FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES

MÁSTER EN CIENCIAS ACTUARIALES Y FINANCIERAS

TRABAJO DE FIN DE MÁSTER

TÍTULO: *Determinantes de la Rentabilidad en Cooperativas de ahorro y crédito de Ecuador. Un análisis mediante Machine Learning.*

AUTOR: *Freddy Alejandro Oquendo Torres*

TUTORA: *María Jesús Segovia Vargas*

CURSO ACADÉMICO: 2020-*2021*

CONVOCATORIA: *Septiembre*

Tabla de contenido

Índice de Gráficos:

Índice de tablas:

Resumen

El presente trabajo busca evidenciar empíricamente los principales determinantes que influyen en la rentabilidad de las Cooperativas de ahorro y crédito (COAC) de Ecuador. Para este fin, se implementarán modelos de aprendizaje automático como el árbol de clasificación CART, el Random Forest y el Gradient Boosting Machine. Mediante el empleo de estos modelos se busca llegar al objetivo principal del trabajo que es predecir los determinantes de la rentabilidad y evaluar cuál de las metodologías empleadas es más eficiente en cuanto a la clasificación de las COAC en Ecuador en rentables o no rentables. Se ha elaborado una base de datos que cuenta con un total de 510 observaciones para el año 2020. Para el análisis se han seleccionado dos variables dependientes que son el ROE y la Rentabilidad de la cartera de microcrédito y, como variables independientes, y de acuerdo con la teoría revisada, se han utilizado ratios financieros específicos para este tipo de entidades que miden la liquidez, solvencia, calidad del crédito, eficiencia y tamaño de la entidad. El presente trabajo llegó a la conclusión que, de los modelos utilizados, el modelo de Gradient Boosting Machine es el que mejor predice la rentabilidad, tanto en el caso del ROE como para la rentabilidad de la cartera de microcrédito. Además, las variables relacionadas con el tamaño de la entidad y el crédito son las que más influyen a la hora de clasificar una entidad como rentable o no rentable.

Palabras clave:

Rentabilidad; Sistema financiero ecuatoriano; Árbol de clasificación; Random Forest; Gradient Boosting Machine.

Abstract

This work seeks to empirically demonstrate the main determinants that influence the profitability of savings and credit cooperatives in Ecuador. For this purpose, machine learning models will be implemented such as: Classification tree, Random Forest, and Gradient Boosting Machine through which it is sought to reach the main objective of the work, which is to predict the determinants of profitability and evaluate which of the methodologies used is more efficient regarding the classification of COACs in Ecuador as profitable or unprofitable. The analyzed database has a total of 510 observations for the year 2020. The dependent variables will be ROE and Profitability of the microcredit portfolio; As independent variables and in accordance with the revised theory, internal or specific factors of the financial institution will be used, which will be proportions that measure the liquidity, solvency, credit quality, efficiency, and size of the entity. The present work concluded that the Gradient Boosting Machine is the one that best predicts profitability. And the variables related to the size of the entity and the credit are the ones that most influence when classifying an entity as profitable or unprofitable.

Key Words:

Profitability; Ecuadorian financial system; Classification Tree; Random Forest; Gradient Boosting Machine.

1. Introducción:

El crecimiento de las entidades financieras junto con la estabilidad de la economía se ha vuelto un factor importante de análisis en los últimos años. De acuerdo con Hollis y Sweetman (1998) la estabilidad en una entidad financiera es una condición necesaria para el buen funcionamiento de estas. Por otro lado, Ledgerwood (1999) hace referencia a los principales riesgos que puede tener una entidad de ahorro y crédito, mencionando que algunas de las entidades en su tiempo de permanencia en el mercado no alcanzan un grado mínimo de eficiencia financiera necesario para cubrir sus costes, y que en ciertos casos no tienen un adecuada gestión de sus ganancias provocando futuros problemas de liquidez y solvencia. Por lo tanto, el análisis de la rentabilidad de las entidades financieras es de primordial importancia para la buena gestión y desempeño de su estructura económica.

El objetivo del presente trabajo es implementar diferentes modelos de Machine Learning que permitan predecir qué variables son las que más influyen en la rentabilidad de empresas del Sector Financiero Popular y Solidario (SFPS) de Ecuador. Para lograr dicho objetivo se han extraído los datos económico-financieros de las entidades financieras que constituyen el SFPS y se ajustarán diferentes modelos para saber cuál de ellos clasifica mejor entre entidades rentables y no rentables.

Para analizar empíricamente los objetivos planteados, se implementará modelos de aprendizaje supervisado como el Árbol de clasificación, Random Forest, y Gradient Boosting Machine. Con respecto a las variables que se utilizarán, se aplicarán los 3 modelos para la variable dependiente Rentabilidad sobre el patrimonio (ROE) y los mismos modelos para la variable dependiente Rentabilidad de la cartera de microcrédito (RCM). Las variables independientes serán ratios financieros que serán descritos más adelante y son los comúnmente seleccionados por la literatura contable.

Para calcular los ratios financieros, se han obtenido los balances financieros mensuales que son publicados por la Superintendencia de Economía popular y solidaria de Ecuador (SEPS, 2021), siendo este el organismo que controla las Cooperativas de Ahorro y Crédito

(en adelante COACS). Cabe mencionar que la información contenida en los balances será al cierre de diciembre de 2020. Sin embargo, La población total se compone de las entidades que estén únicamente con estado actual de "Activas" al 31 de marzo de 2021, dando una base de datos final de 510 entidades.

La importancia de analizar este sector del sistema financiero de Ecuador radica en: i) la influencia de las COAC en el crecimiento económico; ii) La importancia que puede tener el rendimiento de las COACS en la estabilidad del sistema financiero (cabe destacar que muchas de ellas mantienen problemas financieros y el órgano rector (SEPS) decide su liquidación, pudiendo conducir a un pánico bancario o riesgo sistémico) ; y, iii) actualmente no se han realizado investigaciones para el sistema financiero de Ecuador sobre la rentabilidad utilizando modelos de Machine Learning, principalmente existen investigaciones utilizando metodologías como la CAMEL[1].

La presente investigación se compone de seis secciones, siendo la primera esta introducción, mediante las cuales se intenta cumplir con los objetivos propuestos. En la segunda sección se indica el marco teórico de la rentabilidad financiera, seguido del contexto del sistema financiero actual en Ecuador. En la cuarta sección se analizará la metodología que será usada para verificar los determinantes de la rentabilidad. Posteriormente se analizarán los resultados obtenidos para finalizar con las conclusiones del trabajo.

2. Marco Teórico

En el siguiente apartado se hablará sobre la constitución del sistema financiero ecuatoriano, y una descripción de la literatura previa realizada relacionada con los determinantes de la rentabilidad en un sistema financiero. En general, de acuerdo con la literatura revisada, numerosos autores como: Smirlock (1985); Berger (1995) Lawrence y Joe (1999); Demirgüç-Kunt y Huizinga (2000); Bakar y Tahir (2009); Erdal y Karahanoglu (2016); Gonzáles (2019). Hacen énfasis y proponen que los determinantes

[1] CAMEL: Sistema de medición utilizado en entidades financieras donde su propósito es tener una evaluación de la solidez financiera de una entidad basándose en indicadores de Capital, activos, manejo de la entidad, ganancias y utilidades (Dang, 2011).

de la rentabilidad pueden dividirse en factores internos y factores externos.

2.1 Sistema financiero de Ecuador

El Codigo Orgánico Monetario y Financiero, en el Titulo II *"Sistema financiero nacional"*, Capitulo 2 *"Integración del sistema financiero nacional"*, menciona que el sistema financiero nacional de Ecuador está conformado por el sector financiero público, sector financiero privado y el sector financiero popular y solidario. El sector financiero público se encuentra conformado por bancos y corporaciones y en el sector financiero privado se puede encontrar bancos múltiples o bancos especializados. Por otro lado, el sector financiero popular y solidario (SFPS) se compone de: Cooperativas de Ahorro y Crédito (COAC); Mutualistas, Cajas centrales, Banco comunal y Cajas de ahorro.

Actualmente, en el sector financiero popular y solidario se encuentran activas un total de 510 COAC. Estas 510 entidades, de acuerdo con la resolución No. 038-2015-F para la segmentación de entidades del SFPS emitida por la Junta de política y regulación monetaria y financiera, se segmentan de acuerdo con su nivel de activos de la siguiente manera (Tabla 1):

Tabla 1 : Segmentos de COACS

Segmento	Activos (USD)	No. COACS
1	Mayor a 80.000.000	35
2	Mayor a 20.000.000 hasta 80.000.000	46
3	Mayor a 5.000.000 hasta 20.000.000	84
4	Mayor a 1.000.000 hasta 5.000.000	162
5	Hasta 1.000.000	183

Fuente: SEPS

Según datos de la SEPS a diciembre de 2020 el total del sector de COAC alcanza un valor de 16.694 millones de dólares en activos, los cuales representan el 17.27% del PIB de la economía de Ecuador. Por otro lado, según datos del Banco Central del Ecuador, las obligaciones con el público representan el 13.55% del PIB (13.095 millones de dólares). Es importante mencionar que desde el año 2011, año en el cual la SEPS se encarga de monitorear y controlar la estabilidad de las COAC, se han dado un total de 98 liquidaciones de entidades por lo que se vuelve indispensable el análisis de la rentabilidad de este sector.

2.2 Rentabilidad sistema financiero

Con respecto a la revisión de la literatura, los estudios previos realizados sobre los determinantes de la rentabilidad en un sistema financiero han evidenciado que pueden realizarse mediante métodos paramétricos y mediante métodos no paramétricos. Se comenzará describiendo las principales evidencias de los métodos paramétricos y luego de los trabajos no paramétricos. Cabe mencionar que en el presente trabajo se usaran principalmente los últimos detallados.

Con respecto a los métodos paramétricos, la base teórica para analizar los determinantes de la rentabilidad nace mediante la utilización de la hipótesis de estructura-conducta-desempeño "SCP" y la hipótesis de estructura eficiente "ES". La hipótesis "SCP" fue propuesta por Bain (1951) y menciona que existe una relación positiva entre la concentración de empresas en un mercado y la rentabilidad. Por otro lado, la hipótesis "ES", que fue desarrollada por Demsetz (1973) y Peltzman (1977), sugiere que la relación entre concentración y rentabilidad es reflejada por el tamaño de las empresas alcanzando así mayores beneficios por ser eficientes y no por una concentración o coalición en un mercado.

A partir de estas teorías, varios autores han realizado estudios para verificar la búsqueda de determinantes de rentabilidad empresarial en diferentes sectores económicos. Por parte del sector financiero, los determinantes de la rentabilidad tienen un mayor desarrollo para el sector bancario. Esta temática ha sido estudiada en países desarrollados, como Estados Unidos por Heggestad et al. (1976), Canadá, Europa Occidental y Japón por Short (1979), haciendo referencia a la coalición en la estructura del sistema financiero. Por otro lado, autores como Smirlock (1985) y Berger (1995) estudiaron la hipótesis "ES" para ver la rentabilidad en el sector financiero, llegando a la conclusión que la coalición de un mercado podría tener una relación negativa con la rentabilidad y que variables que son más específicas de los bancos como ratios de capital, liquidez, tamaño del mercado, influyen en mayor proporción a la rentabilidad.

Siguiendo con la línea anterior, muchos autores concluyen que la rentabilidad del sistema financiero está en función de factores internos y factores externos. Los determinantes internos se pueden definir como aquellos factores que están influenciados por las decisiones de la gestión de las organizaciones con efecto en los resultados operativos de la entidad. Los determinantes externos se pueden dar mediante impactos de variables macroeconómicas.

Utilizando modelos de datos de panel y series de tiempo, autores como Goddard et al. (2004) y Staikouras et al. (2004) realizan un análisis para medir la rentabilidad en diferentes periodos utilizando muestras de bancos de la Unión Europea. Como variable dependiente utilizan el ROE y el ROA, y en sus variables explicativas utilizan: tamaño de los bancos medido como el logaritmo del total de los activos; variables para medir la adecuación del capital como el ratio entre préstamos y activos, capital entre activos; incluye índices de concentración de mercados. Además, usan variables macroeconómicas como el PIB, inflación, tasas de interés. Llegan a la conclusión que las variables más significativas dentro de su modelo son las variables que miden la adecuación del capital.

Haslem (1968) y Lawrence y Joe (1999) realizan diferentes análisis estadísticos para determinar la rentabilidad en bancos de Estados Unidos. En sus estudios incluyen diferentes ratios financieros segmentados por categorías como tamaño, localización, gestión y calidad de activos. Llegan la conclusión de que de los ratios financieros utilizados los que más influencia tienen en la rentabilidad son: rendimiento obtenido de préstamos, tasas de interés y el volumen de depósitos. Y, que el tamaño de los bancos en este caso es significativo, pero tiene un impacto negativo con respecto a la rentabilidad.

Demirgüç-Kunt y Huizinga (2000) utilizan datos de países de la OCDE para medir el impacto de la estructura del sistema financiero en la rentabilidad bancaria. Utilizan modelos de regresión lineal tomando el ROE como variable dependiente y entre las variables explicativas incluye el total de activos dividido para el PIB proporcionando una medida del tamaño total del sector bancario. Esta variable es significativa pero negativa,

haciendo alusión que mientras mejor estructura tenga un sistema financiero habrá más competitividad por lo que la rentabilidad será más baja. Por otro lado, variables como el volumen de créditos para el PIB y las inversiones para el PIB las incluye mencionando que estas variables suelen ser significativas para países en desarrollo. Finamente, incluyen variables macroeconómicas como la inflación y las tasas de interés e indican que estas variables son significativas en su modelo mostrando una relación positiva entre la inflación, tasas de interés y la rentabilidad del sistema financiero.

Lapo y Tello (2021) en su investigación realizan un modelo estadístico de mínimos cuadrados parciales para medir la influencia del capital en la rentabilidad bancaria de Ecuador. Utilizando como variable explicativa el ROE y como variables explicativas la estructura de capital o solvencia el nivel de capital, el ratio de activo para patrimonio y el ratio del pasivo para patrimonio. Concluyeron que la rentabilidad de los bancos en su mayoría depende de su capital, ya que este puede generar efectos beneficiosos o adversos según sea el caso. Por ello es importante, que no solo se garanticen las operaciones bancarias, sino que también, se puedan determinar las decisiones de estructura de capital que aumenten la rentabilidad del banco para seguir operando y crecer de forma competitiva

Con respecto a los métodos no paramétricos, se identificaron algunos estudios relacionados con los determinantes de la rentabilidad en un sistema financiero, si bien existen varios trabajos realizados para temas financieros, con respecto a la rentabilidad aún no hay muchos trabajos. A continuación, se citan los principales.

Gonzáles (2019) realiza un estudio para ver los determinantes que influyen en la rentabilidad de los bancos de Panamá mediante técnicas de árboles de decisión y redes neuronales. El objetivo de su trabajo fue inferir que variables influyen en la rentabilidad bancaria y evaluar cuál de las técnicas que se utilizaron se ajusta de mejor manera. Para lo mismo utilizó una muestra de 46 bancos y los segmentó para realizar el “training data”, “validation data” y “test data” para el caso de los árboles de decisión utilizó datos de años anteriores (2015-2017) y para el caso de la red neuronal utilizó la misma muestra, pero los separo en tres partes (80%, 10%,10%). Para realizar su predicción,

utilizó el ROA como variable dependiente de sus modelos y se concluye que las variables más significativas utilizando un árbol de decisión son la calidad del crédito, la liquidez y el tamaño del banco. Por parte de la red neuronal las variables que determinan la rentabilidad son la solvencia, tamaño, liquidez y calidad del crédito, siendo esta última técnica la que obtiene un mayor poder de predicción.

Uddin et al. (2020) realizan un modelo de Random Forest para la detección de default en el riego de crédito utilizando una base de datos de un banco comercial de china. Utilizan un total de 81 variables entre financieras, no financieras y macroeconómicas. Llegando a la conclusión de que las variables que más influyen en su modelo son las de solvencia y rentabilidad. Además, también conluyen que el modelo Random Forest es una técnica eficiente de clasificación para el problema que analizan.

Bakar y Tahir (2009) realizan una investigación para predecir el rendimiento de los bancos en Malasia en el periodo 2001-2006, utilizando el ROA para medir la rentabilidad de los bancos, y como variables independientes a ratios financieros y variables macroeconómicas. Los ratios financieros considerados son de liquidez, riesgo de crédito, relación coste-ingreso y el total de los activos. Como metodología aplican una red neuronal, indicando que del total de la muestra el 80% se utilizó para realizar el "training data", el 10% para "testing data" y el otro 10% para el "validation data". Los resultados muestran que las siete variables independientes predicen que el 61,9% de los bancos son rentables.

Erdal y Karahanoglu (2016) utilizan como modelos de predicción el Random Forest, y árboles de decisión para una muestra de 13 bancos de Turquía con 10 variables financieras, entre ellas, Patrimonio/Activos; Prestamos/Activos; Créditos totales/ Total de activos; Ingresos/ activos, Activos Líquidos/Total de activos. Como variable dependiente utilizan el ROE en el periodo comprendido ente 2002 al 2014, si bien su objetivo no era el de determinar qué variable explica mejor la rentabilidad. A través de la utilización de los estadísticos como el R^2 y el error cuadrático medio, llegan a la conclusión de que se tiene una mejor predicción con el Random forest que con el árbol de decisión.

En conclusión, la revisión de la literatura respalda que existen variables que pueden influir en la rentabilidad, y concretamente, los estudios empíricos demuestran que hay una relación entre factores internos como los indicadores financieros tomados de balances y la rentabilidad financiera. Además, para establecer los determinantes se pueden aplicar distintos modelos, paramétricos o no paramétricos. En el presente estudio nos centraremos únicamente en los métodos no paramétricos.

3. Contexto

3.1 Comportamiento financiero del Sector

Según datos de la SEPS a diciembre de 2020, las COAC en su totalidad alcanzaron la suma de US$ 16.691 millones en el total de los activos, posición que, al compararla con diciembre de 2019, se observa un incremento del 10,92%. (US$ 1.643 millones), originado por el incremento del segmento uno. Visto lo anterior, pero de acuerdo con cada segmento de las COACS, el segmento 1 representa el 79,47% del sistema, el segmento 2 el 11,92%, el segmento 3 representa el 5,49%, el segmento 4 el 2,60%; y, finalmente el segmento 5 el 0,51%. Lo mencionado en el párrafo anterior puede ser verificado en el apartado de anexos (tabla 19).

Dentro de la estructura del activo (tabla 20 - anexos) se puede evidenciar que la cuenta que más aporta al total es la cartera de créditos con 67%, seguido por fondos disponibles con 14% y, por las inversiones con 11%. La cartera de créditos bruta del total de COAC muestra una concentración de 84% en dos líneas de negocio: consumo prioritario con 46% y microcrédito con 38%, seguida en importancia por la cartera de crédito inmobiliario con 9%, por la cartera de consumo ordinario con 5%; y, la diferencia se distribuye en los demás segmentos de crédito. Cabe puntualizar, que las COAC del segmento 1 y 2 privilegian el consumo prioritario (en promedio 47%), mientras que, para el resto de los segmentos de 3, 4 y 5 (58%) se inclinan por el microcrédito. (SEPS, 2021)

Tabla 2 : Distribución Cartera de Créditos

CARTERA	SEGMENTO 1	SEGMENTO 2	SEGMENTO 3	SEGMENTO 4	SEGMENTO 5	TOTAL COAC
COMERCIAL PRIORITARIO	1,41%	1,00%	0,89%	0,32%	1,20%	1,30%
CONSUMO PRIORITARIO	47,50%	46,89%	37,02%	32,96%	37,63%	46,37%
INMOBILIARIO	10,41%	5,43%	2,05%	1,27%	3,45%	8,99%
MICROCRÉDITO	34,87%	42,95%	57,79%	62,85%	54,03%	38,10%
PRODUCTIVO	0,22%	0,00%	0,04%	0,10%	0,29%	0,18%
CONSUMO ORDINARIO	5,53%	3,71%	2,12%	2,43%	3,25%	5,00%
COMERCIAL ORDINARIO	0,03%	0,00%	0,00%	0,01%	0,00%	0,02%
VIVIENDA DE INTERÉS PÚBLICO	0,03%	0,00%	0,02%	0,00%	0,15%	0,03%
EDUCATIVO	0,00%	0,02%	0,07%	0,05%	0,00%	0,01%
TOTAL	100%	100%	100%	100%	100%	100%

Fuente: Elaboración propia

Con respecto a la cuenta de los pasivos, según datos de la SEPS a diciembre de 2020 el pasivo total fue de US$ 14.284 millones, mostrando un crecimiento de 11,62% comparado con los resultados alcanzados en similar período del año 2019 (gráfico 1). De los segmentos que conforman este sector, en términos relativos, el de mayor incremento entre los periodos mencionados fue el segmento 1 (14,52%) y el segmento 2 (5,38%). Lo mismo que puede ser evidenciado en la tabla 19 en el apartado de anexos.

La principal cuenta del pasivo correspondió a las obligaciones con el público (tabla 22 - Anexos) y las mismas representaron 92%. Las fuentes secundarias de captación fueron las obligaciones financieras (5,2%); y, las cuentas por pagar (2,8%). En la cuenta de obligaciones con el público, los depósitos a la vista representan 28,46%, depósitos a plazo 69,76% y, la diferencia se reparte entre depósitos de garantía y restringidos. Por parte de los depósitos a plazo, se concentran en depósitos de 31 a 90 días (26,35%), de 91 a 180 días (22,56%), de 181 a 360 días (27,38%), de 1 a 30 días (18,16%) y, mayores a 360 días (5,54%) (véase la Tabla 3):

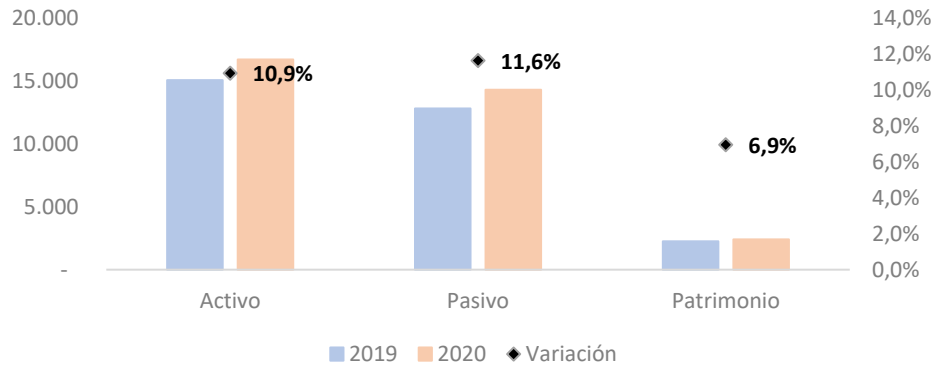
Tabla 3 : Distribución Depósitos a plazo

PLAZO DEPÓSITOS	SEGMENTO 1	SEGMENTO 2	SEGMENTO 3	SEGMENTO 4	SEGMENTO 5	Total COAC
De 1 a 30 días	18,57%	16,48%	16,68%	14,05%	12,03%	18,16%
De 31 a 90 días	26,76%	24,39%	25,83%	21,28%	17,67%	26,35%
De 91 a 180 días	22,76%	22,04%	20,39%	21,77%	23,27%	22,56%
De 181 a 360 días	27,01%	29,34%	28,47%	29,91%	32,88%	27,38%
De más de 361 días	4,90%	7,75%	8,62%	12,93%	14,05%	5,54%
Depósitos por confirmar	0,01%	0,00%	0,01%	0,06%	0,10%	0,01%
TOTAL	**100%**	**100%**	**100%**	**100%**	**100%**	**100%**

Fuente: Elaboración propia

Con respecto al patrimonio, el gráfico 1 muestra que, a diciembre de 2020 según datos de la SEPS, alcanzó la suma de US$ 2.407 millones reflejando un aumento de 6,95% en comparación al valor registrado en 2019 (US$ 2.251 millones). Este grupo se compuso por capital social (33,68%), reservas (55,99%) y superávit por valuaciones (6,69%).

Gráfico 1: Activo Pasivo Patrimonio

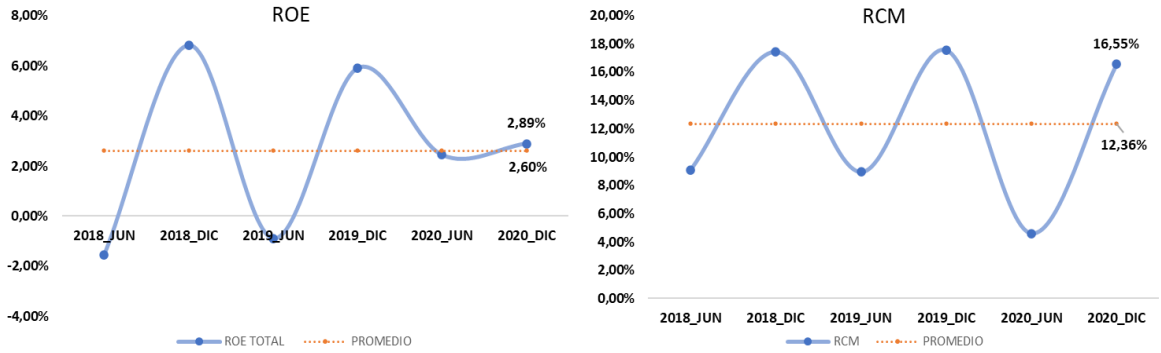


Fuente: Elaboración propia

3.2 Rentabilidad sistema financiero

De acuerdo con Brealey et.al (1996) la rentabilidad intenta medir el modo en que una entidad, después de haber realizado su actividad fundamental de ventas o prestaciones de servicios, que en el caso de las COACS será la intermediación financiera, es capaz de generar utilidad. Con respecto a las COAC, según la SEPS (2020) el ROE alcanzó el valor de 2,89% ubicándose por encima del promedio del periodo entre junio de 2018 a diciembre de 2020 (2,60%). Por otro lado, la RCM alcanzó un valor de 16,55% a diciembre de 2020 situándose por encima del promedio del periodo (12,36%).

Gráfico 2: Evolución ROE y RCM



Fuente: Elaboración propia

4. Metodología

Para poder llegar al objetivo de hallar qué variables determinan si una COAC es rentable o no es rentable, se implementarán diferentes modelos de aprendizaje supervisado como son los árboles de clasificación, el Random Forest y Gradient Boosting Machine. Se aplicarán los 3 modelos mencionados anteriormente para la variable de Rentabilidad sobre el patrimonio (ROE) y para la variable Rentabilidad de la cartera de microcrédito (RCM). En cuanto a las variables independientes serán ratios financieros, los cuales fueron seleccionados basándose en estudios previos realizados (Smirlock (1985); Berger (1995); Lawrence y Joe (1999); Demirgüç-Kunt y Huizinga (2000); Bakar y Tahir (2009); Erdal y Karahanoglu (2016); Gonzáles (2019)).

4.1 Descripción de la muestra

La información se obtendrá de los balances financieros mensuales que son publicados por la SEPS con fecha de diciembre 2020. La muestra se compone de la totalidad de COAC que forman parte del sector financiero popular y solidario que, al 31 de marzo de 2021 y solo se tendrán en cuenta aquellas en las que figuren con estado de "Activas", dando una base de datos total de 510 entidades[2]. A partir de la muestra mencionada se calcularon un total de 28 indicadores financieros dentro de los cuales están las dos variables dependientes y las independientes.

4.2 Descripción de variables

4.2.1 Variable Dependiente

Como se ha mencionado previamente, las variables a utilizar como dependientes serán el ROE y RCM. Según la ficha metodológica realizada por la SEPS en el año 2017, el ROE mide el nivel de retorno generado por el patrimonio invertido por los accionistas de la entidad financiera, mientras que la Rentabilidad de la cartera de microcrédito[3] se refiere al rendimiento que tiene la cartera de microcrédito, sujeto a una banda maduración, es

[2] En el apartado de Anexos se encuentra a manera de ejemplo una muestra de la base de datos utilizada.
[3] Cartera de Microcrédito: Es el otorgado a una persona natural o jurídica con un nivel de ventas anuales inferior o igual a USD 100,000.00, o a un grupo de prestatarios, destinado a financiar actividades de producción y/o comercialización en pequeña escala, cuya fuente principal de pago la constituye el producto de las ventas o ingresos generados por dichas actividades. (SEPS, 2017)

decir en función del rango del vencimiento futuro de las operaciones. Las citadas rentabilidades se calculan de la siguiente manera:

$$ROE = \frac{Ingresos - Gastos}{Patrimonio\ Promedio}$$

$$RCM = \frac{Intereses\ cartera\ microcrédito\ prioritario}{Total\ cartera\ microcrédito}$$

Para poder clasificar a las COAC en entidades rentables o no rentables se realizó un análisis descriptivo de las dos variables dependientes para obtener un punto de corte y así poder clasificar las entidades. En la siguiente tabla se puede observar los principales estadísticos para el ROE y la RCM

Tabla 4 : Estadísticos descriptivos

MEDIDA	ROE	RCM
MEDIA	0,0101	0,1485
MEDIANA	0,0063	0,1618
DESVIACION	0,0649	0,1355
PRIMER CUARTIL	-	0,0459
TERCER CUARTIL	0,0311	0,1942
MINIMO	-0,4933	-
MAXIMO	0,1740	1,4115
ASIMETRÍA	-3,8641	3,6928
CURTOSIS	23,7158	26,8119

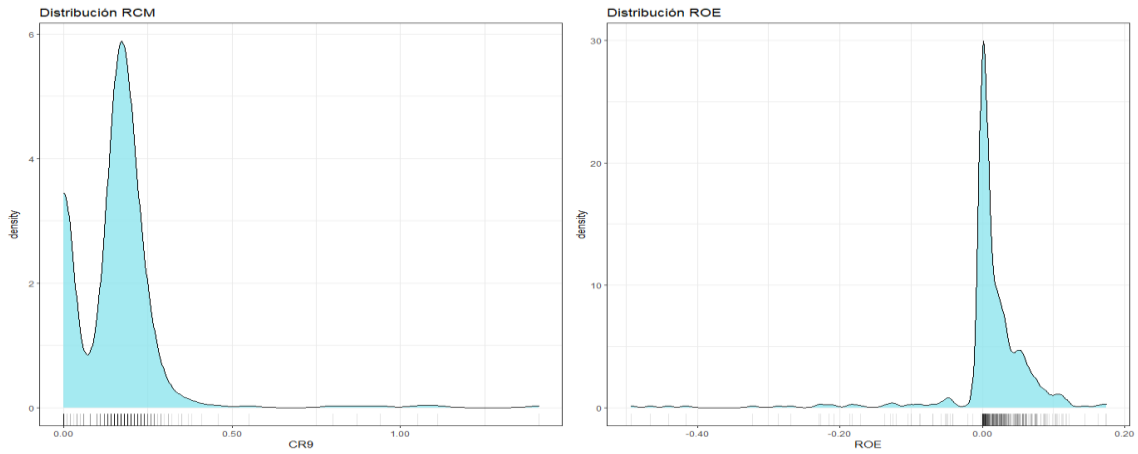
Fuente: Elaboración propia

De acuerdo con la tabla 4 se puede observar que las variables dependientes no se distribuyen normalmente, lo cual podría significar problemas a la hora de realizar ciertos modelos estadísticos. Sin embargo, el presente análisis al ser mediante métodos no paramétricos, el mismo no influirá en los resultados finales del modelo. Así se puede observar que el caso del ROE presenta una asimetría negativa, lo que indica que los datos no se distribuyen de manera uniforme alrededor del promedio y en este caso los datos están distribuidos a la derecha del promedio. Por otro lado, la curtosis presenta un valor positivo dando como resultado una distribución leptocúrtica, lo que quiere decir que existe una gran concentración de los datos alrededor de su media.

Por otro lado, la RCM presenta un coeficiente de asimetría positivo lo que indica que los datos se encuentran a la izquierda del promedio y estos no se distribuyen

uniformemente alrededor de la media. Sin embargo, la curtosis, al igual que el ROE, al ser positiva presenta una distribución leptocúrtica. En el gráfico 3 se puede observar lo mencionado.

Gráfico 3: Distribución RCM y ROE



Fuente: Elaboración propia

Para poder discretizar las variables dependientes y poder implementar los modelos de clasificación, tomando en cuenta la distribución de la variable y de acuerdo con estudios implementados por Gonzáles, Correa, y Acosta (2002); y, Gonzáles (2019) para el caso del ROE se clasificarán las entidades mediante el signo de su indicador, para el caso del RCM se eligió a la media como medida de discretización, las entidades que sean mayores o igual al promedio de la RCM serán clasificadas como rentables, en caso contrario serán no rentables.

Tabla 5 : Clasificación variable dependiente

VARIABLE	ROE	RCM
RENTABLE	339	311
NO RENTABLE	171	199
TOTAL	**510**	**510**

Fuente: Elaboración propia

4.2.2 Variables Independientes

Con respecto a las variables independientes[4], se tomó como base la ficha metodológica realizada por la SEPS en el año 2017 y los diferentes textos mencionados en el apartado del marco teórico. Como primer paso se realizó un análisis descriptivo para ver el

[4] Las variables independientes son detallas en el apartado de Anexos

comportamiento de las variables independientes, para luego ver la correlación que existe entre sí, obteniendo los siguientes resultados:

Tabla 6 : Descriptivos variables independientes

INDICADOR	MÍNIMO	MÁXIMO	MEDIA	DESVIACIÓN	VARIANZA	ASIMETRÍA	CURTOSIS
SOL1	-411707,37	42256,40	-592,15	18562,78	344576741,57	-21,42	475,36
SOL2	-351,43	2206,49	20,77	117,70	13852,57	13,36	237,27
SOL3	0,00	9452,04	21,19	418,67	175286,48	22,54	508,79
SOL4	-2,77	0,88	0,24	0,21	0,04	-5,18	82,84
SOL5	-0,67	8,01	0,45	0,72	0,52	5,90	47,19
EFG1	-2455,04	4372,47	108,20	297,77	88668,24	4,27	104,76
EFG2	0,00	84,90	8,13	6,39	40,89	6,17	59,02
CAA1	0,00	1378,87	127,49	100,08	10016,68	8,12	85,22
CAA2	0,00	136,20	81,80	19,76	390,53	-1,94	4,11
CAA3	-36,20	100,00	18,20	19,76	390,53	1,94	4,11
CR1	0,00	100,00	12,85	17,89	320,07	2,85	8,91
CR2	0,00	175233,00	582,76	8190,82	67089463,04	19,87	413,39
CR3	0,00	614,54	35,75	35,48	1258,93	9,75	144,98
CR4	0,00	50,03	14,49	5,35	28,57	0,96	7,15
CR5	0,00	56,88	5,64	3,70	13,71	5,18	71,12
CR6	-26,09	42,90	8,85	5,08	25,85	1,16	13,34
CR7	0,00	3,37	0,10	0,24	0,06	11,14	139,95
CR8	0,00	19,45	0,15	1,13	1,29	13,52	201,92
EFF1	-2229,91	48,35	-6,95	100,43	10085,68	-21,46	474,37
EFF2	-649,40	834,56	-2,62	54,44	2963,55	2,63	157,16
LIQ1	0,00	709,98	38,45	48,77	2378,44	7,57	84,78
LIQ2	-1,63	0,94	0,64	0,19	0,04	-4,58	43,20
LIQ3	-2,23	15,99	1,11	0,95	0,91	9,14	126,18
T1	7,62	21,65	14,75	2,17	4,73	0,31	0,22
T2	0,00	19,38	13,13	2,20	4,86	-0,90	6,66
T3	0,00	19,37	12,43	2,54	6,43	-0,51	1,98

Fuente: Elaboración propia

En cuanto a los coeficientes de correlación de las variables independientes, se puede verificar que muchas de las variables independientes que se desean introducir para el análisis están fuertemente correlacionadas. Por lo tanto, de las 26 variables independientes se realizó un análisis de componentes principales para poder reducir las dimensiones de la base de datos y presentar variables más estandarizadas.

Tabla 7 : Matriz de correlaciones

	SOL1	SOL2	SOL3	SOL4	SOL5	EFG1	EFG2	CAA1	CAA2	CAA3	CR1	CR2	CR3	CR4	CR5	CR6	CR7	CR8	EFF1	EFF2	LIQ1	LIQ2	LIQ3	T1	T2	T3
SOL1	1,00	0,02	0,00	0,02	0,02	0,00	0,03	0,01	-0,04	0,04	0,02	0,00	-0,01	0,01	0,00	0,01	0,00	0,00	0,00	0,00	0,01	0,01	0,02	-0,07	-0,06	-0,06
SOL2	0,02	1,00	0,00	-0,04	-0,04	0,01	0,00	-0,14	-0,42	0,42	0,48	-0,01	0,11	-0,18	-0,10	-0,12	0,07	0,00	0,00	0,56	0,00	-0,09	-0,03	-0,21	-0,23	-0,26
SOL3	0,00	0,00	1,00	-0,20	-0,05	0,20	0,22	-0,05	-0,10	0,10	-0,02	0,00	-0,04	-0,08	-0,07	-0,03	0,01	0,00	-0,04	0,05	-0,04	-0,08	-0,05	-0,10	-0,28	-0,11
SOL4	0,02	-0,04	-0,20	1,00	0,67	0,08	-0,11	0,47	-0,06	0,06	0,01	0,03	-0,17	0,06	-0,11	0,14	0,09	-0,02	0,01	-0,04	0,16	0,06	0,36	-0,24	0,07	-0,22
SOL5	0,02	-0,04	-0,05	0,67	1,00	0,04	0,11	0,82	-0,12	0,12	0,11	0,06	-0,16	-0,06	-0,12	0,02	0,04	-0,03	0,00	0,00	0,25	-0,09	0,55	-0,32	-0,12	-0,34
EFG1	0,00	0,01	0,20	0,08	0,04	1,00	-0,06	0,00	-0,04	0,04	-0,01	0,00	-0,04	-0,11	0,01	-0,12	-0,10	-0,01	0,13	0,11	0,07	0,06	0,09	-0,03	-0,03	-0,04
EFG2	0,03	0,00	0,22	-0,11	0,11	-0,06	1,00	0,04	-0,30	0,30	0,32	0,05	-0,11	0,30	-0,12	0,40	0,46	0,01	-0,61	-0,36	0,03	-0,29	-0,01	-0,44	-0,45	-0,37
CAA1	0,01	-0,14	-0,05	0,47	0,82	0,00	0,04	1,00	0,27	-0,27	-0,22	0,04	-0,18	0,02	0,15	-0,10	-0,06	-0,04	0,08	0,04	0,41	0,04	0,58	-0,13	0,01	-0,11
CAA2	-0,04	-0,42	-0,10	-0,06	-0,12	-0,04	-0,30	0,27	1,00	-1,00	-0,81	0,04	-0,18	0,10	0,22	-0,06	-0,26	-0,09	0,26	0,11	-0,01	0,45	0,06	0,44	0,42	0,53
CAA3	0,04	0,42	0,10	0,06	0,12	0,04	0,30	-0,27	-1,00	1,00	0,81	-0,04	0,18	-0,10	-0,22	0,06	0,26	0,09	-0,26	-0,11	0,01	-0,45	-0,06	-0,44	-0,42	-0,53
CR1	0,02	0,48	-0,02	0,01	0,11	-0,01	0,32	-0,22	-0,81	0,81	1,00	-0,05	0,16	-0,15	-0,21	0,00	0,44	0,12	-0,30	-0,12	0,00	-0,45	-0,09	-0,41	-0,40	-0,50
CR2	0,00	-0,01	0,00	0,03	0,06	0,00	0,05	0,04	0,04	-0,04	-0,05	1,00	-0,04	0,02	-0,05	0,06	-0,02	-0,01	0,00	0,01	-0,01	0,01	0,03	-0,02	-0,01	-0,01
CR3	-0,01	0,11	-0,04	-0,17	-0,16	-0,04	-0,11	-0,18	-0,18	0,18	0,16	-0,04	1,00	-0,07	0,24	-0,25	0,04	-0,04	-0,22	-0,19	-0,04	-0,06	-0,14	0,14	0,06	0,06
CR4	0,01	-0,18	-0,08	0,06	-0,06	-0,11	0,30	0,02	0,10	-0,10	-0,15	0,02	-0,07	1,00	0,42	0,75	0,14	-0,01	-0,26	-0,12	0,09	-0,04	-0,13	0,13	0,16	0,30
CR5	0,00	-0,10	-0,07	-0,11	-0,12	0,01	-0,12	0,15	0,22	-0,22	-0,21	-0,05	0,24	0,42	1,00	-0,29	-0,09	-0,07	0,01	-0,03	0,38	0,18	0,08	0,31	0,27	0,38
CR6	0,01	-0,12	-0,03	0,14	0,02	-0,12	0,40	-0,10	-0,06	0,06	0,00	0,06	-0,25	0,75	-0,29	1,00	0,21	0,04	-0,28	-0,10	-0,19	-0,17	-0,19	-0,09	-0,03	0,03
CR7	0,00	0,07	0,01	0,09	0,04	-0,10	0,46	-0,06	-0,26	0,26	0,44	-0,02	0,04	0,14	-0,09	0,21	1,00	0,11	-0,63	-0,45	-0,03	-0,65	-0,22	-0,15	-0,13	-0,16
CR8	0,00	0,00	0,00	-0,02	-0,03	-0,01	0,01	-0,04	-0,09	0,09	0,12	-0,01	-0,04	-0,01	-0,07	0,04	0,11	1,00	-0,01	-0,07	-0,03	-0,14	-0,05	-0,04	-0,05	-0,05
EFF1	0,00	0,00	-0,04	0,01	0,00	0,13	-0,61	0,08	0,26	-0,26	-0,30	0,00	-0,22	-0,26	0,01	-0,28	-0,63	-0,01	1,00	0,40	0,04	0,24	0,08	0,12	0,12	0,14
EFF2	0,00	0,56	0,05	-0,04	0,00	0,11	-0,36	0,04	0,11	-0,11	-0,12	0,01	-0,19	-0,12	-0,03	-0,10	-0,45	-0,07	0,40	1,00	0,01	0,30	0,08	0,03	0,00	0,01
LIQ1	0,01	0,00	-0,04	0,16	0,25	0,07	0,03	0,41	-0,01	0,01	0,00	-0,01	-0,04	0,09	0,38	-0,19	-0,03	-0,03	0,04	0,01	1,00	-0,08	0,46	-0,19	-0,13	-0,17
LIQ2	0,01	-0,09	-0,08	0,06	-0,09	0,06	-0,29	0,04	0,45	-0,45	-0,45	0,01	-0,06	-0,04	0,18	-0,17	-0,65	-0,14	0,24	0,30	-0,08	1,00	0,29	0,15	0,16	0,21
LIQ3	0,02	-0,03	-0,05	0,36	0,55	0,09	-0,01	0,58	0,06	-0,06	-0,09	0,03	-0,14	-0,13	0,08	-0,19	-0,22	-0,05	0,08	0,08	0,46	0,29	1,00	-0,20	-0,08	-0,21
T1	-0,07	-0,21	-0,10	-0,24	-0,32	-0,03	-0,44	-0,13	0,44	-0,44	-0,41	-0,02	0,14	0,13	0,31	-0,09	-0,15	-0,04	0,12	0,03	-0,19	0,15	-0,20	1,00	0,91	0,95
T2	-0,06	-0,23	-0,28	0,07	-0,12	-0,03	-0,45	0,01	0,42	-0,42	-0,40	-0,01	0,06	0,16	0,27	-0,03	-0,13	-0,05	0,12	0,00	-0,13	0,16	-0,08	0,91	1,00	0,88
T3	-0,06	-0,26	-0,11	-0,22	-0,34	-0,04	-0,37	-0,11	0,53	-0,53	-0,50	-0,01	0,06	0,30	0,38	0,03	-0,16	-0,05	0,14	0,01	-0,17	0,21	-0,21	0,95	0,88	1,00

Fuente: Elaboración propia

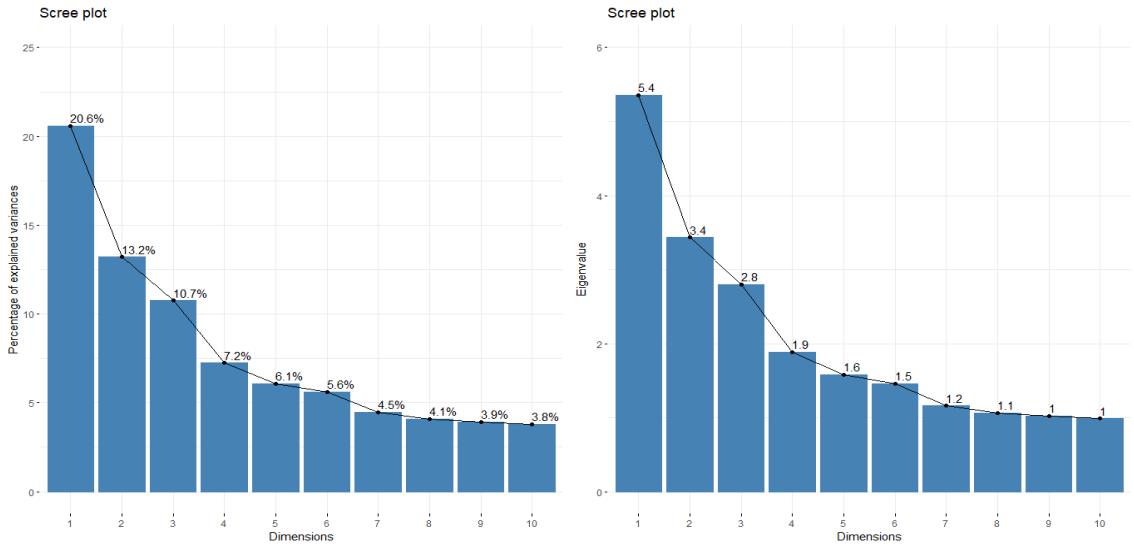
4.3 Análisis de componentes principales

Según Jolliffe (2002) el Análisis de Componentes Principales es una de las técnicas multivariantes más conocida y eficaz para reducir la dimensión de una matriz de datos de alta dimensión. Esto debido a que las componentes principales se obtienen como una combinación lineal de las variables originales. Según González et al. (2015) el análisis de componentes se define como una técnica para aprovechar las relaciones existentes entre diferentes variables, para poder tratar el problema de la correlación de variables independientes, evitando perder información y pudiendo de esta manera explicar mejor la variabilidad de los datos. El concepto de mayor cantidad de información se relaciona con el de mayor varianza. Lo que quiere decir que cuanto mayor sea la variabilidad de los datos, se considera que existe una mayor cantidad de información.

En este caso, el objetivo será del número total de variables independientes (26) obtener un número de componentes menor, el cual abarque la mayor cantidad de información. Para lograrlo, se realizó un análisis de componentes principales mediante el programa RStudio, teniendo como método de selección a los componentes que su autovalor sea

mayor o igual a 1 y que el conjunto de componentes elegidos expliquen más del 70% de la varianza de los datos. Como resultado, se obtuvo un total de 10 componentes los cuales explican el 78.10% de la varianza como puede verse con mayor claridad en el Gráfico 4.

Gráfico 4: PCA Varianza explicada y autovalores



Fuente: Elaboración propia

A la hora de realizar un análisis de componentes principales uno de los objetivos es ver qué variables son las que más influyen en cada componente. La composición de cada componente puede ayudarnos a nombrar cada dimensión del concepto que estamos creando. Esto nos ayudara para entender con mayor claridad las variables que son determinantes a la hora de clasificar a las entidades como rentables o no rentables.

Para explicar el agrupamiento de cada dimensión con cada variable utilizada se procedió a realizar una rotación de los factores con el método varimax. El objetivo de la rotación es obtener una solución más interpretable. De este modo, cada dimensión debe tener unos pocos pesos altos y otros que se aproximen a cero, de tal manera que se pueda identificar qué variable pertenece a cada componente. La ventaja de utilizar el método es que hace que cada componente mantenga correlaciones altas con pocas variables y bajas correlaciones con el resto de las variables. La siguiente tabla indica la agrupación de las variables en cada componente y en el apartado de anexos se puede verificar la matriz de componentes rotada (Tabla 24).

Tabla 8 : Resumen componentes

COMPONENTE	IDENTIFICACIÓN	VARIABLES	VARIANZA EXPLICADA
COMP.1	Vulnerabilidad	SOL2; CAA2; CAA3; CR1	19,85%
COMP.2	Tamaño	T1; T2; T3	12,97%
COMP.3	Solvencia	SOL4; SOL5; CAA1; LIQ3	10,40%
COMP.4	Eficiencia Financiera	EFG2; CR7; EFF1; EFF2; LIQ2	6,97%
COMP.5	Crédito 1	CR4; CR6	6,06%
COMP.6	Liquidez	CR5; LIQ1	5,75%
COMP.7	Gestión	SOL3; EFG1	4,47%
COMP.8	Crédito 2	CR3; CR8;	4,15%
COMP.9	Cobertura de cartera	CR2	3,85%
COMP.10	Suficiencia patrimonial	SOL1	3,66%

Fuente: Elaboración propia

4.4 Árbol de clasificación

El árbol utilizado en el presente estudio es el de "Classification and Regression Tree" (CART) desarrollado por Breiman et al (1984). Es un tipo de algoritmo de aprendizaje supervisado, principalmente usado en problemas de clasificación. Las variables de entrada y salida en el modelo pueden ser categóricas o continuas. Uno de sus objetivos es dividir el espacio de predictores en regiones distintas y no superpuestas para poder crear un modelo de clasificación con diferentes nodos.

Un árbol de clasificación consta de tres tipos de nodos: raíz, interno y terminal. El nodo principal es llamado raíz y contiene a todas las observaciones. A partir de este se pueden dividir en dos ramas, las que pueden ser un nodo interno (si se sigue subdividiendo) o terminal (si ya no tiene más subdivisiones). Cada nodo viene descrito por el subconjunto de la muestra que contiene. Por lo que el nodo busca el mejor valor de predicción para cada posible combinación de características y se queda con aquellos que producen el menor grado de diversidad. Según Breiman et al. (1984), la diversidad de un nodo está en relación con el valor de la función de impureza en el mismo. Por lo que, de acuerdo con lo anterior, se puede definir varias funciones de impureza, dentro de las cuales la más utilizada es la de Gini.

Se introduce el índice de impureza de un determinado grupo como la suma de la probabilidad de la coocurrencia de dos valores distintos. El índice de Gini se generaliza dentro de un grupo t mediante la siguiente fórmula:

$$i(t) = \sum_{i\neq 1}^{J(J-1)} p(j)p(i) = 1 - \sum_{j=1}^{J} p(j)^2$$

Ecuación (1)

Dado que CART en el caso de clasificación propone una clasificación binaria, el criterio para la elección de un modelo será la de hallar aquella división de categorías, dentro de cada variable utilizada, donde se produzca el mayor índice de Gini. Para determinar la variable que habrá de producir el nuevo nodo se compararan los índices de Gini obtenidos en la división de todas las variables del modelo y, aquella que obtenga el mayor índice, deberá producir la siguiente segmentación o nuevo nodo en el árbol (Breiman, 1984).

Con la segmentación producida en el nodo anterior se debe repetir, incluyendo todas las variables independientes. Esto permite que una misma variable pueda producir distintas segmentaciones de forma sucesiva, siempre que cumpla con el criterio de obtener una mejora del índice. Este procedimiento se deberá iterar sucesivamente, obteniendo nuevos nodos, hasta que se llegue a un número determinado de nodos donde el error de clasificación sea mínimo. Uno de los problemas de crear un árbol de clasificación será el sobreajuste del modelo. Para poder evitar este sobreajuste se necesita que el algoritmo limite el crecimiento del número de nodos hasta que el mismo sea óptimo (Breiman, 1984).

4.5 Random Forest

Breiman (2001) propone una manera de mejorar las predicciones y el sobreajuste que en ciertos casos puede producir el árbol de clasificación. Propone usar “ensemble methods” denominados de esta manera porque a partir de los datos del árbol de clasificación se crean nuevas submuestras y su eligen nuevos algoritmos. Dentro de los modelos de ensemble se encuentran los algoritmos “Bagging” en los que se ajustan múltiples modelos, cada uno con un subconjunto distinto de los datos de entrenamiento. Para predecir todos los modelos que forman el modelo general, cada

uno participa aportando su predicción. Por lo que, como valor final, se toma la media de todas las predicciones (regresión) o la clase más frecuente (clasificación).

El método que utiliza el algoritmo de Random forest para realizar el re-muestreo se denomina Bootstrap. Es una técnica que fue propuesta por Efron (1979) en donde se puede inferir una determinada distribución a través de muestras repetidas extraídas de la propia muestra, con reposición. Entonces, lo que hace el algoritmo de Random forest es realizar un modelo para la muestra de entrenamiento con el subconjunto de datos que seleccionó el Bootstrap.

Según Cutler et al. (2011) el Random forest será un conjunto de árboles de clasificación en donde cada árbol depende de un conjunto de variables aleatorias. De manera más específica, para un vector aleatorio de p-dimensiones denominado $X = (X1, \ldots\ldots, Xp)^T$ el cual representa los predictores y existe una variable dependiente denominada Y; se asume una distribución conjunta que es desconocida como: $P_{xy}(X, Y)$.

El objetivo es encontrar una función de predicción $f(x)$ que pueda predecir de manera correcta a Y,donde la función de predicción será determinada por una función de pérdida $L\,(Y, f(x))$ mediante la cual se busca minimizar el valor esperado de dicha función de pérdida $E_{xy}(L\,(Y, f(x)))$). Para el caso de modelos de clasificación y, teniendo en cuenta de que será una función dicotómica, mediante la minimización se obtendrá:

$$f(x) = argmax\ P(Y = y|X = x)$$

Ecuación (2)

4.6 Gradient Boosting Machine

El algoritmo de Gradient Boosting fue propuesto por Friedman (2001) y (2002) y, al igual que el random forest, pertenece a los metodos de ensemble. Sin embargo, en este caso se ajustan secuencialmente múltiples modelos sencillos, llamados *weak learners*, de forma que cada modelo aprende de los errores del anterior. En el caso de Gradient Boosting, los weak learners se consiguen utilizando árboles de decisión con un número mínimo de ramificaciones. Como valor final, al igual que en el Random forest, se toma la media de todas las predicciones para el caso de regresión o la clase más frecuente

para el caso de clasificación. De acuerdo con Friedman et al. (2009) para la obtención del algoritmo, se comienza ajustando un weak learner denominado f_1 con el que inicialmente se podrá predecir la variable respuesta Y. Una vez obtenido la primera estimación se obtienen los errores de la predicción: $Y - f_1(x)$. A partir de esto se genera un segundo weak learner f_2 que intentará predecir los errores del modelo f_1 , y así sucesivamente, obteniendo lo siguiente:

$$f_1(x) \approx y$$
$$f_2(x) \approx y - f_1(x)$$
$$f_3(x) \approx y - f_1(x) - f_2(x)$$

Ecuación (3)

Este proceso se repetirá "m" veces de tal manera que cada nuevo weak learner intentará minimizar los errores de los modelos anteriores.

5. Resultados ROE

En el siguiente apartado se detallan los resultados obtenidos para ver qué modelo, Árbol de clasificación, Random Forest o Gradient Boosting Machine, clasifica de mejor manera las COAC en Ecuador, utilizando al ROE como variable dependiente. Los resultados fueron obtenidos mediante el programa Rstudio.

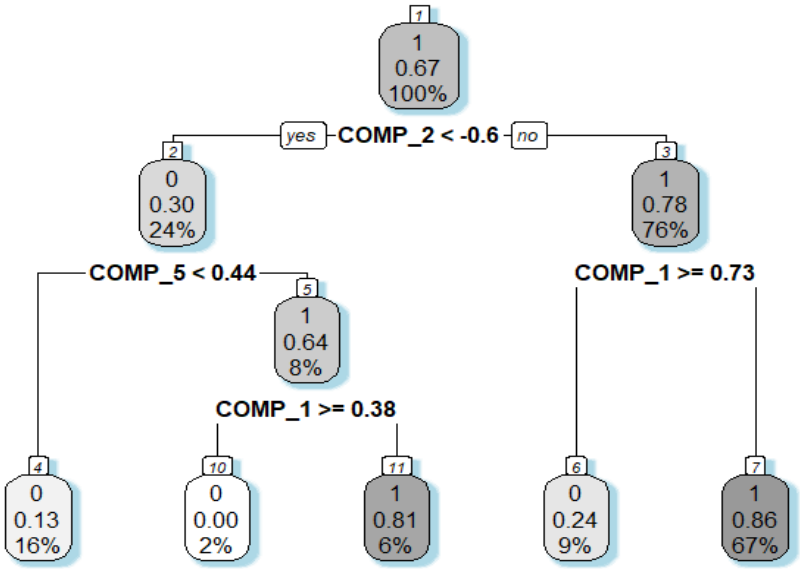
Para poder realizar los diferentes algoritmos de Machine Learning, se dividió la base de datos total en dos submuestras: Una muestra de entrenamiento utilizada para desarrollar el algoritmo y una muestra para realizar validaciones del modelo creado. Del total de la base datos, de manera aleatoria se asignó el 80% de la base total al "Training Data" y el 20% al "Test data".

5.1 Árbol de clasificación ROE

Para obtener un modelo óptimo mediante el árbol de clasificación CART para el año 2020 se realizó un árbol de clasificación sin especificar ningún parámetro extra para luego realizar una optimización mediante una "poda" y así obtener un modelo sin sobreajuste con una mejor clasificación.

Para el caso del ROE, el árbol inicial utilizó 4 componentes para explicar la rentabilidad de las COAC dando como resultado una precisión en el ajuste de los datos del 85,94%. Por otro lado, al realizar la optimización del árbol de clasificación mediante el ajuste de los parámetros, se realizó por medio del número de nodos que minimiza el error de predicción, obteniéndose un árbol que utiliza 3 componentes para explicar al ROE como variable dependiente y, en este caso con un nivel de predicción del 84,38%. Pese a que el modelo inicial tiene mayor precisión se decidió mantener el modelo óptimo por su mayor interpretación. Por lo tanto, el modelo que evita el sobreajuste en la clasificación será el modelo una vez realizada la "poda". Cabe mencionar que en el apartado de anexos se encuentra detallado el código R para su obtención.

Gráfico 5: Árbol de clasificación ROE



Fuente: Elaboración propia

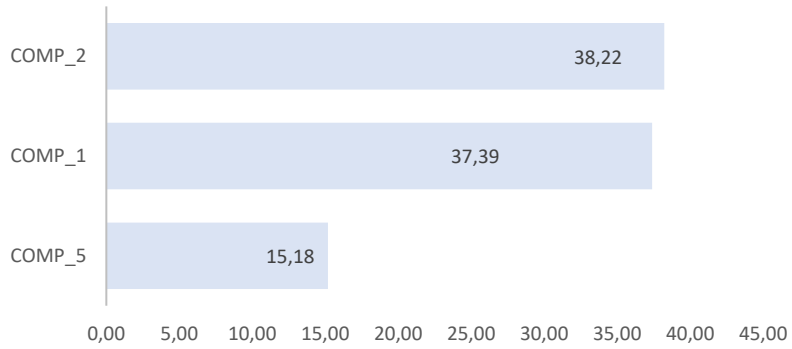
El árbol de clasificación presentado en el gráfico 5, presenta un nodo raíz identificado con la variable COMP_2. Además, existen nodos intermedios identificados con las variables COMP_5 y COMP_1. Finalmente, existen 5 nodos terminales en donde se pueden identificar la clasificación final de las COAC en rentables o no rentables. Las reglas son las siguientes: para el caso que se quiera identificar cuando será rentable una entidad a través del nodo terminal 11 y del 7. Con respecto al nodo terminal 7, se indica que el 67% de COACS son clasificadas como rentables si cumplen que la COMP_2 >= -0.5993 & COMP_1 < 0.7269. Por otro lado, el nodo terminal 11 indica que el 6 % de las

entidades son clasificadas como rentables si cumplen que COMP_2 < 0.6 & COMP_5>=0.444 & COMP_1< 0.3791.

Para el caso en que las COACS no sean rentables, se puede observar su clasificación en los nodos terminales 4, 6 y 10. Con respecto al nodo terminal 4 se indica que el 16 % de COACS son clasificadas como no rentables si cumplen que la COMP_2 < -0.5993 & COMP_5 < 0.444. Con respecto al nodo terminal 6, este indica que el 9% de entidades son clasificadas como no rentables si cumplen que COMP_2 >= -0.5993 & COMP_1 >= 0.7269. Finalmente, el nodo terminal 10 señala que un 2% de entidades son clasificadas como no rentables cuando COMP_2 < -0.5993 & COMP_5 >= 0.444 & COMP_1 >= 0.3791.

Uno de los objetivos del trabajo es ver que variables son las que más influyen en el sector de las COAC en Ecuador, para lo mismo en el Gráfico 6 se observa que, para el modelo de árbol de decisión, la variable formada por la componente 2 (Tamaño) es la que mayor significancia tiene a la hora de clasificar el modelo.

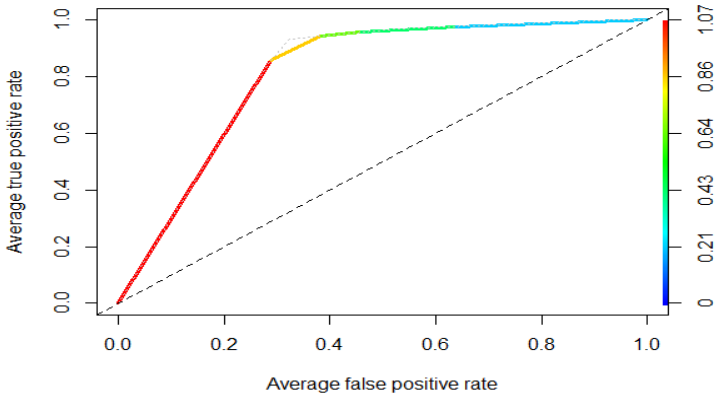
Gráfico 6: Variables relevantes para clasificar ROE según CART



Fuente: Elaboración propia

Adicionalmente, para medir el ajuste global del modelo se utilizó a la curva ROC. De acuerdo con Fawcett (2006), la curva ROC es un gráfico bidimensional en donde se señala la relación entre la especificidad obtenida en la matriz de confusión y la sensibilidad. Una forma de analizar la curva ROC o la mayor exactitud para clasificar será cuando la curva se desplace "hacia arriba y a la izquierda". Es decir que mientras más se acerque la curva al límite superior izquierdo, mejor ajuste tendrá el modelo.

Gráfico 7: ROC Curve Árbol de clasificación ROE

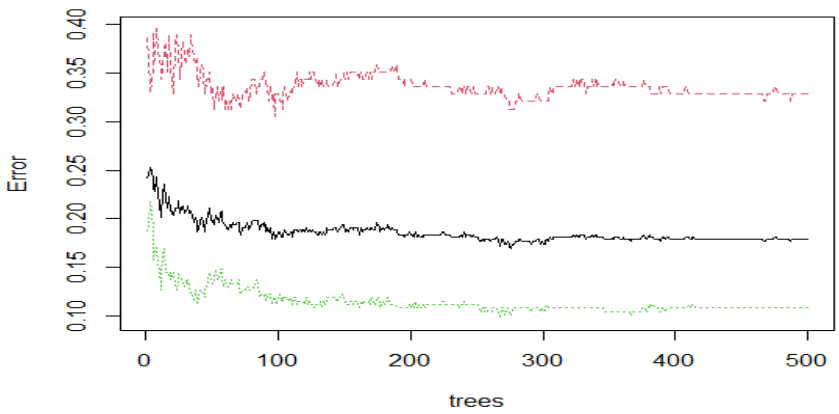


Fuente: Elaboración propia

5.2 Random Forest ROE

Como se mencionó el apartado de la metodología, el Random forest se construye a partir de diferentes arboles de decisión mediante submuestras realizadas por Bootstraping. Así, para la obtención del Random Forest para el ROE se realizó un modelo base y luego se ajustó este modelo buscando los parámetros óptimos (número de árboles y número de variables utilizadas en cada división) tratando de obtener una mejor clasificación y minimizar el error. El modelo inicial contó con 500 árboles de clasificación y 3 variables en cada división, obteniendo una precisión en el ajuste de los datos del 86,36%. Para elegir los parámetros óptimos, se puede observar en el gráfico 8 que a partir de los 300 árboles el modelo comienza a estabilizarse. En consecuencia, se eligió el número mínimo de variables en cada división en el cual se minimizaría el error de clasificación, dando como resultado 3 variables. Cabe mencionar que la obtención de este modelo en R puede verificarse en el apartado de anexos.

Gráfico 8: Número de árboles óptimo Random Forest

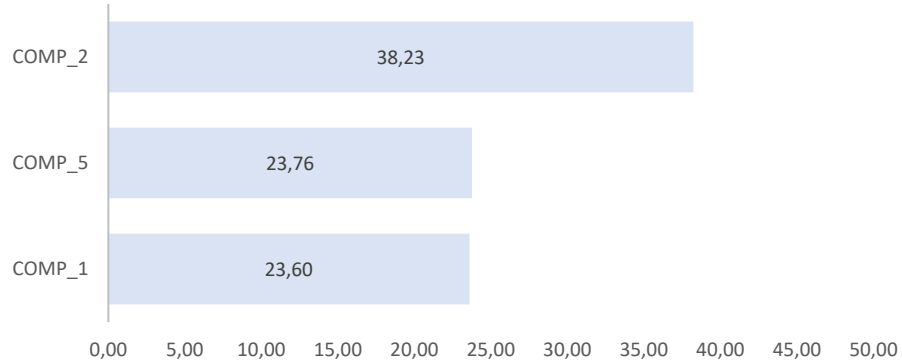


Fuente: Elaboración propia

En este caso al realizar un nuevo entrenamiento de datos con los parámetros óptimos encontrados se obtuvo una precisión en el ajuste de los datos del 87,88%. La interpretación de los resultados del Random Forest se realiza mediante la importancia que tienen las variables a la hora de clasificar y también mediante estadísticos que muestran la bondad de la clasificación del modelo.

La importancia de las variables estará ligada al error del modelo por lo que la importancia de las variables se da en el sentido de cuanto afectaría al ajuste del modelo si una variable no es incluida. En el gráfico 9 se puede observar que se da más importancia al componente 2 formado por las variables de tamaño: logaritmo de los ingresos por intereses; logaritmo del patrimonio y logaritmo de los activos. En los puestos 2 y 3 se encuentran las componentes 5 y 1 formadas por variables referentes a crédito y vulnerabilidad, respectivamente.

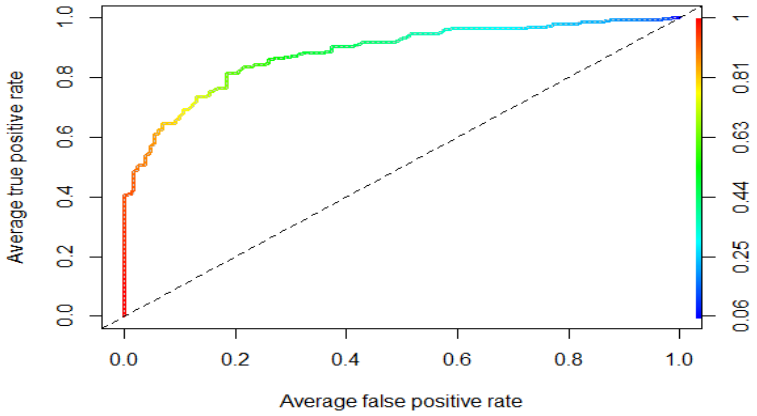
Gráfico 9: Variables relevantes para clasificar ROE según RANDOM FOREST



Fuente: Elaboración propia

En este caso, en el grafico 10 se expresa la curva ROC obtenida para el modelo óptimo, en la cual se puede apreciar que se acerca al límite superior izquierdo por lo cual tendría un buen nivel de ajuste. Cabe mencionar que en los siguientes apartados se tomara en cuenta el índice del área bajo la curva para un análisis más específico.

Gráfico 10: ROC Curve Random Forest ROE

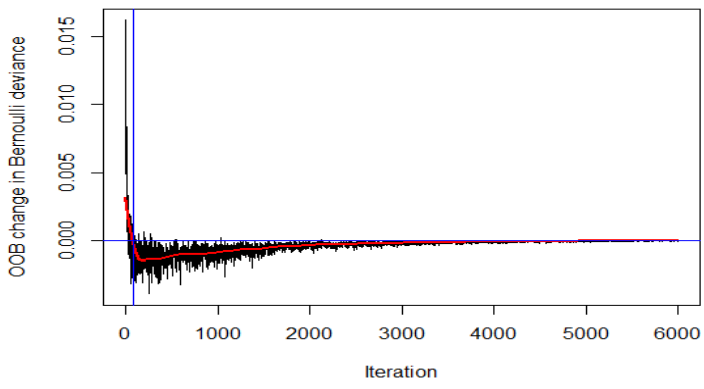


Fuente: Elaboración propia

5.3 Gradient Boosting Machine ROE

La idea de un GBM es entrenar diferentes modelos de forma secuencial, de tal manera que cada modelo ajuste los errores de los modelos anteriores. En este caso para la modelización del GBM se inició calculando un modelo base con 6000 árboles para que a partir de este modelo poder realizar una ajuste de los hiperparámetros y así obtener un número opimo de árboles que minimicen el error de estimación. El modelo inicial dio como resultado una precisión en el ajuste de los datos del 88,71%. Pese a que el modelo a primera vista ajusta muy bien los datos, se realizó la optimización mencionada anteriormente. En el gráfico 11 se puede observar el proceso de iteración o de "parada temprana" que permitirá saber cuántas iteraciones o árboles se necesitaran en el modelo hasta el punto de que el error de validación se reduzca y comience a estabilizarse.

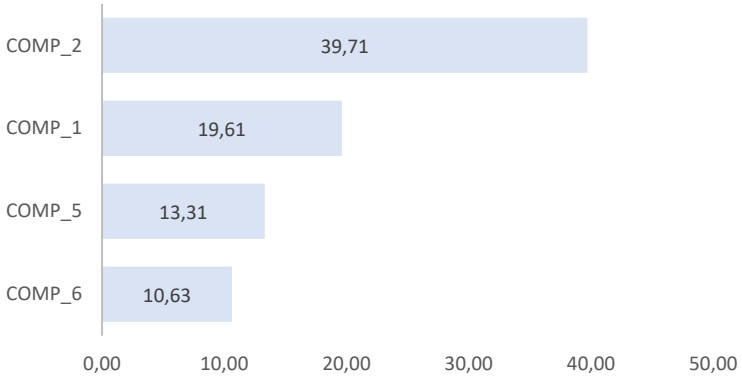
Gráfico 11: Modelo óptimo GBM - ROE



Fuente: Elaboración propia

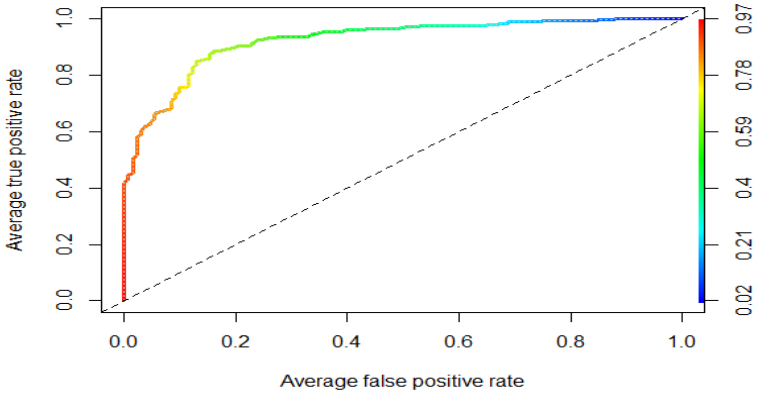
De acuerdo lo identificado, se requerirán 78 iteraciones para obtener un modelo optimo, en el cual se obtuvo una precisión en el ajuste de los datos del 95,16%. Cabe mencionar que el código de la modelización del GBM se encuentra especificado en el apartado de anexos. En el algoritmo GBM se incluyen la importancia de las variables como medida de interpretación del modelo. En este caso la medida de importancia se basará en la mejora del modelo al incluir o eliminar las variables. Al igual que los dos modelos anteriores la variable que más influye en el modelo (Gráfico 12) es la componente número 2 (Tamaño), seguido de la componente 1 y la componente 5. Cabe destacar que el modelo también da cierta importancia a la componente 6 que está formada por variables referentes a liquidez. De la misma manera en el gráfico 13 se observa la curva ROC que, al igual que los modelos anteriores, se implementó para saber el ajuste que tendrá el modelo y más adelante se utilizará el área bajo la curva para obtener un índice más exacto.

Gráfico 12: Variables relevantes para clasificar ROE según GBM



Fuente: Elaboración propia

Gráfico 13: ROC Curve GBM ROE



Fuente: Elaboración propia

6. Resultados Rentabilidad cartera microcrédito

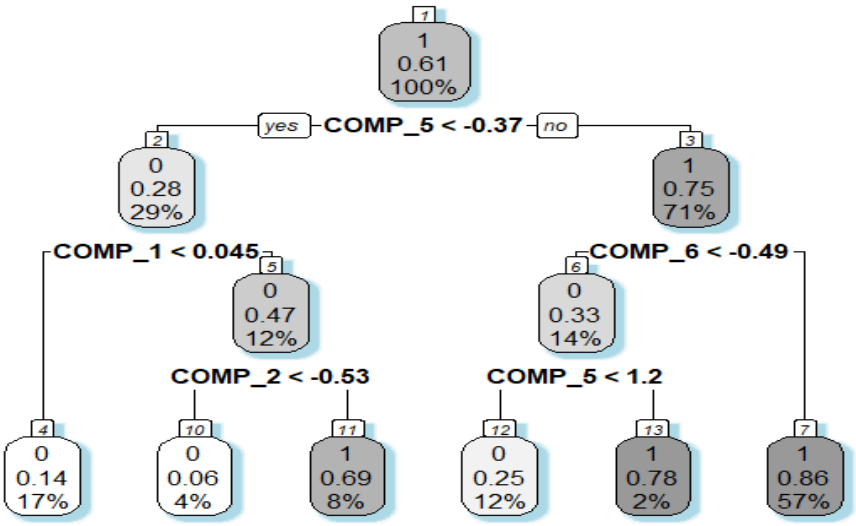
En el siguiente apartado, se detallan los resultados obtenidos para ver qué modelo, Árbol de clasificación, Random Forest o Gradient Boosting Machine, clasifica de mejor manera las COAC en Ecuador, utilizando la RCM como variable dependiente. Al igual que los modelos utilizados para el ROE, también se dividió la base de datos total en dos submuestras: una muestra de entrenamiento utilizada para desarrollar el algoritmo y una muestra para realizar validaciones del modelo creado. Del total de la base datos, de manera aleatoria se asignó el 80% de la base total al "Training Data" y el 20% al "Test data".

6.1 Árbol de clasificación RCM

Para obtener un modelo óptimo mediante un árbol de clasificación se inició realizando un modelo sin especificar ningún parámetro extra para luego realizar una optimización mediante una "poda" y así obtener un modelo sin sobreajuste con una mejor clasificación.

Para el caso del RCM, el árbol inicial utilizó 4 componentes para explicar la rentabilidad de la cartera de microcrédito de las COAC dando como resultado una precisión en el ajuste de los datos del 80%. Por otro lado, al realizar la optimización del árbol de clasificación mediante el ajuste de los parámetros, que en este caso se realizó por medio del número de nodos que minimiza el error de predicción. Se obtuvo un árbol que igualmente utiliza 4 componentes para explicar la rentabilidad de la cartera de microcrédito como variable dependiente, pero en este caso con menos nodos terminales lo que simplificará el análisis. El modelo final obtuvo una predicción del 77%. Por lo cual, el modelo óptimo que evita el sobreajuste en la clasificación será el modelo una vez se haya realizado la "poda" (véase Gráfico 14, y su posterior descripción). Cabe mencionar que en el apartado de anexos se encuentra detallado el primer modelo junto con el código R para su obtención.

Gráfico 14: Árbol de clasificación RCM



Fuente: Elaboración propia

El árbol de clasificación para el RCM presenta un nodo raíz identificado con la variable COMP_5, existen tres nodos intermedios identificados con la variable COMP_1, COMP_6 y COMP_2. Y, existen 6 nodos terminales en donde se pueden identificar la clasificación final de las COAC en rentables o no rentables. A continuación, se explican las reglas obtenidas.

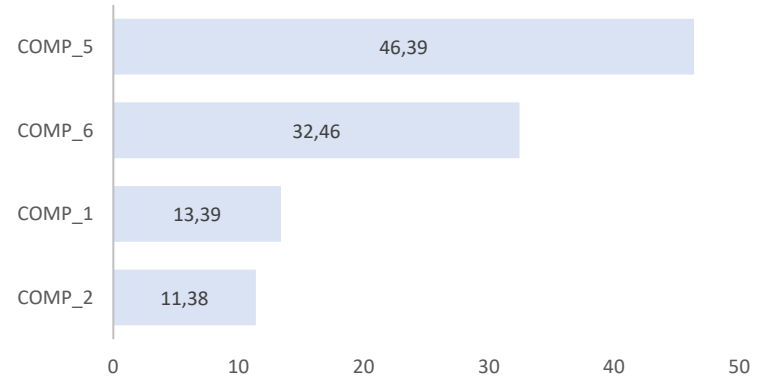
Para identificar cuando será rentable una entidad se han de analizar los nodos terminales 7, 13 y 11. El nodo terminal 7 indica que el 57% de COACS son clasificadas como rentables si cumplen que COMP_5>=-0.3749 & COMP_6>=-0.4909. Con respecto al nodo terminal 13 se indica que existirán un 2% de COAC clasificadas como rentables si COMP_5>=-0.3749 & COMP_6< -0.4909 & COMP_5>=1.205. Finalmente, el nodo terminal 11 indica que el 8% de las entidades serán rentables cuando COMP_5< -0.3749 & COMP_1>=0.04459 & COMP_2>=-0.5312.

Para el caso en que las COACS no sean rentables, se puede observar su clasificación en los nodos terminales 12, 10 y 4. Con respecto al nodo terminal 12 se indica que el 12% de COACS son clasificadas como no rentables si cumplen que COMP_5>=-0.3749 & COMP_6< -0.4909 & COMP_5< 1.205. Con respecto al nodo terminal 10, este indica que el 4% de entidades son clasificadas como no rentables si cumplen que COMP_5< -0.3749

& COMP_1>=0.04459 & COMP_2< -0.5312. Finalmente, el nodo terminal 4 indica que 17% las entidades no serán rentables en el caso que COMP_5< -0.3749 & COMP_1< 0.04459.

En el Gráfico 15 se observa que, para el árbol de decisión, la variable formada por la componente 5 relacionada con variables de crédito (tasa de interés activa implícita y spread de crédito) es la que mayor significancia tiene a la hora de la clasificación y por lo tanto no deberían eliminarse del modelo ya que afectaría en gran medida a su capacidad predictiva. Adicionalmente, la componente 6 (liquidez), la componente 1 y la componente 2 son las que la siguen en importancia para el modelo desarrollado.

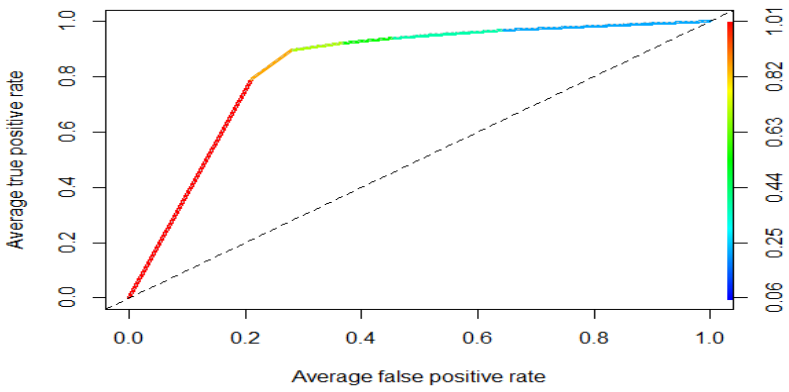
Gráfico 15: Variables relevantes para clasificar RCM según Árbol de clasificación



Fuente: Elaboración propia

De manera similar a los modelos anteriores, se obtuvo la curva ROC para el modelo final del árbol de decisión (véase Gráfico 16).

Gráfico 16: ROC Curve Árbol de clasificación RCM

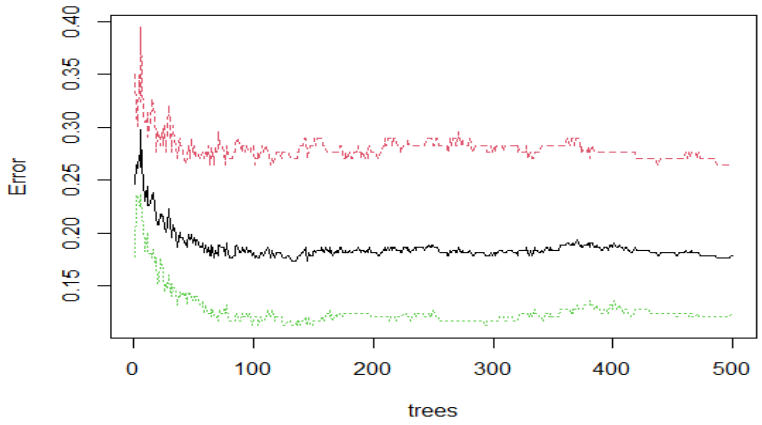


Fuente: Elaboración propia

6.2 Random Forest RCM

Para la obtención del Random Forest para el RCM se realizó un modelo base y luego se ajustó este modelo buscando los parámetros óptimos (número de árboles y numero de variables utilizadas en cada división) obteniendo una mejor clasificación y minimizar el error. El modelo inicial conto con 500 árboles de clasificación y 3 variables en cada división, obteniendo una precisión en el ajuste de los datos del 83%. Para elegir los parámetros óptimos, se puede observar en el gráfico 17 que a partir de los 300 árboles el modelo comienza a estabilizarse y adicional a esto se eligió el número mínimo de variables en cada división en el cual se minimizaría el error de clasificación, dando como resultado 2 variables (véase el apartado ANEXOS para verificar el código R de generación del modelo).

Gráfico 17: Número de árboles óptimo RCM



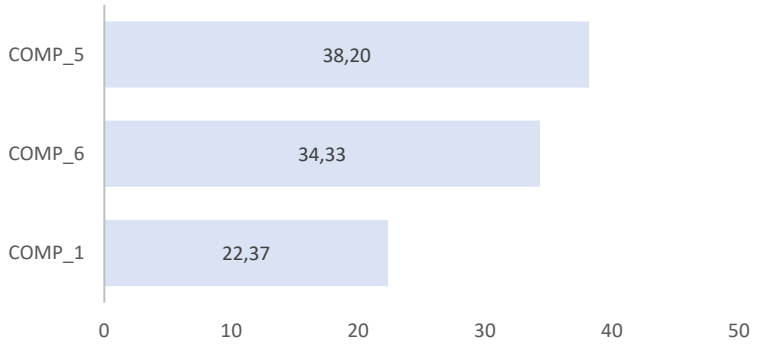
Fuente: Elaboración propia

Al realizar un nuevo entrenamiento de datos con los parámetros óptimos encontrados se obtuvo una precisión en el ajuste de los datos del 85,25%. La interpretación de los resultados del Random Forest se realiza comprobando la importancia que tienen las variables a la hora de clasificar y también mediante la curva ROC (gráfico 19) comprobando su proximidad a la parte superior izquierda lo que permite verificar que el modelo tiene un buen ajuste.

Por otro lado, la importancia de las variables estará ligada al error del modelo. En el gráfico 18 se puede observar que el modelo Random Forest da más importancia al componente 5 formado por las variables de crédito. En los puestos 2 y 3 se encuentran las componentes 6 y 1, formadas por variables referentes a la Liquidez y Vulnerabilidad

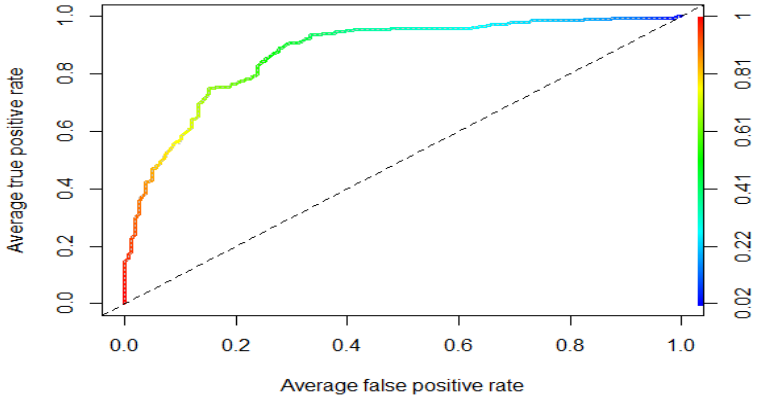
respectivamente.

Gráfico 18: Variables relevantes para clasificar RCM según RANDOM FOREST



Fuente: Elaboración propia

Gráfico 19: ROC Curve Random Forest RCM

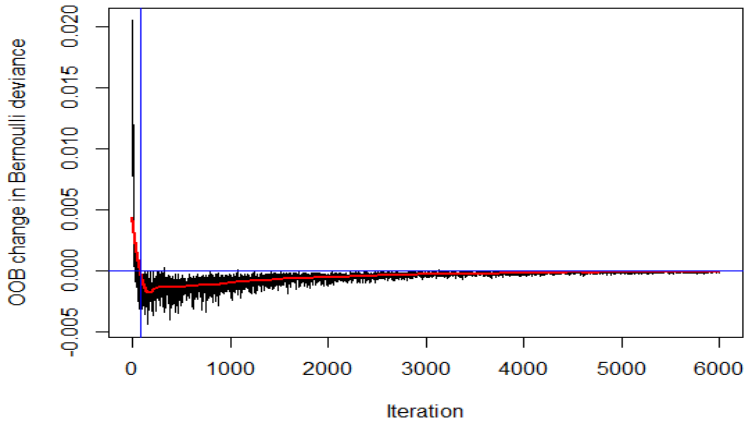


Fuente: Elaboración propia

6.3 Gradient Boosting Machine RCM

Para la modelización del GBM se calculó inicialmente un modelo base con 6000 árboles para que, a partir de este modelo, poder realizar una ajuste de los hiperparámetros y así obtener un número óptimo de árboles que minimicen el error de estimación. El modelo inicial obtuvo una precisión en el ajuste de los datos del 75,81%. Pese a que el modelo a primera vista ajusta muy bien los datos, se realizó la optimización mencionada anteriormente. En el gráfico siguiente (Gráfico 20) se puede observar el proceso de iteración o de "parada temprana" que permitirá saber cuántas iteraciones o árboles se necesitaran en el modelo hasta el punto de que el error de validación se reduzca y comience a estabilizarse.

Gráfico 20: Modelo óptimo GBM - ROE

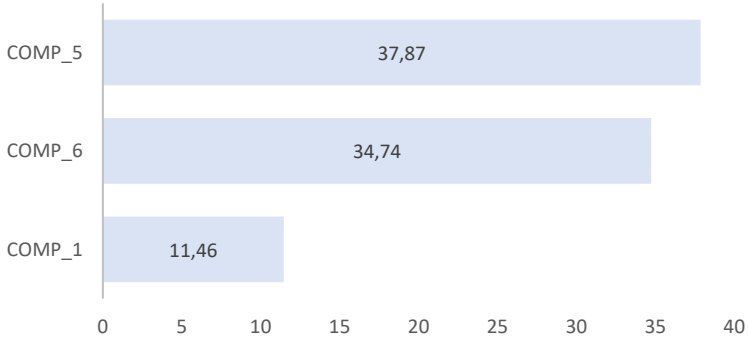


Fuente: Elaboración propia

De acuerdo lo identificado en el gráfico, se requerirán 81 iteraciones para obtener un modelo optimo. Dicho modelo obtuvo una precisión en el ajuste de los datos del 82,26% (el código de la modelización del GBM se encuentra especificado en el apartado de anexos).

En el algoritmo GBM se incluyen la importancia de las variables como medida de interpretación del modelo. En este caso la medida de importancia se puede analizar viendo la mejora del modelo al incluir o eliminar las variables. Al igual que los dos modelos anteriores la variable que más influye en el modelo (Gráfico 21) es la componente número 5 (Tamaño), seguido de la componente 6 y la componente 1.

Gráfico 21: Variables relevantes para clasificar RCM según GBM

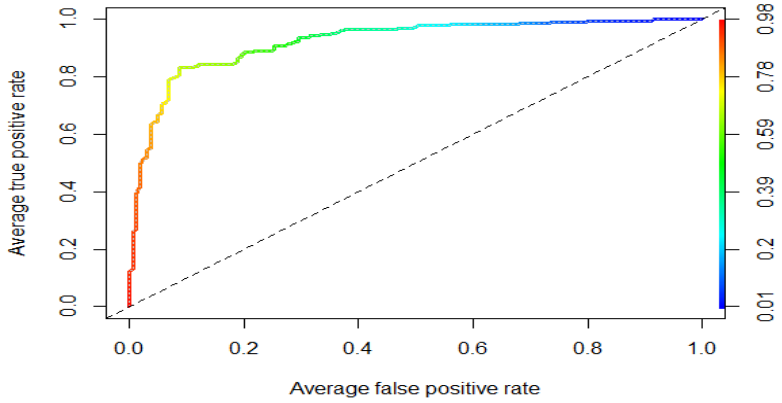


Fuente: Elaboración propia

De la misma manera al igual que los modelos anteriores, se implementó un gráfico donde se muestra la curva ROC para poder ver si el modelo ajusta bien los datos. Se

evidencia que la curva se aproxima a la esquina superior izquierda por lo que se obtiene un buen ajuste global. En los siguientes apartados se calculó el área bajo la curva para poder obtener de manera más exacta el nivel de precisión del modelo (Tabla 13).

Gráfico 22: ROC Curve GBM RCM



Fuente: Elaboración propia

Como conclusión del presente apartado, se puede mencionar que de los 3 modelos analizados para el caso del ROE dan una mayor importancia a la componente 2, la cual contiene a variables de tamaño como logaritmo del patrimonio, logaritmo de los activos y el volumen de negocio (logaritmo de ingresos por intereses). Una justificación para ver la relación directa entre la rentabilidad y el tamaño de una entidad se ve en la medida que las entidades grandes podrían tener a su disposición mayores oportunidades técnicas del mercado y se podrían beneficiar de las economías de escala. Además, tienen mayor poder de negociación ante clientes y tienen mejor perspectivas de financiamiento. (Baumol, 1982)

Por parte de los modelos relacionados a la RCM se verificó que en los 3 modelos la variable que más importancia tiene en los modelos es la componente 5 formada por variables de crédito: tasa de interés activa implícita y el spread de crédito. Como justificación del resultado obtenido se puede decir que las COAC al estar su cartera principalmente compuesta de microcrédito, deberán tener una tasa de interés que genere un cierto nivel de rentabilidad y que permita más captaciones para la entidad y, de esta forma, mediante la intermediación financiera puedan generar más recursos. De la misma manera la diferencia entre la tasa activa y pasiva que mantengan las COAC

deben estar acorde con la generación de rentabilidad para que las entidades tengan una buena salud financiera.

7. Comparación de modelos

7.1 Comparación de modelos ROE

Una vez implementado los diferentes algoritmos de aprendizaje es necesario analizar la capacidad de predicción de los modelos realizados mediante diferentes métricas. Las métricas más frecuentes que han sido utilizadas indistintamente por investigadores en estudios para evaluar modelos de clasificación son: Matriz de confusión, accuracy (precisión), recall (sensibilidad), curva ROC y AUC. (Catal, 2012)

Con respecto a la matriz de confusión, se trata de una tabla de frecuencias donde las filas pertenecen a la clase predicha y las columnas a la clase verdadera, representando el número de predicciones de cada clase mutuamente excluyentes. En las tablas 9, 10 y 11 se puede observar la matriz de confusión para cada modelo realizado tomando en cuenta al ROE como variable dependiente. Para los 3 modelos entrenados en el caso del modelo inicial y el modelo con parámetros óptimos, el algoritmo presenta errores de tipo 1 clasificando a las COAC como rentables cuando en realidad no lo son. Sin embargo, se observa que este error es mucho menor para el Random Forest.

Tabla 9 : Matriz de confusión árbol de clasificación

MODELO INICIAL ROE			
Clase	**No rentable**	**Rentable**	**Total Clases**
No rentable	23	8	31
Rentable	9	55	64
			95

MODELO FINAL ROE (HIPERPARÁMETROS)			
Clase	**No rentable**	**Rentable**	**Total Clases**
No rentable	22	9	31
Rentable	10	54	64
			95

Fuente: Elaboración propia

Tabla 10 : Matriz de confusión Random Forest

MODELO INICIAL ROE			
Clase	**No rentable**	**Rentable**	**Total Clases**
No rentable	31	5	36
Rentable	9	57	66
			102

MODELO FINAL ROE (HIPERPARÁMETROS)			
Clase	**No rentable**	**Rentable**	**Total Clases**
No rentable	32	4	36
Rentable	8	58	66
			102

Fuente: Elaboración propia

Tabla 11 : Matriz de confusión GBM

MODELO INICIAL ROE			
Clase	**No rentable**	**Rentable**	**Total Clases**
No rentable	29	11	40
Rentable	7	55	62
			102

MODELO FINAL ROE (HIPERPARÁMETROS)			
Clase	**No rentable**	**Rentable**	**Total Clases**
No rentable	29	11	40
Rentable	3	59	62
			102

Fuente: Elaboración propia

A partir de las matrices de confusión descritas anteriormente se pueden obtener diferentes métricas como la precisión que según Gu et al. (2009) es una de las métricas más utilizadas por su comprensión para evaluar la efectividad general del algoritmo. Otra medida será la sensibilidad que es una medida que da a conocer la porción de casos positivos que fueron correctamente clasificados. Y, la especificidad que es una medida que da a conocer la porción de casos negativos que fueron clasificados correctamente (Drzewiecki, 2017).

Tabla 12 : Métricas clasificación modelos ROE

MODELO	METRICA	MODELO INICIAL	MODELO FINAL
ARBOL DE CLASIFICACIÓN	Especificidad	71,88%	68,75%
	Sensitividad	87,30%	85,71%
	Precisión	85,94%	84,38%
RANDOM FOREST	Especificidad	77,50%	80,00%
	Sensitividad	91,94%	93,55%
	Precisión	86,36%	87,88%
GBM	Especificidad	80,56%	90,63%
	Sensitividad	83,33%	84,29%
	Precisión	88,71%	95,16%

Fuente: Elaboración propia

De acuerdo con los datos presentados en la tabla 12 se puede evidenciar que, para las 3 metodologías utilizadas, el modelo final implementado con los hiperparámetros obtiene la mejor precisión del modelo. De los 3 modelos realizados el modelo GBM es el que mejor ajusta los datos.

Adicionalmente, se puede realizar un análisis para ver que predícela bondad de la predicción del modelo de manera general. Una vez visto la curva ROC en el apartado anterior para cada modelo, se puede calcular a partir de esta el área bajo la curva AUC. De manera general un desplazamiento de la curva ROC "hacia arriba y a la izquierda sugiere un mejor ajuste de un modelo". Por lo cual, el área bajo la curva AUC se puede

emplear como un índice conveniente de la exactitud global de la prueba: la exactitud máxima correspondería a un valor de AUC más cercana de 1 y la mínima a uno de 0.5. La tabla 13 recoge el índice AUC para los modelos finales realizados teniendo al ROE como variable dependiente. Siendo así que el área bajo la curva para el modelo GBM es de 0.9241, afirmando que el modelo que mejor ajusta o predice los datos será el realizado mediante el algoritmo de GBM.

Tabla 13 : Área bajo la curva ROC - ROE

MODELO	MODELO INICIAL
ARBOL DE CLASIFICACION	0,8153
RANDOM FOREST	0,8824
GBM	0,9241

Fuente: Elaboración propia

7.2 Comparación de modelos RCM

Al igual que en el apartado anterior, para la implementación de los modelos tomando en cuenta a la RCM como variable dependiente, se calculó diferentes indicadores para saber que modelo es el que mejor predice los datos. En las siguientes tablas se puede observar la matriz de confusión de los 3 modelos.

Tabla 14 : Matriz de confusión árbol de clasificación

MODELO INICIAL RCM			
Clase	**No rentable**	**Rentable**	**Total Clases**
No rentable	26	7	33
Rentable	12	50	62
			95

MODELO FINAL RCM (HIPERPARÁMETROS)			
Clase	**No rentable**	**Rentable**	**Total Clases**
No rentable	24	6	30
Rentable	14	51	65
			95

Fuente: Elaboración propia

Tabla 15 : Matriz de confusión Random Forest

MODELO INICIAL RCM			
Clase	**No rentable**	**Rentable**	**Total Clases**
No rentable	30	13	43
Rentable	10	49	59
			102

MODELO FINAL RCM (HIPERPARÁMETROS)			
Clase	**No rentable**	**Rentable**	**Total Clases**
No rentable	31	10	41
Rentable	9	52	61
			102

Fuente: Elaboración propia

Tabla 16 : Matriz de confusión GBM

MODELO INICIAL RCM			
Clase	**No rentable**	**Rentable**	**Total Clases**
No rentable	30	10	40
Rentable	15	47	62
			102

MODELO FINAL RCM (HIPERPARÁMETROS)			
Clase	**No rentable**	**Rentable**	**Total Clases**
No rentable	26	14	40
Rentable	11	51	62
			102

Fuente: Elaboración propia

A partir de los 3 modelos realizados se obtuvieron diferentes métricas de clasificación las cuales servirán para poder comparar los modelos. Cabe mencionar que en los 3 modelos finales se puede observar un error de tipo 1 clasificando a las COAC como rentables cuando en realidad no lo son. Adicionalmente, se observa que este error es mucho menor para el Random Forest. Con respecto a la precisión del modelo en la tabla 17se puede observar la comparativa de los 3 modelos, y en este caso se observa que el Random Forest es el que obtendrá un mejor nivel de precisión en el ajuste de los datos.

Tabla 17 : Métricas clasificación modelos ROE

MODELO	METRICA	MODELO INICIAL	MODELO FINAL
ARBOL DE CLASIFICACIÓN	Especificidad	68,42%	63,16%
	Sensitividad	87,72%	89,47%
	Precisión	80,65%	78,46%
RANDOM FOREST	Especificidad	75,00%	77,50%
	Sensitividad	79,03%	83,87%
	Precisión	83,05%	85,25%
GBM	Especificidad	66,67%	70,27%
	Sensitividad	82,46%	78,46%
	Precisión	75,81%	82,26%

Fuente: Elaboración propia

Para obtener una mejor decisión y saber qué modelo es el que mejor ajusta los datos de manera global se calculó el área bajo la curva ROC de los 3 modelos, obteniendo que en este caso el modelo con mejor ajuste será el GBM con un índice del 0.9227.

Tabla 18 : Área bajo la curva ROC - ROE

MODELO	MODELO FINAL
ARBOL DE CLASIFICACION	0,8360
RANDOM FOREST	0,8742
GBM	0,9227

Fuente: Elaboración propia

8. Conclusiones

Las técnicas de aprendizaje automático utilizadas en este trabajo como los árboles de clasificación, Random Forest y Gradient Boosting Machine son técnicas de carácter no paramétrico. Es decir, no requieren seguir una determinada distribución o cumplir con determinadas hipótesis. Además, son modelos muy robustos y tienen una fácil aplicación a cualquier ámbito de investigación.

De los 3 modelos elegidos, el que mejor nivel de interpretación ofrece es el árbol de decisión ya que mediante las reglas de decisión obtenidas se puede discriminar fácilmente a la muestra de la base de datos. Sin embargo, sufre de los problemas de sesgo y varianza, esto se da porque al construir un árbol pequeño se obtendrá un modelo con baja varianza y alto sesgo. Sin embargo, cuando se incrementa la complejidad de la metodología, se encontrará una disminución en el error de predicción esto se da debido a que se obtiene un sesgo más bajo en la metodología a medida que aumenta su dificultad. Por lo cual, un modelo óptimo debe mantener un balance entre estos dos tipos de errores. A esto se le conoce como "trade-off" entre errores de sesgo y varianza.

La implementación de los modelos de ensemble mejoran los problemas que presentan los árboles de clasificación. Esto es debido a que los modelos son un conjunto de árboles de decisión que intentan mejorar la predicción ya sea por Bootstrap (Random Forest) o por la mejora de los errores de predicción de modelos anteriores (GBM). Sin embargo, el nivel de interpretación de los modelos de ensemble se hace de diferente manera ya que no se pueden visualizar las reglas de decisión como un árbol normal. Como hemos visto, se puede hacer mediante las variables más importantes que afectan al modelo

Para el caso del ROE, de los tres modelos analizados se concluyó que el modelo que mejor clasifica a la variable dependiente (ROE) es el modelo GBM con un índice del área bajo la curva AUC del 0,9241. Sin embargo, los modelos del árbol de clasificación y Random forest también muestran un nivel alto de ajuste con 0,8153 y 0,8824 respectivamente.

Con respecto a la importancia de las variables, los modelos dan una mayor importancia a la componente número 2, la cual contiene a variables de tamaño como Logaritmo del patrimonio, logaritmo de los activos y el volumen de negocio (logaritmo de ingresos por intereses). Una justificación para ver la relación directa entre la rentabilidad y el tamaño de una entidad se ve en la medida que las entidades grandes podrían tener a su disposición mayores oportunidades técnicas del mercado, se podrían beneficiar de las

economías de escala. Además, tienen mayor poder de negociación ante clientes y tienen mejor perspectivas de financiamiento. (Baumol, 1982)

Para el caso del RCM, De los tres modelos analizados se concluyó que el modelo que mejor clasifica a la variable dependiente (RCM) será el modelo GBM con un índice del área bajo la curva AUC del 0,9227. Sin embargo, y al igual que con la otra variable dependiente, los modelos del árbol de clasificación y Random forest también muestran un nivel alto de ajuste con 0,8360 y 0,8742 respectivamente.

De acuerdo con la importancia de las variables utilizando al RCM como variable independiente, se verificó que en los 3 modelos la variable que más importancia tiene en los modelos es la componente 5 formada por variables de crédito: tasa de interés activa implícita y el spread de crédito. Como justificación del resultado obtenido se puede decir que las COAC al estar compuesta su cartera principalmente de microcrédito, deberán tener una tasa de interés que genere un cierto nivel de rentabilidad y que genere más captaciones para la entidad permitiendo que mediante la intermediación financiera puedan generar más recursos. Y, de la misma manera, la diferencia entre la tasa activa y pasiva que mantengan las COAC deben estar acorde con la generación de rentabilidad para que las entidades tengan una buena salud financiera.

Con la obtención de la importancia de las variables de los diferentes modelos realizados, se comprobó la relación de los resultados con lo expuesto en el apartado teórico sobre la estructura eficiente en un mercado financiero. Según la misma, se sugiere que la rentabilidad estará condicionada por el tamaño de la entidad financiera y a partir de esta, una gestión eficiente permitirá obtener beneficios económicos por medio de la intermediación.

Bibliografía

Albertazzi, H., & Gambacorta, L. (2009). *Bank profitability and the business cycle.* Journal of Financial Stability, Volume 5, Issue 4, Pages 393-409.

Altamirano, A. (2018). *Modelo de diagnóstico para medir el desempeño financiero en las Cooperativas de Ahorro y Creédito de Ecuador.* Buenos Aires: Revista de investigación en modelos financieros.

Bain, J. (1951). *Relation of Profit Rate to Industry Concentration: American Manufacturing, 1936-1940.* Oxford: The Quarterly Journal of Economics, Vol. 65, No. 3, pp. 293 -324.

Bakar, N., & Tahir, I. (2009). *Applying Multiple Linear Regression and Neural Network to Predict Bank performance.* International Business Research, Vol 2(4), 176-183.

Baumol, W. (1982). *Business behavior, value andgrowth.* New York: Macmillan, 1959. 164 p.

BCE. (01 de 04 de 2021). *Banco Central del Ecuador*. Obtenido de Banco Central del Ecuador: https://www.bce.fin.ec/

Berger, A. (1995). *The Profit-Structure Relationship in Banking--Tests of Market-Power and Efficient Structure Hypotheses.* Ohio: Journal of Money, Credit and Banking, Vol. 27, No. 2.

Brealey, R., Myers, S., & Marcus, A. (1996). *Principios de direccion financiera.* Madrid: Mcgraw Hill Editorial .

Breiman, L. (2001). *Random Forests.* Machine Learning, 45, 5–32, 2001.

Breiman, L., Friedman, J., Stone, C., & Olshen, R. (1984). *Classification and Regression Trees (Wadsworth Statistics/Probability).* Chapman and Hall/CRC; N.º 1 edición (1 enero 1984).

Catal, C. (2012). *Performance Evaluation Metrics for Software.* Acta Polytechnica Hungarica .

Codigo Orgánico Monetario y Financiero. (2014). *Codigo Orgánico Monetario y Financiero.* Quito.

(2011). *Cutler, Adele; Stevens, John; Cutler, David.* Machine Learning 45(1):157-176.

Dang, U. (2011). *THE CAMEL RATING SYSTEM IN BANKING SUPERVISION: A CASE STUDY.* Arcada University of Applied Sciences, International Business.

Demirgüç-Kunt, A., & Huizinga, H. (2000). *Financial structure and bank profitability.*

Washington: The World Bank.

Demsetz, H. (1973). *Indusry structure, Market rivalry and public policy .* The Journal of Law and Economics, Vol. 16 No. 1, pp. 1-9.

Drzewiecki, W. (2017). *Thorough statistical comparison of machine learning regression models and their ensembles for sub-pixel imperviousness and imperviousness change mapping.* Geodesy an Cartography, 66(2), 171-209.

Efron, B. (1979). *Bootstrap Methods: Another Look at the Jackknife.* Ann. Statist. 7(1): 1-26 (January, 1979).

Erdal, H., & Karahanoglu, I. (2016). *Bagging ensemble models for bank profitability: An emprical research on Turkish development and investment banks.* Applied soft computing 49, 861-867.

Fawcett, T. (2006). *An introduction to ROC analysis.* Pattern Recognition Letters 27 (2006) 861–874.

Friedman, J. (2001). *Greedy function approximation: A gradient boosting Machine.* The Annals of Statistics, 29(5), 1189-1232.

Friedman, J. (2002). *Stochastic gradient boosting.* Computational Statistics & Data Analysis, 38, 367-378.

Goddard, J., Molyneux, P., & Wilson, J. (2004). *THE PROFITABILITY OF EUROPEAN BANKS:A CROSS-SECTIONAL AND DYNAMIC PANEL ANALYSIS.* The Manchester School Vol 72 No. 3.

Gonzáles Perez, A., Rodríguez Correa, A., & Acosta Molina, M. (2002). *Factores determinantes de la rentabilidad financiera de las PYMES.* Revista Española De Financiación Y Contabilidad, 31(112), 395-429.

Gonzáles, L. (2019). *Análisis de rentabilidad del sistema bancario panameño .* Madrid: Universidad Complutense de Madrid.

González, N., & Taborda, A. (2015). *ANÁLISIS DE COMPONENTES PRINCIPALES SPARSE: Formulación, algoritmos e implicaciones en el análisis de datos.*

Gu, Q., Zhu, L., & Cai, Z. (2009). *Evaluation Measures of the Classification Performance of Imbalanced Data Sets.* Computational Intelligence and Intelligent Systems, 461-471.

Haslem, J. (1968). *A Statistical Analysis of the Relative Profitability of Commercial Banks.* The journal of finance.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer Series in Statistics.

Heggestad, A., & Mingo, J. (1976). *Prices, Nonprices, and Concentration In Commercial Banking.* Ohio: Journal of Money, Credit and Banking .

Hollis, A., & Sweetman, A. (1998). *Microcredit: What Can we Learn from the Past? .* World Development.

Jolliffe, T. (2002). *Principal Component Analysis, Second Edition.* Encyclopedia of Statistics in Behavioral Science, 30(3), 487.

Lapo, M., & Tello, M. (2021). *Rentabilidad, capital y riesgo crediticio en bancos ecuatorianos.* Investigación Administrativa 50(127):18-39.

Lawrence, S., & Joe, Z. (1999). *Profitability and Marketability of the Top 55 U.S. Commercial Banks.* Maryland: Managment Science.

Ledgerwood, J. (1999). *Sustainable Banking with the poor, Microfinance handbook.* Washington D.C: The World Bank.

Ley Orgánica de Economía Popular y Solidaria . (2011). *Ley Orgánica de Economía Popular y Solidaria .* Quito.

Peltzman, S. (1977). *The Gains and Losses from Industrial Concentration.* Journal of Law & Economics, 20(2), 229-264.

SEPS. (Abril de 2017). *Super Intendencia de Economia Popuar y Solidaria.* Obtenido de Fichas Metodológicas de Indicadores Financieros: https://www.seps.gob.ec/documents/20181/594508/NOTA+TE%CC%81CNICA+PARA+PUBLICAR+-FICHA+METODOLOGICAS+DE+INDICADORES.pdf/a71e5ed1-7fae-4013-a78d-425243db4cfa

SEPS. (01 de Abril de 2021). *Superintendencia de Economía Popular y Solidaria .* Obtenido de Superintendencia de Economía Popular y Solidaria : https://www.seps.gob.ec/#

Short, B. (1979). *The relation between commercial bank profit rates and banking concentration in Canada, Western Europe, and Japan.* Journal of Banking & Finance Volume 3, Issue 3, Pages 209-219.

Smirlock, M. (1985). *Evidence on the (Non) Relationship between Concentration and Profitability in Banking.* Ohio: Journal of Money, Credit and Banking ,Vol. 17, No. 1.

Staikouras, C., & Wood, G. (2004). *The Determinants Of European Bank Profitability.* International Business & Economics Research Journal (IBER).

Uddin, M., Habib, T., Chi, G., & Al Janabi, M. (2020). *Leveraging random forest in micro-enterprises credit riskmodelling for accuracy and interpretability.* International Journal of Finance and Economics. 2020; 1– 17.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory.* New York: Springer.

9. Anexos

Tabla 19 : Activo – Pasivo – Patrimonio

2020						
CUENTA	SEGMENTO 1	SEGMENTO 2	SEGMENTO 3	SEGMENTO 4	SEGMENTO 5	TOTAL COACS
ACTIVO	13.265,62	1.990,36	916,90	434,68	84,35	16.691,91
PASIVOS	11.482,97	1.650,40	746,17	341,59	62,63	14.283,76
PATRIMONIO	1.782,65	339,96	170,72	92,69	21,75	2.407,76
2019						
CUENTA	SEGMENTO 1	SEGMENTO 2	SEGMENTO 3	SEGMENTO 4	SEGMENTO 5	TOTAL COACS
ACTIVO	11.670,69	1.894,20	951,69	446,63	84,91	15.048,11
PASIVOS	10.027,09	1.566,12	784,32	354,90	64,29	12.796,71
PATRIMONIO	1.643,60	328,08	167,37	91,73	20,61	2.251,39
VARIACIÓN						
CUENTA	SEGMENTO 1	SEGMENTO 2	SEGMENTO 3	SEGMENTO 4	SEGMENTO 5	TOTAL COACS
ACTIVO	13,67%	5,08%	-3,66%	-2,67%	-0,66%	10,92%
PASIVOS	14,52%	5,38%	-4,86%	-3,75%	-2,59%	11,62%
PATRIMONIO	8,46%	3,62%	2,00%	1,05%	5,51%	6,95%

Fuente: Elaboración propia

Tabla 20 : Representación Activo – Pasivo – Patrimonio

CUENTA	SEGMENTO 1	SEGMENTO 2	SEGMENTO 3	SEGMENTO 4	SEGMENTO 5	TOTAL COACS
ACTIVO	79,47%	11,92%	5,49%	2,60%	0,51%	100,00%
PASIVO	80,39%	11,55%	5,22%	2,39%	0,44%	100,00%
PATRIMONIO	74,04%	14,12%	7,09%	3,85%	0,90%	100,00%

Fuente: Elaboración propia

Tabla 21 : Estructura Activo

CUENTA	SEGMENTO 1	SEGMENTO 2	SEGMENTO 3	SEGMENTO 4	SEGMENTO 5	TOTAL
FONDOS DISPONIBLES	1.967,17	199,21	105,23	54,34	10,94	2.336,89
OPERACIONES INTERFINANCIERAS ACTIVAS	-	-	-	-	0,00	0,00
INVERSIONES	1.576,81	160,32	58,23	25,00	3,91	1.824,27
CARTERA DE CRÉDITOS	8.656,45	1.444,65	667,59	312,33	58,62	11.139,65
DEUDORES POR ACEPTACIÓN	-	-	-	0,07	-	0,07
CUENTAS POR COBRAR	383,32	59,78	28,03	16,51	4,43	492,07
PROPIEDADES Y EQUIPO	320,51	69,67	42,59	18,98	4,22	455,97
OTROS ACTIVOS	361,35	56,73	15,22	7,46	2,22	442,98
ACTIVO	**13.265,62**	**1.990,36**	**916,90**	**434,68**	**84,35**	**16.691,91**

Fuente: Elaboración propia

Tabla 22 : Estructura Pasivo

CUENTA	SEGMENTO 1	SEGMENTO 2	SEGMENTO 3	SEGMENTO 4	SEGMENTO 5	TOTAL	REP %
OBLIGACIONES CON EL PÚBLICO	10.673,95	1.398,42	670,86	295,17	55,51	13.093,92	91,7%
OBLIGACIONES INMEDIATAS	1,18	0,08	0,14	0,04	0,01	1,44	0,0%
CUENTAS POR PAGAR	304,15	51,67	23,67	12,29	3,30	395,07	2,8%
OBLIGACIONES FINANCIERAS	465,18	195,16	48,80	32,72	3,22	745,08	5,2%
ACEPTACIONES EN CIRCULACIÓN	-	-	-	0,00	-	0,00	0,0%
POR OTROS CREDITOS DEL FONDO DE LIQUIDEZ	-	-	-	-	-	-	0,0%
VALORES EN CIRCULACIÓN	-	-	-	0,00	-	0,00	0,0%
OBLIGACIONES CONVERTIBLES EN ACCIONES Y APORTES	-	-	-	-	-	-	0,0%
OTROS PASIVOS	38,52	5,06	2,70	1,36	0,59	48,23	0,3%
PASIVO	11.482,97	1.650,40	746,17	341,59	62,63	14.283,76	100,0%

Fuente: Elaboración propia

Tabla 23 : Detalle Variables Independientes

VARIABLE	CÁLCULO	NOTACIÓN
	SOLVENCIA	
Suficiencia patrimonial	Patrimonio + Resultados/Activos Inmovilizados Netos	SOL1
Vulnerabilidad del patrimonio	Cartera Improductiva /Patrimonio + Resultados	SOL2
Pérdidas de capital social	Pérdidas Acumuladas - Reservas / Capital Social	SOL3
Solvencia 4	Patrimonio /Activo	SOL4
Solvencia 5	Patrimonio/Pasivo	SOL5
	EFICIENCIA GTO OPERATIVO	
Eficiencia gasto operativo 1	Gastos operativos / Margen neto financiero	EFG1
Eficiencia gasto operativo 2	Gastos operativos/ Activo	EFG2
	CALIDAD DE ACTIVOS	
Calidad de activos 1	Activos productivos / Pasivos con costo	CAA1
Proporción Activo productivo	Activo productivo/ Activo total	CAA2
proporción Activo improductivo	Activos improductivos netos / Activo total	CAA3
	CRÉDITO	
Morosidad de cartera	Cartera improductiva / Cartera total	CR1
Cobertura de cartera	Provisiones de cartera / Cartera improductiva	CR2
Margen neto de intereses	Intereses ganados - Intereses causados	CR3
TI activa implícita	Intereses de cartera/ Cartera total	CR4
TI pasiva implícita	Intereses causados de depósitos/ Depósitos - Operaciones de reporto	CR5
Spread	Tasa activa implícita - Tasa pasiva implícita	CR6
Crédito 7	Provisión prestamos/préstamo total	CR7
Crédito 8	Provisión crédito consumo/ Préstamo consumo	CR8
Rendimiento cartera microcrédito	Cartera de microcrédito/ Cartera microcrédito por vencer + refinanciada por vencer + reestructurada por vencer	CR10
	EFICIENCIA FINANCIERA	
Eficiencia financiera 1	Margen de intermediación / Activos productivos	EFF1
Eficiencia financiera 2	Margen de intermediación / Patrimonio	EFF2
	LIQUIDEZ	
Liquidez 1	Fondos disponibles / Depósitos corto plazo	LIQ1
Liquidez 2	Cartera bruta/Activo total	LIQ2
Liquidez 3	Cartera bruta/Depósitos totales	LIQ3
	TAMAÑO	
Tamaño 1	Logaritmo activos	T1
Tamaño 2	Logaritmo patrimonio	T2
Volumen negocio	Logaritmo de ingresos por intereses	T3

Fuente: SEPS

Tabla 24 : Detalle ACP

COMPONENTE	AUTOVALOR	% VARIANZA	% ACUMULADO
1	5,35	20,596	20,596
2	3,43	13,205	33,801
3	2,79	10,747	44,547
4	1,88	7,241	51,788
5	1,58	6,090	57,878
6	1,46	5,612	63,490
7	1,16	4,453	67,943
8	1,06	4,080	72,023
9	1,02	3,927	75,950
10	0,99	3,794	79,744

Fuente: Elaboración propia

Tabla 25 : Matriz de componentes rotada

VARIABLE	COMP.1	COMP.2	COMP.3	COMP.4	COMP.5	COMP.6	COMP.7	COMP.8	COMP.9	COMP.10
SOL1	0,038	-0,069	-0,004	0,019	0,041	0,042	-0,001	0,077	0,160	0,825
SOL2	0,663	-0,122	-0,106	0,475	-0,031	0,087	-0,107	-0,055	0,240	-0,217
SOL3	-0,047	-0,244	-0,195	-0,068	-0,021	0,025	0,723	-0,007	0,144	-0,143
SOL4	0,109	0,064	0,821	0,007	0,133	-0,182	-0,073	-0,033	-0,108	0,099
SOL5	0,086	-0,140	0,914	-0,057	-0,005	-0,012	0,016	0,010	0,073	-0,037
EFG1	0,093	0,101	0,128	0,133	-0,057	-0,008	0,788	-0,032	-0,117	0,120
EFG2	0,130	-0,509	-0,046	-0,522	0,420	0,065	0,114	-0,059	0,166	-0,093
CAA1	-0,260	-0,071	0,810	-0,014	-0,040	0,277	-0,009	0,052	0,140	-0,114
CAA2	-0,886	0,259	0,006	0,123	-0,007	0,071	-0,089	-0,009	0,095	-0,094
CAA3	0,886	-0,259	-0,006	-0,123	0,007	-0,071	0,089	0,009	-0,095	0,094
CR1	0,845	-0,251	-0,024	-0,180	-0,064	-0,050	-0,061	0,094	-0,004	-0,020
CR2	-0,055	0,018	0,062	-0,013	0,000	-0,079	0,006	-0,040	0,845	0,161
CR3	0,300	0,205	-0,237	-0,273	-0,356	0,248	-0,131	-0,362	-0,071	0,132
CR4	-0,075	0,167	-0,042	-0,138	0,885	0,312	-0,067	-0,042	-0,042	0,066
CR5	-0,125	0,300	-0,074	-0,018	0,037	0,836	-0,022	-0,118	-0,077	0,072
CR6	0,013	-0,043	0,009	-0,132	0,903	-0,281	-0,054	0,041	0,012	0,017
CR7	0,342	-0,023	0,004	-0,689	0,158	-0,034	-0,065	0,229	0,165	-0,223
CR8	0,055	-0,024	-0,070	-0,026	-0,020	0,024	-0,056	0,787	-0,065	0,113
EFF1	-0,246	0,114	0,069	0,675	-0,272	-0,079	0,088	0,223	-0,123	0,098
EFF2	0,089	-0,002	-0,013	0,878	0,047	0,045	0,039	-0,047	0,190	-0,186
LIQ1	0,007	-0,176	0,359	0,033	-0,048	0,729	0,045	0,140	-0,024	-0,024
LIQ2	-0,472	0,006	0,013	0,465	-0,052	0,035	-0,072	-0,464	-0,128	0,204
LIQ3	-0,131	-0,206	0,665	0,137	-0,154	0,292	0,029	-0,136	-0,008	0,054
T1	-0,245	0,900	-0,219	0,026	-0,034	0,060	0,012	-0,030	0,037	-0,062
T2	-0,210	0,932	0,049	0,022	0,029	-0,008	-0,098	-0,030	-0,015	-0,014
T3	-0,353	0,855	-0,228	0,030	0,126	0,107	-0,005	-0,025	0,022	-0,037

Fuente: Elaboración propia

9.1 Código R –Árbol de clasificación

```
#########################################################################################################################
#########################_____________ARBOL DE CLASIFICACIÓN - TFM - FREDDY OQUENDO_____________#########################
#########################################################################################################################

##########---LIBRERIAS---#########
library(caret)
library(tidyverse)
library(rpart)
library(rpart.plot)
library(ModelMetrics)
library(tree)
library(MLmetrics)
library(ROCR)
library(Epi)
library(rattle)

##########---CARGAMOS Y SEGMENTAMOS LA BASE DE DATOS ROE ---##########

BDD_ROE <- read.csv2("C:/Users/Freddy Oquendo/Desktop/TFM MCAF/MODELO/BDD_ROE.csv", stringsAsFactors = T)

set.seed(123)
asignacion <- sample(1:2, size = nrow(BDD_ROE), prob = c(0.8,0.2) ,replace = TRUE)
BDD_TRAIN_ROE <- BDD_ROE[asignacion == 1, ]
BDD_TEST_ROE <- BDD_ROE[asignacion == 2, ]

##########---CARGAMOS Y SEGMENTAMOS LA BASE DE DATOS RCM---##########

BDD_RCM <- read.csv2("C:/Users/Freddy Oquendo/Desktop/TFM MCAF/MODELO/BDD_RCM.csv", stringsAsFactors = T)

set.seed(123)
asignacion <- sample(1:2, size = nrow(BDD_RCM), prob = c(0.8,0.2) ,replace = TRUE)
BDD_TRAIN_RCM <- BDD_RCM[asignacion == 1, ]
BDD_TEST_RCM <- BDD_RCM[asignacion == 2, ]

##########---ARBOL DE CLASIFICACIÓN ROE---##########

###----MODELO INICIAL ROE----###

  # En este primer modelo entrenamos un árbol sin especificar ningún parametro extra

set.seed(1234)
M1_ROE <- rpart(formula = ROE ~ COMP_1+ COMP_2  + COMP_3 +  COMP_4 + COMP_5 +
              COMP_6 + COMP_7 + COMP_8 + COMP_9 + COMP_10, data= BDD_TRAIN_ROE, method  = "class")

  # Resultado inicial del arbol
rpart.plot(M1_ROE,box.palette="Grays", shadow.col="lightblue", nn=TRUE, type = 2, clip.right.labs = T)
summary(M1_ROE)
```

```
  #Número óptimo de nodos
printcp(M1_ROE)
plotcp(M1_ROE)

  #Importancia de variables
M1_ROE$variable.importance
barplot(t(M1_ROE$variable.importance),horiz=F)

  #matriz de confusion Modelo Inicial
Prediccion_M1_ROE <- predict(M1_ROE, newdata= BDD_TEST_ROE, type = "class")
CM1<-table(Prediccion_M1_ROE, BDD_TEST_ROE$ROE)

  #Specificidad = A/(A + C)
(Specificidad_M1_ROE <- (23/(23 + 9)))

  #Sensitividad = D/(B + D)
(Sensitividad_M1_ROE <- (55/(8 + 55)))

  #Precision = D/(D + C)
(Precision_M1_ROE <- (55/(55 + 9)))


###----MODELO FINAL ROE----###

  #Seleccionamos el tamaño optimo del arbol el cual minimiza el error de clasificación
M1_ROE$cptable[which.min(M1_ROE$cptable[,"xerror"]),"CP"]

  #Entrenamos modelo con parametros óptimos

M2_ROE<- prune(M1_ROE, cp=M1_ROE$cptable[which.min(M1_ROE$cptable[,"xerror"]),"CP"])

  #Resultados del modelo
rpart.plot(M2_ROE, box.palette="Grays", shadow.col="lightblue", tweak = 1, extra = 106, nn=TRUE, type = 2,
           under = F, fallen.leaves = T )
summary(M2_ROE)
asRules(M2_ROE)

  #Importancia de variables
M2_ROE$variable.importance
barplot(t(M1_ROE$variable.importance),horiz=F)

  #Matriz de confusion modelo final
Prediccion_M2_ROE <- predict(M2_ROE, newdata= BDD_TEST_ROE, type = "class")
CM2<-table(Prediccion_M2_ROE, BDD_TEST_ROE$ROE)
```

```
  #Especificidad = A/(A + C)
(Especificidad_M2_ROE <- (22/(22 + 10)))

  #Sensitividad = D/(B + D)
(Sensitividad_M2_ROE <- (54/(9 + 54)))

  #Precision = D/(D + C)
(Precision_M2_ROE <- (54/(54 + 10)))

##########################################################################################################################
##########################################################################################################################


##########---ARBOL DE CASIICACION RCM---##########

###----MODELO INICIAL RCM----###

  # En este primer modelo entrenamos un arbol sin especificar ninguna parametro extra

set.seed(023)
M1_RCM <- rpart(formula = RCM ~ COMP_1+ COMP_2  + COMP_3 +  COMP_4 + COMP_5 +
              COMP_6 + COMP_7 + COMP_8 + COMP_9 + COMP_10, data= BDD_TRAIN_RCM, method  = "class")

  # Resultado inicial del arbol
M1_RCM
rpart.plot(M1_RCM,box.palette="Grays", shadow.col="lightblue", tweak = 1, extra = 106, nn=TRUE, type = 2,
           under = F, fallen.leaves = T )
summary(M1_RCM)
rpart.control()

  #Número óptimo de nodos
printcp(M1_RCM)
plotcp(M1_RCM)

  #Importancia de variables
M1_RCM$variable.importance
barplot(t(M1_RCM$variable.importance),horiz=TRUE)

  #Matriz de confusión modelo inicial
Prediccion_M1_RCM <- predict(M1_RCM, newdata= BDD_TEST_RCM, type = "class")
(CM3<-table(Prediccion_M1_RCM, BDD_TEST_RCM$RCM))

  #Specificidad = A/(A + C)
(Specificidad_M1_RCM <- (26/(26 + 12)))

  #Sensitividad = D/(B + D)
(Sensitividad_M1_RCM <- (50/(7 + 50)))

  #Precision = D/(D + C)
(Precision_M3 <- (50/(50 + 12)))

###----MODELO FINAL RCM----###

  #Seleccionamos el tamaño optimo del arbol el cual minimiza el error de clasificación
M1_RCM$cptable[which.min(M1_RCM$cptable[,"xerror"]),"CP"]

  #Entrenamos modelo con parametros optimos
set.seed(0124)
M2_RCM<- prune(M1_RCM, cp=M1_RCM$cptable[which.min(M1_RCM$cptable[,"xerror"]),"CP"])

  #Resultados del modelo
rpart.plot(M2_RCM, box.palette="Grays", shadow.col="lightblue", tweak = 1, extra = 106, nn=TRUE, type = 2,
           under = F, fallen.leaves = T)
summary(M2_RCM)
asRules(M2_RCM)

  #Importancia de variables
M2_RCM$variable.importance
barplot(t(M2_RCM$variable.importance),horiz=TRUE)

  #Matriz de confusion Modelo Final
Prediccion_M2_RCM <- predict(M2_RCM, newdata= BDD_TEST_RCM, type = "class")
(CM4<-table(Prediccion_M2_RCM, BDD_TEST_RCM$RCM))

  #Specificidad = A/(A + C)
(Specificidad_M2_RCM <- (24/(24 +14)))

  #Sensitividad = D/(B + D)
(Sensitividad_M2_RCM <- (51/(6 + 51)))

  #Precision = D/(D + C)
(Precision_M2_RCM <- (51/(51 + 14)))

##########---ROC CURVE y AUC---##########

### ROC CURVE ROE "PODA"

pred_M2_ROE <- prediction(predict(M2_ROE, type = "prob")[, 2], BDD_TRAIN_ROE$ROE)
performance_M2_ROE <- performance(pred_M2_ROE,measure="tpr",x.measure="fpr")

plot(performance_M2_ROE, avg= "threshold", colorize=TRUE, lwd= 3)+
plot(performance_M2_ROE,lty=3,col="grey78",add=TRUE)+
abline(a=0, b=1, lty=2, lwd=1,col="black")


auc_M2_ROE <- performance(pred_M2_ROE, measure = "auc")
(auc_M2_ROE <- auc_M2_ROE@y.values[[1]])

### ROC CURVE RCM "PODA"

pred_M2_RCM <- prediction(predict(M2_RCM, type = "prob")[, 2], BDD_TRAIN_RCM$RCM)
performance_M2_RCM <- performance(pred_M2_RCM,measure="tpr",x.measure="fpr")

plot(performance_M2_RCM, avg= "threshold", colorize=TRUE, lwd= 3) +
plot(performance_M2_RCM,lty=3,col="grey78",add=TRUE) +
abline(a=0, b=1, lty=2, lwd=1,col="black")

auc_M2_RCM <- performance(pred_M2_RCM, measure = "auc")
(auc_M2_RCM <- auc_M2_RCM@y.values[[1]])
```

9.2 Código R – Random Forest

```
#########################################################################################################################
#########################_________________RANDOM FOREST - TFM - FREDDY OQUENDO_________________#########################
#########################################################################################################################

##########---LIBRERIAS---#########

library(caret)
library(randomForest)
library(mlr)
library(varImp)
library(dplyr)
library(tidymodels)
library(ROCR)
library(MLmetrics)

##########---CARGAMOS Y SEGMENTAMOS LA BASE DE DATOS ROE ---##########

BDD_ROE <- read.csv2("C:/Users/Freddy Oquendo/Desktop/TFM MCAF/MODELO/BDD_ROE.csv")
BDD_ROE$ROE <- as.factor(BDD_ROE$ROE)

set.seed(107)
asignacion <- sample(1:2, size = nrow(BDD_ROE), prob = c(0.8,0.2) ,replace = TRUE)
BDD_TRAIN_ROE <- BDD_ROE[asignacion == 1, ]
BDD_TEST_ROE <- BDD_ROE[asignacion == 2, ]

##########---CARGAMOS Y SEGMENTAMOS LA BASE DE DATOS RCM---##########

BDD_RCM <- read.csv2("C:/Users/Freddy Oquendo/Desktop/TFM MCAF/MODELO/BDD_RCM.csv")
BDD_RCM$RCM <- as.factor(BDD_RCM$RCM)

set.seed(107)
asignacion_RCM <- sample(1:2, size = nrow(BDD_ROE), prob = c(0.8,0.2) ,replace = TRUE)
BDD_TRAIN_RCM <- BDD_RCM[asignacion_RCM == 1, ]
BDD_TEST_RCM <- BDD_RCM[asignacion_RCM == 2, ]

##########---RANDOM FOREST ROE---##########

###----MODELO INICIAL ROE----###

set.seed(108)
RF_ROE <- randomForest(ROE ~ COMP_1+ COMP_2 + COMP_3 +  COMP_4 + COMP_5 +
                         COMP_6 + COMP_7 +  COMP_8 + COMP_9 + COMP_10, data = BDD_TRAIN_ROE,
                       importance=TRUE, proximity=TRUE, na.action=na.fail)
```

```
  ## PRINCIPALES RESULTADOS

print(RF_ROE)
attributes(RF_ROE)

  ## MATRIZ DE CONFUSION TRAIN DATA

predict_RF1 <- predict(RF_ROE, BDD_TRAIN_ROE)
confusionMatrix(predict_RF1, BDD_TRAIN_ROE$ROE)
table(predict_RF1, BDD_TRAIN_ROE$ROE)

  ## PROBAMOS MODELO CON TEST DATA

predict_RF1_2 <- predict(RF_ROE, BDD_TEST_ROE)
confusionMatrix(predict_RF1_2, BDD_TEST_ROE$ROE)
table(predict_RF1_2, BDD_TEST_ROE$ROE)


###----MODELO FINAL ROE----###

  ## OBTENEMOS HYPERPARAMETROS

plot(RF_ROE)
(res <- tuneRF(x = subset(BDD_TRAIN_ROE, select = -ROE), y = BDD_TRAIN_ROE$ROE, ntreeTry = 300))
(mtry_opt <- res[,"mtry"][which.min(res[,"OOBError"])])


  ## NUEVO MODELO CON HYPERPARAMETROS
(RF_ROE_2 <- randomForest(ROE ~ COMP_1+ COMP_2  + COMP_3 +  COMP_4 + COMP_5 +
                            COMP_6 + COMP_7 + COMP_8 + COMP_9 + COMP_10,
                          data = BDD_TRAIN_ROE, importance=TRUE, proximity=TRUE, na.action=na.fail,
                          mtry=3, ntree=300))
print(RF_ROE_2)

  ## MATRIZ DE CONFUSION TRAIN DATA

predict_RF2 <- predict(RF_ROE_2, BDD_TRAIN_ROE)
confusionMatrix(predict_RF2, BDD_TRAIN_ROE$ROE)
table(predict_RF2, BDD_TRAIN_ROE$ROE)

  ## PROBAMOS MODELO CON TEST DATA

predict_RF2_2 <- predict(RF_ROE_2, BDD_TEST_ROE)
confusionMatrix(predict_RF2_2, BDD_TEST_ROE$ROE)
table(predict_RF2_2, BDD_TEST_ROE$ROE)
```

```
  ## IMPORTANCIA DE VARIABLES

varImpPlot(RF_ROE_2)
importance(RF_ROE_2)

##########################################################################################################################################
##########################################################################################################################################

##########---RANDOM FOREST RCM---##########

###----MODELO INICIAL RCM----###

set.seed(105)
(RF_RCM <- randomForest(RCM ~ COMP_1+ COMP_2  + COMP_3 +  COMP_4 + COMP_5 +
                          COMP_6 + COMP_7 +  COMP_8 + COMP_9 + COMP_10, data = BDD_TRAIN_RCM,
                        importance=TRUE, proximity=TRUE, na.action=na.fail))

  ## PRINCIPALES RESULTADOS

print(RF_RCM)
attributes(RF_RCM)

  ## MATRIZ DE CONFUSION TRAIN DATA

predict_RCM1 <- predict(RF_RCM, BDD_TRAIN_RCM)
confusionMatrix(predict_RCM1, BDD_TRAIN_RCM$RCM)
table(predict_RCM1, BDD_TRAIN_RCM$RCM)

  ## PROBAMOS MODELO CON TEST DATA

predict_RCM1_2 <- predict(RF_RCM, BDD_TEST_RCM)
confusionMatrix(predict_RCM1_2, BDD_TEST_RCM$RCM)
table(predict_RCM1_2, BDD_TEST_RCM$RCM)

###----MODELO FINAL RCM----###

  ## OBTENEMOS HYPERPARAMETROS
plot(RF_RCM)
set.seed(105)
(res_RCM <- tuneRF(x = subset(BDD_TRAIN_RCM, select = -RCM), y = BDD_TRAIN_RCM$RCM, ntreeTry = 150))
(mtry_opt_RCM <- res_RCM[,"mtry"][which.min(res_RCM[,"OOBError"])])

  ## NUEVO MODELO CON HYPERPARAMETROS
(RF_RCM_2 <- randomForest(RCM ~ COMP_1+ COMP_2  + COMP_3 +  COMP_4 + COMP_5 +
                            COMP_6 + COMP_7 + COMP_8 + COMP_9 + COMP_10,
                          data = BDD_TRAIN_RCM, importance=TRUE, proximity=TRUE, na.action=na.fail,
                          mtry=2, ntree=150))
```

```
  ## MATRIZ DE CONFUSION TRAIN DATA

predict_RCM2 <- predict(RF_RCM_2, BDD_TRAIN_RCM)
confusionMatrix(predict_RCM2, BDD_TRAIN_RCM$RCM)
table(predict_RCM2, BDD_TRAIN_RCM$RCM)

  ## PROBAMOS MODELO CON TEST DATA

predict_RCM2_2 <- predict(RF_RCM_2, BDD_TEST_RCM)
confusionMatrix(predict_RCM2_2, BDD_TEST_RCM$RCM)
table(predict_RCM2_2, BDD_TEST_RCM$RCM)

  ## IMPORTANCIA DE VARIABLES

varImpPlot(RF_RCM_2)
importance(RF_RCM_2, sort=T)

##########---ROC CURVE y AUC---##########

  ## ROC CURVE MODELO FINAL ROE


pred_RF2_ROE <- prediction(predict(RF_ROE_2, type = "prob")[, 2], BDD_TRAIN_ROE$ROE)
performance_RF2_ROE <- ROCR::performance(pred_RF2_ROE,"tpr","fpr")

plot(performance_RF2_ROE, avg= "threshold", colorize=TRUE, lwd= 3)+
  plot(performance_RF2_ROE,lty=3,col="grey78",add=TRUE) +
  abline(a=0, b=1, lty=2, lwd=1,col="black")

auc_RF_ROE <- ROCR::performance(pred_RF2_ROE, measure = "auc")
(auc_RF_ROE <- auc_RF_ROE@y.values[[1]])

  ## ROC CURVE MODEL FINAL RCM

pred_RF_RCM <- prediction(predict(RF_RCM_2, type = "prob")[, 2], BDD_TRAIN_RCM$RCM)
performance_RF_RCM <- ROCR::performance(pred_RF_RCM,measure="tpr",x.measure="fpr")

plot(performance_RF_RCM, avg= "threshold", colorize=TRUE, lwd= 3) +
  plot(performance_RF_RCM,lty=3,col="grey78",add=TRUE) +
  abline(a=0, b=1, lty=2, lwd=1,col="black")

auc_RF_RCM <-  ROCR::performance(pred_RF_RCM, measure = "auc")
(auc_RF_RCM <- auc_RF_RCM@y.values[[1]])
```

9.3 Código R – Gradient Boosting Machine

```
#########################################################################################################################
####################_________________GRADIENT BOOSTING MACHINE - TFM - FREDDY OQUENDO_________________#####################
#########################################################################################################################

##########---LIBRERIAS---#########

library(caret)
library(gbm)
library(mlr)
library(varImp)
library(dplyr)
library(ROCR)
library(MLmetrics)
library(pROC)

##########---CARGAMOS Y SEGMENTAMOS LA BASE DE DATOS ROE ---##########

BDD_ROE <- read.csv2("C:/Users/Freddy Oquendo/Desktop/TFM MCAF/MODELO/BDD_ROE.csv", stringsAsFactors = T)
set.seed(107)
asignacion <- sample(1:2, size = nrow(BDD_ROE), prob = c(0.8,0.2) ,replace = TRUE)
BDD_TRAIN_ROE <- BDD_ROE[asignacion == 1, ]
BDD_TEST_ROE <- BDD_ROE[asignacion == 2, ]

##########---CARGAMOS Y SEGMENTAMOS LA BASE DE DATOS RCM---##########

BDD_RCM <- read.csv2("C:/Users/Freddy Oquendo/Desktop/TFM MCAF/MODELO/BDD_RCM.csv", stringsAsFactors = T)
set.seed(107)
asignacion <- sample(1:2, size = nrow(BDD_RCM), prob = c(0.8,0.2) ,replace = TRUE)
BDD_TRAIN_RCM <- BDD_RCM[asignacion == 1, ]
BDD_TEST_RCM <- BDD_RCM[asignacion == 2, ]

##########---GBM ROE---##########

###----MODELO INICIAL ROE----###

set.seed(10)
GBM_ROE <- gbm(ROE ~ COMP_1+ COMP_2 + COMP_3 +  COMP_4 + COMP_5 +
                          COMP_6 + COMP_7 +  COMP_8 + COMP_9 + COMP_10, data = BDD_TRAIN_ROE,
                          distribution = "bernoulli", n.trees = 6000)

  ## PRINCIPALES RESULTADOS

print(GBM_ROE)
summary(GBM_ROE)


  ## MATRIZ DE CONFUSION TRAIN DATA

BDD_TRAIN_ROE$ROE <- as.factor(BDD_TRAIN_ROE$ROE)
BDD_TEST_ROE$ROE <- as.factor(BDD_TEST_ROE$ROE)

predict_GBM1 <- predict.gbm(object = GBM_ROE, newdata = BDD_TRAIN_ROE, n.trees = 6000, type = "response")
predict_GBM1 <- round(predict_GBM1)

confusionMatrix(BDD_TRAIN_ROE$ROE, as.factor(predict_GBM1))
table(BDD_TRAIN_ROE$ROE, as.factor(predict_GBM1))

  ## PROBAMOS MODELO CON TEST DATA

predict_GBM2 <- predict.gbm(object = GBM_ROE, newdata = BDD_TEST_ROE, n.trees = 6000, type = "response")
predict_GBM2 <- round(predict_GBM2)

confusionMatrix(BDD_TEST_ROE$ROE, as.factor(predict_GBM2))
table(BDD_TEST_ROE$ROE, as.factor(predict_GBM2))

  ## OBTENEMOS HIPERPARAMETROS PARA MEJOR AJUSTE

ntree_opt_oob <- gbm.perf(object = GBM_ROE,  method =  "OOB", oobag.curve = TRUE)
print(paste0("Optimal n.trees (OOB Estimate): ", ntree_opt_oob))


###----MODELO FINAL ROE----###

BDD_TRAIN_ROE <- BDD_ROE[asignacion == 1, ]
BDD_TEST_ROE <- BDD_ROE[asignacion == 2, ]

set.seed(11)
GBM_ROE_2 <- gbm(ROE ~ COMP_1+ COMP_2 + COMP_3 +  COMP_4 + COMP_5 +
                   COMP_6 + COMP_7 +  COMP_8 + COMP_9 + COMP_10, data = BDD_TRAIN_ROE,
                 distribution = "bernoulli", n.trees = 78)

  ## PRINCIPALES RESULTADOS

print(GBM_ROE_2)
summary(GBM_ROE_2)
```

```
  ## MATRIZ DE CONFUSION TRAIN DATA

BDD_TRAIN_ROE$ROE <- as.factor(BDD_TRAIN_ROE$ROE)
BDD_TEST_ROE$ROE <- as.factor(BDD_TEST_ROE$ROE)

predict_GBM3 <- predict.gbm(object = GBM_ROE_2, newdata = BDD_TRAIN_ROE, n.trees = 78, type = "response")
predict_GBM3 <- round(predict_GBM3)

confusionMatrix(BDD_TRAIN_ROE$ROE, as.factor(predict_GBM3))
table(BDD_TRAIN_ROE$ROE, as.factor(predict_GBM3))

  ## PROBAMOS MODELO CON TEST DATA

predict_GBM4 <- predict.gbm(object = GBM_ROE_2, newdata = BDD_TEST_ROE, n.trees = 78, type = "response")
predict_GBM4 <- round(predict_GBM4)

confusionMatrix(BDD_TEST_ROE$ROE, as.factor(predict_GBM4))
table(BDD_TEST_ROE$ROE, as.factor(predict_GBM4))

#########################################################################################################################
#########################################################################################################################

##########---GBM RCM---##########

###----MODELO INICIAL RCM----###

set.seed(20)
GBM_RCM <- gbm(RCM ~ COMP_1+ COMP_2 + COMP_3 +  COMP_4 + COMP_5 +
                 COMP_6 + COMP_7 +  COMP_8 + COMP_9 + COMP_10, data = BDD_TRAIN_RCM,
               distribution = "bernoulli", n.trees = 6000)

  ## PRINCIPALES RESULTADOS

print(GBM_RCM)
summary(GBM_RCM)

  ## MATRIZ DE CONFUSION TRAIN DATA

BDD_TRAIN_RCM$RCM <- as.factor(BDD_TRAIN_RCM$RCM)
BDD_TEST_RCM$RCM <- as.factor(BDD_TEST_RCM$RCM)

predict_RCM1 <- predict.gbm(object = GBM_RCM, newdata = BDD_TRAIN_RCM, n.trees = 5000, type = "response")
predict_RCM1 <- round(predict_RCM1)

confusionMatrix(BDD_TRAIN_RCM$RCM, as.factor(predict_RCM1))
table(BDD_TRAIN_RCM$RCM, as.factor(predict_RCM1))


  ## PROBAMOS MODELO CON TEST DATA

predict_RCM2 <- predict.gbm(object = GBM_RCM, newdata = BDD_TEST_RCM, n.trees = 6000, type = "response")
predict_RCM2 <- round(predict_RCM2)

confusionMatrix(BDD_TEST_RCM$RCM, as.factor(predict_RCM2))
table(BDD_TEST_RCM$RCM, as.factor(predict_RCM2))

  ## OBTENEMOS HIPERPARAMETROS PARA MEJOR AJUSTE

ntree_opt_oob_RCM <- gbm.perf(object = GBM_RCM, method =  "OOB", oobag.curve = TRUE)
print(paste0("Optimal n.trees (OOB Estimate): ", ntree_opt_oob_RCM))

###----MODELO FINAL RCM----###

BDD_TRAIN_RCM <- BDD_RCM[asignacion == 1, ]
BDD_TEST_RCM <- BDD_RCM[asignacion == 2, ]

set.seed(15)
GBM_RCM_2 <- gbm(RCM ~ COMP_1+ COMP_2 + COMP_3 +  COMP_4 + COMP_5 +
                   COMP_6 + COMP_7 +  COMP_8 + COMP_9 + COMP_10, data = BDD_TRAIN_RCM,
                 distribution = "bernoulli", n.trees = 81)

  ## PRINCIPALES RESULTADOS

print(GBM_RCM_2)
summary(GBM_RCM_2)

  ## MATRIZ DE CONFUSION TRAIN DATA

BDD_TRAIN_RCM$RCM <- as.factor(BDD_TRAIN_RCM$RCM)
BDD_TEST_RCM$RCM <- as.factor(BDD_TEST_RCM$RCM)

predict_RCM3 <- predict.gbm(object = GBM_RCM_2, newdata = BDD_TRAIN_RCM, n.trees = 81, type = "response")
predict_RCM3 <- round(predict_RCM3)

confusionMatrix(BDD_TRAIN_RCM$RCM, as.factor(predict_RCM3))
table(BDD_TRAIN_RCM$RCM, as.factor(predict_RCM3))

  ## PROBAMOS MODELO CON TEST DATA

predict_RCM4 <- predict.gbm(object = GBM_RCM_2, newdata = BDD_TEST_RCM, n.trees = 81, type = "response")
predict_RCM4 <- round(predict_RCM4)

confusionMatrix(BDD_TEST_RCM$RCM, as.factor(predict_RCM4))
table(BDD_TEST_RCM$RCM, as.factor(predict_RCM4))
```

```
##########---ROC CURVE y AUC---##########

## ROC CURVE MODELO FINAL ROE

pred_GBM_ROE <- prediction(predict(GBM_ROE_2, type = "response"), BDD_TRAIN_ROE$ROE)
performance_GBM_ROE <- ROCR::performance(pred_GBM_ROE,"tpr","fpr")

plot(performance_GBM_ROE, avg= "threshold", colorize=TRUE, lwd= 3)+
  plot(performance_GBM_ROE,lty=3,col="grey78",add=TRUE) +
  abline(a=0, b=1, lty=2, lwd=1,col="black")

auc_GBM_ROE <- ROCR::performance(pred_GBM_ROE, measure = "auc")
(auc_GBM_ROE <- auc_GBM_ROE@y.values[[1]])

## ROC CURVE MODELO FINAL RCM

pred_GBM_RCM <- prediction(predict(GBM_RCM_2, type = "response"), BDD_TRAIN_RCM$RCM)
performance_GBM_RCM <- ROCR::performance(pred_GBM_RCM,measure="tpr",x.measure="fpr")

plot(performance_GBM_RCM, avg= "threshold", colorize=TRUE, lwd= 3) +
  plot(performance_GBM_RCM,lty=3,col="grey78",add=TRUE) +
  abline(a=0, b=1, lty=2, lwd=1,col="black")

auc_GBM_RCM <-  ROCR::performance(pred_GBM_RCM, measure = "auc")
(auc_GBM_RCM <- auc_GBM_RCM@y.values[[1]])
```

9.4 Código R – Análisis de componentes principales

```
#########################################################################################################################
################################_______________ACP - TFM - FREDDY OQUENDO_______________#################################
#########################################################################################################################

##########---LIBRERIAS---#########
library(tidyverse)
library(skimr)
library(GGally)
library(ggplot2)
library(ggcorrplot)
library(FactoMineR)
library(factoextra)
library(ggpubr)

#### CARGAMOS BASE DE DATOS ####
BDD<- read.csv2("C:/Users/Freddy Oquendo/Desktop/TFM MCAF/MODELO/BDD_AED.csv", stringsAsFactors = T)
datos <- BDD %>%select(-c(ENTIDADES,ROE, RCM))

#### CORRELACIONES ####
corr_datos <- datos %>%
  cor(use = "pairwise") %>%
  round(1)

ggcorrplot(corr_datos, type = "upper", lab = T, show.legend = F, hc.order = T)

#### PCA ####
PCA_1 <- PCA(X = datos, scale.unit = TRUE, graph = T, ncp = 10)

####  REPRESENTACION PCA ####
G1<- fviz_screeplot(PCA_1, addlabels = TRUE, ylim = c(0, 25))
G2<-fviz_screeplot(PCA_1, choice= c("eigenvalue"), addlabels = TRUE, ylim = c(0, 6))
ggarrange(G1, G2, nrow = 1, align = "v")

#### COMPOSICION COMPONENTES ####
fviz_contrib(PCA_1, choice = "var", axes = 1, top = 10)
fviz_contrib(PCA_1, choice = "var", axes = 2, top = 10)
fviz_contrib(PCA_1, choice = "var", axes = 3, top = 10)
fviz_contrib(PCA_1, choice = "var", axes = 4, top = 10)
fviz_contrib(PCA_1, choice = "var", axes = 5, top = 10)
fviz_contrib(PCA_1, choice = "var", axes = 6, top = 10)
fviz_contrib(PCA_1, choice = "var", axes = 7, top = 10)
fviz_contrib(PCA_1, choice = "var", axes = 8, top = 10)
fviz_contrib(PCA_1, choice = "var", axes = 9, top = 10)
fviz_contrib(PCA_1, choice = "var", axes = 10, top = 10)
```

9.5 Muestra Base de datos

ENTIDADES	RCM	ROE	SOL1	SOL2	SOL3	SOL4	SOL5	EFG1	EFG2	CAA1	CAA2	CAA3	CR1	CR2	CR3	CR4	CR5	CR6	CR7	CR8	EFF1	EFF2	LIQ1	LIQ2	LIQ3	T1	T2	T3
C1	0,00	0,02	278,87	-2,27	0,00	0,30	0,43	78,44	3,81	129,06	89,15	10,85	1,22	179,86	38,13	11,41	4,77	6,64	0,02	0,02	1,17	3,49	25,03	0,67	0,97	17,22	16,01	14,77
C2	0,14	0,00	95,55	-3,48	0,00	0,10	0,11	108,09	6,01	99,79	87,96	12,04	1,81	125,13	49,96	15,20	7,18	8,02	0,02	0,02	-0,51	-4,53	20,67	0,73	0,88	18,34	16,03	16,21
C3	0,17	0,04	1616,08	-20,81	0,00	0,16	0,19	99,07	4,49	121,31	97,59	2,41	2,36	292,59	43,58	15,72	7,64	8,08	0,07	0,04	0,04	0,26	21,60	0,63	0,89	18,78	16,95	16,70
C4	0,13	0,04	192,85	-8,67	0,00	0,13	0,15	96,72	5,16	109,59	92,48	7,52	2,64	153,66	42,33	13,34	6,08	7,26	0,04	0,01	0,19	1,38	23,63	0,72	0,93	18,24	16,18	16,04
C5	0,17	0,01	474,54	-6,01	0,00	0,14	0,16	101,97	6,16	114,26	95,74	4,26	4,67	120,43	47,38	15,86	8,07	7,79	0,06	0,05	-0,12	-0,87	13,26	0,77	0,92	17,81	15,82	15,86
C6	0,00	0,02	212,11	-0,34	0,00	0,18	0,22	89,04	4,29	113,72	91,31	8,69	2,53	103,02	45,96	11,12	5,16	5,96	0,03	0,03	0,58	2,91	15,46	0,76	0,95	15,18	13,47	12,77
C7	0,17	0,05	151,92	-2,18	0,00	0,10	0,11	94,14	3,91	102,96	89,94	10,06	3,60	109,27	51,59	14,07	6,05	8,02	0,04	0,03	0,27	2,36	29,57	0,62	0,71	21,65	19,38	19,37
C8	0,00	0,05	740,49	-4,70	0,00	0,13	0,15	86,96	5,28	119,54	97,89	2,11	1,41	151,74	47,90	14,79	7,36	7,43	0,02	0,01	0,81	6,17	21,68	0,79	0,97	15,64	13,59	13,57
C9	0,04	0,02	4541,81	-13,92	0,00	0,18	0,21	91,02	4,81	122,44	99,37	0,63	3,45	176,17	36,73	11,94	5,77	6,16	0,06	0,06	0,48	2,68	73,15	0,82	1,09	15,50	13,77	13,38
C10	0,14	0,01	3122,72	-33,84	0,00	0,13	0,16	97,93	3,46	115,69	97,11	2,89	3,23	296,84	46,02	13,86	6,37	7,49	0,11	0,05	0,08	0,54	37,10	0,58	0,74	20,80	18,80	18,57
C11	0,16	0,00	124,20	-5,90	0,00	0,08	0,09	107,06	5,12	103,11	92,57	7,43	2,02	130,56	52,79	14,27	6,80	7,47	0,03	0,01	-0,36	-4,23	22,10	0,72	0,82	18,57	16,04	16,40
C12	0,16	0,01	227,91	-0,93	0,00	0,13	0,15	103,47	4,19	110,91	93,60	6,40	4,44	103,68	53,23	14,16	7,28	6,88	0,05	0,04	-0,15	-1,06	27,14	0,68	0,82	17,48	15,46	15,30
C13	0,16	0,01	217,18	0,72	0,00	0,15	0,17	116,08	6,41	110,98	92,82	7,18	4,66	97,07	38,31	15,69	5,89	9,81	0,05	0,05	-0,96	-6,04	26,98	0,70	0,86	16,43	14,51	14,35
C14	0,15	0,06	272,15	-12,49	0,00	0,10	0,11	97,79	4,63	109,17	95,86	4,14	2,31	160,81	48,61	14,65	7,60	7,05	0,04	0,10	0,11	1,09	24,13	0,79	0,96	16,88	14,54	14,83
C15	0,14	0,01	325,59	-3,91	0,00	0,13	0,14	122,18	4,95	111,71	94,99	5,01	2,65	124,79	47,45	14,04	6,29	7,75	0,03	0,03	-0,95	-7,15	27,72	0,70	0,83	16,92	14,85	14,73
C16	0,19	0,06	804,12	-11,11	0,00	0,13	0,15	68,05	3,77	118,05	97,73	2,27	3,93	148,75	48,14	15,70	7,49	8,21	0,06	0,04	1,81	13,50	32,93	0,67	0,81	16,28	14,25	14,23
C17	0,00	0,00	169,89	3,50	0,00	0,15	0,18	99,83	11,87	102,61	85,94	14,06	10,91	94,38	27,64	18,74	5,44	13,30	0,11	0,12	0,02	0,13	32,15	0,69	0,83	12,37	10,49	10,57
C18	0,13	0,00	796,22	0,98	0,00	0,23	0,30	122,11	8,11	130,44	95,58	4,42	3,07	91,63	42,02	11,79	7,31	4,48	0,03	0,03	-1,54	-6,35	32,74	0,83	1,28	15,14	13,68	12,97
C19	0,15	0,07	7283,53	-14,82	0,00	0,19	0,23	51,59	1,95	129,69	99,33	0,67	4,00	183,05	42,48	13,30	9,34	3,96	0,08	0,05	1,84	9,71	13,48	0,72	1,30	17,30	15,63	15,19
C20	0,00	0,00	312,28	3,28	0,00	0,29	0,41	88,68	9,52	124,57	85,52	14,48	3,65	68,13	13,55	11,73	1,94	9,79	0,03	0,03	1,38	4,43	14,67	0,80	1,16	12,36	11,11	10,03
C21	0,14	0,02	476,97	0,96	0,00	0,37	0,59	86,22	6,72	156,37	91,27	8,73	8,38	94,49	18,85	12,04	3,38	8,66	0,09	0,09	1,18	2,90	17,11	0,65	1,12	13,69	12,70	11,43
C22	0,02	0,00	65,42	64,52	0,00	0,10	0,11	117,69	9,56	93,94	83,90	16,10	13,54	50,00	40,47	12,52	5,58	6,94	0,07	0,17	-1,66	-14,22	16,74	0,77	0,94	14,93	12,66	12,74
C23	0,15	0,11	892,65	-3,90	0,00	0,21	0,27	51,11	3,58	128,74	97,09	2,91	8,23	110,50	35,75	14,18	8,00	6,19	0,10	0,09	3,52	16,21	24,78	0,78	1,27	15,40	13,85	13,42
C24	0,16	0,03	911,58	-1,83	0,00	0,17	0,20	102,53	4,70	119,59	97,27	2,73	9,12	103,62	45,88	14,65	9,35	5,29	0,10	0,16	-0,12	-0,69	28,70	0,76	1,11	14,81	13,03	12,84
C25	0,22	0,00	67,41	60,72	0,00	0,13	0,14	98,91	8,09	79,59	68,94	31,06	15,70	13,25	31,96	15,83	3,48	12,35	0,02	0,00	0,13	0,76	25,60	0,54	0,63	12,78	10,70	10,41

ENTIDADES	RCM	ROE	SOL1	SOL2	SOL3	SOL4	SOL5	EFG1	EFG2	CAA1	CAA2	CAA3	CR1	CR2	CR3	CR4	CR5	CR6	CR7	CR8	EFF1	EFF2	LIQ1	LIQ2	LIQ3	T1	T2	T3
C26	0,26	-0,13	41,77	98,55	0,00	0,13	0,15	167,99	4,41	79,37	67,68	32,32	32,38	13,83	74,21	13,84	7,19	6,65	0,05	0,03	-2,64	-13,41	29,21	0,43	0,50	14,69	12,67	12,19
C27	0,14	0,00	-1142,86	-11,47	0,00	0,18	0,22	97,81	4,90	126,00	100,75	-0,75	0,00	0,00	46,64	14,05	8,75	5,30	0,03	0,02	0,11	0,62	42,64	0,59	1,08	13,76	12,03	11,48
C28	0,22	-0,10	151,83	1,57	0,00	0,08	0,09	116,80	8,98	101,76	92,17	7,83	2,09	93,24	49,16	17,83	8,65	9,18	0,02	0,00	-1,40	-16,09	10,94	0,86	0,95	14,81	12,29	12,97
C29	0,24	0,00	110,32	81,08	38,42	0,27	0,37	91,73	8,86	98,69	70,67	29,33	36,23	6,31	41,38	15,55	7,37	8,18	0,02	0,02	1,01	3,24	72,51	0,63	0,88	13,01	11,71	10,95
C30	0,18	0,03	889,37	-15,85	0,00	0,16	0,19	86,65	4,41	120,62	97,71	2,29	4,96	158,50	43,12	14,15	7,31	6,84	0,09	0,05	0,70	4,34	26,48	0,72	0,96	18,03	16,17	15,96
C31	0,00	0,02	80,48	48,34	0,00	0,32	0,47	107,93	7,56	90,48	59,65	40,35	28,69	8,02	15,49	13,92	2,05	11,87	0,02	0,03	-0,93	-1,73	26,95	0,56	0,88	15,17	14,03	12,70
C32	0,15	0,02	115,49	2,39	0,00	0,11	0,13	98,84	4,30	103,21	89,98	10,02	3,11	87,25	53,03	14,01	7,04	6,97	0,03	0,01	0,06	0,45	26,31	0,64	0,86	16,03	13,84	13,72
C33	0,18	0,01	-411707,37	-25,39	0,00	0,14	0,17	101,58	3,24	118,33	98,68	1,32	3,82	239,05	51,32	14,75	7,03	7,72	0,10	0,04	-0,05	-0,35	38,33	0,56	0,68	19,09	17,14	16,90
C34	0,13	-0,22	215,68	-0,08	0,00	0,15	0,17	390,54	5,39	110,75	92,49	7,51	9,54	100,14	51,15	13,61	7,56	6,05	0,11	0,05	-4,34	-27,48	30,30	0,73	0,87	17,36	15,44	15,27
C35	0,16	0,00	335,27	5,28	0,00	0,51	1,04	113,76	5,97	167,10	77,09	22,91	21,49	85,72	27,72	10,90	5,76	5,14	0,23	0,26	-0,94	-1,42	51,42	0,56	1,20	15,95	15,28	13,60
C36	0,18	0,03	174,52	4,61	0,00	0,22	0,29	92,92	8,05	111,53	84,24	15,76	6,97	79,86	25,98	17,52	4,55	12,98	0,06	0,01	0,73	2,74	34,46	0,65	0,86	15,73	14,23	13,70
C37	0,14	0,07	602,77	-8,99	0,00	0,15	0,18	76,54	5,22	118,17	96,51	3,49	1,08	273,84	33,19	14,25	5,95	8,30	0,03	0,03	1,66	10,33	42,96	0,70	1,13	15,50	13,64	13,31
C38	0,15	0,00	161,38	43,19	0,00	0,51	1,03	118,43	7,60	132,08	61,75	38,25	30,63	15,05	13,11	10,87	2,61	8,27	0,05	0,00	-1,73	-2,31	29,73	0,77	1,64	11,90	11,22	9,53
C39	0,19	0,01	141,01	30,39	0,00	0,28	0,38	98,51	8,91	107,07	76,00	24,00	16,22	33,80	28,57	17,57	5,74	11,84	0,06	8,45	0,18	0,49	26,67	0,70	1,01	14,50	13,21	12,52
C40	0,00	0,00	3,04	47,94	86,10	0,52	0,53	-922,88	84,90	0,00	0,00	100,00	100,00	76,89	205,65	50,03	7,13	42,90	3,33	0,00	-2229,91	-439,44	0,06	-0,03	-0,04	10,85	10,19	7,39
C41	0,00	0,00	-9596,02	-1,04	0,00	0,77	3,02	384,20	4,22	375,03	82,99	17,01	0,00	0,00	0,00	1,06	0,00	1,06	0,01	0,00	-3,82	-4,02	104,47	0,76	3,44	8,71	8,45	3,91
C42	0,18	0,02	702,61	-4,20	0,00	0,32	0,47	93,63	10,54	142,15	94,00	6,00	7,28	120,26	25,62	18,29	7,18	11,11	0,10	0,00	0,76	2,24	21,27	0,75	1,21	13,19	12,05	11,44
C43	0,00	0,00	78,20	111,71	4,84	0,42	0,66	1822,38	16,82	32,30	17,84	82,16	72,24	8,67	52,09	0,72	0,43	0,29	0,07	0,00	-58,28	-42,60	165,44	0,55	1,00	11,22	10,34	5,82
C44	0,24	0,03	389,21	-12,49	0,00	0,18	0,21	110,04	5,01	117,57	94,63	5,37	4,37	165,50	39,39	16,37	6,43	9,95	0,08	0,06	-0,48	-2,61	25,20	0,65	0,81	16,23	14,49	14,20
C45	0,17	0,06	400,88	-11,55	0,00	0,15	0,18	67,01	2,66	115,01	94,59	5,41	4,08	160,52	50,58	13,90	7,01	6,89	0,07	0,05	1,38	8,45	21,52	0,63	0,79	19,57	17,71	17,37
C46	0,00	0,00	530,85	8,96	0,00	0,21	0,28	67,49	5,85	134,82	94,99	5,01	3,22	25,97	27,19	9,34	3,27	6,07	0,01	0,01	2,95	13,17	13,09	0,89	1,27	13,86	12,31	11,39
C47	0,00	0,00	201,97	15,04	0,00	0,33	0,50	89,43	8,90	144,38	76,31	23,69	17,96	42,54	7,00	14,02	0,92	13,11	0,08	0,09	1,51	2,92	37,58	0,42	0,79	12,48	11,37	9,81
C48	0,14	0,03	227,57	-7,91	0,00	0,15	0,18	87,33	6,02	108,49	89,74	10,26	2,90	150,27	44,10	15,05	7,98	7,07	0,05	0,04	0,97	5,73	26,48	0,76	1,09	14,10	12,22	12,03
C49	0,12	0,04	-1030,30	-51,82	0,00	0,10	0,11	91,05	3,72	114,25	100,35	-0,35	6,42	194,67	31,47	12,89	4,47	8,42	0,14	0,15	0,36	3,57	20,94	0,65	0,76	17,04	14,76	14,93
C50	0,16	0,03	732,28	-3,79	0,00	0,18	0,22	80,22	5,97	122,37	96,95	3,05	5,51	113,45	40,24	15,25	8,14	7,10	0,07	0,04	1,52	8,05	27,78	0,82	1,13	14,90	13,20	12,97