**Informe Final**

**Aplicaciones de Datamining en Ciencia y Tecnología**

**Integrantes:** Mario Rossi, Fernando Menéndez, Fabio Zilberman y Juan Ignacio Etcheberry Mason

28 de Octubre de 2018.

**CRISP DM**

**Introducción**

El presente trabajo practico consiste en aplicar la metodología CRISP-DM (*Cross Industry Standard Process for Data Mining*) para el análisis e identificación de estrellas pertenecientes al clúster abierto de la Hiades. CRISP-DM es una de la metodología más utilizada en los proyectos de implementación de minería de datos y está dividida en cuatro niveles de abstracción organizados en forma jerárquica y consta en seis fases distintas, algunas de las cuales son bidireccionales. A continuación, se listan y se describen los distintos pasos y los correspondientes resultados obtenidos:

1. **Fase de Comprensión de Dominio**

* *Objetivos*

El objetivo del presente proyecto es hallar de la forma más fiable estrellas que pertenezcan al clúster *Hiades* a partir de información obtenida de los catálogos *Hipparcos* y *Tycho* que representan los productos primarios de la misión astrométrica *Hipparcos* de la Agencia Espacial Europea (ESA).

* *Evaluar situación*

Para el desarrollo de este proyecto se utilizará información obtenida a partir de un análisis preliminar (pre-informe) realizado cruzando datos obtenidos de la base de datos astronómica *SIMBAD* con los datos del catálogo *Hipparcos*. Esta comparación nos permitió identificar una lista de 50 estrellas presentes en el catálogo de *Hipparcos* que corresponden a elementos del clúster abierto *Hiades*. Estas estrellas serán utilizadas como validación para identificar elementos similares basados en sus características de posición, movimiento y espectro de emisión.

* *Objetivos del Data Mining*

En particular para este proyecto se aplicarán distintos algoritmos de aprendizaje no supervisado. Se utilizará clustering por particiones de modo de agrupar los elementos de los distintos *Datasets* analizados en conjuntos lo más homogéneos posible. Esperamos que esta estrategia nos permita identificar uno o pocos clústeres que contengan a la gran mayoría de las estrellas *Hiades* y que además engloben a otras estrellas que puedan representar candidatas a pertenecer al cluster de las *Hiades*. En este sentido, si bien los algoritmos de clustering son métodos de aprendizaje no supervisado, en el contexto del presente proyecto se aplicará una estrategia de aprendizaje semi supervisado que nos permita utilizar los algoritmos de clustering como clasificadores.

* Plan de proyecto

El proyecto se dividirá en las siguientes etapas para facilitar su organización:

• *Etapa 1*: Análisis de la estructura de los datos y la información de las bases de datos.

• *Etapa 2*: Preparación de los datos (selección, limpieza, conversión y formateo, si fuera necesario) para facilitar la minería de datos sobre ellos.

• *Etapa 3*: Aplicación de distintas técnicas de modelado (en nuestro caso utilizaremos sólo clustering) y ejecución de las mismas sobre los datos. En una primera instancia se analizará el *Dataset* *Hipparcos* y luego el *Dataset* *Thyco*.

• *Etapa 4*: Análisis de los resultados obtenidos en la etapa anterior, si fuera necesario repetir la etapa 3.

• *Etapa 5*: Producción de un informe con los resultados obtenidos en función de los objetivos propuestos.

1. **Fase de Comprensión inicial de Datos**

Esta fase consiste en familiarizarse con los datos y estudiar su calidad y estructura, así como identificar las relaciones más evidentes para formular las primeras hipótesis

* *Colección inicial de Datos*

Los datos utilizados para el desarrollo del trabajo derivan de dos catálogos de estrellas (*Hipparcos* y *Thyco*) que contienen información relativa a estrellas obtenidas por la misión astrométrica *Hipparcos* en el marco del programa científico de la Agencia Europea Espacial (ESA). En particular los datos disponibles corresponden a una región acotada del espacio donde se sabe se encuentran las estrella del clúster abierto de las *Hiades*.

* *Describir los datos*

Cada uno de los Datasets contiene distintas variables relativas a la posición, movimiento y espectro de emisión de las distintas estrellas: RA= ascensión recta, DE = declinación, Plx = paralaje, pmRA = movimiento propio en ascensión recta, pmDE= movimiento propio en declinación, B-V = diferencia de magnitud entre filtro Johnson B (400 - 500 nm rango del azul) y filtro Johnson V (500 - 600 nm corresponde con lo que ve el ojo humano), BT = magnitud entre 400 - 500 nm del sistema fotométrico de Tycho, VT= Magnitud entre 500 - 600 nm del sistema fotométrico de *Tycho*. Las mediciones del catálogo *Hipparcos* (alrededor de 2600 estrellas) fueron realizadas con mayor precisión que las obtenidas en el catálogo *Thyco* (alrededor de 16000). Los datos del catálogo *Hipparcos* fueron utilizados en un análisis preliminar (pre-informe), donde, utilizando sólo criterios de semejanza posicional basados en los parámetros: ascensión recta (RA) y declinación (DE). Mediante este análisis se identificaron y clasificaron una lista de 50 estrellas que correspondían al clúster abierto de las *Hiades* que representan nuestros patrones de referencia para la identificación de candidatas *Hiades* en los dos *Datasets*.

* *Exploración de los datos*

En una primera instancia realizamos una análisis exploratorio para ver la distribución y tipo de variables presentes en los dos *Datasets*.

### Data Frame Summary

\*\*hip\*\*

\*\*N:\*\* 2655

**+----+----------------+---------------------------------+----------------------+----------+---------+**

| No | Variable | Stats / Values | Freqs (% of Valid) | Valid | Missing | **+====+================+=================================+======================+==========+=========+**

| 1 | HIP\ | mean (sd) : 21210.44 (4320.41)\ | 2655 distinct values | 2655\ | 0\ |

| | [numeric] | min < med < max :\ | | (100%) | (0%) |

| | | 13510 < 21089 < 28887\ | | | |

| | | IQR (CV) : 7310 (0.2) | | | |

**+----+----------------+---------------------------------+----------------------+----------+---------+**

| 2 | RA\_J2000\ | mean (sd) : 4.54 (0.91)\ | 2655 distinct values | 2655\ | 0\ |

| | [numeric] | min < med < max :\ | | (100%) | (0%) |

| | | 2.9 < 4.52 < 6.1\ | | | |

| | | IQR (CV) : 1.57 (0.2) | | | |

**+----+----------------+---------------------------------+----------------------+----------+---------+**

| 3 | DE\_J2000\ | mean (sd) : 13.9 (6.16)\ | 2655 distinct values | 2655\ | 0\ |

| | [numeric] | min < med < max :\ | | (100%) | (0%) |

| | | 3.41 < 13.89 < 24.5\ | | | |

| | | IQR (CV) : 10.72 (0.44) | | | |

**+----+----------------+---------------------------------+----------------------+----------+---------+**

| 4 | Plx\ | mean (sd) : 8.83 (10.07)\ | 1361 distinct values | 2655\ | 0\ |

| | [numeric] | min < med < max :\ | | (100%) | (0%) |

| | | 0.03 < 6.06 < 172.78\ | | | |

| | | IQR (CV) : 6.42 (1.14) | | | |

**+----+----------------+---------------------------------+----------------------+----------+---------+**

| 5 | pmRA\ | mean (sd) : 22.5 (74.69)\ | 2245 distinct values | 2655\ | 0\ |

| | [numeric] | min < med < max :\ | | (100%) | (0%) |

| | | -330.66 < 4.78 < 1999.05\ | | | |

| | | IQR (CV) : 28.99 (3.32) | | | |

**+----+----------------+---------------------------------+----------------------+----------+---------+**

| 6 | pmDE\ | mean (sd) : -29.4 (72.37)\ | 2196 distinct values | 2655\ | 0\ |

| | [numeric] | min < med < max :\ | | (100%) | (0%) |

| | | -1570.64 < -13.19 < 238.42\ | | | |

| | | IQR (CV) : 26.9 (-2.46) | | | |

**+----+----------------+---------------------------------+----------------------+----------+---------+**

| 7 | Vmag\ | mean (sd) : 8.34 (1.51)\ | 656 distinct values | 2655\ | 0\ |

| | [numeric] | min < med < max :\ | | (100%) | (0%) |

| | | 0.45 < 8.42 < 12.66\ | | | |

| | | IQR (CV) : 1.8 (0.18) | | | |

**+----+----------------+---------------------------------+----------------------+----------+---------+**

| 8 | B-V\ | mean (sd) : 0.63 (0.46)\ | 1269 distinct values | 2640\ | 15\ |

| | [numeric] | min < med < max :\ | | (99.44%) | (0.56%) |

| | | -0.22 < 0.56 < 3.1\ | | | |

| | | IQR (CV) : 0.66 (0.73) | | | |

**+----+----------------+---------------------------------+----------------------+----------+---------+**

| 9 | Symbad\_Hyades\ | | 2605 (98.1%)\ | 2655\ | 0\ |

| | [logical] | | 50 ( 1.9%) | (100%) | (0%) |

**+----+----------------+---------------------------------+----------------------+----------+---------+**

### Data Frame Summary

\*\*tyc\*\*

\*\*N:\*\*

16258

+----+--------------+-----------------------------------+-----------------------+----------+----------+

| No | Variable | Stats / Values | Freqs (% of Valid) | Valid | Missing | +====+==============+===================================+=======================+==========+==========+

| 1 | recno\ | mean (sd) : 73094.02 (45031.95)\ | 16258 distinct values | 16258\ | 0\ |

| | [numeric] | min < med < max :\ | | (100%) | (0%) |

| | | 3569 < 57089.5 < 156305\ | | | |

| | | IQR (CV) : 52262.5 (0.62) | | | |

+----+--------------+-----------------------------------+-----------------------+----------+----------+

| 2 | TYCID1\ | mean (sd) : 897.7 (535.94)\ | 283 distinct values | 16258\ | 0\ |

| | [numeric] | min < med < max :\ | | (100%) | (0%) |

| | | 51 < 723 < 1868\ | | | |

| | | IQR (CV) : 638 (0.6) | | | |

+----+--------------+-----------------------------------+-----------------------+----------+----------+

| 3 | TYCID2\ | mean (sd) : 893.51 (635.11)\ | 2616 distinct values | 16258\ | 0\ |

| | [numeric] | min < med < max :\ | | (100%) | (0%) |

| | | 1 < 789 < 3944\ | | | |

| | | IQR (CV) : 842 (0.71) | | | |

+----+--------------+-----------------------------------+-----------------------+----------+----------+

| 4 | TYCID3\ | mean (sd) : 1 (0.02)\ | 1 : 16254 (100.0%)\ | 16258\ | 0\ |

| | [numeric] | min < med < max :\ | 2 : 4 ( 0.0%) | (100%) | (0%) |

| | | 1 < 1 < 2\ | | | |

| | | IQR (CV) : 0 (0.02) | | | |

+----+--------------+-----------------------------------+-----------------------+----------+----------+

| 5 | RA\_J2000\_24\ | mean (sd) : 4.86 (0.91)\ | 16258 distinct values | 16258\ | 0\ |

| | [numeric] | min < med < max :\ | | (100%) | (0%) |

| | | 2.9 < 5.06 < 6.1\ | | | |

| | | IQR (CV) : 1.49 (0.19) | | | |

+----+--------------+-----------------------------------+-----------------------+----------+----------+

| 6 | DE\_J2000\ | mean (sd) : 13.97 (6.24)\ | 16258 distinct values | 16258\ | 0\ |

| | [numeric] | min < med < max :\ | | (100%) | (0%) |

| | | 3.4 < 14.05 < 24.5\ | | | |

| | | IQR (CV) : 11.02 (0.45) | | | |

+----+--------------+-----------------------------------+-----------------------+----------+----------+

| 7 | pmRA\ | mean (sd) : 6.26 (23.53)\ | 1480 distinct values | 16258\ | 0\ |

| | [numeric] | min < med < max :\ | | (100%) | (0%) |

| | | -149.3 < 1.8 < 198.2\ | | | |

| | | IQR (CV) : 13 (3.76) | | | |

+----+--------------+-----------------------------------+-----------------------+----------+----------+

| 8 | pmDE\ | mean (sd) : -12.44 (20.92)\ | 1349 distinct values | 16258\ | 0\ |

| | [numeric] | min < med < max :\ | | (100%) | (0%) |

| | | -199.9 < -7.6 < 177\ | | | |

| | | IQR (CV) : 14.2 (-1.68) | | | |

+----+--------------+-----------------------------------+-----------------------+----------+----------+

| 9 | BT\ | mean (sd) : 10.74 (1.22)\ | 4343 distinct values | 16258\ | 0\ |

| | [numeric] | min < med < max :\ | | (100%) | (0%) |

| | | 2.79 < 11.04 < 12.85\ | | | |

| | | IQR (CV) : 1.36 (0.11) | | | |

+----+--------------+-----------------------------------+-----------------------+----------+----------+

| 10 | VT\ | mean (sd) : 9.94 (1.1)\ | 4048 distinct values | 16258\ | 0\ |

| | [numeric] | min < med < max :\ | | (100%) | (0%) |

| | | 0.77 < 10.22 < 11.95\ | | | |

| | | IQR (CV) : 1.23 (0.11) | | | |

+----+--------------+-----------------------------------+-----------------------+----------+----------+

| 11 | V\ | mean (sd) : 9.87 (1.1)\ | 15844 distinct values | 16258\ | 0\ |

| | [numeric] | min < med < max :\ | | (100%) | (0%) |

| | | 0.58 < 10.14 < 11.93\ | | | |

| | | IQR (CV) : 1.24 (0.11) | | | |

+----+--------------+-----------------------------------+-----------------------+----------+----------+

| 12 | B-V\ | mean (sd) : 0.68 (0.45)\ | 3439 distinct values | 16258\ | 0\ |

| | [numeric] | min < med < max :\ | | (100%) | (0%) |

| | | -0.46 < 0.56 < 3.39\ | | | |

| | | IQR (CV) : 0.66 (0.67) | | | |

+----+--------------+-----------------------------------+-----------------------+----------+----------+

| 13 | HD\ | mean (sd) : 137082.43 (117395.4)\ | 5751 distinct values | 5758\ | 10500\ |

| | [numeric] | min < med < max :\ | | (35.42%) | (64.58%) |

| | | 18019 < 39727.5 < 287468\ | | | |

| | | IQR (CV) : 219645 (0.86) | | | |

+----+--------------+-----------------------------------+-----------------------+----------+----------+

| 14 | HIP\ | mean (sd) : 21327.99 (4326.46)\ | 2470 distinct values | 2483\ | 13775\ |

| | [numeric] | min < med < max :\ | | (15.27%) | (84.73%) |

| | | 13526 < 21257 < 28882\ | | | |

| | | IQR (CV) : 7368.5 (0.2) | | | |

+----+--------------+-----------------------------------+-----------------------+----------+----------+

| 15 | Plx\ | mean (sd) : 7.99 (6.45)\ | 1135 distinct values | 2253\ | 14005\ |

| | [numeric] | min < med < max :\ | | (13.86%) | (86.14%) |

| | | 1.01 < 6.09 < 87.9\ | | | |

| | | IQR (CV) : 5.52 (0.81) | | | |

+----+--------------+-----------------------------------+-----------------------+----------+----------+

* *Verificar calidad de los datos*

Todas las variables de los dos *Datasets* analizados (salvo “*Symbad\_Hyades*” del Dataset Hipparcos que creamos nosotros) son numéricas y la gran mayoría presenta el 100% de datos validos. Existen solo 15 registros de *Dataset Hipparcos* (0.56%) que presentan datos faltantes en la variable ***B-V***. Por el contrario, en el *Dataset Thyco* observamosque la variable ***Plx*** presenta un número muy alto de faltantes (86.14%). Una situación similar presentan las variables ***HD*** y ***HIP*** del *Dataset Thyco*. Estas dos ultimas variables representan la ID cruzada del catálogo *Thyco* con el catálogo HD y Hipparcos respectivamente, y en el caso de la ultima sirve de referencia para poder combinar los datos de los dos *Datasets* analizados.

1. **Fase de Preparación de los datos (/02.Preprocesamiento/pre\_procesamiento.rmd )**

En esta fase de la metodología se trata de preparar los datos para adecuarlos a las técnicas de minería de datos que se van a emplear sobre ellos. Esto implica seleccionar el subconjunto de datos que se va a utilizar, limpiarlos para mejorar su calidad, añadir nuevos datos a partir de los existentes y darles el formato requerido por la herramienta de modelado.

* *Obtener conjunto inicial de datos*

Los *Datasets* antes descriptos representan nuestro conjunto de datos iniciales y cabe aclarar que para la clusterización no se incluyen las variables que dan información de ID cruzada. Para ambos casos sin embargo, preservamos el campo *symbad hyade*s para luego mapear las *Hiades* por clúster cómo método de validación.

* *Limpieza de los datos*

En análisis de clúster, un método usualmente aceptado es remover instancias que posean faltantes de la clusterización inicial. Luego con clúster bien definidos, se suele intentar mapear las instancias con faltantes a los clúster más parecidos.

Con esta posibilidad en mente, removemos 15 estrellas del *Dataset* de *Hipparcos* para trabajar con un conjunto de datos completos a la hora de clusterizar. En este sentido, para el caso del *Dataset* *Thyco*, la variable ***Plx*** presenta un número muy elevado de faltantes (86.14%) y por lo tanto no se la consideró para el análisis de clusterización. Asimismo, se procedió a remover las columnas de IDs otros catálogos del *Dataset* *Thyco*, ***TYCID1***, ***TYCID2*** y ***TYCID3***, ***HD*** y ***HIP***.

* *Integración de los datos*

Hemos considerado que no es necesario la creación de nuevas estructuras (campos, registros, etc). Sin embargo, con el fin de no repetir el procedimiento para los dos *Datasets*, eliminamos las estrellas del catálogo *Hipparcos* del *Dataset* *Thyco*,excepto las estrellas Hiades que servirán como grupo de control para el análisis.

* *Formateo de los datos*

Para ambos *Datasets*, se procedió a escalar y estandarizar las variables para evitar que las dimensiones de estas incidan o afecten las medidas de distancia dentro del análisis de clúster.

Para las variables RA\_J2000 y DE\_J2000, se realizará un escalado especial que permita conservar las proporciones entre ambas variables, ya que consideramos que una “estandarización” para estas dos podría resultar en una posible distorsión de la medición que se posee de la proyección de la estrella en el espacio.

1. **Fase de Modelado (/03.Modelo\_Evaluacion/modelo\_evaluacion.Rmd** para **Hipparcos y /03.Modelo\_Evaluacion/tycocloud.R** para **Thyco)**

En esta fase de la metodología se escogerá la técnica (o técnicas) más apropiadas para los objetivos marcados de la minería de datos.

* *Seleccionar la técnica del modelo*

Debido a que las consignas especificas del trabajo práctico nos piden realizar un análisis de clusterización no jerarquica utilizamos en una primera instancia pruebas sólo con el *Dataset Hipparcos* con el método de K-medias y El método de clustering por prototipos PAM (*Partition around Medoids*).

* *Generar diseño de prueba*

Par

* *Construir el modelo*
* *Analizar el modelo*

1. Evaluación
2. Evaluación de los resultados
3. Proceso de revisión
4. Próximos pasos
5. Conclusiones
6. Apéndice
7. **Punto adicional I – Estrategia de búsqueda y complejidad algorítmica**

En un proceso de búsqueda de equivalencias entre estrellas de dos lotes, donde el patrón de equivalencia es la similitud de coordenadas, utilizaremos la distancia entre las coordenadas de ambos lotes para determinar cuáles son los pares correspondientes.

El algoritmo más trivial es medir las distancias de todas las estrellas de un lote contra todas las del otro. Elegir el par con la menor distancia de todas. Se quitan ambas y se realiza toda la comparación nuevamente.

Dos lotes de estrellas: A y B

La cantidad de estrellas en el lote A es NA, y la cantidad de estrellas del lote B es NB

* 1. Calculamos NA x NB distancias, armamos una tabla con ID de estrella A, ID de estrella B y la distancia
  2. Ordenamos la tabla por distancias de menor a mayor
  3. Elegimos el primer elemento. Marcamos la estrella del lote A y la del lote B como "par"
  4. Quitamos de la tabla todas las filas que contengan esas estrellas, con el resto de la tabla volvemos al punto 3

La complejidad del proceso es:

1. un cálculo de distancias de tamaño NA \* NB
2. recorrido de una tabla de distancias de NA \* NB filas

El algoritmo de comparación por grillas propuesto en el TP, propone asignar a cada estrella de ambos a la celda de una grilla utilizando la discretización de los valores de las coordenadas para determinar la celda.

Luego realizaremos los cálculos de distancias entre estrellas de ambos lotes solo de las celdas próximas.

Dos lotes de estrellas: A y B

La cantidad de estrellas en el lote A es NA, y la cantidad de estrellas del lote B es NB

La grilla es de K filas y K columnas

La zona de vecindad es una grilla pequeña de L \* L celdas

1. Discretizamos las coordenadas de las NA + NB estrellas y las asignamos a la celda correspondiente, en promedio cada celda recibe NA/K^2 y NB/K^2 estrellas

2. Las zonas de vecindad son de a lo sumo L \* L celdas. Ya que las de los bordes tienen menos celdas.

3. Calculamos las distancias entre cada estrella del Lote A, y las estrellas de las celdas vecinas a esta. La tabla de distancias tiene NA \* NB/K^2 \* L^2 filas

4. Elegimos el primer elemento. Marcamos la estrella del lote A y la del lote B como "par"

5. Quitamos de la tabla todas las filas que contengan esas estrellas, con el resto de la tabla volvemos al punto 6

La complejidad del proceso es:

* 1. discretización de (NA + NB) \* 2 coordenadas
  2. un cálculo de distancias de tamaño NA \* NB \* L^2 / K^2
  3. recorrido de una tabla de distancias de NA \* NB \* L^2 / K^2 filas

Por lo tanto, cuanto mayor sea la relación entre K y L (K/L), más eficiente y menos costoso en tiempo de cálculo resultara este segundo algoritmo.

Generalizando para un tamaño teórico de lotes de N, el proceso de comparación por grilla va a ser de complejidad O(N^2 \* L^2 / K^2) comparado contra O(N^2)

Comparando Symbad (NA = 178) con Hipparcos (NB = 2655)

Todos contra Todos:

Time difference of 5.795763 mins

472590 distancias

Time difference of 5.858445 mins total

50x50 celdas / 3x3 celdas

Time difference of 4.019846 secs

2394 distancias

Time difference of 4.202339 secs

197 veces más rápido en cantidad de distancias calculadas (comparado vs Todos contra Todos)

83.6 veces más rápido en tiempo de computo (comparado vs Todos contra Todos)

30x30 celdas / 3x3 celdas

Time difference of 5.907489 secs

6281 distancias

Time difference of 5.998217 secs

75 veces más rápido en cantidad de distancias calculadas (comparado vs Todos contra Todos)

58.7 veces más rápido en tiempo de computo (comparado vs Todos contra Todos)

50x50 celdas / 5x5 celdas

Time difference of 7.251834 secs

6292 distancias

Time difference of 7.328629 secs

75 veces más rápido en cantidad de distancias calculadas (comparado vs Todos contra Todos)

47.9 veces más rápido en tiempo de computo (comparado vs Todos contra Todos)

1. **Punto adicional II – Análisis Datos Faltantes en Tycho**
2. Punto adicional III
3. Punto adicional IV
4. Bibliografía