

“REDES NEURONALES APLICADAS A LA INFLACIÓN”

CRISTIAN LUCIANO SALTO

INGENIERÍA EN INFORMÁTICA

FACULTAD DE INGENIERÍA

AÑO 2016



Redes neuronales aplicadas a la inflación

Cristian Luciano Salto
Ingeniería en informática
Facultad de ingeniería
Año 2016

Título

Ingeniero en Informática

Profesor guía

Nombre: Dra. Alicia Pérez

Firma: _____

Tribunal Evaluador

Nombre: _____

Firma: _____

Nombre: _____

Firma: _____

Nombre: _____

Firma: _____

Alumno

Nombre: Cristian Luciano Salto

Firma: _____

Fecha de exposición: ____/____/____

Dedicatorias

A mis padres Adriana y Miguel que siempre me apoyaron incondicionalmente.

A mi abuela María Cecilia, en su memoria.

Agradecimientos

A la Dra. Alicia Pérez, mi profesora y tutora, a quien admiro y respeto. Agradezco su paciencia y apoyo que me han servido para realizar el trabajo y crecer profesionalmente.

A mis familiares, amigos, profesores, compañeros y todas aquellas personas que me acompañaron durante mis estudios universitarios.

Índice

Índice de tablas	8
Índice de ilustraciones	9
Abstract.....	11
1 Introducción.....	12
1.1 Descripción del problema.....	12
1.2 Objetivos.....	12
1.3 Motivación.....	12
1.4 Organización del trabajo.....	13
2 Marco teórico.....	14
2.1 El proceso de extracción del conocimiento	14
2.2 Minería de datos	15
2.3 Metodología de minería de datos.....	17
2.3.1 CRISP-DM	17
2.4 Redes Neuronales	18
2.4.1 Función de entrada.....	18
2.4.2 Función de activación.....	19
2.4.3 Valor de salida	19
2.5 Estructuras de las redes neuronales	19
2.6 El aprendizaje en las redes neuronales artificiales	20
2.7 Series temporales.....	20
2.7.1 Extracción de patrones secuenciales.....	21
2.7.2 Clasificación con datos secuenciales y ventanas deslizantes	21
2.7.3 Pronóstico de series temporales con redes neuronales artificiales	22
2.8 Resumen	24
3 Comprensión del negocio	25
3.1 Determinación de objetivos del proyecto	25
3.1.1 Contexto.....	25
3.1.2 Casos de éxito	28
3.1.3 Solución propuesta	28
3.1.4 Objetivos del proyecto.....	29
3.1.5 Criterios de éxito del proyecto.....	29
3.2 Evaluación de la situación	29
3.2.1 Fuentes de datos y conocimiento.....	29
3.2.2 Requerimientos, presunciones y restricciones	29
3.3 Plan de proyecto	30

3.3.1	Alcance	30
3.3.2	Estructura de desglose del trabajo (EDT).....	31
3.3.3	Lista de actividades	32
3.3.4	Requisitos de recursos de la actividad	33
3.3.5	Cronograma del proyecto	35
3.3.6	Gestión de costo y adquisiciones.....	36
3.3.7	Descripciones de cargo	37
3.3.8	Registro de riesgos.....	38
3.4	Alternativas tecnológicas de solución al problema	39
3.4.1	Alternativas de hardware	39
3.4.2	Alternativas de sistemas operativos.....	39
3.4.3	Alternativas de software para minería de datos.....	40
3.4.4	Elección de las alternativas tecnológicas.....	40
3.5	Análisis costo-beneficio del proyecto.....	41
3.6	Resumen	42
4	Comprensión de los datos.....	43
4.1	Recolección de datos iniciales	43
4.2	Descripción de datos.....	44
4.2.1	Informe Monetario Mensual.....	44
4.2.2	Informe Panorama global commodities.....	47
4.2.3	Tasas de desocupación.....	48
4.2.4	Oferta y demanda.....	49
4.2.5	Producto bruto interno a precio de mercado.....	50
4.2.6	Tipo de cambio nominal promedio mensual	51
4.2.7	Índice de precios al consumidor	52
4.3	Exploración de datos	52
4.3.1	Comparación de la tasa de desocupación y el índice de precios al consumidor	52
4.3.2	Distribución de la base monetaria	55
4.4	Verificación de calidad de los datos	56
4.4.1	Leche (por libras).....	56
4.4.2	Carne vacuna (por tonelada).....	57
4.4.3	Azúcar (en libras)	58
4.4.4	Cheques cancelatorios	58
4.4.5	Stock de pases Activos	59
4.4.6	Etanol (por galón) y Biodiesel (por galón).....	60
4.4.7	Plata (en oz. Troy)	61

4.4.8	Tratamiento de los valores nulos y/o erróneos	61
4.5	Resumen	62
5	Preparación de los datos	63
5.1	Limpieza y selección de datos	63
5.1.1	Atributos correspondientes a periodos de tiempo.....	63
5.1.2	Datos agregados.....	63
5.1.3	Otros atributos que no serán considerados en el análisis.....	64
5.1.4	Datos seleccionados.....	64
5.2	Construcción de datos.....	67
5.2.1	Período.....	67
5.2.2	Unificación de moneda económica.....	67
5.3	Formateo de datos.....	68
5.4	Integración de datos.....	68
5.5	Vista minable	68
5.6	Resumen	71
6	Modelado.....	72
6.1	Ventanas de tiempo y validación de ventanas deslizantes	72
6.1.1	Ventanas de tiempo en RapidMiner	72
6.1.2	Validación de ventanas deslizantes en RapidMiner	74
6.2	Construcción del modelo	75
6.2.1	Primer proceso.....	75
6.2.2	Segundo proceso.....	79
6.2.3	Tercer proceso	86
6.3	Descripción del modelo	88
6.4	Resumen	90
7	Evaluación	91
7.1	Evaluación de los resultados.....	91
7.2	Revisión del proyecto	91
7.2.1	Metodología de minería de datos.....	91
7.2.2	Fuentes de datos y conocimiento.....	92
7.2.3	Software para minería de datos	92
7.2.4	Gestión del proyecto con PMI	92
8	Conclusiones y futuras líneas de investigación	93
	Bibliografía.....	94
	Anexo I	97
	Anexo II.....	103

Anexo III.....	105
Interpolación a trozos (Velásquez Zapateiro & Obeso Fernández, 2007).....	105
Interpolación lineal a trozos	105
Aplicación a los datos trimestrales de la interpolación lineal a trozos	106
Anexo IV	109
Máquinas de vectores soporte (SVM)	109
Desvío estándar.....	110
Análisis de componentes principales (PCA)	110
Relief	111
Normalización de los pesos	113
Anexo V	114
Regresión lineal simple	114
Regresión lineal múltiple.....	115
Anexo VI	116

Índice de tablas

Tabla 1. Requisitos de recursos de las actividades	34
Tabla 2. Costos de recursos humanos.....	36
Tabla 3. Costos de hardware.....	36
Tabla 4. Gastos varios	37
Tabla 5. Línea base de costos	37
Tabla 6. Descripciones de cargo	37
Tabla 7. Matriz de probabilidad e impacto de riesgos.....	39
Tabla 8. Descripción de datos pertenecientes a los informes monetarios mensuales.....	46
Tabla 9. Descripción de datos pertenecientes a los informes de panorama global de commodities.....	48
Tabla 10. Descripción de datos de tasa de desocupación	48
Tabla 11. Descripción de datos de oferta y demanda	49
Tabla 12. Descripción de datos de producto bruto interno a precio de mercado	50
Tabla 13. Descripción de datos de tipo de cambio nominal promedio mensual	51
Tabla 14. Descripción de datos de índice de precios al consumidor	52
Tabla 15. Datos seleccionados.....	67
Tabla 16. Ejemplo de diferencias de medidas monetarias.....	67
Tabla 17. Descripción de datos de la vista minable	71
Tabla 18. Resultados de los rendimientos de los modelos elaborados en el primer proceso ...	78
Tabla 19. Resultados de los rendimientos de los modelos elaborados en el segundo proceso	84
Tabla 20. Comparación del rendimiento de las redes neuronales con otras técnicas.....	85
Tabla 21. Evaluación de los modelos con predicciones a tres meses.....	87
Tabla 22. Evaluación de los modelos con predicciones a seis meses.....	87
Tabla 23. Evaluación de los modelos con predicciones a nueve meses	87
Tabla 24. Evaluación de los modelos con predicciones a doce meses	87
Tabla 25. Encuesta de herramientas de Big Data utilizadas.....	102
Tabla 26. Interpolación lineal a trozos del PBI mensual	108

Índice de ilustraciones

Ilustración 1. Fases del proceso KDD (Hernández Orallo et al, 2004)	15
Ilustración 2. Proceso ideal de minería de datos (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004).....	16
Ilustración 3. Fases de CRISP-DM (Chapman et al, 2000).....	18
Ilustración 4. Funciones de una neurona (Russell & Norvig, 2004)	19
Ilustración 5. Curva de Philips a corto plazo	26
Ilustración 6. Estructura de desglose de trabajo	31
Ilustración 7. Cronograma de actividades	35
Ilustración 8. Proceso de descripción de datos	44
Ilustración 9. Distribución de la tasa de desocupación.....	53
Ilustración 10. Distribución de la variación porcentual.....	53
Ilustración 11. Tendencia de la tasa de desocupación	54
Ilustración 12. Tendencia del índice de precios al consumidor base Abril 2008 = 100.....	54
Ilustración 13. Distribución del índice de precios al consumidor base Abril 2008 = 100.....	55
Ilustración 14. Distribución de la base monetaria	55
Ilustración 15. Distribución de la leche	56
Ilustración 16. Distribución de la carne vacuna	57
Ilustración 17. Histograma de Carne vacuna (por tonelada)	57
Ilustración 18. Diagrama de dispersión del azúcar.....	58
Ilustración 19. Distribución de los cheques cancelatorios.....	58
Ilustración 20. Distribución del stock de pases activos	59
Ilustración 21. Diagrama de caja del stock de pases activos	59
Ilustración 22. Distribución de Etanol (por galón)	60
Ilustración 23. Distribución de Biodiesel (por galón)	60
Ilustración 24. Distribución de la plata en dólares	61
Ilustración 25. Proceso de descripción de datos de la vista minable	71
Ilustración 26. Construcción de ventanas de tamaño 2 para un conjunto de datos	73
Ilustración 27. Iteraciones de una validación de ventanas deslizantes	74
Ilustración 28. Primer proceso para la construcción del modelo.....	75
Ilustración 29. Subproceso del operador Sliding Window Validation	76
Ilustración 30. Proceso de selección de atributos	79
Ilustración 31. Datos de salida de un operador Weights to Data.....	80
Ilustración 32. Datos de salida de un operador Generate Attribute	80
Ilustración 33. Subproceso del operador SP-Weight by SVM	81
Ilustración 34. Subproceso del operador SP-Weight by Deviation	81
Ilustración 35. Subproceso del operador SP-Weight by PCA	81
Ilustración 36. Subproceso del operador SP-Weight by Relief	81
Ilustración 37. Diez mejores atributos ponderados por Weight by SVM.....	82
Ilustración 38. Diez mejores atributos ponderados por Weight by Deviation.....	82
Ilustración 39. Diez mejores atributos ponderados por Weight by PCA.....	82
Ilustración 40. Diez mejores atributos ponderados por Weight by Relief.....	83
Ilustración 41. Segundo proceso para la construcción del modelo.....	83
Ilustración 42. Subproceso del operador Sliding Window Validation	83
Ilustración 43. Gráfico de los pronósticos e Índice de precios al consumidor	86
Ilustración 44. Representación gráfica del modelo de datos elaborado	89
Ilustración 45. Proceso para generar las tendencias de las series	103
Ilustración 46. Operador interno al FitTrend.....	104

Ilustración 47. Curva polinomial extraído de (Zapateiro & Fernández, 2007)	105
Ilustración 48. Polinomio interpolador extraído de (Zapateiro & Fernández, 2007)	105
Ilustración 49. Interpolación lineal a trozos para datos trimestrales	106
Ilustración 50. El margen suave de perdida ajustado para una SVM lineal (Smola & Schölkopf, 2003)	110
Ilustración 51. Algoritmo de Relief	112
Ilustración 52. Algoritmo de Relief para problemas de regresión	113
Ilustración 53. Regresión lineal simple	114

Abstract

En la Argentina, la inflación es uno de los principales problemas de la economía. El presente proyecto aplica el proceso de minería de datos para encontrar patrones en variables económicas que permitan pronosticar la inflación. Para ello fue necesario obtener datos de economía, comprenderlos, prepararlos, elaborar el modelo de datos y analizar los resultados obtenidos. Con el objetivo de construir el modelo de datos se utilizaron redes neuronales y series de tiempo, de esta manera es posible realizar proyecciones de la inflación. Como resultado del proyecto se encuentran relaciones no lineales entre variables monetarias del sistema financiero asociadas al índice de precios al consumidor.

1 Introducción

El objetivo del primer capítulo es introducir al lector en la problemática del proyecto, su solución y los aspectos que motivaron su desarrollo. Luego de esto se presenta la organización del trabajo en los siguientes capítulos.

1.1 Descripción del problema

La inflación, en términos económicos, es el aumento sostenido de los precios. Este fenómeno acarrea numerosas consecuencias negativas: genera pérdida del poder adquisitivo, desalienta la inversión y frena el crecimiento, entre otros efectos.

A fin de poder estabilizar la inflación y controlar los factores que inciden en su crecimiento resulta necesario estudiar su problemática no solo desde el punto de vista económico, sino también, por ejemplo, utilizando otras técnicas que permitan extraer patrones de conocimiento a partir de datos de la economía. En este sentido, el presente trabajo aplicará algoritmos de minería de datos para extraer patrones de conocimiento relacionado con la inflación utilizando datos de la economía argentina.

1.2 Objetivos

El objetivo principal del trabajo es analizar la inflación a través de la aplicación de técnicas de minería de datos. La aplicación de estas técnicas tiene dos objetivos específicos. El primero de ellos es determinar variables económicas asociadas con el aumento de precios. El segundo de los objetivos del trabajo consiste en elaborar modelos de datos para pronosticar en base a estas variables el índice de precios. Es importante aclarar que un índice de precios es un indicador que tiene por objeto medir las variaciones, a lo largo del tiempo, en los precios de un conjunto definido de bienes y servicios.

El cumplimiento de los objetivos requiere de la aplicación de técnicas de procesamiento de datos. En este sentido se utilizarán redes neuronales para construir los modelos de datos en base a variables de la economía argentina.

1.3 Motivación

La elección del tema del proyecto responde a diversas circunstancias. En relación a la problemática, se trata de un problema de interés general que requiere una solución de manera inmediata. Además es posible realizar el proyecto ya que existen datos referidos a la economía en organismos oficiales como ser el INDEC (Instituto Nacional de Estadística y Censos de la República Argentina) y el BCRA (Banco Central de la República Argentina).

En cuanto a la tecnología utilizada, la minería de datos es una tecnología que está en pleno desarrollo como respuesta a la necesidad de procesar grandes volúmenes de datos. Esta nueva actividad es un área multidisciplinaria que agrupa distintas ramas del conocimiento como estadística, base de datos, algoritmos, inteligencia artificial y visualización.

Por lo que se refiere al tipo de algoritmo utilizado para realizar minería de datos, las redes neuronales son un paradigma de aprendizaje y procesamiento de datos inspirado en el funcionamiento de los sistemas nerviosos biológicos. Numerosos trabajos se han desarrollado utilizando este tipo de algoritmos, destacándose su capacidad de establecer relaciones no lineales en el conjunto de datos.

En definitiva, resulta atractivo utilizar este tipo de herramienta para analizar un fenómeno económico-social en un marco de un proceso de minería de datos.

1.4 Organización del trabajo

Luego del presente capítulo, la organización del trabajo se realiza de la siguiente manera:

- En el segundo capítulo se presentan los aspectos teóricos ligados a la solución del problema del proyecto, es decir el proceso, la metodología y los algoritmos de minería de datos.
- Los siguientes cinco capítulos corresponden a la aplicación de la metodología CRISP-DM al proyecto, el capítulo 3 contiene la etapa de Comprensión del Negocio, aquí se realiza un análisis de la problemática y el contexto en el que se desarrolla el proyecto. Estos análisis culminan con la elaboración del plan del proyecto.
- En el capítulo 4 son desarrolladas las actividades vinculadas a la fase de Comprensión de Datos. En esta parte del proyecto se ejecutan las tareas de recolección de datos y selección de los atributos que no contienen valores nulos o erróneos.
- Siguiendo con la metodología, en el capítulo 5 se presentan las actividades relacionadas a la etapa de Preparación de Datos. El objetivo de estas tareas es obtener el conjunto de datos que serán utilizados en la etapa de modelado. Para esto es necesario acondicionar los datos, es decir, decidir qué datos serán utilizados, realizar las operaciones de atributos derivados o nuevos atributos y finalmente, integrar los datos de distintas fuentes.
- La cuarta etapa de la metodología, Modelado, está documentada en el capítulo 6. Aquí se elabora y prueba el modelo de datos. A fin de obtener el mejor modelo de datos, se desarrollaron sucesivos experimentos, los cuales están incluidos en esta división del proyecto.
- Finalmente los capítulos 7 y 8, contienen la evaluación del proyecto, conclusiones y futuras líneas de investigación.

2 Marco teórico

En este segundo capítulo del proyecto se desarrollarán los conceptos teóricos que están relacionados a la solución del problema del trabajo. En este sentido es importante prestar atención al proceso de minería de datos y a las técnicas utilizadas para elaborar los modelos de datos. Los aspectos teóricos vinculados a la inflación serán desarrollados más adelante, véase Sección 3.1.1.

2.1 El proceso de extracción del conocimiento

“El proceso de extracción del conocimiento es el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos” (Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996). Las propiedades deseables del conocimiento extraído son:

- Valido: los patrones deben ser precisos para datos usados en su obtención, y para datos nuevos.
- Novedoso: deben aportar algo desconocido para el sistema y para el usuario.
- Potencialmente útil: la información debe conducir a acciones que reporten algún tipo de beneficio para el usuario.
- Comprensible: los patrones obtenidos deben posibilitar su interpretación, revisión, validación y uso.

El descubrimiento de conocimiento en base de datos (Knowledge Discovery in Databases, KDD) es un proceso complejo que incluye no solo la obtención de los modelos o patrones (el objetivo de la minería de datos), sino también la evaluación y posible interpretación de los mismos.

El proceso de KDD es iterativo e interactivo. Es iterativo ya que la salida de algunas de las fases puede hacer volver a pasos anteriores y porque a menudo son necesarias varias interacciones para extraer conocimiento de alta calidad. Es interactivo porque el usuario, o más generalmente un experto en el dominio del problema, debe ayudar en la preparación de los datos, validación del conocimiento extraído, etc (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004).

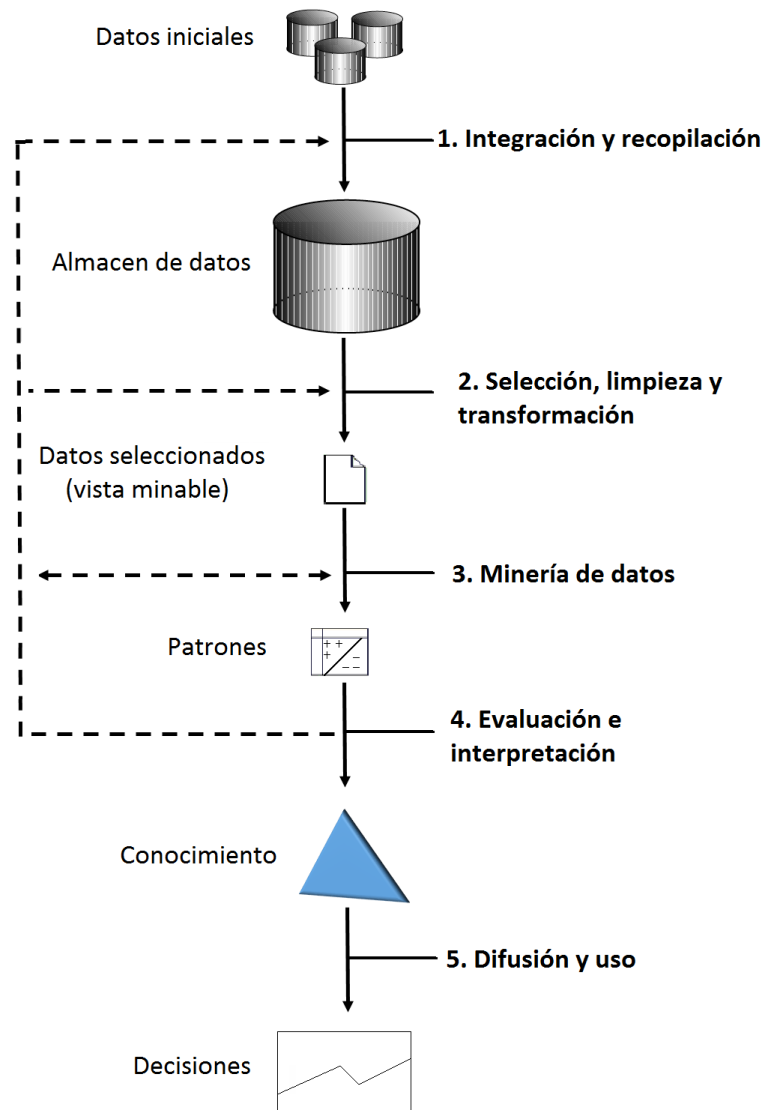


Ilustración 1. Fases del proceso KDD (Hernández Orallo et al, 2004)

El proceso de KDD se organiza en torno a cinco fases. En la fase de integración y recopilación de datos se determinan las fuentes de información que pueden ser útiles y donde conseguirlas. A continuación, se transforman todos los datos a un formato común. Luego, en la fase de selección, limpieza y transformación se eliminan o corrigen los datos incorrectos y se decide la estrategia a seguir con los datos incompletos. En la fase de minería de datos se decide cual es la tarea a realizar (clasificar, predecir, agrupar, etc.) y se elige el método que se va a utilizar. En la fase de evaluación e implementación se evalúan los patrones y se analizan por los expertos, y si es necesario se vuelve a las fases anteriores para una nueva iteración. Finalmente, en la fase de difusión se hace uso del nuevo conocimiento y se hace partícipe de el a todos los posibles usuarios (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004).

2.2 Minería de datos

La minería de datos es un paso esencial de un proceso más amplio cuyo objetivo es el descubrimiento de conocimiento en bases de datos.

Se puede definir a la minería de datos como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos.

La minería de datos es un proceso porque está formada por un conjunto de acciones y actividades interrelacionadas realizadas para extraer conocimiento.

Por otra parte la minería de datos puede ser aplicable a cualquier tipo de repositorio de datos, sean bases de datos relacionales, almacenes de datos (data warehouse), bases de datos transaccionales, sistemas avanzados de bases de datos, archivos planos y la World Wide Web. Los sistemas de bases de datos avanzadas incluyen bases de datos objeto-relacionales y bases de datos específicas orientadas a la aplicación, tales como bases de datos espaciales o bases de datos de series de tiempo.

Durante la fase de minería de datos es necesario tomar una serie de decisiones:

- Determinar qué tipo de tarea de minería es el más apropiado. Las tareas pueden ser predictivas o descriptivas. Los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, denominadas variables objetivo o dependientes, usando otras variables o campos de la base de datos a las que se conocen como variables independientes o predictivas. Los modelos descriptivos, en cambio, identifican patrones que explican o resumen los datos, es decir, sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos.
- Elegir el tipo de modelo. Por ejemplo, para una tarea de clasificación se puede utilizar un árbol de decisión.
- Elegir el algoritmo de minería que resuelva la tarea y obtenga el tipo de modelo que estamos buscando (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004).

Una vez definida la tarea y la técnica de minería de datos, la etapa de minería de datos se podría resumir a través de la siguiente imagen:

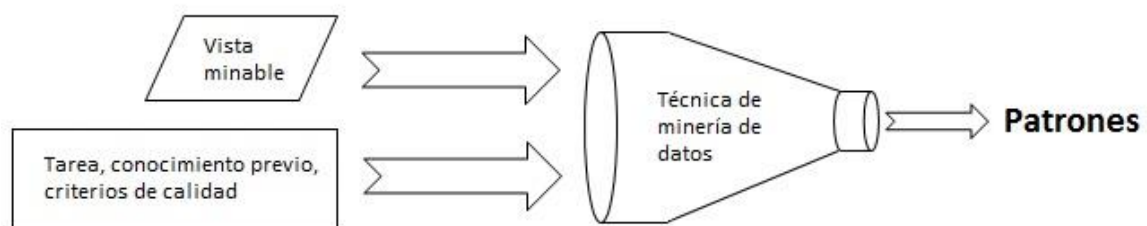


Ilustración 2. Proceso ideal de minería de datos (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004)

A simple vista, la extracción de patrones se logra aplicando la técnica de minería de datos al conjunto de datos obtenidos. Sin embargo este proceso presenta algunas dificultades. La primera de ellas es que los algoritmos utilizados para extraer patrones son computacionalmente costosos; mientras mayor sea el volumen de datos y la complejidad de los patrones a extraer, mayor serán los recursos necesarios para realizar este proceso.

Por otro lado, es necesario contar con conocimientos para elegir e implementar el método en un ordenador. Existen varios métodos de minería de datos dependiendo del tipo de problema. A la hora de implementar estos métodos es necesario conocer el funcionamiento de los algoritmos para poder seleccionar de manera correcta los parámetros que requiera el mismo.

Además de los datos y la tarea, existen otros aspectos que influyen en el aprendizaje, que suelen denominarse conjuntamente bias. Quizás el bias que más influye en esta complejidad sea la manera de expresar o definir los patrones (bias del lenguaje). Por ejemplo, no es lo mismo una regresión lineal que una regresión realizada por una red neuronal multicapa. Ambos métodos permiten realizar la misma tarea, pero la expresividad y la mayor capacidad de la segunda se paga, de alguna manera, con un mayor tiempo de espera para obtener el modelo. El conocimiento previo es otro tipo de bias, que puede ayudar a refinar el espacio de búsqueda (bias de búsqueda) (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004).

2.3 Metodología de minería de datos

2.3.1 CRISP-DM

CRISP-DM de Cross Industry Standard Process for Data Mining es un modelo de proceso de minería de datos. Está compuesto por seis fases (Chapman, y otros, 2000):

- **Comprensión del negocio.** Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto y exigencias desde una perspectiva de negocio, luego convirtiendo este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.
- **Comprensión de los datos.** La fase de entendimiento de datos comienza con la colección de datos inicial y continua con las actividades que le permiten familiarizar primero con los datos, identificar los problemas de calidad de datos, descubrir los primeros conocimientos en los datos, y/o descubrir subconjuntos interesantes para formular hipótesis en cuanto a la información oculta.
- **Preparación de los datos.** La fase de preparación de los datos cubre todas las actividades necesarias para construir el conjunto de datos final (los datos que serán provisto en las herramientas de modelado) de los datos en brutos iniciales. Las tareas de preparación de datos probablemente van a ser realizadas muchas veces y no en cualquier orden prescripto. Las tareas incluyen la selección de tablas, registros, y atributos, así como la transformación y la limpieza de datos para las herramientas que modelan.
- **Modelado.** En esta fase, varias técnicas de modelado son seleccionadas y aplicadas, y sus parámetros son calibrados a valores óptimos. Típicamente hay varias técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de datos. Por lo tanto, volver a la fase de preparación de datos es a menudo necesario.
- **Evaluación.** Antes de proceder al despliegue final del modelo, es importante evaluar a fondo ello y la revisión de los pasos ejecutados para crearlo, para comparar el modelo correctamente obtenido con los objetivos de negocio. Un objetivo clave es determinar si hay alguna cuestión importante de negocio que no ha sido suficientemente considerada. En el final de esta fase, una decisión en el uso de los resultados de minería de datos debería ser obtenida.
- **Implantación.** Dependiendo de los requerimientos, la fase de implantación puede ser tan simple como la generación de un informe o tan compleja como la realización repetida de un proceso cruzado de minería de datos a través de la empresa. En muchos casos, es el cliente, no el analista de datos, quien lleva el paso de desarrollo. Sin embargo, incluso si el analista realizara el esfuerzo de despliegue, esto es importante para el cliente para entender de frente que acciones necesita para ser ejecutadas en orden para hacer uso de los modelos creados actualmente.

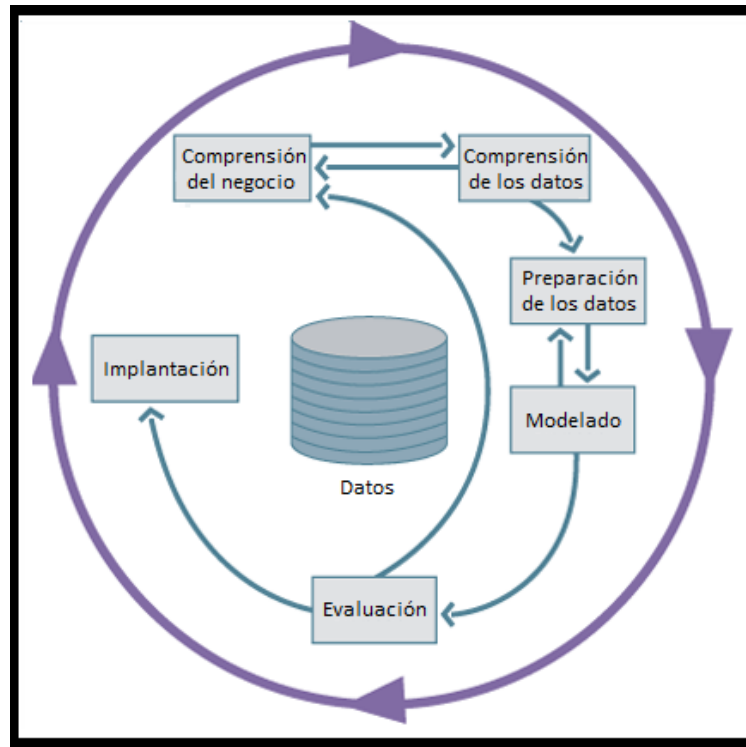


Ilustración 3. Fases de CRISP-DM (Chapman et al, 2000)

2.4 Redes Neuronales

Para descubrir patrones a partir de los datos es necesario aplicar técnicas de minería de datos. De acuerdo al problema de minería de datos, se aplica una técnica o método de minería de datos. Así para los problemas de predicción se pueden resolver mediante redes neuronales.

Las redes neuronales son técnicas de aprendizaje y procesamiento de datos. Están formadas por nodos de entrada, nodos de procesamiento y salida. Para poder realizar el aprendizaje se aplica un proceso de entrenamiento a las redes neuronales, en este periodo se ajustan los valores utilizados en los nodos para obtener las salidas.

2.4.1 Función de entrada

Las redes de neuronas están compuestas de nodos o unidades conectadas a través de conexiones dirigidas. Una conexión de la unidad j a la unidad i sirve para propagar la activación a_j de j a i . Además cada conexión tiene un peso numérico $W_{j,i}$ asociado, que determina la fuerza y el signo de la conexión. Cada unidad i primero calcula una suma ponderada de sus entradas (Russell & Norvig, 2004):

$$in_i = \sum_{j=0}^n W_{j,i} a_j$$

Los valores de entrada se multiplican por los pesos correspondientes. Por consiguiente, los pesos que generalmente no están restringidos cambian la medida de influencia que tienen los valores de entrada. Es decir, que permiten que un gran valor de entrada tenga solamente una pequeña influencia, si estos son lo suficientemente pequeños.

Además la sumatoria de las entradas pesadas, también se puede aplicar:

- Productoria de las entradas pesadas: es el producto de todos los valores de entrada a la neurona, multiplicados por sus correspondientes pesos.

$$in_i = \prod_{j=0}^n W_{j,i} a_j$$

- Máximo de las entradas pesadas: solamente toma en consideración el valor de entrada más fuerte, previamente multiplicado por su peso correspondiente.

$$in_i = \text{Max}(W_{j,i} a_j)$$

2.4.2 Función de activación

Luego aplica una función de activación g a esta suma para producir la salida (Russell & Norvig, 2004):

$$a_i = g(in_i) = g\left(\sum_{j=0}^n W_{j,i} a_j\right)$$

La función de activación g se diseña con dos objetivos. Primero, la unidad estará activa, cuando su valor es cercano a 1, es decir, se proporcionan las entradas correctas; mientras que estará inactiva (cercana a 0) cuando se den las entradas erróneas. Segundo, la activación tiene que ser no lineal, en otro caso la red neuronal en su totalidad se colapsaría con una sencilla función lineal. Dos posibles funciones g son la función umbral y la función sigmoide. La función sigmoide tiene la ventaja de ser diferenciable, lo cual es importante para el algoritmo del aprendizaje de los pesos.

2.4.3 Valor de salida

El último componente de una neurona es la salida (Matich, 2001). El valor resultante de la función de activación será el valor que se transfiere a las neuronas vinculadas.

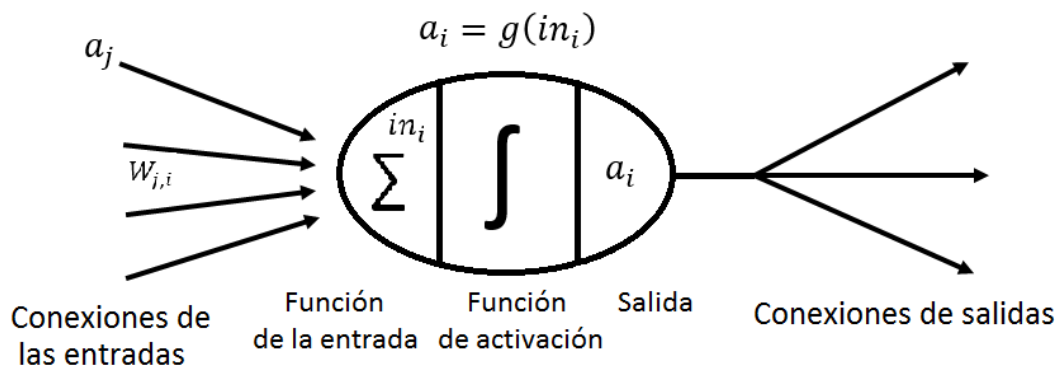


Ilustración 4. Funciones de una neurona (Russell & Norvig, 2004)

2.5 Estructuras de las redes neuronales

Existen dos categorías principales de estructuras de redes neuronales: acíclicas o redes con alimentación hacia adelante y cíclicas o redes recurrentes. Una red con alimentación hacia adelante representa una función de sus entradas actuales; de este modo, no tiene otro estado interno que no sea el de sus propios pesos. Por otro lado, una red recurrente permite que sus salidas alimenten sus propias entradas. Esto significa que los niveles de activación de la red

forman un sistema dinámico que debe alcanzar un estado estable, exhibir oscilaciones o incluso un comportamiento caótico.

Las redes neuronales con alimentación hacia delante normalmente se organizan en capas, de forma que cada unidad recibe entradas únicamente de las unidades de la capa que la precede inmediatamente. Una red con todas las entradas conectadas directamente a las salidas se denomina red neuronal de una sola capa, o red perceptrón.

Las redes con alimentación hacia delante que poseen una o varias capas ocultas, se denominan redes neuronales multicapa. La ventaja de añadir capas ocultas es que amplía el espacio de hipótesis que puede representar la red (Russell & Norvig, 2004).

2.6 El aprendizaje en las redes neuronales artificiales

Los cambios en los pesos de una red neuronal dan lugar al aprendizaje. Estos se producen para modelar los cambios en el rendimiento de las uniones funcionales de las neuronas. Hay dos tipos principales de aprendizaje en RNA (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004):

Aprendizaje supervisado. Con este tipo de aprendizaje, proporcionamos a la red un conjunto de datos de entrada y la respuesta correcta. El conjunto de datos de entrada es propagado hacia delante hasta que la activación alcanza las neuronas de la capa de salida. Entonces podemos comparar la respuesta calculada por la red con aquella que se desea obtener, el valor real u objetivo. Entonces se ajustan los pesos para asegurar que la red produzca de una manera más probable una respuesta correcta en el caso de que se vuelva a presentar el mismo o similar patrón de entrada. Este tipo de aprendizaje será útil especialmente para las tareas de regresión y clasificación.

Aprendizaje no supervisado. Solo se proporciona a la red un conjunto de datos de entrada. La red debe auto-organizarse (es decir, auto enseñarse) dependiendo de algún tipo de estructura existente en el conjunto de datos de entrada. Típicamente esta estructura suele deberse a redundancia o agrupamiento en el conjunto de datos. Este tipo de aprendizaje es útil para las tareas de agrupamiento y reducción de dimensionalidad.

2.7 Series temporales

Una serie temporal es un conjunto de observaciones referidas a una magnitud y ordenadas en el tiempo. Sea X la magnitud bajo estudio y sea x_t el valor de la magnitud X en el instante de tiempo t . Entonces el conjunto $\{x_t\}_{t=1,\dots,T}$: $\{x_1, \dots, x_T\}$ es una serie temporal.

Las series temporales pueden ser analizadas con una finalidad descriptiva, si solo se pretende describir el comportamiento registrado en el pasado, explicativa, si se intenta probar estadísticamente la existencia de relaciones dinámicas causa-efecto entre variables, o predictiva, cuando el objeto es reducir el grado de incertidumbre sobre el futuro a partir del conocimiento del pasado. Para comprender el modo en que una serie temporal puede ser útil en alguno de estos sentidos, conviene, en primer lugar, identificar los componentes que dan lugar a los valores observados en el tiempo, y en segundo lugar, considerar diferentes modelos tanto de cada uno de dichos componentes como del modo en que éstos interactúan (Cáceres Hernández, 2006).

Una serie temporal normalmente se descompone en cuatro elementos o movimientos principales (Han, Kamber, & Pei, 2001):

- Movimientos a largo plazo o tendencias: estos movimientos indican el comportamiento general de la serie en un periodo largo de tiempo. Ayudan a identificar cual es la tendencia que sigue o ha seguido la serie. Por ejemplo, podemos ver que en tiempos de bonanza económica existe una tendencia al alza en el número de venta de coches.
- Variaciones cíclicas: representan ciclos que presentan las series. Estas variaciones cíclicas pueden o no ser periódicas. Es decir, los ciclos pueden no ser completamente iguales después de periodos idénticos. Un ejemplo de estos ciclos son las oscilaciones que presentan las cotizaciones en bolsas de los valores de una compañía. Normalmente, cuando se conoce una buena noticia se produce un inmediato incremento en el valor de las acciones, que con el tiempo se convierte en una bajada por la recogida de beneficios.
- Movimientos estacionales: estos movimientos se deben a eventos que ocurren con una frecuencia establecida y constante. Por ejemplo, el número de juguetes vendidos se incrementan considerablemente en el período de navidad, o el número de ingresos en urgencias es siempre más alto las noches de los fines de semana que las noches entre semana.
- Movimientos aleatorios o irregulares: estos movimientos representan el comportamiento de la serie debido a eventos aleatorios o semi-aleatorios. Por ejemplo, tormentas fuertes, atentados terroristas, irrupción de un virus desconocido, etc.

2.7.1 Extracción de patrones secuenciales

La minería de datos secuenciales se define como la extracción de patrones frecuentes relacionados con el tiempo u otro tipo de secuencia. A diferencia de las series temporales, el momento preciso no es tan relevante, sino que estos eventos se producen secuencialmente. Además, el objetivo no suele ser cómo seguirá una serie sino analizar muchos individuos que tienen un comportamiento secuencial. Un ejemplo de patrón secuencial podría ser “un cliente que se compra una bicicleta, probablemente se comprará un casco en menos de tres meses”. Gran cantidad de datos pueden considerarse como datos secuenciales: transacciones comerciales, acceso a páginas web por un cliente, recorrido de un cliente por las secciones de un supermercado, etc (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004).

2.7.2 Clasificación con datos secuenciales y ventanas deslizantes

El problema de clasificación secuencial puede ser formulado de la siguiente manera. Sea $\{(x_i, y_i)\}$, $1 \leq i \leq N$ un conjunto de N ejemplos. Cada ejemplo es un par de secuencias (x_i, y_i) , donde $x_i = \langle x_{i,1}, x_{i,2}, x_{i,3}, \dots, x_{i,T_i} \rangle$, y $y_i = \langle y_{i,1}, y_{i,2}, y_{i,3}, \dots, y_{i,T_i} \rangle$. Por ejemplo, el caso del reconocimiento de la estructura de frases de acuerdo con el tipo de las palabras. En castellano es bastante extraño encontrar una frase con la siguiente estructura (adjetivo adjetivo verbo verbo). En este caso del reconocimiento de la estructura de una frase, podríamos tener $x_i = \langle \text{Juan come naranjas dulces} \rangle$, y $y_i = \langle \text{nombre verbo nombre adjetivo} \rangle$. El objetivo del aprendizaje supervisado secuencial es la inducción de un modelo h que pueda predecir una secuencia “clase” $y = h(x)$ dada una secuencia de entrada x . Es decir, la clase no es un atributo, sino una secuencia de atributos (Dietterich, 2002).

Una primera técnica que se ha utilizado es la denominada de ventanas deslizantes. Este método permite convertir un problema de aprendizaje supervisado secuencial en un problema clásico de aprendizaje supervisado. Para ello, se utiliza un clasificador de ventana h_w que asigna a una ventana de entrada una salida individual y_i . Un posible ejemplo de ventana de tamaño w sería

$< x_{i,t-d}, x_{i,t-d+1}, \dots, x_{i,t}, \dots, x_{i,t+d-1}, x_{i,t+d} >$ donde $d = (\frac{w-1}{2})$. A partir de esa ventana h_v predeciría el valor y_i . El clasificador ventana se entrena convirtiendo cada ejemplo secuencial (x_i, y_i) en ventanas para después poder aplicar un algoritmo de aprendizaje supervisado. Finalmente, para clasificar una secuencia nueva x' , se divide esta secuencia en ventanas, se aplica el clasificador ventana produciendo un valor y_i' por cada ventana, y por último se concatenan estos valores para formar la secuencia de salida y' . Otra técnica similar se basa en utilizar ventanas recurrentes. En una ventana recurrente, el valor predicho en $y_{i,t}$ se utiliza también para realizar la precisión de $y_{i,t+1}$. Otras aproximaciones se basan en la utilización de modelos ocultos de Markov. Estos modelos representan métodos probabilísticos de cómo se generan las secuencias x_i e y_i (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004).

2.7.3 Pronóstico de series temporales con redes neuronales artificiales

Recientemente, las redes neuronales artificiales se han utilizado en la mayoría de aplicaciones de predicción de series de tiempo. Los siguientes factores conducen el uso generalizado de las redes neuronales artificiales como método de análisis de series temporales (Hakan Aladag & Egrioglu, 2012):

- El análisis puede hacerse independientemente si las series de tiempo tienen estructura curvilínea o lineal.
- Cuando se compara con los métodos convencionales de predicción de series temporales, el método de la red neuronal artificial proporciona los mejores resultados.
- Los modelos de series de tiempo convencionales sólo pueden ser utilizadas para ciertas estructuras curvilíneas, no son flexibles en general. En cambio utilizando la red neuronal artificial, no importa que estructura curvilínea presente la serie temporal.
- La teoría de redes neuronales artificiales no es compleja como en los métodos convencionales de pronóstico de series de tiempo, y es fácil de entender.

El uso de redes neuronales artificiales con alimentación hacia delante y redes neuronales con retroalimentación en predicción de series temporales se pueden resumir en siete pasos.

Paso 1. Selección de la función de activación y pre-procesamiento de los datos

Inicialmente se selecciona el tipo de función de activación para ser utilizado en la capa oculta y de salida de la red neuronal artificial. Una de las funciones de activación utilizada es la función de activación logística. La función de activación logística se da como sigue:

$$f(x) = \frac{1}{1 + \exp(-x)}$$

Los datos se convierten en el rango adecuado para la función de activación. Si es que se utilizará la función de activación logística, para mostrar el valor de entrada x_i se puede convertir al rango (0,1) como,

$$x'_i = \frac{x_i - \text{Min}(x_i)}{\text{Max}(x_i) - \text{Min}(x_i)}$$

donde, $\text{Max}(x_i)$ representa el máximo valor de entrada y $\text{Min}(x_i)$ representa el valor de entrada mínimo.

Paso 2. Organización de datos

En este paso, se determina la cantidad de datos de entrenamiento y la cantidad de datos de prueba del conjunto de datos.

Paso 3. Modelado

El modelo de redes neuronales artificiales se establece mediante la determinación del número de nodos de entrada, el número de capas ocultas, el número de nodos en las capas ocultas, el número de nodos en la capa de salida, la función de activación usado en la unidad, el algoritmo de entrenamiento y los parámetros del algoritmo y el criterio de desempeño.

Paso 4. Generación de valores de entrada

Los valores de entrada de la red neuronal artificial son retrasos de series de tiempo. Sea m , el número de unidades en la capa de entrada, entonces para la serie de tiempo X_t el retraso será generado como $X_{t-1}, X_{t-2}, \dots, X_{t-m}$.

Paso 5. Cálculo de los mejores valores de peso

Los mejores valores de peso se encuentran en el entrenamiento establecido a través del algoritmo de aprendizaje seleccionado. Los valores de salida de la red neuronal artificial son calculados.

Paso 6. Cálculo de los criterios de rendimiento

Se obtienen las predicciones de la red neuronal artificial sobre el conjunto de prueba. El modelo obtenido en los pasos anteriores es aplicado al conjunto de prueba. Los valores resultantes constituyen predicciones del conjunto de entrenamiento y de prueba establecido respectivamente. Sobre la base de las diferencias entre el pronóstico de un conjunto de pruebas y los datos de prueba, se calcula el criterio de rendimiento seleccionado. Hay varios criterios de rendimiento utilizados para determinar que tan bien la red neuronal artificial aprendió la relación en los datos. Uno de los criterios de rendimiento más utilizadas es la raíz del error cuadrático medio (RMSE). Sea z_p el valor de salida de la red neuronal artificial, d_p valor objetivo de salida y p , el crecimiento del conjunto de prueba, RMSE se puede calcular con la siguiente ecuación:

$$RMSE = \sqrt{\frac{\sum_{p=1}^P (d_p - z_p)^2}{p}}$$

Paso 7. Predicción

Por último, se obtienen los valores de pronóstico para el futuro usando los mejores valores de los pesos obtenidos en el paso 5.

2.8 Resumen

El proceso de extracción del conocimiento tiene como objetivo identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos. Este proceso cuenta con distintas etapas, de las cuales la más importante es la de minería de datos ya que aquí se define el tipo de tarea a realizar, el algoritmo a utilizar y se elabora el modelo de datos. Uno de los algoritmos que más se utilizan para realizar predicciones son las redes neuronales. Las redes neuronales son técnicas de aprendizaje y procesamiento de datos inspiradas en la forma que funciona el sistema nervioso de los animales. En el caso de las tareas de predicciones, es importante también la utilización de series de tiempo si los datos representan observaciones tomadas en secuencia a intervalos regulares de tiempo.

3 Comprensión del negocio

La primera fase de la metodología CRISP-DM tiene como objetivo enfocarse en los propósitos y requerimientos del proyecto desde la perspectiva del negocio. De esta manera es necesario centrarse en la problemática del proyecto, identificar los aspectos que influirán en el mismo y elaborar el plan del proyecto.

3.1 Determinación de objetivos del proyecto

El objetivo de esta sección es analizar la problemática de la inflación, presentar algunos ejemplos del análisis de este fenómeno, proponer una solución y definir los criterios de éxito de la misma.

3.1.1 Contexto

“La inflación es un proceso de suba continuada en el nivel general de precios de la economía” (Díaz, 2011). Es un proceso continuado porque ocurre durante un intervalo de tiempo definido. Lo que sube es el nivel general de precios de una economía, no un subconjunto de los bienes o servicios de una economía.

En la Argentina, la inflación es uno de los problemas de la economía de los últimos años. El mayor perjuicio de este problema radica en la pérdida de poder adquisitivo de las personas. Al mantenerse constante los ingresos, si los precios de los bienes o servicios aumentan, disminuye la capacidad de compra.

Otro problema que genera la inflación es la imposibilidad de realizar proyecciones de la economía a mediano y largo plazo. Lo cual repercute de manera directa en las inversiones realizadas en el país.

A continuación se presentan dos teorías que explican cómo a partir de otras variables macroeconómicas puede explicarse el fenómeno de la inflación.

3.1.1.1 La teoría cuantitativa del dinero (Fernández Arufe, 2006)

La teoría cuantitativa del dinero permite analizar relaciones entre la tasa de crecimiento del dinero, la tasa de inflación y la tasa de crecimiento de la renta real.

La siguiente identidad permite explicar la teoría cuantitativa del dinero:

$$V = T / M$$

siendo: V = la velocidad de circulación del dinero (entendida como el número de veces que cada unidad monetaria es utilizada para realizar transacciones en el período considerado), M = la cantidad de dinero (en pesos) y T = el flujo de las transacciones reales de bienes y servicios (en pesos).

Por otro lado, se define la renta nominal $\tilde{Y}$ como el producto del nivel de precios P y el nivel de producción dado de pleno empleo, Y (Dornbusch, 1981):

$$\tilde{Y} = P \cdot Y$$

El flujo de las transacciones T puede ser estimado a través de la renta nominal, ya que esta última constituye el valor de la producción en pesos durante un tiempo determinado. Utilizando a la renta nominal $\tilde{Y}$ como una aproximación de las transacciones T , se obtiene:

$$V = (P \cdot Y) / M$$

Si V , velocidad de circulación del dinero, permanece constante, las variaciones en la oferta monetaria M se traducen en variaciones proporcionales de la renta nominal. Dado que la renta nominal se define en función del nivel general de precios P y el nivel de producción Y ; si se considera que en el corto plazo el nivel de producción Y es fijo, la Teoría Cuantitativa del Dinero implica que el nivel de precios es proporcional a la cantidad de dinero M . Las variaciones del dinero M se traducen en variaciones del nivel general de precios P .

3.1.1.2 La curva de Philips a corto plazo (Krugman & Wells, 2007)

La curva de Philips a corto plazo es la relación inversa que se establece a corto plazo entre la tasa de desempleo y la tasa de inflación.

La explicación de la relación inversa entre la tasa de desempleo y la tasa de inflación se debe a que cuando aumenta el nivel general de precios, esto provoca un incremento del PBI real. Además, entre el PBI real y la tasa de desempleo se establece una relación inversa: la ley de Okun establece que cuando el PBI real es superior a la producción potencial, la tasa de desempleo es inferior a la que alcanzaría si el PBI real estuviera por debajo de la producción potencial. De modo que un aumento del nivel de precios está asociado a un aumento del PBI real, lo que tiende a disminuir la tasa de desempleo.

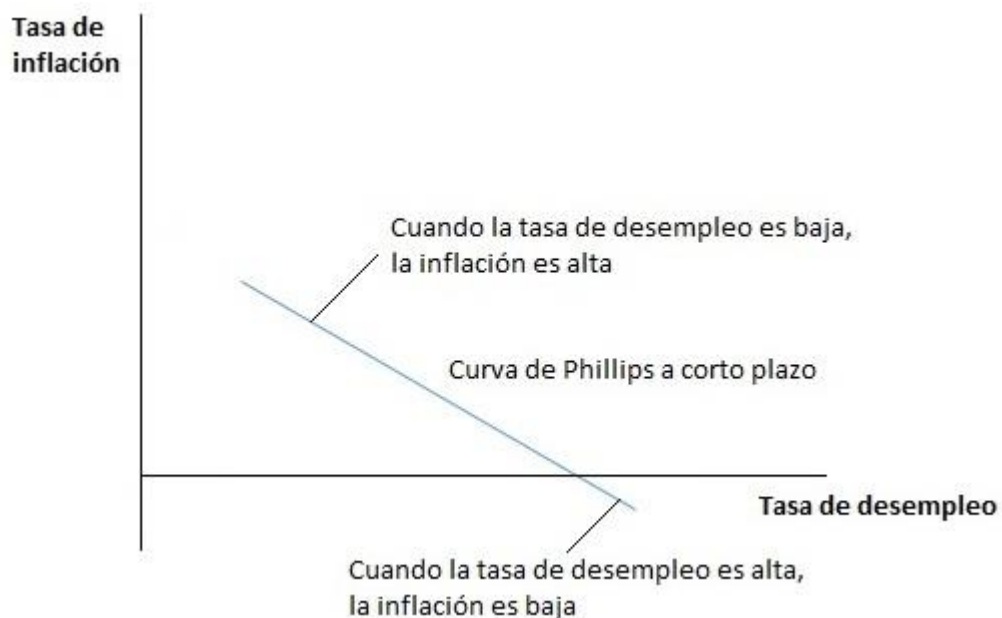


Ilustración 5. Curva de Philips a corto plazo

3.1.1.3 Cálculo de la inflación en la Argentina

En esta sección se presenta el cálculo de la inflación en la Argentina a través del IPC (Índice de precios al consumidor). La información obtenida pertenece a las siguientes fuentes: (INDEC, 2002), (INDEC, 2008;2009; 2010; 2011; 2012; 2013; 2014; 2015), (INDEC, 2014).

¿Qué es un Índice de Precios?

Un Índice de Precios es un indicador que tiene por objeto medir las variaciones, a través del tiempo, en los precios de un conjunto definido de bienes y servicios. El INDEC, Instituto Nacional de Estadísticas y Censos, elabora el índice de precios al consumidor, que mide la variación promedio de los precios minoristas de un conjunto de bienes y servicios que representan el consumo de los hogares en un período específico.

¿Cómo se estructura un Número Índice de Precios?

Se estructura en un Nivel General y en subíndices de menor nivel de agregación. En todos los casos, se refiere a:

- Un período base, generalmente el año en que se determina la estructura de ponderaciones del índice teniendo en cuenta la importancia relativa de cada uno de los bienes y/o servicios que incluye esa estructura. Para dar cuenta de las variaciones de los precios, se le asigna al índice del año base el número 100. Esto significa que cada índice mensual expresará la relación entre los precios relevados ese mes y los promedios vigentes en el año base.
- Una población de referencia.
- Una región geográfica definida.

¿Por qué cambian las bases de los Índices y qué significa ‘empalmar’ una serie?

Con el transcurso del tiempo, el conjunto de bienes y servicios considerados en los índices de precios pueden ir perdiendo representatividad. Los hogares van cambiando su estructura de consumo: dejan de consumir determinados bienes o servicios o los reemplazan por otros; los productores también modifican el tipo de bienes que ofrecen en el mercado; se presentan cambios en las características de las viviendas que se construyen y en las técnicas de construcción aplicadas, etc. En el curso de esos cambios, los números índices van perdiendo su capacidad para representar la realidad y se vuelve necesario modificar su base evaluando la introducción de transformaciones en alguno/s de los siguientes niveles:

- Los bienes y/o servicios que lo integran y su importancia relativa.
- La población de referencia.
- La cobertura geográfica.
- El sistema de relevamiento de precios.
- Las fórmulas de cálculo.

Al modificar la base de un índice se produce una ruptura en la continuidad de la serie, que desde el punto de vista teórico no admite solución cuando la modificación responde a alguno de los niveles citados. La “ruptura” significa que el nuevo índice de precios posee una representatividad cualitativamente diferente del anterior.

No obstante, como muchas veces es necesario contar con series continuas que permitan medir la variación de precios en períodos en donde el índice cambia de base, el INDEC aplica un procedimiento que se denomina “empalme” y que permite unir ambas series.

Los empalmes se desarrollan, habitualmente, para el máximo nivel de agregación de los índices, es decir para el Nivel General y sus principales aperturas

Reseña de los índices de precios al consumidor utilizados

Entre Abril del 2008 y Enero de 2014 el índice de precios al consumidor empleado trabajaba con una base mensual equivalente al índice de precios de Abril de 2008. Esta serie se interrumpió en febrero de 2014 por el cambio del año base del indicador y su actualización metodológica.

Desde Febrero 2014 el INDEC difunde el Índice de Precios al Consumidor Nacional urbano (IPCNu) con base IV trimestre 2013=100, que constituye un avance comparado con el anterior IPC-GBA que incluía los consumos de familias del Área Metropolitana de Buenos Aires

(AMBA). El IPCNu se basa en considerar una canasta de bienes y servicios que representa el comportamiento de los consumidores en otras regiones de Argentina, es decir, tiene cobertura nacional.

A partir del mes de octubre de 2014 el INDEC publicó el Índice de Precios al Consumidor Nacional urbano base octubre 2013 – septiembre 2014 = 100 (IPCNu oct13-sep14=100). Este indicador es continuación del IPCNu, cambiando de una base de un trimestre a una base de 12 meses.

3.1.2 Casos de éxito

A continuación se mencionarán dos casos de aplicación de Redes Neuronales al proceso inflacionario.

3.1.2.1 Aplicación de Redes Neuronales Artificiales para el Análisis de la Inflación en Bolivia (Urquiza Rojas & MENDIVIL GOMEZ, 2011)

En un trabajo de investigación desarrollado en Bolivia, se aplicó las Redes Neuronales Artificiales para estudiar empíricamente a la inflación. Para realizar esta aplicación se tomaron en cuenta una serie de variables como ser la emisión monetaria, variación del tipo de cambio nominal de compra, el índice global de actividad económica (IGAE) y dos variables de rezago para la inflación.

La conclusión obtenida al final del proyecto fue que las Redes Neuronales Artificiales representan una alternativa para obtener resultados confiables debido a que se pueden obtener relaciones de tipo no lineal entre las variables.

3.1.2.2 Predicción de series temporales con redes neuronales: una aplicación a la inflación colombiana (Santana, 2006)

La idea principal del trabajo, desarrollado en Colombia, fue implementar series temporales con redes neuronales para el análisis de la inflación, como una técnica alternativa a las utilizadas hasta ese entonces. Las técnicas implementadas anteriormente para este tipo de estudio fueron el método SARIMA y el suavizamiento exponencial.

Los resultados obtenidos del proyecto fueron que las predicciones obtenidas a través de redes neuronales tienden a ser más precisas que el método SARIMA y el suavizamiento exponencial.

3.1.3 Solución propuesta

El presente proyecto propone analizar la inflación aplicando algoritmos de Redes Neuronales para obtener predicciones en base a distintas variables de la economía. Para aplicar este algoritmo de predicción, el proyecto se adecuará a la metodología CRISP-DM, en el marco de un proceso de minería de datos. En vista que el proyecto no se aplica a una organización muchas de las fases y tareas de CRISP-DM se adaptaran al proyecto. Por otro lado el proyecto se gestionara siguiendo los fundamentos para la dirección del proyecto propuesto por PMI (Project Management Institute).

3.1.4 Objetivos del proyecto

El presente proyecto de investigación tiene como objetivo general analizar el proceso inflacionario a través de la aplicación de minería de datos.

Los objetivos específicos del proyecto son:

- Determinar variables económicas vinculadas con el aumento de precios.
- Aplicar técnicas de predicción, Redes Neuronales, para predecir valores de la inflación a través de variables económicas asociadas con el aumento de precios.

3.1.5 Criterios de éxito del proyecto

Los criterios de éxito del presente proyecto son:

- Comprender y aplicar los conceptos de Minería de Datos.
- Generar modelos de predicción con la mayor precisión posible.
- Predecir valores a corto plazo de la inflación, con los modelos generados.
- Comparar los modelos de predicción de Redes Neuronales con otros algoritmos de predicción.

3.2 Evaluación de la situación

Evaluar la situación es una actividad que tiene como objetivo determinar diversos factores que influirán en el desarrollo del proyecto. Para realizar un proyecto de minería de datos es indispensable disponer de datos, además es importante contar con conocimientos técnicos, relacionados al procesamiento de datos, y del negocio al que se aplicará minería de datos. Otros aspectos que influyen en la mayoría de los proyectos también serán detallados.

3.2.1 Fuentes de datos y conocimiento

Las fuentes de datos para el proyecto serán entidades públicas. Los organismos públicos a los que se recurrirá para obtener los datos son el BCRA (Banco Central de la República Argentina) y el INDEC (Instituto Nacional de Estadísticas y Censos). El Banco Central de la República Argentina es el organismo rector del sistema financiero de la Argentina, encargado de la política monetaria del país, mientras que el INDEC es un organismo público que depende del Ministerio de Economía y Finanzas Públicas de la Nación y ejerce la dirección superior de actividades estadísticas que se realizan en el territorio de la República Argentina.

Por otro lado, las fuentes de conocimiento partirán del juicio de expertos para desarrollar las actividades del proyecto. El juicio de expertos es el conocimiento y/o experiencia de diferentes disciplinas de una persona o un grupo de individuos.

3.2.2 Requerimientos, presunciones y restricciones

3.2.2.1 Requerimientos

- Elaborar informe final de los resultados obtenidos por la aplicación de las Redes Neuronales a la inflación.
- Gestionar el alcance del proyecto. Realizar las actividades necesarias para garantizar que el proyecto incluya todo el trabajo necesario para completarlo con éxito.
- Gestionar el tiempo del proyecto. Realizar la planificación y ejecución adecuada de las tareas para asegurar que el proyecto finalice en el plazo definido.
- Gestionar los costos del proyecto. Estimar, presupuestar y controlar los costos del proyecto.

- Gestionar la calidad del proyecto. Ejecutar y controlar las actividades tendientes a acondicionar los datos para su análisis.
- Gestionar los recursos humanos del proyecto. Adquirir, desarrollar y dirigir el equipo del proyecto.
- Gestionar las comunicaciones del proyecto. Generar una comunicación eficiente entre los diferentes interesados del proyecto.
- Gestionar los riesgos del proyecto. Identificar, analizar, planificar las respuestas a los riesgos, ejecutar las respuestas y controlarlas.
- Gestionar las adquisiciones del proyecto. Efectuar y administrar las adquisiciones de productos, servicios o resultados que es necesario obtener fuera del equipo del proyecto.

3.2.2.2 Presunciones

- Los datos económicos para el proyecto de investigación serán referidos a los años 2008, 2009, 2010, 2011, 2012, 2013, 2014 y 2015.
- Las principales fuentes de información de datos serán el INDEC (Instituto Nacional de Estadísticas y Censos) y el BCRA (Banco Central de la República Argentina).

3.2.2.3 Restricciones

- El presente proyecto deberá realizarse en el equipamiento hardware disponible. Las características del hardware existente son las siguientes:
 - CPU con procesador 2.5 GHz
 - 4 Gb de Memoria RAM
 - 750 Gb de Disco duro
- Los formatos de archivos de los datos serán plantillas de software disponible o archivos de texto (.txt).
- El proyecto no deberá extenderse a un plazo mayor de dos años. Según el Reglamento del Alumno vigente de la facultad, el alumno dispone de un plazo máximo de 2 (dos) años para finalizar el proyecto.
- El proyecto se realizara únicamente en el horario por la mañana, desde 8:00 am a 12:00.
- Recursos humanos disponibles. El proyecto será realizado por una persona quien se encargara de cumplir con los distintos roles identificados (véase Sección 3.3.4).

3.3 Plan de proyecto

Este apartado contiene la planificación del proyecto. Para elaborar la planificación se partió de una descripción del proyecto a realizar, luego en base a esto se definieron los paquetes de trabajo y las actividades necesarias para alcanzar los objetivos del proyecto. A continuación se identificaron los recursos requeridos y se estimó la duración temporal para cada una de las actividades definidas. Finalmente, el registro de riesgo contiene el plan de respuesta a los riesgos identificados.

3.3.1 Alcance

El presente proyecto de grado tiene como fin estudiar las redes neuronales y su aplicación al análisis de la inflación en la economía argentina.

El proyecto se realizará siguiendo un subconjunto de fundamentos establecidos por PMI para la formulación y dirección de proyectos. La metodología a seguir para la aplicación de las Redes Neuronales será la de CRISP-DM. Como el proyecto no se aplica a una organización en particular, muchas de las fases y tareas de la metodología se adaptarán de manera conveniente.

El entregable definitivo será el informe generado como resultado de la aplicación del proceso de minería de datos al análisis de la inflación.

El objetivo, los criterios de éxito, como así también los requerimientos, restricciones y presunciones fueron definidas previamente en “Objetivos del proyecto” y “Evaluación de la situación actual”.

3.3.2 Estructura de desglose del trabajo (EDT)

La estructura de desglose del trabajo se definió utilizando la metodología CRISP-DM. Esta metodología se estructura en seis fases: Comprensión del negocio, Comprensión de los datos, Preparación de los datos, Modelado, Evaluación e Implantación. La EDT es una estructura jerárquica que contiene todos los paquetes de trabajo que deben realizarse para completar el proyecto.

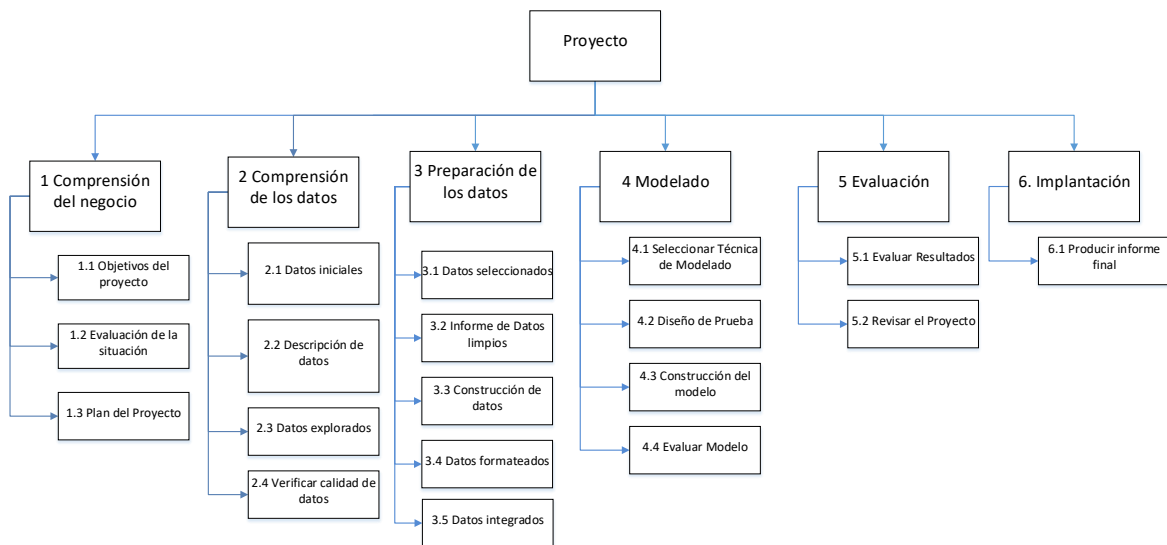


Ilustración 6. Estructura de desglose de trabajo

3.3.2.1 Diccionario de la EDT

A continuación se presenta el diccionario de la EDT, su objetivo es proporcionar una descripción de los paquetes de trabajo definidos en la estructura de desglose de trabajo.

- 1 Comprensión del negocio. Esta fase inicial se enfoca en la comprensión de los objetivos de proyecto y exigencias desde una perspectiva de negocio, luego convirtiendo este conocimiento de los datos en la definición de un problema de minería de datos y en un plan preliminar diseñado para alcanzar los objetivos.
 - 1.1 Objetivos del proyecto. Determinar el objetivo general del proyecto, y sus objetivos específicos.
 - 1.2 Evaluación de la situación. Investigación más detallada sobre recursos, restricciones, presunciones y otros factores a ser considerados en el desarrollo del proyecto.
 - 1.3 Plan del Proyecto. Elaboración de un plan de acción para alcanzar objetivos de negocio y de aplicación de minería de datos.
- 2 Comprensión de los datos. La fase de entendimiento de datos comienza con la colección de datos inicial y continua con las actividades que le permiten familiarizar primero con los datos, identificar los problemas de calidad de datos, descubrir los primeros conocimientos en los datos, y/o descubrir subconjuntos interesantes para formar hipótesis en cuanto a la información oculta.

- 2.1 Datos iniciales. Recolección de datos iniciales.
- 2.2 Descripción de datos. Examinar propiedades de los datos.
- 2.3 Datos explorados. Conocer los datos para tomar decisiones adecuadas al crear los modelos.
- 2.4 Verificar calidad de los datos. Determinar si los datos están completos y son correctos.
- 3 Preparación de los datos. La fase de preparación de datos cubre todas las actividades necesarias para construir el conjunto de datos final (los datos que serán provistos en las herramientas de modelado) de los datos en brutos iniciales.
 - 3.1 Datos seleccionados. Seleccionar los datos que se utilizarán en el modelo.
 - 3.2 Informe de datos limpios. Selección de datos limpios, para elevar la calidad de los datos al nivel requerido para la aplicación del modelo.
 - 3.3 Construcción de datos. Operaciones de preparación de los datos como atributos derivados o ingreso de nuevos registros.
 - 3.4 Datos formateados. Modificaciones sintácticas, requerida por la herramienta de modelado, hechas a los datos
 - 3.5 Datos integrados. Creación de nuevos registros a partir de múltiples tablas o registros.
- 4 Modelado. En esta fase varias se aplica una técnica de modelado y sus parámetros son calibrados a valores óptimos.
 - 4.1 Seleccionar técnicas de modelado.
 - 4.2 Diseño de prueba. Elaboración e implementación de un plan de entrenamiento del modelo.
 - 4.3 Construcción del modelo. Ejecutar la herramienta de modelado sobre el conjunto de datos preparados para crear un modelo.
 - 4.4 Evaluar Modelo. Evaluar el modelo según objetivos del negocio y metas de minería de datos.
- 5 Evaluación. Consiste en evaluar y revisar los pasos efectuados para comparar el modelo obtenido con los objetivos de negocio.
 - 5.1 Evaluar resultados. Evaluar el grado al que el modelo responde a los objetivos de negocio.
 - 5.2 Revisar el proyecto. Evaluar lo que fue correcto y lo que se equivocó, lo que fue bien hecho y lo que necesita para ser mejorado.
- 6 Implantación. La última fase del proyecto consiste en la elaboración de un informe final con los resultados obtenidos de la aplicación de minería de datos.
 - 6.1 Producir informe final. Informe con un resumen del proyecto y los resultados obtenidos.

3.3.3 Lista de actividades

Las siguientes actividades representan el esfuerzo necesario para realizar los paquetes de trabajos definidos en la EDT.

Fase 1: Comprensión del negocio

- Determinar los objetivos del proyecto
- Evaluar la situación actual
- Producir el plan del proyecto

Fase 2: Comprensión de los datos

- Recolectar datos iniciales
- Describir los datos
- Explorar los datos
- Verificar la calidad de los datos

Fase 3: Preparación de los datos

- Seleccionar datos
- Limpiar datos
- Construir datos
- Formatear datos
- Integrar datos
- Preparar datos¹

Fase 4: Modelado

- Escoger técnica de modelado
- Generar la prueba de diseño
- Construir el modelo
- Evaluar el modelo
- Elaborar modelo²

Fase 5: Evaluación

- Evaluar los resultados
- Revisar el proyecto

Fase 6: Implantación

- Producir informe final

3.3.4 Requisitos de recursos de la actividad

Una vez definidas las actividades para desarrollar el proyecto, es necesario determinar los tipos de recursos humanos, materiales o equipos y la cantidad de los recursos necesarios. Este paso es importante desarrollarlo para poder realizar las estimaciones de las duraciones de cada actividad y los costos asociados al proyecto.

A continuación se identifican el tipo y la cantidad de los recursos necesarios para las diferentes actividades.

¹ Implica realizar en una nueva iteración las tareas de preparación de los datos: seleccionar, limpiar, construir, integrar, formatear y preparar datos.

² Implica realizar en una nueva iteración las tareas de preparación de los datos: escoger técnica de modelado, generar la prueba de diseño, construir el modelo y evaluar el modelo.

Actividad	Recurso	Cantidad
Determinar los objetivos del proyecto	Director de proyecto	1
	Computadora	1
Evaluar la situación actual	Director de proyecto	1
	Computadora	1
Producir el plan del proyecto	Director de proyecto	1
	Computadora	1
Recolectar datos iniciales	Analista de datos	1
	Analista económico – financiero	1
	Computadora	1
Describir los datos	Analista de datos	1
	Analista económico – financiero	1
	Computadora	1
Explorar los datos	Analista de datos	1
	Estadístico	1
	Computadora	1
Verificar la calidad de los datos	Analista de datos	1
	Estadístico	1
	Computadora	1
Seleccionar datos	Analista de datos	1
	Computadora	1
Limpiar datos	Analista de datos	1
	Computadora	1
Construir datos	Analista de datos	1
	Computadora	1
Formatear datos	Analista de datos	1
	Computadora	1
Integrar datos	Analista de datos	1
	Computadora	1
Preparar datos	Analista de datos	1
	Computadora	1
Escoger técnica de modelado	Analista de datos	1
	Computadora	1
Generar la prueba de diseño	Analista de datos	1
	Computadora	1
Construir el modelo	Analista de datos	1
	Computadora	1
Evaluar el modelo	Analista de datos	1
	Analista económico – financiero	1
	Estadístico	1
Elaborar el modelo	Analista de datos	1
	Computadora	1
Evaluar los resultados	Analista de datos	1
	Computadora	1
Revisar el proyecto	Director de proyecto	1
	Computadora	1
Producir informe final	Director de proyecto	1
	Computadora	1

Tabla 1. Requisitos de recursos de las actividades

3.3.5 Cronograma del proyecto

A través del juicio de expertos y la información sobre la distribución del esfuerzo en las diferentes etapas de la metodología CRISP-DM se procedió a estimar la duración de las actividades. En el siguiente diagrama de Gantt se representan las actividades con sus fechas de inicio y finalización esperada.

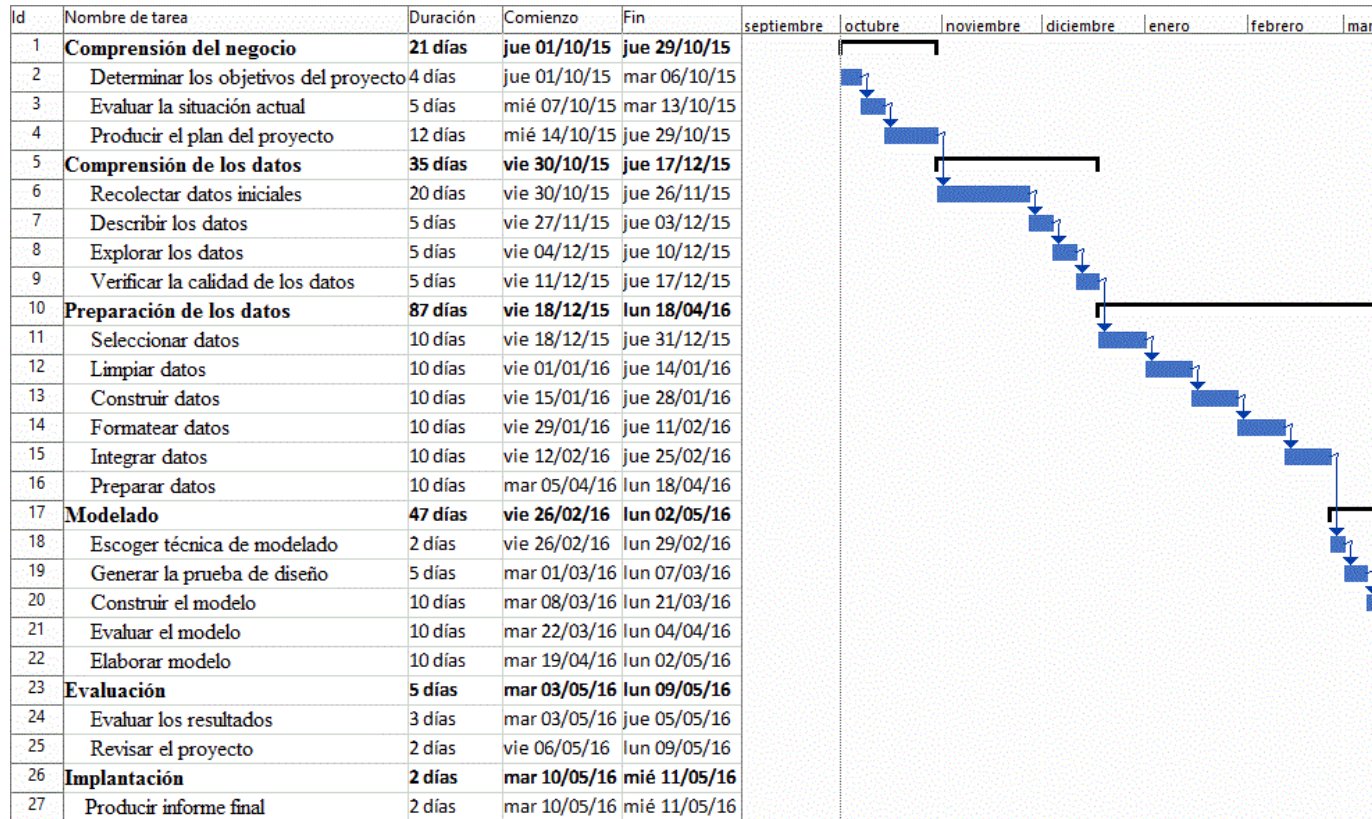


Ilustración 7. Cronograma de actividades

3.3.6 Gestión de costo y adquisiciones

Para realizar la gestión de costo y adquisiciones se aplicó la técnica de la estimación paramétrica. *La estimación paramétrica utiliza una relación entre los datos históricos y otras variables para calcular una estimación de parámetros de una actividad tales como el costo, presupuesto y duración* (PMI, 2008). A través de los valores obtenidos por horas de los honorarios desde distintas fuentes y la utilización de unidades de tiempo se logró obtener estimaciones con un buen nivel de exactitud.

La gestión de costo y adquisiciones comprende los costos vinculados a recursos humanos, los costos de hardware y otros gastos varios del proyecto.

3.3.6.1 Línea base de costos de recursos humanos

La estimación de los costos de recursos humanos se presenta en la siguiente tabla. En ella se identifica el tipo de recurso humano, los honorarios por horas hombre, los días laborales a lo largo del proyecto, la cantidad de horas laborales por día y el costo asociado al recurso.

Recurso	\$ / Hora H	Días	Horas H. / Día	Total Horas H.	Costo
Director del proyecto	\$156,00 ³	25	4	100	\$15.600,00
Analista de datos	\$104,00	135	4	540	\$56.160,00
Analista económico – financiero	\$400,00 ⁴	5	4	20	\$8.000,00
Estadístico	\$95,00	10	4	40	\$3.800,00
TOTAL					\$83.560,00

Tabla 2. Costos de recursos humanos

3.3.6.2 Costos de Hardware

Los costos de hardware hacen referencia a las adquisiciones necesarias a realizar para desarrollar el proyecto. En la siguiente tabla se detallan el hardware necesario, la cantidad de recursos necesarios, su precio unitario y su precio total.

Ítem	Cantidad	Precio unitario	Precio total
Disco externo USB 1 Tb	1	\$899,00	\$899,00
TOTAL			\$899,00

Tabla 3. Costos de hardware

Es importante aclarar que el proyecto se desarrollará utilizando las computadoras disponibles, por eso no se incluyen en el presupuesto de hardware. El equipamiento informático disponible está compuesto por dos computadoras con las características mencionadas en el Apartado 3.2.2.3. El equipo de hardware disponible cuenta con salida a internet, por lo que el servicio de internet tampoco se considerara en los gastos varios.

³ (Consejo Profesional de Agrimensores, Ingenieros y Profesiones Afines, 2015)

⁴ (Consejo Profesional de Ciencias Económicas de Salta, 2015)

3.3.6.3 Gastos varios

A continuación se detallan los gastos varios del proyecto.

Concepto	Cantidad de meses	Costo mensual	Costo
Papelería y útiles de oficina	7	\$55,00	\$385,00
Transporte	7	\$900,00	\$6.300,00
Servicio de telefonía	7	\$300,00	\$2.100,00
TOTAL			\$8.785,00

Tabla 4. Gastos varios

3.3.6.4 Línea base de costos

Finalmente, en la siguiente tabla se indican los costos del proyecto.

Concepto	Costo
Costos de recursos humanos	\$83.560,00
Costos de hardware	\$899,00
Gastos varios	\$8.785,00
TOTAL	\$93.244,00

Tabla 5. Línea base de costos

3.3.7 Descripciones de cargo

A través de un formato tipo texto se documentan las responsabilidades de los miembros del equipo, identificando rol-responsabilidad-autoridad para cada integrante del proyecto. Es importante aclarar que los distintos roles identificados a continuación serán llevados a cabo por una única persona, quien es el autor de proyecto.

Rol	Responsabilidad	Autoridad
Director del proyecto	Realizar la planificación, ejecución y control de las actividades del proyecto.	
Analista de datos	Responsable de desarrollar las tareas de minería de datos, recolectar los datos, comprenderlos, prepararlos, generar el modelo de minería de datos y evaluar el modelo.	Director del proyecto
Analista económico – financiero	Encargado en brindar asesoramiento económico y financiero acerca de las variables económicas del país.	Director del proyecto
Estadístico	Aplicar sus conocimientos para analizar los datos obtenidos y evaluar los modelos generados.	Director del proyecto

Tabla 6. Descripciones de cargo

3.3.8 Registro de riesgos

El principal resultado de la gestión de riesgos es el registro de riesgos, en él se identifican y analizan los riesgos, y se planifica las respuestas a cada uno de ellos. Para identificar los riesgos se utilizó el juicio de expertos y se realizó una revisión de la documentación del proyecto, particularmente los requerimientos presunciones y restricciones. Mientras que para planificar las respuestas a los riesgos identificados se optó por mitigar los riesgos, es decir, reducir la probabilidad o el impacto del riesgo en el proyecto. A continuación se mencionan los riesgos identificados, como sus planes de respuestas.

- Escasos datos económicos de fuentes públicas. Para poder aplicar minería de datos, es necesario contar con una importante cantidad de datos, la imposibilidad de contar con todos los datos necesarios puede perjudicar a los resultados del proyecto. Para garantizar que el proyecto sea viable, será necesario consultar a otras fuentes de datos privadas.
- Fuentes de datos ambiguos. Si se trabajan con datos de fuentes públicas y privadas, pueden existir diferencias en los datos obtenidos, debido a la incidencia política que hay en la economía. Para evitar esto será necesario obtener la totalidad de los datos de una fuente pública o privada. O bien, en caso de considerar ambas alternativas de obtención de datos, será importante que la consultora privada este alineada a la perspectiva política y económica del gobierno.
- Problemas de comprensión de datos económicos. Muchos datos de economía tal vez resulten desconocidos para la realización del proyecto. A fin de mitigar este riesgo, será necesario consultar a especialistas sobre economía.
- Finalización del proyecto en un periodo mayor a dos años. Por cuestiones académicas, el proyecto debe finalizarse antes de los dos años a partir de Diciembre del 2014. El plazo de finalización del proyecto es una restricción que debe cumplirse, para evitar demoras en la finalización del proyecto se realizara un periódico control acerca del estado de avance del mismo, por otro lado se revisaran los criterios utilizados a la hora de elaborar el calendario de actividades.
- Desperfectos en equipamiento informático. El presente proyecto se desarrollara sobre el equipamiento informático disponible, una falla en el software o en el hardware puede perjudicar el proyecto perdiendo datos críticos y haciendo que las actividades no se realicen en el plazo definido. Para evitar esto se realizaran copias periódicas de seguridad de los diferentes documentos del proyecto. Estas copias se almacenaran en dispositivos externos y otros medios de almacenamiento.

3.3.8.1 Matriz de probabilidad e impacto

Luego de identificar y elaborar los planes de respuesta a los riesgos, se procede a realizar el análisis cualitativo de los riesgos. En la tabla a continuación, matriz de probabilidad e impacto, se analiza la probabilidad de ocurrencia de los riesgos como el impacto de los mismos sobre los objetivos del proyecto. La probabilidad de los riesgos se define en el intervalo $[0;1]$. El impacto definido para cada riesgo corresponde a un valor en la escala entre 1 y 5. La valoración representa el producto entre la probabilidad y el impacto de cada riesgo, finalmente la categorización se realiza en función de los valores obtenidos como valoración.

Riesgo	Probabilidad	Impacto	Valoración	Categorización
Problemas de comprensión de datos económicos	0.8	4	3.2	Alta
Escasos datos económicos de fuentes públicas	0.6	3	1.8	Media
Fuentes de datos ambiguos	0.4	2	0.8	Media
Finalización del proyecto en un periodo mayor a dos años	0.2	5	1.0	Baja
Desperfectos en equipamiento informático	0.05	2	0.1	Baja

Tabla 7. Matriz de probabilidad e impacto de riesgos

3.4 Alternativas tecnológicas de solución al problema

El proyecto requiere de un equipo informático con un sistema operativo y un software de minería de datos. En algunos de estos requerimientos es posible optar entre distintas alternativas, en otros casos es importante utilizar los recursos disponibles. El análisis de cada situación y la justificación de la elección de una alternativa están detallados en este apartado.

3.4.1 Alternativas de hardware

El proyecto se realizara utilizando el hardware disponible, por lo tanto las alternativas tecnológicas del hardware no es objeto de análisis. El equipo informático disponible es el apropiado para desarrollar el proyecto, el mismo cuenta con salida a internet y con las siguientes características:

- CPU con procesador 2.5 GHz de dos núcleos
- 4 Gb de Memoria RAM
- 750 Gb de Disco duro

3.4.2 Alternativas de sistemas operativos

A continuación se analizaran las alternativas de sistemas operativos, todas las alternativas de análisis corresponden a versiones de 64 bits para poder aprovechar la memoria RAM, un recurso importante para realizar el proceso de minería de datos.

Microsoft Windows 7. Es la versión más utilizada de los sistemas operativos de Microsoft, no presenta incompatibilidades a nivel de hardware y software, por otra parte es estable y se caracteriza por su facilidad de uso.

Microsoft Windows 8. Es una versión más reciente de los sistemas operativos de Microsoft, cuenta con importantes diferencias a nivel de navegabilidad y funcionalidades de sus antecesores, actualmente a nivel empresarial no fue aceptado, ya que cuenta con incompatibilidades a nivel de hardware y software.

Microsoft Windows 8.1. Es una actualización que representa una mejora de Microsoft Windows 8 debido a sus destacadas funcionalidades de búsqueda y administración de archivos.

Microsoft Windows 10. La última versión de los sistemas operativos de Microsoft, combina las fortalezas de sus dos versiones anteriores, permitiendo la accesibilidad del Windows 7 y lo moderno del Windows 8.

Ubuntu. De los sistemas operativos libre es una de las alternativas más adecuada, su última versión es Ubuntu 16.04. Este sistema operativo se caracteriza por su seguridad, estabilidad y accesibilidad.

3.4.3 Alternativas de software para minería de datos

SQL Server. Su versión más reciente SQL Server 2014 ofrece un rendimiento fiable, rápida obtención de información a partir de cualquier tipo de datos. Además de las soluciones locales, ofrece soluciones en la nube.

Weka. Es una plataforma de software para el aprendizaje automático y la minería de datos, escrito en Java, es software libre y abierto distribuido bajo la licencia GNU-GPL.

RapidMiner. Es una aplicación desarrollada para el análisis y la minería de datos. Permite trabajar mediante el encaminamiento de operadores a través de un entorno gráfico. Su última versión RapidMiner Studio 6.1 es paga, pero la versión RapidMiner Community 5.3 es de distribución libre y código abierto.

R. Es un entorno de software libre para el procesamiento estadístico de datos. Se compila y ejecuta en una amplia variedad de plataformas UNIX, Windows y MacOS. R dispone de una amplia gama de paquetes que pueden ser utilizados para el análisis de datos.

3.4.4 Elección de las alternativas tecnológicas

3.4.4.1 Sistema operativo

Luego de analizar ventajas y desventajas de las alternativas de sistemas operativos, se escogió realizar el trabajo a través de Microsoft Windows 8.1, las razones de esta elección son las siguientes:

Las posibilidades de Microsoft Windows 7 y Microsoft Windows 8 fueron descartadas ya que las versiones más recientes (Windows 8.1 y Windows 10) cuentan con mejoras en materia de seguridad y optimización del funcionamiento del sistema operativo, además de nuevas funcionalidades

Por otro lado, Ubuntu es un sistema operativo de distribución libre y con buenas características de rendimiento, pero para desarrollar el proyecto será necesario cambiar el entorno de trabajo y adaptarse a él, lo cual puede representar un riesgo sobre la planificación temporal del proyecto.

Es importante mencionar que el equipamiento disponible cuenta con licencia de Microsoft Windows 8.1, por lo tanto no será necesario considerar este ítem en los costos del proyecto. La actualización de esta versión a Windows 10 no fue considerada ya que ambos sistemas operativos tienen similares prestaciones en relación a las necesidades del proyecto.

3.4.4.2 Software para minería de datos

Considerando las características de las alternativas para software de minería de datos y contando con experiencia en el manejo de estas herramientas se eligió desarrollar el proyecto con RapidMiner 5.3.015.

La alternativa de Weka fue rápidamente descartada debido a que posee una interfaz gráfica de usuario que no permite una buena usabilidad del sistema, como la posee RapidMiner. Por otro

lado, RapidMiner cuenta con extensiones que permiten incorporar los algoritmos de exploración de Weka.

Si bien SQL Server constituye una alternativa a tener en cuenta, para contar con esta herramienta será necesario adquirir su licencia, lo cual representa un costo más a tener en cuenta en el presupuesto del proyecto.

Finalmente, se descartó la alternativa de R ya que nunca fue utilizada anteriormente y su curva de aprendizaje es de pendiente más elevada que RapidMiner.

Teniendo en cuenta que se escogió por la alternativa de Microsoft Windows 8.1, RapidMiner no presenta incompatibilidades con este sistema operativo, lo que garantiza su correcto funcionamiento.

En el Anexo I se adjunta una encuesta realizada por el sitio KDnuggets sobre las herramientas más utilizadas de minería de datos.

3.5 Análisis costo-beneficio del proyecto

El costo final del proyecto se estima en \$93.244,00 (noventa y tres mil doscientos cuarenta y cuatro pesos) según lo presupuestado. Al tratarse de un proyecto de investigación los beneficios resultan dificultosos cuantificarse monetariamente. Aun así se pueden mencionar los aspectos positivos del desarrollo de este proyecto. Los beneficios se detallan a continuación.

Estimular el desarrollo de proyectos. Muchos proyectos fracasan porque resulta difícil presupuestar los costos debido al problema inflacionario, las predicciones de la inflación pueden ayudar a los gestores de proyectos a determinar sus costos con mayor precisión.

Mejorar la toma de decisiones. Determinar de manera anticipada la inflación ayuda a la toma de decisiones para la gestión de negocios, administración pública y gestión de proyectos para lograr sus objetivos.

Contribuir al análisis de una problemática económica como la inflación a través de técnicas de análisis de datos. Mediante distintas variables de la economía local es posible generar modelos de predicción de la inflación que contribuyan a explicar teorías económicas.

Comprender la metodología CRISP-DM a través de la aplicación en el proyecto al proceso de minería de datos, documentando todas las actividades realizadas. Si bien muchas de las tareas de la metodología se adaptan convenientemente debido a que el proyecto no se aplica a una organización en particular, resultara interesante para otros proyectos de investigación similares la experiencia práctica de esta investigación.

Comprender el funcionamiento y registrar la utilización de la herramienta RapidMiner al proceso de extracción del conocimiento. La documentación paso a paso del trabajo con RapidMiner contribuirá a futuros trabajos de minería de datos como una guía interesante a tener en cuenta.

3.6 Resumen

La inflación es un proceso de suba continuada en el nivel general de precios de la economía. En la Argentina la inflación es uno de los mayores problemas económicos que debe solucionar el gobierno actual. Con el objetivo de analizar y poder estimar la inflación en otros países se desarrollaron trabajos que aplicaron técnicas de minería de datos al estudio de esta problemática. En el presente capítulo se define una propuesta para investigar el aumento en el índice de precios local a través de un proceso de análisis de datos. De esta manera se define el contexto del trabajo a desarrollar, la metodología y recursos a utilizar, las actividades a desarrollar y otros factores que influirán en el proyecto.

4 Comprensión de los datos

A continuación se realiza la comprensión de los datos siguiendo la metodología CRISP-DM. En esta etapa del proyecto, se procede a recolectar los datos de las distintas fuentes, y realizar un análisis preliminar que luego será profundizado en la siguiente etapa. El objetivo de esta etapa es familiarizarse con el conjunto de datos y encontrar en ellos problemas de calidad.

4.1 Recolección de datos iniciales

Los datos extraídos para realizar el análisis de la inflación fueron seleccionados teniendo en cuenta las teorías mencionadas en el marco teórico y además se incluyó datos sobre precios internacionales de materias primas agrícolas más importantes de la Argentina. El periodo al cual pertenecen los datos extraídos está comprendido entre Enero 2008 y Julio de 2015.

Los datos para el análisis fueron extraídos de fuentes de acceso público como el Banco Central de la República Argentina, en adelante BCRA, y el Instituto Nacional de Estadística y Censos (INDEC).

Para obtener los datos del BCRA se accedió a los anexos estadísticos de las siguientes publicaciones:

- El Informe Monetario Mensual: *analiza el comportamiento de las principales variables monetarias y aquellas vinculadas al mercado de capitales, como los medios de pago, el circulante en poder del público, préstamos, depósitos, tasas de interés, reservas (cifras provisionales sujetas a cambio de valuación), tipo de cambio, acciones o títulos públicos.*
- El Panorama Global de Commodities: *brinda información mensual sobre la evolución de los precios internacionales y principales novedades de los mercados de las materias primas agrícolas más relevantes para la economía argentina. Asimismo, ofrece información sobre los mercados conexos, tales como los costos de producción y transporte, los biocombustibles y sobre los posicionamientos financieros en mercados de commodities.*
- Tipo de Cambio Nominal Promedio Mensual (TCNPM). Define una cotización promedio del dólar por cada mes.

Por otro lado, los datos extraídos del INDEC fueron obtenidos de las siguientes secciones:

- Empleo y desempleo. *Presenta los principales indicadores del mercado de trabajo, que permiten hacer un seguimiento de la evolución de las tasas de actividad, empleo, desocupación y subocupación.*
- Agregados macroeconómicos. *Esto incluye estimaciones del Producto Interno Bruto en términos constantes, además de cálculos relacionados con la oferta y demanda agregada y sus componentes. Los indicadores presentados permiten medir la producción de bienes y servicios finales dentro de nuestro territorio. Por el lado de la oferta, se presentan datos del valor agregado que genera cada sector económico, y por el lado de la demanda, datos relacionados con el consumo privado, consumo público, inversión interna bruta fija, exportaciones, importaciones y variación de existencias.*
- Índice de precios al consumidor. *Los índices de precios al consumidor miden la variación de precios de los bienes y servicios representativos del gasto de consumo de los hogares.*

Los informes monetarios mensuales y de panorama global de commodities se encuentran en formato Portable Document Format (PDF), por lo que fue necesaria su carga manual en un archivo de hoja de cálculo de Microsoft Excel. Por otro lado, el tipo de cambio nominal

promedio mensual y los datos extraídos del INDEC fueron obtenidos en hojas de cálculo de Microsoft Excel.

4.2 Descripción de datos

A continuación se presenta el informe de descripción de los datos obtenidos. El análisis se detalla a través de diferentes tablas, en cada una de ellas se indica el rol de cada atributo, su nombre, el tipo de dato, el rango, la cantidad de valores faltantes y algunas estadísticas. Estas últimas dependen del tipo de dato, en el caso de los numéricos se detalla el promedio del conjunto de datos, mientras que con los datos polinominales se mencionan los valores más frecuentes y los menos frecuentes.

Los datos fueron obtenidos a través de la herramienta de minería de datos RapidMiner, mediante un proceso que únicamente cuenta con el operador Read Excel. Cada tabla estaba almacenada en diferentes ficheros generados utilizando Microsoft Excel.

Luego de ejecutar el proceso, los datos a mostrar fueron obtenidos de la vista de metadatos. En la siguiente imagen se ilustra el proceso.

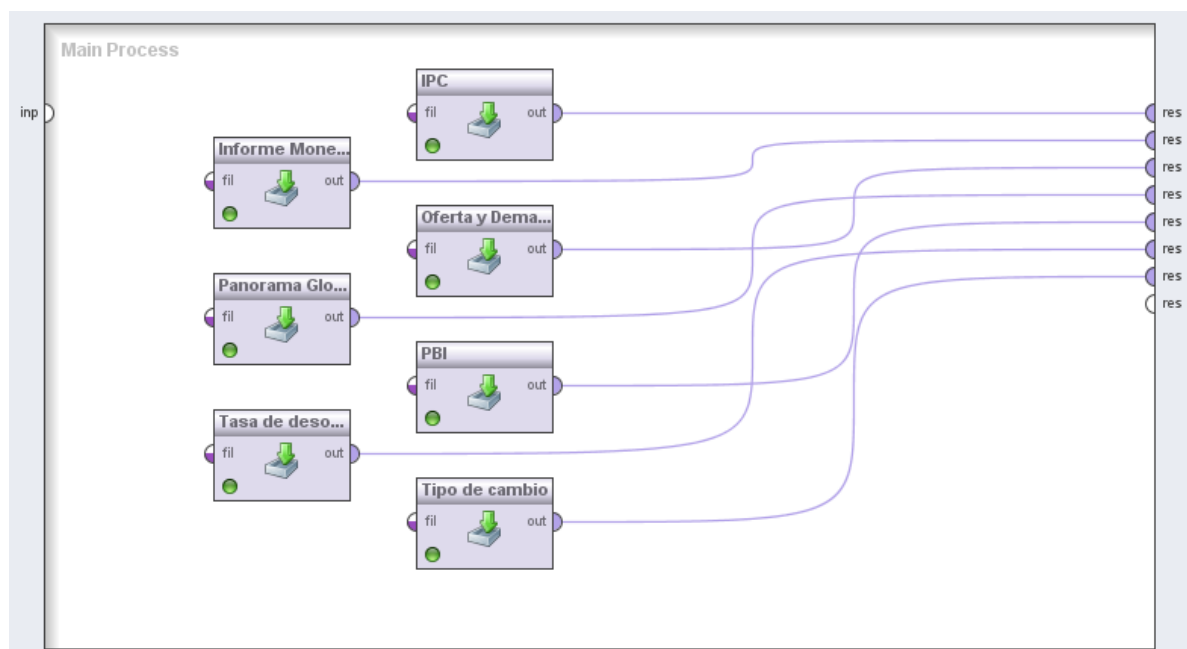


Ilustración 8. Proceso de descripción de datos

4.2.1 Informe Monetario Mensual

La tabla de descripción de datos de los informes monetarios detalla las características de los datos obtenidos de los informes monetarios mensuales. Estos datos corresponden a informes emitidos por el BCRA durante el periodo 2008 a 2015 (BCRA, 2008;2009; 2010; 2011; 2012; 2013; 2014; 2015).

Las cifras de los indicadores monetarios y financieros se encuentran expresadas en millones de pesos. El periodo de tiempo al cual corresponden los datos es mensual.

Rol	Nombre	Tipo de dato	Estadística	Rango	Valores faltantes
regular	Año	integer	avg = 2011.379 +/- 2.136	[2008 ; 2015]	0
regular	Mes	polynomial	mode = Abril (8), least = Julio (7)	Abril (8), Mayo (8), Junio (8), Julio (7), Agosto (7), Septiembre (7), Octubre (7), Noviembre (7), Diciembre (7), Enero (7), Febrero (7), Marzo (7)	0
regular	Base monetaria	integer	avg = 229917.701 +/- 117909.151	[98040 ; 488049]	0
regular	Circulación monetaria	integer	avg = 176788.563 +/- 90413.043	[72003 ; 375416]	0
regular	En poder del público	integer	avg = 158268.966 +/- 81593.247	[65697 ; 338545]	0
regular	En entidades financieras	integer	avg = 18518.011 +/- 8947.885	[6306 ; 38976]	0
regular	Cheques cancelatorios	numeric	avg = 2.405 +/- 2.332	[0 ; 15]	31
regular	Cuenta corriente en el BCRA	integer	avg = 53129.172 +/- 27608.287	[22961 ; 112633]	0
regular	Stock de pases Pasivos	integer	avg = 14322.322 +/- 4612.063	[4682 ; 28111]	0
regular	Stock de pases Activos	integer	avg = 110.920 +/- 412.083	[0 ; 2621]	0
regular	LEBAC	integer	avg = 84926.678 +/- 84342.450	[8161 ; 328337]	0
regular	NOBAC	integer	avg = 19539.920 +/- 15302.318	[0 ; 54676]	0
regular	Reservas internacionales del BCRA	integer	avg = 42829.207 +/- 8006.101	[27332 ; 52427]	0
regular	Depósitos del sector privado y del sector público en pesos	integer	avg = 441637.379 +/- 228607.547	[191539 ; 956618]	0
regular	Cuenta corriente	integer	avg = 127300.954 +/- 66331.708	[50130 ; 294239]	0
regular	Caja de ahorro	integer	avg = 86466.057 +/- 47872.291	[35505 ; 199290]	0
regular	Plazo fijo no ajustable por CER	integer	avg = 209478.023 +/- 108318.782	[82197 ; 442612]	0
regular	Plazo fijo ajustable por CER	integer	avg = 95.713 +/- 242.950	[5 ; 1390]	0
regular	Otros depósitos	integer	avg = 18425 +/- 7602.676	[10282 ; 36890]	0
regular	Depósitos del sector privado	integer	avg = 316578.195 +/- 175940.577	[135862 ; 750477]	0
regular	Depósitos del sector público	integer	avg = 125162.575 +/- 54219.092	[52233 ; 244514]	0

regular	Depósitos del sector privado y del sector público en dólares	integer	avg = 11060.264 +/- 2846.013	[7922 ; 16176]	0
regular	Préstamos al sector privado y al sector público en pesos	integer	avg = 302255.241 +/- 173917.516	[107007 ; 665225]	0
regular	Préstamos al sector privado	integer	avg = 275014.368 +/- 162316.945	[96575 ; 615442]	0
regular	Adelantos	integer	avg = 37554 +/- 20447.609	[15415 ; 78374]	0
regular	Documentos	integer	avg = 60476.839 +/- 40396.596	[18479 ; 143677]	0
regular	Hipotecarios	integer	avg = 29102 +/- 11633.186	[15101 ; 49054]	0
regular	Prendarios	integer	avg = 17626.793 +/- 10197.292	[6389 ; 34251]	0
regular	Personales	integer	avg = 62491.115 +/- 33097.565	[24084 ; 134113]	0
regular	Tarjetas de crédito	integer	avg = 48002.391 +/- 34184.951	[12202 ; 132011]	0
regular	Otros	integer	avg = 19761.368 +/- 13165.657	[4397 ; 43963]	0
regular	Préstamos al sector público	integer	avg = 27240.920 +/- 11972.445	[9388 ; 50907]	0
regular	Préstamos al sector privado y al sector público en dólares	integer	avg = 6016.701 +/- 1873.314	[3333 ; 9893]	0
regular	M1	integer	avg = 285571.460 +/- 147426.766	[115828 ; 616500]	0
regular	M2	integer	avg = 372033.460 +/- 195206.944	[152485 ; 815789]	0
regular	M3	integer	avg = 600006.138 +/- 310028.463	[257592 ; 1295167]	0

Tabla 8. Descripción de datos pertenecientes a los informes monetarios mensuales

4.2.1.1 Explicación de algunas variables

Las variables cuya comprensión a partir del nombre resulta dificultosa, se explican a continuación:

- En poder del público: dinero en poder del público.
- En entidades financieras: dinero en entidades financieras.
- Cheques cancelatorios: medio para la cancelación de obligaciones de dar sumas de dinero.
- LEBAC: letras del Banco Central.
- NOBAC: notas del Banco Central.
- Otros: otros tipos de préstamos al sector privado.
- M1: es el dinero que circula en la economía, incluyendo el dinero en poder del público, los cheques cancelatorios y el dinero en cuentas corrientes.
- M0, es el dinero que circula en la economía y se define como la cantidad de billetes y monedas en manos de los ciudadanos, además del dinero que los bancos tiene en sus cajas, y depositado con el banco central.

- M2: medios de pago, comprende el circulante en poder del público, los cheques cancelatorios en pesos y los depósitos a la vista en pesos del sector público y privado no financiero.
- M3: agregado amplio en pesos, comprende el circulante en poder del público, los cheques cancelatorios en pesos y el total de depósitos en pesos del sector público y privado no financiero

4.2.2 Informe Panorama global commodities

La tabla de descripción de datos de los informes de panorama global de commodities detalla las características de los datos obtenidos de los informes de coyuntura mencionados. Los datos son los precios de las commodities, los cuales corresponden a informes emitidos por el BCRA durante el periodo 2008 a 2015 (BCRA, 2008;2009; 2010; 2011; 2012; 2013; 2014; 2015).

Los valores de los indicadores del panorama global de commodities se encuentran expresados en dólares. El periodo de tiempo al cual corresponden los datos es mensual.

Rol	Nombre	Tipo de dato	Estadísticas	Rango	Valores faltantes
regular	Año	integer	avg = 2011.379 +/- 2.136	[2008 ; 2015]	0
regular	Mes	polynomial	mode = Abril (8), least = Julio (7)	Abril (8), Mayo (8), Junio (8), Julio (7), Agosto (7), Septiembre (7), Octubre (7), Noviembre (7), Diciembre (7), Enero (7), Febrero (7), Marzo (7)	0
regular	Soja (por tonelada)	numeric	avg = 445.061 +/- 82.264	[313 ; 626]	4
regular	Pellets de soja (por tonelada)	numeric	avg = 427.973 +/- 87.638	[293.600 ; 619.100]	4
regular	Aceite de soja (por tonelada)	numeric	avg = 927.980 +/- 192.402	[620.500 ; 1251.600]	4
regular	Maíz (por tonelada)	numeric	avg = 196.417 +/- 60.275	[120.400 ; 309.300]	4
regular	Trigo (por tonelada)	numeric	avg = 237.787 +/- 46.281	[149 ; 319]	4
regular	Arroz (por tonelada)	numeric	avg = 514.876 +/- 78.372	[375.900 ; 764.800]	4
regular	Carne vacuna (por tonelada)	numeric	avg = 2669.118 +/- 1621.744	[40.400 ; 4315.500]	5
regular	Leche (por libras)	numeric	avg = 15.858 +/- 3.165	[8.100 ; 21.700]	28
regular	Azucar (por libras)	real	avg = 0.695 +/- 2.416	[0.100 ; 13.100]	8
regular	Petróleo WTI (U.S. Barrels)	numeric	avg = 83.964 +/- 18.741	[39.100 ; 116.700]	4
regular	Petróleo Brent (U.S. Barrels)	numeric	avg = 91.622 +/- 24.256	[28.100 ; 126.300]	4

regular	Etanol (por galón)	numeric	avg = 2.458 +/- 0.435	[1.580 ; 3.530]	7
regular	Biodiesel (por galón)	numeric	avg = 3.089 +/- 0.515	[1.830 ; 3.700]	7
regular	Oro (por oz.troy)	numeric	avg = 1309.337 +/- 271.314	[760.600 ; 1776.600]	4
regular	Plata (por oz. Troy)	numeric	avg = 22.396 +/- 8.607	[-1 ; 43.100]	4
regular	Cobre (por tonelada)	numeric	avg = 7073.296 +/- 1487.071	[3136 ; 9872.100]	4
regular	Aluminio (por tonelada)	numeric	avg = 2001.119 +/- 308.799	[1337.300 ; 2762.600]	4

Tabla 9. Descripción de datos pertenecientes a los informes de panorama global de commodities

4.2.3 Tasas de desocupación

La tabla de descripción de datos de tasas de desocupación detalla las características de los datos obtenidos sobre el desempleo en Argentina. Estos datos corresponden a informes de prensa emitidos por el INDEC durante el periodo 2008 a 2015 (INDEC, 2008;2009; 2010; 2011; 2012; 2013; 2014; 2015).

Los valores de la tasa de desempleo están expresados en porcentaje, en relación a la población económicamente activa. El periodo de tiempo al cual corresponden los datos es trimestral.

Rol	Nombre	Tipo de dato	Estadísticas	Rango	Valores faltantes
regular	Trimestre	polynomial	mode = 2008 I (1), least = 2008 I (1)	2008 I (1), 2008 II (1), 2008 III (1), 2008 IV (1), 2009 I (1), 2009 II (1), 2009 III (1), 2009 IV (1), 2010 I (1), 2010 II (1), 2010 III (1), 2010 IV (1), 2011 I (1), 2011 II (1), 2011 III (1), 2011 IV (1), 2012 I (1), 2012 II (1), 2012 III (1), 2012 IV (1), 2013 I (1), 2013 II (1), 2013 III (1), 2013 IV (1), 2014 I (1), 2014 II (1), 2014 III (1), 2014 IV (1), 2015 I (1), 2015 II (1)	0
regular	Tasa de desocupación	real	avg = 7.520 +/- 0.662	[6.400 ; 9.100]	0

Tabla 10. Descripción de datos de tasa de desocupación

4.2.4 Oferta y demanda

La siguiente tabla especifica las características de los datos obtenidos sobre la oferta y demanda de bienes y servicios agregada y sus componentes. Por el lado de la oferta, se presentan datos relacionados al PBI y a las importaciones, por el lado de la demanda, datos relacionados con el consumo y las exportaciones. Estos datos corresponden a informes de prensa emitidos por el INDEC durante el periodo 2008 a 2015 (INDEC, 2008;2009; 2010; 2011; 2012; 2013; 2014; 2015).

Los valores están expresados en pesos. El periodo de tiempo al cual corresponden los datos es trimestral.

Rol	Nombre	Tipo de dato	Estadísticas	Rango	Valores faltantes
regular	Trimestre	polynomial	mode = 2008 I (1), least = 2008 I (1)	2008 I (1), 2008 II (1), 2008 III (1), 2008 IV (1), 2009 I (1), 2009 II (1), 2009 III (1), 2009 IV (1), 2010 I (1), 2010 II (1), 2010 III (1), 2010 IV (1), 2011 I (1), 2011 II (1), 2011 III (1), 2011 IV (1), 2012 I (1), 2012 II (1), 2012 III (1), 2012 IV (1), 2013 I (1), 2013 II (1), 2013 III (1), 2013 IV (1), 2014 I (1), 2014 II (1), 2014 III (1), 2014 IV (1), 2015 I (1), 2015 II (1)	0
regular	Producto Interno Bruto a precios de mercado	integer	avg = 806292008.400 +/- 77416105.992	[663315046 ; 913640350]	0
regular	Importaciones	integer	avg = 164342034.367 +/- 27207942.195	[105225458 ; 212917230]	0
regular	Oferta y Demanda	integer	avg = 970634042.633 +/- 99887975.128	[773020190 ; 1089762516]	0
regular	Consumo Individual (Hogares)	integer	avg = 576872936.867 +/- 65677471.087	[461612563 ; 656310091]	0
regular	Consumo Colectivo (Público)	integer	avg = 83027770.800 +/- 12115345.832	[61512079 ; 104564327]	0
regular	Exportaciones	integer	avg = 147269904 +/- 16787220.133	[115032967 ; 174457011]	0
regular	Formación Bruta de Capital Fijo	integer	avg = 166853568.733 +/- 21422709.283	[123908571 ; 204644270]	0

Tabla 11. Descripción de datos de oferta y demanda

4.2.5 Producto bruto interno a precio de mercado

La siguiente tabla especifica las características de los datos obtenidos sobre el producto bruto interno a precio de mercado. Estos datos permiten analizar la composición del producto bruto interno, a partir de otras variables macroeconómicas. Los mismos corresponden a informes de prensa emitidos por el INDEC durante el periodo 2008 a 2015 (INDEC, 2008;2009; 2010; 2011; 2012; 2013; 2014; 2015).

Los valores están expresados en pesos. El periodo de tiempo al cual corresponden los datos es trimestral.

Rol	Nombre	Tipo de dato	Estadísticas	Rango	Valores faltantes
regular	Trimestre	polynomial	mode = 2008 I (1), least = 2008 I (1)	2008 I (1), 2008 II (1), 2008 III (1), 2008 IV (1), 2009 I (1), 2009 II (1), 2009 III (1), 2009 IV (1), 2010 I (1), 2010 II (1), 2010 III (1), 2010 IV (1), 2011 I (1), 2011 II (1), 2011 III (1), 2011 IV (1), 2012 I (1), 2012 II (1), 2012 III (1), 2012 IV (1), 2013 I (1), 2013 II (1), 2013 III (1), 2013 IV (1), 2014 I (1), 2014 II (1), 2014 III (1), 2014 IV (1), 2015 I (1), 2015 II (1)	0
regular	Valor Agregado Bruto	integer	avg = 703684441.800 +/- 68453656.412	[574117492 ; 797867316]	0
regular	Impuestos a los Productos Netos de Subsidios	integer	avg = 41835029.767 +/- 3406961.749	[36289166 ; 47613056]	0
regular	Derechos de Importación	integer	avg = 6907185.300 +/- 1186157.187	[4340905 ; 8991133]	0
regular	Impuesto al Valor Agregado	integer	avg = 53865351.533 +/- 6064094.381	[42309192 ; 61259770]	0
regular	Producto Interno Bruto a precios de mercado	integer	avg = 806292008.400 +/- 77416105.992	[663315046 ; 913640350]	0

Tabla 12. Descripción de datos de producto bruto interno a precio de mercado

4.2.6 Tipo de cambio nominal promedio mensual

La tabla de descripción del tipo de cambio nominal promedio mensual describe a los datos vinculados a la cotización del dólar. Los mismos corresponden a informes de prensa emitidos por el BCRA durante el periodo 2008 a 2015 (BCRA, 2008;2009; 2010; 2011; 2012; 2013; 2014; 2015).

Los valores están expresados en relación peso/dólar. El periodo de tiempo al cual corresponden los datos es mensual.

Rol	Nombre	Tipo de dato	Estadísticas	Rango	Valores faltantes
regular	Año	integer	avg = 2011.379 +/- 2.136	[2008 ; 2015]	0
regular	Mes	polynomial	mode = Abril (8), least = Julio (7)	Abril (8), Mayo (8), Junio (8), Julio (7), Agosto (7), Septiembre (7), Octubre (7), Noviembre (7), Diciembre (7), Enero (7), Febrero (7), Marzo (7)	0
regular	Tipo de cambio nominal promedio mensual	real	avg = 5.063 +/- 1.828	[3.022 ; 9.042]	0

Tabla 13. Descripción de datos de tipo de cambio nominal promedio mensual

4.2.7 Índice de precios al consumidor

La tabla de descripción del índice de precios al consumidor describe los datos vinculados a la inflación. Los mismos corresponden a informes de prensa emitidos por el INDEC durante el periodo 2008 a 2015 (INDEC, 2008;2009; 2010; 2011; 2012; 2013; 2014; 2015).

Los valores representan índices calculados a partir de una base previamente definida (Ver Sección 3.1.1.3). Además se incluye la variación porcentual de la inflación en relación al mes anterior. El periodo de tiempo al cual corresponden los datos es mensual.

Rol	Nombre	Tipo de dato	Estadísticas	Rango	Valores faltantes
regular	Año	integer	avg = 2011.379 +/- 2.136	[2008 ; 2015]	0
regular	Mes	polynomial	mode = Abril (8), least = Julio (7)	Abril (8), Mayo (8), Junio (8), Julio (7), Agosto (7), Septiembre (7), Octubre (7), Noviembre (7), Diciembre (7), Enero (7), Febrero (7), Marzo (7)	0
regular	Variación porcentual	real	avg = 0.924 +/- 0.524	[0.331 ; 3.649]	0
regular	Índice de precios al consumidor base Abril 2008 = 100	numeric	avg = 127.772 +/- 19.964	[100 ; 166.840]	18
regular	Índice de precios al consumidor base IV trimestre 2013 = 100	real	avg = 114.704 +/- 6.613	[102.780 ; 123.170]	77
regular	Índice de precios al consumidor base oct 2013-sep 2014=100	real	avg = 109.630 +/- 7.688	[94.840 ; 120.980]	69

Tabla 14. Descripción de datos de índice de precios al consumidor

4.3 Exploración de datos

Para profundizar el análisis de los datos previo a su preparación, se realizó el análisis de algunos atributos.

4.3.1 Comparación de la tasa de desocupación y el índice de precios al consumidor

Según la curva de Philips, a corto plazo existe una relación inversa entre la tasa de desempleo y la tasa de inflación. Para analizar esta teoría se procedió a realizar una comparación entre ambos atributos en el periodo comprendido entre enero de 2009 y diciembre de 2014 (Ilustración 9 e Ilustración 10).

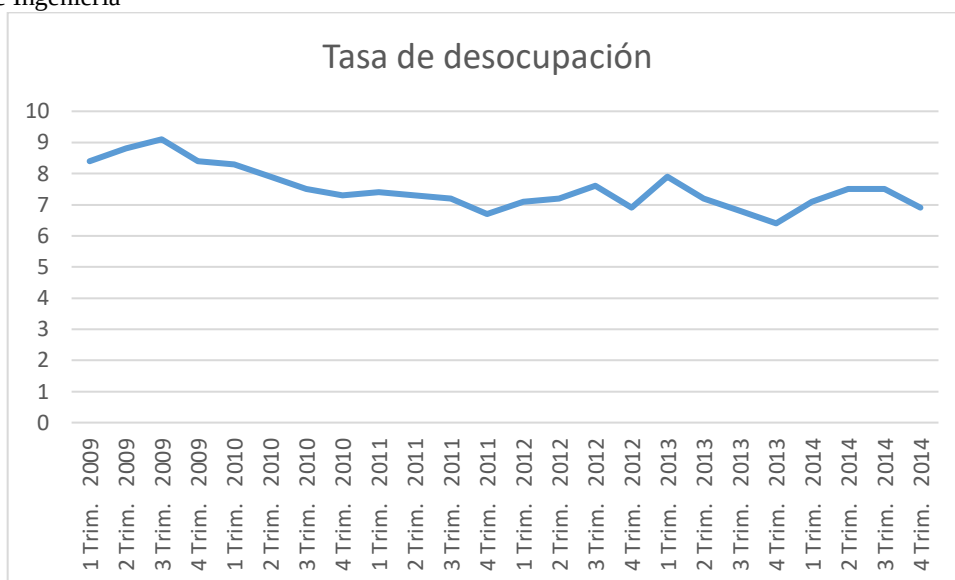


Ilustración 9. Distribución de la tasa de desocupación

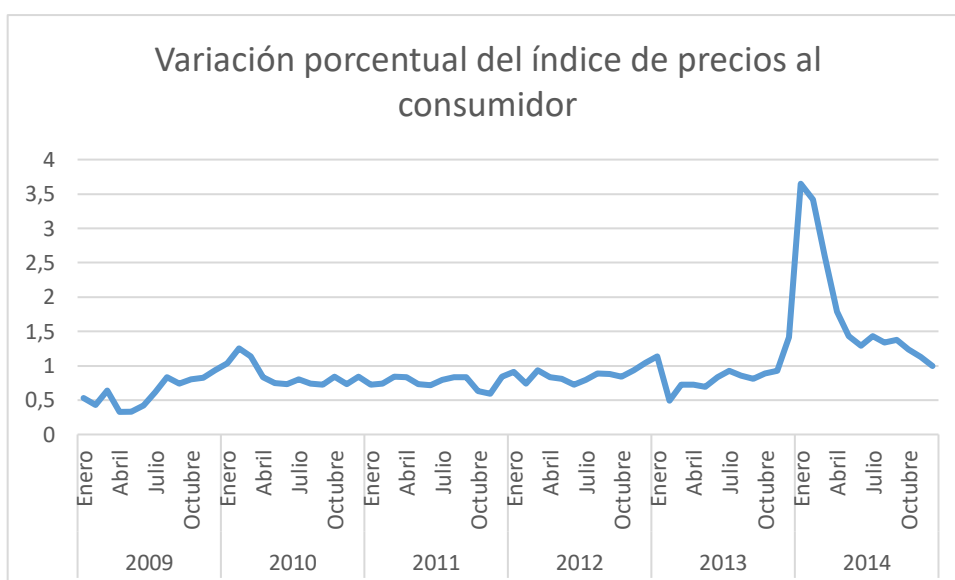


Ilustración 10. Distribución de la variación porcentual

A través de los gráficos no se puede apreciar una relación lineal entre ambas tasas, pero se puede mencionar que en el período del segundo trimestre del 2009, los valores de la tasa de inflación son los menores en el intervalo analizado, en el mismo período (2^{do} trimestre de 2009) se registran valores muy altos de la tasa de desocupación. Además, en el cuarto trimestre de 2013 la tasa de desocupación registro su menor valor, lo que coincide con un crecimiento en la tasa de inflación.

Sin embargo, si se compara la tasa de desocupación con el Índice de precios al consumidor de base Abril 2008 = 100 se puede observar que existe una relación inversamente proporcional, debido a que mientras la tasa de desocupación disminuye a lo largo del periodo analizado, el índice de precios al consumidor aumenta. Esto se puede apreciar en los siguientes gráficos de tendencias, la explicación del proceso que genera ambas ilustraciones se encuentra en el Anexo II.

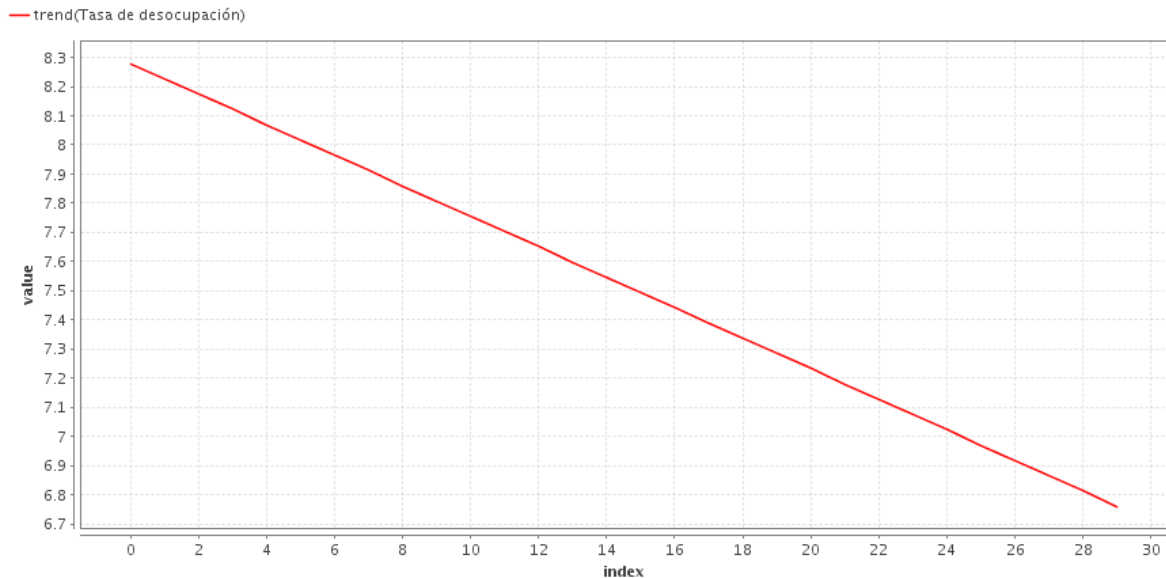


Ilustración 11. Tendencia de la tasa de desocupación

En el gráfico de tendencia de la tasa de desocupación los datos son trimestrales, mientras que en la ilustración de tendencia del índice de precios al consumidor los datos son mensuales, en este caso el número de mes corresponde a los meses de cada dato de la serie. Es decir, el mes 1 corresponde al primer mes, el mes 2 al segundo mes y así sucesivamente.

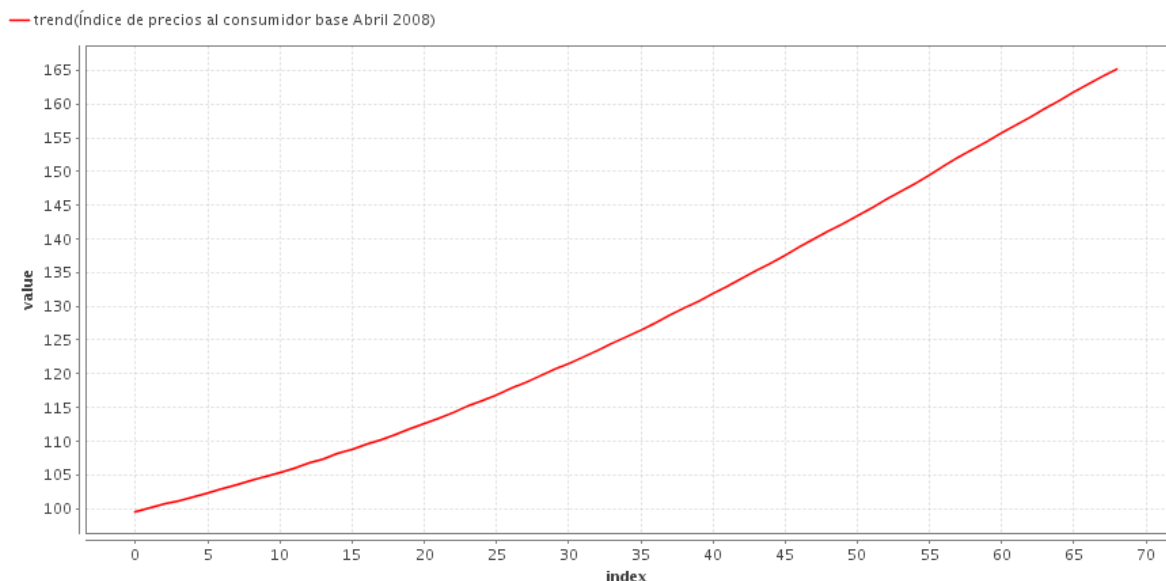


Ilustración 12. Tendencia del índice de precios al consumidor base Abril 2008 = 100

Por lo explicado anteriormente, se decide tomar como referencia de la inflación al índice de precios y no a su variación porcentual. En el periodo entre enero de 2008 y junio de 2015 existen tres mediciones diferentes del índice de precios al consumidor (base Abril 2008 = 100, base IV trimestre 2013 = 100 y base oct 2013-sep 2014=100), cada uno es utilizada en diferentes períodos. Para realizar el análisis de la inflación se decidió utilizar la primera.

La elección de la base Abril 2008 = 100, se debe a que es la base de los índices de precios al consumidor que mayores datos contiene, las bases restantes (base IV trimestre 2013 = 100 y base oct 2013-sep 2014=100) contienen menos registros de datos.

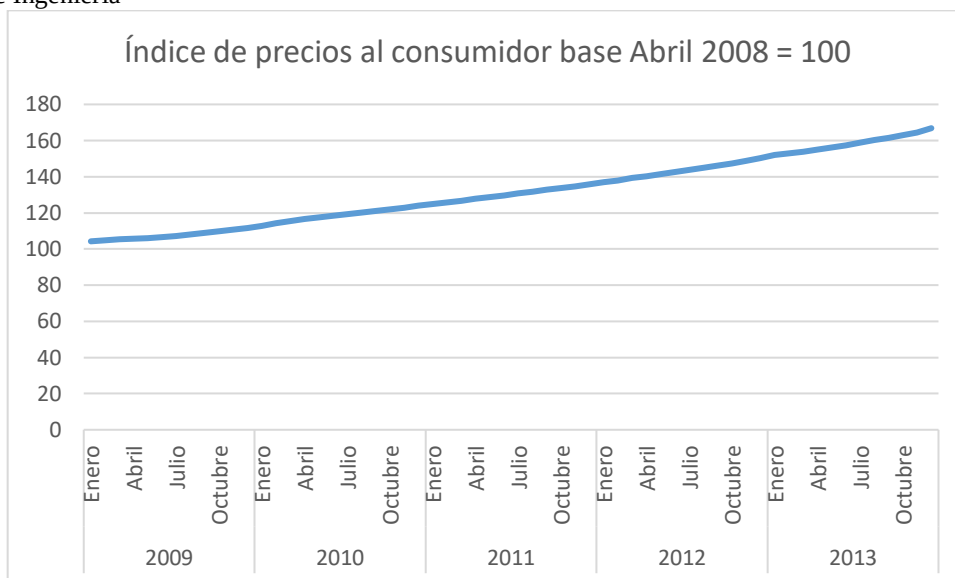


Ilustración 13. Distribución del índice de precios al consumidor base Abril 2008 = 100

Debido a que la relación inversamente proporcional se da entre la tasa de desempleo y el Índice de precios al consumidor base Abril 2008, se puede considerar como objeto de análisis al Índice de precios al consumidor base Abril 2008 y no a su variación porcentual. El atributo de objeto de análisis será el que actúa como un atributo de destino para los operadores de aprendizaje, en este caso el operador de redes neuronales.

Si el análisis de datos se realiza a través de este índice de precios, será necesario considerar estimar los valores faltantes de esta variable para el periodo entre enero de 2009 a junio de 2015, o bien filtrar estos registros.

4.3.2 Distribución de la base monetaria

La base monetaria está constituida por la Circulación monetaria y la Cuenta corriente en el BCRA. Por otro lado, la circulación monetaria está formada por el dinero en poder del público, el dinero en entidades financieras y los cheques cancelatorios.

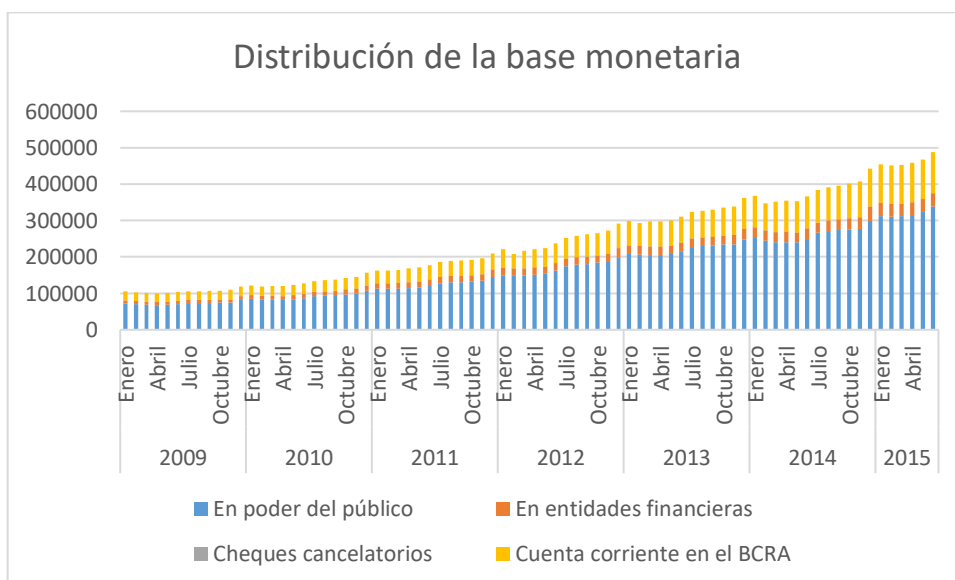


Ilustración 14. Distribución de la base monetaria

4.4 Verificación de calidad de los datos

A continuación se presenta el informe de calidad de datos, para realizar este análisis se recurrió a la visualización a través de gráficos de línea, de dispersión y diagramas de cajas. Esto permite revelar inconsistencias en los conjuntos de datos.

Es importante mencionar que los datos pertenecientes a la tabla Panorama Global de Commodities contienen valores nulos entre Abril y Julio de 2008. No obstante, se realiza el análisis para aquellos atributos que contienen valores faltantes en el resto de los registros.

En primer lugar se presentarán los gráficos de los atributos que contienen valores faltantes y/o valores erróneos. Luego, un análisis de cada caso y finalmente las conclusiones y fundamentos sobre las decisiones tomadas en las distintas situaciones.

4.4.1 Leche (por libras)

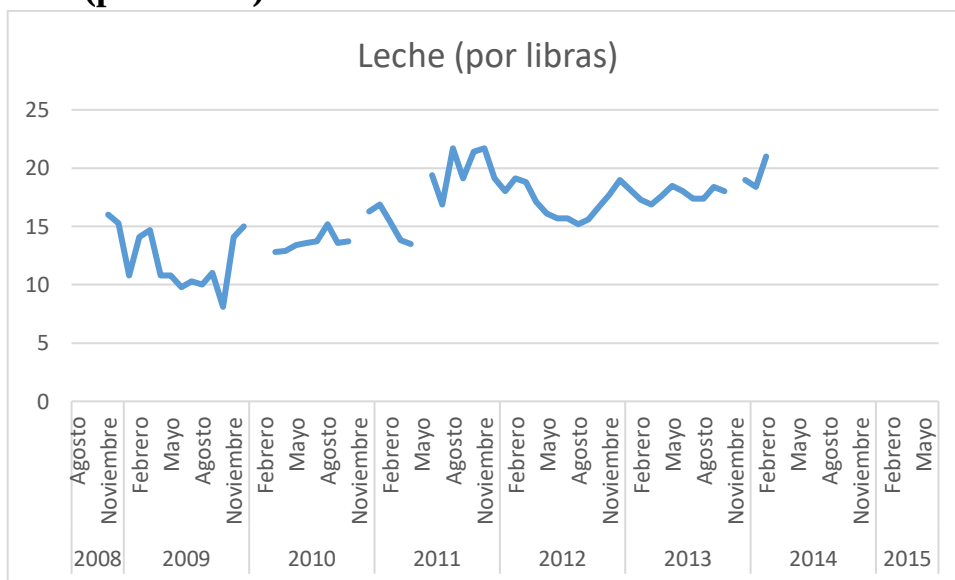


Ilustración 15. Distribución de la leche

La distribución de la cotización del precio de la leche en libras posee valores faltantes en distintos periodos de tiempo. La existencia de valores faltantes, en este caso, se debe a que el conjunto de datos está incompleto, es decir no existe información sobre estos valores en los periodos de tiempo indicados por el gráfico.

4.4.2 Carne vacuna (por tonelada)

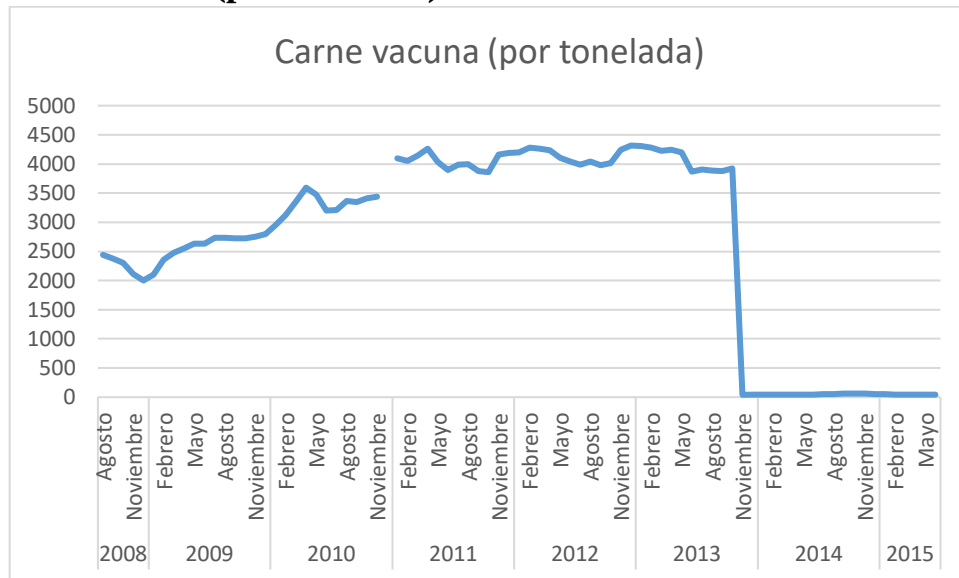


Ilustración 16. Distribución de la carne vacuna

La distribución de la cotización de la carne vacuna registra un valor faltante correspondiente a Diciembre del 2010 y a partir de Noviembre del 2013 existen valores que aparentemente son atípicos como se puede apreciar en el gráfico. Para hacer un análisis más completo del conjunto de datos se realizó un histograma, mediante este grafico se puede apreciar lo separado que están los datos en dos subconjuntos. Estos valores si bien pueden representar datos de la realidad, también pueden generar dificultades al momento de elaborar el modelo de datos, particularmente en el caso de las redes neuronales.

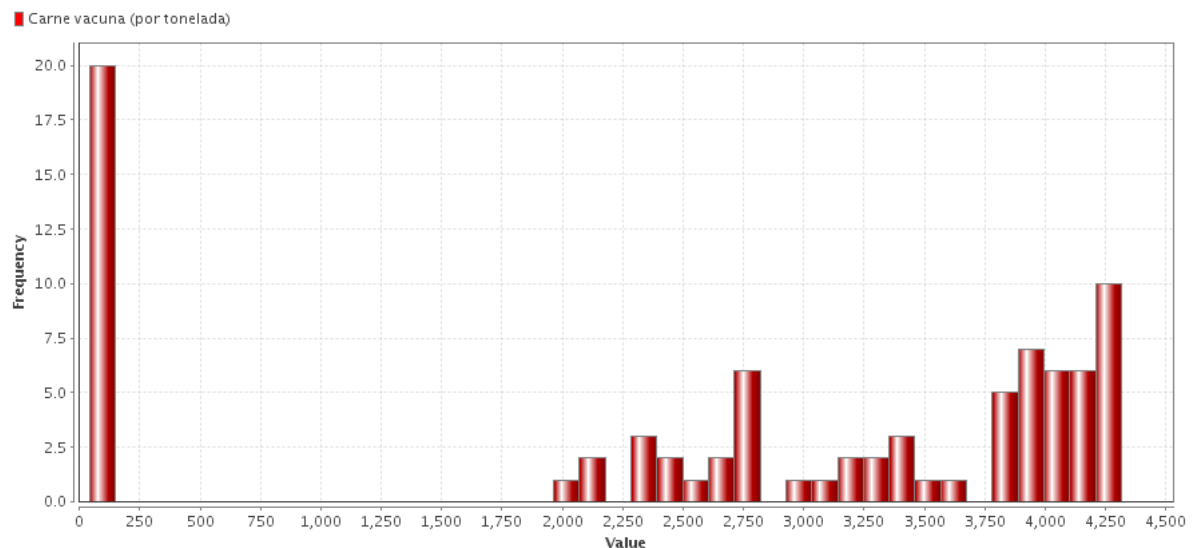


Ilustración 17. Histograma de Carne vacuna (por tonelada)

4.4.3 Azúcar (en libras)

El conjunto de datos pertenecientes al azúcar en libras presenta valores faltantes y valores anómalos, según se puede apreciar en el gráfico de dispersión.

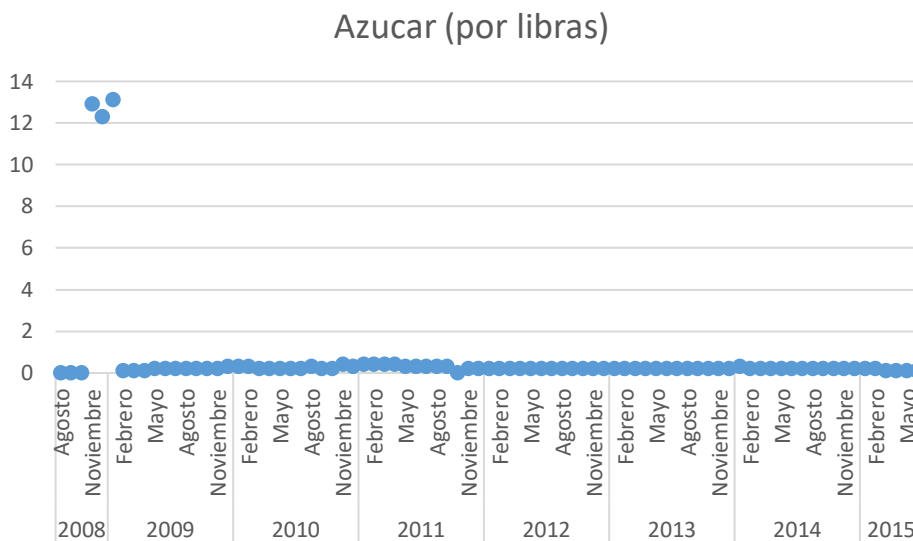


Ilustración 18. Diagrama de dispersión del azúcar

4.4.4 Cheques cancelatorios

El cheque cancelatorio es un medio para la cancelación de obligaciones de dar sumas de dinero. A diferencia de los demás tipos de cheques, los cheques cancelatorios son emitidos por el BCRA (quien los entrega a los bancos en consignación a medida que las entidades registran pedidos del público). Tienen una validez de 30 días corridos desde la fecha que figura en el documento, y pueden tener hasta dos endosos. Además, tienen la particularidad de ser un instrumento bimonetario, ya que pueden ser reclamados al banco en pesos o en dólares según lo prefiera su legítimo tenedor, y son emitidos con sólo tres denominaciones: \$5.000, \$10.000 y \$50.000. (BCRA, 2015)

La distribución de los cheques cancelatorios permite observar que contiene valores faltantes entre Abril de 2008 y Octubre de 2010.

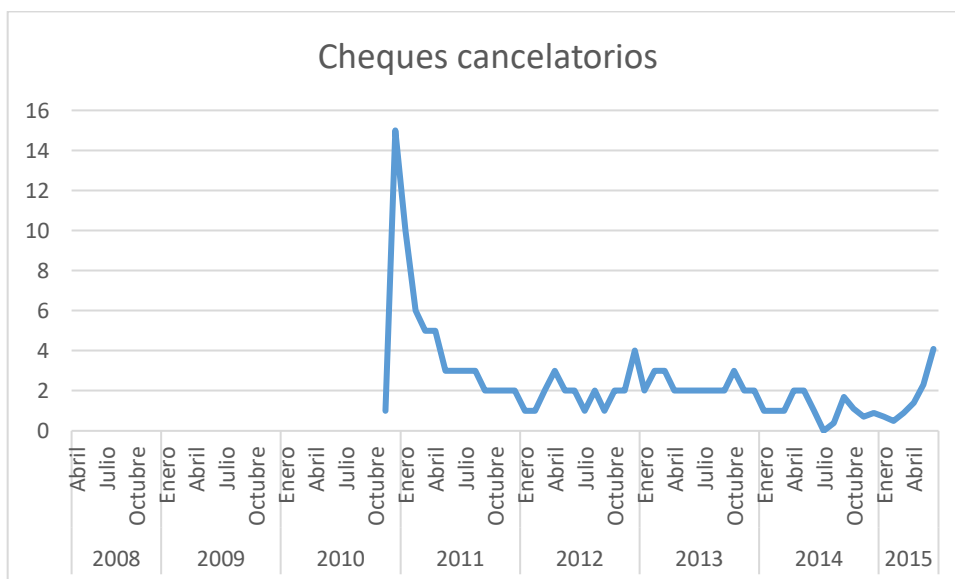


Ilustración 19. Distribución de los cheques cancelatorios

4.4.5 Stock de pases Activos

La distribución de la variable de Stock de pases activos contiene muchos valores equivalentes a 0, y otros que resultan muy alejados a éstos.

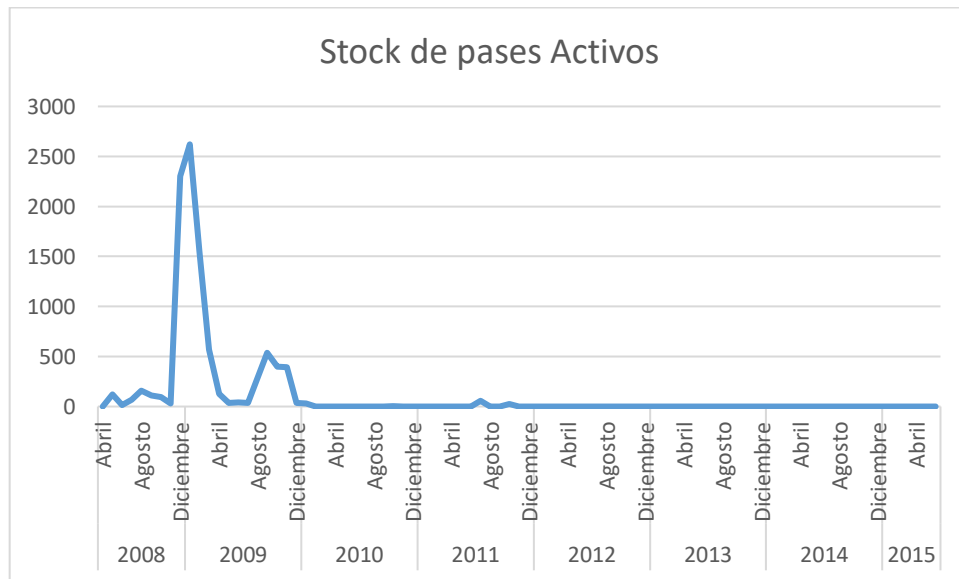


Ilustración 20. Distribución del stock de pases activos

Para profundizar el análisis, se incluye un diagrama de cajas, en el cual se pueden observar muchos valores fuera del rango de valores típicos.

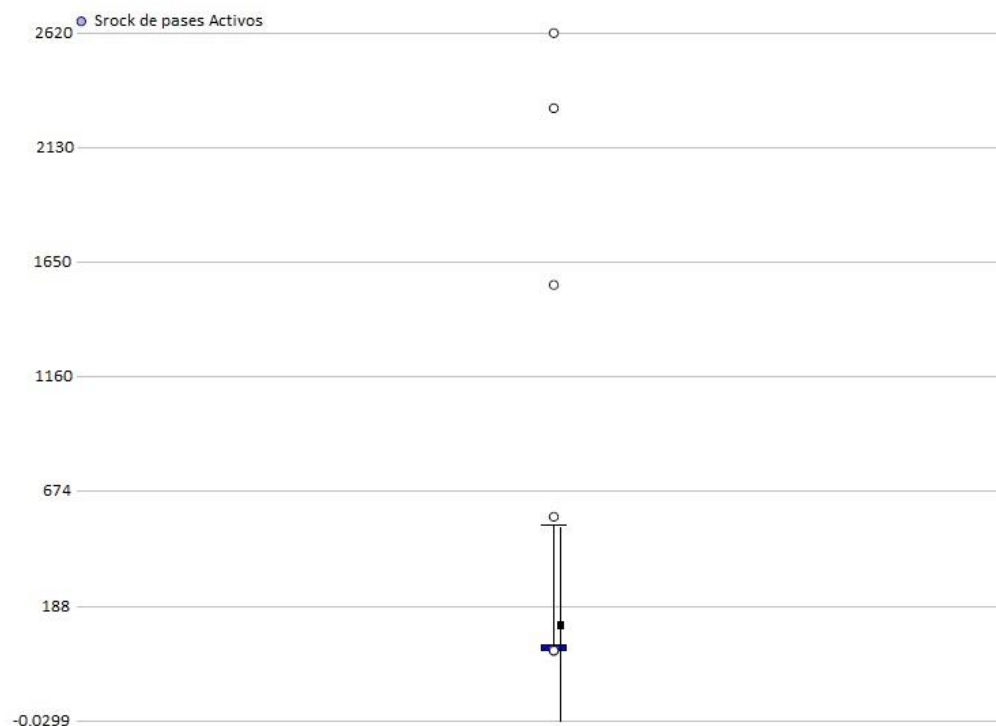


Ilustración 21. Diagrama de caja del stock de pases activos

4.4.6 Etanol (por galón) y Biodiesel (por galón)

Ambos atributos contienen valores faltantes en sus tres primeros registros. Si bien se podría considerar utilizar el conjunto de datos con estos valores faltantes, esto puede ser perjudicial para el algoritmo encargado de generar el modelo, ya que los datos son numéricos y el algoritmo trabaja ponderando estos valores.

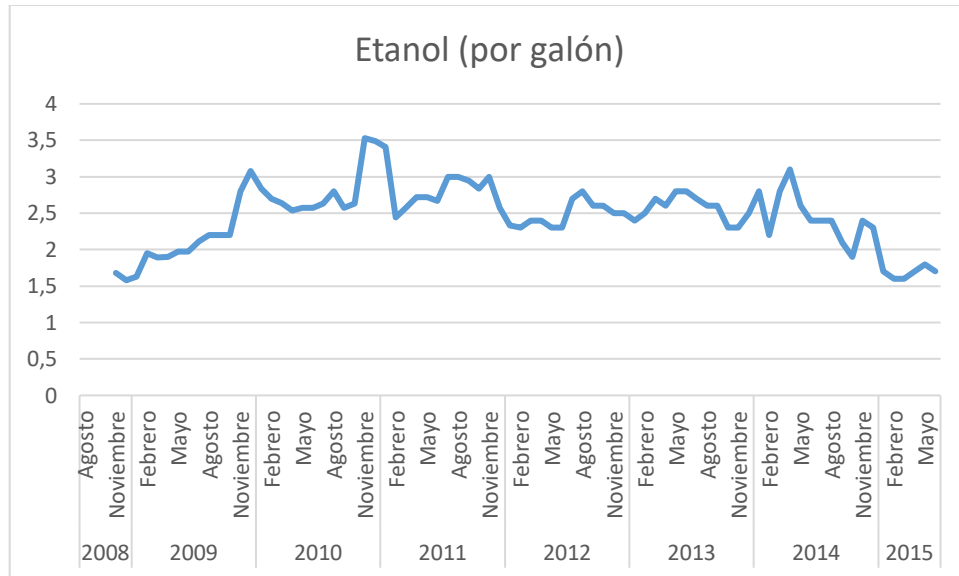


Ilustración 22. Distribución de Etanol (por galón)

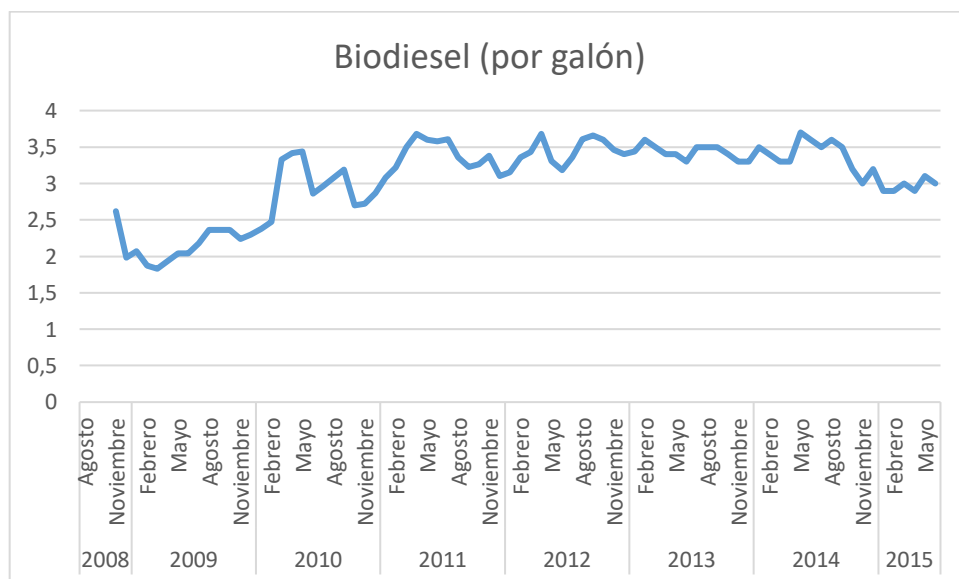


Ilustración 23. Distribución de Biodiesel (por galón)

4.4.7 Plata (en oz. Troy)

El grafico lineal que representa el precio de la plata en dólares contiene un valor negativo y otros que se encuentran alejados de la tendencia de la serie.

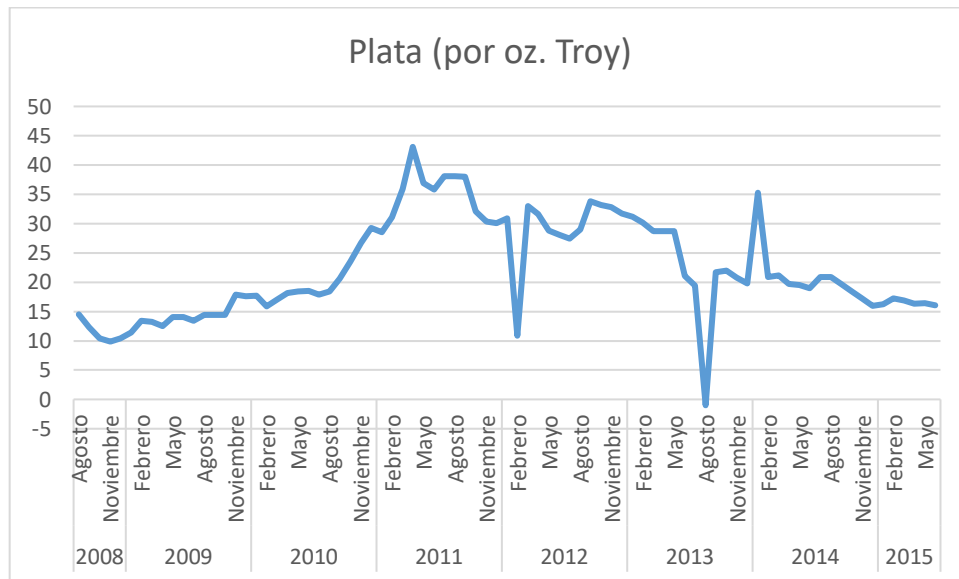


Ilustración 24. Distribución de la plata en dólares

4.4.8 Tratamiento de los valores nulos y/o erróneos

Para poder considerar los datos de la tabla de panorama global de commodities será necesario filtrar las tuplas de los meses correspondientes a Abril, Mayo, Junio y Julio de 2008. En dichos meses existen valores nulos para todos los atributos de esta tabla.

Otros registros a filtrar, serán los correspondientes al periodo entre Enero de 2009 y Junio de 2015, ya que no resulta posible realizar un empalme entre la serie Índice de precios al consumidor de base Abril 2008 = 100 con las siguientes series (Índice de precios al consumidor base IV trimestre 2013 = 100, e Índice de precios al consumidor base oct 2013-sep 2014=100), debido a que estas series se refieren a distintas áreas geográficas del país. Mientras la primera solo considera los consumos de familias del Área Metropolitana de Buenos Aires, las siguientes abarcan todo el territorio nacional. Por otro lado, el periodo de referencia es distinto, en la primera toma como base a un mes, mientras que las restantes contemplan un trimestre y un año respectivamente.

Luego de analizar el resto de los casos en los que existen valores nulos y/o erróneos, se consideraron las siguientes alternativas de tratamiento para estos atributos:

- Utilizar una variable booleana para los datos nulos no resulta útil debido a que todos los datos son de tipo numérico.
- Filtrar las tuplas con valores faltantes y/o erróneos tampoco es solución ya que sería necesario eliminar muchas filas y se perderían muchos datos.
- Reemplazar los valores nulos o erróneos por valores estimados no garantiza que los nuevos valores sean precisos.

Finalmente se decidió eliminar los atributos que presentan estos tipos de problema.

El principal motivo de esta decisión radica en que los datos son de tipo numérico y el algoritmo a utilizar, las redes neuronales, trabaja con ajustes de pesos de acuerdo a los valores de entrada. Los valores nulos o erróneos pueden afectar de manera considerable el modelo generado.

Los atributos a eliminar son: Leche (en libras), Carne vacuna (por tonelada), Azúcar (en libras), Cheques cancelatorios, Stock de pases Activos y Plata (en oz. Troy).

Otros atributos que deberán ser descartados son “Índice de precios al consumidor base VI trimestre 2013 = 100”, “Índice de precios al consumidor base octubre 2013 – septiembre 2014 = 100” y “Variación porcentual”, según se explicó en la exploración de datos.

4.5 Resumen

Los datos que serán analizados fueron obtenidos de dos organismos públicos, el BCRA y el INDEC. Estos datos fueron obtenidos a través de informes en formato Portable Document Format (PDF) y en hojas de cálculo de Microsoft Excel. En el caso de los informes en PDF fue necesaria su carga manual a hojas de cálculo de Microsoft Excel. Una vez extraídos los datos se realizaron los primeros análisis de los mismos. En primer lugar se estudiaron algunas relaciones teóricas entre los datos y luego se consideraron los casos de atributos que contienen valores faltantes o erróneos. El resultado de todas estas tareas fue obtener un conjunto de datos que no tienen valores faltantes o erróneos.

5 Preparación de los datos

Durante la etapa de preparación de datos se realizan las transformaciones necesarias de los datos antes de su procesamiento. El objetivo de la preparación de datos es obtener la descripción de los datos que serán utilizados para elaborar el modelo en la etapa siguiente.

5.1 Limpieza y selección de datos

En la etapa anterior del proyecto, cuando se realizó la verificación de calidad de los datos, se elaboró un análisis inicial de los datos, en esta sección el análisis incluye otros aspectos de los datos que deben ser estudiados antes de realizar el procesamiento de los mismos.

A continuación se presenta el estudio de algunos aspectos de los datos, y luego la selección de los mismos a través de una tabla.

5.1.1 Atributos correspondientes a periodos de tiempo

En el caso de los atributos que hacen referencia al periodo de tiempo, solo se consideraran el año y el mes de la primera tabla, las restantes serán solo para identificar el periodo y realizar la integración de datos. En el caso de los datos trimestrales, primero se realizará la interpolación de datos para obtener los datos mensuales de tales atributos.

5.1.2 Datos agregados

Los datos agregados son aquellos que se obtienen a partir de otros datos mediante una operación matemática, como una suma o una multiplicación. Es necesario filtrar este tipo de datos, ya que puede influir para generar el modelo de datos.

Seguidamente, se detallan los atributos necesarios a filtrar y su correspondiente descomposición aritmética.

Base Monetaria = Circulación monetaria + Cuenta corriente en el BCRA

Circulación monetaria = En poder del público + En entidades financieras + Cheques Cancelatorios

Depósitos del sector privado y del sector público en pesos = Depósitos del sector privado + Depósitos del sector público

Préstamos al sector privado y al sector público en pesos = Préstamos al sector privado + Préstamos al sector público

Préstamos al sector privado = Adelantos + Documentos + Hipotecarios + Prendarios + Personales + Tarjetas de crédito + Otros

M1 = En poder del público + Cheques cancelatorios + Cuenta corriente

M2 = En poder del público + Cheques cancelatorios + Cuenta corriente + Caja de ahorro

M3 = En poder del público + Cheques cancelatorios + Depósitos del sector privado y del sector público en pesos

Oferta y Demanda = Producto Interno Bruto a precios de mercado + Importaciones

Producto Interno Bruto a precios de mercado = Valor Agregado Bruto + Impuestos a los Productos Netos de Subsidios + Derechos de Importación + Impuesto al Valor Agregado
Variación porcentual

5.1.3 Otros atributos que no serán considerados en el análisis

El atributo “Producto Interno Bruto a precios de mercado” se encuentra tanto en la tabla “Producto Interno Bruto a precios de mercado” como en la tabla “Oferta y demanda”. Para evitar que existan atributos repetidos es necesario filtrar uno de los atributos.

“Tipo de cambio nominal promedio mensual” solo será utilizado para convertir precios en dólares a la moneda nacional. Luego será descartado, ya que no se lo considerará en el análisis.

De la tabla “Índice de precios al consumidor”, la columna variación porcentual será omitida, debido a que el análisis de la inflación se realizara a través del “Índice de precios al consumidor base Abril 2008 = 100”.

5.1.4 Datos seleccionados

Los datos seleccionados corresponden al periodo Agosto de 2008 a Diciembre del 2013. A continuación se menciona la decisión de incluir o no cada dato y el motivo de la decisión.

Tabla	Atributo	Decisión	Motivo
Informe Monetario Mensual	Año	Incluir en el análisis	Atributo importante para el análisis
	Mes	Incluir en el análisis	Atributo importante para el análisis
	Base monetaria	Excluir del análisis	Atributo que contiene datos agregados
	Circulación monetaria	Excluir del análisis	Atributo que contiene datos agregados
	En poder del público	Incluir en el análisis	Atributo importante para el análisis
	En entidades financieras	Incluir en el análisis	Atributo importante para el análisis
	Cheques cancelatorios	Excluir del análisis	Contiene valores nulos y/o erróneos
	Cuenta corriente en el BCRA	Incluir en el análisis	Atributo importante para el análisis
	Stock de pases Pasivos	Incluir en el análisis	Atributo importante para el análisis
	Stock de pases Activos	Excluir del análisis	Contiene valores nulos y/o erróneos
	LEBAC	Incluir en el análisis	Atributo importante para el análisis
	NOBAC	Incluir en el análisis	Atributo importante para el análisis
	Reservas internacionales del BCRA	Incluir en el análisis	Atributo importante para el análisis
	Depósitos del sector privado y del sector público en pesos	Excluir del análisis	Atributo que contiene datos agregados
	Cuenta corriente	Incluir en el análisis	Atributo importante para el análisis
	Caja de ahorro	Incluir en el análisis	Atributo importante para el análisis
	Plazo fijo no ajustable por CER	Incluir en el análisis	Atributo importante para el análisis
	Plazo fijo ajustable por CER	Incluir en el análisis	Atributo importante para el análisis
	Otros depósitos	Incluir en el análisis	Atributo importante para el análisis

	Depósitos del sector privado	Incluir en el análisis	Atributo importante para el análisis
	Depósitos del sector público	Incluir en el análisis	Atributo importante para el análisis
	Depósitos del sector privado y del sector público en dólares	Incluir en el análisis	Atributo importante para el análisis
	Préstamos al sector privado y al sector público en pesos	Excluir del análisis	Atributo que contiene datos agregados
	Préstamos al sector privado	Excluir del análisis	Atributo que contiene datos agregados
	Adelantos	Incluir en el análisis	Atributo importante para el análisis
	Documentos	Incluir en el análisis	Atributo importante para el análisis
	Hipotecarios	Incluir en el análisis	Atributo importante para el análisis
	Prendarios	Incluir en el análisis	Atributo importante para el análisis
	Personales	Incluir en el análisis	Atributo importante para el análisis
	Tarjetas de crédito	Incluir en el análisis	Atributo importante para el análisis
	Otros	Incluir en el análisis	Atributo importante para el análisis
	Préstamos al sector público	Incluir en el análisis	Atributo importante para el análisis
	Préstamos al sector privado y al sector público en dólares	Incluir en el análisis	Atributo importante para el análisis
	M1	Excluir del análisis	Atributo que contiene datos agregados
	M2	Excluir del análisis	Atributo que contiene datos agregados
	M3	Excluir del análisis	Atributo que contiene datos agregados
Panorama Global de Commodities	Año	Excluir del análisis	Atributo que se refiere a período de tiempo
	Mes	Excluir del análisis	Atributo que se refiere a período de tiempo
	Soja (por tonelada)	Incluir en el análisis	Atributo importante para el análisis
	Pellets de soja (por tonelada)	Incluir en el análisis	Atributo importante para el análisis
	Aceite de soja (por tonelada)	Incluir en el análisis	Atributo importante para el análisis
	Maíz (por tonelada)	Incluir en el análisis	Atributo importante para el análisis
	Trigo (por tonelada)	Incluir en el análisis	Atributo importante para el análisis
	Arroz (por tonelada)	Incluir en el análisis	Atributo importante para el análisis
	Carne vacuna (por tonelada)	Excluir del análisis	Contiene valores nulos y/o erróneos
	Leche (por libras)	Excluir del análisis	Contiene valores nulos y/o erróneos
	Azúcar (por libras)	Excluir del análisis	Contiene valores nulos y/o erróneos
	Petróleo WTI (U.S. Barrels)	Incluir en el análisis	Atributo importante para el análisis

	Petróleo Brent (U.S. Barrels)	Incluir en el análisis	Atributo importante para el análisis
	Etanol (por galón)	Excluir del análisis	Contiene valores nulos y/o erróneos
	Biodiesel (por galón)	Excluir del análisis	Contiene valores nulos y/o erróneos
	Oro (por oz.troy)	Incluir en el análisis	Atributo importante para el análisis
	Plata (por oz. Troy)	Excluir del análisis	Contiene valores nulos y/o erróneos
	Cobre (por tonelada)	Incluir en el análisis	Atributo importante para el análisis
	Aluminio (por tonelada)	Incluir en el análisis	Atributo importante para el análisis
Tasas de desocupación	Trimestre	Excluir del análisis	Atributo que se refiere a período de tiempo
	Tasa de desocupación	Incluir en el análisis	Atributo importante para el análisis
Oferta y demanda	Trimestre	Excluir del análisis	Atributo que se refiere a período de tiempo
	Producto Interno Bruto a precios de mercado	Excluir del análisis	Atributo que contiene datos agregados
	Importaciones	Incluir en el análisis	Atributo importante para el análisis
	Oferta y Demanda	Excluir del análisis	Atributo que contiene datos agregados
	Consumo Individual (Hogares)	Incluir en el análisis	Atributo importante para el análisis
	Consumo Colectivo (Público)	Incluir en el análisis	Atributo importante para el análisis
	Exportaciones	Incluir en el análisis	Atributo importante para el análisis
	Formación Bruta de Capital Fijo	Incluir en el análisis	Atributo importante para el análisis
Producto Bruto Interno a precio de mercado	Trimestre	Excluir del análisis	Atributo que se refiere a período de tiempo
	Valor Agregado Bruto	Incluir en el análisis	Atributo importante para el análisis
	Impuestos a los Productos Netos de Subsidios	Incluir en el análisis	Atributo importante para el análisis
	Derechos de Importación	Incluir en el análisis	Atributo importante para el análisis
	Impuesto al Valor Agregado	Incluir en el análisis	Atributo importante para el análisis
	Producto Interno Bruto a precios de mercado	Excluir del análisis	Atributo repetido en otra tabla
Tipo de cambio nominal promedio mensual	Año	Excluir del análisis	Atributo que se refiere a período de tiempo
	Mes	Excluir del análisis	Atributo que se refiere a período de tiempo
	Tipo de cambio nominal promedio mensual	Excluir del análisis	Atributo considerado solamente para obtener datos calculados
Índice de precios al consumidor	Año	Excluir del análisis	Atributo que se refiere a período de tiempo
	Mes	Excluir del análisis	Atributo que se refiere a período de tiempo

	Variación porcentual	Excluir del análisis	Atributo no utilizado como referencia de la inflación
	Índice de precios al consumidor base Abril 2008 = 100	Incluir en el análisis	Atributo importante para el análisis
	Índice de precios al consumidor base IV trimestre 2013 = 100	Excluir del análisis	Contiene valores nulos y/o erróneos
	Índice de precios al consumidor base oct 2013-sep 2014=100	Excluir del análisis	Contiene valores nulos y/o erróneos

Tabla 15. Datos seleccionados

5.2 Construcción de datos

En esta sección se construyen nuevos datos a partir de otros. El objetivo de la construcción de datos es garantizar que los datos a integrar no sean inconsistentes ni tengan distintos formatos.

5.2.1 Período

Hasta ahora el periodo temporal de los datos fue trabajado con dos atributos Año y Mes. Se añadirá un atributo denominado Periodo para poder trabajar con un único atributo, para lograr esto se realizará la siguiente transformación:

En primer lugar los valores mensuales serán expresados a través de números del 1 al 12 según corresponde. Luego se realizara una concatenación entre el valor de Año y el nuevo valor de Mes de la siguiente manera: Año/Mes.

Así por ejemplo a partir del Mes de Noviembre del año 2008, el valor de Periodo será de 2008/11.

5.2.2 Unificación de moneda económica

Los datos económicos se encuentran expresados en diferentes medidas monetarias (en pesos, millones de pesos y en dólares). Es necesario realizar una unificación de medidas antes del análisis de los datos.

Atributo	Medida monetaria
Importaciones	Peso
Exportaciones	Peso
Impuestos a los Productos Netos de Subsidios	Peso
Cuenta corriente	Millones de pesos
Caja de ahorro	Millones de pesos
Hipotecarios	Millones de pesos
Arroz (por tonelada)	Dólar
Oro (por oz.troy)	Dólar
Cobre (por tonelada)	Dólar

Tabla 16. Ejemplo de diferencias de medidas monetarias

Los datos serán convertidos a pesos, la conversión de los datos en millones de pesos no presenta mayores inconvenientes, ya que a través de la correspondiente operación multiplicativa se obtendrán la conversión necesaria.

En el caso de los datos expresados en dólares, será necesario recurrir al atributo “Tipo de cambio nominal promedio mensual” que expresa la cotización promedio del dólar para cada mes.

Sea $A_{dolar}(Mi)$ el valor de un atributo expresado en dólares correspondiente al Mi y sea $TCPM(Mi)$ la cotización del dólar para el mes Mi , entonces el valor aproximado del atributo expresado en dólares se obtiene mediante:

$$A_{peso}(Mi) = A_{dolar}(Mi) * TCPM(Mi)$$

5.3 Formateo de datos

“El formateo de datos se refiere a modificaciones principalmente sintácticas hechas a los datos que no cambian su significado, pero podrán ser requeridas por la herramienta de modelado” (Chapman, y otros, 2000).

Para poder realizar el análisis a los datos obtenidos será necesario que todos hagan referencia al mismo periodo de tiempo. Actualmente se disponen de datos mensuales y trimestrales. Para realizar la conversión de los datos trimestrales a datos mensuales se utilizara una técnica de interpolación, denominada interpolación lineal a trozos. La explicación teórica del procedimiento y la adaptación del mismo a los datos disponibles se encuentra detallada en el Anexo III.

5.4 Integración de datos

Una vez completada la interpolación de datos trimestrales, se procedió a filtrar los atributos y las tuplas que no serán considerados en el análisis de datos. La integración de datos se efectuó a través de Microsoft Excel.

5.5 Vista minable

Luego de realizar las transformaciones necesarias a los datos, previo a su procesamiento, se obtiene como resultado de estas transformaciones a la vista minable. La vista minable constituye la entrada a la siguiente etapa de minería de datos.

A continuación se presenta la descripción de los datos que se utilizarán para elaborar el modelo de datos.

Rol	Nombre	Tipo de dato	Estadísticas	Rango	Valores faltantes
regular	Índice de precios al consumidor base Abril 2008 = 100	real	avg = 129.429 +/- 19.373	[102.050 ; 166.840]	0
regular	Período	polynomial	mode = 2008/8 (1), least = 2008/8 (1)	2008/10 (1), 2008/11 (1), 2008/12 (1), 2008/8 (1), 2008/9 (1), 2009/1 (1), 2009/10 (1), 2009/11 (1), 2009/12 (1), 2009/2 (1), 2009/3 (1), 2009/4 (1), 2009/5 (1), 2009/6 (1), 2009/7 (1), 2009/8 (1), 2009/9 (1), 2010/1 (1), 2010/10 (1),	0

				2010/11 (1), 2010/12 (1), 2010/2 (1), 2010/3 (1), 2010/4 (1), 2010/5 (1), 2010/6 (1), ... and 15 more ... , 2011/8 (1), 2011/9 (1), 2012/1 (1), 2012/10 (1), 2012/11 (1), 2012/12 (1), 2012/2 (1), 2012/3 (1), 2012/4 (1), 2012/5 (1), 2012/6 (1), 2012/7 (1), 2012/8 (1), 2012/9 (1), 2013/1 (1), 2013/10 (1), 2013/11 (1), 2013/12 (1), 2013/2 (1), 2013/3 (1), 2013/4 (1), 2013/5 (1), 2013/6 (1), 2013/7 (1), 2013/8 (1), 2013/9 (1)	
regular	En poder del público	real	avg = 130168246153.846 +/- 56245356611.966	[66939000000 ; 247157000000]	0
regular	En entidades financieras	real	avg = 15561369230.769 +/- 6040997042.524	[6845000000 ; 29936000000]	0
regular	Cuenta corriente en el BCRA	real	avg = 43035138461.538 +/- 18351927901.004	[22961000000 ; 84484000000]	0
regular	Stock de pases Pasivos	real	avg = 14078384615.385 +/- 4333705886.321	[6351000000 ; 28110000000]	0
regular	LEBAC	real	avg = 49063846153.846 +/- 29603535520.605	[8161000000 ; 112745000000]	0
regular	NOBAC	real	avg = 23076676923.077 +/- 15325285178.941	[2573000000 ; 54676000000]	0
regular	Reservas internacionales del BCRA	real	avg = 46075892307.692 +/- 4905940914.223	[30612000000 ; 52427000000]	0
regular	Cuenta corriente	real	avg = 102097830769.231 +/- 37353575597.997	[53535000000 ; 190977000000]	0
regular	Caja de ahorro	real	avg = 68812723076.923 +/- 30119914487.505	[35651000000 ; 140224000000]	0
regular	Plazo fijo no ajustable por CER	real	avg = 172469676923.077 +/- 76111727648.838	[82197000000 ; 331312000000]	0
regular	Plazo fijo ajustable por CER	real	avg = 61846153.846 +/- 127717000.480	[6000000 ; 612000000]	0

regular	Otros depósitos	real	avg = 15437169230.769 +/- 4487932470.697	[10282000000 ; 26774000000]	0
regular	Depósitos del sector privado	real	avg = 250098169230.769 +/- 106899022687.530	[135862000000 ; 484435000000]	0
regular	Depósitos del sector público	real	avg = 108747738461.538 +/- 41045639240.925	[58290000000 ; 186208000000]	0
regular	Depósitos del sector privado y del sector público en dólares	real	avg = 11863676923.077 +/- 2859860027.758	[8034000000 ; 16176000000]	0
regular	Adelantos	real	avg = 30418307692.308 +/- 14493328268.900	[15738000000 ; 59604000000]	0
regular	Documentos	real	avg = 45625092307.692 +/- 26620925308.934	[18479000000 ; 111404000000]	0
regular	Hipotecarios	real	avg = 25098984615.385 +/- 8231870892.475	[16414000000 ; 43074000000]	0
regular	Prendarios	real	avg = 14167723076.923 +/- 7577164662.878	[6885000000 ; 31297000000]	0
regular	Personales	real	avg = 50841338461.538 +/- 22379228939.938	[26366000000 ; 98484000000]	0
regular	Tarjetas de crédito	real	avg = 34782400000 +/- 19281369052.955	[13208000000 ; 80411000000]	0
regular	Otros	real	avg = 15465815384.615 +/- 9897626989.733	[4802000000 ; 35427000000]	0
regular	Préstamos al sector público	real	avg = 23747369230.769 +/- 9306858063.696	[9388000000 ; 41123000000]	0
regular	Préstamos al sector privado y al sector público en dólares	real	avg = 6635815384.615 +/- 1720694317.459	[3694000000 ; 9893000000]	0
regular	Soja (por tonelada)	real	avg = 1963.678 +/- 615.078	[1043.452 ; 3107.140]	0
regular	Pellets de soja (por tonelada)	real	avg = 1809.370 +/- 646.079	[950.830 ; 3438.897]	0
regular	Aceite de soja (por tonelada)	real	avg = 4182.177 +/- 1161.293	[2212.732 ; 5697.867]	0
regular	Maíz (por tonelada)	real	avg = 906.406 +/- 347.531	[447.680 ; 1478.936]	0
regular	Trigo (por tonelada)	real	avg = 1042.124 +/- 322.072	[585.052 ; 1607.085]	0
regular	Arroz (por tonelada)	real	avg = 2297.708 +/- 342.370	[1727.670 ; 2878.831]	0
regular	Petróleo WTI (U.S. Barrels)	real	avg = 369.819 +/- 116.218	[137.300 ; 617.384]	0
regular	Petróleo Brent (U.S. Barrels)	real	avg = 408.235 +/- 147.981	[105.885 ; 700.165]	0
regular	Oro (por oz.troy)	real	avg = 5756.942 +/- 1857.812	[2532.170 ; 8270.100]	0
regular	Cobre (por tonelada)	real	avg = 31157.543 +/- 8992.473	[10733.378 ; 45584.029]	0
regular	Aluminio (por tonelada)	real	avg = 8642.900 +/- 1637.814	[4698.758 ; 10971.358]	0
regular	Tasa de desocupación	real	avg = 7.568 +/- 0.653	[6.400 ; 9.100]	0
regular	Importaciones	real	avg = 166017568.251 +/- 30199383.988	[105225458 ; 212917230]	0
regular	Consumo Individual (Hogares)	real	avg = 569832479.092 +/- 59470952.209	[476635549 ; 656310091]	0

regular	Consumo Colectivo (Público)	real	avg = 81529680.179 +/- 9456385.766	[63895100 ; 101862935]	0
regular	Exportaciones	real	avg = 150805379.800 +/- 14054763.999	[124105977 ; 174457011]	0
regular	Formación Bruta de Capital Fijo	real	avg = 166743182.569 +/- 23265933.570	[123908571 ; 204644270]	0
regular	Valor Agregado Bruto	real	avg = 696714128.595 +/- 60492656.536	[574117492 ; 793047795]	0
regular	Impuestos a los Productos Netos de Subsidios	real	avg = 41679100.174 +/- 2827920.311	[36289166 ; 47112507]	0
regular	Derechos de Importación	real	avg = 6966346.467 +/- 1283664.359	[4340905 ; 8991133]	0
regular	Impuesto al Valor Agregado	real	avg = 53169486.282 +/- 5145921.799	[44474003 ; 60389536]	0

Tabla 17. Descripción de datos de la vista minable

La descripción de datos de la vista minable se obtuvo a través de un proceso de RapidMiner que solo contenía un operador Read Excel con los datos a procesar.

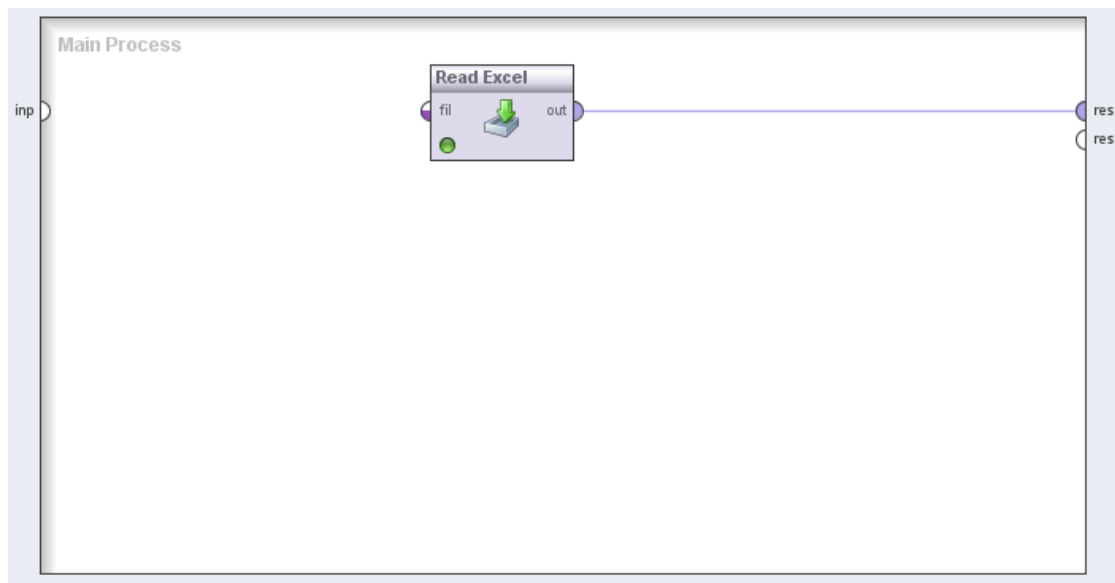


Ilustración 25. Proceso de descripción de datos de la vista minable

5.6 Resumen

La vista minable es el resultado de realizar sucesivas veces las actividades de limpieza, selección, construcción, formateo e integración de datos. Es importante hacer varios ciclos de estas tareas ya que cada iteración ayuda a mejorar la calidad de los datos y el éxito de la siguiente etapa depende de qué tan bien preparados están los datos.

6 Modelado

En esta sección se desarrollan los conceptos utilizados para la construcción y evaluación del modelo de datos, los diferentes procesos realizados para elaborar los modelos, las evaluaciones de los modelos y las conclusiones de los resultados obtenidos.

6.1 Ventanas de tiempo y validación de ventanas deslizantes

Antes de explicar los procesos que se implementaron, es necesario mencionar dos elementos importantes utilizados para elaborar y probar el modelo de datos, estos son la generación de ventanas de tiempo y la validación de ventanas deslizantes.

Dado un conjunto inicial de ejemplos, se pueden generar nuevos ejemplos utilizando ventanas de tiempo. Luego, con los ejemplos generados se utiliza la técnica de validación de ventanas deslizantes para elaborar y evaluar el modelo de datos.

Generar ventanas de tiempo es importante ya que esta reestructuración de los datos posibilita realizar predicciones a futuro de la serie de tiempo, por ejemplo tomando los datos de diferentes variables económicas de los meses de Enero y Febrero se podría predecir el índice de precios del mes de Marzo. Por otro lado la validación de ventanas deslizantes permite utilizar los ejemplos generados para entrenar y probar el modelo a través de distintas técnicas de evaluación de rendimiento. En las siguientes secciones se explican con más profundidad la generación de ventanas de tiempo y la validación de ventanas deslizantes.

6.1.1 Ventanas de tiempo en RapidMiner

RapidMiner permite generar ventanas de series de tiempo para series univariadas (con un solo atributo) o multivariadas (con varios atributos). Para crear ventanas de tiempos de series multivariadas es necesario emplear el operador Windowing.

El operador Windowing se utiliza para generar ejemplos a partir de un conjunto de ejemplos iniciales que contienen los datos de una serie de tiempo. Los nuevos ejemplos se crean al ir desplazando una ventana sobre el conjunto de datos inicial. Esta reestructuración de datos es útil para poder posteriormente hacer predicciones de la serie de tiempo.

Supongamos que tenemos un conjunto de datos inicial en que hay dos atributos (Cuenta corriente y Caja de ahorro) y un atributo que se desea predecir (Índice de precios). Trabajando con ventanas de tamaño 2, para predecir el índice de precios en el instante t , se utilizan los valores de los atributos Cuenta Corriente, Caja de ahorro e Índice de precios de los instantes $t-1$ y $t-2$. Al aplicar el operador Windowing, Rapidminer genera el conjunto de atributos desplazados en el tiempo, en este caso Cuenta corriente-0, Cuenta corriente-1, Caja de ahorro-0, Caja de ahorro-1, Índice de precios-0 e Índice de precios-1 que se corresponden con los instantes $t-1$ y $t-2$ respectivamente.

Conjunto de ejemplos iniciales

Row No.	Período	Cuenta corriente	Caja de ahorro	Índice de precios
1	2008/8	53535000000	36510000000	102.050
2	2008/9	59194000000	36409000000	102.570
3	2008/10	58825000000	36069000000	103.010
4	2008/11	56880000000	35651000000	103.360
5	2008/12	64915000000	37356000000	103.710
6	2009/1	67894000000	37917000000	104.260
7	2009/2	63247000000	37529000000	104.710
8	2009/3	61662000000	36684000000	105.380
9	2009/4	61108000000	37254000000	105.730
10	2009/5	62751000000	37440000000	106.080
11	2009/6	65389000000	38809000000	106.530
12	2009/7	61709000000	40442000000	107.190
13	2009/8	62861000000	38880000000	108.080
14	2009/9	66124000000	39192000000	108.880
15	2009/10	65681000000	39482000000	109.750
16	2009/11	65919000000	39954000000	110.660
17	2009/12	71844000000	43347000000	111.690
18	2010/1	74987000000	44248000000	112.850

Conjunto de ejemplos generados

Período	label	Cuenta corriente-1	Cuenta corriente-0	Caja de ahorro-1	Caja de ahorro-0	Índice de precios-1	Índice de precios-0	
2008/9	103.010	53535000000	59194000000	36510000000	36409000000	102.050	102.570	
2008/10	103.360	59194000000	58825000000	36409000000	36069000000	102.570	103.010	
2008/11	103.710	58825000000	56880000000	36069000000	35651000000	103.010	103.360	
2008/12	104.260	56880000000	64915000000	35651000000	37356000000	103.360	103.710	
2009/1	104.710	64915000000	67894000000	37356000000	37917000000	103.710	104.260	Predicción Febrero
2009/2	105.380	67894000000	63247000000	37917000000	37529000000	104.260	104.710	Marzo
2009/3	105.730	63247000000	61662000000	37529000000	36684000000	104.710	105.380	Abril
2009/4	106.080	61662000000	61108000000	36684000000	37254000000	105.380	105.730	
2009/5	106.530	61108000000	62751000000	37254000000	37440000000	105.730	106.080	
2009/6	107.190	62751000000	65389000000	37440000000	38809000000	106.080	106.530	
2009/7	108.080	65389000000	61709000000	38809000000	40442000000	106.530	107.190	
2009/8	108.880	61709000000	62861000000	40442000000	38880000000	107.190	108.080	
2009/9	109.750	62861000000	66124000000	38880000000	39192000000	108.080	108.880	
2009/10	110.660	66124000000	65681000000	39192000000	39482000000	108.880	109.750	
2009/11	111.690	65681000000	65919000000	39482000000	39954000000	109.750	110.660	
2009/12	112.850	65919000000	71844000000	39954000000	43347000000	110.660	111.690	
2010/1	114.260	71844000000	74987000000	43347000000	44248000000	111.690	112.850	
2010/2	115.560	74987000000	75261000000	44248000000	44357000000	112.850	114.260	

Ilustración 26. Construcción de ventanas de tamaño 2 para un conjunto de datos

El par de atributos Cuenta corriente-1 y Cuenta corriente-0, se corresponden con los valores de Cuenta corriente en los dos ejemplos de la ventana. Lo mismo ocurre con los otros atributos, a excepción de Período, ya que este último actúa como identificador del conjunto de ejemplos.

El conjunto de ejemplos generado además contiene un atributo denominado label. Los valores de este atributo coinciden con los valores del atributo que se desea predecir, el Índice de precios. El período del índice de precios que toma cada ventana depende del valor que se defina como horizonte. En este caso el valor del horizonte es de 1, por eso es que cada ventana toma como valor de label al valor del índice de precios en el instante t. Si el valor del horizonte fuera de 2, el valor de label correspondería al valor del índice de precios en el instante t+1.

Dado que el tamaño de la ventana es de 2, para predecir el índice de precios de un mes se utilizarán los datos de los dos meses anteriores, así por ejemplo se utilizan los datos de Enero y Febrero para predecir el valor del índice de precios del mes de Marzo. Si el tamaño de la ventana fuera de 3, se utilizarían los datos de Diciembre, Enero y Febrero para predecir el valor de Marzo.

6.1.2 Validación de ventanas deslizantes en RapidMiner

Antes de seleccionar el modelo adecuado, es necesario definir un procedimiento para probar la calidad de las diversas alternativas. Por ejemplo, en tareas de minería de datos supervisadas como la clasificación, es común usar tasas de error como medidas de calidad para modelos de minería de datos. Por lo tanto, el diseño de prueba especifica que el conjunto de datos debería ser separado en el entrenamiento y en el conjunto de prueba. El modelo está construido sobre el conjunto de entrenamiento y su calidad estimada sobre el conjunto de prueba (Chapman, y otros, 2000).

En este caso, para evaluar la calidad del modelo de datos elaborado será necesario utilizar una técnica de evaluación de modelos predictivos que se ajuste al método de series de tiempo que se aplicará. En este sentido, la validación con ventanas deslizantes se aplica a la situación en cuestión. La validación con ventanas deslizantes utiliza una cierta ventana de ejemplos para el entrenamiento y utiliza otra ventana para las pruebas. El proceso se realiza sucesivas veces utilizando todos los ejemplos. Sea T la cantidad total de ejemplos, E la cantidad de ejemplos para el entrenamiento, P la cantidad de ejemplos de prueba y D el tamaño del deslizamiento de la ventana de entrenamiento, siendo $E + P \leq T$; en la primera iteración, los primeros E ejemplos son utilizados para el entrenamiento y los siguientes P para la prueba. Luego, en la segunda iteración, la ventana de entrenamiento avanza sobre D ejemplos, mientras que la ventana de prueba se desplaza sobre 1 ejemplo, así el conjunto de entrenamiento estará formado desde el primer ejemplo al ejemplo $E+D$, mientras que el conjunto de prueba incluirá los P ejemplos a partir del ejemplo $E+D$. El proceso se repite hasta utilizar el último ejemplo en el conjunto de prueba. Las mediciones de rendimiento (error cuadrático medio, error absoluto, u otra) se calculan en cada iteración y se promedian al final. Este promedio indicará qué tan acertado es el modelo de datos en relación al conjunto de datos. En la siguiente ilustración se puede apreciar cómo son utilizados los ejemplos para realizar el entrenamiento y la prueba del modelo, en este caso el conjunto de entrenamiento está formado por 6 ejemplos y el de prueba por 2, mientras que el valor del tamaño del deslizamiento de la ventana de entrenamiento es de 1. Estos valores corresponden a los parámetros “training window width”, “test window width” y “training window step size” del operador Sliding Window Validation en RapidMiner.

Iteración 1	Iteración 2	Iteración 3
Ejemplo 1	Ejemplo 1	Ejemplo 1
Ejemplo 2	Ejemplo 2	Ejemplo 2
Ejemplo 3	Ejemplo 3	Ejemplo 3
Ejemplo 4	Ejemplo 4	Ejemplo 4
Ejemplo 5	Ejemplo 5	Ejemplo 5
Ejemplo 6	Ejemplo 6	Ejemplo 6
Ejemplo 7	Ejemplo 7	Ejemplo 7
Ejemplo 8	Ejemplo 8	Ejemplo 8
Ejemplo 9	Ejemplo 9	Ejemplo 9
Ejemplo 10	Ejemplo 10	Ejemplo 10

Conjunto de entrenamiento – Conjunto de prueba

Ilustración 27. Iteraciones de una validación de ventanas deslizantes

El operador mencionado tiene un cuarto parámetro, “horizon”, que determina cual será el primer ejemplo de la ventana de entrenamiento. En el caso anterior su valor es de 1, ya que la ventana de prueba empieza a partir del primer ejemplo luego de la ventana de entrenamiento. Si su valor fuera de 2, las ventanas de prueba abarcarían los ejemplos a partir del segundo ejemplo siguiente a la ventana de entrenamiento.

6.2 Construcción del modelo

La construcción del modelo se realizó en varias iteraciones, seguidamente se presentan los diferentes procesos elaborados y los resultados obtenidos para mostrar cómo se fueron escogiendo los operadores y parámetros para obtener el mejor modelo de datos. El conjunto de datos utilizados cuenta con 65 registros, que corresponden al periodo entre Agosto del 2008 y Diciembre del 2013, y 47 atributos.

6.2.1 Primer proceso

El primer proceso se elaboró con todos los datos disponibles, sin filtrar atributos, a través del siguiente proceso. El objetivo de este experimento era analizar si resultaba necesario filtrar algunos atributos o se podría obtener un buen modelo con todos los datos. Una restricción del proyecto es la capacidad de procesamiento del equipo informático disponible y muchos datos para procesar puede que ocasionen algún inconveniente. Para definir los parámetros de los operadores se prestó principal atención al operador de redes neuronales y la cantidad de ventanas generadas a través de Windowing.

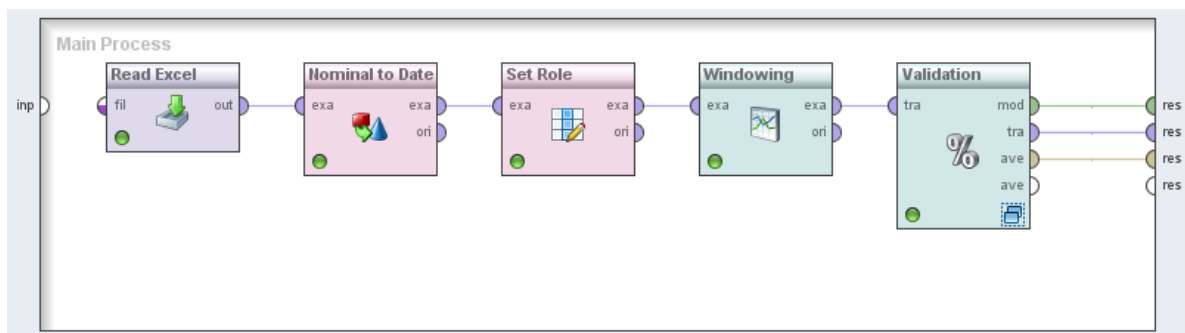


Ilustración 28. Primer proceso para la construcción del modelo

En este proceso, los datos son leídos a través del operador Read Excel, luego en el operador Nominal to Date convierte los datos del atributo “Periodo” a un tipo de dato “fecha”. Los parámetros principales del operador Nominal to Date son los siguientes:

- Attribute name: Período
- Data type: date
- Date format: yyyy/MM

Posteriormente el operador Set Role define al atributo Periodo como identificador del conjunto de datos, los valores de las variables de este operador son los siguientes:

- Attribute name: Período
- Target role: id

Siguiendo con el proceso, el operador Windowing crea una serie de ejemplos a partir de una serie temporal. En este operador se definieron los siguientes parámetros:

- Series representation: encode_series_by_examples. De esta manera cada registro del conjunto de datos es un ejemplo de la serie temporal y corresponde a distintos instantes de tiempo de la serie.
- Window size: especifica el tamaño de las ventanas, es decir la cantidad de períodos de la serie temporal que se tomarán para construir un ejemplo del conjunto de entrenamiento, como se explicó en la Sección 6.1.1. En principio el tamaño de la ventana será de 1, luego definido todos los parámetros de los operadores del proceso, también se ejecutará el proceso para ventanas de tamaño 2 y 3.
- Step size: define el tamaño del desplazamiento de la ventana en cada iteración. Para aprovechar mejor los datos se utilizará el valor de 1.
- Horizon: determina la distancia entre el último valor de la ventana y el valor a predecir. En este caso será de 1, ya que se desea estimar el valor más próximo a la ventana.
- Create single attributes: true
- Create label: true
- Select label by dimension: false
- Label attribute: Índice de precios al consumidor base Abril 2008 = 100
- Add incomplete windows: false
- Stop on too small dataset: true

Finalmente, las variables de Sliding Window Validation (véase Sección 6.1.2) son las siguientes:

- Training window width: 20. Determina la cantidad de ejemplos que se utilizarán para el entrenamiento en cada iteración.
- Training window step size: 1. Define el número de ejemplos que la ventana se desliza en cada iteración.
- Test window width: 5. Determina la cantidad de ejemplos que se utilizarán para la prueba en cada iteración.
- Horizon: 1. Especifica el incremento desde el último ejemplo de la ventana de entrenamiento al primer ejemplo de la ventana de prueba.
- Cumulative training: true
- Average performances only: true

El operador Sliding Window Validation es anidado, en la siguiente imagen se ilustra el proceso que contiene este.

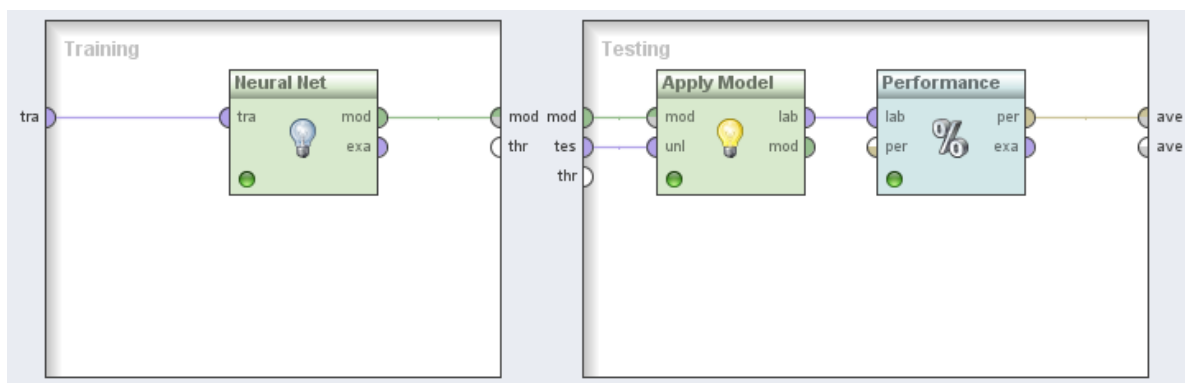


Ilustración 29. Subproceso del operador Sliding Window Validation

En el subproceso de entrenamiento se colocó un operador Neural Net, mientras que en el proceso de prueba se colocó un operador Apply Model y un Forecasting Performance. El objetivo de esto es generar el modelo con el operador de redes neuronales, luego aplicarlo y evaluar los resultados del modelo en el conjunto de prueba.

Los parámetros de Apply Model son los definidos por defecto y los de Forecasting Performance son los siguientes:

- Horizon: 1. Define la distancia para el cálculo de las medidas de rendimiento de la predicción, su valor debe ser menor al tamaño de la ventana de entrenamiento.
- Main criterion: prediction-trend-accuracy. Determina el criterio utilizado para la comparación de los vectores de rendimiento. Prediction trend accuracy mide la media de veces que una predicción de regresión fue capaz de predecir correctamente la tendencia de la regresión, es decir el promedio de veces que el modelo estimó adecuadamente el comportamiento (crecimiento o disminución) de la serie a través del tiempo.
- Skip undefined labels: true
- Use example weights: true

A continuación se presentan las razones de la elección de los valores de las distintas variables de Neural Net.

6.2.1.1 Selección de los parámetros para el operador Neural Net

Para seleccionar los valores de las variables se recurrió a distintas fuentes de conocimiento, en la mayoría de los casos se partió de la experiencia de expertos, en otros casos se consideró los valores definidos por defecto en RapidMiner, además de realizar distintas pruebas para comparar los resultados obtenidos. Seguidamente se expondrán los valores de cada parámetro con su justificación correspondiente:

- Hidden layers. Este parámetro describe el nombre y el tamaño de todas las capas ocultas. Para este proceso se utilizó una única capa oculta. Según (Hernández Orallo, Ramírez Quintana, & Ferri Ramírez, 2004) uno de los aspectos más complejos es determinar el número de neuronas de la capa oculta, ya que pocas neuronas pueden hacer que el concepto a aprender no se pueda capturar mientras que demasiadas neuronas pueden sobreajustar los datos.

Para este parámetro RapidMiner define por defecto una capa oculta con una cantidad de nodos equivalente a $(\text{número de atributos} + \text{número de clases}) / 2 + 1$, en este caso el número de atributos es equivalente a 50 y el número de clases es de 2, por lo que debería generar una capa de 27 nodos. Ya que cada capa oculta añade un nodo más que no está conectado a la capa anterior, en este caso la capa de entrada, pero sí a la capa siguiente, la de salida, la cantidad de nodos ocultos final es de 28.

- Training cycles. Este parámetro especifica la cantidad de veces que se ajustarán los pesos de cada conexión luego de comparar los valores de salida y de las respuestas correctas. En este caso también se optó por su valor por defecto que es de 500.
- Learning rate y momentum

En (Martín Diazaraque, 2008), se menciona lo siguiente sobre estos parámetros: el valor de la tasa de aprendizaje (η) controla el tamaño del cambio de los pesos en cada iteración. Se deben evitar dos extremos: un ritmo de aprendizaje demasiado pequeño puede ocasionar una disminución importante en la velocidad de convergencia y la posibilidad de acabar atrapado en un mínimo local; en cambio, un ritmo de aprendizaje

demasiado grande puede conducir a inestabilidades en la función de error, lo cual evitará que se produzca la convergencia debido a que se darán saltos en torno al mínimo sin alcanzarlo.

Por tanto, se recomienda elegir un ritmo de aprendizaje lo más grande posible sin que provoque grandes oscilaciones. En general, el valor de la tasa de aprendizaje suele estar comprendida entre 0.05 y 0.5. El factor momento (α) acelera la convergencia de los pesos. Suele tomar un valor próximo a 1 (por ejemplo, 0.9).

En RapidMiner los valores por defecto son 0.3 para learning rate y 0.2 para momentum. Luego de efectuar algunas pruebas trabajando con una ventana de tamaño 1 del operador Windowing, se llegó a la conclusión que los valores de Learning rate: 0.5 y Momentum: 0.4 son los que dan mejores resultados.

- Decay. Indica si el coeficiente de aprendizaje (learning rate) debe reducirse durante la etapa de aprendizaje, por defecto su valor es falso. En este caso su valor será de verdadero.
- Shuffle. Indica si los datos deben ser barajados antes del aprendizaje. A pesar de la utilización de memoria se recomienda realizar este proceso. Se utilizará su valor por defecto que es verdadero.
- Normalize. El operador de red neuronal utiliza una función sigmoide habitual como la función de activación. Por lo tanto, el rango de valores de los atributos debe hacerse a escala en -1 y +1. Esto se puede hacer mediante el parámetro Normalize. La normalización se realiza antes del aprendizaje. A pesar de que aumenta el tiempo de ejecución, es necesario en la mayoría de los casos. Su valor por defecto, verdadero, será utilizado
- Error épsilon. Valor por el cual la optimización se detiene si el error es menor al valor indicado. Se utiliza el valor definido por defecto 1.0E-5.
- Use local random seed. Indica si una semilla aleatoria local debe ser utilizada para la aleatorización. Se utilizará su valor por defecto que es false.

6.2.1.2 Resultados del primer proceso

Los resultados de este primer proceso se resumen en la siguiente tabla

Tamaño de la ventana	Numero de ejemplos	Número de atributos regulares	Performance	Tiempo de ejecución
1	64	46	0.581 +/- 0.482	31 segundos
2	63	92	0.474 +/- 0.490	1 minuto y 47 segundos
3	62	138	-	-

Tabla 18. Resultados de los rendimientos de los modelos elaborados en el primer proceso

Para obtener los resultados, en cada iteración del operador Sliding Window Validation se calcula la media de veces que una predicción acertó la tendencia de la regresión, luego estos valores se promedian y se calcula el desvío estándar.

Apreciando la tabla se puede sacar algunas conclusiones, al ser más grande el tamaño de la ventana, mayor será la cantidad de los atributos regulares como se mencionó anteriormente. Mientras mayor es la cantidad de atributos, el proceso requerirá mayores recursos de procesamiento y demandará más tiempo para devolver los resultados. Sin embargo, trabajar con una ventana más grande no asegura que el modelo generado sea mejor.

En cuanto a los resultados obtenidos de rendimiento del modelo no son muy buenos, 0.581 y 0.474 indican que en el 58,1% (cuando el tamaño de la ventana es de 1) y 47,4% (cuando el

tamaño de la ventana es de 2) de los casos se estimó correctamente la tendencia o dirección general de la serie a través del tiempo. Lo que se está analizando en este caso no es la precisión del modelo en cuanto a los valores del índice de precios, sino el comportamiento de la tendencia, es decir, si el índice de precios crece o decrece de un periodo al otro. Para analizar la precisión del modelo en relación a los valores del índice de precios es necesario otro tipo de criterios de evaluación del modelo, como ser el error cuadrático medio, el error absoluto o el error relativo. Estos criterios se tendrán en cuenta en el próximo proceso.

Los valores de 0.482 y 0.490 indican el desvío estándar de las medias de veces que una predicción estimó correctamente la tendencia de la regresión. Estos valores indican que los promedios de aciertos sobre el comportamiento de la serie en cada iteración son muy dispersos, es decir existen iteraciones que son muy precisos (con valores cercanos a 1) y otras que son muy imprecisos (con valores cercanos a 0).

Para una ventana de tamaño 3, la ejecución superaba la hora y no generaba los resultados.

6.2.2 Segundo proceso

En el segundo proceso se decidió por filtrar algunos atributos para poder hacer un análisis con una ventana de tamaño 3. Para filtrar los atributos se utilizaron diferentes operadores que asignan un peso a cada atributo, una explicación más detallada de cada operador se encuentra en el Anexo IV. Los operadores utilizados fueron los siguientes:

- SP-Weight by SVM: calcula la relevancia de los atributos a través de algoritmos de máquina de vectores soporte.
- SP-Weight by Deviation: define la importancia de los atributos en base a sus desvíos estándar.
- SP-Weight by PCA: utiliza el análisis de componentes principales (PCA) para establecer la relevancia de los atributos
- SP-Weight by Relief: la calidad de un atributo es estimada de acuerdo a qué tan bien los valores del atributo permiten diferenciar los valores del atributo objetivo.

A fin de generar los valores de los operadores mencionados en un mismo proceso fue necesario utilizar un Multiply que asigna su objeto de entrada a distintos subprocesos. Cada subproceso genera los pesos de los atributos según una de las técnicas mencionadas.

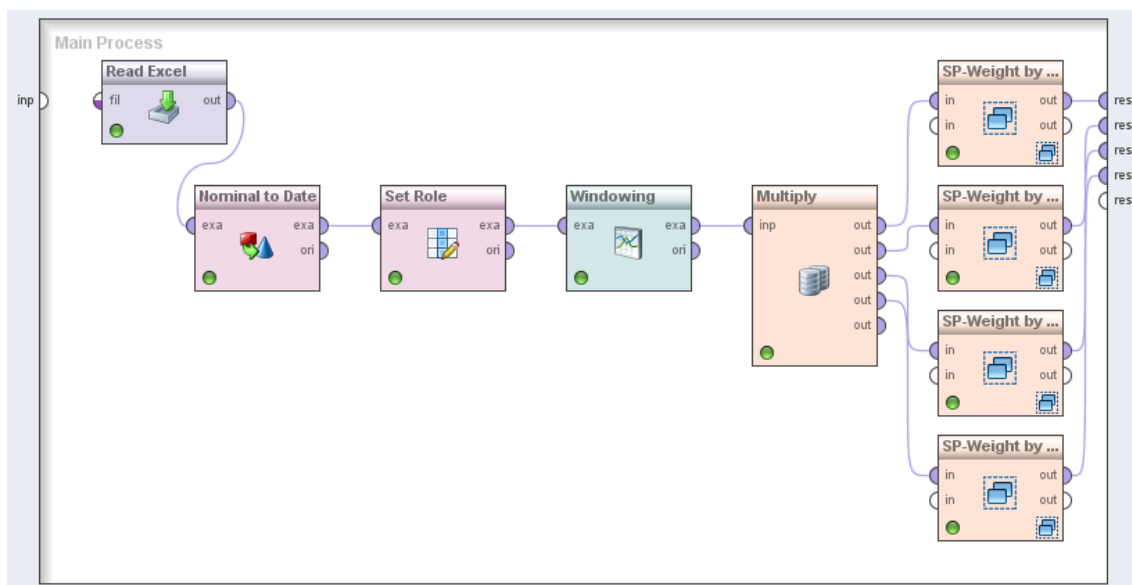


Ilustración 30. Proceso de selección de atributos

La lógica de cada subproceso es la misma, primero los atributos son ponderados según la técnica seleccionada, luego los pesos de los atributos son convertidos en datos con el operador Weights to data para poder manipularlos posteriormente.

Hasta el momento lo que se está ponderando son grupos de atributos que corresponden a valores de un atributo en 3 instantes de tiempo ya que se está trabajando con una ventana de tamaño 3. Así por ejemplo en este momento se está estimando la importancia de LEBAC-0, LEBAC-1 y LEBAC-2, estos nuevos atributos corresponden a valores del atributo LEBAC en tres periodos de tiempo, el atributo LEBAC-2 corresponde al primero, LEBAC-1 al segundo y finalmente LEBAC-0 al último. La siguiente imagen que corresponde al conjunto de salida de un operador Weights to Data.

Row No.	Attribute	Weight ▼
1	LEBAC-0	1
2	LEBAC-1	0.964
3	Préstamos al sector público-2	0.946
4	Préstamos al sector público-1	0.932
5	Préstamos al sector público-0	0.913
6	LEBAC-2	0.910
7	Índice de precios al consumidor base Abril 2008 = 100-0	0.880
8	Índice de precios al consumidor base Abril 2008 = 100-1	0.876
9	Petróleo WTI (U.S. Barrels)-0	0.876
10	Índice de precios al consumidor base Abril 2008 = 100-2	0.873
11	Tarjetas de crédito-2	0.854
12	Tarjetas de crédito-1	0.846
13	Petróleo WTI (U.S. Barrels)-1	0.840
14	Cuenta corriente-2	0.838
15	Oro (por oz.troy)-2	0.836

Ilustración 31. Datos de salida de un operador Weights to Data

Para poder obtener la ponderación sobre el atributo independientemente del instante de tiempo, que en este caso sería LEBAC se añadió un operador Generate attribute que genera un nuevo atributo denominado “Atributo”, a partir del atributo “Attribute” de la imagen anterior, cuya expresión matemática es “cut(Attribute,0,index(Attribute,"-"))”. El objetivo de esta operación es cortar de cada “Attribute” los dos últimos caracteres, que en este caso toman tres valores “-0”, “-1” y “-2”. En la siguiente imagen se puede apreciar el nuevo atributo generado.

Row No.	Attribute	Atributo	Weight ▼
1	LEBAC-0	LEBAC	1
2	LEBAC-1	LEBAC	0.964
3	Préstamos al sector público-2	Préstamos al sector público	0.946
4	Préstamos al sector público-1	Préstamos al sector público	0.932
5	Préstamos al sector público-0	Préstamos al sector público	0.913
6	LEBAC-2	LEBAC	0.910
7	Índice de precios al consumidor base Abril 2008 = 100-0	Índice de precios al consumidor base Abril 2008 = 100	0.880
8	Índice de precios al consumidor base Abril 2008 = 100-1	Índice de precios al consumidor base Abril 2008 = 100	0.876
9	Petróleo WTI (U.S. Barrels)-0	Petróleo WTI (U.S. Barrels)	0.876
10	Índice de precios al consumidor base Abril 2008 = 100-2	Índice de precios al consumidor base Abril 2008 = 100	0.873
11	Tarjetas de crédito-2	Tarjetas de crédito	0.854
12	Tarjetas de crédito-1	Tarjetas de crédito	0.846
13	Petróleo WTI (U.S. Barrels)-1	Petróleo WTI (U.S. Barrels)	0.840
14	Cuenta corriente-2	Cuenta corriente	0.838
15	Oro (por oz.troy)-2	Oro (por oz.troy)	0.836

Ilustración 32. Datos de salida de un operador Generate Attribute

Finalmente, el operador Aggregate realiza la operación suma del atributo Weight, los pesos generados por el operador Weight correspondiente, agrupando los datos según el nuevo atributo.

A continuación se muestran los subprocesos del proceso anterior.

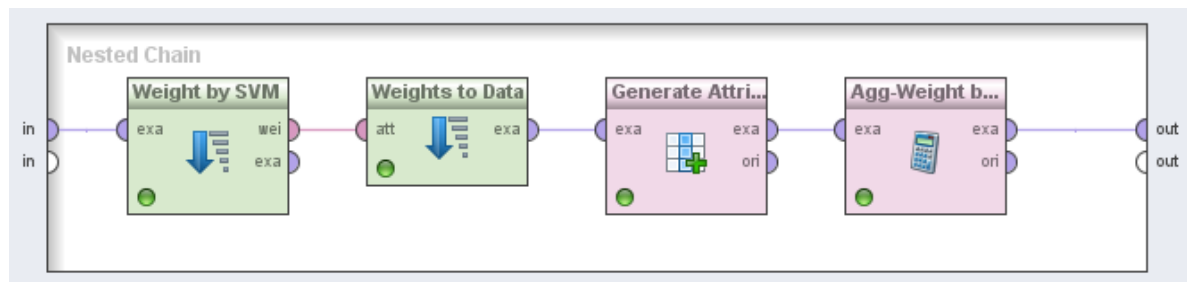


Ilustración 33. Subproceso del operador SP-Weight by SVM

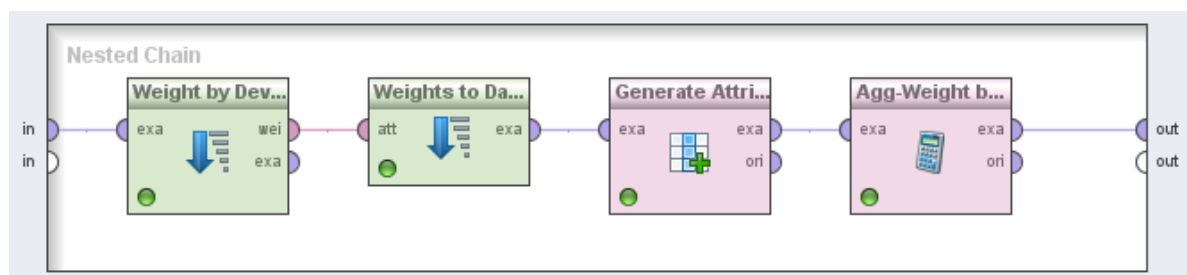


Ilustración 34. Subproceso del operador SP-Weight by Deviation

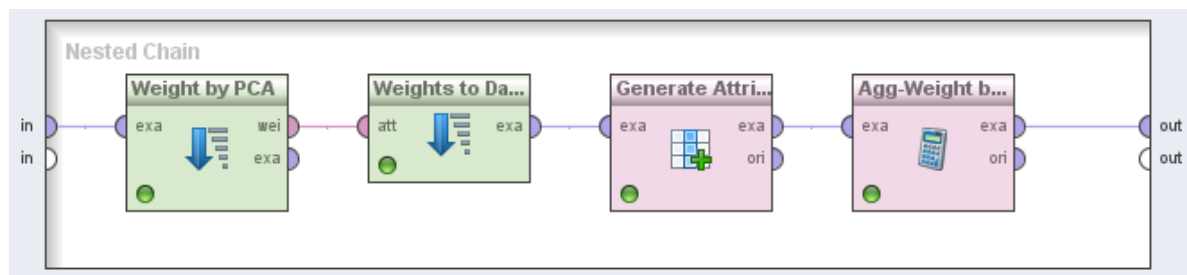


Ilustración 35. Subproceso del operador SP-Weight by PCA

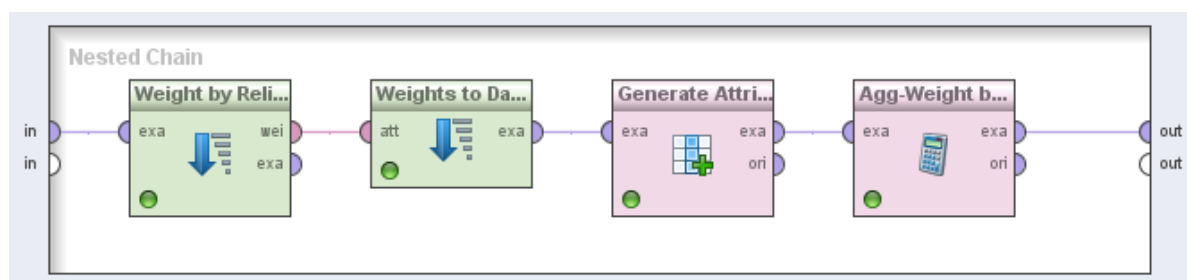


Ilustración 36. Subproceso del operador SP-Weight by Relief

Los diez atributos más ponderados por cada técnica se muestran en las siguientes imágenes.

Row No.	Atributo	sum(Weight) ▼
24	LEBAC	2.874
38	Préstamos al sector público	2.791
46	Índice de precios al consumidor base Abril 2008 = 100	2.629
33	Petróleo WTI (U.S. Barrels)	2.536
42	Tarjetas de crédito	2.532
9	Cuenta corriente	2.472
16	En entidades financieras	2.470
17	En poder del público	2.415
31	Personales	2.407
11	Depósitos del sector privado	2.405

Ilustración 37. Diez mejores atributos ponderados por Weight by SVM

Row No.	Atributo	sum(Weight) ▼
11	Depósitos del sector privado	2.911
35	Plazo fijo no ajustable por CER	2.103
17	En poder del público	1.544
13	Depósitos del sector público	1.136
9	Cuenta corriente	1.009
5	Caja de ahorro	0.817
24	LEBAC	0.815
15	Documentos	0.711
31	Personales	0.610
42	Tarjetas de crédito	0.519

Ilustración 38. Diez mejores atributos ponderados por Weight by Deviation

Row No.	Atributo	sum(Weight) ▼
11	Depósitos del sector privado	2.910
35	Plazo fijo no ajustable por CER	2.095
17	En poder del público	1.542
13	Depósitos del sector público	1.119
9	Cuenta corriente	1.002
5	Caja de ahorro	0.814
24	LEBAC	0.714
15	Documentos	0.706
31	Personales	0.609
42	Tarjetas de crédito	0.518

Ilustración 39. Diez mejores atributos ponderados por Weight by PCA

Row No.	Atributo	sum(Weight) ▼
18	Exportaciones	2.771
39	Reservas internacionales del BCRA	2.759
23	Impuestos a los Productos Netos de Subsidios	2.568
43	Tasa de desocupación	2.481
37	Préstamos al sector privado y al sector público en dólares	2.088
30	Pellets de soja (por tonelada)	1.978
26	NOBAC	1.885
36	Prendarios	1.856
24	LEBAC	1.832
19	Formación Bruta de Capital Fijo	1.808

Ilustración 40. Diez mejores atributos ponderados por Weight by Relief

Luego de obtener los atributos más relevantes se procedió a realizar el siguiente proceso, para elaborarlo se escogieron los 5 mejores atributos de cada técnica. La parte superior de este es similar al primero que contaba con todos los atributos. A diferencia de él se ingresó un operador Select Attribute para seleccionar los atributos más relevantes.

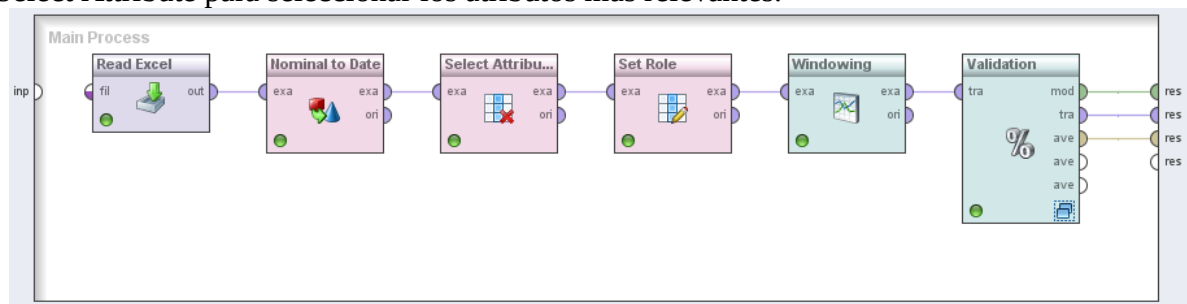


Ilustración 41. Segundo proceso para la construcción del modelo

El subproceso del operador Validation es el que se muestra a continuación.

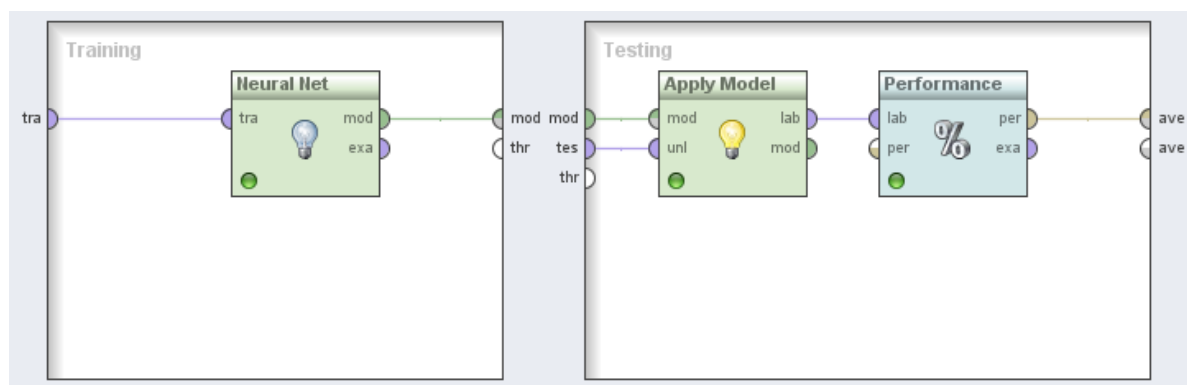


Ilustración 42. Subproceso del operador Sliding Window Validation

En este subproceso el operador Neural Net cuenta con los siguientes parámetros:

- Hidden layers: 0
- Training cycles: 500
- Learning rate: 0.3
- Momentum: 0.2
- Decay: false
- Shuffle: true
- Normalize: true

- Error epsilon: 1.0E-5
- Use local random seed: false

Luego de ejecutar el proceso para las 4 tipos de ponderación de atributos, se reemplazó el operador Forecasting Performance del subproceso de Sliding Window Validation por un operador Performance (Regression) a fin de evaluar el modelo con otros criterios de evaluación. Los criterios seleccionados en este operador fueron los siguientes:

- `root_mean_squared_error`: calcula el error cuadrático medio de las predicciones en relación a los valores reales.
- `absolute_error`: calcula el error absoluto de las predicciones en relación a los valores reales.
- `relative_error`: calcula el error relativo de las predicciones en relación a los valores reales.

6.2.2.1 Resultados

Los resultados del proceso se resumen en la siguiente tabla.

Selección de atributos Criterio de performance	SVM	Deviation	PCA	Relief
<code>prediction_trend_accuracy</code>	0.566 +/- 0.375	0.691 +/- 0.332	0.691 +/- 0.332	0.579 +/- 0.349
<code>root_mean_squared_error</code>	1.135 +/- 0.566	0.916 +/- 0.426	0.916 +/- 0.426	1.253 +/- 0.825
<code>absolute_error</code>	1.023 +/- 0.552	0.821 +/- 0.416	0.821 +/- 0.416	1.083 +/- 0.760
<code>relative_error</code>	0.72% +/- 0.38%	0.58% +/- 0.28%	0.58% +/- 0.28%	0.76% +/- 0.52%

Tabla 19. Resultados de los rendimientos de los modelos elaborados en el segundo proceso

Los 5 mejores atributos seleccionados por Weight by Deviation y Weight by PCA son los mismos, por eso tienen idénticos resultados en los criterios de performance. Según estos criterios el modelo generado con los atributos escogidos por los mencionados operadores es el mejor. Este modelo tiene el `prediction_trend_accuracy` más alto, por lo que estima mejor la tendencia de la regresión. En cuanto a las mediciones de los errores, tiene los valores más bajos, por lo que los valores predichos son muy cercanos a los valores reales.

Los indicadores de error cuadrático medio, error absoluto y error relativo permiten apreciar la diferencia entre los valores reales y los valores que el modelo predice.

El error cuadrático medio (RMSE o root mean squared error) de un estimador mide el promedio de los errores al cuadrado, es decir, la diferencia entre el estimador y lo que se estima. Sean $\bar{x}_i$ el conjunto de predicciones y x_i el conjunto de valores verdaderos, entonces el error cuadrático medio está definido por:

$$RMSE = \frac{1}{n} \sum_{i=1}^n (\bar{x}_i - x_i)^2$$

Por otro lado, el error absoluto es la diferencia entre el valor real y el valor estimado. Finalmente, el error relativo es el cociente entre el error absoluto y el valor exacto. Sus fórmulas son las siguientes:

$$Error\ absoluto = \frac{1}{n} \sum_{i=1}^n |\bar{x}_i - x_i|$$

$$Error\ relativo = \frac{1}{n} \sum_{i=1}^n \frac{|\bar{x}_i - x_i|}{\bar{x}_i}$$

En el mejor modelo generado, el error absoluto es del orden de 0.8 en datos que son de orden 100, esto indica que los valores estimados por el modelo son muy buenos. A la misma conclusión se puede llegar analizando los valores obtenidos de error cuadrático medio y error relativo.

Si bien el modelo generado aparenta ser bueno, resulta llamativo que no pueda predecir con una exactitud del 100% la tendencia de la regresión, ya que siempre el índice de precios al consumidor aumenta. Para comparar el rendimiento de las redes neuronales con otras técnicas de predicción se ejecutó el proceso anterior, reemplazando el operador de redes neuronales con operadores de otras técnicas de predicción. Estas técnicas son las máquinas de vectores soporte (Support Vector Machine) y la regresión lineal. En el Anexo VI se encuentra una explicación del funcionamiento de las máquinas de vectores soporte, mientras que en el Anexo V encontrará una descripción de la regresión lineal. El subconjunto de atributos utilizados para realizar los modelos son En poder del público, Cuenta corriente, Plazo fijo no ajustable por CER, Depósitos del sector público y Depósitos del sector privado, además del Índice de precios al consumidor base Abril 2008 = 100 y Período.

Los resultados obtenidos se presentan en la siguiente tabla.

Técnica utilizada Criterio de performance	Redes neuronales	SVM	Regresión lineal
prediction_trend_accuracy	0.691 +/- 0.332	1.000 +/- 0.000	0.914 +/- 0.175
root_mean_squared_error	0.916 +/- 0.426	1.700 +/- 1.332	3.778 +/- 1.392
absolute_error	0.821 +/- 0.416	1.583 +/- 1.264	3.363 +/- 1.286
relative_error	0.58% +/- 0.28%	1.18% +/- 1.02%	2.42% +/- 1.06%

Tabla 20. Comparación del rendimiento de las redes neuronales con otras técnicas

Observando los valores promedio de los errores pareciera que las redes neuronales predican valores más cercanos a los reales ya que sus valores promedios son los más pequeños. Sin embargo, analizando los intervalos de confianza del error cuadrático medio para las redes neuronales y las máquinas de vectores soporte, se puede apreciar que las máquinas de vectores soporte tienen un intervalo más grande (0.368; 3.032), mientras que el de las redes neuronales es más acotado (0.49; 1.342). El intervalo de las máquinas de vectores soporte tiene valores más cercanos a cero, esto quiere decir que en algunas predicciones los valores estimados por

las máquinas de vectores soportes son más cercanos a los reales que los pronosticados por las redes neuronales.

Por otro lado, en las predicciones de la tendencia de la serie de tiempo las redes neuronales tienen los peores resultados de las tres técnicas.

Con el objetivo de continuar analizando los resultados, seguidamente se presenta un gráfico comparativo, en el que se puede considerar el comportamiento del índice de precios y las predicciones de los modelos obtenidos. Los datos utilizados para construir el gráfico pertenecen al período comprendido entre Octubre de 2008 y Noviembre de 2013.

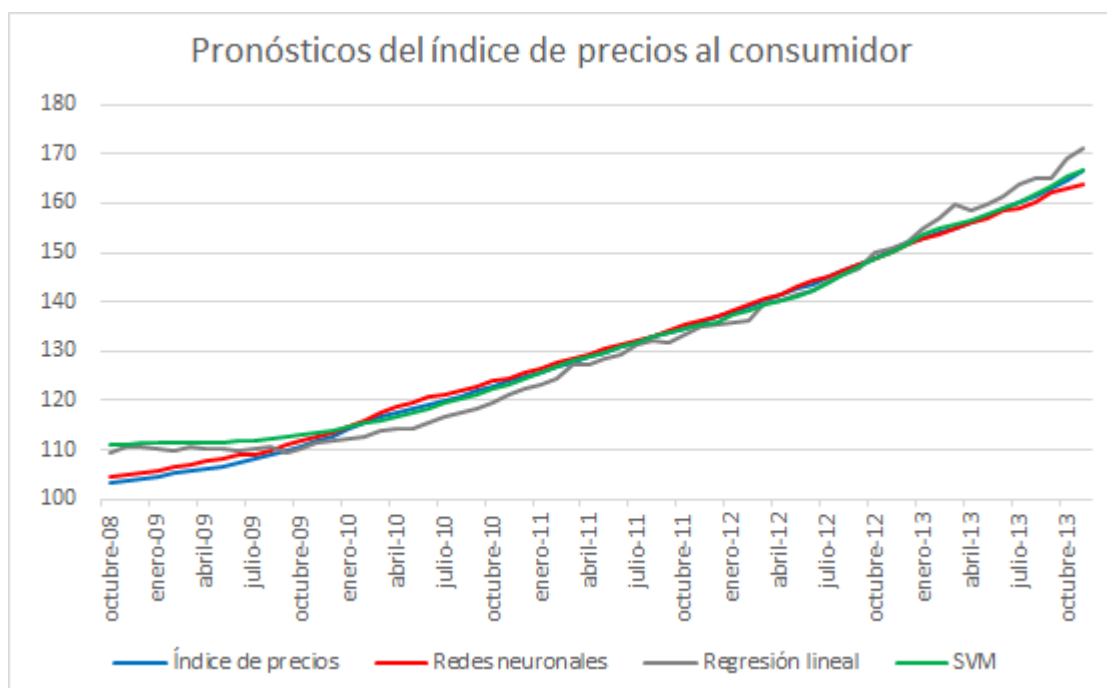


Ilustración 43. Gráfico de los pronósticos e Índice de precios al consumidor

Al igual que los resultados de los criterios de rendimiento analizados anteriormente, se puede apreciar gráficamente que los valores pronosticados por las redes neuronales y las máquinas de vectores soporte se acercan notablemente a los valores reales del índice de precios. Así mismo, mediante la ilustración es posible resaltar el comportamiento de las máquinas de vectores soporte. En las primeras predicciones (entre octubre de 2008 y septiembre de 2009) sus estimaciones están muy alejadas del índice de precios real. Sin embargo, luego de este intervalo de tiempo el rendimiento de sus predicciones mejora formidablemente.

Hasta ahora lo que se analizó fue la inflación del periodo inmediatamente posterior a los datos que se utilizaron de entrenamiento. A continuación se procede a mostrar lo que ocurre cuando se predicen valores del índice de precios con una anticipación de tres, seis, nueve y doce meses. En todos los casos se compara el rendimiento de las redes neuronales con las máquinas de vectores soporte y la regresión lineal.

6.2.3 Tercer proceso

Para poder realizar el experimento de predecir el índice de precios al consumidor con una anticipación de 3, 6, 9 y 12 meses, tomando datos de un trimestre, es necesario modificar el valor del parámetro horizon del operador Windowing. Este parámetro define la distancia entre la ventana y el valor a predecir. A fin de generar el modelo de datos solo se tuvieron en cuenta los atributos que fueron mejor ponderados por los operadores Weight by Deviation y Weight

by PCA (véase Tabla 19): “Depósitos del sector privado”, “Plazo fijo no ajustable por CER”, “En poder del público”, “Depósitos del sector público” y “Cuenta corriente”. Con los mencionados atributos se generó el mejor modelo de datos utilizando redes neuronales en el Apartado 6.2.2.

6.2.3.1 Resultados

Luego de ejecutar el proceso para cada uno de los casos se obtuvo los siguientes resultados:

Técnica utilizada Criterio de performance	Redes neuronales	SVM	Regresión lineal
prediction_trend_accuracy	0.507 +/- 0.351	1.000 +/- 0.000	0.903 +/- 0.179
root_mean_squared_error	1.341 +/- 0.740	1.720 +/- 1.418	3.863 +/- 1.306
absolute_error	1.192 +/- 0.695	1.583 +/- 1.349	3.430 +/- 1.260
relative_error	0.84% +/- 0.49%	1.17% +/- 1.06%	2.45% +/- 1.00%

Tabla 21. Evaluación de los modelos con predicciones a tres meses

Técnica utilizada Criterio de performance	Redes neuronales	SVM	Regresión lineal
prediction_trend_accuracy	0.250 +/- 0.302	1.000 +/- 0.000	0.795 +/- 0.311
root_mean_squared_error	1.901 +/- 0.884	1.763 +/- 1.298	4.186 +/- 2.418
absolute_error	1.680 +/- 0.806	1.630 +/- 1.228	3.577 +/- 2.219
relative_error	1.17% +/- 0.57%	1.18% +/- 0.95%	2.54% +/- 1.72%

Tabla 22. Evaluación de los modelos con predicciones a seis meses

Técnica utilizada Criterio de performance	Redes neuronales	SVM	Regresión lineal
prediction_trend_accuracy	0.233 +/- 0.329	0.967 +/- 0.085	0.717 +/- 0.369
root_mean_squared_error	2.503 +/- 1.395	1.694 +/- 1.050	4.286 +/- 1.568
absolute_error	2.235 +/- 1.309	1.552 +/- 0.990	3.729 +/- 1.492
relative_error	1.53% +/- 0.87%	1.10% +/- 0.76%	2.62% +/- 1.20%

Tabla 23. Evaluación de los modelos con predicciones a nueve meses

Técnica utilizada Criterio de performance	Redes neuronales	SVM	Regresión lineal
prediction_trend_accuracy	0.213 +/- 0.243	0.935 +/- 0.146	0.657 +/- 0.386
root_mean_squared_error	2.365 +/- 1.147	1.629 +/- 1.127	3.980 +/- 1.362
absolute_error	2.083 +/- 1.054	1.496 +/- 1.062	3.509 +/- 1.340
relative_error	1.41% +/- 0.69%	1.05% +/- 0.79%	2.43% +/- 1.02%

Tabla 24. Evaluación de los modelos con predicciones a doce meses

Analizando los resultados de las distintas tablas se pueden sacar las siguientes conclusiones:

- Al igual que en el ensayo anterior, aparentemente las redes neuronales parecen tener un mejor rendimiento que las máquinas de vectores soporte con predicciones a tres meses, si se analiza solamente los errores promedios; sin embargo, analizando los intervalos de confianza de los errores cuadráticos se puede apreciar que el intervalo de las máquinas

de vectores soporte (0.302; 3.138) tiene valores más cercanos a cero que el de las redes neuronales (0.601; 2.081). Esto quiere decir que las máquinas de vectores soporte en algunos casos son más precisas que las redes neuronales.

- Las redes neuronales tienen las peores predicciones de la tendencia de la serie. En estos experimentos como en el realizado anteriormente (véase Tabla 20), el criterio de evaluación prediction trend accuracy no es bueno. En todos los casos presenta valores menores a las máquinas de vectores soporte y la regresión lineal.
- Para predicciones mayores a 3 meses tienen un mejor rendimiento las máquinas de vectores soporte, no solo en valores relativos a los errores de la predicción sino también en la tendencia. De hecho presenta el mejor rendimiento para predecir la tendencia de la serie de tiempo.
- Las predicciones de la regresión lineal son muy imprecisas, aunque puede estimar la tendencia de la serie mejor que las redes neuronales.

6.3 Descripción del modelo

Luego de elaborar un modelo de datos de redes neuronales, evaluar su rendimiento de acuerdo a distintos criterios de evaluación y comparar su desempeño con otros modelos creados a partir de otras técnicas, se procede a realizar una descripción de la estructura de la red neuronal y una interpretación de los resultados en términos de negocio del modelo elaborado. Para realizar estos análisis se estudiará el modelo generado para predecir el índice de precios a partir de los datos económicos de tres meses anteriores (véase Sección 6.2.2). La Ilustración 44 contiene el modelo de datos de las redes neuronales gráficamente, mientras que en el Anexo VI está detallado el modelo generado en base a los pesos de las conexiones entre las neuronas.

Mediante el gráfico mencionado anteriormente es posible apreciar el tipo de técnica empleado, una red neuronal multicapa (véase Apartado 2.5). Las redes con alimentación hacia adelante representan una función de sus entradas actuales y no tienen memoria a corto plazo, a diferencia de las redes recurrentes o cíclicas. La topología de la red neuronal se compone por tres capas: de entrada, oculta y de salida. La cantidad de nodos de la capa de entrada es equivalente a la cantidad de atributos seleccionados más un nodo umbral. La capa oculta contiene doce nodos, de los cuales uno de ellos no está conectado a la capa precedente. Finalmente, la capa de salida contiene un único nodo que representa al atributo objetivo, el índice de precios al consumidor. Es importante mencionar que el algoritmo utilizado para el aprendizaje de la red neuronal es la propagación hacia atrás (del inglés backpropagation), a través de este algoritmo es posible reducir el error de la capa de salida.

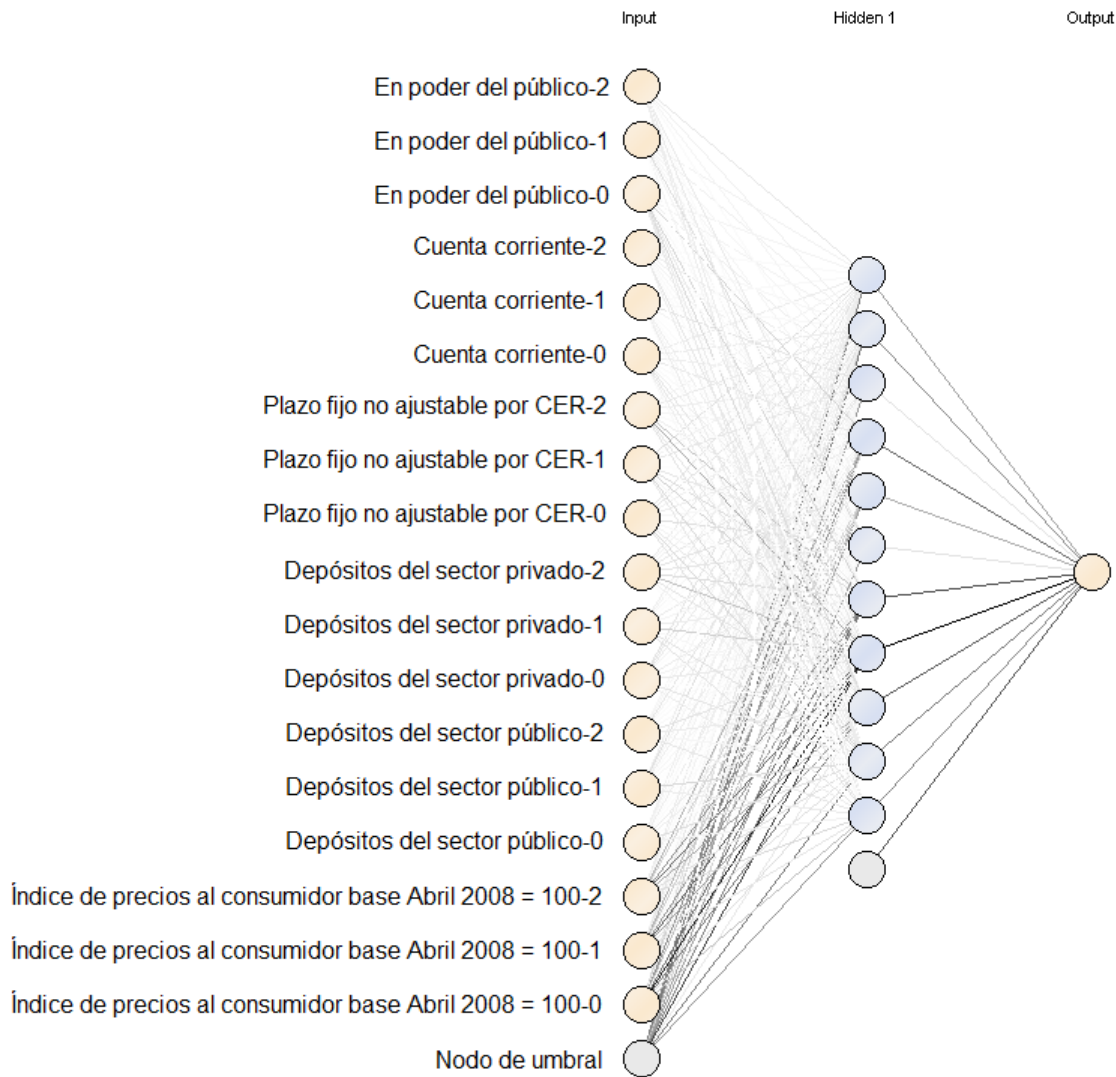


Ilustración 44. Representación gráfica del modelo de datos elaborado

El modelo elaborado de las redes neuronales contiene los atributos que fueron seleccionados como mejores atributos (véase Tabla 19). A continuación se presenta una descripción de cada atributo utilizado para elaborar el modelo de datos y las relaciones encontradas entre estos atributos para poder predecir el índice de precios al consumidor.

En poder del público representa el dinero circulante en poder del público. En (Wende, 2012) se menciona que una tendencia de crecimiento del dinero que circula en la economía puede generar una presión tanto en el tipo de cambio como en la inflación. Además, el artículo menciona una teoría que Irvin Fisher formuló en 1911, en la cual son dos los fenómenos que influyen en el nivel de precios, particularmente cuando no crece la producción: el aumento de la oferta de dinero y la velocidad de circulación del mismo. Por lo tanto, si la gente busca gastar rápidamente lo que tiene en sus bolsillos porque le “quema” ante la pérdida de poder adquisitivo, la consecuencia es que termina presionando aún más la inflación.

Cuenta corriente constituye el dinero en cuenta corriente en pesos. El dinero en cuenta corriente con el dinero en poder del público y los cheques cancelatorios conforman el agregado económico M1, tal como se vio en la Sección 5.1.2. Este agregado económico representa la oferta monetaria más líquida, ya que está constituido por dinero con el cual se puede realizar pagos de manera inmediata y sin costo. La variable cheques cancelatorios no fue tomada en

cuenta para la elaboración del modelo de datos, ya que contaba con valores faltantes como se vio en el Apartado 4.4.4. De esta manera, el modelo de datos elaborado cuenta con las dos variables más importantes que constituyen el agregado M1.

El plazo fijo ajustable por CER es aquel cuyo capital se actualiza según la variación del CER (coeficiente de estabilización de referencia). Este coeficiente es elaborado por el BCRA en base al aumento de la inflación. A diferencia de los plazos fijos ajustables por CER, los plazos fijos tradicionales o no ajustables por CER son aquellos que aseguran una tasa de interés en un tiempo determinado.

Finalmente, Depósitos del sector público constituye los depósitos del sector público en pesos y Depósitos del sector privado los depósitos del sector privado en pesos. Estas dos variables con “En poder del público” y “Cheques cancelatorios” integran el agregado monetario M3. Este agregado a diferencia del agregado M1 contiene también el dinero “menos líquido”, es decir, aquel que no puede ser utilizado para realizar pagos de manera inmediata.

Se puede señalar que los atributos seleccionados para elaborar el modelo de datos representan variables monetarias del sistema financiero. En un principio, cuando se realizó la recolección de datos, se tomaron datos monetarios, macroeconómicos, de commodities, de empleo y desempleo entre otros. Luego de realizar la selección de los atributos más relevantes para pronosticar el índice de precios solo quedaron los atributos referidos a variables monetarias de la economía local. Esto es importante mencionar, ya que permite comprender la importancia de la política monetaria del país para solucionar la problemática de la inflación.

Otro aspecto interesante es que Cuenta corriente y Plazo fijo no ajustable por CER representan dinero tanto del sector público como del sector privado, estos atributos corresponden a dos tipos de depósitos que se pueden realizar en un banco. Los totales de los depósitos realizados de acuerdo al tipo de actividad económica están incluidos en los atributos Depósitos del sector privado y Depósitos del sector público. Con los atributos disponibles no es posible determinar la cantidad de dinero en cuenta corriente del sector privado y la cantidad de dinero en cuenta corriente del sector público, lo mismo ocurre con Plazo fijo no ajustable por CER y los otros tipos de depósitos. Lo ideal sería poder descomponer estos atributos para obtener un mayor nivel de granularidad en los datos.

6.4 Resumen

El éxito en la etapa de modelado depende de diversas cuestiones, es importante el trabajo previo con los datos en las etapas anteriores del proyecto para poder elaborar un modelo de datos de buena calidad. Además es necesario contar con alguna experiencia en el uso de la herramienta de minería de datos, en este caso RapidMiner; esto permite diseñar de manera eficiente el procesamiento de los datos. También resulta significativo el conocimiento de los distintos algoritmos utilizados a través de los operadores de RapidMiner.

7 Evaluación

Es la etapa final del proyecto, aquí se realizan las evaluaciones del trabajo elaborado. El objetivo es determinar si las actividades realizadas permiten alcanzar los objetivos del proyecto. También es necesario realizar una evaluación de los recursos utilizados, restricciones, presunciones, y otros factores que influenciaron la realización del trabajo.

7.1 Evaluación de los resultados

Para evaluar los resultados del proyecto se tendrán en cuentas dos aspectos importantes: los criterios de éxito y las restricciones del proyecto.

Comenzando por los criterios de éxito, ejecutar el proyecto permitió comprender y aplicar los conceptos de minería de datos. El resultado del proyecto es una aplicación del proceso de minería de datos siguiendo la metodología CRISP-DM. Para desarrollar las actividades fue necesario recurrir a distintas fuentes de conocimiento como ser el juicio de expertos y bibliografía relacionada a la temática.

El siguiente punto a mencionar son los modelos de datos. Estos fueron elaborados utilizando técnicas de procesamiento de datos como las redes neuronales. Luego sus pruebas fueron realizadas teniendo en cuenta dos criterios de evaluación, la tendencia del índice de precios al consumidor y los errores entre los valores reales y los pronosticados. La tendencia del índice de precios fue pronosticada de mejor manera con otras técnicas, mientras que si se analiza el rendimiento en base a los errores entre los valores reales y los pronosticados el desempeño de las redes neuronales fue muy bueno.

En cuanto a las restricciones del proyecto, las dos que más influyeron fueron el equipo informático y los recursos humanos disponibles. El proyecto se vio limitado con su equipo informático a la hora de elaborar el modelo de datos, la capacidad de procesamiento de los recursos informáticos disponibles no permitía elaborar modelos con un gran número de atributos, por eso es que se decidió seleccionar los atributos más relevantes del conjunto de datos para elaborar el modelo.

Para finalizar, es necesario mencionar que el proyecto elaborado no entrega como resultado proyecciones actuales de la inflación. Realizar esto implica volver a desarrollar las distintas actividades del proceso de minería de datos con datos actuales. La ejecución de estas actividades requiere de tiempo, y estos tiempos se acortan si se precisan resultados de manera inmediata. Para poder hacer frente a estas restricciones es necesario contar con una mayor cantidad de recursos humanos, de esta manera se podrán acortar los tiempos de las tareas a desarrollar.

7.2 Revisión del proyecto

La revisión del proyecto se realizó analizando diversos factores que influenciaron en el mismo.

7.2.1 Metodología de minería de datos

Para desplegar el proyecto se desarrollaron las seis fases del modelo de referencia CRISP-DM, no obstante es importante mencionar que de la última etapa, 'Implementación', solo se realizó una tarea, producir el informe final, ya que el resto de las actividades no se aplican al proyecto en cuestión. Dado que el objetivo de estas actividades es permitir que el cliente pueda utilizar el modelo de datos, y el modelo elaborado en este proyecto no será utilizado por ningún cliente, no se desarrollaron las mismas.

Por lo que se refiere a las tareas realizadas, se destaca que la secuencia de las mismas no es lineal, sino más bien iterativa. Es necesario en muchos casos volver a fases anteriores para rehacer ciertas actividades. En el proyecto, las etapas que más requirieron de este tipo de acciones fueron las de Comprensión y Preparación de datos. Estas etapas son importantes ya que en ellas se realiza la preparación de los datos para posteriormente utilizarlos al construir el modelo de datos.

7.2.2 Fuentes de datos y conocimiento

Las fuentes de datos y el conocimiento de expertos estuvieron disponibles al momento de requerirlos.

7.2.3 Software para minería de datos

La elección de RapidMiner como software de minería de datos fue una decisión acertada. Tres aspectos que se destacan de esta herramienta son su interfaz gráfica, las herramientas de visualización de datos y la documentación disponible.

7.2.4 Gestión del proyecto con PMI

La planificación y ejecución del proyecto fue realizada según las buenas prácticas del Project Management Institute (PMI), encaminar el proyecto con estas prácticas fue beneficioso en la medida que contribuyó a una mayor organización de las actividades desarrolladas.

8 Conclusiones y futuras líneas de investigación

Uno de los principales problemáticas de la economía argentina es el constante aumento de los precios de los productos. Para comprender este fenómeno resulta importante asociar el comportamiento del índice de precios con otras variables de la economía local. En este sentido, el trabajo realizado permite encontrar variables financieras del sistema económico que se relacionan con el crecimiento del índice de precios al consumidor. Tomando datos económicos de estas variables es posible anticipar el comportamiento del índice de precios.

Para realizar el análisis se consideró al Índice de precios base Abril 2008 = 100 como indicador de la inflación en el periodo comprendido entre Agosto del 2008 y Diciembre del 2013, los datos procesados fueron obtenidos de fuentes oficiales como el BCRA y el INDEC.

El proceso de minería de datos se desarrolló siguiendo el modelo de referencia CRISP-DM, esta guía es de gran ayuda para quienes desean realizar minería de datos y no tienen mucha experiencia en este tipo de actividades. En muchos casos fue necesario ejecutar varias veces una misma tarea, esto es indicado como necesario en la metodología ya que contribuye a obtener un mejor resultado final. Las tareas de modelado son las que mayor esfuerzo requirieron; en muchos casos, fue necesario investigar sobre el funcionamiento de distintos operadores y sus parámetros.

En cuanto a los experimentos efectuados, es posible verificar que las redes neuronales son una buena herramienta a la hora de pronosticar la inflación u otra variable económica. El trabajo realizado permite ver que los pronósticos a corto plazo utilizando redes neuronales son muy buenos. Aun así, sería interesante desarrollar modelos de datos que contengan datos con una mayor granularidad como se vio en el Apartado 6.3.

En relación a la comparación del rendimiento de las redes neuronales en contraste a otras técnicas, resultó muy interesante el desempeño de las máquinas de vectores soporte. Con esta técnica es posible predecir con un alto grado de certeza la tendencia del índice de precios. Resultaría atractivo poder combinar tanto redes neuronales como máquinas de vectores soporte para elaborar modelos de predicciones a alguna problemática en particular.

Para finalizar, se mencionan algunas aplicaciones de la minería de datos en el ámbito económico – financiero que podrían ser de interés para otros proyectos de análisis de datos:

- Predicciones de indicadores financieros
- Análisis de tendencias de indicadores financieros
- Detección de blanqueamiento de dinero y delitos financieros
- Análisis de componentes principales para determinar factores de riesgos en inversiones.

Bibliografía⁵

- BCRA. (2008;2009; 2010; 2011; 2012; 2013; 2014; 2015). *Informe monetario mensual*.
Obtenido de <http://www.bcr.gov.ar/Publicaciones/pubcoy040302.asp>
- BCRA. (2008;2009; 2010; 2011; 2012; 2013; 2014; 2015). *Panorama global de commodities*.
Obtenido de <http://www.bcr.gov.ar/Publicaciones/pubcoy040500.asp>
- BCRA. (2008;2009; 2010; 2011; 2012; 2013; 2014; 2015). *Tipo de cambio nominal promedio mensual*. Obtenido de <http://www.bcr.gov.ar/Estadisticas/estser030502.asp>
- BCRA. (2015). *Medios de pago - Cheque cancelatorio*. Obtenido de
http://www.bcr.gov.ar/medios_pago/medche030100.asp
- Bennett, K., & Mangasarian, O. (1992). Robust linear programming discrimination of two linearly inseparable sets. *Optimization methods and software*, 1(1), 23-34.
- Boser, B., Guyon, I., & Vladimir, V. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory* (págs. 144-152). New York, USA: ACM.
- Cáceres Hernández, J. J. (2006). *Conceptos básicos de estadística para ciencias sociales*. Madrid, España: Delta Publicaciones.
- Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide. *CRISP-DM consortium*.
- Consejo Profesional de Agrimensores, Ingenieros y Profesiones Afines. (2015). *Honorarios de Informática*. Obtenido de <http://www.copaipa.org.ar/informatica/>
- Consejo Profesional de Ciencias Económicas de Salta. (2015). *Honorarios mínimos sugeridos*. Obtenido de http://www.consejosalta.org.ar/wp-content/uploads/honorarios_sugeridos_afiche_2015.pdf
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297.
- Díaz, R. A. (2011). Inflación: Definición, causas y efectos. *Boletín Plural, n° 48 de la Facultad de Humanidades Ciencias Sociales y de la Salud, Universidad Nacional de Santiago del Estero*. Santiago del Estero: Artículo de opinión.
- Dietterich, T. (2002). Machine Learning for Sequential Data: A Review. En I. T. Caelli (Ed.), *Structural, Syntactic, and Statistical Pattern Recognition; Lecture Notes in Computer Science* (Vol. 2396, págs. 15-30). Springer-Verlag.
- Dornbusch, R. (1981). *La macroeconomía de una economía abierta*. Barcelona: Antoni Bosch Editor, S.A.
- Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining (American Association for Artificial Intelligence)*. Menlo Park, CA, USA.
- Fernández Arufe, J. (2006). *Principios de política económica: ejercicios de test y cuestiones resueltas*. Madrid: Delta.

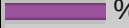
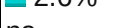
⁵ Páginas web vigentes al 9/05/2016

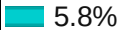
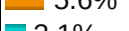
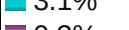
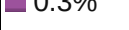
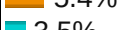
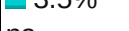
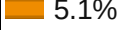
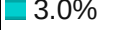
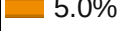
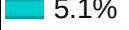
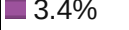
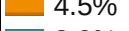
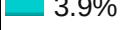
- Grané Chávez, A. (2002). *Análisis de Componentes Principales*. Madrid: Universidad Carlos III, Departamento de Estadística. Obtenido de http://halweb.uc3m.es/esp/Personal/personas/agrane/ficheros_docencia/MULTIVARIANT/slides_comp_reducido.pdf
- Hakan Aladag, C., & Egrioglu, E. (2012). *Advances in time series forecasting*. Oak Park, Illinois: Bentham Ebooks.
- Han, J., Kamber, M., & Pei, J. (2001). *Data Mining: Concepts and Techniques* (Primera ed.). San Francisco: Morgan Kaufmann Publishers.
- Hernández Orallo, J., Ramírez Quintana, J., & Ferri Ramírez, C. (2004). *Introducción a la Minería de Datos*. Madrid, España: Pearson Prentice Hall.
- INDEC. (2002). *Cómo usar un índice de precios*. Obtenido de <http://www.estadistica.chubut.gov.ar/home/archivos/menuTematico/indice-de-precios/como-usar.pdf>
- INDEC. (2008;2009; 2010; 2011; 2012; 2013; 2014; 2015). *Empleo y desempleo*. Obtenido de http://www.indec.gov.ar/nivel4_default.asp?id_tema_1=4&id_tema_2=31&id_tema_3=58
- INDEC. (2008;2009; 2010; 2011; 2012; 2013; 2014; 2015). *Índice de precios al consumidor*. Obtenido de http://www.indec.gov.ar/informesdeprensa.asp?id_tema_1=3&id_tema_2=5&id_tema_3=31
- INDEC. (2008;2009; 2010; 2011; 2012; 2013; 2014; 2015). *Oferta y demanda*. Obtenido de <http://200.51.91.244/cnarg/agregados.php>
- INDEC. (2008;2009; 2010; 2011; 2012; 2013; 2014; 2015). *Producto bruto interno a precio de mercado*. Obtenido de <http://200.51.91.244/cnarg/agregados.php>
- INDEC. (2014). *Sobre el empalme de la series en el IPCNu*. Obtenido de http://www.indec.gov.ar/ftp/cuadros/economia/nota_metodologica_ipcnu_empalme.pdf
- KDnuggets. (2015). *R leads RapidMiner, Python catches up, Big Data tools grow, Spark ignites*. Obtenido de <http://www.kdnuggets.com/2015/05/poll-r-rapidminer-python-big-data-spark.html>
- Krugman, P. R., & Wells, R. (2007). *Introducción a la Economía. Macroeconomía*. Barcelona: Reverté.
- Martín Diazaraque, J. M. (2008). *Introducción a las redes neuronales: Materiales, Curso Data Mining, Universidad Carlos III de Madrid*. Obtenido de <http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/tema3dm.pdf>
- Matich, D. J. (2001). *Redes Neuronales: Conceptos básicos y Aplicaciones*. Obtenido de http://www.fro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monograis/matich-redesneuronales.pdf

- Pedroza, H., & Dicoyskyi, L. (2006). *Sistema de análisis estadístico con SPSS*. Managua, Nicaragua: Instituto Interamericano de Tecnología Agrícola (IICA) – Instituto de Tecnología Agropecuaria (INTA).
- PMI. (2008). *Guía de los Fundamentos para la Dirección de Proyectos (Guía del PMBOK)* (Cuarta ed.). Pennsylvania, USA: Project Management Institute, Inc.
- Ruiz Sánchez, R. (2005). *Selección de atributos mediante proyecciones*. Obtenido de www.lsi.us.es/docs/doctorado/proyectos_tesis/proyecto-memoria-roberto.pdf
- Russell, S., & Norvig, P. (2004). *Inteligencia Artificial un enfoque moderno* (Segunda ed.). Madrid, España: Pearson Prentice Hall.
- Santana, J. (2006). Predicción de series temporales con redes neuronales: una aplicación a la inflación colombiana. *Revista Colombiana de Estadística*, 29(1), 77-92.
- Smola, A., & Schölkopf, B. (2003). *A Tutorial on Support Vector Regression*. Obtenido de <http://alex.smola.org/papers/2003/SmoSch03b.pdf>
- Urquiza Rojas, D., & Mendivil Gomez, F. (2011). Aplicación de Redes Neuronales Artificiales para el Análisis de la Inflación en Bolivia. *4to. Encuentro de Economistas de Bolivia*.
- Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag: New York.
- Velásquez Zapateiro, J., & Obeso Fernández, V. (2007). *Análisis numérico. Notas de clase*. Barranquilla, Colombia: Ediciones Uninorte.
- Wende, P. (2012). El dinero en poder del público subió 36%: \$ 24.500 millones más. *Ámbito financiero*. Obtenido de <http://www.ambito.com/diario/646582-el-dinero-en-poder-del-publico-subio-36--24500-millones-mas>

Anexo I

A continuación se presenta una encuesta realizada por el sitio KDnuggets sobre las herramientas más utilizadas de minería de datos. En la encuesta se distinguen software libre del comercial y los porcentajes de utilización de los años 2013, 2014 y 2015. Muchas de las personas encuestadas utilizó más de una herramienta de minería de datos, por lo tanto, también se indica el porcentaje de las personas que solamente utilizó cada herramienta.

¿Qué herramientas de Big Data, minería de datos, software de Ciencia datos utilizó en los últimos 12 meses para un proyecto real? [2759 votantes] (KDnuggets, 2015)	
Leyenda: Rojo: Libre / herramientas Open Source verdes: Herramientas Comerciales Fucsia: Herramientas Hadoop / Big Data	 % usuarios en 2015  % usuarios en 2014  % usuarios en 2013
R (1293), 3.6% solamente	 46.9%  38.5%  37.4%
RapidMiner (870), 13.7% solamente	 31.5%  44.2%  39.2%
SQL (853), 0% solamente	 30.9%  25.3%
Python (837), 0% solamente	 30.3%  19.5%  13.3%
Excel (631), 0% solamente	 22.9%  25.8%  28.0%
KNIME (553), 6.7% solamente	 20%  15.0%  5.9%
Hadoop (507), 0% solamente	 18.4%  12.7%  9.3%
Tableau (341), 0% solamente	 12.4%  9.1%  6.3%
SAS base (313), 0.6% solamente	 11.3%  10.9%  10.7%
Spark (311), 0% solamente	 11.3%  2.6% na

Weka (310), 0% solamente	 11.2%  17.0%  14.3%
SAS Enterprise Miner (302), 3.6% solamente	 10.9%  7.2%  5.9%
Microsoft SQL Server (268), 0% solamente	 9.7%  10.5%  7.0%
MATLAB (243), 0% solamente	 8.8%  8.4%  9.9%
scikit-learn (229), 0% solamente	 8.3% na na
Unix shell/awk/gawk (221), 0% solamente	 8.0%  5.8% na
IBM SPSS Statistics (213), 0% solamente	 7.7%  7.7%  8.7%
IBM SPSS Modeler (197), 7.1% solamente	 7.1%  5.7%  6.1%
Alteryx (155), 39.4% solamente	 5.6%  3.1%  0.3%
Pig (150), 0% solamente	 5.4%  3.5% na
Other programming languages (140), 0% solamente	 5.1%  3.0% na
Other free analytics/data mining tools (138), 0% solamente	 5.0%  5.1%  3.4%
Other Hadoop/HDFS-based tools (125), 0% solamente	 4.5%  3.9% na

TIBCO Spotfire (119), 11.8% solamente	 4.3%  2.8%  1.4%
Rattle (117), 0.9% solamente	 4.2%  4.9%  4.5%
QlikView (116), 0% solamente	 4.2%  3.0%  2.4%
Revolution Analytics (now part of Microsoft) (109), 0% solamente	 4.0%  9.1%  4.5%
Microsoft Azure ML (102), 1.0% solamente	 3.7% na na
Microsoft Power BI (98), 0% solamente	 3.6% na na
MLlib (91), 0% solamente	 3.3%  1.0% na
JMP (86), 0% solamente	 3.1%  3.8%  4.1%
SAP (including former KXEN) (82), 26.8% solamente	 3.0%  6.8%  1.4%
Perl (79), 0% solamente	 2.9%  3.0% na
Mahout (76), 0% solamente	 2.8%  2.5% na
Pentaho (74), 0% solamente	 2.7%  2.6% na
Other paid analytics/data mining/data science software (66), 6.1% solamente	 2.4%  1.9%  2.4%

Salford SPM/CART/Random Forests/MARS/TreeNet (64), 43.8% solamente	 2.3%  3.6%  2.2%
Gnu Octave (64), 0% solamente	 2.3%  3.9%  2.9%
IBM Watson Analytics (57), 0% solamente	 2.1% na na
Ayasdi (56), 10.7% solamente	 2.0% na na
Dataiku (56), 7.1% solamente	 2.0% na na
Actian (56), 7.1% solamente	 2.0%  0.5% na
H2O (0xdata) (55), 0% solamente	 2.0%  0.2% na
Orange (53), 0% solamente	 1.9%  3.4%  3.6%
Mathematica (52), 0% solamente	 1.9%  2.3%  2.1%
IBM Cognos (51), 0% solamente	 1.8%  1.8%  2.4%
Dell (including StatSoft) (47), 19.1% solamente	 1.7%  1.7%  9.0%
XLSTAT for Excel (42), 0% solamente	 1.5%  1.2%  0.9%
Stata (36), 2.8% solamente	 1.3%  1.4%  2.1%

Lexalytics (35), 28.6% solamente	1.3% na na
Vowpal Wabbit (35), 0% solamente	1.3% na na
C4.5/C5.0/See5 (35), 0% solamente	1.3% 1.5% 1.1%
Julia (31), 3.2% solamente	1.1% 0.8% na
Splunk/ Hunk (30), 0% solamente	1.1% 0.7% na
Datameer (26), 0% solamente	0.9% 0.8% na
MicroStrategy (24), 0% solamente	0.9% na na
BigML (23), 0% solamente	0.8% 0.9% na
Zementis (22), 31.8% solamente	0.8% 0.8% 0.9%
Miner3D (22), 9.1% solamente	0.8% 0.9% 1.8%
Oracle Data Miner (22), 0% solamente	0.8% 2.2% 1.0%
Amazon Machine Learning (20), 5.0% solamente	0.7% na na
F# (18), 0% solamente	0.7% 0.5% 0.7%

BayesiaLab (16), 12.5% solamente	 0.6%  4.1%  1.0%
Dato (former Graphlab) (15), 6.7% solamente	 0.5%  0.9% na
Clojure (13), 0% solamente	 0.5%  0.5% na
Alpine Data Labs (13), 0% solamente	 0.5%  2.7% na
Angoss (11), 18.2% solamente	 0.4%  0.4%  0.3%
Lavastorm (10), 0% solamente	 0.4%  0.3%  0.4%
Lisp (10), 0% solamente	 0.4%  0.3% na

Tabla 25. Encuesta de herramientas de Big Data utilizadas

Anexo II

Para generar los gráficos de las tendencias de la Tasa de desocupación y el Índice de precios al consumidor base Abril 2008 = 100 se realizó el siguiente proceso en RapidMiner.

Los datos son extraídos a través de dos operadores ReadExcel (RE_TasaDesocupacion y RE_IndiceDePrecios). Posteriormente, los pertenecientes a la tasa de desocupación, son procesados en un operador Fit Trend para generar la tendencia.

En el caso de los datos del índice de precios, es necesario agregar un atributo adicional y filtrar algunos ejemplos. Se requiere generar un atributo adicional para poder representar los meses en el gráfico de la tendencia. Para realizar esto se generó un atributo de rol “id” en RapidMiner a través del operador “Generate ID”, luego a este atributo se cambió el rol “id” a un rol “regular” mediante el operador “Set Role”. El proceso continúa con el operador “Rename” que modifica el nombre del atributo “id” generado a “Número de mes” además de cambiar el nombre del atributo “Índice de precios al consumidor base Abril 2008 = 100” por “Índice de precios al consumidor base Abril 2008”, ya que es necesario eliminar el operador “=” del nombre del atributo para filtrar los ejemplos en la siguiente etapa del proceso. Finalmente, el operador “Filter Examples” elimina los registros donde el atributo “Índice de precios al consumidor base Abril 2008” es nulo.

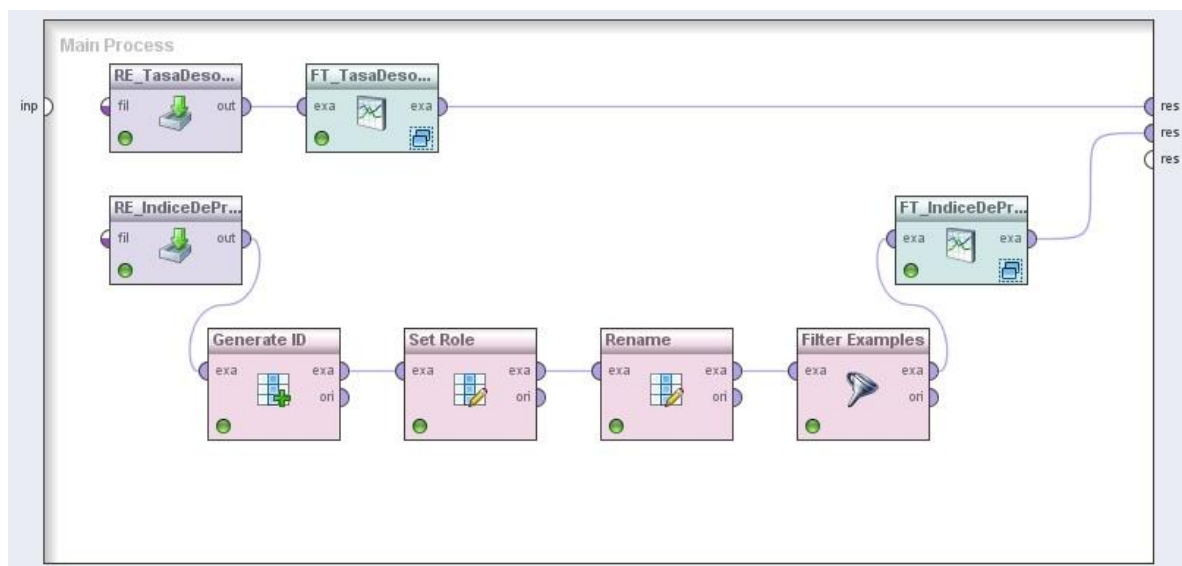


Ilustración 45. Proceso para generar las tendencias de las series

Fit Trend es un operador anidado que requiere de otro operador para estimar las tendencias, en ambos casos (FT_TasaDesocupacion y FT_IndiceDePrecios) se utilizó el operador de regresión lineal como se muestra más adelante. Es importante mencionar que se utilizó el operador de regresión lineal para obtener una tendencia lineal como resultado.

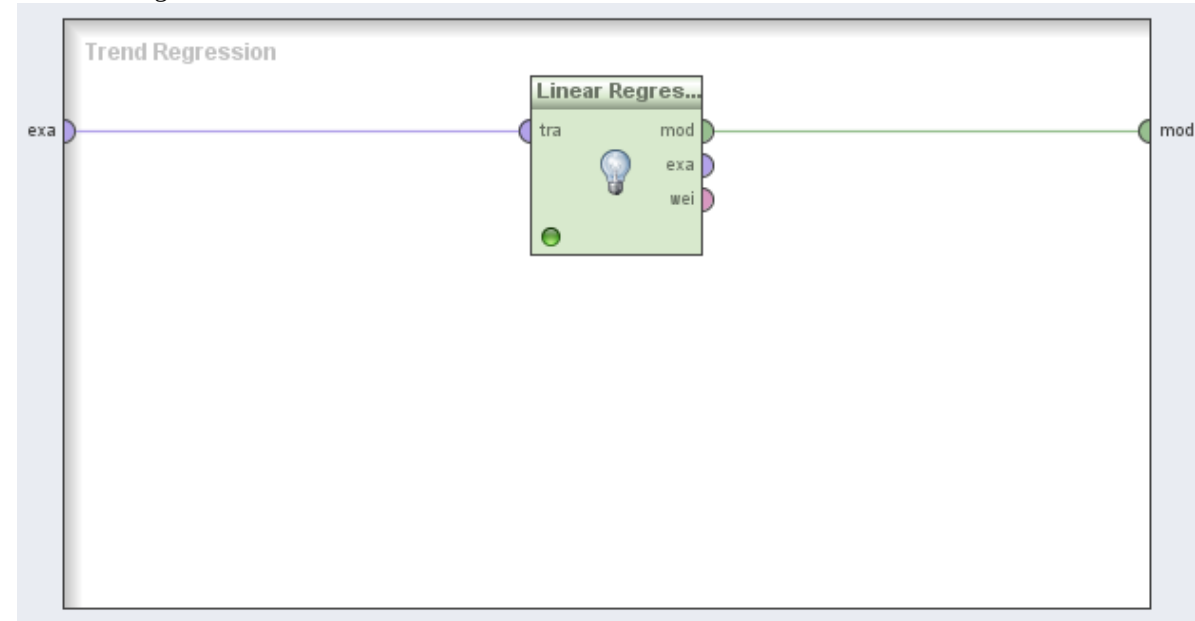


Ilustración 46. Operador interno al FitTrend

Anexo III

A continuación se presenta una introducción teórica del procedimiento interpolación lineal a trozos, luego una adaptación del mismo a los datos disponibles y finalmente una aplicación de ejemplo a un conjunto de datos.

Interpolación a trozos (Velásquez Zapateiro & Obeso Fernández, 2007)

En un polinomio interpolador de grado n es posible tener $n-1$ extremos relativos, lo cual trae como consecuencia que tenga muchas oscilaciones o fluctuaciones al pasar por los puntos dados. Una alternativa para evitar dichas fluctuaciones es generar polinomios $S_k(x)$ que solo interpolen dos nodos consecutivos y luego unirlos en sus extremos. El conjunto $\{S_k(x)\}$ forma una curva polinomial a trozos o spline, que notaremos $S(x)$.

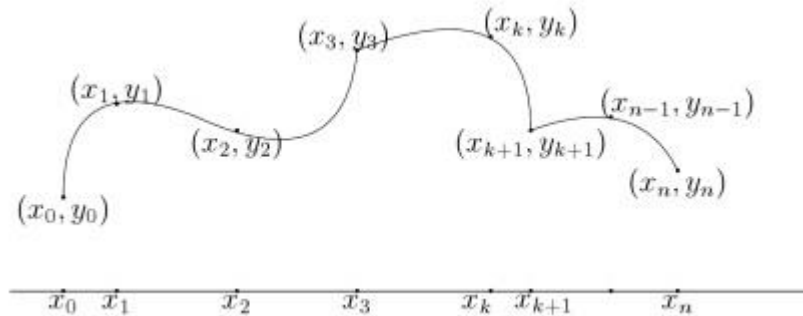


Ilustración 47. Curva polinomial extraído de (Zapateiro & Fernández, 2007)

Interpolación lineal a trozos

Sean $(x_0, y_0), (x_1, y_1), (x_2, y_2), \dots, (x_k, y_k), (x_{k+1}, y_{k+1}), \dots, (x_n, y_n)$ los puntos por los cuales debe pasar el polinomio interpolador; el polinomio más simple que se puede contruir es el polinomio de grado uno, el cual produce una línea poligonal.

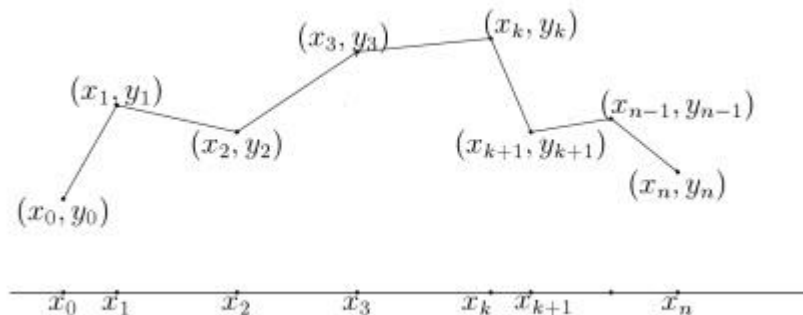


Ilustración 48. Polinomio interpolador extraído de (Zapateiro & Fernández, 2007)

Por los polinomios interpoladores de Lagrange podemos determinar cada uno de los segmentos que forman esta línea poligonal. Si $S_k(x)$ es el k -ésimo segmento que une los puntos $(x_k, y_k), (x_{k+1}, y_{k+1})$, se tiene entonces que

$$S_k(x) = y_k \frac{x - x_{k+1}}{x_k - x_{k+1}} + y_{k+1} \frac{x - x_k}{x_k - x_{k+1}}$$

o también

$$S_k(x) = y_k + d_k(x - x_k), \quad d_k = \frac{y_{k+1} - y_k}{x_{k+1} - x_k}$$

Que es la ecuación de la recta que pasa por los puntos dados, luego entonces

$$S(x) \begin{cases} y_0 + d_0(x - x_0) & \text{si } x \in [x_0, x_1] \\ y_1 + d_1(x - x_1) & \text{si } x \in [x_1, x_2] \\ \vdots & \vdots \\ y_k + d_k(x - x_k) & \text{si } x \in [x_k, x_{k+1}] \\ \vdots & \vdots \\ y_{n-1} + d_{n-1}(x - x_{n-1}) & \text{si } x \in [x_{n-1}, x_n] \end{cases}$$

Aplicación a los datos trimestrales de la interpolación lineal a trozos

Los datos trimestrales corresponden al promedio de cada trimestre. Si se asigna este promedio al segundo mes de cada trimestre, se podría hacer la interpolación de la siguiente manera; el valor mensual del tercer mes del trimestre n y el valor del primer mes del trimestre n+1 podrían ser obtenidos a través de los valores de los segundos meses de los trimestres n y n+1.

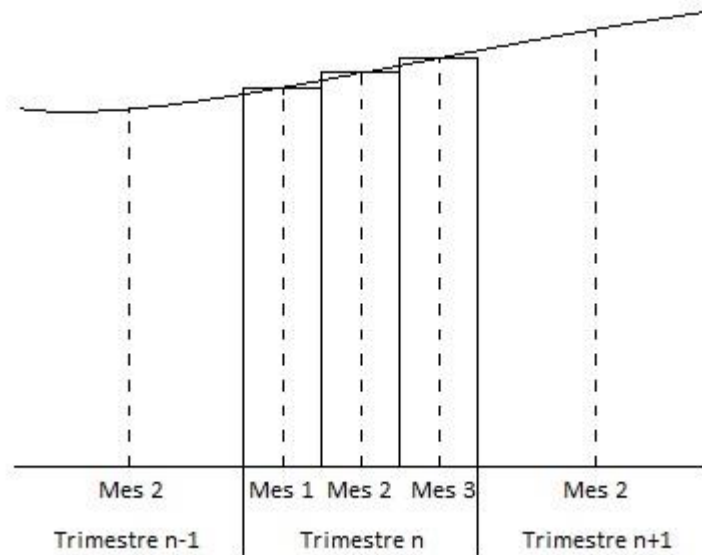


Ilustración 49. Interpolación lineal a trozos para datos trimestrales

Para poder desarrollar la explicación se toma como ejemplo a una variable de la cual solo se disponen datos trimestrales, en este caso el Producto Interno Bruto a precios de mercado.

Sean $PBI(T1)$, $PBI(T2)$, $PBI(T3)$, ..., $PBI(TN)$ los valores del producto bruto interno a precio de mercado obtenidos para los trimestres T1, T2, T3, ..., TN.

Sean $PBI_1(T1)$, $PBI_2(T1)$, $PBI_3(T1)$ los valores del producto bruto interno mensual para el trimestre 1.

Dado que $PBI(T1) = PBI_2(T1)$, entonces los valores de PBI para los meses 1 y 3 se pueden obtener a través de la interpolación lineal a trozos.

$$PBI_1(K) = PBI(K - 1) + \frac{PBI(K) - PBI(K - 1)}{A} C_2$$

$$PBI_3(K) = PBI(K) + \frac{PBI(K+1) - PBI(K)}{A} C_1$$

Donde A es la cantidad de meses entre el mes 2 del trimestre k y el mes 2 del trimestre k-1 o el trimestre k+1. Tanto para $PBI_1(K)$ como para $PBI_3(K)$ el valor de A será de 3. Por otro lado, el valor C_1 corresponde a la cantidad de meses entre el mes 3 y el mes 2 del trimestre k, así también el valor C_2 representa a la cantidad de meses entre el mes 1 del trimestre k, y el segundo mes del trimestre k-1. En el primer caso el valor de C_1 será de 1 y en el segundo C_2 será equivalente a 2.

De esta manera los valores estimados mensuales se obtendrían a partir de las formulas:

$$PBI_1(K) = PBI(K-1) + 2 \frac{PBI(K) - PBI(K-1)}{3}$$

$$PBI_3(K) = PBI(K) + \frac{PBI(K+1) - PBI(K)}{3}$$

A continuación se presenta, a modo de demostración, un ejemplo de la interpolación para datos trimestrales del PBI.

Año	Trimestre	Producto Interno Bruto a precios de mercado	Mes 1	Mes 2	Mes 3
2008	1	665.476.403		665.476.403	687.280.719
2008	2	730.889.351	709.085.035	730.889.351	725.510.412
2008	3	714.752.534	720.131.473	714.752.534	713.948.583
2008	4	712.340.681	713.144.632	712.340.681	695.998.803
2009	1	663.315.046	679.656.924	663.315.046	677.273.651
2009	2	705.190.860	691.232.255	705.190.860	704.309.268
2009	3	702.546.083	703.427.675	702.546.083	719.637.189
2009	4	753.819.400	736.728.294	753.819.400	735.371.269
2010	1	698.475.008	716.923.139	698.475.008	730.024.740
2010	2	793.124.205	761.574.473	793.124.205	788.496.826
2010	3	779.242.067	783.869.446	779.242.067	793.169.724
2010	4	821.025.039	807.097.382	821.025.039	804.332.391
2011	1	770.947.094	787.639.742	770.947.094	799.018.456
2011	2	855.161.180	827.089.818	855.161.180	851.364.842
2011	3	843.772.166	847.568.504	843.772.166	856.276.026
2011	4	881.283.746	868.779.886	881.283.746	855.669.554
2012	1	804.441.171	830.055.363	804.441.171	817.338.034
2012	2	843.131.761	830.234.898	843.131.761	843.588.588
2012	3	844.502.241	844.045.414	844.502.241	858.320.600
2012	4	885.957.317	872.138.958	885.957.317	862.290.679
2013	1	814.957.403	838.624.041	814.957.403	838.904.482
2013	2	886.798.641	862.851.562	886.798.641	882.082.025
2013	3	872.648.792	877.365.408	872.648.792	882.131.120
2013	4	901.095.775	891.613.447	901.095.775	874.652.682
2014	1	821.766.495	848.209.588	821.766.495	845.538.266
2014	2	893.081.809	869.310.038	893.081.809	885.721.157

2014	3	870.999.853	878.360.505	870.999.853	882.472.407
2014	4	905.417.516	893.944.962	905.417.516	883.265.099
2015	1	838.960.265	861.112.682	838.960.265	863.853.627
2015	2	913.640.350	888.746.988	913.640.350	

Tabla 26. Interpolación lineal a trozos del PBI mensual

Anexo IV

Máquinas de vectores soporte (SVM)

Las máquinas de vectores soporte (SVM, del inglés Support Vector Machines) tienen su origen en los trabajos sobre la teoría del aprendizaje estadístico y fueron introducidas en los años 90 por Vapnik y sus colaboradores (Boser, Guyon, & Vladimir, 1992) y (Cortes & Vapnik, 1995). Aunque originariamente las SVMs fueron pensadas para resolver problemas de clasificación binaria, actualmente se utilizan para resolver otros tipos de problemas (regresión, agrupamiento, multclasificación). El problema de regresión mediante lo que se conoce como SVR (del inglés Support Vector Regression machines). A continuación se presenta una explicación extraída de (Smola & Schölkopf, 2003) sobre las máquinas de vectores soporte aplicadas a la regresión.

Supongamos que tenemos un conjunto de datos de entrenamiento $\{(x_1, y_1), \dots, (x_n, y_n)\} \subset X \times \mathbb{R}$, donde X denota el espacio de los patrones de entrada (por ejemplo, $X = \mathbb{R}^d$). Estos podrían ser, por ejemplo, los tipos de cambio de alguna moneda medido en días posteriores junto con sus correspondientes indicadores econométricos. En la regresión (Vapnik, 1995), nuestro objetivo es encontrar una función $f(x)$ que tiene como máximo una desviación ε del y_i objetivo obtenido para todos los datos de entrenamiento, y al mismo tiempo sea lo más llana posible. En otras palabras, no importan los errores siempre y cuando sean menores a ε , pero no se aceptara una desviación mayor a este valor. Esto puede ser importante si se quiere estar seguro de no perder más dinero que ε cuando se trata de los tipos de cambio, por ejemplo.

Para empezar se tomara el caso de una función lineal de la forma:

$$f(x) = \langle w, x \rangle + b, \quad \text{con } w \in X, b \in \mathbb{R} \quad (1)$$

donde $\langle \cdot, \cdot \rangle$ denota el producto vectorial en X . La llanura en este caso significa que se busca un valor pequeño de w . Una manera para asegurarse de esto es minimizar la norma, por ejemplo $\|w\|^2 = \langle w, w \rangle$. Podemos escribir este problema como un problema de optimización convexa.

$$\begin{aligned} &\text{Minimizar } \frac{1}{2} \|w\|^2 \\ &\text{Sujeto a } \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon \end{cases} \end{aligned} \quad (2)$$

La suposición tácita en (2) era que la función f existe y se aproxima a todos los pares (x_i, y_i) con una precisión ε , o en otras palabras que el problema de optimización convexa es factible.

A veces, sin embargo, este no puede ser, o también se desea permitir algunos errores. De forma análoga a la función de pérdida de “margen suave” (Bennett & Mangasarian, 1992) que fue adaptada para las máquinas de vectores soporte por Cortes y Vapnik, se puede introducir variables de holgura ξ_i, ξ_i^* para hacer frente a las limitaciones de lo contrario no sería factible el problema de optimización (2). Por lo tanto llegamos a la formula declarada en (Vapnik, 1995).

$$\begin{aligned} &\text{Minimizar } \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \\ &\text{Sujeto a } \begin{cases} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{aligned} \quad (3)$$

La constante $C > 0$ determina el equilibrio entre la llanura de f y hasta que valor son toleradas las desviaciones más grande que ε . Esto corresponde a tratar con una función de pérdida denominada ε -insensible $|\xi|_\varepsilon$ descrita por:

$$|\xi|_\varepsilon = \begin{cases} 0, & |\xi| \leq \varepsilon \\ |\xi| - \varepsilon, & \text{de otra forma} \end{cases} \quad (4)$$

La Ilustración 50 representa la situación gráficamente. Solo los puntos fuera de la región sombreada contribuyen al costo en la medida, ya que las desviaciones se penalizan de forma lineal. Resulta que en la mayoría de los casos, el problema de optimización (3) se puede resolver con más facilidad en su doble formulación. Por otra parte la formulación dual proporciona la clave para extender las máquinas de vectores soporte a funciones no lineales.

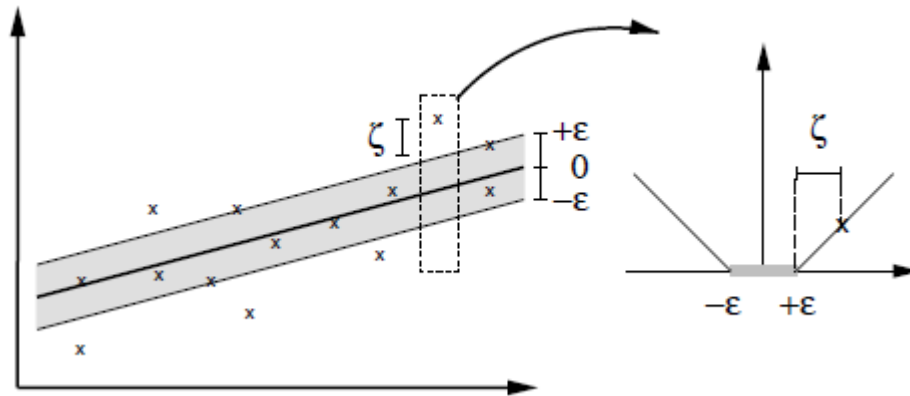


Ilustración 50. El margen suave de pérdida ajustado para una SVM lineal (Smola & Schölkopf, 2003)

Desvío estándar

El operador Weight by Deviation calcula la relevancia de los atributos de acuerdo a su desvío estándar, en estadística una variable que tiene un desvío estándar grande representa una variable que contiene mucha información. Por este motivo es que se puede jerarquizar la importancia de cada variable de acuerdo a su desvío estándar. El desvío estándar σ de un conjunto de valores x_i esta dado por la siguiente formula:

$$\sigma = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$$

donde n es el tamaño del conjunto de datos y $\bar{x}$ es la media del conjunto de datos.

Análisis de componentes principales (PCA)

Según la documentación de RapidMiner, el análisis de componentes principales es un procedimiento matemático que utiliza una transformación ortogonal para convertir un conjunto de observaciones de atributos correlacionados posiblemente en un conjunto de valores de atributos no correlacionadas llamadas componentes principales. El número de componentes principales es menor o igual que el número de atributos originales. Esta transformación se define de tal manera que la varianza del primer componente es tan alta como sea posible, y cada componente subsiguiente a su vez tiene la mayor varianza posible bajo la restricción que debe ser ortogonal (correlacionado con) los componentes anteriores.

A continuación se presenta una explicación de la transformación ortogonal en base a las explicaciones encontradas en (Grané Chávez, 2002).

Sea un conjunto de atributos numéricos $Y_1, Y_2, \dots, Y_m$, sea S su matriz de covarianza donde $S_{(i,j)} = Cov(Y_i, Y_j)$

Puesto que $S \geq 0$ y simétrica, su descomposición espectral es

$$S = T\Lambda T'$$

Donde $T'T = TT' = I$, con $T = [\vec{v}_1, \dots, \vec{v}_m]$ y $\Lambda = diag(\lambda_1, \dots, \lambda_m)$ con $\lambda_1 > \dots > \lambda_m$

Las componentes principales de Y son las nuevas variables

$$X_i = Y_i \vec{v}_i, \quad i = 1, \dots, m$$

Para cada i , la nueva variable X_i se constituye a partir del i -ésimo autovector de S .

Cálculo de los autovalores y autovectores de la matriz S

Para calcular los autovalores de la matriz S , es necesario calcular los valores de λ_i que cumplan la siguiente condición:

$$\det(S - \lambda I) = 0, \quad \text{donde } \Lambda = \lambda I = diag(\lambda_1, \dots, \lambda_m)$$

Luego para obtener los autovectores, para cada valor obtenido de λ_i es necesario obtener los vectores $\vec{v}_i$ que cumplan la ecuación:

$$(S - \lambda_i I)\vec{v}_i = 0$$

Es importante aclarar que los vectores $\vec{v}_i$ deben estar normalizados.

Relief

Relief es un algoritmo inspirado en el aprendizaje basado en casos que intenta obtener los atributos estadísticamente más relevantes. El algoritmo se basa en asignar un peso a cada atributo y seleccionar los atributos cuyo peso supera un umbral prefijado. Para este propósito, dada una instancia, Relief busca en el conjunto de datos dos vecinos, el más cercano de la misma clase (near-hit) y el de clase distinta (near-miss). El peso asociado a un atributo se modifica a partir de la distancia euclídea entre el valor del atributo de la instancia y el valor del mismo atributo de los vecinos encontrados. Las instancias se escogen aleatoriamente del conjunto de datos, un número determinado de veces. Cuando el número de instancias de la base de datos es pequeño, se realiza el proceso para cada una de las instancias, no existiendo aleatoriedad. Relief favorece a los atributos correlacionados sobre los relevantes, por lo que no está garantizado un conjunto óptimo. ReliefF es una versión mejorada en la que se seleccionan las k instancias más parecidas de la misma y diferente clase, respectivamente) y se toma su media (Ruiz Sánchez, 2005).

La posterior ilustración muestra la lógica del algoritmo Relief, en ella $W[A]$ es el vector de pesos para cada atributo A , n el tamaño del conjunto de datos y m la cantidad de atributos.

```
establecer todos los pesos  $W[A] := 0$ ;  
for  $i := 1$  to  $n$  do  
    begin  
        seleccionar aleatoriamente una instancia;  
        encontrar los vecinos hit H y miss M;  
        for  $A := 1$  to  $m$  do  
             $W[A] := W[A] - \text{diff}(A, R, H)/n + \text{diff}(A, R, M)/n$ ;  
        end;
```

Ilustración 51. Algoritmo de Relief

En la Ilustración 51 la distancia entre ejemplos se calcula de la siguiente manera, si A es un atributo discreto:

$$\text{diff}(A, X, Y) = \begin{cases} 0, & \text{si } \text{valor}(A, X) = \text{valor}(A, Y) \\ 1, & \text{de otra forma} \end{cases}$$

En el caso que A sea un atributo continuo, la diferencia se estima a través de la siguiente ecuación:

$$\text{diff}(A, X, Y) = \frac{|\text{valor}(A, X) - \text{valor}(A, Y)|}{\max(A) - \min(A)}$$

En problemas de regresión los valores del atributo objetivo son continuos, por lo tanto los vecinos más cercanos no se pueden utilizar. En lugar de exigir el conocimiento exacto de si dos instancias pertenecen a la misma clase o no, podemos introducir una especie de probabilidad si los valores objetivos de dos casos son diferentes. Esta probabilidad puede ser modelada con la distancia relativa entre los valores del atributo objetivo de los dos casos.

```

establecer todos los pesos  $N_{dc}$ ,  $N_{dA}[A]$ ,  $N_{dc\&dA}[A]$ ,  $W[A]$  en 0;

for  $i := 1$  to  $n$  do
  begin
    seleccionar aleatoriamente una instancia  $R_i$ ;
    seleccionar  $k$  instancias  $I_j$  cercanas a  $R_i$ ;

    for  $j := 1$  to  $k$  do
      begin
         $N_{dc} := N_{dc} + |clase(R_i) - clase(I_j)|/k$ ;

        for  $A := 1$  to  $m$  do
          begin
             $N_{dA}[A] := N_{dA}[A] + diff(A, R_i, I_j)/k$ ;
             $N_{dc\&dA}[A] := N_{dc\&dA}[A] + |clase(R_i) - clase(I_j)| * diff(A, R_i, I_j)/k$ ;
          end;
        end;
      end;
    end;

  for  $A := 1$  to  $m$  do
    begin
       $W[A] := N_{dc\&dA}[A]/N_{dc} - (N_{dA}[A] - N_{dc\&dA}[A])/(n - N_{dc})$ ;
    end;
  end;
end;

```

Ilustración 52. Algoritmo de Relief para problemas de regresión

La Ilustración anterior muestra el pseudocódigo del algoritmo Relief cuando el atributo objetivo es continuo, las variables en este procedimiento son las siguientes:

N_{dc} : representa el valor acumulado de la distancia relativa entre los valores del atributo objetivo de dos instancias distintas.

$N_{dA}[A]$: es el vector que contiene el grado en que los valores de los atributos son diferentes entre las instancias.

$N_{dc\&dA}[A]$: es el vector que contiene el grado en que los valores de los atributos son diferentes entre las instancias dado la distancia entre los valores de la clase.

$W[A]$: es el vector que contiene los pesos de los atributos.

Normalización de los pesos

Luego de calcular los pesos de los atributos según alguna de las técnicas mencionadas, los operadores permiten al usuario la normalización de los pesos a valores entre 0 y 1, para realizar la normalización es necesario obtener el máximo y el menor valor del conjunto de pesos de los atributos. Luego, el valor de cada peso normalizado W_i se obtiene a partir del peso sin normalizar w_i de la siguiente forma:

$$W_i = \frac{w_i - \text{Min}(w)}{\text{Max}(w) - \text{Min}(w)}$$

Anexo V

Regresión lineal simple

El concepto de regresión se refiere al “cuantum” o “cantidad de cambio” que experimenta una variable dependiente (Y), en relación al cambio de una unidad de una variable independiente (X). La regresión es un concepto estadístico estrechamente vinculado al concepto de correlación; mientras que la regresión estudia la naturaleza de la relación entre dos variables dependientes, la correlación estudia la estrechez de la relación entre esas dos variables una dependiente de la otra. Al igual que con otros procedimientos estadísticos, destaca que en la regresión lineal se desea realizar una inferencia estadística partiendo de los valores muestrales obtenidos; por lo tanto, se deben cumplir ciertos requisitos, que en el caso de la regresión línea son los siguientes (Pedroza & Dicovskyi, 2006):

- 1) Normalidad y homogeneidad de varianzas en la variable dependiente (Y) del modelo para los valores fijos de la variable independiente (X).
- 2) Independencia de las observaciones de Y .
- 3) Linealidad en la relación entre las variables.

El modelo de regresión simple es el siguiente

$$Y_1 = B_0 + B_1X_1 + e_1$$

donde:

Y_1 : es la variable dependiente.

B_0 : es la ordenada en el origen, o bien es el intercepto.

B_1 : es la pendiente de la recta de regresión.

e_1 : es el término del error, es decir la diferencia entre los valores predichos por la regresión y los valores reales.

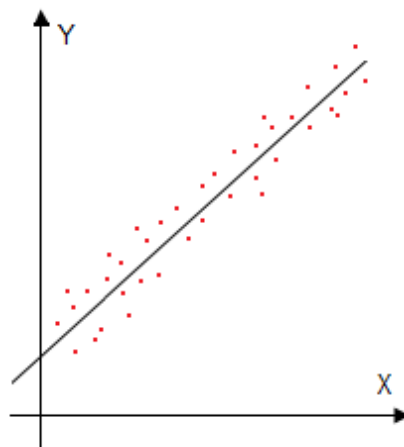


Ilustración 53. Regresión lineal simple

Regresión lineal múltiple

La regresión lineal múltiple es una extensión del modelo simple al que se incorporan dos o más variables independientes. Este modelo puede ser expresado como (Pedroza & Dicovskyi, 2006):

$$Y_i = B_0 + B_1X_{1i} + B_2X_{2i} + B_3X_{3i} + \dots + B_pX_{pi} + e_1$$

donde

X_{pi} : son las variables independientes.

B : son los parámetros estandarizados desconocidos.

e_i : son los términos de residuos o errores, de media = 0 y varianza constante.

Anexo VI

Inmediatamente se incluye el modelo de redes neuronales generado a través de la herramienta de minería de datos.

ImprovedNeuralNet

Hidden 1

=====

Node 1 (Sigmoid)

En poder del público-2: 0.091

En poder del público-1: 0.048

En poder del público-0: 0.040

Cuenta corriente-2: -0.019

Cuenta corriente-1: -0.024

Cuenta corriente-0: 0.066

Plazo fijo no ajustable por CER-2: 0.046

Plazo fijo no ajustable por CER-1: 0.120

Plazo fijo no ajustable por CER-0: 0.102

Depósitos del sector privado-2: 0.080

Depósitos del sector privado-1: 0.065

Depósitos del sector privado-0: 0.104

Depósitos del sector público-2: 0.039

Depósitos del sector público-1: 0.071

Depósitos del sector público-0: 0.059

Índice de precios al consumidor base Abril 2008 = 100-2: 0.075

Índice de precios al consumidor base Abril 2008 = 100-1: 0.046

Índice de precios al consumidor base Abril 2008 = 100-0: 0.125

Bias: -0.465

Node 2 (Sigmoid)

En poder del público-2: 0.057

En poder del público-1: 0.058

En poder del público-0: -0.046

Cuenta corriente-2: 0.029

Cuenta corriente-1: 0.021

Cuenta corriente-0: 0.070

Plazo fijo no ajustable por CER-1: 0.053

Plazo fijo no ajustable por CER-0: 0.103

Depósitos del sector privado-2: 0.049

Depósitos del sector privado-1: 0.050

Depósitos del sector privado-0: 0.100

Depósitos del sector público-2: 0.033

Depósitos del sector público-1: 0.121

Depósitos del sector público-0: 0.050

Índice de precios al consumidor base Abril 2008 = 100-2: 0.054

Índice de precios al consumidor base Abril 2008 = 100-1: 0.094

Índice de precios al consumidor base Abril 2008 = 100-0: 0.183

Bias: -0.511

Node 3 (Sigmoid)

En poder del público-2: 0.041

En poder del público-1: 0.037

En poder del público-0: 0.059

Cuenta corriente-2: 0.065

Cuenta corriente-1: 0.099

Cuenta corriente-0: 0.059

Plazo fijo no ajustable por CER-2: 0.064

Plazo fijo no ajustable por CER-1: 0.090

Plazo fijo no ajustable por CER-0: 0.056

Depósitos del sector privado-2: 0.051

Depósitos del sector privado-1: 0.032

Depósitos del sector privado-0: 0.083

Depósitos del sector público-2: 0.096

Depósitos del sector público-1: 0.042

Depósitos del sector público-0: 0.021

Índice de precios al consumidor base Abril 2008 = 100-2: 0.019

Índice de precios al consumidor base Abril 2008 = 100-1: -0.052

Índice de precios al consumidor base Abril 2008 = 100-0: -0.036

Bias: -0.376

En poder del público-2: -0.020
En poder del público-1: -0.038
En poder del público-0: -0.164
Cuenta corriente-2: -0.053
Cuenta corriente-1: -0.045
Cuenta corriente-0: 0.085
Plazo fijo no ajustable por CER-2: -0.040
Plazo fijo no ajustable por CER-1: -0.002
Plazo fijo no ajustable por CER-0: 0.070
Depósitos del sector privado-2: -0.088
Depósitos del sector privado-1: -0.008
Depósitos del sector privado-0: 0.044
Depósitos del sector público-2: -0.037
Depósitos del sector público-1: 0.069
Depósitos del sector público-0: 0.068
Índice de precios al consumidor base Abril 2008 = 100-2: 0.290
Índice de precios al consumidor base Abril 2008 = 100-1: 0.296
Índice de precios al consumidor base Abril 2008 = 100-0: 0.429
Bias: -0.176

Node 5 (Sigmoid)

En poder del público-2: 0.000
En poder del público-1: 0.039
En poder del público-0: 0.009
Cuenta corriente-2: 0.046
Cuenta corriente-1: -0.006
Cuenta corriente-0: -0.018
Plazo fijo no ajustable por CER-2: 0.019
Plazo fijo no ajustable por CER-1: -0.008
Plazo fijo no ajustable por CER-0: -0.030
Depósitos del sector privado-2: 0.100
Depósitos del sector privado-1: 0.077
Depósitos del sector privado-0: 0.006

Depósitos del sector público-1: 0.018

Depósitos del sector público-0: 0.013

Índice de precios al consumidor base Abril 2008 = 100-2: -0.175

Índice de precios al consumidor base Abril 2008 = 100-1: -0.197

Índice de precios al consumidor base Abril 2008 = 100-0: -0.265

Bias: -0.358

Node 6 (Sigmoid)

En poder del público-2: 0.075

En poder del público-1: 0.124

En poder del público-0: 0.024

Cuenta corriente-2: 0.032

Cuenta corriente-1: 0.014

Cuenta corriente-0: 0.022

Plazo fijo no ajustable por CER-2: 0.063

Plazo fijo no ajustable por CER-1: 0.091

Plazo fijo no ajustable por CER-0: 0.120

Depósitos del sector privado-2: 0.093

Depósitos del sector privado-1: 0.043

Depósitos del sector privado-0: 0.072

Depósitos del sector público-2: 0.019

Depósitos del sector público-1: 0.034

Depósitos del sector público-0: 0.014

Índice de precios al consumidor base Abril 2008 = 100-2: -0.008

Índice de precios al consumidor base Abril 2008 = 100-1: -0.035

Índice de precios al consumidor base Abril 2008 = 100-0: -0.011

Bias: -0.368

Node 7 (Sigmoid)

En poder del público-2: 0.015

En poder del público-1: 0.043

En poder del público-0: 0.151

Cuenta corriente-2: -0.020

Cuenta corriente-0: -0.065

Plazo fijo no ajustable por CER-2: 0.220

Plazo fijo no ajustable por CER-1: 0.134

Plazo fijo no ajustable por CER-0: 0.023

Depósitos del sector privado-2: 0.128

Depósitos del sector privado-1: 0.067

Depósitos del sector privado-0: 0.044

Depósitos del sector público-2: 0.023

Depósitos del sector público-1: -0.030

Depósitos del sector público-0: 0.074

Índice de precios al consumidor base Abril 2008 = 100-2: -0.417

Índice de precios al consumidor base Abril 2008 = 100-1: -0.407

Índice de precios al consumidor base Abril 2008 = 100-0: -0.545

Bias: -0.473

Node 8 (Sigmoid)

En poder del público-2: 0.079

En poder del público-1: -0.000

En poder del público-0: 0.083

Cuenta corriente-2: -0.015

Cuenta corriente-1: 0.059

Cuenta corriente-0: -0.100

Plazo fijo no ajustable por CER-2: 0.327

Plazo fijo no ajustable por CER-1: 0.172

Plazo fijo no ajustable por CER-0: 0.046

Depósitos del sector privado-2: 0.231

Depósitos del sector privado-1: 0.181

Depósitos del sector privado-0: 0.080

Depósitos del sector público-2: 0.102

Depósitos del sector público-1: -0.002

Depósitos del sector público-0: 0.064

Índice de precios al consumidor base Abril 2008 = 100-2: -0.571

Índice de precios al consumidor base Abril 2008 = 100-1: -0.622

Índice de precios al consumidor base Abril 2008 = 100-0: -0.857

Node 9 (Sigmoid)

En poder del público-2: 0.005
En poder del público-1: -0.039
En poder del público-0: -0.135
Cuenta corriente-2: -0.089
Cuenta corriente-1: -0.038
Cuenta corriente-0: 0.092
Plazo fijo no ajustable por CER-2: -0.097
Plazo fijo no ajustable por CER-1: -0.037
Plazo fijo no ajustable por CER-0: 0.091
Depósitos del sector privado-2: -0.099
Depósitos del sector privado-1: -0.031
Depósitos del sector privado-0: 0.006
Depósitos del sector público-2: -0.070
Depósitos del sector público-1: -0.007
Depósitos del sector público-0: 0.012
Índice de precios al consumidor base Abril 2008 = 100-2: 0.330
Índice de precios al consumidor base Abril 2008 = 100-1: 0.381
Índice de precios al consumidor base Abril 2008 = 100-0: 0.506
Bias: -0.085

Node 10 (Sigmoid)

En poder del público-2: 0.068
En poder del público-1: 0.001
En poder del público-0: -0.077
Cuenta corriente-2: -0.064
Cuenta corriente-1: -0.072
Cuenta corriente-0: 0.064
Plazo fijo no ajustable por CER-2: 0.105
Plazo fijo no ajustable por CER-1: 0.148
Plazo fijo no ajustable por CER-0: 0.146
Depósitos del sector privado-2: 0.054

Depósitos del sector privado-0: 0.093

Depósitos del sector público-2: 0.035

Depósitos del sector público-1: 0.145

Depósitos del sector público-0: 0.121

Índice de precios al consumidor base Abril 2008 = 100-2: 0.115

Índice de precios al consumidor base Abril 2008 = 100-1: 0.113

Índice de precios al consumidor base Abril 2008 = 100-0: 0.126

Bias: -0.516

Node 11 (Sigmoid)

En poder del público-2: 0.069

En poder del público-1: -0.003

En poder del público-0: -0.084

Cuenta corriente-2: -0.002

Cuenta corriente-1: 0.017

Cuenta corriente-0: 0.023

Plazo fijo no ajustable por CER-2: 0.003

Plazo fijo no ajustable por CER-1: 0.092

Plazo fijo no ajustable por CER-0: 0.108

Depósitos del sector privado-2: 0.009

Depósitos del sector privado-1: 0.102

Depósitos del sector privado-0: 0.067

Depósitos del sector público-2: 0.086

Depósitos del sector público-1: 0.071

Depósitos del sector público-0: 0.052

Índice de precios al consumidor base Abril 2008 = 100-2: 0.137

Índice de precios al consumidor base Abril 2008 = 100-1: 0.098

Índice de precios al consumidor base Abril 2008 = 100-0: 0.235

Bias: -0.420

=====

Regression (Linear)

Node 1: 0.360

Node 2: 0.442

Node 3: 0.109

Node 4: 0.567

Node 5: -0.313

Node 6: 0.104

Node 7: -0.673

Node 8: -0.962

Node 9: 0.625

Node 10: 0.496

Node 11: 0.457

Threshold: -0.667