Universidad de San Andrés

Escuela de Administración y Negocios

Magister en Finanzas

Modelos de *machine learning* aplicados a la estimación de la probabilidad de *default* de entidades bancarias del sistema bancario argentino

Autor: Mariano Ezequiel Bozzi

DNI: 28.280.898

Director de Tesis: Mg. Ing. Federico Filgueira

Ciudad Autónoma de Buenos Aires, Marzo de 2023

Resumen

A lo largo del trabajo se desarrollan distintos modelos de *machine learning* con el fin de generar modelos predictivos asociados al *default* de entidades bancarias. A tal efecto, se lleva a cabo un relevamiento de la información disponible en el BCRA[1], cuyo resultado final es la generación de una base de información con decenas de indicadores. Ésta es, junto a la metodología desarrollada, la piedra angular del modelo desarrollado. El objetivo final de dicho modelo es cuantificar, dado un conjunto de valores de un *set* de indicadores, la probabilidad de *default* que posee la entidad bancaria que se analice. Por otro lado, en la sección correspondiente se detalla el modo en que se aplica el aprendizaje automático supervisado al desarrollo del modelo predictivo en cuestión y, por último, se resumen los resultados alcanzados a partir del abordaje de los parámetros obtenidos y su poder explicativo.

Palabras clave: *Machine Learning*, Probabilidad de *default*, PD, BCRA, Entidades Bancarias Argentinas.

[1]El BCRA es una entidad autárquica del Estado Nacional Argentino. Tiene por finalidad promover, en la medida de sus facultades y en el marco de las políticas establecidas por el gobierno nacional, la estabilidad monetaria, la estabilidad financiera, el empleo y el desarrollo económico con equidad social.

Agradecimientos

Esta tesis es fruto de un trabajo arduo y de gran crecimiento personal. Es importante destacar mi agradecimiento hacia las personas que ayudaron a sortear las dificultades durante la investigación.

En primer término a mi tutor Federico Filgueira, por creer en mi, y haberme brindado su guía, compromiso y dedicación sin lo cual este trabajo no hubiese sido posible. Además, es mi deseo agradecer a la Universidad de San Andrés y al cuerpo docente por su excelencia académica y por haber provisto el ámbito necesario para crecer tanto en lo profesional y como en lo personal, especialmente a Gabriel Basaluzzo y Elsa Cortina por todo su apoyo a lo largo del proceso de aprendizaje.

Por último, quiero agradecer al personal de la Biblioteca Prebish del BCRA por la información suministrada, haciendo una mención especial para Nicolás Rodríguez, cuya predisposición y colaboración en la búsqueda de información fueron claves para alcanzar los resultados del presente trabajo.

Índice

1. Introducción

Los sistemas bancarios se encuentran sujetos, a nivel global, a una serie de regulaciones aplicadas por las respectivas autoridades monetarias de cada país. Uno de sus objetivos es mitigar los riesgos en sus diferentes aspectos en balance con los niveles de rentabilidad.

En el caso de la Argentina, la ley 24.144 (Carta Orgánica) y la ley 21.526 establecen al Banco Central de la República Argentina (BCRA) como entidad de aplicación. Dentro de sus funciones y facultades, se destaca la de regular el funcionamiento del sistema bancario, aplicar la Ley de Entidades Financieras y las normas que, en su consecuencia, se dicten. Asimismo, supervisa las actividades financieras y cambiarias por intermedio de la Superintendencia de Entidades Financieras y Cambiarias (SEFyC), la que dependerá directamente del presidente de la institución. Dentro de sus facultades, el superintendente deberá tener a disposición del Directorio y de las autoridades competentes información sobre la calificación de las entidades financieras y criterios utilizados para dicha calificación. Del mismo modo, puede cancelar la autorización para operar, aprobar los planes de regularización y/o saneamiento de las entidades, implementar y aplicar las leyes antes mencionadas, y establecer los requisitos que deben cumplir los auditores de las entidades financieras y cambiarias.

A fin de regular el funcionamiento del sistema bancario el BCRA cuenta con una serie de atribuciones, dentro de las cuales se destaca la emisión de normas referidas a lo que se denomina Régimen Informativo[2](RI). Este es un conjunto de normas a través de las cuales el BCRA requiere información con carácter de declaración jurada a entidades bancarias y a otros entes sujetos a su contralor, relacionada con su situación económica y financiera, el cumplimiento de regulaciones y sus actividades. En particular se destaca el RI - Plan de Cuentas[3], cuyo objetivo es generar el marco regulatorio mínimo sobre el cual las entidades establecen su contabilidad. Bajo dicho encuadre,

[2] http://www.bcra.gov.ar/SistemasFinancierosYdePagos/RRII-preguntas-frecuentes.asp

[3] http://www.bcra.gov.ar/pdfs/texord/manual.pdf

el BCRA, a través de sus diferentes medios, hace pública la información contable de cada entidad para cada período.

En este trabajo se desarrolla una metodología propia con el fin de demostrar los indicadores y modelos predictivos que permitan obtener las probabilidades de *default* a fin de diferenciar aquellas entidades bancarias que se encuentran en situación de *default* de aquellas que se encuentran en situación de no *default*. Esto se realiza a partir de la información contable pública, que es utilizada para la construcción de bases de datos de saldos por rubro contables significativos de cada entidad para cada período bajo estudio. Sustentado en las bases de datos del trabajo, y tomando en consideración los rubros contables relevantes y la relación entre estos, se define el universo de indicadores. Una vez definido el universo de indicadores, se lleva adelante un ejercicio comparativo con los indicadores significativos definidos por Fernández-Sainz and Llaugel (2011), donde se realiza un primer análisis de *default* de las entidades bancarias argentinas y su grado de certeza en función del estudio mencionado. Posteriormente, se construyen modelos predictivos utilizando la técnica de *machine learning* con el objetivo de generar la probabilidad de *default* a fin de obtener la separabilidad de entidades bancarias en *default* de aquellas cuya situación sea de no *default*. Para lograr este objetivo se lleva adelante una metodología propia que usa las bases de datos construidas y los indicadores definidos para obtener, a partir de herramientas de *machine learning*, indicadores relevantes que permitan la posterior modelización y obtención de las respectivas probabilidades de *default*. Por último, se utilizan herramientas de evaluación a partir de las que se obtienen los modelos e indicadores que cumplen con el objetivo de generación de probabilidad de *default*.

La estructura del trabajo es de la siguiente manera: la sección 2 contiene la revisión de la literatura; en la sección 3 se presentan el origen de los datos, la información que es utilizada y el pre-procesamiento que se lleva adelante, la definición de los periodos que serán analizados y la construcción de las bases de datos; la sección 4 contiene una breve reseña de los modelos que se utilizan y las herramientas de evaluación de dichos modelos; en la sección 5 se ana-

lizan los resultados sobre la replicación del trabajo de Fernández-Sainz and Llaugel (2011); en la sección 6 se desarrolla una metodología propia mediante la utilización de modelos de definición de parámetros y de modelos predictivos de aprendizaje automático supervisado; en la sección 7 se describen las conclusiones arribadas y en la sección 8 las posibles futuras discusiones.

2. Revisión literaria

Desde la década de 1970, el sistema financiero mundial discute, tanto desde un punto de vista teórico como práctico, la forma de medir la fortaleza de los sistemas bancarios. Para satisfacer esa necesidad de medición se observa la evolución en las distintas técnicas estadísticas aplicadas al análisis financiero. El objetivo, desde sus orígenes, es desarrollar un modelo de alertas tempranas donde surja de manera clara una probabilidad de *default* en función de los indicadores obtenidos de la información contable pública. La construcción de alertas tempranas comienza con el trabajo de Altman (1977) donde se incorpora por primera vez el discriminante lineal para la predicción en las asociaciones de préstamo y ahorro.

Luego, Martin (1977) profundiza el trabajo de Altman (1977) comparando el resultado de indicadores entre el modelo *logit* y su relación con el modelo del discriminante lineal. En este caso, lo que busca el autor es estimar los coeficientes mediante el modelo *logit* para reproducir probabilidades altas en los casos con problemas y bajas para aquellos bancos sin dificultades. El autor concluye que, para los datos utilizados en términos de clasificación entre bancos con problemas y sin problemas, ambos modelos presentaron similares resultados.

En Berg and Hexeberg (1994) se realiza un trabajo similar sobre el sistema bancario noruego. Se utilizó específicamente para el análisis de la crisis 1988-1992, la cual socavó de manera relevante la industria bancaria de Noruega con 25 de 150 unidades financieras en problemas. Berg utiliza la información contable de los bancos para ese periodo y realiza un análisis con los indicadores desarrollados previamente en el trabajo de Martin (1977). Aplica como modelo la regresión logística y encuentra, al igual que su par de los Estados Unidos, que los indicadores de adecuación de capital y calidad de activos presentan relevancia en un número de estudios. La conclusión general de este informe sostiene que los bancos en problemas de Noruega presentaron características similares a las de los bancos de Estados Unidos analizadas por Martin (1977).

En Argentina, la crisis que se desarrolla en 2001 lleva a la búsqueda de modelos econométricos que estimen la probabilidad de *default*. Para esto, Klein (2019) propone la construcción de un modelo *probit* utilizando información pública del BCRA correspondiente a diciembre de 2001 con el propósito de analizar la fragilidad de los bancos comerciales. Para ello dirigió su trabajo con el fin de replicar la metodología CAMEL (por las iniciales en inglés: Capital, Activo, Administración, Rentabilidad, Liquidez y Sensibilidad), utilizada por el BCRA. Esta se basa en realizar una evaluación global de la situación de una entidad, lo que da como resultado una calificación numérica que toma valores entre 1 y 5 (1 corresponde a la entidad con mejor calificación y 5 a la de peor calificación) en carácter de confidencial. El BCRA hace uso de esta metodología, ya que otorga un marco general para la evaluación de la solidez de las entidades supervisadas por parte de la SEFyC y del cumplimiento de las leyes y normas vigentes. El sistema de calificación considera características y funciones, incluyendo factores cualitativos y cuantitativos comunes a todas las categorías de entidades. Una de las consecuencias derivadas de la calificación otorgada es el nivel de capital regulatorio requerido por parte de la entidad de aplicación sobre la entidad supervisada. En este contexto debemos aclarar que, dentro de las regulaciones a las que se encuentran sujetas las entidades financieras, se destacan las normas sobre el capital regulatorio bancario. Uno de los aspectos prominentes de dicha regulación busca como objetivo analizar la suficiencia de capital como soporte de la absorción de perdidas inesperadas a la que cada entidad se viera expuesta. En este sentido, los acuerdos macro a nivel mundial establecen lineamientos objetivos sobre requisitos de capital. Estos son aplicados por las respectivas entidades de cada jurisdicción en función del apetito al riesgo que cada entidad lleve adelante según su modelo de negocio.

Calificación asignada	Valor de "k"
1	1
2	1,03
3	1,08
4	1,13
5	1,19

Cuadro 1: Factor vinculado a la calificación asignada a la entidad según la evaluación efectuada por la SEFyC.

Del Cuadro 1 observamos “k" el cual es el factor vinculado a la calificación asignada a la entidad según la evaluación efectuada por la SEFyC. Tal como surge del Cuadro 1, a medida que empeoran su calificación, dicho factor aumenta con lo cual las entidades bancarias deberán incrementar los requerimientos de capital regulatorio dado el escenario que presentan y el grado de exposición ante pérdidas inesperadas. De allí que Klein (2019) busque replicar dicha metodología mediante el modelo *probit*, basándose en la construcción de indicadores financieros que evalúen desde el punto de vista de la capitalización, calidad de activos, eficiencia, rentabilidad y liquidez. Como conclusión obtiene que el modelo presenta una capacidad de predicción explicativa del 94,4 %. El autor resalta lo consistente de los resultados con el escenario macro-económico y financiero que configuró la crisis sistémica del 2001.

Posteriormente, con el devenir y el mayor desarrollo de nuevos métodos, se busca la aplicación de dichas herramientas al desarrollo de mejores y más ajustados sistemas de alertas tempranas. Tal es el caso de un estudio Min and Lee (2005) donde se aplica la técnica de *support vector machine* (SVM)Vapnik (1998) al inconveniente de la predicción de bancos en situación de *default* y no *default*. En dicho caso se busca una solución global con mejor ajuste de generalización, con un universo acotado de datos de entrenamiento, en contraposición a las redes neuronales, las cuales deben contar con grandes cantidades de datos a fin de su entrenamiento. De esta forma, SVM utiliza un modelo lineal Min and Lee (2005) donde concluyen que el método puede ser aplicado

a un sistema de clasificación con pequeña cantidad de datos y evitando problemas de sobreajuste. Los resultados empíricos del trabajo demuestran que la técnica de SVM produjo mejores resultados que las técnicas de regresión logística, análisis por discriminante multivariable y *back-propagation neural network* (BPN).

Tras los logros del método SVM, se suma un nuevo trabajo Fernández-Sainz and Llaugel (2011), que compara dos diferentes métodos: la regresión logistica y SVM Vapnik (1998). Ambos métodos se aplican a la información contable de los bancos comerciales durante la crisis bancaria ocurrida entre 2002 y 2004 en República Dominicana. Este trabajo reconoce que dentro de los indicadores más utilizados en los países industrializados para determinar bancos en problemas, están aquellos que se dan por la relación capital-activo. Mientras que, en contraposición, para economías emergentes como los países de América Latina y Asia del Este, estos indicadores tienen un pobre comportamiento y pueden no ser efectivos a la hora de reflejar su condición financiera. Esto se debe a posibles problemas contables, lo cual abre un análisis y una pregunta relevante para futuros estudios.

Indicador	Referencia
OCIF / Activos	I6
Depósitos / (Pasivos + PN neto de resultados)	I8
(Capital social + Reserva legal) / (Pasivo + PN)	I10
Préstamos Moneda Extranjera / Préstamos Brutos	I14
Disponibilidades / Depósitos	I17
(Disponibilidades + Títulos Públicos y privados) / Activo	I19
(Préstamos + Títulos públicos y privados) / (Depósitos + OOIF)	I20
Gastos en personal / Gastos de Administración	I22
Gastos de Administración / (Gastos de Administración + Egreso financieros)	I24

Cuadro 2: Indicadores financieros trabajo de Fernández-Sainz and Llaugel (2011)

El trabajo de Fernández-Sainz and Llaugel (2011) utiliza los datos que provienen de los estados financieros mensuales de 14 bancos comerciales de Re-

publica Dominicana en el periodo noviembre 2002 - diciembre 2004. A partir de dicha información, se elaboraron indicadores que fueron utilizados y puestos a competir por dos modelos: el *logit* y SVM. Los resultados del trabajo demuestran que el modelo *logit* no es capaz de identificar apropiadamente los bancos a punto de fallar en el período previo a la crisis, mientras el modelo SVM supera la regresión logística tanto en el período previo a la crisis como posterior a esta. Resalta asimismo que los modelos basados en SVM pueden ser reentrenados cuando se detectan nuevos patrones en el comportamiento de los indicadores bancarios. En este trabajo se encuentra, en los indicadores que se describen en el Cuadro 2, un adecuado nivel de predicción.

Finalmente se destaca el trabajo de Trujillo Fernández (2017), quien realiza un análisis con la técnica de *machine learning* o aprendizaje automático. Esta técnica es una disciplina cuyo objetivo reside en que los modelos aprendan de manera automática buscando identificar patrones complejos dentro de una gran cantidad de datos obtenidos mediante la experiencia o instrucciones predefinidas.Trujillo Fernández (2017) llega a la conclusión de que las técnicas de *machine learning* pueden aportar soluciones más eficientes a la estimación del *credit scoring*.

En este trabajo, el punto de partida es el armado de bases de datos con información contable de las entidades bancarias argentinas. Sobre dichas bases, y una vez definido el universo de indicadores, se busca replicar el grado de predicción de los indicadores definidos como relevantes en el trabajo de Fernández-Sainz and Llaugel (2011). Luego se construirá un conjunto de indicadores con metodología propia. A partir de estos, con la aplicación de metodología de aprendizaje automático supervisado, se buscan los que mejor se ajusten con la predicción de *default* de las entidades bancarias.

3. Construcción de las bases de datos

3.1. Origen de datos y preprocesamiento

En función del objetivo, y en el marco de los bancos del sistema bancario argentino, se recurrió a las bases de datos del BCRA desde el 01 de enero de 2001 hasta el 31 de diciembre de 2019. Del proceso de exploración de datos se destacan criterios de información divergentes a lo largo del periodo 2001 a la fecha en su modo de publicación. La información previa al 01 de enero de 2003 se encuentra disponible mediante soporte en Compact Disc suministrado por parte de la Biblioteca Prebish[4] del BCRA de manera mensual. La información desde 01 de enero de 2003 y hasta el 31 de diciembre 2010 se encuentra disponible en la pagina web[5], con información sobre el detalle mensual publicada al 31 de diciembre de cada uno de los años. Finalmente, desde 01 de enero de 2011 a la fecha, la información publicada en el BCRA se encuentra detallada y publicada mensualmente. Esta diferencia de criterios en la calidad de la información lleva a una división del trabajo del periodo 01 de enero de 2001 al 31 de diciembre de 2019 en tres ventanas temporales de estudio.

3.2. Definiciones adoptadas

De acuerdo con los criterios divergentes de información, el trabajo se divide en tres ventanas de análisis. La primera ventana temporal analiza la información contable referida a los bancos en *default* por el periodo 01 de enero de 2002 al

[4]La Biblioteca Prebisch se especializa en economía, y se dedica preferente atención a los temas monetarios y financieros. Sus fondos bibliográficos cubren, además, áreas de historia económica, estadística, finanzas, derecho bancario y demás materias relacionadas.

[5]En la pagina del BCRA se obtiene la información en la dirección `http://www.bcra.gov.ar/PublicacionesEstadisticas/Entidades_financieras.asp` Dentro de esta dirección encontramos información actualizada que el BCRA publica mensualmente un archivo rar, donde se encuentra, dentro de la carpeta *Info_Hist*, un archivo denominado *Bajas.txt*, que posee la integridad de las entidades cuya autorización fue dada de baja. Por otro lado dentro de la carpeta *Entfin*, sub carpeta *Tec_cont* encontramos el archivo *imput.txt*, el cual posee balances detallados de los últimos 20 años de todas las entidades financieras autorizadas a funcionar al momento de cada una de las publicaciones.

31 de diciembre de 2003. La segunda ventana temporal analiza la información contable de los bancos en *default* por el periodo 01 de enero de 2007 al 31 de diciembre de 2010. La tercera ventana temporal analiza la información contable de los bancos en *default* por el periodo 01 de enero de 2014 al 31 de diciembre de 2019.

Dentro de las características de la información, encontramos como criterio que toda información contable histórica de bancos, cuya autorización para funcionar fuera rechazada, es eliminada de los registros publicados por el BCRA con posterioridad a la fecha de *default*. Consecuentemente con el objetivo de confeccionar las bases de datos, se debe recurrir a la información que fuera hecha pública con fecha anterior al retiro de cada una de las autorizaciones correspondientes. Adicionalmente, de la base de información del BCRA, se obtiene que la fecha de cesación de la autorización dista de la última información contable pública. Esto genera que se defina como momento T-00 a aquel periodo que se corresponde con el último periodo contable informado y publicado. Tomando como referencia dicha fecha se analiza la información contable en T-00, T-03, donde se refiere a la información contable correspondiente a los tres meses previos a T-00 y T-06 referido a la información contable correspondiente a los 6 meses previos al momento T-00. A fin de determinar la separabilidad de los bancos en *default* y no *default*, se asigna una etiqueta donde los bancos en problemas se identifican con la variable *target* 1, mientras a aquellos sin problemas se les asigna la variable *target* 0.

La primera ventana temporal se compone de los bancos cuya autorización para funcionar es revocada por parte del BCRA entre el 01 de enero de 2002 y el 31 de diciembre de 2003. La última información con la que se cuenta se corresponde a los balances publicados y cerrados al 31 de diciembre de 2001, 30 de septiembre de 2001 y 30 de junio de 2001. En el Cuadro 3 se detallan los bancos detectados como bancos en *default* para la primer ventana. Dentro del Cuadro 3 se expone la fecha en que el BCRA retiró la autorización para operar y la última información contable pública disponible previa a la quita de la autorización que define el momento T-00.

Bancos en *default* 2002 - 2003				
Banco	Fecha suspensión	T-00	T-03	T-06
Banco Banco General de Negocios S.A.	12/04/2002	200112	200109	200106
Banco Scotiabank Quilmes S.A.	18/04/2002	200112	200109	200106
Banco Suquía S.A.	20/05/2002	200112	200109	200106
Banco Biesel S.A.	20/05/2002	200112	200109	200106
Banco de Entre Ríos S.A.	16/09/2002	200112	200109	200106
Banco de la Edificadora de Olavarría S.A.	17/10/2002	200112	200109	200106
Banco Velox	28/10/2002	200112	200109	200106
Banco Municipal de La Plata S.A.	14/08/2003	200112	200109	200106
Banco Bansud S.A.	19/12/2003	200112	200109	200106

Cuadro 3: Bancos en *default* 2002 - 2003

Dentro del preprocesamiento de la información, se detecta la existencia de información de meses que no se corresponden al periodo que se declara como informado. Ante esta situación se depuran los datos a fin de contar con información concerniente al periodo de estudio.

La segunda ventana temporal presenta los bancos cuya autorización para funcionar es revocada por el BCRA entre el 01 de enero de 2007 y el 31 de diciembre de 2010. El Cuadro 4 presenta el detalle de bancos clasificados en *default*, y se detalla la fecha en la que fue revocada la autorización para operar y la última información contable publicada por el BCRA[6].

La tercera ventana temporal comprende los bancos cuya autorización fuera revocada desde 01 de enero de 2014 hasta el 31 de diciembre de 2019. Con posterioridad al 01 de enero de 2010, el BCRA suministra información con mayor periodicidad en su página web, lo que permite que la distancia entre

[6]En este caso, a diferencia de la ventana uno, las fechas en T-00 son dispares entre las diferentes entidades, con lo cual la base de los bancos definidos como sanos se obtienen de cada uno de los períodos T-NN de forma aleatoria y sin reposición.

Bancos en *default* 2007 - 2010				
Banco	Fecha suspensión	T-00	T-03	T-06
Hexagon Bank Argentina S.A.	02/03/2007	200612	200609	200606
Banco Banex S.A.	16/10/2007	200612	200609	200606
Nuevo Banco Suquía S.A.	16/10/2007	200612	200609	200606
Nuevo Banco Bisel S.A.	06/08/2009	200812	200809	200806
Banco Regional de Cuyo S.A.	01/11/2010	200912	200909	200906

Cuadro 4: Bancos en *default* ventana dos

Bancos en *default* 2014 - 2019				
Banco	Fecha suspensión	T-00	T-03	T-06
Banco Privado de Inversiones S.A.	19/03/2014	201311	201308	201305
Banco BC Sociedad Anónima	01/12/2017	201711	201708	201705
Banco Finansur S.A.	09/03/2018	201712	201709	201706
Banco del Tucumán S.A.	15/10/2019	201906	201903	201812

Cuadro 5: Bancos en *default* ventana tres

la fecha de *default* y el momento T-00 se reduzca notoriamente. Esto permite obtener información oportuna[7].

Dentro del pre-procesamiento se tiene en cuenta la existencia de la calidad de información que brindan los bancos de inversión y filiales de bancos del exterior. Dichas entidades exponen divergencias en la información respecto de bancos universales. Atento a estas características se opta por analizar los bancos de carácter universal[8].

Para cada una de las ventanas definidas y cada período T-NN se organiza la información contable relevante en formato de tablas de acuerdo con el plan

[7]En este caso, al igual que en la ventana dos, las fechas en T-00 son dispares entre las diferentes entidades, con lo cual la base de los bancos definidos como sin problemas se obtienen de cada uno de los períodos T-NN de forma aleatoria y sin reposición.

[8]En la ventana tres no fueron tomados para el análisis bancos cuya característica distintiva es corresponderse a filiales del exterior de bancos de inversión y no poseer el carácter de bancos universales dentro del sistema financiero argentino (The Royal Bank of Scotland N.V. y MUFG Bank, Ltd.

de cuentas del BCRA[9]. Posteriormente, con dicha información se arman los índices (ver Cuadro 6 y apéndice A) sobre los cuales se realizan el análisis y la construcción de indicadores relevantes[10].

La estimación de *default* entorno a las bases de datos elaboradas se lleva adelante mediante el criterio de aprendizaje supervisado. Con ese objetivo se realiza un pre-procesamiento de datos, y una posterior modelación y aplicación de las herramientas de evaluación. Para ello se utiliza el lenguaje de programación *Python*, a través de la consola *Jupyter Lab*, con los modulos *Pandas*, *Numpy*, *Sklearn* y *Matplotlib*.

Grupos de indicadores financieros	
Grupo	Indicadores
Estructura de activos	I01, I02, I03, I04, I06, I07, I39, I40, I42.
Estructura de fondeo	I08, I09, I10, I11, I34, I38.
Cartera de préstamos	I12, I13, I14, I16, I31, I32, I33, I35, I36, I37.
Liquidez	I17, I18, I19.
Eficiencia	I21, I22, I24, I27, I29, I30, I41, I43.
Otros	I20, I34a, I3a.

Cuadro 6: Grupo de indicadores financieros

[9] `http://www.bcra.gov.ar/Pdfs/Texord/plandecuentas.pdf`

[10] Ventana tres. Fueron excluidos los bancos The Royal Bank of Scotland N.V. (5), MUFG Bank, Ltd. (18), Banco do Brasil S.A. (46), Bank of America Illinois (262), American Express Bank Ltd. Sociedad Anónima (295), BANCO DEL SOL S.A.(310), BACS Banco de credito y securitizacion S.A.(340), Banco de servicios financieros S.A. (332), RCI Banque S.A. (339). Ventana dos. Fueron excluidos Banco Mercurio S.A. (293), Banco Cofidis S.A. (335),Bank of America Illinois (262), ING Bank N.V. (294). Ventana uno. Fueron excluidos Banco Mercurio S.A. (293), Banco do Estado de Sao Paulo S.A. (236), The Chase Manhattan Bank N.A. (42), Mercobank S.A. (326), Banco Exterior de América S.A. (260), Banco 1784 S.A. (178), Banco de Catamarca (92), Kookmin Bank, Sucursal Argentina (329), Banco San Miguel de Tucumán S.A. (327), Banco Cofidis S.A. (335), Morgan Guaranty Trust Company of New York (165) y Banco Exterior de América S.A. (260)

4. Modelos

El trabajo se lleva adelante haciendo uso de la técnica de aprendizaje automático descripta por Trujillo Fernández (2017). Dentro de esta técnica se encuentran diferentes métodos de trabajo. En nuestro caso, el método que se utiliza se denomina de aprendizaje supervisado. Este método consiste en que, partiendo de una muestra construida por n pares de variables *(X,Y)*, se construye una función con la cual dado un vector de entrada X, se pueda predecir con cierto grado de confianza la variable $Y = f(x)$. En nuestro caso de estudio, se lleva adelante el aprendizaje supervisado de clasificación, el cual es empleado cuando la variable es discreta o categórica. En este caso los bancos en *default* poseen la categoría 1. Sobre la base del aprendizaje supervisado, se aplican cuatro modelos y son puestos a competir respecto de un grado de certeza sobre la variable dependiente mediante 5 herramientas de evaluación de modelos. A continuación se describen brevemente los modelos utilizados.

4.1. Análisis Discriminante Cuadrático (QDA)

El análisis discriminante es una técnica estadística cuya finalidad es analizar si existen diferencias significativas entre k grupos de objetos. En nuestro caso bancos en situación de *default* y situación de no *default*, teniendo en cuenta un conjunto de variables medidas sobre los mismos que expliquen en qué sentido se dan y se faciliten procesos de clasificación sistemática de nuevas observaciones de origen desconocido en uno de los grupos. En el caso bajo estudio se aplica *Quadratic Disciminant Analisys* (QDA), el cual es un tipo mas general de *Linear Discriminant Analisys* (LDA), ya que trabaja bajo los supuestos de multinormalidad de las variables regresoras, sin embargo relaja el supuesto que todas las categorías k tengan la misma estructura de covarianzas. Ante este cambio de supuesto éste método da lugar a fronteras discriminantes cuadráticas.

$$P(X/y = k)\ siendo\ k = 1, 0, \tag{1}$$

En 1, encontramos en el primer término P$(X/y = k)$ lo que representa la probabilidad "a priori" para cada X dado y de pertencer a una clase k.

Sobre la base de lo descripto en 1 las predicciones pueden ser obtenidas a través de la aplicación de la regla de Bayes para cada muestra de entrenamiento $x \in R^n$ donde n está dado por la cantidad de indicadores.

El modelo de Bayes se utiliza para el cálculo de la probabilidad de $Yi, P(Yi|x)$

$$P(Y_i|x) = \frac{P(x|Y_i)P(Y_i)}{P(x)}, \tag{2}$$

donde

$$P(x) = \sum_{i=1}^{N} P(x|Y_i)P(Y_i),$$

de este modo

$$P(Y_i = k|x) = \frac{P(x|Y_i = k)P(Y_i = k)}{P(x)} = \frac{P(x|Y_i = k)P(Y_i = k)}{\sum_l P(x|Y_i = l)P(Y_i = l)}. \tag{3}$$

En el caso bajo estudio, Y se define como los bancos en *default* o no *default*. Esto se logra a través de maximizar la varianza entre los k grupos y minimizar la varianza dentro de los grupos. Para el caso de LDA y QDA asume que las observaciones para cada una de las clases k están determinadas mediante una distribución gausiana multivariada.

$$P(x|y = k) = \frac{1}{(2\pi)^{d/2} \left\|\Sigma_k\right\|} exp\left(-\frac{1}{2}(x - u_k)^t \Sigma^- 1_k (x - u_k)\right). \tag{4}$$

Sin embargo a diferencia de LDA, en la caso de QDA asume que cada clase k posee su matriz de covarianza. Por lo que asume que $X \backsim N(u_k, \Sigma_k)$, donde Σ_k es la matriz de covarianza para cada clase k

4.2. Regresión logística

El modelo de regresión logística, o simplemente *logit*, es utilizado para el caso de clasificación binaria, donde la variable dependiente presenta dos categorías que representan la ocurrencia y la no ocurrencia del acontecimiento definido por la variable, lo cual se realiza una vez que se obtienen los pesos a través del método de máxima verosimilitud (ML). De acuerdo con ML los pesos elegidos son aquellos que maximizan la probabilidad de observar la clasificación buscada, en este caso los bancos en *default*. Una de las ventajas del modelo *logit* se encuentra que no necesita suponer la multinormalidad de las variables regresoras, así como tampoco la igualdad de matrices de covarianzas de los dos grupos. Por lo tanto, necesita menos supuestos que el análisis discriminante. El análisis *logit* incorpora los efectos no lineales, y utiliza la función logística acumulativa para predecir la probabilidad de *default*, siendo

$$P(Y_i = 1) = P_i = \frac{1}{1 + e^{-H_i}} = \frac{1}{1 + e^{-(\beta_1 X_{i1} + \beta_2 X_{i2} \ldots + \beta_n X_{in})}}. \tag{5}$$

donde

$$H_i = \sum_{j=1}^{n} \beta_j X_{ij}.$$

4.3. *Cross-Validation*

Al crear un modelo lo entrena con un conjunto de datos de entrenamiento y con otro conjunto se busca comprobar su evaluación. Sin embargo esta situación no siempre es viable, ya sea por falta de datos, imposibilidad de

obtener un muestreo adecuado o cualquier otro impedimento. Para ello se suele aplicar una estrategia denominada de *Cross-Validation* o Validación Cruzada. Técnica empleada para evaluar el rendimiento de un modelo y garantizar que los resultados son independientes de la partición entre los datos de entrenamiento y testeo.

Este procedimiento consiste en la búsqueda de parámetros del modelo mediante una primera división del universo de datos en datos de entrenamiento y datos de testeo, para luego dividir al conjunto de datos de entrenamiento en k particiones aleatorias llamadas *folds*. Tras esto, y siempre sobre el conjunto de datos de entrenamiento, se utiliza k-1 *folds* para entrenar el modelo y la partición restante (prueba) se emplea en la evaluación del modelo que se entrena. Este procedimiento se repite k veces, cambiando cada vez la partición de prueba. El modelo resultante es validado finalmente con la parte de datos que fuera definida al inicio como datos de testeo y separada de los datos de entrenamiento.

4.4. *Support Vector Machine*

La metodología conocida como *Support Vector Machine* (SVM) es un método no paramétrico que consiste en modelos matemáticos de optimización, desarrollados para implementar clasificación de patrones. El clasificador adquiere la facultad de distinguir entre diferentes patrones a través del desarrollo de funciones de decisión. De esta forma para entrenar al clasificador se debe maximizar el desempeño usando datos de entrenamiento. En el caso de estudio donde el problema es de dos clases, un SVM se entrena de manera que la función de decisión maximice la habilidad de generalización, buscando el mejor conjunto de parámetros que permita separar los bancos solventes de los bancos en problemas. La habilidad de generalización depende fuertemente de la localización del hiperplano de separación, donde se usa la distancia euclidiana entre un dato de entrenamiento X y el hiperplano, buscando maximizar la separación entre grupos.

5. Herramientas de evaluación

A fin de analizar el rendimiento de los distintos modelos descritos en el punto anterior, se lleva adelante la comparación de diferentes métricas. En este trabajo se utilizaron métricas que se focalizan en el análisis de problemas de clasificación binaria, donde son evaluados solo los casos determinados como 1 (en *default*). Las métricas utilizadas en el caso de estudio son *accuracy*, *precision*, *recall*, AUC-Curva ROC, la matriz de confusión y *F1-Score*.

5.1. *Accuracy*

La métrica *accuracy* mide el porcentaje de observaciones en la que el modelo ha acertado. Es la métrica más básica de desempeño de los modelos, pero no es la más eficiente, motivado en que el resultado de esta métrica es la proporción de resultados verdaderos positivos y negativos en el número total de casos.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{6}$$

donde:

TP= *True Positives* (Verdaderos positivos),

TN= *True Negatives* (Verdaderos Negativos),

FP= *False Positives* (Falsos Positivos),

FN= *False Negatives* (Falsos Negativos).

En el caso de tener clases desbalanceadas, si el modelo predice siempre la clase mayoritaria, siempre tendría un excelente nivel de *accuracy*. Esta situación ocurre en el nivel de un análisis de datos descriptivos y puede confundir al observador, por eso es necesario analizar el nivel de *accuracy* de la mano de otros indicadores.

5.2. *Precision* y *Recall*

Precision mide la habilidad del clasificador de no definir como positivo una muestra que es negativa, mientras *recall* mide la habilidad del modelo para clasificar la mayor cantidad de positivos de la muestra analizada.

$$Precision = \frac{TP}{TP + FP}. \tag{7}$$

$$Recall = \frac{TP}{TP + FN}. \tag{8}$$

5.3. Curva ROC-AUC

El área bajo la Curva Característica Operativa del Receptor (ROC por sus siglas en ingles) expone resultados entre 0.5 y 1. Otorga una medida sobre la habilidad de discriminación del modelo sobre los dos resultados bajo análisis, donde expone la probabilidad de detección de verdaderos positivos (VP) y falsos positivos (FP) para la integridad de los posibles resultados analizados.

donde:

ROC = 0.5 Esto sugiere no tener poder de discriminación,
0.5 <ROC <0.7 Se considera un nivel acotado de discriminación,
0.7 <ROC <0.8 Se considera un nivel aceptable de discriminación,
0.8 <ROC <0.9 Se considera un nivel excelente de discriminación,
ROC >0.9 Se considera un nivel sobresaliente de discriminación.

5.4. *F1-score*

La medida *F1-Score* se define como una media armónica de *precision* y *recall*. Esta métrica tiene la capacidad de tener en cuenta tanto los falsos positivos

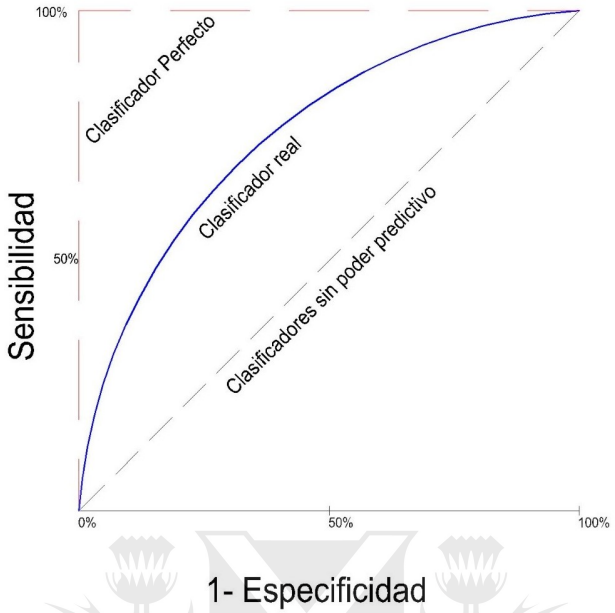

Figura 1: Curva ROC.

como los falsos negativos. Así, un *F1-Score* de 1 indicará una *precision* y un *recall* perfectos. Por lo tanto, un clasificador exacto al momento de discriminar las clases. Esta métrica resulta realmente útil en los casos donde existen clases distribuidas de forma desigual, para la aplicación del método de clasificación sobre los diferentes modelos descritos, como ocurre en el caso de este trabajo.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}. \tag{9}$$

El método de *F1-score* otorga una métrica por cada tipo de resultado, en este caso tendremos entonces, un F1-score por los 0 y un *F1-Score* por los 1. Dentro de esta medida de bondad de ajuste existe el *F1-score weight*, el cual busca exponer un resultado teniendo en cuenta el peso ponderado de cada una de las clases reales sobre la muestra analizada lo que arroja un valor único.

6. Análisis empírico

6.1. Asunciones generales

Sobre la base de indicadores conformados y en cada una de las ventanas descritas, como primer punto se realiza un análisis en función de los indicadores determinados como relevantes del trabajo de Fernández-Sainz and Llaugel (2011) (ver Cuadro 2).

A fin de realizar el armado de las diferentes ventanas temporales, se encuentra tanto en la ventana dos como en la ventana tres un fuerte desbalanceo de clases. En estas, los bancos en *default* representan un porcentaje acotado respecto de los bancos en no *default*. Atentos a esta circunstancia, se procede a realizar, en primera medida, un sorteo de entidades-periodo sin problemas de manera aleatoria y, a su vez, aumenta la población de bancos en problemas, con lo cual se procede a duplicar las entidades bancarias en problemas en ambas ventanas a fin de subsanar el fuerte desbalanceo. Esta técnica se denomina *oversampling*.

Una vez definidas las ventanas, y para cada uno de los periodos establecidos T-00, T-03, y T-06, para cada caso se realizó una separación de los datos a ser analizados en un set de datos de entrenamiento y otro de testeo. La relación se determina en 70/30, teniendo en cuenta mantener el balanceo de clases, debido a la disparidad existente. Se define como variable *target* 1 a los bancos en *default*, mientras que 0 corresponde a bancos no *default*.

6.2. Resultados sobre trabajo de Fernández-Sainz and Llaugel (2011)

Sobre las bases de datos se llevó adelante la aplicación de cuatro modelos: uno mediante *Quadratic Disciminant Analisys* (QDA), un segundo por Regresión Logística, un terceros mediante Regresión Logística con Validación Cruzada y

un cuarto modelo mediante SVM. Se utilizaron los indicadores de Fernández-Sainz and Llaugel (2011) sobre el conjunto de datos definido, tomando como parámetro de ajuste el área bajo la curva (AUC) o curva ROC.

Ventana Período	QDA	RL [a]	RLCV [b]	SMV
Ventana 2002 - 2003 T-00	0,50	0,58	0.74	0.80
Ventana 2002 - 2003 T-03	0,50	0,56	0.50	0.50
Ventana 2002 - 2003 T-06	0,50	0,49	0.46	0.50
Ventana 2007 - 2010 T-00	0,50	0,46	0.56	0.50
Ventana 2007 - 2010 T-03	0,50	0,54	0.56	0.50
Ventana 2007 - 2010 T-06	0,50	0,65	0.62	0.48
Ventana 2014 - 2019 T-00	1,00	0,71	0.91	0.50
Ventana 2014 - 2019 T-03	1,00	0,95	0.95	0.75
Ventana 2014 - 2019 T-06	1,00	0,88	0.73	0.75

Cuadro 7: Resultados - AUC con aplicación de indicadores según Fernández-Sainz and Llaugel (2011), ver Figuras 2 a 13

[a] Regresión Logística
[b] Regresión Logística con validación cruzada

El Cuadro 7 expone de manera clara el modelo de Regresión Logística mediante Validación Cruzada como el modelo de mayor consistencia y mejor ajuste en función de las variables determinadas como resultado de la tesis de Fernández-Sainz and Llaugel (2011). Asimismo, el mejor ajuste surge en el periodo T-00 para cada de una de las tres ventanas puestas a estudio. Consecuentemente encontramos que, a medida que la información analizada toma más distancia del momento del *default*, hay un menor ajuste en el rendimiento de todos los modelos.

7. Desarrollo de metodología propia

7.1. Preselección de indicadores propios

A partir de los resultados descritos para los indicadores de Fernández-Sainz and Llaugel (2011), y en función de los definidos en apéndice A, se procede

a realizar la construcción de indicadores propios para cada una de las ventanas definidas y períodos T-NN sobre la base de modelos de aprendizaje supervisado.

El procedimiento aplicado se basa en llevar adelante una metodología de búsqueda de los parámetros que mejor ajusten tomando como medida de bondad de ajuste el área bajo la curva definida por *Receiver Operating Characteristic* (ROC)[11]. Esto nos permite tener una adecuada medida de ajuste en función de la positividad del estudio. La metodología de búsqueda de parámetros comienza con la aplicación de un modelo de selección de variables denominado *Recursive feature elimination cross-validation (RFECV)*. Este es un estimador de parámetros recursivo con validación cruzada estratificada a fin de mantener el balance de clases que utiliza como medida de bondad de ajuste al AUC. Su funcionamiento radica en eliminar las variables no significativas considerando un menor conjunto de parámetros de entrenamiento, hasta obtener el conjunto de valores que arroje la mejor medida de ajuste medido a través de una función de costo (AUC).

Posteriormente, y a modo de competencia, se aplica el modelo denominado *Sequential Feature Selector* (SFS), comunmente denominado *stepwise*. Este es un modelo de aprendizaje supervisado que opera seleccionando una variable por vez y midiendo su valor predictivo. Genera una cantidad de variables predeterminada en función de su poder predictivo. En función a la cantidad de indicadores determinados como óptimos en el sistema RFECV, y a fin de generar una comparabilidad de los indicadores relevantes, se genera una competencia en los indicadores resultantes de cada modelo de selección.

[11] El análisis de la curva ROC proporciona herramientas para seleccionar los modelos posiblemente óptimos y descartar modelos subóptimos independientemente del coste de la distribución de las dos clases sobre las que se decide. La curva ROC es también independiente de la distribución de las clases en la población. El análisis ROC se relaciona de forma directa y natural con el análisis de coste/beneficio en toma de decisiones diagnósticas.

7.2. Desarrollo de modelos mediante métodos de *Machine Learning*

Como paso siguiente se lleva adelante la aplicación de los cuatro modelos definidos: *Quadratic Disciminant Analisys* (QDA), Regresión Logística, Regresión Logística con validación cruzada y SVM, sobre los parámetros encontrados. De esta forma, para cada una de las ventanas y cada uno de los periodos T-NN se obtuvieron los resultados descritos en el Cuadro 8.

Ventana Período	QDA	RL [a]	RLCV [b]	SMV
Ventana 2002 - 2003 T-00	0,50	0,74	0.74	0.50
Ventana 2002 - 2003 T-03	0,50	0,72	0.64	0.50
Ventana 2002 - 2003 T-06	0,50	0,40	0.46	0.50
Ventana 2007 - 2010 T-00	0,67	0,90	0.94	0.50
Ventana 2007 - 2010 T-03	0,50	0,56	0.92	0.50
Ventana 2007 - 2010 T-06	0,50	0,54	0.62	0.50
Ventana 2014 - 2019 T-00	1,00	0,93	0.95	0.50
Ventana 2014 - 2019 T-03	1,00	1,00	0.96	0.75
Ventana 2014 - 2019 T-06	1,00	0,98	0.95	0.96

Cuadro 8: Resultados - AUC indicadores propios

[a] Regresión Logística
[b] Regresión Logística con validación cruzada

De acuerdo con los resultados obtenidos, al igual que el análisis con indicadores de Fernández-Sainz and Llaugel (2011), el rendimiento del modelo basado en Regresión Logística, mediante el sistema de validación cruzada, con estratificación y balance de clases, surge como el mejor modelo a aplicar. Dado que el objeto es saber qué indicadores permiten una separabilidad de los bancos en *default* de aquellos no *default*, se procedió a aplicar un tercer paso en la metodología. Este último paso consiste en la búsqueda de los parámetros relevantes según modelo de Regresión Logística con Validación Cruzada donde se aplican nuevamente los modelos con los indicadores determinados como materiales para cada ventana. Los resultados se demuestran en Cuadro 9.

Ventana Período	QDA	RL [a]	RLCV [b]	SMV
Ventana 2002 - 2003 T-00	0,77	0,65	0.74	0.50
Ventana 2002 - 2003 T-03	0,75	0,64	0.64	0.72
Ventana 2002 - 2003 T-06	0,54	0,37	0.43	0.50
Ventana 2007 - 2010 T-00	0,50	0,81	0.92	0.46
Ventana 2007 - 2010 T-03	0,50	0,71	0.94	0.65
Ventana 2007 - 2010 T-06	0,50	0,83	0.62	0.67
Ventana 2014 - 2019 T-00	0,75	0,98	0.96	0.50
Ventana 2014 - 2019 T-03	1,00	1,00	0.96	0.50
Ventana 2014 - 2019 T-06	1,00	0,95	0.95	0.98

Cuadro 9: Resultados - AUC con indicadores acotados

[a]Regresión Logística
[b]Regresión Logística con validación cruzada

Si bien dentro del análisis se toma en cuenta el período comprendido entre 2002 al 2019, dados los cambios de contexto macro, en suma a la necesidad de búsqueda de parámetros que sean aplicables al momento histórico actual, se pone foco en la ventana tres (periodo 2014 -2019) para el análisis de indicadores y modelos de *machine learning*. Sobre la base de los indicadores definidos en el apéndice A el trabajo concluye para la ventana de análisis 2014-2019 un conjunto de indicadores consistentes para los tres períodos T, tal como se describe del Cuadro 10.

Ventana - Período	Indicadores significativos
Ventana 2002 - 2003 - T-00	I1, I17, I36, I39
Ventana 2002 - 2003 - T-03	I8, I9, I13, I33, I34a
Ventana 2002 - 2003 - T-06	I11, I27, I31, I34a
Ventana 2007 - 2010 - T-00	I4, I8, I10, I12, I22, I36
Ventana 2007 - 2010 - T-03	I4, I6, I9, I12, I22, I36
Ventana 2007 - 2010 - T-06	I4, I6, I9, I27, I36, I39
Ventana 2014 - 2019 - T-00	I1, I2, I3, I9, I11, I29, I32, I33, I40
Ventana 2014 - 2019 - T-03	I1, I2, I3, I9, I11, I29, I32, I33, I40
Ventana 2014 - 2019 - T-06	I1, I2, I3, I9, I11, I29, I32, I33, I40

Cuadro 10: Definición final de indicadores financieros

Los indicadores mencionados se aplican para cada período T-NN según los cuatro modelos definidos. En el caso particular de la ventana tres (2014-2019) los coeficientes de los indicadores mencionados, encontramos de la composición del financiamiento (I11), signo positivo de los coeficientes para el caso de los bancos en situación de *default*, atribuible a una fuerte pérdida de fondeo estable de terceros mediante el retiro de depósitos. A su vez, esto nos lleva a observar una relación positiva en el coeficiente relativo a fondeo propio respecto de activo (I40), donde un escenario de exiguo apalancamiento refleja una situación de estrés en la toma de financiamiento estable. En suma, presenta signo negativo el coeficiente relacionado al fondeo de corto plazo en OOIF (I9).

En lo que respecta al retorno sobre fondeo propio (I29), presenta relación inversa, donde la mayor parte de los bancos en *default* expusieron utilidad neta positiva en los períodos previos a la declaración de *default*. Los indicadores correspondientes a préstamos (I2 y I32) en su relación con el activo presentan en los coeficientes signo positivo, mientras que la relación de préstamos comercial (I33) presenta signo contrario, donde en las entidades en *default* se destaca una participación marginal de la cartera comercial sobre la integridad de préstamos otorgados. El coeficiente relativo a títulos (I3) sobre activo presenta un signo positivo. Finalmente, en el caso de disponibilidades (I1) la relación del coeficiente para los modelos en T-00 y T-06 resulta de signo positivo, mientras que en T-03 presenta relación negativa.

8. Conclusiones

Como resultado del armado de las bases de datos, y de la aplicación del método de aprendizaje supervisado sobre los indicadores construidos, observamos en primer lugar resultados dispares en la replicación del trabajo de Fernández-Sainz and Llaugel (2011). No obstante, es significativo el adecuado nivel predictivo para la ventana temporal 2014 - 2019. Esto nos permite validar la aplicabilidad de dicho trabajo en el contexto actual de Argentina.

En suma a lo mencionado, y del trabajo que se realiza mediante metodología propia a fin de arribar al conjunto de indicadores para cada una de las ventanas, resaltamos la disparidad de indicadores relevantes para cada ventana temporal. Entendemos que esta situación se da ante los cambios en los contextos macroeconómicos de la Argentina. Adicionalmente a un conjunto de cambios regulatorios constantes determinados por la entidad de aplicación, conjuga un escenario cambiante de indicadores relevantes.

Finalmente se destaca el modelo de Regresión Logística con Validación Cruzada como el modelo de mayor estabilidad y efectividad. Mediante este modelo, para el caso de la ventana temporal 2014 - 2019, encontramos una relación de los indicadores que permiten analizar la conformación de la estructura patrimonial y de sus activos, así como los niveles de rentabilidad y la composición de la cartera de préstamos en relación con los créditos de carácter comercial, que generan un nivel predictivo sobresaliente a la hora de determinar la separabilidad de los bancos en *default* y en no *default* (ver tabla E). A lo mencionado se suma que, dado el contexto en el que fueron analizados (2014 - 2019), entendemos que presentan una adecuada aplicabilidad a fin de realizar un seguimiento actual por parte de la entidad regulatoria a aquellos bancos que demuestren un comportamiento desfavorable en dichos indicadores.

9. Pasos siguientes del trabajo

Entendemos que el trabajo realizado, al ser llevado adelante exclusivamente con información del sistema bancario argentino, presenta particularidades propias de su sistema regulatorio. Consecuentemente, los modelos obtenidos, así como los indicadores relevantes, deben ser actualizados en función de la existencia de nuevos bancos que fueran declarados en *default* en el futuro a fin de poder aplicar el modelo a nuevos escenarios del sistema bancario argentino.

Referencias

Edward I Altman. Predicting performance in the savings and loan association industry. *Journal of Monetary Economics*, 3(4):443–466, 1977.

António Antunes, Diana Bonfim, Nuno Monteiro, and Paulo MM Rodrigues. Early warning indicators of banking crises: exploring new data and tools1. *Economic Bulletin*, page 90, 2014.

BCRA. Capitales mínimos de las entidades financieras, a. URL `http://www.bcra.gov.ar/Pdfs/Texord/t-capmin.pdf`.

BCRA. Lineamientos para la gestión de riesgos en las entidades financieras, b. URL `http://www.bcra.gov.ar/Pdfs/Texord/t-lingeef.pdf`.

Sigbjørn Atle Berg and Barbro Hexeberg. *Early warning indicators for Norwegian banks: A logit analysis of the experiences from the banking crisis*. Number 1994/1. Arbeidsnotat, 1994.

Joel Bessis. *Risk management in banking*. John Wiley & Sons, 2011.

BIS. *Basilea III: marco regulador internacional para los bancos*.

F Buchieri. Crisis bancarias recientes en argentina: Un modelo teórico y evidencia emp'ırica asociada. *Revista Ciencias Económicas. Facultad de Ciencias Económicas, UNL*, 7, 2009.

Ana Fernández-Sainz and Felipe Llaugel. ¿Bancos con problemas? Un Sistema de Alerta Temprana para la Prevención de Crisis Bancarias. Technical report, Instituto de Economía Aplicada a la Empresa de la Universidad del País Vasco, 2011.

Zan Huang, Hsinchun Chen, Chia-Jung Hsu, Wun-Hwa Chen, and Soushan Wu. Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision support systems*, 37(4):543–558, 2004.

Felipe Klein. Estimación de la probabilidad de default: un modelo probit para los bancos argentinos. *Ensayos de Política Económica*, 2(2):88–115, 2019.

Kin Keung Lai, Lean Yu, Shouyang Wang, and Ligang Zhou. Credit risk analysis using a reliability-based neural network ensemble model. In *International Conference on Artificial Neural Networks*, pages 682–690. Springer, 2006.

Daniel Martin. Early warning of bank failure: A logit regression approach. *Journal of Banking & Finance*, 1(3):249–276, 1977.

David A Mermelstein. Hacia un indicador de vulnerabilidad bancaria basado en pruebas de estrés. Technical report, Serie Documentos de Trabajo, 2017.

Jae H Min and Young-Chan Lee. Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert systems with applications*, 28(4):603–614, 2005.

Alfredo B. Roisenzvit. Understanding and regulating banks: A new paradigm. 2011.

Kyung-Shik Shin, Taik Soo Lee, and Hyun-jung Kim. An application of support vector machines in bankruptcy prediction model. *Expert systems with applications*, 28(1):127–135, 2005.

Daniel K Tarullo et al. The evolution of capital regulation: a speech at the clearing house business meeting and conference, new york, new york, nov. 9, 2011. Technical report, 2011.

David Trujillo Fernández. Aplicación de metodologías machine learning en la gestión de riesgo de crédito. 2017.

Dennis G Uyemura and Donald R Van Deventer. *Financial risk management in banking: the theory & application of asset & liability management*. Bankers Publishing Company, 1993.

Vladimir Vapnik. The support vector method of function estimation. In *Nonlinear modeling*, pages 55–85. Springer, 1998.

David West. Neural network credit scoring models. *Computers & Operations Research*, 27(11-12):1131–1152, 2000.

A. Indicadores financieros

Indicador	Referencia
Disponibilidades / Activo	I1
Prestamos Netos / Activo	I2
Títulos Públicos y privados / Activo	I3
Activos inmovilizados / Activo	I4
OCIF / Activos	I6
Prestamos Netos + Tít. Públicos y privados / Activo	I7
Depósitos / (Pasivos + PN neto de resultados)	I8
OOIF / Pasivo	I9
(Capital social + Reserva legal) / (Pasivo + PN neto de resultados)	I10
PN neto de resultados / (Pasivo + PN neto de resultados)	I11
Prestamos Netos / Prestamos Brutos	I12
Prestamos Moneda Local / Prestamos Brutos	I13
Prestamos Moneda Extranjera / Prestamos Brutos	I14
Previsiones / Prestamos Brutos	I16
Disponibilidades / Depósitos	I17
Disponibilidades / (Depósitos + OOIF)	I18
(Disponibilidades + Tít. públicos y privados) / Activo	I19
(Préstamos + Tit públicos y privados) / (Depósitos + OOIF)	I20
Gastos Administrativos / Depósitos	I21
Gastos en personal / Gastos de Administración	I22
Gastos de Adm. / (Gastos de Administración + Egreso financieros)	I24
Egreso financiero / (Depósitos a plazo en moneda local + Depósitos a plazo en moneda extranjera)	I27
Utilidad neta / PN	I29
Gastos de Administración / Ingresos financieros	I30
Prestamos sector público / Prestamos Brutos	I31
Prestamos sector privado / Prestamos Brutos	I32
Prestamos comerciales / Prestamos Brutos	I33
Capital social / Activo	I34
Depósitos a plazo en moneda local / Depósitos en moneda local	I35
Depósitos a plazo en ME / Depósitos en moneda ME	I36
Prestamos Brutos / Depósitos	I37
Pasivo / PN neto de resultado	I38
Activo / Pasivo	I39
PN neto de resultados / Activo	I40
Utilidad neta / Activo	I41
Activos inmovilizados / PN	I42
Cargo por incobrabilidad / Resultado Operativo	I43
RPC / Activo	I34
Títulos públicos y privados / RPC	I3a

B. Código en Python

El codigo conjuntamente con las bases de datos se encuentra en el repositorio `https://github.com/mbozzi80/Tesis.git`

C. Gráficos

C.1. Gráficos análisis indicadores Fernández-Sainz and Llaugel (2011)

C.1.1. Ventana 3 en T-00

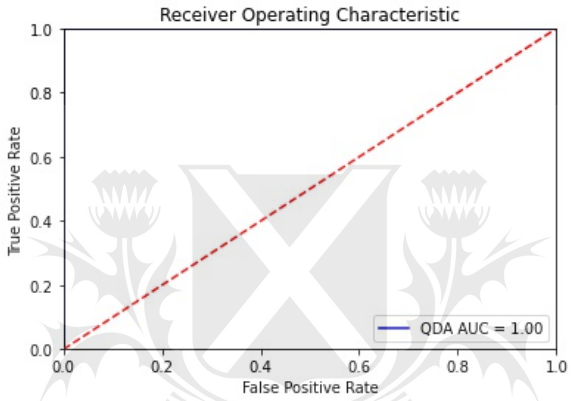

Figura 2: AUC bajo modelo Discriminante Lineal Cuadrático

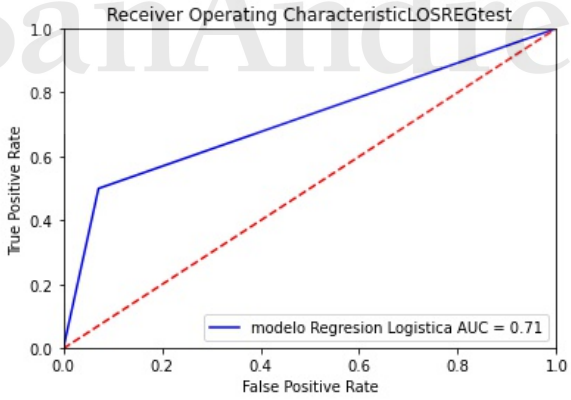

Figura 3: AUC bajo modelo Regresión Logística

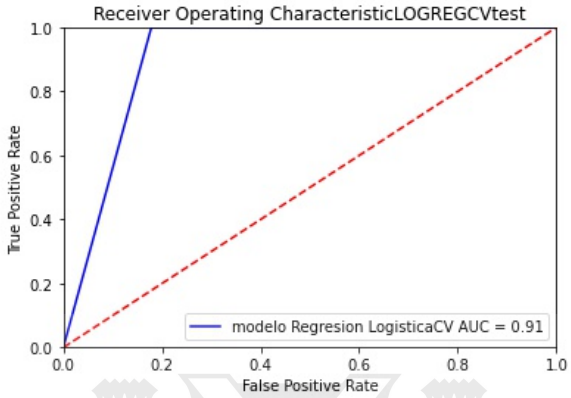

Figura 4: AUC bajo modelo Regresión Logística con Validación Cruzada

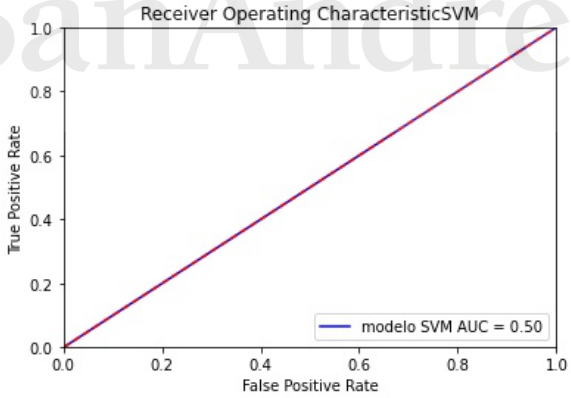

Figura 5: AUC bajo modelo SVM

C.1.2. Ventana 3 en T-03

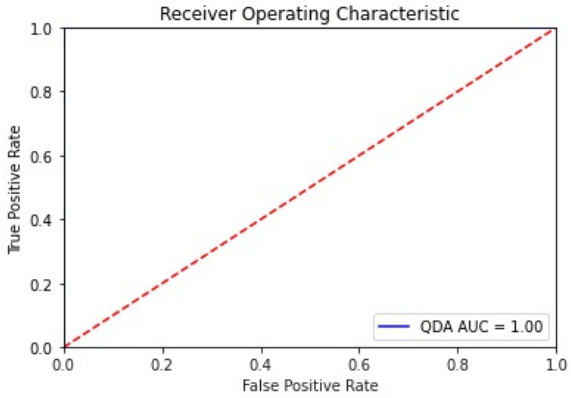

Figura 6: AUC bajo modelo Discriminante Lineal Cuadrático

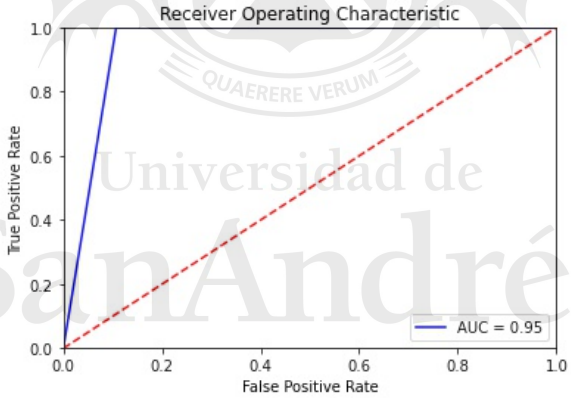

Figura 7: AUC bajo modelo Regresión Logística

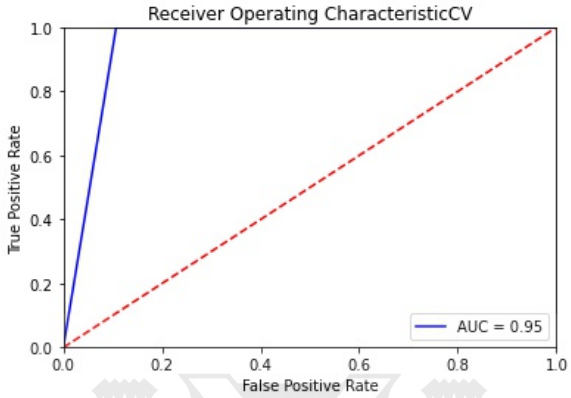

Figura 8: AUC bajo modelo Regresión Logística con Validación Cruzada

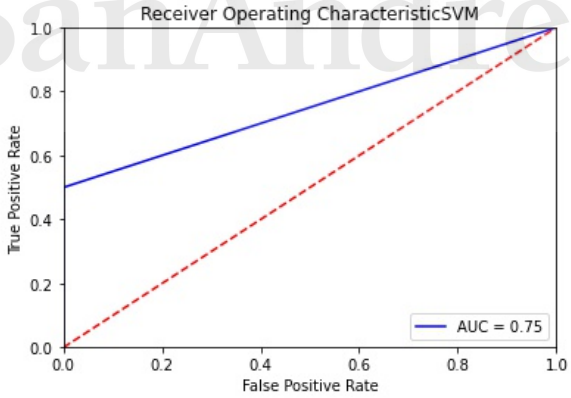

Figura 9: AUC bajo modelo SVM

C.1.3. Ventana 3 en T-06

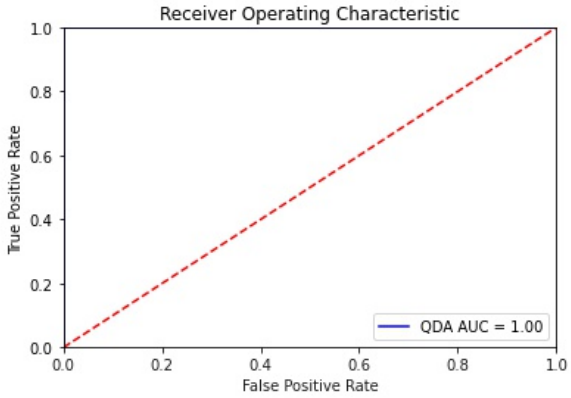

Figura 10: AUC bajo modelo Discriminante Lineal Cuadrático

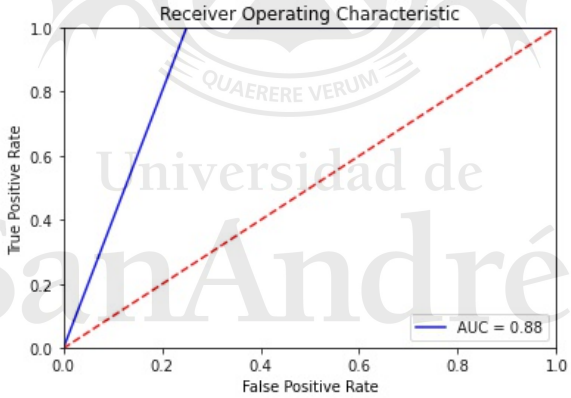

Figura 11: AUC bajo modelo Regresión Logística

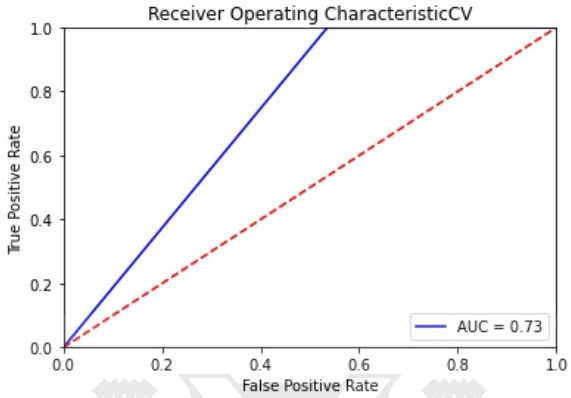

Figura 12: AUC bajo modelo Regresión Logística con Validación Cruzada

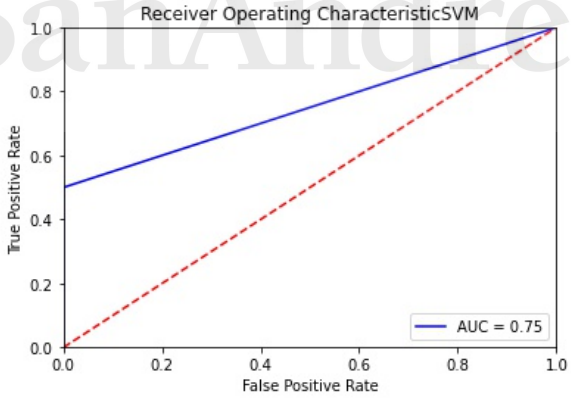

Figura 13: AUC bajo modelo SVM

C.2. Gráficos análisis indicadores propios

C.2.1. Ventana 3 en T-00

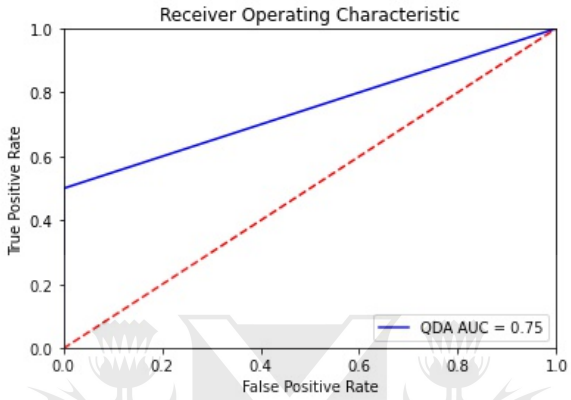

Figura 14: AUC bajo modelo Discriminante Lineal Cuadrático

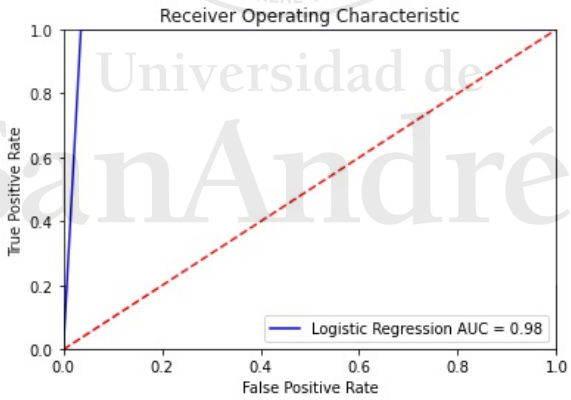

Figura 15: AUC bajo modelo Regresión Logística

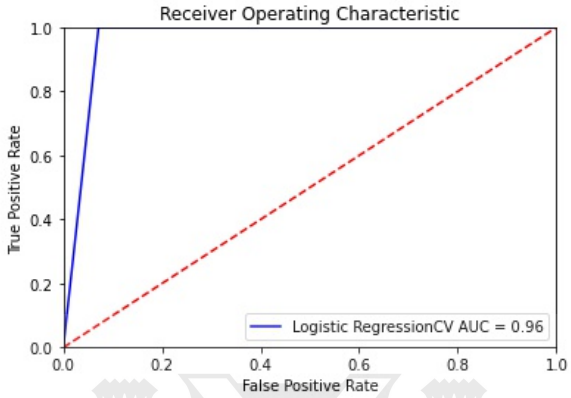

Figura 16: AUC bajo modelo Regresión Logística con Validación Cruzada

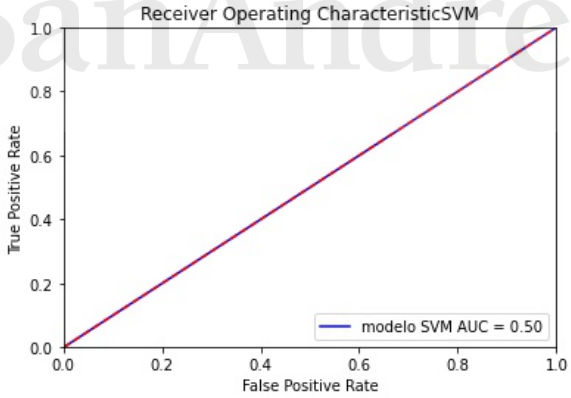

Figura 17: AUC bajo modelo SVM

C.2.2. Ventana 3 en T-03

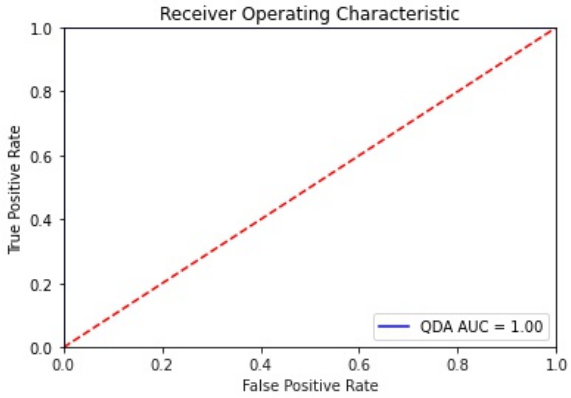

Figura 18: AUC bajo modelo Discriminante Lineal Cuadrático

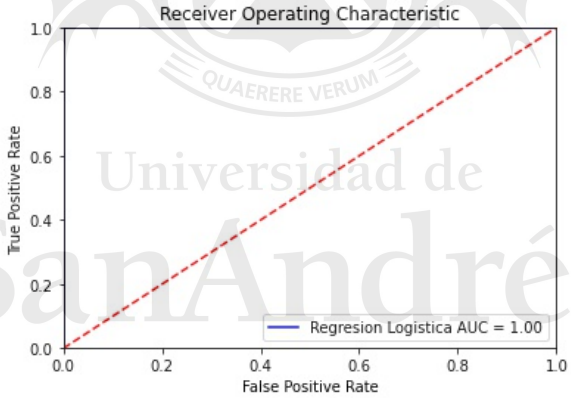

Figura 19: AUC bajo modelo Regresión Logística

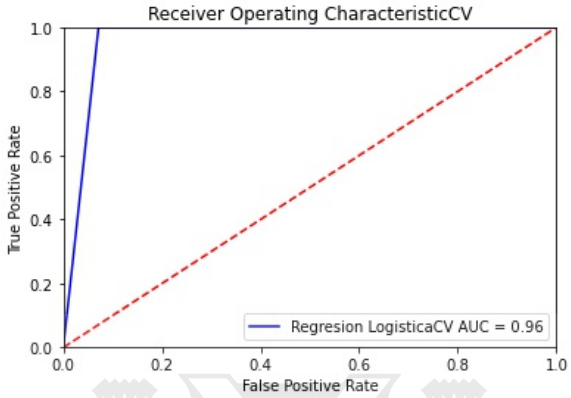

Figura 20: AUC bajo modelo Regresión Logística con Validación Cruzada

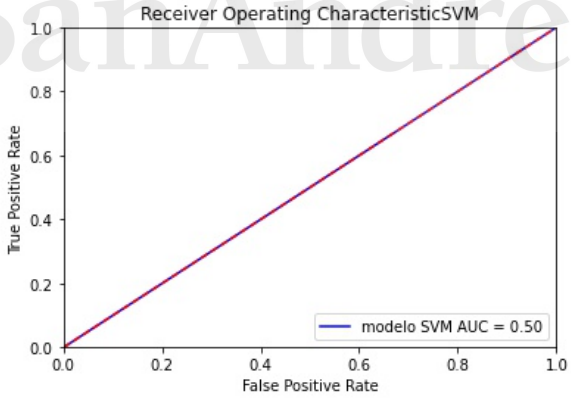

Figura 21: AUC bajo modelo SVM

C.2.3. Ventana 3 en T-06

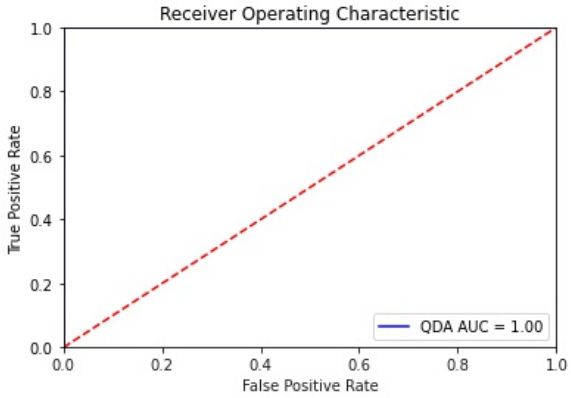

Figura 22: AUC bajo modelo Discriminante Lineal Cuadrático

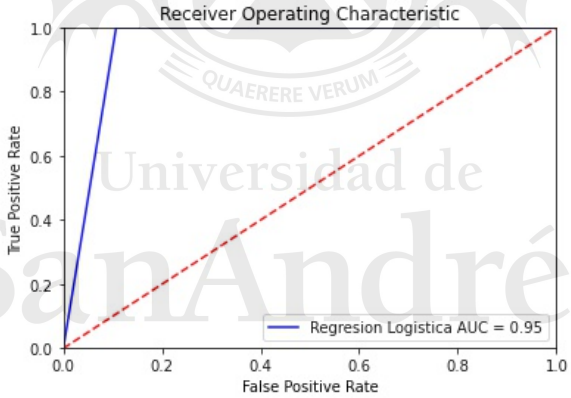

Figura 23: AUC bajo modelo Regresión Logística

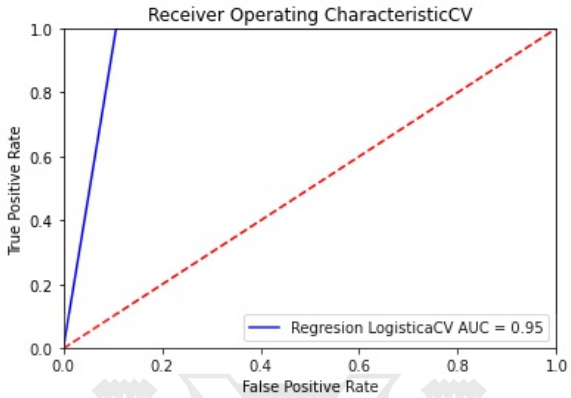

Figura 24: AUC bajo modelo Regresión Logística con VC

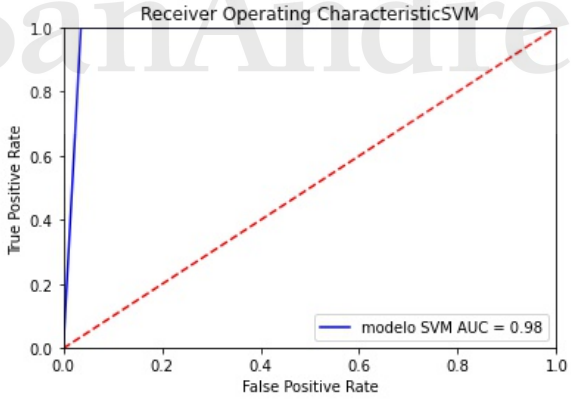

Figura 25: AUC bajo modelo SVM

D. Tabla de resultados indicadores Fernández-Sainz and Llaugel (2011)

	Accuracy	*Precisión*	*Recall*	*AUC-ROC*	*F1score*	*F1score-w*
Ventana 1- T00 - QDA	0.84	0.00	0.00	0.50	0.00	0.77
Ventana 1- T00 - LOGREG	0.53	0.20	0.67	**0.58**	0.31	0.59
Ventana 1- T00 - LOGREGCV	0.79	0.40	0.67	**0.74**	**0.50**	0.81
Ventana 1- T00 - SVM	0.89	0.67	0.67	**0.80**	0.67	0.89
Ventana 1- T03 - QDA	0.90	0.00	0.00	0.50	0.00	0.85
Ventana 1- T03 - LOGREG	0.60	0.12	0.50	0.56	0.20	0.68
Ventana 1- T03 - LOGREGCV	0.90	0.00	0.00	0.50	0.00	0.85
Ventana 1- T03 - SVM	0.90	0.00	0.00	0.50	0.00	0.85
Ventana 1- T06 - QDA	0.90	0.00	0.00	0.50	0.00	0.85
Ventana 1- T06 - LOGREG	0.47	0.10	0.50	0.49	0.17	0.57
Ventana 1- T06 - LOGREGCV	0.42	0.10	0.50	0.46	0.15	0.52
Ventana 1- T06 - SVM	0.89	0.00	0.00	0.50	0.00	0.85
Ventana 2- T00 - QDA	0.89	0.00	0.00	0.50	0.00	0.84
Ventana 2- T00 - LOGREG	0.30	0.10	0.67	0.46	0.17	0.36
Ventana 2- T00 - LOGREGCV	0.74	0.17	0.33	0.56	0.22	0.78
Ventana 2- T00 - SVM	0.89	0.00	0.00	0.50	0.00	0.84
Ventana 2- T03 - QDA	0.89	0.00	0.00	0.50	0.00	0.84
Ventana 2- T03 - LOGREG	0.44	0.12	0.67	0.54	0.21	0.53
Ventana 2- T03 - LOGREGCV	0.74	0.17	0.33	0.56	0.22	0.78
Ventana 2- T03 - SVM	0.89	0.00	0.00	0.50	0.00	0.84
Ventana 2- T06 - QDA	0.89	0.00	0.00	0.50	0.00	0.84
Ventana 2- T06 - LOGREG	0.63	0.18	0.67	0.65	0.29	0.70
Ventana 2- T06 - LOGREGCV	0.85	0.33	0.33	0.62	0.33	0.85
Ventana 2- T06 - SVM	0.85	0.00	0.00	0.48	0.00	0.82
Ventana 3- T00 - QDA	1.00	1.00	1.00	**1.00**	1.00	1.00
Ventana 3- T00 - LOGREG	0.90	0.33	0.50	**0.71**	0.40	0.91
Ventana 3- T00 - LOGREGCV	0.83	0.29	1.00	**0.91**	**0.44**	0.87
Ventana 3- T00 - SVM	0.93	0.00	0.00	0.50	0.00	0.90
Ventana 3- T03 - QDA	1.00	1.00	1.00	**1.00**	1.00	1.00
Ventana 3- T03 - LOGREG	0.90	**0.40**	**1.00**	**0.95**	0.57	0.92
Ventana 3- T03 - LOGREGCV	0.90	0.40	**1.00**	**0.95**	**0.57**	0.92
Ventana 3- T03 - SVM	0.97	1.00	0.50	0.75	0.67	0.96
Ventana 3- T06 - QDA	1.00	1.00	1.00	**1.00**	1.00	1.00
Ventana 3- T06 - LOGREG	0.76	0.22	1.00	**0.88**	**0.36**	0.82
Ventana 3- T06 - LOGREGCV	0.50	0.12	1.00	**0.73**	**0.21**	0.61
Ventana 3- T06 - SVM	0.97	1.00	0.50	**0.75**	0.67	0.96

E. Tabla de resultados indicadores propios

	Accuracy	*Precision*	*Recall*	*AUC-ROC*	*F1score*	*F1score-w*
Ventana 1- T00 - QDA	0.84	0.50	0.67	**0.77**	0.57	0.85
Ventana 1- T00 - LOGREG	0.63	0.25	0.67	**0.65**	0.36	0.68
Ventana 1- T00 - LOGREGCV	0.79	0.40	0.67	**0.74**	0.50	0.81
Ventana 1- T00 - SVM	0.84	0.00	0.00	0.50	0.00	0.77
Ventana 1- T03 - QDA	0.95	1.00	0.50	**0.75**	0.67	0.94
Ventana 1- T03 - LOGREG	0.75	0.20	0.50	**0.64**	0.33	0.83
Ventana 1- T03 - LOGREGCV	0.75	0.20	0.50	**0.64**	0.50	0.90
Ventana 1- T03 - SVM	0.90	0.50	0.50	**0.72**	0.40	0.86
Ventana 1- T06 - QDA	0.58	0.12	0.50	0.54	0.20	0.66
Ventana 1- T06 - LOGREG	0.26	0.07	0.50	0.37	0.12	0.34
Ventana 1- T06 - LOGREGCV	0.38	0.08	0.00	0.43	0.14	0.46
Ventana 1- T06 - SVM	0.90	0.00	0.00	0.50	0.00	0.85
Ventana 2- T00 - QDA	0.89	0.00	0.00	0.50	0.00	0.84
Ventana 2- T00 - LOGREG	0.67	0.25	1.00	**0.81**	0.40	0.73
Ventana 2- T00 - LOGREGCV	0.85	0.43	1.00	**0.92**	0.60	0.87
Ventana 2- T00 - SVM	0.81	0.00	0.00	0.46	0.00	0.80
Ventana 2- T03 - QDA	0.89	0.00	0.00	0.50	0.00	0.84
Ventana 2- T03 - LOGREG	0.48	0.18	1.00	**0.71**	0.30	0.56
Ventana 2- T03 - LOGREGCV	0.89	0.50	1.00	**0.94**	0.67	0.90
Ventana 2- T03 - SVM	0.89	0.50	0.33	0.65	0.00	0.88
Ventana 2- T06 - QDA	0.89	0.00	0.00	0.50	0.00	0.84
Ventana 2- T06 - LOGREG	0.70	0.27	1.00	**0.83**	0.43	0.76
Ventana 2- T06 - LOGREGCV	0.85	0.33	0.33	**0.62**	0.33	0.85
Ventana 2- T06 - SVM	0.89	0.00	0.00	0.67	0.00	0.84
Ventana 3- T00 - QDA	0.97	1.00	0.50	**0.75**	0.67	0.96
Ventana 3- T00 - LOGREG	0.98	0.67	1.00	**0.98**	0.80	0.97
Ventana 3- T00 - LOGREGCV	0.97	0.50	1.00	**0.96**	0.67	0.94
Ventana 3- T00 - SVM	0.93	0.00	0.00	0.50	0.00	0.90
Ventana 3- T03 - QDA	1.00	1.00	1.00	**1.00**	1.00	1.00
Ventana 3- T03 - LOGREG	1.00	1.00	1.00	**1.00**	1.00	1.00
Ventana 3- T03 - LOGREGCV	0.93	0.50	1.00	**0.96**	0.67	0.94
Ventana 3- T03 - SVM	0.93	0.00	0.00	0.50	0.00	0.90
Ventana 3- T06 - QDA	1.00	1.00	1.00	1.00	1.00	1.00
Ventana 3- T06 - LOGREG	0.90	0.40	1.00	**0.95**	0.57	0.92
Ventana 3- T06 - LOGREGCV	0.90	0.40	1.00	**0.95**	0.57	0.92
Ventana 3- T06 - SVM	0.96	0.67	1.00	**0.98**	0.67	0.94