



**“APLICACIÓN DE ÁRBOLES DE DECISIÓN PARA LA
ESTIMACIÓN DEL ESCENARIO ECONÓMICO Y LA
ESTIMACIÓN DE MOVIMIENTO LA TASA DE
INTERÉS EN CHILE”**

**TESIS PARA OPTAR AL GRADO DE
MAGÍSTER EN FINANZAS**

**Alumno: Carlos Dupouy Berrios
Profesor Guía: David Díaz Solís Ph.D.**

Santiago, Julio 2014

Tabla de Contenidos

1. RESUMEN	1
2. INTRODUCCIÓN	3
3. MARCO TEÓRICO	6
3.1 EMISIÓN Y POLÍTICA MONETARIA EN CHILE.....	6
3.2 INSTRUMENTOS DE DEUDA EN CHILE	9
3.3 METODOLOGÍAS DE ESTIMACIÓN DE UNA CURVA CERO CUPÓN.....	13
3.4 INTERPRETACIÓN DE LA CURVA Y SU RELACIÓN CON LAS VARIABLES ECONÓMICAS	15
3.4.1 Hipótesis de expectativas	15
3.4.2 Hipótesis de Preferencia por liquidez.....	15
3.4.3 Segmentación de Mercado	16
3.5 MINERÍA DE DATOS	17
3.5.1 Árboles de Decisión (Decision Trees, DT).....	19
3.5.2 Algoritmo ID3 (Iterative Dichotomies 3).....	20
3.5.3 Algoritmo C4.5.....	21
3.5.4 Limitaciones del algoritmo C4.5.....	22
3.5.5 Algoritmo C5.0.....	23
3.5.6 Redes Neuronales (Artificial Neural Networks ANN).....	23
3.6 LOS USOS MÁS FRECUENTES DE TÉCNICAS DE MINERÍA DE DATOS APLICADAS A LAS FINANZAS	25
3.6.1. Predicción del mercado accionario	25
3.6.2 Detección de fraudes.....	26
3.6.3 Predicción mercado de divisas.....	27
3.6.4 Administración de Portafolios	27
3.7 TRABAJOS PREVIOS DE ESTIMACIÓN DE TASAS DE INTERÉS CON HERRAMIENTAS DE MINERÍA DE DATOS.....	28
3.8 PROPUESTA DE INVESTIGACIÓN, MOTIVACIONES Y OBJETIVOS	30
4 METODOLOGÍA	32
4.1 RECOLECCIÓN Y COMPRESIÓN DE LA DATA	34
4.1.1 Descripción de las variables y su proceso de obtención.....	34
4.1.2 Comprensión de los datos y construcción de los escenarios económicos.....	39
4.2 PREPARACIÓN DE LA DATA	45
4.3 MODELADO: SECCIÓN PREDICTIVA DE LA TASA A CINCO AÑOS.....	46
4.4 IMPLEMENTACIÓN Y RESULTADOS DE LOS MODELOS	48
4.4.1 Modelo Conglomerado-árbol I	48
4.4.2 Modelo Escenario manual-árbol II.....	57
4.4.3 Modelo III árbol simple.....	58
4.4.4 Modelo IV Red Neuronal.....	59
4.4.5 Modelo V Auto Regresivo.....	60
4.4.6 Resumen resultados.....	63
5. TEST COMPARATIVO DE LA CALIDAD DE LA PREDICCIÓN DE DOS MODELOS	64

5.1 RESULTADOS DEL TEST DM	66
6 USO DE LOS MODELOS EN UNA SIMULACIÓN DE INVERSIÓN	67
7. DISCUSIÓN.....	70
8 CONCLUSIONES Y PRÓXIMOS DESARROLLOS	72
9 REFERENCIAS	77
10 ANEXOS	83

1. Resumen

El objetivo de este trabajo consiste en la proyección de la estimación de la tasa libre de riesgo para el mercado Chileno utilizando herramientas de Minería de Datos. Primero se realiza un análisis de descriptivo de las variables de manera de identificar una relación causal entre variables relevantes para el mercado chileno e identificar escenarios económicos que en una segunda etapa, permitan discriminar la data para modelar árboles de decisión sobre cada uno de ellos. Los árboles de decisión tienen la ventaja de manejar una gran cantidad de variables junto a sus relaciones no lineales, siendo capaces de describir la dinámica de las tasa en términos de sus variables explicativas. Una vez construidos los diferentes modelos de árbol se mide su capacidad predictiva y se comparan con un modelo de redes neuronales y un modelo econométrico básico. El rendimiento de los árboles resulta ser superior al del resto y la configuración de los escenarios resulta ser útil para la comprensión de la dinámica de la economía chilena sirviendo de punto de partida para simulaciones y proyección de otras variables económico-financieras.

2. Introducción

La tasa libre de riesgo en Chile corresponde al rendimiento de deuda emitida por la autoridad monetaria y su riesgo crediticio corresponde al menor disponible en el mercado. Para Chile esta deuda está conformada por instrumentos emitidos por el Banco Central de Chile y la Tesorería general de la República a diferentes plazos, en pesos y unidades de fomento. Las tasas implícitas a diferentes plazos de estas emisiones conforman la estructura libre de riesgo la que conforma unos de los indicadores más relevantes de la economía explicando su importancia ya que de ella es posible interpretar una serie de factores que resumen el estado de la economía, tales como el premio por riesgo, expectativas de inversión y el equilibrio entre oferta y demanda de endeudamiento a diferentes plazos en el futuro.

La proyección de la estructura de tasas libre de riesgo es de gran interés para los distintos agentes del mercado, sin embargo hasta tiempos recientes no había sido un tema de investigación muy abordado, quizás debido a los malos resultados obtenidos en los primeros estudios en los que se modelaba la tasa como el resultado de una función que dependía de algunos factores de riesgo (Vasicek 1977 [82], Cox et al. 1985 [19], Dai and Singleton 2000 [21]), incluso en el año 2002 Duffee [27] muestra que las proyecciones de este tipo de modelos son de inferior calidad a los modelos *random-walk*. Estudios recientes basados sobre la reformulación dinámica del modelo de tres factores de Nelson & Siegel 1987 [62] han resultado ser bastante más efectivos y si bien estos tres factores son parsimoniosos e interpretables como: la curvatura, pendiente y nivel de la estructura de tasas a diferentes plazos, todos estos modelos son puramente estadísticos, presentando pocos fundamentos macroeconómicos y adoleciendo de relaciones causales provenientes de características específicas del mercado como en ocasiones cuando la autoridad monetaria usa la tasa de interés como una herramienta activa de política.

Otra desventaja de los modelos econométricos tiene que ver con el criterio de selección de variables. Si bien se pueden seleccionar variables explicativas con una base teórica o económica que la sustenta, muchas veces estas variables resultan ser irrelevantes o poco "significativas". Por el contrario, el cuidado de no sobre parametrizar el modelo puede significar la omisión de variables relevantes con su consecuente impacto en el modelo. Si bien la econometría cuenta con herramientas que permiten la selección del modelo más adecuado, estas metodologías están basadas en criterios que no están exentas de error y más relevante aun, no nos dicen si una variable pasa a ser significativa cuando otra de ellas alcanza un nivel específico o cuando se da un escenario extremo.

Con el avance de las tecnologías informáticas a partir de la década de los 80 junto a la acumulación de grandes cantidades de información en formato de bases de datos, se crearon y desarrollaron nuevas técnicas de análisis en las que la información era analizada sin la existencia de una hipótesis previa y por el contrario, se “aprendía” de la información ("Knowledge Discovery in Databases" o KDD [70]), comenzando a estudiar cómo interpretarla de mejor forma con el objetivo de encontrar los patrones que estaban escondidos en ella. El proceso antes descrito corresponde a la minería de datos [80] la cual está sustentada sobre la capacidad de aprendizaje de las máquinas y su capacidad para almacenar, clasificar e identificar patrones sobre grandes volúmenes de información.

Este proceso de obtención de conocimiento se caracteriza por ser estructurado e iterativo en el que es posible aplicar distintos modelos o algoritmos dependiendo de la naturaleza del problema. Los tipos de problemas pueden ser de clasificación, segmentación, asociación o regresión. Dentro de las técnicas de clasificación se encuentran las técnicas de árboles de decisión los cuales se caracterizan por ser una herramienta que jerarquiza las variables independientes en base a su poder explicativo de la variable objetivo con la capacidad de modelar relaciones no lineales de alta complejidad dada la cantidad de variables y permitiendo a su vez, describir el camino que sigue la variable explicada mostrando su dinámica hasta llegar resultado final.

El objetivo de esta tesis es utilizar este proceso estructurado con el objetivo de construir un modelo de árboles de decisión para predecir el movimiento de la tasa a cinco años libre de riesgo en Chile. Previo a esto se realiza un análisis descriptivo de las variables seleccionadas para este estudio de manera de encontrar patrones en la data que permitan configurar lo que llamaremos un “escenario económico” que utilizaremos como una variable de entrada en algunos diseños de árbol para la proyección de la tasa.

La organización de esta tesis comienza con el capítulo 3 que consiste en un marco teórico que describe y explica todas las metodologías y supuestos utilizados en este trabajo, incluyendo una descripción detallada del funcionamiento del mercado de renta fija libre de riesgo Chileno y de sus instrumentos. Continúa con una revisión de las técnicas de estimación de curvas de rendimiento y la teoría asociada a la dinámica de tasas de interés.

También se incluye una revisión teórica del proceso de minería de datos junto sus algoritmos utilizados en este trabajo y finalmente se incluye una revisión bibliográfica de la aplicación de las técnicas de minería de datos en las finanzas y se explican las motivaciones de esta tesis a la luz de los trabajos ya realizados en este tema.

En el capítulo 4 se implementa el proceso de minería de datos partiendo con la obtención de la data y la definición del problema. Aquí se definen las variables que serán incluidas en los modelos y se explica su importancia para el problema. Posteriormente se sigue con la sección de comprensión de la data en la que se realiza un análisis descriptivo de las relaciones entre las variables a lo largo de la historia estudiada, finalmente, se aplican técnicas de clasificación y segmentación en la búsqueda de condiciones económica-financieras consistentes en el tiempo que permitan resumir un periodo un “escenario económico”.

Este capítulo continúa con la preparación de la data y modelado donde se definen los modelos que serán contrastados y sus características. A continuación se prepara la data que servirá de entrada a estos modelos en términos de rezagos, muestreo, conversión y transformación. Para terminar este capítulo, se exponen y analizan los resultados de cada de los modelos mostrando los ajustes para mejorar su rendimiento.

En el capítulo 5 se incluye una sección que contrasta la significancia estadística de estos resultados con el test de Diebold y Mariano de manera de comprobar la robustez y significancia estadística de los resultados obtenidos.

En el capítulo 6 los modelos se ponen a prueba en un ejercicio de rentabilidad de cartera donde compiten bajo diferentes supuestos y estrategias de inversión. Finalmente, el capítulo 7 presenta las Conclusiones y una discusión de las principales implicancias de lo presentado en esta tesis.

3. Marco Teórico

Esta sección expone desde una perspectiva teórica las distintas metodologías y supuestos utilizados en este trabajo, junto a los estudios que los sustentan. Dado que este trabajo se centra en el mercado chileno de renta fija, primero se presenta una descripción de la emisión de deuda libre de riesgo en Chile, su historia y las razones que motivan su emisión. A continuación se detallan los tipos de instrumentos emitidos, su estructura y funcionamiento junto a una serie de estadísticas que permiten describir el mercado.

Considerando las condiciones específicas del mercado chileno en términos de tamaño y profundidad, se presenta una base teórica de estimación de curvas de interés de manera de obtener la variable objetivo desde la información de mercado secundario y adicionalmente, se presentan las teorías asociadas a la dinámica de las tasas de interés para fundamentar y explicar los resultados. Posteriormente, se detalla y explica el proceso de minería de datos junto a las herramientas de proyección que serán utilizadas en este trabajo, dándole un especial énfasis en las técnicas de árboles de decisión, explicando sus ventajas y desafíos.

A continuación se expone el trabajo y los avances existentes en la aplicación de técnicas de minería de datos a los problemas en finanzas, explicando sus usos actuales y potenciales junto a una revisión detallada de los trabajos sobre la proyección de tasas de interés, tema central de este trabajo.

A la luz de las evidencias expuestas en la revisión teórica, se exponen las motivaciones y objetivos de este trabajo. Adicionalmente, se plantean las preguntas e hipótesis a ser contestadas en base a la aplicación del proceso de minería de datos, sus herramientas proyectivas, y la explicación de los resultados sobre la base teórica de las tasas de interés y el conocimiento de la estructura del mercado chileno.

3.1 Emisión y Política Monetaria en Chile

En Chile el Banco Central de Chile (BCCCh) se creó en 1925 [9], con un directorio conformado por diez personas y con una personalidad jurídica de derecho público que es independiente del gobierno. Esta independencia fue perfeccionada con una serie de sucesivas leyes entre las que destacan la de 1960 donde se modificó la elección y composición del directorio, se crearon los cargos de Presidente y Gerente General para la administración del Banco y se fusionó al Banco Central con la Comisión de Cambios Internacionales, facultándolo para regular las operaciones de cambio internacional y comercio exterior. En la ley de 1975 se crea el “Consejo Monetario” que era

un organismo de nivel ministerial y sus funciones eran fijar la política monetaria, crediticia y arancelaria.

Actualmente, los objetivos de esta autonomía son la estabilidad de la moneda, es decir, el control de la inflación y la estabilidad de precios en la economía, y segundo, el normal funcionamiento de los pagos internos y externos lo que significa que el banco debe velar por el normal funcionamiento de los pagos, evitando posibles situaciones de liquidez del sistema financiero.

Para lograr los objetivos anteriores, el BCCh cuenta con tres políticas: la monetaria, la cambiaria y la financiera. La primera se focaliza en proteger el valor de la moneda nacional con una meta inflación que actualmente se encuentra en un 3% de inflación anual (con más/menos 1% de desviación), para esto controla la cantidad de dinero en la economía y vela que su "precio" asegurando que la tasa interbancaria (TIB), se mantenga cercana a la tasa de política monetaria (TPM), a través de operaciones de liquidez expansivas o restrictivas llamadas Operaciones de Mercado Abierto, compras/venta con pacto de retro compra, líneas de crédito y depósitos de liquidez.

La política cambiaria, dado el término de régimen de bandas de tipo de cambio en septiembre de 1999, se focaliza en la posibilidad de una intervención cuando el valor de la divisa está muy desviada de su valor de equilibrio, tal como ocurrió en enero del 2011 con un programa de compra de USD 12 mil millones [10].

Finalmente, la política financiera se focaliza en garantizar un sector bancario sólido y seguro que asegure el funcionamiento del sistema de pagos. Bajo esta política, el BCCh ha buscado completar y profundizar los mercados a través de la emisión de bonos en pesos y UF, junto con utilizarlos con un medio política monetaria, como medio de esterilización de la liquidez en pesos frente a escenarios de intervención del tipo de cambio.

En general, para la emisión de instrumentos existe un programa anual de licitaciones de bonos que es comunicado los primeros días del año y en los que especifican los montos e instrumentos a emitir [7]. Estas licitaciones corresponden al mercado primario de bonos donde pueden participar instituciones bancarias, sociedades financieras, administradoras de fondos de pensiones, administradoras de fondos mutuos y compañías de seguros (Capítulo IV.B.6.2 Compendio de Normas Financieras Banco Central de Chile [8]). El BCCH también realiza la licitación de bonos de la TGR en su papel de agente fiscal y los participantes autorizados son los mismos.

Si bien el programa de emisión de deuda es anual, este programa puede sufrir modificaciones explicadas por los distintos usos que la autoridad le puede dar a la emisión de deuda de largo plazo:

1.- La deuda de largo emitida por la autoridad permite señalar donde está el nivel de tasas en la economía en su rol de “creadores de mercado” estableciendo *benchmarks* que se caracterizan por un tamaño de emisión, regularidad de monos y plazos de emisión y liquidez que difícilmente se verá afectada por participantes individuales en el mercado. En Chile las emisiones más largas en con protección inflacionaria (Indizado a la unidad de fomento, UF) son bastante utilizadas por inversionistas institucionales, mayoritariamente fondos de pensiones que calzan sus pasivos de largo plazo con esta deuda.

2.- La emisión de deuda permite el desarrollo de otros mercados ya que esta deuda puede ser utilizada como colateral en el mercado repo. Adicionalmente, genera una masa crítica para el mercado en términos de la disponibilidad de instrumentos líquidos tanto por los volúmenes de emisión como porque a su vez el BCCH recibe estos instrumentos a cambio de liquidez instantánea a través de operaciones colateralizadas.

La apertura de nuevos mercados mundiales junto a las de nuevas oportunidades de inversión ha incidido en una modernización y profesionalización del mercado financiero chileno. Por otro lado existe el interés inherente de cada país de realizar un diagnóstico continuo de su economía de manera de asegurar su estabilidad financiera, para esto los gobiernos producen gran cantidad de información de la economía y de sus cuentas nacionales velando por el bienestar económico de su población.

También existe un mercado secundario, que corresponde a todas las operaciones realizadas entre personas jurídicas o naturales de bonos emitidos en el mercado primario, más las emisiones de bonos de agentes privados y sus respectivas operaciones en el mercado secundario. Estas operaciones son realizadas en las bolsas (remate electrónico o calce de posiciones) y el mercado "*over the counter*" (donde existen solo dos partes que llegan a un acuerdo).

Después de la quiebra de los bancos en 1981, Chile comenzó un proceso de liberalización y profundización de los mercados con mejoras sustantivas en su regulación. Entre los hitos más importantes destacan la ley de bancos de 1997 que aplica el estándar internacional Basilea I, la nueva regulación a las administradoras de Fondos de pensiones que permite las inversiones en el extranjero, la reforma al Mercado de Capitales I, la nominalización del banco Central en 2001 y la libre flotación del tipo de cambio en 1999. Todos estos avances en el mercado chileno permitieron

su rápido desarrollo, generándose a su vez un mayor volumen información financiera de mejor calidad y con menos distorsiones permitiendo realizar proyecciones cada vez de mejor calidad con frecuencia intra-diaria.

3.2 Instrumentos de Deuda en Chile

Un instrumento de deuda es una obligación en el que su emisor obtiene fondos del mercado y a cambio, este se compromete a regresar una serie de pagos futuros en fechas específicas. Esta obligación puede ser transada en el mercado por lo que emisor debe realizar los pagos prometidos al que sea el dueño en ese momento. En general, la estructura de estas obligaciones está basadas en unos pocos parámetros que caracterizan sus flujos asociados y tabla de desarrollo o calendario de flujos de manera de que sean relativamente estandarizados y por lo tanto fáciles de valorizar y transar. Estas características son la moneda, la tasa cupón (fija o variable, simple o compuesta), la frecuencia de los pagos (anual, semestral, etc.), la base de conteo de días para intereses (el devengo desde el ultimo cupón), la fecha de emisión, la fecha de vencimiento y cuándo se regresa el principal o valor facial de la del instrumento. Existen variadas combinaciones y consideraciones adicionales de las características mencionadas más arriba en esta sección, se detallara la matemática y características asociadas a los instrumentos libres de riesgo emitidos por el BCCh y la TGR actualmente transados en la Bolsa de Comercio de Santiago. En Chile los instrumentos son emitidos con las siguientes características:

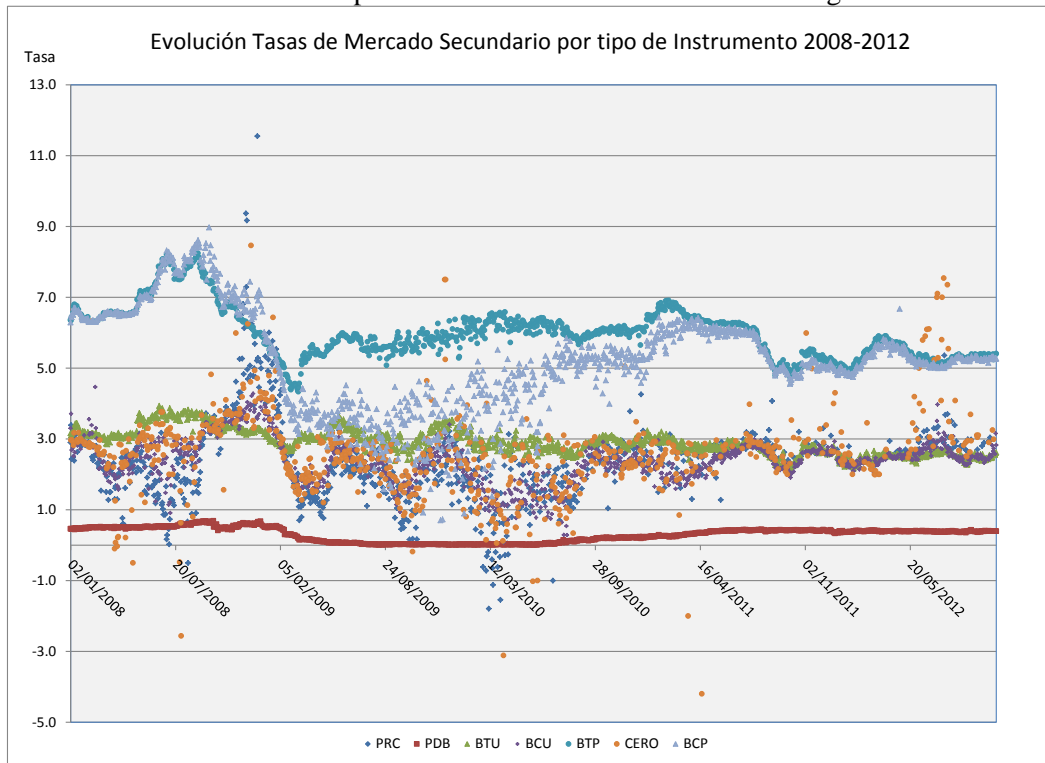
Tabla 1: Instrumentos soberanos libre de riesgo en Chile

Instrumento	PRC	CERO	BCU	BCP	PDBC	BTU	BTP
Emisor	BCCh	BCCh	BCCh	BCCh	BCCh	TGR	TGR
Moneda	UF	UF	UF	PESO	PESO	UF	PESO
Tasa cupón	8.50%	Sin cupón	Entre 3% y 6%	Entre 4% y 6%	Sin cupón	Entre 4.5% y 2.1%	6%
Frecuencia cupón	Semestral	Al vencimiento	Semestral	Semestral	Al vencimiento	Semestral	Semestral
Conteo Intereses	ACT/360	ACT/360	ACT/360	ACT/360	sin intereses	ACT/360	ACT/360
Amortización	En cada cupón	Al vencimiento	Al vencimiento	Al vencimiento	Al vencimiento	Al vencimiento	Al vencimiento
Plazo	8,12,15,20 años	8,12,15,20 años	5,10,20 y 30 años	5,10,20 y 30 años	1,3,9 y 12 meses	5,10,20 y 30 años	5,10,20 y 30 años

Fuente: Características de los Instrumentos del Mercado Financiero Nacional, BCCh 2005 [6]

El siguiente gráfico muestra las tasas de las transacciones en la Bolsa de Comercio de Santiago. Se observa que la cantidad de operaciones totales no es tan alta, sin embargo la periodicidad de las operaciones ha mejorado de la mano de la los inversionistas institucionales y los fondos de pensiones.

Gráfico 1: Tir de mercado operaciones Bolsa de Comercio de Santiago 2008-2012



Fuente: Bolsa de Comercio de Santiago

Los cuadros siguientes permiten caracterizar el mercado secundario forma mostrando la cantidad de operaciones anuales abiertas en peso y unidades de fomento:

Tabla 2: Operaciones Bonos en Peso

Año	Numero Operaciones Pesos			
	0-1Y	1-5Y	5-10Y	Total
2008	7,707	1,928	1,820	11,455
2009	10,407	3,068	3,131	16,606
2010	7,265	2,563	1,749	11,577
2011	5,641	2,377	1,812	9,830
2012	2,702	1,273	998	4,973

Tabla 3: Operaciones Bonos en UF

AÑO	Numero Operaciones UF				
	0-1Y	1-5Y	5-10Y	10+Y	TOTAL
2008	1,049	8,234	4,124	3,462	16,869
2009	1,151	10,983	4,730	2,930	19,794
2010	700	8,699	4,532	2,104	16,035
2011	559	7,062	3,666	1,825	13,112
2012	605	3,378	2,219	1,203	7,405

Las tablas muestran los distintos plazos en los que es posible encontrar emisiones en peso y UF, resalta el bajo nivel (hablar sobre la nominalización) de emisión de deuda de corto plazo en UF junto a la progresiva disminución de operaciones a plazos más largos en pesos, lo que muestra la poca liquidez de algunos sectores de la curva y revela el uso que se le da a la denominación de los instrumentos.

En general ha aumentado la cantidad de operaciones en el mercado secundario, sin embargo, aún no existen operaciones a todos los plazos con frecuencia diaria lo que genera incertidumbre en algunos tramos de la curva cero cupón en peso y UF.

La estructura de estos instrumentos se puede resumir en tres tipos:

1.- **Bonos Bullet:** Son aquellos bonos en que la amortización o nominal ocurre íntegramente en su último flujo o vencimiento, estos instrumentos se valorizan de la siguiente forma:

$$P = \frac{\sum_{n=1}^N C_n + A_N}{(1 + r_n)} \quad \text{Ecuación (1)}$$

Donde P es el precio; C_n el cupón n ; A_N es la amortización en el último flujo; r_n la tasa cero cupón para el periodo n . En el mercado, los instrumentos se transan con la TIR o *yield* que cuantifica el rendimiento implícito del instrumento si pagamos su precio y lo mantenemos al vencimiento:

$$P = \frac{\sum_{n=1}^N C_n + A_N}{(1 + y)^{n/365}} \quad \text{Ecuación (2)}$$

Si bien la TIR nos da un rendimiento total al vencimiento y en condiciones normales sería un valor mayor que cero, es necesario evaluar si este rendimiento es mayor o menor que el rendimiento por intereses, lo que corresponde al valor nominal del bono, más los intereses devengados desde el último cupón:

$$V_{Par} = A_N (1 + r_c f) \quad \text{Ecuación (3)}$$

Donde r_c es la tasa cupón del bono y f es la porción del cupón ya devengado que depende de la convención del conteo de días, composición de la tasa y frecuencia del cupón. El valor porcentual obtenido de la razón del precio representa el valor de un bono expresado como el porcentaje de su valor justo si no existen cambios de tasa de mercado ni en su riesgo de crédito:

$$P_{\%} = \frac{P}{V_{Par}} \quad \text{Ecuación (4)}$$

Por lo que un bono a un $P_{\%}$ menor a 100, significa que este instrumento se transa a descuento, por el contrario, un valor mayor a 100 indica que se transa con un premio. Los instrumentos BCU, BCP, BTP y BTU son bonos bullet.

2.- Bonos con amortización periódica: Son bonos en el que el cupón está compuesto de amortización e intereses. Para el mercado local estos bonos corresponden a los bonos del tipo PRC,

para los cuales es necesario calcular una tabla de desarrollo la que depende de la tasa cupón, la fecha de emisión y el corte de la emisión:

Tabla 4: Tabla de desarrollo PRC-12
NEMOTECNICO: PRC-
5B0401

FECHA CUPON	CUPO N	INTER ES	PRINCIPA L	OUTSTANDIN G
01/10/2001	6.032	3.253	2.779	97.221
01/04/2002	6.032	3.145	2.887	94.334
01/10/2002	6.032	3.069	2.963	91.371
01/04/2003	6.032	2.956	3.076	88.295
01/10/2003	6.032	2.872	3.16	85.135
01/04/2004	6.032	2.769	3.263	81.872
01/10/2004	6.032	2.663	3.369	78.503
01/04/2005	6.032	2.54	3.492	75.011
01/10/2005	6.032	2.44	3.592	71.419
01/04/2006	6.032	2.31	3.722	67.697
01/10/2006	6.032	2.202	3.83	63.867
01/04/2007	6.032	2.066	3.966	59.901
01/10/2007	6.032	1.949	4.083	55.818
01/04/2008	6.032	1.816	4.216	51.602
01/10/2008	6.032	1.679	4.353	47.249
01/04/2009	6.032	1.528	4.504	42.745
01/10/2009	6.032	1.39	4.642	38.103
01/04/2010	6.032	1.233	4.799	33.304
01/10/2010	6.032	1.083	4.949	28.355
01/04/2011	6.032	0.917	5.115	23.24
01/10/2011	6.032	0.756	5.276	17.964
01/04/2012	6.032	0.584	5.448	12.516
01/10/2012	6.032	0.407	5.625	6.891
01/04/2013	7.114	0.223	6.891	0

Fuente: Elaboración Propia

Esta tabla de desarrollo se construye progresivamente calculando primero el cupón total:

$$C_n = \frac{Cte (1 + t_c)^{0.5}}{\left(1 - \frac{1}{(1 + (1 + t_c)^{0.5})^n}\right)} \quad \text{Ecuación (5)}$$

Con C_n = Cupón, Cte = Corte de instrumento y t_c = Tasa cupón. Donde una parte de este cupón corresponde a intereses, calculados con un conteo de días 30/360:

$$I_n = Ins_{n-1} (1 + t_c)^{d/360} - Ins_{n-1} \quad \text{Ecuación (6)}$$

y una porción a Amortización del capital:

$$A_n = C_n - I_n \quad \text{Ecuación (7)}$$

Con esto tenemos la primera fila de la tabla y nos deja en pie para calcular el saldo insoluto Ins_n para comenzar la segunda (el primer Ins_n en 100 por que aun no se amortiza nada):

$$Ins_n = Ins_{n-1} - A_n \quad \text{Ecuación (8)}$$

La tabla anterior muestra que la relación amortización/ interés es decreciente a medida que nos acercamos al vencimiento, el ultimo capón es diferente ya que se paga todo el saldo de capital pendiente. Para valorar estos instrumentos se utiliza las mismas fórmulas utilizadas para el bono *bullet*:

$$P = \frac{\sum_{n=1}^N C_n + A_n}{(1 + r_n)} \quad y \quad P = \frac{\sum_{n=1}^N C_n + A_n}{(1 + Tera)} \quad \text{Ecuación (9)}$$

3.3 Metodologías de estimación de una curva cero cupón

La curva de rendimiento ha sido tema de investigación de innumerables trabajos, incluso en Chile. Existen variadas alternativas de ajuste para el cálculo de una curva cero cupón empezando por una simple interpolación lineal o *bootstrapping* para completar los plazos faltantes, hasta modelos que son soluciones de ecuaciones diferenciales. En general todas estas metodologías ajustan y conectan tasas cero cupón a cualquier plazo permitiéndonos descontar flujos a su valor presente, observar el premio por riesgo a distintos plazos, o también, inferir sobre las expectativas económicas por parte de los distintos agentes (Estrella y Hardouvelis, 1991 [30]).

Nelson y Siegel (1987) [62] desarrollaron una metodología caracterizada por ser un modelo parsimonioso o de pocos parámetros en la que su forma funcional depende de factores modelados sobre la base de criterios económicos. Este modelo parte sobre la base de que la tasa forward t es igual a:

$$f(t) = \varphi_0 + \varphi_1 e^{\left(\frac{-t}{\lambda}\right)} + \varphi_2 \frac{t}{\lambda} e^{\left(\frac{-t}{\lambda}\right)} \quad \text{Ecuación (10)}$$

Con $\varphi_0, \varphi_1, \varphi_2$ y λ parámetros a estimar con mínimos cuadrados ordinarios (MCO) u otra metodología Diebold and Li (2006) [25]. En este modelo, la constante representa el nivel de la tasa de largo plazo, el segundo factor que corresponde a una función de decaimiento exponencial de la pendiente a la curva (Si $\varphi_1 > 0$ entonces la función exponencial es monótona creciente y $\varphi_1 > 0$ es

decreciente), Finalmente, El tercer término es la multiplicación de un polinomio por una función de decaimiento exponencial que dependiendo del $\varphi_2 > 0$ genera una joroba, o un valle en caso contrario, es decir, la curvatura. La tasa instantánea puede ser expresar como una suma ponderada de las tasas forward, que de manera continua corresponde a:

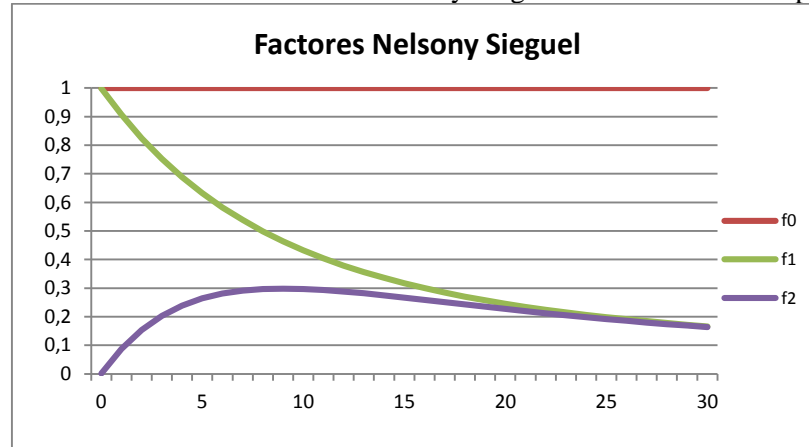
$$r(t) = \int_0^t f(u) du \quad \text{Ecuación (11)}$$

Evaluando esta función en t y resolviendo la ecuación diferencial queda:

$$r(t) = \varphi_0 + \varphi_1 \lambda \frac{\left(1 - e^{\left(\frac{-t}{\lambda}\right)}\right)}{t} + \varphi_2 \lambda \frac{\left(1 - e^{\left(\frac{-t}{\lambda}\right)}\right)}{1 - e^{\left(\frac{-t}{\lambda}\right)}} \quad \text{Ecuación (12)}$$

Si dejamos λ constante en cinco y graficamos cada una de estas tres componentes en función del tiempo:

Gráfico 2: Factores curva de Nelson y Siegel en función del tiempo



En este gráfico podemos observar claramente los tres factores que explican el ajuste. Cuando $t \rightarrow \infty$ las funciones f_1 y f_2 tienden a cero y lo único que queda es el nivel o f_0 , es decir que el valor de φ_0 representa el nivel de la tasa de largo plazo. Cuando $t \rightarrow 0$ solo el componente de curvatura o f_2 se hace cero y en el límite la tasa es $(\varphi_0 + \varphi_1)$. El segundo factor ponderado por φ_1 representa la pendiente de la curva de rendimiento. Finalmente φ_1 controla la velocidad a la que la curvatura y

la pendiente caen a cero, lo que junto a λ que controla en que punto de t estar localizado el máximo o mínimo de la curvatura explican la forma final de la curva de rendimiento.

3.4 Interpretación de la Curva y su relación con las variables económicas

La curva cero cupón muestra la relación entre rendimiento y madurez que los agentes económicos tienen de instrumentos financieros con la misma calidad crediticia. Su nivel, pendiente y curvatura representan el actuar de distintas variables sobre las cuales se pueden inferir futuros escenarios económicos. La curva cero cupón ha sido ampliamente utilizada como variables indicativas del curso y la efectividad de la política monetaria por parte de los bancos centrales y también como la base sobre la cual los distintos agentes económicos proyectan la evolución de variables económicas, valoran sus activos y optimizan sus portafolios. Las teorías que permiten interpretar y explicar su relación con las distintas variables económicas se pueden resumir en la hipótesis de las expectativas y la de segmentación de mercado las cuales dependiendo de los supuestos utilizados de mercado

3.4.1 Hipótesis de expectativas: Los precursores de esta teoría fueron Fisher y Hicks (1939) [38] y consiste en que la curva de rendimiento refleja las expectativas de los inversionistas sobre las tasas de interés y de inflación a futura. Bajo esta hipótesis y suponiendo un mercado sin imperfecciones (Las tasas están arbitradas, no existen costos de transacción, existe perfecta información por parte de los agentes y los mercados no están segmentados) las tasas de plazos mayores de la curva son una ponderación entre las tasas cortas actuales y esperadas en el mercado. Es decir, si se cree que las tasas de interés no cambiarán en el futuro implica que debiera dar lo mismo tomar dos depósitos a seis meses versus tomar uno a un año, lo que llevaría a una curva plana.

Si los agentes esperan una mayor inflación y por lo tanto mayores tasas de interés en el futuro, naturalmente los agentes se segmentarán en los que quieren invertir a plazos menores y poder reinvertir a tasas mayores en el futuro contra aquellos que preferirán endeudarse a largo plazo a una tasa menor. El punto de equilibrio de estas preferencias lleva a que necesariamente las tasas largas sean mayores a las cortas. Lo anterior sugiere que cuando el mercado espere un aumento de las tasas futuras de corto plazo implicará un aumento de las tasas de plazo mayores y viceversa ya que estas tasas más largas son un promedio ponderado entre las tasas cortas actuales y las esperadas.

3.4.2 Hipótesis de Preferencia por liquidez: El precursor de esta teoría fue Hicks (1939) [38] y consiste en que para un individuo averso al riesgo un instrumento a largo plazo tiene tasas que tienden a ser mayores que a corto plazo ya que los instrumentos a mayor plazo tienen menor liquidez y son muy sensibles a los desplazamientos de las tasas de interés. Por tanto, la curva de

rendimiento tenderá a ser ascendente. Esta teoría es más bien complementaria a la anterior ya es compatible con la presencia de expectativas, sin embargo, la curva igual debería tener pendiente positiva aun si el mercado no espera cambios en la tasa de interés.

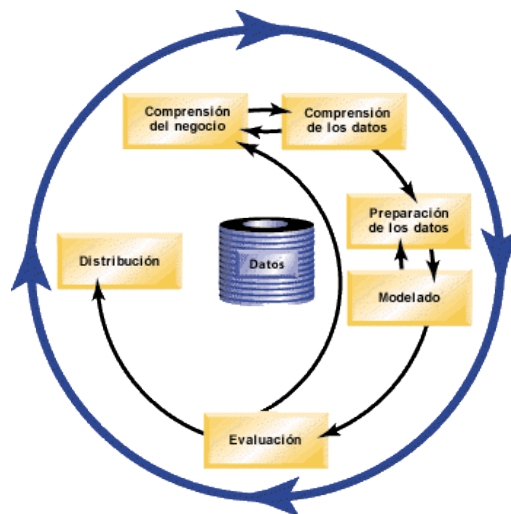
3.4.3 Segmentación de Mercado: Sugiere que el mercado está dividido en segmentos de la curva donde cada actor del mercado se ubica en el plazo donde eliminan el riesgo sistemático calzando sus activos con sus pasivos por lo que el punto de equilibrio para cada plazo está determinado por la oferta y demanda de préstamos de cada segmento. La relación general entre las tasas vigentes en cada segmento determina la pendiente de la curva de rendimiento.

3.5 Minería de Datos

Se considera Minería de datos (Data Mining, DM) o *knowledge discovery* [80] como un proceso computacional que analiza grandes volúmenes de información en la búsqueda de patrones y relaciones sistemáticas con un proceso que utiliza métodos de aprendizaje, inteligencia artificial y estadística para resumir, categorizar y fundamentar nuevo conocimiento que pueda ser útil para tomar decisiones. El objetivo final del DM es poder predecir una variable o comportamiento siendo esta la principal razón de que este proceso tenga tantas aplicaciones en distintos campos.

Existe una serie de paquetes informáticos especializados en DM (Microsoft, IBM, Oracle, etc.) que son capaces de aplicar diversas transformaciones a la data, corregirla, resumirla, modelarla e incluso en base a criterios estadísticos son capaces de proponer cual es modelo más adecuado para el problema. Si bien estos software son bastantes avanzados, el factor interpretativo del usuario al utilizar estas herramientas es fundamental y depende en gran parte del grado de "conocimiento del negocio" que este tenga de manera de poder definir una hipótesis inicial. Este punto inicial en el proceso de obtención de conocimiento es iterativo y se resume en una serie de pasos bien definidos que se detallan a continuación:

Figura 1: Proceso de Minería de Datos



FUENTE: IBM CRISP DM

i) **La definición del Problema:** En esta etapa se identifican los objetivos del estudio entendiendo el problema y definiendo las variables relevantes. Esta parte no es necesaria la utilización de ningún software.

ii) **Exploración y preparación de la Data:** En esta parte del proceso se estudia la data disponible se evalúa que información entregan, en general se le aplican estadísticos que la describan, se determina su calidad y finalmente se le transforma y "limpia" para usarlos como variables de entrada en los modelos.

iii) **Modelamiento:** En esta etapa se seleccionan y aplican distintos modelos sobre la data se calibran sus parámetros a sus valores óptimos, generalmente es necesario volver a la etapa anterior para perfeccionar la data de entrada.

iv) **Evaluación:** Aquí se someten a evaluación los modelos estudiando los pasos que llevaron a sus resultados, verificando si estos cumplen los objetivos del estudio intentando identificar si hay factores de importancia que no fueron considerados en el modelo. En esta etapa ya es posible definir si es posible utilizar la información generada por el DM para tomar decisiones sobre la problemática abordada.

v) **Implementación:** En esta etapa los resultados son utilizados y exportados a reportes o a otras bases de datos.

Las técnicas de minería de datos para modelar la data son variadas y su elección a veces no resulta fácil ya que todos los modelos dan resultados distintos. Lo bueno es que no es necesario estar limitado a un solo algoritmo para obtener soluciones, en relativamente fácil aplicar varios de ellos y ver cual o cuales modelos ofrecen una vista completa de la variable explicada. Los modelos más comunes en DM son:

i) **Algoritmos de Clasificación:** Se utilizan para predecir una o más variables discretas basándose en distintos atributos de la data, es decir, buscan predecir por ejemplo si algo va a suceder si ocurre una serie de otros factores. Las funciones de este tipo son algoritmos de árboles de decisión, algoritmos de agrupación (*o clustering*), redes neuronales y Clasificador de Bayes (*o Naive Bayes classifier*).

ii) **Algoritmos de Regresión:** Es lo mismo que las variables de clasificación pero cambia el tipo de variable dependiente, aquí se predice una variable de tipo continua. Los algoritmos de regresión son capaces de determinar que variables de entradas son relevantes en la predicción, estas variables de

entrada o explicativas pueden ser continuas o discretas o categóricas y una combinación de estas va a simular el resultado de la variable explicada. Las funciones de este tipo son algoritmos de árboles de decisión, Algoritmos de regresión (Como MCO o Máxima verosimilitud) en corte transversal o series de tiempo.

iii) **Algoritmos de segmentación:** Estas funciones agrupan la data en grupos o clústeres (*clusters*) que comparten ciertas características o tienen propiedades similares. Los métodos de agrupación son variados y van desde la agrupación genérica donde se identifican características que aparecen juntas con frecuencia sin ninguna noción preconcebida de un patrón existente, hasta metodologías que detectan un centro en la data y agrupa la variable normalizándolas y clasificándolas respecto de la "distancia" a cada uno de estos centros. Las funciones de este tipo son algoritmos de árboles de decisión, algoritmos de agrupación.

iv) **Algoritmos de Asociación:** Son algoritmos que buscan correlaciones o asociaciones en la data, es decir se buscan elementos que suelen suceder juntos, por ejemplo al asociar los productos comprados en un supermercado (cerveza y papas fritas). Este tipo de análisis es posible dado el gran poder computacional disponible pero se debe tener cuidado de poder identificar esas relaciones que sean espurias o mera casualidad. Las funciones de este tipo son algoritmos de árboles de decisión y técnicas de asociación usando todas las combinaciones posibles de todas la variables y seleccionando aquellas que más se repiten para después algunos test de independencia un ejemplo de esto es el *Apriori algorithm* (Quinlan 1993) [71].

v) **Algoritmos de Secuencia:** Estas funciones encuentran secuencias típicas de sucesos de datos, es decir encuentran patrones sobre una lista de objetos, importando su orden. Un ejemplo simple se da cuando una persona ve la primera parte de una película y en consecuencia ve después la segunda parte.

En este trabajo se enfoca principalmente en dos algoritmos de clasificación que son los árboles de decisión y las redes Neuronales:

3.5.1 Árboles de Decisión (*Decision Trees, DT*)

Esta técnica predictiva de clasificación consiste en una división jerárquica y secuencial del problema en el que cada una de estas divisiones o nodos describen gráficamente las decisiones posibles y por lo tanto los resultados de las distintas combinaciones de decisiones y eventos. A cada evento se le asignan probabilidades y a cada una de las ramas se le determina un resultado. Los

árboles representan “reglas” las que pueden ser entendidas por lo humanos con la ventaja de que el conocimiento lo genera el mismo árbol y no parte de la premisa de un experto en el tema.

La mayoría de los algoritmos utilizados para construir un árbol son variaciones de uno genérico llamado “*Greedy algorithm*” que básicamente va desde la raíz hacia abajo (*Top-Down*) buscando de manera recursiva los atributos que generan el mejor árbol hasta encontrar el óptimo global con una estructura de árbol lo más simple posible. Los algoritmos más conocidos son el ID3 el C4.5 (Quinlan 1993), C5.0 y CART (*Classification And Regression Trees*). La diferencia entre los 3 primeros y CART es la posibilidad del último de obtener valores reales como resultados versus valores discretos en el resto de los modelos. Las principales características de estos algoritmos es su capacidad de procesar un gran volumen de información de manera eficiente y que pueden manejar el ruido (error en los valores o en la clasificación de estos) que pudiese existir en los datos de entrenamiento. En general los distintos modelos se diferencian en el algoritmo que clasifica los distintos atributos del árbol y la eficiencia de estos en con el objetivo de obtener un mejor árbol que sea lo más simple posible.

3.5.2 Algoritmo ID3 (Iterative Dichotomies 3):

Este algoritmo fue propuesto por J Ross Quinlan 1975 en su libro “*Machine learning*” vol. 1. Básicamente ID3 construye un árbol de decisión (DT) desde un set fijo de “ejemplos”, el DT generado se usa para clasificar futuros ejemplos. Cada ejemplo tiene varios atributos que pertenecen a una clase (como los valores sí o no). Los nodos de “hoja” del árbol (*leaf nodes*) contienen el nombre de la clase, mientras que los nodos “no-hoja” son los nodos de decisión donde cada uno de ellos (cada rama) corresponde un posible valor del atributo. Cada nodo de decisión es una prueba del atributo con otro árbol que comienza a partir de él. El algoritmo ID3 utiliza el criterio de la “ganancia de información” para decidir que atributo va en cada nodo de decisión. Esta medida estadística mide que tan bien un atributo divide los ejemplos de entrenamiento en cada una de las clases seleccionando aquella con más información (información útil para separar). Para definir “ganancia de información” primero debemos definir el concepto entropía que básicamente corresponde a la cantidad de incertidumbre en un atributo.

Dada una colección S (data sobre la cual se calcula la entropía que cambia con cada iteración) con c posibles resultados:

$$Entropía(S) = \sum -p_x \log_2 p_x \quad \text{Ecuación (13)}$$

Donde x es el set de clases en S y p_x es la proporción de S que pertenece a la clase x . Cuando $Entropía(S) = 0$, entonces el set S está perfectamente clasificado, es decir, que todos los elementos de S están en la misma clase (si es 1 es que están clasificados en forma aleatoria). De la definición anterior se desprende que la “ganancia de información” (*Information Gain* “IG”) corresponde a la diferencia en entropía entre antes de haber separado la data S por un atributo versus después de hacerlo, es decir, en cuanto se redujo la incertidumbre en el set S después de dividirla en el atributo “A”:

$$IG(S, A) = Entropía(S) - \sum \left(\frac{|S_v|}{|S|} \right) Entropía(x) \quad \text{Ecuación (14)}$$

Donde S_v es el set de S para el cual el atributo A tiene el valor v . Los elementos $|S_v|$ y $|S|$ corresponden al número de observaciones en S_v y S respectivamente.

3.5.3 Algoritmo C4.5:

Este algoritmo fue desarrollado por Ross Quinlan en 1993 [71] y básicamente es una versión avanzada del algoritmo ID3 en el que se incluyen las siguientes capacidades o ventajas:

- a.- Manejo de valores continuos y discretos: Para manejar atributos continuos lo que hace el algoritmo es crear un umbral para después dividir el atributo entre aquellos que están sobre y bajo el umbral. Esta característica es fundamental para este estudio donde la mayoría de los valores son continuos y sus umbrales pueden ser de alta relevancia.
- b.- Tiene la capacidad de manejar valores de atributos faltantes: En el caso de un atributo faltante el algoritmo usa una ponderación de valores y probabilidades en vez de valores cercanos o comunes. Esta probabilidad se obtiene directamente de las frecuencias observadas para esa instancia, por lo que se puede decir que el algoritmo C4.5 usa la clasificación más probable calculada como la suma de los pesos de las frecuencias de los atributos.
- c.- Es capaz de generar un set de reglas que son mucho más fáciles de interpretar para cualquier tipo de árbol.

d.- Este algoritmo construye un gran árbol y lo concluye con una “poda” de las ramas de manera de simplificarlo de manera de generar resultados más fáciles de entender y haciéndolo menos dependiente de la data de prueba.

3.5.4 Limitaciones del algoritmo C4.5

Aunque este algoritmo es uno de los más populares este presenta algunas deficiencias entre las que podemos nombrar:

a.- La presencia de ramas vacías: A veces los arboles presentan ramas que tienen nodos con casi cero valores o muy cercanos a él. Estos nodos no ayudan a construir reglas ni tampoco con la clasificación de los atributos y solo hacen que el árbol sea más grande y más complejo.

b.- Ramas poco significativas: Esto sucede cuando el número de atributos discretos crean la misma cantidad de ramas para construir el árbol, generando ramas que no son relevantes en la tarea de clasificación disminuyendo su poder predictivo y generando el problema del sobre ajuste.

c.- Sobreajuste (*Over fitting*): Este problema sucede cuando el algoritmo selecciona información con características poco usuales, causando fragmentación en el proceso de distribución. La fragmentación se define como la creación de nodos estadísticamente no significativos por los que pasan muy pocas muestras. Crear un árbol que clasifique todos los datos de entrenamiento perfectamente puede no llevarnos a mejor generalización de los datos con características poco usuales. Este problema se manifiesta generalmente con data que presenta ruido.

Para solucionar este problema, este algoritmo cuenta con las técnicas de pre y post podado del árbol. El pre podado consiste en detener el crecimiento del árbol en su construcción cuando no hay suficientes datos para obtener resultados confiables. En el caso del post podado, una vez generado todo el árbol, se eliminan todos aquellos sub árboles que no tengan suficiente evidencia reemplazándola con la clase de la mayoría de los elementos que quedan o con la distribución de probabilidades de la clase. Para seleccionar cuales sub árboles deben ser recortados se utilizan los siguientes métodos:

1. Validación cruzada: Consiste en separar algunos datos de entrenamiento el árbol para evaluar la utilidad de los sub árboles.
2. Test estadístico: Utilizar un test estadístico sobre los datos de entrenamiento de manera de poder identificar información con origen aleatorio.

3. Largo de descripción mínima (*Minimum description Length MDL*): Consiste en determinar si la complejidad adicional del árbol resulta mejor que simplemente recordar las excepciones resultantes del recorte.

3.5.5 Algoritmo C5.0:

Al igual que sus predecesores, este algoritmo construye los árboles en base a un conjunto de datos de entrenamiento optimizado bajo el criterio de ganancia de información y corresponde a una evolución de su versión anterior, el algoritmo C4.5. Las mayores ventajas de esta versión tienen que ver con la eficiencia en el tiempo de construcción de árbol, el uso de memoria y la obtención de árboles considerablemente más pequeños que en el C4.5 con la misma capacidad predictiva. Adicionalmente, tiene la opción de ponderar algunos atributos de manera de enfocar la construcción del árbol y se pueden utilizar un aprendizaje penalizado en que es posible asignar un costo a los posibles resultados o matriz de resultados (*Cost sensitive algorithm*) (Weiss, McCarthy , Zabar 2008)[83].

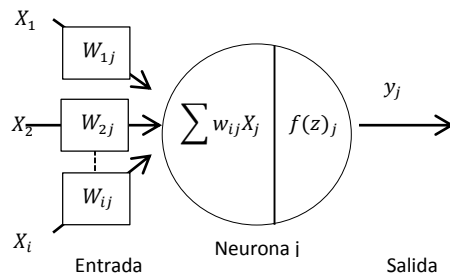
Los arboles obtenidos en este trabajo están modelados con el algoritmo C5.0 para poder utilizar gran cantidad de información, maneje mejor las variables continuas y los arboles obtenidos sean más simples y fáciles de entender.

3.5.6 Redes Neuronales (*Artificial Neural Networks ANN*)

Un modelo de redes neuronales, llamado así por su semejanza al funcionamiento de las células del sistema nervioso, consiste en una densa red de unidades interconectadas donde cada una de estas unidades recibe un número de valores de entrada, las procesa y produce un valor único de salida. Tanto las entradas como la respuesta pueden provenir o pueden servir de entrada a otra unidad. Estas unidades reciben el nombre de nodos o neuronas las que están interconectadas por enlaces de comunicación los cuales tienen ponderadores los que son capaces de almacenar conocimiento y multiplicar la señal ponderada (conocimiento) al resto de la red.

La gran ventaja de las redes neuronales proviene del uso de un gran número de estas neuronas entrenándose en paralelo y generando una red organizada en capas de manera adaptativa que aprende de la experiencia creando un modelamiento propio del problema, representado por la cantidad de nodos en cada capa y las interconexiones entre ellos lo que le da la forma a red

Figura 2: Esquema funcionamiento Red Neuronal



Donde y_j representa a la única salida de la neurona i , $f()$ es la función de transferencia, W_{ij} es la ponderación de la entrada j en la neurona i o lo que podríamos llamar el “umbral de activación” y finalmente X_j corresponde al valor de entrada j .

La unidad básica anterior se organiza de forma jerárquica formando capas. Estas capas se caracterizan por recibir la información de la misma fuente (la data de entrada u otra capa) y todas sus salidas van también al mismo destino (la data de salida u otra capa). Lo anterior implica que existe capa de entrada, capas intermedias u ocultas y capas de salida de la red. En estos términos las redes neuronales pueden ser monocapa o multicapa.

La distinción de estos tipos de estructura tiene que ver con el uso que se le puede dar a la red. En el caso de las redes multicapa llamadas así por contar con más de una capa son llamadas modelo heteroasociativas. Se caracterizan por su capacidad de aproximarse al recorrido de cualquier función, es decir que es capaz de aprender casi cualquier relación entre el conjunto de datos de entrada y el de salida. Funcionan aprendiendo en parejas de datos, una entrada (A_i) y una salida (B_i) de manera que cuando asocian entradas con diferentes salidas necesita a lo menos dos capas, una para captar y retener todas las posibles entradas y otra capa para mantener todas las salidas asociadas a cada entrada, si no existiese al menos otra capa la información se perdería la información inicial al obtenerse la salida asociada. Es necesario retener la información de entrada en la primera capa para poder acceder múltiples veces a ella.

Por su parte una red mono capa también llamadas redes auto asociativas consiste en una capa donde todas sus neuronas están interconectadas. Funcionan aprendiendo de los datos de entrada de manera que cuando se le presenta un dato de entrada, este buscará dentro de sus datos almacenados y responderá con el que más se asemeja. En general este tipo de redes se utiliza en tareas de filtrado de información para la reconstrucción de datos o encontrando relaciones similares dentro de una base de datos.

El funcionamiento de la red neuronal es simple donde múltiples procesadores elementales trabajando en paralelo que se conectan entre si y se van adaptando lo que significa que la red va ajustando los pesos de las interconexiones para alcanzar los requerimientos de desempeño del problema basado en sets de entrenamiento. Este proceso de ajuste o entrenamiento de la red puede ser supervisado o no supervisado.

El aprendizaje supervisado consiste en ajustar los pesos de la red dándole los *inputs* y la respuesta correcta de manera que la salida de la red sea lo más parecida posible. La información de entrada se propaga a través de la red hasta que llega a las neuronas de salida, si la respuesta es igual a la real entonces no es necesario un cambio en la red, pero si al comparar las respuestas son diferentes, entonces los pesos se ajustan para intentar de que la red obtenga mejor resultado posible si recibe una entrada de datos similar. Esta supervisión del resultado le su nombre a este tipo de entrenamiento.

En el caso del aprendizaje no supervisado sólo se le provee la data de entrada haciendo que la red se auto organice o que aprenda sola dependiendo de la estructura de esta data de entrada, es decir ajuste los pesos de la red hasta que encuentre algún tipo de redundancia en la data de entrada o en sub conjuntos de ella.

3.6 Los usos más frecuentes de técnicas de minería de datos aplicadas a las finanzas

Dada la gran cantidad de información generada por los mercados financieros, así como su calidad y frecuencia muchos autores comenzaron a aplicar las técnicas de minería de datos a problemas financieros clásicos dada su capacidad para manejar las complejas relaciones no lineales existentes entre las variables, la estacionalidad o la presencia de quiebres estructurales.

En la literatura es posible encontrar una gran cantidad de trabajos de minería de datos aplicados a finanzas, sin embargo los temas abordados son más limitados y se pueden resumir de la siguiente manera:

3.6.1. Predicción del mercado accionario: En general las técnicas de regresión utilizadas en este tipo de problemas estaban limitadas a la captura de relaciones lineales entre las variables seleccionadas. En el trabajo del año 2011 el autor Soni, S. [76] se realiza una extensa revisión de la literatura reciente respecto de las técnicas de *machine learning* para la predicción de mercados accionarios concluyendo que la técnica predominante son las redes neuronales (RN). Las razones que esgrime este autor son que las RN son capaces de encontrar la relación entre la variable

dependiente y la independiente incluso si estas son altamente complejas gracias a la utilización de funciones de aproximación siendo y a la existencia de una gran cantidad de data de entrenamiento. Otra razón que esgrime el autor, es el hecho de que las RN tienen la capacidad de la “generalización” en la que después del entrenamiento las neuronas tienen la capacidad de identificar nuevos patrones incluso si estos estuvieron ausentes en la data de utilizada en su entrenamiento, lo que las hace una buena herramienta proyectiva. Finalmente el autor explica que las RN son consideradas funciones de aproximación general y está probado que las RN *Multilayer Perceptron* (MLP) son capaces de aproximar cualquier función continua permitiendo aprender de la relación entre la variables independiente de su nivel de complejidad.

En 2010 los autores Dase, R., y Pawar, D (2010) [22] también realizaron una revisión bibliográfica de este tema exponiendo la gran cantidad de trabajos al respecto y la indudable capacidad de las redes neuronales para encontrar patrones no lineales en la variables destacando el éxito de algunos trabajos en la predicción del signo de algunos índices accionarios. Un buen ejemplo de lo anterior es el trabajo de Enke, D., & Thawornwong, S. (2005) [29] donde predicen el signo del movimiento del índice S&P 500 con RN. Para esto utilizaron variables económicas y financieras como los dividendos por acción de cada mes para el S&P, el índice de inflación (CPI), el índice de producción industrial (IP), el *yield* del *t-bill* a diferentes plazos y la cantidad de dinero en la economía (M1) por nombrar algunas de las 31 variables utilizadas en un periodo de tiempo que va desde enero de 1976 hasta diciembre de 1999. Dado las distintas frecuencias de la información algunas variables como el M1 o el CPI tenían dos meses de desfase. Posteriormente, se utilizó un criterio de “ganancia de información” (Quinlan 1993) [71] para seleccionar cuales de esas variables contienen mayor información predictiva y utilizarlas como entrada a la RN. De las 31 variables iniciales sólo 15 quedaron seleccionadas como entradas a los modelos de RN las cuales fueron normalizadas con los valores +1/-1 tratando de minimizar el efecto de la diferencia de magnitud de las diferentes variables y también aumentar la efectividad del algoritmo de aprendizaje. Una de las conclusiones más importantes de este trabajo es que si bien en general la mayoría de las proyecciones financieras consisten en estimar de manera exacta el precio de un activo, los autores sugieren que la unión de una estrategia de compra y venta guiada por una estimación de la dirección del cambio en el precio puede ser más efectivo y rentable, mencionando una serie de estudios que lo avalan como los de Aggarwal and Demaskey (1997) [3], Maberly (1986) [58] y Wu and Zhang (1997) [85] entre otros.

3.6.2 Detección de fraudes: Las técnicas de minería de datos son ampliamente usadas en la detección de fraudes en las tarjetas de crédito o el lavado de dinero, su ventajas en la detección de

patrones inusuales en grandes volúmenes de información y en una limitada cantidad de tiempo explican el gran desarrollo de esta área. En general las técnicas utilizadas en la detección de fraudes tienen que lidiar con información donde la información legítima es mucho mayor a la fraudulenta y donde la información útil es escasa y se diferencia de la data habitual por lo que se utiliza el análisis de *outliers*. En el trabajo de Sharma, A., & Panigrahi, P. K. (2012). [74] se realiza una completa y extensa bibliografía de los trabajos realizados al respecto revelando que las técnicas de minería de datos más utilizadas son las Redes Neuronales, Modelos de regresión, *Fuzzy Logic* y sistemas experto en conjunto con algoritmos genéticos. En el Anexo II es posible encontrar las tablas extraídas de este trabajo con el detalle de cada uno de los trabajos en este tema, su objetivo principal y su respectiva referencia al trabajo original que se encuentra en la bibliografía de este trabajo.

3.6.3 Predicción mercado de divisas: Este mercado corresponde a uno de los más líquidos del mundo, destacando por ser un buffer o regulador de los términos de intercambio entre las economías. En general los trabajos se están más enfocados a estimaciones de corto plazo sobre la base teórica del análisis técnico observándose buenos resultados mu lo puede estar explicado por la existencia de *volatility clusters* que consisten en la tendencia a periodos persistentes de alta o baja volatilidad en el tiempo. En el trabajo de Garg A. (2012) [35] el autor utiliza una mezcla de técnicas de minería de datos como los bosques aleatorios y los árboles de regresión junto a modelos econométricos GARCH para modelar los cambios en la volatilidad. Los resultados de este trabajo muestran una capacidad de predicción limitada pero superior a un AR(1), especialmente para horizontes más largos. Por otra parte, en el trabajo de Peramunetilleke, D., & Wong, R. K. (2002) [66] los autores realizan la estimación utilizando técnicas de *text mining*, capturando los encabezados de las noticias en tiempo real, evaluando su impacto y prediciendo el movimiento intra-día de la paridad, a diferencia del resto de los estudios, aquí no sólo se estudian los efectos sobre la paridad sino que también los posibles significado de este movimiento. Este estudio reporta un performance superior a un camino aleatorio e incluso mejor a uno de redes neuronales.

3.6.4 Administración de Portafolios: La teoría moderna de portafolio intenta maximizar el retorno y minimizar el riesgo de una cartera (Markowitz, Harry M. 1959) [59]. Lo anterior, junto a los modelos de valoración de activos como el *capital asset pricing model* (CAPM) o el *arbitrage pricing theory* (APT) son la base para la optimización de un portafolio de inversión. Los últimos años se han incorporado técnicas de minería de datos a este proceso de optimización. Un ejemplo de lo anterior es el trabajo de iu, K. C., & Xu, L. (2003) [42] en el que aplican una optimización dinámica de portafolios, es decir una optimización que va tomando los últimos precios del mercado

para calcular los pesos óptimos dentro del portafolio. La técnica utilizada en este trabajo es llamada *temporal factor analysis* (TFA) que consiste una forma alternativa de implementar APT considerando factores escondidos que si afectan al portafolio y por lo tanto a los pesos de los activos dentro de él sobre la base de ya no optimizar la frontera eficiente en términos de la media y varianza si no que optimizando el *Sharpe ratio*. Otro trabajo a destacar es el de Chapados (2010) [17] donde se muestra la transformación del problema Markoviano de optimización de portafolios en un proceso de aprendizaje supervisado a través del algoritmo de búsqueda *K-path* y *RN* optimizando también el *sharpe ratio*.

3.7 Trabajos previos de estimación de tasas de interés con herramientas de minería de datos

Si bien la bibliografía muestra que son otras las temáticas financieras que utilizan más frecuentemente herramientas de minería de datos, existen trabajos previos para modelar y predecir tasas de interés. La razón de esto radica en la importancia de esta variable en la economía como herramienta de política monetaria, señalizador de riesgo y expectativas futuras o simplemente como el precio del dinero. Otra razón que los explica es la falla de los modelos estadísticos lineales para predecir la tasa, donde la minería de datos se asoma como una alternativa para solucionar este problema. En general, las técnicas más utilizadas en estos estudios corresponden a las redes neuronales (RN), *Support Vector Machines* (SVM) y razonamiento basado en casos (CBR) donde se comparan las distintas metodologías separando una muestra de entrenamiento, para construir el modelo y otra de prueba para probar su rendimiento.

En el trabajo de Joo y Han (2000) [44] se utilizan RN para estimar la tasa de interés en Estados Unidos, considerando el hecho de lo que los autores llaman “puntos de cambio” de la serie, que ellos explican cómo cambios estructurales producto de la aplicación de la política monetaria por parte del gobierno. Es por lo anterior que en una primera etapa se dedican a identificar estos puntos de cambio a través del test estadístico de Pettitt (1979) [69] de manera de obtener una segmentación que en la segunda etapa, se utilizará para entrenar redes diferentes de manera que le sea más fácil encontrar los patrones relevantes que ayuden a predecir la tasa de interés. Las variables utilizadas en este trabajo fueron seleccionadas por estrecha relación teórica con la tasa representada por la ecuación de Fisher, que incluye la cantidad de dinero, la variación de la inflación, el interés real esperado y la producción industrial. Los autores argumentan que la combinación de la detección de los puntos cambio con RN es estadísticamente superior a las RN puras.

En el trabajo de Jacovides, (2008) [43] utiliza y compara RN y SVM en la estimación a seis meses de la tasa de interés para el Reino Unido (UK) a diferentes plazos. La variable utilizada para predecir corresponde íntegramente a la variable objetivo el día de hoy argumentando que es la facilidad de su obtención (UK es un mercado altamente desarrollado) versus la dificultad en la obtención de variables macroeconómicas. Los resultados muestran una buena capacidad predictiva, pero menciona que mucho mejor aún es la capacidad predictiva de la dirección de la tasa con tasas de acierto de hasta un 75%. En términos comparativos, el performance de SVM es superior al de RN.

Otro trabajo interesante en este tema es el de Kim y Noh (1997) [49] en la que los autores utilizan modelos sobre la base RN y razonamiento basado en casos (CBR) de manera independiente y en conjunto. La técnica CBR consiste en que el sistema va buscando casos de las variables en que el resultado ha sido el mismo (o parecido) y ante cada diferencia va corrigiendo levemente la combinación de las variables explicativas, lo relevante de lo anterior, es que eventualmente esta técnica reduce las combinaciones de entrada a la red neuronal si las dos técnicas son usadas en conjunto, por un lado, mejorando el rendimiento de la red y por el otro entregando una explicación al patrón encontrado (a través del CBR). El ejercicio involucró la incorporación de variables económicas tales como el índice de inflación, el de producción industrial, cantidad de dinero, permisos de construcción e índices accionarios con data mensual desde 1981 a 1989. Lo anterior se aplicó a la economía de EEUU y Korea con resultados disímiles. Mientras en EEUU el rendimiento de RN en conjunto con CBR supera a un *random walk*, ninguno de los modelos lo hace de manera estadísticamente significativa para Korea.

En el trabajo de Zimmermann (2002) [88] se utilizan tres variaciones de las redes neuronales para estimar la curva de rendimiento de Alemania en 3 y 6 meses. Los modelos utilizados son las RN convencionales, La RN con corrección de errores y la RN de tres capas optimizada con un proceso de podados para extraer las estructuras que no varían en el tiempo y por lo tanto no ayudan a predecir la curva. Para contrastar sus resultados los autores realizan un ejercicio que consiste en una estrategia de compra y venta en base a las estimaciones de cada uno de los modelos, mostrando su retorno sobre la inversión (ROI). Los resultados muestran que si es posible optimizar el rendimiento de las RN, especialmente en el caso con corrección de errores.

Otro trabajo que destaca por ser una aproximación diferente del tema corresponde al trabajo de Hong y Han (2002) [39] en el que utilizan información de noticias desde un servidor en internet las cuales a través de un sistema experto que consiste en una serie de reglas incluidas en una “base de conocimiento” que discrimina si la noticia es positiva o negativa, sirviendo de entrada a una RN que

incluye adicionalmente variables económicas y financieras como la cantidad de dinero, crecimiento económico, tasa de política monetaria, tasa de desempleo y precios de materias primas. Los resultados muestran una mejora del rendimiento respecto de una RN convencional y argumentan la importancia de presentar una explicación causal de cada noticia en el mercado.

En el trabajo de Vela (2013) [20] se estiman las curvas de rendimiento para diferentes países de latino américa y EEUU, incluyendo Chile para después realizar una proyección utilizando métodos econométricos y redes neuronales a diferentes plazos entre 1 y 6 meses hacia adelante. En general el método de ajuste a la curva con el que se obtuvieron los mejores resultados de estimación fue con Nelson y Siegel, excepto para EEUU. En términos comparativos, las RN presentaron un rendimiento superior al resto de los modelos, sin embargo para algunas curvas como la chilena y la colombiana las RN no superaron al modelo *random walk* para algunos plazos. Por último el autor destaca la capacidad predictiva de las RN para predecir la parte corta (hasta tres meses) de la curva chilena.

Finalmente, en el trabajo de Muñoz, Moreno (2014) [61] se proyecta la tasa de interés con diferentes herramientas centrándose en los SVM, RN y aboles de decisión. En sus conclusiones, los autores destacan a los arboles de decisión argumentando que permiten describir la reglas que explican el resultado de la variable proyectada versus los otros modelos que son “cajas negras”. En términos de rendimiento, el modelo que obtiene mejores resultados son la RN, sin embargo el porcentaje de aciertos alcanza apenas el 53%.

Los trabajos anteriores resumen y caracterizan el trabajo realizado en este tema mostrando en general una éxito respecto de las metodologías convencional utilizadas para predecir la tasa de interés.

3.8 Propuesta de investigación, motivaciones y objetivos

A la luz de la información expuesta en esta sección, a continuación se exponen las razones que justifican este trabajo junto a las brechas que esta investigación permitirá cerrar respecto de lo que actualmente se ha realizado en este tema. Las motivaciones para un trabajo sobre la proyección de tasas de interés con herramientas de minería de datos son:

- a.- Es posible observar que el mercado chileno se ha desarrollado enormemente en los últimos años de la mano de una mejor regulación, la presencia activa de grandes participantes e instituciones sólidas. Sin embargo lo anterior el mercado de renta fija es aún

un mercado en desarrollo resultando de mucha relevancia poder proyectar las tasas de interés entendiendo la dinámica de las variables económico financieras que la explican.

b.- A la luz de los trabajos realizados en la proyección de la tasa de interés con herramientas de minería de datos se puede desprender que si bien los trabajos en general son exitosos en superar los modelos de camino aleatorio, estos no obtienen un rendimiento excepcional. Sin embargo lo anterior, cuando estas herramientas se utilizan en combinación con conocimiento teórico previo, el rendimiento de los modelos mejora enormemente (Hong y Han (2002) [39], Kim y Noh (1997) [49]).

c.- Las metodologías más utilizadas como RN o SVM, no son capaces de explicar sus resultados con fundamentos que hagan sentido, teniendo un problema fundamental al momento de argumentar su uso, aun cuando el rendimiento sea de ellos sea superior a las metodologías convencionales.

d.- Los estudios en general se enfocan en la proyección del valor exacto de la tasa de interés, enfocándose en reducir los errores con respecto del valor real con el costo en la predicción correcta del signo del movimiento. En general cuando se utilizan estos modelos para invertir o utilizar un escenario probable más que el valor exacto de la variable proyectada interesa si esta subirá o bajará y si se espera que ese movimiento sea importante o no.

e.- Son pocos los trabajos realizados sobre la estimación de tasas de interés con herramientas de minería de datos (Vela (2013) [20] y Muñoz, Moreno 2014 [61]) en Chile y menos que utilicen arboles de decisión como herramienta proyectiva la que cuenta con la ventaja de poder mostrar las reglas en base a las variables explicativas con que se obtiene el valor.

f.- En Chile, las Instituciones como el Banco Central de Chile (BCCh), la Tesorería General de La Republica (TGR) junto a la información de mercado secundario producen y entregan de manera oportuna una gran cantidad de información económica financiera la que es suficiente para alimentar modelos intensivos en el uso de datos, lo que sumado a las capacidades de software y hardware actuales hacen factible su implementación.

De las motivaciones y razones que justifican este trabajo nacen las hipótesis y objetivos de este trabajo que su vez permitirán cerrar las brechas respecto de lo realizado hasta hoy de manera se ser

un aporte al tema es cuestión. El objetivo de este trabajo es responder las siguientes interrogantes de investigación:

- i. ¿Cuáles son las variables económica-financieras más importantes que ayudan a predecir la tasa de interés en Chile?
- ii. ¿Qué variables son las que afectan a la tendencia de la tasa de interés y cuales más bien su movimiento de corto plazo?
- iii. ¿Cambia la relación entre las variables dependiendo de la coyuntura económica que en que se encuentra el país y como esto afecta la estimación de la tasa de interés?
- iv. ¿Qué tan bueno es el rendimiento de los árboles de decisión como herramienta predictiva de la tasa de interés en comparación a otras herramientas de minería de datos? ¿Cómo lo es respecto de un modelo simple de regresión? ¿Es estadísticamente significativa esta diferencia?
- v. ¿Es posible mejorar el rendimiento de los arboles a través de la aplicación de ajustes al modelo en base al conocimiento previo del mercado chileno? ¿Existen sinergias al combinar estas dos herramientas?
- vi. ¿Cómo funciona el modelo proyectivo en una simulación de estrategia de inversión de portafolio?

Para contestar estas preguntas, en la siguiente sección se describe paso a paso las tareas y desarrollos para contestar estas interrogantes.

4 Metodología

Esta sección expone las metodologías utilizadas en la proyección de las tasas cero cupón utilizando minería de datos para el mercado chileno. En la implementación también se fundamenta la elección de los modelos, supuestos y dificultades encontradas en el proceso de manera de mostrar las ventajas y desventajas de la utilización versus otras metodologías de estimación. Para lo anterior se siguió la metodología CRISP-DM descrita en el marco teórico.

El primer paso de la metodología CRISP-DM corresponde a la **comprensión del negocio y el planteamiento del problema**. Ambos aspectos ya han sido cubiertos en la sección anterior, el primero con la breve descripción del mercado de renta fija en Chile junto a una revisión de los trabajos realizados en minería de datos y la proyección de tasas de interés. Adicionalmente, se definen las motivaciones y objetivos del presente trabajo cubriendo el planteamiento del problema.

El segundo paso de la metodología involucra la **recolección y comprensión de la data** en el que se analizan las variables desde un enfoque teórico, estadístico y descriptivo con el objetivo de comprender los datos y poder darle respuesta a las preguntas relacionadas con las variables explicativas de la tasa de interés. Adicionalmente, se realiza un análisis de conglomerados sobre variables económicas relevantes para el mercado chileno para modelar una variable representativa del ciclo económico que llamaremos “escenario económico”. Para términos comparativos, se intenta modelar la misma variable sobre la base de discriminar el ciclo económico identificando los eventos relevantes en la economía durante el periodo de estudio.

Una vez generadas todas las variables relevantes para la modelación, en el tercer paso corresponde al de **preparación de la data** en el que transforma y modifican las variables de manera de permitir su entrada en el modelo y posteriormente mejorar su rendimiento. Las transformaciones de las variables incluyen su discretización, normalización y rezago. Por otro lado el perfeccionamiento del rendimiento del modelo es un proceso iterativo en el que involucra agregar o quitar variables o rezagos de ella y modificaciones a la transformación de las variables.

La cuarta etapa del proceso corresponde al **modelamiento** que es donde definen y diseñan las distintas herramientas o modelos que se correrán con la data. En esta etapa se definen las muestras de entrenamiento y prueba, también se optimizan los modelos tratando de identificar los elementos relevantes que más impactan en su rendimiento. Es importante mencionar que en términos de contrastar la utilidad de algunas variables como el “escenario económico” se han diseñado modelos que la incluyen y la versión que no dejando todo lo demás constante.

En la cuarta etapa del proceso se **evalúan** los resultados de los modelos sobre una misma muestra de prueba respecto de un *benchmark* consistente en un modelo autoregresivo AR(1), llamado también modelo sencillo (*naive*) de manera de contrastar si el modelo cuenta con algún poder de predicción. Adicionalmente los diferentes modelos son comparados, rankeados y se evalúa estadísticamente a través de test de errores de Diebold y Mariano si estas diferencias entre los modelos son significativas.

En la última y quinta etapa corresponde a la etapa de **explotación** en la cual se hacen competir los modelos en una simulación con el objetivo de dimensionar su real utilidad e identificar posibles problemas en su uso real.

A continuación se desarrolla en detalle cada una de las etapas descritas más arriba respondiendo cada una de las interrogantes planteadas en al final del marco teórico.

4.1 Recolección y comprensión de la data

La realidad actual de los mercados internacionales cada vez más globalizados y modernos ha derivado en el desarrollo de plataformas tecnológicas cada vez más avanzadas, con la capacidad de almacenar grandes cantidades de información disponibles incluso en un país en de mercados financieros poco desarrollado como el de Chile. Sus principales proveedores de la información económica son el Banco Central de Chile (BCCCh) y el Instituto Nacional de Estadísticas (INE) los cuales entregan el pulso de la actividad industrial, mercado laboral, cuantías nacionales y estadísticas monetarias y financieras que son entregadas en general con frecuencia mensual.

4.1.1 Descripción de las variables y su proceso de obtención

Lo primera tarea fue construcción de la tasa de interés a cero cupón a 5 años ya que esta se encuentra implícita en las transacciones de mercado de estos instrumentos. Para esto se utilizó el ajuste de curva de Nelson y Siegel (1987) [62] adaptado al mercado local, caracterizado por tener poca profundidad para algunos tramos de la curva y tipos de instrumento. Esta adaptación consistió en subdividir la curva en tramos de tres años y utilizar operaciones de días anteriores (Con rezago máximo de días hábiles) para completar los tramos sin transacciones mejorando de esta manera el ajuste de la curva.

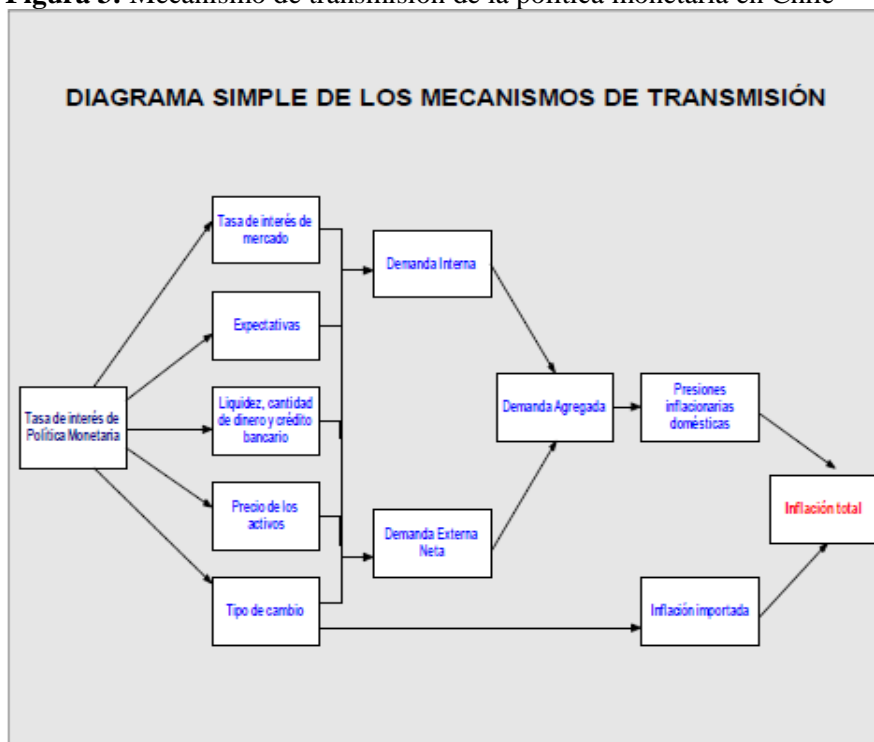
Para la proyección de las tasas utilizando DM es necesaria una base de datos que contenga todas aquellas variables que puedan tener una relación directa o indirecta con la curva de tasas libre de riesgo. Para esto, se realizó un estudio para determinar las variables económicas y financieras que podrían servir como la información de entrada a los modelos, basándose en fundamentos económicos y/o la evidencia de una significancia estadística que permita explicar o fundamentar la evolución de estas tasas. Es por esto que en la primera etapa del proceso se describen las variables y se fundamenta su relevancia en la proyección de la tasa. Las variables utilizadas junto a la justificación de su uso son:

Tasa de Política Monetaria (TPM): Corresponde a la tasa objetivo de la implementación de la política monetaria y que sirve de ancla y el nivel para todo el resto de las tasas de interés del mercado local. Para que esta tasa cumpla con su definición, el Banco Central influye sobre la tasa de mercado a la que los bancos se prestan entre sí (Tasa interbancaria o TAB) a través de operaciones de mercado abierto (OMA) que afectan la liquidez del mercado. La TPM es evaluada periódicamente en las reuniones de política y en sus ajustes son consideradas una serie de variables económicas relacionadas al objetivo último de controlar la inflación. Las reuniones de política monetaria generan expectativas en el mercado y movimientos de los agentes previo al posible cambio de la TPM que afectara de manera inmediata la tasa de corto plazo. En general el banco

central está constantemente generando y entregando una descripción de la coyuntura económica actual que justifica los posibles cambios de la TPM en sentido y magnitud por lo que esta variable puede tener una buena capacidad predictiva en la parte corta de la curva cero cupón y posteriormente al resto de la curva.

El siguiente diagrama resume de forma gráfica la manera en que se transmiten los distintos efectos ante un cambio en la política económica en la economía:

Figura 3: Mecanismo de transmisión de la política monetaria en Chile



Fuente: Política Monetaria del Banco Central de Chile: Objetivos y Transmisión.

Tasa del Tesoro de Estados Unidos a 10 Anos (T10Y): El nivel de tasas de principal economía del mundo está relacionada directamente con la curva de rendimientos en Chile principalmente porque no existen mayores restricciones al flujo de capitales de entrada ni salida al país y porque existe libre flotación cambiaria. Actualmente la T10Y presenta rendimientos históricamente bajos producto de una política monetaria expansiva y los programas de emisión de deuda para asegurar la liquidez (*Quantitative easing*), ambas medidas para afrontar la crisis sub-prime que estallo en 2008. Bajo este contexto existe evidencia de una mayor cantidad de flujo capitales extranjeros hacia países emergentes, especialmente aquellos con niveles de tasa de interés mayores. Si bien esta relación y efecto pudiese no ser explícito en el corto plazo, esta variable puede ayudar a identificar cambios de tendencia de la variable proyectada.

Índice selectivo de precios de acciones (IPSA): Este es el principal índice accionario chileno. Está compuesto por las 40 acciones de mayor presencia bursátil ponderadas en base a su capitalización y el número de acciones. La idea de incluir un índice accionario en la estimación tiene dos justificaciones. La primera tiene que ver con que la tendencia de IPSA refleja las expectativas económicas a futuro, si vemos a las empresas como el valor presente de sus proyectos futuros, la expectativa de buenos proyectos a mediano y largo plazo implica buenas perspectivas económicas y viceversa, dado que corresponde a un índice representativo del mercado, su evolución está más representado por el riesgo sistemático de la economía. La segunda justificación tiene que ver con que en general las acciones son un instrumento de inversión sustitutos con respecto a los bonos, es decir los agentes se mueven entre estos tipos de instrumentos para optimizar sus portafolios y aprovechar oportunidades de mercado (Fly to quality y Fly to risk).

Paridad Peso Dólar (CLP/USD): En Chile existe un régimen de flotación cambiaria en el cual la paridad se mueve libremente pero el Banco Central se reserva la posibilidad de intervención en situaciones de excepción. La razón de esto consiste en que si la paridad se mantiene fija esta pasa a ser el ancla para los precios y el nivel de inflación. Si los agentes especulan contra la moneda la autoridad debe estar dispuesta a comprar o vender cualquier cantidad de divisas perdiendo el control sobre la cantidad de dinero local y por lo tanto sobre la tasa de interés. De esta manera la autoridad pierde el control sobre la política monetaria. Por otro lado, un tipo de cambio flexible permite que la autoridad si pueda tener un control sobre la cantidad de dinero y por ende sobre la tasa de interés.

En Chile existe una gran volatilidad del tipo de cambio debido a que es un economía mayoritariamente exportadora y abierta haciendo que el tipo de cambio sea el regulador, sin embargo un dólar desalineado de su nivel de largo plazo puede ser perjudicial para la economía del país por lo que el Banco Central se reserva la posibilidad de intervenir el mercado, pero siempre emitiendo deuda para esterilizarla y no afectar la tasa de interés. La relación entre el tipo de cambio y las tasas de interés se remonta los estudios Frenkel (1981) [34] y Dornbusch (1978) [26] los cuales postulan que la relación entre estas variables viene de dos teorías: La paridad del poder de compra (PPP) y la paridad de intereses. La PPP consiste en que la paridad mantendrá el equilibrio del poder de compra de bienes y servicios en ambas economías, si sube la tasa nominal producto de expectativas de un nivel de inflación mayor en Chile que en Estados Unidos (la inflación implica que con el mismo dinero se puedan comprar menos cosas) se debería observar una depreciación del peso y viceversa. Por otro lado la paridad de intereses postula que una mayor tasa de interés relativa traerá mayores flujos de divisas al país en busca de mayores rendimientos lo que hará que el tipo de cambio reaccione apreciándose y viceversa.

Indicador Mensual de Actividad Económica (IMACEC): Como su nombre lo indica este corresponde a un indicador mensual que mide la actividad económica del país capturando la mayoría de las actividades productivas que conforman al cálculo del Producto Interno Bruto. Este indicador es visto como una estimación del PIB en el corto plazo, presenta problemas de volatilidad en algunas cuentas pero en general su evolución permite obtener una buena estimación del ritmo económico del país y por lo tanto es una de las variables esperadas por los agentes económicos para tomar decisiones. Una evaluación positiva de este indicador debiera ser indicador de una política monetaria más restrictiva, junto a la migración de capitales a instrumentos más riesgosos.

Unidad de Fomento (UF): Es una unidad de cuenta que se reajusta de acuerdo al Índice de Precios al Consumidor que de manera simple calcula cuánto varía el costo de una canasta representativa, con el fin de capturar el nivel de precios en la economía.

Swap Peso-Cámara UF y Peso (SPC-UF y SPC-CLP): Un swap es un instrumento financiero en la que dos contrapartes se intercambian flujos en fechas determinadas en el que uno paga una tasa fija y conocida al comienzo del contrato y el otro paga una tasa variable conocida al momento del flujo. En la práctica las dos contrapartes se traspasan el flujo neto en las fechas determinadas reduciendo de esta manera el riesgo de crédito de las contrapartes. Su función consiste principalmente en la transformación de un pasivo o inversión de renta fija a variable y viceversa.

En Chile estos instrumentos han tenido un fuerte crecimiento entre los bancos de la plaza desde el año 2002 a partir de la nominalización de la política monetaria, estas operaciones se realizan mayoritariamente *over the counter (OTC)*, existiendo alternativamente algunos *brokers* que ofrecen puntas. Los SPC son usados para cobertura y especulación pero también ha sido el vehículo a través del cual bancos y fondos extranjeros pueden exponerse a las tasas de interés locales sin restricciones tributarias ni legales. Su estructura consiste en una tasa swap fija contra una tasa variable que varía según el promedio de la Tasa interbancaria (TIB) sobre el mismo plazo y nominal. La TIB es computada por el Banco Central y corresponde al promedio de las tasas de mercado a un día a la cual los bancos se prestan entre sí. En el caso de los SPC-UF la tasa variable corresponde a la conversión en UF de la tasa.

La estructura de pagos de la parte fija de un SPC es muy similar a la de un BCP o BCU, lo que sumado su bajo a su bajo riesgo de crédito tiene un comportamiento muy parecido al de los bonos libres de riesgo BCP y BCU. El spread entre la tasa SPC y un BCP es conocida como swap spread en el que una expectativa de una alza de tasas implica que más agentes captaran recursos a tasa fija y pagaran variable lo que empujaría la tasa SPC hacia abajo, aumentando el spread. De manera

contraria, expectativas de baja de tasas incentivarán a los agentes comprar los bonos libres de riesgo y captarán SPC a tasa variable para colocar a tasa fija lo que haría bajar las dos tasas haciendo que el spread no varíe o baje. (Varela 2007)[81]

Colocaciones BCU y BCP en el Mercado Primario: El Banco Central publica anualmente su programa de emisión de deuda, sin embargo esta puede sufrir modificaciones producto de la coyuntura económica del momento, estas sorpresas afectan el nivel de tasas reales y nominales. Adicionalmente la relación entre la oferta y la demanda en cada colocación expresada por el cupo adjudicado versus la demanda de la licitación también tiene una relación directa en las tasas (Batarce 2009) [12].

Emerging Market Bond Index Chile (EMBI Chile): Este índice busca reflejar una estimación del costo de endeudamiento y percepción de riesgo de un país por parte de los inversionistas. Este índice es calculado por el banco de inversión JP Morgan y cubre una gran cantidad de países, de hecho Chile es uno de alrededor de 30 países que componen el índice global de países emergentes. La motivación para incluirlo se basa en que es un buen indicador del riesgo exigido por parte de los inversionistas de todo el mundo respecto de la deuda soberana, siendo bastante sensible a los flujos de capitales de inversión desde y hacia Chile, elemento relevante ya que desde el 2008 Estados Unidos comenzó a proveer liquidez a los mercados y se cree que buena parte de esta liquidez migró a países emergentes producto de las buenas condiciones crediticias, especialmente a Chile.

Colocaciones de Bancos en papeles BCCH y TGR: La superintendencia de Bancos e Instituciones financieras publica una agregación de una serie de estadísticas del sistema de bancos local entre ellas la cantidad invertida de los bancos en instrumentos del BCCH y TGR, lo que da una estimación de la demanda de papeles de un actor importante del mercado chileno. Aunque su frecuencia es mensual, esta variable muestra una tendencia en la composición de la cartera de los bancos respecto de los papeles libres de riesgo.

Finalmente, el siguiente cuadro de correlaciones sobre las variables continuas busca ilustrar el sentido y la magnitud de las relaciones en la data en uso. Se observa una alta relación entre las tasas swap a distintos plazos y una elevada correlación entre el IPSA, el EMBI y la unidad de fomento.

Tabla 5: Matriz de correlaciones variables económico-financieras 2006-2012

Matriz de Correlaciones	T10_USD	CLP	IPSA	Swap camara _1y	Swap camara _5y	Swap camara _10y	CLIMMOMS	TPM	PETROLEO	COBRE	EMBI_CL_CORP	EMBI_CL_SOB	CLIMSA	UF	CDS_CL_5Y	VIX	TIB	FPD	FPL	COMP_INF	CLP_5Y	CLP_5Y_BLP
T10_USD	1.00	0.19	-0.65	0.22	0.57	0.69	0.01	0.07	-0.25	-0.09	-0.88	-0.45	-0.87	-0.92	-0.63	-0.33	0.07	-0.50	-0.04	0.03	0.42	0.36
CLP		1.00	-0.57	-0.10	-0.11	-0.07	-0.08	0.05	-0.71	-0.85	-0.47	0.56	-0.47	-0.18	0.50	0.59	0.05	-0.03	-0.03	-0.17	-0.20	-0.22
IPSA			1.00	-0.25	-0.39	-0.43	0.08	-0.30	0.45	0.61	0.89	-0.12	0.86	0.74	0.09	-0.13	-0.30	0.44	0.05	0.03	-0.15	-0.13
Swap camara _1y				1.00	0.81	0.62	-0.11	0.93	0.35	0.19	-0.28	0.01	-0.07	-0.27	-0.20	-0.02	0.93	-0.24	0.11	0.33	0.69	0.77
Swap camara _5y					1.00	0.95	-0.09	0.61	0.37	0.23	-0.51	-0.25	-0.37	-0.50	-0.48	-0.20	0.60	-0.33	0.05	0.41	0.90	0.91
Swap camara _10y						1.00	-0.07	0.38	0.28	0.19	-0.57	-0.36	-0.50	-0.59	-0.55	-0.25	0.37	-0.35	0.02	0.37	0.88	0.86
CLIMMOMS							1.00	-0.17	0.01	0.06	0.05	-0.14	0.09	-0.03	-0.12	-0.04	-0.17	0.10	0.01	-0.04	-0.10	-0.11
TPM								1.00	0.21	0.01	-0.25	0.24	-0.03	-0.17	0.04	0.15	1.00	-0.18	0.09	0.22	0.44	0.55
PETROLEO									1.00	0.71	0.43	-0.17	0.52	0.36	-0.19	-0.22	0.21	0.14	0.05	0.44	0.51	0.53
COBRE										1.00	0.44	-0.56	0.48	0.19	-0.48	-0.53	0.01	0.07	0.04	0.19	0.33	0.35
EMBI_CL_CORP											1.00	0.08	0.96	0.91	0.32	0.02	-0.25	0.53	0.03	-0.01	-0.30	-0.27
EMBI_CL_SOB												1.00	0.14	0.45	0.91	0.86	0.24	0.23	0.04	0.08	-0.18	-0.16
CLIMSA													1.00	0.89	0.33	0.06	-0.03	0.05	0.06	-0.21	-0.15	
UF														1.00	0.64	0.37	-0.17	0.57	0.03	0.06	-0.28	-0.25
CDS_CL_5Y															1.00	0.83	0.03	0.38	0.03	-0.03	-0.37	-0.35
VIX																1.00	0.14	0.21	0.05	0.08	-0.11	-0.11
TIB																	1.00	-0.19	0.10	0.22	0.43	0.54
FPD																		1.00	-0.11	0.17	-0.22	-0.20
FPL																			1.00	0.13	0.08	0.09
COMP_INF																				1.00	0.50	0.50
CLP_5Y																					1.00	0.98
CLP_5Y_BLP																						1.00

Adicionalmente es posible encontrar una tabla de estadística descriptiva de cada una de las variables (Anexo I).

4.1.2 Comprensión de los datos y construcción de los escenarios económicos

En esta sección se analiza la data recopilada y se realiza un análisis descriptivo de las variables de manera de identificar una relación causal entre variables relevantes para el mercado chileno. Adicionalmente, esta información permite configurar una coyuntura económica identificando si la economía se encuentra en la parte del ciclo de crisis, debilitamiento o recuperación con el objetivo principal de responder las primeras tres preguntas de este trabajo. (Sección 3.8)

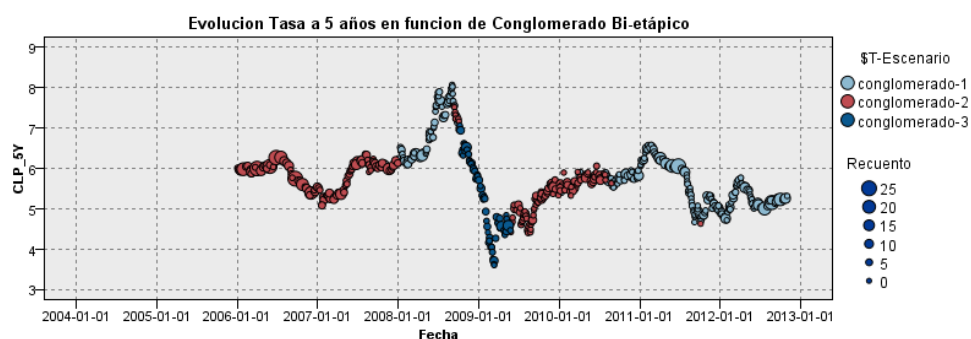
Tanto las variables económicas como las financieras cuentan con relaciones fundamentales entre sí, existen estudios que sugieren que el premio por plazos varía en el tiempo dependiendo de variables macroeconómicas, (ej Ludvigson and Ng(2009) [57], Barillas (2010) [11] and Joslin et. al. (2009) [45]) evidenciando que la dinámica de las tasas es distinta dependiendo del contexto económico predominante. Si vemos este argumento desde el punto de vista de comportamiento humano la teoría de la aversión al riesgo respalda este punto. Esta teoría plantea que no es de esperar que todos los individuos tengan las mismas actitudes frente al riesgo, sin embargo Shiller, Robert J. 1981 [75] encontró que en condiciones de neutralidad al riesgo, la volatilidad máxima de las acciones es

mucho menor que lo que se puede encontrar en el mercado y que ante un mayor nivel de aversión al riesgo, entonces mayor es la volatilidad de las acciones.

Del análisis de la data, es posible concluir que es que existen dos grupos de variables, las variables económicas y las variables financieras. Las variables económicas son aquellas que permiten identificar un escenario o coyuntura económica para Chile y en son variables de tendencia con frecuencia mensual que están menos dominadas por la volatilidad proveniente de eventos aislados. La fundamentación de esta separación se basa en la hipótesis de que la relación entre las variables relevantes para la estimación de la tasa cambia de dependiendo de si existe una situación de crisis, de debilitamiento o de fortalecimiento económico. Para contrastar la existencia de esta segregación de las variables se realizó un análisis descriptivo de las variables utilizando técnicas minería de datos buscando principalmente agrupar el tipo de relaciones que se da entre las variables en distintos periodos del tiempo. Específicamente, se realizó un análisis de conglomerados bi-etápico utilizando como función objetivo la tasa de 5 años. El resultado es la creación de tres conglomerados en los que si consideramos el nivel de las variables dentro de cada uno y la coincidencia en la ocurrencia con eventos conocidos como una crisis, o los máximos de la bolsa (recuperación) es posible categorizarlos y llamarlos escenarios de crisis, debilitamiento o recuperación.

En el grafico3 es posible observar la evolución de la variable “escenario” en función de la tasa a cinco años:

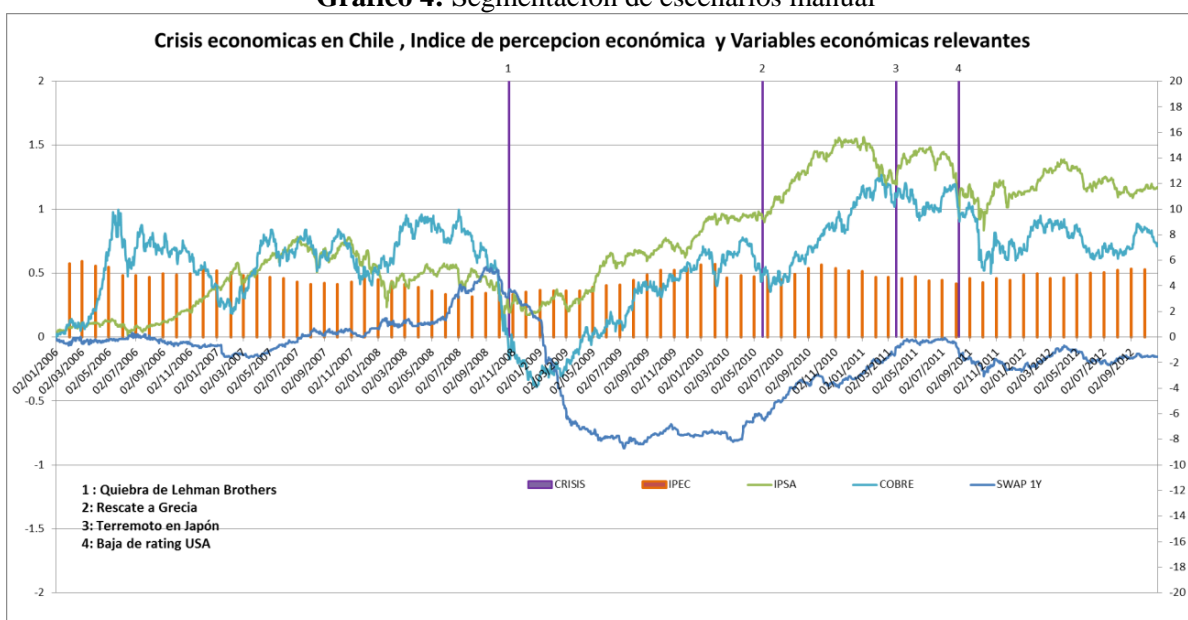
Grafico 3: Tasa cero cupón a 5 años y conglomerado al que pertenecen



El resultado anterior considera relaciones no lineales entre las variables, por lo que su comprensión puede presentar ciertas complicaciones, para esto se creó una versión “humana” de los escenarios construida sobre la base de la identificación de las crisis económico financieras en Chile desde el 2006, con sus respectivos periodos de debilitamiento y recuperación.

Si bien las crisis le dan el orden cronológico a las otras etapas, el problema mayor consiste en identificar cuando termina cada una de ellas para dar paso a la siguiente. Para esto se utilizó el Índice de Percepción Económica del BCCh (IPEC) que corresponde a un índice que captura la percepción económica a través de una encuesta a sus agentes más importantes. Son los cambios en la tendencia de este índice entre dos escenarios de crisis las que delimitan el largo de los escenarios. El siguiente gráfico muestra un detalle de las crisis económicas del periodo, la evolución del IPEC y tres de las variables económicas más importantes del mercado chileno.

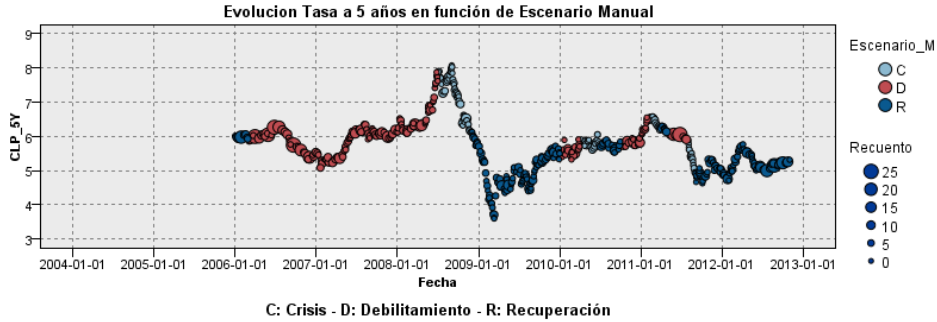
Gráfico 4: Segmentación de escenarios manual



En el gráfico anterior se puede observar que el IPEC tiene máximos locales después de cada periodo de crisis que marcan el término de un periodo de recuperación y el comienzo del periodo de debilitamiento hasta llegar a un nuevo periodo de crisis que por lo general son cortos y pronunciados.

Considerando las reglas de selección de escenario antes mencionado, la evolución de la tasa de cinco años en función de este ‘escenario manual’ queda representada de la forma:

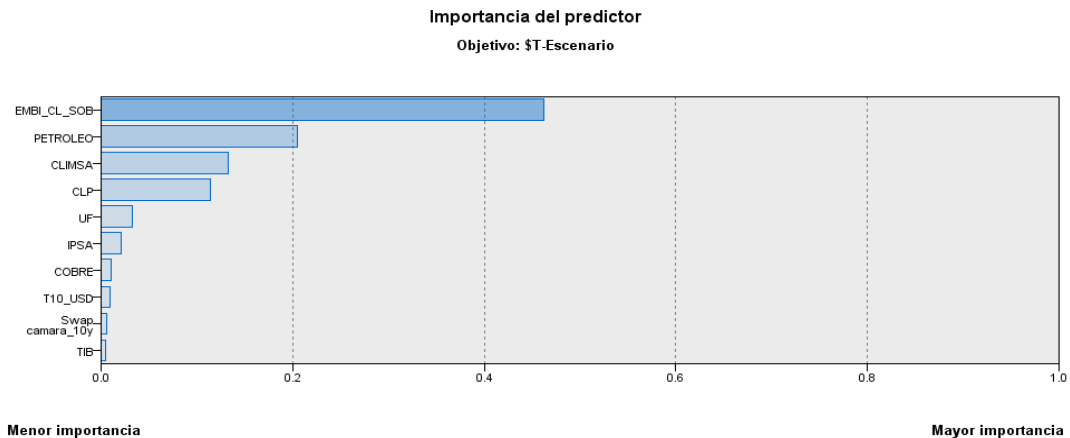
Grafico 5: Tasa cero cupón a 5 años y conglomerado manual al que pertenecen



Como podemos observar, para los periodos de mayor volatilidad o de crisis se ven reflejados contemporáneamente en tanto en el análisis de conglomerados como en los escenarios construidos manualmente, sin embargo los otros dos escenarios presentan diferencias entre las dos metodologías principalmente por el supuesto del término de un escenario asociado al cambio de la tendencia del IPEC. Otro elemento a tomar en cuenta es que al parecer la crisis asociada al terremoto en Japón no fue tan relevante para el variable objetivo en el caso del análisis de conglomerado observándose que a partir de la crisis asociada al techo de la deuda en USA no hay más variaciones del escenario quedando en fijo en debilitamiento.

Para analizar que variables son más relevantes en los escenarios creados por el análisis de conglomerados se aplicó un árbol de decisión que muestra un camino lógico de como las variables van explicado el escenario actual y como a partir de ciertos umbrales de esas variables las relaciones entre ellas cambian haciéndose relevantes entrando nuevas variables en la explicación.

Figura 4: Importancia del predictor del conglomerado



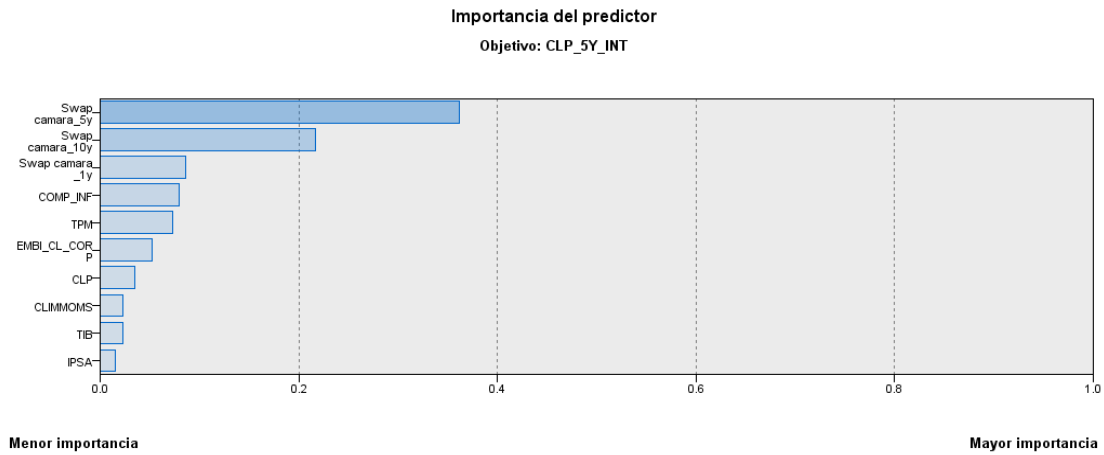
El resultado del árbol, muestra que las principales variables que explican el escenario corresponden a al grupo de las variables económicas tales con el índice soberano EMBI en primer lugar también son relevantes el precio del petróleo y el IMACEC (CLIMSA), quedando más atrás el peso, el IPSA y el precio del cobre.

El EMBI soberano de para Chile está construido refleja las expectativas de los inversionistas respecto del premio por riesgo de la economía Chilena por parte de los inversionistas del resto del mundo, de hecho es un conocido proxy del riesgo país. Una característica importante de este indicador es que resume no sólo las condiciones económicas internas del país si no mejor aún mide el impacto directo de las crisis económicas internacionales. La gran mayoría de las crisis afrontadas por Chile en las últimas dos décadas han sido originadas en los mercados internacionales y la peor fue la crisis asiática de 1998. (Massad 1998 [60])

El IMACEC corresponde a una variable que mide la actividad industrial interna por lo que su tendencia ayuda a definir un escenario. En el caso del dólar es más difícil de explicar sin embargo es relevante mencionar que Chile es un país con tratados de comercio con casi todos los países del mundo y presenta un sistema de libre flotación. En general se confirma un dólar apreciado frente a los escenarios de crisis por ser una moneda refugio y por el cierre de las líneas que sufrieron los bancos de la plaza como el la crisis sub prime del 2008. Por otro lado un dólar depreciado tiende a estimular las importaciones y en consecuencia el consumo, lo que se asocia con un escenario más positivo hasta que esta alza en el consumo no sea un motivo del BCCh para subir la tasa. En sentido opuesto Chile en su calidad de exportador se ve afectado por un dólar depreciado por lo que un dólar bajo debiera estar asociado a un escenario negativo. Con lo anterior se hace un poco complicado utilizar el dólar como una variable que discrimine el escenario, especialmente cuando no se está en una situación de crisis o de crecimiento marcado.

En el caso del precio del cobre está asociado al ingreso de la parte más importante de la producción industrial del país y del gobierno, impactando directamente en la estimación del gasto fiscal. Lo anterior implica que un precio del cobre alto debiera estar asociado con un escenario favorable para Chile y viceversa. Ahora bien, para contestar la pregunta de cuáles son las variables que más influyen en la tasa de interés construimos un árbol de decisión descriptivo utilizando como variable objetivo a la tasa de interés a cinco años y se obtiene el siguiente resultado:

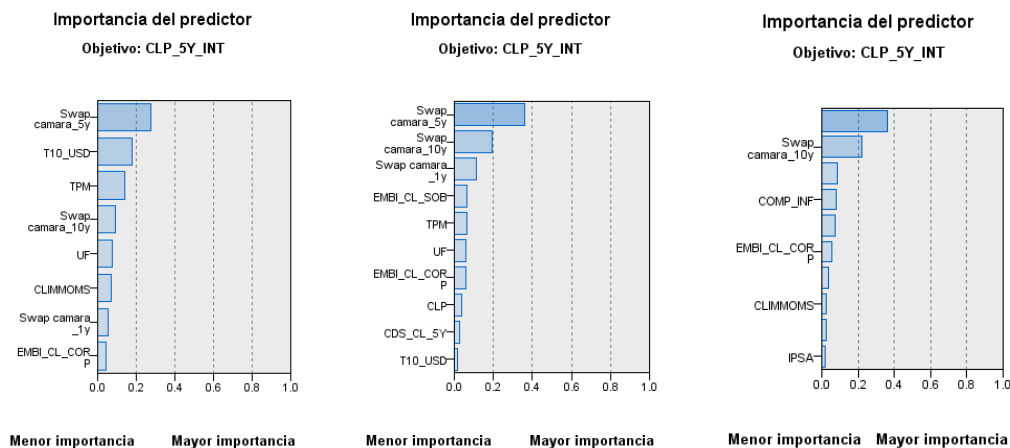
Figura 5: Importancia del predictor de la tasa a 5 años



Es posible observar en este caso que las variables que explican mejor a la variable objetivo son aquellas del grupo de las variables de mercado, destacando la tasa swap promedio cámara que está muy relacionada con la tasa libre de riesgo del mismo plazo.

La capacidad explicativa de estos dos grupos de variables evidencia que si bien ambos grupos sirven en la estimación de la tasa de 5 años, el grupo de variables económicas entrega un contexto económico donde las variables de mercado interactúan de manera distinta, para ilustrar esto se construyó un árbol descriptivo con la tasa a 5 años como objetivo pero esta vez se segmentó la data para crear un árbol por cada escenario generado por el análisis de conglomerado:

Figura 6: Importancia del predictor de la tasa a 5 años por conglomerado



Lo que se observa es que dentro de cada conglomerado el grupo de variables económicas explica bastante menos que las del grupo de mercado, pero más importante aún, la importancia de las

variables de mercado para explicar la tasa de 5 años es distinta para cada escenario. Por ejemplo en el caso del conglomerado 1 que se corresponde con el escenario manual de “debilitamiento”, entra a jugar un rol importante en la tasa de política monetaria, ya que es en este momento cuando se generan las mayores expectativas de ajuste, haciéndose relevante para toda la curva libre de riesgo. Lo nos da un argumento para responder que variables afectan en el corto plazo a la variables y cuales su tendencia de largo plazo y adicionalmente reafirma que la relación entre las variables cambia dependiendo del contexto económico en que se encuentra la economía.

4.2 Preparación de la Data

El foco principal de esta fase corresponde a la preparación y/o conversión de la data previamente seleccionada y comprendida para su uso como entrada en la sección de modelamiento. En el caso específico de este trabajo dado que el objetivo es la proyección a 5 días de la tasa en pesos a 5 años con árboles de decisión, la preparación incluye:

- 1.- Eliminación días no hábiles: En general la data estaba en un formato en la que se repetía para los fines de semana el último dato hábil, excepto para la UF. Se optó por remover los días no hábiles de la serie, ya que podrían influir en el rendimiento del modelo.
- 2.- Rezago de las variables financieras: En general estas variables presenta una frecuencia diaria por lo que se rezago tres veces cada variable con un rezago de 5, 6 y 7 días hábiles.
- 3.- Rezago variables económicas: En este caso dado su complejidad de cálculo y lo agregado de la estadística lo habitual es que estas variables si bien cuenten con una frecuencia diaria, estas estén disponibles con un mes de desfase. Dado lo anterior se rezago tres veces cada variable con un rezago 20, 21 y 22 días hábiles. Lo anterior asegura una aplicación realista del modelo.
- 4.- Conversión variable objetivo en discreta: Uno de los requisitos de los árboles de decisión es que su precisión es sobre variables discretas, por lo que en este caso se crean una nuevas variables objetivo creadas en base a su división en rangos bajo distintos criterios entre los que están un rango fijo, desviaciones respecto a la media, cuartiles o deciles (contienen igual frecuencia) y rangos preestablecidos producto por ejemplo de una decisión táctica de inversión.

5.- Creación de las variables Escenario y Escenario Manual: Estas dos variables proviene de un análisis y pre procesamiento de la data utilizando como entrada para la sección de modelamiento de los árboles.

4.3 Modelado: Sección predictiva de la tasa a cinco años

La proyección de la tasa a cinco años involucra una serie de decisiones que impactarán directamente en la capacidad predictiva del modelo. Las principales decisiones tienen que ver con el tipo de metodología utilizada para predecir y la preparación y/o transformación de las variables tanto dependientes como independientes.

Los paquetes informáticos de minería de datos actuales cuentan con una serie de modelos que se pueden utilizar para proyectar, entre los más utilizados están las redes neuronales, los árboles de decisión, regresiones, reglas de asociación, redes bayesianas, etc. El uso de un modelo u otro depende del tipo de información que se está utilizando, el tipo de resultado buscado y la calidad de la proyección obtenida. En el caso de esta tesis, el interés se focaliza en la idea de la relación entre las variables financieras y económicas en Chile cambia dependiendo del contexto económico por el cual atraviesa la economía, siendo relevante una herramienta de proyección que muestre con detalle que variables está utilizando para la estimación y si sus diferentes niveles cambian su relación entre ellas y con otras variables que pasan a ser relevantes en la estimación.

La herramienta que cumple mejor estos requisitos son los árboles de decisión ya que muestran de manera muy gráfica todo el camino que sigue la variable estimada a través de las variables explicativas y su nivel hasta su estimación. Otra ventaja de su uso es que el árbol resultante puede ser traducido en un conjunto de “reglas” en las que se basan en umbrales, los cuales nos pueden dar información sobre el punto en que una variable comienza a relacionarse distinto con otras variables. Una desventaja de estos modelos es que resulta necesario llevar la variable objetivo a un plano discreto y la decisión de cómo realizar esto puede afectar el grado de predicción del modelo, sin embargo es necesario entender que este tipo de modelos está diseñado para responder preguntas como si la variable subirá o bajará en cierto rango a cierto plazo, versus el resto de los modelos que intentan predecir el valor exacto de la variable estimada. Lo anterior hace un poco difícil y arbitraria la comparación entre el rendimiento de los distintos modelos.

El segundo aspecto relevante a considerar es la preparación y/o modificaciones de las variables que servirán de entrada al modelo. La información recolectada incluye una historia de 6 años con frecuencia diaria, entre el 01 de enero de 2006 y el 31 de octubre de 2012. La mayoría de las variables económicas tiene una frecuencia mensual por lo que se completó el resto del mes con el

último valor disponible, adicionalmente se le dio el mismo tratamiento a todas las variables con valores diarios faltantes.

En general para proyectar una variable con las herramientas de minería de datos lo que se hace es rezagar las variables independientes de manera que la estimación de la variable dependiente en el tiempo “t” quede modelada en función de valores pasados de ellas. En el caso de este estudio se realizaron tres rezagos de cinco días hábiles para todas las variables con frecuencia diaria y tres rezagos de veinte días hábiles de las variables con frecuencia mensual, es decir, que cada variable explicativa se multiplicó por tres. Con lo anterior se asegura no perder ninguna posible relación desfasada entre las variables y/o que la variable explicada dependa de un combinación de varios rezagos de o las variables independientes, el aprovechamiento de estas características es una de las ventajas de los modelos de árbol que es su capacidad para diferenciar la data útil sin el riesgo de sobre parametrizar el modelo.

Para la aplicación de los distintos modelos es necesario separar la data en un conjunto de entrenamiento y otra de comprobación. En el primer conjunto se utiliza para entrenar el modelo y obtener sus parámetros, el segundo conjunto se utiliza para medir el poder predictivo del modelo. Es relevante realizar esta separación ya que permite realizar una evaluación instantánea del modelo permitiendo su mejora constante.

La siguiente tabla resume los modelos estudiados en este trabajo con sus respectivos modelos, etapas y transformación a las variables:

Tabla 6: Modelos Estudio

N	Modelo	Variables Independientes	Variable Dependiente
1	Conglomerado-Árbol	Todas a la vez-Escenario y Financieras-tasa	CLP 5 Y Discreta, 6 cajas.
2	Árbol	Manual-Escenario y Financieras-tasa	CLP 5 Y Discreta, 6 cajas.
3	Árbol	Todas a la vez	CLP 5 Y Discreta, 6 cajas.
4	Red Neuronal	Todas a la vez	CLP 5 Continua
5	Modelo ARIMA	CLP 5 rezagada	CLP 5 Continua

En el modelo 2 se utiliza la variable “escenario-manual” creada manualmente para segmentar los periodos o escenarios y sobre esa segmentación teórica se modela un árbol para cada escenario. Este modelo fue creado para evaluar en qué medida cambia la capacidad de predicción si el escenario es asignado arbitrariamente en base a una regla o una percepción específica como la del autor. El modelo 3 elimina las etapas y utiliza tanto las variables económicas como las financieras rezagadas

para estimar la tasa, la idea es testear nuevamente si un solo árbol con todas las variables entrega una mejor predicción que los arboles por escenario.

En el modelo 4 se utiliza un modelo de redes neuronales para la estimación de la tasa con el objetivo de comparar el rendimiento de esta estimación contra la de los modelos de árboles. A diferencia de la estimación por árboles, las redes neuronales pueden estimar el valor exacto de la variable objetivo y no sólo una transformación discreta como en el caso de los árboles. Sin embargo lo anterior, para términos de comparabilidad la red también estimará los cambios discretos de las tasa a 5 años

Finalmente, se desarrolló un modelo AR(1) para comparar los resultados de los modelos de minería de datos con un modelo autorregresivo puramente econométrico al que se le realizara un análisis estadístico para evaluar sus poder predictivo.

4.4 Implementación y resultados de los modelos

Esta sección corresponde a la parte proyectiva de este trabajo en que se analizan los resultados obtenidos y se muestran las optimizaciones realizadas a los modelos de manera de obtener el mejor resultado posible con el objetivo de responder las últimas dos preguntas de investigación de la sección 3.8. A continuación se desarrolla cada uno de los modelos individualmente para después

4.4.1 Modelo Conglomerado-árbol I

El modelo 1 es resultante de la sección descriptiva de las variables que sugiere que la estimación de la tasa cero cupón a cinco años debiera ser un proceso en dos etapas. En la primera etapa se utiliza como variable objetivo a la variable “escenario” que es la resultante del análisis de conglomerados de la fase descriptiva, que a su vez es estimada utilizando todas las variables disponibles, la razón de esto es que las variables económicas podrían entregar una información rezagada de los cambios de escenario, situación que se soluciona con la incorporación de las variables financieras que cuentan con mayor frecuencia. En la segunda etapa la variable objetivo corresponde a la tasa en pesos a 5 años y se utiliza la variable estimada “t-escenario” creando un árbol de decisión para cada uno de los escenarios, utilizando como variables independientes al grupo de las variables financieras rezagadas.

Etapas I

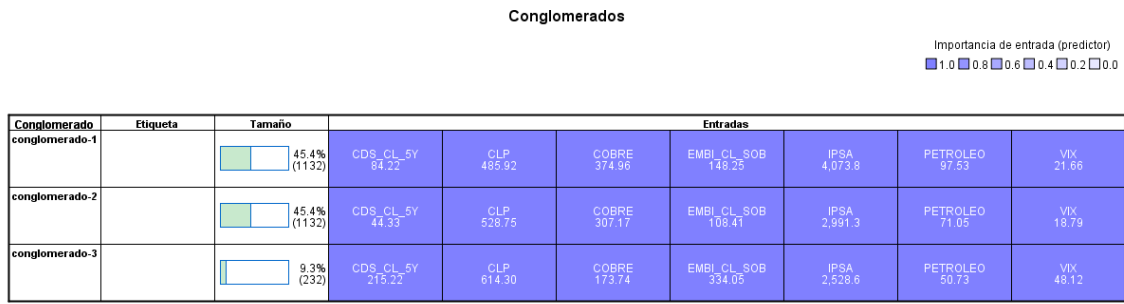
Si bien la evolución histórica de la variable “escenario” nos hace sentido en momentos extremos como los de una crisis esta variable presenta una serie de dificultades entre las que podemos nombrar:

- 1.- Al ser una variable descriptiva puede ser cuestionable en su interpretación. Sucede algo parecido al del análisis de componentes principales (PCA) en el que las variables más bien son factores y dependen mucho de la interpretación que se les pueda dar y lo estables que sean en el tiempo.
- 2.- Las relaciones causales o teóricas entre las variables económicas no siempre se cumplen y estas cambian en el tiempo. Resulta difícil cuantificar en que momento el mercado considera que la persistencia de uno o varios indicadores económicos vislumbra un escenario o el otro.
- 3.- Resulta difícil delimitar el rango en que las variables pasan a estar en otro escenario. Es difícil delimitar o discriminar entre un escenario y otro.
- 4.- En general sólo somos capaces de observar las relaciones lineales entre las variables, perdiendo el componente no lineal que puede ser relevante frente a un gran número de variables.

La respuesta a todas estas dificultades puede estar en la obtención de una variable escenario en base a un análisis histórico de cómo se han relacionado las distintas variables bajo los distintos escenarios económicos identificables y buscar patrones que se repitan en el tiempo. Para este análisis utilizaremos un modelo de conglomerados en dos etapas que básicamente lo que hace es clasificar en conglomerados a aquellas relaciones en común a lo largo del tiempo, para esto utiliza Log-verosimilitud como criterio de distancia entre las variables y criterio bayesiano de Schwarz (BIZ) como criterio de conglomeración. Lo anterior tiene un serie de ventajas, permite que sea el análisis de conglomerados el que defina el escenario evitando la modelación arbitraria, permite que la relación causal entre las variables pueda cambiar en el tiempo y por último, captura todas las relaciones entre las variables no sólo las lineales.

La siguiente tabla muestra bajo el título de Entradas a las variables utilizadas para construir el escenario junto a la media de las sus distribuciones dentro el conglomerado. Adicionalmente se muestra el tamaño relativo y absoluto de cada conglomerado.

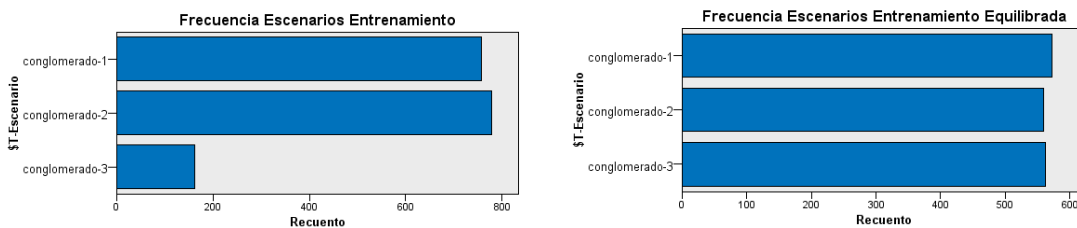
Figura 7: Descripción de conglomerado y sus principales componentes proyectados



De la información anterior es posible observar que los conglomerados son claramente identificables como escenarios económicos, por ejemplo, el conglomerado 1 está caracterizado por buenos niveles de la bolsa, un alto precio del cobre, un bajo nivel del dólar y un nivel de volatilidad (VIX) medio. Por el otro lado el conglomerado-3 presenta niveles de volatilidad altos con un nivel bajo de la bolsa, y el índice de bonos soberanos alto. Por su parte el conglomerado-3 se caracteriza por ser un escenario mucho menos frecuente, que sin embargo no es tan bajo producto de los múltiples episodios de inestabilidad durante los últimos seis años.

Respecto de la estimación de estos conglomerados que a partir de ahora llamaremos “escenarios” utilizaremos un árbol de decisión y utilizaremos como variable de entrada tres de las variables de todas las variables disponibles. Para esto se separó la data de manera aleatoria en un 70% para entrenamiento y el restante 30% para comprobación. Ahora tomando la porción de entrenamiento la tabla muestra que las frecuencias de ocurrencia están dispares lo que puede llevar al modelo a subestimar los resultados extremos, lo que genera un problema cuando se corren algoritmos que aprenden de la data ya que se focalizan de los datos con mayor frecuencia dejando menos desarrolladas las ramas en resultados extremos.

Figura 8: Tablas de frecuencia de evento, efecto al equilibrar.



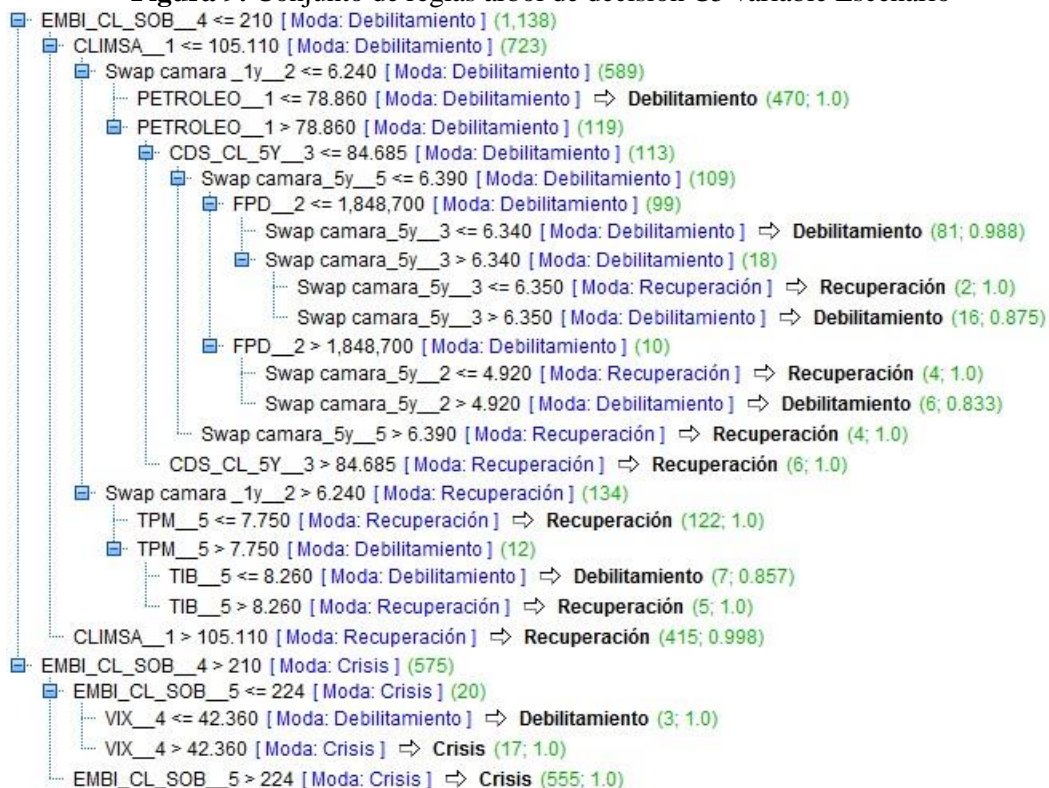
Para corregir lo anterior se equilibra la muestra de esta variable discreta multiplicando la muestra por un factor que equipara la cantidad de ocurrencias, asegurando la correcta estimación del árbol. Con la variable objetivo lista se corre el modelo de árboles de decisión con el algoritmo C5 que corresponde a una evolución del modelo original de Quinlan (ID3) el año 1986., y que permite

manejar una mayor cantidad de variables, incluyendo las continuas y la posibilidad de valores faltantes junto mejora en el criterio de optimización.

El siguiente conjunto de reglas representa el árbol generado utilizando la data de entrenamiento. Se observa que las variables principales del árbol corresponden al grupo de las variables económicas, sin embargo hacia los niveles inferiores comienzan a aparecer las variables financieras y las cotizaciones de algunos *commodities*. Destacan las tasas swap promedio cámara a un año junto a las cotizaciones del cobre y el petróleo en diferentes rezagos. Lo anterior sugiere que frente a cierto escenario explicado por las variables económicas, es la tendencia de las variables financieras la que permite discriminar si estamos en la parte descendente o ascendente del ciclo.

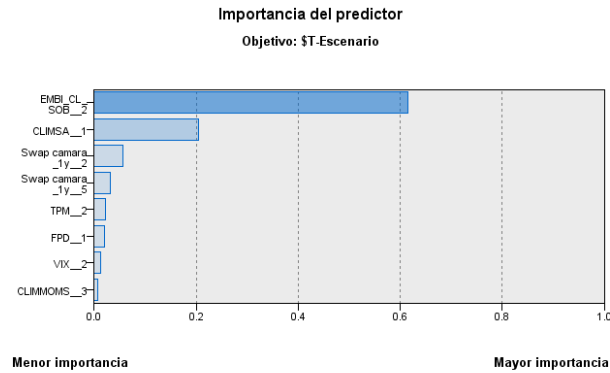
Otro aspecto relevante es la parte del árbol que explica los eventos de crisis. Como podemos ver la variable más importante para predecir una crisis en el modelo corresponde al EMBI que es conocido como la principal medida de riesgo país (EMBI_CL) junto al índice de volatilidad mundial (VIX) identificándose que cuando estas variables alcanzan ciertos niveles o umbrales es bastante probable que se desencadene una crisis. Dada la ventaja del árbol de tener la capacidad de explicar su resultado el siguiente paso sería generar un árbol excluyendo estas dos variables de manera de buscar más patrones entre las variables para detectar un escenario de crisis.

Figura 9: Conjunto de reglas árbol de decisión C5 variable Escenario



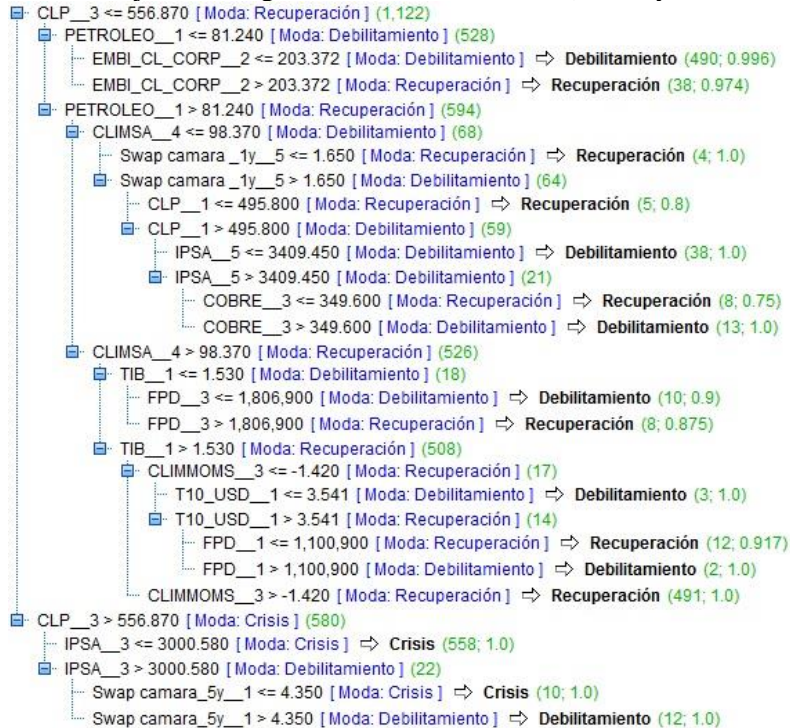
El siguiente grafico muestra la importancia del predictor relativa del modelo evidenciando que el EMBI soberano de Chile explica más del 50% del árbol, en segundo lugar queda el índice de producción industrial (IMACEC) generado por el BCCh.

Figura 10: Importancia del predictor árbol escenario



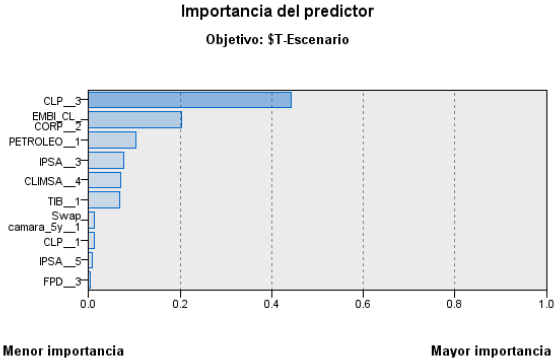
Al excluir ambas variables del árbol, este cambia en su estructura pasando a tomar relevancia en nivel del peso chileno, el nivel de la bolsa y el nivel de las tasas de mercado representado por el swap promedio cámara a cinco años. Es con esto, que el árbol predice una crisis con un peso depreciado y tanto tasas como la bolsa de valores bajas. Es posible ver también los umbrales que generan los distintos escenarios:

Figura 11: Conjunto de reglas árbol de decisión C5 (EMBI y VIX excluidos)



En términos de la importancia del predictor, se puede observar que el grado predictivo de las variables cae, con un nivel del peso spot CLP explicando un poco más del 40% del árbol.

Figura 12: Importancia del predictor árbol escenario



Con variables mencionadas anteriormente sumado al Índice de percepción económica (IPEC) se crearon los escenarios arbitrarios utilizados en el modelo 2.

Tabla 7: Porcentaje y matriz de frecuencia de aciertos Modelo 1 conglomerado-árbol

Predicción Escenarios	Rendimiento	Escenario real	Escenario Estimado		
			Crisis	Debilitamiento	Recuperación
% Correctos	98.61%	Crisis	70	0	0
% Erroneos	1.39%	Debilitamiento	1	342	6
Total Obs.	794	Recuperación	0	4	371

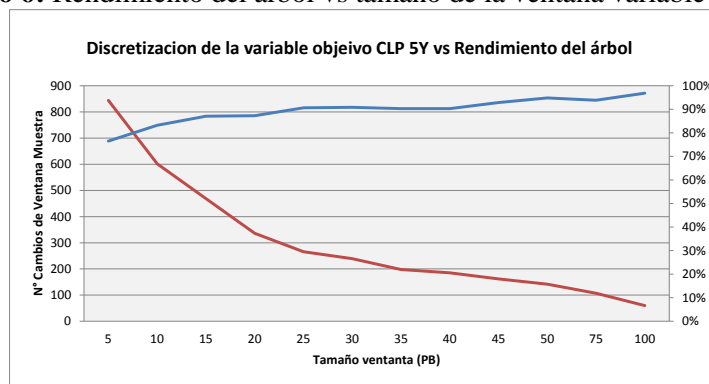
Estos cuadros muestran un análisis del rendimiento del árbol C5 sobre la variable escenario corrido sobre la data de comprobación (se separó el 30% de la data original de manera aleatoria la que es excluyente de la data de entrenamiento, utilizada para crear el modelo). Como se puede observar el rendimiento es alto, se debe a que la variable escenario es una variable que cambia de estado pocas veces en comparación a la cantidad de información incluida en su modelamiento, por lo que el número de veces en que el árbol podría estar erróneo es baja, adicionalmente se debe considerar que se utilizó la misma data para obtener la variable objetivo a través del análisis de conglomerados. Sin embargo lo anterior, esta “estabilidad” de la variable escenario es exactamente la condición base para el desarrollo de la segunda etapa de la estimación, donde se crea un árbol sobre cada uno de los escenarios para las variables financieras.

Etapa II

Con la segmentación de los escenarios a través de la historia, se utiliza esta información para estimar un árbol para cada uno de ellos. El primer paso para crear los arboles es convertir la variable objetivo en una variable discreta, lo que implica la creación de una delimitación o valores de corte que crearan distintos “segmentos” o “cajas”. Esto significa, por ejemplo que si el valor de la observación de la tasa tiene un calor de 4.5% el valor de esta variable será 5 que es la caja que incluye los valores desde el 4% al %5. El criterio para seleccionar el corte de cada presenta un supuesto fuerte respecto del rendimiento predictivo del modelo.

Existen varios criterios de creación de intervalos, entre los que podemos mencionar esta la división por cuartiles o deciles, el número de desviaciones respecto de la media y los de ancho fijo. En el caso de los primeros dos, estos dependen de los valores máximos y mínimos de la muestra utilizada para la estimación lo puede llevar a un error en la capacidad predictiva del modelo ya que los intervalos pueden ser demasiado amplios en periodo de baja volatilidad o viceversa. El autor optó por un criterio de intervalos de ancho fijo ya que hace comparables tanto el rendimiento en el tiempo de la variable objetivo como los distintos arboles correspondientes a cada escenario. Sin embargo lo anterior, la amplitud del intervalo también es un factor importante ya que un intervalo amplio implica que el árbol tendrá que estimar un cambio e caja en un menor número de veces. El siguiente grafico muestra el porcentaje de aciertos y el número de cambios de caja para el modelo 3 (Ver Tabla de Modelos).

Grafico 6: Rendimiento del árbol vs tamaño de la ventana variable objetivo



Como se puede observar, la cantidad de saltos de caja de la variable objetivo decrece rápidamente al aumentar la amplitud de la caja, lo que implica que un aumento en el rendimiento del modelo medido como el número de aciertos al intervalo. Esto tiene que ver con si bien la desviación típica de la tasa 5Y para toda la muestra que es de casi 70 pb. , el cambio diario promedio de la variable es de alrededor de 5 pb. Si bien el rendimiento con 5 pb base es bueno (levemente menor a un 80%

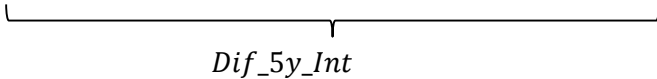
sobre 800 cambios de ventana) el número de cajas es muy alto generándose árboles demasiado extensos con muchos posibles valores. De manera de simplificar lo anterior, se transformó la variable objetivo en dos pasos, el primero:

$$Dif_5y = r_{5y} - r_{5y\ t-5} \quad \text{Ecuación (14)}$$

La ecuación anterior muestra la cantidad de puntos base que cambio la tasa en un horizonte de 5 días. En el segundo paso, se convierte *Dif_5y* a una variable discreta utilizando sólo 6 posibles valores de la siguiente forma:

Si *Dif_5y*

<-10pb	<-5pb	<0pb	=>0pb	>5pb	>10pb
-3	-2	-1	1	2	3

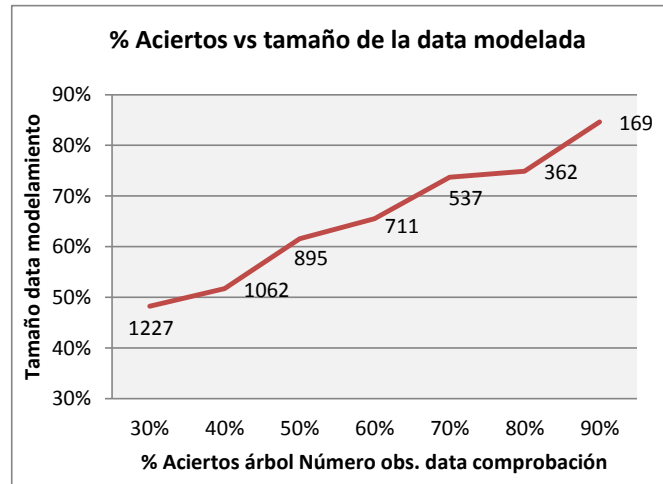


Dif_5y_Int

Con esto el árbol predecirá si la tasa en cinco días más subirá o bajará en tres posibles grados de magnitud. El autor considera que un valor de *Dif_5y_Int* de las dos últimas cajas de ambos extremos permitirían realizar una apuesta concreta. Toda esta preparación de datos adicional es un proceso iterativo del proceso de modelado con el objetivo de mejorar el resultado de la estimación.

Otro elemento que determinante en el rendimiento del modelo es el tamaño de la muestra aleatoria usado para la modelación. Para testear esto se tomó el modelo del árbol simple incluyendo todas variables y se modelo con distintos tamaños de la data de modelamiento:

Grafico 7: Rendimiento del árbol vs tamaño de la data de modelamiento



A pesar que la división de la data entre grupos se realiza de forma aleatoria, para cada uno de los tamaños de la data de modelamiento se corrió los modelos en repetidas oportunidades sobre la data de prueba para cada tamaño de la data modelada y los resultados son consistentes pero no exactos en cada iteración. Si bien la relación entre el performance y el tamaño de la data de modelamiento es ascendente es necesario entender que en el extremo esto juega en contra ya que si bien por un lado más data permite que el modelo entienda mejor todos los posibles resultados, por el otro, el menor tamaño de la data de comprobación puede llevar a una conclusión equivocada respecto del rendimiento del modelo.

El autor seleccionó para contrastar el rendimiento de los modelos un tamaño de la data de un 70% como data de modelamiento y un 30% como data de comprobación ya que 537 observaciones (más de dos años) es un número adecuado para testear el modelo y porque el salto en desde un tamaño de la data desde 60% a 70% es el último incremento marginal significativo (8.19%) hasta el 90%.

El cuadro inferior muestra el rendimiento del árbol tanto utilizando como variable independiente sólo variables financieras y una variante que incluye todas las variables, como se puede observar que el rendimiento de ambos modelos es alto. Adicionalmente los resultados sugieren que el modelo que incluye tanto el rezago de las variables financieras como las económicas muestra un mejor desempeño, la razón de esto se puede deber porque dentro de un escenario, los cambios en la tendencia de algunas variables económicas pueden ayudar a explicar mejor el movimientos de la tasa, por ejemplo el IPSA presenta una correlación negativa respecto de la tasa lo que se explica generalmente como el efecto “*Fly to Quality*” que se caracteriza por que los inversionistas se mueven entre activos riesgosos y libre de riesgo dependiendo del nivel de volatilidad de los mercados.

Tabla 8: Porcentaje de aciertos Modelo 1 conglomerado-árbol

Modelo I	Escenario	Variable Independiente	
		Financieras	Todas
% Correctos	Recuperacion	69.47%	73.19%
	Debilitamiento	66.11%	73.25%
	Crisis	64.41%	76.56%

La tabla anterior muestra una mejora significativa en el porcentaje de correctos para el modelo que utiliza todas las variables, especialmente en el escenario de crisis, reforzando que el modelamiento de los movimientos extremos de la tasas no sólo están explicados por los rezagos de la variable independiente u otras tasas de mercado, sino que es necesaria toda la información económica disponible, se incluyeron todas las variables en el modelado de los restantes modelos. La siguiente

tabla muestra una matriz de coincidencias entre el valor real de *Dif_5y_Int* (las filas) y su estimado (las columnas) para el escenario de crisis, debilitamiento y recuperación:

Tabla 9: Frecuencia de aciertos por caja Modelo 1 conglomerado-árbol
Dif_5y_Int estimada

C D R	<-10pb	<-5pb	<0pb	=>0pb	>5pb	>10pb
<-10pb	5 34 28	2 0 3	5 0 2	1 2 1	0 1 2	0 1 1
<-5pb	0 1 1	8 30 31	0 5 2	1 1 1	0 4 3	0 0 0
<0pb	1 4 1	0 0 6	9 23 26	2 7 4	0 1 0	0 1 3
=>0pb	0 0 7	0 3 3	0 8 2	8 14 30	0 4 0	0 4 2
>5pb	0 0 0	0 2 1	0 2 2	0 2 2	11 41 27	1 2 3
>10pb	1 2 3	0 1 1	0 0 3	1 4 3	0 3 1	8 36 30

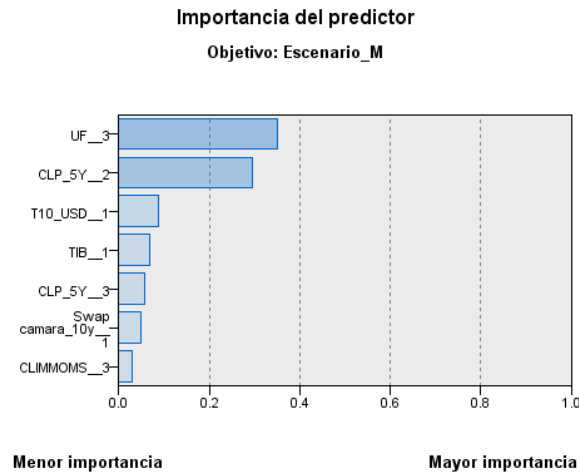
La matriz anterior representa la frecuencia de ocurrencia para cada escenario (Crisis|Debilitamiento|Recuperación) entre el movimiento real de la tasa expresado en el eje vertical versus los resultados estimados por el modelo 1 en el eje horizontal. La diagonal, de la matriz representa la cantidad de veces que el modelo estima correctamente el movimiento de la tasa para cada uno de los escenarios. Por ejemplo la celda ($=>0pb, <0pb$) con el valor 0|8|2 significa que el modelo predijo que el movimiento estaría entre 0 y -5 puntos base, siendo que realmente el movimiento de la tasa estuvo entre 0 y +5 puntos base en 0, 8 y 2 oportunidades para los escenarios de crisis, debilitamiento y recuperación respectivamente.

Se observa que en general el modelo tiene un buen rendimiento, incluso en los escenarios más extremos. Adicionalmente en la mayoría de las oportunidades el modelo falla en la magnitud del movimiento pero no en su sentido por lo que en una apuesta su resultado aun sería favorable.

4.4.2 Modelo Escenario manual-árbol II

Tomando como supuesto el hecho de que la tasa de interés se explica mejor al modelar los arboles dependiendo del escenario económico predominante, se identificó los eventos de crisis entre los años 2006 y 2012, para identificar los puntos bajos del ciclo económico. Adicionalmente se utilizó el índice de percepción económica (IPEC) del BCCh para identificar las recuperaciones y debilitamientos en el ciclo entre estos escenarios de crisis. Se utilizó la variable “Escenario_M” para modelar un árbol, observándose que las variables que mejor explican esta caracterización del ciclo económico son la UF y la tasa a 5 años en distintos rezagos y las tasas en el exterior, representado por la tasa a diez años de un *treasury* norteamericano. La incidencia de esta última variable se explica principalmente porque todas las crisis correspondieron a eventos internacionales que impactaron fuertemente los mayores mercados.

Figura 13: Importancia del predictor árbol escenario manual



El rendimiento de este árbol para el modelo II es superior al 95%, sobre una muestra de comprobación del 30% de la data que es independiente de la utilizada para el modelado. La siguiente tabla muestra el rendimiento de los tres árboles para el modelo II junto a su matriz de coincidencias, en el modelado del árbol se consideró todas las variables disponibles como variables independientes y la data corresponde a una partición de comprobación:

Tabla 10: Porcentaje y matriz de frecuencia de aciertos Modelo 2 Escenario manual-árbol *Dif_5y_Int* estimada

Modelo II	Escenario	Rendimiento	<i>Dif_5y_Int</i> real \ <i>Dif_5y_Int</i> estimada						
			<-10pb	<-5pb	<0pb	=>0pb	>5pb	>10pb	
% Correctos	Recuperacion	79.75%	6 42 27	2 2 1	2 0 1	1 0 2	0 0 0	0 0 1	
	Debilitamiento	76.03%	1 7 4	12 29 37	1 3 3	0 1 1	1 0 0	0 0 1	
	Crisis	76.62%	0 5 4	1 5 2	10 29 33	0 5 1	1 0 2	0 1 0	
				0 3 1	0 2 2	2 3 2	10 30 26	0 3 1	1 2 2
				0 0 1	0 0 0	1 0 2	1 8 3	8 29 30	0 3 2
				0 0 0	1 0 0	0 0 2	0 4 4	2 1 4	13 25 40

Los resultados muestran un rendimiento levemente superior al modelo I, sugiriendo que la elección que la definición del escenario tiene un impacto en el rendimiento del modelo. Lo anterior se puede deber también a que posiblemente esta definición del escenario discrimina mejor el nivel de volatilidad de los mercados y la tendencia de la dirección tasa. Por su parte la matriz de coincidencia muestra que este modelo casi no falla en estimar la dirección de la tasa, lo que en una apuesta sigue siendo favorable.

4.4.3 Modelo III árbol simple.

Este modelo se creó para contrastar el valor agregado de la creación de escenarios económicos y evaluar si al crear un árbol exclusivo sobre cada uno de ellos aportaba un mejor rendimiento al

modelo. El modelo cuenta con el mismo tratamiento de la data en términos de las transformaciones y los rezagos de las variables, tomando todas las variables para estimar el árbol sobre un 70% del total de la data para modelamiento y dejando el restante 30% para comprobación del modelo y sobre el cual se muestran las tablas de resultados expuestas a continuación. En términos de rendimiento, se observa un porcentaje correctos menor al de los dos modelos anteriores, evidenciando que el uso de las variables económicas para subdividir la data temporalmente dependiendo del escenario podría mejorar el rendimiento de los modelos.

La matriz de coincidencias muestra que este modelo presenta más problemas para predecir el signo del movimiento, observándose proporcionalmente un mayor número de ocurrencias en los cuadros más oscuros en comparación a los modelos anteriores. Lo anterior implica que la diferencia aumenta al momento de realizar apuestas con el modelo.

Tabla11: Porcentaje y matriz de frecuencia de aciertos Modelo 3 árbol simple

		<i>Dif_5y_Int</i> estimada						
		N° Obs.	<-10pb	<-5pb	<0pb	=>0pb	>5pb	>10pb
Modelo III	Rendimiento							
	% Correctos		69.53%					
	% Erroneos		30.47%					
	Total Obs.		548					
	<i>Dif_5y_Int</i> real	<-10pb	74	8	3	2	3	6
	<-5pb	3	58	7	2	3	3	
	<0pb	7	9	65	14	2	6	
=>0pb	9	8	13	47	10	2		
>5pb	3	2	3	5	79	7		
>10pb	6	3	2	5	11	58		

4.4.4 Modelo IV Red Neuronal

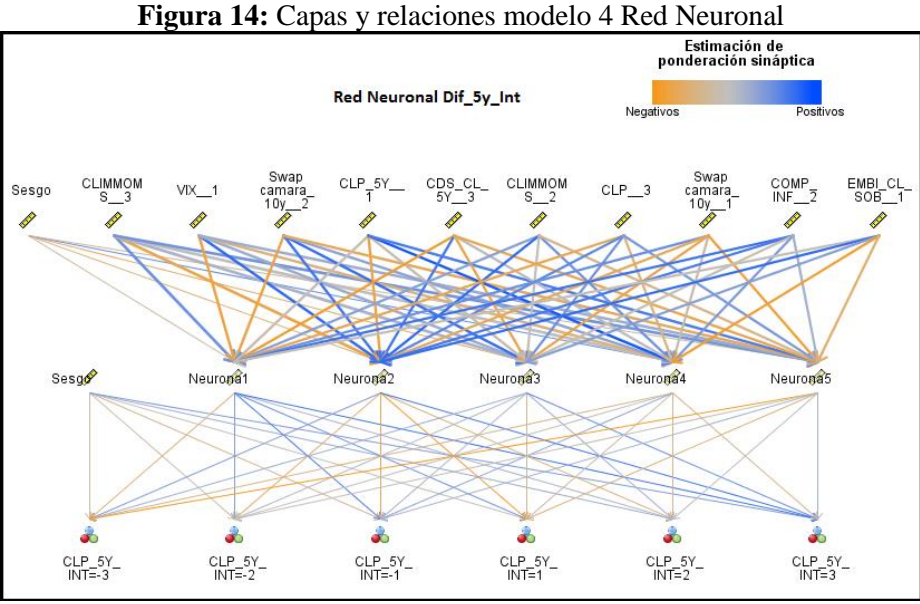
En este modelo se utilizó la misma información pre procesada de los modelos anteriores y se utilizó como variable objetivo a la variable discreta *Dif_5y_Int*, separándose el 70% de la data para crear la red y el 30% para medir su rendimiento, hay que recordar que este modelo no ofrece resultados fácilmente interpretables como en los árboles de decisión. Las siguientes tablas muestran el rendimiento del modelamiento de una red neuronal simple y sus resultados revelan que el rendimiento es muy inferior al de los árboles de decisión.

Tabla 12: Porcentaje y matriz de frecuencia de aciertos Modelo 4 Red Neuronal

		<i>Dif_5y_Int</i> estimada						
		N° Obs.	<-10pb	<-5pb	<0pb	=>0pb	>5pb	>10pb
Modelo IV	Rendimiento							
	% Correctos		29.49%					
	% Erroneos		70.51%					
	Total Obs.		539					
	<i>Dif_5y_Int</i> real	<-10pb	35	5	11	5	20	5
	<-5pb	14	18	26	9	17	4	
	<0pb	14	14	26	12	20	3	
=>0pb	7	12	24	16	21	7		
>5pb	12	18	10	7	44	11		
>10pb	24	6	8	5	29	20		

Quizás este rendimiento inferior viene dado a que este es una red simple o “naive”. Existe la posibilidad de realizar un “boosting” del modelo donde se crean varias redes al mismo tiempo y se van corrigiendo los errores. Queda pendiente un análisis más profundo de este modelo.

A continuación se muestra un diagrama de resultado que muestra los nodos de la red en dos capas, las de entrada que son las variables conocidas y la capa oculta con 5 nodos:



4.4.5 Modelo V Auto Regresivo

El último modelo evaluado corresponde a un modelo econométrico estándar definido por un rezago de la variable objetivo $5y_{Int}$, con los mismos supuestos y tratamiento de la data de entrenamiento y comprobación, la ecuación estimada es la siguiente:

$$CLP_{5y} = \alpha + \beta CLP_{5y_{t-5}} + e \tag{Ecuación (15)}$$

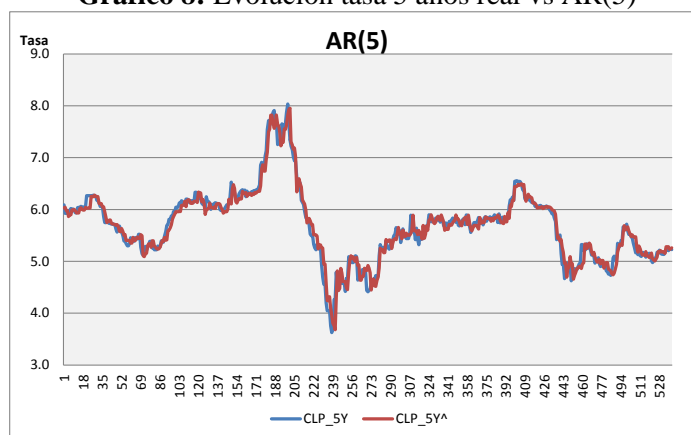
El método de estimación de la regresión anterior corresponde al de mínimos cuadrados (MC). El cuadro resultado de la regresión arroja que el modelo cuenta con un buen ajuste y con coeficientes estadísticamente significativos:

Tabla 13: Coeficientes y estadísticos modelo 5 Auto Regresivo

Modelo V	Coefficient	Std. Error	t-Statistic	Prob.
C	0.15	0.03	4.36	0.00
CLP_5y	0.97	0.01	163.47	0.00
R-squared	0.96	Mean dependent var		5.73
Adjusted R-squared	0.96	S.D. dependent var		0.69
S.E. of regression	0.14	Akaike info criterion		-1.05
Sum squared resid	25.00	Schwarz criterion		-1.04
Log likelihood	638.44	Hannan-Quinn criter.		-1.04
F-statistic	26723.90	Durbin-Watson stat		0.55
Prob(F-statistic)	0.00			

El siguiente grafico muestra la variable estimada versus la real, la variable x corresponde a sólo el número de la observación ya que esta la data esta fue extraída aleatoriamente de la muestra para propósitos de una medición de desempeño equivalente a la del resto de los modelos.

Grafico 8: Evolución tasa 5 años real vs AR(5)



Ahora para evaluar el desempeño del modelo de una manera equivalente al resto de los modelos se realizó la siguiente transformación:

$$CLP_{5y_{def}} = CLP_{5y} - CLP_{5y_{t-5}} \quad \text{Ecuación (16)}$$

Finalmente la variable resultante de la transformación anterior es catalogada según el criterio:

<-10pb	<-5pb	<0pb	=>0pb	>5pb	>10pb
-3	-2	-1	1	2	3

Con lo anterior el rendimiento del modelo muestra ser el más bajo evidenciando que gran parte de la predictibilidad de la variable objetivo proviene de del resto de las variables.

Tabla 14: Porcentaje y matriz de frecuencia de aciertos Modelo 5 Auto Regresivo
Dif_5y_Int estimada

Modelo V		Rendimiento	N° Obs.					
			<-10pb	<-5pb	<0pb	=>0pb	>5pb	>10pb
% Correctos	15.00%							
% Erroneos	85.00%							
Total Obs.	539							

<i>Dif_5y_Int</i> real	N° Obs.					
	<-10pb	<-5pb	<0pb	=>0pb	>5pb	>10pb
<-10pb	0	3	32	62	0	0
<-5pb	0	2	29	46	1	0
<0pb	0	0	39	37	0	0
=>0pb	0	0	51	40	0	0
>5pb	0	0	34	70	0	0
>10pb	0	1	37	53	2	0

El resultado anterior está explicado porque el modelo básicamente multiplica $CLP_{5y_{t-5}}$ por el parámetro 0.97, sumado a una constante (0.15) por lo que en general el resultado del modelo en términos de las cajas será 1 o -1 ya que la volatilidad diaria es menor a 5 pb.

Un planteamiento alternativo del modelo es estimar la siguiente regresión:

$$CLP_{5y_{dif}} = \alpha + \beta CLP_{5y_{dif-1}} \quad \text{Ecuación (17)}$$

El resultado de la regresión anterior muestra una constante poco significativa:

Tabla 15: Coeficientes y estadísticos modelo 5 Auto Regresivo versión modificada

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.00	0.00	0.06	0.96
CLP_5Y_DIF_T_1	0.81	0.02	46.16	0.00
R-squared	0.64	Mean dependent var		0.00
Adjusted R-squared	0.64	S.D. dependent var		0.14
S.E. of regression	0.09	Akaike info criterion		-2.04
Sum squared resid	9.22	Schwarz criterion		-2.03
Log likelihood	1244.55	Hannan-Quinn criter.		-2.04
F-statistic	2130.70	Durbin-Watson stat		1.92
Prob(F-statistic)	0.00			

Al utilizar la data de comprobación el rendimiento de este modelo es el siguiente:

Tabla 16: Porcentaje y matriz de frecuencia de aciertos Modelo 4 Red Neuronal
Dif_5y_Int estimada

Modelo V B		Rendimiento	N° Obs.					
			<-10pb	<-5pb	<0pb	=>0pb	>5pb	>10pb
% Correctos	47.68%							
% Erroneos	52.32%							
Total Obs.	539							

<i>Dif_5y_Int</i> real	N° Obs.					
	<-10pb	<-5pb	<0pb	=>0pb	>5pb	>10pb
<-10pb	58	21	8	5	2	3
<-5pb	6	37	23	9	2	1
<0pb	3	12	34	21	3	3
=>0pb	5	6	18	41	12	9
>5pb	2	5	16	33	40	8
>10pb	1	1	7	14	23	47

El rendimiento de esta regresión es mejor que la anterior, sin embargo se observa un alto número de errores en las columnas que discriminan el signo del movimiento. Aun considerando lo anterior, para términos comparativos utilizaremos este modelo.

4.4.6 Resumen resultados

A continuación se muestra una tabla que resume el porcentaje de aciertos por modelo para la data de prueba ordenados por rendimiento.

Tabla 17: Resumen Porcentaje de aciertos por modelo

Pos	Modelo	% Aciertos
1	Árbol Escenario Manual	77.47%*
2	Árbol Escenario	74.33%*
3	Árbol	69.53%
4	AR	47.68%
5	Red Neuronal	29.49%

* Promedio de rendimiento por escenario

De esta sección se extrae que los árboles de decisión muestran los mejores rendimientos y que este se puede incrementar si se utilizan en conjunto con otras técnicas. Es importante notar que además del rendimiento del modelo los errores de signo son menores por lo que su uso como estrategia de inversión debiera resultar en mejores resultados. Todos los modelos de árbol son superiores al benchmark contrastando con el de Redes neuronales el cual a juicio del autor tiene espacio de mejoramiento.

En la siguiente sección se testeará si estas diferencias de rendimiento son estadísticamente significativas.

5. Test comparativo de la calidad de la predicción de dos modelos

De manera de comparar si un par de modelos presenta una mejor predicción que el otro, se implementó el ampliamente usado test estadístico desarrollado por Diebold – Mariano (DM) en 1995. Este test compara pares de estimaciones utilizando los errores de estimación de cada uno y nos da una idea de si la victoria de un modelo sobre otro es significativa estadísticamente y no se debe solamente a la “buena suerte”. Lo que se realizará es correr el test sobre todos los pares de modelos del estudio.

El test consiste en lo siguiente:

Supongamos que $\{y_t\}$ va ser la serie a ser estimada por los dos modelos. Sean $y_{t+h|t}^1$ y $y_{t+h|t}^2$ las proyecciones del modelo 1 y 2 de y_{t+h} basada en I_t . Con esto generamos los errores proyectados de ambos modelos como:

$$\begin{aligned}\varepsilon_{t+h|t}^1 &= y_{t+h} - y_{t+h|t}^1 \\ \varepsilon_{t+h|t}^2 &= y_{t+h} - y_{t+h|t}^2\end{aligned}\quad \text{Ecuación (18)}$$

El incremento en h está dado por el mismo cálculo para $t = t_0, \dots, T$. Generando las dos series de errores estimados:

$$\{\varepsilon_{t+h|t}^1\}_{t_0}^T, \{\varepsilon_{t+h|t}^2\}_{t_0}^T$$

Debido a que cada paso h proyectado utiliza data superpuesta, los errores $\{\varepsilon_{t+h|t}^1\}_{t_0}^T$ y $\{\varepsilon_{t+h|t}^2\}_{t_0}^T$ presentaran correlación serial. La precisión de cada estimación es medida por una “función de pérdida” expresada como:

$$L(y_{t+h}, y_{t+h|t}^i) = L(\varepsilon_{t+h|t}^i), i = 1, 2 \quad \text{Ecuación (19)}$$

Las funciones L más populares son el error cuadrático y el error absoluto:

$$L(\varepsilon_{t+h|t}^i) = (\varepsilon_{t+h|t}^i)^2 \text{ y } L(\varepsilon_{t+h|t}^i) = |\varepsilon_{t+h|t}^i| \quad \text{Ecuación (20)}$$

Entonces, para determinar si un modelo estima mejor que el otro se testea la hipótesis nula:

$$H_0: E[L(\varepsilon_{t+h|t}^1)] = E[L(\varepsilon_{t+h|t}^2)] \quad \text{Ecuación (21)}$$

Con lo que la hipótesis alternativa queda como:

$$H_1: E[L(\varepsilon_{t+h|t}^1)] \neq E[L(\varepsilon_{t+h|t}^2)] \quad \text{Ecuación (22)}$$

El test DM se basa sobre el diferencial de la función L sobre los errores de ambos modelos llamada función de pérdida:

$$d_t = L(\varepsilon_{t+h|t}^1) - L(\varepsilon_{t+h|t}^2) \quad \text{Ecuación (23)}$$

Si ambos modelos tiene igual precisión, entonces su hipótesis nula sería:

$$H_0: E[d_t] = 0 \quad \text{Ecuación (24)}$$

Finalmente la forma funcional del test DM es:

$$S = \frac{\bar{d}}{(\widehat{avar}(\bar{d}))^{1/2}} = \frac{\bar{d}}{\left(\frac{LRV_{\bar{d}}}{T}\right)^{1/2}} \quad \text{Ecuación (25)}$$

Donde *avar* es la varianza asintótica:

$$\bar{d} = \frac{1}{T_0} \sum_{t=t_0}^T d_t \text{ y } LRV_{\bar{d}} = y_0 + 2 \sum_{j=1}^{\infty} y_j y_j = cov(d_t, d_{t-j}) \quad \text{Ecuación (26)}$$

$LRV_{\bar{d}}$ es una estimación consistente de la varianza asintótica (de largo plazo) de $\sqrt{T}\bar{d}$. Esta varianza de largo plazo es usada en el test por que la muestra de errores diferenciales $\{d_t\}_{t_0}^T$ presenta correlación serial para todos los pasos de h mayores que 1. Diebold y Mariano (1995) muestran que bajo la hipótesis nula de igualdad de precisión:

$$S^A \sim N(0,1)$$

Por lo que rechazamos la hipótesis nula de ambos modelos tienen el mismo rendimiento al 5% de confianza si:

$$|S| > 1.96$$

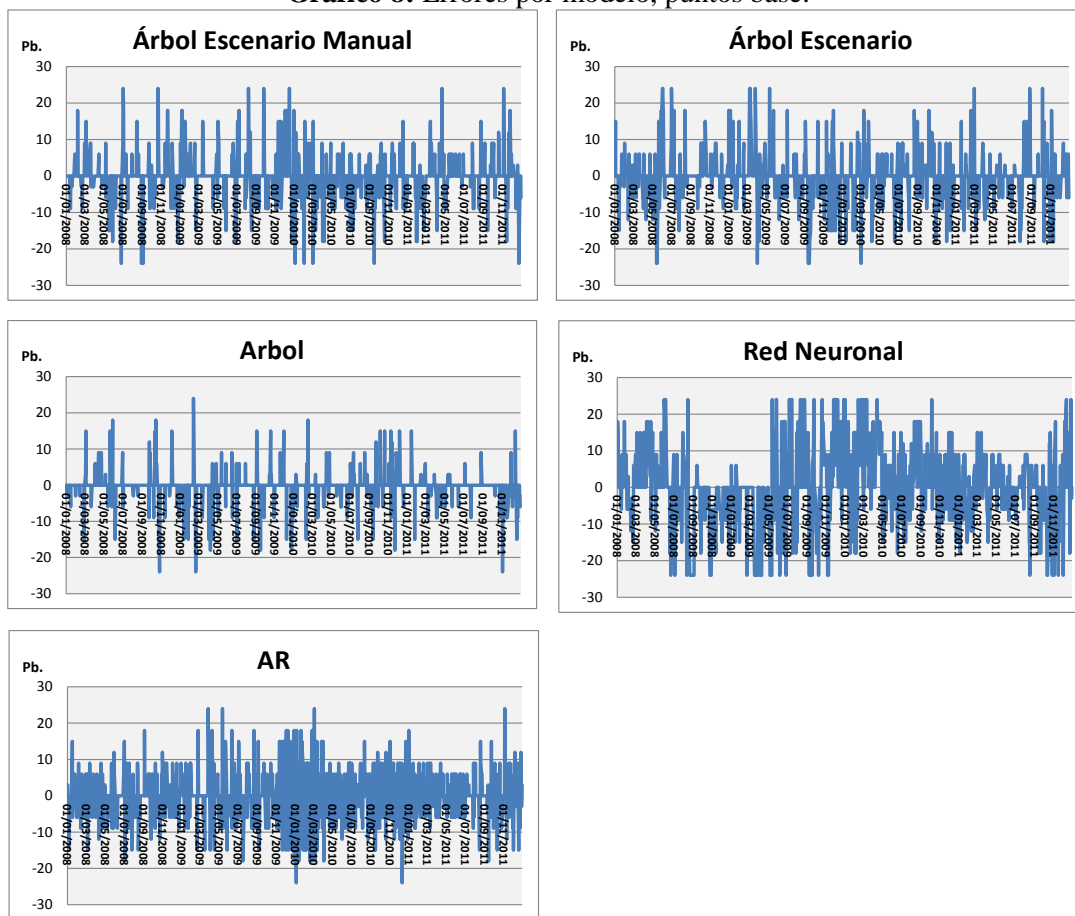
5.1 Resultados del test DM

Los modelos analizados en este trabajo entregan resultados discretos diferencia del modelo ARMA que entrega resultados continuos. Para hacer esto se llevó las estimaciones del modelo ARMA a su equivalente en términos de:

<-10pb	<-5pb	<0pb	=>0pb	>5pb	>10pb
-3	-2	-1	1	2	3

Dado lo anterior, se presentan los gráficos de la evolución de los errores para cada modelo:

Gráfico 8: Errores por modelo, puntos base.



La siguiente tabla muestra el test DM para todos los pares de modelos:

Tabla 17: Test de DM por par de modelos

Modelos	S	p-value	Ganador
Árbol Escenario Manual vs Árbol Escenario	-1.496	0.135	Ninguno
Árbol Escenario Manual vs Arbol	6.037	0.000	Arbol
Árbol Escenario Manual vs Red Neuronal	-10.941	0.000	Árbol Escenario Manual
Árbol Escenario Manual vs AR	-7.144	0.000	Árbol Escenario Manual
Árbol Escenario vs Arbol	8.044	0.000	Arbol
Árbol Escenario vs Red Neuronal	-9.925	0.000	Árbol Escenario
Árbol Escenario vs AR	-4.543	0.000	Árbol Escenario
Arbol vs Red Neuronal	-14.459	0.000	Arbol
Arbol vs AR	-13.537	0.000	Arbol
Red Neuronal vs AR	8.313	0.000	AR

Del análisis anterior se puede notar que los modelos de árbol superan al resto de los modelos y que dentro de los modelos de árbol, modelo de árbol en una etapa sin definición de escenario supera al resto de los modelos de árbol. Para aquellos modelos de árbol que usan el escenario no es posible definir si es mejor uno manual o el generado a través del análisis de conglomerados.

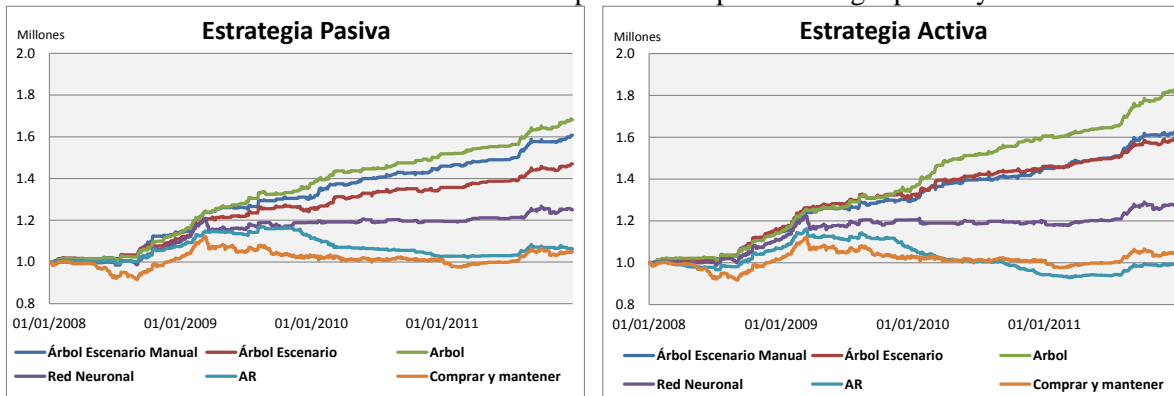
6 Uso de los modelos en una simulación de inversión

El objetivo de esta sección es testear el funcionamiento de los modelos dándole respuesta a la última pregunta de la sección 3.8. El ejercicio planteado consiste en evaluar la gestión de un portafolio con un valor inicial de 1 millón de pesos desde comienzos de 2008 hasta el término del año 2011. Las estrategias están creadas en base que los modelos no predicen el valor exacto que la tasa a 5 años tendrá en cinco días si no que en rangos:

<-10pb	<-5pb	<0pb	=>0pb	>5pb	>10pb
-3	-2	-1	1	2	3

El ejercicio estará dividido en dos tipos de estrategias. La primera llamada “estrategia pasiva” en la que las operaciones se realizan sólo si el modelo predice un cambio de signo con más de una ventana de diferencia y la “estrategia agresiva” que realiza operaciones ante cualquier cambio de caja. Las simulaciones anteriores no permiten venta corta por lo que solo se podrán beneficiar de las caídas de tasas y perder cuando el modelo estime el signo incorrectamente. Los ajustes al portafolio son diarios pero la proyección de la tasa es a 5 días. Lo anterior se debe a la baja cantidad de operaciones en el mercado chileno y porque un horizonte más largo le da una mejor estimación de la tendencia. Por otro lado, los ajustes diarios permiten incorporar los cambios repentinos en la tasa. La rentabilidad viene dada por el grado de sensibilidad en el precio que tiene un PDBC 5, expresada como la duración del bono por cambios en la tasa de mercado.

Grafico 9: Evolución valor de cartera por modelo para estrategia pasiva y activa



En los gráficos anteriores se observa que todos los modelos presentan retornos positivos en el periodo de análisis, sin embargo es el modelo de árbol simple el que obtiene mejor rentabilidad y el AR la peor. Se observa también que los modelos de árbol presentan una rentabilidad parecida y superior al del resto de los modelos.

Tabla 18: Rentabilidades anuales por modelo para la estrategia pasiva

Modelo	Rentabilidad				N operaciones
	2008	2009	2010	2011	
Árbol Escenario Manual	14.6%	14.0%	11.8%	10.1%	334
Árbol Escenario	11.3%	13.0%	7.9%	8.4%	353
Arbol	14.0%	20.8%	10.3%	10.7%	344
Red Neuronal	9.8%	8.2%	0.6%	4.5%	365
AR	7.4%	3.7%	-7.9%	3.5%	284
Comprar y mantener	1.8%	0.7%	-1.7%	3.8%	-

Se observa que durante los años 2008 y 2009 se obtienen las mejores rentabilidades principalmente por la crisis sub prime.

Tabla 19: Rentabilidades anuales por modelo para la estrategia activa

Modelo	Rentabilidad				N operaciones
	2008	2009	2010	2011	
Árbol Escenario Manual	15.3%	12.5%	11.8%	11.7%	472
Árbol Escenario	16.4%	13.9%	10.2%	9.2%	490
Arbol	14.7%	18.8%	17.9%	13.5%	485
Red Neuronal	11.3%	8.2%	-0.6%	6.3%	477
AR	6.8%	-0.7%	-11.2%	5.2%	491
Comprar y mantener	1.8%	0.7%	-1.7%	3.8%	-

En los cuadros anteriores se observa que todos los modelos presentan retornos positivos en el periodo de análisis, sin embargo es el modelo de árbol simple el que obtiene mejor rentabilidad. Adicionalmente se observa que si bien pueden tener errores en la magnitud del movimiento los modelos discriminan bien el signo de la variación, incluso en la estrategia agresiva donde se incluyeron las cajas +1 y -1, observándose un mejor desempeño en todos los modelos.

Se ha simulado también la posibilidad de ir corto en la compra del bono de manera de que el portafolio se beneficie tanto de las bajadas como de las subidas en la tasa.

Grafico 10: Evolución valor de cartera por modelo para estrategia venta corta

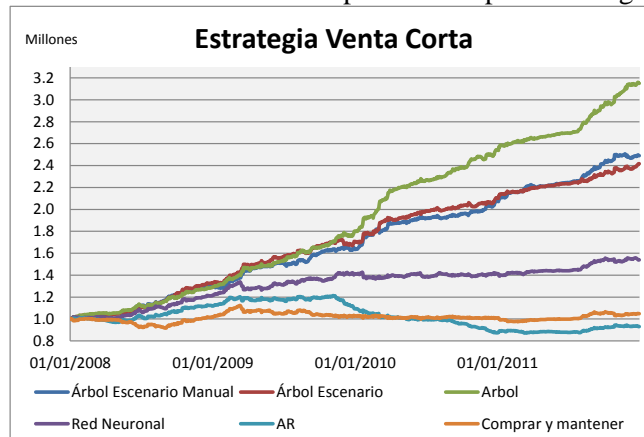


Tabla 20: Rentabilidades anuales por modelo para la estrategia venta corta

Modelo	Rentabilidad				N operaciones
	2008	2009	2010	2011	
Árbol Escenario Manual	30.4%	25.4%	26.9%	20.0%	1330
Árbol Escenario	32.9%	28.4%	23.3%	14.8%	1312
Arbol	29.0%	39.7%	41.1%	23.9%	1317
Red Neuronal	21.6%	16.2%	0.2%	8.7%	1325
AR	11.8%	-2.3%	-19.9%	6.5%	1311
Comprar y mantener	2.0%	0.7%	-1.7%	3.8%	-

Se puede observar un incremento considerable en la cantidad de transacciones explicado por la necesidad de realizar dos transacciones (una venta del bono más una venta en corto del mismo) cuando el modelo proyecta una subida tasas. Es posible notar que el modelo AR no presenta ningún poder predictivo.

Las simulaciones anteriores suponen algunos supuestos los cuales listan a continuación:

- a.- No existen costos de transacción.
- b.- Existe mercado para realizar las operaciones diariamente.
- c.- Se puede realizar venta corta.
- d.- No se incluyeron el devengado ni los pagos de cupones del PDBC.
- e.- No se varió la duración del PDBC se asume de que siempre hay disponible un PDBC 5 en el mercado.

7. Discusión

De acuerdo a los resultados anteriormente expuestos, en esta sección se procederá a contestar de manera sistemática las preguntas de investigación de la sección 3.8 de manera de presentar de manera resumida los principales resultados.

En particular, respecto de la pregunta de investigación i), los resultados del análisis descriptivo de las variables para el mercado chileno muestran que la variable escenario económico está explicada principalmente por el EMBI, el petróleo y el IMACEC quedando más atrás el CLP, el IPSA y el precio del cobre. Estas variables son fundamentales para configurar una coyuntura económica con un énfasis en la sensibilidad de Chile a los eventos externos, determinando que variables son las relevantes para explicar la tasa de interés. Por otro lado, las variables que más explican el árbol de decisión sobre la tasa de interés son variables que tienen relación directa con el mercado de tasas como las swap promedio cámara a diferentes plazos, la compensación inflacionaria y la TPM.

Para la pregunta ii) los resultados muestran que evidencia positiva y estadísticamente significativa entorno al hecho los modelos de árboles en conjunto con la configuración de escenarios (escenario y escenario manual) explicados por variables económicas que incluyen un visión de mediano plazo obtienen los mejores resultados, lo que apoya la hipótesis de que nuestro grupo de variables financieras de más corto plazo y frecuencia diaria obtienen mejores resultados con las variables económicas de frecuencia mensual.

Para la pregunta iii) los resultados del análisis descriptivo muestran que las variables que explican la evolución de la tasa de interés cambian bastante dependiendo del escenario o coyuntura económica reinante, lo que sumado al mejor rendimiento de los modelos de árbol con escenario permiten respaldar el hecho de que la relación entre las variables cambia en el tiempo y más aún, revelando que las variables que más explican a la dinámica de la tasa de interés van cambiando en el tiempo. Un ejemplo es la relevancia que toma la TPM en el escenario de crisis.

En el caso de la pregunta iv) los resultados del modelamiento muestran un porcentaje de aciertos de los modelos de árbol de hasta 77.5% superando tanto al modelo de redes como al modelo auto regresivo (benchmark). Adicionalmente, si consideramos que buena parte de los errores son de magnitud pero no de signo el rendimiento de los árboles es aún mayor.

Para la pregunta v) es indudable que una buena selección de las variables de entrada es fundamental para mejorar el rendimiento del modelo, lo que se demuestra en el rendimiento de los modelos de árbol al compararlos con el modelo auto regresivo. Por otro lado, si bien los modelos de árbol

combinados con la variable escenario obtuvieron mejores resultados en el modelamiento, los resultados de los test de DM no apoyan la hipótesis de que los modelos de árbol con escenario son superiores a la alternativa de un árbol convencional con variables de entrada tanto económicas como financieras. Por otro lado, esto también puede significar que el aprendizaje del árbol convencional sobre la etapa de escenario (o coyuntura económica) es superior al modelamiento en base a un análisis de conglomerados o una segmentación sobre reglas basadas en conocimiento teórico (escenario manual).

Finalmente, para la pregunta vi) la sección 6 muestra que los modelos de árbol demostraron ser útiles en su uso práctico ya que muestran buenos resultados en las distintas pruebas realizadas consistentemente en el tiempo.

8 Conclusiones y próximos desarrollos

Las técnicas de minería de datos ya no tienen una historia tan reciente. Los últimos desarrollos en este campo han ido de la mano con el rápido desarrollo de los computadores y del software, elementos fundamentales sobre el que se sustentan. Hoy en día es común encontrar empresas con sus propios departamentos de inteligencia de negocios que explotan la información generada en su medio para identificar nuevas tendencias, comportamientos o condiciones que potencialmente pueden ser una nueva oportunidad de negocio.

La inclusión de estas técnicas en el campo de las finanzas si tiene una historia reciente y son muchas sus posibles aplicaciones dada las características específicas de las series financieras en términos de volumen, frecuencia, facilidad de obtención y calidad haciéndolas candidatas naturales para su uso en minería de datos para obtener conocimientos de su interacción.

Existe una amplia teoría respecto del uso de la tasa de interés como una variable instrumental de la política monetaria, haciendo evidente una relación entre esta y el ciclo económico de la economía pasando por el control de la inflación y velando por la estabilidad de la moneda. Adicionalmente existen una serie de teorías que explican la dinámica de las tasas de interés como son teoría de la preferencia por liquidez o la teoría de las expectativas que ayudan a relacionar la tasa de interés con variables financieras como las paridades o el precio de las materias primas. Considerando que existe una gran cantidad de información económica-financiera disponible actualmente en Chile junto a bastante literatura que sustenta sus interacciones, en este trabajo se aplican las técnicas de minería de datos a un conjunto de variables económica-financieras de Chile y los mercados internacionales para describir su tendencia e interacción con el objetivo de crear un modelo que pueda predecir la tasa libre de riesgo de cinco años.

La técnica de minería de datos que mejor se amolda a los objetivos anteriores corresponde a los modelos de árboles de decisión los cuales se caracterizan por ser una técnica que jerarquiza las variables independientes en base a su poder explicativo de la variable objetivo y permiten modelar relaciones no lineales de alta complejidad dada la cantidad de variables y permitiendo a su vez, describir el camino que sigue la variable explicada mostrando su dinámica hasta llegar resultado final. Esta característica nos permitió observar claros patrones que van en línea con la teoría económica y financiera así como con las características específicas del mercado chileno. Se observa que la tasa de interés tiene un comportamiento diferente dependiendo del contexto económico financiero u escenario en el que se encuentra la economía, dado lo anterior se intentó definir estos escenarios para poder describirlos, demarcarlos y usarlos para aplicar los árboles de decisión

calibrados especialmente para cada uno de estos escenarios para mejorar la estimación de la tasa a cinco años.

En la definición de estos escenarios se utilizó el análisis de conglomerados sobre todas las variables disponibles llegando a tres conglomerados que asociamos a los posibles estados de la economía ya que presentaban diferencias marcadas respecto del nivel y tendencia de las variables sumado a que una de ellas coincide temporalmente con los periodos en que se desencadenaron algunas crisis.

Posteriormente utilizamos esta variable “escenario” como la variable objetivo de un árbol con el objetivo de que nos ayude a explicar cómo las variables y sus umbrales definían cada escenario. Se observa que las variables más relevantes para definir un escenario son el EMBI soberano de Chile, el índice de actividad económica IMACEC, el dólar y el precio del petróleo siendo estas variables las que más discriminan mejor ante un análisis de conglomerados. Es indudable que el EMBI soberano de Chile está construido para capturar las expectativas de los inversionistas respecto del premio por riesgo de la economía Chilena por lo que tiene una directa relación para explicar el ciclo económico considerando los factores relevantes en los mercados internacionales como por ejemplo el diferencial de tasas con las economías desarrolladas especialmente o las condiciones económicas en Asia al cual Chile presenta una mayor exposición. En el caso del IMACEC esta corresponde a una variable que mide íntegramente el dinamismo de actividad industrial interna por lo que su tendencia ayuda a definir un escenario. En el caso del dólar la posible explicación es un poco más complicada producto de su sensibilidad a una gran cantidad de variables, sin embargo quisiera mencionar que en general se confirma un dólar apreciado frente a los escenarios de crisis por ser una moneda refugio y por el cierre de las líneas que sufrieron los bancos de la plaza como el la crisis sub prime del 2008. Por otro lado un dólar depreciado tiende a estimular las importaciones y en consecuencia el consumo, lo que se asocia con un escenario más positivo hasta que esta alza en el consumo no sea un motivo del BCCh para subir la tasa. En sentido opuesto Chile en su calidad de exportador se ve afectado por un dólar depreciado por lo que un dólar bajo debiera estar asociado a un escenario negativo. Con lo anterior se hace un poco complicado utilizar el dólar como una variable que discrimine el escenario, especialmente cuando no se está en una situación de crisis o de crecimiento marcado. Finalmente, el precio del cobre está asociado al ingreso de la parte más importante de la producción industrial del país y del gobierno, impactando directamente en la estimación del gasto fiscal. Lo anterior implica que un precio del cobre alto debiera estar asociado con un escenario favorable para Chile y viceversa.

Alternativamente, se creó un variable escenario arbitraria sobre la base de la identificación temporal de los escenarios de crisis y los cambios que se dieron a la baja en la tendencia del Índice de percepción

económica como el término del escenario de recuperación. Esta medida alternativa del escenario permitió medir la sensibilidad del árbol frente a cambios en la definición del escenario en términos del rendimiento en la estimación de la tasa de interés.

Una limitación de lo anterior tiene que ver con como se expuso en las secciones anteriores, la dinámica de las relaciones entre las variables cambia dependiendo de la coyuntura económica, pero también con los cambios estructurales o regulatorios del mercado chileno. Si este fuera el caso, sería necesario un nuevo análisis descriptivo de las variables ya incluidas en el modelo, revisar si es necesario agregar nuevas variables para a continuación modelar su interacción con árboles. Por otro lado, el modelamiento de la variable escenario se basa principalmente sobre variables económicas de mediano plazo que no debieran sufrir grandes alteraciones frente a cambios provenientes de cambios en los agentes de mercado o modificaciones a la regulación.

Para medir la capacidad predictiva de estos árboles, se creó una variación del árbol que consiste simplemente en crear un árbol con todas las variables disponibles. Adicionalmente se creó un modelo de redes neuronales y un modelo auto regresivo, llamado también como modelo benchmark.

La modelación de los distintos arboles de decisión no fue un proceso simple ya que estos modelos estiman resultados discretos de la variable objetivo por lo que fue necesario convertirla en rangos fijos. Se observó que el tamaño de estos rangos tiene directa relación con el rendimiento del árbol por lo que se debe manejar esta limitación considerando la volatilidad de la variable objetivo de manera de que el árbol discrimine correctamente los rangos un número de veces considerable.

En el proceso de modelación se descubrió que un factor determinante en el rendimiento de los arboles es el tamaño de la data de entrenamiento del árbol versus el tamaño de la data de comprobación. Es importante entrenar el árbol de manera que aprenda sobre secciones de la data que incluyan eventos extremos de la variable objetivo, pero a su vez, si el conjunto de datos de comprobación es muy pequeño en general el rendimiento del árbol va estar sobreestimado, esta sensibilidad dependerá específicamente de la data que se esté modelando. Este *trade of* descubierto puede significar un problema si no se tiene un mayor conocimiento de la data que se está dejando en cada uno de los conjuntos de la data.

Se observa que bajo la misma data los modelos de árbol presentan un rendimiento bastante superior al de redes neuronales y el auto regresivo, destaca que los modelos si bien presentan errores en la magnitud de la estimación del movimiento de la tasa en general no cometen errores respecto del

signo del movimiento lo que implica un rendimiento significativo en las pruebas de rentabilidad de cartera. Sin embargo lo anterior, este sub rendimiento en el caso de las redes neuronales se puede deber a que solo se utilizó un modelo básico sin utilizar “boosting” que consiste en que una vez concluido el proceso de optimización de la red se toman los errores y se utilizan para construir la red nuevamente de manera iterativa. Queda pendiente desarrollar este modelo al máximo de sus capacidades.

Para verificar la validez de los rendimientos encontrados se utilizó el test de Diebold y Mariano sobre cada par de modelos confirmando estadísticamente un mayor rendimiento de los árboles. Dentro de los modelos de árbol el test muestra que el mejor modelo corresponde a un árbol simple. Lo anterior puede contradecir la conclusión de que la dinámica de la tasa depende del escenario económico o bien que esto lo puede hacer un árbol en una sola etapa siendo capaz de diferenciar mejor el escenario económico. Por su parte el árbol con escenario creado manualmente obtiene un rendimiento levemente superior al del escenario creado en base al análisis de conglomerado siendo esta diferencia estadísticamente significativa. Por último el modelo auto regresivo presentó el peor rendimiento sugiriendo lo significativas que son el resto de las variables para estimar una buena proyección de la tasa a 5 años.

Una limitación de lo anterior, es el grado de comparación del rendimiento del árbol producto de la proyección discreta a la que está limitado el modelo. En este trabajo se realizaron transformaciones a los modelos continuos para hacerlos comparables lo que puede haber significado que el rendimientos de estos modelos se viera afectado o no sea justa la comparación. Para aminorar este factor, se intentó discretizar los modelos tomando utilizando ventanas fijas y no dependientes de la dispersión de los datos.

En conclusión, podríamos decir que este trabajo fue exitoso y contribuye tanto a la academia como a la práctica en términos de la utilización de árboles de decisión para describir la dinámica de las distintas variables frente a los distintos escenarios económicos y utilizar este conocimiento para modelar un árbol proyectivo de la tasa con una buena capacidad de predicción y siendo fácil de implementar.

Los desarrollos futuros involucran profundizar la investigación descriptiva de las variables para configurar escenarios y entender la evolución en la relación de las variables relevantes del mercado chileno por ejemplo en la predicción de futuras crisis. Otro desarrollo relacionado con lo anterior es el estudio de la próxima implementación de la normativa basada en Basilea III sobre los

requerimientos de liquidez de los bancos que teóricamente debiera afectar de buena forma la relación entre las variables en periodos de crisis.

En términos del uso de los árboles como medida predictiva, este estudio debiera extenderse a otros países tanto desarrollados como países emergentes para testear el uso de árboles en conjunto con la estimación de escenarios y comparándolos a su vez con otras herramientas proyectivas.

9 Referencias

- [1] Abbot, J. L., Park, Y., & Parker, S. (2000). The effects of audit committee activity and independence on corporate fraud. *Managerial Finance*, 26(11), 55–67.
- [2] Adedara, O., & Longe, O. B. (2012). Forecasting Portfolio Investment Using Data Mining. *African Journal of Computing & ICT*, 5(4).
- [3] Aggarwal, R., & Demaskey, A. (1997). Using derivatives in major currencies for cross-hedging currency risks in Asian emerging markets. *Journal of Future Markets*, 17, 781–796.
- [4] Ammar, Salwa, Wright, R., & Selden, S. (2000). Ranking State Financial Management: A Multilevel Fuzzy Rule-based System. *Decision Sciences*. 31: 449-481. 2000.
- [5] Bai, B., Yen, J. & Yang, X. (2008). False financial statements: characteristics of China's listed companies and CART detecting approach, *International Journal of Information Technology & Decision Making* 7 (2) 339–359.
- [6] Banco Central de Chile (2005), Características de los Instrumentos del Mercado Financiero Nacional.
- [7] Banco Central de Chile, Calendario Anual de Emisión de Títulos de Deuda: <http://www.bcentral.cl/prensa/notas-prensa/pdf/03012013.pdf>
- [8] Banco Central de Chile, Capítulo IV.B.6.2 Compendio de Normas Financieras. <http://www.bcentral.cl/normativa/normas-financieras/pdf/indice.pdf>
- [9] Banco Central de Chile, Evolución Histórica: <http://www.bcentral.cl/acerca/funciones/02.htm>
- [10] Banco Central de Chile, Programa de compra de divisas Comunicadi de consejo 3 de enero 2011: <http://www.bcentral.cl/prensa/comunicados-consejo/otros-temas/03012011.pdf>
- [11] Barillas (2010) Can We Exploit Predictability in Bond Markets?, Manuscript, New York University
- [12] Batarce (2009) Efectos de la Emisión de Bonos del Banco Central Sobre las Tasas de Interés, Banco Central de Chile Documentos de Trabajo.
- [13] Beasley, M. (1996). An empirical analysis of the relation between board of director composition and financial statement fraud. *The Accounting Review*, 71(4), 443–466.
- [14] Bell, T., & Carcello, J. (2000). A decision aid for assessing the likelihood of fraudulent financial reporting. *Auditing: A Journal of Practice & Theory*, 9(1), 169–178.
- [15] Cerullo, M. J., Cerullo, V. (1999). Using neural networks to predict financial reporting fraud, *Computer Fraud & Security* May/June (1999) 14–17.

- [16] Chai, W., Hoogs, B.K., & Verschueren, B.T. (2006). Fuzzy Ranking of Financial Statements for Fraud detection. In proceeding of International Conference on Fuzzy System, (2006), 152–158.
- [17] Chapados, N. (2010) "Sequential Machine Learning Approaches for Portfolio Management." Département d'informatique et de recherche opérationnelle Faculté des arts et des sciences. Doctoral thesis. November 2009.
- [18] Chen, G., Zhanjia L., & Feng, H. (2007). Empirical Study on detecting financial statements Fraud- based on empirical data of public companies. [J] Auditing Study, 2007.
- [19] Cox, J. C., Ingersoll Jr, J. E., & Ross, S. A. (1985). A theory of the term structure of interest rates. *Econometrica: Journal of the Econometric Society*, 385-407.
- [20] D. Vela (2013). Forecasting Latin-American yield curves: An artificial neural network approach, Borradores de Economía, Banco de la Republica Colombia, Num. 761
- [21] Dai, Q., & Singleton, K. (2003). Term structure dynamics in theory and reality. *Review of Financial Studies*, 16(3), 631-678.
- [22] Dase, R. K., & Pawar, D. D. (2010). Application of Artificial Neural Network for stock market predictions: A review of literature. *International Journal of Machine Intelligence*, 2(2), 14-17.
- [23] Deshmukh, A., & Talluru, L. (1998). A rule-based fuzzy reasoning system for assessing the risk of management fraud, *International Journal of Intelligent Systems in Accounting, Finance & Management* 7 (4) 223–241.
- [24] Deshmukh, A., Romine, J., & Siegel, P.H. (1997). Measurement and combination of red flags to assess the risk of management fraud: a fuzzy set approach, *Managerial Finance* 23 (6) 35–48.
- [25] Diebold, F. X., & Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of econometrics*, 130(2), 337-364.
- [26] Dornbusch, R. (1979). Monetary policy under exchange rate flexibility.
- [27] Duffee, G. R. (2002). Term premia and interest rate forecasts in affine models. *The Journal of Finance*, 57(1), 405-443.
- [28] Eining, M. M., Jones, D. R., & Loebbecke, J. K. (1997). Reliance on decision aids: an examination of auditors' assessment of management fraud. *Auditing: A Journal of Practice and Theory*, 16(2), 1–19.
- [29] Enke, D., & Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with applications*, 29(4), 927-940.
- [30] Estrella, A., & Hardouvelis, G. A. (1991). The term structure as a predictor of real economic activity. *The Journal of Finance*, 46(2), 555-576.

- [31] Fanning, K., & Cogger, K. (1998). Neural network detection of management fraud using published financial data. *International Journal of Intelligent Systems in Accounting, Finance & Management*, vol. 7, no. 1, pp. 21- 24, 1998.
- [32] Fanning, K., Cogger, K., & Srivastava, R. (1995). Detection of management fraud: a neural network approach. *International Journal of Intelligent Systems in Accounting, Finance & Management*, 1. 4, no. 2, pp. 113– 26, June 1995.
- [33] Feroz, E.H., Kwon, T.M., Pastena, V., & Park, K.J. (2000). The efficacy of red flags in predicting the SEC's targets: an artificial neural networks approach, *International Journal of Intelligent Systems in Accounting, Finance, and Management* 9 (3) 145–157.
- [34] Frenkel, Jacob A. (1981) Collapse of Purchasing Power Parities during the 1970s., *European Economic Review*, Vol. XVI, No. 1, pp. 145-165.
- [35] Garg A. (2012).Forecasting exchange rates using machine learning models with time-varying volatility, Master Thesis in Statistics and Data Mining from Linköpings universitet/Statistik
- [36] Green, B. P., & Choi, J. H. (1997). Assessing the risk of management fraud through neural-network technology. *Auditing: A Journal of Practice and Theory*, 16(1), 14–28.
- [37] Hansen, J. V., McDonald, J. B., Messier, W. F., & Bell, T. B. (1996). A generalized qualitative—response model and the analysis of management fraud. *Management Science*, 42(7), 1022–1032.
- [38] Hicks, J. R. (1939). *Value and Capital: An Inquiry Into Some Fundamental Principles of Economic Theory*. [Mit Schaubildern und Einem Mathematischen Anhang]. Clarendon Press.
- [39] Hong, T., & Han, I. (2002). Knowledge-based data mining of news information on the Internet using cognitive maps and neural networks. *Expert Systems with Applications*, 23(1), 1-8.
- [40] Hoogs, B., Kiehl, T., Lacombe, C., & Senturk, D. (2007). A genetic algorithm approach to detecting temporal patterns indicative of financial statement fraud, *Intelligent Systems in Accounting, Finance and Management*, 2007, vol. 15: 41-56.
- [41] Huang, X. (2006). *Research on Public Company Accounting Fraud and regulation-from perspective of protecting investors [D]*. Xiamen: Xiamen University, 2006.
- [42] Iu, K. C., & Xu, L. (2003). Optimizing financial portfolios from the perspective of mining temporal structures of stock returns. In *Machine Learning and Data Mining in Pattern Recognition* (pp.266-275). Springer Berlin Heidelberg.
- [43] Jacovides, A. (2008). *Forecasting Interest Rates from the Term Structure: Support Vector Machines Vs Neural Networks* (Doctoral dissertation, University of Nottingham).
- [44] Joo, K., and Han, I., (2000). “Artificial Neural Networks supported by Change-Point Detection for Interest Rates Forecasting”. Graduate School of Management, Korea Advanced Institute of Science and Technology.

- [45] Joslin, S., Priebisch, M., & Singleton, K. J. (2009). Risk premium accounting in macro-dynamic term structure models. Manuscript, Stanford University.
- [46] Juszczak, P., Adams, N.M., Hand, D.J., Whitrow, C., & Weston, D.J. (2008). Off-the-peg and bespoke classifiers for fraud detection, *Computational Statistics and Data Analysis*, vol. 52 (9): 4521-4532.
- [47] Kapardis, M. K., Christodoulou, C. & Agathocleous, M. (2010). Neural networks: the panacea in fraud detection? *Managerial Auditing Journal*, 25, 659-678.
- [48] Kiehl, T. R., Hoogs, B. K., & LaComb, C. A. (2005). Evolving Multi-Variate Time-Series Patterns for the Discrimination of Fraudulent Financial Filings. In *Proc. of Genetic and Evolutionary Computation Conference*, 2005.
- [49] Kim, S. H., & Noh, H. J. (1997). Predictability of interest rates using data mining tools: a comparative analysis of Korea and the US. *Expert Systems with Applications*, 13(2), 85-95.
- [50] Kirkos, E., Spathis, C., & Manolopoulos, Y. (2007). Data mining techniques for the detection of fraudulent financial statements, *Expert Systems with Applications* 32 (4) (2007) 995–1003.
- [51] Koskivaara, E. (2000). Different pre-processing models for financial accounts when using neural networks for auditing, *Proceedings of the 8th European Conference on Information Systems*, vol.1, 2000, pp. 326–3328, Vienna, Austria.
- [52] Kotsiantis, S., Koumanakos, E., Tzelepis, D. & Tampakas, V. (2006). Forecasting fraudulent financial statements using data mining, *International Journal of Computational Intelligence* 3 (2) 104–110.
- [53] Lenard, M. J., & Alam, P. (2004). The use of fuzzy logic and expert reasoning for knowledge management and discovery of financial reporting fraud. *Organizational data mining: Leveraging enterprise data resources for optimal performance*.
- [54] Lenard, M. J., Watkins, A.L., and Alam, P. (2007). Effective use of integrated decision making: An advanced technology model for evaluating fraud in service-based computer and technology firms. *The Journal of Emerging Technologies in Accounting* 4(1): 123-137.
- [55] Lin, J. W., Hwang, M. I., & Becker, J. D. (2003). A Fuzzy Neural Network for Assessing the Risk of Fraudulent Financial Reporting. *Managerial Auditing Journal*, 2003, 18(8):657-665.
- [56] Liou, F. M. (2008). Fraudulent financial reporting detection and business failure prediction models: a comparison. *Managerial Auditing Journal* Vol. 23 No. 7, pp. 650-662.
- [57] Ludvigson, S. C., & Ng, S. (2009). A factor analysis of bond risk premia (No. w15188). National Bureau of Economic Research.
- [58] Maberly, E. D. (1986). The informational content of the interday price change with respect to stock index futures, *Journal of Futures Markets*, 6, 385–395.

- [59] Markowitz, Harry M. (1959). *Portfolio Selection: Efficient Diversification of Investments*. New York: John Wiley & Sons.
- [60] Massad, Carlos (1998) *La Crisis de Asia y sus consecuencias sobre la Economía Chilena*, Documetos del Banco Central de Chile.
- [61] Muñoz, Moreno (2014) *Aplicación de Herramientas de Data Mining en la Predicción de la Tasa de Interés en Chile*, Universidad de Chile Escuela de Postgrado de Economía y Negocios.
- [62] Nelson, C. R., & Siegel, A. F. (1987). Parsimonious modeling of yield curves. *Journal of business*, 473-489.
- [63] Owusu-Ansah, S., Moyes, G.D. , Oyelere, P.B., Hay, P. (2002). An empirical analysis of the likelihood of detecting fraud in New Zealand, *Managerial Auditing Journal* 17 (4) 192–204.
- [64] Pacheco, R., Martins, A., Barcia, R.M., & Khator, S. (1996). A hybrid intelligent system applied to financial statement analysis, *Proceedings of the 5th IEEE conference on Fuzzy Systems*, vol. 2, 1996, pp. 1007–10128, New Orleans, LA, USA.
- [65] Pathak, J., Vidyarthi, N., & Summers, S. L. (2005). A fuzzy-based algorithm for auditors to detect elements of fraud in settled insurance claims, *Managerial Auditing Journal* 20 (6) (2005) 632–644.
- [66] Peramunetilleke, D., & Wong, R. K. (2002). Currency exchange rate forecasting from news headlines. *Australian Computer Science Communications*, 24(2), 131-139.
- [67] Perols, J. (2011). *Financial Statement Fraud Detection: An Analysis of Statistical and Machine Learning Algorithms*, *Auditing: A Journal of Practice and Theory*, Vol. 30(2), pp. 19-50.
- [68] Persons O.S. (1995). Using Financial Statement Data to Identify Factors Associated with Fraudulent Financial Reporting. *Journal of Applied Business Research*, 1995, 11(3):38-46.
- [69] Pettitt, A. N. (1979). A non-parametric approach to the change-point problem. *Applied statistics*, 126-135.
- [70] Piatetsky-Shapiro, G. (1990). Knowledge discovery in real databases: A report on the IJCAI-89 Workshop. *AI magazine*, 11(4), 68.
- [71] Quinlan J. R. (1993) *C4.5: programs for machine learning*, Morgan Kaufmann Publishers Inc., San Francisco, CA.
- [72] Ravisankar P., Ravi V., Rao G.R. & Bose I. (2011). Detection of financial statement fraud and feature selection using data mining techniques. *Decision Support System*, 50, 491-500.
- [73] Ren, H. (2006). *Investigation Research on Public company financial report fraud*. [D] Dalian: Dongbei University of Finance, 2006.
- [74] Sharma, A., & Panigrahi, P. K. (2012). A Review of Financial Accounting Fraud Detection based on Data Mining Techniques, *International Journal of Computer Applications*, 39.

- [75] Shiller, Robert J. (1981) Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends? *American Economic Review* 71(3), pp. 421–436.
- [76] Soni, S. (2011). Applications of ANNs in stock market prediction: a survey. *International Journal of Computer Science & Engineering Technology*, 2(3), 71-83.
- [77] Spathis, C. T. (2002). Detecting false financial statements using published data: some evidence from Greece, *Managerial Auditing Journal* 17 (4) (2002) 179–191.
- [78] Spathis, C., Doumpos, M., & Zopounidis, C. (2002). Detecting falsified financial statements: a comparative study using multicriteria analysis and multivariate statistical techniques. *The European Accounting Review*, 11(3), 509–535.
- [79] Summers, S. L., & Sweeney, J. T. (1998). Fraudulent misstated financial statements and insider trading: an empirical analysis. *The Accounting Review*, 73(1), 131–146.
- [80] Tom Mitchell (1997), *Machine Learning*, ISBN: 0070428077 Páginas 3-4.
- [81] Varela (2007) *Mercados de Derivados: Swap de Tasas Promedio Cámara y Seguro Inflación*, N 56, Documentos de estudio del Banco Central de Chile.
- [82] Vasicek, O. (1977). An equilibrium characterization of the term structure. *Journal of financial economics*, 5(2), 177-188.
- [83] Weiss, G. M., McCarthy, K., & Zabar, B. (2007, June). Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs?. In *DMIN* (pp. 35-41).
- [84] Welch, J., Reeves, T. E., & Welch, S. T. (1998). Using a genetic algorithm-based classifier system for modeling auditor decision behavior in a fraud setting, *International Journal of Intelligent Systems in Accounting, Finance & Management* 7 (3) (1998) 173–186.
- [85] Wu, Y., & Zhang, H. (1997). Forward premiums as unbiased predictors of future currency depreciation: a non-parametric analysis. *Journal of International Money and Finance*, 16, 609–623.
- [86] Yuan, J., Yuan, C., Deng, Y., & Yuan, C. (2008). The effects of manager compensation and market competition on financial fraud in public companies: an empirical study in China, *International Journal of Management* 25 (2) (2008) 322–335.
- [87] Zhou, W. and Kapoor, G. (2011). Detecting evolutionary financial statement fraud. *Decision Support Systems*, v. 50, n. 3, p. 570-575.
- [88] Zimmermann, H. (2002). Yield Curve Forecasting by Error Correction Neural Networks and Partial Learning. Bruges: ESANN'2002 proceedings - European Symposium on Artificial Neural Networks.

10 Anexos

10.1 Anexo 1: Estadística descriptiva de la Data.

Analisis Descriptivo Data

Campo	Gráfico de muestras	Medida	Min	Máx	Suma	Rango	Media	Err. tip. de la media	Desv. típica	Varianza	Asimetría	Err. tip. de asimetría	Kurtosis	Err. tip. de Kurtosis	Unicos	Válidos
1 Fecha		Continua	2008-01-01	2012-10-31	--	215568000.000	--	--	--	--	--	--	--	--	--	2496
2 T10_USD		Continua	1.388	5.295	8680.942	3.907	3.478	0.021	1.025	1.051	-0.203	0.049	-0.958	0.098	--	2496
3 CLP		Continua	429.550	682.750	1291122.310	253.200	517.277	0.844	42.147	1776.371	1.180	0.049	2.193	0.098	--	2496
4 IPSA		Continua	1939.600	5040.970	8584907.300	3101.370	3439.226	17.428	870.704	758125.311	0.122	0.049	-1.229	0.098	--	2496
5 Swap canara_1yr		Continua	0.760	9.010	11790.960	8.250	4.724	0.038	1.877	3.522	-0.463	0.049	-0.270	0.098	--	2496
6 Swap canara_5yr		Continua	3.050	8.150	14113.710	5.100	5.655	0.018	0.882	0.777	0.228	0.049	-0.115	0.098	--	2496
7 Swap canara_10yr		Continua	3.840	8.110	15028.310	4.270	6.021	0.015	0.726	0.528	0.120	0.049	-0.224	0.098	--	2496
8 CLINDOMS		Continua	-3.770	6.280	824.190	10.050	0.330	0.022	1.095	1.200	0.889	0.049	12.076	0.098	--	2496
9 TPM		Continua	0.500	8.250	10852.000	7.750	4.348	0.043	2.149	4.620	-0.514	0.049	-0.506	0.098	--	2496
10 PETROLEO		Continua	33.870	145.290	202602.500	111.420	81.171	0.399	19.944	397.756	0.368	0.049	0.260	0.098	--	2496
11 COBRE		Continua	124.750	462.850	812480.800	338.100	325.513	1.412	70.535	4975.154	-0.801	0.049	0.328	0.098	--	2496
12 EMBL_CL_CORP		Continua	144.558	241.896	461240.885	97.338	184.792	0.564	28.169	793.520	0.392	0.049	-1.224	0.098	--	2496
13 EMBL_CL_SOB		Continua	67.000	411.000	368033.000	344.000	147.449	1.390	69.457	4824.343	1.874	0.049	3.570	0.098	--	2496
14 CLINSA		Continua	89.340	118.140	255008.060	28.800	102.167	0.155	7.726	59.693	0.433	0.049	-0.828	0.098	--	2496
15 UF		Continua	17910.460	22732.790	51211038.720	4822.330	20517.243	29.804	1489.022	221786.295	-0.412	0.049	-1.129	0.098	--	2496
16 CDS_CL_5Y		Continua	12.498	322.962	195447.390	310.464	78.304	1.153	57.590	3316.664	1.399	0.049	2.298	0.098	--	2496
17 VIX		Continua	9.880	80.860	56954.310	70.970	22.818	0.218	10.886	118.465	1.925	0.049	4.839	0.098	--	2496
18 TIB		Continua	0.300	8.380	10837.775	8.080	4.342	0.043	2.169	4.706	-0.543	0.049	-0.502	0.098	--	2496
19 FPD		Continua	0.000	4225700.000	1919312000.000	4225700.000	768955.128	13952.631	697073.226	48591082350.244	1.147	0.049	1.340	0.098	--	2496
20 FPL		Continua	0.000	355429.000	109497004.063	383429.000	41465.146	1441.979	72041.261	5189943273.266	3.049	0.049	11.867	0.098	--	2496
21 COMP_INF		Continua	0.000	5.667	7907.610	5.667	3.168	0.017	0.864	0.747	-1.432	0.049	5.765	0.098	--	2496
22 CLP_5Y		Continua	3.599	8.068	14287.926	4.469	5.724	0.014	0.677	0.458	0.422	0.049	1.200	0.098	--	2496
23 Escenario_M		Nominal	--	--	--	--	--	--	--	--	--	--	--	--	3	2496
24 CLP_5Y_BLP		Continua	3.180	8.450	14312.780	5.270	5.734	0.016	0.810	0.657	0.223	0.049	1.335	0.098	--	2496

10.2 Anexo 2: Investigación realizada sobre detección de fraudes con técnicas de minería de datos. (Fuente: Sharma, A., & Panigrahi, P. K. (2012). [74])

Investigación en Redes Neuronales en la detección de fraudes

No .	Autor	Técnica Minería de Datos	Objetivo Principal	Referencia
1	Fanning, Cogger and Srivastava (1995)	Neural Networks	To use neural networks to develop a model for detecting managerial fraud	[32]
2	Green and Choi (1997)	Neural Networks	To develop a neural network fraud classification model employing endogenous financial data in corporate fraud	[36]
3	Fanning and Cogger (1998)	Neural Networks	To use neural networks to develop a model for detecting managerial fraud	[31]
4	Cerullo and Cerullo (1999)	Neural Networks	To use neural networks to predict the occurrence of corporate fraud at the management level	[15]
5	Koskivaara (2000)	Neural Networks	To investigate the impact of various pre-processing models on the forecast capability of neural network for auditing financial accounts	[51]
6	Feroz et al. (2000)	Neural Networks	To predict the possible fraudsters and accounting manipulations	[33]
7	Lin, Hwang and Becker (2003)	Fuzzy Neural Network, Logistic Model	To evaluate the utility of an integrated fuzzy neural network model for corporate fraud detection	[55]
8	Kotsiantis et al. (2006)	Decision Trees, Neural Networks, Bayesian Belief Network, K-Nearest Neighbour	To apply a hybrid decision support system using stacking variant methodology to detect fraudulent financial statements	[52]
9	Kirkos et al. (2007)	Neural Networks, Decision Trees, Bayesian Belief Network	To explore the effectiveness of neural networks, decision trees and Bayesian belief networks in detecting fraudulent financial statements (FFS) and to identify factors associated with FFS	[50]
10	Fen-May Liou (2008)	Neural Networks	To build detection/prediction models for detecting fraudulent financial reporting	[56]
11	M Krambia-Kapardis et al. (2010)	Neural Networks	To test the use of artificial neural networks as a tool in fraud detection	[47]
12	Ravisankar et al. (2011)	Neural Network, Support Vector Machines	To identify companies that resort to financial statement fraud	[72]
13	Perols (2011)	Neural Networks, Support Vector Machines	To compares the performance of popular statistical and machine learning models in detecting financial statement fraud	[67]
14	Zhou and Kapoor (2011)	Neural Networks, Bayesian Networks	To detect financial statement fraud with exploring a self-adaptive framework	[87]

			(based on a response surface model) with domain knowledge	
--	--	--	---	--

Investigación en Expert System y Algoritmos genéticos en la detección de fraudes

No.	Autor	Técnica de Minería de Datos	Objetivo Principal	Referencia
1	Pacheco et al. (1996)	Hybrid intelligent system with NN and fuzzy expert system	To diagnose financial problems in companies	[64]
2	Eining, Jones and Loebbecke (1997)	Expert System	To build an expert system applying the Logit statistical model to enhance user engagement and increase reliance on the aid	[28]
3	Welch, Reeves and Welch (1998)	Evolutionary algorithms (genetic algorithms)	To use genetic algorithms to aid the decisions of Defense Contractor Audit Agency (DCAA) auditors when they are estimating the likelihood of contracts fraud	[84]
4	Kiehl, Hoogs and LaComb (2005)	Genetic Algorithm	To automatically detect financial statement fraud	[48]
5	Hoogs et al. (2007)	Genetic Algorithm	To detect financial statement fraud based on anomaly scores as a metrics for characterizing corporate financial behavior	[40]
6	Juszczak et al. (2008)	Supervised and Semi-Supervised Classification	To detect financial statement fraud	[46]

Investigación en Modelos de Regresión en la detección de fraudes

No.	Autor	Técnica de Minería de Datos	Objetivo Principal	Referencia
1	Persons (1995)	Logistic Model	To detect financial reporting frauds	[68]
2	Beasley (1996)	Logit Regression Analysis	To predict the presence of financial statement fraud	[13]
3	Hansen et al. (1996)	Probit And Logit Techniques	To use Probit and Logit techniques to predict fraud	[37]
5	Summers and Sweeney (1998)	Logit Regression Analysis	To investigate the relationship between insider trading and fraud	[79]
4	Abbot, Park and Parker (2000)	Statistical Regression Analysis	To examine if the existence of an independent audit committee mitigates the likelihood of fraud	[1]
6	Bell and Carcello (2000)	Logistic Model	To develop a logistic regression model to estimate fraudulent financial reporting for an audit client	[14]
7	Spathis (2002)	Logistic Regression	To use logistic regression to examine published data and develop a model to detect the factors associated with FFS	[77]
8	Spathis, Doumpos, and	Logistic Regression	To develop a model to identify factors associated with fraudulent financial	[78]

	Zopounidis (2002)		statement	
9	Owusu-Ansah et al. (2002)	Logistic Regression Models	To explore the Logit regression model to detect corporate fraud in New Zealand	[63]
10	Xuemin Huang (2006)	Regression Analysis Using Logit Model	To analyze financial indexes which can predict financial fraud	[41]
11	Haisong Ren (2006)	Logistic Analysis And Clustering Analysis	To establish a detecting model of fraud which can be used for empirical analysis of financial indexes	[73]
12	Guoxin et al. (2007)	Logistic Regression	To develop accounting fraud detecting model	[18]
13	Bai, Yen and Yang (2008)	Classification And Regression Trees (CART)	To introduce classification and regression trees to identify and predict the impact of fraudulent financial statements	[5]
14	Yuan et al. (2008)	Logistic Regression Models	To employ a logistic regression model to test the effects of managerial compensation and market competition on financial fraud among listed companies in China	[86]
15	Fen-May Liou (2008)	Logistic Regression, Classification Trees	To build detection/prediction models for detecting fraudulent financial reporting	[56]
16	Ravisankar et al. (2011)	Logistic Regression	To identify companies that resort to financial statement fraud	[72]
17	Perols (2011)	Logistic Regression, C4.5	To compares the performance of popular statistical and machine learning models in detecting financial statement fraud	[67]
18	Zhou and Kapoor (2011)	Regression, Decision Tree	To detect financial statement fraud	[87]

Investigación en Modelos de Regresión en la detección de fraudes

No.	Autor	Técnica de Minería de Datos	Objetivo Principal	Referencia
1	Deshmukh, Romine and Siegel (1997)	Fuzzy Logic	To provide a fuzzy sets model to assess the risk of managerial fraud	[24]
2	Deshmukh and Talluru (1998)	Rule-Based Fuzzy Reasoning System	To build a rule-based fuzzy reasoning system to assess the risk of managerial fraud	[23]
3	Ammar, Wright and Selden (2000)	Fuzzy Logic	To use fuzzy set theory to represent imprecision in evaluated information and judgments	[4]
4	Lenard and Alam (2004)	Fuzzy Logic and Expert Reasoning	To develop fuzzy logic model to develop clusters for different statements	[53]

			representing red flags in the detection of fraud	
5	Pathak, Vidyarthi and Summers (2005)	Fuzzy Logic and Expert System	To identify fraud in settled claims	[65]
6	Chai, Hoogs and Verschueren (2006)	Fuzzy Logic	To convert binary classification rules learned from a genetic Algorithm to a fuzzy score for financial data fraud rule matching	[16]
7	Lenard, Watkins and Alam (2007)	Fuzzy Logic	To detect financial statement fraud using fuzzy logic	[54]