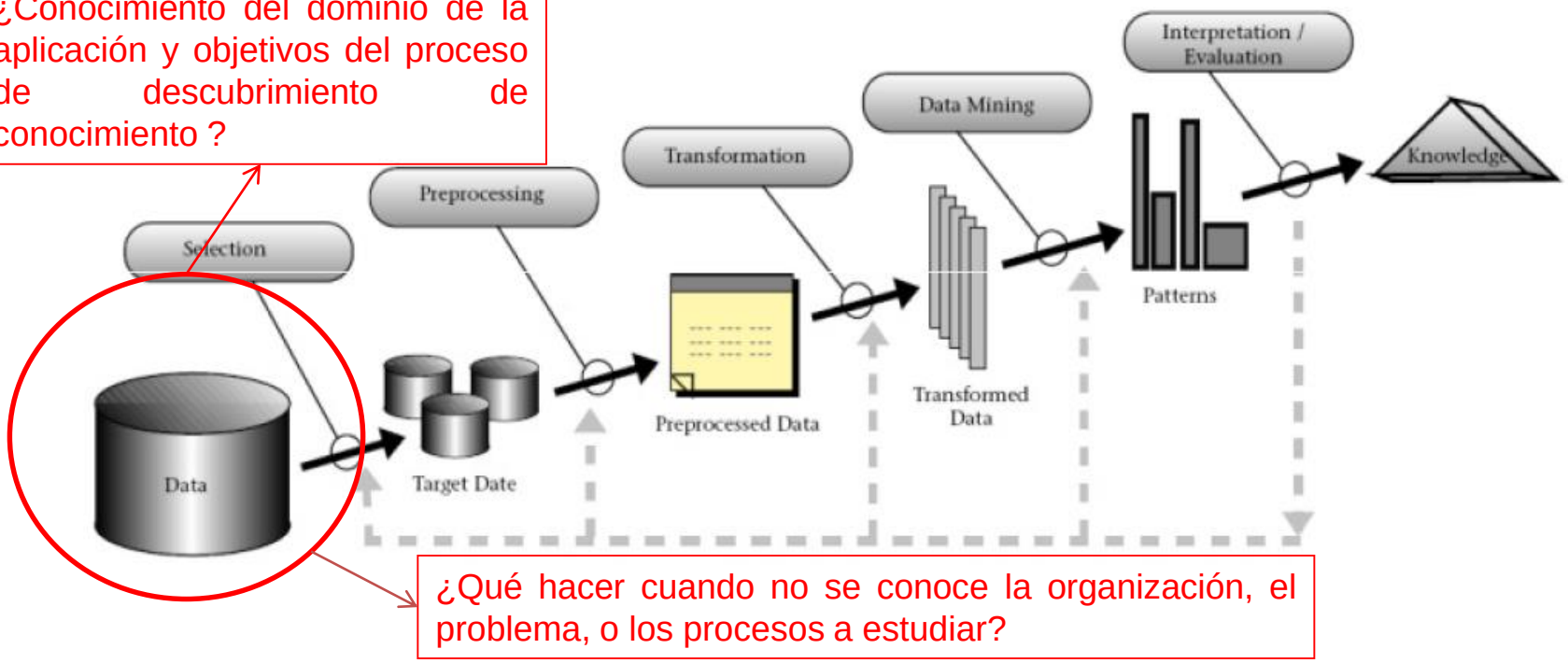
MIDANO

“Metodología para el desarrollo de aplicaciones de minería de datos basada en el análisis organizacional”

MIDANO

¿Conocimiento del dominio de la aplicación y objetivos del proceso de descubrimiento de conocimiento ?

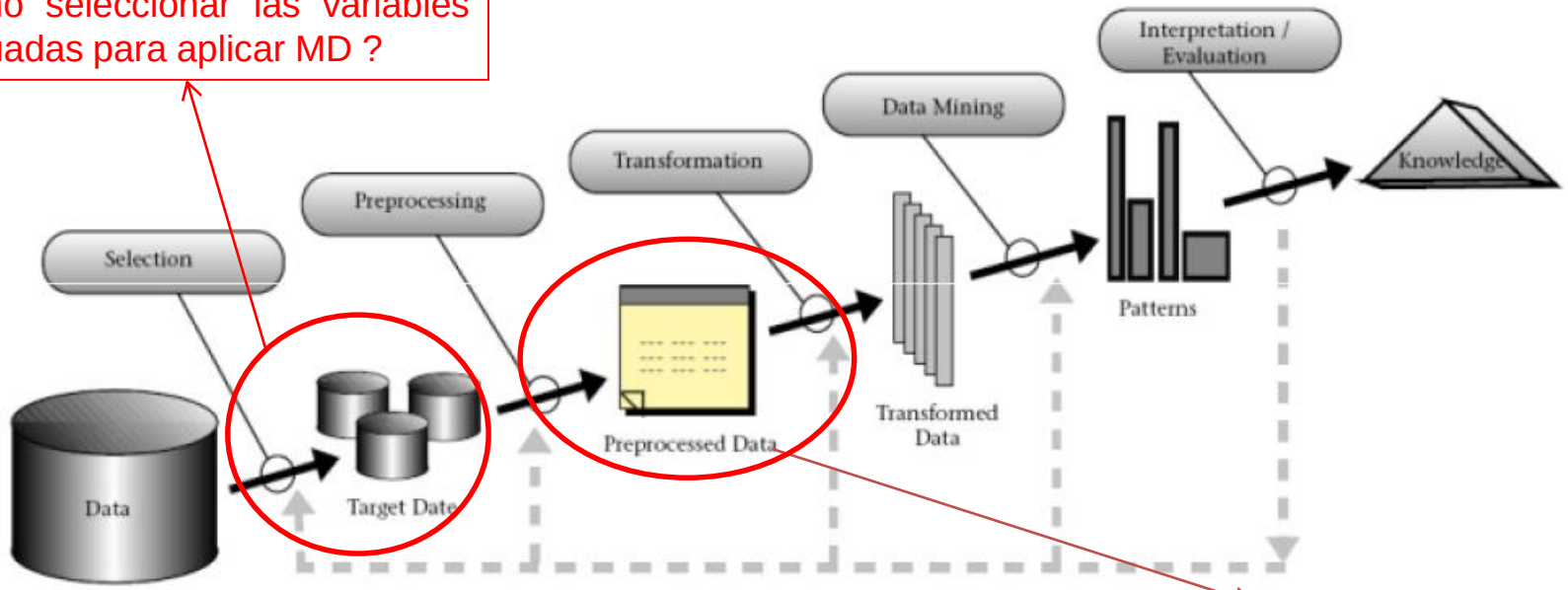


¿Qué hacer cuando no se conoce la organización, el problema, o los procesos a estudiar?

“Metodología para el desarrollo de aplicaciones de minería de datos basada en el análisis organizacional”

MIDANO

¿Cómo seleccionar las variables adecuadas para aplicar MD ?



¿Cómo realizar el procesamiento de datos?

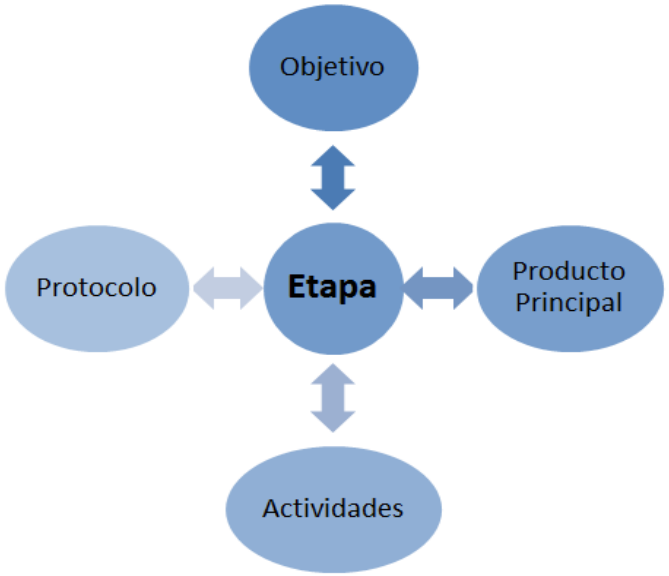
“Metodología para el desarrollo de aplicaciones de minería de datos basada en el análisis organizacional”

MIDANO

MIDANO consta de tres fases.

MIDANO

Cada fase de la metodología está dividida en etapas, los elementos que describen cada etapa. En cada etapa se especifican cuatro aspectos principales, como se describe a continuación.



Fase 1: Conocimiento de la Organización

Esta fase tiene como finalidad realizar un proceso de ingeniería de conocimiento, orientado a organizaciones/empresas, de las cuales no se conoce o se tiene poca información del (de los) problema(s), o los procesos a estudiar.

Etapa 1: Conocimiento de la Organización

1. Objetivo

- Conocer la organización/empresa, sus objetivos, procesos, objetos y actores

2. Protocolo de la Fase:

- Descripción de los elementos de la institución/empresa y sus características. Objetivos, Procesos , Objetos y Actores.
- Descripción de las relaciones entre estos elementos.
- Organización de estos elementos.

Etapa 2: Caracterización detallada de los procesos de la organización

1. Objetivo

- Conocer los procesos sobre los cuales se puede enfocar el proyecto de minería de datos.

2. Protocolo de la Fase:

- Familiarización con los procesos sobre los cuales se puede realizar la ingeniería de conocimiento
- Identificación de la fuente de conocimiento
- Familiarización con los ambientes computacionales donde se encuentran los datos a ser utilizados en cada proceso.

Etapa 3: Análisis de factibilidad y selección del proceso

1. Objetivo

- Analizar los procesos con la información proporcionada/recogida, con la finalidad de conocer la factibilidad de la aplicación de la minera de datos sobre cada uno de ellos

2. Protocolo de la Fase:

- Revisión de los procesos propuestos por los expertos
- Disponibilidad del experto o grupo de expertos
- Análisis de las fuentes de información sobre los procesos

Etapa 4: Análisis para caracterizar las posibles tareas de Minería de Datos

1. Objetivo

- Caracterizar las posibles tareas de minería de datos a realizar en el(los) proceso(s) seleccionado(s) en la fase anterior (objetivos, requerimientos, factibilidad, etc.), con la finalidad de escoger las tareas de MD de interés a desarrollar.

2. Protocolo de la Fase:

- Selección y descripción de los actores.
- Descripción de los escenarios actuales y posibles escenarios futuros de la institución/empresa.
- Especificación de los requerimientos para los posibles escenarios futuros (donde se puedan aplicar tarea(s) de MD)
- Elaboración de los casos de uso para los requerimientos funcionales

Etapa 5: Formalización del Problema/tareas de Minería de Datos (MD)

1. Objetivo

- Definir el(los) problema(s) formales de MD.

2. Protocolo de la fase

- Desarrollo de un informe, con la conceptualización del proceso a estudiar, la caracterización de sus problemáticas operacionales y del uso de la MD en dicho proceso.

Fase 2: Preparación de Datos

Para aplicar MD sobre un problema en específico, es necesario contar con un historial de datos asociado al problema en estudio.

Para realizar tareas de MD es necesario tener los datos integrados en una sola vista, la cual comúnmente se conoce como *Vista Minable*. Existen dos tipos de vista minable:

- **Vista Minable Conceptual (VMC):** describe en detalle cada una de las variables a tomar en cuenta para la tarea de MD, para cada escenario futuro seleccionado (proveniente de la primera fase de MIDANO).
- **Vista Minable Operativa (VMO):** Es el resultado de cargar los datos del historial y de realizar la etapa de tratamiento de datos, basado en la información de la VMC. La VMO se traduce a lo que se conoce como Vista Minable en la literatura, para realizar tareas de MD.

Fase 2: Preparación de Datos

- En esta fase se plantea realizar la preparación de los datos desarrollando dos etapas. Los productos más resaltantes de esta fase son las vistas minables (conceptual y operativa) y las variables objetivos.

Etapa 1: Caracterización de los datos del dominio de la aplicación

a. Objetivos

- Ubicar y comprender los datos asociados a el(los) escenario(s) futuro(s)
- Construir una VMC que tenga las variables de interés para el caso de estudio
- Construir una VMO inicial
- Definir la(s) variable(s) objetivo(s) en la vista minable operativa

b. Protocolo de la etapa

- Comprender la fuente de datos de entrada
- Generar la VMC y la VMO inicial
- Integración de los datos de entrada
- Definir la(s) variable(s) objetivo(s)

Etapa 1: Caracterización de los datos del dominio de la aplicación

c. **Productos principales**

- Documento que describe las características de los repositorios donde se encuentran los datos
- Documento que describe la VMC, la cual es presentada en una tabla descriptiva.
- Archivo donde esta almacenada la VMO
- Documento que describe las características de la(s) variable(s) objetivo(s)

Etapa 2: Tratamiento de datos

a. Objetivos

Esta etapa se centra en generar datos de calidad, es decir, sin anomalías, sin inconsistencias de formato, sin capturas erróneas, sin campos vacíos; aplicando métodos de limpieza, transformación y reducción sobre la vista minable operativa.

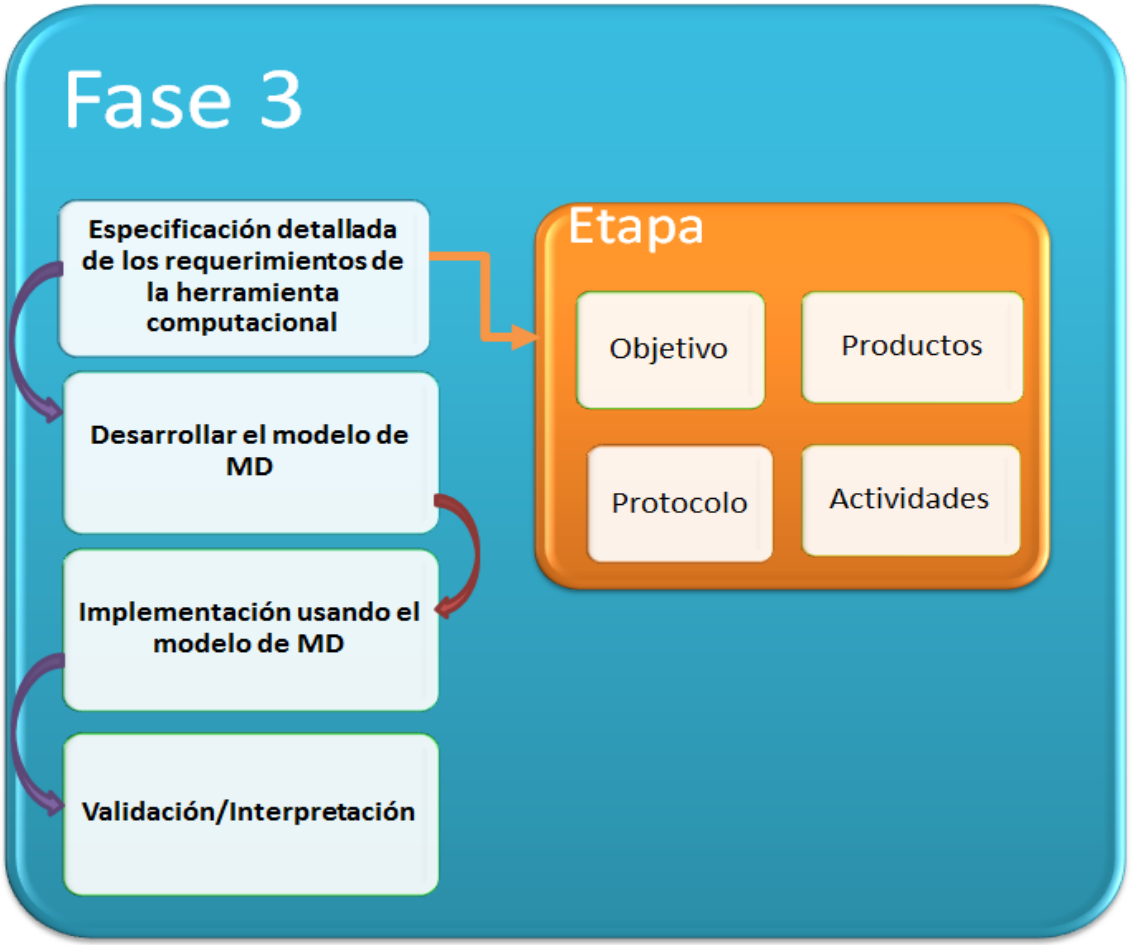
b. Protocolo de la etapa

- Limpieza
- Transformación
- Reducción

c. Productos principales

- VMO depurada
- Documento descriptivo de los tratamientos realizados usando tablas descriptivas con información pertinente.

Fase 3: Desarrollo de herramientas de MD



Etapa 1: Especificación detallada de los requerimientos de la herramienta computacional

a. Objetivos

captar los requerimientos no funcionales.

b. Protocolo de la etapa

- Requisitos de interfaz de usuario,
- Interfaces de software,
- Requerimientos de desempeño,
- Adicionalmente se pueden mencionar: de portabilidad, costos, rendimiento, accesibilidad, entre otros.

Etapa 2: Desarrollar el modelo de MD

a. Objetivos

escoger el modelo de MD resultante de la comparación de varias técnicas para una misma tarea.

b. Protocolo de la etapa

- Selección del Software para realizar las tareas de MD
- Escoger la técnica de MD para la tarea identificada.
- Definir cuáles son los datos de entrenamiento y de prueba dispuestos en la vista minable,
- Comenzar a realizar pruebas sobre la vista minable, para ir llenando la tabla comparativa de las técnicas de MD.
- Definir una estrategia para la validación de la técnica seleccionada, aplicarla y observar el rendimiento.
- Realizar las correcciones necesarias
- Repetir el procedimiento de ser necesario

Etapa 3: Implementación usando el modelo de MD

a. **Objetivos**

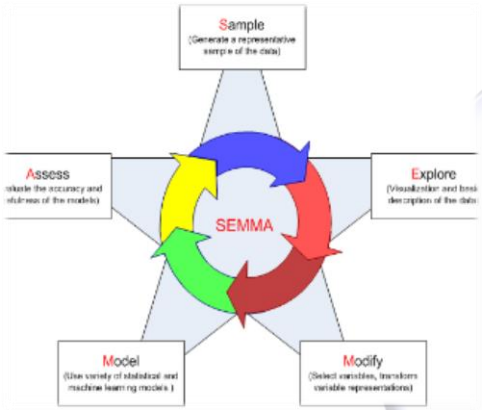
   Realizar la herramienta de MD con el modelo seleccionado.

Etapa 4: Validación/Interpretación

a. **Objetivos**

   Validar la herramienta de MD.

Fase 3: Desarrollo de herramientas de MD



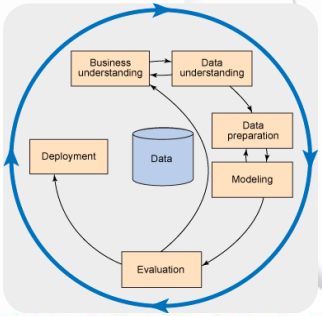
SEMMA

- Orientado a la parte técnica
- Carece de un análisis del problema.

Se puede usar cualquier metodología para esta fase de desarrollo de tareas de MD, mezclada con las fases de MIDANO

CRISP-DM

- Proceso continuo y progresivo del proceso de creación
- Más utilizado por empresas que trabajan con DM



CATALYST

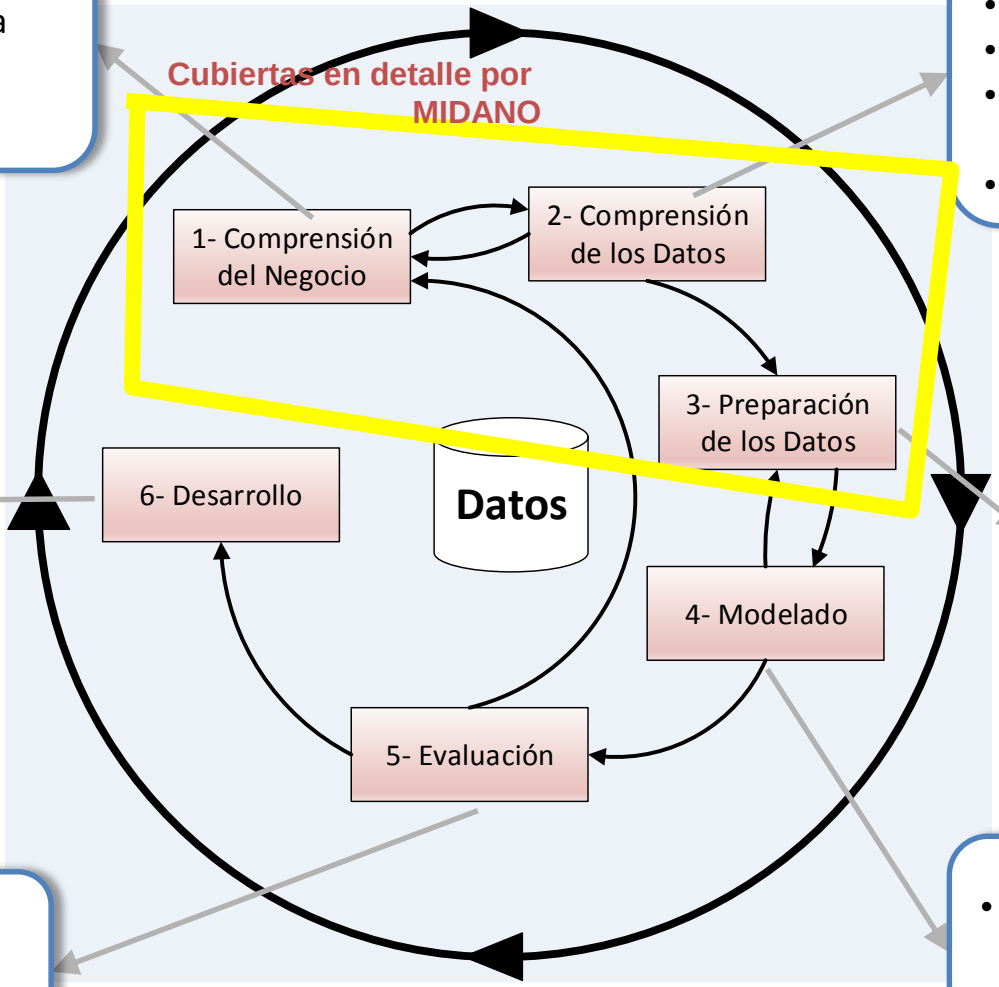
- Estructura en “boxes”
- Primer Modelo: Analiza el problema.
- Segundo Modelo: Solución en el aspecto técnico.

Fase 3: Desarrollo de herramientas de MD

CRISP-DM

- Objetivos y criterios de éxito del negocio y de la MD
- Plan del Proyecto.

- Análisis inicial de datos
- Recolección
- Descripción
- Identificación de problemas
- Verificación de calidad



- Plan para el desarrollo
- Informe final
- Presentación final
- Revisión general del proyecto

- Selección de datos
- Preparar, limpiar y/o construir datos
- Generar nuevos registros
- Integrar o formatear datos

- Evaluar el modelo
- Decisión sobre el modelo.

- Selección de técnica de modelado
- Obtener el modelo.

UNIVERSIDAD
DE LOS ANDES
MERIDA VENEZUELA

Ejemplo uso MIDANO

Fase 1: Conocimiento de la Organización

Caso de Estudio: Empresa Petrolera

Etapa 1: Conocimiento de la organización:

Se trata de una empresa que se encarga de la exploración, extracción, producción, mejoramiento y comercialización de crudo extrapesado.

Etapa 2: Caracterización de los procesos de la organización

La cadena de valor de la empresa se muestra en la siguiente figura, donde el proceso principal objeto de estudio se concentra en la tercera etapa de la cadena de valor.

Para el grupo de expertos, una de las etapas más importantes para obtener el producto deseado es la refinación, llevada a cabo en lo que se conoce como “complejo mejorador”.

Fase 1: Conocimiento de la Organización

Caso de Estudio: Empresa Petrolera

Etapa 3: Selección del Proceso

Se estudió cada uno de los subproceso (objetivos, actividades, productos, etc.), y se obtuvo la interacción entre ellos.

En la tabla se ilustra este proceso de priorización y selección, considerando sólo los dos procesos que resultaron mejor ponderados en este caso de estudio.

Criterios	CDU	DCU
Importancia para la organización	5	5
Propósito de la MD	5	5
Interacciones entre procesos	2	4
Procesos dependientes	5	3
Importancia de la calidad del producto	4	4
Seguridad Industrial	4	5
Replicabilidad de la herramienta desarrollada	5	4
Cantidad de Expertos	5	5
Fuentes de información	5	5
Confidencialidad de la información	3	3
¿Qué información se recoge del proceso para ser almacenada?	5	5
Con que frecuencia se recoge la información almacenada	4	4
¿Qué herramientas se cuentan, para recolectar y manipular la información?	4	4
Total sin ponderación	56	56
Total ponderado	83	76

Descripción del escenario futuro

El escenario futuro seleccionado es para **predecir la calidad del producto y optimizar la cantidad de nafta a la salida de la columna destilador atmosférico.**

Fase 1: Conocimiento de la Organización

Caso de Estudio: Empresa Petrolera

Etapa 4: Análisis para caracterizar las posibles tareas de Minería de Datos (MD)

Descripción del escenario actual

Resultados que se obtienen	Actor(es) asociado(s)	Variables Asociadas	Actividades que se realizan
Gasoil directo (SRGO), nafta pesada y residuo atmosférica.	• Expertos asociados al proceso • Ingenieros de Procesos • Operadores • Unidad de destilación atmosférica	• Tren de precalentamiento: temperatura de la carga. • Desaladores: tiempo para el asentamiento y separación del agua del petróleo, presión. • Hornos de crudo: temperatura • Columna de crudo: presión, temperatura, rata de vapor de despojamiento.	• Carga del crudo. • Precalentamiento del crudo diluido. • Desalado. • Precalentamiento del crudo desalado. • Generación de cortes de crudo en la columna.

Fase 1: Conocimiento de la Organización

Caso de Estudio: Empresa Petrolera

Descripción del escenario futuro

Resultados que se desean obtener	Actor(es) asociado(s)	Variables Asociadas	Actividades de MD que se realizarían	Funcionalidades nuevas
Predicción de la calidad del producto, para optimizar el proceso	• Expertos asociados al proceso • Operadores • Columna de crudo	Presión, temperatura de tope y rata de vapor de despojamiento de la columna de crudo.	Predicción	• Predicción de las características del producto, según las condiciones de funcionamiento de la torre de crudo. • Ayudar a optimizar el proceso de producción, generando información para orientar a los actores en la toma de decisiones con la predicción (es) resultante(s).

Descripción del escenario futuro

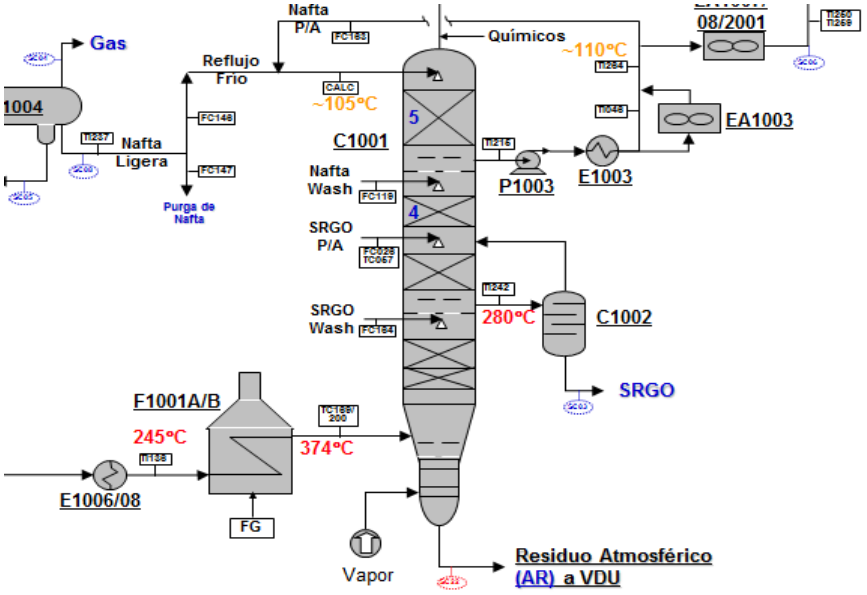
El escenario futuro seleccionado es para **predecir la calidad del producto y optimizar la cantidad de nafta a la salida de la columna destilador atmosférico.**

Fase 2: Preparación de datos

Etapa 1. Dominio de la aplicación

- ***Comprensión de los datos de entrada***

  A través de los diagramas de instrumentos de la planta, se determinaron cuáles son los datos asociados a las variables, se tomaron las variables más importantes asociadas al escenario futuro



Fase 2: Preparación de datos

Etapa 1. Dominio de la aplicación

- ***Construcción de la VMC***

  A partir del escenario futuro escogido, y con apoyo del grupo de expertos, se construyó una VMC con las características de la Tabla 1. Debido a que la misma cuenta con más de cien variables, sólo se presenta una pequeña muestra.

Tabla 1. Muestra de la VMC del escenario seleccionado

Variable	Descripción	Dependencia	Observaciones
11FC900	Flujo de nafta de lavado para la preflash	Identificar relación con FY119	Controlada
11PI1005	Presión de entrada de nafta wash	–	No es relevante para el estudio
11PI001A	Presión tope de la columna preflash	–	–

Fase 2: Preparación de datos

Etapa 1. Dominio de la aplicación

- ***Construcción de la VMO***

  Se carga el historial en un archivo con las variables obtenidas en la VMC.

  Los datos proporcionados por la empresa fueron entregados en formato Excel, donde todos los datos están integrados en un documento menos una variable de laboratorio, ya que la misma, es tomada con una frecuencia diferente a las demás variables.

Fase 2: Preparación de datos

Etapa 1. Dominio de la aplicación

- ***Integración de los datos de entrada***

Fecha	11_FI158T_PNT	....	11_FC010_MEAS
01.01.2009 0:00:00	320.139504	....	39.03201294
01.01.2009 0:05:00	318.8554796		39.03201294
01.01.2009 0:10:00	315.9257853		39.03201294
01.01.2009 0:15:00	316.9394877		39.03201294
01.01.2009 0:20:00	316.2324899		39.03201294
01.01.2009 0:25:00	318.2673392		39.03201294
01.01.2009 0:30:00	311.0020414		39.03201294
01.01.2009 0:35:00	314.7039024		39.03201294
.	.	.	.
.	.	.	.
.	.	.	.

(a) Formato de la tabla de datos con las variables asociadas a los sensores de la planta

Fecha	[°API]
02.07.2008 05:00:00	45.9
03.07.2008 05:00:00	46.1
04.07.2008 05:00:00	46.1
05.07.2008 05:00:00	46.2
06.07.2008 05:00:00	46.4
07.07.2008 05:00:00	45.8
08.07.2008 05:00:00	46
09.07.2008 05:00:00	45.6
10.07.2008 05:00:00	45.1
11.07.2008 05:00:00	45.4
12.07.2008 05:00:00	45.3
13.07.2008 05:00:00	45.6
.	.
.	.
.	.

(b) Formato de la tabla de datos de gravedad API (medición de laboratorio)

Fase 2: Preparación de datos

Etapa 1. Dominio de la aplicación

- ***Construcción de la VMO***

	11_FI158T_PNT	51_FT006_PNT
02.01.2009 23:30:00	314.7672055	1393.200177
02.01.2009 23:35:00	313.6730738	1396.853361
02.01.2009 23:40:00	317.3760808	1391.633283
02.01.2009 23:45:00	314.5253747	1391.253645
02.01.2009 23:50:00	315.1430386	1398.356516
02.01.2009 23:55:00	311.9205457	1400.912088
03.01.2009 0:00:00	312.5063793	1392.555884
03.01.2009 0:05:00	312.7566352	1394.478128
03.01.2009 0:10:00	312.8069345	1399.825388
03.01.2009 0:15:00	312.0659453	1401.837267

	Gravedad API a 60 °F
	[°API]
02.01.2009 05:00:00	47
03.01.2009 05:00:00	46.7
04.01.2009 05:00:00	46.8
05.01.2009 05:00:00	47.1
06.01.2009 05:00:00	48.6
07.01.2009 05:00:00	46.9

Vista minable operativa(VMO)

	11_FI158T_PNT	51_FT006_PNT	[°API]
02.01.2009 23:30:00	314.7672055	1393.200177	47
02.01.2009 23:35:00	313.6730738	1396.853361	47
02.01.2009 23:40:00	317.3760808	1391.633283	47
02.01.2009 23:45:00	314.5253747	1391.253645	47
02.01.2009 23:50:00	315.1430386	1398.356516	47
02.01.2009 23:55:00	311.9205457	1400.912088	47
03.01.2009 0:00:00	312.5063793	1392.555884	46.7
03.01.2009 0:05:00	312.7566352	1394.478128	46.7
03.01.2009 0:10:00	312.8069345	1399.825388	46.7
03.01.2009 0:15:00	312.0659453	1401.837267	46.7

Fase 2: Preparación de datos

Etapa 1. Dominio de la aplicación

- ***Definir las variables objetivo***

  Observar el(los) objetivo(s) de cada una de las variables en el escenario futuro seleccionado. Con el escenario futuro seleccionado se obtiene lo siguiente:

  - Escenario futuro: Producir la mayor cantidad de Nafta a 46 API
  - Funcionalidades nuevas: Predicción del API del producto, según las condiciones de funcionamiento de la torre de destilación.

Tabla 2. Variables objetivos

Variables objetivo	Observaciones
API NAFTA	Predecir el api de la nafta
11FC158	Maximizar el flujo de nafta producto a 46 api

Fase 2: Preparación de datos

Etapa 2. Tratamiento de Datos

- ***Limpieza***

Se ubicaron las variables con más errores en la VMO. Los resultados obtenidos son reflejados en la Tabla 3.

Tabla 3. Variables con mas errores en la VMO

PERIODO	Ene-Mar 2008	Abril-Jul 2008	Jul-Sept 2009	Oct-Dic 2010	Abril-Jul 2011	Ener-Marzo 2012	Oct-Dic 2013
NOMBRE DE LA VARIABLE	% Error	% Error	% Error	% Error	% Error	% Error	% Error
11_FC158_MEAS	0,0063593	92,43361416	100	100	100	100	100
11_FC044_MEAS	15,42766296	60,76037112	22,43088159	100	100	0,004852014	100
11_FC300_MEAS	0,0063593	18,76932921	4,478724197	100	100	100	100
11_FC119_MEAS	0,0063593	39,43158793	4,478724197	100	100	100	100
11_FC133_MEAS	99,9936407	99,7493868	100	0,0054638	14,629017772	33,915574496	0,007715454

Fase 2: Preparación de datos

Etapa 2. Tratamiento de Datos

- ***Limpieza***

La Tabla fue evaluada con los expertos del proceso con la finalidad de definir acciones que se podrían tomar. Las acciones tomadas son resumidas en la tabla 4, donde se describe la justificación de cada acción realizada.

Tabla 4. Acciones tomadas con las variables con más anomalías en la VMO

NOMBRE	JUSTIFICACIÓN	ACCIÓN
11_FC158_MEAS	Se puede eliminar porque es el mismo registro que el 11_FI158T_PNT	Eliminar de la VMO
11_FC044_MEAS	Se puede eliminar del estudio (es una línea de arranque o usada en operaciones muy puntuales).	Eliminar de la VMO
11_FC300_MEAS	Preferiblemente incluirla, si esta 11_FT300_PNT, se puede eliminar	Estudiar
11_FC119_MEAS	Preferiblemente incluirla, si esta FC119, se puede eliminar	Estudiar
11_FC133_MEAS	Se puede eliminar	Eliminar de la VMO

Fase 2: Preparación de datos

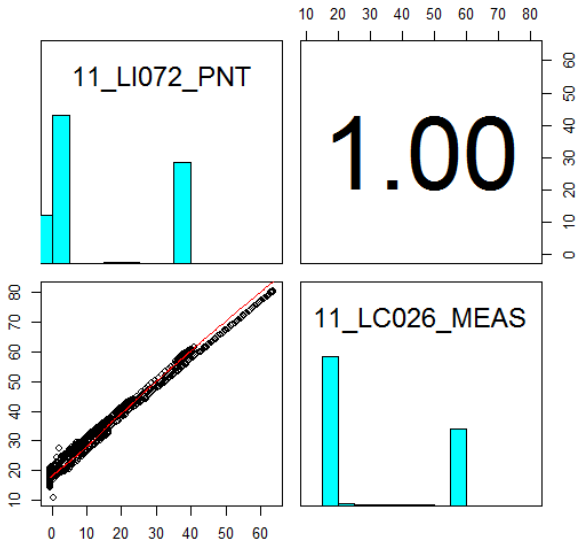
Etapa 2. Tratamiento de Datos

- ***Limpieza***

Para las variables que tienen dependencias con otras variables, se construyeron modelos lineales para sustituir los datos dañados por el resultado de estas relaciones.

Dichas dependencias pueden ser apoyadas y justificadas usando un gráfico de dispersión. En este caso, se seleccionó un diagrama de dispersión con las siguientes características:

- El histograma y nombre de cada variable (ver la diagonal de la Figura).
- La distribución de los puntos entre las dos variables y la curva regresada (parte inferior izquierda de la Figura).
- El coeficiente de correlación entre parejas de variables (parte superior derecha de la Figura).



- ***Transformación***

En este estudio no aplica el proceso de transformación, debido a que toda la data se encuentra en un formato consistente de unidades y magnitudes en las variables.

Fase 2: Preparación de datos

Etapa 2. Tratamiento de Datos

- ***Reducción***

Se realizaron análisis estadísticos entre variables que el experto identificó con dependencias en la VMC y además se construyó una matriz con la correlación entre todas las variables, con la finalidad de identificar las variables altamente correlacionadas.

	11_FI158T_PNT	11_FIC011_PNT	11_FC010_MEAS	11_FC159_MEAS	11SC01_BSW_NUM	11SC24_BSW_NUM
11_FI158T_PNT	1	0.368111972669728	0.484318010844852	-0.204051244869186	0.144851822147737	-0.311201934354127
11_FIC011_PNT	0.368111972669728	1	0.210678189668377	0.180986820619784	0.270251411195036	-0.067983791041037
11_FC010_MEAS	0.484318010844852	0.210678189668377	1	0.0075735692671944	-0.066805623663014	0.0505010255310379
11_FC159_MEAS	-0.204051244869186	0.180986820619784	0.0075735692671944	1	0.373898847616421	0.865146726620298
11SC01_BSW_NUM	0.144851822147737	0.270251411195036	-0.066805623663014	0.373898847616421	1	0.340144245505393
11SC24_BSW_NUM	-0.311201934354127	-0.067983791041037	0.0505010255310379	0.865146726620298	0.340144245505393	1
11_PI1001A_PNT	0.149797893568095	-0.230486068559374	0.0795532368827437	-0.530813442822542	-0.169561318779577	-0.265363516836238
11_PDI1001_PNT	-0.3513783530082	-0.064687446799048	0.0146327207575092	0.898505092755542	0.280231203226639	0.987591989101072
11_FC069_MEAS	0.836782714193593	0.429652170296929	0.412503957853951	-0.246747320588408	0.069211016952507	-0.427705464814407
11_TC167_MEAS	0.540859492533276	0.254940613271293	0.297113202094931	0.136683941679193	0.121901253006149	0.0642377397275748
11_TI168_PNT	0.528217711006828	0.251589322326238	0.294849147655103	0.16455588536907	0.131377400543795	0.0938984439071776
11_FC097_MEAS	0.863314754277501	0.4603093266029	0.453315700501899	-0.146716637712636	0.198073044334871	-0.327059949655631
11_TI205_PNT	0.465570226005517	0.231037998971888	0.128340774528971	-0.247838628436702	-0.262645490028422	-0.465553704306329
11_PI149_PNT	-0.023605088600139	0.24201692083265	0.0128921387555581	0.479588931261072	-0.030646242732858	0.183565008277804
11_PDI148_PNT	0.290003045342975	-0.090345184479632	0.0422040129610041	-0.978816278934173	-0.367572420350681	-0.911512964571754
11_FC164_MEAS	0.520628675400118	0.241198936667788	0.327886765160563	-0.021969684525066	-0.060401548555582	-0.074430658678374
11_FC114_MEAS	0.669735791599634	0.377566765068302	0.232778574068234	-0.481099372744771	0.152729477350047	-0.612534965210986
11_TI206_PNT	0.379966415482969	0.0902448084933138	0.0880606872816109	-0.68915732489563	-0.431286570684153	-0.781755090555928
11_LC016_MEAS	0.210108572264961	0.265240077290013	0.125599091378307	0.366193900688599	0.410047378697821	0.228435524378336
11_FC148_MEAS	-0.55524114311874	-0.255163977546328	-0.316329956267743	-0.083064170826386	-0.187975808467312	-0.041865764739969
11_FI149_PNT	-0.429285166720203	-0.103822072026452	-0.107453649503538	0.80189622301047	0.150822460062886	0.833467964006083
11_TI204_PNT	0.716377153818456	0.303070770590223	0.344980429883973	-0.184537891817327	0.009581408992425	-0.2871877823792

Se realizó la reducción de la VMO sobre las variables con alta correlación. La VMO inicial contaba con 97 variables, después de realizar la limpieza y reducción de datos la VMO final cuenta con 33 variables.

Fase 2: Preparación de datos

Etapa 2. Tratamiento de Datos

- ***Reducción***

Tabla 5. Reducción de Variables

Nombre	Resultado	Acción	Justificación
11_PI1001B_PNT	Correlación alta con 11_PI1001A_PNT	ELIMINAR	Variable altamente correlacionada, donde ambas aportan la misma información
11_PI157A_PNT/ 11_PI157B_PNT	Correlación alta con 11_PI149_PNT	ELIMINAR	Variable altamente correlacionada, donde las tres aportan la misma información
11_PI156_PNT	Correlación alta con 11_PI155_PNT	ELIMINAR	Variable altamente correlacionada, donde ambas aportan la misma información
. . .			

Se realizó la reducción de la VMO sobre las variables con alta correlación. La VMO inicial contaba con 97 variables, después de realizar la limpieza y reducción de datos la VMO final cuenta con 33 variables.

Fase 3: Desarrollo de herramientas de MD

Etapa 1: Especificación detallada de los requerimientos de la herramienta computacional

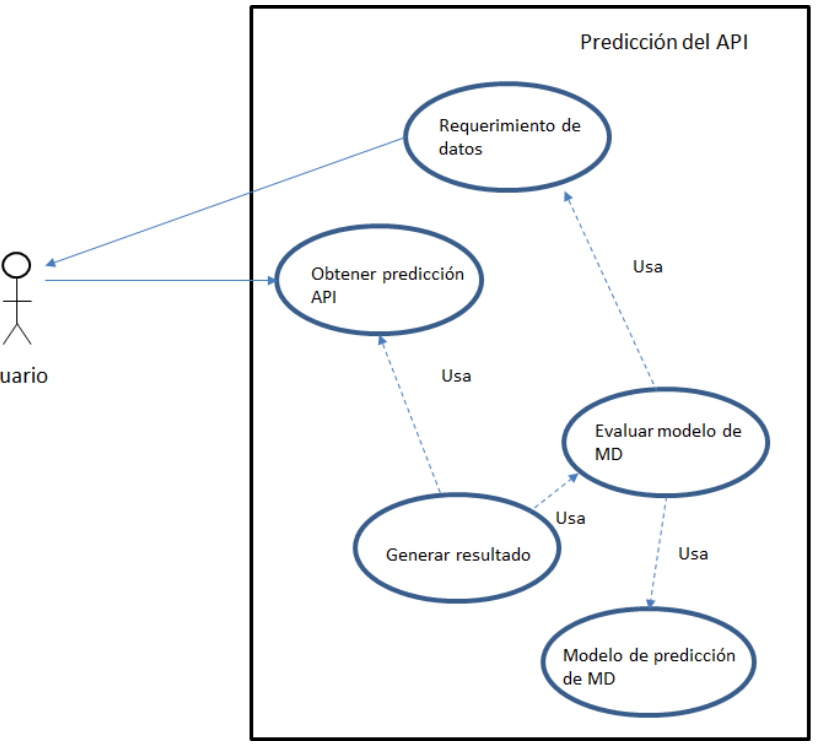
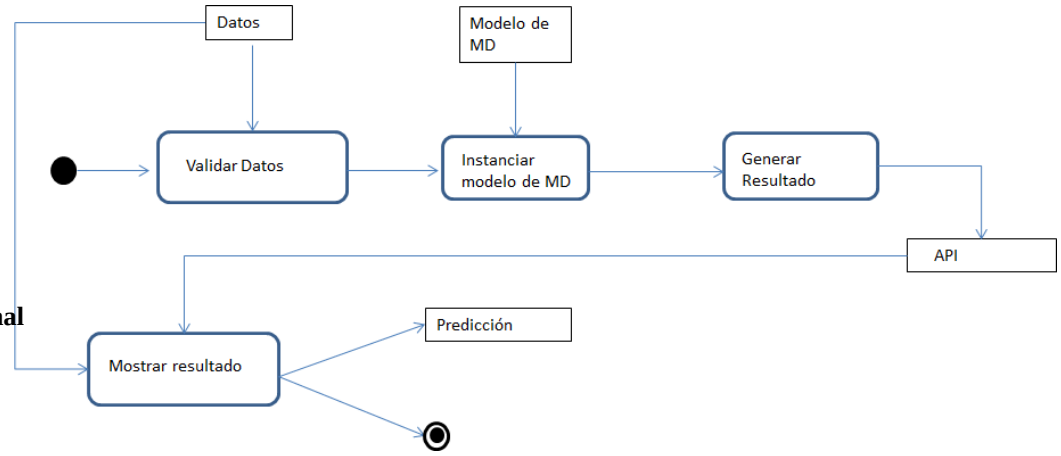


Diagrama de actividades de la herramienta computacional para la predicción del API

Entre los requisitos no funcionales:

- Requisitos de interfaz de usuario.
- Interfaces de software.
- Requerimientos de desempeño.
- Otros: de portabilidad, costos, accesibilidad, etc.

Caso de uso para la predicción del caso de estudio



Fase 3: Desarrollo de herramientas de MD

Etapa 2: Desarrollar el modelo de MD

Predicción del API de nafta:

1. Selección del Software para realizar las tareas de MD: Weka.
2. Escoger la tarea de MD para el escenario futuro: predicción.
3. Definir cuáles son los datos de entrenamiento y de prueba dispuestos en la vista minable: Los datos de entrenamiento son el 70% de los registros de la VMO y el resto será utilizado para realizar una validación cruzada[2] con Weka
4. Selección del algoritmo de MD: Los algoritmos considerados a evaluar son en general algoritmos basados en regresión lineal y redes neuronales.

Tabla 6. Algoritmos evaluados para la predicción del API de nafta

Algoritmo	Mean absolute error	Root mean squared error
LinearRegression	0.3546	0.503
RBFNetwork	0.4964	0.6953
SimpleLinearRegression	0.4422	0.6198
PaceRegression	0.3562	0.5028
IsotonicRegression	0.4293	0.5923

Validar

7. Modelo de MD: modelo que expresa las relaciones, a través de fórmulas y reglas, entre las variables del proceso. La ecuación obtenida para la predicción del API obtenida con *LinearRegression* viene dada por

API = 0.0032 * 11_FI158T_PNT + 0.0006 * 11_FIC011_PNT - 0.004* 11_FC010_MEAS -0.1673 * 11SC01_BSW_NUM +2.8329 * 11_PI149_PNT -0.1857 * 11_FC114_MEAS +0.0481 * 11_TI206_PNT - 0.0027 * 11_LC016_MEAS -0.0087 * 11_FC148_MEAS + 0.0042 * ….

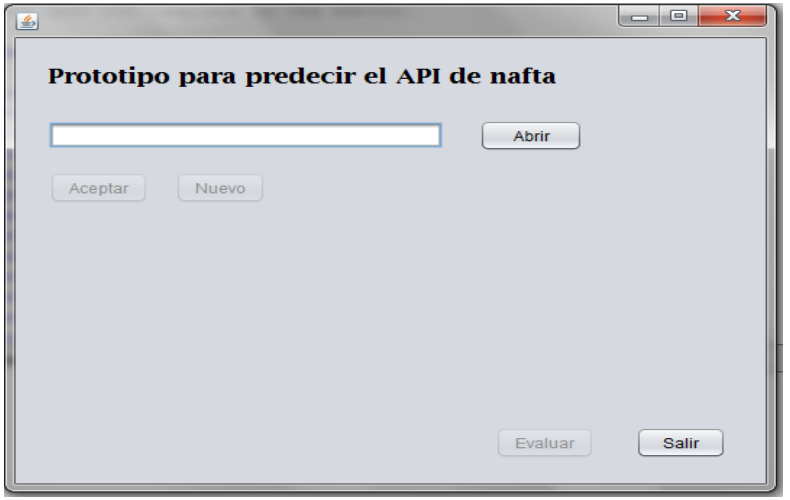
Fase 3: Desarrollo de herramientas de MD

Etapa 3: Implementación usando el modelo de MD

- **Formato de la entrada**

	A	B	C	D	E	F	G	H	I	J
1	Fecha	11_FI158T_PI	11_FIC011_P	11_FC010_M	11SC01_BSW	11SC24_BSW	11_TC167_M	11_TI205_PN	11_PI149_PN	11_FC114_M
2	-	303.373624	156.608638	40.409008	2.23353173	0.03146036	376.358974	370.675293	1.15274024	7.02434635
3	-	303.980151	151.507827	40.409008	2.23392856	0.03146106	376.532866	370.675293	1.15274024	7.02434635
4	-	304.417303	142.997793	40.409008	2.23432538	0.03146176	376.694006	370.675293	1.15274024	7.02434635
5	-	297.576434	141.570275	40.409008	2.23472221	0.03146246	376.845872	370.675293	1.15274024	7.02434635
6	-	301.650863	142.512228	40.409008	2.23511903	0.03146316	377.019765	370.675293	1.15274024	7.02434635
7	-	300.597413	161.583226	40.409008	2.23551586	0.03146386	377.193657	370.675293	1.15274024	7.02434635
8	-	300.120383	240.557088	40.409008	2.23591268	0.03146456	377.367549	370.675293	1.15274024	7.02434635
9	-	295.659939	244.557774	40.409008	2.23630951	0.03146526	377.541442	370.675293	1.15274024	7.02434635
10	-	296.365927	248.684239	40.409008	2.23670634	0.03146596	377.690216	370.675293	1.15274024	7.02434635
11	-	299.257376	243.036145	40.409008	2.23710316	0.03146666	377.587813	370.675293	1.15274024	7.02434635
12	-	297.409396	244.189429	40.409008	2.23749999	0.03146736	377.413921	370.675293	1.15274024	7.02434635
13	-	299.060158	243.92957	40.409008	2.23789681	0.03146806	377.240028	370.675293	1.15274024	7.02434635

Variables de la VMO



- **Salida**