Prueba de Minería de Datos

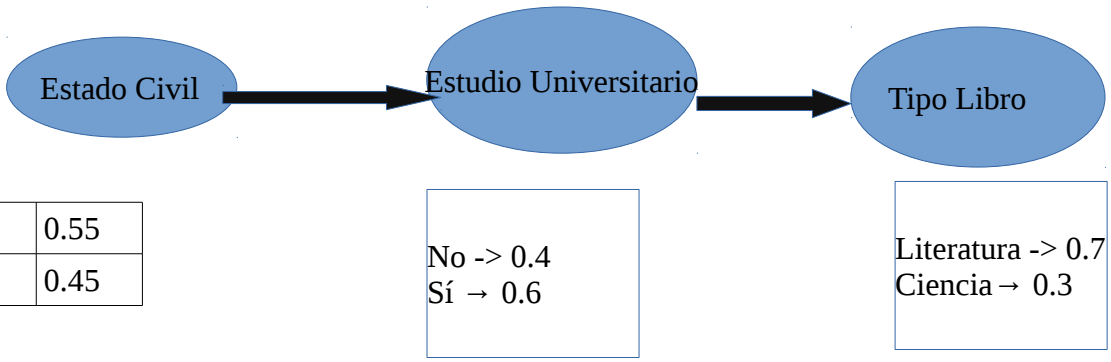
Nombre: Juan Erazo

Explicar con un ejemplo cómo se usa el teorema de Bayes Navis para armar un clasificador

Una empresa de venta de libros quiere ofrecer el nuevo texto que les ha llegado. Y para ello, se tiene un historial de los gustos de cada cliente con el gusto del libro:

Cliente	**Estado Civil**	**Estudios universitarios**	**Libro de literatura**	**Libro de ciencia**
Cliente1	S	No	Sí	No
Cliente2	C	Sí	No	Sí
…	.	..	..	...

Para utilizar el método Bayes Naive se clasifica a los clientes asignándoles una probabilidad de compra basándose en el historial que se tiene

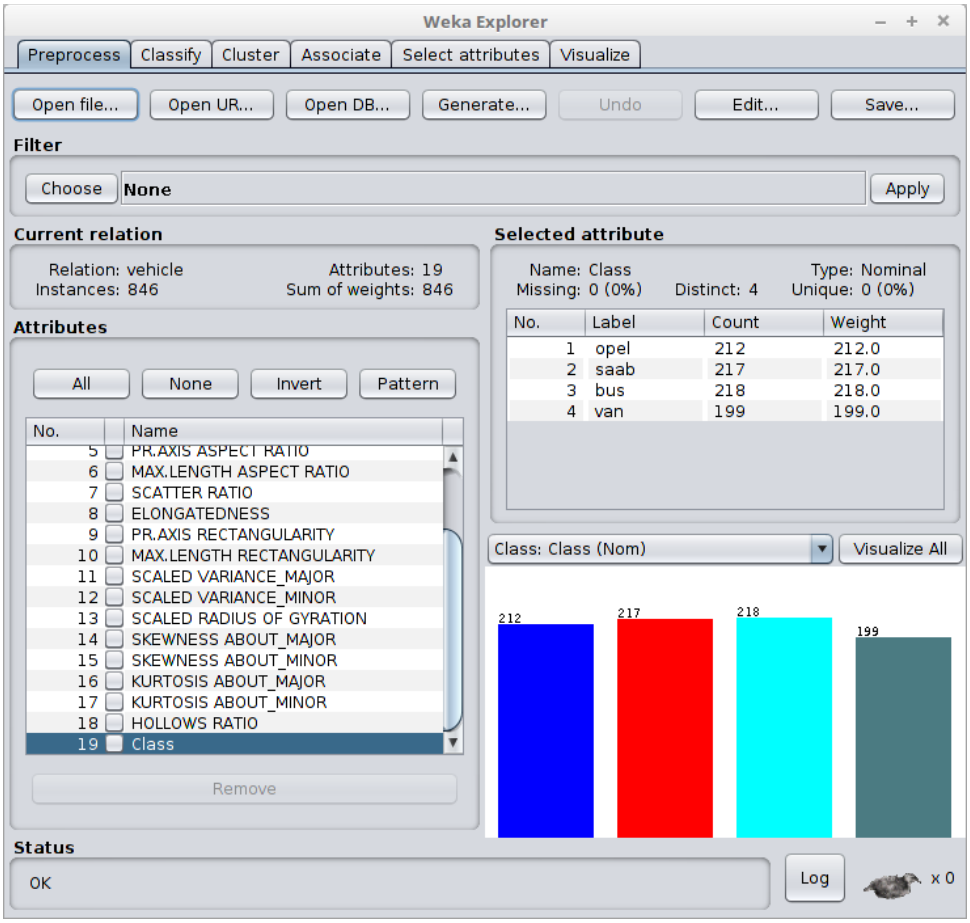


Y se multiplica cada probabilidad, con ese valor se podrá predecir si un cliente está interesado en el nuevo libro que se tiene

Explicar con un ejemplo un algoritmo de clustering

El algoritmo EM (Esperanza-Maximización) consiste en asignar a cada instancia una distribución de probabilidad de pertenencia a cada cluster. El algoritmo puede decidir cuántos clusters crear basado en validación cruzada o se le puede especificar a priori cuantos debe generar. EM Utiliza el modelo Gaussiano finito de mezclas, asumiendo que todos los atributos son variables aleatorias independientes.

Para realizar un ejemplo se utilizará la herramienta **Weka** en donde cargaremos un archivo de extensión **.arff** que tendrá un conjunto de datos que corresponden a las características de de diferentes tipos de vehículo:



Los atributos que tendrá cada vehículo son:

- COMPACTNESS
- CIRCULARITY
- DISTANCE CIRCULARITY
- RADIUS RATIO
- PR.AXIS ASPECT RATIO
- MAX.LENGTH ASPECT RATIO
- Sca ATTER RATIO
- ELONGATEDNESS
- PR.AXIS RECTANGULARITY
- MAX.LENGTH RECTANGULARITY
- SCALED VARIANCE_MAJOR
- SCALED VARIANCE_MINOR
- SCALED RADIUS OF GYRATION
- SKEWNESS ABOUT_MAJOR
- SKEWNESS ABOUT_MINOR
- KURTOSIS ABOUT_MAJOR
- KURTOSIS ABOUT_MINOR
- HOLLOWS RATIO

Y los tipo de vehículo son:

- OPEL
- SAAB

- VAN
- BUS

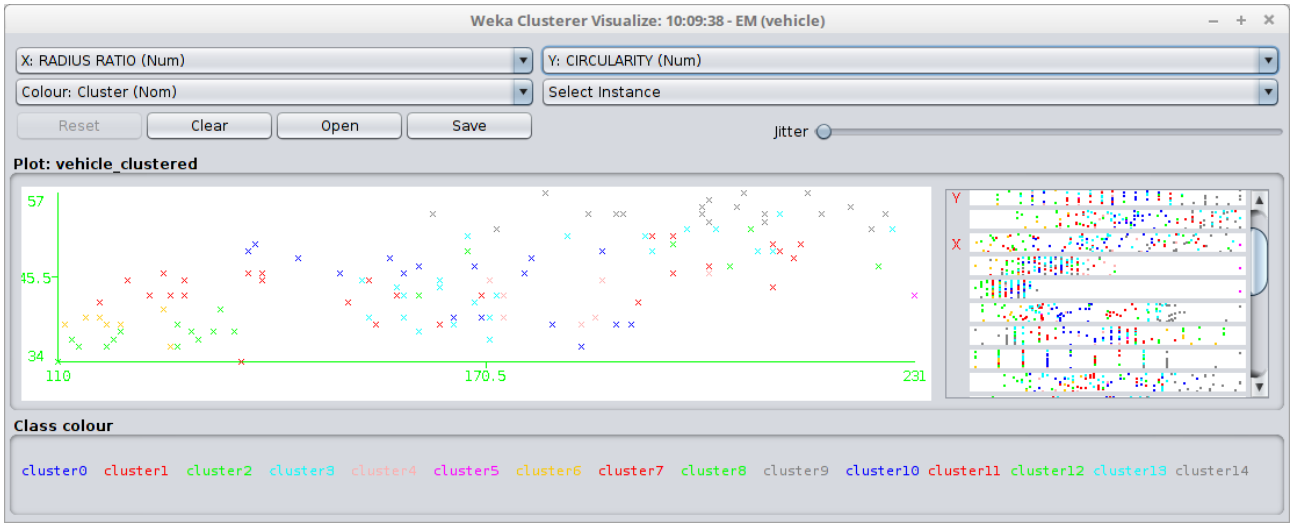
En vista de que es más elaborado requiere muchas más operaciones y se demorará de acuerdo al número de instancias con que se trabaje.

Entonces, este algoritmo trabajó con 846 instancias del archivo .arff para buscar el número de grupos apropiados. Se demoró 30.79 segundos en total para obtener 15 clusters:

Clustered Instances

0	11 (9%)
1	13 (10%)
2	6 (5%)
3	15 (12%)
4	7 (6%)
5	1 (1%)
6	7 (6%)
7	10 (8%)
8	4 (3%)
9	16 (13%)
10	8 (6%)
11	7 (6%)
12	9 (7%)
13	10 (8%)
14	3 (2%)

En el siguiente gráfico se puede apreciar los 14 clústers dispersos en las variables RADIUS RATIO y CIRCULARITY:



Y si tomamos las característica de cierto clúster, se puede apreciar de la siguiente manera: