



Minería de Datos

Profesor: Andrés Medina
Ayudante: Leonel Vega
Aplicación: Customer Analytics.

Consideraciones:

- a) Tarea grupal (máximo 5 alumnos).
- b) Se debe adjuntar el notebook de python y un informe autocontenido de no más de 3 páginas presentando los resultados más interesantes de acuerdo al contexto del problema. Asegurarse que el código funcione, de lo contrario no se evaluará.
- c) **Plazo: 01 de octubre 23:59**

Ustedes trabajan en el área de riesgo de una conocida empresa crediticia chilena que entrega créditos a agricultores. Si bien, Dentro de sus tareas es desarrollar la construcción de modelos predictivos (en particular, modelos de comportamiento crediticio o Behavioral Scoring), sin embargo, en esta etapa, *a la gerencia le interesa conocer la cartera de clientes en el sentido de poder generar segmentaciones y ofrecer productos diferenciados. De esta forma, se pretende aumentar el revenue de la estrategia comercial.* Para ello dispone de una serie de variables asociadas a los clientes antiguos, incluyendo el resultado del crédito (si el cliente fue *bueno* o *malo*).

Para esta labor, esta contra el tiempo, y su manager también por lo que este, le pide lo siguiente:

1. Realice un pre-procesamiento a los datos. Esto implica realizar un proceso de limpieza, selección, y transformación de variables. Justifique cada una de las medidas adoptadas en el análisis.
2. Describa en forma breve el modelo K-means para que lo entienda una persona sin conocimiento en el tema de la ingeniería. Seleccione y modifique, justificadamente, las variables pertinentes, y determine el mejor número de clusters que se puede obtener utilizando el método K-means. Además, analice los clusters generados y determine si existe algún cluster relevante.
3. Describa en forma breve el modelo DBSCAN para que lo entienda una persona sin conocimiento en el tema de la ingeniería. Seleccione y modifique, justificadamente, las variables pertinentes, y determine un buen set de parámetros para aplicar el método DBSCAN. Además, analice los clusters generados y determine si existe algún cluster relevante.

4. Describa en forma breve el modelo clustering jerárquico aglomerativo, con alguna de las variaciones pertinentes, para que lo entienda una persona sin conocimiento en el tema de la ingeniería. Seleccione y modifique, justificadamente, las variables pertinentes, y aplique el algoritmo de clustering jerárquico aglomerativo con la variación seleccionada. Analice el dendrograma y seleccione, justificadamente, un número de cluster pertinente. Además, analice los clusters generados y determine si existe algún cluster relevante.
5. Describa en forma breve el modelo C-means para que lo entienda una persona sin conocimiento en el tema de la ingeniería. Seleccione y modifique, justificadamente, las variables pertinentes, y determine el mejor número de clusters que se puede obtener utilizando el método C-means. Además, analice los clusters generados y determine si existe algún cluster relevante.
6. De todas las alternativas anteriores, seleccione la mejor estrategia encontrada de acuerdo a métricas de performance y sin descuidar el sentido de negocio de la segmentación. Desarrolle un informe con estos resultados.

Cabe destacar que el informe generado, se presentará en un comite ejecutivo donde los gerentes de productos financieros e inteligencia de negocios no conoce nada de clustering. Por lo tanto, debe ser redactado en un lenguaje claro.

1. Anexo

La base contiene 72.894 clientes.

Variable	Descripción	Tipo Variable
Sol	Número de solicitud	ID
Cliente	Código de Id. Cliente	ID
region	Región del Cliente	Catégorica
RUBRO	Rubro de la empresa	Catégorica
Aval	Si tiene aval para el crédito	Binaria
Garantía	Si crédito posee garantía	Binaria
Valor_Garantía	Valor de la garantía (UF)	Numérica
Edad	Edad del cliente	Numérica
Tenencia	Propiedad del terreno.	Catégorica
Predios	Cantidad de predios del cliente	Numérica
monto_coloc	Monto del crédito	Numérica
Plazo	Plazo en meses del crédito	Numérica
Cred_Hist	Cantidad total de créditos históricos que ha tenido el cliente	Numérica
Cred_Vig	Cantidad de créditos vigentes que el cliente tiene actualmente	Numérica
Mora_Max	Mora máxima en días que ha tenido en algún crédito histórico	Numérica
Defaulter	Si cayó en cartera vencida dentro de los primeros 12 meses del crédito	Binaria, Obj.