# Estadística
# Proyecto Fase 2

**Carlos Bermudez Porto** C.BERMUDEZ@ESTUDIANTES.MATCOM.UH.CU
*Grupo C412*

**Leynier Gutiérrez González** L.GUTIERREZ@ESTUDIANTES.MATCOM.UH.CU
*Grupo C412*

**Tony Raúl Blanco Fernández** T.BLANCO@ESTUDIANTES.MATCOM.UH.CU
*Grupo C411*

**Tutor(es):**
Lic. Dalia Diaz Sistachs, *Departamento de Matemática Aplicada*

**Palabras Clave:** Estadística, Regresión, Lineal, Reducción, Dimensión, ANOVA
**Tema:** Estadística, Regresión Lineal, ANOVA.

## 1. Introducción

El siguiente trabajo corresponde a la investigación de los autores como evaluación de la asignatura de Estadística de la carrera de Ciencia de la Computación de la Facultad de Matemática y Computación de la Universidad de La Habana.

Los datos utilizados para la investigación son sacados de una base de datos del FIFA 15, la cual contiene todos los jugadores existentes en este videojuego, en la que cada jugador tiene varias estadísticas de sus habilidades como valoración general *(overall)*, potencial *(potential)*, entre otras. Además de la nacionalidad, pie preferido, club al que pertenece, etc.

Se decidió analizar el comportamiento de los jugadores de clase mundial, por lo que se redujo el análisis a los jugadores que pertenecen a los equipos de la fase de grupos de la UEFA Champions League de la temporada 2014-2015, ya que en esta competición participan los mejores clubes de futbol de las mejores ligas, considerado la competición de clubes más importante.

En las pruebas de hipótesis se asumió un nivel de significación del 5%, por tanto, fue posible aceptar las hipótesis nulas con una probabilidad de error menor o igual a 0.05.

### 1.1 Selección y descripción general de los datos

De todas las variables disponibles en la base de datos, se seleccionaron 17 porque se consideró son las que mejor describen la forma de jugar de un futbolista. Estas son:

- overall
- potential
- attacking_finishing
- attacking_short_passing
- skill_ball_control
- skill_fk_accuracy
- skill_long_passing
- movement_agility
- movement_balance
- movement_sprint_speed
- power_shot_power
- power_stamina
- power_long_shots
- mentality_interceptions
- mentality_vision
- defending_marking
- defending_sliding_tackle

Se realizó un análisis general de esas variables utilizando la función *skim* de la biblioteca *skimr*. El resultado se puede ver en la Fig. 1.

```
-- Variable type: numeric -------------------------------------------------------------------------
# A tibble: 17 x 11
   skim_variable             n_missing complete_rate  mean    sd    p0   p25   p50   p75  p100 hist
 * <chr>                         <int>         <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <chr>
 1 overall                           0             1  72.9  8.31    49    68    74    79    93 ▁▃▇▆▁
 2 potential                         0             1  78.1  6.10    58    74    79    83    95 ▁▂▆▇▂
 3 attacking_finishing               0             1  52.9 19.3     20    36    56    68    95 ▇▆▆▆▂
 4 attacking_short_passing           0             1  65.8 16.5     20    60    70    77    95 ▁▂▅▇▃
 5 skill_ball_control                0             1  66.5 17.7     20    61    72    78    96 ▁▂▅▇▃
 6 skill_fk_accuracy                 0             1  51.5 18.2     20    36    51    67    93 ▇▆▆▅▂
 7 skill_long_passing                0             1  60.4 16.6     21    52    64    72    95 ▁▂▅▇▃
 8 movement_agility                  0             1  67.8 14.8     25    59    70    78    94 ▁▂▅▇▃
 9 movement_balance                  0             1  65.5 14.4     27    57    66    76    97 ▁▂▅▇▃
10 movement_sprint_speed             0             1  70.2 13.5     29    64  72.5    79    96 ▁▂▅▇▃
11 power_shot_power                  0             1  64.2 17.2     20    55    69    77    94 ▁▂▅▇▃
12 power_stamina                     0             1  67.3 16.3     21    61    71    78    95 ▁▂▅▇▃
13 power_long_shots                  0             1  56.4 19.0     20    41  60.5    72    93 ▇▅▅▆▃
14 mentality_interceptions           0             1  53.5 22.4     20    29    57    74    93 ▇▃▃▆▃
15 mentality_vision                  0             1  58.9 18.2     20    47    62    73    96 ▁▃▅▇▃
16 defending_marking                 0             1  49.6 22.0     20    25  50.5    71    90 ▇▃▃▆▃
17 defending_sliding_tackle          0             1  51.6 22.3     20    25    56    73    95 ▇▃▃▆▃
```

Figure 1: Resultado de aplicar la función *skim*

Como resultado del análisis general se detectó que la media de las habilidades de los jugadores es baja incluso para los mejores equipos de las mejores ligas de futbol. A partir de cálculo de los cuartiles el 75% de los jugadores están por debajo de 80 puntos (exceptuando el caso del *potential* que alcanza 83) lo cual quiere decir que los jugadores considerados clase mundial corresponden solo a un 15%.

## 2. Regresión Lineal y Análisis de Componentes Principales

Para aplicar la regresión es necesario realizar un análisis de las correlaciones de las variables seleccionadas.

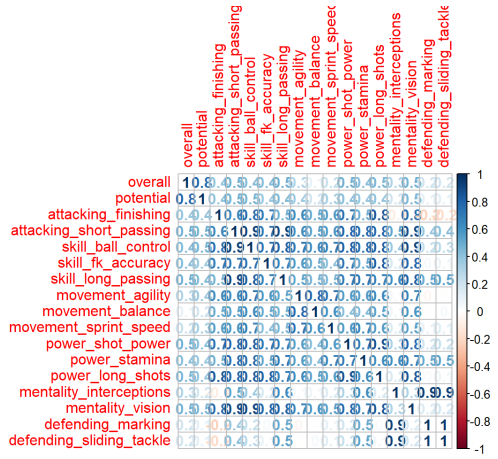


Figure 2: Correlación de todas las variables

Como resultado del análisis de la Fig. 2, se detectó una alta correlación entre las variables, por lo que no sería correcto aplicar una regresión lineal, por lo tanto, se aplicó la técnica de reducción de dimensiones para obtener las componentes principales que sean independientes entre sí y que sean dependientes de una variable respuesta elegida.

### 2.1 Análisis de Componentes Principales

En un inicio se seleccionó como variable respuesta a *overall*, por lo que se hizo necesario extraer esa variable antes de realizar el análisis de las componentes principales.

```
Importance of components:
                          PC1    PC2     PC3     PC4     PC5     PC6     PC7     PC8
Standard deviation     2.9998 1.8015 1.02576 0.87163 0.74757 0.53447 0.47729 0.44698
Proportion of Variance 0.5624 0.2028 0.06576 0.04748 0.03493 0.01785 0.01424 0.01249
Cumulative Proportion  0.5624 0.7653 0.83102 0.87850 0.91343 0.93129 0.94552 0.95801
                           PC9    PC10    PC11    PC12    PC13    PC14    PC15    PC16
Standard deviation     0.38960 0.37894 0.31884 0.29301 0.25416 0.24031 0.19436 0.1697
Proportion of Variance 0.00949 0.00897 0.00635 0.00537 0.00404 0.00361 0.00236 0.0018
Cumulative Proportion  0.96750 0.97647 0.98283 0.98819 0.99223 0.99584 0.99820 1.0000
```

Figure 3: Análisis de Componentes Principales

Aplicando el criterio de *Kaiser*, fue posible quedarse con las primeras tres componentes. Pasando a crear una matriz con la variable respuesta junta las tres componentes principales seleccionadas y analizando sus correlaciones.


Figure 4: Correlación 1 de las Componentes Principales

Como se puede ver en la Fig. 4 la variable respuesta escogida no tiene dependencia con las componentes principales seleccionadas, luego se realizó un análisis similar iterando por todas las variables seleccionándolas como variable respuesta en cada paso se obtiene que la variable *mentality_interceptions* tiene una mayor correlación con las componentes principales seleccionadas (referirse a la Fig. 5) por lo que era una mejor opción a variable respuesta.

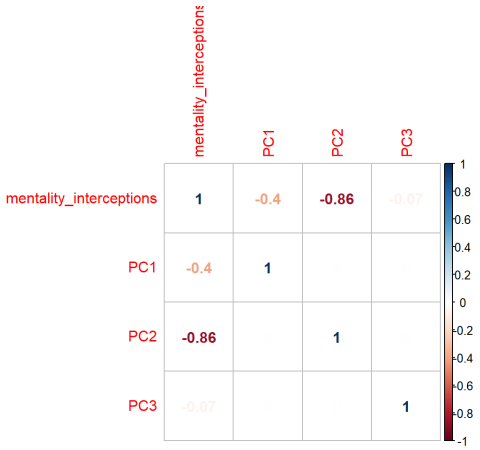


Figure 5: Correlación 2 de las Componentes Principales

#### 2.1.1 Descripción de las Componentes

- **PC1:** La componente se caracteriza por una presencia negativa de *overall*, *potential*, *attacking_finishing*, *attacking_short_passing*, *skill_ball_control*, *skill_fk_accuracy*, *skill_long_passing*, *movement_agility*, *movement_balance* y *movement_sprint_speed*. Se puede interpretar como que

son los jugadores no tan habilidosos, los cuales, como vimos en análisis anteriores, son la mayoría.

- **PC2:** La componente se caracteriza por una presencia positiva de *attacking_finishing* y una presencia negativa de *defending_marking* y de *defending_sliding_tackle*. Se puede interpretar como los jugadores de la posición de delantero centro. Ya que su función es marcar goles y no tienen mucha obligación defensiva.

- **PC3:** La componente se caracteriza por una presencia positiva de *overall* y *potential*, y una presencia negativa de *movement_balance*. Se puede interpretar como los jugadores habilidosos pero con tendencia a caer mucho en el campo, ya sea por excesiva cantidad de faltas o por engañar al arbitro.

### 2.2 Regresión Lineal

Teniendo como la variable respuesta a *mentality_interceptions* y como variables independientes las componentes principales se realizó un modelo de regresión lineal múltiple, a continuación, se muestra la información del modelo realizado.

```
Call:
lm(formula = "mentality_interceptions ~ .", data = reg_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.19642 -0.15499  0.00418  0.16314  1.11200

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.072e-16  1.025e-02   0.000        1
PC1         -1.331e-01  3.399e-03 -39.149  < 2e-16 ***
PC2         -5.466e-01  6.486e-03 -84.271  < 2e-16 ***
PC3         -5.582e-02  8.720e-03  -6.401 2.59e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2951 on 824 degrees of freedom
Multiple R-squared:  0.9133,    Adjusted R-squared:  0.9129
F-statistic:  2892 on 3 and 824 DF,  p-value: < 2.2e-16
```

Figure 6: Resultado del Modelo de Regresión Lineal

De la información de la Fig. 6 se deduce la siguiente ecuación para determinar *mentality_interceptions* en base a las componentes principales.

$$mentality_{interceptions} = -0.13PC_1 - 0.55PC_2 - 0.06PC_3$$

Uno de los factores a tener en cuenta es la estandarización de los datos durante el proceso de clasificación, por lo que su efecto se ve reflejado en la ausencia del término independiente en la ecuación de regresión.

La precisión del modelo, medida en términos del valor del $Ajusted\ R - squared$ es de 0.9129 lo cual es bastante alto tomando en consideración el desprecio de datos resultante del análisis de las componentes principales. El $p - value$ de la prueba de $F - statistic$ es menor que 0.05 por lo que podemos asegurar que el modelo produce resultados. El error residual es de 0.2951.

#### 2.2.1 Análisis de los supuestos

Los supuestos de la regresión linear son los siguientes:

1. Las variables independientes no están correlacionadas.
2. La media y la suma de los errores es cero.
3. Los errores son independientes.
4. Los errores tienen distribución normal.
5. La varianza de los errores es constante.

El **primer supuesto** se cumple porque las variables independientes utilizadas son resultado del análisis de componentes principales.

El **segundo supuesto** también se cumple, siendo la media y la suma de los errores extremadamente cercanos a cero. (Fig. 7)

```
[1] "Media de los Errores"
[1] 8.170144e-19
[1] "Suma de los Errores"
[1] 6.921547e-16
```

Figure 7: Media y suma de los errores

El **tercer supuesto**, que los errores sean independientes, como se puede ver la prueba de *Durbin-Watson* no es significativa, siendo el $p - value$ mayor que 0.05, no se puede rechazar la hipótesis nula, por tanto, se puede deducir que los errores son independientes entre sí. (Fig. 8)

```
        Durbin-Watson test

data:  model
DW = 1.999, p-value = 0.4814
alternative hypothesis: true autocorrelation is greater than 0
```

Figure 8: Prueba de Durbin-Watson

El **cuarto supuesto**, que los errores tengan una distribución normal, se analizó en primera instancia el histograma de los errores (Fig. 9) y el Q-Q Plot (Fig. 10).

A partir de la interpretación del histograma (Fig. 9) se pudo observar una distribución normal, pero no tanto a partir de la interpretación del Q-Q Plot (Fig. 10), por lo que se realizó una prueba de $Shapiro - Wilk$ (Fig. 11).

Como se puede ver la prueba de $Shapiro - Wilk$ (Fig. 11) es significativa, siendo el $p - value$ menor que 0.05, se puede rechazar la hipótesis nula, por tanto, se puede deducir que los errores no siguen una distribución normal, incumpliendo con los supuestos, y, por tanto, el modelo no es de utilidad.

Al incumplirse uno de los supuestos, el modelo deja de tener utilidad, pero aun así, se decidió analizar la homocedasticidad en este trabajo.

El **quito supuesto**, que la varianza de los errores se constante, se analizó en primera instancia el gráfico de los Residuos Estandarizados (Fig. 12).

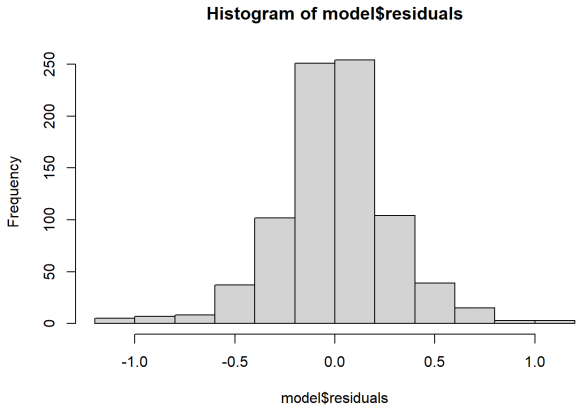


Figure 9: Histograma de los errores

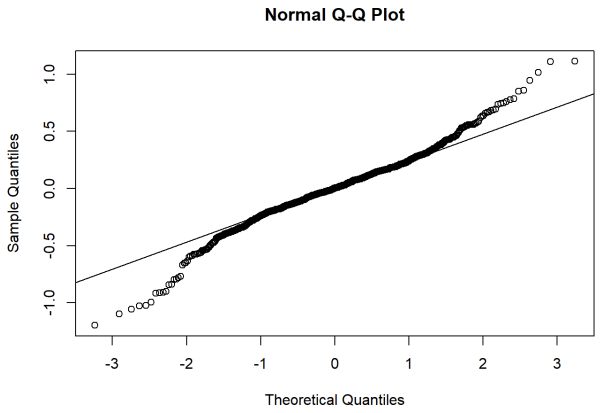


Figure 10: QQ-Plot de los errores

A partir de la interpretación del gráfico de los Residuos Estandarizados (Fig. 12) se pudo observar una homocedasticidad en los errores, pero no quedó muy claro, por lo que se realizó una prueba de $Breushc - Pagan$ (Fig. 13) para confirmarlo.

Como se pudo ver, la prueba de $Breushc - Pagan$ (Fig. 13) es significativa, siendo el $p - value$ menor que 0.05, se pudo rechazar la hipótesis nula, por tanto, se pudo deducir que no hay heteroscedasticidad en los errores (de donde se dedujo que hay homocedasticidad)

```
        Shapiro-Wilk normality test

data:  model$residuals
W = 0.97401, p-value = 5.602e-11
```

Figure 11: Prueba de Shapiro-Wilk

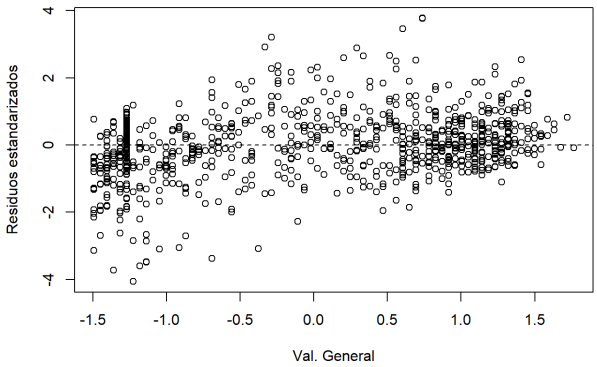


Figure 12: Residuos Estandarizados

```
        studentized Breusch-Pagan test

data:  model
BP = 47.667, df = 3, p-value = 2.507e-10
```

Figure 13: Prueba de Breushc-Pagan

## 3. ANOVA

Para el ANOVA o análisis de varianzas se decidió analizar si el pie preferido influye en la habilidad de deslizarse de los jugadores. A continuación se muestra el resultado del modelo (Fig. 14).

```
             Df Sum Sq Mean Sq F value  Pr(>F)
Fo            1   4176    4176   8.443 0.00376 **
Residuals   826 408605     495
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 14: Resultado de ANOVA

Como el $p-value$ es menor que la significación prefijada de 0.05, se rechaza la hipótesis nula y se acepta que el pie preferido influye en la habilidad para deslizarse.

### 3.1 Análisis de los supuestos

1. Los errores siguen una distribución normal con media cero.

2. Los errores son independientes entre sí.

3. Los errores tienen la misma varianza.

El **primer supuesto**, que los errores tengan una distribución normal con media cero, se analizó en primera instancia el histograma de los errores (Fig. 15) y el Q-Q Plot (Fig. 16).

A partir de la interpretación del histograma (Fig. 15) y del Q-Q Plot (Fig. 16) se pudo observar que la distribución no es normal, por lo que se realizó una prueba de $Shapiro - Wilk$ (Fig. 17) para confírmalo.

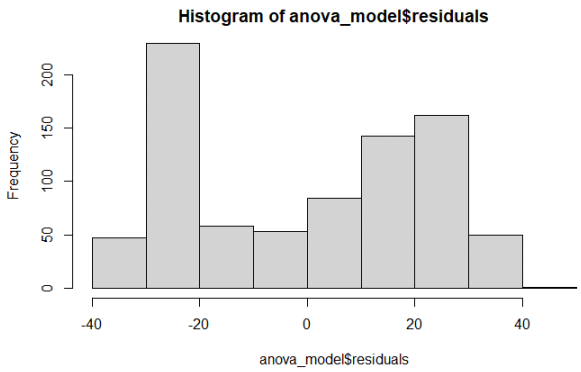


Figure 15: Histograma de los errores del modelo de ANOVA

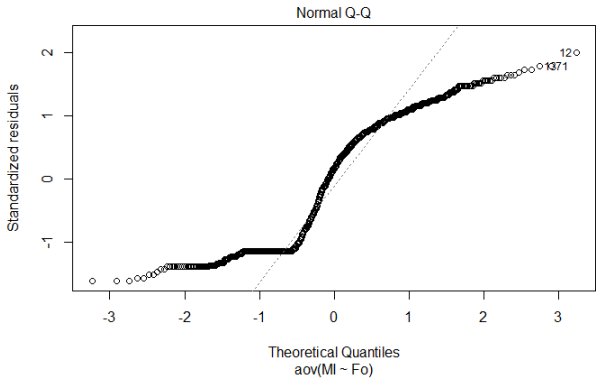


Figure 16: Q-Q Plot del modelo de ANOVA

Como se puede ver la prueba de $Shapiro-Wilk$ (Fig. 17) es significativa, siendo el $p-value$ menor que 0.05, se puede rechazar la hipótesis nula, por tanto, se puede deducir que los errores no siguen una distribución normal, incumpliendo con los supuestos, y, por tanto, el modelo no es de utilidad.

Al incumplirse uno de los supuestos, el modelo deja de tener utilidad, pero, aun así, se decidió verificar los restantes supuestos.

El **segundo supuesto**, que los errores sean independientes entre sí, como se puede ver la prueba de $Durbin-Watson$ (Fig. 18) es significativa, siendo el $p-value$ menor que 0.05, se puede rechazar la hipótesis nula, por tanto, se puede deducir que los errores no son independientes entre sí, incumpliendo con los supuestos, y, por tanto, el modelo no es de utilidad.

El **tercer supuesto**, que los errores tengan la misma

```
        Shapiro-Wilk normality test

data:  anova_model$residuals
W = 0.89162, p-value < 2.2e-16
```

Figure 17: Prueba de Shapiro-Wilk

```
        Durbin-Watson test

data:  anova_model
DW = 1.8366, p-value = 0.009299
alternative hypothesis: true autocorrelation is greater than 0
```

Figure 18: Prueba de Durbin-Watson

varianza, como se puede ver la prueba de $Bartlett$ (Fig. 19) no es significativa, siendo el $p-value$ mayor que 0.05, no se puede rechazar la hipótesis nula, por tanto, se puede deducir que los errores tienen la misma varianza, o sea se cumple la homocedasticidad.

```
        Bartlett test of homogeneity of variances

data:  anova_model$residuals and data_anova$Fo
Bartlett's K-squared = 0.013901, df = 1, p-value = 0.9061
```

Figure 19: Prueba de Bartlett

## 4. Clústeres

En el siguiente análisis utilizando clústeres se tomó como referencia los resultados obtenidos del análisis de las componentes principales, en las que según el críterio del porcentaje el resultado era dos componentes principales y según el críterio de Kaiser el resultado era tres componentes principales.

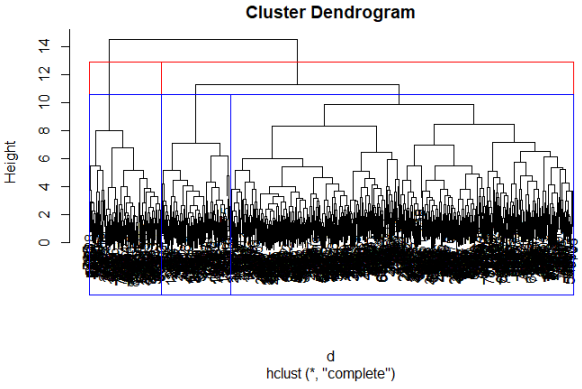


Figure 20: Dendograma con 2 y 3 clústeres

Analizando los dendrogramas (Fig. 20) se puede observar que tanto para 2 como para 3 clústeres existe uno de los clústeres que tiene la mayoría de los datos, por lo que se analizó aumentar la cantidad de clústeres a 4, 5, 6 y 7 para analizar el comportamiento con estas cantidades.

Del análisis del dendograma (Fig. 21) se eligió 6 clústeres como medidor principal debido a que con 4 y 5 clústeres permanecía una diferencia sustancial en cuanto a la cantidad de datos en cada clúster, y con 7 clústeres no se observó una mejora sobre a elección de 6 clústeres. Tomando en cuenta que se están segmentando en grupos a jugadores de futbol según sus

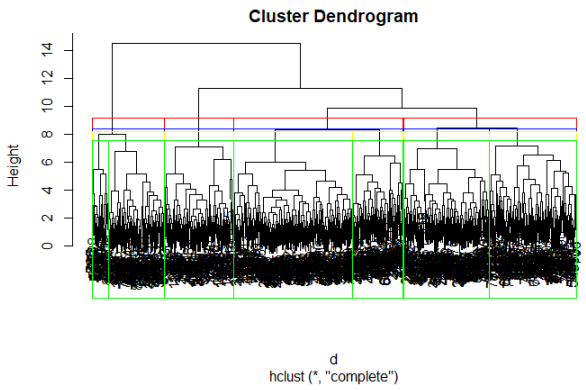


Figure 21: Dendograma con 4, 5, 6 y 7 clústeres

características, 6 grupos no es una división exagerada, ya que los jugadores se pueden dividir por posiciones en el campo, estilo: portero, defensas centrales, laterales, mediocampistas defensivos, medio campistas ofencivos y delanteros. No quiere decir que el análisis con 6 clústeres determine estos grupos, solamente se quisó ilustrar que una clasificación en 6 grupos a los futbolistas no es una división extraña.

## 5. K-Medias

Se realizó un análisis utilizando el algoritmo de K-medias. Se utilizaron 6 particiones basándose en el resultado del análisis de los clústeres.

Del análisis de los resultado de K-medias (Fig. 22) se obtuvo que el porciento de similitud es del 69% aproximadamente, lo cual es bastante bueno para la naturaleza de los datos que se analizaron. La cantidad de elementos en cada partición fue de 94, 168, 207, 98, 122, 139 respectivamente.

## 6. Árbol de Decisión

En un inicio, debido a que el modelo de regresión lineal no fue útil, ya que se incumplieron sus supuestos, se decidió realzar un análisis utilizando árboles de decisión, en primer árbol de decisión se realizó utilizando las 17 variables seleccionadas y como variable respuesta a *overall*, el resultado se puede observar en la Fig. 23, pero el error es del 99% por lo que se intento con un segundo árbol de decisión utilizando como variable respuesta *mentality_interceptions* y como variables independientes a las componentes principales, nuevamente el error fue alto, cercano al 98%. (Fig. 24)

## 7. Conclusiones

Se realizó un análisis sobre los datos obtenidos, aplicando una regresión lineal junto a un análisis de componente principales, además de un análisis de varianza o ANOVA. En ambos casos, los supuestos de los modelos se incumplieron, pero se pudo mostrar la aplicación de las principales técnicas de regresión y análisis de

```
K-means clustering with 6 clusters of sizes 96, 131, 83, 158, 209, 151

Cluster means:
     overall  potential attacking_finishing attacking_short_passing skill_ball_control
1 -0.2429575 -0.5144791          -1.4553263             -2.1956346        -2.29124647
2 -0.6969268 -0.3539542           0.6626823             -0.1815608         0.17342625
3 -1.6754556 -1.3501545          -0.9238194             -0.5890226        -0.61240001
4  0.1769127  0.1135452          -0.7299853              0.1564573        -0.05260782
5  0.5132358  0.3411253           0.3379355              0.7265980         0.61020175
6  0.7845407  0.7853342           1.1542127              0.7077816         0.85331056
  skill_fk_accuracy skill_long_passing movement_agility movement_balance
1        -1.4371836         -1.9134466       -1.4816150       -1.41789902
2        -0.0549479         -0.5325294        0.3130685        0.24048743
3        -0.6811272         -0.5313792       -0.5725308       -0.08669947
4        -0.5179197          0.2807979       -0.5070835       -0.52321272
5         0.6533200          0.8416485        0.4095844        0.40739286
6         0.9734357          0.5118284        0.9487363        0.72405967
  movement_sprint_speed power_shot_power power_stamina power_long_shots
1           -1.78180462      -1.90494270    -2.0251754       -1.6498170
2            0.30670100       0.17235068    -0.2595015        0.2321690
3           -0.29623829      -0.92878263    -0.4329806       -0.9670902
4           -0.06482423      -0.07319423     0.2880507       -0.5223537
5            0.24016725       0.64959172     0.7790527        0.7124626
6            0.76496995       0.74957309     0.3709596        0.9394966
  mentality_interceptions mentality_vision defending_marking defending_sliding_tackle
1              -1.3196089     -1.873676221       -1.1202818              -1.1938849
2              -0.9771977     -0.005733225       -0.9450496              -0.9598935
3              -0.0362921     -0.731398119        0.1191498               0.1981726
4               0.9713536     -0.337620982        1.1061957               1.0218146
5               0.9457328      0.737637165        0.8617595               0.9027570
6              -0.6187048      0.930516299       -0.8836261              -0.8358427

Clustering vector:
  [1] 6 6 6 6 1 6 6 5 6 6 6 5 5 5 6 5 6 6 1 5 6 6 6 6 5 5 4 6 5 5 6 6 6 5 5 4 1 6 4 5 6 6 6
 [44] 4 5 1 1 5 5 5 4 6 6 6 5 4 5 6 5 6 5 5 6 4 5 6 1 4 4 5 4 4 5 5 6 1 6 5 4 5 1 6 6 6 6 1
 [87] 6 5 6 6 6 5 5 1 6 6 4 4 6 4 6 6 1 1 5 1 5 6 1 5 5 5 5 6 5 5 5 1 4 4 6 6 6 5 5 5 6 5 4 5
[130] 6 5 6 6 5 6 4 6 5 1 6 6 5 5 5 6 4 2 4 4 4 5 4 4 5 6 6 5 6 6 6 6 6 1 6 4 6 1 2 5 5 5 6
[173] 6 4 6 5 6 6 6 5 5 5 6 5 4 5 5 5 5 5 6 4 6 5 6 5 5 5 6 5 4 5 4 6 5 5 5 5 4 5 6 1 1 5 5 5 5 1
[216] 5 6 4 5 1 1 5 4 5 5 6 5 4 6 6 6 6 6 4 6 4 4 2 5 5 5 4 6 6 1 5 6 5 5 4 5 6 2 4 5 4 5 6
[259] 5 1 1 4 5 6 6 5 4 5 5 4 4 4 6 1 5 6 2 4 5 5 5 6 5 2 5 6 1 5 6 5 1 4 1 4 6 4 5 5 4 4 5 1
[302] 4 5 5 4 5 4 5 5 6 5 4 5 5 5 4 2 5 5 5 6 4 5 6 6 6 5 4 5 5 4 6 1 6 6 4 5 2 5 5 5 1 6 4
[345] 6 6 5 5 4 6 5 1 5 6 4 6 1 6 1 5 5 6 1 1 5 6 5 5 4 6 4 4 5 1 1 6 5 6 6 5 6 5 6 4 6 6 6
[388] 4 4 6 4 1 6 5 5 6 2 2 5 5 5 5 4 4 6 4 6 2 4 6 5 5 4 6 6 4 4 6 4 6 4 2 6 6 5 4 4 6 2 5
[431] 2 5 4 1 5 1 2 2 1 5 4 5 5 4 2 5 5 6 5 5 6 4 2 4 2 4 5 1 4 6 6 4 4 1 2 2 5 4 2 5 4 5 2
[474] 4 2 5 5 5 5 1 4 2 5 6 5 5 4 2 5 5 5 2 4 4 5 2 4 4 4 4 2 6 4 6 4 2 4 5 2 5 4 4 4 2 5 4
[517] 4 4 2 5 4 6 1 2 5 4 4 2 5 5 4 4 6 5 2 4 1 2 5 5 5 4 2 4 1 1 4 5 5 1 2 4 5 5 2 2 5 2 4
[560] 4 5 2 2 4 5 5 2 2 4 1 1 2 2 4 2 1 1 4 1 1 4 2 5 5 2 2 2 6 2 2 6 1 4 2 5 4 2 4 2 4 4 1
[603] 5 2 2 1 2 4 4 3 5 4 4 5 5 5 2 4 4 2 4 5 1 4 2 5 4 5 4 2 2 4 4 4 2 2 3 1 4 3 2 2 4 1 1
[646] 2 2 2 4 2 2 3 2 3 2 2 4 3 2 2 1 5 2 3 2 2 2 3 1 4 4 2 3 2 4 2 2 4 1 3 1 1 2 2 2 3 4 2
[689] 3 3 5 3 2 1 2 3 3 2 1 2 2 3 2 3 3 4 2 3 2 3 2 4 3 2 2 1 3 3 1 2 2 3 3 2 2 1 1 2 2 1 1 2
[732] 2 2 3 2 2 3 3 2 2 2 3 1 1 3 3 3 3 2 2 3 2 3 2 2 2 2 3 3 2 3 1 3 2 2 1 1 3 3 1 2 3 3 3
[775] 1 3 1 3 3 3 3 3 3 1 2 3 3 3 2 1 2 3 3 3 1 3 3 3 1 1 1 3 1 3 3 1 1 1 1 3 3 3 1 1 3 3 3 3 3
[818] 3 3 3 3 3 3 3 3 3 3 3

Within cluster sum of squares by cluster:
[1] 452.0284 755.8749 543.3536 909.6469 952.8335 711.4896
 (between_SS / total_SS =  69.2 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"         "iter"         "ifault"
```

Figure 22: Resultado de K-medias

varianzas para un uso posterior en otras circunstancias y/o datos.

## 8. Referencias


- Conferencia 5 - Correlación
- Conferencia 6 - Regresión Lineal Simple
- Conferencia 7 - Regresión Lineal Múltiple
- Conferencia 8 - ANOVA
- Conferencia 9 - ACP
- Conferencia 10 - Clúster, Árboles de Decisión

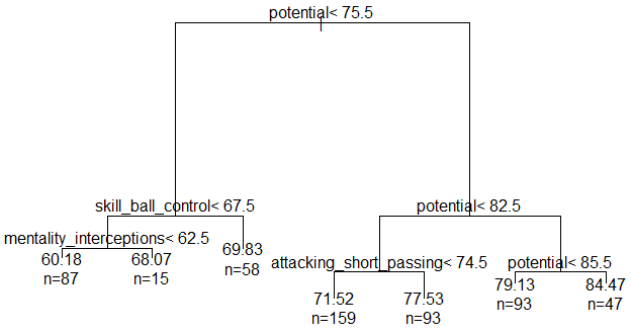


Figure 23: Árbol de Decisión con *overall*

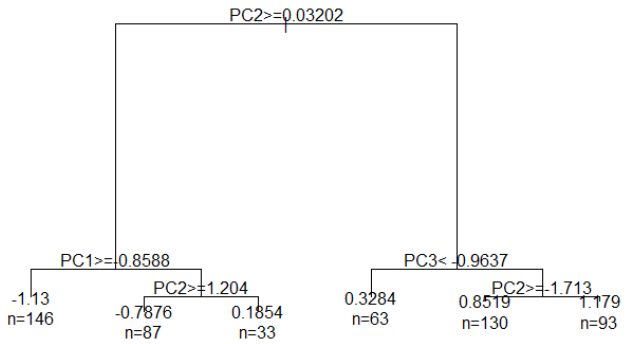


Figure 24: Árbol de Decisión con *mentality_interceptions*