Sistema de Recomendación Crediticia

Juan Camilo Campos
Daniela Martínez Duarte
Javier Alejandro Velasco

INTRODUCCIÓN

Actualmente uno de los proyectos del Grupo Bancolombia es brindar funcionalidades nuevas a sus usuarios dentro del marco de los *Personal Financial Managers* (PFM). Para ello se dispone de un set de datos sobre las transacciones por PSE de distintos usuarios, realizadas en una ventana de tiempo de aproximadamente 2 años. Con el fin de proveer una solución para esta problemática desde la analítica de datos, las siguientes fases del modelo de proceso CRISP-DM fueron completadas (ver Fig. 0.1): Comprensión del negocio (ver sección 1), Comprensión de datos (ver sección 2), Preparación de datos (ver sección 3) y Modelado (ver sección 4). De acuerdo con esto, en las siguientes secciones se presentan las actividades y análisis realizados en cada una de estas fases. Teniendo en cuenta dichos análisis, se presenta nuestra propuesta de valor titulada "Sistema de Recomendación Crediticia", la cual busca asesorar a los clientes en el manejo de sus deudas, así como proveer al banco de clientes objetivos para productos crediticios.

Figura 0.1: Fases de la metodología CRISP

1. Comprensión del negocio

Los datos entregados en este reto corresponden a transacciones realizadas por clientes persona del banco vía [PSE]. Estas transacciones, a diferencia de aquellas realizadas vía[POS], no cuentan con un código [MCC] atado a la transacción, que permite conocer la categoría de comercio a la que pertenece el establecimiento de comercio donde se realiza la transacción. Adicionalmente, muchas de estas transferencias por PSE corresponden a transferencias de pagos de servicios públicos, seguros, colegios, arrendamientos, y otros gastos que pueden ser denominados como «gastos grandes». En el marco de un sistema de gestión de finanzas personales, poder categorizar adecuadamente las transacciones que se realizan por PSE es de suma importancia para contar con un panorama completo de la actividad de gastos de los clientes. Por esto, nuestro propósito será categorizar la mayor cantidad de transacciones posibles que se realizan por PSE. Además, aprovecharemos la categorización de transacciones para crear un modelo que logre predecir, mes a mes, los gastos por deudas que tendrán los usuarios y junto con sus ingresos poder evaluar la "salud crediticia" de cada cliente.

Para un cliente natural, es de vital importancia tener siempre presente su capacidad de endeudamiento para lograr administrar adecuadamente sus gastos. Si bien, en las aplicaciones de PFM las personas pueden recibir alertas de acuerdo con las metas de ahorro fijadas y ver información de sus gastos, no cuentan con un mecanismo automatizado que pueda predecir el gasto de pago de deudas y hacer las recomendaciones adecuadas. Por su parte, el banco puede usar esta información para enfocar sus campañas de productos crediticios a aquellos que se prediga no van a utilizar toda su capacidad financiera en los siguientes meses.

2. Comprensión de los datos

Como parte del set de datos para el análisis, se dispuso de dos tablas. La primera tabla es una tabla de clientes que cuenta la descripción de 300 mil clientes (persona). Los atributos de esta tabla son: id cliente, segmento, ocupación, tipo vivienda, nivel académico, estado civil, genero, edad, e ingreso rango. La segunda tabla es de transacciones la cual cuenta con 14 millones de registros (uno para cada transacción), realizados entre septiembre de 2016 y octubre de 2018. Los atributos de esta tabla son: id transaccion, id cliente, fecha, hora, valor trx, ref1, ref 2, ref3, sector, subsector, descripcion y descripcion2. Las siguientes tablas 2.1 y 2.2 muestran el porcentaje de valores ausentes en cada atributo de cada tabla.

Con las tablas descritas, se tiene una idea global de los datos disponibles que se tienen en el reto y del porcentaje de datos nulos en cada columna. Por ende, con esta información se puede proceder a preparar los datos y posteriormente, a analizarlos detalladamente.

Columna	Porcentaje NaN
id cliente	0.0
segmento	0.0
ocupacion	2.01
tipo vivienda	50.86
nivel academico	13.08
estado civil	1.94
genero	1.68
edad	0
ingreso rango	2.18

Cuadro 2.1: Porcentaje campos vacios por atributo (Tabla Clientes)

Columna	Porcentaje NaN
id trn ach	0.03
id cliente	0.5
fecha	10.7
hora	0.5
valor trx	0.5
ref1	3.0
ref2	42.7
ref3	100.0
sector	71.917082
subsector	71.917082
descripcion	71.917082
descripcion2	99.812615

Cuadro 2.2: Porcentaje campos vacios por atributo (Tabla Transacciones)

3. PREPARACIÓN DE LOS DATOS

3.1. TABLA CLIENTES

La fase de preparación de los datos inicia con la limpieza de las tablas. Para la tabla de clientes se decidió eliminar la columna “tipo de vivienda” debido a que su porcentaje de valores faltantes es mayor al 50 %. Seguido esto se procedió a revisar la coherencia de los datos, por lo cual, se eliminaron los clientes con edades fuera del rango de 10 años a 100 años, al considerarse estas edades como ruido.

Con base en la nueva tabla de clientes, se realizó un estudio descriptivo de la tabla de clientes (ver notebook Data Clients Cleaning). Dentro de los hallazgos más importante en este análisis descriptivo es que la mayoría de los clientes de la tabla son personas solteras, con estudios universitarios, empleadas y con edades entre los 24 y los 37 años. Además, el conjunto de

clientes se encuentra muy parejo en sexo (51 % mujeres y 49 % hombres). Las siguientes figuras muestran la distribución del nivel académico de los clientes (Fig. 3.1) junto con la distribución de su rango de ingreso (Fig. 3.2).

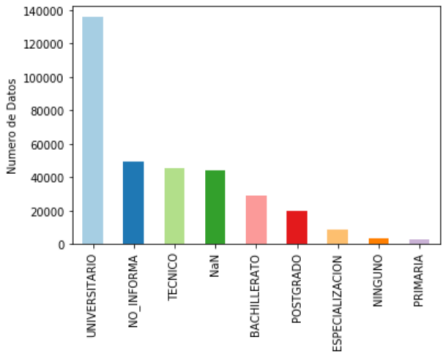

Figura 3.1: Histograma nivel de estudio

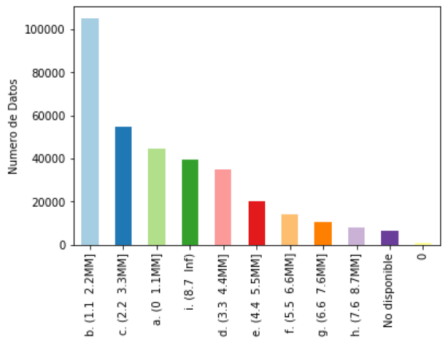

Figura 3.2: Histograma rango de ingresos

3.2. Tabla Transacciones

Al igual que con la tabla de clientes, lo primero que se realizó en esta etapa con la tabla de transacciones fue el manejo de los datos faltantes (ver notebook Data Transactions Cleaning). De acuerdo con la tabla 2.2, la columna **ref3** no contiene información por lo cual se procede a eliminarla. También se decide eliminar todas las transacciones que no contengan información en los campos **id transaccion**, **id cliente**, **fecha**, o **valor trx**, al considerarse esta información indispensable en la definición de una transacción. Además, se convirtió el campo fecha a formato datetime, y para el campo **hora** que contiene hora minutos y segundos, se decidió solo conservar la hora.

Los campos **ref1** y **ref2** son diligenciados por el recaudador, por lo cual son campos que contienen mucho ruido. Lo primero que se hace es considerar las referencias que esten

definidas como “\\n”, “\\N” o “nan” como campos vacios. Seguido esto, a los textos de ambos campos se les aplica el siguiente procedimiento de limpieza:

1. Se convierte a minúsculas.
2. Se eliminan tildes y caracteres extraños (fuera del código ASCII).
3. Se eliminan las palabras: ti, cc, ce, idc, ti, nit, pp, pc, tpni, y null.
4. Se eliminan espacios al inicio y al final del texto.
5. Se eliminan stop words.

Como nuestro propósito es clasificar las transacciones, se decide eliminar aquellas transacciones que después de este proceso de limpieza no contengan información ni en **ref1**, ni en **ref2**, ni en el campo **descripción**. Finalmente, se unen los campos **ref1** y **ref2** en un solo campo llamado **ref**. Se termina de limpiar el dataset eliminando transacciones con valores atípicos. Por esto toda transacción con valores menores a 1600 se elimina, junto con las transacciones que se consideren outliers (mayores a el tercer cuartil más 1.5 veces el rango intercuartil de los datos). De este proceso de limpieza se termina con 8’677.3260 transacciones, aproximadamente el 61 % de los datos iniciales.

3.3. Limpieza de palabras campo ref

Con el objetivo de terminar de limpiar el campo **ref** se eliminan las palabras que aparecen menos de 10 veces dentro de este campo en todo el dataset (ver notebook Data TransactionsCleaning Words). Esto se hace con el propósito de eliminar cadenas de texto sin sentido, nombres propios, direcciones e-mail, etc. Además, palabras como pago y factura también se eliminan porque tienen alta frecuencia y no serán útiles para la clasificación. De hecho, estas palabras pueden generar un sesgo en el modelo.

Una vez eliminadas las palabras deseadas se eliminan las transacciones que terminaron con el campo ref vacío al realizar este proceso. En este momento el dataset de transacciones queda con un total de 8’300 238, y es este dataset el que buscara categorizar cada una de sus transacciones.

4. Modelado

En esta parte de la metología CRISP, se exponen en primera instancia los métodos utilizados para clasificar los datos por categorías. Posteriormente, tomando como base los datos clasificados se exponen los modelos de predicción usados para estimar el pago de deudas de los clientes. Cabe resaltar, que se evaluan distintas técninas en ambos casos, luego se verifica su desempeño y finalmente, se determina cual se ajusta más a las necesidades del negocio. Para explicar cómo fue el proceso del modelado, se divide esta sección en cuatro partes denominadas: Clasificación de los datos, modelo de predicción de categorías, detección de comunidades de clientes y modelo de predicción de gasto en pago de deudas.

4.1. CLASIFICACIÓN DE LOS DATOS

Para la clasificación de los datos primero nos concentramos en las transacciones con datos no nulos en los campos sector, subsector y descripción que corresponde al 33 %. Estas transacciones llegan a usuarios de Bancolombia, entonces nuestra idea es aprovechar la descripción que ha realizado Bancolombia sobre estas transacciones, para asignarles una categoría. Hecho esto, aprovechamos estos datos para crear un modelo que a partir del campo **ref** prediga la categoría de cada transacción. La Fig. 4.1 ilustra el procedimiento desarrollado (ver notebook Data Transactions Description).

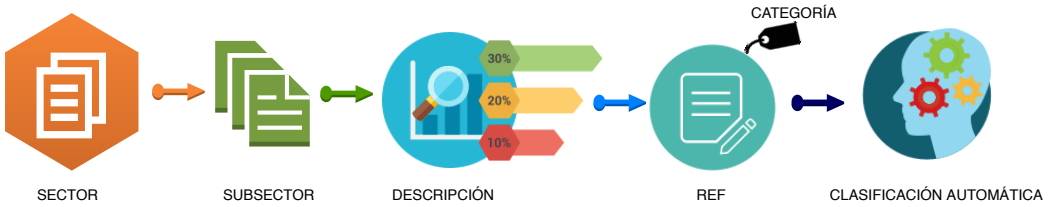

Figura 4.1: Proceso para la clasificación de los datos

Categoría	**Datos asociados**
Gobierno	126
Entretenimiento	112
Subscripción	52
Educación	46
Transporte	26
TyC	26
Pago de Deudas	25
Otros	10
Hogar	16
Seguros	11
Viajes	2
Salud	2
Moda	1
Comida	1

Cuadro 4.1: Número de referencias distintas clasificadas en cada categoría

Para definir las categorías a utilizar para la clasificación, se tomó como base aquellas propuestas por el equipo de Bancolombia de las transacciones POS, y se agregaron unas propias. En la tabla 4.1 se describen las categorías usadas. Definidas las categorías, se procede a analizar en detalle los subsectores que han sido asignados por Bancolombia por medio del histograma que se muestra en la figura 4.2. Se puede apreciar que aquellos subsectores en los que se concentra la mayor parte de los datos corresponden a bancos, telefonía fija, valor agregado, electricidad y administración central. Por otra parte, algunos subsectores secundarios, que

contienen información que se considera también relevantes son municipios, cajas de compensación, establecimientos educativos, obras de infraestructura, eps y salud prepagada y servicios a personas.

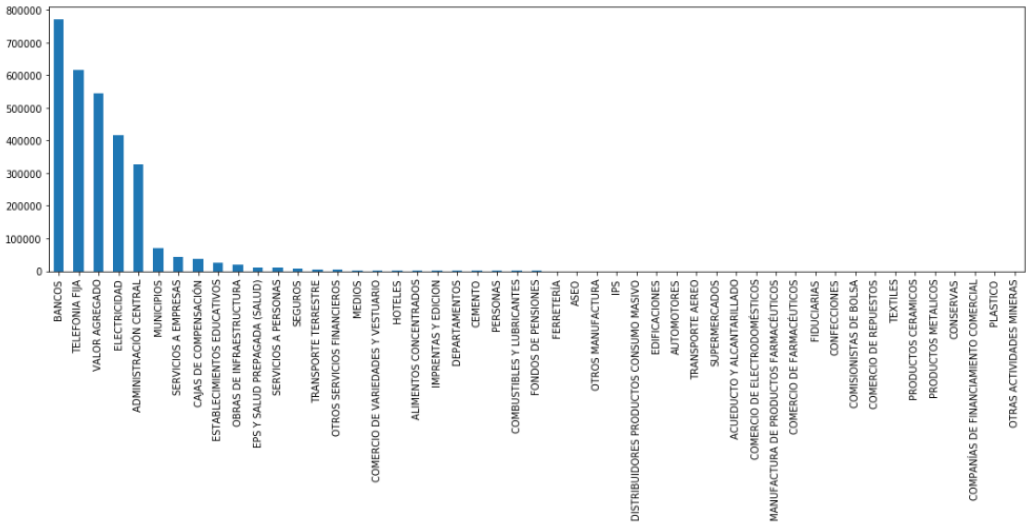

Figura 4.2: Histograma Subsectores

Considerando este resultado, se decide revisar el número de descripciones distintas que se tienen de los subsectores mencionados anteriormente, para determinar cuales categorías usar y como hacer la clasificación de las transacciones de cada subsector. Se encuentra que los subsectores que agrupan la mayor parte de la información estudiada cuentan con una única descripción. Esto se puede entender por que Bancolombia cuenta con información de los clientes a los que se les realizó la transacción, pero no especificamente sobre cual fue el objetivo final de la misma. Por ende, se decide explorar las referencias únicas en el campo creado a partir de **ref1** y **ref2** denominado **ref** (ver notebook Transactions+Categorization).

Para cada subsector se filtran aquellas referencias que son únicas y capturan el 90 % de las transacciones del subsector y se les asigna una categoría basados en la descripción. Para algunos subsectores “ADMINISTRACIÓN CENTRAL” es sencillo hacer la clasificación, porque, aunque contiene 136 referencias distintas, todas se deben de clasificar en la categoría Gobierno, sin revisarlas en detalle. Sin embargo, algunos subsectores como “BANCOS” contienen un gran número referencias distintas, que corresponden a categorías distintas. Por esto las referencias que capturan el 90 % de las transacciones de “BANCOS” de deben clasificar de forma manual una a una.

Junto con la clasificación de las referencias correspondientes a las transacciones con descripción, se clasifican manualmente las referencias más utilizadas en el dataset que no contiene descripción. La tabla 4.1 muestra el número de referencias distintas clasificadas en cada cate-

goría.

Debido a que se hizó una exhaustiva limpieza de los datos, al unir las categorias con todo el set de datos aproximadamente 5,500,000 quedan clasificados de acuerdo a su referencia. Ahora bien, aproximadamente 2,700,000 quedan sin clasificación y por ende, se busca un modelo de predicción para clasificar texto que sea capaz de clasificar cada transacción apartir del campo referencia (ver notebook Clasify+all+data).

4.2. Modelo de predicción de categorías

Para realizar el modelo de predicción de las categorías se usarón las referencias ya categorizadas. Lo primero es definir la forma de representar el texto que se encuentra en el campo ref. Por esto decidimos usar una técnica conocida como Tf-idf vectorizer, en donde cada texto se representa mediante un vector que indica cuantas veces se encuentra cada palabra en el texto. Ademas, para aquellas palabras comunes, se reduce su peso usando frecuencia inversa. Con los textos vectorizados, se procedió a entrenar un clasificador Bayesiano ingenuo (Naive Bayes) y una máquina de soporte vectorial para realizar la prediccion. Cabe resaltar que antes de realizar la vectorizacion de los texto se realizo un "stemming", para llevar las palabras a su raiz y buscar amplicar la capacidad de generalización de los modelos predictivos.

4.2.1. Clasificador Bayesiano Ingenuo (Naive Bayes)

Es un método de aprendizaje supervisado que usa medidas de probabilidad para hacer la clasificación. Este método asume que cada atributo siendo clasificado es independiente de los demás, por ende cada atributo contribuye independientemente a la probabilidad de que un elemento sea clasificado en determinado grupo. Para el caso del texto, lo que hace es que de acuerdo al número de veces que se repite una palabra determina la probabilidad que un elemento pertenezca a una categoría, asumiendo que las palabras no guardan una relación.

4.2.2. Máquinas de soporte vectorial

Es un método de aprendizaje supervisado que consiste en crear un espacio de atributos dimensional, donde el atributo se refiere a la importancia de una palabra en particular. Los datos son mapeados posteriormente a uno de estos espacios asociados a una categoría.

En este caso, ambos métodos se implementaron mediante librerías de sklearn (ver notebook Clasify+all+data). Para evitar que los clasificadores se vieran sesgados hacia las categorías con mayor número de referencias clasificadas, fue necesario balancear los datos. El balance se realizó, repitiendo datos de las categorías con menor numero de apariciones. Posteriormente, se verificaron los resultados usando cross-validation, y mediendo el valor-F1 que indica el desempeño de la predicción. En la tabla 4.2 se puede observar que para ambos métodos se obtiene que el valor-F1 es superior a 0.9 en todos los casos.

Ambos clasificadores se usaron para asignarle una categoría a cada una de las 2,700,000 de transacciones que no se encuentran categorizadas. En ambos casos se evidenció que

NB	SVM
0.96102989	0.96430667
0.97048853	0.97399221
0.99349305	1.
0.96060039	0.96703653
0.96774115	0.97739123

Cuadro 4.2: Valor-F1 de la validación cruzada

para un gran número de transacciones la clasificación no era la deseada. Por esto, se decidió analizar la probabilidad de predicción (probabilidad de que la transacción pertenezca a la categoría asignada) de los algoritmos en cada una de las transacciones a predecir. La Fig 4.3 muestra el comportamiento de esta probabilidad de predicción para la máquina de soporte vectorial. Note que al ser 15 categorias, si la probabilidad de predicción es cercana a 0,066(1/15), dicha predicción no es confiable. Es decir, valores cercanos a 0,066 indican que todas las categorías tienen probalidades cercanas de ser la asignada. Por esta razón se decide que para las transacciones con probabilidad de predicción menor a 0,1 se asumirá que el clasificador no genera una clasificación confiable, es decir, no es capaz de categorizarlas. En la Fig 4.3, se puede observar que aproximadamente para 1'200.000 transacciones la clasificación de la máquina de soporte vectorial no es confiable. Este número aumenta cuando se prueba el clasificador Bayesiano ingenuo, por lo cual se decide utilizar la clasificación dada por la máquina de soporte vectorial.

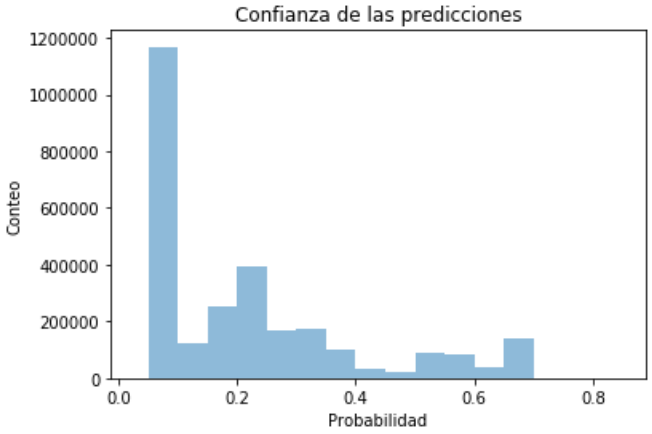

Figura 4.3: Histograma de confianza de las predicciones

Finalmente de las 8'300.238 transacciones que estan disponibles para categorizar nuestro procedimiento es capaz de categorizar 6'983.666 de transacciones, lo que corresponde a un 82 %.

4.3. Detección de comunidades de clientes

Partiendo de la categorización de las transacciones, utilizando la clasificación manual y el modelo de predicción de categorías, se buscó identificar grupos de clientes con base en diferentes parámetros. Para esto, se abordaron dos enfoques principales: el primero se concentra en la agrupación de clientes a partir de sus gastos promedio por categoría de transacción; el segundo enfoque busca agrupar los clientes a partir de sus características demográficas (e.g. edad, género o rango de ingresos).

4.3.1. Agrupación por categorías de gastos

Al abordar la agrupación de clientes según sus gastos promedio por categoría de transacción, se procesaron los datos de transacciones categorizadas de forma que se obtuvieron todas las transacciones por categoría para cada cliente. Con esto, se calcularon los valores promedio de los gastos de cada cliente en cada una de las categorías y se creó una matriz que relacionara estos gastos promedio con el cliente correspondiente, como se ve en la Fig. 4.4. Se realizó un análisis descriptivo de esta matriz para así conocer también el comportamiento general de las categorías, respecto a los gastos promedio por cliente, de aquí se obtuvo que, en promedio, la categoría en la que más gastan los clientes es en *pago de deudas*.

clasificacion	Comida	Educacion	Entretenimiento	Gobierno	Hogar	Indefinido	Moda	Otros	Pago de Deudas	Salud
id_cliente										
1	0.0	0.000000	0.000000	0.000000	82570.870000	0.000000	0.0	0.000000	0.000000	0.000000
10	0.0	0.000000	0.000000	201806.926111	0.000000	22519.390000	0.0	191968.613333	0.000000	0.000000
100	0.0	0.000000	0.000000	92687.010000	0.000000	224878.255000	0.0	0.000000	355267.984286	128132.740000
1000	0.0	0.000000	0.000000	0.000000	200216.191154	236231.861818	0.0	200944.858667	0.000000	0.000000
10000	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000
100000	0.0	0.000000	0.000000	41008.300000	0.000000	136339.406250	0.0	0.000000	0.000000	0.000000
100001	0.0	0.000000	0.000000	193318.417500	0.000000	98392.548000	0.0	154147.210000	107520.234286	96633.290000
100002	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000
100003	0.0	0.000000	0.000000	16304.286923	0.000000	129564.902857	0.0	0.000000	382452.230000	0.000000
100004	0.0	0.000000	0.000000	15011.425714	0.000000	0.000000	0.0	0.000000	0.000000	0.000000
100005	0.0	0.000000	0.000000	14803.800000	0.000000	14889.460000	0.0	0.000000	0.000000	0.000000
100006	0.0	0.000000	0.000000	309830.910000	466750.242500	0.000000	0.0	0.000000	160172.700000	0.000000
100007	0.0	0.000000	0.000000	0.000000	0.000000	173705.950000	0.0	0.000000	0.000000	0.000000
100008	0.0	0.000000	0.000000	152890.290000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000
100009	0.0	410296.530000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000
10001	0.0	0.000000	0.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.000000	0.000000
100010	0.0	118496.150000	4747.200000	57481.933333	260978.690000	184716.895000	0.0	160291.420000	0.000000	0.000000

Figura 4.4: Matriz de gastos promedio de clientes vs. categorías

Una vez obtenida la matriz de gastos promedio por categoría para cada cliente, cada característica fue escalada a un rango de 0 a 1, con el fin de obtener un mejor desempeño de los modelos. El algoritmo utilizado para realizar la agrupación fue el de K-means. También se intentaron otras técnicas como *agglomerative clustering* y DBSCAN, sin embargo, los resultados no fueron exitosos ya que su ejecución toma cantidades de tiempo y de memoria excesiva de acuerdo al hardware disponible.

Este algoritmo de K-means se concentra en dividir las muestras en grupos de igual varianza, minimizando la suma de cuadrados entre el mismo grupo. Una de sus desventajas principales es que requiere la especificación previa del número de grupos en que se busca dividir los datos, sin embargo, posee la ventaja de escalar bien a un gran número de muestras con buenos resultados.

Para aplicar el algoritmo de k-means, se debió utilizar una técnica que permitiese encontrar el número óptimo de grupos en los cuales dividir los datos. Dicha técnica fue el método *elbow*, el cual consiste en llevar a cabo la agrupación para diferentes valores de k (número de grupos) y posteriormente graficar los mismos contra la métrica de calificación del algoritmo (*score*) para cada agrupación; en la gráfica resultante, se busca el valor de k a partir del cual la métrica deja de crecer significativamente, este se manifiesta como un ángulo en la curva plasmada, o codo (ver Fig. 4.5). Aquel valor de k se toma como el número de grupos óptimo para el conjunto de datos.

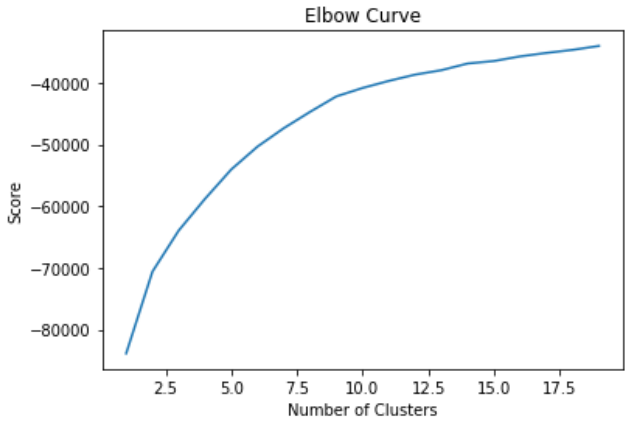

Figura 4.5: Gráfica del método *elbow* para la agrupación por tipos de gastos

Habiendo encontrado el valor óptimo para el número de grupos, de 9 en este caso, se procede a evaluar el modelo obtenido en el conjunto de datos, con base en esto se obtuvo la distribución de clientes en cada uno de los grupos, como se ve en la Fig. 4.6. En esta distibución se puede observar que existe una desigualdad marcada en la concentración de clientes de uno de los grupos respecto a los demás, contando este con *129284* clientes, mientras que las concentraciones de los otros grupos oscilan entre *12524* y *33286*.

Se obtuvieron las características principales de los grupos resultantes, con lo que se hicieron las siguientes observaciones importantes:

- El grupo con la mayor concentración de clientes es aquel en que el gasto promedio total es menor, con un valor de $313.107.
- La distribución de hombres y mujeres por grupo es aproximadamente equitativa.
- El grupo más grande es aquel con la edad promedio más baja.

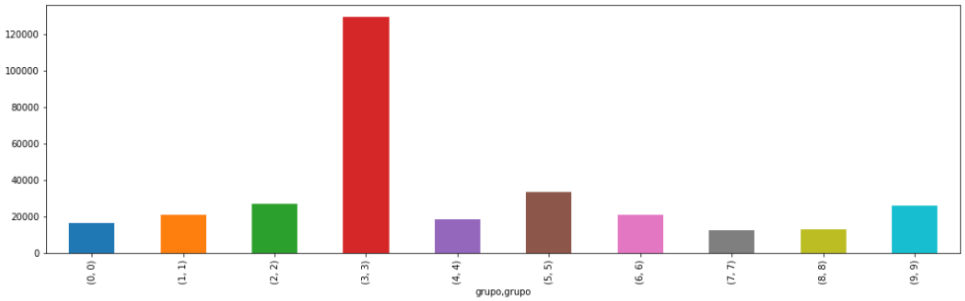

Figura 4.6: Distribución de clientes en grupos según tipos de gastos

- El grupo más pequeño es aquel con la edad promedio más alta.
- El grupo más grande es aquel con el ingreso promedio más bajo.
- La ocupación predominante en todos los grupos es *empleado*, seguida de *independiente*
- El grupo más grande posee la mayor concentración de estudiantes.
- En el grupo más grande, las categorías con mayor y menor gasto promedio son *tecnología y comunicaciones* y *pago de deudas*, respectivamente.

4.3.2. Agrupación por características demográficas

Habiendo llevado a cabo la agrupación de los clientes según sus gastos promedio por categoría de transacción, se pasó a tomar ahora como parámetros de entrada al algoritmo de agrupamiento, las características demográficas disponibles: ocupación, nivel académico, estado civil, género, edad y rango de ingresos. Para ello, se ejecutó un preprocesamiento de los datos, para tratar con valores nulos y codificar las características no numéricas, de manera que pudiesen ser alimentadas al algoritmo de agrupamiento. La codificación de características se hizo de forma que las diferentes etiquetas posibles en las categorías no cuantitativas son numeradas sucesivamente, como se muestra en la Fig. 4.7

	ocupacion	nivel_academico	estado_civil	genero	edad	ingreso_rango
307765	JUBILADO	UNIVERSITARIO	VIUDO	M	92.0	7.0
30636	JUBILADO	POSTGRADO	CASADO	M	87.0	9.0
107698	RENTISTA_DE_CAPITAL	UNIVERSITARIO	CASADO	M	89.0	1.5
15429	JUBILADO	BACHILLERATO	CASADO	M	91.0	2.5
46053	JUBILADO	UNIVERSITARIO	CASADO	M	90.0	6.0
184383	JUBILADO	UNIVERSITARIO	SOLTERO	M	88.0	5.0

	ocupacion	nivel_academico	estado_civil	genero	edad	ingreso_rango
307765	JUBILADO	UNIVERSITARIO	VIUDO	M	92.0	7.0
30636	JUBILADO	POSTGRADO	CASADO	M	87.0	9.0
107698	RENTISTA_DE_CAPITAL	UNIVERSITARIO	CASADO	M	89.0	1.5
15429	JUBILADO	BACHILLERATO	CASADO	M	91.0	2.5
46053	JUBILADO	UNIVERSITARIO	CASADO	M	90.0	6.0
184383	JUBILADO	UNIVERSITARIO	SOLTERO	M	88.0	5.0

Figura 4.7: Ejemplo de codificación de etiquetas

Ya con los datos codificados, se aplicó el algoritmo de agrupación *k-means*, en el que se obtuvo un número de grupos óptimo de 3, de acuerdo con el método *elbow*. Los grupos

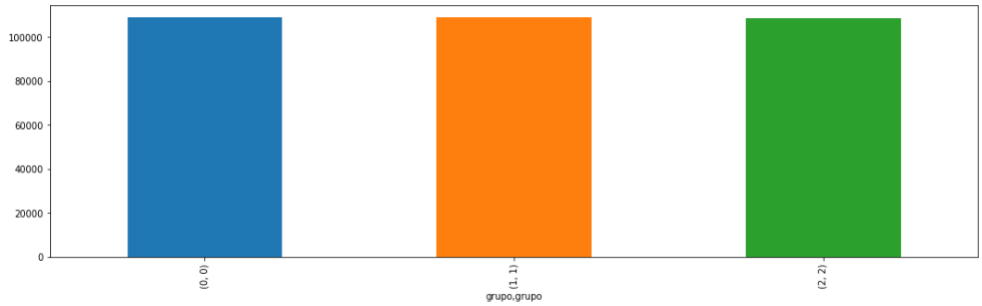

Figura 4.8: Distribución de clientes en grupos según características demográficas

resultantes cuentan con concentraciones de clientes aproximadamente iguales, como se ve en la Fig. 4.8.

Posteriormente se obtuvieron las características principales de cada uno de los grupos, con lo que se llegó a las siguientes conclusiones:

- El grupo cuyo gasto promedio total es menor, es aquel con una edad promedio menor.
- El grupo cuyo gasto promedio total es mayor, es aquel con una edad promedio más alta.
- Como es de esperarse, el grupo con menor gasto promedio total tiene el menor ingreso promedio, y aquel con gasto promedio total mayor, cuenta con el mayor ingreso promedio.
- La ocupación con mayor concentración en todos los grupos es *empleado*, seguida en 2 de los 3 grupos por *independiente*.
- En el grupo con gasto promedio menor se encuentra la mayor concentración de estudiantes, siendo la segunda ocupación con mayor concentración en el grupo.
- El grupo con mayor gasto promedio tiene una concentración muy baja de estudiantes, comparado con los otros grupos.

4.4. Modelo de predicción de gasto en pago de deudas

Dentro del marco de recomendación crediticia, se busca saber qué clientes son aptos para adquirir deudas, tanto con el objetivo de realizar recomendaciones a los mismos sobre el manejo de sus deudas, por medio del software de PFM, como para la identificación de estos como clientes potenciales para los productos crediticios del banco. Para lograr esto, se plantea aplicar técnicas de predicción de gastos en materia de pago de deudas a cada uno de los grupos identificados en la etapa de detección de comunidades de clientes, cuyos resultados, aplicados a cada cliente y en conjunto con los datos de ingresos del mismo, pueden ser utilizados para aproximar su capacidad de endeudamiento.

4.4.1. CONSTRUCCIÓN DEL CONJUNTO DE DATOS

En este orden de ideas, se tomaron las transacciones categorizadas y se llevó a cabo un filtrado de las mismas para obtener únicamente aquellas correspondientes a pago de deudas, agrupadas de acuerdo con las comunidades identificadas. En este caso se utilizaron, a modo de prueba, las comunidades resultantes del agrupamiento por características demográficas, esto debido a que se tiene una menor cantidad de grupos que los obtenidos a partir de los gastos promedio por categoría, lo que a su vez resulta en un número de clientes por grupo más alto. Estas dos características en conjunto permiten llevar a cabo pruebas con un número más bajo de modelos, teniendo un mejor ajuste de cada uno de ellos a su comunidad correspondiente.

Las transacciones fueron agrupadas de manera que se obtengan las transacciones mensuales de cada cliente, para así calcular el gasto total mensual del mismo en pago de deudas. Sobre este gasto total mensual por cliente, se aplicó un algoritmo de detección y eliminación de valores atípicos, con el fin de eliminar ruido resultante de conjuntos de transacciones fuera de lo común, sobre todo casos en los que se presentaban muchas transacciones por mes para el mismo cliente, lo que generaba valores totales de pago de deudas muy elevados. Una vez realizada la eliminación de datos atípicos, se aplicó un umbral de filtrado para descartar los clientes que contaran con un valor de pago de deudas diferente de 0 en un número de meses muy bajo, ya que se consideran como usuarios no regulares que afectan el ajuste del modelo al comportamiento real.

Así, se obtuvo la matriz de pago de deudas mensual para cada cliente del grupo, como se observa en la Fig. 4.9.

id_cliente	Sep/2016	Oct/2016	Nov/2016	Dec/2016	Jan/2017	Feb/2017	Mar/2017	Apr/2017	May/2017	Jun/2017	...	Jan/2018	Feb/2018	Mar/2018
521	534850.01	293938.91	352606.50	344401.96	306255.32	381990.25	267094.11	288079.27	437676.82	235730.27	...	112480.84	125631.77	117843.28
3141	667186.96	695661.32	678682.35	681938.44	704663.52	685749.94	713580.42	309272.14	299319.47	706459.55	...	714521.47	770176.88	750118.47
3774	375865.22	148140.80	483523.92	245993.74	150117.79	155243.93	255021.54	163398.53	201042.55	246268.26	...	189352.16	181557.46	303508.27
4360	164345.09	181335.20	0.00	181799.04	197030.81	191879.03	201159.22	272839.45	537132.00	498081.02	...	220983.21	231960.39	392907.07
4541	162088.95	179212.63	158381.44	157992.98	329541.22	161563.51	153065.78	166457.61	479390.33	307571.86	...	398692.68	480257.15	34629.11
4810	0.00	118575.56	110930.74	117378.70	127885.33	130024.43	133571.35	136661.00	136992.33	137765.94	...	132789.32	145245.37	142647.55
5849	207987.71	205092.63	203684.55	545492.59	216423.51	222985.92	218497.24	218466.16	224342.98	417410.50	...	259587.61	255746.56	259211.24
6259	68931.99	70419.40	145747.91	133778.26	137134.20	257003.70	462667.24	607492.54	469141.62	493846.31	...	163381.07	167639.31	311544.91
6397	198913.46	191339.36	191192.23	198681.64	214243.89	171222.68	223366.43	207782.88	204309.21	201860.55	...	285364.38	213982.64	221974.35
6455	87339.45	98007.38	87171.85	87984.73	89534.32	97122.58	98449.51	141637.43	103715.39	111156.12	...	100377.34	161158.45	76731.44
7076	233910.45	108877.83	558451.00	517530.93	108669.23	356724.82	580528.45	397919.19	373580.64	73537.10	...	80534.10	520686.39	386755.76

Figura 4.9: Matriz de pago mensual de deudas para cada cliente

Con esta matriz, se construyó el conjunto de datos a usar en el modelo de predicción. La construcción del conjunto de datos se llevó a cabo mediante la formación de muestras para cada cliente en las que se toma como salida cada uno de los meses de los cuales se poseían datos de los 3 meses anteriores y del mes correspondiente para el año anterior, los cuales servirían como entrada de la muestra. Se debe tener en cuenta que esta ventana de tiempo se definió con base en diversas pruebas con los modelos de predicción. Así, al contar con registros de transacciones desde septiembre de 2016 hasta octubre de 2018, las muestras se construyeron tomando como salida los meses comprendidos entre septiembre de 2017 y

septiembre de 2018, el mes de octubre de 2018 no se tuvo en cuenta ya que no se poseen los datos completos de las transacciones para este mes. Por lo tanto, la estructura de las muestras en el conjunto de datos es la mostrada en la tabla 4.3. Todas las muestras obtenidas para cada cliente se utilizan en la construcción del conjunto de datos, el que contendrá entonces un número de muestras igual al número de meses considerados (13), multiplicado por el número de clientes muestreados.

Entrada	**Salida**
Gasto mes $k-3$	Gasto mes k
Gasto mes $k-2$	
Gasto mes $k-1$	
Gasto mes $k-12$	

Cuadro 4.3: Estructura de las muestras para la predicción de gastos.

Estos datos se escalan a un rango entre 0 y 1, ya que los modelos de regresión usualmente utilizados para la predicción de variables continuas se ven afectados por las magnitudes de los valores. Posteriormente, el conjunto de datos se divide en un subconjunto de entrenamiento y uno de pruebas, con proporciones de 75% y 25% del total de muestras en el conjunto, respectivamente.

4.4.2. Implementación del método de predicción

Para la predicción de gastos dentro de cada grupo se probaron diferentes técnicas de regresión como lo son *stochastic gradient descent*, *ridge* y máquinas de soporte vectorial con diferentes tipos de *kernels*, perceptrón multicapa y regresión lineal. Sin embargo, algunas de estas técnicas mostraron no ser adecuadas para el problema, ya que el tiempo requerido para el entrenamiento de los modelos es muy elevado y los resultados obtenidos son similares a los de técnicas más ligeras. Finalmente, los modelos que generaron mejores resultados con tiempos de entrenamiento manejables fueron *perceptrón* multicapa y *regresión lineal*.

Perceptrón multicapa Para el entrenamiento del perceptrón multicapa se probaron diferentes configuraciones de la red neuronal en lo que respecta a profundidad (número de capas), cantidad de neuronas en cada capa, funciones de activación, factores de entrenamiento, entre otros parámetros. Se observó que las diferencias entre los resultados para estas configuraciones son pequeñas, presumiblemente debido al número reducido de características de entrada al modelo por cada muestra. Aun así, los mejores resultados se obtuvieron para una configuración con una capa oculta de 3 neuronas, función de activación sigmoide y una tasa de entrenamiento inicial de 0.001, que disminuye a medida que se entrena el modelo.

Regresión lineal El modelo de regresión lineal se ajustó a los datos de entrada con diferentes ajustes en sus parámetros, como la configuración de la búsqueda de la intersección con el eje y y el número de iteraciones, la configuración final habilita la búsqueda de la intersección y toma una iteración por muestra. Los resultados de este modelo fueron similares

a los obtenidos con el perceptrón multicapa, no obstante, este último fue ligeramente superior.

Se decidió utilizar como modelo definitivo el perceptrón multicapa, ya que este, además de obtener resultados un poco mejores que la regresión lineal, posee mejores características de escalado en cuanto a número de características y relaciones no lineales entre las mismas y la salida, siendo más adecuado para datos complejos que la regresión lineal,

4.4.3. Resultados obtenidos

La evaluación de los modelos, mencionada previamente, se llevó a cabo mediante la obtención de diferentes métricas y representaciones visuales sobre el conjunto de datos de prueba, las cuales se listan a continuación:

- *R2 Score*.
- *Explained Variance Score*.
- *Error absoluto medio*.
- *Distribución del error absoluto*: consiste en una gráfica que relaciona los valores de errores absolutos de las predicciones con los valores reales correspondientes. Así, es posible visualizar qué valores de gastos es más difícil predecir para el modelo.
- *Histograma del error absoluto*: es una representación en barras del conteo de valores de error absoluto en rangos determinados. En este caso, se tomaron 10 rangos de errores entre $0 y $900.000, cada uno con un ancho de $100.000.

A continuación se muestran los resultados obtenidos para estas métricas sobre los datos del primer grupo demográfico, con diferentes umbrales de filtrado de los clientes respecto al número de meses con pagos de deudas diferentes a cero.

Umbral de 3 meses El umbral menos exigente usado es aquel en que los clientes requieren solamente de 3 meses con pagos de deudas diferentes de 0. Es uno de los umbrales en que mayor proporción de errores se presenta, ya que existen muchos datos iguales a cero para los cuales el modelo predice un gasto mayor. Esto es natural, ya que es poco probable que el modelo haga una predicción exacta de 0. En un ambiente real de una aplicación de PFM, el cliente debe ingresar constantemente sus gastos mes a mes, por lo que es inusual encontrar meses con gastos en pago de deudas iguales a 0, reduciendo de manera considerable la cantidad de estos errores.

- *R2 Score* = 0.2057.
- *Explained Variance Score* = 0.2058.
- *Error absoluto medio* = $129.105.

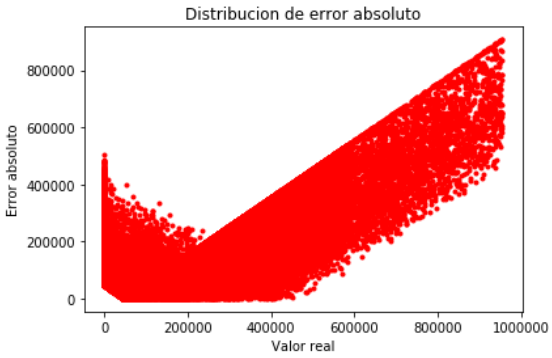

Figura 4.10: Distribución de error absoluto con un umbral de 3 meses

- *Distribución del error absoluto*: Para este umbral, el menor error absouluto se encuentra en las muestras cuyo valor real es cercano a $250.000 (ver Fig. 4.10).
- *Histograma del error absoluto*: En el histograma de la Fig. 4.11 se nota cómo la mayor concentración de errores se encuentra para las muestras cuyo valor real es menor a $100.000, conteniendo este rango aproximadamente al 60 % de los errores presentados. Esto es, *para aproximadamente el 60 % de los datos, el valor del pago de deudas mensual se puede predecir con un error máximo de $100.000.*

Figura 4.11: Histograma de error absoluto con un umbral de 3 meses

UMBRAL DE 12 MESES Es el umbral intermedio, en el que los clientes deben contar con pagos de deudas no nulos en aproximadamenta la mitad del periodo de muestreo. De los umbrales probados, es con el que se obtiene el peor desempeño del modelo con base en la mayoría de métricas, ya que al no existir tantos datos nulos como en el umbral de 3, el modelo no se ajusta a este comportamiento, lo cual afecta aún más su desempeño respecto a los mismos, que aún representan una cantidad considerable en la muestra.

- *R2 Score* = 0.1567.

- *Explained Variance Score* = 0.1564.
- *Error absoluto medio* = $167.494.
- *Distribución del error absoluto*: Para este umbral, el menor error absouluto se encuentra en las muestras cuyo valor real es cercano a $300.000 (ver Fig. 4.12).

Figura 4.12: Distribución de error absoluto con un umbral de 12 meses

- *Histograma del error absoluto*: En el histograma (Fig. 4.13) disminuye la diferencia en la concentración de errores en el rango de $0 a $100.000 respecto a otros rangos, en comparación con el umbral anterior.

Figura 4.13: Histograma de error absoluto con un umbral de 12 meses

UMBRAL DE 24 MESES Este es el umbral más exigente de los utilizados para las pruebas, requiere que los clientes cuenten con pagos de deudas en 24 meses o más, es decir, sólo pueden tener un mes con pago de deudas igual a cero en todo el periodo de la muestra. Con la aplicación de este umbral se elimina la mayoría de los valores nulos de las muestras, generándose una mejoría considerable en el desempeño del modelo.

- *R2 Score* = 0.3478.
- *Explained Variance Score* = 0.3482.
- *Error absoluto medio* = $99.026.
- *Distribución del error absoluto*: En este caso se puede ver cómo, a pesar de la concentración de errores en el valor real de 0, existe una mayor densidad de puntos en la parte baja de la gráfica, es decir, en los valores de error bajos (ver Fig. 4.14).

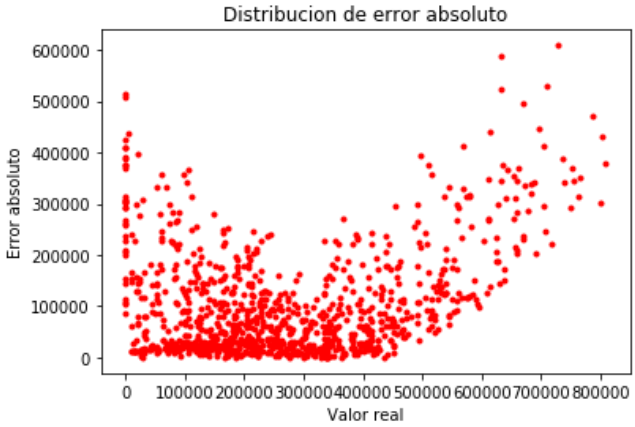

Figura 4.14: Distribución de error absoluto con un umbral de 24 meses

- *Histograma del error absoluto*: En el histograma (Fig. 4.15) los errores absolutos se concentran de nuevo en el rango de $0 a $100.000, representando un gran porcentaje del total, lo que significa una buena calidad de las predicciones.

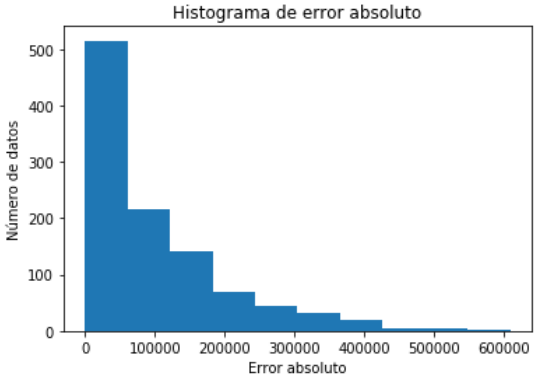

Figura 4.15: Histograma de error absoluto con un umbral de 24 meses

Así pues, es evidente cómo la presencia de valores nulos en los pagos de deudas mensuales de los clientes, los cuales no se esperan en una situación real de uso de un software de PFM, es la principal causa de errores de predicción en el modelo utilizado, lo cual se repite de manera consistente para los modelos de las dos comunidades de clientes restantes. Es evidente además que al reducir el número de valores nulos en las muestras, el desempeño de la predicción mejorará, disminuyendo el error absoluto. Además, se puede ver cómo, incluso con los errores inducidos por estos valores en el umbral más bajo, el sistema es capaz de predecir, con un error máximo de $100.000 en más del 60% de los casos, los gastos en pago de deudas que tendrán los clientes en un mes determinado. Poniendo esto en contexto respecto al objetivo de identificar clientes potenciales para productos crediticios, es claro que este error es pequeño comparado con las magnitudes de las deudas que el cliente debe ser capaz de asumir para adquirir uno de estos productos, por lo que la predicción obtenida es de utilidad para esta aplicación.

5. SISTEMAS DE RECOMENDACION CREDITICIA

Para un cliente natural, es de vital importancia tener siempre presente su capacidad de endeudamiento para lograr administrar adecuadamente sus gastos. Si bien, en las aplicaciones de PFM las personas pueden recibir alertas de acuerdo con las metas de ahorro fijadas y ver información de sus gastos, no cuentan con un mecanismo automatizado que pueda predecir el gasto de pago de deudas y hacer las recomendaciones adecuadas. Por su parte, el banco puede usar esta información para enfocar sus campañas de productos crediticios a aquellos que se prediga no van a utilizar toda su capacidad financiera en los siguientes meses. Así, con todo el desarrollo expuesto en las anteriores secciones proponemos un "Sistema de recomendación crediticia" que busca ser útil tanto a los usuarios de una aplicación PFM como para el banco.

Aprovechando el modelo predictivo de pago de deudas, el cual tiene un error menor a 100,000 para más de un 60% de los clientes en el peor de los casos, se propone evaluar si ese monto estimado está por debajo, o por encima de la capacidad crediticia de cada persona (determinada por sus ingresos, que se conoce). Se espera entonces que el sistema, dependiendo del caso, tome alguna de las siguientes acciones:

- *Predicción de monto **Pago de Deudas** por debajo de capacidad de endeudamiento*: si esto sucede, al banco se le sugiere ofrecer incentivos para que el cliente adquiera un mayor monto en crédito.
- *Predicción de monto **Pago de Deudas** por encima de capacidad de endeudamiento*: si esto sucede, la aplicación PFM deberá generar alertas en los gastos crediticios de la persona durante el mes, buscando que este adquiera un endudamiento saludable.

Finalmente, cabe mencionar que el modelo de predicción propuesto se puede aplicar a los gastos totales del cliente o a sus gastos en las demás categorías, lo que permite, entre otras cosas, aproximar mejor la capacidad de endeudamiento del cliente y dar más recomendaciones a este.