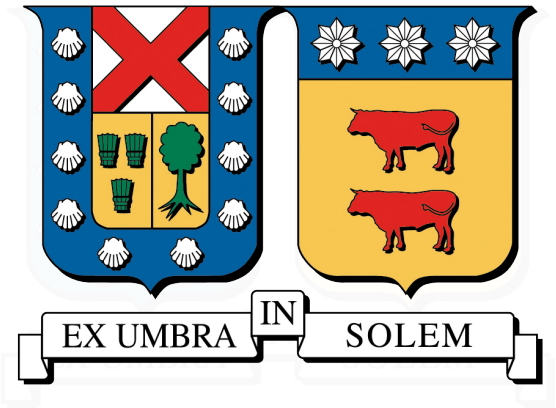
UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE INDUSTRIAS
VALPARAISO - CHILE



PREDICCIÓN DE RENDIMIENTOS DE ACCIONES EN EL MERCADO CHILENO MEDIANTE COMBINACIÓN DE CLASIFICADORES

DYNY MARCELO HUENUHUEQUE SEPULVEDA

MEMORIA PARA OPTAR AL TÍTULO DE
INGENIERO CIVIL INDUSTRIAL

PROFESOR GUÍA : WERNER KRISTJANPOLLER
PROFESOR CORREFERENTE : RODOLFO SALAZAR

ENERO 2018

Índice de Contenidos

1 | Problema de Investigación

A lo largo de la historia, y en el último tiempo las finanzas y economía se han caracterizado por constantes cambios en su comportamiento y entendimiento. El aumento de tal entendimiento junto con su impacto en el área financiera y en la economía de todos los hogares y negocios ha aumentado el interés de las personas para involucrarse más y más en sus temas relacionados. Dado lo anterior las personas dedicadas al mundo de las finanzas son quienes se encargan de buscar continuamente nuevos modelos, más eficientes y exactos para predecir rendimientos y comportamientos de los distintos instrumentos financieros que varían y que concluyen sobre el estado teórico del mercado.

Respecto a ello uno de los factores indicativos más importantes del mercado bursátil global y específicamente el chileno, es la predicción exacta de los rendimientos bursátiles de los inversores. Por tanto cualquier sistema computarizado que tenga la capacidad de predecir con precisión los rendimientos de las acciones es muy útil para la economía y útil para sus inversores.

Muchos de los estudios previos en diferentes mercados del mundo han dedicado sus esfuerzos en predecir la rentabilidad de las acciones, emparejando la información disponible (como pueden ser series temporales) junto a los rendimientos de las existencias utilizando como supuesto la posibilidad de por ejemplo una regresión lineal directa. No obstante no existe actualmente evidencia suficiente que enlace una relación lineal directa, es decir, donde el rendimiento de las acciones y la cantidad de información sea perfectamente concordante. Como resultado de lo anterior se abre la posibilidad de emplear modelos no lineales para explicar la incertidumbre resultante de una ecuación de regresión frente al rendimiento real de las existencias, y por tanto alcanzar un mejor desempeño en su

predicción.

En el último tiempo los avances en el área mencionada se centran en utilizar técnicas de aprendizaje automático, es el caso de redes neuronales, árboles de regresión y otros, ello para construir modelos "más seguros" de predicción en el rendimiento de las acciones.

Dentro de los estudios relacionados el método de redes neuronales ha sido el mejor considerado, pues como resultado en muchos de estos estudios las redes neuronales demuestran superar en muchas formas las técnicas estadísticas como regresiones.

En esta área donde el enfoque es el aprendizaje automático y el reconocimiento de algún comportamiento y metodología, la combinación de múltiples métodos de clasificación de a poco ha demostrado tener mejores resultados que una clasificación con único método. Más cuanto sorprende que dicha idea no ha sido totalmente probada en la predicción de la rentabilidad de acciones en los diferentes mercados financieros. Como existen variadas estrategias de combinación usadas para unir clasificadores, mencionándose Bagging (bootstrap aggregation) y decisión basada en votación por mayoría, no hay una respuesta definitiva de si la combinación de clasificadores es el mejor enfoque, por tanto estudiarla para predecir rendimientos y ver su cercanía a la realidad actual del mercado resulta una buena forma de aplicar y medir certeza de metodologías actuales en un mercado creciente y estable como lo es el chileno.

1.1. Objetivo General

Examinar el desempeño en la predicción de rendimiento de acciones en el Mercado Chileno, mediante técnicas que combinan conjuntos de clasificadores homogéneos y heterogéneos, con el fin de reducir la incerteza al momento de optar por inversiones.

1.2. Objetivo Específicos

- Determinar que combinación de clasificadores, como redes neuronales, arboles de regresión y regresión logística, aplicar para entregar el mejor resultado en predicción.
- Comprobar factibilidad de aplicar Redes Neuronales y su aprendizaje automático para predecir rendimiento de inversión en diferentes acciones del mercado chileno.
- Analizar la cantidad y tipos de indicadores del mercado chileno a utilizar en los diferentes modelos de predicción y si ellos funcionan bien con el aprendizaje automático.

2 | Revisión Bibliográfica

A continuación se presenta la revisión bibliográfica, donde se analizaran estudios relacionados al aprendizaje automático (ocupando para ello redes neuronales):

- Kim et al. (2006) en sus estudios realizados MJ. Kim realiza un análisis a la metodología aplicada durante las últimas décadas donde se señala que el uso de tecnologías dedicadas al aprendizaje automático ha estado en constante crecimiento. Iniciándose en sus orígenes en metodologías de análisis individual, pero que con el paso de los años se especializa en buscar formas de combinar los métodos de clasificación como maquinas de vectores de soporte SVMs y metodología Boosting, donde esta ultima la aborda para realizar una comparación entre una metodología únicamente de Redes Neuronales, otros métodos de combinación y su nivel de precisión asociado.

  Pero que se reconoce como boosting; un meta algoritmo de aprendizaje automático y cuyo trabajo es reducir el sesgo y varianza mediante un aprendizaje supervisado. Naciendo del cuestionamiento de si el conjunto de clasificadores débiles pudiese crear un clasificador robusto, donde este último tiene un mejor desempeño que un clasificador débil, pues sus clasificaciones se aproximan aún más a las verdaderas clases.

  El boosting consiste en combinar resultados de variados clasificadores débiles, como en su caso de estudio Redes Neuronales donde agrupa diversos clasificadores de este tipo, que muchos presentan resultados débiles, pero que una vez reunidos cambian su estructura de pesos, es decir, aquellos que son mal clasificados ganaron peso y aquellos clasificados correctamente perdían peso. De dicha manera los clasificadores débiles se centran de mejor manera en los casos que en su momento fueron mal

clasificados.

Como resultado de su investigación que ya se mencionaba únicamente centrada en Redes Neuronales y su combinación mediante Boosting, bagging y otros métodos, llegando a concluir que no existen grandes diferencias entre un método de combinación y otro, pero que de manera muy específica en la predicción financiera en momentos de crisis aquel que mejor resultados generales muestra es el Bagging y no boosting.

- Tsai y Chen (2010) hace referencia a la existencia de los riesgos al momento que una institución financiera emite préstamos de consumo, y es por ello que resulta fuertemente necesario desarrollar métodos de clasificación de acuerdo a los estándares existentes. Por tanto en su trabajo pretende enfocar los esfuerzos en comparar diversos modelos heterogéneos de aprendizaje automático para calificar los créditos. Los modelos desarrollados basan las predicciones de clasificación en modelos básicos, para luego mediante "Clustering" proporcionar mayor precisión al modelo bajo estudio, lo cual implicaría reducir las tasas de error y permitir con ello obtener el máximo beneficio al momento de entregar un crédito.

  Los métodos de clasificación "base" utilizados son la Regresión Logística y las Redes Neuronales, que mediante diferentes formas de ser combinadas entregan un resultado nuevo al entregado inicialmente.

  En la practica el hallazgo obtenido señala que el ensamblar modelos de clasificación de forma heterogénea entrega mejores resultados que la combinación homogénea o inclusive que el mejor modelo utilizado de manera individual, demostrando tener mejores tasas de precisión en la predicción y menores tasas de error en termino de calificación crediticia.

  Además es certero en señalar que en base al modelo desarrollado la mejor combinación consiste en RL +NN, generándose respuestas que para una institución financiera le puede permitir en un futuro con confianza llegar a emitir préstamos de consumo.

  Cabe señalar que este estudio considera variadas técnicas populares de ensamblaje

híbrido, pero que desde el punto de vista práctico resulta imposible realizar un estudio exhaustivo de cada una de las técnicas existentes y por tanto se aplica un criterio bastante subjetivo en cuanto a decidir que metodologías estudiar.

- Lin et al. (2015) presenta un enfoque nuevo a los métodos de clasificación estándar, presentado como clasificación dual, el cual tiene por objetivo mejorar el rendimiento en la clasificación a partir de la selección de instancias. Ello implico el escoger dos clasificadores y entrenarlos con datos indicados como datos buenos y datos ruidosos, pero que al momento de pasar por el clasificador es responsabilidad del mismo el acomodarlos al grupo correcto.

  La necesidad de agrupar el tipo de dato que se está trabajando previo a utilizar un segundo clasificador permite eliminar aquellos datos sindicados como ruidosos y de tal forma entrenar un conjunto reducido de datos, pero calificados como "correctos", lo que produciría teóricamente un rendimiento final mejor que el clasificador de referencia utilizado con el conjunto original.

  Sin embargo existen numerosos algoritmos clasificadores que son dependientes de la existencia de los datos atípicos por ser parte de una variable totalmente necesaria de pertenecer al modelo.

  Lo anterior se ve agravado por el hecho de que resulta muy complicado definir cuando un valor puede resultar como atípico si se considera la existencia de muchas clases de conjuntos de datos.

  Específicamente por tal motivo que plantea un modelo que ensambla dos clasificadores a modo de asegurar que los datos considerados como atípicos al extraerse de la muestra no sean simplemente un falso positivo.

  Los datos resultantes muestran que al trabajar sobre una base de datos a gran escala o una muestra pequeña, se reduce el riesgo de sobre seleccionar datos atípicos al asignar dos instancias de clasificación, permitiendo que el resultado final de estudio presente mejoras en la predicción, comparado a si se utilizase la base de muestra original.

- Tsai et al. (2014) tiene como objetivo el estudio de la predicción de la instancia conocida como bancarrota, mediante el uso de minería de datos y aprendizaje automático de modo de que la precisión en tal predicción resulte lo mayor posible.

  Y es que en muchas otras literaturas se ha desarrollado la comparación de metodologías de combinación de clasificadores y demostrado que su rendimiento es superior al de muchos clasificadores únicos.

  Existe lamentablemente tres problemas críticos al momento de construir los clasificadores en un conjunto pues su rendimiento puede verse afectado de diversas formas, mencionándose en primer lugar el escoger correctamente la metodología de combinación, que se adecue correctamente a la naturaleza de los datos de estudio, en segundo lugar se tiene el saber que clasificadores individuales se utilizaran para realizar la combinación y en qué cantidades.

  Y es por tal motivo que dentro del trabajo a desarrollarse se seleccionaran igualmente que en este estudio un número limitado de clasificadores, tres para ser exacto y siempre incluyendo las neuronas multicapa (MLP), ello por sus múltiples usos en estudios de variado tipo y presentar una reconocida certeza en la predicción. De la investigación realizándose únicamente ensambles de clasificadores de manera homogénea, es que se comprueba la mejora en la predicción de instancias de bancarrota, comparándose con modelos individuales, demostrando que es posible mediante el aprendizaje automático y reuniendo un número acotado de clasificadores reducir el error de predicciones y evitar caer en decisiones mal justificadas.

  Se denota además que a medida que aumenta la cantidad de clasificadores utilizados en la combinación reduce en alguna medida la precisión general, ello por sobreentrenar la base de datos al enfrentar las respuestas dadas por tantos métodos.

- Kimoto et al. (1990) analiza un sistema de predicción para prever los tiempos de compra y venta de acciones en la bolsa de Tokio. Para realizar aquello se basa en redes neuronales modulares, desarrollando un algoritmo de aprendizaje que fuese lo más precisas posible.

  La predicción realizada obtuvo un excelente beneficio al momento de someterla a un ejercicio simulado. La manera en la cual fue construido el modelo adecuaba las diferentes reglas de fluctuación del precio de las acciones al hacer análisis de clúster.

  El trabajo en especifico permite asegurar que mediante la correcta definición de la estructura de combinación de clasificadores se puede alcanzar una precisión tal que la predisposición invertir "riesgosamente" aumenta, pues asegura que la mayor parte de los escenarios pueden ser considerados y previstos.

- Ortiz (2008) y Luis Ayala Jiménez (2009) señalan que en la actualidad no existe ningún mercado libre de más fácil acceso que el mercado bursátil, ello debido a la inexistencia de barreras de entrada. Y es por esa razón que se plantea la posibilidad de creación de un portafolio de inversión de lo más adecuado posible, a modo de escoger aquellas empresas que maximicen la rentabilidad mediante la aplicación de técnicas como redes neuronales, las cuales tratan de resolver problemas estándar como optimización, reconocimiento y generalización.

  Las redes neuronales no son consideradas como las metodologías de decisión más antigua, pues previa a ellas se encontraban técnicas clásicas, como el Análisis Fundamental y posteriormente el Análisis Técnico, pero que rápidamente se fueron quedando opacadas por la alta velocidad de procesamiento de de las NN.

  Como resultados del trabajo de predicción en redes neuronales artificiales se confirma la aplicabilidad practica en el mercado bursátil por sobre los métodos usuales, ya que no depende su funcionamiento de los supuestos teóricos que utilizan los otros mecanismos. Su certeza en precisión demostrada por el cuadrado del error indica que el modelo basado en las redes neuronales tiene una mejor certeza y facilidad de re aplicabilidad una vez realizado el modelo.

- Sepúlveda y Correa (2013) señala omo el modelo de regresión lineal resulta ser uno de los métodos más utilizados al momento de proyectar múltiples situaciones. Muchas de sus ventajas hacen de esta metodología útil para su uso, por su facilidad de interpretación, de elaboración y ser reducido en costos. Su facilidad de uso y uso reiterativo hace caer a muchos de sus usuarios en el error de aplicarlo en situaciones erróneas, como es el caso de respuestas de carácter discreto, donde de muchas maneras se intenta llegar a un resultado adecuado, dañando un modelo que por naturaleza de aplicación ya se encontraba dañado.

  Por tal motivo se pasa a realizar la comparación con una metodología conocida como arboles de clasificación y regresión (CART), modelo que igualmente utiliza datos históricos para construir un árbol que clasifica y predice nuevos datos. Como ventaja primordial es que por naturaleza permite utilizar tanto variables numéricas como categóricas.

  El CART de manera muy general es conocido como una partición recursiva para construir un modelo de regresión en cada instancia o nodo en la cual se encuentran los datos, dividiendo la base entrante en una subdividida que por reglas de decisión avanzan por una rama u otra al nodo siguiente.

  Del estudio se concluye que cuando se comparan predicciones resultantes de CART y las mismas para Regresión Lineal, ya sea mediante regresión cuadrática o trigonométrica, el error de la predicción del primer modelo es levemente mayor al segundo, pero que en la media de los datos la varianza resulta ser bastante similar.

  El CART por tanto es una alternativa que permite al usuario dar una primera impresión de cómo es el comportamiento de las variables dentro de modelo, ello para luego aplicar un modelo que asegure mejor precisión de respuesta.

- Villada et al. (2016) realiza un nuevo estudio que pretende analizar la fiabilidad y certeza en la predicción de Redes neuronales aplicadas a una situación tan variable como es el precio de un bien. El objetivo de este modelo es predecir los precios de cierre del oro en el mercado de Londres, basándose en la data existente del banco central de Colombia. El planteamiento inicial consiste en agregar variadas configuraciones de RNA con propagación hacia adelante, tomando como entradas indicadores económicos variados.

  Como cualquier modelo de aprendizaje automático utiliza una serie histórica para el entrenamiento, parte de la cual será guardada para posteriormente realizar una predicción y validación del modelo.

  Sus resultados muestran una predicción en el precio del oro bastante acertada a la realidad, destacando la capacidad de la metodología de modelar sistemas complejos y generar errores menores tanto a nivel de predicción como entrenamiento.

  El asegurar la ventaja de predicción de una metodología como redes neuronales permite incentivar su uso en otras ares de trabajo que pudiesen generar un impacto tan grande en el mercado, como aquel que se presentara en el trabajo actual relacionado a rentabilidades de empresas chilenas.

3 | Marco Teórico

3.1. Antecedentes

En el desarrollo se utiliza como base principal el trabajo realizado por Tsai et al. (2011), el cual realiza la predicción de rendimientos en base a conjuntos de clasificadores (tanto homogéneos como heterogéneos), aplicando dicho procedimiento de predicción a la industria tecnológica de Taiwan, con datos extraídos del Taiwan Economic Journal (TEJ).

Respecto a estudios relacionados, son las redes neuronales el modelo más utilizado en la predicción de los precios de acciones y sus rendimientos, ello puesto que últimamente son las redes neuronales las que demuestran funcionar mejor que los métodos estadísticos usuales.

Dentro de aquellos que puedan entregar mayor fuerza al trabajo a realizar destacan los estudios relacionados de Kim et al. (2006) y Hassan et al. (2007), donde del mismo modo aplicado en el primer trabajo mencionado se realizan combinaciones o ensamblajes de clasificadores.

Tratándose de Kim, se realiza una predicción centrada de manera casi exclusiva en la combinación de múltiples redes neuronales, ello mediante variados métodos de combinación, y aplicando metodologías de decisión como mayoría de votos y otros. Por tanto en cuanto a su trabajo se puede concluir como directamente enfocado en la investigación de un único clasificador y sus diferentes formas de combinación, sirviendo como referencia directa en cuanto a aplicabilidad y ejecución en el modelo homogéneo que se encontrara bajo análisis.

Hassan et al. (2007) y otros trabajos como Albanis y Batchelor (2007), proceden a diferen-

cia de las investigación anterior a combinar ahora no solo redes neuronales, sino más bien se enfocan en clasificadores heterogéneos, ello quiere decir combinar diferentes métodos de clasificación, llegando a ser uno de los actuales estudios que demuestra directamente que dichos clasificadores combinados de manera heterogénea superan a los clasificadores individuales al momento de predecir retornos.

Se menciona el trabajo realizado por Wang et al. (2009), donde enfocan sus esfuerzos en agrupar la gran información financiera que existe, y dentro de las cuales se mide el desempeño de las empresas, con el objetivo de que aquellos interesados en su evolución a través del tiempo como auditores, accionistas, acreedores y analistas financieros puedan utilizar dicha información para distinguir entre el rendimiento posible de una empresa u otra. Para lo anterior se utiliza como método común de análisis los arboles de decisión, ello por ser rápidos y simples de aplicar e interpretar.

Una cantidad de indicadores cercano a los 50 de las mejores empresas en China se somete al análisis estándar de árbol de decisión, buscando mejorar las conclusiones obtenidas al compararlo con el método de combinación de modelos llamado Bagging.

Como resultado empírico se muestra que a diferencia de los modelos estándar, muchas veces estudiados como lo es el árbol de decisión muestran ser inferiores si se compara con su versión mejorada mediante ensamblaje por Bagging, donde este último mejora no solo la precisión del modelo base, también reduce el tiempo de análisis computacional que conlleva todo el proceso.

Mas se asegura que la predicción del modelo puede seguir mejorando, siempre y cuando desde el mercado Chino se pueda obtener una mejor data histórica con el paso del tiempo (que solo comenzaba su registro completo desde el año 2000), y con esto corroborar que el análisis no es estático para el periodo analizado y que por tanto puede avanzar en perfeccionamiento tal cual cambia paulatinamente el mercado.

Y no es solo el mercado de acciones donde se realiza el estudio de las predicciones mediante el aprendizaje automático, pues Jaque (2014) lleva la investigación de modelos de redes neuronales al sector de las AFP en Chile, acotando su estudio a una entidad especifica “AFP CUPRUM”. En el análisis realizado se trabaja en un plazo comprendido desde el 2003 al

2008, comprobando semanalmente como es que cambia el signo de variación en los valores de la cuota de sus multifondos asociados.

Aplicando el modelo desarrollado en Redes Neuronales Ward, se observa que la predicción para los distintos fondos de AFP varía del mismo modo que lo hace el porcentaje de retorno a una inversión ficticia funcionando bajo el mismo modelo.

Se demuestra teóricamente que la capacidad de predicción del modelo permite generar rentabilidades al generar una guía de inversión activa para un inversionista previsional, pues en términos de esa rentabilidad casi todos los fondos logran superar el método de inversión pasiva, excluyendo de tal logro solo al Fondo E.

Nuevamente se da lamentablemente el caso que la certeza asegurada del modelo solo se basa en la muestra y su validación, por tanto no confirma en su totalidad que los valores de las cuotas de los fondos de AFP sigan un camino previsible, pero si asegura que dicha previsibilidad se produzca en algún grado.

De las empresas y acciones del mercado común se traslada la investigación a Turquía (Erdal y Karahanoğlu (2016)), al sector bancario, donde como en muchos otros países existen los conocidos bancos de desarrollo e inversión, los cuales no pueden funcionar como un banco estándar, cuyo principal fuente financiera son los depósitos, por tanto al pertenecer a un subgrupo de bancos en Turquía dichos bancos no pueden aceptar tales depósitos. Lo que hace a tales entidades un objeto de estudio interesante para predecir sus rentabilidades.

En este mercado se utiliza datos financieros trimestrales desde el 2002 al 2014, de 13 bancos, con el fin de comprobar el uso potencial de Bagging, al ser reconocido como uno de los métodos de combinación de clasificadores más populares.

Los tres clasificadores base utilizados son del tipo “arboles de decisión”, demostrando mediante el estudio realizado que el error cuadrado medio tiende a crecer a medida que se utiliza mayor cantidad de clasificadores individuales en el bagging.

Pese a lo anterior si es utilizado un número adecuado de clasificadores el modelo resulta ser prometedor por sobre los modelos básicos, una vez que se encuentre sorteado el efecto

recurrente y negativo de los ruidos, con su correspondiente eliminación.

Propuesto el modelo y frente a las dificultades mencionadas se señala la posibilidad de aplicación de otras metodologías de clasificación junto a las de árboles de decisión en el ensamblaje como lo son las Redes Neuronales.

Un trabajo de lo más cercano a lo planteado en el informe, realizado por Zheng (2006), sugiere la aplicación de metodologías de ensamblaje distintas sobre clasificadores básicos y cuyo alto nivel de precisión sea altamente reconocido, como lo son las redes neuronales, la regresión y la regresión logística.

El estudio propone la aplicación inicial de los clasificadores a modo de registrar su rendimiento individual, para luego unificarlos mediante Boosting y Bagging, que teóricamente son técnicas que mejoran el rendimiento de los algoritmos de aprendizaje.

Las técnicas utilizan como base de datos la información histórica financiera de empresas como Apple Computer Inc. (AAPPL), Microsoft Corp. (MSFT), y otras obtenidas de una base de datos financiera a lo largo de 13 años. Como resultado de aplicar combinación de clasificadores base para predecir el comportamiento financiero de las acciones, se comprueba un aumento en la precisión superior a otros métodos de clasificación como regresión logística. Se comprueba además con datos fuera de la muestra que la diferencia de predicciones se reduce drásticamente al unir modelos y obtener una conclusión global.

Por último se procedió a comparar ambos modelos de ensamblaje, quedando por una diferencia mínima como mejor predictor (ello verificando la varianza en el resultado final) la metodología Bagging.

Los estudios mencionados anteriormente resulta muy útil, pues reunir dichos trabajos realizados y fusionarlos se adecua perfectamente al trabajo planteado a realizar, y de esta manera comparar ensamblajes de clasificadores heterogéneos versus ensamblaje de clasificadores homogéneos basados en redes neuronales (MLP) e incluso compararlos con las metodologías de clasificación aplicadas de manera individual, al momento de predecir los retornos de acciones en el Mercado Chileno, e identificar qué tipo de combinación o

aplicación de clasificadores presenta los mejores resultados.

Como ya es recurrente el identificar trabajos que han dedicado sus esfuerzos en demostrar la superioridad de los conjuntos de clasificadores sobre los clasificadores individuales, resultara aún más interesante el proponer y comparar las nuevas formas de unir los diferentes métodos y compararlos en cuanto a resultados, ello implica usar tanto combinación homogénea como heterogénea y llegar a obtener la conclusión más confiable sobre la inversión en nuestro mercado.

3.2. Terminologías y definiciones

3.2.1. Coeficientes financieros

Término utilizado en finanzas y sistemas financieros para medir y/o evaluar la situación financiera y el desempeño de una empresa. El periodo de tiempo en que se analiza la empresa es fundamental para determinar el estado financiero y cuidadosamente estudiar su salud.

El estudio de índices financieros sirve generalmente para realizar dos tipos de comparaciones: primero es comparar un índice actual con su índice pasado correspondiente o con los probables para el futuro, y en segunda instancia para comparar dichos índices con los de otras empresas similares.

El análisis de aspectos como razón, las tendencias e indicaciones de buenas o malas prácticas en un negocio no resulta complicado de identificar. Respecto a lo anterior se procede a reunir y estudiar los datos tanto históricos como actuales y se obtiene la valoración de la empresa, y con ello predecir bajo una probabilidad de certeza la evolución en el precio de sus acciones, por ejemplo.

3.2.2. Indicadores Económicos

Un indicador económico, o indicador de negocios es una estadística (se utiliza en base a los datos estadísticos) respecto la economía, para mostrar las tendencias generales de la economía. Por tanto, los indicadores económicos permiten analizar el desempeño económico y hacer predicciones de resultados futuros. De lo anterior que una de sus grandes aplicaciones sea el estudio de los ciclos económicos. Existen tres tipos de indicadores económicos, que son los indicadores coincidentes (del ciclo económico), indicadores adelantados e indicadores retardados[1].

- **Los indicadores coincidentes:** se obtienen al mismo momento en que ocurre la actividad económica relacionada, por tanto van en conjunto con el ciclo económico. Se puede utilizar un índice coincidente para identificar las fechas de picos y valles en el ciclo económico.

- **Los indicadores adelantados:** (por ejemplo, precio de las acciones y tasas de interés) poseen valores predictivos que tienden a cambiar antes de que ocurra y se muestre en la actividad económica general. Lo anterior hace que se consideren únicamente como predictores a corto plazo de la economía. Un ejemplo claro es el mercado de valores, el cual de manera general comienza a disminuir antes de que la economía en su conjunto lo haga, y por tanto empiece a mejorar antes que la economía comience a recuperarse de una depresión.

- **Los indicadores retardados:** (por ejemplo, la tasa de desempleo) son los únicos indicadores que se hacen evidentes después de la actividad económica general. Por ejemplo, la tasa de desempleo normalmente disminuye dos o tres trimestres después de un repunte en la economía general.

[1] http://www.enciclopediafinanciera.com/indicadores-economicos.htm

3.2.3. Análisis técnico

El análisis técnico tiene sus origines en Estados Unidos a finales del siglo XIX con Charles Henry Dow, quien crea la Teoría de Dow, la cual adquirió un impulso vertiginoso con la aplicación de Ralph Nelson Elliott dentro de los mercados accionarios con su Teoría de la Ondas de Elliott, y que posteriormente termina extendiéndose al mercado de futuro. Ello no implica que sus principios y herramientas no puedan ser aplicables al estudio de las gráficas de cualquier instrumento financiero.

Dentro del análisis bursátil, el análisis técnico se puede definir como "el estudio de la acción del mercado", lo anterior principalmente mediante el uso de gráficas, con el objetivo de predecir tendencias futuras en el precio. Indicadores como el Índice de Fuerza Relativa (Relative Strength Index o RSI) resultan útiles para pronosticar las tendencias de precios y las decisiones de inversión en el mercado.

El análisis técnico tiene tres principales fuentes de información, mencionándose:

- Precio o cotización: Variable que resulta ser la más importante dentro de la acción del mercado. Normalmente representada a través de una gráfica de barras, en la parte superior de la misma.

- Volumen: Correspondiente a la cantidad de unidades o contratos operados durante un cierto periodo de tiempo. Se representa como una barra vertical bajo la gráfica de cotizaciones.

- Interés abierto: Utilizado primordialmente en futuros y opciones, representando el número de contratos que permanecen abiertos al cierre de un periodo. Representado como una línea continua ubicada entre el precio y el volumen.

El análisis técnico queda dividido en dos grandes categorías, siendo la primera aquella que realiza un análisis gráfico, utilizando estrictamente información revelada en los gráficos, sin utilizar herramientas adicionales. En segundo lugar se encuentra el análisis técnico en su sentido más estricto, pues emplea indicadores calculados en función de diferentes

variables características del comportamiento de los valores analizados.

Si lo comparamos con su principal contraste que es el análisis fundamental, se encuentra una gran diferencia pues el análisis técnico no se centra en el "valor intrínseco" de una población, sino más bien en las extrapolaciones de los patrones de precios históricos.

3.2.4. Clasificación

El proceso de clasificar consiste en asignar el objeto bajo estudio a una de las clases existentes y definidas con anterioridad. Dichos objetos pueden ser definidos por una o varias características, que pueden ir desde tamaño hasta características visuales como el color.

Para clasificar objetos se hace necesario definir con exactitud las fronteras entre cada una de las clases. Normalmente las fronteras se calcularan mediante un proceso de entrenamiento el cual mediante las características de una serie de prototipos establece las diferentes clases. Se hablara de fronteras por claridad siempre que el clasificador infiera las reglas de decisión durante el entrenamiento.

De lo anterior resultara más claro definir que clasificar consiste en asignar un objeto desconocido a la clase, en la cual las características usadas durante el entrenamiento tienen mejor correspondencia con las características del objeto.

Independientemente del tipo de clasificador utilizado, el procedimiento resulta constar de la misma serie de pasos:

- Se reúnen muestras de objetos de estudio, con clases conocidas. Se escoge un conjunto de características, conocido como vector de características, adecuado al modelo que se pretende implementar, obteniéndose finalmente las características de dichas muestras (prototipos).

- El conjunto de vectores construido se utiliza para entrenar el clasificador seleccionado. Definiéndose con ello las fronteras entre cada una de las clases.

- De los objetos desconocidos que se pretende clasificar se extrae la misma información utilizada en las muestras.

- El clasificador ya entrenado utiliza las fronteras establecidas durante el entrenamiento para definir en qué clase será acomodado el vector característica de los nuevos objetos bajo análisis.

Para decidir a qué clase pertenecerá un objeto es necesario conocer las diferencias existentes entre cada clase, y por tanto las similitudes entre los miembros de una misma clase, y para conocer ello es necesario el análisis mencionando de las características de los objetos muestra.

Las características se agruparan en un vector características, el cual a su vez define el espacio de características que se representa en un eje coordenado de cualquier dimensión.

Una vez creado el espacio de características se observa cómo es que los vectores se disponen en tal espacio, buscando consigo conglomerados (también conocidos como clusters) de tal forma que ellos se puedan separar.

Es justamente la tarea de un clasificador el separar los clusters durante el proceso de entrenamiento y asignar de tal manera un vector de características de entrada a cada clase conocida.

3.2.5. Clasificadores individuales

Para construir un solo clasificador, tal cual se definió anteriormente se utiliza un cierto número de muestras de entrenamiento, compuesto de un par de vectores de características (es decir, factores relacionados) y la etiqueta de la clase a la cual se asociara, dejándose definido cada clase con el fin de entrenar el clasificador. En el caso puntual bajo estudio, el rotulo de clase asociado a la muestra de entrenamiento se representara según rendimientos de valores positivos o negativos (0,1).

En otras palabras, la tarea de aprendizaje (entrenamiento) es calcular un clasificador o modelo que se aproxime al mapeo realizado entre los ejemplos de entrada y salida, de manera que la etiqueta asociada a las muestras de entrenamiento se realice con cierto nivel de precisión.

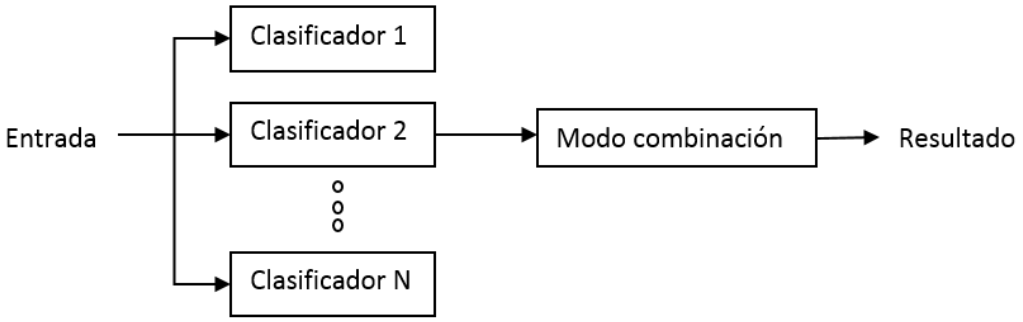
Después de que el clasificador es generado o entrenado, entonces procede a ser capaz de

clasificar una instancia desconocida dentro de una de las etiquetas de clase aprendidas en las muestras de entrenamiento, permitiendo con esto realizar tareas de predicción. Más específicamente, el clasificador calcula la similitud con cada una de las clases entrenadas y asigna la instancia (objeto de análisis) no clasificada a la clase con el porcentaje de similitud más alto.

3.2.6. Conjunto de clasificadores

Un enfoque de clasificación basado en la combinación de múltiples clasificadores individuales, permitiendo con ello obtener clasificadores de alta precisión combinando algunos menos precisos. Los conjuntos o ensambles de clasificación están destinados a mejorar el rendimiento de un clasificador individual. Es decir, la combinación resulta ser capaz de complementar los errores cometidos por los clasificadores en forma individual en diferentes partes del espacio de entrada. Por lo tanto, se supone el rendimiento de ensamblajes de clasificadores mejor que uno de los mejores clasificadores individuales utilizados de manera aislada

Figura 3.1: Ensamblaje de clasificadores



Para construir conjuntos de clasificadores, el conjunto de entrenamiento elegido se utiliza para hacer practicar una serie de técnicas de clasificación. A continuación, el conjunto de pruebas o validación se utiliza para probar los clasificadores de manera individual, de manera que cada uno de los clasificadores producirá una salida resultante a los valores

proporcionados.

Finalmente se escoge un módulo o método de combinación, el cual procesa los resultados obtenidos por cada uno de los clasificadores, generando una salida final la cual se convertirá en el resultado de la clasificación.

Según la literatura, existe una serie de métodos de combinación, mencionándose voto por mayoría, bagging, redes Bayesianas, boosting que da mayor peso a un voto, e inclusive crear nuevos clasificadores que decidan como combinar los resultados (Stacking).

En West et al. (2005) *"Neural network ensemble strategies for financial decision applications, Computers and Operations Research"*, se comparan varios métodos de combinación y en sus resultados se muestra que el método de Bagging proporciona un mejor rendimiento que otros en la predicción de crisis financieras. Por tal motivo utilizaremos en la preparación del documento metodologías de voto mayoritario y métodos de bagging para realizar los ensambles necesarios.

Tabla 3.1: Indicadores financieros y económicos.

Ratios Financieros	
Estructura de capital	Razón deuda activo
Capacidad de Amortización	Liquidez corriente
	Liquidez Ácida
Capacidad operacional del negocio	Rotación de activos
	Rotación de activos fijos
	Rotación de inventarios
	Rotación sobre Cobros
Rentabilidad	EBITDA
	NAV
	ROE
Flujo de caja	Cash flow ratio
Otros	Return on total assets growth ratio
Indicadores económicos	Tasa interés de la facilidad de deposito
	Tipos de cambio (pesos por dólar)
	TPM (%)
	M1
	IPC General
	Tasa desempleo (%)
	Chile: Índice de bonos de gobierno (GBI0E)
	IGPA
	IPSA
	IPI Minería
	IPI Manufactura

3.3. Análisis del Mercado

El mercado chileno se caracterizaba antes de los años 2000 por ser parte de una economía cerrada al comercio mundial y ser regida fundamentalmente por el gobierno del país. Lo anterior involucraba completamente al sector financiero, donde se controlaban tasas de interés, asignaciones crediticias y el banco estaba casi en su totalidad en poder del Estado.

Con el paso de los años las tasas de interés fueron liberadas, se eliminó el control al crédito y el sector bancario paso a mano de privados. Pero como el país no estaba preparado para un cambio tan radical, y debido a la poca regulación que rigiese el cambio se presentó ante los ojos del país la crisis financiera de 1982. La crisis hizo entrar en razón a la población, por lo cual tras superada la situación se mejora el marco regulatorio con una nueva ley de bancos, la cual supervisaría y regularía el correcto funcionamiento del sistema bancario y financiero.

Entre los años 1996 y 2000 el sistema financiero inicia un proceso de liberación y adecuación, autorizándose la inversión en Capitales Extranjeros.

Chile sigue su avance en los años posterior al 2000 donde se promulgan leyes como la Ley de Opas, que mejora la protección a accionistas menores, y se implementa la Reforma al Mercado de Capitales, que entrego incentivos al ahorro, entre otras medida.

En cuanto al mercado accionario, el tamaño de éste ha tenido un crecimiento continuo, desde los años 1996, donde la cantidad de acciones negociadas de acuerdo al PIB fue en crecimiento constante, con un par de caídas para llegar al año 2007 comenzar a presentar el primer indicio de una fuerte caída, caída que se ha ido acentuando durante los últimos años[2].

[2] http://databank.bancomundial.org/data/reports.aspx?source=indicadores-del-desarrollo-mundial#

Chile hoy tiene uno de los sistemas financieros más desarrollados en cuanto a tamaño se refiere si se compara con las economías aún emergentes, sin embargo sería excesivo señalar que nos acercamos a niveles como los que presentan países asiáticos, o países desarrollados como Estados Unidos o Canadá.

Entonces el grado de desarrollo financiero de Chile es todavía menor que el existente en otras economías como ya se mencionaba, ello implica posibilidades aún de crecimiento dentro del sistema financiero y con ello mejorar el crecimiento económico del país.

Pero a pesar del notable desarrollo del sistema financiero de Chile continua siendo débil frente a los shocks en las economías extranjeras. Siendo lo anterior una causa probable de la falta de liquidez en los mercados donde las asimetrías de información y los costos de transacción son muy altos. Y es claro que esas son unas de las primeras causas que generan incertidumbre sobre el verdadero valor de los activos, cuyo diferencial en precio asociado finalmente es sobrellevado en su mayoría por el comprador.

3.3.1. Análisis Financiero

Debido a los cambios en el entorno empresarial, los inversionistas y partes interesadas se ven en la necesidad de adquirir conocimientos mucho más altos que el resto de la competencia, a modo que les permita tomar decisiones rápidas y oportunas, ello requiere de la aplicación de herramientas que se abordaran a lo largo del informe y que permitirán generar una pauta de acción.

Las organizaciones por lo general se encuentran con problemas financieros que no siempre son fáciles de manejar, y por tanto debe enfrentarse constantemente a costos financieros,

riesgos, toma de decisiones poco efectivas, entre otros.

Para evitar o reducir el impacto de los efectos de los problemas financieros es indispensable conocer los principales indicadores financieros y económicos como base en la toma de buenas decisiones. De ello nace el fenómeno de análisis financiero, que le permite determinar no solo su situación financiera, también la del resto del mercado, y mediante su cálculo e interpretación lograr ajustar el desempeño operativo de la organización.

Entre los indicadores financieros utilizados en la metodología a utilizar más adelante en la proyección de rentabilidad de acciones empresariales se puede mencionar:

Indicadores de liquidez y solvencia

Primero resultara necesario hacer la diferenciación entre liquidez y solvencia, siendo la liquidez aquella que implica mantener el efectivo necesario para cumplir o pagar compromisos adquiridos con anterioridad; mientras que solvencia se enfoca en mantener los bienes y recursos necesarios para resguardar las deudas adquiridas, sean o no en efectivo.

Resumiendo lo anterior, tener liquidez implica tener la capacidad de pagar de manera inmediata a los acreedores; en tanto solvencia es la capacidad de responder al corto plazo los compromisos contraídos.

- **Liquidez corriente:** también conocida como razón circulante, y que determina si una empresa tiene la capacidad para cancelar sus deudas en el corto plazo, relacionando mediante una división los activos circulantes con los pasivos circulantes. El que la razón circulante sea alta no implica una disponibilidad inmediata para atender las operaciones de la empresa, pues el valor de dicho indicador puede estar centrado en un inventario (por ejemplo) que no puede ser vendido rápidamente y por tanto tener baja liquidez.

  Si una empresa presenta problemas financieros, suele aumentar sus deudas o atender sus compromisos con mayor lentitud, ubicando el pasivo circulante por sobre el activo, dando un bajo indicador y por tanto reduciendo uno de los más claros representantes de la liquidez de una compañía.

- **Liquidez ácida:** o razón ácida, mide la suficiencia o no que posee una empresa para pagar de forma inmediata sus deudas en un momento dado; similar al indicador anterior solo que ahora no se incluye el inventario, considerado como el activo menos liquido de los activos circulantes.

Indicadores de eficiencia en la actividad empresarial

En una actividad empresarial es fundamental conocer el cómo se utilizan los insumos, los activos y como se gestionan los procesos; para ello es importante señalar que la eficiencia es referida a la relación entre el valor final del producto generado y los factores que permitieron obtenerlo.

- **Rotación de activos totales:** indica la capacidad que posee una compañía de utilizar todos sus activos en la obtención de ingresos, es decir la eficiencia con que se manejan los activos a modo de generar mayores ventas. Su forma de interpretar corresponde al número de veces que una empresa renueva sus activos totales durante su ejercicio. A mayor rotación de activos totales mayor será el resultado en nivel de eficiencia sobre el uso de la empresa de los bienes que posee.

- **Rotación de activos fijos:** su lectura similar al indicador anterior, solo que ahora la eficiencia de la empresa estará referida a la capacidad de la empresa para generar ingresos mediante inversión realizada en activos fijos (edificios, instalaciones, equipos, maquinaria). Su forma de lectura asociada correspondería al número de veces que una empresa renueva sus activos fijos en un año.

- **Rotación de inventarios:** indicador que expresa el número de veces que rota el inventario en un año, y que constituyen la cantidad mínima de artículos disponibles que se requiere para satisfacer la demanda de los clientes. Otra forma de interpretarse corresponde a la capacidad de la empresa de convertir el inventario en efectivo o cuentas por cobrar.

- **Rotación sobre cobros:** en este caso lo que se mide es el número de rotaciones de las compras o ventas a crédito .Así a diferencia de la rotación clientes o proveedores, que son otros indicadores utilizables, la rotación sobre cobros distingue entre operaciones realizadas al contado o a crédito, se centra exclusivamente en lo que significa crédito y por tanto se puede alinear a una mejor gestión.

Indicadores de rentabilidad

Estos indicadores sirven para comparar los resultados de la empresa con distintas partidas del balance o de la cuentas de ganancias y pérdidas. Por tanto miden el nivel de eficiencia en la utilización de los activos de la empresa en relación a la gestión de sus operaciones.

- **EBITDA:** utilizado frecuentemente para valorar la capacidad de una empresa de generar beneficios considerando únicamente su actividad productiva, puesto indica el resultado obtenido directamente por la explotación del negocio, ya que no incluye todos los gastos de la empresa. Su cálculo resulta de restar a los ingresos los gastos, excluyendo gastos financieros.

- **NAV:** equivale al valor de mercado de cierre de todos los valores dentro de un portafolio, después de deducir todas las obligaciones incluyendo gastos y honorarios. ROE: relaciona los beneficios obtenidos netos en una determinada operación de inversión con los recursos necesarios para obtenerla. Se puede considerar como una forma de valorar la ganancia obtenida sobre los recursos empleados.

- **Flujo de caja:** Se puede definir como la acumulación neta de activos líquidos en un periodo determinado y que por tanto constituye un indicador importante al medir la liquidez de la empresa.

- **Ratio flujo de efectivo:** el índice de flujo de efectivo operativo es una medida de cuan bien cubierto se encuentran los pasivos corrientes por el flujo de caja generado por las operaciones de la compañía. El índice mencionado puede medir la liquidez de una compañía a corto plazo y se considera una medida más precisa al considerar únicamente aquello generado por la operación, cuyo valor no puede manipularse.

Otros:

- **Ratio de crecimiento de retorno de los activos totales:** índice que mide las ganancias de una compañía antes de intereses e impuestos frente a sus activos netos totales. El índice se considera un indicador de eficacia con que una empresa utiliza sus activos para generar ganancias antes de que se paguen las obligaciones contractuales.

- **Indicadores económicos:** Los indicadores económicos son valores estadísticos que muestran el comportamiento de la economía, y que por tanto ayudan a analizar y prever el comportamiento de la misma. Dentro de los indicadores utilizados destacan la Tasa de interés de depósito, tasa de cambio, la TPM, IPC, Índice de desempleo, y activos totales de las compañías cotizadas, entre otros; todos extraídos de la base de datos del Banco Central[3] durante un periodo de análisis que se divide en trimestres desde el año 2011 al 2017.

[3] https://si3.bcentral.cl/

3.4. Base de datos

Dentro de las variables que específicamente se utilizan en el actual trabajo, se menciona a diferencias de otros trabajos previos, solo Ratios de tipo financiero, a los que se suman indicadores económicos. No se consideraran indicadores técnicos, de modo de ser fiel al modelo replicado en Chile del trabajo ya referenciado, además de considerar que aquellos indicadores principalmente son utilizados para la predicción de precios y retornos a corto plazo. Reforzando lo anterior el hecho que el análisis está enfocado en trimestres, espacio donde difícilmente se desenvolverían de buena forma los indicadores técnicos. (Tabla 3.1) Destacar que como los datos obtenidos suelen ser valores económicos y/o indicadores porcentuales se procede antes de realizar el moldeamiento de los diferentes clasificadores a normalizar cada uno de los valores. Los indicadores financieros se obtienen de base de datos del mercado chileno, de acuerdo a las 100 mejores empresas calificadas y registradas con transacciones dentro del país. Indiferente del tipo de mercado en que se desempeñase cada una de las compañías, aun cuando cabe destacar que dentro de los indicadores se hace referencia al mercado minero y manufacturero cuya incidencia a nivel global de mercado chileno es elevada.

La base de datos a la que se hace referencia se encuentra conformada por indicadores trimestrales, ello por demostrarse su buena representatividad a nivel general de mercado (Callen et al. (1996)), iniciándose desde el cuarto trimestre del 2011, hasta el primer trimestre del 2017. Como resultado de lo anterior la base de datos queda conformada por 2222 datos, de los cuales se procede a eliminar aquellos cuyos indicadores a ser utilizados no se encuentren registrados en el sistema por un periodo de tiempo muy elevado o de forma reiterada. Ello implica que de la cantidad inicial se redujese a un total de 1231 instancias completamente registradas para cada indicador, valor del cual 527 se clasifican como retornos positivos y 704 como negativos).

El valor del retorno de cada trimestre analizado se encuentra regido por el ROA, que corresponde al uno de los indicadores financieros más importantes y utilizados por las empresas como forma de medición de su rentabilidad, también conocido como Return on

Assets.

En forma resumida corresponde a la relación entre el beneficio logrado en un determinado período y los activos totales de una empresa. Utilizado para medir la eficiencia de los activos totales de una compañía, independientemente de las fuentes de financiación utilizada y de la carga fiscal que el país pueda aplicar en el giro principal de la empresa.

En definitiva el ROA mide la capacidad para generar retorno a su propia inversión (como un todo) por ellos mismos. Su método de cálculo de manera general quedaría del siguiente modo

$$ROA = \frac{BENEFICIO\ NETO}{ACTIVOS\ TOTALES} \tag{3.1}$$

Los retornos generados por el tipo de clasificador a utilizar se distinguen según una etiqueta. Para las salidas entonces se entregara una etiqueta de acuerdo a si el retorno proyectado por el sistema se encuentra por encima o por debajo del valor esperado de retorno. En nuestro caso, al ser un país con economía bastante estable resulta complicado encontrar retornos que pudiesen ser directamente o buenos o malos. Por tal motivo se implementa el criterio del valor promedio. Definiendo un valor promedio al indicador objetivo, y donde si el valor proyectado se encuentra por sobre dicho valor se le asigna la etiqueta “0” (cero), o lo que es lo mismo una rentabilidad positiva. El caso contrario es que el valor proyectado resultase encontrarse por debajo del valor esperado (promedio) y por tanto se le asigna la etiqueta “1” como retorno negativo.

3.5. Modelos de predicción

Del acercamiento previo donde se mencionaba los tres tipos de clasificadores a utilizar, en este trabajo se dará uso a perceptores multi-capa (MLP) que corresponden a un modelo de redes neuronales, arboles de regresión y decisión (CART), y la Regresión Logística (LR).

.

3.5.1. Las redes neuronales (MLP)

De manera conceptual, una red neuronal puede ser explicada como un modelo del tipo matemático que pretende simular el comportamiento del cerebro como un procesador de información que permite analizar, comprender y dirigir una gran cantidad de procesos a una velocidad enorme, uniendo y separando todo aquello que recibe del entorno a modo de compararlo con lo aprendido en épocas previas y de dicha manera dar respuestas adecuadas-

.

Por tanto la intención fundamental de las redes neuronales se puede resumir en generar un poderoso sistema computacional que de manera paralela a otros procesos pueda resolver problemas altamente complejos.

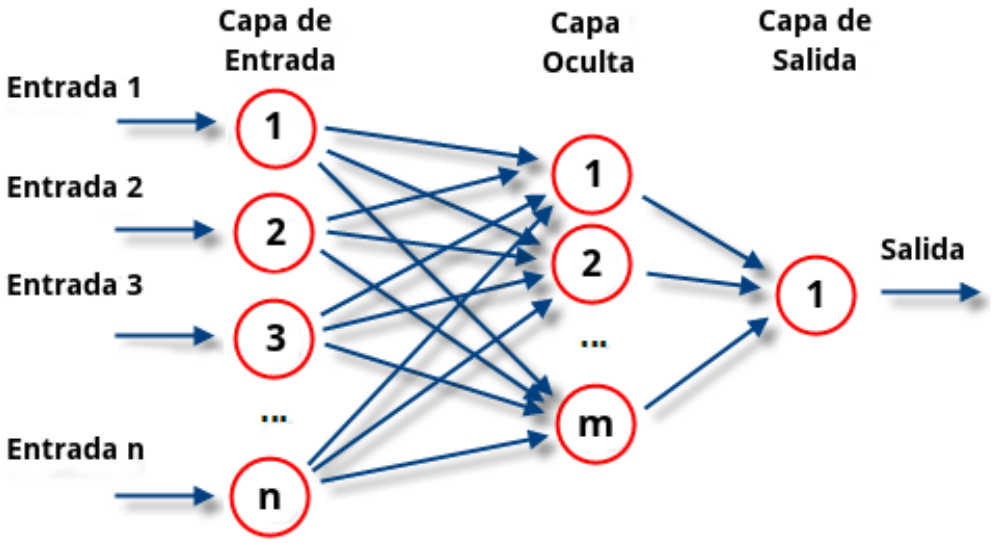
Organizada de una manera tal que su funcionamiento visto con una perspectiva más biológica busca ser la fiel representación del funcionamiento de un cerebro humano, interconectando neuronas y formando una compleja capa o red.

3.5.1.1. Estructura de una red neuronal

La estructura mencionada hace referencia a la forma de organizarse o disponerse las neuronas dentro de una red neuronal. Aclarado lo anterior se puede mencionar que las neuronas artificiales se organizan formando diferentes capas., capas que interconectadas entre si conforman el complejo mundo de las Redes Neuronales.

La estructura básica y muy similar a la que se utiliza dentro del procesamiento a presentar consiste en una capa de entrada, una capa oculta y una capa de salida.

Figura 3.2: Red neuronal clásica



La estructura entonces corresponde a un conjunto finito de neuronas que se encuentran interconectadas, dichas neuronas siguen procesos específicos basados en una función matemática definida.

Estos procesos son enumerados de 1 a N, donde cada proceso recibe uno o varios output resultantes de otras neuronas y que se convertirán en inputs para la función a construir. Lo anterior se convierte en un proceso continuo pues al recibir un input la función deriva en un output que será interconectado al siguiente proceso.

Los procesos que se desarrollan en cada capa son realizados por un grupo de neuronas que comparten los mismos inputs, y por tanto concentraran sus resultados al final de la red o hacia la neurona o capa más próxima.

- **Capa de entrada o input:** Corresponde a la primera etapa de toda la red, pues es en tal capa donde se recibe la información proveniente de las fuentes externas y se encarga de trasladarla al siguiente nivel. Lo mencionado es su única función y por tanto entre sus características mencionar que no procesa la información y es de carácter único, el número de neuronas posibles en esta capa se encuentra definido por la cantidad de inputs.

- **Capas ocultas:** En esta etapa se reconoce el procesamiento de la información recibida por las capas input. El procesamiento del cual se hace referencia corresponde única y precisamente a una función del tipo matemática la cual es definida con antelación como la función encargada de activar las neuronas que igualmente son previamente definidas. Como es de esperar una red neuronal puede poseer más de una capa oculta, y entre ellas pueden comunicarse de manera tanto secuencial como paralela.

- **Capa de salida u output:** Es la capa conformada por las neuronas encargadas de proporcionar la respuesta al problema planteado. Igual que la capa de entrada en el caso de la salida el número de neuronas que en ella se encuentran está definido por la cantidad de outputs.

  Igual que en las capas ocultas en la capa de salida se aplica una función de activación a los datos que se recibe, antes de pasar a entregar los resultados del modelamiento.

- **Pesos sináptico:** El peso sináptico usualmente representado como w_{ij} define la fuerza que presenta una conexión sináptica entre una neurona y otra, la neurona previa (pre sináptica) i y la neurona posterior (post sináptica) j.

  Los pesos sinápticos son una especie de símil a los pesos en una regresión, y puede tomar valores positivos, negativos o cero. En caso de que una entrada sea positiva, el peso sináptico actúa como un excitador, mientras que un peso sináptico negativo correspondería a un inhibidor. En caso de encontrarnos frente a un peso sináptico igual a cero, la comunicación entre neuronas no existe.

- **Regla de propagación:** La regla de propagación determina el potencial resultante de la interacción de la neurona i con las neuronas vecinas. El potencial resultante h_i se puede expresar de la siguiente manera:

$$h_i = \sigma_i(w_{ij}, x_j(t)) \tag{3.2}$$

Siendo la regla de propagación más simple y utilizada se propone el realizar una

suma de las entradas ponderadas con sus pesos sinápticos correspondientes:

$$h_i(t) = \sum_j w_{ij} \cdot x_j(t) \tag{3.3}$$

- **Función de activación:** La función de activación es la que determina el estado de activación actual de la neurona en base al potencial resultante h_i y al estado de activación anterior de la neurona $a_i(t-1)$. El estado de activación de la neurona para un determinado instante de tiempo t puede ser expresado de la siguiente manera:

$$a_i(t) = f_i(a_i(t-1), h_i(t)) \tag{3.4}$$

- **Perceptrón:** Este modelo tiene una importancia enorme, puesto fue el primer modelo en poseer un mecanismo de entrenamiento que permite determinar automáticamente los pesos sinápticos que clasifican correctamente a conjunto de patrones a partir del conjunto de entrenamiento.

  La arquitectura del perceptrón está compuesta por dos capas de neuronas, una de entrada y una de salida. La capa de entrada es la que recibe la información proveniente del exterior y la transmite a las neuronas sin realizar ningún tipo de operación sobre la señal de entrada.

  El algoritmo de entrenamiento del perceptrón se encuentra dentro de los denominados algoritmos por corrección de errores. Este tipo de algoritmo ajusta los pesos de manera proporcional a la diferencia entre la salida real y la proporcionada por la red, con el fin de minimizar el error producido por la red.

  Se puede demostrar que este método de entrenamiento converge siempre en un tiempo finito, y con independencia de los pesos de partida, siempre y cuando la función sea linealmente separable. El principal problema de este método de entrenamiento

es que cuando la función a representar no es linealmente separable el proceso de entrenamiento oscilara y nunca alcanzara la solución. Por lo anterior una función no separable linealmente no puede ser representada por un perceptrón.

- **Perceptrón Multicapa:** El perceptrón multicapa es una extensión del perceptrón simple. Su estructura está definida por un conjunto de capas ocultas, una capa de entrada y una de salida. No existen restricciones sobre la función de activación aunque generalmente se suele utilizar función sigmoidea, con rango [0,1].

$$y = \frac{1}{1 + e^{-x}} \tag{3.5}$$

La operación de un perceptrón multicapa con una única capa oculta puede ser resumida de la siguiente manera:

$$z_k = \sum_j w'_{kj} y_i - \theta'_i = \sum_j w'_{kj} f\left(\sum_i w_{ji} x_i - \theta_i\right) - \theta'_i \tag{3.6}$$

3.6. Arboles de clasificación y regresión (CART)

Los métodos basados en arboles de clasificación (mismo caso arboles de decisión) son muy utilizados actualmente en minería de datos, pudiéndose utilizar para clasificación y regresión. Dichos métodos se derivan de una metodología existente de manera previa conocida como automatic interaction detection. Su utilidad destaca en la exploración inicial de datos y resulta ser bastante apropiado cuando hay un número significativo de datos y existe incertidumbre en la manera en la cual las variables explicativas debiesen introducirse y afectar el modelo. Lo anterior da a considerar la existencia de un muy buen método de clasificación, pero es de importancia mencionar la herramienta no asegura ser una

herramienta demasiado precisa de análisis.

Si el conjunto de datos a ser analizado es pequeño, resultara bastante complicado que la metodología revele la estructura existente entre ellos, de modo que su mejor aplicación tal cual se mencionó previamente se encuentra en base de datos grandes, donde métodos convencionales de regresión suelen fallar o complicar un análisis acertado.

Entre los problemas donde su uso es más recurrente suele mencionarse la regresión con una variable dependiente y continua, regresiones del tipo binaria (donde un modelo de regresión estándar no aplica), problemas de clasificación con categorías múltiples ordinales y nominales.

Como ventajas primordiales se señala que el resultado entregado no es variante frente a una transformación monótona de las variables explicativas utilizadas. La metodología se adapta fácilmente donde aparecen datos "perdidos" sin necesidad de eliminar la observación completa, y puede incluir tanto modelos de clasificación generales como modelos de regresión.

Como desventaja se encuentra que el modelo global resultante puede no ser óptimo, pues solo se asegura que cada subdivisión del árbol es óptima. Cuando un árbol es demasiado grande se pierde en alguna medida las interacciones con su elemento predecesor, lo cual genera que las predicciones tengan muchas veces cierto carácter de "desconocido".

Suponemos entonces que la muestra bajo análisis y por ende su porción de entrenamiento involucra toda la información que refleja al grupo en cada caso y por tanto permitirá construir un adecuado criterio de clasificación.

3.6.1. Funcionamiento

Se comienza con un nodo de partida, y se procede a dividir el conjunto de datos entrantes en dos partes homogéneas utilizando una de las variables involucradas. Dicha variable será escogida de manera que la partición de datos se realice de la manera más homogénea posible.

Se elige, por ejemplo la variable x1 y se determina un punto de corte, por ejemplo “d” a modo que se pueda separar el conjunto en dos partes, aquellos con x1 mayor a “d” y lo que sean menores o igual a tal valor. Por tanto de nodo inicial se generan dos caminos, según el criterio antes mencionado, para que luego en cada nodo objetivo se repita el mismo proceso de seleccionar una variable y un punto de corte para volver a dividir la muestra.

El proceso concluye una vez que la metodología determina que cada observación fuese clasificada correctamente en su grupo.

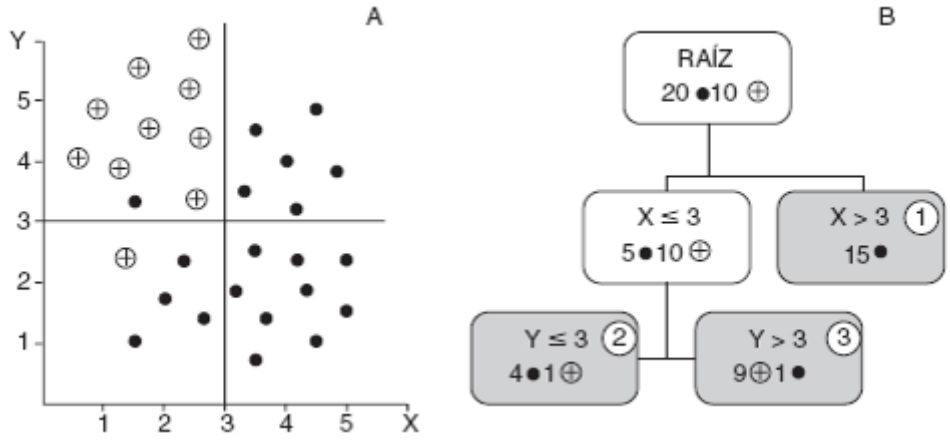
En la construcción de un modelo CART por tanto será necesario:

- Seleccionar las variables de análisis y sus puntos respectivos de corte para realizar las divisiones mencionadas.
- Definir un criterio que permita concluir si un nodo es terminal o si debe continuar dividiendo.
- La asignación de una clase al momento de definir un nodo terminal.

Se puede asignar a cada nodo la cantidad de subconjuntos o datos que pueden pasar por él. Para ello debe estar premeditadamente indicada la variable que va a utilizarse en la partición del nudo, para luego definir el número de observaciones que tendrán paso, mediante una medida de entropía o diversidad probable que llega a cada nodo.

Como la variable que se asigna o introduce en un nodo es aquella que minimiza la heterogeneidad resultante de la división en el nodo, es el mismo modelo el cual una vez minimizado lo suficiente la variabilidad decida el momento de detener el paso de datos.

Figura 3.3: Esquema de partición en la construcción de un árbol.



3.7. LR o regresión Logística

Los métodos de regresión de variable dependiente cualitativa abarcan diferentes modelos que de una forma u otra tratan de explicar y predecir una característica cualitativa a partir de los datos de otras variables conocidas, las cuales pueden ser del tipo cualitativa o cuantitativa.

La característica que se quiere explicar puede ser del tipo binomial (más frecuente), una cualidad que puede ser multinomial, una característica que puede representar modelos ordenados y modelos que pueden resultar del tipo anidado.

Como es conocido el concepto de regresión hace referencia a una fórmula matemática que traduce la relación entre diferentes variables a un resultado o respuesta final. Esta función a diferencia del método de regresión lineal, utiliza normalmente el método de mínimos cuadrados y aritméticamente hablando resulta en un funcionamiento fluido y natural.

Cuando la variable independiente solo puede tomar dos valores (ocurre o no ocurre una instancia), al evaluar la función para valores específicos se obtendrá un resultado diferente de 0 o 1 , lo cual pierde todo el sentido de respuesta y es en dicho caso donde la regresión lineal es inmediatamente descartada, mientras que la Regresión Logística se ajusta bastante bien a la situación.

La función logística encuentra para cada prueba y según los valores proporcionados de

una base de datos como independiente, existe una probabilidad definida que el resultado entregado se adecue correctamente al resultado de estudio. Una transformación logarítmica de dicha función, la cual recibe el nombre de Logit, consiste en convertir la probabilidad en ventajosa.

Esta metodología se utiliza para definir si una o más variables explican una variable que toma finalmente un valor cualitativo.

Su utilización en la predicción es su uso más frecuente y conocido, enmarcado en estudios de diferente naturaleza, destacando la capacidad de predecir riesgos, y servir para estimar la fuerza de asociación de un factor o variable independiente.

Como no es pretensión de trabajo presente el desarrollar a profundidad todas las posibilidades de interpretación y forma de aplicación de cada uno de los modelos, puesto que cada metodología mencionada y no mencionada se restringirá de manera inicial pero finalmente será aplicada por un software computacional. Por lo que de manera muy superficial cuando se realiza una RL lo que se pretende es estimar el valor de los pesos que acompaña cada variable independiente dentro de una ecuación $(B_o, B_1, B_2, \ldots B_k)$, tal cual sigue:

$$Y = B_o + B_1 X_1 + B_2 X_2 + \ldots + B_k X_k \tag{3.7}$$

Donde la ecuación 3.7 tiene a Y como el logaritmo natural del objetivo o escenario resultado del estudio; B_o es la ordenada en el origen de la función de regresión; $B_1, B_2, \ldots, B_k$ representan los coeficientes que determinan el impacto de las variables independientes en el modelo.

Si la base de datos integrada al modelo se ajusta de manera satisfactoria al modelo proyectado mediante el conjunto de prueba, se puede decir que el modelo tiene la capacidad de explicar o inclusive proyectar una respuesta de manera muy segura.

Cabe destacar que los coeficientes B_k son iguales al logaritmo natural del coeficiente original del modelo planteado., por lo que si se quiere obtener el valor original será necesario utilizar el logaritmo inverso.

El enfoque principal del trabajo se encuentra en el modelamiento en primera instancia de

la red neuronal MLP, pues resulta ser uno de los métodos más utilizados y con resultados bastante aceptables.

Para aplicara MLP se necesita considerar dos condiciones o formas de trabajo a aplicar:

Primero se debe escoger la arquitectura concreta de la red, donde para hacer referencia y ser fiel al trabajo realizado por Fong Tsai y su equipo de trabajo. En dicha investigación se escoge:

- Como es posible el sobre-entrenamiento durante la etapa misma de entrenamiento, es que se diseñan 7 diferentes tipos de épocas de aprendizaje, las que incluyen 150, 300, 1000, 1500, 3000, y 5000.
- Misma situación al momento de escoger la cantidad de nodos ocultos, donde se incluye 5, 10, 15, 20, 25 y 30.

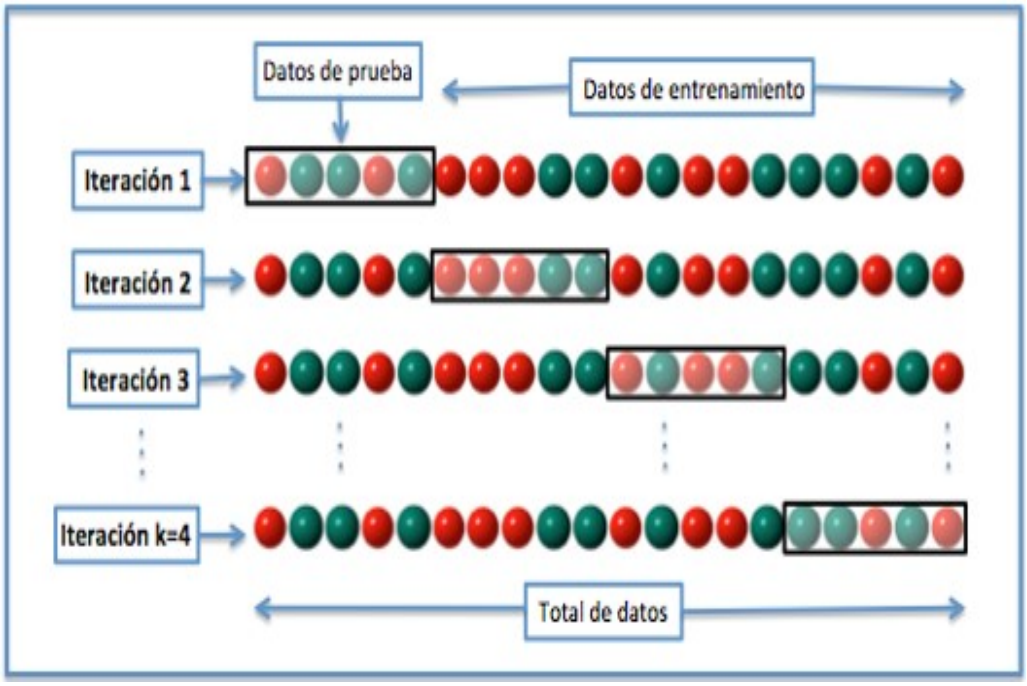
Lo anterior resumido en el modelo básico de red neuronal donde se poseen 3 capas, entrada, capa oculta, y capa de salida.

En segundo punto se decide aplicar validación cruzada al modelo, donde dicha validación cruzada es una técnica utilizada comúnmente para evaluar los resultados de un análisis del tipo estadístico y garantizar de alguna forma que son independientes de la partición entre los datos que se le entregan durante el entrenamiento y prueba.

3.8. Validación cruzada

La **validación cruzada** consiste en dividir en K subconjuntos el conjunto completo de datos utilizado para la construcción de la red neuronal. Por tanto uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento. El proceso de validación será repetido durante k iteraciones, con cada uno de los posibles subconjuntos de datos de prueba. Finalmente mediante una media aritmética de los resultados de cada iteración se obtiene un único resultado.

Figura 3.4: Validación Cruzada



Nosotros utilizaremos validación cruzada con 5 subconjuntos pues la variabilidad de las muestras puede llegar a afectar de algún modo el rendimiento de los tres modelos presentados. Se debe tener presente que no se utiliza el estándar de 10 subconjuntos pues se generaría una muestra demasiado pequeña cada vez que se quisiera realizar la validación, a lo que se suma el extenso tiempo de análisis frente a las otras variaciones en el modelo como el numero alto de nodos.

Utilizando el método de validación cruzada de 5 subconjuntos, lo que se hace es dividir la base de datos en 5 partes iguales. Independiente de cual sea la parte seleccionada, las 4 restantes se utilizan para entrenar la red. El segmento restante es ejecutado para validar el modelo.

Como resultado de lo anterior es que cada parte será entrenada y probada 5 veces.

Como resumen de la estructura generada se generaran 210 diferentes modelos de MLP (7 épocas de aprendizaje X 6 números diferentes de nodos X 5 etapas en validación cruzada). Quedando finalmente con el modelo cuyo indicador resultante presente el mejor rendimiento.

4 | Metodología

Haciendo referencia a lo planteado inicialmente el objetivo del trabajo presente es comparar las metodologías de clasificación, pero no enfocando la vista ente encontrar un método de predicción más acertado que otro, lo cual puede ser situacional 100 % a la forma de plantear cada uno de los tres modelo. Es por lo anterior que en primera fase nos enfocaremos en estudiar los clasificadores por separado.

En segunda fase el interés será ensamblar (unir) los métodos de clasificación en distintas formas, mediante bagging tanto homogéneamente, ello es repetir el mismo tipo de clasificador con distintos parámetros y mediante la conclusión de cada uno de los clasificadores y por metodología de mayoría de votos se seleccionara el resultado al conjunto de validación y con ello el rendimiento del modelo en conjunto.

En última instancia se procederá a unir los modelos nuevamente por método de bagging pero ahora de manera heterogénea, ello es por ejemplo unir la metodología de CART, MLP, y LR en un modelo general y verificar su rendimiento.

4.1. Método de Evaluación

Para evaluar el rendimiento de los clasificadores tanto de manera individual como en ensamble se utilizara la matriz de confusión, la cual resume de manera directa el número de aciertos y errores de la predicción respecto al valor esperado.

A la misma matriz se procede a agregar un indicador de precisión promedio, que obtiene el rendimiento del modelo al realizar una ponderación de los valores acertados sobre el universo total de valores proyectado.

Tabla 4.1: Matriz de confusión

Matriz de confusión	**Predicción**	
Dato observado	Positivo	Negativo
Positivo	**Verdadero Positivo (A)**	Falso Negativo (B)
Negativo	Falso Positivo (C)	**Verdadero Negativo (D)**

$$\textbf{Precisión Promedio} = \frac{A + D}{A + B + C + D} \tag{4.1}$$

El resumen de rendimiento se presenta entonces mediante la matriz de confusión (Tabla 4.1), donde el valor **A** corresponde a aquellas predicciones que resultan en nuestro caso con retorno positivo y su valor real coincide exactamente con ser positivo; **B** son predicciones que dieron negativas y su valor resultaba ser positivo, ello implica en términos simples un falso negativo; **C** corresponde a predicciones positivas pero con valor real negativo; y finalmente **D** que son aquellos resultados que tanto predichos como reales resultan ser negativos.

5 | Resultados experimentales

5.1. Clasificadores individuales

Se presenta un resumen de los clasificadores trabajados cada uno de manera individual, representando en cada caso su efectividad al momento de concertar los datos esperados con aquellos proyectados.

Tabla 5.1: Resultados MLP

Resultados MLP			**Precisión Promedio**
	0	**1**	
0	**77,31 %**	22,69 %	**76,23 %**
1	24,85 %	**75,15 %**	

Tabla 5.2: Resultados CART

Resultados CART			**Precisión Promedio**
	0	**1**	
0	**91,60 %**	8,40 %	**84,55 %**
1	22,49 %	**77,51 %**	

Tabla 5.3: Resultados LR

Resultados LR			**Precisión Promedio**
	0	**1**	
0	**60,50 %**	39,50 %	**59,30 %**
1	21,89 %	**78,11 %**	

Estos resultados permitirán generar la base de comparación a realizarse entre los ensambles homogéneos y heterogéneos. Por tanto también se puede generar un pequeño acercamiento a la comparación de metodologías de ensamble homogéneamente y heterogéneamente, mediante mayoría de votos y bagging respectivamente.

5.2. Ensamble de clasificadores homogéneos y Heterogéneos

Destacar en primer lugar que en la elaboración de cada uno de los modelos se utiliza el programa WEKA de la universidad de Waikato en Nueva Zelanda, que corresponde a un software de libre licencia para la construcción y desarrollo de diversos métodos de aprendizaje.

En el diseño del ensamble de clasificadores, se utilizara los tres tipos de clasificadores antes mencionados MLP, CART y LR. Ellos se combinaran de acuerdo a su rendimiento resultante la validación cruzada de 5 etapas, donde se selecciona al momento de realizar la agrupación aquellos con mejor rendimiento general

Tabla 5.4: Modelos de Red Neuronal (MLP) con mejor rendimiento

Rank	**Retorno positivo (%)**	**Retorno negativo (%)**	**Promedio**	N° **de nodos ocultos**	**Etapas de aprendizaje**
1st	77,31 %	75,15 %	76,23 %	5	150
2nd	74,79 %	73,96 %	74,38 %	20	150
3rd	76,47 %	71,60 %	74,03 %	10	150

5.3. Mayoría de votos

Para el caso de los clasificadores a ser ensamblados homogéneamente se selecciona los mejores modelos de acuerdo al nivel de precisión demostrado en la etapa de validación, para realizarse un resumen de acuerdo a la cantidad de aciertos y errores que se presentan al momento de cruzar los resultados obtenidos con aquellos esperados. Para el caso de MLP, aquellos tres modelos que poseen el índice de precisión más alto quedaran representativos sobre los otros 210 modelos seleccionados. (Tabla 5.4)

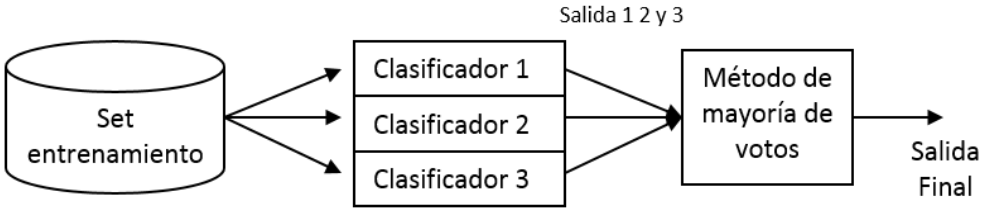
La estrategia planteada consiste una vez identificados los tres mejores modelos MLP hacer la comparación una vez realizados los ensambles de clasificadores homogéneos.

Recordar que como el método de ensamble a utilizar es el criterio de decisión de **mayoría de votos**, el cual ensamblara los tres mejores modelos del universo generado en MLP.

En el caso de ensamble heterogéneo por mayoría de votos se realizara un cruce directo entre los mejores modelos resultantes de MLP, CART y LR.

Para mayoría de votos y la metodología de ensamble independiente si es homogénea o heterogénea, el resultado final se basara únicamente en las etiquetas de salida de los tres modelos que reciben dos votos. Por ejemplo un conjunto de clasificadores heterogéneos, donde se tiene que el MLP y LR arrojan un “0” y el modelo CART con su mejor resultado emite un “1”, el resultado para el conjunto de los tres clasificadores resulta ser “0”. Jaque (2014)

Figura 5.1: Ensamble de clasificadores por mayoría de votos



Al momento de realizarse el ensamblaje tanto homogéneo como heterogéneo, donde en este último se consideran las 3 metodologías de clasificación mencionadas (siempre aquellas con mejor rendimiento), se obtiene el siguiente resumen.

Tabla 5.5: Rendimiento del ensamblaje Homogéneo de 3, 5 y 7 MLP's

Sets de entrenamiento	Retorno positivo %	Retorno negativo %	Precisión promedio
3	54,62 %	79,29 %	66,96 %
5	58,82 %	86,39 %	72,61 %
7	59,66 %	85,21 %	72,44 %

Tabla 5.6: Rendimiento del ensamblaje Heterogéneo MLP, CART y LR

No de set de entrenamiento	Retorno positivo (%)	Retorno negativo (%)	Precisión promedio	Estrategia de combinación
3	83,19 %	75,74 %	79,47 %	MLP x 1, LR x 1, CART x 1

La tabla 5.5 hace un pequeño acercamiento al ensamblaje de clasificadores, en este caso mediante "mayoría de votos", y se convierte en la primera instancia que va dejando claro cómo que las metodologías efectivamente se ven afectadas al momento de unificarlas. Destacar que mejora la precisión al proyectar retornos negativos, pero reduce bastante su rendimiento en la predicción de retornos positivos, afectando la precisión promedio, que se acerca pero no logra superar la precisión promedio del mejor MLP funcionando de manera individual.

Distinto es el caso de ensamble Heterogéneo (Tabla 5.6), donde la precisión de cada instancia (positiva, negativa y general) de retornos se ve mejorada respecto al modelo individual de clasificación.

5.4. Bagging

Es considerada una de las metodologías más antiguas, simple y bien conocida por permitir crear un ensamble de clasificadores. En bagging la diversidad se obtiene construyendo cada clasificador con un conjunto de ejemplos diferente que se obtienen seleccionando elementos mediante reemplazo del conjunto de ejemplos original. Por lo tanto, para construir un ensamble basado en bagging formamos diferentes conjuntos de ejemplos y le aplicamos a cada uno de ellos el algoritmo base. West et al. (2005)

Para el método de combinación bagging necesitaremos un conjunto de clasificadores independientes a tomar en cuenta para la construcción del modelo, es por tal motivo que se procede a construir 3, 5 y 7 combinaciones de clasificadores tanto homogéneos como heterogéneos para la aplicación de bagging.

Como ejemplo tenemos el MLP, en el cual se utiliza el mismo criterio presentado en "mayoría de votos", donde se escogerán ahora los 3, 5 y 7 mejores modelos obtenidos del universo de 210. Ello con el fin de replicar lo realizado por el otro método de ensamblaje y poder realizar un contraste directo de metodología de ensamblaje pero a nivel homogéneo.

De lo anterior:

Tabla 5.7: Rendimiento de Ensamblaje Homogéneo de 3, 5, 7 MLP

Sets de entrenamiento	Retorno positivo %	Retorno negativo %	Precisión promedio
3	39,50 %	95,86 %	67,68 %
5	44,54 %	96,45 %	70,49 %
7	43,70 %	96,45 %	70,07 %

La tabla 5.7 resume el resultado obtenido de realizar un ensamblaje de tipo homogéneo, pues se combina únicamente metodología de clasificación MLP, seleccionando los 3, 5 y 7 modelos con mejor desempeño.

De ella se puede observar que la precisión en el cálculo de retornos positivos se ve bastante reducida respecto al mejor modelo de MPL individual, misma situación que se genero con el ensamblaje homogéneo por "mayoría de votos", más si se continúa analizando, el efecto es inverso en el cálculo de retornos efectivamente negativo, cuya precisión aumenta

considerablemente, pero que aún no logra sobrellevar el resultado de la precisión general.

De este cuadro no se puede obtener información en gran profundidad, pero si nos permite señalar que el modelo en combinación ya presenta una notable mejora en la precisión de resultados negativos respecto de la metodología de clasificación aplicada de manera individual.

Ahora si este cuadro es comparado directamente con el modelo homogéneo de ensamblaje mediante Mayoría de Votos, existe una mejora considerable en la correspondencia de retornos negativos proyectados con los reales, mas no logra mejorar lo suficiente el retorno positivo , lo que lleva aparejado un cercano pero no mejor resultado general.

En el caso de combinar metodologías de clasificación, o lo que es lo mismo ensamble heterogéneo, mediante bagging, se utilizara 3 tipos de set combinados. Donde la estrategia de combinación es directamente aleatoria, pero sigue la lógica de siempre ir en incremento de MLP hasta llegar a los mismos 3 mejores modelos de regresión multicapa.

Tabla 5.8: Rendimiento de Ensamblaje Heterogéneo de 3, 5, 7 MLP's

N° **de set de entrenamiento**	**Retorno positivo (%)**	**Retorno negativo (%)**	**Precisión promedio**	**Estrategia de combinación**
3	81,51 %	89,94 %	85,73 %	MLP x 1, LR x1, CART x 1
5	75,63 %	91,12 %	83,38 %	MLP x 2, LR x2, CART x 1
7	59,66 %	95,27 %	77,47 %	MLP x 3, LR x2, CART x 2

Se observa de la tabla 5.8 que el mejor porcentaje de aciertos en retornos positivos y negativos se ven aparejados en el primer set de entrenamiento (MLP x 1, LR x 1, CART x 1), el cual demuestra tener una precisión de representatividad respecto al resultado real bastante alta, más aún parece corresponder al mejor resultado general de predicción. Ello comparándolo tanto con los modelos combinados por "mayoría de votos"" homogénea y heterogéneamente, e inclusive con los modelos realizados de manera individual de cualquier tipo.

6 | Conclusiones

El problema de predecir el rendimiento de las acciones ha sido un tema importante de análisis y estudio durante muchos años en el área financiera. El avance en la tecnología informática ha permitido que muchos estudios recientes utilicen técnicas de aprendizaje de máquinas como redes neuronales y árboles de regresión para predecir los retornos de variadas existencias.

Dentro de las metodologías utilizadas para obtener las diferentes predicciones, las conclusiones suelen ser muy variadas en cuanto a rendimiento de dicha predicción, destacándose las mencionadas como Redes neuronales. Y es en esta área del aprendizaje automático donde surgen los conjuntos de clasificadores, es decir, la combinación de múltiples clasificadores, que han demostrado ser un método superior a los clasificadores individuales. Con el fin de construir un mejor modelo para predecir el rendimiento de las acciones en el mercado chileno de manera eficaz y eficiente es que se planteó inicialmente esté trabajo, donde como objetivo se investigar el rendimiento promedio de tal predicción que utiliza el método de ensamblar clasificadores para analizar el rendimiento de diferentes empresas. En particular, se consideran los métodos heterogéneos de ensamble, donde mediante una metodología conocida como “Bagging” y la clasificación referida a la “mayoría de votos”, donde mediante simulación de clasificadores tanto homogéneos como heterogéneos se buscó verificar su beneficio superior sobre una predicción con metodología de clasificación individual.

Es por lo anterior que el rendimiento se proyectó desde tres tipos de clasificadores (metodologías), comparados entre ellos y al momento de ensamblarlos de dos formas para dar

una conclusión de retorno de inversión que puede resultar ser positiva o negativa.

Al momento de utilizar conjuntos de clasificadores se presentó la preferencia de utilizar combinación "homogénea" de MLP (por ejemplo, un conjunto generado en base a 3 MLP diferentes) y conjuntos de clasificadores "heterogéneos" (por ejemplo, un conjunto que mezcla MLP, árboles de regresión y regresión logística), en diferentes proporciones.

La manera primordial de concluir sobre el rendimiento de la predicción de retorno se basó primordialmente en examinar la cantidad de rendimientos acertados positivos y negativos por sobre un universo de datos con resultado conocido.

Mediante el software computacional WEKA se concluyen rendimientos de inversión, asignándoles una etiqueta que se compara con la etiqueta conocida externamente al modelo.

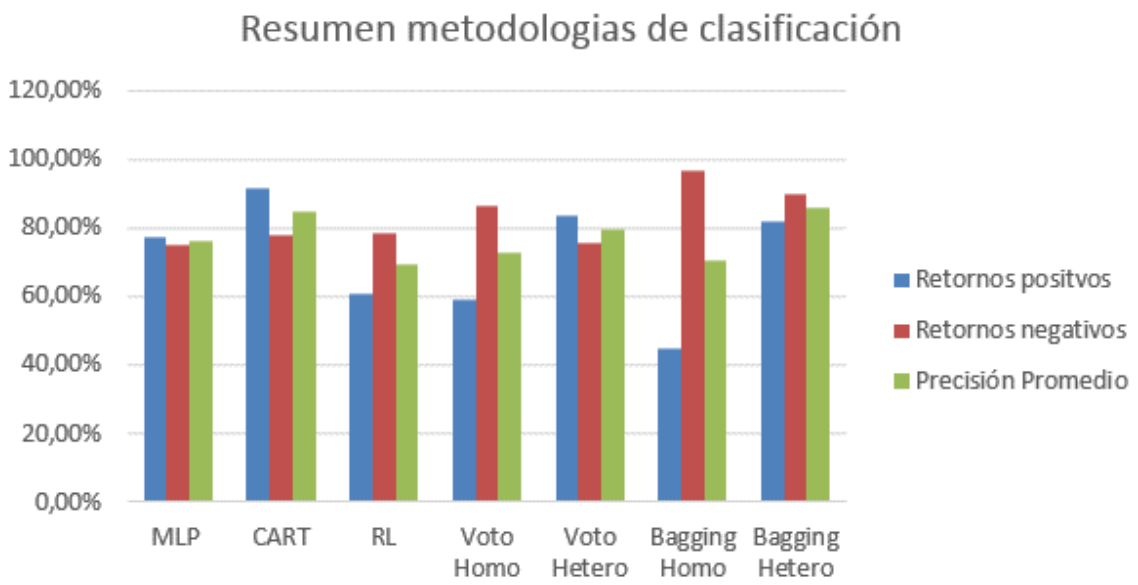
Tabla 6.1: Resumen de metodologías aplicadas con ranking

	Retornos positivos	**Retornos negativos**	**Precisión Promedio**	**Ranking**
MLP	77,31 %	75,15 %	76,23 %	4
CART	91,60 %	77,51 %	84,56 %	2
RL	60,50 %	78,11 %	69,31 %	7
Voto Homo	58,82 %	86,39 %	72,61 %	5
Voto Hetero	83,19 %	75,74 %	79,47 %	3
Bagging Homo	44,54 %	96,45 %	70,49 %	6
Bagging Hetero	81,51 %	89,94 %	85,73 %	1

Se observa de la Tabla 6.1 que de entre todos los modelos trabajados de manera individual, aquel que resulta en una precisión mayor resulta ser el CART, pero tal cual se menciona al definir dicho modelo, existe una lata posibilidad de que este método por naturaleza del mismo, se encuentre mal representado.

Pese a lo anterior aquel que recibe la mejor calificación, seguido por la metodología CART es el modelo de Redes neuronales multicapa (MLP), que reafirma lo planteado inicialmente de poseer una alta capacidad predictiva.

Figura 6.1: Gráfica resumen conjunto de metodologías



De la Figura 6.1 se puede observar que aquella metodología predominante en acertar retornos negativos es el ensamblaje mediante Bagging de manera homogénea, pero que pese a llevar un buen grado de acierto se ve aplacada enormemente por la disminución en la calidad de predicción de retornos positivos.

La metodología individual de clasificación MLP (mejor modelo obtenido), sigue un patrón de predicción bastante certero de igual forma en retornos positivos como negativos, lo cual lo re-afirma como metodología de predicción estrella en este tipo de estudios.

A nivel de precisión promedio el modelo a ser considerado como el "mejor" por sobre los demás resulta ser el ensamblaje por Bagging Heterogéneo. Ello implica que ensamblar clasificadores permite reducir enormemente la cantidad de errores en la predicción, y posiciona dicho procedimiento por sobre la aplicación de una clasificación individual, la cual por variadas condiciones se puede ver sobre o infravalorada.

De lo anterior el mercado Chileno, con retornos en general beneficiosos a la inversión parece ser posible de ser representado mediante una metodología de clasificación de aprendizaje automático como lo es MLP y de mejor modo su combinación con otros métodos de clasificación.

Si se comparan los resultados obtenidos al realizar el análisis en el mercado Chileno con

aquellos trabajos enfocados en otros mercados del mundo se obtiene que:

Respecto al mercado de Taiwán, del trabajo de Tsai et al. (2011), sus resultados señalan que una metodología de ensamblaje conjunta presenta mejores resultados a nivel de predicción que las metodologías básicas como redes neuronales, ello pues disminuye de manera significativa la varianza entre los retornos proyectados y los esperados. Más aún la mejor predicción la obtiene al realizar dicho ensamblaje mediante la metodología Bagging, superando la "mayoría de votos" o el estandarizado método de Buy and Hold. De la aplicabilidad en este mercado igualmente ocurre en el mercado Chileno, no se asegura al 100 % la aplicabilidad del modelo de manera real, pero si se presenta la posibilidad de guiar de cierta forma la decisión al momento de invertir e indica que de alguna forma el comportamiento regular del mercado es predecible por factores financieros y/o económicos.

Del mercado de Turquía y su análisis relacionado al trabajo de Erdal y Karahanoğlu (2016) donde aplica únicamente arboles de decisión en la proyección de rentabilidades bancarias, el resultado obtenido en el mercado Chileno termina siendo finalmente el mismo, pues tras realizarse el ensamblaje por Bagging la precisión se mejora de manera importante, confirmándose ya en un tercer mercado la primera parte de la investigación que planteaba el mejor funcionamiento de una metodología de ensamblaje, y complementándose una propuesta realizada del trabajo en Turquía, sobre aplicar bagging pero de forma Heterogénea.

Finalmente al comparar los resultados obtenidos con los de China del trabajo de Wang et al. (2009) el bagging se reafirma como metodología superior a las metodologías de predicción usuales como arboles de decisión, ello para predecir el comportamiento del mercado de inversiones, el cual en principio seguiría una lógica de predicción basada en datos históricos.

Resultará necesario mencionar que a diferencia de muchos de los trabajos mencionados donde el bagging como metodología de ensamblaje resultaba ser la que mejor resultados (respecto a precisión) presentaba, en el trabajo presente se integra la comparación entre dicha metodología aplicada a clasificadores reunidos de manera heterogénea y homogénea, quedando como "mejor" modelo de predicción aquel que utilice combinación heterogénea

(de un número acotado de clasificadores), específicamente de Redes Neuronales, arboles de clasificación y Regresión logística.

Bibliografía

Albanis, George y Batchelor, Roy (2007). Combining heterogeneous classifiers for stock selection. *Intelligent Systems in Accounting, Finance and Management*, 15(1-2), 1–21. 3.1

Callen, Jeffrey L; Kwan, Clarence CY; Yip, Patrick CY; y Yuan, Yufei (1996). Neural network forecasting of quarterly accounting earnings. *International Journal of Forecasting*, 12(4), 475–482. 3.4

Erdal, Hamit y Karahanoğlu, İlhami (2016). Bagging ensemble models for bank profitability: An emprical research on turkish development and investment banks. *Applied Soft Computing*, 49, 861–867. 3.1, 6

Hassan, Md Rafiul; Nath, Baikunth; y Kirley, Michael (2007). A fusion model of hmm, ann and ga for stock market forecasting. *Expert systems with Applications*, 33(1), 171–180. 3.1

Jaque, Mauricio Alarcón (2014). *MODELOS DE REDES NEURONALES APLICADO EN LA PREDICCIÓN DEL SIGNO DE LOS FONDOS DE AFP CUPRUM (2014)*. PhD thesis, uchile. 3.1, 5.3

Kim, Myoung-Jong; Min, Sung-Hwan; y Han, Ingoo (2006). An evolutionary approach to the combination of multiple classifiers to predict a stock price index. *Expert Systems with Applications*, 31(2), 241–247. 2, 3.1

Kimoto, Takashi; Asakawa, Kazuo; Yoda, Morio; y Takeoka, Masakazu (1990). Stock market prediction system with modular neural networks. In *Neural Networks, 1990., 1990 IJCNN International Joint Conference on* (pp. 1–6).: IEEE. 2

Lin, Wei-Chao; Tsai, Chih-Fong; Ke, Shih-Wen; y You, Mon-Loon (2015). On learning dual classifiers for better data classification. *Applied Soft Computing*, 37, 296–302. 2

Luis Ayala Jiménez, Sebastián Letelier González, Pablo Zagal Morgado (2009). *Modelo de Redes Neuronales para la Predicción de la Variación del Valor de la Acción de First Solar (2009)*. PhD thesis, uchile. 2

Ortiz, Luis Eduardo Meneses (2008). *Modelo para estructurar portafolios de inversiones en acciones mediante redes neuronales (2008)*. PhD thesis, UTP. 2

Sepúlveda, Juan Felipe Díaz y Correa, Juan Carlos (2013). Comparación entre árboles de regresión cart y regresión lineal. *Comunicaciones en Estadística*, 6(2), 175–195. 2

Tsai, Chih-Fong y Chen, Ming-Lun (2010). Credit rating by hybrid machine learning techniques. *Applied soft computing*, 10(2), 374–380. 2

Tsai, Chih-Fong; Hsu, Yu-Feng; y Yen, David C (2014). A comparative study of classifier ensembles for bankruptcy prediction. *Applied Soft Computing*, 24, 977–984. 2

Tsai, Chih-Fong; Lin, Yuah-Chiao; Yen, David C; y Chen, Yan-Min (2011). Predicting stock returns by classifier ensembles. *Applied Soft Computing*, 11(2), 2452–2459. 3.1, 6

Villada, Fernando; Muñoz, Nicolás; y García-Quintero, Edwin (2016). Redes neuronales artificiales aplicadas a la predicción del precio del oro. *Información tecnológica*, 27(5), 143–150. 2

Wang, Huacheng; Jiang, Yanxia; y Wang, Hui (2009). Stock return prediction based on bagging-decision tree. In *Grey Systems and Intelligent Services, 2009. GSIS 2009. IEEE International Conference on* (pp. 1575–1580).: IEEE. 3.1, 6

West, David; Dellana, Scott; y Qian, Jingxia (2005). Neural network ensemble strategies for financial decision applications. *Computers & operations research*, 32(10), 2543–2559. 3.2.6, 5.4

Zheng, Zhuo (2006). *Boosting and bagging of neural networks with applications to financial time series*. Technical report, Working paper, Department of Statistics, University of Chicago, Tech. Rep. 3.1

Enciclopedia Financiera. (2017). *Indicadores Económicos.* http://www.enciclopediafinanciera.com/indicadores-economicos.htm [Consultado el 1 de Diciembre de 2017]

Banco Central de Chile. (2017). *Base de Datos Estadisticos.*

https://si3.bcentral.cl/ [Consultado el 1 de Diciembre de 2017]

MathWorks. (2018). *Prepare the Predictor Data and Prepare the Response Data.* https://la.mathworks.com/help/stats/framework-for-ensemble-learning.html?requestedDomain=true&nocookie=true#bsvjyz_ [Consultado el 14 de Enero de 2018]