UNIVERSIDAD NACIONAL DE COLOMBIA

Categorización del riesgo de crédito para clientes del sector empresarial aplicando técnicas de clasificación estadística

Diana Catalina Peña Vásquez

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadísticas
Medellín, Colombia
2023

Categorización del riesgo de crédito para clientes del sector empresarial aplicando técnicas de clasificación estadística

Diana Catalina Peña Vásquez

Trabajo Final de Maestría en Ciencias - Estadística presentada como requisito parcial para optar al título de:
Magister en Ciencias - Estadística

Director:
Mauricio Alejandro Mazo Lopera
Doctor en Ciencias - Estadística

Tipo de Línea:
Profundización

Universidad Nacional de Colombia
Facultad de Ciencias, Escuela de Estadística
Medellín, Colombia
2023

Dedicatoria a mi madre que siempre estuvo apoyándome, brindándome su bendición y motivándome para sacar esta meta adelante.

Diana Catalina Peña Vásquez

Agradecimientos

En este camino de conocimientos estadísticos, una ciencia alucinante que nos lleva a encontrar mundos tan diversos en el aprovechamiento de la información, doy gracias a Dios, por mostrarme este camino y creer en que todo se puede con perseverancia y dedicación.

También un agradecimiento muy especial a mi familia que estuvo siempre apoyándome, a los profesores que aportaron en éste camino profesional y a todas aquellas personas con las que compartí conocimientos y que me aportaron en este proceso tan valioso.

Mil gracias !!!

Resumen

Categorización del riesgo de crédito para clientes del sector empresarial aplicando técnicas de clasificación estadística.

Un elemento importante en el apalancamiento de las industrias o empresas es el crédito, el cual posibilita los planes de producción, de inversión, la adquisición de bienes o servicios o simplemente el alivio de estrés financiero en las organizaciones. Esta herramienta es muy utilizada por las instituciones financieras, las cuales cuentan con metodologías, políticas e información importante para otorgar el crédito. Sin embargo, cuando son las empresas del sector real las que se enfrentan a mecanismos de venta a plazo o de financiación, no se cuenta con todos los elementos para hacer operaciones de venta estratégicas y controladas, que no solo le permitan aumentar la participación en el mercado sino también evitar posibles pérdidas económicas. Con el fin de administrar mejor lo que en el argot financiero es llamado *riesgo de crédito*, se planteará para una empresa de prestación de servicios públicos que tiene clientes del sector industrial, empresarial y comercial, una metodología para valorar el riesgo de crédito por medio de la aplicación de técnicas estadísticas de clasificación de dichos clientes. Se analizará la aplicación de algunos métodos estadísticos, tales como la regresión logística, regresión logística multinomial, regresión logística multinomial ordinal y máquinas de soporte vectorial.
Para la construcción de la metodología, se calcularon indicadores financieros, se recopiló información de morosidad interna para las empresas y se consultó información complementaria de algunos atributos geográficos y modelos alternos al interior de la compañía.

Se presentan herramientas importantes para la selección de variables, balanceo de muestras y medidas de desempeño para los modelos aplicados. Entre los resultados fue necesario realizar comparaciones individuales para los modelos que presentan variables dicotómicas y los que presentan más de dos niveles en la etiqueta. Para la variable dicotómica Impago o Default, se destacaron los modelos con balanceo Oversampling tanto para la regresión logística como para el modelo de máquinas de soporte con kernel radial. Mientras que para la variable relacionada con la clasificación en tres niveles del riesgo, los modelos con mejores resultados correspondieron al modelo de máquinas de soporte con Kernel lineal o el modelo de regresión multinomial.

Palabras clave: Riesgo de Crédito, Impago, modelos de clasificación, matrices de trancisión, indicadores financieros, Probabilidad de Incumplimiento.

Abstract

Credit risk categorization for clients in the business sector using statistical classification techniques.

An important element in the leverage of industries or companies is credit, which enables production plans, investment, the acquisition of goods or services or simply the relief of financial stress in organizations. This tool is very used by financial institutions which have methodologies, policies and important information to grant the credit. However, when it is the companies of the in the real sector, those that face mechanisms of sale in installments or financing, do not have all the elements to make strategic and controlled sales, which not only allow you to increase market share but also avoid possible economic losses.

In order to better manage what in financial jargon is called credit risk, will be propose for a company that provides public services that has clients from the industrial, business and commercial sector, a methodology to value credit risk through the application of statistical classification techniques. Will be analyzed the application of some statistical methods, such as logistic regression, multinomial logistic regression, ordinal multinomial logistic regression, and support machines vector.

For the construction of the methodology, financial indicators, internal payment age information of the companies were compiled and complementary information of some geographical attributes and alternative models within the company was consulted.

Important tools are presented for the selection of variables, sample balancing and performance measures for the applied models. Among the results, it was necessary to make individual comparisons for the models that present dichotomous variables and those that present more than two levels in the label. For the Default dichotomous variable, the models with Oversampling balancing were highlighted both for the logistic regression and for the model of support machines with radial kernel. While for the three-level rating variable, the models with the best results corresponded to the support machine model with linear Kernel or the multinomial regression model.

Keywords: Credit Risk, Default, Classification models, Transition matrices, Financial indicators, Probability of Default.

Contenido

Agradecimientos vii

Resumen ix

Contenido xi

Lista de figuras xii

Lista de Figuras xiii

Lista de tablas xiii

Lista de Tablas xiv

1 Introducción 1

2 Marco teórico 4

2.1 Consideraciones sobre el riesgo crédito . . . 4

2.2 Modelos de regresión . . . 6

2.2.1 Regresión logística binomial . . . 6

2.2.2 Regresión logística multinomial . . . 8

2.2.3 Regresión logística ordinal . . . 9

2.3 Máquinas de vectores de soporte . . . 10

2.3.1 Clasificación usando la separación por medio de hiperplano . . . 11

2.3.2 Kernel lineal . . . 11

2.3.3 Kernel polinómico . . . 12

2.3.4 Kernel radial . . . 13

2.4 Tratamiento para base de datos . . . 14

2.4.1 Métodos de selección de variables . . . 15

2.4.2 Alternativas para datos desbalanceados . . . 16

2.4.3 Submuestreo . . . 16

2.4.4 Sobremuestreo . . . 17

2.5 Evaluación de los modelos de clasificación . . . 17

2.5.1 Clasificación binaria . . . 19

2.5.2 Medidas clasificación binaria 19
2.5.3 Medidas clasificación multiclase 20
2.5.4 Medidas clasificación multiclase 22

Introducción **1**

3 Análisis exploratorio de los datos **24**
3.1 Descripción de la base de datos 24
3.1.1 Variables respuesta 24
3.1.2 Variable de riesgo de 3 niveles 24
3.1.3 Variable dicótoma Impago 26
3.2 Variables predictoras 27
3.3 Depuración y análisis descriptivo de la base de datos 30
3.3.1 Análisis descriptivo variables respuesta 30
3.3.2 Análisis descriptivo variables predictoras 31
3.3.3 Resumen Univariado con la variable Impago 43

4 Modelos de clasificación - Riesgo de crédito **45**
4.1 Preselección de variables 45
4.1.1 Algoritmos de selección - Mejor subconjunto, selección hacia adelante y selección hacia atrás 45
4.1.2 Análisis de correlación de variables y de varianza ANOVA 47
4.2 Partición de la base de datos 49
4.3 Modelo logístico binomial sin balanceo de datos 50
4.4 Modelo logístico binomial con métodos de balanceo 52
4.4.1 Modelo con metodología Submuestreo 52
4.4.2 Modelo con metodología Sobremuestreo 53
4.4.3 Modelo Multinomial con variable respuesta Clasificación de riesgo . . 55
4.4.4 Modelo Multinomial ordinal con variable respuesta Clasificación de riesgo 59
4.4.5 Modelo máquinas de soporte con variable respuesta dicótoma 61
4.4.6 Modelo máquinas de soporte con variable respuesta policótoma 65

5 Conclusiones y recomendaciones **68**

6 Referencias **70**

Lista de Figuras

2-1 Función logística. Fuente (Arias, 2011). 7
2-2 Probabilidades acumuladas: Regresión ordinal. Fuente (Agresti, 2003). 10
2-3 SVM kernel lineal. Fuente (James, Witten, Hastie, y Tibshirani, 2013). . . . 12
2-4 SVM kernel polinómico. Fuente (James et al., 2013). 13
2-5 SVM Kernel radial. Fuente (James et al., 2013). 14
2-6 Submuestreo, basada en (Gonzalez, 2019). 16
2-7 Sobremuestreo, basada en (Gonzalez, 2019). 17
2-8 Distribución de clases, basada en (Moreno Valencia, 2012). 18
2-9 Curva ROC, basada en (Moreno Valencia, 2012). 18

3-1 Componentes variable Clasificación - 3 niveles. 25
3-2 Distribución variables respuesta. 30
3-3 Variable Segmento. 31
3-4 Variable Sector. 32
3-5 Variable Buró. 32
3-6 Variable Árbol. 33
3-7 Variable Tamaño de la empresa. 33
3-8 Variable Región de operación de la empresa. 34
3-9 Variable estatus operacional. 35
3-10 Mora - análisis descriptivo. 37
3-11 Probabilidad de incumplimiento - análisis descriptivo. 38
3-12 Índice Altman - análisis descriptivo. 39
3-13 Cartera total - análisis descriptivo. 40
3-14 Activos totales - análisis descriptivo. 40
3-15 Máximo PDI - análisis descriptivo. 41
3-16 Valor provisión - análisis descriptivo. 42
3-17 Concentración endeudamiento - análisis descriptivo. 42

4-1 Selección mejor subconjunto para la variable Impago. 46
4-2 Selección hacia adelante para la variable Clasificación de Riesgo 46
4-3 Función STRATIFIED. Basada en Thomas (2020). 49

Lista de Tablas

2-1 Matriz de confusión de 2×2. 19
2-2 Matriz de confusión de 3×3. 20

3-1 Ejemplo matriz de transición. 26
3-2 Variables Respuesta. 27
3-3 Variables predictoras. 28
3-4 Variables predictoras. 28
3-5 Variables predictoras. 29
3-6 Variables Categóricas. 36
3-7 Edades de Mora. 37
3-8 Franjas Altman. 39
3-9 Resumen de variables numéricas. 43

4-1 Primera matriz de correlaciones. 47
4-2 Matriz de correlaciones variables numéricas. 47
4-3 Selección de variables. 48
4-4 Proporción de la partición Variables Respuesta. 50
4-5 Proporción de la partición Variables predictoras. 50
4-6 Modelo logístico Impago con datos sin balanceo. 51
4-7 Modelo logístico datos sin balancear. 52
4-8 Modelo logístico Impago con datos método Submuestreo. 53
4-9 Modelo logístico datos con método Submuestreo. 53
4-10 Modelo logístico Impago con datos método sobremuestreo. 54
4-11 Modelo logístico datos con método Sobremuestreo. 54
4-12 Modelo Multinomial con variable clasificación de riesgo. 57
4-13 Modelo Multinomial con variable Clasificación de riesgo - valores P. 57
4-14 Matriz de confusión modelo Multinomial, Datos entrenamiento. 58
4-15 Matriz de confusión modelo Multinomial, Datos testeo. 58
4-16 Sensibilidad modelo Multinomial. 58
4-17 Especificidad modelo Multinomial. 58
4-18 Clasificación de Riesgo en tres escalas. 59
4-19 Clasificación de Riesgo en tres escalas. 60
4-20 Matriz de confusión modelo Multinomial Ordinal, Datos entrenamiento. . . . 60

4-21 Matriz de confusión modelo Multinomial Ordinal, Datos testeo. 60
4-22 Sensibilidad modelo Multinomial Ordinal. 61
4-23 Especificidad modelo Multinomial. 61
4-24 Hiperparámetro Costo con Submuestreo. 62
4-25 Hiperparámetro Costo con Sobremuestreo. 62
4-26 Matrices de confusión para todos los kernels - Submuestreo. 63
4-27 Matrices de confusión para todos los kernels - Sobremuestreo. 63
4-28 Comparación diferentes metodologías - SVM - variable Impago. 64
4-29 Validación cruzada hiperparámentro con variable de 3 niveles. 65
4-30 (Train) - Matriz de confusión modelo SVM 3 niveles - Kernel lineal. 66
4-31 (Testeo) - Matriz de confusión modelo SVM 3 niveles - Kernel lineal. 66
4-32 (Train) - Matriz de confusión modelo SVM 3 niveles - Kernel polinomial. . . 66
4-33 (Testeo) - Matriz de confusión modelo SVM 3 niveles - Kernel polinomial. . . 66
4-34 (Train) - Matriz de confusión modelo SVM 3 niveles - Kernel radial. 66
4-35 (Testeo) - Matriz de confusión modelo SVM 3 niveles - Kernel radial. 66
4-36 Comparación diferentes metodologías - SVM - variable Impago. 67

1 Introducción

En un ambiente competitivo, cuantificar y gestionar los diferentes tipos de riesgos es uno de los objetivos más importantes en cualquier tipo de organización, puesto que permiten evitar la posibilidad de que se pueda incurrir en algún tipo de pérdida y proponer estrategias para lograr la sostenibilidad en el largo plazo.

La gestión del riesgo "... *es un método lógico y sistemático para el establecimiento del contexto, identificación, análisis, evaluación, tratamiento, monitoreo y comunicación de los riesgos asociados con cualquier actividad, función o proceso, de forma que posibilite que las organizaciones minimicen pérdidas y maximicen oportunidades*" (Linares Galván, 2010).

Cualquiera que sea el objeto social de una empresa, uno de los riesgos importantes para gestionar, es el *riesgo de crédito*, el cual ha sido abordado de forma importante por las entidades financieras. Estas no solo cuentan con metodologías para cuantificarlo, sino también con normatividad e instituciones que facilitan las políticas y pautas para gestionarlo.

Para empresas del sector real, el *riesgo de crédito* se origina desde la misma facturación de los servicios o productos que ofrece o cuando se dan alternativas de venta financiada o de venta a plazos. Sin embargo, al comparar los avances que se tienen para administrar el riesgo de crédito respecto al sector financiero, no se cuentan aún con mecanismos suficientes, regulación asociada o metodologías estandarizadas para prevenir el incumplimiento de pago.

Aún así, las empresas del sector real [1], cada vez están más interesadas en lograr su posicionamiento y crecer no solo a través sus productos y prestación de servicios, sino también a través de una colocación inteligente de las ventas por medio de decisiones oportunas y estratégicas. Evitar o administrar el riesgo de crédito, cada vez es más usual en cualquier tipo de empresa.

Con el fin determinar el riesgo de crédito para los clientes de una empresa del sector real, se propone una metodología de clasificación, en la cual se recolecte información clave asociada a los clientes tales como: la información financiera, por medio de estados financieros y construcción de indicadores, e información de cartera, en la cual se pueda capturar comportamiento de pago, morosidad entre otros atributos.

[1] Sector Real: Allí pertenecen todos los sectores económicos, exceptuando el sector financiero y monetario. `https://enciclopedia.banrepcultural.org/index.php/Sector_real/`

Cuando no se cuenta con una variable respuesta que clasifique el riesgo de crédito de manera explícita, se pueden explorar diferentes alternativas que incluyan la evaluación de indicadores de conocimiento del cliente, análisis de solvencia financiera, o historial de pago interno o externo si hay autorizaciones para obtenerlo en centrales de riesgo.

Para el caso de los empresas a analizar, se consideró una metodología empírica de la compañía, en la que se clasifica el riesgo en la siguiente escala (Riesgo Bajo, Riesgo Medio y Riesgo Alto). Verona Martel (2007), tiene en cuenta éste tipo de aplicaciones y analiza la solvencia de una entidad por medio de la relación entre la clasificación de crédito y la probabilidad de incumplimiento de las obligaciones derivadas de la deuda, es decir, a mayor clasificación de crédito, menor probabilidad de incumplimiento.

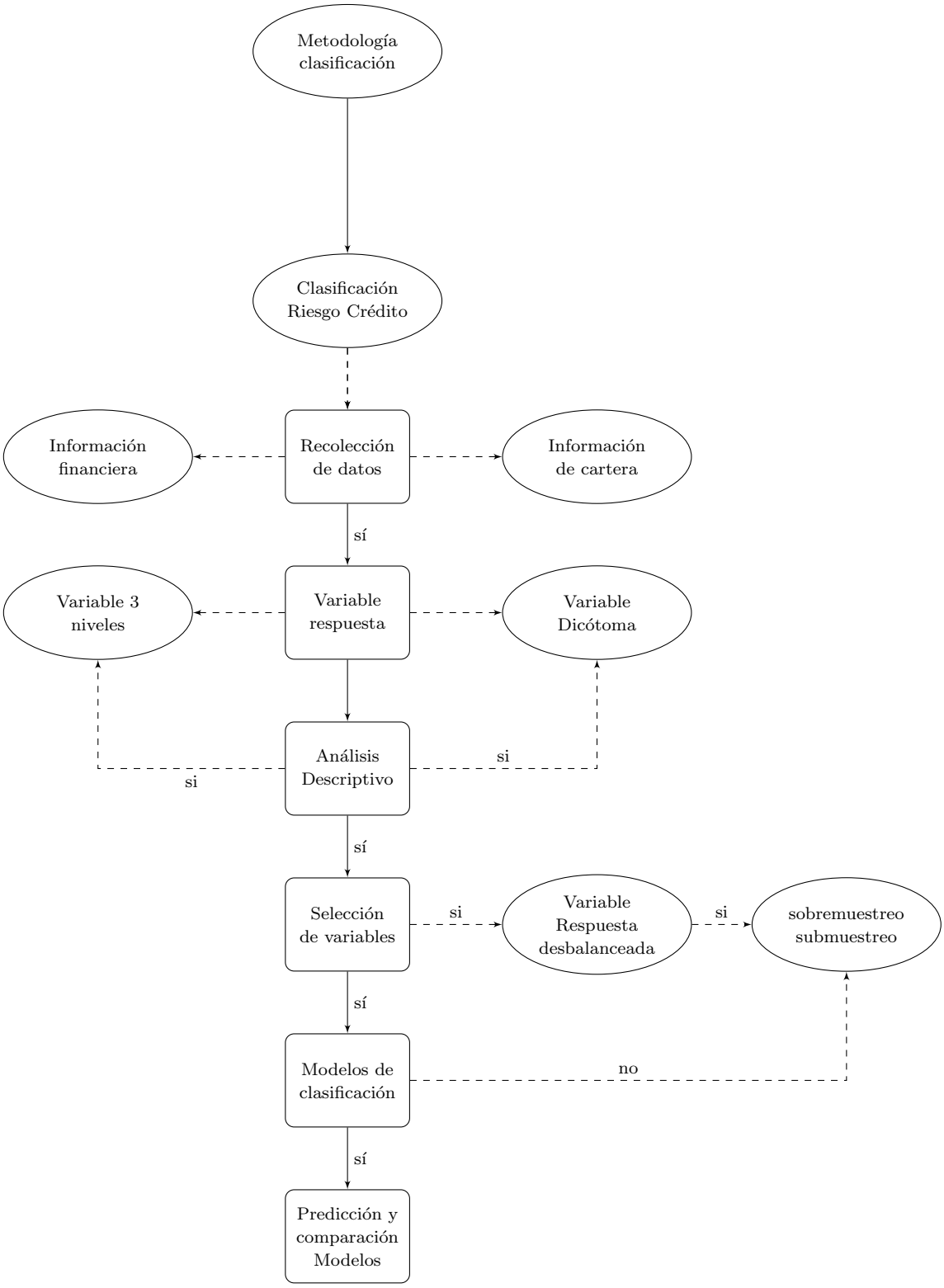
Alternativamente, cuando no hay un punto de partida apriori en la clasificación de riesgo de crédito para los clientes que se requieren analizar, el impago asociado al incumplimiento de las obligaciones de pago, se constituye en una variable respuesta importante para la clasificación del riesgo de crédito.

Para determinar la variable impago, el uso de las matrices de transición de acuerdo con las alturas de mora que presentan los clientes, permiten como veremos, obtener una respuesta dicótoma con valor igual a cero, para las empresas que no presentan el comportamiento de Impago, y de 1 para las que efectivamente sí lo tienen.

Una vez contemplada la información de los clientes, es importante realizar los diferentes análisis descriptivos e implementar métodos de selección de variables. Para los casos en los que la variable respuesta esté desbalanceada (número de observaciones pertenecientes a un grupo o clase es significativamente mayor que las pertenecientes al otro grupo), se recurren a técnicas para ajustar la distribución de clases.

A partir de lo anterior se analizan las distintas técnicas de clasificación. Para el presente trabajo de profundización, se aplica el modelo de regresión logística y máquinas de vectores de soporte para la variable dicótoma Impago, y modelos de clasificación multinomial, multinomial ordinal y máquinas de vectores de soporte para la variable de tres niveles de riesgo.

Finalmente el uso de los indicadores tales como la exactitud, sensibilidad y especificidad sobre datos de entrenamiento y testeo, permiten determinar la predicción de los diferentes modelos.


Metodología clasificación
Clasificación Riesgo Crédito
Información financiera
Recolección de datos
Información de cartera
sí
Variable 3 niveles
Variable respuesta
Variable Dicótoma
si
si
Análisis Descriptivo
sí
Selección de variables
si
Variable Respuesta desbalanceada
si
sobremuestreo submuestreo
sí
Modelos de clasificación
no
sí
Predicción y comparación Modelos

2 Marco teórico

En el presente capítulo el lector encontrará en la Sección 2.1, definiciones y consideraciones importantes sobre el riesgo de crédito, teniendo en cuenta que se trata del foco de atención para el presente trabajo, luego en la Sección 2.2, se presentan los modelos de regresión en el cual se explorará por medio del modelo de regresión logística, multinomial y ordinal, las opciones de clasificación para variables dicotómicas, de tres o más niveles de clasificación y como opción, un algoritmo de aprendizaje muy útil en este tipo de estudios (máquinas de vectores de soporte) . En el Capítulo 2.4 se presentan opciones sobre el tratamiento de los datos para la selección de las variables predictoras y balance de las clases en la muestra; finalmente en el Capítulo 2.5 la aplicación de las medidas de desempeño para comprobar la predicción de los modelos.

2.1. Consideraciones sobre el riesgo crédito

A continuación, se relacionarán algunas definiciones y variables importantes sobre el riesgo de crédito (Norman, 2010):

- **Impago**: Evento en el que el deudor deja de realizar los pagos de sus compromisos de deuda, lo que puede dar lugar a un evento de incumplimiento.
- **Tasas de recuperación (ω)**: Porcentaje de recuperación que toma valores entre 0 y 1, que se obtiene cuando hay opciones de garantía o mecanismos de cobertura de la deuda.
- **El capital expuesto (F)**: Corresponde al valor de la deuda, cuando se presenta el evento de impago, el cual se define como: $(1 - \omega)F$.
- **Evento de impago (D)**: Dado el intervalo de tiempo $[0, T]$, o período de vigencias del crédito, dicho evento está definido como $I(\tau \leq T) = 1$, el cual corresponde al momento en el que el deudor deja de atender sus compromisos de pagos. El evento de impago está dado por $D = I(\tau \leq T)$, el cual corresponde a una probabilidad entre cero y uno.
- **Morosidad de la cartera**: Dependiendo del comportamiento de pago de los clientes, de las políticas o de los referenciamientos financieros, se puede definir un punto de

morosidad para determinar la existencia del incumplimiento. Por ejemplo, a partir de los 30 días de mora.

- **Pérdida anual de cartera**: Reuniendo los términos anteriores, la pérdida anual o deterioro de la cartera, está dada por la siguiente expresión:

$$L = F(1-\omega)I(\tau \leq T). \tag{2-1}$$

  Reemplazando con el Impago, se obtienen las pérdidas anuales en cada periodo:

$$L = \sum_{j=1}^{N} F_j(1-\omega_j)D_j.$$

- **Riesgo de crédito**: Pérdidas derivadas del potencial incumplimiento de las obligaciones financieras por parte de los acreditados (Sanchez-Roger, Oliver-Alfonso, y Sanchís-Pedregosa, 2020).

La Superintendencia Financiera de Colombia, (Superintendencia Financiera, 1995), enuncia la siguiente definición de "Riesgo de Crédito" en un contexto más general: "*posibilidad de que una entidad financiera incurra en pérdidas y se disminuya el valor de sus activos, como consecuencia de que sus deudores fallen en el cumplimiento oportuno o cumplan imperfectamente los términos acordados en los contratos de crédito*".

Algunos autores han explorado desde el punto de vista de la solvencia financiera el riesgo de crédito; entre los trabajos más referenciados en la literatura, se encuentra el del autor Edward Altman (1966), quien por medio de un modelo de análisis de discriminante, y del enfoque desarrollado por Fisher (1936), utiliza indicadores de desempeño financiero para dar un puntaje y clasificar a las empresas.

Desde el sistema financiero, el riesgo de crédito también ha sido un elemento ampliamente estudiado a raíz de crisis económicas tales como las del 1998 y 2008, que dieron lugar a regulaciones y políticas de gestión como Basilea I, II y II, que plantean un conjunto de acuerdos para uniformar la regulación bancaria entre los países a nivel internacional (OVF, 2021). También desde el punto de vista contable, organismos como el IASB (International Accounting Standards Board por sus siglas en inglés)(Dictionary, 2023) , han establecido un conjunto de normas contables de aplicación internacional (NIC 39 y NIIF 9 - instrumentos financieros), que han venido evolucionando para brindar pautas para revelar en los estados financieros el deterioro de la cartera y exigiendo requisitos para registrar de forma fidedigna, los ingresos reales de las empresas. Esto exige cada vez más a todo tipo de empresas incluidas las del sector financiero, real, comercial o manufacturero, que cuenten con herramientas para identificar el riesgo de crédito de manera fiable, a fin de que se puedan tomar decisiones

controladas y estrategias de gestión preventivas.

Como un componente importante en el riesgo de crédito, es necesario determinar la presencia del evento de Impago. Para esto, no es suficiente con la identificación de un incumplimiento en un período determinado, se requiere que haya consistencia histórica en este tipo de sucesos para varios periodos de mora. Por esta razón, autores como Trueck y Rachev (2009), sugieren el uso de la aplicación de cadenas de Markov mediante matrices de transición, las cuales son útiles para encontrar los periodos o puntos de mora que dan lugar a la presencia del Impago en la carteras.

2.2. Modelos de regresión

La regresión logística, es uno de los métodos de clasificación más utilizados entre las técnicas estadísticas supervisadas; ésta permite predecir el resultado de una variable categórica en función de sus variables predictoras, las cuales pueden ser continuas o categóricas.

Existen tres tipos básicos de regresión logística (Blissett, 2017):

- **Regresión logística binomial**: En ésta regresión solo existen dos tipos de predicción de acuerdo con la variable respuesta. Para éste caso de aplicación, se calcula la probabilidad de que las empresas entren en Impago dado un perfil correspondiente, asociando éste evento, con un conjunto de variables predictoras. Con $Y = 1$, se detecta la presencia de Impago y con Con $Y = 0$, la ausencia de Impago.
- **Regresión logística multinomial**: Con esta opción podemos encontrar la predicción cuando hay más de dos niveles en la variable respuesta.
- **Regresión logística ordinal**: Permite encontrar la predicción para una variable respuesta de más de dos niveles, pero en éste caso se considera el orden de sus diferentes escalas.

2.2.1. Regresión logística binomial

La regresión logística binomial teniendo en cuenta el uso de múltiples predictores $X_1, X_2, X3, \ldots, X_p$, es una extensión de la regresión lineal simple en el cual se predice una variable respuesta de carácter binario Y (James et al., 2013). Se define de la siguiente manera:

$$Y = \frac{e^{\beta_0+\beta_1 X_1+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}}, \tag{2-2}$$

donde $\beta_0, \beta_1, \ldots, \beta_p$, son los parámetros del modelo.

Luego de estimar sus parámetros por medio del método de máxima verosimilitud, obtenemos la razón de posibilidades del suceso:

$$\frac{Y}{1-Y} = e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}. \tag{2-3}$$

La expresión (2-3), puede tomar valores entre cero e infinito. Los valores de la *razón de posibilidades* cercanos a cero y a infinito, significan probabilidades de incumplimiento muy bajas y muy altas respectivamente.

Como se aprecia en la Figura (**2-1**), cuando nos encontramos con valores de probabilidad superiores a 0,5 para la variable respuesta, el evento se clasificará con el valor de 1, pero si en el caso contrario la probabilidad es inferior a 0,5, el evento se clasificará como no presencia del mismo.

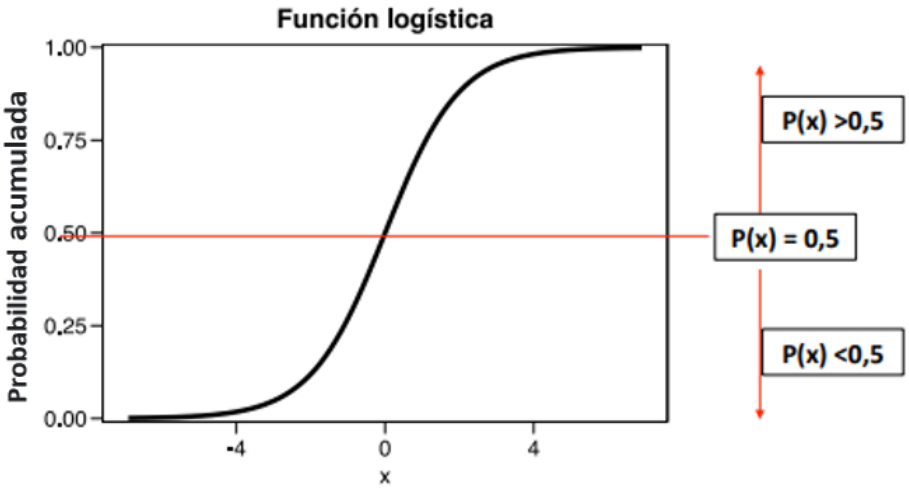


Figura 2-1: Función logística. Fuente (Arias, 2011).

Método de estimación

La intuición básica detrás del uso del método de máxima verosimilitud para ajustar el modelo de regresión logística, es estimar cada uno de los coeficientes $\beta_0, \beta_1, \ldots, \beta_p$, de forma que la probabilidad para cada individuo sea la máxima en (2-4). En otras palabras, encontrar los parámetros que en (2-4), arrojen números cercanos a uno para los individuos que pertenecen al evento observado y a un número cercano a cero, para los individuos que no pertenecen al evento (James et al., 2013).

$$l(\beta_0, \beta_1, \ldots, \beta_p) = \prod_{i:y_i=1} p(X_1^T, \ldots, X_p^T) \prod_{i:y_i=0} (1 - p(X_1^T, \ldots, X_p^T)). \tag{2-4}$$

2.2.2. Regresión logística multinomial

La regresión logística multinomial, fue introducida por Hosmer, Jovanovic, y Lemeshow (1989), como una extensión de la regresión logística binomial. Predice el modelo, cuando la variable respuesta tiene más de dos niveles (politómica), o categorías mutuamente excluyentes que forman una clase.

Bajo el supuesto de que cada uno de los n ensayos de la variable respuesta sean independientes e idénticamente distribuidas, se puede obtener como resultado cualquiera de las c categorías. Si el ensayo tiene un resultado en la categoría j, el resultado será $y_{ij} = 1$, de lo contrario el resultado será $y_{ij} = 0$, de modo que $\boldsymbol{y}_i = (y_{i1}, y_{i2}, \ldots, y_{ic})^\top$, representa ensayos multinomiales, con $\sum \pi_i = 1$, donde:

$$\pi_i = Y = \frac{e^{\beta_0+\beta_1 X_1+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p}}. \tag{2-5}$$

Sea $n_j = \sum_i y_{ij}$, el número de ensayos en la categoría j, $(n_1, n_2, , \ldots, n_c)$, los recuentos con distribución multinomial y sea $\pi_j = P(Y_{ij} = 1)$, las probabilidades en cada categoría j. La función masa de probabilidad de la distribución multinomial estará dada por la siguiente expresión (Agresti, 2003):

$$p(n_1, n_2, \ldots, n_{c-1}) \quad = \quad \left(\frac{n_!}{n_1! n_2! \cdots n_c!}\right) \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_c^{n_c}. \tag{2-6}$$

Dado $\sum_j n_j = n$, este corresponde a un espacio $c-1$ dimensional con $n_c = n-(n_1+\cdots+n_1)$, la distribución binomial es el caso especial cuando $c = 2$.

Para la estimación de los parámetros, primero se obtiene las estimaciones de π, donde la función masa de probabilidad de la distribución multinomial es proporcional a:

$$\prod_j \pi_j^{n_j}, \tag{2-7}$$

donde para todo $\pi_j \geqslant 0$ y $\sum_j \pi_j = 1$. El estimador de máxima verosimilitud (ML), corresponde a las probabilidades π_j, que maximizan (2-7).

La función de máxima verosimilitud multinomial está dada por:

$$L(\pi) = \sum_j n_j \log(\pi_j). \tag{2-8}$$

Simplificando se obtiene:

$$\frac{\hat{\pi_j}}{\hat{\pi_c}} = \frac{n_j}{n_c}. \tag{2-9}$$

Para calcular el riesgo relativo, un valor de $\pi_1 - \pi_2$ puede tener mayor importancia cuando ambos π_i, son cercanos a cero o a uno, que en el caso contrario.

El riesgo relativo está dado por $\frac{\pi_1}{\pi_2}$ que puede ser cualquier número real no negativo. Un riesgo relativo de 1, corresponde a independencia. La razón de posibilidades, para una probabilidad de éxito π, esta dado por:

$$\Omega = \frac{\pi}{(1-\pi)}. \tag{2-10}$$

Dichas probabilidades son no negativas, con $\Omega > 1$, cuando es más probable el éxito que el fracaso.

2.2.3. Regresión logística ordinal

Cuando el objetivo es establecer una clasificación de la variable categórica de forma *ordinal*, se utilizan los modelos logits **acumulados**. La probabilidad acumulada de una variable Y, es la probabilidad de que Y, sea menor que un determinado valor j, de modo que para cada categoría j, se define el *logit acumulado* de la siguiente manera (Agresti, 2003):

$$P(Y \leq j|x) = \pi_1(x) + \cdots + \pi_j(x), \quad \text{con} \quad j = 1, \ldots, J. \tag{2-11}$$

Las probabilidades acumuladas, reflejan el orden de las categorías:

$$P(Y \leq 1) \leq P(Y \leq 2) \leq \cdots \leq P(Y \leq J) = 1. \tag{2-12}$$

La función logística acumulada está definida como:

$$\begin{aligned} \text{logit}[P(Y \leq j|x)] &= \log\left(\frac{P(Y \leq j|x)}{1 - P(Y \leq j|x)}\right) \\ &= \log\left(\frac{\pi_1(x) + \cdots + \pi_j(x)}{1 - \pi_1(x) - \cdots - \pi_j(x)}\right), \end{aligned} \tag{2-13}$$

con $j = 1, ..., J$. Cada logit acumulado utiliza todas las categorías de la respuesta J. El indicador de razón de posibilidades para el logit$[P(Y \leq j|x)]$, utilizando simultáneamente todos los logits acumulados $(J-1)$, se representa de la siguiente manera:

$$\text{logit}[P(Y \leq j|x)] = \alpha_j + \boldsymbol{\beta}^T x, \quad \text{con} \quad j = 1, \ldots, J-1. \tag{2-14}$$

Cada logit acumulado tiene su propia intercepción. El término α_j**,** es creciente en j, porque $P(Y \leq j|x)$ aumenta en j para un x dado y porque el logit es una función creciente de $P(Y \leq j|x)$.

Aquí el modelo asume los mismos efectos para cada logit. La Figura (**2-2**), presenta para un $J = 4$, la curva logística con el mismo efecto para cada una de las tres probabilidades acumuladas en las 4 categorías de respuesta:

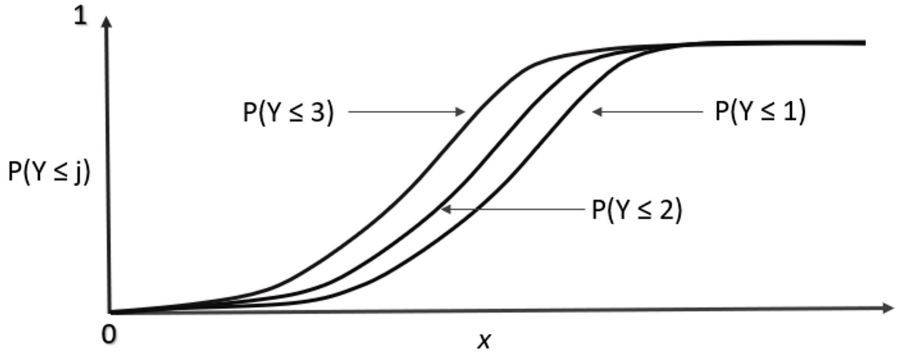


Figura 2-2: Probabilidades acumuladas: Regresión ordinal. Fuente (Agresti, 2003).

La función de verosimilitud para la estimación de los parámetros en el caso de una regresión logística ordinal es la siguiente (Agresti, 2003):

$$\begin{aligned}\prod_{i=1}^{n}\left[\prod_{j=1}^{j}\pi_j(x_i)^{y_{ij}}\right] &= \prod_{i=1}^{n}\left\{\prod_{j=1}^{j}[P(Y\leq j|x_i)-P(Y\leq j-1|x_i)\right\}^{y_{ij}} \\ &= \prod_{i=1}^{n}\left\{\prod_{j=1}^{j}\left[\frac{\exp(\alpha_j+\beta^T X_i)}{1+\exp(\alpha_j+\beta^T X_i)}-\frac{\exp(\alpha_{j-1}+\beta^T X_i)}{1+\exp(\alpha_{j-1}+\beta^T X_i)}\right]^{y_{ij}}\right\}.\end{aligned}$$

2.3. Máquinas de vectores de soporte

Otra de las técnicas matemáticas que se estudiaron para la aplicación del objeto de estudio en la propuesta de trabajo de grado, es el conjunto de algoritmos de aprendizaje supervisado llamado: máquinas de vector de soporte, desarrollado por Vapnik, Golowich, Smola, et al. (1997), muy utilizado también en problemas de clasificación.

Partiendo de un conjunto de datos de entrenamiento que corresponden a una segmentación de la base de datos para la elaboración del modelo, es posible etiquetar las clases y predecir a cual corresponde la muestra nueva. El objetivo es separar las muestras lo más amplio posible dentro de un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad, definido como el vector entre los puntos de las clases existentes más cercanas a las que se les llama *vectores de soporte*. Al exponer las muestras en dicho modelo en función de los espacios a los que pertenezcan pueden ser clasificadas de una u otra clase. Una buena separación entre las clases, permitirá la clasificación idónea.

2.3.1. Clasificación usando la separación por medio de hiperplano

En un espacio $p-dimensional$, un *hiperplano* se define como un subespacio plano y afín de dimensiones $p-1$. Por ejemplo, en un espacio de dos dimensiones, el hiperplano es un subespacio de dimensión uno, es decir, una recta. En un espacio tridimensional, un hiperplano es un subespacio de dimensión dos, es decir un plano. Para dimensiones $p \geq 3$ no es posible visualizar un hiperplano, pero el concepto de subespacio con $p-1$ tiene el mismo entendimiento (Rodrigo, 2017).

Considerando los parámetros $\beta_0, \beta_1, ..., \beta_p$ y en el caso en el que se cumple la igualdad $X = (X_1, X_2, ..., X_p)$, es decir son puntos del hiperplano, la ecuación generalizada a $p-$ dimensiones es la siguiente:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = 0. \tag{2-15}$$

Cuando los puntos $X_1, X_2, ..., X_p$, no caen exactamente al interior del hiperplano, la anterior ecuación se representará con el signo “>” ó “<”. Es decir, a un lado o al otro del hiperplano.

Las máquina de soporte vectorial (SVM de sus siglas en en inglés) es una extensión del clasificador de vectores de soporte, que resulta de ampliar el espacio de características de una manera específica, utilizando *kernels*. Los límites de separación lineales generados en el espacio aumentado se convierten en límites de separación no lineales al proyectarlos en el espacio original.

La generalización de los *kernel* se obtiene con el producto interno de dos observaciones x_i y $x_{i'}$ de la forma:

$$K(x_i, x_{i'}), \tag{2-16}$$

donde K es la función correspondiente al kernel, que identifica el tipo de separación de las observaciones. Existen diferentes tipos de *kernel* que se resumirán a continuación.

2.3.2. Kernel lineal

El clasificador para un vector de soporte con kernel lineal está representado por:

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i \left\langle x, x_i \right\rangle, \tag{2-17}$$

donde hay n parámetros por cada una por observaciones de entrenamiento: $\alpha_i, i = 1, \ldots, n$. Para la estimación de los parámetros α_i y β_0, se requiere en $\binom{n}{2}$, productos internos entre los pares de observaciones de entrenamiento $\left\langle x, x_i^{'} \right\rangle$. La ecuación (2-17), brinda el número

de pares entre un conjunto de entrenamiento. Para los casos en que α_i es diferente de cero, se obtiene los vectores de soporte para la solución:

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} x_{i'j}. \tag{2-18}$$

El kernel lineal, cuantifica la similitud de un par de observaciones utilizando la correlación de Pearson (estándar).

La Figura (**2-3**), corresponde al clasificador de máquinas de vectores de soporte (SVM) con kernel lineal. Como se observa, existen dos clases de observaciones en las cuales se trata de establecer un hiperplano que las separe de la mejor manera posible, para el conjunto de los datos de entrenamiento. Sin embargo al observar la separación que hace el hiperplano, se detalla que no separa muy bien las clases, puesto que no se alcanza a observar para todas ellas una distancia importante entre las líneas discontinuas, y muchas de la misma clase aparecen a lado y lado del hiperplano.

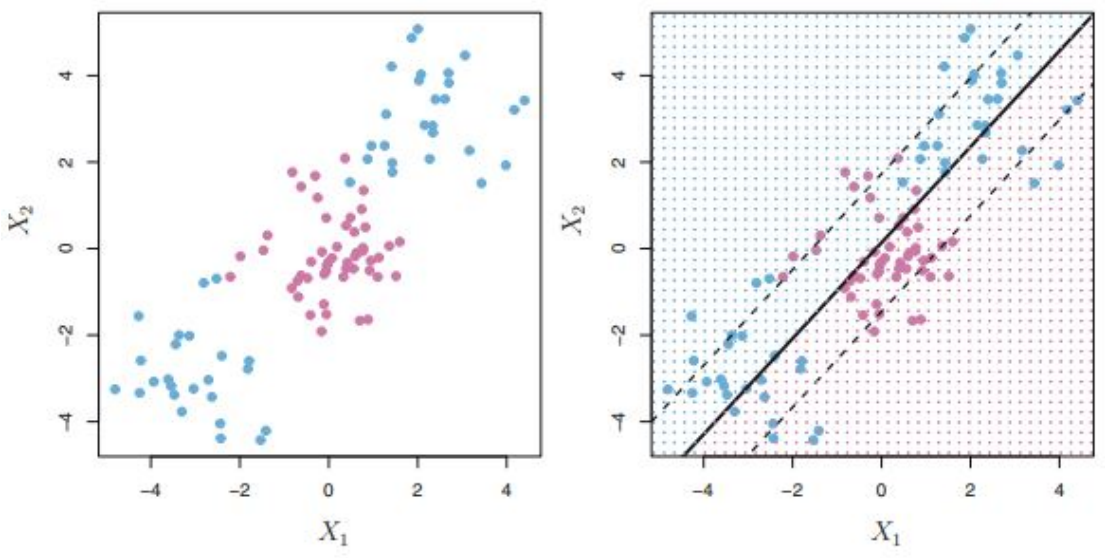


Figura 2-3: SVM kernel lineal. Fuente (James et al., 2013).

2.3.3. Kernel polinómico

De la expresión anterior, se puede reemplazar $\sum_{j=1}^{p} x_{ij}\ x_{i'j}$, con la siguiente fórmula:

$$K(x_i, x_{i'}) = \left(1 + \sum_{j=1}^{p} x_{ij}, x_{i'j}\right)^d. \tag{2-19}$$

Esto se conoce como núcleo polinomial de grado d, donde d es un número entero positivo. El uso de un núcleo de este tipo con $d > 1$, en el algoritmo del clasificador de vectores de soporte, conduce a una forma más *flexible*, del límite de decisión. Es decir, se ajusta un vector de soporte, a un clasificador en un espacio de mayor dimensión que involucra polinomios de grado d, en lugar de en el espacio de características original.

En la Figura (**2-4**), se generan diferentes límites de decisión de carácter no lineal. Cuando el parámetro d aumenta, se expanden los límites no lineales. Para éste kernel se observa, que si dicho parámetro d aumenta o los limites se prolongan demasiado, ya no se encuentra una separación idónea para las clases, presentándose así problemas de sobreajuste.

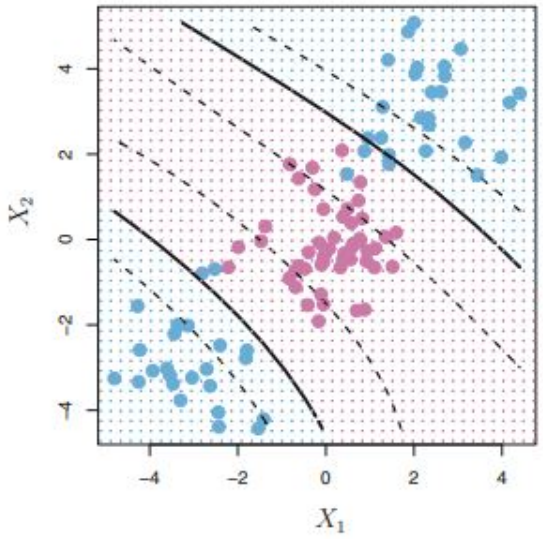


Figura 2-4: SVM kernel polinómico. Fuente (James et al., 2013).

2.3.4. Kernel radial

En éste caso cualquier kernel, es capaz de capturar el límite de decisión. Toma la siguiente expresión:

$$K(x_i, x_{i'}) = \exp\left\{-\gamma \sum_{j=1}^{p}(x_{ij} - x_{i'j})^2\right\}, \tag{2-20}$$

donde γ, es una constante positiva.

En términos de una distancia euclidiana, lo que está dentro de la sumatoria en la expresión (2-20), puede traer valores significativamente grandes y al aplicar el exponencial al parámetro negativo γ, el resultado se traducirá en una distancia lo suficientemente pequeña. Para

el caso del kernel radial, γ es el parámetro encargado de controlar la separación. Al observar la Figura (**2-5**), los puntos violetas rodeados por el hiperplano, identifican una separación ideal del resto de las observaciones de color azul.

Sin embargo en este caso, pueden encontrarse diferentes kernels para los cuales se requiere explorar por medio de validación cruzada, diferentes opciones de hiperparámetros que den lugar al valor óptimo. El kernel radial tiene un comportamiento muy local, ya que las observaciones de entrenamiento cercanas inciden bastante bien en la etiqueta de clase de una observación de prueba (James et al., 2013).

Existen diferentes posibles kernels en el proceso, y para encontrar el valor óptimo del hiperparámetro, se requiere aplicar validación cruzada. El kernel radial tiene un comportamiento muy local, ya que las observaciones de entrenamiento cercanas, inciden bastante bien en la etiqueta de clase de una observación de prueba.

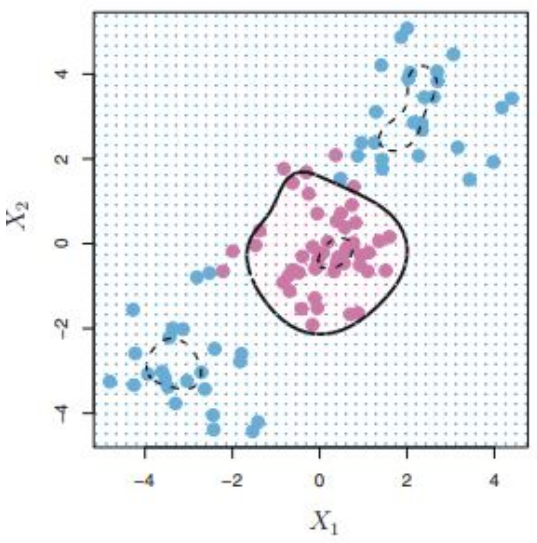


Figura 2-5: SVM Kernel radial. Fuente (James et al., 2013).

2.4. Tratamiento para base de datos

Teniendo en cuenta como veremos en la exploración de los datos, la proporción en la variable respuesta con más individuos enmarcados con una etiqueta que en otra, requerirá el uso de técnicas para balancear la muestra objetivo. En esta sección se describirá algunos métodos usados para la selección de variables, al igual que opciones de muestreo para nivelar las observaciones en las clases dicotómicas.

2.4.1. Métodos de selección de variables

Permiten seleccionar de un conjunto de variables regresoras, aquellas que son más relevantes para explicar la variabilidad de la variable respuesta. Para el presente trabajo de aplicación se utilizaron las siguientes:

- *Mejor Subconjunto*: En el cual se realizan diferentes iteraciones para encontrar los mejores subconjuntos de modelos para una, dos, tres o más variables (incluyendo el intercepto). De acuerdo con la dimensionalidad de la base de datos, podríamos definir un k determinado de variables para realizar el proceso iterativo. Para dicho algoritmo se utilizó la función del lenguaje Rstudio "regsubsets" del paquete "leaps"(Lumley, 2020).

- *Selección hacia adelante*: Algoritmo que va agregando variable por variable de manera secuencial del conjunto total de variables, iniciando con las que mayor correlación tenga con la variable respuesta, hasta identificar aquellas que efectivamente aportan o ponderan más en el modelo. Al igual que el anterior, se utiliza la función "regsubsets" con el método "forward".

- *Selección hacia atrás*: Considera inicialmente todas las variables y van descartando las variables que menos aportan al modelo en relación con la variable respuesta. Para este caso se utiliza también la función "regsubsets", pero esta vez con el método "backward".

Con el fin de determinar cuales es el conjunto de variables que más se ajustan a la variable respuesta, se utilizan las siguientes medidas de sobreajuste, eligiendo gráficamente aquellas que menor valor de Cp, AIC y BIC presentan, o para el caso del $R^2_{Ajustado}$, el menor valor de ésta medida.

Para un modelo con d variables regresoras:

$$C_p = \frac{1}{n}\left(SS_{res} + 2d\hat{\sigma^2}\right), \tag{2-21}$$

$$AIC = \frac{1}{n\hat{\sigma^2}}\left(SS_{RES} + 2d\hat{\sigma^2}\right), \tag{2-22}$$

$$R^2_{Ajustado} = 1 - \frac{\frac{SS_{RES}}{(n-d-1)}}{\frac{SS_T}{n-1}}. \tag{2-23}$$

De manera complementaria, el análisis de varianza ANOVA permite validar la selección. Allí se utiliza el estadístico de prueba F, iniciando con un modelo reducido e incorporando poco

a poco las variables en modelos más complejos. La hipótesis nula para este caso, será que el modelo final es suficiente para explicar los datos, versus la hipótesis alternativa que sugiere que un modelo saturado o complejo, no aportará al modelo (James et al., 2013).

2.4.2. Alternativas para datos desbalanceados

Un conjunto de datos o muestras de datos desbalanceados, se presenta cuando un número de observaciones correspondientes a un conjunto o clase, es muchísimo mayor que el resto de las demás clases. Lo anterior puede representar problemas de sobreajuste, cuando los modelos que se apliquen den como resultados una predicción para la clase que más represente información.

Para solucionar el *desbalance*, se puede optar por completar la muestra, de tal manera que se obtenga un mayor aporte para la clase minoritaria. Sin embargo, lo anterior no siempre es viable en ciertos eventos en los que la distribución de la muestra, no evidencia una mejora en el balance.

Como alternativa al desbalance en los datos, se han utilizado herramientas de remuestreo del conjunto de datos, en el que se equilibre la información tomando más datos de la clase minoritaria o disminuyendo las muestras de la clase mayoritaria para obtener así un equilibrio en las clases.

2.4.3. Submuestreo

Modifica la distribución del conjunto de datos, disminuyendo el número de muestras de la clase mayoritaria. Tiene como desventaja que puede eliminar muestras potencialmente importantes en los modelos de clasificación. Se recomienda en los casos en los que el conjunto de datos es de gran tamaño reduciendo la muestra y por ende, mejorando el tiempo de procesamiento, ver Figura (**2-6**).

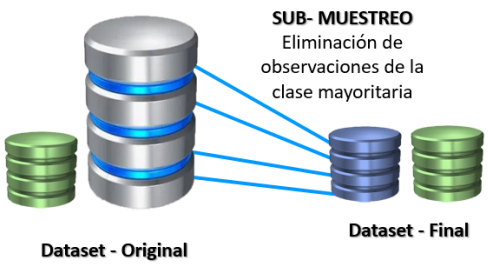


Figura 2-6: Submuestreo, basada en (Gonzalez, 2019).

2.4.4. Sobremuestreo

Modifica la distribución del conjunto de datos, aumentando el número de muestras de la clase minoritaria. Tiene como desventaja que pueden crearse muestras que no necesariamente provengan de la distribución original, generando así distorsión, también puede generar sobreajuste o problemas de procesamiento durante el proceso de clasificación, ver Figura (**2-7**).

Figura 2-7: Sobremuestreo, basada en (Gonzalez, 2019).

Otros Algoritmos utilizados son los algoritmos Híbridos, que consisten en combinar ambas técnicas, ver (Hadad, Evin, y Drozdowicz, 2009).

2.5. Evaluación de los modelos de clasificación

Teniendo en cuenta que los modelos que se exploran corresponden a modelos de clasificación, las indicadores de desempeño para comparar los resultados se dan a través de las medidas de exactitud, sensibilidad y especificidad.

Para analizar el poder discriminatorio de los modelos que se aplicarán, la curva ROC (acrónimo de Receiver Operating Characteristic), se usa como una representación gráfica de la sensibilidad frente a la especificidad para un sistema clasificador. Es una herramienta muy útil, puesto que ayudará a evaluar las observaciones que representan el caso positivo, versus los que no lo presentan.

La Figura (**2-8**), muestra que no se puede observar una separación clara en ambas distribuciones (clase positiva y clase negativa). Ambas distribuciones se traslapan y como se observa en el medio, hay un punto C que corresponde al umbral o punto de corte ideal de clasificación.

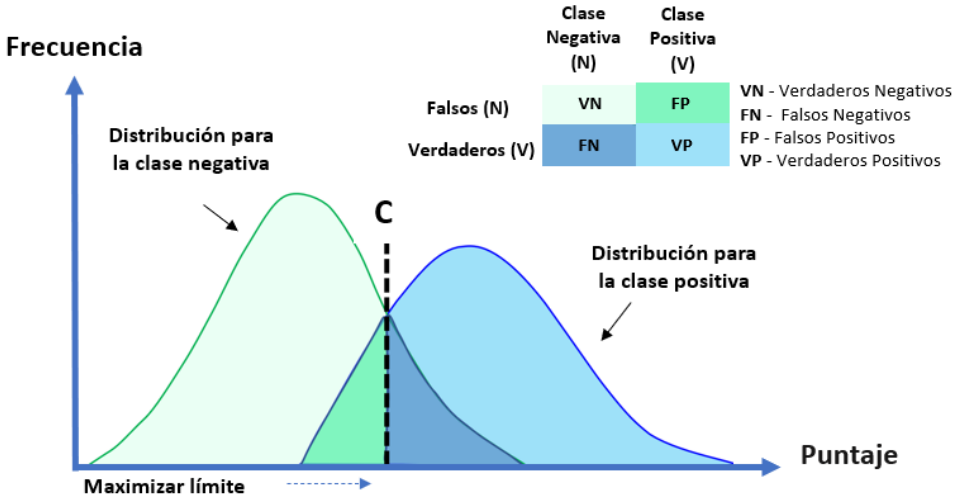


Figura 2-8: Distribución de clases, basada en (Moreno Valencia, 2012).

La Figura (**2-9**), presenta la curva ROC que se construye a través de una matriz con tablas frecuencias. Las predicciones dependen de un umbral C y en el caso que la probabilidad supere dicho umbral se clasifica en la categoría de la clase observada, asignándole el valor de 1.

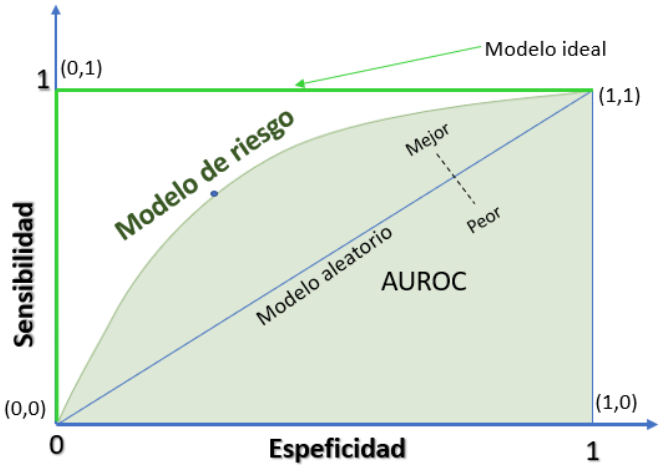


Figura 2-9: Curva ROC, basada en (Moreno Valencia, 2012).

En la esquina superior izquierda de la Figura (**2-9**), el arreglo (0,1) representa el método ideal de predicción, con un 100 % de la sensibilidad y un 100 % de especificidad (ningún falso positivo). Por el contrario, una clasificación totalmente aleatoria, daría un punto a lo largo

de la línea diagonal comprendida desde el punto $(0, 0)$ hasta $(1, 1)$, e indicaría que el modelo no tendría poder discriminatorio. Es ideal que el área bajo la curva ROC (AUC), sea lo más amplia posible para acercarnos a un umbral o punto de corte en el que se pronostique bien al individuo que representa la clase positiva, cuando realmente se presentó, y aquellas que pertenecen a la clase negativa, cuando efectivamente lo son. Por lo tanto, entre mayor sea el valor del AUC, mejor será el pronóstico del modelo. Un valor cercano al 0,5 (50 % del área total) será un valor no recomendado.

2.5.1. Clasificación binaria

Considerando la variable respuesta de tipo Binario. Impago (1) y no Impago (0), la matriz de confusión como se presenta en la Tabla (**2-1**):

Tabla 2-1: Matriz de confusión de 2×2.

Clase Actual	**Predicción Modelo**	
	Negativos (0)	**Positivos (1)**
Negativos (0)	Verdaderos Negativos (VN)	Falsos Positivos (FP)
Positivos (1)	Falsos Negativos (FN)	Verdaderos Positivos (VP)

donde,

- **VP** - Si el caso es positivo y se clasifica positivo se dice positivo cierto.
- **FN** - Si el caso es positivo y se clasifica negativo se dice falso negativo.
- **VN** - Si el caso es negativo y se clasifica negativo se dice negativo cierto.
- **FP**- Si el caso es negativo y se clasifica positivo se dice falso positivo.

2.5.2. Medidas clasificación binaria

- **Exactitud**: Es la capacidad que tiene el modelo de clasificar la clase positiva y la clase negativa. Se formula de la siguiente manera:

$$\text{Exactitud} = \frac{VP + VN}{VP + VP + FP + FN}. \tag{2-24}$$

- **Sensibilidad**: Corresponde a la proporción de verdaderos positivos. Es decir la capacidad de identificar los casos positivos en el caso de que si se presenten:

$$\text{Sensibilidad} = \frac{VP}{VP + FN}. \tag{2-25}$$

- **Especificidad:** Es la capacidad que tiene el modelo de identificar la clase negativa, cuando realmente se den. Es decir de identificar los Verdaderos negativos (VN):

$$\text{Espeficidad} = \frac{VN}{VN + VP}. \tag{2-26}$$

- **Exactitud equilibrada:** Es una medida alterna útil, cuando las clases están desbalanceadas (Statistical Odds & Ends, 2020).

$$\text{Sensibilidad} = \frac{\text{Sesnsibilidad} + \text{Exactitud}}{2}. \tag{2-27}$$

2.5.3. Medidas clasificación multiclase

Aplica para una clasificación multiclase ejemplo (A, B y C), 3×3, Tabla (**2-2**).

Tabla 2-2: Matriz de confusión de 3×3.

Pred.Modelo	**Clase Actual**		
	CLASE (A)	**CLASE (B)**	**CLASE (C)**
CLASE (A)	VP_A	E_{BA}	E_{CA}
CLASE (B)	E_{AB}	VP_B	E_{CB}
CLASE (C)	E_{AC}	E_{BC}	VP_C
	TA	**TB**	**TC**

Con el fin de explicar las medidas multiclases, por facilidad, se ilustrará con las escalas A, B y C. En la Tabla (**2-2**), cada componente de la matrix, se describe así:

Elementos de la diagonal:

- VP_A - Número de muestras positivas verdaderas en la clase A. Es decir, número de muestras que se clasificaron correctamente de la muestra A.
- VP_B - Número de muestras positivas verdaderas en la clase B. Es decir, número de muestras que se clasificaron correctamente de la muestra B.
- VP_C - Número de muestras positivas verdaderas en la clase C. Es decir, número de muestras que se clasificaron correctamente de la muestra C.

Elementos por fuera de la diagonal:

- EB_A - Muestras de la clase A, que se clasificaron de manera errónea en la clase B.
- EC_A - Muestras de la clase A, que se clasificaron de manera errónea en la clase C.
- EA_B - Muestras de la clase B, que se clasificaron de manera errónea en la clase A.
- EC_B - Muestras de la clase B, que se clasificaron de manera errónea en la clase C.
- EA_C - Muestras de la clase C, que se clasificaron de manera errónea en la clase A.
- EB_C - Muestras de la clase C, que se clasificaron de manera errónea en la clase B.

Definiciones importantes para las medidas de clasificación.

Falsos Negativos:

- FN_A **- Falso negativo de la clase A**: Se calculará $FNA = EAB + EAC$. Esto quiere decir que es la suma de todas las muestras de la clase A, que se clasificaron incorrectamente como B o C.
- FN_B **- Falso negativo de la clase B**: Se calculará $FNB = EBA + EBC$. Esto quiere decir que es la suma de todas las muestras de la clase B, que se clasificaron incorrectamente como A o C.
- FN_C **- Falso negativo de la clase C**: Se calculará $FNC = ECA + ECB$. Esto quiere decir que es la suma de todas las muestras de la clase C, que se clasificaron incorrectamente como B o C.

Falsos positivos:

- FP_A **- Falso negativo de la clase A**: $FP_A = EBA + ECA$. Esto quiere decir que es la suma de todas las muestras de la clase A, que se clasificaron correctamente como B o C.
- FP_B **- Falso negativo de la clase B**: $FP_B = EAB + ECB$. Esto quiere decir que es la suma de todas las muestras de la clase B, que se clasificaron correctamente como A o C.
- FP_C **- Falso negativo de la clase C**: $FP_A = EAC + EBC$. Esto quiere decir que es la suma de todas las muestras de la clase A, que se clasificaron correctamente como B o C.

Es decir, que para cualquier Falso Negativo, la clase que se encuentra en una columna, se calcula sumando los errores de dicha clase.

En el caso de los Falsos Positivos, para cualquier predicha que se encuentre en la fila representa la suma de todos los errores en esa fila.

Con una matriz con dimensiones $m \times m$, hay m correctas clasificaciones y $M2 - m$ posibles errores.

- VN_A **- Verdaderos negativos clase A**: Se calcula como $VN_A = VP_B + EC_B + EB_C + VP_c)$.
- VN_B **- Verdaderos negativos clase B**: Se calcula como $VN_B = VP_A + EC_A + EA_C + VP_C$.
- VN_C **- Verdaderos negativos clase C**: Se calcula como $VN_C = VP_A + EB_A + EA_B + VP_B$.

Totalidad Clases, Se obtiene sumando todos los componentes de la matriz o de las clases $TA + TB + TC$.

2.5.4. Medidas clasificación multiclase

Exactitud: Es la capacidad que tiene el modelo de clasificar correctamente las 3 clases. Se formula de la siguiente manera:

$$\text{Exactitud} = \frac{VP_A + VP_B + VP_C}{TA + TB + TC}. \tag{2-28}$$

Sensibilidad: La sensibilidad se calcula para cada clase:

$$\text{Sensibilidad (A)} = \frac{VP_A}{VP_A + FN_A}. \tag{2-29}$$

$$\text{Sensibilidad (B)} = \frac{VP_B}{VP_B + FN_B}. \tag{2-30}$$

$$\text{Sensibilidad (C)} = \frac{VP_C}{VP_C + FN_C}. \tag{2-31}$$

Especificidad: Al igual que la sensibilidad se calcula para cada clase:

$$\text{Especificidad (A)} = \frac{VN_A}{VN_A + FP_A}. \tag{2-32}$$

$$\text{Especificidad (B)} = \frac{VN_B}{VN_B + FP_B}. \tag{2-33}$$

$$\text{Especificidad (C)} = \frac{VN_C}{VN_C + FP_C}. \tag{2-34}$$

3 Análisis exploratorio de los datos

3.1. Descripción de la base de datos

La base de datos objeto de estudio, cuenta con 40 variables y 12.175 registros que corresponden cada una a empresas atendidas por la entidad prestadora de servicios públicos. Para los análisis, se toma la información del periodo 2021-12 (Diciembre del 2021). Es considerado éste mes como referencia, dado que es un mes de cierre fiscal y contiene información financiera completa y auditada para la mayoría de las empresas que están obligadas a reportar información de cierre de año.

3.1.1. Variables respuesta

Cada una de las 12.175 empresas presenta dos tipos de variables respuesta: **Impago**, que posee una clasificación binaria 0 y 1, para No Impago e Impago respectivamente, y **Clasificación de Riesgo** con escalas de Riesgo Bajo, Riesgo Medio y Riesgo Alto.

3.1.2. Variable de riesgo de 3 niveles

Dicha metodología se asemeja a un clasificación de riesgo de crédito que en escalas ordinales suelen trabajar las calificadoras de riesgos y algunas entidades financieras. Es una metodología que se ha construido de manera empírica, en la cual no se considera de manera a priori elementos o técnicas de clasificación estadística. Allí se computaron los siguientes dos componentes que se resumen en un plano cartesiano para reflejar tres escalas de Riesgo: Riesgo Alto, Riesgo Medio y Riesgo bajo. Cada componente como veremos a continuación, Figura (**3-1**), tiene unos rangos específicos que al promediar, da como resultado, las escalas mencionadas.

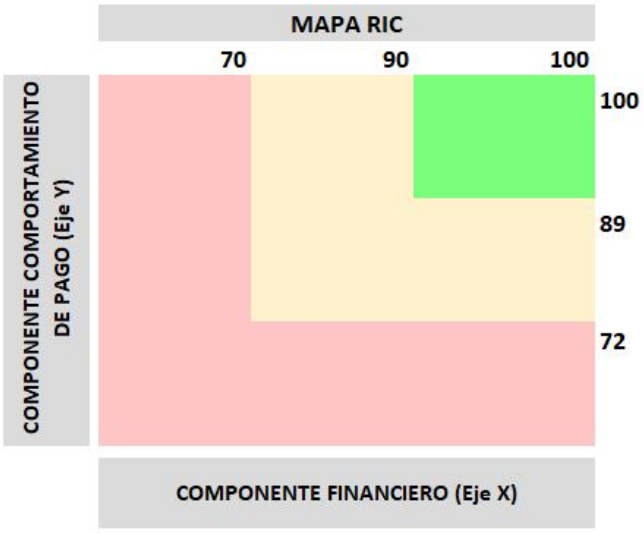


Figura 3-1: Componentes variable Clasificación - 3 niveles.

- *Componente de pago*: El resultado final se ubica en el eje Y del plano cartesiano y considera el score (puntaje) de la provisión de cartera. Es obtenido de modelo propio de la compañía.

- *Componente de financiero*: El resultado final se ubica en el eje X del plano cartesiano. Aquí se ponderan tres elementos.

  - El índice Altman que corresponde a un modelo discriminate propuesto por Edward Altman en el año 1968 (Trujillo Ospina, Belalcázar Grisales, et al., 2016).
  - El resultado de la utilidad operacional positiva o no en los estados financieros.
  - El estatus de la empresa en términos de si la empresa opera normalmente o si se acoge a alguno de los regímenes de insolvencia empresarial en la república de Colombia - ley 1116 de 2006 (De la República, 2006).

Para cada una de los dos anteriores componentes, se realiza un escalado, en el que se acotaron los datos considerando el límite máximo y el mínimo de la variable. Se obtiene para ambos un resultado de cero a cien:

$$\text{Variable}_{\text{escalada}} = \frac{X - X_{min}}{X_{max} - X_{min}}. \tag{3-1}$$

Finalmente, se promedia los dos componentes y se clasifican en riesgo Alto, Medio o Bajo de acuerdo con escalas definidas de manera empírica, según la dispersión de puntos encontrados en el plano cartesiano.

3.1.3. Variable dicótoma Impago

La segunda variable que se construyó durante el proceso de análisis del presente estudio, teniendo en cuenta que la literatura del análisis del riesgo de crédito, esta asociado de manera usual al **Impago** y a su probabilidad de impago. Se consideró un recorrido de los últimos doce meses anteriores a Diciembre 2021. Aquellas empresas que presentaron incumplimiento, se les asignó el valor de uno (1), por su parte, aquellas empresas que no presentaron incumplimiento, se les asignó el valor de cero (0). Dicha variable, fue construida a partir del comportamiento mes a mes del rodamiento de cartera, por medio del análisis de matrices de transición.

En la Tabla (**3-1**) se considera para cada empresa, la comparación de la mora en el mes actual con la que presentaba en el mes inmediatamente anterior. Allí el mes anterior se ubica en el rótulo horizontal de las edades y el mes actual en el rótulo vertical de las edades. Se resume en las celdas internas, el rodamiento para los clientes, el cual puede reflejar tres tipos de comportamientos:

- Cantidad de empresas que conservan la misma morosidad o la misma franja de mora en los meses de comparación. (Franja Amarilla).
- Cantidad de empresas que se trasladan a una edad mayor en los meses de comparación. (Franja rosa).
- Cantidad de empresas que se trasladan a una edad menor en los meses de comparación. (Franja verde).

Tabla 3-1: Ejemplo matriz de transición.

		Mes Anterior					
		Al día	**1-30**	**31-60**	**61-90**	**May90**	**Total**
Mes actual	**Al día**	500	7	0	0	0	**507**
	1-30	30	45	2	0	0	**77**
	31-60	8	30	7	1	0	**46**
	61-90	1	0	55	2	1	**58**
	May90	1	1	2	6	1	**11**
	Total	**540**	**83**	**66**	**9**	**1**	**699**
	Deterioro	7 %	37 %	86 %	67 %	100 %	65,34 %

Finalmente se obtuvo un comportamiento histórico de rodamiento, promediando 12 matrices de los meses anteriores a diciembre 2021. Allí se calculó el nivel de deterioro en términos de la cantidad de empresas que en cada edad de mora se ubicaron en la franja rosa en comparación con el cantidad total de empresas en la franja de color. Aquella

franja en la que el porcentaje de rodamiento supera el 50 %, es considerada como el referente de Impago. Para la base de datos analizada, el punto de incumplimiento se detectó a partir de los 60 días de mora.

Con la anterior información, se realizó la distinción de las empresas que presentaron dicho incumplimiento, marcando con "1" el resultado de la variable respuesta y con "0" aquellas empresas que no presentaron incumplimiento.

A continuación la Tabla (**3-2**), expone las características de ambas variables respuesta:

Tabla 3-2: Variables Respuesta.

Variables	**Grupo**	**Variables**	**Descripción**	**Tipo de variable**
Respuesta	De tres Niveles	Calificación financiera de riesgo	Resultado del método empírica de clasificación de riesgo en tres escalas: R. Bajo (1), R.Medio (2) y R.Alto (3)	**Cualitativa Ordinal**
	De dos niveles	Impago últimos 12 meses	Impago (1), No Impago (0). Análisis de matriz de transición	**Cualitativa, Dicotómica**

3.2. Variables predictoras

Las variables analizadas se clasifican en tres grupos:

- *Atributos de los clientes*: Asociada con características de ubicación, sector empresarial, estado de operación y tipo de segmento al que pertenece la empresa. Este último, hace parte de una clasificación interna de la compañía, basada en los ingresos originados por los clientes y que es útil para la focalización en la atención comercial de la compañía.
- *Información financiera*: Tomada del repositorio de información de la empresa. Para el periodo fiscal diciembre 2021 y de los estados financieros de situación financiera y estado de resultados, se calcularon algunos indicadores financieros que de acuerdo a políticas internas han sido relevantes en estudios de otorgamiento y cuyos rubros puedan ser obtenidos de manera integral para todas las empresas a analizar. Lo anterior es fundamental, debido a que en la literatura financiera, existen diversidad de indicadores de liquidez, solvencia y rentabilidad, pero no todos pueden ser obtenidos fácilmente y de manera uniforme en los estados financieros.
- *Información de cartera y modelos alternos*: Incluye variables como la mora, el valor facturado, la cartera y modelos alternos (provisión y otros de clasificación de pago).

A continuación las Tablas (**3-3**), (**3-5**) y (**3-4**), exponen las características de las variables predictoras de acuerdo con los grupos mencionados anteriormente:

Tabla 3-3: Variables predictoras.

Variables	Grupo	Variables	Descripción	Tipo de variable
Predictoras	Atributos de las empresas	Segmento	*Segmentos*: Empresa, Grandes Clientes y Gobierno. Dependen del nivel de ingresos que aportan (políticas de la compañía).	Cualitativa nominal
		Regiones	Ubicación regional de la empresa	Cualitativa nominal
		Sector	Basado en los macro sectores: Comercio, Manufactura y servicios	Cualitativa nominal
		Tamaño Empresas	Dec. 957 de 2019 del Gno Nal, define el tamaño según el nivel de ingresos y tipo de actividad	Cualitativa ordinal
		Estado Operacional	Incluye la clasificación de la empresa: Operacional, o si está en procesos de insolvencia o reestruct. (ley 1116 de 2006)	Cualitativa nominal

Tabla 3-4: Variables predictoras.

Variables	Grupo	Variables	Descripción	Tipo de Variable
Predictoras	Info de cartera, facturación y otros modelos	Mora Actual	Edad de mora en la que se encuentra la empresa. Son 15 edades de tipo ordinal	**Contínuas**
		Valor Cuenta de Cobro	Valor facturado en el mes para la empresa	
		Cartera Total	Incluye tanto la cartera corriente como la diferida de la empresa	
		Cartera diferida	Cartera diferida total asociada a la empresa	
		ke	Capital expuesto base para el cálculo de la provisión de cartera	
		Prov	Valor de provisión de cartera	
		Máx Pdi-Epm	Pérdida dado el incumplimiento. Se consideró el máximo por empresa	
		Pi-Indiv	Probab de incumplimiento. Se expresa en términos porcentuales. Es uno de los componentes del cálculo de pérdidas esperadas - norma NIIF 9	
		Buró	Comportamiento de pago en el último Año, para los servicios individuales prestados a los clientes. Niveles, donde (1) es el de mejor calificación, (2) es el de calificación media y (3) el de menor calificación	**Categórica ordinal**
		Árbol De Clasificación	Modelo interno para la segmentación de los clientes en la gestión de la cobranza. Los primeros niveles requieren menos esfuerzo de gestión y los últimos más	

Nota: Teniendo en cuenta la variabilidad de los datos, variables como la cta de cobro, carteras, Ctal Expuesto, valor provisión y valores primarios de los estados de resultados: Activos,pasivos y patrimonio, se transforman a logaritmo, para lograr capturar la dispersión de las variables.
Para la variable Altman, se toma como referencia la versión Z_2, Berrío Guzmán y Cabeza de Vergara (2003).

Tabla 3-5: Variables predictoras.

Variables	Grupo	Variables	Descripción	Tipo de variable
Predictoras	Información Financiera	Activos Totales	Obtenidos del estado de situación financiera y del estado de resultados	**Contínuas**
		Activos Corrientes		
		Total Patrimonio		
		Pasivos Totales		
		Pasivos Corrientes		
		Total Ingr Operativo		
		Ebit		
		Ganancia Neta		
		Índice Altman. En tres escalas.	$Z_2 = 6{,}56.\frac{\text{Capital de trab}}{\text{Total Activos}} + 3{,}26.\frac{\text{Utilidad retenidas}}{\text{Total Activos}} + 6{,}72.\frac{\text{UAII}}{\text{Total Activos}} + 1{,}05.\frac{\text{Tot Patrimonio}}{\text{Pasivo Total}}$	
		Razón Corriente	$Raz\ cte = \frac{\text{Activo Corriente}}{\text{Pasivo Corriente}}$	
		Solvencia Patrimonial	$Sol\ Pat = \frac{\text{Total Patrimonio}}{\text{Pasivo Total}}$	
		Concentración Endeudamiento	$Con\ E = \frac{\text{Pasivo Corriente}}{\text{Pasivo Total}}$	
		Capital de Trabajo	$\text{CT} = \text{Activo Cte} - \text{Pasivo Cte}$	
		Eficiencia Operativa	$\text{EO} = \frac{\text{Ingreso Operativo}}{\text{Activo No Corriente}}$	
		Apalancamiento Corto Plazo	$Ap\ CP = \frac{\text{Pasivo Corriente}}{\text{Total Patrimonio}}$	
		Capital Neto de Trabajo	$KNW = \frac{\text{Activo Cte-Pasivo Cte}}{\text{Pasivo Cte}}$	
		Rentabilidad del Activo	$ROA = \frac{\text{Ganancia Neta}}{\text{Total Activos}}$	
		Rentabilidad del Patrimonio	$ROE = \frac{\text{Ganancia Neta}}{\text{Total Patrimonio}}$	
		Rentabilidad del Activo	$ROA = \frac{\text{Ganancia Neta}}{\text{Total Activos}}$	
		Margen Neto	$MgNeto = \frac{\text{Ganancia Neta}}{\text{Ingreso Operativo}}$	
		Endeudamiento	$E = \frac{\text{Pasivo Total}}{\text{Total Activos}}$	
		Marg Solvencia	$MgSolv = \frac{\text{Pasivo Total}}{\text{Total Patrimonio}}$	

3.3. Depuración y análisis descriptivo de la base de datos

Como se mencionó anteriormente, la muestra cuenta con un total de 12.175 empresas. Sin embargo, luego de la realizar auditoría de variables en su calidad y datos faltantes, se llega a un total de 11.983 empresas a analizar.

En cuanto a las variables de la base de datos, se cuentan con 40 variables que incluyen dos variables respuestas y el resto corresponden a variables predictoras relacionadas con la cartera, atributos de las empresas y variables financieras mencionadas en el aparte anterior.

A continuación se detallará el análisis descriptivo para las variables respuesta y predictoras.

3.3.1. Análisis descriptivo variables respuesta

La Figura (**3-2**) cuadrante izquierdo, corresponde a la clasificación que arroja el método empírico. Allí se observa que la mayoría de las empresas para el período de análisis, fueron clasificadas en riesgo Bajo y Medio con un 65,34 % y 32,6 % respectivamente. Del lado derecho, observamos la clasificación obtenida a través de las matrices de transición con la variable Impago, en el cual se obtiene un 83,35 % en la categoría No Impago (0) y un 16,65 % en la categoría Impago (1). Si bien no se observa problemas graves de desbalance en la base de datos, se explorará una metodología de balanceo para la clasificación dicótoma Impago, con el fin de evaluar si hay mejores resultados en la clasificación.

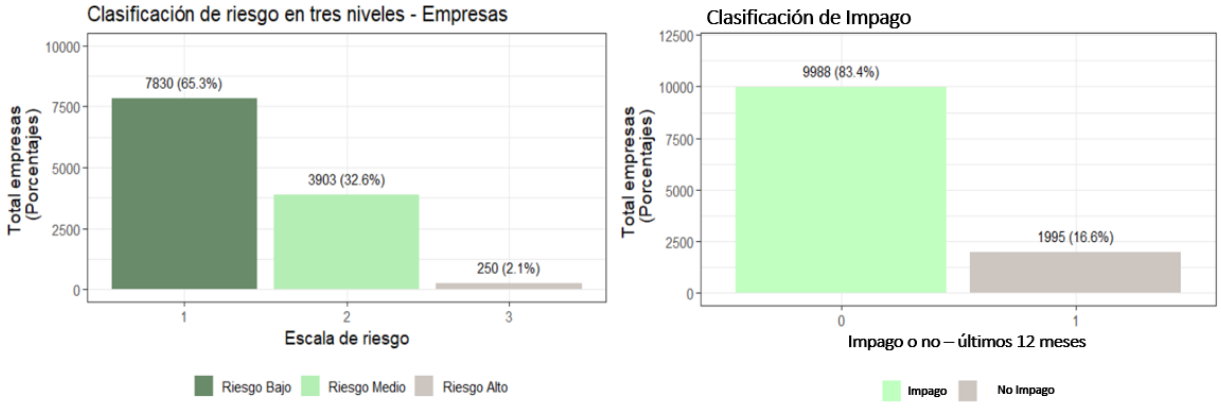


Figura 3-2: Distribución variables respuesta.

3.3.2. Análisis descriptivo variables predictoras

Análisis descriptivo variables categóricas

A continuación, se realiza el análisis univariado de las variables predictoras de carácter categórico:

La variable Segmento

La Figura (**3-3**), muestra teniendo en cuenta el total de empresas en cada categoría, que el segmento *Gobierno*, es el que mayor proporción presenta en incumplimiento con un 60,71 %, seguido del segmento *Grandes Clientes* con un 25,04 % y finalmente el segmento *Empresas* con un 15,91 %. Sin embargo, es posible que el desbalance en la cantidad de empresas por cada clase de segmento influyan en el anterior comportamiento, dado que por ejemplo el segmento Empresas contiene el 89,52 % del total de observaciones en Impago.

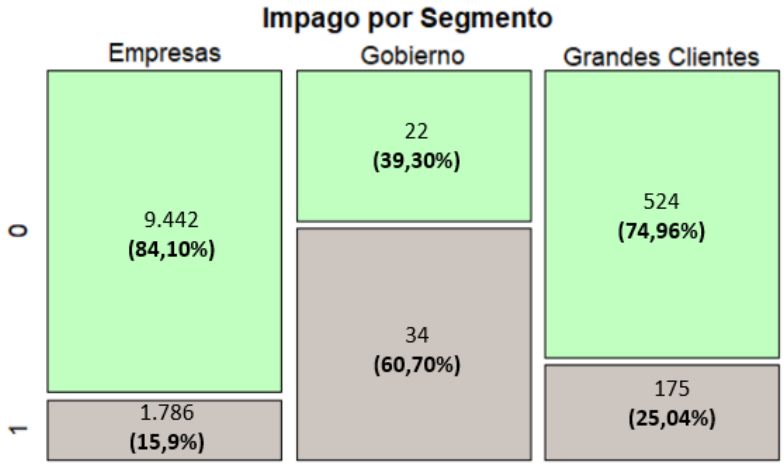


Figura 3-3: Variable Segmento.

La variable Sector

Teniendo en cuenta el total de empresas en cada categoría, en la Figura (**3-4**), el sector Manufactura es el que mayor proporción presenta en incumplimiento o Impago con un 20,90 %, seguido del sector servicios con un 16,72 % y finalmente el sector Comercio con un 10,57 %. Respecto al total de empresas que entraron en Impago (1.995), los sectores Servicios y Manufactura presentan un porcentaje de participación del 42,76 % y 42,06 %, respectivamente ver **Tabla (3-6)**.

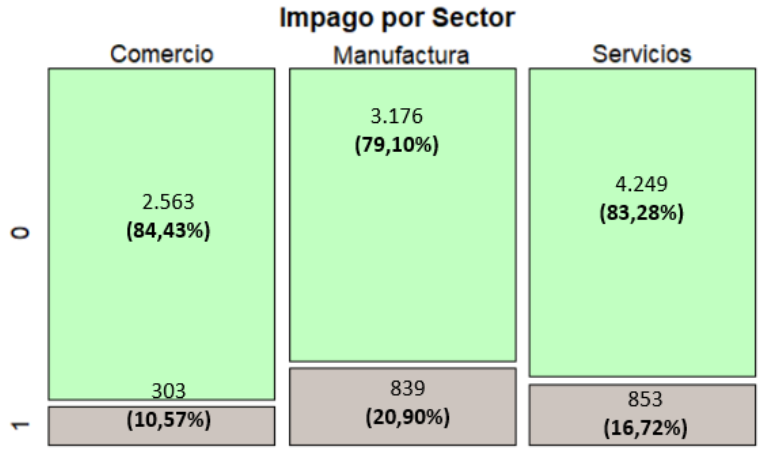


Figura 3-4: Variable Sector.

Variable Buró

Teniendo en cuenta que esta variable considera el comportamiento de pago de las empresas, en la Figura (**3-5**), es congruente observar que las empresas mejor calificadas son las que menor proporción de deterioro presentan con un 5,62 %, mientras que aquellas que presentan un nivel (3), son las que mayor proporción presentan con un 76,65 %.

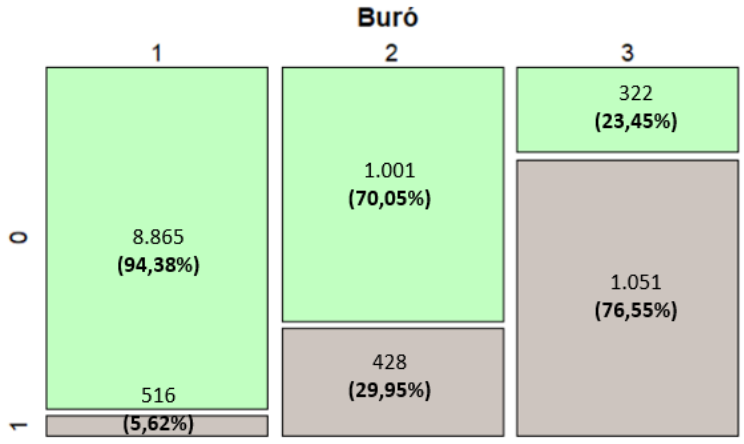


Figura 3-5: Variable Buró.

Variable árbol de clasificación

Los dos primeros niveles como se observa en la Figura (**3-6**), son los que menor proporción de Impago presentan, con el 14,57 % y 10,73 %, respectivamente. Lo anterior tiene sentido, considerando que estos son los que menor dedicación cobranza requieren,

de acuerdo con las políticas de la empresa. Sin embargo entre los últimos niveles, el correspondiente al quinto es el que mayor proporción tiene de los clientes que entran en Impago 62,32 %. La compañía como se observa en **Tabla (3-6**), tiene del total de clientes que entraron en Impago (1.995) el 62,31 % en los niveles (1) y (2) de priorización de cobranza.

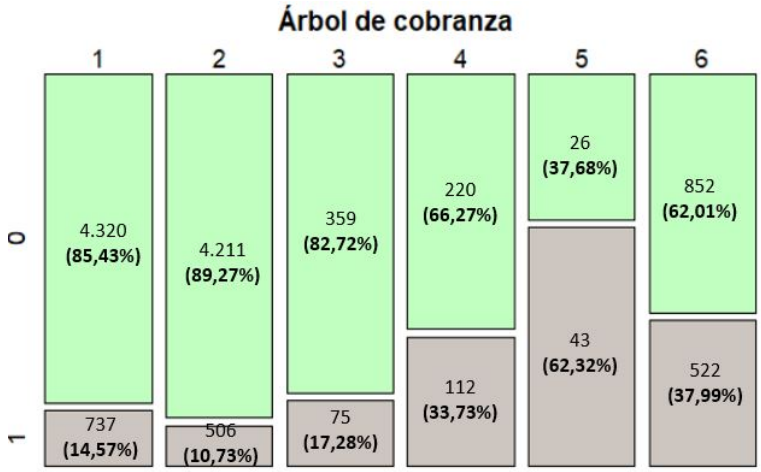


Figura 3-6: Variable Árbol.

Tamaño de la empresa

Llama la atención como se observa en la Figura (**3-7**), que las grandes y medianas empresas, son las que mayor proporción Impago (1) presentan, con un 26,07 % y 18,51 %. Por su parte las empresas pequeñas y microempresas, presentaron una proporción del 13,77 % y del 15,92 %.

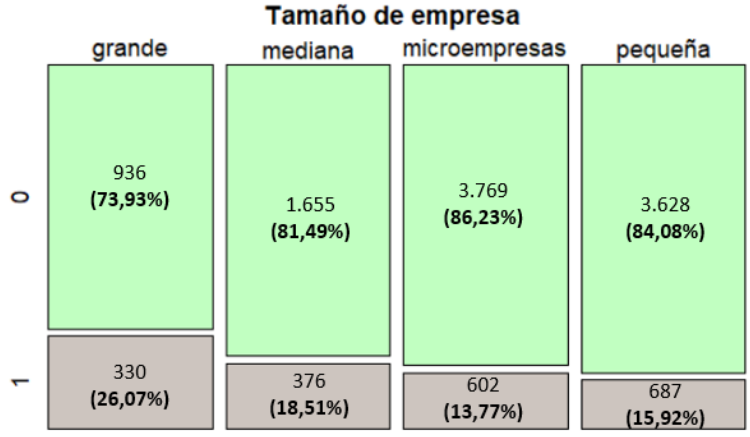


Figura 3-7: Variable Tamaño de la empresa.

Región de operación de la empresa

La empresas objeto de estudio, pertenecen a diversas regiones a nivel nacional. Con el fin de tener un mejor análisis de ésta variable, se tiene en cuenta de forma independiente las Regiones de Antioquia y de Centro país y el resto de regiones se agrupan en la categoría Otras. En la Figura (**3-8**), se observa que teniendo en cuenta el total de empresas que entraron en Impago respecto al total de la muestra, todas las regiones presentan porcentajes muy similares. Antioquia representa el 16,40 %, la región centro el 19,57 % y el resto de las regiones el 16,93 %.

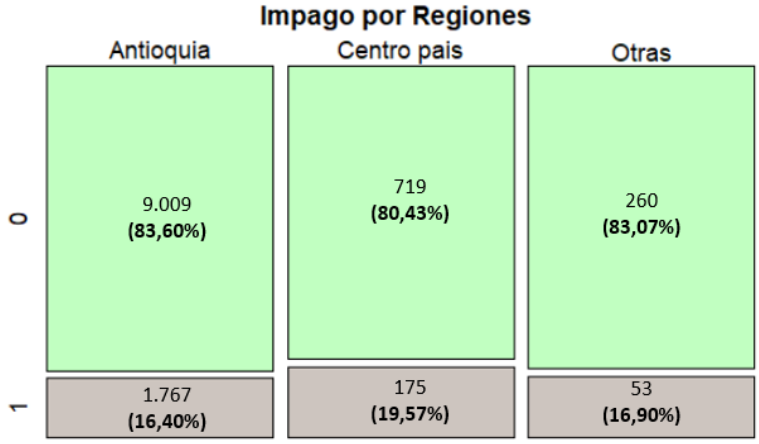


Figura 3-8: Variable Región de operación de la empresa.

Sin embargo, es posible que esta variable no influya en la significancia de los modelos, debido a que el porcentaje de empresas por región que presentaron Impago respecto al total de empresas que entraron en dicho incumplimiento (1.995), tiene una participación muy grande en la región Antioquia 88,57 %. Es decir teniendo en cuenta que las regiones en las que se concentra la prestación de servicios es Antioquia, esto puede dar lugar a excluir dicha variable del análisis, por no tener una buena distribución en todas las clases.

Estatus operacional de las empresas

En la Figura (**3-9**) las empresas que se encuentran en liquidación, son las que mayor proporción de Impago presentan con un 2 %, seguidas de las que se encuentran en reorganización y en operación con un $23,12\,\%$. Sin embargo del total de empresas analizadas, la mayoría se encuentran en estado operacional con un 95 %, por lo que el desbalance en las categorías puede dar lugar a que la variable no presente significancia en la selección de variables en la **Tabla (3-6)**.

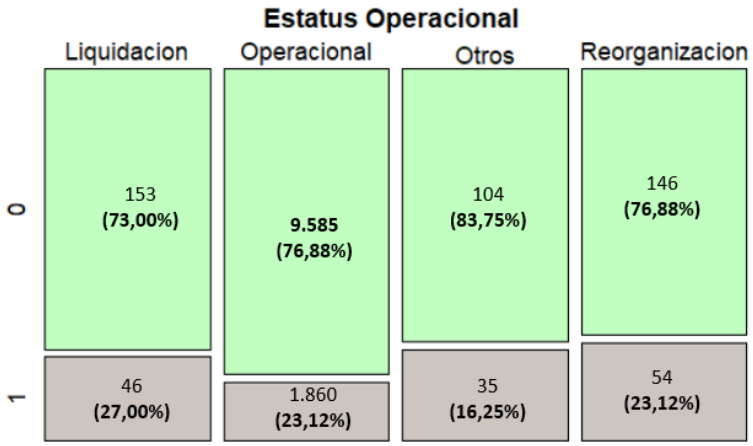


Figura 3-9: Variable estatus operacional.

Cuadro resumen variables categóricas.

La Tabla (**3-6**), muestra para cada variable, el número y porcentaje de empresas que presentaron Impago y las que no, así como la proporción de empresas con Impago en cada categoría, teniendo en cuenta el total de empresas evaluadas. Las últimas 3 columnas registran los estadístico Chi cuadrado con su valor P, y los criterios de información de Akaike (AIC) y Seudo R^2 respectivamente.

Todas las variables por medio del estadístico Chi cuadrado, arrojan un P-valor menor con un nivel de significancia $\alpha = 0{,}05$, indicando que existe una asociación entre cada una de las variables categóricas y la variable respuesta Impago. Cada variable se dispone en la tabla, desde las que menor criterio AIC y mayor Seudo R^2 presentan.

Por ejemplo, entre las variables que pueden presentar mayor ajuste en los modelos ya que presentan menor criterio AIC y mayor Seudo R^2, son las variables el Buró y Árbol que como vimos en la descripción de variables, corresponden a criterios de calificación de cartera para los clientes. En éstas variables a medida que sus niveles aumentan, incrementa la proporción de caer en Impago.

Tabla 3-6: Variables Categóricas.

Variables	No Impago		Impago		Total	Pro.Impago	Test χ^2	AIC	Seudo.R^2
	N. Empr	%	N. Empr	%	N	Pro.Impago	Test χ^2 (Vlr P)		
Buró									
Nivel 1	8.665	86,75 %	516	25,86 %	9.181	5,62 %	4536,8		
Nivel 2	1.001	10,02 %	428	21,45 %	1.429	29,95 %			
Nivel 3	322	3,22 %	1.051	52,68 %	1.373	76,55 %	(2,2e-16)		
	9.988		**1.995**		**11.983**			**7219.7**	**0.332**
Árbol Cobranza									
Nivel 1	4.320	43,25 %	737	36,94 %	5.057	14,57 %			
Nivel 2	4.211	42,16 %	506	25,36 %	4.717	10,73 %	759.57		
Nivel 3	359	3,59 %	75	3,76 %	434	17,28 %			
Nivel 4	220	2,20 %	112	5,61 %	332	33,73 %			
Nivel 5	26	0,26 %	43	2,16 %	69	62,32 %	(2,2e-16)		
Nivel 6	852	8,53 %	522	26,17 %	1.374	37,99 %			
	9.988		**1.995**		**11.983**			**10.167**	**0,059**
Sector									
Comercio	2.563	25,66 %	303	15.19 %	2.866	10,57 %	128,49		
Manufactura	3.176	31,80 %	839	42.06 %	4.015	20,90 %			
Servicios	4.249	42,54 %	853	42.76 %	5.102	16,72 %	(2,20e-16)		
	9.988		**1.995**		**11.983**			**10.633**	**0,0125**
Tamaño Empresas									
Grande	936	9,37 %	330	16.54 %	1.266	26,07 %			
Mediana	1.655	16,57 %	376	18.85 %	2.031	18,51 %	113.7		
Pequeña	3.769	37,74 %	602	30.18 %	4.371	13,77 %			
MicroEmp	3.628	36,32 %	687	34.44 %	4.315	15,92 %	(2,2e-16)		
	9.988		**1.995**		**11.983**			**10.694**	**0,009**
Segmento									
Empresas	9.442	94,53 %	1.786	89.52 %	11.228	15,91 %	118,25		
Gobierno	22	0,22 %	34	1,70 %	56	60,71 %			
Grandes Cl	524	5,25 %	175	8,77 %	699	25,04 %	(2,2e-16)		
	9.988		**1.995**		**11.983**			**10.706**	**0,008**
Regiones									
Antioquia	9.009	90,20 %	1.767	88,57 %	10.776	16,40 %	6,0246		
CentroPais	719	7,20 %	175	8,77 %	894	19,57 %			
Otras	260	2,60 %	53	2,66 %	313	16,93 %	0,04918		
	9.988		**1.995**		**11.983**			**10.791**	**0,0005**
Clasificac Est.Operac									
Operacional	9.585	95,97 %	1.860	93.23 %	11.445	16,25 %	35,138		
En Liquid	153	1,53 %	46	2,31 %	199	23.12 %			
En Reestructu	146	1,46 %	54	2.71 %	200	27,00 %	(2,34e-05)		
Otros	104	1,04 %	35	1.75 %	139	25,18 %			
	9.988		**1.995**		**11.983**			**10.767**	**0,003**

Análisis descriptivo variables numéricas

Se describe a continuación, las variables numéricas que podrían aportar a los modelos:

Mora de la empresa

La mora actual asigna una edad de 0 a 15 niveles que corresponden a las siguientes clases. Tabla (**3-7**):

Como se observa en la Figura (**3-10**), las empresas que no caen en Impago se concentran en las primeras edades, mientras que las que si presentaron el evento, presentaron una dispersión en todas los rangos de edades; se observa que algunas observaciones se concentran en las primeras edades y otras en las últimas edades. Los clientes que

entran en Impago, tienen una edad de mora promedio de 5 (franja de 121 a 150 días); por el contrario para los clientes que no presentaron Impago, la edad promedio es de cero, lo que es consistente cuando no se presentan incumplimientos. Las mediana para las empresas en Impago se ubican en la edad 1 (entre 1 a 30 días) y la desviación con respecto a la media es mayor para los clientes que presentaron Impago, ver la **Tabla (3-9)**.

Tabla 3-7: Edades de Mora.

Clase	Niveles
Corriente	0
1-30	1
31-60	2
61-90	3
91-120	4
121-150	5
151-180	6
181-210	7
211-240	8
241-270	9
271-300	10
301-330	11
331-360	12
361-720	13
271-1080	14
Mayor 1080	15

Esta variable tiene un valor P menor que el nivel de significancia de 0.05, por lo que resulta significativa, además entre todas las variables cuantitativas es la que mayor indicador Seudo R^2 arroja con un 0.2735204.

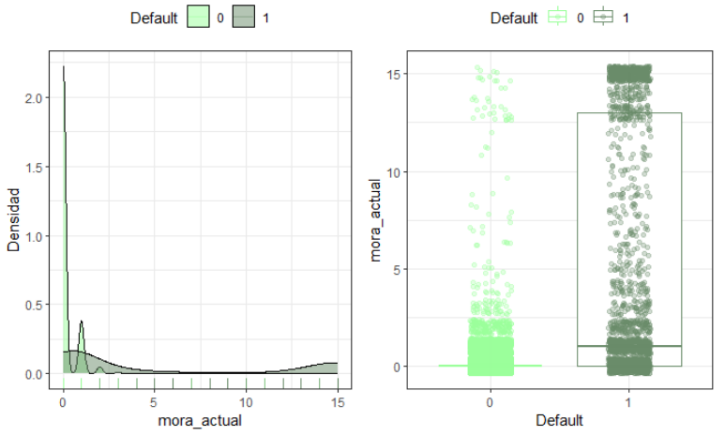


Figura 3-10: Mora - análisis descriptivo.

Probabilidad de incumplimiento

En la siguiente Figura (**3-11**), no se alcanza a apreciar de manera clara la distribución y la dispersión de las probabilidades. De acuerdo con el resumen expuesto en la **Tabla (3-9)**, las empresas que no evidenciaron Impago tuvieron una media por debajo del 0,3 % con una mediana del 0,066 %, mientras que para las empresas que presentaron Impago, la media se ubica al rededor del 4,23 %. Lo anterior significa que la mayoría de las empresas objeto de estudio, presentan probabilidades de incumplimiento bajas o no representativas, incluso para las empresas que presentan Impago.

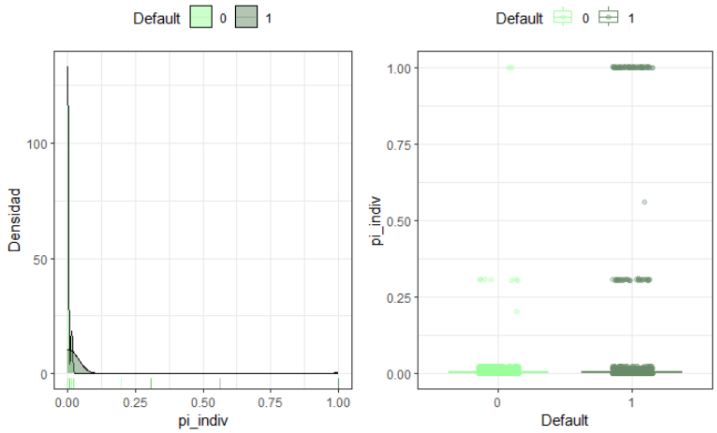


Figura 3-11: Probabilidad de incumplimiento - análisis descriptivo.

Índice Altman

En la base de datos se calculó para todas las empresas, la versión de índice Altman correspondiente al Z-core (Z2), dado su uso sugerido para empresas comerciales y de servicio. Dicho indicador incluye el cálculo de indicadores económicos y clasifica la insolvencia para las empresas, de acuerdo con las siguientes franjas de riesgo (Trujillo Ospina et al., 2016):

Se acotó el índice Altman a cero, cuando éste arrojó valores negativos y se acotó a cuatro, cuando el resultado se encontró por encima de dicho valor. Lo anterior para evitar la dispersión alta en el indicador y con el finde visualizar mejor los resultados, respetando las escalas propuestas por el autor. La siguiente es la clasificación de solvencia en riesgo, de acuerdo con las siguientes franjas en la Tabla (**3-8**).

Para la muestra, se observa en la Figura (**3-12**) una media superior para las empresas que están categorizadas en No Impago 2,65 %, que para aquellas empresas que se

Tabla 3-8: Franjas Altman.

	Riesgo Alto	**Riesgo Medio**	**Riesgo Bajo**
Valor del indicador	$< 1{,}1$	$\geqslant 1{,}1$ y $\leq 2{,}6$	$> 2{,}6$

encuentran en Impago con un 2,32 %. Lo mismo ocurre con el valor de la mediana, ver la **Tabla (3-6**). La dispersión sobre todas las franjas del indicador es visible para las empresas que presentaron Impago o no, visualizándose en el último cuántil una mayor cantidad de empresas que no tienen Impago, con resultados del indicador superiores a 2,6, es decir,empresas con mejor solvencia financiera.

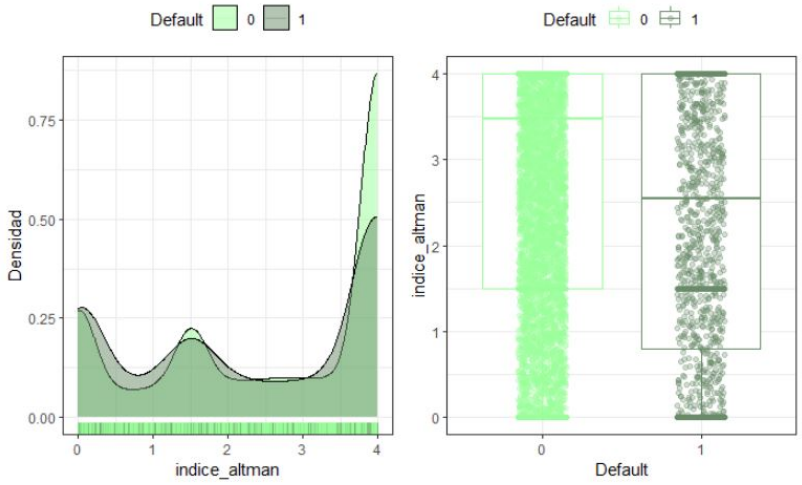


Figura 3-12: Índice Altman - análisis descriptivo.

Cartera Total

La muestra contiene desde empresas que tienen saldo cero, hasta empresas que tienen saldo de al rededor los 21 mil millones de pesos. En cuanto al valor de la cartera que está expresada en logaritmo natural, Figura (**3-13**), los clientes que no presentaron Impago, presentan una cartera promedio de 4.87, que en medidas de pesos corresponde a 5,53 millones. Por su parte las empresas que entraron en Impago, tenían en promedio 10,73 o 43,533 millones con una desviación estándar mayor para estos clientes. Para los dos resultados posibles (Impago y no Impago), la distribución es muy similar, incluso con mayor concentración de observaciones en el último cuántil.

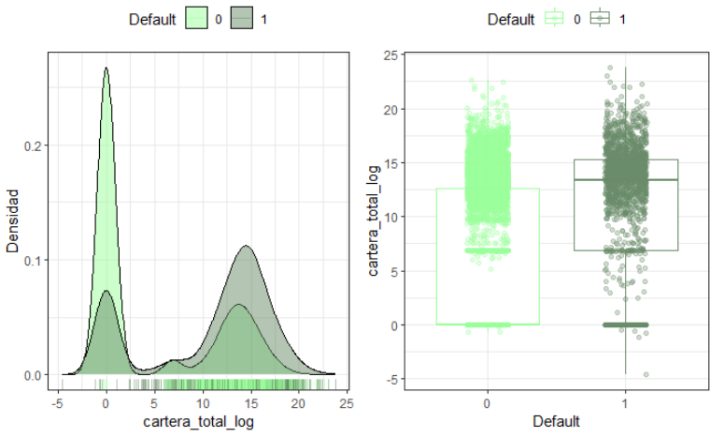


Figura 3-13: Cartera total - análisis descriptivo.

Activos totales

En la Figura (**3-14**), la variable Activos totales, expresada en logaritmo natural, presenta para los clientes en Impago un mayor valor en promedio de los activos de (22.96 en logaritmo natural o 703.536 mill), que los que no presentaron Impago (22,01 en logaritmo natural o 61.613 millones). También para los clientes que se encuentran en Impago, la dispersión de los archivos es más alta.

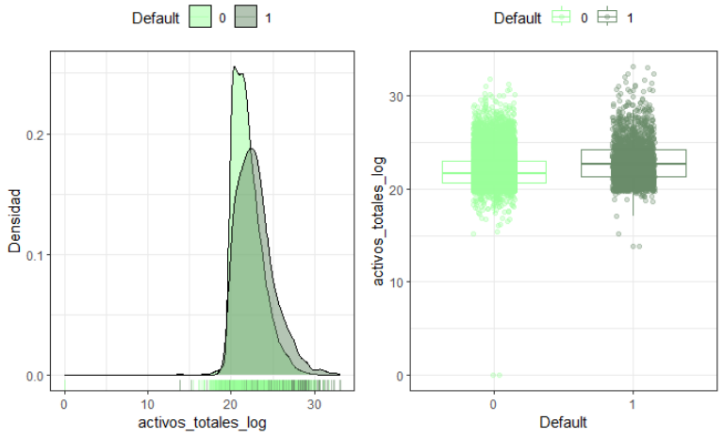


Figura 3-14: Activos totales - análisis descriptivo.

Pérdida dada el incumplimiento - PDI

Se asignó como nombre de la variable el prefijo “max”, dado que se tomó el valor máximo del total de servicios prestados a los clientes o empresas.

Es el segundo componente para el cálculo del deterioro, y es calculado en el modelo de provisión de cartera que tiene la compañía. En éste caso se considera el porcentaje que ya no es posible recuperar dado que el cliente ya entró en Impago, de allí que sus porcentajes sean tan altos. De acuerdo con las edades de mora de las empresas, se asignan 3 porcentajes de PDI: Para edades menores a 420 días un 63 %, para edades entre 421 y 540 días de mora un 80 % y para edades superiores a 541 días de mora un 100 %. Estos cortes son los que se ven de manera clara en la Figura (**3-15**), lado derecho del gráfico de caja y bigotes. Algunos se sitúan en cero porque son empresas que no presentan saldo de cartera.

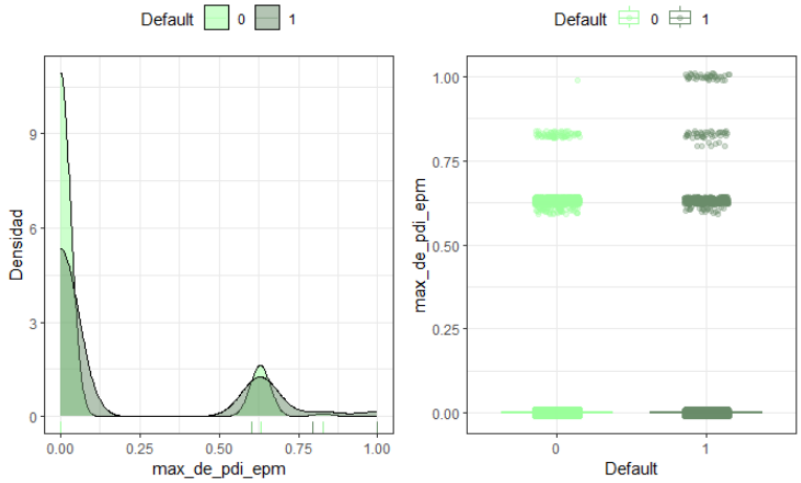


Figura 3-15: Máximo PDI - análisis descriptivo.

Valor provisión.

El valor de provisión es una variable que resulta multiplicar el valor del capital expuesto con el porcentaje de PI (Probabilidad de incumplimiento) y porcentaje del PDI (Pérdida dada el incumplimiento). En la Figura (**3-16**), la variable está en términos de logaritmo natural y presenta un valor promedio más alto para los clientes que con Impago (1.87 en logaritmo natural o 1,86 mill). La mediana es igual en ambos casos, ver Tabla (3-9).

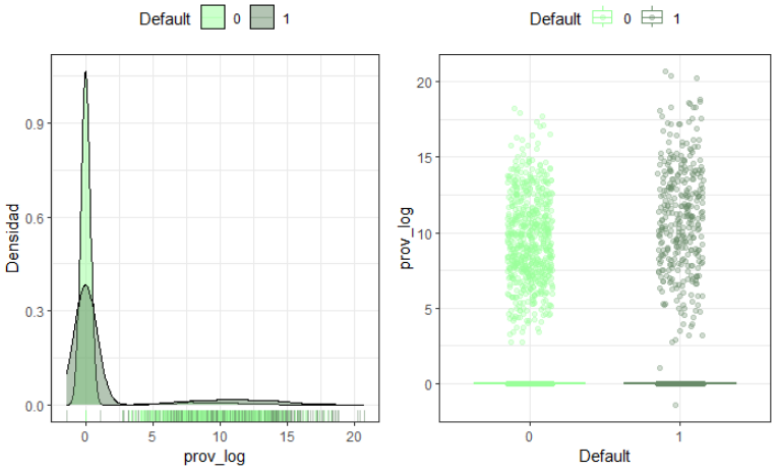


Figura 3-16: Valor provisión - análisis descriptivo.

Concentración endeudamiento

En la Figura (**3-17**), se observan diferentes porcentajes del indicador, llamando la atención un mayor volumen de porcentajes altos en empresas que no presentaron la marca de Impago resultado de las matrices de transición. Los porcentajes en promedio de empresas con comportamiento de no Impago e Impago fueron similares con un 67,2 % y 60,5 % respectivamente, pero con una mediana más alta para las empresas que no tuvieron incumplimiento 80,5 %. Es posible que éste tipo de indicador financiero, se asocie con incumplimientos a otros compromisos de deuda que tengan los clientes, pero con buen comportamiento de pago con la compañía.

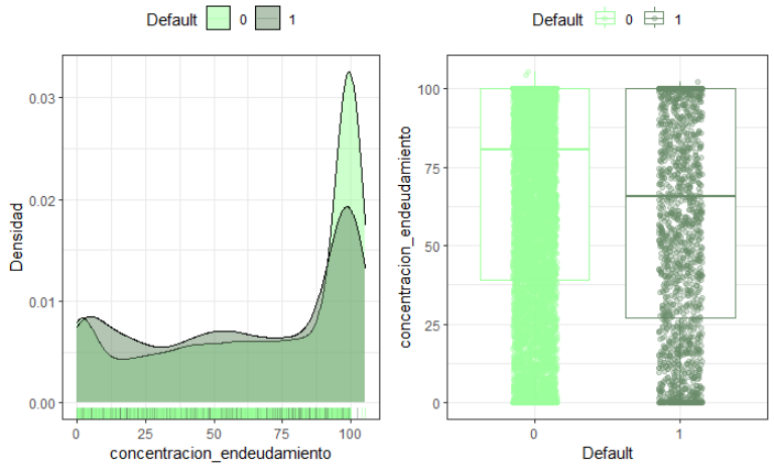


Figura 3-17: Concentración endeudamiento - análisis descriptivo.

3.3.3. Resumen Univariado con la variable Impago

En la Tabla (**3-9**), se expresa para cada una de las variables cuantitativas, el resultado del análisis univariado con la variable respuesta Impago. Se muestra el valor promedio, mediana y desviación estándar para las empresas que entran en Impago o no. Asimismo, en orden descendente las variables que mayor significancia tienen considerando el valor P de 0.05 y aquellas que representan un mayor Seudo-R^2 y menor criterio de información de Akaike (AIC) respectivamente.

Tabla 3-9: Resumen de variables numéricas.

Var	No Impago			Impago			Valor P			
	$\bar{X}$	XMediana	σ	$\bar{X}$	XMediana	σ	AIC	$Pr(>\|z\|)$		S-R^2
Mora	0	0	1	5	1	6	7843,6	$< 2e-16$	***	0,274
C.T(l)	4,869	0	1,029	10,793	13,359	6,188	9608,5	$< 2e-16$	***	0,11
Act.T(l)	22,007	21,672	1,8	22,965	22,669	2,272	10408	$< 2e-16$	***	0,036
PI	0,003	0,001	0,017	0,042	0,001	0,182	10466	$< 2e-16$	***	0,031
Prov.(l)	0,618	0	2,457	1,873	0	4,384	10547	$< 2e-16$	***	0,023
KE.(l)	0,923	0	3,577	2,566	0	5,805	10574	$< 2e-16$	***	0,02
Ct.D(l)	0,481	0	2,441	1,594	0	4,405	10601	$< 2e-16$	***	0,018
M.PDI	0,088	0	0,221	0,156	0	0,293	10668	$< 2e-16$	***	0,012
Altman	2,657	3,476	1,53	2,328	2,542	1,606	10722	$< 2e-16$	***	0,007
P.Ct(l)	19,338	20,239	5,216	20,295	21,095	5,475	10733	$3,09e-13$	***	0,006
Cc.E	67,209	80,496	35,594	60,532	65,737	36,287	10739	$3,43e-14$	***	0,005
C.N.W	33508	1248	560929	426099	2800	6118308	10753	0,000149	***	0,004
Ende	0,508	0,492	0,723	0,615	0,553	1,487	10774	0,000402	***	0,002
Ebit(l)	13,763	18,359	9,132	13,482	18,534	9,78	10776	0,000614	***	0,002
C.W	11674	544	289173	66745	781	1003865	10780	0,00461	**	0,001
Mg.Op	-0,036	0,049	3,961	-1,057	0,047	30,372	10784	0,00562	**	0,001
G.Nt(l)	14,385	18,159	8,439	13,756	18,222	9,284	10786	0,00282	**	0,001
Roa	0,043	0,034	0,362	0,015	0,016	0,252	10787	0,0165	*	0,001
Act.C(l)	13,4	10,421	6,787	13,828	11,32	6,758	10789	0,0101	*	0,001
CC	1,739	0	4,962	2	0	5,281	10791	0,034	*	0
IO(l)	20,43	21,439	5,567	20,706	21,697	6,078	10791	0,0472	*	0
R.cte	9,355	1,74	68,985	13,422	1,575	143,867	10792	0,0678	.	0
Mg.Solv	4,302	0,852	239,277	13,997	0,938	278,949	10792	0,0895	.	0
Ef.Oper	1,447	0,976	5,396	1,862	0,492	38,68	10794	0,362		7E-05
Mg.Nto	-0,161	0,034	15,162	0,16	0,031	52,499	10795	0,607		3E-05
Roe	0,143	0,082	1,806	0,122	0,05	2,344	10795	0,64		2E-05
Patr(l)	20,396	20,831	4,423	20,446	21,478	6,026	10795	0,663		2E-05
S.Patr	10,304	0,888	150,273	9,407	0,671	145,395	10795	0,807		6E-06
Ap.CP	3,939	0,493	121,065	4,285	0,476	119,047	10795	0,907		1E-06

Nemotecnia: Mora-(Edad de mora), C.T(l)-(Cartera Total Log), Act.T(l)-(Activos Totales log), PI-(Prob de incump) Prov.(l)-(Vlor Provisión log), KE.(l)-(Capital expuesto Log), Ct.D(l)-(Cartera Diferida Log), M.PDI-(Máximo PDI) Altman-(Altman), P.Ct(l)-(Pasivos corrientes log), Cc.E-(Concentración endeud), C.N.W-(Capital Neto Operativo) Ende-(Endeudamiento), Ebit(l)-(Utilidad Operativa Ebit), C.W-(Capital de Trabajo), Mg.Op-(Margen Operativo) G.Nt(l)-(Ganancia Neta), Act.C(l)-(Activos Corrientes log), CC-(Cuenta de Cobro), IO(l)-(Ingreso Operativo) R.cte-(Raxón Corriente), Mg.Solv-(Margen solvencia), Ef.Oper-(Eficiencia Oper), Mg.Nto-(Margen Neto) Patr(l)-(Patrimonio), S.Patr-(Solvencia Patrimonial), Ap.CP-(Apalancamiento de corto plazo)

Complementando los análisis para las variables numéricas, elementos como el capital expuesto y cartera diferida que están relacionadas o incluidas en el valor de la cartera total, presentan un comportamiento similar a esta variable, ya que para empresas que no presentan Impago, se evidencia menores saldos y menor desviación estándar que las empresas que si lo tienen. La variable endeudamiento, presenta menores promedios y medianas en las empresas que no presentaron Impago, que para aquellas que si lo generaron, lo cual es consistente para empresas que administran mejor la estructura de deuda. Por su parte, variables como el Ebit y ganancia neta que está relacionada con la utilidad operacional y ganancia final de las empresas al igual que sus respectivos

márgenes operativos y netos, no muestran una clara separación para el promedio y mediana entre las empresas que presentaron Impago y las que no, por lo que no se puede asociar un comportamiento usual al incumplimiento y posiblemente no ser significativas en el proceso de selección de variables. Como se observa en los resúmenes descriptivos, las variables que parten de la rentabilidad del activo ROA y el resto de variables, ya no presentan significancia en el análisis univariado con la variable respuesta Impago, y esto se detecta en mayores niveles del criterio AIC, valores P no significativos y resultados del seudo R^2 menos aportantes.

4 Modelos de clasificación - Riesgo de crédito

Como parte de la metodología para medir el riesgo de crédito, en las secciones anteriores se describió como se obtuvo las variables respuesta, sus características y el análisis descriptivo tanto para las variables predictoras numéricas como categóricas. A continuación se muestra dentro del proceso de análisis, como seleccionar las variables que pueden aportar más en los diferentes modelos, opciones de equilibrio de muestras para clases desbalanceadas, la aplicación de algunos modelos de clasificación y finalmente las medidas de desempeño para comparar los diferentes modelos.

4.1. Preselección de variables

4.1.1. Algoritmos de selección - Mejor subconjunto, selección hacia adelante y selección hacia atrás

Teniendo en cuenta el análisis descriptivo tanto para variables categóricas como numéricas (29 variables), iniciamos el análisis, con las siguientes variables regresoras para aplicar los algoritmos de selección expuestos en la Sección **2.4.1**, que consisten en la selección del mejor subconjunto, selección hacia adelante y selección hacia atrás:

Variables: Edad de mora, Cartera total (log), Activos totales (log), Probabilidad de incump, Valor de provisión (log), Capital Expuesto (KE), Cartera diferida (log), Máximo PDI, Indicador Altman, Pasivos corrientes (log), Concentración endeudamiento, Capital de trabajo Neto, Endeudamiento, Utilidades Operativas Ebit (log), Capital de trabajo, Margen operativo, Ganancia Neta (log), Roa, Activos Corriente (log), Cuenta de Cobro, Buró, Árbol de cobranza, Sector, Tamaño Empresa, Segmento, Regiones, Clasificación Estado Operacional.

En los tres casos, se definió tomar 20 variables, como la cantidad a seleccionar tanto en variables numéricas como categóricas. Por medio del paquete leaps de R (Lumley, 2020) , se obtuvieron las siguientes visualizaciones gráficas que detallan en cada algoritmo las variables sugeridas, teniendo en cuenta las medidas de ajustes mencionadas en el

marco teórico. El objetivo es identificar tanto para la variable respuesta Impago, como para la variable de tres niveles (Clasificación de Riesgo), aquellas variables que se encuentran en tope de la gráfica, el cual se señaló con una línea naranja. Adicionalmente se identificaron con líneas discontinuas, aquellas variables que sugería cada algoritmo.

A continuación, en la Figura (**4-1**) se ilustra para el caso de la variable Impago, las variables que fueron identificadas en el algoritmo de mejor subconjunto:

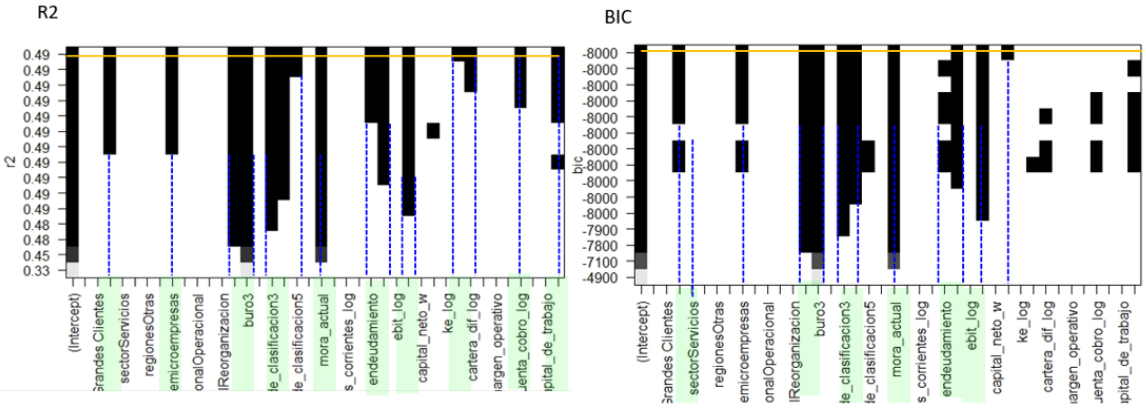


Figura 4-1: Selección mejor subconjunto para la variable Impago.

En la siguiente Figura (**4-2**), para el caso de la variable Clasificación de Riesgo de Crédito, la técnica de selección hacia adelante sugiere las siguientes variables:

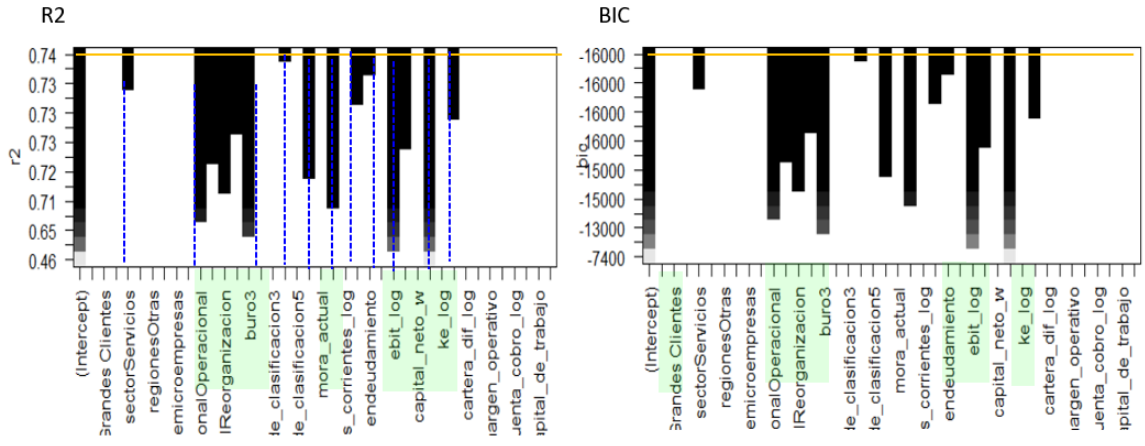


Figura 4-2: Selección hacia adelante para la variable Clasificación de Riesgo

4.1.2. Análisis de correlación de variables y de varianza ANOVA

Con las variables numéricas sugeridas en el apartado anterior, se realiza el análisis de correlaciones para evitar considerar variables que presenten Multicolinealidad. De acuerdo con la Tabla (**4-1**), las variables Capital de trabajo y Capital de trabajo neto operativo, presentan una fuerte correlación del (0.8), seguida de las variables Ganancia neta y Ebit con una correlación del (0.66), finalmente la última asociación que destaca, se presenta entre las variables Cartera diferida y Valor de la factura o cuenta de cobro con una correlación del (0.45). El resto de las variables presentan una asociación inferior al 40 %.

Tabla 4-1: Primera matriz de correlaciones.

Var	Mora	C.T	Act.T	PI	Cart.D	Alt	Cc.E	C.N.W	Ende	Ebit	C.W	G.Nt	Act.C	Fact
mora	1	0,27	0,18	0,25	0,16	-0,04	-0,06	0,06	0,01	0,02	0,06	0	-0,03	0,04
C.T	0,27	1	0,16	0,14	0,19	-0,09	-0,08	0,05	0,03	-0,03	0,04	-0,01	-0,02	0,08
Act.T	0,18	0,16	1	0,16	0,2	0,03	-0,1	0,17	0	0,31	0,17	0,3	0	0,3
PI	0,25	0,14	0,16	1	0,28	-0,03	-0,02	0,15	0,09	0,03	0,09	0,02	-0,02	0,18
Cart.D	0,16	0,19	0,2	0,28	1	-0,04	-0,04	0,11	0,04	0,05	0,08	0,04	-0,02	0,45
Alt	-0,04	-0,09	0,03	-0,03	-0,04	1	0,02	-0,02	-0,2	0,34	0,01	0,36	-0,05	0,01
Cc.E	-0,06	-0,08	-0,1	-0,02	-0,04	0,02	1	0,02	0	0,1	0	0,14	0,04	-0,01
C.N.W	0,06	0,05	0,17	0,15	0,11	-0,02	0,02	1	0,01	0,05	0,8	0,05	0,02	0,05
Ende	0,01	0,03	0	0,09	0,04	-0,2	0	0,01	1	-0,04	0	-0,06	0,07	0,01
Ebit	0,02	-0,03	0,31	0,03	0,05	0,34	0,1	0,05	-0,04	1	0,06	0,66	0	0,11
C.W	0,06	0,04	0,17	0,09	0,08	0,01	0	0,8	0	0,06	1	0,06	0,02	0,04
G.Nt	0	-0,01	0,3	0,02	0,04	0,36	0,14	0,05	-0,06	0,66	0,06	1	-0,01	0,11
Act.C	-0,03	-0,02	0	-0,02	-0,02	-0,05	0,04	0,02	0,07	0	0,02	-0,01	1	-0,05
Fact	0,04	0,08	0,3	0,18	0,45	0,01	-0,01	0,05	0,01	0,11	0,04	0,11	-0,05	1

Nemotecnia: Mora (Mora Actual), C.T (Cartera total), Act.T (Activos totales), PI (Probabilidad de incumplim), Cart.D (Cartera Diferida) Alt (Índice Altman), Cc.E (Concentración Endeudamiento) , C.N.W (Capital neto de Trabajo), Ende (Endeudamiento), Ebit C.W (capital de trabajo), G.Nt (Ganancia Neta) ,Act.C (Activos corrientes) , Fact (Cuenta de Cobro o facturación)

Como complemento a los algoritmos anteriores y al estudio de altas correlaciones en algunas variables, se realiza el análisis de varianza ANOVA. Allí se fueron generando diferentes modelos iniciando con un modelo reducido y en cada uno se iba agregando variables de modo que en el modelo completo, todas las variables tuvieran significancia con un valor P menor que un nivel de significancia de 0.05. La función ANOVA permite comprobar el aporte de todos los modelos agregados.

Con las variables observadas en los análisis anteriores, se comprueba que no se presenta una fuerte asociación lineal entre las variables.

Tabla 4-2: Matriz de correlaciones variables numéricas.

Variables	Mora.Act	Cart.Tot	Activ tot	PI	Ind.Altman	Vlr fact
Mora.Act	1	0,27	0,18	0,25	-0,04	0,04
Cart.Tot	0,27	1	0,16	0,14	-0,09	0,08
Activ tot	0,18	0,16	1	0,16	0,03	0,3
PI	0,25	0,14	0,16	1	-0,03	0,18
Ind.Altman	-0,04	-0,09	0,03	-0,03	1	0,01
Vlr fact	0,04	0,08	0,3	0,18	0,01	1

La siguiente Tabla (**4-3**), presenta el resumen de los métodos de selección y de extracción de variables. En la primera columna se identifican si las variables son continuas o

categóricas, en la segunda el nombre de las variables, en la tercera y cuarta columna, se marcan las variables que arrojaron los algoritmos de mejor subconjunto, selección hacia adelante y selección hacia atrás, tanto para las variables Impago como para la de tres niveles Clasificación, en la quinta columna el resumen de la selección común para las variables respuestas, en la sexta las variables descartadas por alta correlación y en la séptima, las variables que finalmente fueron seleccionadas para la aplicación de los diferentes modelos que veremos en la siguiente sección y que fueron validadas por medio del método ANOVA.

Tabla 4-3: Selección de variables.

	Variables	V.Impago	V.Clasif	Mét selecc	Correlación alta	Modelos-Anova
Numéricas	Edad de mora)	✓		✓		✓
	Cartera total(l)	✓		✓		✓
	Activos tot(l)	✓		✓		✓
	Prob de incump		✓	✓		✓
	Valor de prov(l)					
	Cap Exp-(l)					
	Cartera Dif(l)	✓	✓	✓		
	Máximo PDI					
	Ind Altman		✓	✓		✓
	Pasivos ctes (l)					
	Conc endeudam		✓	✓		
	Cptal W Neto				⊗	
	Endeudamiento	✓	✓	✓		
	Ebit(l)	✓	✓	✓	⊗	
	Ctal de W				⊗	
	Mg operativo					
	Ganancia Nta(l)	✓		✓	⊗	
	Roa					
	Activos Ctes(l)	✓		✓		
	Cta de Cobro	✓	✓	✓	⊗	✓
Categóricas	Buró	✓	✓	✓		✓
	Árbol cobranza	✓	✓	✓		✓
	Sector	✓	✓	✓		✓
	Tamaño Empresa	✓	✓	✓		✓
	Segmento					
	Regiones					
	Estado Operac		✓	✓		

La variables que se consideran para la aplicación de los diferentes modelos son las siguientes: Edad de mora, Cartera total, Activos totales, Probabilidad de incumplimiento, Indicador Altman, Cuenta de Cobro, Buró, Árbol de cobranza, Sector y Tamaño Empresa.

4.2. Partición de la base de datos

Teniendo en cuenta que se evaluarán dos variables respuesta: Clasificación dicotómica (No Impago, Impago) y clasificación de tres niveles (Riesgo Bajo, Riesgo Medio y Riesgo Alto), se realiza dos particiones independientes de la base de datos, considerando cada una de las variable respuesta. Así mismo, se mostrará las diferentes medidas de desempeño, como lo ejemplifica Tharwat (2021), dependiendo de las metodologías de clasificación con una matriz 2×2 para la clasificación dicotómica y matriz 3×3, para la clasificación en tres niveles.

La base de datos con los procesos de selección y validación, inició con 11 variables (incluyendo la variable respuesta) y 11,983 registros que corresponden a las empresas a evaluar. Se particiona la muestra con un 80 %, para el entrenamiento de los modelos y un 20 % para la validación cruzada o comprobación en la efectividad de los mismos.

Se utilizó la función "stratified", del paquete splitstackshape (Mahto, 2019), que garantiza de cada clase o estrato, una selección homogénea de acuerdo con el porcentaje seleccionado. Como se muestra en la siguiente Figura (**4-3**), se toma igual proporción de elementos, de acuerdo a cada clase, en éste caso caras.

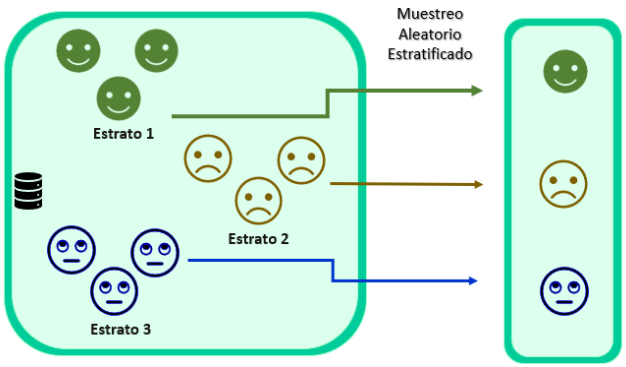


Figura 4-3: Función STRATIFIED. Basada en Thomas (2020).

La distribución del proceso de partición para las variables respuesta y predictoras como se ilustra a continuación en las siguiente Tabla (**4-4**) y Tabla (**4-5**), presentan la

misma proporción para cada una de sus categorías de acuerdo con los porcentajes de entrenamiento y validación:

Tabla 4-4: Proporción de la partición Variables Respuesta.

Variable 3 Niveles	Total	Train	Test
Clas.Riesgo	**11.983**	**9586**	**2397**
Partición	**100 %**	**80 %**	**20 %**
1	7830	6264	1566
1	3903	3122	781
3	250	200	50

Variable 2 Niveles	Total	Train	Test
Impago	**11.983**	**9586**	**2397**
Partición	**100 %**	**80 %**	**20 %**
0	9988	7990	1998
1	1995	1596	399

Tabla 4-5: Proporción de la partición Variables predictoras.

Variable	Total	Train	Test	Variable	Total	Train	Test
Mora-Actual	**11.983**	**9586**	**2397**	**PI-indv**	**11.983**	**9586**	**2397**
Partición	**100 %**	**80 %**	**20 %**	**Partición**	**100 %**	**80 %**	**20 %**
Corriente	8876	7114	1762	**0 %**	10162	8117	2045
1-30	1821	1440	381	**1 %**	77	62	15
31-60	320	267	53	**1 %**	1532	1241	291
61-90	105	80	25	**2 %**	97	74	23
91-120	69	50	19	**20 %**	1	1	0
121-150	43	29	14	**31 %**	45	33	12
151-180	29	23	6	**56 %**	1	0	1
181-210	26	23	3	**100 %**	68	58	10
211-240	25	17	8	**SECTOR**	11983	9586	2397
241-270	25	22	3	**Partición**	100 %	80 %	20 %
271-300	16	14	2	**Comercio**	2866	2301	565
301-330	21	14	7	**Manufactura**	4015	3197	818
331-360	16	14	2	**Servicios**	5102	4088	1014
361-720	115	94	21	**BURÓ**	11983	9586	2397
721-1080	75	64	11	**Partición**	100 %	80 %	20 %
MAYOR 1081	401	321	80	**Nivel 1**	9181	7328	1853
Árbol de clasif	**11.983**	**9586**	**2397**	**Nivel 2**	1429	1151	278
Partición	100 %	80 %	20 %	**Nivel 3**	1373	1107	266
Nivel 1	5057	4027	1030	**Tamaño Empresa**	11983	9586	2397
Nivel 2	25	22	3	**Partición**	100 %	80 %	20 %
Nivel 3	434	354	80	**Grande**	1266	1010	256
Nivel 4	332	261	71	**Mediana**	2031	1619	412
Nivel 5	69	55	14	**Pequeña**	4371	3502	869
Nivel 6	1374	1115	259	**Microempresa**	4315	3455	860

4.3. Modelo logístico binomial sin balanceo de datos

El objetivo será inicialmente analizar como son los resultados de implementación del modelo logístico, sin la metodología de balanceo para la base de datos y considerando

las variables predictoras que fueron identificadas con los métodos de selección. Para ajustar el modelo logístico múltiple con la respuesta de tipo dicótoma Impago, hacemos uso de la función "glm" correspondiente a modelos lineales generalizados, y delimitamos el uso para la aplicación del modelo binomial, indicando en el parámetro "family", el modelo requerido.

El modelo GLM ajusta la variable respuesta en términos de una función de enlace, los coeficientes del modelo logit se interpretan como el logaritmo de la razón de posibilidades:

Tabla 4-6: Modelo logístico Impago con datos sin balanceo.

Coeficientes	Estimación	Error Est	Z Valor	P Valor	$Pr(>\|z\|)$	Exp(Estim)
(Intercept)	-9,616	0,828	-11,609	< 2e-16	***	6,7e-05
Act Tot log	0,251	0,031	8,009	1,16e-15	***	1,285
Árbol Clas2	0,234	0,102	2,305	0,021	*	1,264
Árbol Clas3	-0,835	0,199	-4,19	0	***	0,434
Árbol Clas4	-0,554	0,187	-2,961	0,003	**	0,575
Árbol Clas5	0,556	0,369	1,505	0,132		1,744
Árbol Clas6	0,039	0,12	0,323	0,747		1,04
Buro2	2,016	0,109	18,42	< 2e-16	***	7,508
Buro3	3,209	0,115	27,915	< 2e-16	***	24,754
Cart total log	0,03	0,007	4,541	0,000006	***	1,03
Índice altman	-0,053	0,026	-2,082	0,037305	*	0,948
Mora actual	0,369	0,018	20,275	< 2e-16	***	1,446
PI	3,206	0,983	3,262	0,001	**	24,68
SectorManufactura	0,433	0,115	3,753	0	***	1,542
SectorServicios	0,219	0,115	1,906	0,057	.	1,245
Tam.Empr mediana	0,307	0,175	1,752	0,08	.	1,359
Tam.Empr micro	0,775	0,202	3,845	0	***	2,171
Tam.Empr pequeña	0,494	0,191	2,586	0,01	**	1,639
Fact log	-0,032	0,009	-3,559	0	***	0,969

Para la muestra de entrenamiento en la Tabla (**4-6**), todas las variables resultan ser significativas en el modelo, puesto que presentan un $P - valor \leq 0,05$. De acuerdo con el resultado del estadístico Z (obtenido como la diferencia entre valor estimado y cero dividido por la desviación estándar tal como lo establecen Hogg, McKean, y Craig (2013) en el Teorema 6.4.1), las variables que más aportan en el modelo son las que están relacionadas con el comportamiento de la cartera, tales como el nivel de calificación Buró que evalúa el comportamiento de pago a un año y la Mora actual que identifica la edad de mora en las que se encuentran las empresas. En términos generales, las variables que presentan signo positivo, tienen un efecto negativo, puesto que dan lugar a que el evento de Impago aumente y viceversa.

Por ejemplo, por cada unidad que se incrementa la variable Mora (o Rango de mora), se espera que el logaritmo de la razón de posibilidad de la variable Impago se incremente en promedio $0,36918$ unidades. En otras palabras, una empresa que pre-

sente más nivel de mora, mayor probabilidad de Impago tendrá. Al exponenciar los coeficientes, por cada unidad que se incrementa la variable edad de mora, la razón de posibilidad de que una empresa entre en Impago se incrementan en promedio 1,45 unidades: $e^{0,36918} = 1,45$.

Para evaluar la idoneidad del modelo en la Tabla (**4-7**), se analizan las medidas de desempeño. Allí cuando comparamos el número de empresas observadas con la información resultante de la predicción, se obtuvo una capacidad predictiva general del modelo o *Exactitud* del 78 %. De otro lado, la especificidad (proporción de empresas que efectivamente no se clasifican en Impago, es del 97,2 %). Sin embargo, la sensibilidad o la capacidad para detectar los empresas que entran en Impago, es del 58,8 %. En este caso, el valor del umbral o el punto de corte que maximiza la sensibilidad (o evento de Impago) y minimiza los falsos positivos (especificidad), es del 57,29 %.

Base de datos	Sensibilidad	Especificidad	Exactitud	Exactitud_bal
Sin balanceo	0,588	0,972	0,908	0,78

Tabla 4-7: Modelo logístico datos sin balancear.

Todo lo anterior indica, que el modelo sin aplicar métodos de balanceo es moderadamente sensible para detectar las empresas que efectivamente entrarán en Impago.

4.4. Modelo logístico binomial con métodos de balanceo

En la base de datos, se tiene un porcentaje de Impago correspondiente al 16,65 % y del 83,65 % de empresas sin Impago. Si bien no se cuenta con problemas graves de desbalance en el conjunto de datos, es posible que se pueda mejorar la predicción en la clasificación de los datos. A continuación se aplicarán las metodologías de balance en la muestra (Submuestreo y Sobremuestreo) y se compararán los resultados.

4.4.1. Modelo con metodología Submuestreo

Cuando se aplica el modelo logístico binomial con ésta metodología, disminuye la muestra de la clase minoritaria (Impago).

En la Tabla (**4-8**), se presenta al igual que el modelo sin balanceo, los mismos signos en los valores estimados, valor del estadístico Z y valor P de cada variable, pero los

coeficientes son en la mayoría de las variables mayores. Cuando se comparan los valores de significancia, las variables como la probabilidad de incumplimiento, sectores a los que pertenecen las empresas y algunos niveles del árbol de clasificación, dejan de ser significativos con $P - valor \geq 0,05$. En este caso se corre el riesgo, de que al no ser una base datos suficientemente grande, se hayan eliminado muestras importantes en el modelo.

En cuanto a las medidas de desempeño, la Tabla (**4-9**), muestra con relación al modelo sin balanceo, que mejoran indicadores como la exactitud y la sensibilidad en un 7,81 % y 30,93 % respectivamente, sin embargo la especificidad, disminuye en un 6,20 %. También el valor del umbral disminuye y se reubica en este caso en un 34,93 %.

Tabla 4-8: Modelo logístico Impago con datos método Submuestreo.

Coeficientes	**Estimación**	**Error Est**	**Z Valor**	**P Valor**	$Pr(>\|z\|)$	**Exp(Estim)**
(Intercept)	-8,821	1,294	-6,815	9,41e-12	***	1,48e-04
Act Tot log	0,277	0,049	5,62	1,91e-08	***	1,319
Árbol Clas2	0,205	0,149	1,381	0,167		1,228
Árbol Clas3	-1,136	0,306	-3,712	0	***	0,321
Árbol Clas4	-0,356	0,318	-1,118	0,264		0,7
Árbol Clas5	2,182	1,241	1,758	0,079	.	8,864
Árbol Clas6	0,127	0,194	0,658	0,511		1,135
Buro2	2,089	0,17	12,298	< 2e-16	***	8,077
Buro3	3,343	0,217	15,384	< 2e-16	***	28,304
Cart total log	0,033	0,01	3,342	0,001	***	1,034
Índice altman	-0,088	0,039	-2,257	0,024	*	0,916
Mora actual	0,593	0,065	9,127	< 2e-16	***	1,809
PI	5,455	3,904	1,397	0,162		233,925
SectorManufactura	0,303	0,169	1,791	0,073	.	1,354
SectorServicios	0,08	0,168	0,479	0,632		1,083
Tam.Empr mediana	0,607	0,268	2,268	0,023	*	1,835
Tam.Empr micro	1,007	0,317	3,178	0,001	**	2,737
Tam.Empr pequeña	0,744	0,296	2,512	0,012	*	2,104
Fact log	-0,033	0,014	-2,379	0,017	*	0,968

Base de datos	Sensibilidad	Especificidad	Exactitud	Exactitud_bal
Submuestreo	0,771	0,912	0,842	0,842

Tabla 4-9: Modelo logístico datos con método Submuestreo.

4.4.2. Modelo con metodología Sobremuestreo

Cuando se aplica el modelo logístico binomial con la metodología de Sobremuestreo, incrementa la muestra con base a la clase mayoritaria No Impago. Al comparar los resultados con los modelos anteriores, se observa en la Tabla (**4-10**), valores de los coeficientes muy parecidos al modelo sin balanceo, en la mayoría de los casos un poco

mayores, donde el efecto en el incremento de alguna de las variables, también incrementa los casos de Impago. En este modelo a diferencia del modelo con balanceo Submuestreo, todas las variables presentan un nivel de significancia importante, con $P - valor \leq 0,05$.

En cuanto a las medidas de desempeño del modelo, la Tabla (**4-11**) presenta para los indicadores exactitud y la sensibilidad, mejores resultados respecto al modelo sin balanceo y que la metodología anterior. Ambos aumentaron en un 8,57 % y 35 % respectivamente. El valor del umbral en este caso es del 37,60 %.

Tabla 4-10: Modelo logístico Impago con datos método sobremuestreo.

Coeficientes	**Estimación**	**Error Est**	**Z Valor**	**P Valor**	$Pr(>\|z\|)$	**Exp(Estim)**
(Intercept)	-8,432	0,555	-15,199	< 2e-16	***	2,18e-04
Act Tot log	0,268	0,021	12,682	< 2e-16	***	1,307
Árbol Clas2	0,178	0,063	2,808	0,004981	**	1,195
Árbol Clas3	-1,023	0,134	-7,626	2,43e-14	***	0,36
Árbol Clas4	-0,503	0,14	-3,586	0,000335	***	0,605
Árbol Clas5	1,393	0,345	4,043	0,0000528	***	4,027
Árbol Clas6	0,111	0,085	1,305	0,19178		1,117
Buro2	2,03	0,075	27,22	< 2e-16	***	7,614
Buro3	3,255	0,092	35,436	< 2e-16	***	25,92
Cart total log	0,029	0,004	6,745	1,53e-11	***	1,029
Índice altman	-0,065	0,017	-3,828	0,000129	***	0,937
Mora actual	0,463	0,023	19,964	2e-16	***	1,589
PI	3,545	1,091	3,249	0,001159	**	34,64
SectorManufac	0,476	0,073	6,493	8,42e-11	***	1,61
SectorServicios	0,192	0,073	2,646	0,008141	**	1,212
Tam.Empr.media	0,377	0,115	3,289	0,001007	**	1,458
Tam.Empr.micro	0,771	0,134	5,739	9,5e-09	***	2,162
Tam.Empr.peq	0,46	0,126	3,664	0,000248	***	1,584
Fact log	-0,022	0,006	-3,928	0,0000858	***	0,978

Base de datos	Sensibilidad	Especificidad	Exactitud	Exactitud_bal
Sobremuestreo	0,795	0,899	0,848	0,848

Tabla 4-11: Modelo logístico datos con método Sobremuestreo.

El indicador de *Exactitud* o AUC, presenta en la Tabla (**4-11**) un porcentaje muy similar al obtenido en la metodología de Sobremuestreo 84,8 %, pero en esta oportunidad el modelo es más sensible en la predicción para detectar el Impago en las empresas.

4.4.3. Modelo Multinomial con variable respuesta Clasificación de riesgo

En éste caso se mide la capacidad de clasificación del modelo multinomial, con la variable de 3 escalas de Clasificación de riesgo. Recordemos que dicha variable tiene las siguiente clasificación: Riesgo Bajo, codificado con (1), Riesgo Bajo, codificado con (2) y Riesgo Alto, codificado con (3).

La regresión logística multinomial, se lleva a cabo cuando la variable respuesta es nominal con más de dos niveles, midiendo el efecto de los predictores sobre la probabilidad de éxito en cada categoría, en comparación con la categoría de referencia. La categoría de referencia corresponde al Riesgo bajo, codificado como (1).

Las probabilidades se obtienen de la siguiente manera:

$$\text{Log.mult} = \log\left(\frac{\pi = \text{Riesgo Medio}}{\pi = \text{Riesgo Bajo}} = x\beta\right), \tag{4-1}$$

$$\text{Log.mult} = \log\left(\frac{\pi = \text{Riesgo Alto}}{\pi = \text{Riesgo Bajo}} = x\beta\right). \tag{4-2}$$

El resumen del modelo, presenta una columna de coeficientes y otro correspondiente a los errores estándar. Cada columna tiene una fila de valores correspondientes a una ecuación modelo.

En la salida del resumen del modelo, la columna de coeficientes, muestran para la primera fila, la comparación del riesgo Medio con el riesgo de referencia que en éste caso corresponde al (Riesgo Bajo) y en la segunda fila el Riesgo Alto con la línea base (Riesgo Bajo). En la mayoría de los casos, dichos valores son mayores en el nivel de riesgo 3.

Los resultados en la Tabla (**4-12**), no muestran el resultado de los coeficientes para el nivel de referencia. Cada coeficiente muestra las pendientes de las regresiones logísticas de cada variable, respecto a dicho nivel de referencia. Sin embargo en el intercepto, se incluye el logit del nivel de referencia 1 (riesgo bajo) de las variables predictoras, lo anterior significa que los demás coeficientes expresan los diferenciales del logit en los otros dos niveles de cada variable predictora. En cuanto al estimador Residual Deviance: $7143,182$, indica que el resultado de este valor dividido entre dos, es decir $3571,59$, corresponde al historial de iteraciones del modelo.

Al igual que en el modelo logístico, los coeficientes que presentan signos positivos, influyen negativamente en la escala de riesgo que tendrá la empresa, cuando aumenten unidades de dichas variables, la escala de riesgo podrá estar en Riesgo Medio (2), o en Riesgo Alto (3). Para las variables que presentan coeficientes con signos negativos, un aumento de unidades en dicha variable, influirá en que la escala de riesgo para la empresa sea mejor, pasando de Riesgo Alto a medio, o de Riesgo Alto a riesgo bajo según el caso.

Por ejemplo, para la variable probabilidad de incumplimiento, un aumento en una unidad de dicha variable, aumentará $13,52714$ el logit de que la empresa se encuentre en el Riesgo Medio, respecto al riesgo Bajo. De manera análoga, aumentar una unidad de la variable probabilidad de incumplimiento, aumenta en $21,77802$ el logit de que la empresa se encuentre en el Riesgo Alto, respecto al riesgo Bajo.

Ahora si interpretamos el riesgo relativo por medio del inverso de los logs, se tendrá por ejemplo en la variable índice Altman la siguiente interpretación: El riesgo relativo de que la empresa, se encuentre en riesgo medio versus en riesgo Alto, por una unidad que incremente el indicador Altman, es $0,25744403$. Para el caso de la variable edad de mora, el valor exponencial de su coeficiente es de $e^{(0,007729274)} = 1{,}0077$, lo significa que por cada incremento en una unidad de la edad de mora que presente la empresa, la posibilidad de que esté en riesgo medio en lugar de un riesgo bajo se incrementa un 0,0077, manteniendo las demás variables constantes.

A continuación se expone la significancia de las variables en el modelo. De acuerdo con los resultados de los valores P: Activos totales, cartera total, mora actual y algunos niveles del árbol de clasificación y tamaño de las empresas, dejan de ser significativos si comparamos con los modelos anteriores.

Tabla 4-12: Modelo Multinomial con variable clasificación de riesgo.

	Estimación		**Error Estándar**	
	2	**3**	**2**	**3**
(Intercept)	1,111371	-1,822048	0,6850201	1,7934201
Act Tot log	0,002637145	0,011888589	0,02624665	0,06751081
Árbol Clas2	-0,03456075	-0,19506227	0,07821092	0,21884487
Árbol Clas3	0,560742	-0,5049864	0,2057594	0,7512474
Árbol Clas4	0,8195046	1,1495027	0,2031672	0,4157228
Árbol Clas5	1,120205	1,449137	0,3960932	0,7829952
Árbol Clas6	1,190465	1,124732	0,1135195	0,2660763
Buro2	0,6009153	0,3288897	0,1234285	0,3250501
Buro3	2,247986	2,735333	0,1290573	0,2859084
Cart total log	0,004257361	0,015312907	0,005667441	0,015437803
Índice altman	-1,356953	-3,874118	0,02749374	0,27166246
Mora actual	0,007729274	-0,005083344	0,01137942	0,02603391
PI	13,52714	21,77802	2,744184	2,874818
SectorManufactura	0,2984479	0,1241063	0,09146838	0,26283394
SectorServicios	-0,2988276	-0,5073623	0,08876448	0,25613685
Tam.Empr mediana	0,2647423	0,7455968	0,1550197	0,5002351
Tam.Empr micro	1,126588	2,22175	0,1757293	0,5201492
Tam.Empr pequeña	0,6481281	1,1558177	0,1666686	0,5166714
Fact log	-0,0215595	-0,06219291	0,007904792	0,024071857

Tabla 4-13: Modelo Multinomial con variable Clasificación de riesgo - valores P.

	Puntuación Z		**Valores** P			
	2	**3**	**2**	**3**	**2**	**3**
(Intercept)	3,039	0,162	0	0	***	***
Act Tot log	1,003	1,012	0,92	0,86		
Árbol Clas2	0,966	0,823	0,659	0,373		
Árbol Clas3	1,752	0,604	0,006	0,501	***	
Árbol Clas4	2,269	3,157	0	0,006	***	**
Árbol Clas5	3,065	4,259	0,005	0,064	**	*
Árbol Clas6	3,289	3,079	0	0	***	***
Buro2	1,824	1,389	0	0,312	***	
Buro3	9,469	15,415	0	0	***	***
Cart total log	1,004	1,015	0,453	0,321		
Índice altman	0,257	0,021	0	0	***	***
Mora actual	1,008	0,995	0,497	0,845		
PI	749480,8	2871263620,8	0	0	***	***
SectorManufactura	1,348	1,132	0,105	0,31		
SectorServicios	0,742	0,602	0,001	0,637	***	
Tam.Empr mediana	1,303	2,108	0,001	0,048	***	**
Tam.Empr micro	3,085	9,223	0,088	0,136	*	
Tam.Empr pequeña	1,912	3,177	0	0,025	***	**
Fact log	0,979	0,94	0,006	0,01	***	***

La Tabla (**4-14**) y Tabla (**4-15**), corresponden a las matrices de confusión de los datos de entrenamiento y testeo respectivamente.

	1	2	3
1	5649	615	0
2	757	2355	10
3	0	168	32

Tabla 4-14: Matriz de confusión modelo Multinomial, Datos entrenamiento.

	1	2	3
1	1398	168	0
2	195	585	1
3	0	45	5

Tabla 4-15: Matriz de confusión modelo Multinomial, Datos testeo.

Ahora bien, con el fin de poder comparar los modelos aplicados para la variable respuesta de tres niveles, también se tiene en cuenta las mediciones de predicción general, y las métricas de sensibilidad y especificidad. Para el modelo en la Tabla (**4-12**), el indicador AUC o predicción general del modelo, corresponde a un 83 % tanto en la base de datos de entrenamiento como la de testeo.

Como se trata de un modelo politómico o cuya variable respuesta presenta más de tres niveles de clasificación, la Tabla (**4-16**) referencia los porcentajes de sensibilidad (1), sensibilidad (2), sensibilidad (3) y sensibilidad promedio, correspondientes a las Escalas de riesgo (1,2 y 3), respectivamente. No se puede calcular un indicador general de sensibilidad para el modelo, pero se considera el promedio de los anteriores, con el fin de homologar a un indicador de sensibilidad general. De igual manera la Tabla (**4-17**), presenta el mismo esquema para la especificidad en función de los 3 niveles de riesgo.

Base de datos	Sensibilidad (1)	Sensibilidad (2)	Sensibilidad (3)	Prom Sensibilidad
Multin train	0,88	0,75	0,76	0,80
Multin test	0,88	0,73	0,83	0,815

Tabla 4-16: Sensibilidad modelo Multinomial.

Base de datos	Especificidad (1)	Especificidad (2)	Especificidad (3)	Prom Especif
Multin train	0,81	0,88	0,98	0,89
Multin test	0,79	0,88	0,98	0,88

Tabla 4-17: Especificidad modelo Multinomial.

Los porcentajes anteriores son importantes, teniendo en cuenta que en esta variable respuesta presentaba muy poca información para la predicción de la escala C (Correspondiente al Riesgo Alto). Ver Tabla (**4-16**).

Clasificación	Cantidad	%
1	7830	65,34 %
2	3903	21,57 %
3	250	2,09 %

Tabla 4-18: Clasificación de Riesgo en tres escalas.

4.4.4. Modelo Multinomial ordinal con variable respuesta Clasificación de riesgo

En esta sección se presenta la capacidad de predicción aplicando el modelo Multinomial ordinal, teniendo en cuenta la variable respuesta de clasificación de riesgo que posee características de escalas secuenciales u ordenadas.

Para construir este modelo, se utilizó la función en el software estadístico R "polr - Ordered Logistic or Probit Regression", del paquete MASS (Venables y Ripley, 2002) que es apropiada para estimar éste tipo de modelos. Se indica el comando "Hess=TRUE", para permitir que la salida del modelo muestre la matriz de información observada de la optimización que se usa para obtener errores estándar. También es importante iniciar con la ordenación de la variable respuesta.

La Tabla (**4-19**) muestra los resultados estimados de los coeficientes de salida de la regresión, los cuales van acompañados de sus errores estándar, valores del estadístico t, estimaciones para las dos intersecciones, desviación residual y AIC. Al tomar como ejemplo la variable Cartera total, las probabilidades de que se presente un Riesgo Alto o Medio frente a un Riesgo "Bajo" son $1,006$ veces mayores cuando se da un aumento en el valor de la cartera total.

Tabla 4-19: Clasificación de Riesgo en tres escalas.

Coefficients:	**Estimate:**	**Exp(Estimate) - OR**	**Std. Error**	*t*-valor	$Pr(>\|z\|)$	
Act Tot log	0,003	1,003	0,024	0,124	0,902	
Árbol Clas2	-0,062	0,94	0,074	-0,833	0,405	
Árbol Clas3	0,39	1,477	0,19	2,05	0,04	*
Árbol Clas4	0,735	2,085	0,171	4,306	1,7e-05	***
Árbol Clas5	0,828	2,289	0,316	2,625	0,009	**
Árbol Clas6	0,92	2,508	0,1	9,231	2,7e-20	***
Buro2	0,533	1,704	0,113	4,724	2,3e-06	***
Buro3	1,902	6,701	0,111	17,157	5,6e-66	***
Cart total log	0,006	1,006	0,005	1,216	0,224	
Índice altman	-1,332	0,264	0,025	-52,909	0	***
Mora actual	0,006	1,006	0,01	0,642	0,521	
PI	5,233	187,285	0,381	13,727	7e-43	***
SectorManufactura	0,244	1,277	0,085	2,867	0,004	***
SectorServicios	-0,254	0,776	0,083	-3,049	0,002	***
Tam.Empr mediana	0,272	1,312	0,145	1,871	0,061	*
Tam.Empr micro	1,091	2,978	0,163	6,695	2,2e-11	***
Tam.Empr pequeña	0,603	1,827	0,155	3,882	0	***
Fact log	-0,015	0,985	0,007	-2,228	0,026	***
1|2	-1,212	0,298	0,629	-1,927	0,054	*
2|3	4,294	73,258	0,637	6,738	1,6e-11	***

Para analizar la predicción del modelo y comparar más adelante los indicadores de AUC (Exactitud), sensibilidad y especificidad, obtenemos en la Tabla (**4-20**) y Tabla (**4-21**), las matrices de confusión de los datos de entrenamiento y testeo respectivamente para el modelo multinomial ordinal. El nivel de predicción general del presente modelo, arroja un valor de muy similar al modelo Multinomial, y corresponde a un 83,74 %.

	1	2	3
1	5587	698	0
2	677	2401	160
3	0	23	40

Tabla 4-20: Matriz de confusión modelo Multinomial Ordinal, Datos entrenamiento.

	1	2	3
1	1387	180	0
2	179	598	45
3	0	3	5

Tabla 4-21: Matriz de confusión modelo Multinomial Ordinal, Datos testeo.

Algo similar sucede con el indicador de sensibilidad para clase 1 y 2 que son ligeramente más altos en un 1 %. Sin embargo llama la atención como para la clase 3, el modelo es poco sensible para predecir las empresas que realmente entran en la clase de Riesgo Alto, Tabla (**4-22**).

Base de datos	Sensibilidad (1)	Sensibilidad (2)	Sensibilidad (3)	Prom Especif
Multin train	0,90	0,77	0,2	0,62
Multin test	0,89	0,77	0,1	0,58

Tabla 4-22: Sensibilidad modelo Multinomial Ordinal.

Para el caso de la especificidad, el modelo ordinal es capaz de identificar mejor los verdaderos negativos, ver Tabla (**4-23**):

Base de datos	Especificidad (1)	Especificidad (2)	Especificidad (3)	Prom Especif
Multin train	0,80	0,87	0,99	0,89
Multin test	0,79	0,86	0,99	0,88

Tabla 4-23: Especificidad modelo Multinomial.

4.4.5. Modelo máquinas de soporte con variable respuesta dicótoma

Modelo máquinas de soporte para Impago con datos balanceados

Retomando los porcentajes que se consideraron para la base de datos de entrenamiento 80 % y de testeo 20 %, se inicia con la ejecución del modelo, sobre las bases de datos balanceadas con submuestreo y sobremuestreo, que como vimos en la sección del modelo (4) logístico binomial, fueron los que mayor porcentaje arrojaron para indicadores como Exactitud, Especificidad y Sensibilidad.

Teniendo en cuenta que de manera apriori no se conocía que tipo de kernel aplicaba para la base de datos, se utilizaron las opciones que ofrece el paquiete SVM, kernel: Lineal, Polinomial y radial, con el fin de identificar que tipo de kernel se ajustaba mejor.

Hiperparámetros para datos de Submuestreo y Sobremuestreo

El modelo SVM debe contar inicialmente, con la identificación de un hiperparámetro C (Costeo), que controle el sesgo-varianza y la capacidad predictiva del modelo, ya que con ésto se logrará una mejor separación de las observaciones de manera previa. Para esto, se utilizó la metodología de validación cruzada para definir el mejor valor del hiperparámetro y la función “tune” del paquete e1071 (Meyer, Dimitriadou, Hornik,

Weingessel, y Leisch, 2022).

Para el caso de los datos balanceados con el método de Submuestreo, se obtuvo que el costo con menor error de validación y dispersión corresponde a un $C = 0{,}1$ y para la base de datos con el método de Sobremuestreo, el valor fue de $C = 5$, ver Tabla (**4-24**) y Tabla (**4-25**).

Cost	Error	Dispersión
0,001	0,2016866	0,02886589
0,01	0,1640763	0,01665619
0,1	0,1511535	0,01880157
1	0,1515456	0,01941905
5	0,1515456	0,01941905
10	0,1515456	0,01941905
15	0,1515456	0,01941905
20	0,1515456	0,01941905

Tabla 4-24: Hiperparámetro Costo con Submuestreo.

Cost	Error	Dispersión
0,001	0,1788934	0,01383793
0,01	0,1720124	0,01295159
0,1	0,1630161	0,01255178
1	0,1617646	0,01209798
5	0,1611389	0,0126172
10	0,1612954	0,01237729
15	0,1611389	0,0126172
20	0,1614518	0,01216254

Tabla 4-25: Hiperparámetro Costo con Sobremuestreo.

Matrices de confusión para Submuestreo y Sobremuestreo

A continuación se muestra para la base de datos balanceada con Submuestreo y Sobremuestreo en cada uno de los kernels, el resultado de las matrices de confusión y métricas de predicción de exactitud, sensibilidad y especificidad, tal como se observa en la Tabla (**4-26**) y Tabla (**4-29**) respectivamente:

Submuestreo

Data de entrenamiento			**Data de testeo**		
	0	**1**		**0**	**1**
0	1157	268	**0**	290	78
1	120	1009	**1**	29	241
Kernel Lineal			**Kernel Lineal**		
Data de entrenamiento			**Data de testeo**		
	0	**1**		**0**	**1**
0	1267	826	**0**	317	227
1	10	451	**1**	2	92
Kernel Polinómico			**Kernel Polinómico**		
Data de entrenamiento			**Data de testeo**		
	0	**1**		**0**	**1**
0	1183	314	**0**	294	89
1	94	963	**1**	25	230
Kernel radial			**Kernel radial**		

Tabla 4-26: Matrices de confusión para todos los kernels - Submuestreo.

Sobremuestreo

Data de entrenamiento			**Data de testeo**		
	0	**1**		**0**	**1**
0	5733	1388	**0**	1432	166
1	659	5004	**1**	340	1258
Kernel Lineal			**Kernel Lineal**		
Data de entrenamiento			**Data de testeo**		
	0	**1**		**0**	**1**
0	5745	1301	**0**	1424	314
1	647	5091	**1**	318	1280
Kernel Polinómico			**Kernel Polinómico**		
Data de entrenamiento			**Data de testeo**		
	0	**1**		**0**	**1**
0	5745	1301	**0**	1425	293
1	647	5091	**1**	173	1305
Kernel radial			**Kernel radial**		

Tabla 4-27: Matrices de confusión para todos los kernels - Sobremuestreo.

Métricas de predicción para Submuestreo y Sobremuestreo

A continuación se presenta el resumen para la variable respuesta Impago, con los resultados correspondientes a las medidas de clasificación en cada una de las bases de datos balanceados y de acuerdo a los diferentes kernels: Lineal, polinómico y Radial, se muestran en la Tabla (**4-28**).

Los modelos aplicados a la variable dicotómica Impago presentan los siguientes resultados:

El modelo logístico presenta un nivel de predicción general (Exactitud) superior al 78 %, incluso si la muestra presenta desbalance en sus clases, sin embargo estima mejor los verdaderos casos en los que no se presenta el Impago que los verdaderos, lo que sugiere utilizar las muestras balanceadas que presentan mejor predicción en los resultados verdaderos de Impago, que representa una señal más importante para controlarlos.

Los modelos con Submuestreo, predicen mejor el indicador de Especificidad, mientras que los modelos con Sobremuestreo, presentan mejores porcentajes en los indicadores de sensibilidad. De los modelos aplicados para la variable Impago, los mejores resultados se observan en los modelos de máquina de soporte, destacándose el correspondiente al kernel lineal. Sin embargo no son tan distantes del modelo con Sobremuestreo de la regresión logística que presenta igualmente resultados muy aceptables en los indicadores de desempeño de sensibilidad, especificidad y Exactitud (AUC).

Modelo	**Base datos**	**Sensibilidad**	**Especificidad**	**Exactitud**
Logístico	Sin Balancear	0,5889724	0,9724725	0,7807225
	Submuestreo	0,7711599	0,9122257	0,8416928
	Sobremuestreo	0,795995	0,8992491	0,847622
SMV Lineal	Submuestreo	0,7554859	0,9090909	0,8322884
	Sobremuestreo	0,787234	0,8961202	0,8416771
SMV polinomial	Submuestreo	0,7084639	0,9216301	0,815047
	Sobremuestreo	0,8035044	0,8911139	0,8473091
Radial	Submuestreo	0,7210031	0,9216301	0,8213166
	Sobremuestreo	0,8166458	0,8917397	0,8541927

Tabla 4-28: Comparación diferentes metodologías - SVM - variable Impago.

4.4.6. Modelo máquinas de soporte con variable respuesta policótoma

En esta ocasión se aplica para la variable respuesta Clasificación de 3 niveles de riesgo, el modelo de máquinas de soporte.

Partiendo de la misma proporción definida en los anteriores modelos, se cuenta con datos de entrenamiento 80 % y de testeo 20 %, respectivamente. Solo que en este caso dado que no hay posibilidades de balancear los datos, se trabaja con datos estratificados.

Hiperparámetros para datos de clasificación de escalas de riesgos

Se utiliza por medio del paquete caret (Kuhn, 2021), un método de validación cruzada con 10 particiones y 5 repeticiones. De cuerdo con la metodología, se promedió los valores de Exactitud obtenidos para cada valor del hiperparámetro, se identifica cual es el mejor de los valores mostrados en la Tabla (**4-29**) y se reajusta el modelo empleando todas las observaciones de entrenamiento y el mejor valor del hiperparámetro.

Los valores C probados a un $C = 10$, consigue los mejores resultados con un indicador de Exactitud de 0.8380549.

Cost	Exactitud	Kappa
0,001	0,8311066	0,6341593
0,01	0,835676	0,6456975
0,1	0,837554	0,6494058
0,5	0,8379921	0,650509
1	0,8379922	0,6505797
10	0,8380549	0,6507184

Tabla 4-29: Validación cruzada hiperparámentro con variable de 3 niveles.

Para analizar la predicción del modelo y comparar más adelante los indicadores de exactitud (AUC), sensibilidad y especificidad, obtenemos en la Tabla (**4-34**) a la Tabla (**4-35**), las matrices de confusión de los datos de entrenamiento y testeo respectivamente, para el modelo de máquinas de soporte politómica con un kernel lineal. En el eje horizontal se comparan los niveles actuales, versus los predichos en el eje vertical:

	1	2	3
1	5560	644	0
2	704	2468	176
3	0	10	24

Tabla 4-30: (Train) - Matriz de confusión modelo SVM 3 niveles - Kernel lineal.

	1	2	3
1	1382	184	0
2	157	624	0
3	0	45	5

Tabla 4-31: (Testeo) - Matriz de confusión modelo SVM 3 niveles - Kernel lineal.

Para un kernel polinomial las matrices son las siguientes:

	1	2	3
1	5843	845	0
2	421	2277	154
3	0	0	46

Tabla 4-32: (Train) - Matriz de confusión modelo SVM 3 niveles - Kernel polinomial.

	1	2	3
1	1430	136	0
2	241	538	2
3	0	44	6

Tabla 4-33: (Testeo) - Matriz de confusión modelo SVM 3 niveles - Kernel polinomial.

Para un kernel radial las matrices son las siguientes:

	1	2	3
1	5718	668	0
2	546	2451	159
3	0	3	41

Tabla 4-34: (Train) - Matriz de confusión modelo SVM 3 niveles - Kernel radial.

	1	2	3
1	1402	164	0
2	205	575	1
3	0	41	9

Tabla 4-35: (Testeo) - Matriz de confusión modelo SVM 3 niveles - Kernel radial.

Medidas de desempeño para variable 3 niveles en SVM

Utilizando la variable respuesta de 3 niveles, se obtuvieron los siguientes resultados mostrados en la Tabla (**4-36**). Allí para la variable Clasificación del Riesgo correspondiente a la etiqueta de 3 clases, el nivel de predicción general es muy similar en todos los modelos con resultados superiores al 82 %, destacándose el modelo de máquinas de soporte con Kernel Lineal y el modelo ordinal con un $83,9\,\%$ y un $83,02\,\%$ respectivamente. Sin embargo, para el modelo de máquinas de soporte con kernel lineal, existen resultados más estables en todos los niveles de sensibilidad de cada clase, lo que arroja un promedio en el indicador de sensibilidad bastante aceptable del $87,65\,\%$. Por su parte los indicadores de especificidad también muestran porcentajes similares con valor promedio superior al $88,12\,\%$ en todos los niveles.

Modelo	**Sensib (A)**	**Sensib (B)**	**Sensib (C)**	**Prom.sens**	**Exactitud**
Multinomial	0,8776	0,7331	0,8333	0,8147	0.8294
Ordinal	0,8857	0,7657	0,1	0.5838	0.8302
SMV.lineal	0,898	0,7315	1	0.8765	0.8390
SMV.polinomial	0,8558	0,7493	0,75	0.7850	0.8235
SMV.radial	0,8724	0,7372	0,9	0.8365	0.8285
	Especif (A)	**Especif** (B)	**Especif** (C)	**Promedio.Esp**	
Multinomial	0,791	0,8774	0,9812	0,8832	
Ordinal	0,7834	0,8614	0,9987	0,8812	
SMV.lineal	0,7855	0,8983	0,9812	0,8884	
SMV.polinomial	0,8127	0,8553	0,9816	0,8832	
SMV.radial	0,7924	0,8726	0,9828	0,8826	

Tabla 4-36: Comparación diferentes metodologías - SVM - variable Impago.

Como se observa, tanto para la variable dicotómica Impago como para la variable de 3 niveles correspondiente a la clasificación de riesgo, se obtuvieron medidas de desempeño con buena predicción general en la mayoría de los modelos. Estos mejoran en el caso de variable Impago para los indicadores de sensibilidad cuando se aplica el balanceo de la muestra con Sobremuestreo. Para el caso de la variable de Clasificación de Riesgo que es de tres clases, la medida general de aciertos Exactitud, está por encima del 82 %. Allí el modelo aplicado con máquinas de soporte con kernel lineal presentó un indicador de sensibilidad más alto en comparación con las otras metodologías.

5 Conclusiones y recomendaciones

Proponer una metodología para clasificar el riesgo de crédito, permitió no solo explorar alternativas desde las metodologías de clasificación estadística, sino también poder determinar por medio de los análisis descriptivos que tanta relación hay entre la variable respuesta y sus variables explicativas. Es importante que exista una buena conexión entre éstas, para que los resultados sean congruentes y haya un buen punto de partida en el análisis.

Para la variable dicótoma Impago, todos los modelos de clasificación presentaron una exactitud general por encima del 78 %, siendo menor el resultado en el modelo logístico sin datos balanceados. Éste último un modelo fue poco sensible para identificar los casos de las empresas que realmente entran en Impago, con un 58 %.

Al aplicar las metodologías de balanceo sobre la variable Impago, tales como Sobremuestreo o Submuestreo, éstas arrojaron indicadores muy buenos no solo en la predicción general, sino también en medidas como la sensibilidad con un porcentaje de alrededor del 80 %.

Así mismo, al aplicar metodologías no supervisadas como máquinas de soporte vectorial, también se obtuvieron resultados muy altos superiores al 88 %, tanto para las bases de datos balanceados con Submuestreo como Sobremuestreo. Los modelos con Sobremuestreo se destacan en indicadores de sensibilidad identificando mejor la presencia del Impago, por su parte los modelos con Submuestreo predicen mejor los verdaderos negativos o con No Impago

Para el caso de la variable de tres niveles correspondiente a la clasificación de Riesgo de Crédito, la escala de riesgo alto solo participaba con un 2 % en el total de los datos. Sin embargo, no se pudo obtener opciones claras en la literatura, para el balanceo de datos con más de dos escalas en la variable respuesta. Los modelos de clasificación Multinomial, Multinomial ordinal, y máquinas de soporte, presentaron porcentajes de clasificación general superiores al 82 %. En el caso del modelo Multinomial ordinal, presentaron una baja sensibilidad en la detección de los casos reales de deterioro en

las empresas para la escala de Riesgo alto.

Métricas, para identificar el poder de clasificación de los modelos, para variables politómicas, no son muy usuales en la literatura, pero permiten lograr alternativas de comparación con los modelos de dos repuestas. Aunque se puede lograr un indicador de exactitud general para los modelos de dos y tres respuestas, indicadores como la sensibilidad y Especificidad se obtienen en el caso de la variable respuesta de tres niveles por cada una de las categorías, lo que no permite de manera técnica una comparación con los resultados obtenidos de una variable de dos niveles. Se optó finalmente por tomar el promedio de la sensibilidad y especificidad de la variable respuesta de riesgo en tres niveles.

Cada uno de los modelos, presentan ventajas y desventajas en los resultados, al final se deben ponderar elementos de tiempo de procesamiento, facilidad en la obtención de los datos y evitar posibles sobreajustes. Aunque modelos de tipo no supervisados obtuvieron mejores resultados, el tiempo de procesamiento puede significar un elemento de decisión para optar por modelos como el modelo logístico binomial con datos balanceados - Sobremuestreo, que permitió obtener los resultados en la clasificación y cuyas variables resultaron significativas de manera importante. Adicionalmente modelos no supervisados, no permiten analizar de manera clara la influencia de las variables predictoras.

Un trabajo posterior importante para los modelos logísticos será la aplicación de tablas de desempeño en la identificación de las empresas con impago o no, con el fin de poder obtener las clasificaciones en escala de más de dos niveles y lograr una mayor segmentación de los resultados.

6 Referencias

Agresti, A. (2003). *Categorical data analysis*. John Wiley Sons.

Altman, E. (1966). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. , *23(4)*(1), 589–609.

Arias, M. (2011). Fundamentos de medicina basada en la evidencia análisis de las causas de la variabilidad en la práctica médica. *Evidencias Pediatricas*, 7–21. Descargado de `https://evidenciasenpediatria.es/articulo/5578/enlace`

Berrío Guzmán, D., y Cabeza de Vergara, L. (2003). Verificación y adaptación del modelo de altman a la superintendencia de sociedades de colombia.

Blissett, R. (2017). Logistic regression in R. *Obtenido de net: https://rpubs.com/rslbliss/r_logistic_ws*.

De la República, C. (2006). Ley 1116 de 2006. *Diario Oficial*(46.494).

Dictionary, C. (2023). Cambridge advanced learner's dictionary. *Recuperado de: https://dictionary.cambridge.org/es/*.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of eugenics*, *7*(2), 179–188.

Gonzalez, L. (2019). Conjunto de datos desbalanceados). *Obtenido de net: https://aprendeia.com/conjunto-de-datos-desbalanceado/*.

Hadad, A., Evin, D., y Drozdowicz, B. (2009). Modelo para el tratamiento de datos desbalanceados basado en redes neuronales autoorganizadas. En *XVII Congreso Argentino de Bioingeniería, Rosario, Santa Fe.*

Hogg, R. V., McKean, J. W., y Craig, A. T. (2013). *Introduction to mathematical statistics* (Seventh ed.). Pearson.

Hosmer, D. W., Jovanovic, B., y Lemeshow, S. (1989). Best subsets logistic regression. *Biometrics*, *Vol.45*(N°4), 1265–1270.

James, G., Witten, D., Hastie, T., y Tibshirani, R. (2013). *An introduction to statistical learning with applications in R* (Vol. 112). Springer.

Kuhn, M. (2021). caret: Classification and regression training [Manual de software informático]. Descargado de `https://CRAN.R-project.org/package=caret` (R package version 6.0-90)

Linares Galván, J. E. (2010). Gestión del riesgo: pautas generales ofrecidas por la norma técnica colombiana–NTC 5254 95. *Apuntes contables*, *Vol.11*.

Lumley, T. (2020). leaps: Regression subset selection [Manual de software informático]. Descargado de `https://CRAN.R-project.org/package=leaps` (R package

version 3.1)

Mahto, A. (2019). splitstackshape: Stack and reshape datasets after splitting concatenated values [Manual de software informático]. Descargado de `https://CRAN.R-project.org/package=splitstackshape` (R package version 1.4.8)

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., y Leisch, F. (2022). e1071: Misc functions of the department of statistics, probability theory group (formerly: E1071), tu wien [Manual de software informático]. Descargado de `https://CRAN.R-project.org/package=e1071` (R package version 1.7-12)

Moreno Valencia, S. (2012). El modelo logit mixto para la construcción de un scoring de crédito. *Escuela de Estadística*.

Norman, G. G. (2010). Islr: Gestion cuantitativa de riesgo de credito, con aplicaciones en R [Manual de software informático].

Rodrigo, J. A. (2017). Máquinas de vector soporte (support vector machines, svms). *Obtenido de cienciadedatos, Abril. net: https://www. cienciadedatos. net/documentos/34_maquinas_de_vector_soporte_support_vector_machines*.

Sanchez-Roger, M., Oliver-Alfonso, M. D., y Sanchís-Pedregosa, C. (2020). Capacidad total de absorción de pérdidas–hacia una metodología simple y eficiente. *Cuadernos de Gestión*, *20*(2), 199–222.

Statistical Odds & Ends. (2020). What is balanced accuracy? [Manual de software informático]. Descargado de `https://tinyurl.com/jsvezz8k`

Superintendencia Financiera, I., de Colombia Capítulo. (1995). Reglas relativas a la gestión del riesgo crediticio. *Circular Externa*, *100*(95), 1–31.

OVF, O. V. d. F. (2021). Normas de Basilea. *Obtenido de net: https://observatoriodefinanzas.com/wp-content/uploads/NORMAS-DE-BASILEA.pdf*.

Tharwat, A. (2021). Classification assessment methods. *Applied computing and informatics*, *17*(1), 168–192.

Thomas, L. (2020). Stratified sampling definition, guide examples). *Obtenido de net: https://www.scribbr.com/methodology/stratified-sampling/* .

Trueck, S., y Rachev, S. T. (2009). *Rating based modeling of credit risk: theory and application of migration matrices*. Academic press.

Trujillo Ospina, A., Belalcázar Grisales, R., y cols. (2016). *¿Es el modelo z-score de altman un buen predictor de la situación financiera de las pymes en colombia?* (Tesis Doctoral no publicada). Universidad EAFIT.

Vapnik, V., Golowich, S. E., Smola, A., y cols. (1997). Support vector method for function approximation, regression estimation, and signal processing. *Advances in neural information processing systems*, *Vol.9*, 281–287.

Venables, W. N., y Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth ed.). New York: Springer. Descargado de `https://www.stats.ox.ac.uk/pub/MASS4/` (ISBN 0-387-95457)

Verona Martel, M. C. (2007). El rating como evaluación de la calidad crediticia de las empresas. *Innovar*, *17*(29), 195–196.