
UNIVERSIDAD DE SEVILLA

UNIVERSIDAD DE SEVILLA
MASTER EN INGENIERÍA DEL SOFTWARE: CLOUD DATOS Y GESTIÓN DE LAS TECNOLOGÍAS

FUNDAMENTOS DE INGENIERÍA DE DATOS

APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS A CONJUNTO DE DATOS BANCARIO DE DIFERENTES CLIENTES PARA PREDECIR SI SERÁN PERSUADIDOS O NO.

BOUCHAIB, EL ASRI
ENRIQUEZ CAICEDO, BRIAN DANILO
ESTRADA NIETO, JUAN CARLOS
PALMA, JUAN JOSÉ
PREEL, LUCAS
SEVILLA, Enero, 2022

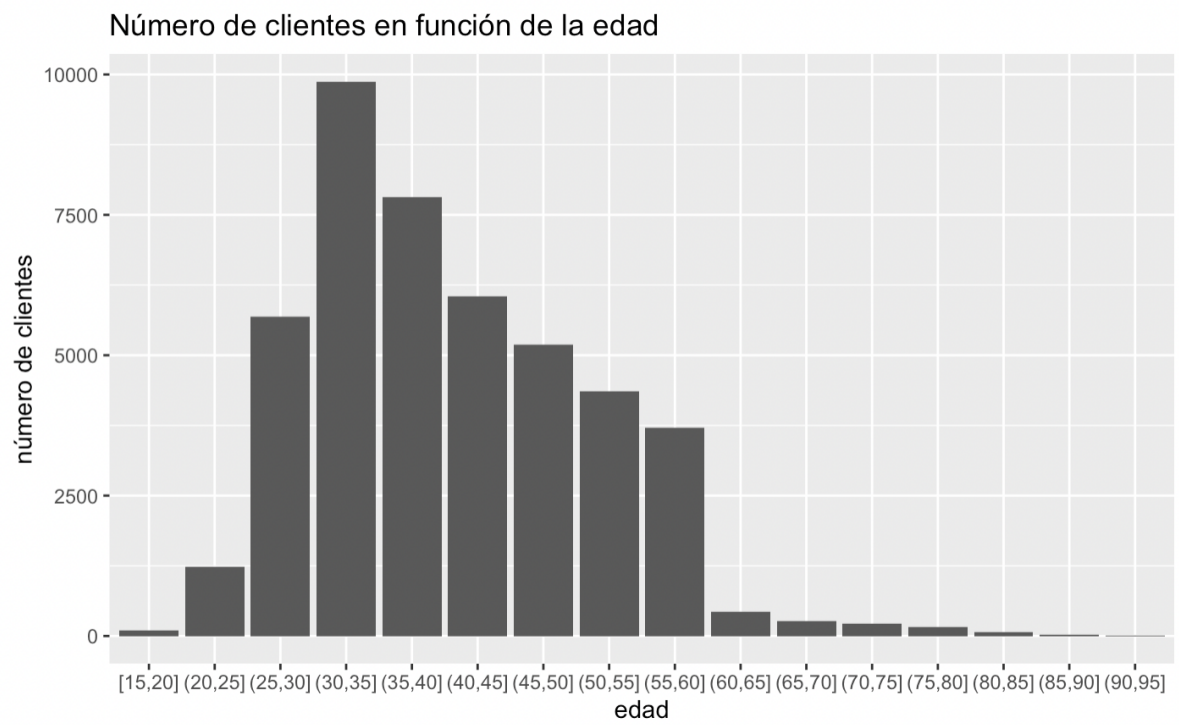
ÍNDICE

DATASET

El conjunto de datos que elegimos es un conjunto de datos bancarios de marketing el cual brinda información sobre una campaña de una entidad bancaria; estos datos deben servir para mejorar campañas futuras a realizar.
Para lograr este objetivo se deben identificar patrones que nos ayuden a concluir el desarrollo de futuras estrategias de marketing.

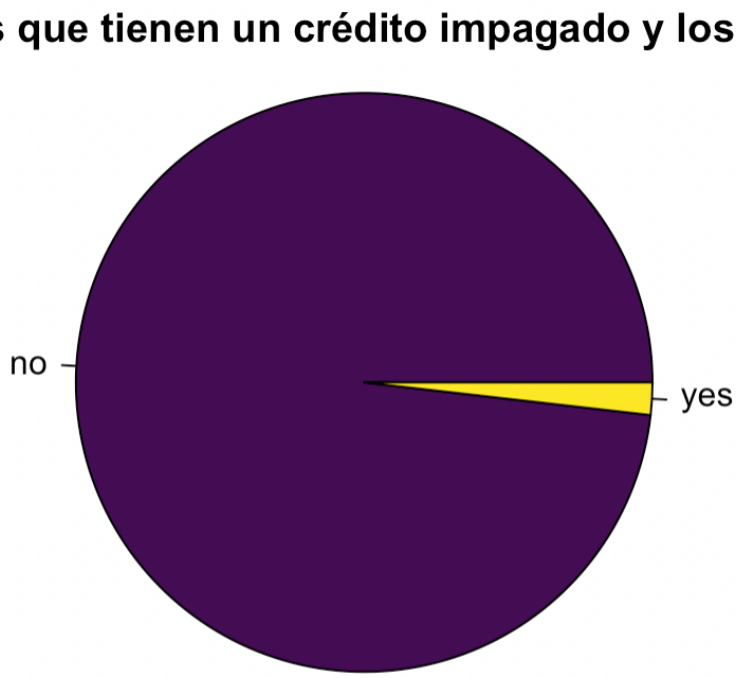
EXPLORACIÓN DE LOS DATOS

El dataset está compuesto por 17 atributos, a los cuales a todos se les realizó visualización pero se destacan: “age”, “balance”, “default”; ya que en ellas se ven reflejadas las diferentes formas de visualización de datos utilizadas en todos los atributos.



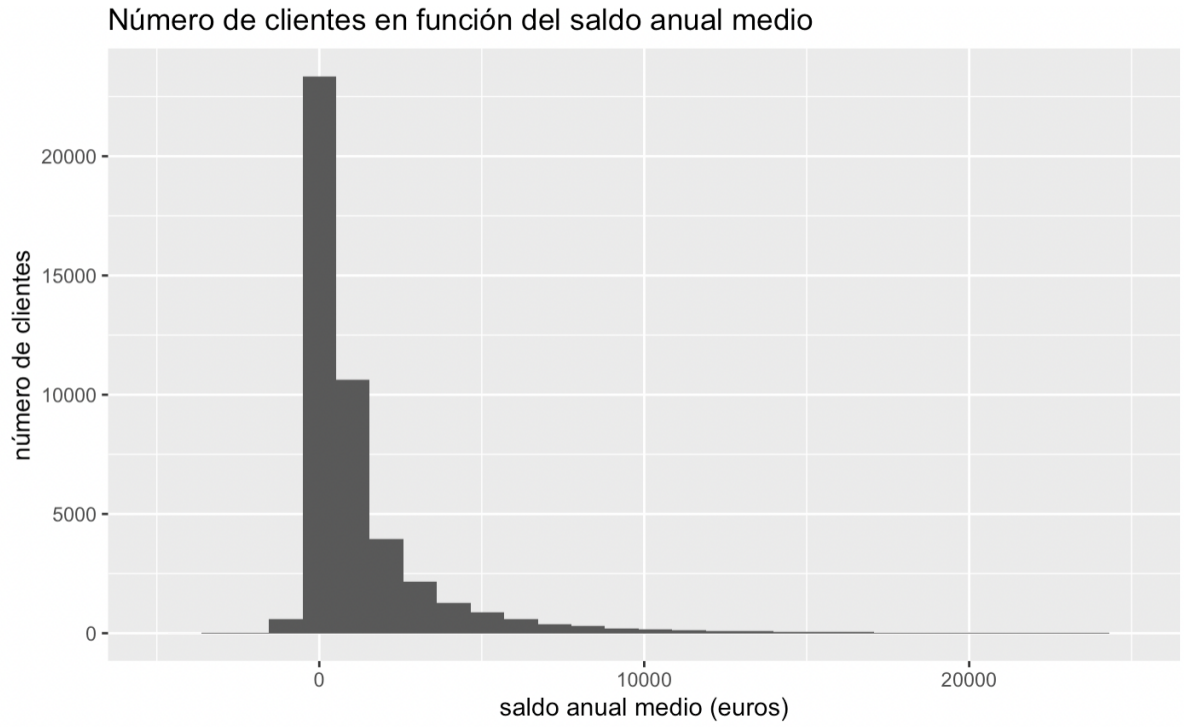
La visualización del atributo age se agrupó en un rango de cinco, en un gráfico de barras y se concluye lo siguiente:

- En la gráfica se puede observar que la mayoría de los clientes están en el rango de edad entre los 25 a 60 años.
- El rango de edad con más clientes es el de 30 a 35 años.
- Se deduce que los clientes son personas con una vida laboral activa.



La visualización del atributo default se realizó en una gráfico de torta y se concluye lo siguiente:

- Se observa que la mayoría de clientes tienen crédito con cartera sana.



A la visualización del atributo balance se le dió un rango mínimo de 5000 y máximo de 25000 en el eje X, esto se representó en un gráfico de barras y se concluye lo siguiente:

- Hay clientes con saldos anuales promedios mayores a 20000 y se pueden considerar como outliers.
- El rango del saldo anual medio con mayor número de clientes está entre 0 y 2500 euros.

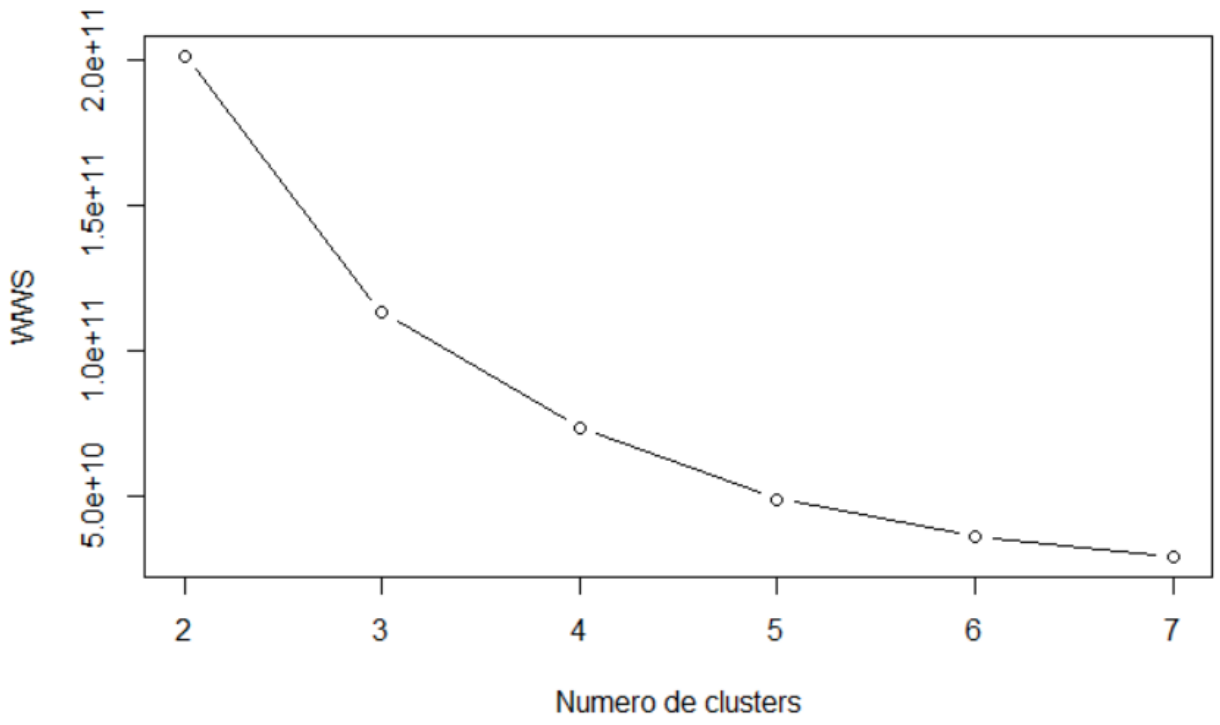
APLICACIÓN DE ALGORITMOS DE CLUSTERING

Se ha decidido utilizar técnicas de clustering para verificar si se pueden encontrar agrupaciones en los datos que pudieran aportar algún conocimiento. Para ello se procedió a hacer uso de 2 técnicas de clasificación aprendidas durante las clases de teoría.

Kmeans

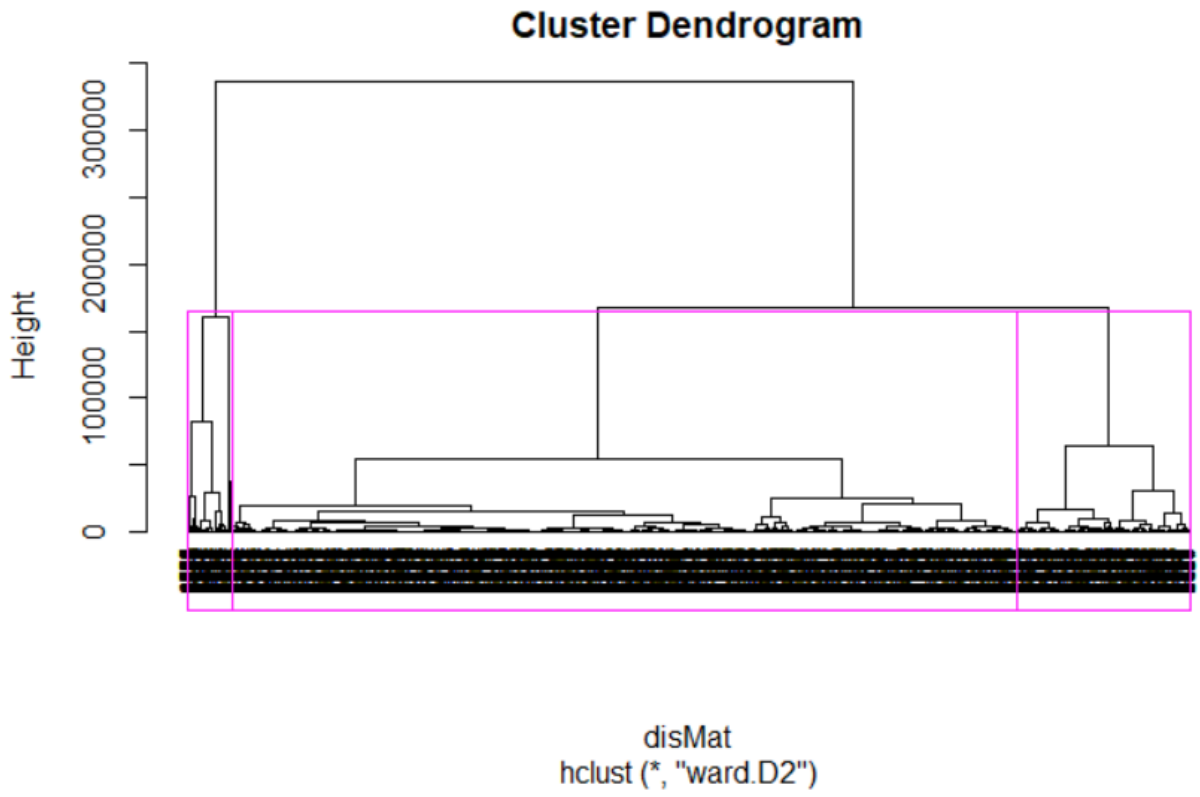
Como es bien sabido este es un algoritmo no supervisado que agrupa objetos en k grupos basándose en sus características minimizando la suma de distancias entre cada objeto y el centroide de su grupo o cluster. Se suele usar la distancia cuadrática.

En este caso se hizo uso del algoritmo para obtener la métrica Within cluster Sum of Square, esta mide la distancia entre datos que pertenecen a un mismo cluster, lo que realmente se puede obtener de esta, es el número adecuado de clusters tomando como referencia el “codo” de la siguiente gráfica.



Jerárquico

Se graficó un dendrograma con los datos de prueba para poder visualizar mejor la información y así poner en práctica lo aprendido en clases. Se ha utilizado un k = 3 ya que ese fue el valor óptimo obtenido con el algoritmo kmeans.



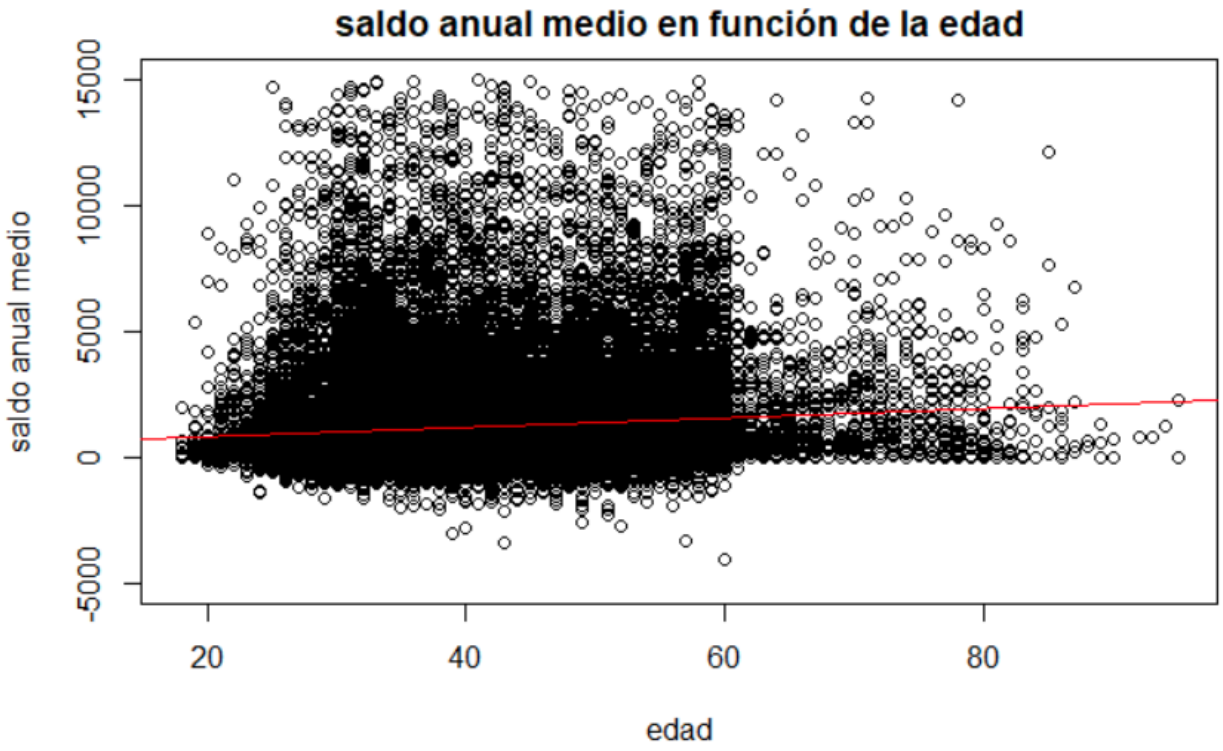
Se pueden observar 3 grandes grupos pero no se pueden ver los detalles de los datos porque son muchos, a pesar de haberse usado los datos de prueba.

TÉCNICAS DE REGRESIÓN

Regresión Simple

Se busca verificar si existe una relación entre la edad y el balance o saldo anual medio.

Para aplicar esta técnica de aprendizaje supervisado se tuvo que hacer un preprocesamiento, eliminando algunos valores outliers con la función filter, después se procedió a graficar las variables balance y edad, dibujando en ellas la pendiente o línea de regresión lineal.



Se pudo observar una cierta tendencia que a mayor edad mayor balance o saldo en la cuenta.

Después se procedió a verificar las métricas para ver cuán bueno es este resultado. La métrica R_square obtuvo un resultado de 0.00981882902467823,

bastante cercano a 0, lo que quiere decir que utilizar esta regresión para predecir el balance en base a la edad no sería muy acertado. Mientras el valor de esa métrica sea más cercano a 1 será más acertada la predicción, mientras más cercano a cero, menos acertada será.

Regresión multivariable

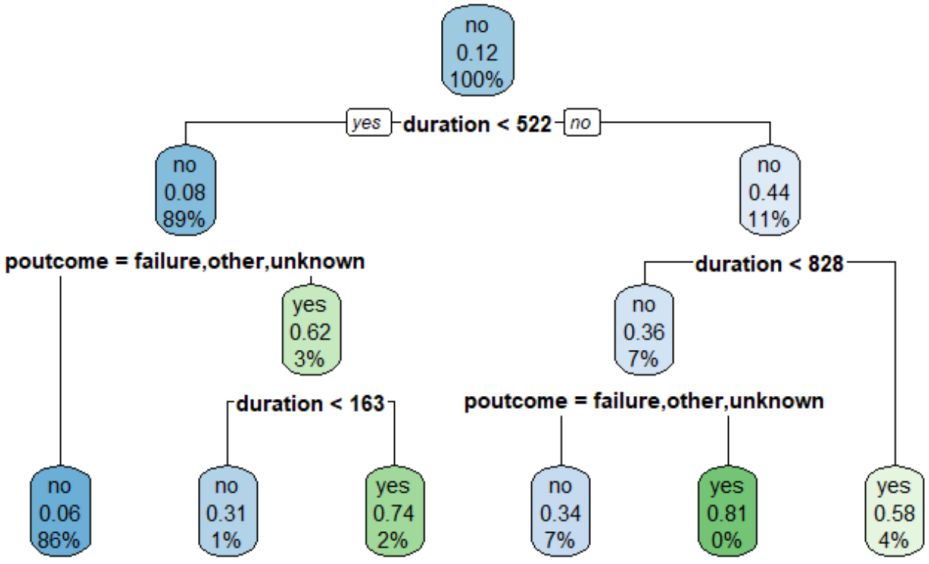
Dado que la regresión lineal simple(en base a una sola variable independiente) no fue exitosa, Ahora se intentará predecir el balance pero mirando a todas las otras variables y no sólo la edad.

El valor de R_squared es 0.06, es decir, que solo el 6% de la variación del saldo anual medio se puede explicar con las variables de entrada, lo que significa que el modelo no se ajusta bien a los datos. Se puede inferir que las otras variables no tienen mucha influencia en el saldo anual medio o balance en la cuenta.

CLASIFICACIÓN

Árbol de decisión

Rpart (Rercursive Partitioning And Regression Trees) es una librería que permite entrenar árboles de decisión para hacer predicciones. Se va a utilizar para construir un árbol de decisión que permite predecir el valor de la variable "y" que representa la respuesta del cliente (si o no) para suscribir a un depósito de término fijo.



Se puede ver que la duración de la llamada y el éxito (o no) de la campaña de marketing anterior tienen mucho impacto en el resultado.

Al incluir una precisión del 100% el resultado es fiable en un 90%, este resultado parece bueno, pero no significa que el clasificador es bueno : por ejemplo si se tiene 90% de los clientes que dicen "no", un clasificador no confiable se encuentra clasificando todos los clientes como clientes que van a decir "no" y aunque tiene 90% de precisión, es necesario tener en cuenta la matriz de confusión para verificar que el algoritmo funciona bien se revisa el resultado:

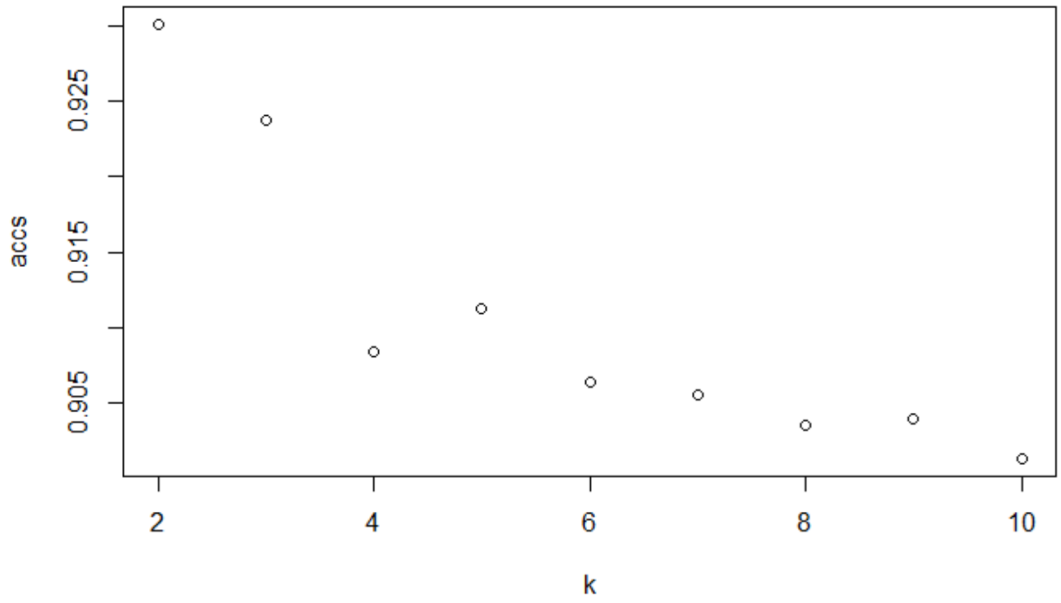
	Predicciones	
	No	Si
No	3896	104
Si	348	173

Se puede ver que el clasificador no es relevante para unas situaciones pero para otras sí, y clasifica bastante bien tanto los clientes que dan una respuesta positiva como los que dan una respuesta negativa.

KNN

Ahora se va a utilizar el algoritmo Knn (K-Nearest-Neighbors), para hacer la clasificación.

- Primero se convertirán las variables categóricas en variables numéricas porque con el algoritmo KNN hace falta calcular la distancia entre los datos y para realizarlo se necesitan variables numéricas.
- Segundo se va a aplicar el algoritmo, pero para eso hace falta elegir con qué valor de "k" (número de vecinos más cercanos) se va a aplicar el algoritmo.
- Tercero se exploran distintos valores ya que de antemano no conocemos cuál es el mejor valor para k, se aplican en el algoritmo con varios valores de k y de esta manera elegir el valor que da los mejores resultados. Dentro del código puede observarse que la ejecución tarda varios minutos debido al entrenamiento de varios clasificadores para el hallazgo del correcto.



En el gráfico se observa que hay resultados óptimos (precisión superior al 90%) con todos los valores de k, y hay una precisión de más de 92% con k=2 y k=3. Se va a calcular con más detalle la precisión que se obtiene con el mejor k.

Conclusión:

Teniendo en cuenta que el rendimiento de los clasificadores y observando que el rendimiento del clasificador KNN es tremendo bache más alto que el del árbol de decisión concluyendo lo siguiente:

- El árbol de decisión no es concluyente en su totalidad permite revisar la posible respuesta del impacto de la campaña realizada pero de acuerdo al objetivo no es simplemente la revisión de un sí o un no y es estático para próximas revisiones con un 90% de fiabilidad en los resultados planteados.
- El clasificador KNN por medio de su variable k, permite la variabilidad y realizar un ajuste en el modelo que busca el ajuste de datos permitiendo alcanzar entre un 92% - 93% de fiabilidad en los resultados planteados, y con el plus de que podría utilizarse para predicciones futuras.

De acuerdo a las conclusiones encontradas en este ejercicio el ajuste del set de datos objeto de este informe KNN se ajusta correctamente y su reuso puede tenerse en cuenta en caso de búsqueda de información adicional del set.

BIGML - Association Discovery

Qué es BIGML y para qué sirve.

BIGML es una compañía fundada, en 2011, en Oregon (Estados Unidos) por Francisco J. Martín, licenciado por la Universitat Politècnica de València (UPV), y el profesor Tom Dietterich, pionero en definir el machine learning -aprendizaje automático- como disciplina académica, con el objetivo de "democratizar" y hacer esta parte de la inteligencia artificial "accesible para todo el mundo" (Fontanillo, 2018).

BigML ofrece machine learning como servicio, lo cual ha reducido las barreras de entrada en cuanto a dificultad técnica y nivel de formación necesario para su manejo.

Association Discovery

Es una técnica de aprendizaje no supervisado que busca correlaciones interesantes o patrones en grandes volúmenes de datos. Es ampliamente usada en el mundo del comercio para descubrir patrones de compras y poder ofrecer a los clientes productos adecuados de una manera más efectiva.

Partes de una regla de asociación

Las reglas de asociación son declaraciones del tipo “if-then”, que ayudan a mostrar la probabilidad de las relaciones entre los elementos de datos, por lo cual se compone de 2 partes:

1) Un antecedente (si).
2) Un consecuente (entonces).

Métricas

Las métricas más significativas son el soporte y la confianza. Para ilustrar de una manera más fácil de entender se hará uso del siguiente ejemplo sobre ventas en un supermercado extraído de “Reglas de asociación” (2020).

ID	Leche	Pan	Mantequilla	Cerveza
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

La tabla representa una pequeña base de datos D que contiene los ítems, donde el código '1' se interpreta como que el producto (ítem) correspondiente está presente en la transacción y el código '0' significa que dicho producto no está presente.

Un ejemplo de regla para el supermercado podría ser: {Leche, Pan} => {Mantequilla}
Significa que si el cliente compró 'leche' y 'pan' también compró 'mantequilla'.

De lo anterior podemos definir las siguientes variables para facilitar la demostración de los cálculos.

X = {Leche, Pan}
Y = {Mantequilla}

- **Soporte**

El soporte de un conjunto de ítems X en una base de datos D se define como la proporción de transacciones en la base de datos que contiene dicho conjunto de ítems. Como la combinación de ítems X aparece 2 veces en un total de 5 transacciones que tiene la base de datos D, el soporte de X sería:

$$sop(X) = \frac{|X|}{|D|}$$

$$sop(\{Leche,\ Pan\}) = \frac{2}{5} = 0.4$$

Es decir, el soporte es del 40% (2 de cada 5 transacciones).

- **Confianza**

La confianza puede interpretarse como la probabilidad de encontrar la parte derecha de una regla condicionada a que se encuentre también la parte izquierda. Se calcula dividiendo el soporte de las transacciones donde ocurre tanto antecedente como consecuente entre el soporte de las transacciones donde ocurre el antecedente.

Se debe calcular el soporte de $sop(\{Leche,\ Pan\} \cup \{Mantequilla\})$, ya que de las 5 transacciones que hay en la base de datos, solo 1 contiene tanto el antecedente como el consecuente, se puede calcular:

$$sop(X \cup Y) = \frac{|X \cup Y|}{|D|}$$

$$sop(\{Leche,\ Pan\} \cup \{Mantequilla\}) = \frac{1}{5} = 0.2$$

Finalmente, basado en los valores obtenidos anteriormente, se calcula la confianza:

$$conf(X => Y) = \frac{sop(X \cup Y)}{sop(X)}$$

$$conf(\{Leche,\ Pan\} => \{Mantequilla\}) = \frac{sop(\{Leche, Pan\} \cup \{Mantequilla\})}{sop(\{Leche, Pan\})} = \frac{0.2}{0.4} = 0.5$$

Este cálculo significa que el 50% de las reglas de la base de datos que contienen 'leche' y 'pan' en el antecedente también tienen 'mantequilla' en el

consecuente; en otras palabras, que la regla $\{Leche,\ Pan\} \Rightarrow \{Mantequilla\}$ es cierta en el 50% de los casos.

- **Lift**

Mide la proporción entre el soporte observado y el soporte esperado si X e Y fueran independientes y a partir de esa proporción permite sacar algunas conclusiones.

Para el ejemplo la fórmula de lift es la siguiente:

$$lift(X \Rightarrow Y) = \frac{sop(X \cup Y)}{sop(X)\ x\ sop(Y)}$$

$$lift(\{Leche,\ Pan\} \Rightarrow \{Mantequilla\}) = \frac{sop(\{Leche, Pan\} \cup \{Mantequilla\})}{sop(\{Leche, Pan\})\ x\ sop(Mantequilla)}$$

$$lift(\{Leche,\ Pan\} \Rightarrow \{Mantequilla\}) = \frac{0.2}{0.4\ x\ 0.4} = 1.25$$

Como bien especifican en “Association rule learning” (2021), si la regla tuviera un lift de 1, implicaría que la probabilidad de ocurrencia del antecedente y la del consecuente son independientes entre sí. Cuando dos eventos son independientes entre sí, no se puede establecer una regla que involucre a esos dos eventos.

Si el lift es > 1, eso nos permite saber el grado en que esas dos ocurrencias son dependientes entre sí, y hace que esas reglas sean potencialmente útiles para predecir el consecuente en conjuntos de datos futuros.

Si el lift es < 1, eso nos permite saber que los elementos se sustituyen entre sí. Esto significa que la presencia de un elemento tiene un efecto negativo sobre la presencia de otro elemento y viceversa.

Casos de uso de las reglas de asociación

- **Medicina**

Dado que muchas enfermedades comparten síntomas, esta técnica de minería de datos puede ser usada para encontrar correlaciones entre una serie de síntomas y una enfermedad.

- **Comercio**

Este es el uso más habitual, se utiliza la información de las compras para encontrar patrones de asociación o grupos de artículos que son comprados juntos con frecuencia y así ajustar la estrategia de mercadeo y ventas.

- **Experiencia de Usuario en Software (UX)**

Los desarrolladores pueden analizar los comportamientos de los usuarios al interactuar con los sitios web o aplicaciones, y una vez se han detectado tendencias se pueden diseñar los sitios para lograr engagement o participación mediante un llamado a la acción dirigido.

- **Entretenimiento**

Servicios de video o audio bajo demanda suelen usar reglas de asociación para descubrir patrones de consumo y fortalecer sus motores de recomendación o bien para organizar el contenido de manera que se muestre opciones interesantes según un usuario dado.

Uso de Association Discovery en BIGML

Función 1 click Association Discovery en BIGML

Identifica co relaciones entre los datos y esto permite revisar comportamientos con un objetivo puntual y marcado, esto se realiza a través de un “sí” y un “entonces” es decir, sí sucede algo entonces se toma esta determinación, el comportamiento lo podemos observar de la siguiente manera:

fecha	cliente	cuenta	canal	tipo	Código postal	Monto de compra
10/06/2020	Rob	089	online	a	41006	2.500
03/06/2020	Ash	070	debit	a	42007	1.200
04/08/2020	Eric	064	debit	b	41221	450
10/08/2020	Rob	089	online	b	41006	700
11/01/2020	Rob	089	online	a	47023	1.350
12/25/2020	Ash	070	debit	a	42007	500
12/28/2020	Ash	070	online	b	42007	720

Teniendo en cuenta lo anterior, al analizar los registros del cliente Rob entendemos que por la realización de sus transacciones es 41006, esto puede determinar una asociación de donde se realizan las transacciones de este cliente y a partir de dichas asociaciones intuir que la tercera transacción se trata de un fraude por la ***asociación incorrecta*** del código postal.

Las reglas de asociación se utilizan para encontrar correlaciones y co-ocurrencias entre conjuntos de datos. Se utilizan idealmente para la explicación de patrones en datos de repositorios de información aparentemente independientes, como bases de datos

relacionales y bases de datos transaccionales. El acto de usar reglas de asociación a veces se denomina "minería de reglas de asociación" o "asociaciones de minería".

¿Cómo funciona dentro de BIGML?

Con configuración de parámetros avanzados teniendo en cuenta los siguientes ítems:

a. Configuraciones básicas (Basic Configurations)

1. Número máximo de asociaciones
   Este puede darse entre 1 y 500, y viene detallada por las posibles relaciones exponenciales, a mayor cantidad de campos mayores posibilidades de crecer las relaciones y mayor cantidad de asociaciones tardará más tiempo en calcular.
2. Max items in antecedent
   Debido a que hay un sí y luego un entonces, se podría considerar que el número máximo de antecedentes este se puede establecer valores entre 1 y 10 moviendo la cantidad de elementos al máximo. Un mayor número de ítems producirá naturalmente reglas de asociación más complejas. Sin embargo, el conjunto de elementos consecuente siempre contendrá un elemento.
3. Search Strategy
   A raíz de los descubrimientos realizados permite priorizar en las asociaciones objetivo, entonces los valores por encima de los hallazgos encontrados serán analizados usando herramientas para análisis de datos predeterminados. Para Search Strategy el apalancamiento es vital, ya que es una de las medidas que da resultados relevantes en la mayoría de los casos. Al utilizar apalancamiento, se encontrarán asociaciones de elementos que ocurren con más frecuencia en su conjunto de datos. Del lado contrario se encuentra la elevación (Lift) de los datos de está manera se encontrarán asociaciones de elementos menos frecuentes en su conjunto de datos, pero fuertemente relacionados entre sí. La estrategia

que elija debe ser coherente con el objetivo final.

4. Search for complementary items

   Complementary ítems genera una selección asociada contraria como se muestra a continuación:

   - Sencillas: para el ítem (café), el complemento sería (NO café).
   - Compuesta: (leche, café) → (azúcar), reglas complementarias como (leche, NO café) → (chocolate)

5. Search for missing items

   De acuerdo al objetivo buscado se puede establecer que los ítems perdidos sean tenidos en cuenta como una asociación de datos válidos estableciendo una regla previa para su aplicación.

b. Tipos de datos (Data types)

1. Numeric: datos establecidos como numéricos dentro del dataset
2. Categorical: datos con un significado puntual y específico (yes, no, rojo, verde)
3. Date time: variables de tiempo de todo tipo y pueden encontrarse en multiformato y con diferentes medidas (fecha corta, fecha larga, días de la semana, número de días)
4. Text: parámetro natural de texto
5. Ítems: son como text, sin embargo son una lista estructurada con una característica específica.

c. Configuraciones avanzadas (Advanced Configurations)

1. Measure: permite adecuar parámetros para cada regla establecida como: support o leverage.
2. Minimun significance: establecer correctamente los mínimos permite disminuir los datos basuras o datos desperdicios; a mayor valor, menos restrictivo es en busca de un objetivo puntual dentro del dataset.
3. Consequent: busca uno o más valores específicos de acuerdo a un determinado criterio.

4. Discretization: son la cantidad de valores soportados, desde tipos numéricos, tamaño, y tipo de valor.
5. Sampling: mediante un generador de números aleatorios, que significa que dos muestras del mismo conjunto de datos probablemente serán diferentes incluso cuando se usen las mismas tasas y rangos de filas. Se elige entre un muestreo aleatorio o un muestreo determinista. Si se escoge muestreo determinista, el generador de números aleatorios siempre utilizará la misma semilla, produciendo así resultados repetibles. Esto le permite trabajar con muestras idénticas del mismo conjunto de datos.

Bibliografía

Fontanillo, Olivia. (07/10/2018). BigML acerca el 'machine learning' a la empresa y la universidad. El Economista. Recuperado de: https://www.eleconomista.es/valenciana/noticias/9436203/10/18/BigML-acerca-el-machine-learning-a-la-empresa-y-la-universidad.html

Martinez, J. (2017, November 12). *Bank Marketing Dataset* . Predicting Term Deposit Suscriptions. https://www.kaggle.com/janiobachmann/bank-marketing-dataset

Reglas de asociación. (22/05/2020). *Wikipedia, La enciclopedia libre*. Fecha de consulta: 08:50, enero 3, 2022. Recuperado de: https://es.wikipedia.org/w/index.php?title=Reglas_de_asociaci%C3%B3n&oldid=126273483

Association rule learning. (26/12/2021). *Wikipedia, La enciclopedia libre*. Fecha de consulta: 20:38, enero 3, 2022. Recuperado de: https://en.wikipedia.org/w/index.php?title=Association_rule_learning&oldid=1062056903

Lutkevich, Ben. (September 2020). Association Rules. Search Business Analytics. Tech Target. Recuperado de: https://searchbusinessanalytics.techtarget.com/definition/association-rules-in-data-mining

Rodríguez González, Ansel & Martínez-Trinidad, José Francisco & Carrasco-Ochoa, Jesús & Ruiz-Shulcloper, José. (2009). Minería de Reglas de Asociación sobre Datos Mezclados. Recuperado de: (PDF) Minería de Reglas de Asociación sobre Datos Mezclados