

K-MEANS CLUSTERING

Rodrigo Yerena Herrera | Universidad Modelo | Escuela de ingeniería | IDTS | 8vo semestre | Inteligencia de negocios | Prof. Luis Jorge Novelo Novelo

# Introducción:

En este proyecto se hará una demostración del uso del algoritmo “K-Means Clustering” en un problema aplicado.

Esta idea surgió a raíz de mi clase “Inteligencia de negocios” en donde pudimos aprender como funciona este algoritmo en este entorno d ellos negocios, de cómo es posible clasificar los datos de una base de datos para conocer datos que puedan servir para hacer graficaciones y estadísticos para un análisis mas profundo según sea requerido.

# DATA:

Para este proyecto se utilizará la base de datos de los pasajeros del Titanic. El documento fue recolectado de una página web el cual está indicado en las referencias de este documento.

Este documento indica lo siguiente:

* ID del pasajero
* Si sobrevivió o no sobrevivió
* Número de clase
* Nombre completo del pasajero
* Sexo
* Edad
* Familiares o cónyuges de los pasajeros
* Pasajeros con niños
* Número de boleto
* Tarifa
* Camarote
* Puerto de embarque

# Metodología:

Utilizaremos Python como lenguaje de programación y Jupyter Notebook como entorno programático junto con la librería “Pandas” y la extensión de “Sklearn”: “KMeans” el cual es el algoritmo de agrupamiento de datos, el cual será ejecutado para obtener resultados de los grupos en un formato de datos que faciliten la interpretación de los mismo.

Primero se descargan las librerías que utilizaremos en el código, en este caso son “Pandas” y “Skelarn.KMeans”

Posteriormente se carga al código el archivo “csv” con la información.

La idea del K-Means en este proyecto es que identifique las clasificaciones que existen en este archivo de datos y los agrupe, para que luego puedan verse de una manera fácil, el cual pueda servir para hacer distintos análisis estadísticos dentro de un estudio más profundo.

# Resultados:

Como resultado se obtiene un nuevo archivo “csv” simplificado con los datos que son relevantes para el análisis de esta base de datos. Se agregó una nueva columna donde se especifica a qué grupo pertenece según los halla clasificado.

### GRUPO 1:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Sobrevivió | No. Clase | Nombre del pasajero | GRUPO |
| 233 | 0 | 2 | Sjostedt, Mr. Ernst Adolf | 0 |
| 234 | 1 | 3 | Asplund, Miss. Lillian Gertrud | 0 |
| 235 | 0 | 2 | Leyson, Mr. Robert William Norman | 0 |
| 237 | 0 | 2 | Hold, Mr. Stephen | 0 |
| 238 | 1 | 2 | Collyer, Miss. Marjorie "Lottie" | 0 |
| 239 | 0 | 2 | Pengelly, Mr. Frederick William | 0 |
| 240 | 0 | 2 | Hunt, Mr. George Henry | 0 |
| 243 | 0 | 2 | Coleridge, Mr. Reginald Charles | 0 |
| 244 | 0 | 3 | Maenpaa, Mr. Matti Alexanteri | 0 |
| 245 | 0 | 3 | Attalah, Mr. Sleiman | 0 |
| 246 | 0 | 1 | Minahan, Dr. William Edward | 0 |
| 247 | 0 | 3 | Lindahl, Miss. Agda Thorilda Viktoria | 0 |
| 248 | 1 | 2 | Hamalainen, Mrs. William (Anna) | 0 |

### GRUPO 2:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Sobrevivió | No. Clase | Nombre del pasajero | GRUPO |
| 676 | 0 | 3 | Edvardsson, Mr. Gustaf Hjalmar | 1 |
| 677 | 0 | 3 | Sawyer, Mr. Frederick Charles | 1 |
| 678 | 1 | 3 | Turja, Miss. Anna Sofia | 1 |
| 679 | 0 | 3 | Goodwin, Mrs. Frederick (Augusta Tyler) | 1 |
| 680 | 1 | 1 | Cardeza, Mr. Thomas Drake Martinez | 1 |
| 682 | 1 | 1 | Hassab, Mr. Hammad | 1 |
| 683 | 0 | 3 | Olsvigen, Mr. Thor Anderson | 1 |
| 684 | 0 | 3 | Goodwin, Mr. Charles Edward | 1 |
| 685 | 0 | 2 | Brown, Mr. Thomas William Solomon | 1 |
| 686 | 0 | 2 | Laroche, Mr. Joseph Philippe Lemercier | 1 |
| 687 | 0 | 3 | Panula, Mr. Jaako Arnold | 1 |
| 688 | 0 | 3 | Dakic, Mr. Branko | 1 |
| 689 | 0 | 3 | Fischer, Mr. Eberhard Thelander | 1 |

### GRUPO 3:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | Sobrevivió | No. Clase | Nombre del pasajero | GRUPO |
| 1 | 0 | 3 | Braund, Mr. Owen Harris | 2 |
| 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Thayer) | 2 |
| 3 | 1 | 3 | Heikkinen, Miss. Laina | 2 |
| 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | 2 |
| 5 | 0 | 3 | Allen, Mr. William Henry | 2 |
| 7 | 0 | 1 | McCarthy, Mr. Timothy J | 2 |
| 8 | 0 | 3 | Palsson, Master. Gosta Leonard | 2 |
| 9 | 1 | 3 | Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg) | 2 |
| 10 | 1 | 2 | Nasser, Mrs. Nicholas (Adele Achem) | 2 |
| 11 | 1 | 3 | Sandstrom, Miss. Marguerite Rut | 2 |
| 12 | 1 | 1 | Bonnell, Miss. Elizabeth | 2 |
| 13 | 0 | 3 | Saundercock, Mr. William Henry | 2 |

### GRUPO 4:

|  |  |  |  |
| --- | --- | --- | --- |
| 456 | 1 | 3 | Jalsevac, Mr. Ivan |
| 457 | 0 | 1 | Millet, Mr. Francis Davis |
| 459 | 1 | 2 | Toomey, Miss. Ellen |
| 461 | 1 | 1 | Anderson, Mr. Harry |
| 462 | 0 | 3 | Morley, Mr. William |
| 463 | 0 | 1 | Gee, Mr. Arthur H |
| 464 | 0 | 2 | Milling, Mr. Jacob Christian |
| 466 | 0 | 3 | Goncalves, Mr. Manuel Estanslas |
| 468 | 0 | 1 | Smart, Mr. John Montgomery |
| 470 | 1 | 3 | Baclini, Miss. Helene Barbara |
| 472 | 0 | 3 | Cacic, Mr. Luka |
| 473 | 1 | 2 | West, Mrs. Edwy Arthur (Ada Mary Worth) |
| 474 | 1 | 2 | Jerwan, Mrs. Amin S (Marie Marthe Thuillard) |

# Conclusión:

Al final, ya no es necesario que nosotros hagamos la clasificación de los grupos, aquí ya el programa ya hizo el agrupamiento, ya los categorizó en cuatro diferentes tipos y ya solo es cuestión de analizar cada dato en base al resultado que nos ha lanzado el algoritmo.

Por ejemplo, para la tarifa del boleto, es muy probable que haya un grupo bien definido para esta categoría, al ser un grupo de personas de clase socioeconómica alta.

Otro ejemplo sería corroborar si el grupo de mujeres es diferente del grupo de niños o diferente del grupo de hombre, esto dependiendo del número de agrupamientos, el cual puede ser personalizado en el código.

# Referencias:

[*Titanic - Machine Learning from Disaster | Kaggle*](https://www.kaggle.com/c/titanic/data)