MVA PROJECT INTERPRETATIONS

Index

Content

[1.](#_Toc185281705)  [Introduction](#_Toc185281705)  [1](#_Toc185281705)

[2.](#_Toc185281706)  [Exploratory Analysis](#_Toc185281706)  [2](#_Toc185281706)

[Two data sets are used: one containing general information about the notable NBA players (name, college, team), and another containing their physical characteristics and performance (height, weight, points, rebounds, assists, and TS%).](#_Toc185281707)  [2](#_Toc185281707)

[3.](#_Toc185281708)  [Multivariate Analysis](#_Toc185281708)  [8](#_Toc185281708)

[**a)**](#_Toc185281709)  [**Objective 1: Analyze general patterns in player performance in relation to age and physical variables.**](#_Toc185281709)  [8](#_Toc185281709)

[**b)**](#_Toc185281710)  [**Objective 2: Identify the traits of those players with exceptional metrics.**](#_Toc185281710)  [9](#_Toc185281710)

[**c)**](#_Toc185281711)  [**Objective 3: Compare performance metrics between the two universities with the highest number of outstanding players.**](#_Toc185281711)  [10](#_Toc185281711)

[**d)**](#_Toc185281712)  [**Objective 4: Determine whether performance metrics vary significantly between the 4 teams with the highest number of featured players.**](#_Toc185281712)  [11](#_Toc185281712)

[**e)**](#_Toc185281713)  [**Objective 5: Find components that reduce the dimensionality of the data.**](#_Toc185281713)  [13](#_Toc185281713)

[**f)**](#_Toc185281714)  [**Objective 6: Establish segments that group players according to performance metrics and their physical characteristics.**](#_Toc185281714)  [15](#_Toc185281714)

[**g) Objective**](#_Toc185281715)  [7 **: Analyze the relationship between the levels of defensive rebounds, level of offensive rebounds, level of assists and the levels of points obtained.**](#_Toc185281715)

[h)](#_Toc185281716)  [**Objective 8:** Determine the characteristics of player performance metrics that influence the team's overall game impact.](#_Toc185281716)  [21](#_Toc185281716)

[4.](#_Toc185281717)  [Conclusions](#_Toc185281717)  [23](#_Toc185281717)

# Introduction

The present work aims to perform a multivariate analysis of the performance of outstanding NBA players. This analysis explores general patterns, identifies outliers, and segments players into groups based on their performance and physical characteristics. Additionally, the association between different metrics is evaluated and performances are compared between teams and universities. The variables of special interest are the performance metrics in the game of basketball. For the present work, these variables correspond to the number of games played throughout the season (gp), average number of points scored per game (pts), average number of rebounds per game (reb), average number of assists per game (ast), percentage of available offensive rebounds that the player captured while on the court (oreb\_pct), percentage of available defensive rebounds that the player captured while on the court (dreb\_pct), a measure of the player's shooting efficiency that takes into account free throws, 2- and 3-point shots (ts\_pct), and percentage of field goals from teammates assisted by the player while on the court (ast\_pct). On the other hand, physical variables such as age, height, and weight are available. Finally, categorical variables such as university, school, year of contract, round in which the player plays, and jersey number are also available.

# Exploratory Analysis

Data Description

Two data sets are used: one containing general information about the top NBA players (college, team), and another containing their physical characteristics and performance (height, weight, points, rebounds, assists, and TS%).

Distributions

The distribution of the data according to each variable is presented below.

Gráfico, Histograma

Descripción generada automáticamenteThe age variable presents a positive asymmetric distribution, that is, the ages of the players are concentrated at lower ages. Likewise, it is observed that the age of the players fluctuates between 19 and 40 years, which are concentrated around 27.07 years, while 25% of the players have a maximum age of 24 years; 50%, a maximum age of 27.07 years; and 75%, a maximum age of 30 years.

The height variable presents negative asymmetry, that is, the height of the players is concentrated in high values. Likewise, it is observed that the height of the players fluctuates between 175 and 221 cm, which are concentrated around 200.8 cm, while 25% of the players have a maximum height of 193 cm; 50%, a maximum height of 203.2 cm; and 75%, a maximum height of 108.86 cm.

The weight variable of the players tends to be symmetrical, that is, the weights of the players are concentrated around the average of the weights. Likewise, it is observed that the weight of the players fluctuates between 73.03 and 133.81 kg, which are concentrated around 100.87 kg, while 25% of the players have a maximum weight of 91.17 kg; 50%, a maximum weight of 101.6 kg; and 75%, a maximum weight of 133.81 kg.

Texto

Descripción generada automáticamente

Gráfico

Descripción generada automáticamente con confianza mediaThe variable number of games played presents negative asymmetry, that is, the number of games in which the players participate is concentrated in high values. Likewise, it is observed that the number of games fluctuates between 1 and 82, which are concentrated around 54.92 while 25% of the players participated in a maximum of 39 games; 50% participated in a maximum of 61 games; and 75% participated in a maximum of 76 games.

The points obtained variable presents a positive asymmetric distribution, that is, the points obtained by the players are concentrated in low values. Likewise, it is observed that the points obtained fluctuate between 0 and 28.7, which are concentrated around 7,958 points, while 25% of the players obtained a maximum of 3.7 points; 50% obtained a maximum of 6.7 points; and 75% obtained a maximum of 11.3 points.

The variable number of rebounds presents a positive asymmetric distribution, that is, the rebounds of the players are concentrated in low values. Likewise, it is observed that the rebounds made fluctuate between 0 and 14.4, which are concentrated around 3,522 rebounds, while 25% of the players made a maximum of 2.8 rebounds; 50% made a maximum of 2.8 rebounds; and 75% made a maximum of 4.7 rebounds.

Texto

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamenteThe variable points assists presents a positive asymmetric distribution, that is, the assists of the players are concentrated in low values. Likewise, it is observed that the assists made fluctuate between 0 and 11.1, which are concentrated around 1.809 while 25% of the players made a maximum of 0.5 assists; 50% made a maximum of 1.1 assists; and 75% made a maximum of 2.4 assists.

The variable percentage of offensive rebounds obtained presents a positive asymmetric distribution, that is, the proportion of offensive rebounds of the players is concentrated in low values. Likewise, it is observed that said proportion of rebounds fluctuates between 0 and 0.5, which are concentrated around 0.05517 while 25% of the players made a maximum of 0.02; 50% made a maximum of 0.04; and 75% made a maximum of 2.4.

The percentage of defensive rebounds obtained variable presents a positive asymmetric distribution, that is, the proportion of defensive rebounds of the players is concentrated in low values. Likewise, it is observed that said proportion of rebounds fluctuates between 0 and 0.5, which are concentrated around 0.1434 while 25% of the players made a maximum of 0.094; 50% made a maximum of 0.1360; and 75% made a maximum of 0.187.

Una captura de pantalla de un celular con texto e imagen

Descripción generada automáticamente con confianza media

Gráfico

Descripción generada automáticamente con confianza mediaThe variable percentage of team plays obtained presents a symmetrical distribution, that is, the proportion of plays is concentrated around the mean. Likewise, it is observed that this proportion fluctuates between 0 and 0.368, which are concentrated around 0.1886 while 25% of the players made a maximum of 0.154; 50% made a maximum of 0.187; and 75% made a maximum of 0.222.

The variable shooting efficiency presents a negative asymmetric distribution, that is, the shooting efficiency of the players is concentrated in high values. Likewise, it is observed that said efficiency fluctuates between 0 and 0.931, which are concentrated around 0.5043, while 25% of the players obtained a maximum efficiency of 0.4780; 50% obtained a maximum efficiency of 0.52; and 75% obtained a maximum efficiency of 0.555.

The variable passes from teammates presents a positive asymmetric distribution, that is, passes from teammates to players are concentrated in low values. Likewise, it is observed that these passes fluctuate between 0 and 0.6, which are concentrated around 0.1327 while 25% of the players received a maximum of 0.065 passes; 50% received a maximum of 0.1 passes; and 75% received a maximum of 0.183 passes.

Una captura de pantalla de un celular con texto e imágenes

Descripción generada automáticamente con confianza baja

Regarding the categorical variables, the data set includes the university of origin and the abbreviated name of the team. The largest number of players represented were the universities of Kentucky, Duke and North Carolina. The teams with the largest number of players represented were DAL, CHA, MIN and NYK.

Gráfico

Descripción generada automáticamente

2.4 Missing Values and Imputation

The metric variables did not present missing values, so it was not necessary to carry out the imputation process. The variables year of call-up to the team, jersey number and game round presented missing values. However, they were not included in the study because they do not present any relationship with the other variables.

Interfaz de usuario gráfica

Descripción generada automáticamente

# Multivariate Analysis

The objectives are then developed using multivariate methods.

## **Objective 1 : Analyze general patterns in player performance in relation to age and physical variables.**

This objective is proposed because it is reasonable to assume that performance metrics are associated with physical variables such as age, weight and height. To do this, the correlation matrix specified for the physical variables is used to discover which of the performance metrics is best associated.

It is observed that the variable age correlates positively and in a low but significant way with the score of influence on the team; it correlates negatively and in a low but positive way with the percentage of offensive rebounds, as well as with the percentage of shots of their teammates. This implies that, the older the players, the more positive their impact, the lower the percentage of offensive rebounds and the fewer passes received by their teammates.

Texto

Descripción generada automáticamente

Regarding the variable height, it is observed that it presents a moderate positive correlation with the percentage of offensive and defensive rebounds (both related to ball possession with rebounds) and a moderate negative correlation with assists and passes received by teammates. This implies that the taller the players are, the higher the percentage of offensive and defensive rebounds and the lower the number of assists, as well as passes received by their teammates.

Texto

Descripción generada automáticamente

Regarding the variable weight, it is observed that it presents a moderate positive correlation with the percentage of offensive and defensive rebounds (both related to ball possession with rebounds) and a moderate negative correlation with assists and passes received by teammates. This implies that the taller the players, the higher the percentage of offensive and defensive rebounds and the lower the number of assists, as well as passes received by their teammates.

Texto

Descripción generada automáticamente

## Gráfico, Gráfico de dispersión Descripción generada automáticamente**Objective 2 : Identify the traits of those players with exceptional metrics.**

The aim of this study is to discover whether there are players within the data set with the most outstanding performance metrics. To do this, multivariate outliers will be identified.

The Mahalanobis distance metric was used to identify those players with multivariate atypical characteristics. According to the R report, 236 players are identified with characteristics different from the others. These make up 5.54% of the dataset.

Texto

Descripción generada automáticamente

Next, the maximum values of the players' performance metrics are compared to highlight some of the features of this group of players with atypical values. Class “0” refers to players not considered as atypical, while class “1” refers to players with atypical characteristics. These players have a maximum value lower than the maximum games played of “normal” players. To belong to the group of special players, they must stand out in at least the highest points obtained, the highest number of rebounds, the highest number of assists, and the highest percentage of offensive and defensive rebounds.

|  |  |  |
| --- | --- | --- |
| Cluster | 0 | 1 |
| Gp | 82 | 80 |
| Pts | 28.1 | 28.7 |
| Reb | 12.4 | 14.4 |
| Ast | 9.7 | 11.1 |
| oreb\_pct | 0.183 | 0.5 |
| dreb\_pct | 0.303 | 0.5 |

## **Objective 3 : Compare performance metrics between the two universities with the highest number of outstanding players.**

Through this objective, we intend to determine whether universities with a greater number of outstanding players present a difference between performance metrics. To do this, the T2 Hotelling test was applied.

Kentucky and Duke were identified as the universities with the most standout players.

Texto

Descripción generada automáticamente con confianza media

To apply Hotelling's T2 test, we proceeded to verify whether the performance metrics data fit a Normal distribution, where the following hypotheses are contrasted:

H0: The data of the i-th variable of the j-th university fits a Normal distribution.

H1: The data of the i-th variable of the j-th university does NOT fit a Normal distribution.

The following table shows the p-values for the Kolmogorov test by university. It can be observed that the p-values for the game performance metrics are greater than 0.05, that is, the variables fit the Normal distribution.

|  |  |  |
| --- | --- | --- |
|  | Kentucky | Duke |
| **Variable** | **Pvalue** | **Pvalue** |
| gp | 0.3992 | 0.478 |
| pts | 0.6101 | 0.4093 |
| reb | 0.8768 | 0.4347 |
| ast | 0.05158 | 0.3965 |
| oreb\_pct | 0.4932 | 0.1451 |
| dreb\_pct | 0.8059 | 0.4920 |
| ts\_pct | 0.3673 | 0.6856 |
| ast\_pct | 0.1524 | 0.702 |

With this result, the hypothesis test of the vector of means of performance metrics for both groups is applied. The Hotelling d test is not significant, that is, no differences are found in the vector of means of the performance metrics of the players at the universities of Kentucky and Duke.

Texto

Descripción generada automáticamente con confianza media

## **Objective 4 : Determine whether performance metrics vary significantly between the 4 teams with the highest number of featured players.**

Through this objective, we intend to determine whether universities with a greater number of outstanding players present a difference between the performance metrics. To do this, the MANOVA test was applied.

The teams were identified as the teams with the most outstanding players.

Imagen que contiene Interfaz de usuario gráfica

Descripción generada automáticamente

To apply the MANOVA test, we proceeded to verify whether the performance metrics data fit a Normal distribution, where the following hypotheses are contrasted:

H0: The data of the i-th variable of the j-th team fits a Normal distribution.

H1: The data of the i-th variable of the j-th team does NOT fit a Normal distribution.

The following table shows the p-values for the Kolmogorov test by university. It can be observed that the p-values for the game performance metrics are greater than 0.05, that is, the variables fit the Normal distribution.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **DAL** | **CHA** | **MIN** | **NYK** |
| **Variable** | **Pvalue** | **Pvalue** | **Pvalue** | **Pvalue** |
| gp | 0.6686 | 0.7695 | 0.6845 | 0.7071 |
| pts | 0.3907 | 0.5895 | 0.9849 | 0.3121 |
| reb | 0.9127 | 0.9788 | 0.9127 | 0.5858 |
| ast | 0.4415 | 0.3264 | 0.2219 | 0.1551 |
| oreb\_pct | 0.3017 | 0.6792 | 0.3698 | 0.4165 |
| dreb\_pct | 0.5639 | 0.8928 | 0.6123 | 0.4369 |
| ts\_pct | 0.05395 | 0.1294 | 0.1942 | 0.1386 |
| ast\_pct | 0.8704 | 0.411 | 0.5187 | 0.3995 |

With this result, the MANOVA hypothesis test is applied. The MANOVA test is not significant (p value = 0.807), that is, no difference is found in at least one team with the vector of means of the performance metrics of the players in the DAL, CHA, MIN and NYK teams. Likewise, the individual ANOVAs show p values greater than 0.05.

Texto

Descripción generada automáticamente

Texto, Carta

Descripción generada automáticamente

## **Objective 5 : Find components that reduce the dimensionality of the data.**

To achieve this objective, physical variables such as age, weight and height, as well as performance metrics, were used. To avoid problems with measurement units, the program indicates that the data must be standardized beforehand, while categorical variables are used as supplementary variables.



The eigenvalue chart presents the eigenvalues associated with each component in decreasing order. According to the KMO rule, components with eigenvalues greater than 1 should be retained.

Gráfico, Gráfico de líneas

Descripción generada automáticamente

The unit circle graph shows the relationship between variables and generated components.

Diagrama, Dibujo de ingeniería

Descripción generada automáticamenteAccording to the graph, it is observed that dimension 1 explains 31.35% of the total variability while dimension 2 explains 25.33% of the total variability, together explaining 56.68% of the total variability. Regarding the correlation between variables and components, it is observed that to the right (in Dim 1): player\_weight, player\_height and dreb\_pct are positively correlated with each other and have a strong contribution to the first dimension; variables to the left: ast\_pct and ast have opposite contributions to the variables on the right; variables upwards (Dim 2): reb, gp, pts and ts\_pct and they contribute significantly to the second dimension; and variables towards the center: They have less variance and less influence on Dim 1 and Dim 2. From this, it can be suggested that groups of correlated variables are related to the player's physique such as player\_weight, player\_height; offensive performance such as pts, ts\_pct, reb; and usage statistics such as usg\_pct, net\_rating.

To identify which variables best explain each component, the correlation between components and original variables was calculated and sorted according to the component with which the variables are best associated. For example, the variables player\_height , player\_weight, ast oreb\_pct,dreb\_pct and ast\_pct are most associated with dimension 1.

Texto

Descripción generada automáticamente con confianza media

According to the correlations between variables and scores of the dimensions according to the context of NBA player data:

* Dim.1: Size and Rebounding: This dimension captures characteristics related to the player's physical size (height and weight) and his ability to rebound (offensive and defensive).
* Dim.2: Game Production: This component represents the player's overall contribution in terms of games played, scoring, and shooting efficiency. It indicates the level of activity and performance on the court.
* Dim.3: Efficiency and Offensive Usage. This component describes how a player impacts the team's efficiency when on the court (net rating) and how involved he is in the offense (offensive usage).
* Dim.4: Age and Experience. It relates to the player's experience in the NBA, measured by his age.

## **Objective 6 : Establish segments that group players according to performance metrics and their physical characteristics.**

In order to find the best segments, the clustering methods were compared: cluster after PCA and Kmeans.

Using the cluster method after PCA, 3 segments are obtained, where the first is made up of 118 players, the second of 231 and the third of 120.

Texto

Descripción generada automáticamente

In order to select the best grouping process, the silhouette index was calculated. The first group has a silhouette value of 0.25, the second a silhouette value of 0.04 and the third group has a silhouette value of 0.09.

Texto

Descripción generada automáticamente

For this approach, an average silhouette indicator of 0.1 is obtained. The reason is that a number of observations below zero are observed, i.e., these observations are incorrectly assigned to the cluster.

**Gráfico

Descripción generada automáticamente**

Gráfico, Gráfico de líneas

Descripción generada automáticamenteLikewise, the number of clusters was calculated using k means through the pseudo indicator F. According to the graph, the value of this indicator is maximized in two groups, so that two groups are suggested through k-means.

The silhouette indicator was calculated in the same way. The first group had a silhouette value of 0.24 and the second group had a silhouette value of 0.20.

Texto

Descripción generada automáticamente

For this approach, an average silhouette indicator of 0.22 is obtained. The graph shows that in group 1, the observations are correctly assigned while in group 2 there is a minimum number of incorrectly assigned observations. Thus, the K means method generates better differentiated clusters.

Imagen que contiene Gráfico

Descripción generada automáticamente

In order to characterize the segments found, the means for both groups were calculated.

Texto

Descripción generada automáticamente

Based on the averages of the variables for the segments obtained, the following description can be made:

**Segment 1: This segment includes players with an average age** of 27.10 years, average height of 200.49 cm, and average weight of 100.58 kg, an average of 69.86 games , meaning they participate regularly during the season. Regarding **performance metrics** , this group includes players with 9.89 **points per game (pts)** , indicating a moderate offensive contribution; 4.19 rebounds **per game (reb)** , contributing in both defense and offense; 2.24 **assists per game (ast)** , indicating a reasonable participation in playmaking, -0.40 **Net Rating** , which indicates a close balance between points for and against while on the court, 5.34% **offensive rebound percentage (oreb\_pct)** ; 14.6% in **defensive rebound percentage (dreb\_pct),** suggesting greater defensive responsibility, 19.18% in usage **(usg\_pct)** , indicating that these players have a considerable role within the offense, 53.01%, efficient in their shots and 14.04% in **assists percentage (ast\_pct)** , that is, they actively participate in the distribution of the ball. Based on these characteristics, it is suggested that this segment be nominated as **"Consistent Players"** , since it groups players who have a balanced performance, with good participation in games and significant contributions in several areas of the game.   
**Segment 2: This segment includes players with an** average age of 27 years, similar to segment 1, average height of 201.33 cm and average weight of 101.47 kg, 24.65 **games played on average (gp), that is,** considerably less participation in the season. Regarding the **performance metrics,** in this group, the players present 4.04 points **per game (pts)** , a sign of a low offensive contribution; 2.17 rebounds **per game (reb)** : 2.17, that is, less involved in rebounding; 0.93 **assists per game (ast)** , which means less participation in playmaking; -7.15 **Net Rating** , indicating that their presence on the court is associated with a significant deficit in points; 5.88% in **offensive rebound percentage (oreb\_pct)** , showing a similar behavior to segment 1; 13.82% in **defensive rebound percentage (dreb\_pct),** slightly lower than segment 1; 18.21% in usage **(usg\_pct),** i.e. with less prominence in the offense; 45.20% in shooting efficiency; and 11.69% in **assists percentage (ast\_pct), i.e.** less participation in the distribution of the ball. Based on these characteristics, it is suggested that this segment be nominated as **"Support or Reserve Players"** , because this group concentrates the players with less participation in the season, as well as a reduced impact on the game, both offensively and defensively.

## **Objective 7 : Analyze the relationship between the levels of defensive rebounds, level of offensive rebounds, level of assists and the levels of points obtained.**

To achieve this objective, the MCA was used. To do this, the variables were categorized according to quartiles. Likewise, these categories were nominated according to their characteristics. For example, using the quartiles for the variable points obtained, the levels were categorized into low, medium, high and very high.

Texto, Carta

Descripción generada automáticamente

Similarly, for the attendance variable, they were categorized as very few, few, medium and many.

Texto, Carta

Descripción generada automáticamente

For the variable offensive rebounds it was categorized as: very little, little, medium and many.

Texto

Descripción generada automáticamente

Finally, the defensive rebounds variable was categorized as: very little, little, medium and many.

Texto, Carta

Descripción generada automáticamente

The variability of the variables for this objective is explained 100% through 12 dimensions. With the first 3 dimensions, approximately 42% of the total variability can be explained.

Texto

Descripción generada automáticamente

The contributions report shows that the levels of the variable offensive rebounds and defensive rebounds for few and very few are negatively associated with dimension 1 while the medium and high levels are positively associated with dimension 1. Conversely, the levels of the variable assistance for the levels few and very few are positively related to dimension 1.

Tabla

Descripción generada automáticamente

Regarding the degree of proportion of the total inertia of each categorical variable explained by each dimension, it is noteworthy that dimension 1 mostly explains the level of offensive rebounds (60.9%), dimensions 1 and 3 mostly explain the variability of the level of defensive rebounds (48.3% and 44.9% respectively), dimensions 1 and 2 mostly explain the variability of the level of assists (69.2% and 59.1% respectively), and dimensions 2 and 3 mostly explain the level of points obtained (57.9% and 49.6% respectively).

Texto

Descripción generada automáticamente

The factor map shows that very high levels of points obtained by players are related to very low offensive rebounds, very low defensive rebounds and many assists. On the contrary, low levels of points obtained by players are related to many offensive rebounds, many defensive rebounds and few assists. Finally, high levels of points obtained by players are related to low offensive rebounds, low defensive rebounds and regular assists. This pattern found reflects that the points obtained by players are due to a collective work and not to the individual talent of the players.

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

## **Objective 8:** Determine the characteristics about player performance metrics that influence the team's overall playing impact.

The aim of this study is to discover which variables influence whether the player generated an advantage or disadvantage for his team. When the net\_rating is positive, it indicates that the player is advantageous (191 players) for the team, while a negative sign is evidence that the player becomes a disadvantage for the team (292 players).

Of all the performance metrics, points scored, number of rebounds and number of assists were used. These three were chosen because the other metrics are somehow correlated and this could generate an unstable linear or quadratic discriminant model.

Since the predictors present positive asymmetry, the Box Cox transformation was applied (adding 0.0001 for the Box Cox algorithm to work) so that the data of these variables fit the Normal distribution, where the following hypotheses are contrasted:

H0: The data of the i-th variable of the j-th team fits a Normal distribution.

H1: The data of the i-th variable of the j-th team does NOT fit a Normal distribution.

The following table shows the p-values for the Kolmogorov test according to the impact on the team. It can be observed that the p-values for the transformed game performance metrics are greater than 0.05, i.e. the data of the transformed variables fit the Normal distribution.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | **good** | **Bad** |
| **Variable** | **Lambda** | **Pvalue** | **Pvalue** |
| pts | 0.4242 | 0.9506 | 0.6179 |
| reb | 0.03838 | 0.05054 | 0.6117 |
| ast | Log | 0.2441 | 0.2149 |

Similarly, homogeneity between variance-covariance matrices was verified using the M-Box test. This was done in order to determine whether it is appropriate to apply a linear discriminant model or a quadratic model.

Texto

Descripción generada automáticamente

The Box M test is not significant, i.e. it is assumed that the covariance matrices of the good and bad groups are similar. Therefore, Fisher's linear discriminant analysis is applied.

After estimating the Fisher Linear discriminant model, it is found that the **"bad" group** has an a priori probability of **0.5927** (approximately 59.3%), while the **"good" group** has a probability of **0.4072 (40.7%). This suggests that the "bad"** group is more frequent in the data.

Texto

Descripción generada automáticamente con confianza media

Regarding the average values per group we have that: The **"good" group** has an average of **3.41 in transformed points** , while the **"bad" group** has an average of **2.62** . This suggests that the "good" have more transformed points on average. Similarly, the "good" have an average of **1.60** transformed rebounds, while the "bad" have **1.19** . Here too the "good" stand out. Finally, the most notable difference is in assists. The "good" have an average of **0.42 transformed assists** , while the "bad" have just **0.10** . With this, it is highlighted that players classified as "good" tend to have more points, rebounds and assists.

Texto

Descripción generada automáticamente

Regarding the coefficients, it is noteworthy that: **pts\_boxcox** has the highest coefficient ( **0.412** ), which means that **it is the most influential variable** to differentiate between "bad" and "good"; **reb\_boxcox** (coefficient **0.227** ) also contributes, although to a lesser extent; and **ast\_boxcox** (coefficient **0.138** ) has the least impact, but is still relevant.

Texto

Descripción generada automáticamente

Using Fisher's linear discriminant model, 234 players with a negative effect are correctly classified out of a total of 278 in this class (specificity 84.17%) and 61 players with a positive effect are correctly classified out of a total of 191 (sensitivity 31.94%) in this class. This results in an accuracy of 62.9% ((234+61)/469).

Texto

Descripción generada automáticamente

# Conclusions

From the analysis of NBA data it is concluded that:

* The variables associated with the age of the players have a positive impact on the team, while the variables of percentage of offensive rebounds and passes to the player are related in a negative way. For both height and weight, these variables are positively related to percentage of offensive and defensive rebounds, rebounds, while they are negatively related to number of assists and passes to the player.
* The characteristics that generate players with metrics with atypical values are those players who have a maximum of points obtained, number of rebounds, number of assists, percentage of offensive or defensive rebounds.
* There are no significant differences in performance metrics between Kentucky and Duke, which have a higher number of standout players.
* There are no significant differences in performance metrics between DAL, CHA, MIN, NYK teams, which feature a higher number of featured players.
* Performance metrics, as well as physical variables such as age, height and weight, are explained by 75.8% by the dimensions of Size and Rebounding, Production in Play, Efficiency and Offensive Use and Experience.
* Two segments are generated in which players can be grouped as: consistent players and support and reserve players.
* An association is found between high levels of points obtained with low levels of offensive rebounds, low levels of defensive rebounds and regular levels of assists, which reflects that the points obtained by the players are due to collective work and not to the individual talent of the players.
* The transformed variables for points scored, rebounds and assists influence the player's impact on the team by 62.9%.

Appendix