreza en 0.6352 puntos.

Con este análise, o corpo técnico do Obradoiro CAB sabe onde debe centrar os seus esforzos para mellorar o rendemento colectivo. Entraremos en máis detalle no capítulo de conclusións.

3.4.2 Random Forest

O segundo obxectivo era a procura de quintetos que poderían mellorar o seu rendemento contando cunha maior participación. Como o relevante aquí non era o peso específico das variables, senón a predición final, o modelo escollido foi o Random Forest.

Realizaremos dous cortes distintos por minutos xogados para seleccionar os quintetos. O primeiro corte será de 10 minutos ou máis, sendo este un tempo de xogo considerable para un mesmo grupo de 5 xogadores a partir do cal poder extraer conclusións sen medo ao ruído. O segundo corte será de entre 5 e 10 minutos, e serán os quintetos sobre os que realizaremos predicións para comprobar se o seu rendemento a longo prazo será bo. Estes datos poderán ter algo de ruído, pero mínimo, ao fixar tamén un corte por minutos razoable. Así, co modelo adestrado en base aos primeiros quintetos, predicimos sobre os segundos e observamos que valor obterían se disputaran máis minutos.

Partindo das variables de rendemento e da configuración de tipos de xogadores en cada quinteto (o número de xogadores de cada cluster que hai no quinteto, como explicamos no preprocesado), volvemos tomar a diferenza de puntos como variable resposta. Neste caso, esta variable resposta ten unha distribución peculiar que se pode beneficiar dunha transformación que mellore os resultados:

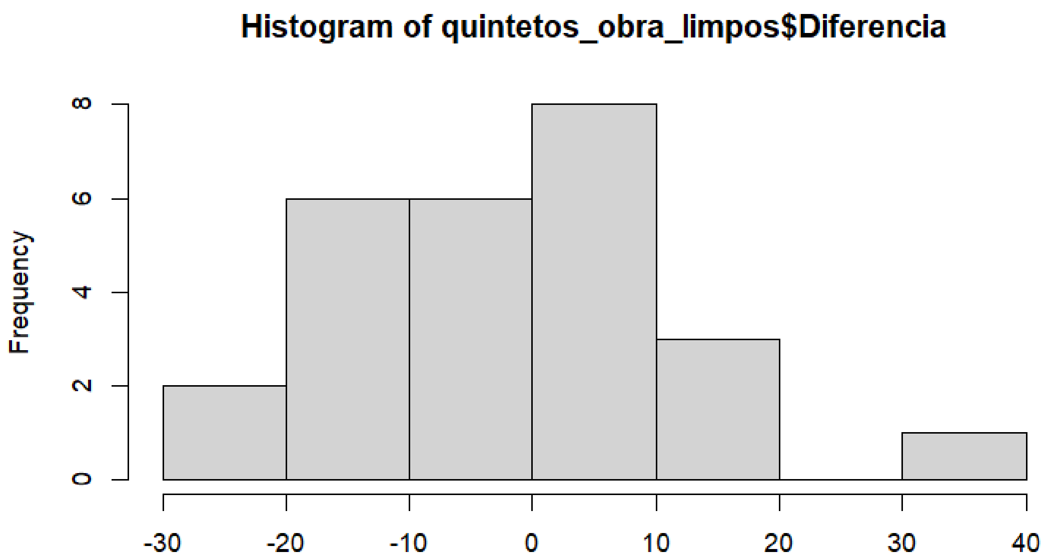

Figura 3.32: Histograma da variable resposta

Esta variable presenta asimetría, xa que a cola da dereita é maior ca cola da esquerda, e ademais presenta un dato atípico na cola maior. Isto nos leva á conclusión de que unha transformación logarítmica tería un efecto positivo, xa que esta é beneficiosa precisamente nestes casos. Non obstante, coma a variable resposta pode tomar valores negativos, houbo que realizar unha pequena transformación previa. Sumamos a todos os valores o mínimo que toma esa variable, e logo aplicamos a función *logp1*, propia de R. Esta función calcula log(1+x) no lugar do logaritmo só, para evitar calcular o logaritmo de 0 no caso da observación con valor mínimo. Así, solucionamos o problema dos valores negativos pero mantemos as diferenzas entre observacións.

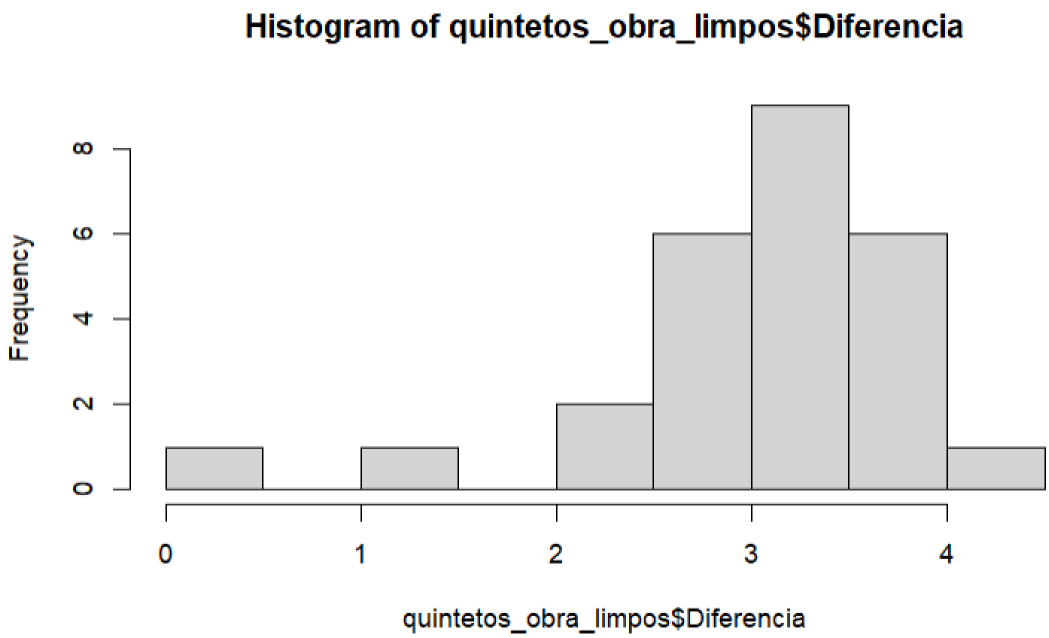

Figura 3.33: Histograma da variable resposta transformada

Ademáis, realizamos unha selección de variables que evite sobreaxuste e retorne mellores resultados, tal e como se explicou no capítulo anterior. Así, as variables que este proceso de selección considerou relevantes foron:

- **RebTot**: Porcentaxe de rebotes totais capturados polo quinteto
- **PPMin**: Puntos por minuto anotados polo quinteto
- **PPT**: Puntos por tiro
- **PPT2**: Puntos por tiro de 2
- **PPT3**: Puntos por tiro de 3
- **RebDef**: Porcentaxe de rebotes defensivos capturados
- **RebOf**: Porcentaxe de rebotes ofensivos capturados polo equipo
- **PPPos**: Puntos por posesión
- **T3Asist**: Ratio de triplas anotados que proveñen de asistencias
- **T2Asist**: Ratio de tiros de 2 puntos anotados que proveñen de asistencias

- **TAsist**: Ratio de tiros anotados que proveñen de asistencias
- **Minutos**: Minutos xogados polo quinteto
- **Asist**: Asistencias logradas polo quinteto
- **TL_Min**: Ratio de tiros libres por minuto
- **TO**: Pérdidas do quinteto
- **Cluster1**: Número de xogadores do clúster 1 presentes no quinteto
- **Cluster2**: Número de xogadores do clúster 2 presentes no quinteto
- **Cluster3**: Número de xogadores do clúster 3 presentes no quinteto
- **Cluster4**: Número de xogadores do clúster 4 presentes no quinteto
- **Cluster5**: Número de xogadores do clúster 5 presentes no quinteto
- **Cluster6**: Número de xogadores do clúster 6 presentes no quinteto

Como vemos, a configuración enteira do quinteto foi considerada relevante para predicir a diferenza.

Adestrando o modelo despois cos datos destas variables e utilizando validación cruzada, o R^2 obtido foi de 0.764382, o que quere dicir que este modelo explica máis dun 76% da varianza na diferenza media. Esta métrica foi calculada desfacendo a transformación dos datos e aplicando a fórmula. Visualizamos agora os resultados, situando no eixo X as observacións e no eixo Y as predicións:

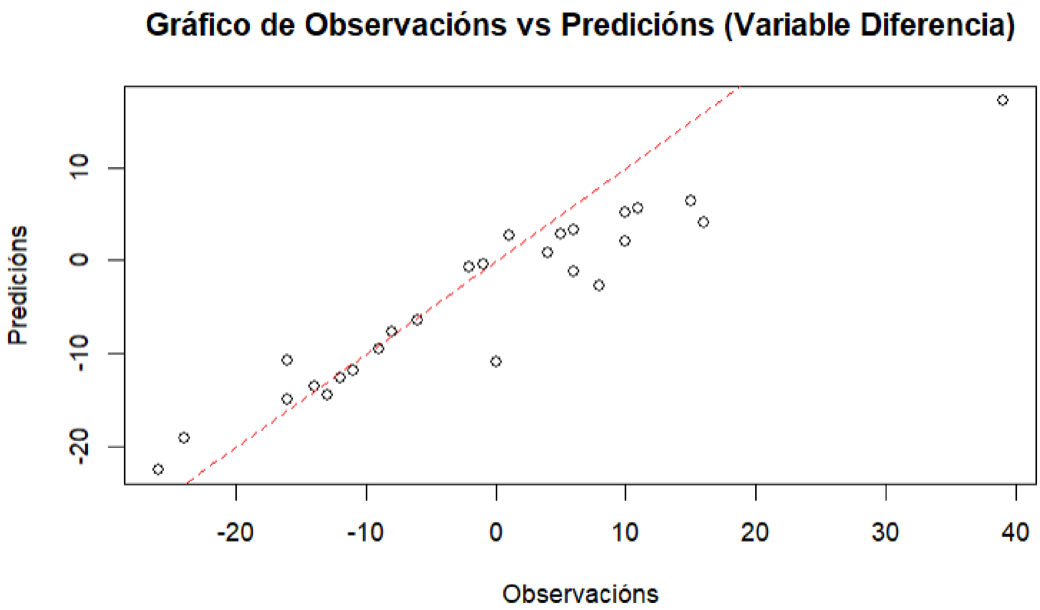

Figura 3.34: Observacións vs Predicións (diferenza total como resposta)

Debuxamos a recta x=y, que representaría unha predición perfecta con respecto ás observacións. Observamos que todas as observacións seguen bastante ben esta recta, menos un caso atípico que obtivo moito maior resultado ca o predito. Decatámonos de que esta diferenza correspóndese co dato atípico que víamos no histograma inicial, e que este quinteto foi o que máis minutos disputou con moita diferenza sobre o segundo, o que pode que estea levando ao modelo a malos resultados nesta predición concreta.

Para mitigar este problema, imos repetir esta análise cos mesmos datos e a mesma transformación sobre a resposta, pero traballando coa diferenza por minuto no lugar da diferenza total. Así, a selección de variables devolveu como relevantes só as seguintes:

- **RebTot**
- **PPMin**
- **TAsist**
- **T3Asist**
- **PPT2**
- **PPT3**
- **Clúster 3**

- **Clúster 5**

Así, considerou relevantes só o número de xogadores de dous clústers e moitas menos variables ca inicialmente. Comprobando o coeficiente de determinación neste caso, vemos que subiu ata un 0.8690698, o que quere dicir que explica aproximadamente un 87% da varianza da diferenza por minuto, un gran resultado. Volvemos a visualizar as predicións contra as observacións, para ver se o problema inicial foi solucionado:

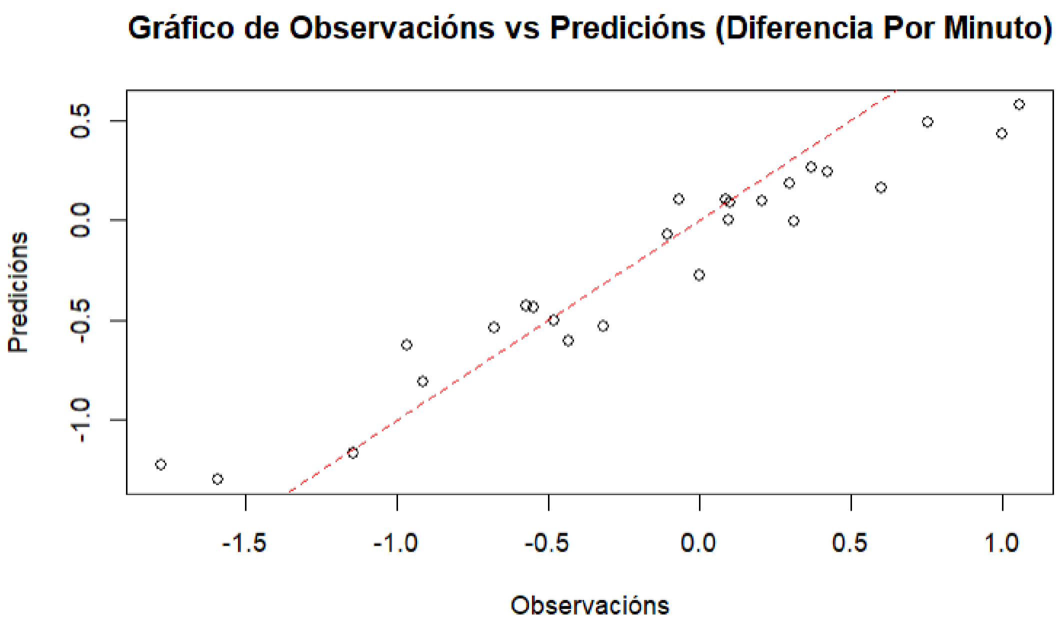

Figura 3.35: Observacións vs Predicións (diferenza por minuto como resposta)

Vemos como agora non hai ningún dato que se afaste demasiado da recta x=y, como si pasaba anteriormente. Adoptamos este modelo coma o mellor para o noso obxectivo e pasamos á fase de predicións sobre os quintetos de menor minutaxe.

Como demostración, imos facer dúas predicións concretas. Obviamente, mostraremos o quinteto con mellor predición de todos os que xogaron menos de 10 minutos, como exemplo de quinteto que podería aportar un bo rendemento se pasase a ter unha importancia maior. Pero ademais, imos realizar unha segunda predición sobre os quintetos formados por configuracións non probadas no primeiro grupo de quintetos, os que xogaron maior tempo. É dicir, imos comprobar que quintetos cunha configuración de clusters que non foi probada máis de 10 minutos aporta maior rendemento. Isto é relevante para o Obradoiro CAB, xa que propón novas combinacións de tipos de xogadores que paguen a pena a nivel de resultado.

No primeiro caso, o quinteto con mellor predición que xogou menos de 10 minutos sería o formado por Thomas Scrubb, Fernando Zurbriggen, Janis Timma, Artem Pustovyi e Rigoberto Mendoza; cunha diferenza esperada de 0.325403. Para facernos unha idea, a media de diferenza por minutos dos quintetos con máis de 10 minutos foi de -0.16558 e a mediana de -0.03309, polo que este quinteto ofrecería un rendemento superior a moitos dos que xogaron máis ca eles. Mostramos visualmente onde se situaría este quinteto, que corresponde á liña azul:

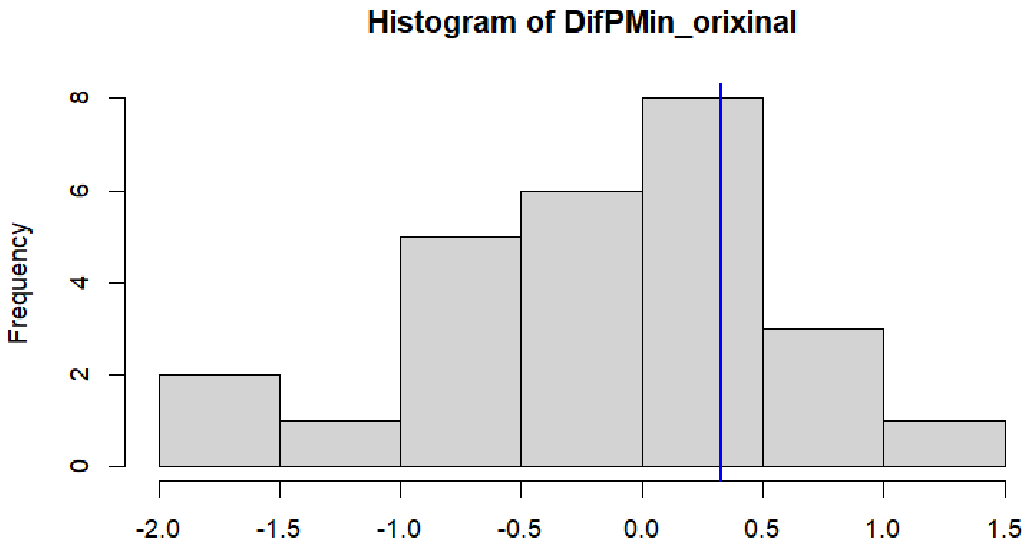

Figura 3.36: Situación do quinteto con maior predición no histograma

Non obstante, aínda que este quinteto concreto non xogou máis de 10 minutos, un quinteto con configuración de clusters idéntica si que o fixo. Buscando a configuración nova con mellor rendemento esperado, obtemos o quinteto formado por Thomas Scrubb, Devon Dotson, Marek Blazevic, Rigoberto Mendoza e Alex Suarez; cunha diferenza esperada de 0.2767701. De novo, este resultado está por enriba da media e a mediana dos quintetos con máis de 10 minutos. Engadimos á visualización anterior unha liña vermella para mostrar o quinteto con mellor predición e configuración nova:

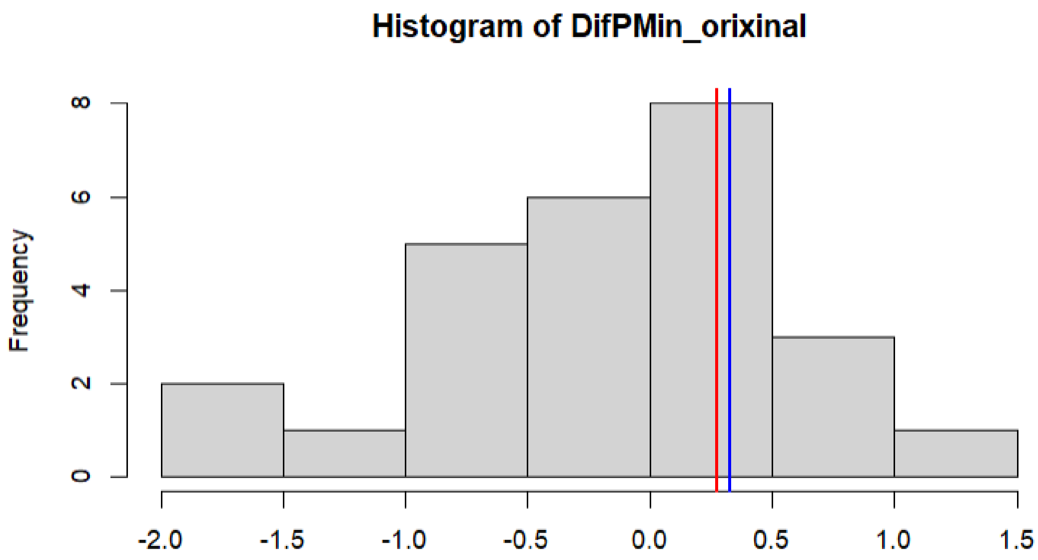

Figura 3.37: Situación do quinteto con maior predición e configuración non probada no histograma

Capítulo 4
Conclusións

DERRADEIRO capítulo da memoria, onde se presentará a situación final do traballo, as leccións aprendidas, a relación coas competencias da titulación en xeral e a mención en particular, posibles liñas futuras,...

Apéndices

Apéndice A

Material adicional

Exemplo de capítulo con formato de apéndice, onde se pode incluír material adicional que non teña cabida no corpo principal do documento, suxeito á limitación de 80 páxinas establecida no regulamento de TFGs.