UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE CIENCIAS

UTILIZACIÓN DE INDICADORES DE IMPACTO SOCIAL EN LA PREDICCIÓN DE MERCADOS FINANCIEROS COMO ALTERNATIVA A LOS ENFOQUES TRADICIONALES

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

ACTUARÍA

P R E S E N T A :

GERALDINE GONZÁLEZ FERNÁNDEZ

DIRECTOR DE TESIS:

ACT. EDGAR DÍAZ ORDÓÑEZ

Ciudad Universitaria, CD. MX., 2021

Dedicado a madrecita por su infinito amor,

paciencia, apoyo

y su linda compañía en cada uno de mis pasos.

Tabla de Contenido

Lista de Figuras

Lista de Tablas

Capítulo 1: Introducción al problema

1.1. Introducción

La popularidad de los modelos enfocados al pronóstico de los mercados financieros que consideran factores emocionales y del comportamiento de la sociedad han aumentado en los últimos años, ya que se ha enfatizado la importancia que el factor emocional puede llegar a representar en la toma de decisiones de los participantes del mercado bursátil.

En este trabajo se analizará el comportamiento del índice Dow Jones Industrial Average (DJIA), este índice es el más antiguo que existe actualmente y consolida el desempeño de 30 grandes empresas que cotizan principalmente en la bolsa de Nueva York (NYSE) pero también incluye algunas que cotizan en NASDAQ. Debido a que NYSE y NASDAQ son los principales mercados bursátiles de Estados Unidos, este índice es una buena opción para dar seguimiento al estado del mercado. Conviene subrayar que los índices bursátiles son un valor numérico, que consolida las variaciones de valor o rentabilidad promedio de las acciones que forman parte de él.

En esta tesis además de trabajar con las variables tradicionales financieras como son: precios de cierre y la volatilidad, se incorporan los encabezados de noticias de periódicos estadounidenses financieros publicados en la red social de Facebook y los índices de tendencia de búsqueda proporcionados por Google Trends. Al incorporar variables de índole social a la construcción del modelo predictivo, se pretende entender cómo reacciona un entorno bursátil ante acciones del comportamiento humano como lo son periodos de euforia o de pánico por parte de sus participantes.

1.2. Estado del Arte

Diariamente se puede leer o escuchar información acerca de los mercados bursátiles, las finanzas y el dinero, estos temas atraen la atención de muchas personas que desean convertirse tanto en inversionistas como de individuos con deseo de entender el comportamiento de los mercados y poder pronosticar las diferentes variables financieras.

La manera tradicional de analizar los mercados es a través del análisis técnico. El análisis técnico es el estudio de los movimientos del mercado, principalmente mediante el uso de gráficos con el propósito de pronosticar futuras tendencias de los precios, es decir, el análisis técnico considera que todos los cambios en el mercado se pueden entender únicamente al analizar los movimientos en los precios. Por otro lado, gracias a los avances en las tecnologías de la información se puede analizar mayor cantidad y diversidad de información, y con esto abrir oportunidad a otros enfoques que pueden explicar con mayor precisión a los mercados financieros.

Algunos autores identifican a Vincent Cho como el primero en dar un enfoque distinto a la tarea de pronosticar mercados bursátiles. Cho (1996) decidió pronosticar mercados bursátiles utilizando información disponible en la web, Cho supone que la información textual contiene no únicamente el efecto (ej. El precio de la acción decae) también puede contener posibles causas del evento (ej. Las acciones decaen debido a un debilitamiento del dólar y de los bonos de la tesorería) por lo que explotar los textos mejoraría la calidad de las variables de entrada.

Cho utilizaba la información actual para pronosticar el futuro inmediato de los mercados accionarios, por otro lado, Víctor Lavrenko (2000) trató de pronosticar tendencias a partir de modelos que representaran patrones de lenguaje que están altamente relacionados con tendencias del mercado. Lavrenko utilizó información de un par de días previos a las tendencias, utilizó la técnica de bolsa de palabras para asociar historias con las tendencias de la serie de tiempo y concluyó que las regresiones lineales son útiles para describir series de tiempo cuando se tiene interés en las fluctuaciones de esta.

Autores como Paul Tetlock (2007) y Bo Zhao (2016) decidieron aprovechar los textos desde la perspectiva del análisis de sentimientos. Ambos notaron que el mercado es sensitivo a la negatividad, en otras palabras, el pesimismo en los medios de información puede ser indicador de una tendencia a la baja sobre los precios de mercado. Asimismo, Tetlock concluyó que un nivel de pesimismo inusualmente alto o bajo conlleva a un alto volumen de negociaciones en el mercado.

Huina Mao (2011) decidió comparar información de múltiples medios como son: encuestas, noticias, Twitter y tendencia de búsquedas en web. Mao resaltó la importancia que tiene el comportamiento y los factores emocionales como el estado de ánimo de la sociedad en los mercados bursátiles, al incorporar diferentes fuentes de información pudo aprovechar desde distintos ángulos el sentimiento que se vive alrededor del mercado, asimismo, observó que el efecto social en el mercado se puede observar en algunas fuentes desde una semana de anticipación.

1.3. Objetivo

En este trabajo se intentará modelar el comportamiento del Dow Jones Industrial Average con base a las fluctuaciones de los precios de cierre, la volatilidad del mercado y una serie de variables construidas a partir de textos de periódicos financieros y los índices de tendencia de búsqueda de Google Trends.

De manera general se analizarán dos puntos:

1) El impacto de añadir información social a modelos de índole bursátil comparando diferentes fuentes de información mediática.
2) La comparación del desempeño de técnicas de Minería de datos y Econometría para la construcción de modelos predictivos financieros.

Habría que decir también que se estima que el 90% de los datos existentes en el universo digital son datos no estructurados. Los datos no estructurados no tienen valor hasta que se identifican y se almacenan de una manera organizada, por lo que este trabajo brindará una perspectiva de como explotar este tipo de información. Se puede añadir que las técnicas descritas en este trabajo se pueden emplear en otros ámbitos.

Finalmente, conviene subrayar que un punto importante al pronosticar series de tiempo financieras es no solo medir la precisión de las estimaciones sino enfocarse en la dirección del cambio que tuvo la serie financiera, por lo que este trabajo se enfocará en captar la tendencia futura del índice.

1.4. Resumen

Este trabajo es de naturaleza interdisciplinaria, es por esto que a lo largo de este documento se podrán encontrar temas asociados a Minería de Datos, Minería de Textos, Análisis de Sentimientos, Econometría y Web Scraping. En resumen, la estructura de este trabajo consta de cinco capítulos y tres anexos, en estos se irá trabajando ordenadamente en los objetivos descritos anteriormente.

El capítulo dos describe los procesos realizados para la recolección de datos de las distintas fuentes empleadas y la construcción de variables. Por otro lado, en este capítulo también se puede encontrar un análisis descriptivo de las variables construidas con el objetivo de entender el comportamiento de las variables disponibles.

El capítulo tres resume los conceptos y la teoría en la que se fundamenta las técnicas que se emplearan para tratar de explicar el DJIA, asimismo, se puede encontrar el listado de métricas de ajuste que se utilizan para comparar a los diferentes modelos.

El capítulo cuatro explica la metodología que se siguió para la construcción de los diferentes modelos que se implementaron en este proyecto, también relata los aspectos destacables que se observaron tras comparar los diferentes enfoques de modelación.

El capítulo cinco amplía la información obtenida del que es considerado el modelo ganador de este trabajo y también incluye las conclusiones y el trabajo futuro que se podría llevar a cabo a partir de este proyecto.

Finalmente, en los anexos se encontrará información extra de algunos aspectos técnicos o teóricos que se emplean en esta tesis.

Capítulo 2: Recolección de Datos

Como se mencionó anteriormente, se desea modelar el comportamiento del DJIA a partir de variables que midan el impacto del comportamiento social en el mercado bursátil, por lo anterior, se requiere obtener datos de distinta índole. Los insumos que se recopilaron para la elaboración de este trabajo abarcan los aspectos financieros, sociales y tecnológicos que se desean evaluar.

La lista de insumos que fueron recopilados para la elaboración de este trabajo se muestra a continuación:

Tabla 1: Descripción de Insumos

Insumo	Descripción
Noticias	Texto de las publicaciones en la red social de Facebook de periódicos estadounidenses de índole financiera.
Google Trends (GIS)	Índice que refleja el volumen de búsquedas en Google de términos financieros.
Diccionario McDonald	Diccionarios de términos para poder realizar la técnica de minería de texto enfocada al análisis de sentimientos.
DJIA	Información histórica del comportamiento del índice Dow Jones Industrial Average.
VIX	Información histórica del comportamiento del índice Chicago Board Options Exchange Market Volatility Index.

2.1. Metodología para la recolección de insumos

A lo largo de esta sección se enlistará a detalle las fuentes de información y el procedimiento utilizado para la extracción de cada uno de los insumos utilizados.

2.1.1. Noticias

Una componente para entender el comportamiento del mercado bursátil contempla el estar bien informado acerca del entorno social, político y financiero que rodean a la operación financiera, para mantenerse actualizado de estos aspectos es necesario revisar las secciones de noticias de los medios de difusión tradicionales tales como radio, televisión y medios impresos como periódicos o revistas. Sin embargo, actualmente gracias a la era digital ha revolucionado la forma en que las personas consultan la información, pues ahora los medios electrónicos de información están ganando terreno a los medios de información tradicionales.[1]

En este caso como se ha mencionado se desea explicar el comportamiento del índice Dow Jones, a partir de distintas variables asociadas al contexto social y financiero que se vive durante la operación del mercado. Las variables clave de este trabajo derivan de las noticias que son publicadas, ya que se desea medir el impacto que causa en el mercado la publicación de determinados tipos de noticias.

El primer paso es la extracción de noticias, este trabajo se centró en analizar el entorno mostrado en los encabezados de noticias. Por cuestiones, de disponibilidad de información se utilizaron los encabezados

[1] Vázquez R. Diarios impresos vs. Diarios digitales. Forbes México. 2014. [consultado 20 de noviembre de 2020]. Disponible en: https://www.forbes.com.mx/diarios-impresos-vs-diarios-digitales/

publicados en las cuentas de Facebook de siete instituciones de suma importancia que se dan a la tarea de difusión de información financiera en Estados Unidos.

Posteriormente, se realizó la transformación de estos textos a datos estructurados que puedan ser explotados y aplicados a la modelación del comportamiento del DJIA. Convertir la información no estructurada del corpus[2] a información estructurada, se consiguió a partir de la aplicación de técnicas de minería de texto.

Las cuentas de Facebook de los periódicos con los que se trabajó se enlistan en la siguiente tabla:

Tabla 2. Lista de journals para consulta de noticias.

Periódico	Perfil de Facebook
The Wall Street Journal	@WSJ
CNN Money	@cnnmoney
Bloomberg	@bloombergbusiness
CNBC	@cnbc
Reuters	@reuters
Forbes	@Forbes
Financial Times	@financialtimes

Metodología para la extracción de noticias

La descarga de las publicaciones de los perfiles de Facebook mencionados se realizó utilizando el software R. Facebook permite utilizar en modo desarrollador una API con la se puede acceder a la información de los perfiles catalogados como página pública, usualmente los perfiles que cumplen con esta característica son aquellos perteneciente a instituciones, figuras públicas o instrumentos gubernamentales, dado lo anterior, es posible acceder a información publicada por los periódicos.

Es importante mencionar que Facebook tiene algunas restricciones respecto a la extracción de las publicaciones, por ejemplo, no se pueden descargar todas las publicaciones asociadas a un perfil sino solo aquellas que Facebook haya habilitado para descarga. El número de publicaciones que cumplen con la característica antes mencionada es variable, por lo que se tiene poco control respecto a la información disponible. Además, no existe un indicador para dar seguimiento al total de publicaciones que se realizaron en determinada fecha, lo anterior sería de utilidad para evaluar la proporción de noticias disponibles en cada día.

La descarga de los datos de Facebook utilizados para este trabajo se realizó en septiembre de 2017, y ha habido cambios en las políticas de uso de la API de Facebook por lo que la metodología utilizada puede ya no ser válida.

La descarga de datos se realizó basándose en las instrucciones de un enlace encontrado en la web[3].

[2] En minería de texto se denomina corpus a la colección de textos a analizar.

[3] Data mining Facebook data using R. Big Data Enthusiast. 2016 [consultado 20 de noviembre de 2020]. Disponible en: https://bigdataenthusiast.wordpress.com/2016/03/19/mining-facebookdata-using-r-facebook-api /

2.1.2. Diccionario de McDonald

Se ha mencionado que se desea extraer información de valor de las noticias, una forma de hacer esto es a partir de la minería de texto, una opción en particular es usando la técnica de análisis de sentimientos.

El análisis de sentimientos se realizó utilizando un diccionario de términos negativos. Mao, Counts y Bollen[4] mencionan dos diccionarios aceptados para la identificación de lenguaje negativo, además, el artículo de Zhao, et al.[5] muestra evidencia que el sentimiento de negatividad en las noticias ha funcionado como mejor variable de entrada en modelos, comparando la negatividad con otros sentimientos.

Por lo anterior, se decidió enfocar este trabajo al análisis del sentimiento de negatividad en los textos. Tras revisar el contenido de los diccionarios de léxico negativo mencionados en el artículo de Mao, se decidió utilizar el de Loughran and McDonald ya que los términos que contiene están más asociados a términos financieros.

El diccionario de McDonald está disponible para descarga en la web[6].

2.1.3. Índice de Google Trends (GIS)

Otra fuente de datos que es utilizada en el artículo de Mao, et al. es aquella que se denomina Google Insights for Search (GIS), cabe mencionar que este recurso ahora es denominado Google Trends. Google Trends es un servicio de Google que brinda información acerca del volumen de búsquedas en la web de diversos términos, este servicio inició sus funciones en 2004.

Esta herramienta brinda un índice en la escala de 0 a 100, donde 100 índica el punto máximo del volumen de búsquedas registradas en el período de tiempo seleccionado y 0 índica que no se registraron búsquedas para el termino seleccionado. Se debe agregar que este valor es calculado de manera semanal con el consolidado de búsquedas de cada termino. La descarga de este insumo se realizó de manera manual a través de la página web de Google Trends[7].

Se debe agregar que los términos de búsqueda utilizados para la realización de este trabajo son los 26 términos financieros mostrados en el artículo de Mao, et al. los cuales son listados a continuación:

[4] H. Mao, S. Counts, and J. Bollen, (2011). Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data., Journal of Computational Science.

[5] B. Zhao, H. Yongji, Y. Chunfeng and H. Yihua, (2016) Stock Market Prediction Exploiting Microblog Sentiment Analysis. Center of Novel Software Technology and Industrialization.

[6] Software Repository for Accounting and Finance [sitio de internet]. University of Notre Dame. [consultado 20 de noviembre de 2020]. Disponible en: https://sraf.nd.edu/textual-analysis/resources/

[7] Google Trends [sitio de internet]. Google. [consultado 20 de noviembre de 2020]. Disponible en: https://trends.google.com.mx/

Tabla 3: Términos para búsqueda

Término de Búsqueda Google Trends
Finance_news financial_news stock_market_news stock_market_today djia Dow Dow_Jones Dow_Jones_Industrial_Average bear_market stock_fall Stock_market_crash SP500 stock stock_market bullish bull_market stock_decline finance financial_market wall_street_news_today wall_street best_stock long_stock stock_price stock_to_buy bearish

Es importante mencionar que la información de Google Trends asociada a cada término se descargó de forma anual y restringida al territorio de Estados Unidos. La descarga anual se realizó con el objetivo de hacer comparable la información, puesto que la construcción de este índice se calcula considerando el valor máximo de búsqueda en la ventana de tiempo por lo que al descargar la información de esta forma se desea mitigar el impacto del crecimiento de usuarios de internet en Estados Unidos.

Posteriormente, se consolidó la información anual para formar una única serie de tiempo por término.

2.1.4. DJIA y VIX

El DJIA es un índice bursátil constituido por las 30 empresas con mayor capitalización bursátil de la Bolsa de Valores de Nueva York (NYSE). Por otro lado, el VIX es un índice que mide la volatilidad del mercado de opciones de Chicago, este índice se calcula utilizando una serie de opciones a corto plazo que tienen como subyacente al índice S&P500. El índice S&P500 contiene a las 500 más grandes empresas que cotizan en las bolsas NYSE o NASDAQ.

Se debe añadir que las empresas del DJIA están contenidas en el S&P500, es por esto que medir la volatilidad general del mercado a través del índice VIX, ayuda a explicar el comportamiento del DJIA.

Estos dos factores de riesgo se descargaron de la plataforma de Yahoo! Finanzas[8].

2.2. Construcción de Variables

Para el desarrollo de este trabajo se crearon una serie de variables a partir de los insumos anteriormente mencionados, esto con el objetivo de extraer información que aporte valor a la tarea de explicar el comportamiento del índice DJIA. En esta sección se detallarán los procedimientos realizados para la construcción de cada una de las variables que participan en el desarrollo de este trabajo.

2.2.1. Corrección al DJIA y al VIX

Dado que se pretende medir el impacto que el entorno social tiene en el comportamiento del mercado, por medio de la publicación de noticias, se desean considerar también las publicaciones realizadas en fines de semana y días no laborables del mercado bursátil. Para trabajar con todos los días de la semana se procedió a realizar una imputación a los índices DJIA y VIX en los días sin información, esto se consiguió a partir de mantener el último valor observado de los índices previo a los días con valores ausentes, un ejemplo de esta imputación se muestra a continuación.

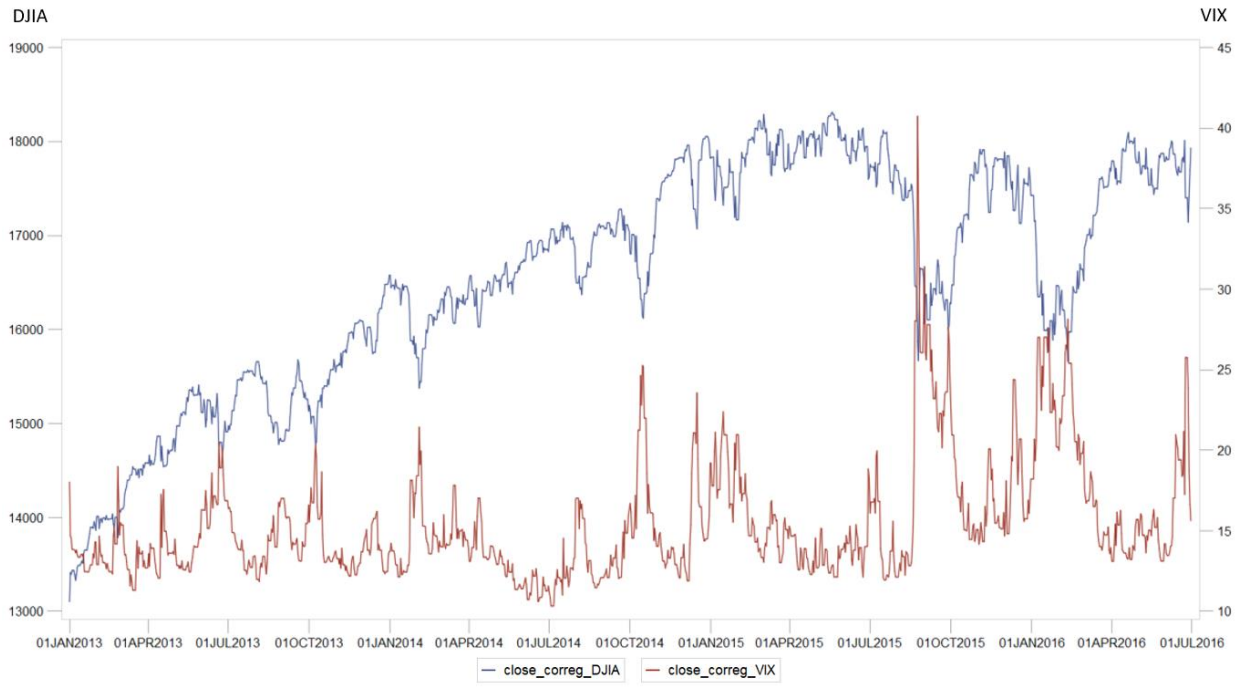
[8] Yahoo! Finance [sitio de internet]. Yahoo! [consultado 20 de noviembre de 2020]. Disponible en: https://es-us.finanzas.yahoo.com/

Figura 1: Corrección de DJIA y VIX

date_completa	close_DJIA	close_correg_DJIA
04JAN2010	10583.95996	10583.95996
05JAN2010	10572.01953	10572.01953
06JAN2010	10573.67969	10573.67969
07JAN2010	10606.86035	10606.86035
08JAN2010	10618.19043	10618.19043
09JAN2010	.	10618.19043
10JAN2010	.	10618.19043
11JAN2010	10663.99023	10663.99023
12JAN2010	10627.25977	10627.25977
13JAN2010	10680.76953	10680.76953
14JAN2010	10710.54981	10710.54981
15JAN2010	10609.65039	10609.65039
16JAN2010	.	10609.65039
17JAN2010	.	10609.65039
18JAN2010	.	10609.65039
19JAN2010	10725.42969	10725.42969
20JAN2010	10603.15039	10603.15039
21JAN2010	10389.87988	10389.87988
22JAN2010	10172.98047	10172.98047
23JAN2010	.	10172.98047

En la siguiente imagen, se puede observar el comportamiento en la ventana de análisis de estos dos índices. En la gráfica se puede observar, que altos valores en el VIX se asocian a caídas en el valor del DJIA.

Figura 2: DJIA y VIX corregidos



Es importante aclarar que a partir de este punto, cada que se mencione que se utiliza la información de DJIA y VIX se estarán ocupando las versiones corregidas.

2.2.2. Log Return

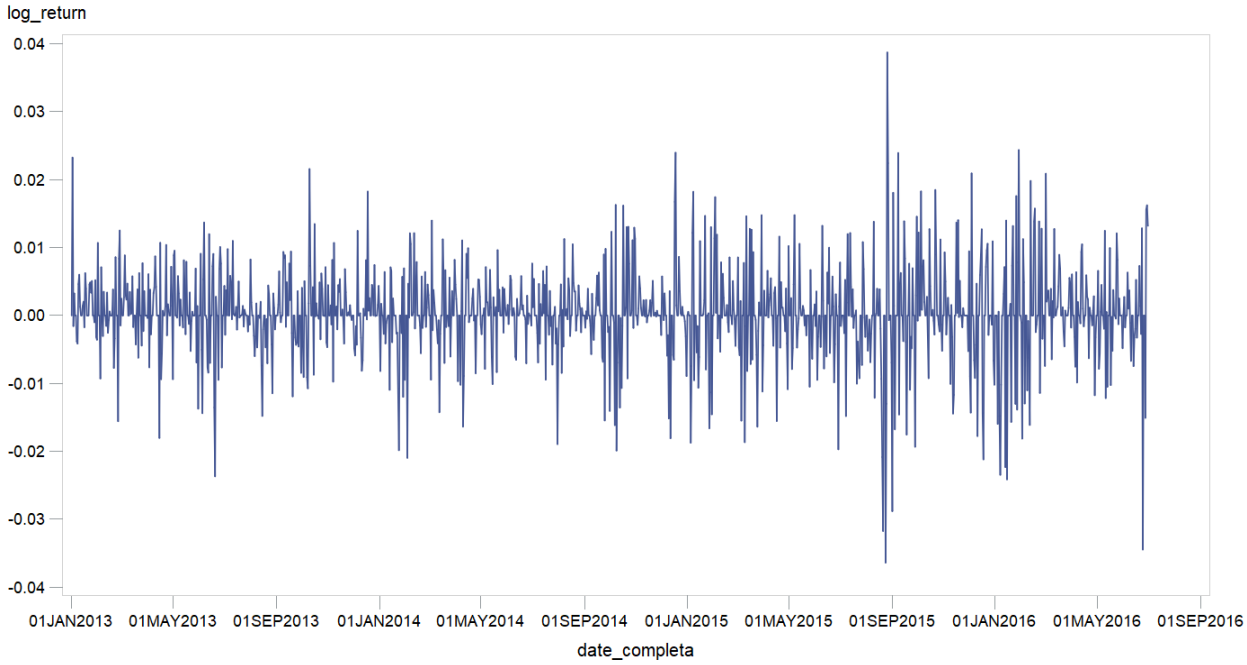
Tal como en el artículo de Mao[9] la variable objetivo de este trabajo es el log return diario del mercado, esta variable se construyó a partir de los valores de cierre del DJIA. El log return ayuda a medir la magnitud y dirección de la tendencia del Dow Jones. La definición para el cálculo del log return es la siguiente:

$$R_{\Delta t} = logS(t) - logS(t - \Delta t)$$

Con $\Delta t = 1$

Gráficamente el histórico de los log-rendimientos se puede observar en la siguiente imagen.

Figura 3: Histórico de Log-Rendimientos



2.2.3. Índice de Negatividad

La construcción de este índice se realizó posterior al proceso de parseo del corpus (para más información ver el apéndice de minería de texto). Los pasos para la construcción de esta variable se detallan a continuación:

[9] H. Mao, S. Counts, and J. Bollen, (2011). Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data., Journal of Computational Science

PASO1. Se contabiliza el número de palabras de valor por noticia.

PASO 2. Se contabiliza el número de palabras negativas por noticia con ayuda del diccionario de McDonald y la P_key de salida del PROC TGPARSE[10].

PASO3. Se cruzan las anteriores tablas para la construcción del índice negativo por noticia.

Solamente se contabilizaron palabras de 31,151 noticias y también se encontró que existen noticias que no contabilizan palabras que aporten información de interés, y por tanto se descartan.

Un ejemplo de noticia que contabiliza cero palabras es la siguiente.

Figura 4. Noticias sin palabras de interés

	message_clean	from_name	message
60	WOWZERZ	CNNMoney	Wowzerz.

El índice de negatividad se construyó primero a nivel noticia, haciendo el cociente del número de palabras asociadas al sentimiento negativo, entre el total de palabras del encabezado. Posteriormente para agregar un único valor diario de este índice, se sumó el índice de negatividad por noticia de las publicadas un mismo día y se dividió entre el número de noticias disponibles en esa fecha.

$$Indice\ de\ Negatividad\ por\ Noticia = \frac{Palabras\ negativas}{Total\ de\ palabras}$$

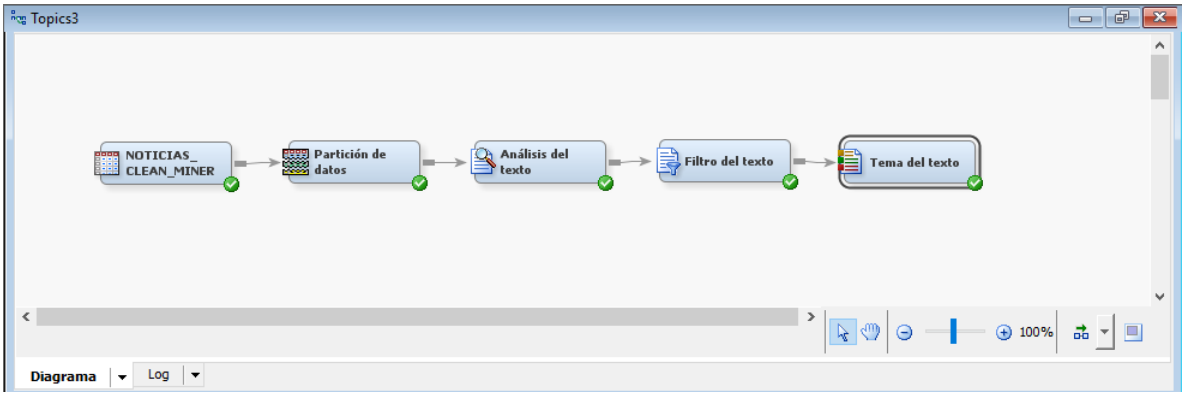
$$Indice\ de\ negatividad\ Diario = \frac{\sum Negative\ Index\ por\ Noticia}{Noticias\ disponibles\ en\ el\ día}$$

2.2.4 Creación de Temas de Texto

La creación de temas de texto se realiza utilizando la herramienta SAS Enterprise Miner con su complemento para Minería de Texto, para la generación de los topics de textos se utilizan los nodos de análisis de texto (**Text Parsing Node**) y temas de texto (**Text Topics Node**). El nodo de **Text Parsing** ayuda a descomponer los textos en palabras, por otro lado, el nodo de **Text Topics** permite analizar palabras para generar conjuntos asociados a un mismo tema. Esta herramienta se maneja por medio de diagramas que estructuran el flujo de las tareas a realizar, en este caso, el diagrama implementado se muestra en la siguiente imagen:

[10] SAS Text Miner 13.1 High-Performance Procedures [Sitio de internet]. SAS Institute Inc. [consultado 20 de noviembre de 2020]. Disponible en: https://support.sas.com/documentation/onlinedoc/txtminer/13.1/tmhpprcref.pdf

Figura 5: Diagrama de creación de topics de texto



En la anterior configuración, el nodo de **Filtro de texto** ayuda a reducir el número de términos arrojados por el análisis de minería de texto. Este nodo ayuda a eliminar términos con baja frecuencia que no llegan a aportar información de valor al modelo.

El **filtrado de términos** se realiza buscando conservar únicamente los términos que sobrepasen el horizonte del peso que debe de tener un término para considerarse un término de valor.

Los pesos de los términos se construyen teniendo como objetivo, medir la importancia de los términos con base en su frecuencia de ocurrencia y también se considera la distribución de un término a lo largo de la colección de documentos.

El primer paso que se realiza previo al filtrado de textos es la construcción de una matriz de frecuencias de términos por documentos:

$$\begin{Bmatrix} F_{11} & F_{12} & \dots & F_{1N} \\ F_{21} & F_{22} & \dots & F_{2N} \\ \vdots & \vdots & \dots & \vdots \\ F_{M1} & F_{M2} & \dots & F_{MN} \end{Bmatrix}$$

donde N es el número de documentos y M el número de términos.

Posteriormente, se construye una matriz de frecuencias ponderadas por términos y documentos:

$$\begin{Bmatrix} W_1*g(F_{11}) & W_1*g(F_{12}) & \dots & W_1*g(F_{1N}) \\ W_2*g(F_{21}) & W_2*g(F_{22}) & \dots & W_2*g(F_{2N}) \\ \vdots & \vdots & \dots & \vdots \\ W_M*g(F_{M1}) & W_M*g(F_{M2}) & \dots & W_M*g(F_{MN}) \end{Bmatrix}$$

donde la función g(.) es una función que ayuda a ponderar la frecuencia de los términos. En este caso se utiliza la función logaritmo ya que esta amortigua el efecto de los términos que ocurren en múltiples ocasiones en un documento.

Por otro lado, los pesos de los términos se construyen a partir del método de Entropía el cual nos ayuda a diferenciar los términos que podrían ser de interés de aquellos que no lo son. Este método tiene como supuesto que un término útil es aquel que ocurre en pocos documentos, pero ocurre múltiples veces en estos.

El método de Entropía asigna el peso de un término de la siguiente manera:

$$w_i = 1 + \sum_j \frac{\left(\frac{f_{ij}}{g_i}\right) * log_2(f_{ij}/g_i)}{log_2(n)}, \qquad con\ j \in \{1,2, \dots N\}$$

Con:

g_i el número de ocurrencias del término en el universo de documentos

n el número de documentos en el universo de análisis

f_{ij} la frecuencia del término i en el documento j, cabe mencionar que log(.)=0 si f_{ij}=0

Este método da mayor peso a los términos que se producen con poca frecuencia en la colección de documentos mediante el uso de la entropía de la información, definiendo el concepto de entropía a partir de la teoría de la información[11]. Comprendiendo superficialmente el concepto de entropía, este se asocia a la cantidad de información promedio que contiene una colección de textos. El concepto de entropía se fundamenta en que las cadenas de texto que más aportan para la comprensión de un escrito son aquellas con menor probabilidad de ocurrencia en el mismo.

Finalmente, tras la asignación de pesos este nodo ordena los términos por su peso de manera descendente y conserva los primeros K términos.

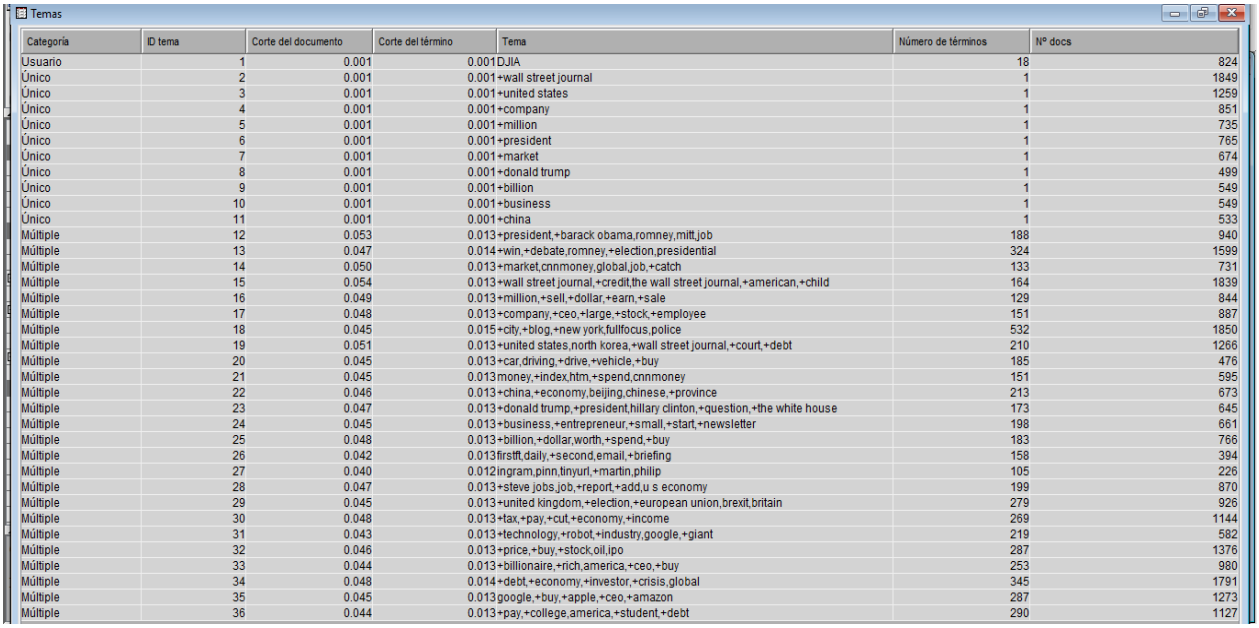
Una vez que se concluye el filtrado de términos, la siguiente tarea se encarga de la construcción de **temas de texto**. Un tema es una asociación de términos que describen un concepto, a los documentos se les coloca una puntuación de si hablan o no de determinado tema. Un mismo documento puede hablar de diversos temas.

Los resultados de realizar esta tarea con las noticias de FB de los periódicos financieros se muestran en la siguiente imagen. Se obtuvieron 36 topics, de los cuales 10 están formados por un único término, 25 contienen una colección de términos, y 1 tema fue generado de forma manual para indicar si la noticia hace referencia o no al índice DJIA.

Se puede notar que los topics de la categoría múltiple, contiene cada uno al menos 100 términos. Asimismo, se puede destacar que al menos 824 noticias mencionan al índice DJIA.

[11] C. Shannon. (1948) A Mathematical Theory of Communication. The Bell System Technical Journal

Figura 6: Resultados de temas de texto



Temas

Categoría	ID tema	Corte del documento	Corte del término	Tema	Número de términos	Nº docs
Usuario	1	0.001	0.001	DJIA	18	824
Único	2	0.001	0.001	+wall street journal	1	1849
Único	3	0.001	0.001	+united states	1	1259
Único	4	0.001	0.001	+company	1	851
Único	5	0.001	0.001	+million	1	735
Único	6	0.001	0.001	+president	1	765
Único	7	0.001	0.001	+market	1	674
Único	8	0.001	0.001	+donald trump	1	499
Único	9	0.001	0.001	+billion	1	549
Único	10	0.001	0.001	+business	1	549
Único	11	0.001	0.001	+china	1	533
Múltiple	12	0.053	0.013	+president,+barack obama,romney,mitt,job	188	940
Múltiple	13	0.047	0.014	+win,+debate,romney,+election,presidential	324	1599
Múltiple	14	0.050	0.013	+market,cnnmoney,global,job,+catch	133	731
Múltiple	15	0.054	0.013	+wall street journal,+credit,the wall street journal,+american,+child	164	1839
Múltiple	16	0.049	0.013	+million,+sell,+dollar,+earn,+sale	129	844
Múltiple	17	0.048	0.013	+company,+ceo,+large,+stock,+employee	151	887
Múltiple	18	0.045	0.015	+city,+blog,+new york,fullfocus,police	532	1850
Múltiple	19	0.051	0.013	+united states,north korea,+wall street journal,+court,+debt	210	1266
Múltiple	20	0.045	0.013	+car,driving,+drive,+vehicle,+buy	185	476
Múltiple	21	0.045	0.013	money,+index,htm,+spend,cnnmoney	151	595
Múltiple	22	0.046	0.013	+china,+economy,beijing,chinese,+province	213	673
Múltiple	23	0.047	0.013	+donald trump,+president,hillary clinton,+question,+the white house	173	645
Múltiple	24	0.045	0.013	+business,+entrepreneur,+small,+start,+newsletter	198	661
Múltiple	25	0.048	0.013	+billion,+dollar,worth,+spend,+buy	183	766
Múltiple	26	0.042	0.013	firstft,daily,+second,email,+briefing	158	394
Múltiple	27	0.040	0.012	ingram,pinn,tinyurl,+martin,philip	105	226
Múltiple	28	0.047	0.013	+steve jobs,job,+report,+add,u s economy	199	870
Múltiple	29	0.045	0.013	+united kingdom,+election,+european union,brexit,britain	279	926
Múltiple	30	0.048	0.013	+tax,+pay,+cut,+economy,+income	269	1144
Múltiple	31	0.043	0.013	+technology,+robot,+industry,google,+giant	219	582
Múltiple	32	0.046	0.013	+price,+buy,+stock,oil,ipo	287	1376
Múltiple	33	0.044	0.013	+billionaire,+rich,america,+ceo,+buy	253	980
Múltiple	34	0.048	0.014	+debt,+economy,+investor,+crisis,global	345	1791
Múltiple	35	0.045	0.013	google,+buy,+apple,+ceo,+amazon	287	1273
Múltiple	36	0.044	0.013	+pay,+college,america,+student,+debt	290	1127

2.2.5. Índice de Google Trends Diario

La publicación de este índice (**2.1.3. Índice de Google Trends (GIS)**) se realiza de manera semanal. Sin embargo, el resto de las variables tiene una frecuencia de información diaria. Por lo anterior, se optó por transformar esta variable a una frecuencia diaria, esta tarea se realizó por medio de splines cúbicos y la utilización del PROC EXPAND[12] de SAS. Los splines cúbicos son curvas cuya definición se basa en polinomios definidos en porciones del intervalo que se desea analizar.

Es decir, para cada par de observaciones semanales consecutivas se construye un polinomio y posteriormente, se evalúan todos los días de la semana con esta función; de esta forma las observaciones se convirtieron en datos diarios.

$$f(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i$$

$$\text{Con } i = \quad i \;\; si\; x_i \leq x < x_{i+1}$$

$$1 \;\; si\; x < x_1$$

$$n \;\; si\; x \geq x_n$$

Se buscó que la estructura de correlación entre los índices descargados se mantuviera al transformar la periodicidad de la información. A continuación, se muestran las matrices de correlación de los términos antes y después de aplicar esta técnica. Se aplicó una escala de colores a las correlaciones para facilitar su interpretación, la escala de colores aplicada se describe en la **Tabla 4**.

[12] SAS® 9.4 and SAS® Viya® 3.4 Programming [sitio de internet]. SAS Institute Inc. [consultado 20 de noviembre de 2020]. Disponible en: https://documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.4&docsetId=etsug&docsetTarget=etsug_expand_overview.htm&locale=es

Tabla 4: Rango de colores para correlaciones.

COLOR	RANGO
	[-1,-1]
	(-1,-0.75]
	(-0.75,-0.5]
	(-0.5,-0.25]
	(-0.25,0.25]
	(0.25,0.5]
	(0.5,0.75]
	(0.75,1)
	[1,1]

Figura 7: Matriz de correlaciones de GIS semanal.

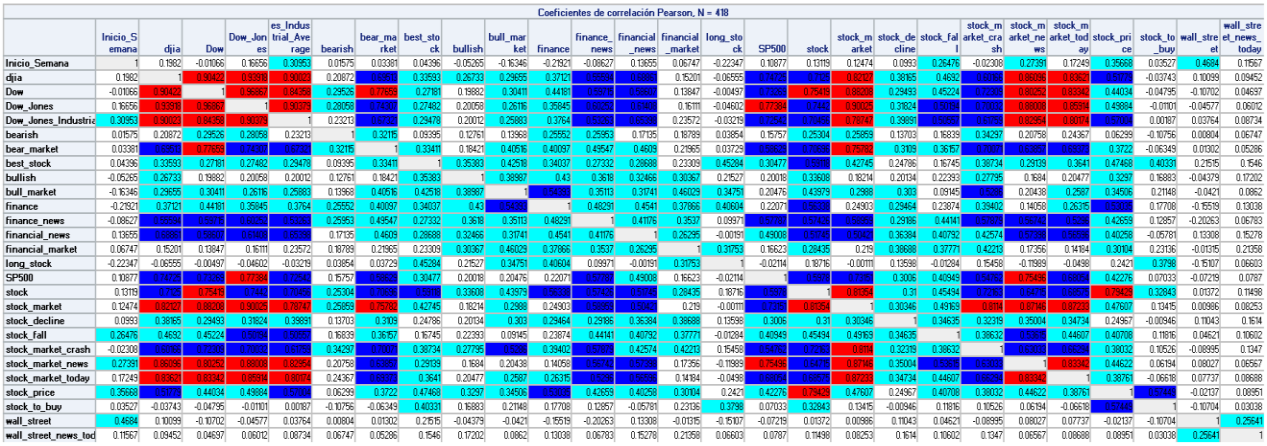
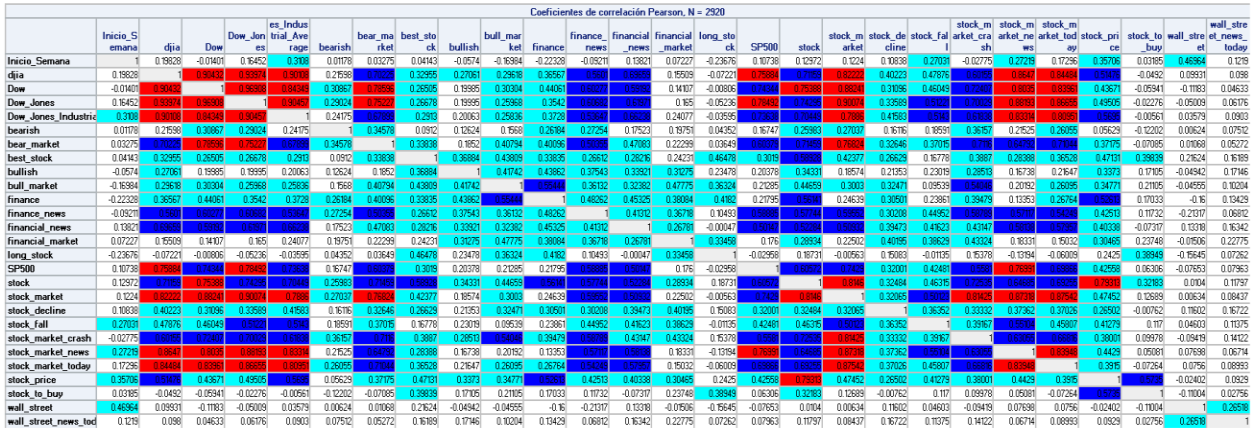


Figura 8: Matriz de correlaciones de GIS diario.



Lo más importante de estas matrices, es señalar que la distribución de colores se mantiene tras la transformación por lo que concluimos que la estructura de correlación entre los diferentes índice se mantiene.

2.2.6 Clúster de Variables

Al contar con una gran cantidad de variables relacionadas como lo son los términos de GIS y los Topics generados, se aplicó una técnica para la reducción de variables. El proceso utilizado fue el PROC VARCLUS[13] de SAS el cuál aplica una técnica de construcción de clúster sobre las variables de entrada.

La teoría que sustenta a este procedimiento, es la construcción de clúster no jerárquicos con las variables de entrada. El análisis de conglomerados de variables implementado en PROC VARCLUS es una alternativa a los métodos multivariados tradicionales, como el análisis de componentes principales y el análisis factorial. Se decidió utilizar esta técnica debido a que en el artículo de D. Pasta y D. Suhr[14], se comparó este procedimiento versus otros enfoques y se observaron resultados satisfactorios a favor de está técnica.

En el PROC VARCLUS inicialmente, todas las variables comienzan en el mismo clúster y un análisis de componentes principales es realizado, si el segundo valor característico es mayor que el punto de corte especificado entonces se divide al clúster de variables en dos. Que el segundo valor característico sea alto indica que al menos dos componentes principales explican un gran monto de variabilidad entre las variables de entrada, por lo que existe más de una dimensión en el clúster.

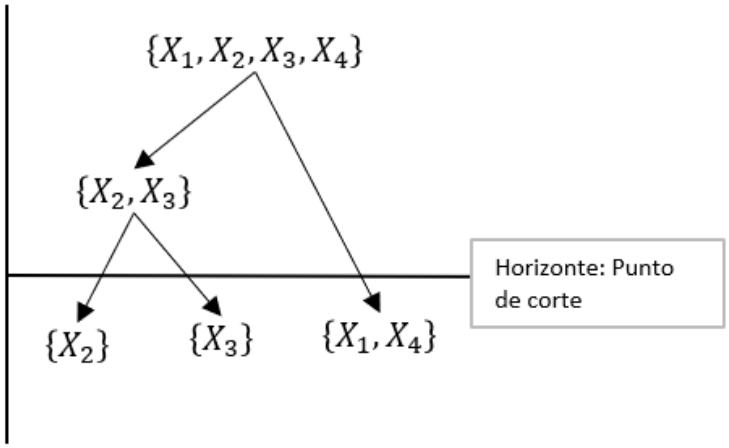
La asociación de variables a cada clúster se realiza bajo la premisa de formar grupos de variables que estén correlacionadas entre ellas y que sean no correlacionadas con las variables de otros grupos.

[13] SAS® 9.4 and SAS® Viya® 3.4 Programming [sitio de internet]. SAS Institute Inc. [consultado 20 de noviembre de 2020]. Disponible en: https://documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.4&docsetId=statug&docsetTarget=statug_varclus_overview.htm&locale=es

[14] D. Pasta y D. Suhr (2004). SAS Users Group International 29: Best Papers and honorable mentions (SUGI 29). SAS Institute

El anterior procedimiento se repite en cada uno de los clústeres formados, hasta que el segundo valor característico del análisis de componentes principales realizado en cada rama, no supere el horizonte marcado.

Figura 9: Clúster de variables



Esta técnica se utiliza para la reducción de dimensiones de los topics de textos y de los índices de búsqueda de Google.

A continuación, se muestran los resultados de aplicar la técnica de clúster de variables a los Topics de términos obtenidos. Los 36 topics se consolidaron en 10 clúster; cabe destacar que los clústeres muestran consistencia al agrupar temas con ideas similares. Por ejemplo, el clúster 1 habla de temas asociados a la economía y al dinero, por otro lado, el clúster 2 habla de asuntos políticos. Se sumó la frecuencia diaria de ocurrencia de cada uno de los topics que conforman el clúster para consolidar la información en los grupos generados.

Tabla 5: Clúster de Topics

Clúster de Topics	Topic de Texto	Descripción del Topic de Textos
Cluster1	News_topic3	+united states
Cluster1	News_topic13	+win,+debate,romney,+election,presidential
Cluster1	News_topic19	+united states,north korea,+wall street journal,+court,+debt
Cluster1	News_topic21	money,+index,htm,+spend,cnnmoney
Cluster1	News_topic27	ingram,pinn,tinyurl,+martin,Philip
Cluster1	News_topic28	+steve jobs,job,+report,+add,u s economy
Cluster1	News_topic30	+tax,+pay,+cut,+economy,+income
Cluster1	News_topic32	+price,+buy,+stock,oil,ipo
Cluster1	News_topic33	+billionaire,+rich,america,+ceo,+buy
Cluster1	News_topic34	+debt,+economy,+investor,+crisis,global
Cluster1	News_topic36	+pay,+college,america,+student,+debt

Clúster de Topics	Topic de Texto	Descripción del Topic de Textos
Cluster2	News_topic6	+president
Cluster2	News_topic8	+donald Trump
Cluster2	News_topic12	+president,+barack obama,romney,mitt,job
Cluster2	News_topic23	+donald trump,+president,hillary clinton,+question,+the white house
Cluster2	News_topic26	firstft,daily,+second,email,+briefing
Cluster2	News_topic29	+united kingdom,+election,+european union,brexit,britain
Cluster3	News_topic2	+wall street journal
Cluster3	News_topic15	+wall street journal,+credit,the wall street journal,+american,+child
Cluster4	News_topic7	+Market
Cluster4	News_topic14	+market,cnnmoney,global,job,+catch
Cluster5	News_topic5	+million
Cluster5	News_topic16	+million,+sell,+dollar,+earn,+sale
Cluster6	News_topic10	+business
Cluster6	News_topic24	+business,+entrepreneur,+small,+start,+newsletter
Cluster7	News_topic11	+china
Cluster7	News_topic22	+china,+economy,beijing,chinese,+province
Cluster8	News_topic1	DJIA
Cluster8	News_topic18	+city,+blog,+new york,fullfocus,police
Cluster8	News_topic20	+car,driving,+drive,+vehicle,+buy
Cluster8	News_topic31	+technology,+robot,+industry,google,+giant
Cluster8	News_topic35	google,+buy,+apple,+ceo,+amazon
Cluster9	News_topic9	+billion
Cluster9	News_topic25	+billion,+dollar,worth,+spend,+buy
Cluster10	News_topic4	+company
Cluster10	News_topic17	+company,+ceo,+large,+stock,+employee

Por otro lado, se aplicó el mismo procedimiento a los índices de búsqueda de Google Trends de los 26 términos, los resultados se listan en la **Tabla 6**. Se obtuvieron 5 conjuntos de términos con ideas similares, por ejemplo, el clúster 4 se enfoca en acciones y el clúster 5 se asocia a un mercado a la baja. Se debe añadir que para consolidar los índices de búsqueda de los términos de cada grupo se promedió el índice GIS de cada término del clúster.

Tabla 6: Clúster de Términos

Clúster de Términos	Término Google Trends
Cluster1	finance_news
Cluster1	financial_news
Cluster1	stock_market_news
Cluster1	stock_market_today
Cluster1	Djia
Cluster1	Dow
Cluster1	Dow_Jones
Cluster1	Dow_Jones_Industrial_Average
Cluster1	bear_market
Cluster1	stock_fall
Cluster1	stock_market_crash
Cluster1	SP500
Cluster1	Stock
Cluster1	stock_market
Cluster2	Bullish
Cluster2	bull_market
Cluster2	stock_decline
Cluster2	Finance
Cluster2	financial_market
Cluster3	wall_street_news_today
Cluster3	wall_street
Cluster4	best_stock
Cluster4	long_stock
Cluster4	stock_price
Cluster4	stock_to_buy
Cluster5	Bearish

2.3. Selección de la ventana de análisis

La descarga de noticias se realizó para el período de enero del 2010 a septiembre de 2017, sin embargo, el período de análisis en el que se enfoca este trabajo está restringido del 1 de enero de 2013 al 30 de junio de 2016 por las razones que se detallan a continuación:

En primer lugar, se buscó identificar un período de estrés en el índice DJIA y corroborar de manera empírica que los textos publicados en períodos cercanos sí mostrarán la situación que estaba viviendo este índice. Se definió que el índice se encontraba en un período de estrés, cuando los cambios en este

exceden los límites estadísticos para que una observación sea considerada extrema, es decir, cuando se sobrepase los siguientes rangos[15]:

$$q_1 - 3IQR \quad y \quad q_3 + 3IQR \qquad donde\ IQR\ es\ el\ rango\ intercuartil$$

En la variable que mide los cambios en el DJIA se observaron los estadísticos de la **Tabla 7**, con esto y las anteriores formulas se considera que un día es extremo si la diferencia es menor a $q_1 - 3 * IQR = -417.122006$ o si es mayor a $q_3 + 3 * IQR = 438.14256$. Los estadísticos de la siguiente tabla son estáticos y se calcularon utilizando toda la información disponible.

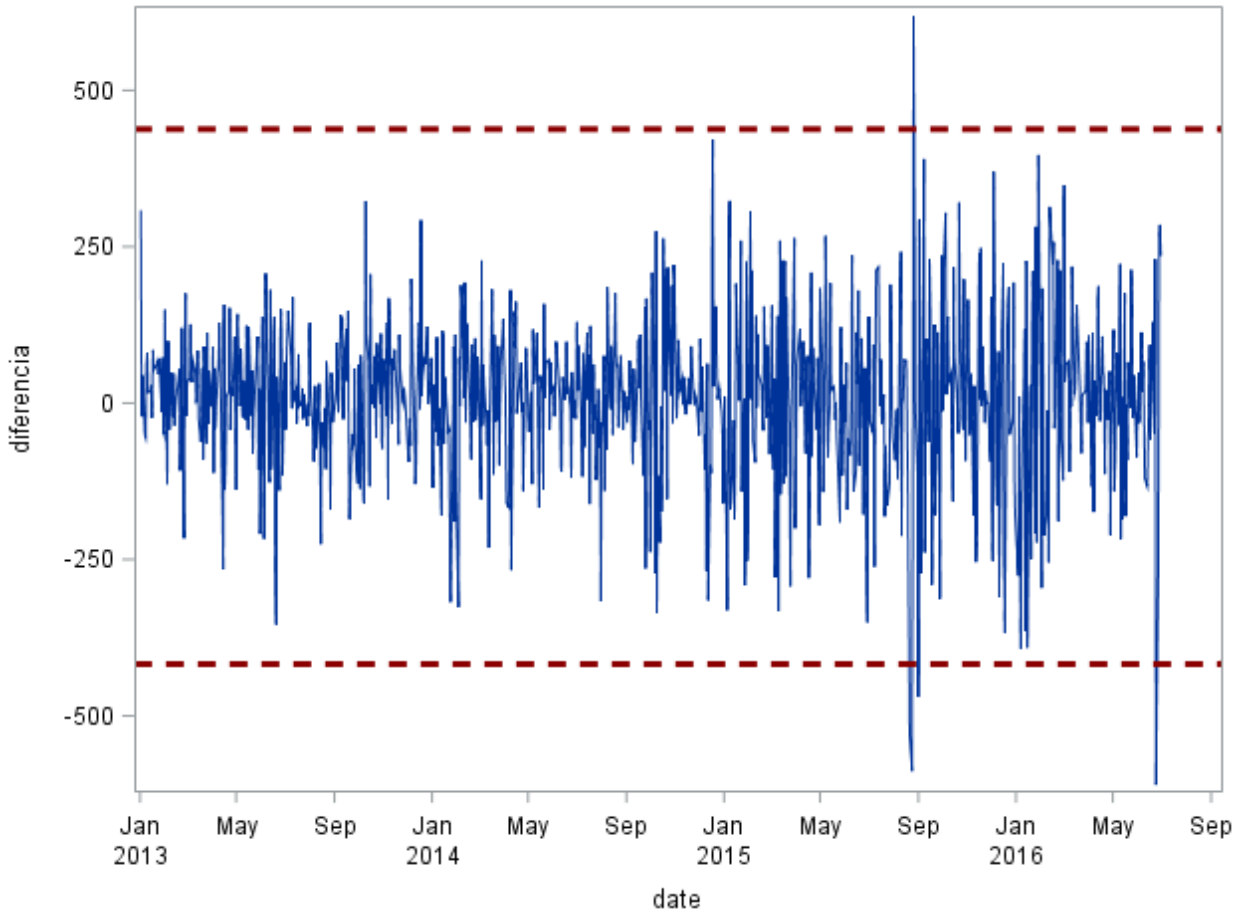
Tabla 7: Estadísticos de la variable de diferencia en el DJIA

Analysis Variable: diferencia			
Mean	**Lower Quartile**	**Median**	**Upper Quartile**
6.0259824	-50.58008	7.41015	71.60058

En la siguiente imagen, se observa gráficamente como se distribuyen las diferencia del DJIA dentro de las franjas para considerar que una observación es extrema.

[15] https://people.richland.edu/james/lecture/m170/ch03-pos.html

Figura 10: Gráfica de diferencias en el DJIA



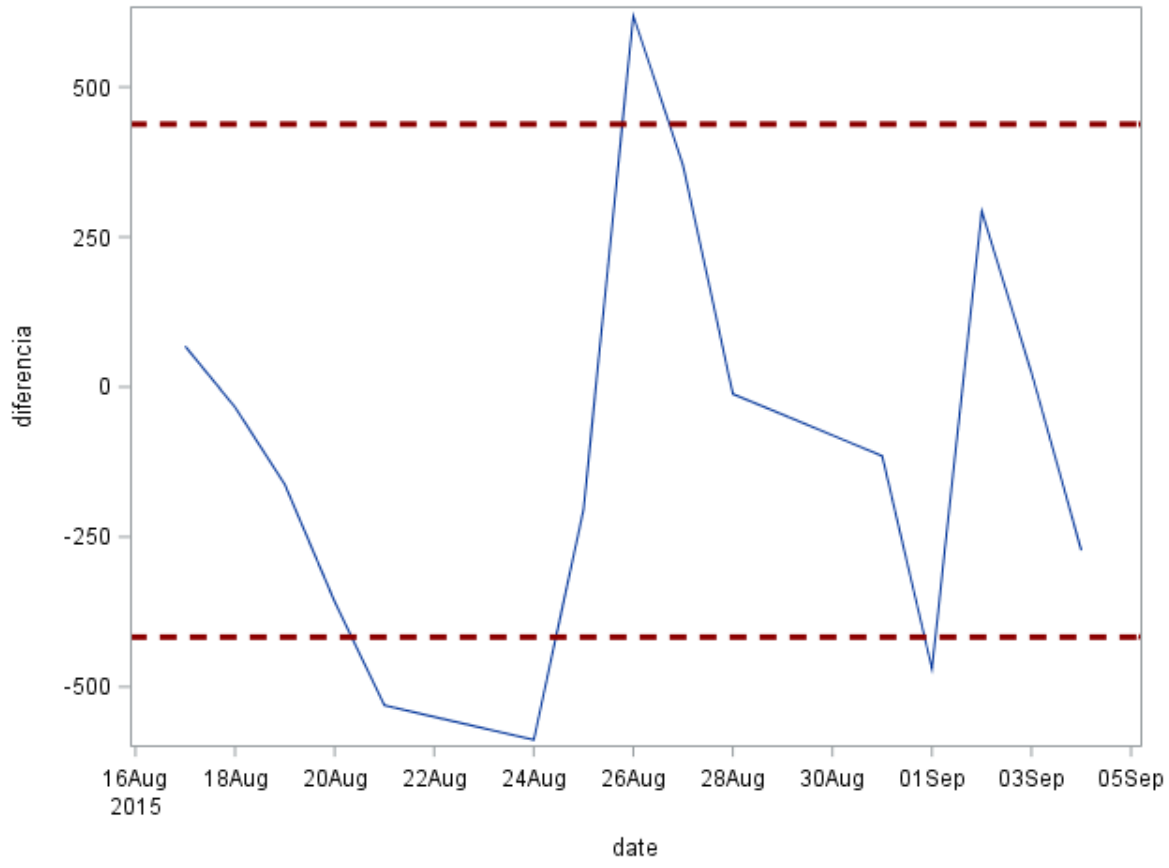
En particular se observó en agosto del 2015 un periodo de aproximadamente dos semanas, donde se presentaron múltiples días con comportamiento extremo, lo cual se asocia a una alta volatilidad en el mercado. En la **Tabla 8** y en la **Figura 11** se muestran las observaciones del DJIA en el período de alta volatilidad identificado.

Tabla 8: Período con alta volatilidad en el índice DJIA

Date	close	volume	lag1_close	diferencia	Flag	f_extremo
18/08/2015	17511.3398	79900000	17545.1797	-33.83985	0	0
19/08/2015	17348.7305	104720000	17511.3398	-162.60937	0	0
20/08/2015	16990.6895	128530000	17348.7305	-358.04102	0	0
21/08/2015	16459.75	225170000	16990.6895	-530.93945	0	1
24/08/2015	15871.3496	293920000	16459.75	-588.40039	0	1
25/08/2015	15666.4404	213220000	15871.3496	-204.90918	0	0

Date	close	volume	lag1_close	diferencia	Flag	f_extremo
26/08/2015	16285.5098	208420000	15666.4404	619.06934	1	1
27/08/2015	16654.7695	171980000	16285.5098	369.25976	1	0
28/08/2015	16643.0098	131790000	16654.7695	-11.75976	0	0
31/08/2015	16528.0293	141440000	16643.0098	-114.98047	0	0
01/09/2015	16058.3496	171390000	16528.0293	-469.67969	0	1
02/09/2015	16351.3799	133480000	16058.3496	293.03027	1	0
03/09/2015	16374.7598	109730000	16351.3799	23.37989	1	0
04/09/2015	16102.3799	127270000	16374.7598	-272.37989	0	0

Figura 11: Período con alta volatilidad en el índice DJIA



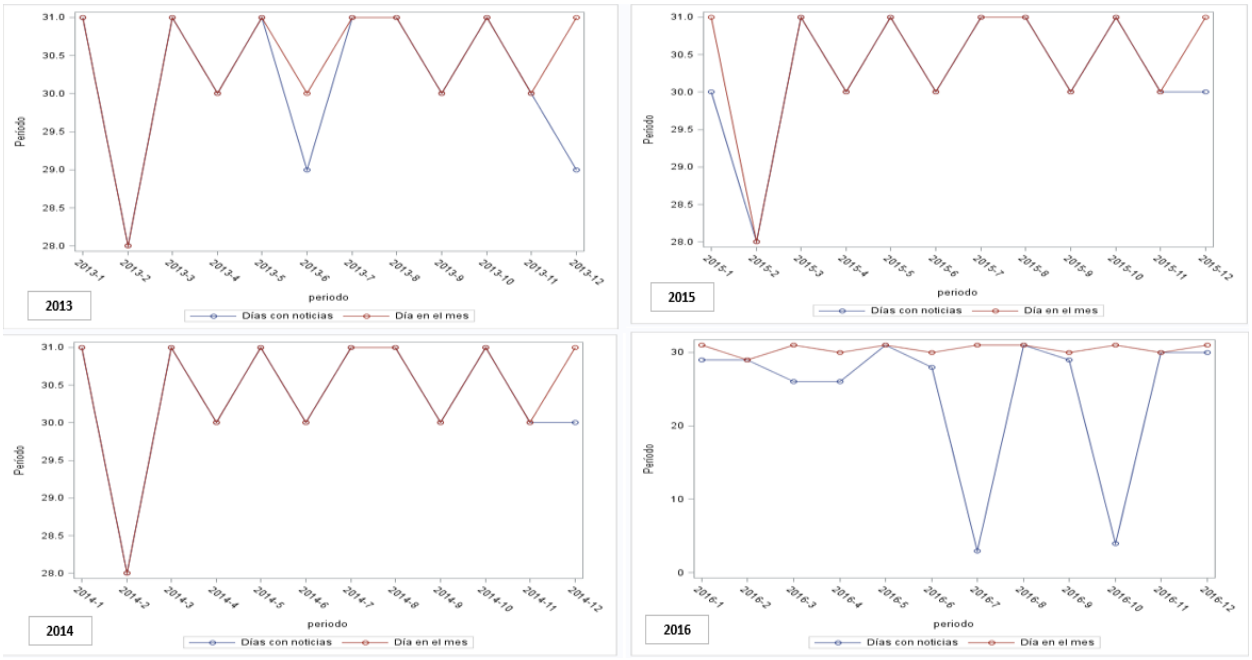
En la **Tabla 9** se muestran algunas de las noticias publicadas en el período de estrés identificado, como se puede observar los textos plasman la condición negativa que está viviendo el mercado.

Tabla 9: Muestra de noticias observadas en el período de estrés

created_time	from_name	Message
25Aug2015	CNBC	Yesterday's Dow looks nothing like today's Dow. Behold, a tale of two markets…
25Aug2015	CNBC	Monday morning's 1,000+ point drop in the Dow was scary and dramatic, but it was nothing like October 1987...
25Aug2015	Financial	A slowing Chinese economy and turbulent stock markets are having an impact on global commodity prices. We explain why.
24Aug2015	Bloomberg	JUST IN: S&P 500 drops 3.9% and falls into a correction for the first time since 2011 http://bloom.bg/1U9ob8z
24Aug2015	Bloomberg	U.S. stocks are bouncing back after this morning's massive plunge http://bloom.bg/1U9ob8z

Asimismo, se validó la completez de la descarga de información. Se construyeron gráficas que comparan el número de días en cada mes con el número de días en los que se descargó información de noticias, se observa que existen meses con información incompleta tal como diciembre de 2013, donde solamente se encontró información para 29 de los 31 días en el mes, también, se notó que a partir de julio de 2016 se encontró que existen meses sin descarga de información. Por lo anterior, se restringió la ventana de estudio de enero de 2013 a junio de 2016.

Figura 12: Completez de la información



2.4. Construcción de ABT

Se denomina ABT (Analytical Base Table, por sus siglas en inglés) a la tabla base para la construcción de un modelo. En este caso la ABT es la consolidación de los resultados de los procesos de construcción de variables que se aplicaron a la información extraída.

Además, de la construcción de las variables anteriormente mencionadas se consolidaron algunas otras para conseguir un nivel de información diaria. El resumen de las variables que finalmente conforman la ABT con su respectivo proceso de construcción se muestra en la siguiente tabla:

Tabla 10: Variables que conforman la ABT

Variable	Fuente	Técnica para construcción.
date_completa		
freq_noti_FBTS	Noticias	Suma
fuentes_disti_FBTS	Noticias	Contar distintos journals
likes_FBTS	Noticias	Suma
comments_FBTS	Noticias	Suma
shares_FBTS	Noticias	Suma
daily_neg_index	Noticias / Diccionario McDonald	Índice de Negatividad
Cluster_Topics1	Noticias	Topics de texto y clúster de variables
Cluster_Topics2	Noticias	Topics de texto y clúster de variables
Cluster_Topics3	Noticias	Topics de texto y clúster de variables
Cluster_Topics4	Noticias	Topics de texto y clúster de variables
Cluster_Topics5	Noticias	Topics de texto y clúster de variables
Cluster_Topics6	Noticias	Topics de texto y clúster de variables
Cluster_Topics7	Noticias	Topics de texto y clúster de variables
Cluster_Topics8	Noticias	Topics de texto y clúster de variables
Cluster_Topics9	Noticias	Topics de texto y clúster de variables
Cluster_Topics10	Noticias	Topics de texto y clúster de variables
GIS_Cluster1_GIS	Google Trends (GIS)	Clúster de variable
GIS_Cluster2_GIS	Google Trends (GIS)	Clúster de variable
GIS_Cluster3_GIS	Google Trends (GIS)	Clúster de variable

Variable	Fuente	Técnica para construcción.
GIS_Cluster4_GIS	Google Trends (GIS)	Clúster de variable
GIS_Cluster5_GIS	Google Trends (GIS)	Clúster de variable
close_correg_DJIA	DJIA	Corrección DJIA
flag_corregida_DJIA	DJIA	$Si\ DJIA_i \geq DJIA_{i-1}\ entonces\ 1$ En otro caso 0
log_return	DJIA	Log return
close_correg_VIX	VIX	Corrección de VIX

Cabe mencionar que en la ABT no existen valores ausentes ya que se realizó una imputación de datos previa a la tarea de modelación. La imputación aplicada en este trabajo consiste en la aplicación de splines cúbicos de manera similar a como se describe en la sección **2.2.5. Índice de Google Trends Diario** de este trabajo, la diferencia se encuentra en que aquí la construcción del polinomio a emplear en la imputación se realiza a partir de los dos valores no nulos que acotan por arriba y por abajo a la fecha del valor ausente.

Tras realizar la anterior técnica de imputación se observó que surgen valores extraños en algunas variables, por ejemplo, se observaron días con valor ausente en la variable de frecuencia de noticias que fueron rellenadas con un valor negativo.

Figura 13: Valores extraños tras imputación.

date_completa	freq_noti_FBTS	fuentes_distl_FBTS	likes_FBTS
26MAR2016	14	4	8350
27MAR2016	2.3870970766	1.3966044704	8755.6655328
28MAR2016	-5.426421921	-0.313489261	8422.5142411
29MAR2016	-9.997584053	-1.27331974	7524.8136397
30MAR2016	-11.88341638	-1.625925515	6236.8312433
31MAR2016	-11.64094598	-1.514345132	4732.8345668
01APR2016	-9.827199888	-1.081617139	3187.0911249
02APR2016	-6.999205185	-0.470780082	1773.8684325

Por lo anterior, se aplicaron las siguientes reglas para garantizar que la información imputada sea congruente con el contexto. Con lo anterior se garantiza, que no haya valores negativos en variables cuya naturaleza es tener valores mayores o iguales a cero, y también se mantiene la característica de que variables discretas no tengan valores imputados con parte decimal.

- Si alguna de las siguientes variables es negativa entonces transformar a 0.
    - freq_noti_FBT
    - fuentes_disti_FBTS
    - likes_FBTS
    - comments_FBTS
    - shares_FBTS
    - daily_neg_index
    - Cluster_Topics1
    - Cluster_Topics2
    - Cluster_Topics3
    - Cluster_Topics4
    - Cluster_Topics5
    - Cluster_Topics6
    - Cluster_Topics7
    - Cluster_Topics8
    - Cluster_Topics9
    - Cluster_Topics10

- Si una de las siguientes de las siguientes variables es no entera entonces aplicar función piso.
    - freq_noti_FBTS
    - comments_FBTS
    - likes_FBTS
    - shares_FBTS
    - Cluster_Topics1
    - Cluster_Topics2
    - Cluster_Topics3
    - Cluster_Topics4
    - Cluster_Topics5
    - Cluster_Topics6
    - Cluster_Topics7
    - Cluster_Topics8
    - Cluster_Topics9
    - Cluster_Topics10
    - fuentes_disti_FBTS

2.5. Exploración de datos

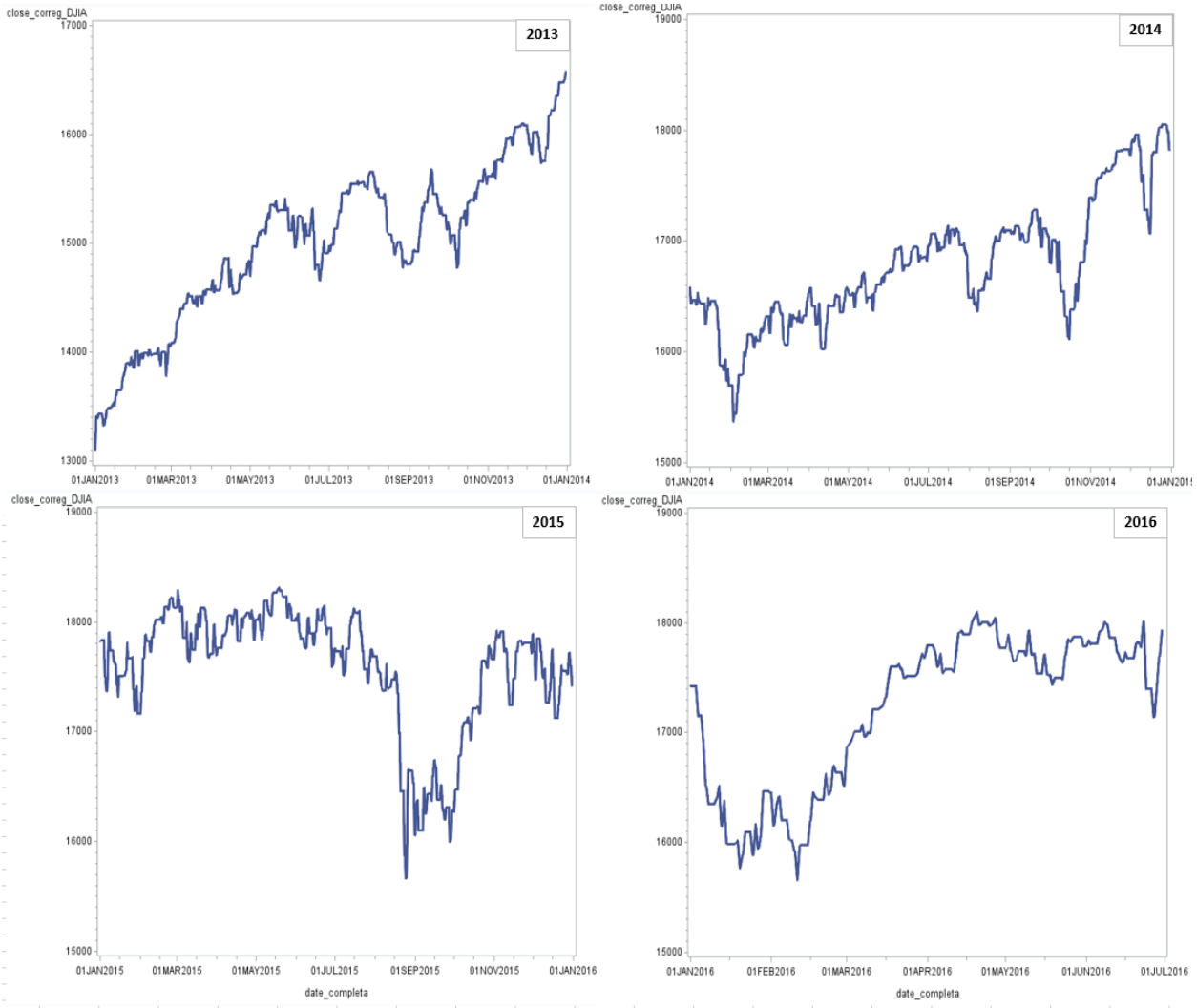
En esta sección se muestran los hallazgos realizados durante la fase de análisis exploratorio de las variables a utilizar en el desarrollo de este trabajo.

2.5.1. Análisis del Índice DJIA

En la

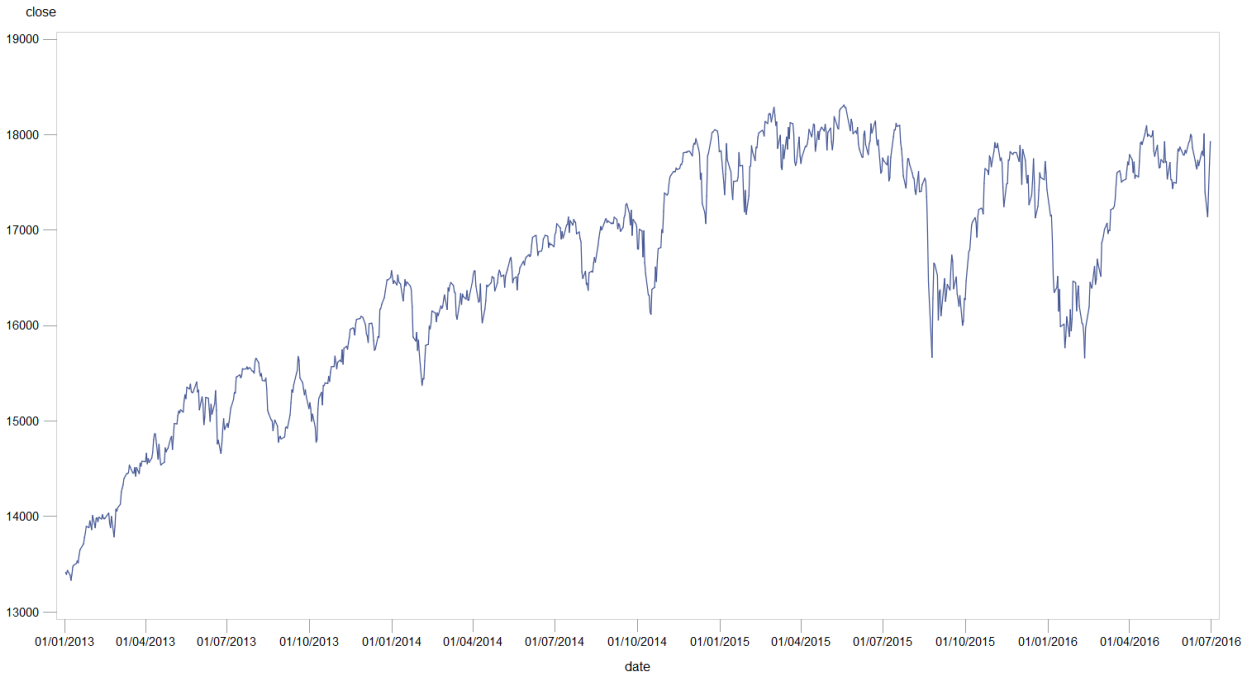
Figura 14 se puede observar que el índice DJIA creció en el 2014, pues existe un cambio en la escala de la gráfica a partir de este año. Asimismo, se pueden observar algunos periodos de recesión en el mercado, por ejemplo, el período más claro se localiza alrededor del mes de septiembre del 2015, lo cual coincide con el periodo de estrés identificado en la **Tabla 8** de la **2.3. Selección de la ventana de análisis**.

Figura 14: Comportamiento del DJIA por año



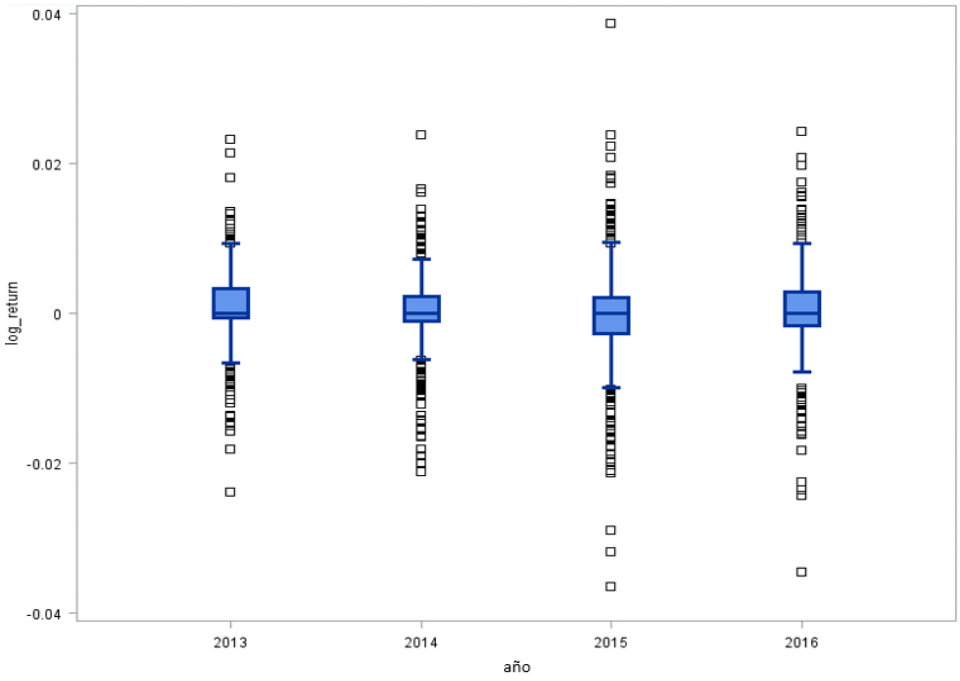
Por otro lado, en el siguiente gráfico se muestra el comportamiento total del índice. En este gráfico, resalta el hecho de que a partir de mediados del 2015 el índice entró en un período de recesión, y aunque posteriormente recuperó puntos no logró alcanzar la tendencia alcista que había mostrado hasta antes del 2015.

Figura 15: Comportamiento del DJIA



Por otro lado, también se analizó el comportamiento de los log-rendimientos del DJIA. Se observó que el año 2015 es el que presenta la mayor volatilidad, pues los bigotes del diagrama de caja son más amplios, asimismo, se observan más observaciones extremas.

Figura 16: Distribución anual del Log_Return



2.5.2. Análisis Univariado

- Noticias FB

Utilizando la API de Facebook para R se descargaron un total de 33,928 noticias provenientes de 7 periódicos distintos. Las noticias se descargaron en la ventana de tiempo del 3 de enero del 2010 al 21 septiembre de 2017, sin embargo, el análisis se restringió de enero del 2013 a junio 2016. La distribución de las noticias en cada uno de los periódicos es homogénea como se muestra en el siguiente cuadro, este hecho es importante pues al análisis no estará sesgado a la opinión de un periódico en particular. Asimismo, cabe señalar que se observaron algunas noticias cuyo único mensaje es NA, fueron 1648 noticias con textos que no aportan información relevante por lo que estas fueron retiradas y se conservaron únicamente 32,280 noticias para el desarrollo de este trabajo, el porcentaje de noticias retiradas fue el 4.8%.

Noticias con mensaje distinto de NA = 32,280

Tabla 11: Distribución de noticias por journal

Journal	Frecuencia	Porcentaje
Bloomberg	4792	14%
CNBC	4975	15%
CNN	4769	14%
FORBES	5016	15%
FINANCIAL TIMES	4576	13%
REUTERS	4750	14%
WALL STREET JOURNAL	5050	15%
TOTAL	33928	100%

Se revisó la unicidad de las noticias descargadas y se observó que existe aproximadamente un 2% de encabezados que se repiten al menos en dos ocasiones y el máximo número de repeticiones que llega a tener una noticia es de diez veces.

Tabla 12: Frecuencia de noticias repetidas

Número de Repeticiones	Noticias	Porcentaje
10	1	0.003%
8	3	0.009%

Número de Repeticiones	Noticias	Porcentaje
7	1	0.003%
6	5	0.015%
5	5	0.015%
4	7	0.022%
3	28	0.087%
2	179	0.555%
1	31714	98.247%
Total	32280	100.00%

A continuación, se muestran algunos ejemplos de las publicaciones repetidas, se puede observar que algunas de estas hacen referencia a anuncios publicados en los periódicos analizados.

Tabla 13: Ejemplo de noticias repetidas

Noticias	Repeticiones
DO YOU AGREE	10
EIGHT STORIES IN PHOTOS FROM AROUND THE WORLD	8
FOLLOW REUTERS ON INSTAGRAM WWW INSTAGRAM COM REUTERS	8
THIS COMPANY COULD POWER THE WORLD FOR 65 BILLION YEARS	8
WHITE HOUSE PRESS SECRETARY SEAN SPICER CONDUCTS THE DAILY BRIEFING WATCH ON FACEBOOKLIVE	7
WANT THE LATEST ON THE MEDIA BUSINESS SIGN UP FOR OUR CNN RELIABLE SOURCES NEWSLETTER HERE<U 2192> HTTP BIT LY 1JVM6U2	6
HACKING HUNGER HOW TO SURVIVE IN A FOOD DESERT	6
IN THE FUTURE MOTORCYCLE RIDERS MAY NOT NEED TO WORRY ABOUT WEARING HELMETS OR PADDED CLOTHING CC BMW VIA CNN TECH	6
DESPITE ITS RETRO ROOTS RECORDS SALES ARE GROWING CNNMONEY S CRISTINA ALESCI HEADS TO BROOKLYN TO VISIT ONE OF SIXTEEN RECORD PRESSING PLANTS IN THE UNITED STATES	6
DID YOUR CITY MAKE THE LIST	6

Noticias	Repeticiones
JUST IN	5
IT CAN CRAWL THROUGH SNOW IT CAN CRAWL THROUGH FIRE AND IT CAN SURVIVE GETTING RUN OVER BY A CAR	5
FROM OUR FRIENDS AT FINS	5
RESEARCHERS AT THE UNIVERSITY OF WASHINGTON HAVE INVENTED A PROTOTYPE PHONE THAT MAKES CALLS USING ENERGY FROM AMBIENT RADIO WAVES AND SUNLIGHT INSTEAD OF A BATTERY	5
THIS FLYING HOTEL CAN BE YOURS FOR 74 000 AN HOUR	5

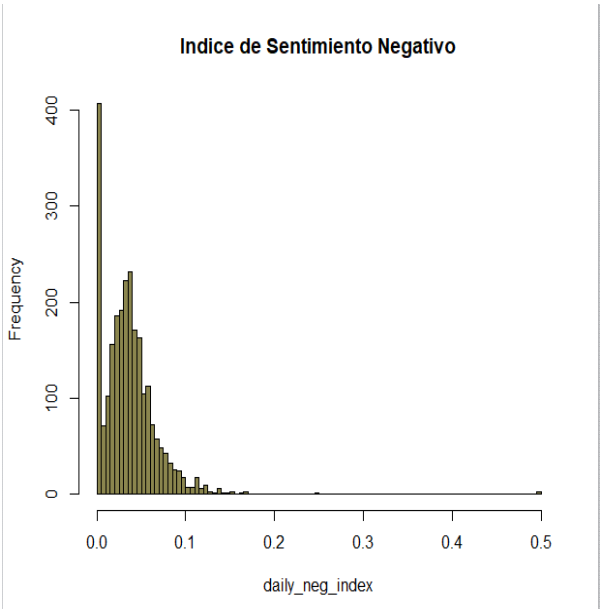
- Negative Index

El índice de negatividad presenta una menor dispersión en la información, sin embargo, hay un pico en valores bajos de esta variable.

Tabla 14: Distribución del índice de sentimiento negativo

Mean	Std Dev	Minimum	Maximum	Lower Quartile	Median	Upper Quartile
0.0373488	0.0356011	0	0.5	0.0166667	0.0333469	0.05

Figura 17: Distribución del índice de sentimiento negativo



Al analizar, el comportamiento del léxico negativo por journal se observa que Reuters es la editorial cuyos textos presentan mayor nivel de negatividad, debido a la cantidad de encabezados con índole negativa que publicó. Además, Reuters es el journal que en promedio utiliza mayor cantidad de palabras negativas por encabezado negativo.

Tabla 15: Negatividad por journal

from_name	ANÁLISIS DE PALABRAS							ÍNDICE DE NEGATIVIDAD					
	Noticias que contabilizaron palabras por Journal	Noticias con palabras negativas	% Noticias Negativas	Total de palabras	negative_words	Palabras por noticia	Palabras negativas por noticia	minimo	media	máximo	q1	mediana	q3
Bloomberg	4168	1243	30%	36037	1804	8.65	1.45	0.00	0.05	1.00	0.00	0.00	0.07
CNBC	4889	1384	28%	44506	1995	9.10	1.44	0.00	0.04	1.00	0.00	0.00	0.06
CNNMoney	4631	1158	25%	39428	1568	8.51	1.35	0.00	0.04	1.00	0.00	0.00	0.03
Financial	4281	1647	38%	55382	2505	12.94	1.52	0.00	0.04	1.00	0.00	0.00	0.07
Forbes	4820	1024	21%	50017	1514	10.38	1.48	0.00	0.03	1.00	0.00	0.00	0.00
Reuters	4428	1900	43%	75208	3335	16.98	1.76	0.00	0.04	1.00	0.00	0.00	0.07
The Wall	4934	1687	34%	64054	2546	12.98	1.51	0.00	0.04	0.50	0.00	0.00	0.06

Entre las palabras que frecuentemente se encuentran en los encabezados con tendencia negativas se hallan los términos: crisis, question y problem, los cuales se asocian a un contexto de incertidumbre que repercute en el comportamiento del mercado.

Tabla 16: Top 10 de palabras negativas por journal

BLOOMBERG	CNBC	CNNMoney	FINANCIAL	FORBES	REUTERS	WSJ
crisis	question	bad	crisis	questions	questions	break
claims	worst	claims	cut	bad	protest	questions
lost	questions	cut	questions	worst	protesters	cut
recession	crisis	unemployment	challenges	question	protests	lost
cut	bad	crisis	problem	cut	force	problems
worst	cut	problem	problems	problem	crisis	bad
poor	lost	recession	worst	break	violence	question
force	problem	questions	concerns	lost	victims	worst
bad	recession	problems	fears	miss	damaged	disagree
dropped	break	fear	lost	lose	injured	victims

- Indicadores De Impacto Social

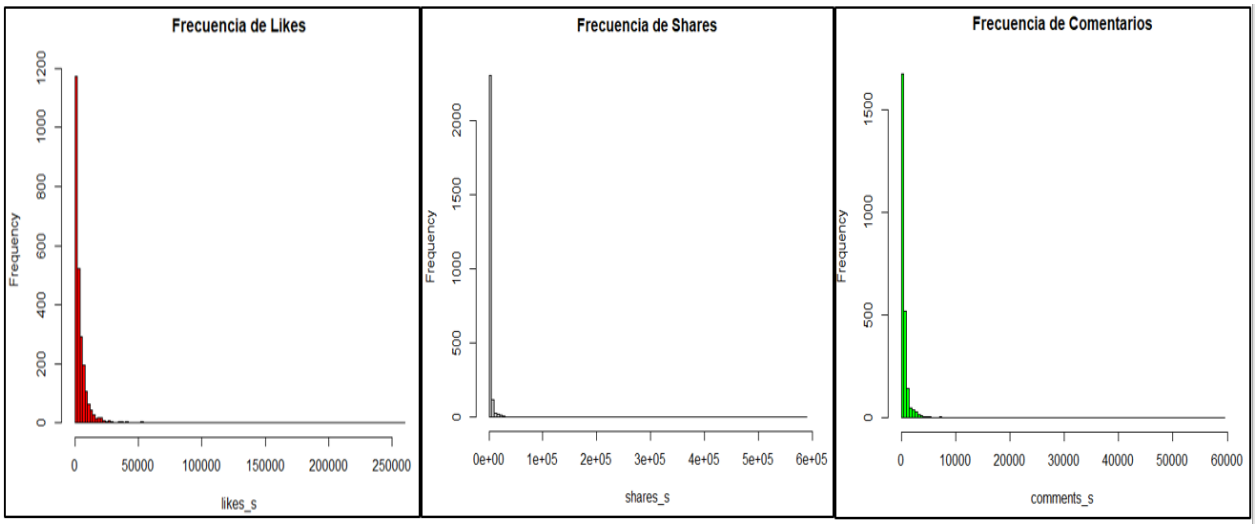
Existen variables asociadas a la respuesta social en Facebook de las publicaciones que se están analizando, estas reacciones son: número de me gusta, número de comentarios y número de veces que se comparte la publicación. En la **Tabla 17** se observan los estadísticos globales de estas variables, cabe destacar el hecho de que existen noticias con cero reacciones.

Tabla 17: Estadísticas de los indicadores de impacto social.

Variable	Mean	Std Dev	Coefficient of variation	Min.	Max.	Q1	Q2	Q3	Miss
Likes	4526.77	10597.34	0.427	1	259601	656	2273.5	5324.5	0
comments	706.58	2220	0.318	0	59427	129	312	644.5	0
Shares	2919.83	17720.81	0.164	0	588089	50.5	718	2010	0

Asimismo, al observar los histogramas de estos indicadores de reacción social se observa que existe una gran dispersión en los datos, y que los tres indicadores presentan una fuerte concentración en valores bajos.

Figura 18: Histogramas de los indicadores de impacto social



2.5.3. Análisis Temporal

- Noticias FB

Otro aspecto que es importante analizar es que la noticias se encuentren bien distribuidas en los días de la semana, en el siguiente cuadro se observa que existe información disponible de manera homogénea a lo largo de la semana, lo cual es buen indicador de una buena distribución de la información.

Tabla 18: Frecuencia de días en la semana donde se encuentran noticias.

WEEKDAY	FREQUENCY	PERCENT
Sunday	334	13.23%
Monday	369	14.62%
Tuesday	366	14.50%

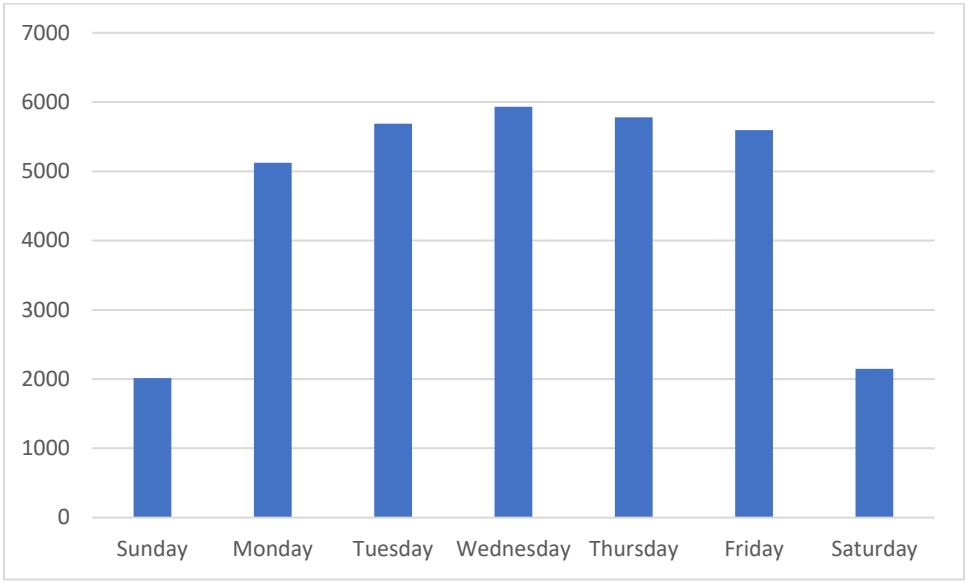
WEEKDAY	FREQUENCY	PERCENT
Wednesday	372	14.74%
Thursday	374	14.82%
Friday	373	14.78%
Saturday	336	13.31%
TOTAL	2524	100%

Por otro lado, también es necesario revisar la frecuencia de noticias que se encuentran en cada uno de los días de la semana. Se puede observar que hay mayor cantidad de noticias en la parte central de la semana, y en los fines de semana es donde se observa menor cantidad de noticias.

Tabla 19: Frecuencia de noticias por día de la semana.

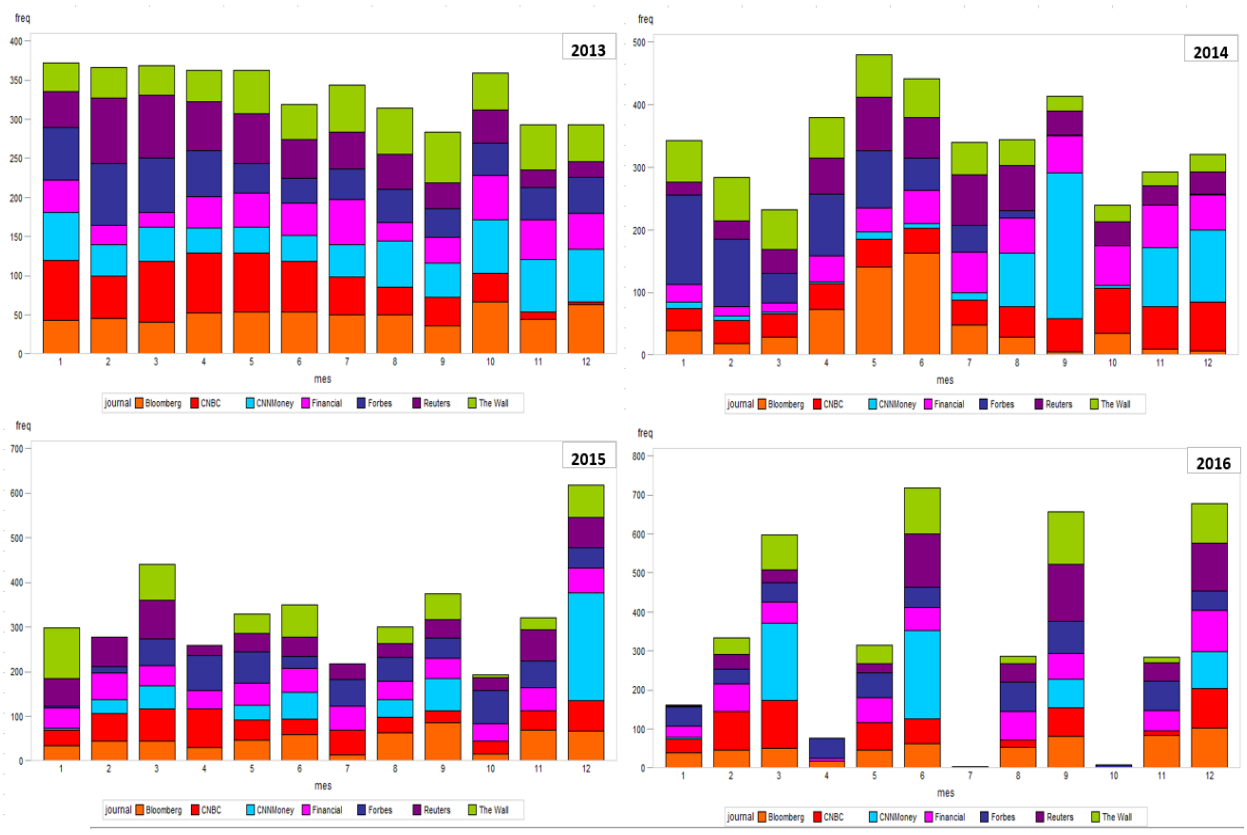
WEEKDAY	FREQUENCY	PERCENT
Sunday	2014	6.24%
Monday	5125	15.88%
Tuesday	5687	17.62%
Wednesday	5930	18.37%
Thursday	5780	17.91%
Friday	5597	17.34%
Saturday	2147	6.65%
TOTAL	32280	100%

Figura 19: Frecuencia de noticias por día de la semana



En la **Figura 20** se puede observar la distribución de noticias a lo largo del tiempo por periódico, cada fuente se representa por un color. Se observa que no existe predominancia global de un solo periódico, aunque sí existen lapsos donde predomina una fuente. Este hecho es importante, ya que no habrá sesgo por el estilo de escritura de una fuente en particular.

Figura 20: Distribución por journal a lo largo del tiempo



2.5.4. Análisis Variables versus DJIA

A continuación, se muestra la matriz de correlaciones entre las distintas variables construidas y el índice DJIA, se puede notar que hay variables correlacionadas con el índice DJIA, lo cual marca a estas variables como posibles variables explicativas. Por otro lado, es necesario añadir que hay niveles de alta correlación entre algunas variables independientes como lo son los indicadores de impacto social, lo cual es alerta de que hay que tener cuidado al incorporar estas variables al modelo para no llegar a tener problemas de multicolinealidad. La escala de colores aplicada es la descrita en la **Tabla *4***.

Figura 21: Matriz de correlaciones parte 1

Coeficientes de correlación Pearson, N = 1277

	close_correg_DJIA	freq_noti_FBTS	fuentes_disti_FBTS	likes_FBTS	comments_FBTS	shares_FBTS	daily_neg_index	Cluster_Topics1	Cluster_Topics2	Cluster_Topics3	Cluster_Topics4	Cluster_Topics5
close_correg_DJIA	1.000	-0.002	-0.190	0.114	0.073	0.073	0.052	-0.282	0.057	-0.263	-0.052	-0.201
freq_noti_FBTS	-0.002	1.000	0.766	0.258	0.178	0.125	-0.094	0.699	0.585	0.401	0.402	0.311
fuentes_disti_FBTS	-0.190	0.766	1.000	0.156	0.099	0.069	-0.156	0.599	0.378	0.442	0.312	0.272
likes_FBTS	0.114	0.258	0.156	1.000	0.753	0.756	-0.030	0.110	0.229	0.064	0.145	0.037
comments_FBTS	0.073	0.178	0.099	0.753	1.000	0.905	-0.029	0.083	0.185	0.009	0.113	0.025
shares_FBTS	0.073	0.125	0.069	0.756	0.905	1.000	-0.013	0.035	0.106	-0.032	0.087	0.015
daily_neg_index	0.052	-0.094	-0.156	-0.030	-0.029	-0.013	1.000	-0.055	0.001	-0.044	-0.022	0.067
Cluster_Topics1	-0.282	0.699	0.599	0.110	0.083	0.035	-0.055	1.000	0.415	0.372	0.330	0.325
Cluster_Topics2	0.057	0.585	0.378	0.229	0.185	0.106	0.001	0.415	1.000	0.112	0.268	0.117
Cluster_Topics3	-0.263	0.401	0.442	0.064	0.009	-0.032	-0.044	0.372	0.112	1.000	0.176	0.256
Cluster_Topics4	-0.052	0.402	0.312	0.145	0.113	0.087	-0.022	0.330	0.268	0.176	1.000	0.125
Cluster_Topics5	-0.201	0.311	0.272	0.037	0.025	0.015	0.067	0.325	0.117	0.256	0.125	1.000
Cluster_Topics6	-0.127	0.318	0.287	0.031	0.022	0.002	-0.031	0.263	0.170	0.160	0.129	0.184
Cluster_Topics7	-0.014	0.248	0.246	0.039	0.008	-0.006	0.094	0.200	0.121	0.241	0.141	0.092
Cluster_Topics8	-0.208	0.653	0.597	0.104	0.053	0.022	-0.021	0.523	0.296	0.406	0.236	0.286
Cluster_Topics9	-0.226	0.299	0.297	0.016	0.009	-0.016	-0.037	0.333	0.127	0.245	0.128	0.259
Cluster_Topics10	-0.143	0.358	0.313	0.035	0.025	0.011	-0.034	0.344	0.164	0.189	0.169	0.189
GIS_Cluster1_GIS	-0.738	-0.041	0.071	-0.096	-0.044	-0.045	-0.027	0.203	-0.061	0.246	0.092	0.138
GIS_Cluster2_GIS	-0.373	-0.015	-0.081	-0.039	-0.007	-0.014	0.054	0.075	0.003	0.092	0.068	0.057
GIS_Cluster3_GIS	0.393	-0.071	-0.198	0.107	0.072	0.068	0.063	-0.208	0.047	-0.176	-0.053	-0.130
GIS_Cluster4_GIS	0.035	-0.085	-0.199	-0.024	-0.011	-0.009	0.048	-0.059	-0.055	0.055	0.020	-0.001
GIS_Cluster5_GIS	-0.324	0.001	0.048	-0.006	-0.041	-0.041	-0.076	0.108	-0.024	0.149	0.063	0.041
close_correg_VIX	-0.002	0.026	-0.057	0.088	0.059	0.067	-0.014	-0.064	0.137	-0.161	0.043	-0.118

Figura 22: Matriz de correlaciones parte 2.

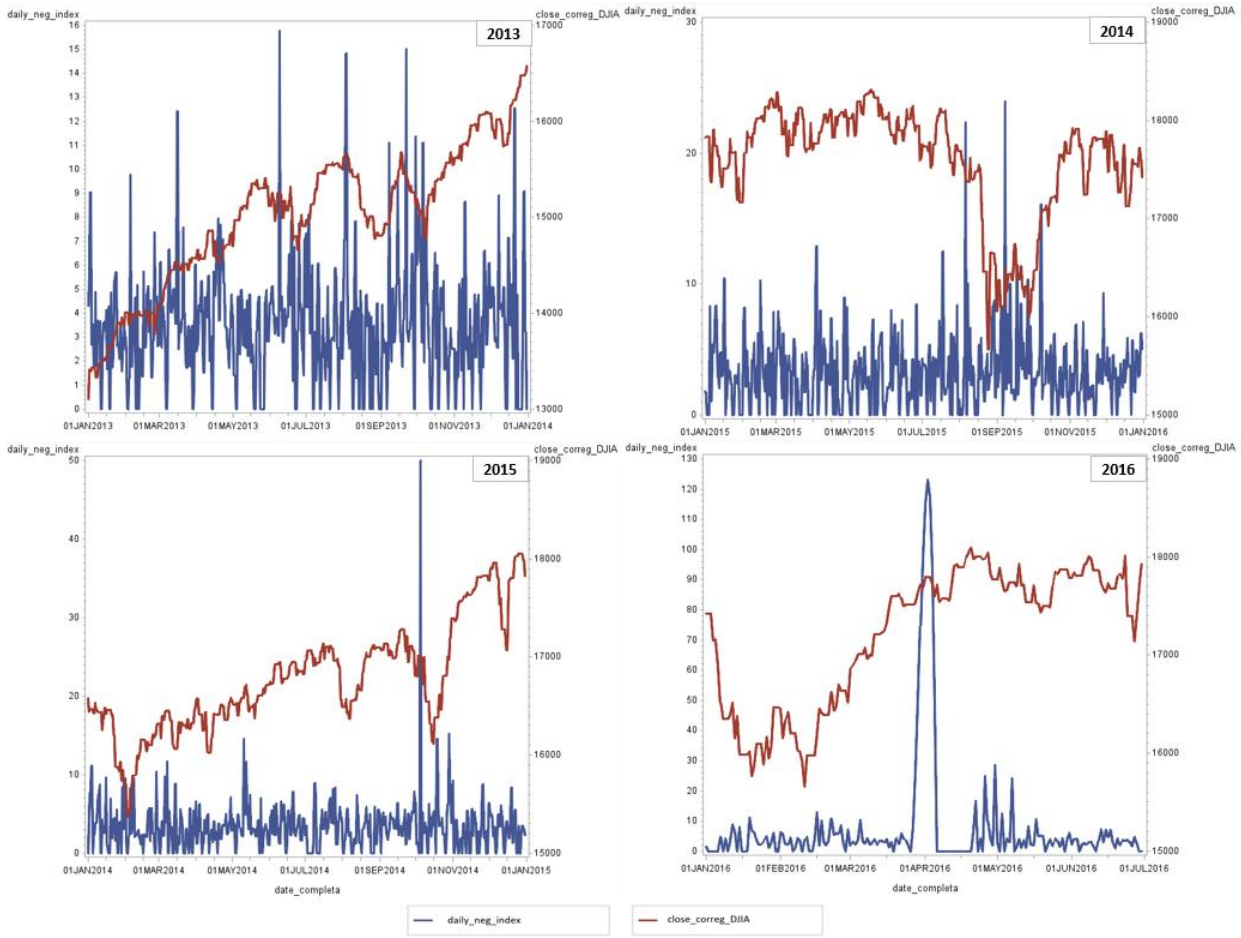
Coeficientes de correlación Pearson, N = 1277

	Cluster_Topics6	Cluster_Topics7	Cluster_Topics8	Cluster_Topics9	Cluster_Topics10	GIS_Cluster1_GIS	GIS_Cluster2_GIS	GIS_Cluster3_GIS	GIS_Cluster4_GIS	GIS_Cluster5_GIS	close_correg_VIX
close_correg_DJIA	-0.127	-0.014	-0.208	-0.226	-0.143	-0.738	-0.373	0.393	0.035	-0.324	-0.002
freq_noti_FBTS	0.318	0.248	0.653	0.299	0.358	-0.041	-0.015	-0.071	-0.085	0.001	0.026
fuentes_disti_FBTS	0.287	0.246	0.597	0.297	0.313	0.071	-0.081	-0.198	-0.199	0.048	-0.057
likes_FBTS	0.031	0.039	0.104	0.016	0.035	-0.096	-0.039	0.107	-0.024	-0.006	0.088
comments_FBTS	0.022	0.008	0.053	0.009	0.025	-0.044	-0.007	0.072	-0.011	-0.041	0.059
shares_FBTS	0.002	-0.006	0.022	-0.016	0.011	-0.045	-0.014	0.068	-0.009	-0.041	0.067
daily_neg_index	-0.031	0.094	-0.021	-0.037	-0.034	-0.027	0.054	0.063	0.048	-0.076	-0.014
Cluster_Topics1	0.263	0.200	0.523	0.333	0.344	0.203	0.075	-0.208	-0.059	0.108	-0.064
Cluster_Topics2	0.170	0.121	0.296	0.127	0.164	-0.061	0.003	0.047	-0.055	-0.024	0.137
Cluster_Topics3	0.160	0.241	0.406	0.245	0.189	0.246	0.092	-0.176	0.055	0.149	-0.161
Cluster_Topics4	0.129	0.141	0.236	0.128	0.169	0.092	0.068	-0.053	0.020	0.063	0.043
Cluster_Topics5	0.184	0.092	0.286	0.259	0.189	0.138	0.057	-0.130	-0.001	0.041	-0.118
Cluster_Topics6	1.000	0.087	0.243	0.139	0.230	0.076	0.021	-0.035	-0.053	0.034	0.002
Cluster_Topics7	0.087	1.000	0.200	0.111	0.092	0.032	0.035	-0.032	0.063	-0.003	-0.014
Cluster_Topics8	0.243	0.200	1.000	0.273	0.358	0.134	0.053	-0.192	-0.078	0.082	-0.102
Cluster_Topics9	0.139	0.111	0.273	1.000	0.209	0.135	0.067	-0.127	-0.046	0.086	-0.092
Cluster_Topics10	0.230	0.092	0.358	0.209	1.000	0.100	0.032	-0.130	-0.059	0.051	-0.053
GIS_Cluster1_GIS	0.076	0.032	0.134	0.135	0.100	1.000	0.647	-0.357	0.444	0.282	0.062
GIS_Cluster2_GIS	0.021	0.035	0.053	0.067	0.032	0.647	1.000	-0.009	0.676	0.223	0.158
GIS_Cluster3_GIS	-0.035	-0.032	-0.192	-0.127	-0.130	-0.357	-0.009	1.000	0.178	-0.014	0.375
GIS_Cluster4_GIS	-0.053	0.063	-0.078	-0.046	-0.059	0.444	0.676	0.178	1.000	0.167	0.150
GIS_Cluster5_GIS	0.034	-0.003	0.082	0.086	0.051	0.282	0.223	-0.014	0.167	1.000	-0.008
close_correg_VIX	0.002	-0.014	-0.102	-0.092	-0.053	0.062	0.158	0.375	0.150	-0.008	1.000

También se realizaron gráficas para ver el comportamiento de las variables versus el comportamiento del DJIA, a continuación, se muestras algunos hallazgos.

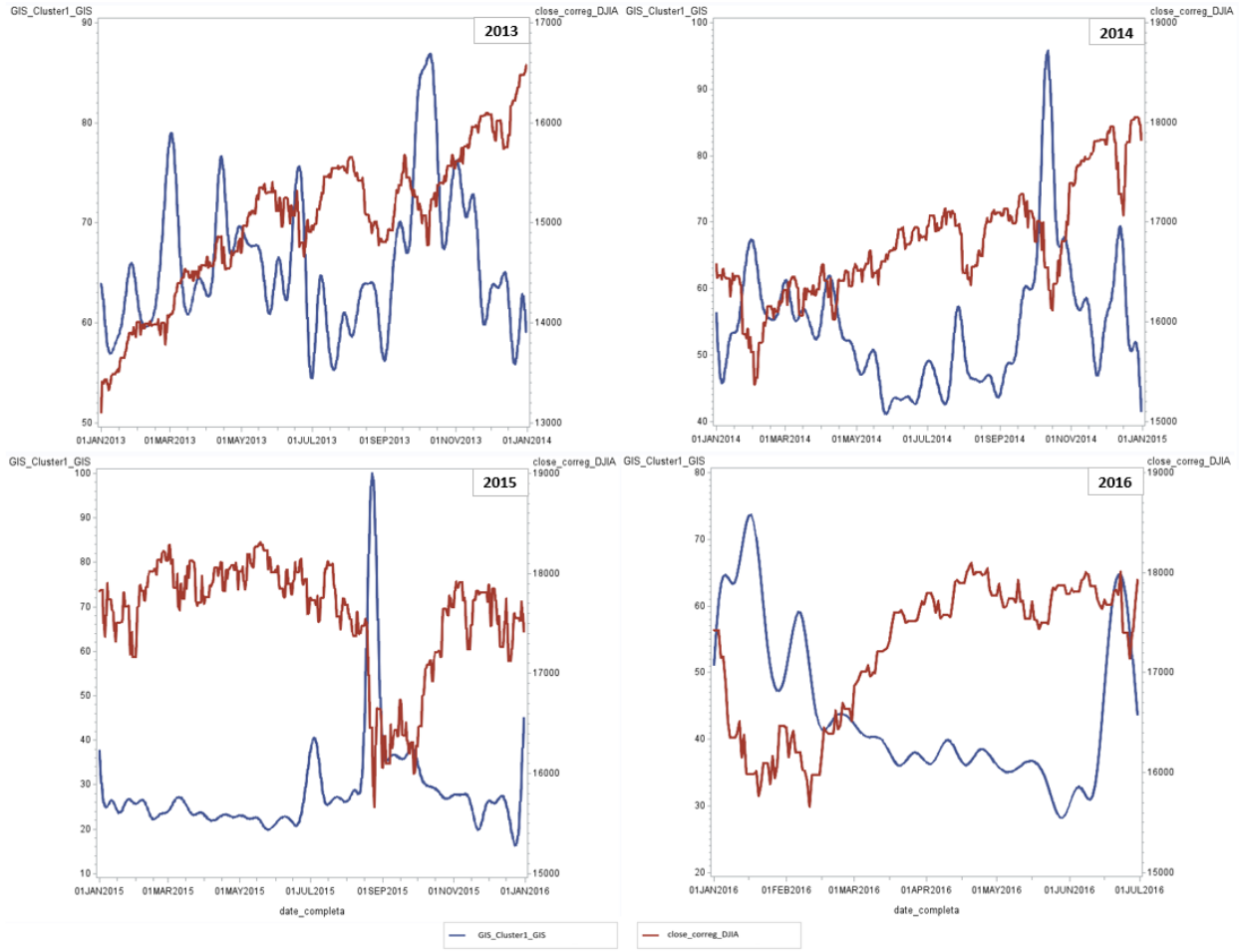
En las gráficas del índice de negatividad versus el DJIA, se puede observar que algunos picos del índice de negatividad ocurren previamente a caídas en el DJIA, por lo que se podría suponer que antes de un periodo de recesión los medios de comunicación empezarán a reflejar un estado de incertidumbre, volatilidad u otros efectos asociados a un entorno negativo.

Figura 23: Índice de negatividad versus Close DJIA



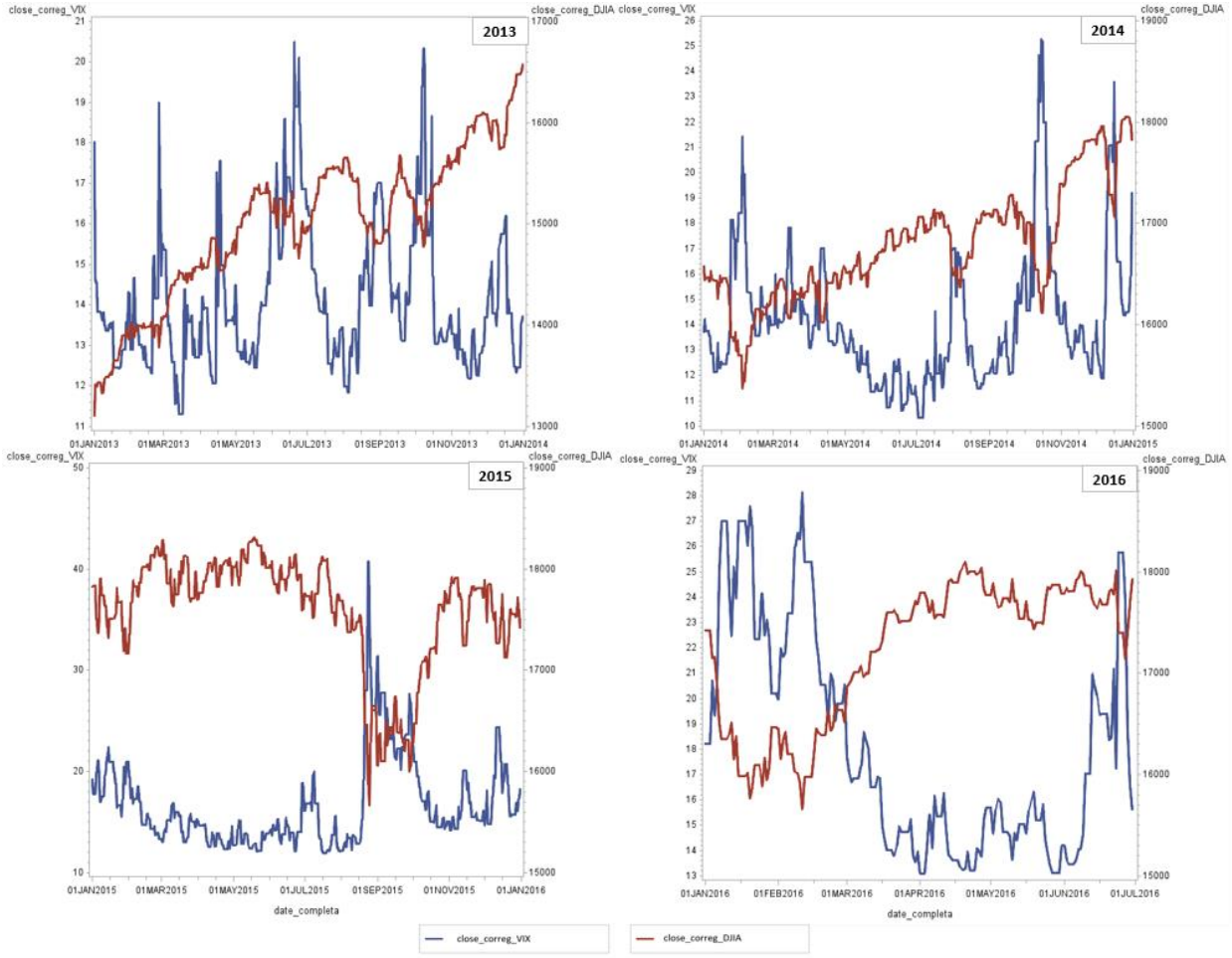
Por otro lado, la gráfica del clúster 1 de términos GIS versus el DJIA, muestra un comportamiento simétrico. El clúster GIS 1 contiene lo términos: finance_news, financial_news, stock_market_news, stock_market_today, Djia, Dow, Dow_Jones, Dow_Jones_Industrial_Average, bear_market, stock_fall, stock_market_crash, SP500, stock, stock_market, este comportamiento simétrico se puede interpretar como que: un aumento en las búsquedas de términos asociados al mercado en Google pueden indicar que el mercado está viviendo una situación a la baja.

Figura 24: Clúster 1 de términos GIS versus Close DJIA



Finalmente, la gráfica del índice VIX versus el DJIA también muestra un comportamiento simétrico, es decir cuando la volatilidad aumenta, el índice DJIA tiende a ir a la baja. Hay que recordar que la volatilidad está asociada a la inestabilidad de los precios en el mercado financiero y esto a su vez con la rentabilidad de los activos. Por lo que, un aumento en la volatilidad conlleva a los inversionistas a tener una postura más estrecha, lo que suele desvalorizar a los activos financieros y desencadenar un periodo con tendencia bajista como se observa en la gráfica.

Figura 25: Índice VIX versus Close DJIA

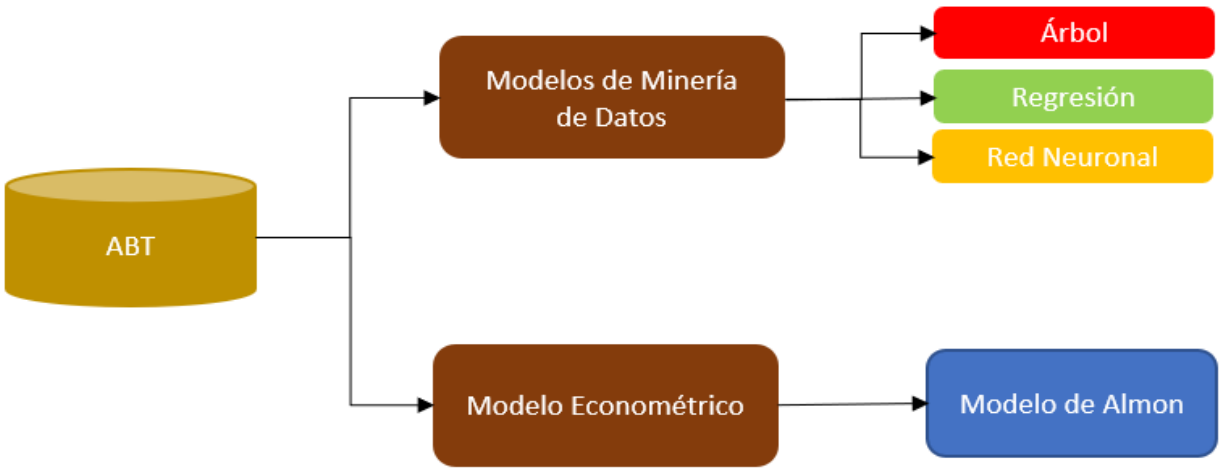


Capítulo 3: Técnicas de Modelación

Durante la realización de este trabajo se emplearon técnicas pertenecientes a dos grandes áreas de conocimiento que son la Minería de Datos y la Econometría. El objetivo de este capítulo es dar una explicación de las técnicas empleadas para pronosticar la tendencia del índice Dow Jones Industrial Average (DJIA).

Se debe agregar que las técnicas de minería de datos empleadas son: ***Árbol de Regresión, Regresión y Red Neuronal***. Por otro lado, el modelo econométrico empleado es el ***Modelo de Almon***.

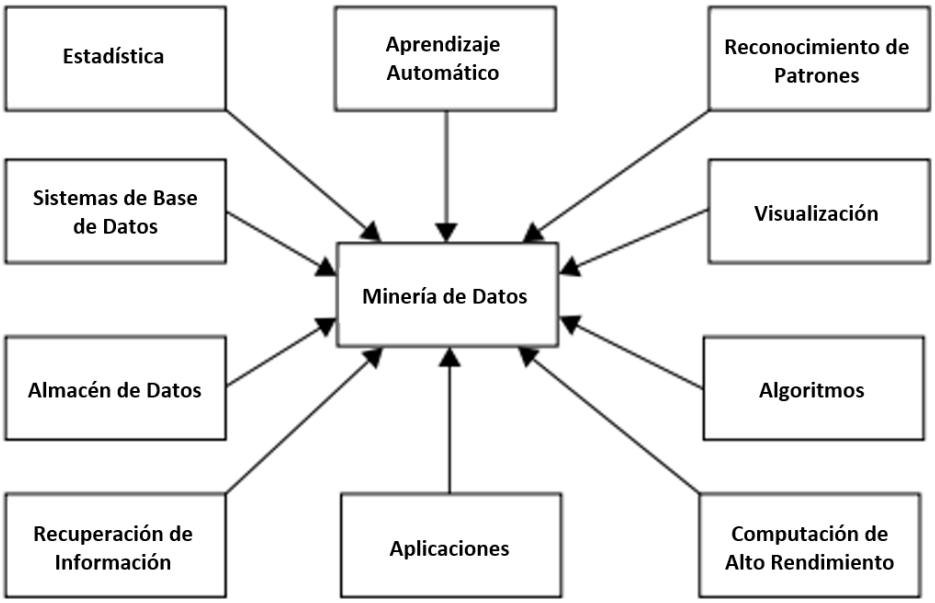
Figura 26: Técnicas de modelación empleadas



3.1. Minería de Datos

La minería de datos hace referencia a la recopilación de técnicas destinadas a la identificación de patrones, anomalías y correlaciones existentes, pero no visibles en un conjunto de datos. La minería de datos incorpora técnicas de diferentes áreas como son: estadística, machine learning, reconocimiento de patrones, etc. Al ser un área interdisciplinaria existe una extensa área de aplicaciones para la minería de datos.

Figura 27: Minería de datos.



La minería de datos se puede dividir en dos grandes enfoques de aprendizaje: supervisado y no supervisado. El aprendizaje supervisado hace referencia a tener observaciones etiquetadas en los datos de entrenamiento del modelo. Por otro lado, el aprendizaje no supervisado o reconocimiento de patrones hace referencia a contar con datos de entrenamiento no etiquetados.

En este trabajo se desea representar la asociación entre las variables de entrada y la variable objetivo, que en este caso es la tendencia medida matemáticamente por los log-rendimiento. Por lo anterior, este trabajo se desarrolla en la rama del aprendizaje supervisado de la minería de datos. Particularmente, en este trabajo se utilizan tres importantes técnicas que son: **Árbol de regresión**, **Regresión Lineal** y **Red neuronal**. Estas técnicas serán el puente para resolver la tarea de pronosticar el DJIA, se decidió emplear estas técnicas ya que se consideran los algoritmos básicos para modelación supervisada, con base en el manual Applied Analytics Using SAS Enterprise Miner[16].

Otra característica importante al construir un modelo es conocer el tipo de pronóstico que se va a realizar, básicamente los tipos de pronóstico se pueden dividir en tres: **decisión**, **ordenamiento** y **estimación**.

- Decisión: Usualmente están asociadas a algún tipo de decisión (p. ej. compra – no compra, dona – no dona), por lo que los problemas de tipo decisión también se conocen como clasificaciones.
- Ordenamiento: Se asocian a una condición de concordancia, casos de alto valor tienen una puntuación alta y viceversa.
- Estimación: Aproximan el valor esperado de la variable objetivo. Se debe resaltar que un pronóstico de tipo estimación puede ser transformado tanto a un pronóstico de tipo decisión como de tipo ranqueo.

[16] P. Christie, J. Georges, J Thompson and C. Wells. (2011). Applied Analytics Using SAS Enterprise Miner Course Notes. SAS Institute.

3.1.1. Árbol de Decisión.

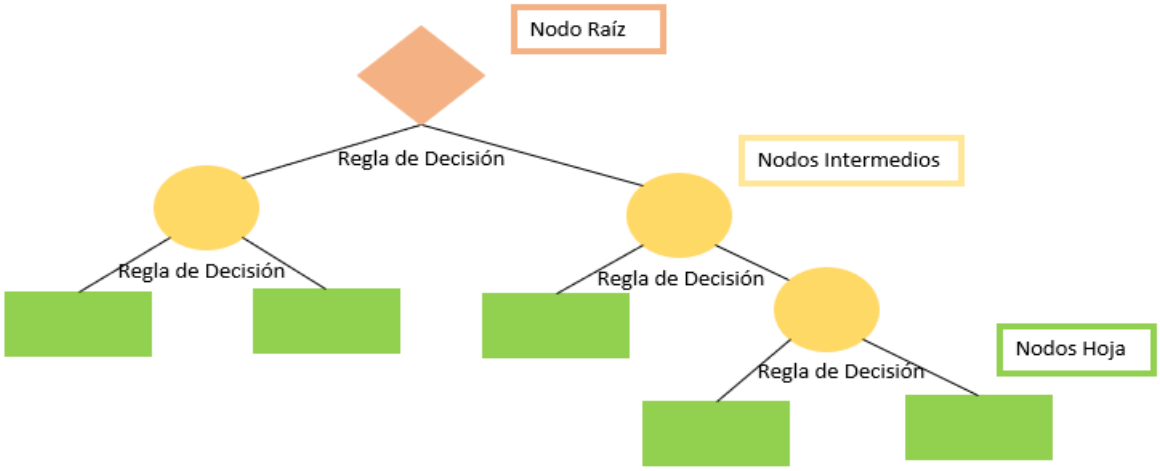
Los árboles de decisión son algoritmos que califican los casos usando **Reglas de Predicción** (Prediction Rules), seleccionan variables importantes para el modelo por medio de **Algoritmo de Búsqueda** (Split-Search) y controlan la complejidad del modelo por medio de **Poda** (Pruning).

Existen dos categorías de árboles, los cuales se diferencian por el tipo de variable objetivo, si la variable es categórica se le conoce como **árbol de clasificación**, si la variable es numérica se le conoce como **árbol de regresión**.

Las reglas del árbol se acomodan en una estructura jerárquica compuesta por nodos y líneas, cada nodo representa una regla de decisión y las líneas el resultado de la regla. La primera regla del árbol se denomina nodo raíz, las reglas siguientes se conocen como nodos interiores, finalmente los nodos con solo una conexión se llaman nodos hoja.

Para evaluar un nuevo caso, se requiere aplicar las reglas definidas a los valores de entrada de la nueva observación, de esta forma eventualmente el nuevo caso caerá en una única hoja del árbol. Los nodos hojas proporcionan la estimación del modelo, es decir todos los valores que se agrupen en una hoja tendrán el mismo valor pronosticado.

Figura 28: Estructura de árbol de decisión



- Construcción del árbol

La selección de variables de entrada útiles, se realiza por medio de **Algoritmo de Búsqueda**. Este algoritmo inicia seleccionando una variable de entrada y buscando el punto óptimo para particionar en caso de una variable continua, o en agrupar los niveles de una variable categórica de tal forma que se generen dos grupos de observaciones. Tras haber realizado la partición (Split) y en este caso por haber trabajado con arboles binarios se utilizaron matrices de 2x2, donde las columnas hacen referencia a las ramas que se formaron (valores pronosticados), y los renglones hacen referencia a la variable objetivo (valores observados)[17].

[17] En caso de una variable objetivo continua se pone un punto de corte para dividir a la población en categorías.

Se evalúa la calidad de la partición al realizar una **prueba $\mathbf{X^2}$** (chi cuadrado) para independencia. Se busca tener valores altos del estadístico pues esto indica que existe diferencia en las proporciones de los casos clasificados en cada hoja, por lo que se tiene una buena partición. La anterior idea se generaliza al utilizar el **valor-p** asociado al estadístico para construir otra métrica que es el **logworth.**

El **logworth** se define como: -log (valor-p de la prueba X^2). Para considerar que un split es adecuado, el logworth asociado debe de superar un horizonte marcado, se considera que el mejor split para una variable de entrada es aquel que mantiene el **logworth** más alto.

Posteriormente, el algoritmo del árbol busca el mejor **split** para cada variable de entrada y compara los respectivos **logworth**. El **split** con el mayor **logworth** se elige para crear la siguiente regla de partición.

El proceso descrito se repite en cada una de las ramas formadas, hasta que ninguna variable cumpla la condición de **split**, cuando se llega a este punto se dice que se obtuvo el árbol maximal.

- Poda del árbol

El árbol maximal es el árbol más complejo que se puede construir a partir de un conjunto de datos de entrenamiento, esta característica de contar con demasiadas reglas de decisión provoca que exista la posibilidad de que él árbol no generalice correctamente el comportamiento de los datos. Cuando sucede lo anterior se dice que el modelo está sobreajustado, el algoritmo con el que cuentan los árboles para evitar este problema es la poda.

Este algoritmo consiste en ir removiendo ramas al árbol maximal, de este modo se obtiene un subconjunto de árboles con determinado nivel de complejidad (número de ramas). Posteriormente, se comparan los árboles pertenecientes a cada nivel de complejidad con base a un estadístico de ajuste y se selecciona al mejor como el representante de ese nivel, este procedimiento se repite en cada subconjunto de árboles con un mismo número de ramas.

Finalmente, se selecciona al árbol más simple con el mayor estadístico de ajuste, este procedimiento se ejemplifica en la siguiente figura.

Figura 29: Algoritmo de poda de un árbol de decisión.

- Ventajas y desventajas de los árboles

En esta sección se enlistan algunas de las ventajas y desventajas de esta técnica.

Ventajas:

- El aspecto visual de los árboles hace a esta técnica fácil de entender e interpretar.
- Los valores ausentes o atípicos no afectan a los árboles de decisión, por lo que requiere menor tiempo de procesamiento de datos.
- Se pueden usar variables de tipo categórico o numérico.
- Los árboles de decisión son un modelo no paramétrico por lo que no requiere ningún supuesto de linealidad entre sus variables, ni tampoco requiere supuestos sobre su distribución.

Desventajas:

- El árbol de decisión tiende a sobre ajustarse por lo que es de suma importancia vigilar el proceso de poda.
- No es adecuado utilizar esta técnica para pronosticar los valores de una variable continua.
- Es más eficiente cuando existe correlación entre las variables a utilizar, si no existe dicha correlación el árbol presenta un pobre desempeño.

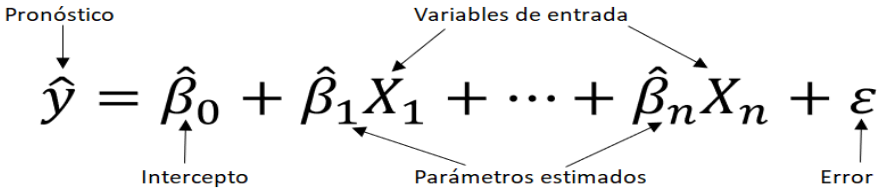
3.1.2. Regresión Lineal

Las regresiones son un modelo paramétrico por lo que asumen una estructura de asociación entre las variables independientes y la variable objetivo.

- Construcción de la regresión

Las regresiones estiman sus nuevos casos usando una ecuación matemática que considera los valores de cada una de las variables de entrada como se muestra en la siguiente imagen.

Figura 30: Ecuación de pronóstico de una regresión



En la anterior combinación lineal de variables de entrada, el intercepto centra el rango de valores de los pronósticos y el resto de los parámetros estimados determinan la fuerza y tendencia entre cada **variable de entrada** y la **variable objetivo**. El intercepto y los parámetros estimados se determinan a modo de minimizar el error cuadrado entre los valores pronosticados y los valores observados de la variable objetivo.

$$Error\ cuadrado = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Método de selección de variables

Se busca encontrar la combinación de variables óptima, es decir, que minimice el error cuadrado. Una primera solución podría ser probar todas las combinaciones de variables posibles, sin embargo, este método es muy exhaustivo e impráctico pues el número de modelos crece de manera exponencial conforme las variables de entrada aumentan, para ser precisos la cantidad de modelos a evaluar en una búsqueda exhaustiva sigue la siguiente relación:

$$Número\ de\ Modelos = 2^k - 1, con\ k\ el\ número\ de\ variables\ de\ entrada.$$

Por lo anterior, las regresiones utilizan métodos secuenciales de selección de variables para formar modelos, específicamente son tres los métodos que se utilizan para realizar la selección. Estos métodos se describen a continuación, sin embargo, hay que mencionar que por practicidad en este trabajo se optó por solo usar el método de **Stepwise**.

- **Forward**: Crea una secuencia de modelos incrementando el nivel de complejidad (número de variables). La secuencia inicia con un modelo de referencia que pronostica a todas las observaciones el valor promedio de la **variable objetivo**. Posteriormente, el algoritmo busca de forma secuencial variables que añadir al modelo para mejorar el modelo del paso anterior.
- **Backward**: Crea una secuencia de modelo decrementando el nivel de complejidad. Este método inicia la secuencia con un modelo saturado, todas las variables participan. Posteriormente, las variables son secuencialmente removidas.

- **Stepwise**: Combina elementos de los métodos anteriores. El proceso inicia como el método **forward** y secuencialmente se añaden variables, sin embargo, este método reevalúa la significancia de las variables incluidas si alguna no cumple esta condición se removerá del modelo.

Es necesario añadir que en las anteriores técnicas de selección las variables se van agregando o removiendo con base en el **valor-p** y un punto de corte de entrada o de permanencia.

- Ventajas y desventajas de las regresiones

Ventajas:

- Es un algoritmo fácil de entender y de explicar.
- Es menos propenso al sobreajuste que los árboles.
- Es fácil de implementar.

Desventajas:

- No puede trabajar con valores ausentes
- Las regresiones hacen buenos pronósticos con casos cerca de los centros de la distribución de cada **variable de entrada**, por lo que no es apropiado utilizar una regresión cuando existen valores de entrada extremos o atípicos.
- Tienen supuesto estrictos

3.1.3. Red Neuronal

La red neuronal es un algoritmo que tiene una fórmula de pronóstico similar al de una regresión, sin embargo, la red neuronal añade un componente extra que la convierte en un algoritmo más flexible. Esta adición permite que una red neuronal entrenada adecuadamente modele prácticamente cualquier asociación entre las variables de entrada y la variable objetivo. La flexibilidad tiene un costo ya que la red neuronal es incapaz de seleccionar variables. La anterior deficiencia se compensa con un método de optimización de la complejidad de la red llamado **stopped training**.

- **Construcción de la red neuronal**

Las redes neuronales al igual que las regresiones pronostican a las observaciones utilizando una ecuación matemática que utiliza los valores asociados a las variables de entrada, como se muestra en las siguientes ecuaciones:

$$\hat{y} = \widehat{W}_{00} + \widehat{W}_{01}H_1 + \widehat{W}_{02}H_2 + \widehat{W}_{03}H_3$$

$$H_1 = \tanh\left(\widehat{W}_{10} + \widehat{W}_{11}X_1 + \widehat{W}_{12}X_2 + \cdots + \widehat{W}_{1n}X_n\right)$$

$$H_2 = \tanh\left(\widehat{W}_{20} + \widehat{W}_{21}X_1 + \widehat{W}_{22}X_2 + \cdots + \widehat{W}_{2n}X_n\right)$$

$$H_3 = \tanh\left(\widehat{W}_{30} + \widehat{W}_{31}X_1 + \widehat{W}_{32}X_2 + \cdots + \widehat{W}_{3n}X_n\right)$$

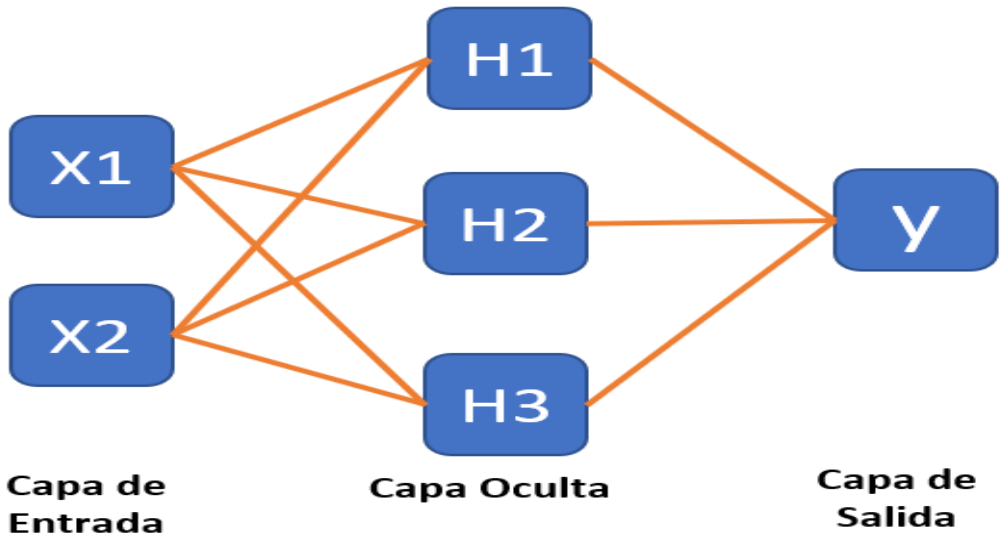
La red neuronal se puede pensar como una regresión con variables de entradas derivadas llamadas unidades ocultas. A su vez las unidades ocultas pueden considerarse regresiones de las variables de entrada originales. Estas pseudo- regresiones incorporan una función de enlace denominada **función de**

activación, la cual flexibiliza al algoritmo ya que le ayuda a modelar a la variable objetivo de forma no lineal.

Por otra parte, la red neuronal es un modelo con raíces biológicas, los nombres de sus componentes difieren a los términos empleados en la regresión, más aún, varios componentes poseen nombres biológicos. Por ejemplo, a las unidades ocultas también se les denomina neuronas. La red neuronal no tiene un intercepto, tiene un **sesgo** (bias). Asimismo, la red neuronal no cuenta con parámetros estimados posee **pesos estimados** (weight estimates).

Las redes neuronales usualmente suelen representarse en diagramas en lugar de ecuaciones, cada una de estas componentes se plasman en capas. La primera capa es la **capa de entrada**, esta conecta a una capa de neuronas llamada **capa oculta**, la cual a su vez se conecta a la capa final llamada **capa objetivo o capa de salida.** A continuación, se muestra un diagrama que representa a una red neuronal con dos variables de entrada y tres unidades ocultas.

Figura 31: Diagrama de red neuronal



La mayor diferencia entre las redes neuronales y las regresiones se ve claramente en el diagrama anterior, y es el mayor número de parámetros a estimar y la compleja relación que hay entre los pesos de las capas ocultas y la variable objetivo.

- **Stopped Training**

Este método ocurre tras la estimación de los pesos de la red. Es necesario mencionar que la estimación de los parámetros es por medio de un proceso de optimización. El proceso de optimización inicia dando valores iniciales a los pesos, luego, el entrenamiento procede a actualizar las estimaciones de modo que se incremente la función de verosimilitud.

El **stopped training** consiste en considerar cada iteración como un modelo diferente. La iteración con el mejor valor en el estadístico de ajuste será considerado el modelo final. El nombre de s**topped training** deriva del hecho de que el modelo final se selecciona como si el entrenamiento de la red se detuviera en la iteración óptima.

- **Ventajas y desventajas de las redes neuronales**

VENTAJAS

- Posee aprendizaje adaptativo, es decir, tiene la capacidad de aprender a realizar tareas con base en los datos iniciales y un periodo de aprendizaje.
- Permite modelar cualquier asociación entre las variables de entrada y la variable objetivo.

DESVENTAJAS

- Difícil de explicar.
- No posee un método de selección de variables.
- Presenta complejidad para estimar los parámetros.

3.2. Econometría

La econometría puede definirse como el análisis cuantitativo de la economía, es decir, esta área se enfoca en analizar sistemáticamente la teoría económica por medio de datos observados. La econometría recopila conocimiento de otras ciencias como lo son: economía, matemáticas y estadística.

En particular, existen modelos económicos que consideran los valores rezagados de los regresores que participan en el modelo, a esta clase de modelos se les denomina **modelos dinámicos o de rezagos distribuidos**. El incorporar valores rezagados ayuda a medir el impacto que una variable provoca en otra por varios periodos de tiempo.

3.2.1. Modelo de Almon

Matemáticamente la relación entre la variable dependiente e independiente en este modelo con rezagos se define de la siguiente manera:

$$Y_t = \alpha + \beta_0 X_t + \cdots + \beta_s X_{t-s} + u_t, \qquad t = 1, 2, \dots, T$$

Donde X_{t-s} denota la (t-s) observación de la variable X, en otras palabras, el modelo de Almon considera que los regresores son la variable X y sus s-valores rezagados. Se debe agregar que este modelo distribuye el efecto de la variable X en Y a través de s- períodos, este efecto se divide en dos, el **efecto a corto plazo** dado por β_0 y el **efecto a largo plazo** que es $(\beta_0 + \beta_1 + \cdots + \beta_s)$.

El modelo de Almon impone una estructura a los parámetros β a estimar, a estos los define como $\beta_i = f(i)$ para i = 0, 1, …, s donde $f(i)$ es una función continua en un intervalo cerrado. Por teoría de cálculo, la función puede ser aproximada por un polinomio de grado r.

$$f(i) = a_0 + a_1 i + \cdots + a_r i^r$$

Por ejemplo, con las anteriores restricciones y la suposición de un polinomio de grado r= 2 los parámetros toman la siguiente forma:

$$\beta_0 = a_0$$

$$\beta_1 = a_0 + a_1 + a_2$$

$$\beta_2 = a_0 + 2a_1 + 4a_2$$

$$\vdots \quad \vdots \quad \vdots$$

$$\beta_s = a_0 + sa_1 + s^r a_2$$

Tras hacer la anterior transformación y sustituir en la ecuación original, el modelo a estimar ahora es el siguiente:

$$Y_t = \alpha + \sum_{i=0}^{s} (a_0 + a_1 i + a_2 i^2) X_{t-i} + u_t$$

$$= \alpha + a_0 \sum_{i=0}^{s} X_{t-i} + a_1 \sum_{i=0}^{s} i X_{t-i} + a_2 \sum_{i=0}^{s} i^2 X_{t-i} + u_t$$

Haciendo un cambio de variable se puede definir $Z_0 = \sum_{i=0}^{s} X_{t-i}$, $Z_1 = \sum_{i=0}^{s} i X_{t-i}$ y $Z_2 = \sum_{i=0}^{s} i^2 X_{t-i}$, y de este modo se reduce la complejidad del modelo ya que en lugar de estimar s- parámetros solo será necesario estimar tres. El ajuste con la transformación de Almon queda como se muestra en la siguiente ecuación:

$$Y_t = \alpha + a_0 Z_0 + a_1 Z_1 + a_2 Z_2 + u_t$$

- **Ventajas y desventajas del modelo de Almon**

VENTAJAS

- Considera la estructura de dependencia en el tiempo de las variables explicativas.
- Los estimadores obtenidos son BLUE (Best Linear Unbiased Estimator).
- Con la estructura del rezago se elimina la multicolinealidad entre los regresores.
- Fácil de estimar y de interpretar.

DESVENTAJAS

- Se pierden observaciones con cada valor rezagado que se incorpora al modelo.
- El imponer una estructura a los coeficientes es un supuesto restrictivo.
- Complejidad para determinar el valor del rezago y del polinomio a emplear.

3.3. Medidas de Ajuste para la Comparación de Modelos

En esta sección se listan los criterios de ajuste que se utilizan para medir el rendimiento y hacer la comparación entre los modelos anteriormente descritos.

3.3.1 Dstat

La principal medida que se empleó para la comparación de los modelos es el Dstat, esta métrica se introduce en el artículo de los científicos Vladamani Ravi, Dadabada Pradeepkumar y Kalyanmoy Deb[18]. Un punto importante al pronosticar series de tiempo financieras es no solo medir la precisión de las estimaciones, sino enfocarse en la dirección del cambio que tuvo la serie financiera.

El Dstat mide el porcentaje de asertividad del modelo para pronosticar la tendencia del índice, y se define de la siguiente manera:

$$Dstat = \frac{1}{N}\sum_{i=1}^{N} a_i * 100\%$$

$$Donde\ a_t = \left\{ \begin{array}{ll} 1, & si\ (y_{t+1} - y_t) * (\hat{y}_{t+1} - \hat{y}_t) \geq 0 \\ 0, & en\ otro\ caso \end{array} \right\}$$

Esta medida acumula cada vez que el valor pronosticado y el estimado coinciden en signo, es decir, cuando el pronostico sigue la dirección que lleva el índice.

3.3.2. Correspondencia entre Valores Observados y Estimados

Además, de conocer el porcentaje de asertividad en la tendencia por medio del Dstat se construyeron tablas de contingencia para analizar a detalle el signo que tienen los aciertos que pronostica el modelo, esto para averiguar si existe un sesgo en el pronóstico de casos al alza o en los casos a la baja. En concreto se construyó una variable que identifica la coincidencia en signo, como se muestra a continuación:

$$mismo\_signo = \left\{ \begin{array}{ll} 1, & si\ y_t \geq 0\ and\ \hat{y}_t \geq 0 \\ 2, & si\ y_t < 0\ and\ \hat{y}_t < 0 \\ 0, & en\ otro\ caso \end{array} \right\}$$

Al analizar la anterior variable se puede conocer si el modelo generaliza de mejor forma las condiciones que conllevan a un mercado a la alza o a la baja.

3.3.3. Error Cuadrático Medio

Este estadístico mide el promedio de los errores al cuadrado, en otras palabras, se fija en la diferencia entre el valor observado y el valor pronosticado de la variable objetivo. El Error Cuadrático Medio (ASE, por sus siglas en inglés) se calcula de la siguiente manera:

$$ASE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2, donde\ n\ es\ el\ número\ de\ observaciones.$$

La selección del mejor modelo bajo este estadístico se realiza por medio de aquel que tenga el menor valor de ASE.

[18] V. Ravi, D. Pradeepkumar and K. Deb, (2017), Financial time series prediction using hybrids of chaos theory, multi-layer perceptron and multi-objective evolutionary algorithms. Swarm and Evolutionary Computation

Capítulo 4: Implementación de modelos

En este capítulo se describirán los principales resultados obtenidos al implementar los diferentes enfoques de modelación que se usaron en la búsqueda de explicar el comportamiento del Índice Dow Jones Industrial Average (DJIA). Es necesario mencionar que los resultados mostrados en este capítulo se estructuran con base en la **Figura 26: Técnicas de modelación empleadas** del capítulo 3, es decir, inicialmente se mostrará la descripción de aplicar el enfoque de minería de datos y posteriormente se listará la perspectiva de la econometría.

4.1. Pre-Implementación de Modelo de Minería de Datos

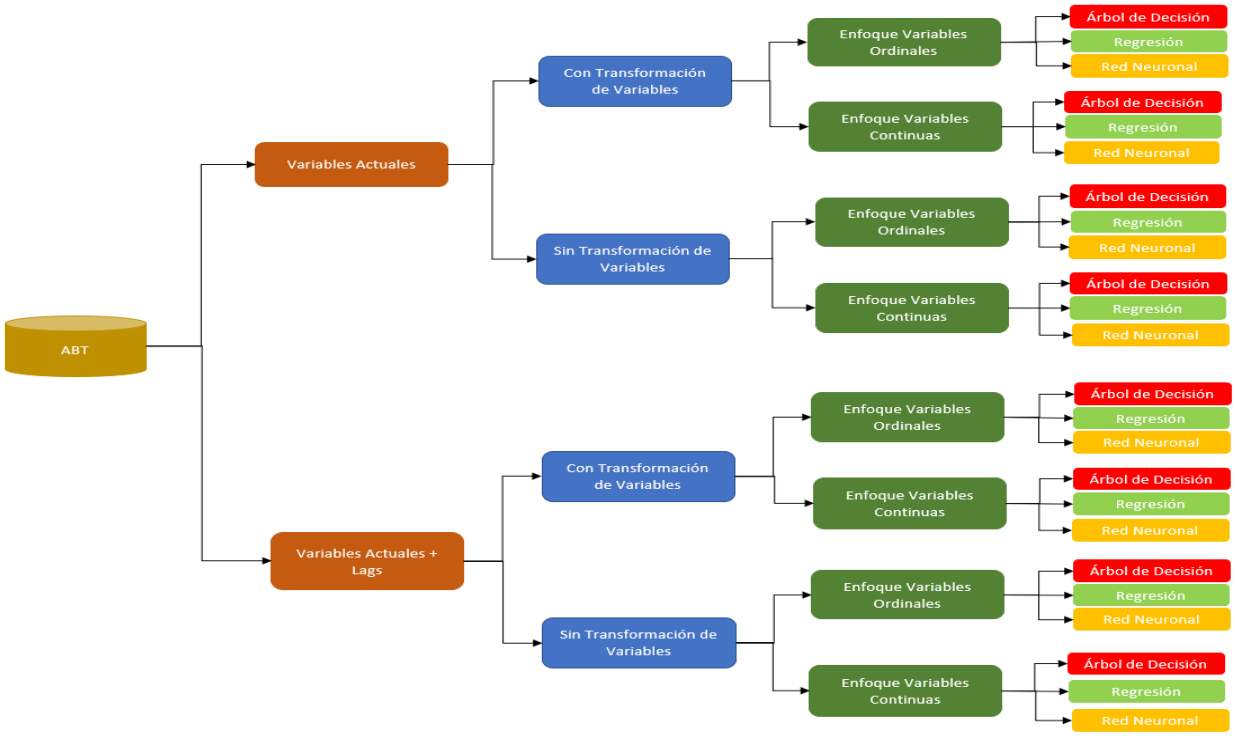
Como se menciona en el capítulo 3 se utilizaron tres técnicas de minería de datos que son: **Árbol de Decisión**, **Regresión** y **Red Neuronal.** Cabe señalar que se aplicaron diferentes enfoques al momento de modelar, con el objetivo de ayudar a los algoritmos a tener una perspectiva diferente de las variables de entrada.

Las técnicas auxiliares a la modelación empleadas son:

- Incorporar valores rezagados (lags) de las variables.
- Aplicar transformaciones a las variables de entrada.
- Analizar la información de variables desde una perspectiva continua y ordinal.

En resumen, la realización de los modelos de minería de datos añadiendo las anteriores técnicas auxiliares se muestra en el siguiente diagrama.

Figura 32: Panorama de modelos de Minería de Datos.



4.1.1. Descripción de Técnicas Auxiliares de Modelación.

En esta sección se dará una descripción general de las técnicas auxiliares que se emplearon durante el proceso de construcción del modelo.

- Variables Rezagadas (lags)

Las variables rezagadas son aquellas cuyos valores pertenecen a un punto anterior en el tiempo.

Tabla 20: Ejemplo de variables rezagadas

Mes	X	Lag1_X	Lag2_X
Enero	1	.	.
Febrero	2	1	.
Marzo	3	2	1
Abril	4	3	2
Mayo	5	4	3
Junio	6	5	4

- Transformación de Variables

Se aplicaron transformaciones a las variables antes de ingresarlas al modelo, las transformaciones se aplicaron desde dos perspectivas.

La primera consistió en aplicar la transformación Box Cox a las variables de tipo continuo y en colapsar los niveles de las variables categóricas con ayuda de un árbol de decisión. Para más detalles del proceso técnico revisar **Anexo 1: Transformación Box Cox** y **Anexo 2: Colapsar Variables Categóricas con Árbol de Decisión**.

Por otro lado, la segunda estrategia consistió en transformar todas las variables de entrada con ayuda del nodo de transformación de variables de la herramienta SAS Enterprise Miner. Este nodo permite crear nuevas variables a partir de una transformación de las existentes. Las transformaciones son útiles cuando se pretende mejorar el ajuste de un modelo a los datos, por ejemplo, las transformaciones ayudan a estabilizar varianzas, a remover comportamientos no lineales, corregir normalidad, etc.

En particular, este nodo tiene la opción de elegir la mejor transformación para cada variable, esto lo consigue al realizar diferentes transformaciones y finalmente utilizar aquella que consiga el **mejor test** $\mathbf{X^2}$ con la **variable objetivo**. En el **Anexo 3: Nodo de Transformación de Variables** se puede conocer mayor detalle del funcionamiento de este nodo.

- Perspectiva Continua y Ordinal de Variables

La variable fuentes_disti_FBTS y las 10 variables asociadas a los clústeres de topics (Cluster_Topics1 - Cluster_Topics10) contienen una serie de valores enteros ordenados, donde mayor número representa una mayor ocurrencia del concepto que miden, por esto, estas variables se pueden asociar a los niveles de medida **Ordinal** o **Continuo**.

Como se ha dicho en el capítulo 3, los modelos llegan a utilizar diferentes definiciones para operar a las variables continuas y categóricas, es por esto que se decidió proporcionar a los modelos la información de estas variables desde ambos enfoques. Con lo anterior, se desea observar los resultados y con esto elegir la perspectiva que beneficia más al ajuste.

En resumen, las estrategias auxiliares que fueron aplicadas a cada una de las variables que conforman la ABT se enlistan en la siguiente tabla[19].

Tabla 21: Aplicación de técnicas auxiliares a variables

Variable	Se aplica Lag	Se transforma la variable	Se analiza desde enfoque ordinal y continuo
date_completa			
freq_noti_FBTS	Sí	Colapsar Niveles	
fuentes_disti_FBTS	Sí		Sí
likes_FBTS	Sí	Colapsar Niveles	
comments_FBTS	Sí	Colapsar Niveles	
shares_FBTS	Sí	Colapsar Niveles	
daily_neg_index	Sí	Box Cox	
Cluster_Topics1	Sí	Colapsar Niveles	Sí
Cluster_Topics2	Sí	Colapsar Niveles	Sí
Cluster_Topics3	Sí	Colapsar Niveles	Sí
Cluster_Topics4	Sí	Colapsar Niveles	Sí
Cluster_Topics5	Sí	Colapsar Niveles	Sí
Cluster_Topics6	Sí	Colapsar Niveles	Sí
Cluster_Topics7	Sí	Colapsar Niveles	Sí
Cluster_Topics8	Sí	Colapsar Niveles	Sí
Cluster_Topics9	Sí	Colapsar Niveles	Sí

[19] Todas las variables de entrada pasaron por el nodo de transformación de variables

Variable	Se aplica Lag	Se transforma la variable	Se analiza desde enfoque ordinal y continuo
Cluster_Topics10	Sí	Colapsar Niveles	Sí
GIS_Cluster1_GIS	Sí	Box Cox	
GIS_Cluster2_GIS	Sí	Box Cox	
GIS_Cluster3_GIS	Sí	Box Cox	
GIS_Cluster4_GIS	Sí	Box Cox	
GIS_Cluster5_GIS	Sí	Box Cox	
log_return	Sí		
close_correg_VIX	Sí		

4.2. Implementación de Modelo de Minería de Datos

Para iniciar la construcción de modelos es importante realizar las transformaciones externas de variables del primer enfoque descrito en la sección de **Transformación de Variables** de este capítulo. Lo que se hizo primero fue aplicar la transformación de la familia Box Cox con el parámetro óptimo de potencia lambda (λ), en el siguiente cuadro se muestra la lambda utilizada para transformar cada variable.

Tabla 22: Lambda óptima para transformación Box Cox

Variable	Lambda
daily_neg_index	0
GIS_Cluster1_GIS	0.75
GIS_Cluster2_GIS	0.5
GIS_Cluster3_GIS	0.75
GIS_Cluster4_GIS	-0.25
GIS_Cluster5_GIS	0.75

Asimismo, siguiendo este enfoque se colapsaron los niveles de las variables categóricas con ayuda de un árbol de decisión, en el siguiente cuadro se muestra el número de niveles y el rango de valores que se consolidaron para cada una de las variables. Se puede notar que cada variable se colapsó en a los más seis niveles.

Tabla 23: Rangos para colapsar variables

Variable Original	Intervalos para colapsar	Nueva variable
freq_noti_FBTS	(-∞, 7) [7, ∞)	tr_freq_noti_FBTS
likes_FBTS	(-∞, 1955) [1955, 6764) [6764, ∞)	tr_likes_FBTS
comments_FBTS	(-∞, 146) [146, 839) [839, ∞)	tr_comments_FBTS
shares_FBTS	(-∞, 388) [388, 670) [670, 985) [985, 1519) [1519, 4245) [4245, ∞)	tr_shares_FBTS
Cluster_Topics1	(-∞, 5) [5, 8) [8, ∞)	tr_Cluster_Topics1
Cluster_Topics2	(-∞, 1) [1, ∞)	tr_Cluster_Topics2
Cluster_Topics3	(-∞, 4) [4, ∞)	tr_Cluster_Topics3
Cluster_Topics4	(-∞, 2) [2, ∞)	tr_Cluster_Topics4
Cluster_Topics5	(-∞, 3) [3, ∞)	tr_Cluster_Topics5
Cluster_Topics6	(-∞, 1) [1, ∞)	tr_Cluster_Topics6
Cluster_Topics7	(-∞, 1) [1, ∞)	tr_Cluster_Topics7
Cluster_Topics8	(-∞, 2) [2,6) [6, ∞)	tr_Cluster_Topics8

Variable Original	Intervalos para colapsar	Nueva variable
Cluster_Topics9	(-∞, 1) [1, ∞)	tr_Cluster_Topics9
Cluster_Topics10	(-∞, 1) [1, ∞)	tr_Cluster_Topics10

Con las anteriores dos transformaciones se conformó una nueva ABT que contiene las versiones transformadas de las variables explicativas, utilizando la ABT original (ver **Tabla 10: Variables que conforman la ABT**), con esta segunda versión se inició la construcción de modelos.

La construcción de los modelos de minería de datos se realizó en la herramienta SAS Enterprise Miner, para la realización de este trabajo se construyeron 21 modelos de minería de datos siguiendo la lógica que se muestra en la **Figura 32: Panorama de modelos de Minería de Datos.** En las siguientes dos imágenes se muestran los diagramas de construcción de estos modelos, como se puede observar, se trabajó de manera separada los modelos que contienen rezagos de los que no.

Figura 33: Implementación de modelos con rezagos.

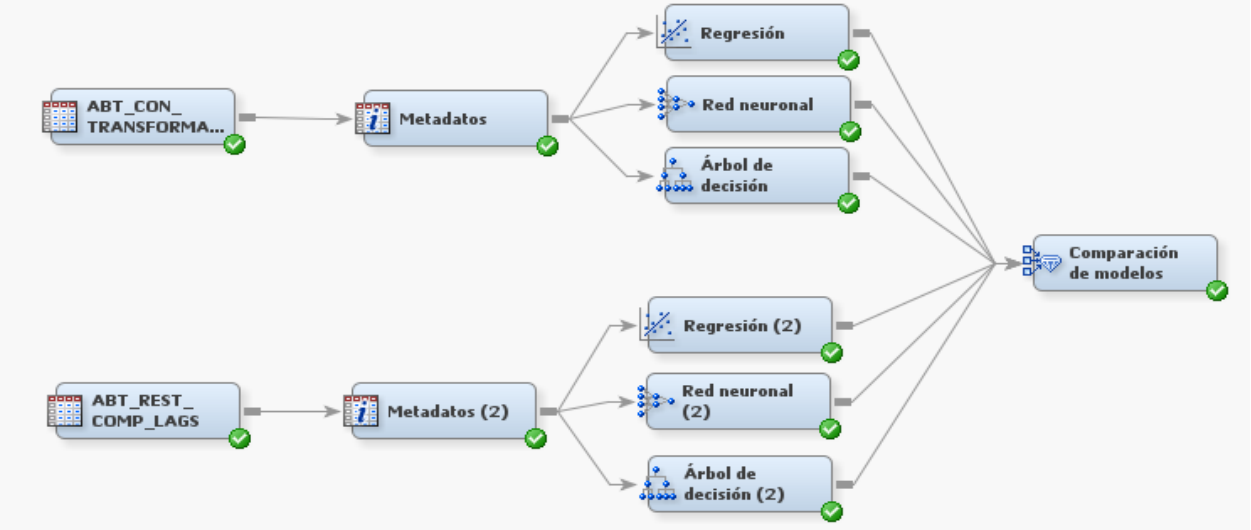
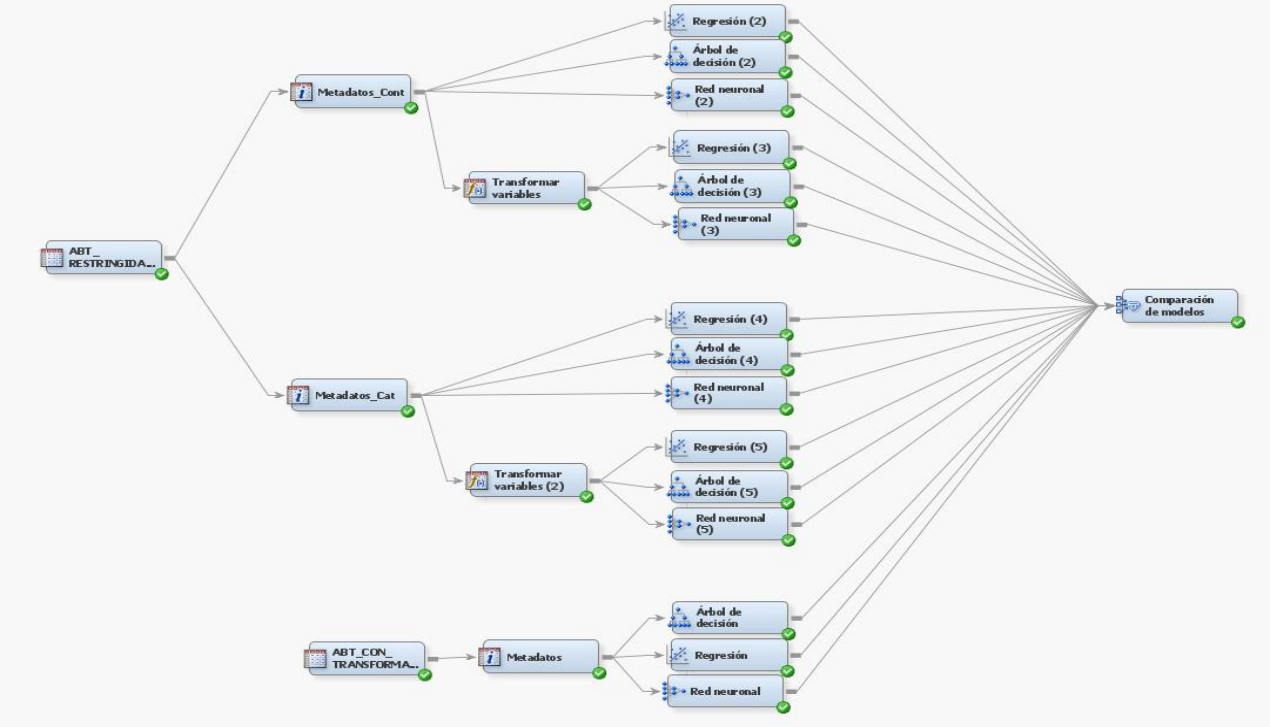


Figura 34: Implementación de modelos sin rezagos



En los siguientes párrafos se hablará de los resultados de los modelos, por practicidad se muestran únicamente los resultados de los tres mejores modelos con y sin rezagos, con base en el criterio del Dstat. A continuación, se describen las características generales de los mejores modelos.

M O D E L O S S I N R E Z A G O S

1. Árbol con enfoque categórico y variables transformadas por nodo Miner.
2. Árbol con enfoque continuo y variables transformadas por nodo Miner.
3. Regresión con enfoque continuo y variables transformadas por nodo Miner.

M O D E L O S C O N R E Z A G O S

1. Regresión con variables originales y rezagos
2. Árbol con variables originales y rezagos
3. Regresión con variables transformadas manualmente y rezagos.

Como se puede observar, los árboles y las regresiones son las técnicas que otorgan el mejor ajuste bajo el criterio del Dstat.

Por otro lado, la **Tabla 24** resume las variables que participan en cada modelo. Se observa que la única variable que forma parte de los seis modelos es el índice de volatilidad (VIX), lo cual se asocia al riesgo inherente del mercado. Dentro de las variables de topics la más popular en los modelos es Cluster_Topics4, la cual está asociada a términos del mercado global y el trabajo, en segundo lugar, de participación se encuentra el Cluster_Topics2 que habla de asuntos presidenciales y política exterior.

También se observa participación de los índices de búsqueda en Google de los conjuntos de términos asociados a las variables GIS_Cluster4_GIS, GIS_Cluster5_GIS y GIS_Cluster2_GIS los cuales hablan de acciones, mercado a la baja y mercado al alta respectivamente.

Tabla 24: Selección de Variables por Modelo

	MODELOS SIN REZAGOS			MODELOS CON REZAGOS		
	1. Árbol con enfoque categórico y variables transformadas por nodo miner.	2. Árbol con enfoque continuo y variables transformadas por nodo miner.	3. Regresión con enfoque continuo y variables transformadas por nodo miner.	1. Regresión con variables originales y rezagos	2. Árbol con variables originales y rezagos	3. Regresión con variables transformadas manualmente y rezagos.
close_correg_VIX	Optimal Binning	Optimal Binning	Optimal Binning	Actual, Lag1, Lag2	Actual, Lag1, Lag2, Lag7	Actual
Cluster_Topics1	x	exp(Cluster_Topics4)				
Cluster_Topics10						Tree Lag7
Cluster_Topics2				Lag2, Lag6	Lag1	Tree Lag6
Cluster_Topics3						Tree Lag3
Cluster_Topics4	x	exp(Cluster_Topics2)	exp(Cluster_Topics2)	Lag2, Lag3, Lag4		Actual
Cluster_Topics5				Lag7		
Cluster_Topics7				Lag5, Lag6		Tree Lag6
Cluster_Topics8					Lag4, Lag6	
comments_FBTS					Lag2	
daily_neg_index	1 / (daily_neg_index + 1)	1 / (daily_neg_index + 1)			Lag3	
freq_noti_FBTS	exp(freq_noti_FBTS)	exp(freq_noti_FBTS)				
GIS_Cluster1_GIS						BoxCox Actual, Lag1, Lag4, Lag7
GIS_Cluster2_GIS			Optimal Binning			
GIS_Cluster3_GIS						
GIS_Cluster4_GIS	Optimal Binning	Optimal Binning	Optimal Binning			
GIS_Cluster5_GIS				Lag2		
likes_FBTS						Tree Lag4
log_return				Lag1, Lag2, Lag5		

Con respecto a los estadísticos de ajuste conseguidos por cada uno de los modelos, estos se resumen en la **Tabla 25**. Claramente los modelos que incorporan rezagos brindan un mayor ajuste que los que no los incorporan.

Tabla 25: Cuadro de comparación de modelos de Minería de Datos

Modelo	Error Cuadrático Medio	Raíz del Error Cuadrático Medio	Dstat	Valores Positivos Acertados	Valores Negativos Acertados
MODELOS SIN REZAGOS					
1. Árbol con enfoque categórico y variables transformadas por nodo miner.	0.00004055	0.006368	60.45419	44.56	15.9
2. Árbol con enfoque continuo y variables transformadas por nodo miner.	0.00004075	0.006383	59.6132	36.52	23.09
3. Regresión con enfoque continuo y variables	0.0000414	0.006435	54.15245	54.15	0

Modelo	Error Cuadrático Medio	Raíz del Error Cuadrático Medio	Dstat	Valores Positivos Acertados	Valores Negativos Acertados
transformadas por nodo miner.					
MODELOS CON REZAGOS					
1. Regresión con variables Originales y retrasos	0.00001195	0.003456	81.7975	46.08	35.72
2. Árbol con variables originales y retrasos	0.00003702	0.006085	63.70876	41.41	22.3
3. Regresión con variables transformadas manualmente y retrasos.	0.00004021	0.006341	62.9124	40.96	21.96

Dentro de los modelos sin rezagos destaca que a pesar de que los errores cuadráticos medios tengan una magnitud similar la tendencia que aciertan varía considerablemente, por ejemplo, entre los modelos 1 y 2 existe solo una unidad de diferencia en el Dstat, sin embargo, el modelo 1 está sesgado a generalizar de mejor forma las observaciones con subidas en el DJIA. Por otro lado, el modelo 3 es completamente inútil ya que no brinda información del comportamiento de los días malos para el índice. Por todo esto, es importante observar más de una métrica para comparar a los modelos.

Por otro lado, en los modelos con rezagos existe un claro ganador que es el modelo 1, este brinda el mayor Dstat mostrando una mejoría de casi 20% con respecto a los otros dos modelos. Además, es necesario resaltar que el modelo 1 es equilibrado en cuanto a generalizar el comportamiento de los días buenos y malos para el índice Dow Jones, esto se concluye tras observar las proporciones de tendencias positivas y negativas que logra acertar.

En conclusión, el mejor modelo del enfoque de minería de datos para explicar el comportamiento del índice Dow Jones Industrial Average es el modelo 1 con rezagos, es decir, es una regresión construida con las variables originales de la ABT y sus valores rezagados.

4.3. Implementación de Modelo de Econometría

Los modelos econométricos se implementaron con ayuda del procedimiento PROC PDLREG[20], este ajusta modelos dinámicos desde el enfoque de Almon. Cómo se mencionaba en la sección 3.2.1. del capítulo 3 el modelo necesita de dos parámetros para ser ajustado: el **número de rezago** y el **grado del polinomio**.

[20] SAS® 9.4 and SAS® Viya® 3.4 Programming [sitio de internet]. SAS Institute Inc. [consultado 20 de noviembre de 2020]. Disponible en: https://documentation.sas.com/?cdcId=pgmsascdc&cdcVersion=9.4_3.4&docsetId=etsug&docsetTarget=etsug_pdlreg_overview.htm&locale=es

Se implementaron modelos utilizando las 22 variables de entrada de la ABT utilizando un número de rezago máximo de 7, se debe agregar que para cada nivel de rezago se probaron todos los grados de polinomio que no excedieran al número de rezago. Por consiguiente, se implementaron 28 modelos por variable sumando un total de 616 modelos econométricos Almon.

A continuación, se muestran los resultados de los seis mejores modelos econométricos obtenidos. Se debe resaltar que estos modelos comparten que son ajustes de la misma variable explicativa, el **índice de volatilidad VIX**. Asimismo, se observa que los mejores seis modelos utilizan los números de rezagos del dos al siete, y tienen este mismo valor como grado de polinomio, se observa que los modelos con dos y tres rezagos son los que muestran ligeramente mejores estadísticos de ajuste.

Tabla 26: Cuadro de comparación de modelos de Econometría

Variable	Grado Polinomio	Lag	Average Squared Error	Root Average Squared Error	Dstat	Valores Positivos Acertados	Valores Negativos Acertados
close_correg_VIX	2	2	0.0000146360	0.003825707	80.57948	57.32	23.26
close_correg_VIX	3	3	0.0000146270	0.00382453	80.34456	56.77	23.57
close_correg_VIX	5	5	0.0000145871	0.003819302	79.71809	56.07	23.65
close_correg_VIX	4	4	0.0000146437	0.003826709	79.71809	56.07	23.65
close_correg_VIX	7	7	0.0000145769	0.003817965	79.32655	55.76	23.57
close_correg_VIX	6	6	0.0000145730	0.003817465	79.09162	55.68	23.41

Por otra parte, también se analizó la significancia de los coeficientes de cada modelo como punto de desempate para seleccionar al mejor modelo econométrico. En la siguiente tabla se muestran los p-value de la prueba t asociados a los parámetros de cada modelo. Se observa que en todos los modelos los parámetros asociados al ajuste del modelo cumplen con la prueba t, sin embargo, únicamente el modelo con dos rezagos cumple también con la significancia de sus parámetros del polinomio.

Por lo anterior, se establece que el mejor modelo econométrico es el de modelo con dos rezagos y polinomio de grado dos.

Tabla 27: Comparación de la significancia por variable de los modelos econométricos

Modelo	Modelo Lag 7 - Polinomio 7	Modelo Lag 6 - Polinomio 6	Modelo Lag 5 - Polinomio 5	Modelo Lag 4 - Polinomio 4	Modelo Lag 3 - Polinomio 3	Modelo Lag 2 - Polinomio 2
Término	**Pr > |t|**	**Pr > |t|**	**Pr > |t|**	**Pr > |t|**	**Pr > |t|**	**Pr > |t|**
PARÁMETROS DEL MODELO						
Intercepto	0.0031	0.0029	0.001	0.0002	0.0002	0.0005
close_correg_VIX**0	0.0112	0.0102	0.004	0.0008	0.0011	0.002
close_correg_VIX**1	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
close_correg_VIX**2	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
close_correg_VIX**3	<.0001	<.0001	<.0001	<.0001	<.0001	
close_correg_VIX**4	<.0001	<.0001	<.0001	<.0001		
close_correg_VIX**5	<.0001	<.0001	<.0001			
close_correg_VIX**6	<.0001	<.0001				
close_correg_VIX**7	0.001					
PARÁMETROS DEL POLINOMIO						
close_correg_VIX(0)	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
close_correg_VIX(1)	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
close_correg_VIX(2)	0.0009	0.0009	0.0015	0.0008	0.0008	0.0013
close_correg_VIX(3)	0.4908	0.504	0.5831	0.6065	0.1611	
close_correg_VIX(4)	0.0197	0.0188	0.0206	0.4567		
close_correg_VIX(5)	0.4887	0.446	0.0087			
close_correg_VIX(6)	0.1657	0.077				
close_correg_VIX(7)	0.9943					

4.4. Comparación de Modelos Campeones de Ambos Enfoques

En esta sección se comparan los modelos ganadores de los enfoques de Minería de Datos y Econometría, con el objetivo de encontrar al mejor modelo que explique el comportamiento del Dow Jones Industrial Average (DJIA).

Recapitulando el modelo seleccionado en el enfoque de minería de datos es una regresión linear con las siguientes variables:

Intercept Cluster_Topics2_2 Cluster_Topics2_6 Cluster_Topics4_2 Cluster_Topics4_3 Cluster_Topics4_4 Cluster_Topics5_7 Cluster_Topics7_5 Cluster_Topics7_6 GIS_Cluster5_GIS_2 close_correg_VIX close_correg_VIX_1 close_correg_VIX_2 log_return_1

NOTA: Las variables con un guion bajo y un número al final indican el número de rezago.

Por otro lado, el modelo escogido como el representante de los modelos econométricos es una regresión con ajuste de Almon que utiliza la variable **close_correg_VIX** con dos rezagos y un polinomio de grado dos.

Tabla 28: Comparación de estadísticos de mejores modelos con enfoque de Minería de Datos y Econometría.

Estadístico de Ajuste	Regresión linear con variables originales y rezagos	Modelo de Almon con dos rezagos y un polinomio de grado dos.
Error Cuadrático Medio	0.00001195	0.000014636
Raíz del Error Cuadrático Medio	0.003456	0.00382571
Dstat	81.7975	80.57948
Valores Positivos Acertados	46.08	57.32
Valores Negativos Acertados	35.72	23.26

Con la intención de encontrar un modelo campeón que explique el comportamiento del índice Dow Jones se utilizó un criterio extra para comparar a los modelos de ambos enfoques. Teniendo en cuenta que ambos modelos son regresiones se decidió validar los supuestos de cada modelo como criterio de desempate.

En resumen, los supuestos y las pruebas que se emplearon se detallan a continuación:

- Media Cero: Prueba T
- Normalidad: Prueba Anderson Darling
- Homocedasticidad: Prueba Brush-Pagan
- Independencia: Prueba Durbin Watson

En el siguiente cuadro se consolidan los resultados de los supuestos que cumple cada uno de los modelos. Se puede observar que el modelo de minería de datos cumple con **Media Cero** e **Independencia** y por otro lado el modelo econométrico únicamente cumple con tener **Media Cero**, al cumplir un supuesto más se designa como **modelo campeón de este trabajo** al modelo de Minería de datos la **Regresión linear con variables originales y rezagos.**

Cada uno de los supuestos proporciona características a los parámetros estimados, usualmente cuando uno de estos no se cumple se aplican técnicas para remediar la violación de cada supuesto. Sin embargo, el alcance de este trabajo no considera la corrección de estos por lo que se dejará la regresión tal como fue implementada.

Tabla 29: Validación de supuestos de mejores modelos con enfoque de Minería de Datos y Econometría.

Supuesto	Regresión linear con variables originales y rezagos	Modelo de Almon con dos rezagos y un polinomio de grado dos.
Media cero	Sí	sí
Normalidad	No	no
Homocedasticidad	No	no
Independencia	Sí	no

Capítulo 5: Modelo Ganador

El modelo ganador es una regresión con variables originales y rezagos, se debe agregar que, las variables con valores enteros se incorporaron desde un enfoque categórico ordinal. Como las variables categóricas se incorporan por niveles, existe una prueba t para cada coeficiente asociado a un valor distinto. En la siguiente tabla se muestra un resumen de la significancia de los niveles de cada variable.

En la **Tabla 30**, se puede observar que existen variables con múltiples niveles no significativos.

Tabla 30: Significancia de los coeficientes del modelo campeón.

Variable	Tipo de Variable	Toda la Variable es Significativa	Número de Niveles	Niveles No Significativos
Intercept	Intercepto	Sí		
Cluster_Topics2_2	Categórica	No	20	11
Cluster_Topics2_6	Categórica	No	20	11
Cluster_Topics4_2	Categórica	No	7	2
Cluster_Topics4_3	Categórica	No	7	5
Cluster_Topics4_4	Categórica	No	7	4
Cluster_Topics5_7	Categórica	No	9	7
Cluster_Topics7_5	Categórica	No	9	6
Cluster_Topics7_6	Categórica	No	9	6
GIS_Cluster5_GIS_2	Intervalo	Sí		
close_correg_VIX	Intervalo	Sí		
close_correg_VIX_1	Intervalo	Sí		
close_correg_VIX_2	Intervalo	Sí		
log_return_1	Intervalo	Sí		
log_return_2	Intervalo	Sí		
log_return_5	Intervalo	Sí		

5.1. Refinamiento del modelo Ganador

Fundamentalmente, los modelos buscan incorporar información estadísticamente significativa y que aporte valor al fenómeno a explicar. Al no cumplir con esta característica se buscó una manera alternativa de incorporar la información de estas variables al modelo, esto se consiguió tras colapsar los niveles de las variables de entrada, siguiendo los rangos que se muestran en la **Tabla 31**. Cabe mencionar, que los rangos se definieron buscando agrupar a los niveles no significativos.

Tabla 31: Reglas para afinar variables del modelo ganador

Variable	Regla para Variable
Cluster_Topics2_2	0-6, 7+
Cluster_Topics2_6	0-6, 7+
Cluster_Topics4_2	0-3, 4+
Cluster_Topics4_3	0-3, 4+
Cluster_Topics4_4	0-3, 4+
Cluster_Topics5_7	0-3, 4+
Cluster_Topics7_5	0-2, 3+
Cluster_Topics7_6	0-2, 3+
GIS_Cluster5_GIS_2	
close_correg_VIX	
close_correg_VIX_1	
close_correg_VIX_2	
log_return_1	
log_return_2	
log_return_5	

Tras colapsar las variables se volvió a ajustar la regresión con las entradas y sus rezagos seleccionados, los resultados de esta nueva serie de pruebas t se muestran en la **Tabla 32**, se debe añadir que a su vez en este nuevo ajuste se flexibilizó a 0.1 el nivel de confianza de la prueba t. A pesar del ajuste realizado las variables **Cluster_Topics4** y **Cluster_Topics5** con sus respectivos rezagos siguen sin ser estadísticamente significativas para el modelo, es por esto que se decidió retirarlas y volver a ajustar la regresión.

Tabla 32: Significancia de los coeficientes del modelo campeón afinado

Variable	Nivel	DF	Estimación	Error estándar	t valor	Pr > |t|	Flag No Significativo
Intercepto		1	0.0022	0.00079	2.76	0.0058	0
Cluster_Topics2_2	1_ 0-6	1	-0.0004	0.000242	-1.79	0.073	0
Cluster_Topics2_6	1_ 0-6	1	-0.0005	0.000244	-2.02	0.0438	0
Cluster_Topics4_2	1_ 3	1	0.0003	0.000229	1.36	0.1746	1

Variable	Nivel	DF	Estimación	Error estándar	t valor	Pr > |t|	Flag No Significativo
Cluster_Topics4_3	1_ 0-3	1	-0.0004	0.000229	-1.53	0.1255	1
Cluster_Topics4_4	1_ 0-3	1	-0.0001	0.00023	-0.55	0.5833	1
Cluster_Topics5_7	1_ 0-3	1	-0.0002	0.000183	-1.1	0.2733	1
Cluster_Topics7_5	1_ 0-2	1	-0.0004	0.000209	-1.71	0.0872	0
Cluster_Topics7_6	1_ 0-2	1	0.0005	0.000209	2.27	0.0233	0
GIS_Cluster5_GIS_2		1	0.0000	6.28E-06	2.18	0.0298	0
close_correg_VIX		1	-0.0048	0.000092	-51.68	<.0001	0
close_correg_VIX_1		1	0.0038	0.000184	20.86	<.0001	0
close_correg_VIX_2		1	0.0008	0.000159	5.07	<.0001	0
log_return_1		1	-0.1091	0.0278	-3.92	<.0001	0
log_return_2		1	-0.0290	0.0162	-1.79	0.0734	0
log_return_5		1	0.0404	0.0162	2.49	0.0128	0

Finalmente, el ajuste de los coeficientes y las pruebas t asociadas a la regresión se muestran en la siguiente tabla. Se desea subrayar que en este ajuste todos los parámetros son estadísticamente significativos por lo que cada componente aporta información al momento de explicar el comportamiento del Dow Jones Industrial Average (DJIA).

Tabla 33: Significancia de los coeficientes del modelo campeón afinado sin variables no significativas

Variable	Nivel	DF	Estimación	Error estándar	t valor	Pr > |t|	Flag No Significativo
Intercepto		1	0.00195	0.00072	2.71	0.0069	0
Cluster_Topics2_2	1_ 0-6	1	-0.00046	0.00024	-1.91	0.056	0
Cluster_Topics2_6	1_ 0-6	1	-0.00049	0.000243	-2.01	0.0451	0
Cluster_Topics7_5	1_ 0-2	1	-0.00037	0.000209	-1.76	0.0789	0
Cluster_Topics7_6	1_ 0-2	1	0.000474	0.000209	2.26	0.0237	0
GIS_Cluster5_GIS_2		1	0.000014	6.28E-06	2.18	0.0291	0
close_correg_VIX		1	-0.00473	0.000091	-51.75	<.0001	0

Variable	Nivel	DF	Estimación	Error estándar	t valor	Pr > |t|	Flag No Significativo
close_correg_VIX_1		1	0.00383	0.000183	20.96	<.0001	0
close_correg_VIX_2		1	0.000793	0.000159	5	<.0001	0
log_return_1		1	-0.1078	0.0278	-3.88	0.0001	0
log_return_2		1	-0.0281	0.0162	-1.74	0.0828	0
log_return_5		1	0.0366	0.0161	2.28	0.023	0

De manera semejante se analizaron los estadísticos de ajuste del modelo ganador y sus versiones afinadas, dichas métricas se muestran en la **Tabla 34**. Se puede observar que las métricas decaen sutilmente, sin embargo, al hacer este ajuste garantizamos que toda la información es de utilidad, al mismo tiempo que se reduce la complejidad del modelo. Otro rasgo importante del modelo afinado sin variables no significativas es que sigue cumpliendo los supuestos de **independencia** y **media cero**.

Tabla 34: Comparación de estadísticos de modelo ganador y su versión afinada.

Estadístico de Ajuste	Regresión linear con variables originales y rezagos	Regresión Afinada	Regresión Afinada quitando variables no significativas
Error Cuadrático Medio	0.00001195	0.000014236	0.000014264
Raíz del Error Cuadrático Medio	0.003456	0.003773062	0.003776771
Dstat	81.7975	80.77361	80.77361
Valores Positivos Acertados	46.08	45.85	45.73
Valores Negativos Acertados	35.72	34.93	35.04

En conclusión, el mejor modelo que se obtuvo en la realización de este trabajo para explicar el Índice Dow Jones es una regresión lineal que tiene los coeficientes que se muestran en la **Tabla 33** y analíticamente sigue la siguiente ecuación:

$$LogReturn_t = 0.00195 - 0.00046 * Cluster\_Topics2_{t-2}$$
$$- 0.00049 * Cluster\_Topics2_{t-6}$$
$$-0.00037 * Cluster\_Topics7_{t-5}$$
$$+0.000474 * Cluster\_Topics7_{t-6}$$
$$+0.000014 * GIS\_Cluster5\_GIS_{t-2}$$
$$-0.00473 * close\_correg\_VIX_t$$
$$+0.00383 * close\_correg\_VIX_{t-1}$$
$$+0.000793 * close\_correg\_VIX_{t-2}$$
$$-0.1078 * \log\ \_return_{t-1}$$
$$-0.0281 * \log\ \_return_{t-2}$$
$$+0.0366 * \log\ \_return_{t-5} + \varepsilon_t$$

Se observa que la variable que mayor impacto genera al modelo es el **log_return** con sus rezagos uno, dos y cinco, cabe destacar que los rezagos uno y dos impactan de manera negativa lo cual indica que la tendencia del mercado suele mantenerse en al menos dos o tres días. Por otro lado, se observa que el rezago más antiguo se asocia de manera positiva por tanto se podría pensar que cada cinco días sucede en pequeño cambio en la tendencia del índice.

La segunda variable que mayor efecto genera en el modelo es el índice de volatilidad el cuál es un indicador ampliamente usado para conocer el comportamiento del mercado. Se puede observar que al igual que el **log_return**, esta variable solo nos ayuda a darnos una idea de cómo estará el índice en los próximos dos días.

Dejando atrás a las variables tradicionales se puede notar que la incorporación del componente mediático aporta información de utilidad a la modelación, en virtud de que el modelo ganador incorpora fuentes externas como son las variables de Google Trends y las variables que se construyeron con minería de texto a partir de los encabezados de los principales periódicos financieros.

Entre las variables mediáticas que participan en el modelo están **Cluster_Topics2** y **Cluster_Topics7**, los cuales hablan de política de Estados Unidos y de la economía China respectivamente. Hay que recordar que Estados Unidos además de ser la nación a la que pertenece el DJIA, es la mayor economía a nivel mundial, seguida de China, por esto tiene sentido que estas variables sean las que más repercuten al comportamiento del índice. Adicionalmente, se debe añadir que el índice DJIA tiene como componentes a empresas como: Apple, Coca-Cola y JP Morgan que tienen presencia mundial, incluida China.

Finalmente, la última variable que participa en este modelo es **GIS_Cluster5_GIS** con el rezago de dos días, esta variable es el índice de búsqueda en Google de términos asociados a un mercado a la baja. El rezago dos indica que un incremento de búsqueda de términos negativos financieros podría desencadenar en una tendencia a la baja del mercado en un lapso de dos días.

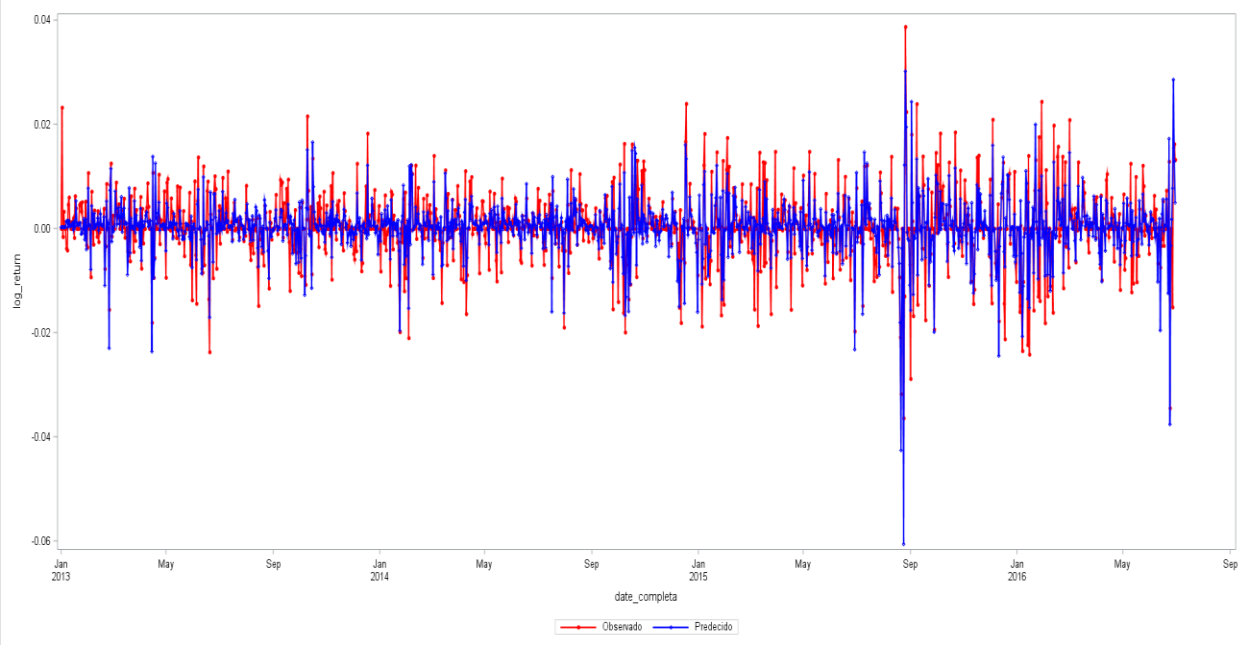
5.2. Analizando resultados del modelo Ganador

En esta sección se analizarán los diferentes resultados obtenidos del modelo ganador para explicar el comportamiento del Dow Jones Industrial Average.

Recordemos que la variable objetivo de este trabajo es el log-rendimiento, este nos permite observar la tendencia del mercado. En la **Figura 35** se presenta la comparación de log-rendimientos observados en la ventana de análisis (Enero 2013 – Junio 2016) versus los log-rendimientos pronosticados.

Se puede observar que los valores pronosticados (línea azul) siguen la tendencia de los observados (línea roja), no obstante, resalta que el modelo capta la tendencia más no la magnitud de los cambios en el índice.

Figura 35: LogReturn observados vs pronosticados.



Se transformaron los valores pronosticados del log return a modo de poder mostrar los pronósticos para el valor de cierre de mercado del DJIA, recordando la definición de mostrada en la sección **1.2.2 Log Return** esta variable se calcula como:

$$R_{\Delta t} = logS(t) - logS(t - \Delta t)$$

En la anterior ecuación S(t) es el precio de cierre del día, por lo que despejando este factor se tiene que:

$$S(t) = e^{R_{\Delta t} - logS(t-\Delta t)}$$

Para hacer esta transformación es necesario contar con un valor inicial del precio de cierre para sustituir en $S(t - \Delta t)$, se tomo el primer valor observado en la ventana de análisis como solución inicial.

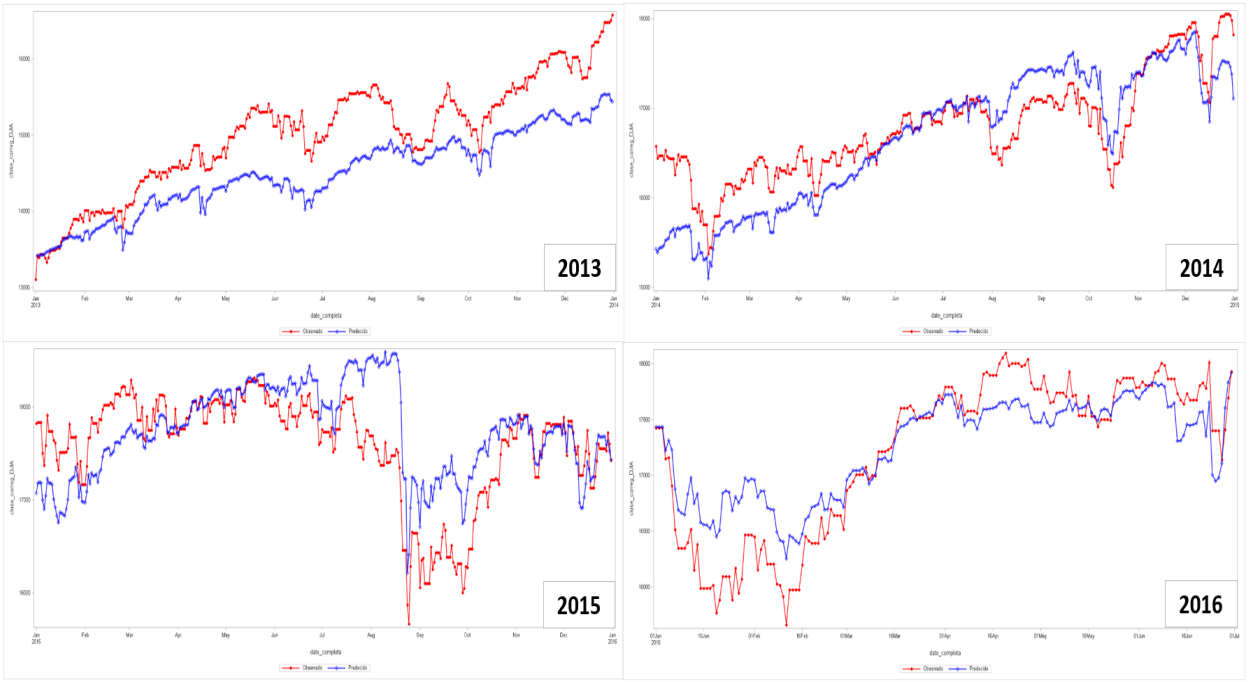
La comparación de la transformación de los valores de cierre pronosticados frente a los observados se muestra en la siguiente imagen. Es preciso señalar que el modelo reacciona oportunamente a los periodos de crisis, por ejemplo, el modelo ajustó las caídas del índice que ocurrieron cercanas al mes de septiembre de los años 2014 y 2015.

Por otro lado, en este gráfico se puede apreciar de mejor forma como el modelo logra explicar la tendencia del índice, en virtud de que se puede apreciar como la línea de valores pronosticados (línea azul) sigue la trayectoria marcada por los valores observados (línea roja).

Figura 36: Valores CLOSE Observados VS Pronósticos.

La siguiente imagen hace un **zoom** por año a la gráfica anterior, para poder apreciar a mayor detalle el ajuste brindado por el modelo ganador en cada uno de los años de análisis.

Figura 37: Valores CLOSE Observados VS Pronosticados por año



5.3. Estabilidad del modelo Ganador

Otro aspecto que se revisó del modelo ganador fue la estabilidad que tiene al utilizarse para evaluar casos distintos a los que fueron utilizados durante su entrenamiento, para esta validación se construyó la ABT para otras ventanas de tiempo. Se tomaron tres periodos de dos meses ajenos al ciclo de análisis (Enero 2013 – Junio 2016), se utilizaron estos rangos de tiempo por disponibilidad de la información.

Los períodos extra de tiempo corresponden a los meses de: Agosto-Septiembre 2016, Noviembre-Diciembre 2016, y Septiembre-Octubre 2012. Se ajustó el modelo ganador en cada una de estas ventanas y se obtuvieron los respectivos estadísticos de ajuste para evaluar el desempeño del modelo.

En la **Tabla 35** muestran los estadísticos de ajuste, se puede observar que en cada una de estas ventanas las métricas decaen menos del 10% con respecto al valor obtenido al evaluar el modelo en la ventana original de análisis.

Tabla 35: Evaluación del modelo ganador en diferentes ventanas de análisis.

Modelo	Average Squared Error	Root Average Squared Error	Dstat	Valores Positivos Acertados	Valores Negativos Acertados
Modelo Ganador Regresión con variables Originales y retrasos	0.00001195	0.003456	81.7975	46.08	35.72
Agosto y Septiembre 2016	0.00001888	0.00434555	75.40984	52.46	22.95
Noviembre y Diciembre 2016	0.00001948	0.004413786	75.40984	63.93	11.48
Septiembre y Octubre de 2012	0.00012267	0.011075439	73.77049	59.02	14.75

En el área de minería de datos el concepto de sobreajuste hace referencia a que el modelo se adapta a los datos de entrenamiento, este concepto también se asocia a si el modelo depende demasiado de características irrelevantes[21].

Algunos autores[22] indican que especializar un modelo agregando demasiadas variables llega a ser contraproducente, debido a que el modelo suele volverse muy específico con los datos de entrenamiento, pero su capacidad para generalizar el comportamiento de los datos y clasificar casos no observados se ve reducido. Por consiguiente, un enfoque estándar para reducir el sobreajuste es optar por modelos con menos variables ya que estos suelen ser más generales, y en segundo lugar sacrificar precisión en el conjunto de entrenamiento, buscando incrementarla en los datos de prueba.

En la sección **5.1. Refinamiento del modelo Ganador** se eliminaron variables redundantes y no significativas para el modelo, el tomar esta acción redujo en aproximadamente un 2% el ajuste del modelo (**Tabla 34**). Sin embargo, se decidió evaluar el modelo ganador original, sin afinar y con las variables no significativas en las ventanas de prueba.

Tabla 36: Resultados de Afinamiento como Estrategia para Evitar Sobreajuste

Modelo	Modelo Afinado Sin Variables No Significativa			Modelo Original		
	Dstat	Valores Positivos Acertados	Valores Negativos Acertados	Dstat	Valores Positivos Acertados	Valores Negativos Acertados
Agosto Septiembre 2016	75.40984	52.46	22.95	65.90909	36.36	29.55
Noviembre y Diciembre 2016	75.40984	63.93	11.48	54.7619	45.24	9.52
Mayo y Junio 2017	75	52.27	22.73	65.90909	50	15.91
Septiembre y Octubre de 2012	73.77049	59.02	14.75	47.5	50.82	6.56

[21] https://towardsdatascience.com/8-simple-techniques-to-prevent-overfitting-4d443da2ef7d
[22] M. Bramer. (2016) Principles of Data Mining. Third Edition. Springer.

Claramente, se puede observar que si el modelo no hubiera sido afinado se tendría un serio problema de sobre ajuste, puesto que el rendimiento del modelo original en las diferentes ventanas decae grandes proporciones. Por consiguiente, se puede concluir que el modelo construido y afinado generaliza adecuadamente el comportamiento del Dow Jones Industrial Average (DJIA).

5.4. Áreas de Oportunidad

Con la intención de evidenciar que el trabajo realizado en este estudio tiene mayor oportunidad de crecimiento, se relatará brevemente un área de crecimiento hacia la que se podría direccionar este trabajo.

Como se había mencionado inicialmente, cada vez existen más modelos que intentan incorporar factores emocionales y del comportamiento de la sociedad al momento de tratar de explicar un fenómeno bursátil, este nuevo enfoque surge en respuesta a observar históricamente la reacción que sucede en un mercado ante periodos de euforia o de pánico por parte de los inversionistas.

Daniel Kahneman[23] describe dos formas que tenemos de pensar: la intuición (sistema 1) y la lógica (sistema 2), él añade que solemos utilizar más el sistema 1. El uso de la intuición suele estar sesgado a creer y confirmar, es decir, si nosotros tenemos una idea sobre como se comportarán los mercados o algunas acciones, solemos buscar información que sustente estos pensamientos.

Por lo anterior, contar con un modelo que analice textos es una buena estrategia para tomar decisiones financieras más completas, ya que una máquina puede procesar una enorme cantidad de información, a diferencia del tiempo que a una persona le tomaría poder leerla. Asimismo, al contar con un modelo que incorporé técnicas de minería de texto se obtiene la capacidad de analizar grandes cantidades de información escrita con una perspectiva objetiva.

El modelo ganador destaca por además de incorporar variables financieras tradicionales, incorporar datos no estructurados de medios sociales (Facebook) y aprovechar la información que estos brindan respecto a situaciones políticas, económicas y sociales y transformar esto a variables que ayuden a evidenciar las repercusiones que estas situaciones generan.

Al ser las situaciones sociales y bursátiles fenómenos de cambio constante, es difícil imaginar que una serie de variables fijas expliquen al mercado por largos periodos de tiempo. Por ejemplo, en el modelo ganador participan las variables **Cluster_Topics2** y **Cluster_Topics7**, que hablan de política de Estados Unidos y de la economía China respectivamente, suena poco probable pensar que estas sean las únicas situaciones que puedan incitar a cambios en el mercado.

Por lo anterior, un área de crecimiento que se observa es reentrenar periódicamente al modelo para observar los temas que cobran importancia a lo largo de tiempo. Como ejercicio se dividió la ventana de análisis (Enero 2013 - Junio 2016) en trimestres y se ajustó el modelo ganador (Regresión con Stepwise) en cada uno de estos períodos disjuntos. En la **Figura 38**, se puede observar que las variables seleccionadas en cada lapso son diferentes, lo cual sostiene la teoría de que el nivel de importancia de los temas sociales es cambiante.

[23] D. Kahneman. (2011). Thinking Fast and Slow. Farrar, Straus and Giroux.

Figura 38: Evolución de las variables en el período de análisis del índice.

	Enero - Marzo 2013	Abril - Junio 2013	Julio - Septiembre 2013	Octubre - Diciembre 2013	Enero - Marzo 2014	Abril - Junio 2014	Julio - Septiembre 2014	Octubre - Diciembre 2014	Enero - Marzo 2015	Abril - Junio 2015	Julio - Septiembre 2015	Octubre - Diciembre 2015	Enero - Marzo 2016	Abril - Junio 2016
Cluster_Topics1	x	x		x	x	x	x	x		x	x			
Cluster_Topics2	x	x	x	x	x			x	x	x	x	x		x
Cluster_Topics3			x	x			x	x	x		x			
Cluster_Topics4		x	x	x	x	x	x	x	x	x	x	x	x	
Cluster_Topics5	x	x	x	x	x		x	x	x	x		x		
Cluster_Topics6			x	x	x	x		x		x	x		x	
Cluster_Topics7	x	x	x	x	x	x		x	x	x	x	x		
Cluster_Topics8	x		x			x	x	x		x	x	x		x
Cluster_Topics9	x	x	x		x	x	x			x	x	x		
Cluster_Topics10	x		x	x	x		x	x		x	x			x
daily_neg_index	x		x	x	x	x				x	x	x		
likes_FBTS		x				x		x		x				
shares_FBTS	x	x	x			x		x		x				x
comments_FBTS	x	x	x	x	x	x		x	x					
freq_noti_FBTS		x		x		x			x	x	x			
fuentes_disti_FBTS						x			x		x			
close_correg_VIX		x	x	x	x					x	x		x	x
log_return	x	x	x		x				x	x	x		x	
GIS_Cluster1_GIS	x	x			x					x			x	
GIS_Cluster2_GIS									x			x		
GIS_Cluster3_GIS														
GIS_Cluster4_GIS					x									
GIS_Cluster5_GIS	x	x	x	x	x									

Vars. Minería de Texto Vars. Medios Sociales Vars. Financieras Vars. Tendencia de Busqueda

Otra área de oportunidad que tiene este trabajo, es incluir información de textos provenientes de otros medios sociales como: Reddit, Twitter o LinkedIn. Lo anterior, debido a que con base a una encuesta realizada en 2016 por el Pew Research Center[24], el 62% de los estadounidenses que ve noticias lo hace a través de redes sociales. Asimismo, en los resultados de esta encuesta se observó que los mayores porcentajes de usuarios que utilizan una red social como fuente de información se encuentran en: Facebook, Reddit, Twitter y LinkedIn.

[24] Gottfried J. and Shearer E. New Use Across Social Media Platforms 2016. Pew Research Center Journalism & Media. [serie en internet] 2016 [consultado 20 de noviembre de 2020]. Disponible en: https://www.journalism.org/2016/05/26/news-use-across-social-media-platforms-2016/

5.5. Conclusiones

Sintetizando, en este trabajo además de construir un modelo que explique el comportamiento de un índice bursátil, se mostró el valor agregado que aporta el utilizar técnicas de minería de texto y minería de datos en la modelación de mercados financieros.

La minería de texto en el contexto de este trabajo sirvió como puente para incorporar aspectos económicos, políticos y sociales, así como el factor emocional a la tarea de explicar el fenómeno del DJIA, a su vez esto permitió medir el papel que estas situaciones generan en el mercado financiero.

Un modelo como el descrito en este trabajo, además de ayudarnos hacer un pronóstico de cómo se comportará un índice, también ayuda a entender la reacción que tiene el mercado ante diferentes circunstancias. Lo anterior coincide con la creciente popularidad de modelos que incorporan factores emocionales, pues estos enfatizan el importante rol que tiene el estado anímico y los sentimientos en la toma de decisiones de los participantes del mercado y el reflejo que tiene el comportamiento humano en los cambios que ocurren en los mercados financieros.

De igual modo, en este trabajo se evidencia el potencial de las técnicas de minería de datos para descubrir relaciones que sirvan de herramienta para explicar fenómenos, tal es el caso del modelo ganador descrito.

Se debe añadir que la incorporación de estas técnicas no intenta remplazar a las variables y técnicas tradicionales, sino se busca brindar apoyo y hacer un cambio de perspectiva que ayude a tener más herramientas que nos permitan entender fenómenos.

Otro aspecto que cabe destacar de este trabajo es la adopción de múltiples medios de información para ampliar la base de conocimiento. En este trabajo además de utilizar las variables financieras tradicionales como son los precios de cierre y la volatilidad del mercado, se incorporación de fuentes de información no convencionales como las reacciones en redes sociales, los encabezados de journals financieros e índices de búsquedas en internet. Tras hacer esta incorporación se infiere que estas fuentes de información tienen valor que aportar para mejorar el pronóstico de modelos financieros.

Finalmente, es necesario resaltar el papel que el acceso a la información y el desarrollo tecnológico representó en el desarrollo de este trabajo, pues gracias a la cantidad de recursos que tenemos hoy en día es que se pudo extraer y procesar toda la información utilizada. Gracias a la tecnología es que ahora tenemos la capacidad de indagar nuevas vertientes de información, pues la tecnología nos brinda un amplio inventario de herramientas. Asimismo, gracias a la capacidad de cómputo es que ahora se puede extraer información de datos no estructurados como son los textos y añadir estos recursos a la generación de modelos.

Anexo 1: Transformación Box Cox

Las transformaciones de Box y Cox son una familia de distribuciones que son usadas para corregir sesgos en la distribución, corregir problemas de heterocedasticidad y dar una forma normal a la distribución de las variables. La transformación Box Cox es llamada en honor a los estadistas George Box y Sir David Roxbee Cox quienes en 1964 colaboraron y desarrollaron esta técnica.

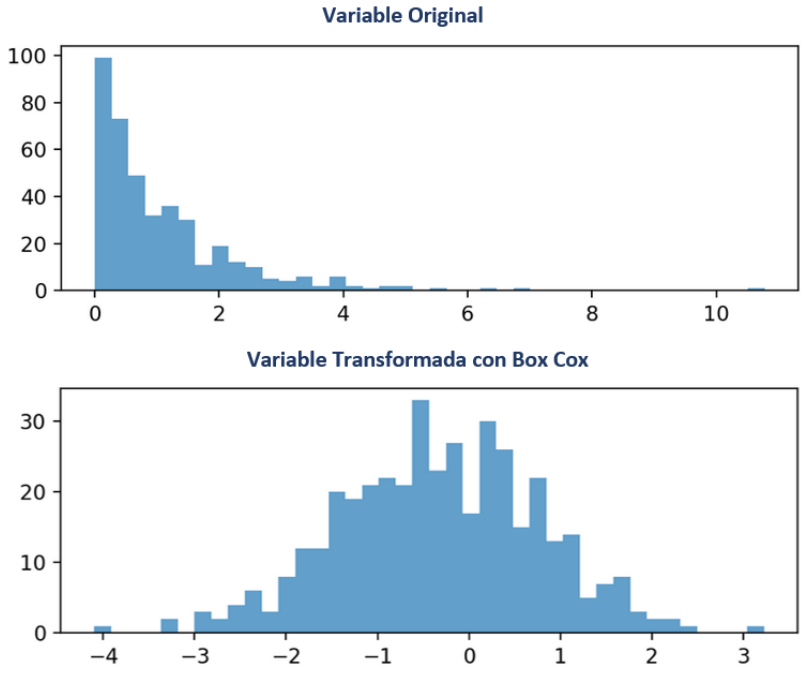
El núcleo de la transformación es el parámetro lambda (λ), este parámetro toma valores en el rango de -5 a 5. Se busca encontrar el valor lambda óptimo, donde el valor óptimo es aquel que consigue la mejor aproximación a la curva de una distribución normal.

La transformación Box Cox tiene la siguiente fórmula:

$$y(\lambda) = \begin{cases} \dfrac{y^{\lambda} - 1}{\lambda}, & si\ \lambda \neq 0; \\ \log y, & si\ \lambda = 0; \end{cases}$$

Cabe resaltar que la transformación solamente funciona con valores positivos.

Figura 39: Ejemplo transformación Box Cox



Anexo 2: Colapsar Variables Categóricas con Árbol de Decisión

Por la manera secuencial de la funcionalidad de la herramienta SAS Enterprise Miner, esta permite utilizar un árbol de decisión para colapsar una variable de entrada en una cantidad N de nodos hoja. Y posteriormente utilizar esta variable colapsada, como entrada de futuros nodos o modelos. Para poder utilizar un árbol de decisión es necesario modificar los siguientes parámetros:

1. Seleccionar únicamente como variable de entrada a aquella que se quiere colapsar.
2. Seleccionar como número máximo de ramas el número de niveles que tiene la variable categórica.
3. Configurar la máxima profundidad a 1.
4. Configurar el nivel de significancia a 1.
5. Configurar el método de subárbol como **Largest,** esto con el objetivo de evitar la post-poda.
6. Indicar que no se haga selección de variables.
7. Definir el rol de entrada para las hojas resultantes.
8. Ejecutar el nodo y posteriormente añadir los nodos de los modelos que se deseen emplear.

Esta uso auxiliar del árbol de decisión fue extraído del manual de Modelado con Árboles de Decisión publicado por SAS Institute en 2012.[25]

[25] L. Rothman and W. Potts. (2012) Decision Tree Modeling Course Notes. SAS Institute.

Anexo 3: Nodo de Transformación de Variables

En este nodo se describe de manera general la funcionalidad y los parámetros del "Nodo de Transformación de Variables" disponible en la herramienta SAS Enterprise Miner.

El nodo de transformación de variables permite crear nuevas variables que son transformaciones de las existentes. Las transformaciones son útiles cuando se desea mejorar el ajuste de un modelo a los datos, debido a que las trasformaciones ayudan a estabilizar varianzas, remover no linealidades y corregir problemas de no-normalidad.

Este nodo divide la forma de transformar con base a si las variables son de intervalo o categóricas, las transformaciones disponibles son las siguientes.

- Variables de Intervalo.

1. **Best:** realiza varias transformaciones y utiliza la transformación que tiene la mejor prueba de Chi Squared con la variable objetivo.
2. **Múltiple:** crea varias transformaciones destinadas a su uso en nodos de selección de variables sucesores.
3. **Log** — transforma usando el logaritmo de la variable.
4. **Registro 10** — transforma utilizando el logaritmo base-10 de la variable.
5. **Raíz cuadrada** — transforma usando la raíz cuadrada de la variable.
6. **Inverso** — transforma usando la inversa de la variable.
7. **Cuadrado** — transforma usando el cuadrado de la variable.
8. **Exponencial** — transforma utilizando el logaritmo exponencial de la variable.
9. **Centrado:** centra los valores de las variables restando la media de cada variable.
10. **Estandarizar:** estandariza la variable restando la media y dividiendo por la desviación estándar.
11. **Rango** — transforma usando un valor escalado de una variable igual a (x - min) / (max - min), donde x es el valor de la variable actual, min es el valor mínimo para esa variable, y max es el valor máximo para esa variable.
12. **Bucket:** los buckets se crean dividiendo los datos en intervalos espaciados uniformemente en función de la diferencia entre los valores máximo y mínimo.
13. **Quantile:** los datos se dividen en grupos con aproximadamente la misma frecuencia en grupos.
14. **Binning óptimo:** los datos se dividen intentando maximizar la relación con la variable objetivo.
15. **Máxima normalidad** —transforma buscando maximizar la normalidad.
16. **Correlación máxima:** transforma buscando maximizar la correlación con la variable objetivo. Esta transformación solo aplica con variables objetivo de tipo intervalo.
17. **Ninguno** — (configuración predeterminada) No se realiza ninguna transformación.

- **Variables Categóricas**

    1. **Niveles raros de grupo:** transformar usando niveles raros.
    2. **Indicadores** ficticios: transformar usando variables dummy. La codificación de variables dummy se aplica a las variables categóricas desde el valor de clase más alto hasta el valor de clase más bajo.
    3. **Ninguno** — (predeterminado) No se realiza ninguna transformación.

Se desea subrayar que la configuración utilizada fue aplicar a todas las variables de entrada de tipo intervalo la mejor transformación y se decidió no transformar con este nodo a las variables categóricas.

Bibliografía

[1] M. Mittermayer and G. Knolmayer,(2006) Text Mining Systems for Market Response to News: A Survey, Institute of Information Systems University of Bern

[2] B. Wüthrich, D. Permunetilleke, S. Leung, V. Cho, J. Zhang and W. Lam, (1998) Daily Prediction of Major Stock Indices from Textual WWW Data., American Association for Artificial Intelligence

[3] H. Mao, S. Counts, and J. Bollen, (2011). Predicting Financial Markets: Comparing Survey, News, Twitter and Search Engine Data., Journal of Computational Science.

[4] P. C. Tetlock. (2007). Giving content to investor sentiment: The role of media in the stock market. Journal of Finance.

[5] B. Zhao, H. Yongji, Y. Chunfeng and H. Yihua, (2016) Stock Market Prediction Exploiting Microblog Sentiment Analysis. Center of Novel Software Technology and Industrialization.

[6] V. Lavrenko, M. Schmill, D. Lawrie, P. Ogilvie, D. Jensen, and J. Allan, (2000), Mining of Concurrent Text and Time Series. Department of Computer Science University of Massachusetts

[7] N. Blasco y S. Ferreruela., (2017), Behavioral Finance: ¿Por qué los inversores se comportan como lo hacen y no como deberían?. Estudios y publicaciones Facultad De Economía y Empresa. Universidad de Zaragoza.

[8] V. Ravi, D. Pradeepkumar and K. Deb, (2017), Financial time series prediction using hybrids of chaos theory, multi-layer perceptron and multi-objective evolutionary algorithms. Swarm and Evolutionary Computation

[9] L. Rothman and W. Potts. (2012) Decision Tree Modeling Course Notes. SAS Institute.

[10] Baltagi B., (2008). Econometrics. Editorial Springer

[11] W. Griffiths, R. Carter and G. Lim.

[12] Bollen J., Mao H. and Zeng X., (2010). Twitter mood predicts the stock market

[13] Randall S. (2016). Strategic Analytics and SAS Using Aggregate Data to Drive Organizational Initiatives. Cary, NC: SAS Institute Inc. (2012). Using SAS for Econometrics Fourth Edition. John Wiley & Sons, Inc.

[14] SAS Institute Inc. (2015). SAS® Text Miner 14.1: Reference Help. Cary, NC: SAS Institute Inc.

[15] D. Gupta. (2018). Applied Analytics through Case Studies Using SAS and R. Apress.

[16] L. Breiman, J. Friedman, R. Olshen and C. Stone. (1998). Classification and Regression Trees. Chapman & Hall/CRC.

[17] J. Han, M. Kamber and J. Pei. (2012). Data Mining: Concepts and Techniques Third Edition. Morgan Kaufmann.

[18] P. Christie, J. Georges, J Thompson and C. Wells. (2011). Applied Analytics Using SAS Enterprise Miner Course Notes. SAS Institute.

[19] J. Han, J. Pei and M. Kamber. (2011). Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems.

[20] D. Pasta y D. Suhr (2004). SAS Users Group International 29: Best Papers and honorable mentions (SUGI 29). SAS Institute