

|  |
| --- |
| **FACULTAD DE MATEMÁTICAS** |
| **Minería de Datos** |

**Unidad II:Procesamiento de Datos**

**ADA 5: *Reducción de Datos.***

*Licenciatura en Actuaria.*

**Integrantes:**

- Álvarez Herrera Samantha

- Ciau Puga Abigail

- Colonia Espinosa Cindy

- Fernández Caro Frida

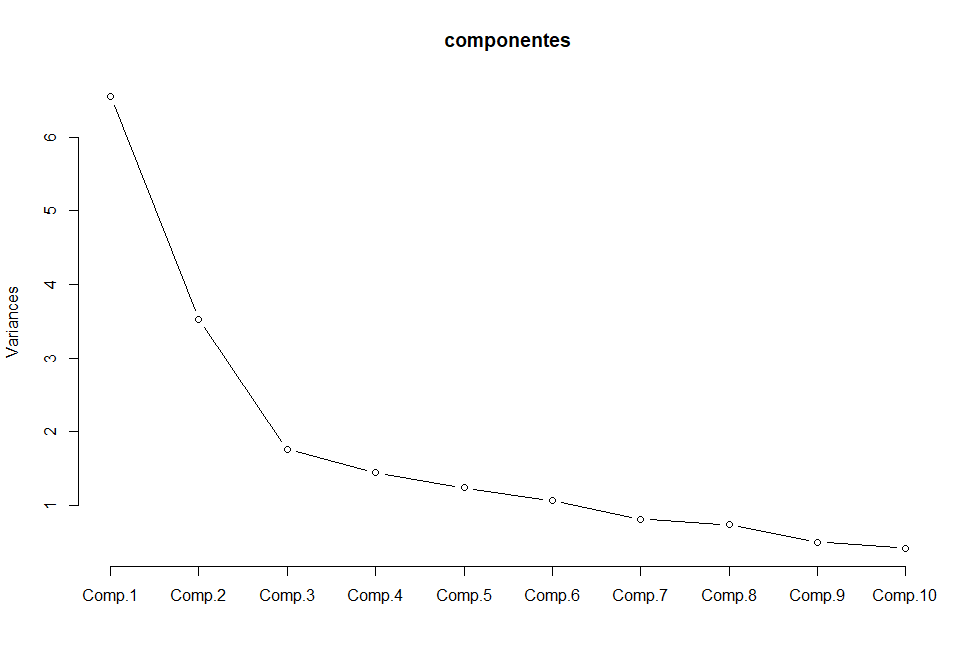
- Padilla Jiménez Meybor

- Sobrino Bermejo Samantha

**M.C. Ernesto Guerrero Lara**

**EJERCICIO 1. Análisis de Componentes Principales**

Inicialmente, la base de datos ppg2018 cuenta con 19 variables a las que aplicaremos el método de componentes principales.

Utilizando el método del codo, decidimos trabajar con tres componentes, ya que es a partir del tercero de estos donde la gráfica se estabiliza, como se observa a continuación: 

La varianza explicada por los 3 componentes seleccionados es del 62.30%

|  |  |  |  |
| --- | --- | --- | --- |
|  | Standard deviation | Proportion of Variance | Cumulative Proportion |
| Comp.1 | 2.5596 | 0.3448 | 0.3448 |
| Comp.2 | 1.8766 | 0.1854 | 0.5302 |
| Comp.3 | 1.3278 | 0.0928 | 0.6230 |
| Comp.4 | 1.2008 | 0.0759 | 0.6988 |
| Comp.5 | 1.1115 | 0.0650 | 0.7639 |
| Comp.6 | 1.0292 | 0.0558 | 0.8196 |
| Comp.7 | 0.9008 | 0.0427 | 0.8623 |
| Comp.8 | 0.8564 | 0.0386 | 0.9009 |
| Comp.9 | 0.7028 | 0.0260 | 0.9269 |
| Comp.10 | 0.6459 | 0.0220 | 0.9489 |
| Comp.11 | 0.5422 | 0.0155 | 0.9643 |
| Comp.12 | 0.4996 | 0.0131 | 0.9775 |
| Comp.13 | 0.4223 | 0.0094 | 0.9869 |
| Comp.14 | 0.3611 | 0.0069 | 0.9937 |
| Comp.15 | 0.3263 | 0.0056 | 0.9993 |
| Comp.16 | 0.0868 | 0.0004 | 0.9997 |
| Comp.17 | 0.0575 | 0.0002 | 0.9999 |
| Comp.18 | 0.0416 | 0.0001 | 1.0000 |
| Comp.19 | 0.0120 | 0.0000 | 1.0000 |

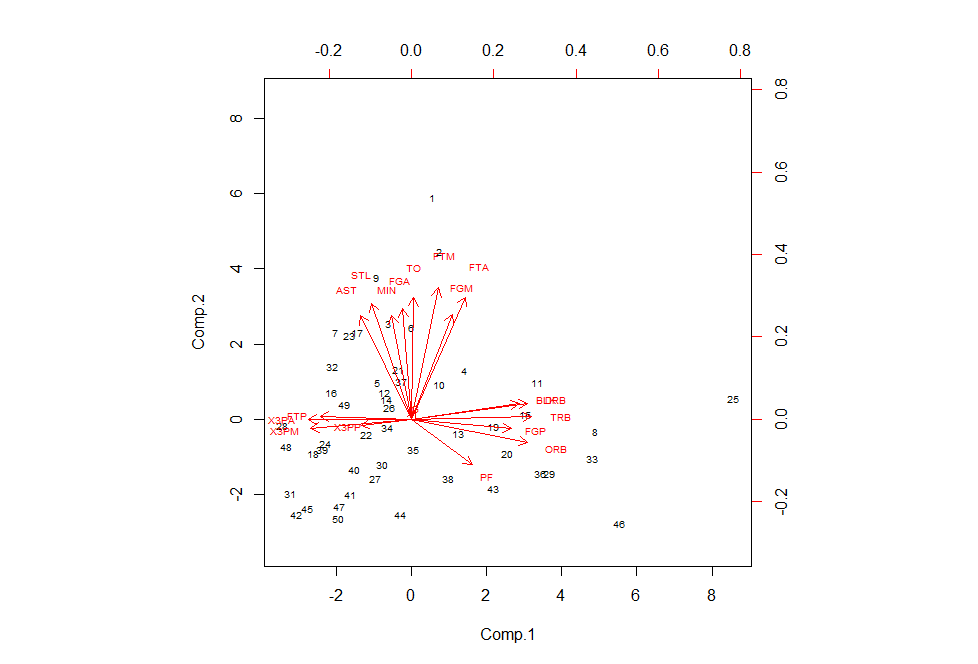
Los pesos que se le asigna a las 19 variables en cada componente quedan como sigue:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Comp.1 | Comp.2 | Comp.3 |
| G | 0.0104 | 0.0262 | 0.1683 |
| MIN | -0.0593 | 0.3143 | 0.2722 |
| FGM | 0.1234 | 0.3192 | 0.4551 |
| FGA | -0.0268 | 0.3359 | 0.4803 |
| FGP | 0.3042 | -0.0287 | -0.0443 |
| FTM | 0.0808 | 0.3985 | -0.2230 |
| FTA | 0.1649 | 0.3700 | -0.2602 |
| FTP | -0.2772 | 0.0089 | 0.1459 |
| X3PM | -0.3056 | -0.0275 | 0.1497 |
| X3PA | -0.3130 | -0.0006 | 0.1262 |
| X3PP | -0.1528 | -0.0184 | 0.1040 |
| ORB | 0.3530 | -0.0712 | 0.1186 |
| DRB | 0.3521 | 0.0489 | 0.1566 |
| TRB | 0.3655 | 0.0069 | 0.1525 |
| AST | -0.1560 | 0.3149 | -0.2122 |
| STL | -0.1223 | 0.3522 | -0.1429 |
| BLK | 0.3270 | 0.0473 | 0.0608 |
| TO | 0.0063 | 0.3686 | -0.3240 |
| PF | 0.1850 | -0.1389 | -0.1844 |

Los cuadros sombreados representan los pesos no significativos en la componente.

De las 19 variables que contaba inicialmente la base de datos, conseguimos reducir su dimensión a 3 variables, que son:

Analizando el gráfico biplot con las primeras dos componentes. Para la primera componente, la mayor parte de las variables tienen peso positivo, sin embargo, la carga debe ser mayor para los pesos negativos los que ocasiona que las nuevas observaciones sean mayormente negativas para la variable . En cuanto a la segunda componente, vemos que la mayor parte de las variables tienen pesos cercanos a cero o negativos, lo que ocasiona que la mayoría de las nuevas observaciones de la variable sean negativas.

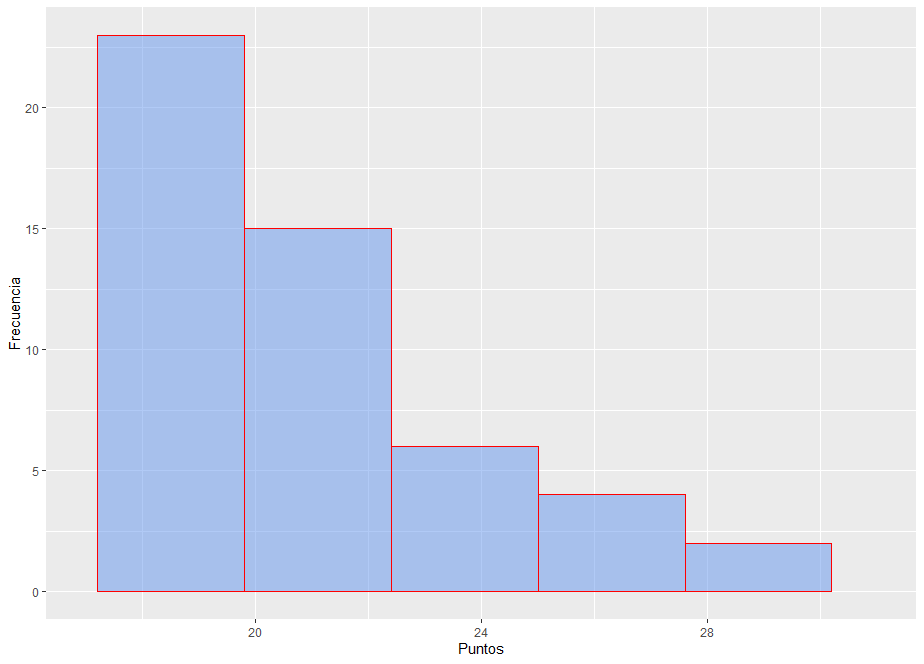


**EJERCICIO 2. Suavizamiento**

Se usó el método de binarización, el cual es un método de arriba-abajo, consiste en suavizar datos numéricos tomando representantes. En este caso se decidió crear 5 clases y el representante de cada grupo es el promedio del máximo y mínimo del rango de cada clase. Los intervalos y los representantes de cada clase resultaron ser los siguientes:

|  |  |  |
| --- | --- | --- |
| Número de clase (Grupo) | Intervalo | Representante |
| 1 | [17.19999 19.8] | 18.49999 |
| 2 | (19.8,22.4] | 21.10000 |
| 3 | (22.4,25] | 23.70000 |
| 4 | (25,27.6] | 26.30000 |
| 5 | (27.60000 30.20000] | 28.90000 |

Los puntos de corte fueron elegidos debido a que la longitud de los intervalos es idóneo pues no es muy grande y los elementos de cada clase estarán bien simbolizados por el representante de cada clase. Cabe destacar que se decidió crear 5 clases debido a que consideramos que la variación de la variable original no era tan amplia como para usar más grupos además consideramos que una clasificación de 5 grupos es entendible para poder explicar tanto visual como verbalmente la variable puntos.



En la gráfica se puede observar que el primer grupo tiene una frecuencia de 23, es decir, 23 jugadores obtuvieron entre 17.19 y 19.8 puntos. Para el segundo grupo, 15 jugadores obtuvieron entre 19.8 y 22.4 puntos, en el tercer grupo 6 jugadores obtuvieron entre 22.4 y 25 puntos. El cuarto grupo tuvo una frecuencia de 4 jugadores cuyos puntos se encuentran entre 25 y 27.6, y el último grupo solo cuenta con dos jugadores cuyos puntos se encuentran entre 27.6 y 30.2.