Escuela de Negocios

Maestría en Dirección de Empresas

TESIS

“Explotación de datos al servicio de la transformación digital”

AUTOR: Gonzalo Mallo

TUTOR: Pablo Roccatagliata

EMBA 2020

Diciembre 2021

Agradecimientos

Quiero agradecer en este trabajo al importante apoyo brindado durante todo este último tiempo a mi familia y especialmente a mi esposa por alentarme en mi desarrollo profesional.

A mis padres por toda la educación y amor que me brindaron ya que sin ellos no hubiera podido desarrollarme.

A mi empresa que me dio la posibilidad de continuar formándome y a mis compañeros del EMBA por el valioso aporte en el intercambio realizado tanto dentro como fuera del aula.

A mi amigo Estanislao Irigoyen, por enseñarme y embarcarme en el mundo de la minería de datos.

Por último, a mi tutor, Pablo Roccatagliata y a la universidad por estar siempre cerca de nuestra formación personal y profesional.

Resumen Ejecutivo

El objetivo del trabajo es explicitar cómo la explotación eficiente de datos contribuye al programa de transformación digital, cuáles son las etapas necesarias para hacerlo de manera sostenible y la gobernanza de estos para garantizar la eficacia a lo largo del tiempo en el marco de empresas operadoras de petróleo y gas que tienen como objetivo la maximización de su producción y minimización de su OPEX.

Primero definiremos las condiciones tecnológicas y procedurales conceptuales para luego desarrollar una aplicación real con el uso de una arquitectura técnica determinada.

Utilizaremos un caso de uso dentro del programa de transformación digital que ejemplifica todo el camino del dato digitalizado y testificará la metodología que usaremos para alcanzar un objetivo particular que será la selección de pozos a ser fracturados para maximizar la producción acumulada.

Como punto de partida para dar orden y tener un marco de referencia se adoptó el estándar de datos PPDM como modelo de datos. En paralelo se especificaron los sistemas fuentes de datos y luego se desarrollaron las interfaces de carga (ETL). Se realizó a posteriori un proceso de caracterización y "data cleansing". Se seleccionaron los datos de un yacimiento para ser usado como piloto.

La plataforma usada para el análisis y predicción de datos fue TIBCO *Spotfire Analyst* & Miner con la cual se logró un modelo estable que predice los pozos con potencial para maximizar su producción. Se destaca la importancia de que para alcanzar el objetivo se debe disponer de la mayor cantidad de datos e información y contar con la ayuda de un equipo interdisciplinario.

Expondremos una aplicación práctica donde se muestra la metodología propuesta y las herramientas usadas para alcanzar este objetivo.

Justificación– Delimitación

La presente investigación es un análisis descriptivo.

Palabras Claves

Analytics, Explotación de Datos, Transformación Digital, E&P, Forecast, Datos

Índice

Contenido

SECCIÓN A: Modelo Conceptual de explotación e Datos

A.1.0 Introducción

El crecimiento exponencial de los datos y la gran variedad de aplicativos de negocios enfrentan a los analistas de negocios al desafío de utilizar de manera eficaz y eficiente los datos producidos, en pos de optimizar el proceso de toma de decisiones basado en información precisa y conocimiento accionable que permita mejorar el retorno de inversión de los proyectos.

La industria del petróleo y gas no es una excepción a esta regla, contando con un gran volumen de datos, con una amplia variedad de aplicativos, procesos y necesidad de toma de decisiones que implican muchas veces la ejecución de presupuestos importantes.

Esta situación en su conjunto genera una complejidad en los procesos de decisión.

El objetivo del trabajo es definir un proceso y sus etapas necesarias para alcanzar una explotación y gobernanza de datos sostenible en la disciplina de E & P. De manera particular se trabajó sobre el objetivo puntual de la generación de sugerencias para la selección de pozos a ser fracturados que logren maximizar la producción acumulada de hidrocarburos a modo de ejemplo de aplicación.

De manera general el proceso definido en este trabajo consta de las siguientes etapas:

- Definición de un equipo de trabajo multidisciplinario.
- Determinación de las necesidades del negocio.
- Determinación de Arquitectura Tecnológica. Y estándar de almacenamiento.
- Definición de la metodología para la explotación de datos.

- Selección de las plataformas tecnológicas que soporten el proceso.
- Implementación de la interfaces de integración y procesamiento
- Desarrollo de modelos analíticos
- Integración del producto desarrollado en el proceso de toma de decisiones.

A.1.1 Estructura del Trabajo

El presente trabajo está estructurado en dos secciones principales de manera de abarcar todo el camino necesario para poder explotar los datos de manera eficiente en los programas de transformación digital.

En la primera sección A se describen cada uno de los hitos necesarios y posibles en forma general para llegar a soluciones implementables dependiendo el tipo de problema que se desee solucionar. De esta manera esta sección, permitirá tener una referencia de distintas opciones que pueden ser usadas como referencia para alcanzar un objetivo.

Comenzaremos describiendo de forma general la infraestructura tecnológica necesaria para ser utilizada en sus distintas modalidades para luego adentrarnos en inteligencia analítica donde describiremos las metodologías que existen y cada uno de los algoritmos que podemos utilizar para poder predecir un objetivo concreto.

La sección B, aplicaremos en forma práctica lo visto en la sección A de manera óptima y de forma particular para poder explotar los datos de manera de agregar valor al proceso de transformación digital.

A.2.0 Infraestructura Tecnológica

La infraestructura de tecnología de la información (TI) hace referencia a los elementos necesarios para operar y gestionar entornos de TI empresariales. La infraestructura de TI puede implementarse en un sistema de cloud computing o en las instalaciones de la empresa dependiendo las necesidades y sobre todo el retorno de la inversión asociado al proyecto que vamos a realizar.

Estos elementos incluyen el hardware, el software, los elementos de red, un sistema operativo (SO) y el almacenamiento de datos. Todos ellos se utilizan para ofrecer servicios y soluciones de TI.

Los productos de infraestructura de TI se pueden descargar como aplicaciones de software que se ejecutan en los recursos de TI actuales (por ejemplo, el almacenamiento definido por software) o como soluciones en línea que ofrecen los proveedores de servicios (por ejemplo, la infraestructura como servicio o IaaS).

A.2.1 Elementos de la infraestructura de TI

A.2.1.1 Hardware

El hardware incluye los servidores, los centros de datos, las computadoras personales, los enrutadores, los conmutadores y otros equipos.

Las instalaciones que almacenan los servidores; enfrían y proporcionan energía a un centro de datos también podrían considerarse parte de la infraestructura.

A.2.1.2 Software

El software hace referencia a las aplicaciones que utiliza la empresa, como los servidores web, los sistemas de gestión de contenido y el sistema operativo (por ejemplo, Linux®). El sistema operativo se encarga de gestionar el hardware y los recursos del sistema y establece las conexiones entre el software y los recursos físicos que ejecutan las tareas.

A.2.1.3 Redes

Los elementos de red interconectados permiten la comunicación, la gestión y las operaciones de red entre los sistemas internos y externos. La red consta de conexión a Internet, habilitación de la red, firewalls y seguridad, al igual que de elementos de hardware, como enrutadores, conmutadores y cables.

A.2.2 Tipos de infraestructura de TI

A.2.2.1 Infraestructura tradicional

En la infraestructura tradicional, las empresas son las propietarias de todos los elementos (como los centros de datos, los sistemas de almacenamiento de datos, entre otros), a los cuales gestionan en sus propias instalaciones. El funcionamiento de esta infraestructura suele considerarse costoso y requiere grandes cantidades de sistemas de hardware (por ejemplo, servidores), así como energía eléctrica y espacio físico.

A.2.2.2 Infraestructura de nube

La infraestructura de nube hace referencia a los elementos y los recursos que se necesitan para el cloud computing. Puede diseñar una nube privada usted mismo utilizando los recursos que se le destinan de forma exclusiva. O bien, puede usar una nube pública a través del alquiler de una

infraestructura de nube de un proveedor de nube, como Alibaba, Amazon, Google, IBM o Microsoft. También puede diseñar una nube híbrida mediante la incorporación de un cierto grado de gestión, organización y portabilidad de las cargas de trabajo en varias nubes.

A.2.2.3 Infraestructura hiper convergente

La infraestructura hiper convergente le permite gestionar los recursos informáticos, de red y de almacenamiento de datos desde una sola interfaz. Así podrá admitir cargas de trabajo más modernas con arquitecturas escalables en el hardware estándar del sector a través de la combinación del almacenamiento de datos y la informática definidos por software.[1]

A.2.3 Gestión de la infraestructura de TI

La gestión de la infraestructura de TI es la coordinación de todos los recursos, los sistemas, las plataformas, las personas y los entornos de TI.

Los tipos más comunes de gestión de la infraestructura tecnológica son los siguientes:

- **Gestión del sistema operativo:** supervisa los entornos que ejecutan el mismo sistema operativo, proporcionando gestión de suscripciones, implementaciones, parches y contenidos.
- **Gestión de la nube:** entrega a los administradores de nubes el control de todo lo que se ejecuta en ellas (los usuarios finales, los datos, las aplicaciones y los servicios), ya que gestiona la recuperación ante desastres, la integración, el uso y las implementaciones de recursos.

[1] 2020, RED HAT, https://www.redhat.com/es/topics/cloud-computing/what-is-it-infrastructure

- **Gestión de la virtualización:** interactúa con los entornos virtuales y el hardware físico subyacente para simplificar la administración de los recursos, mejorar el análisis de los datos y optimizar las operaciones.
- **Gestión de las operaciones de TI**: también se le conoce como gestión de procesos empresariales. Es la práctica con la que se modelan, analizan y optimizan los procesos de esta naturaleza que son continuos o predecibles, o también aquellos que suelen repetirse.
- **Automatización de la TI:** crea instrucciones y procesos repetibles para reemplazar o reducir la interacción humana con los sistemas de TI. También se la conoce como automatización de la infraestructura.
- **Organización de contenedores:** automatiza la implementación, la gestión, la escalabilidad y la conexión en red de los contenedores.
- **Gestión de la configuración:** mantiene los sistemas informáticos, los servidores y el software en un estado deseado y uniforme.
- **Gestión de las API:** distribuye, controla y analiza las interfaces de programación de aplicaciones (API) que conectan las aplicaciones y los datos en las empresas y las nubes.
- **Gestión de riesgos:** identifica y evalúa los riesgos y crea planes para disminuirlos o controlarlos, así como para reducir sus posibles efectos.

A.3.0 Inteligencia analítica

Esta especialidad es el resultado de la confluencia orgánica de diferentes disciplinas tales como la estadística, aprendizaje automático, visualización de información, optimización matemática y bases de datos.

Alternativamente se define como el proceso que tiene como fin “explotar” los datos almacenados en los sistemas de la empresa, “extraer” patrones estables allí ocultos, “expresarlos” en modelos legibles, operativos y accionables para luego “transformarlos” en conocimientos del negocio que mejoren la rentabilidad de este.[2]

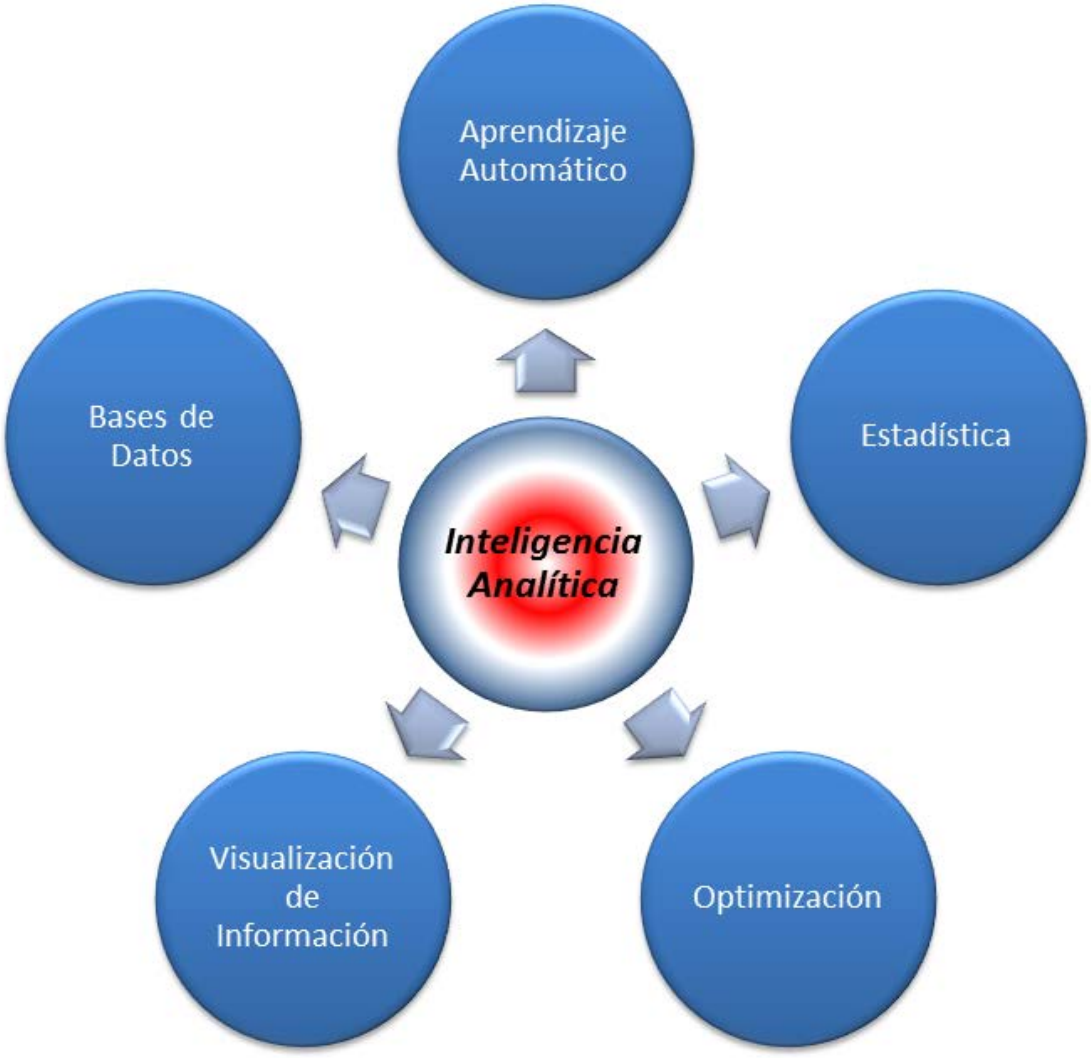


Fig. 1: Componentes de la Inteligencia Analítica

[2] 2014, Keith R. Holdaway. Harness Oil and Gas BIG DATA with Analytics.

Witten & Frank se centra directamente en la minería de datos y la define como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos. Es decir, la tarea fundamental de la minería de datos es encontrar modelos inteligibles a partir de los datos.

Para que este proceso sea efectivo debería ser automático o semiautomático (asistido) y el uso de los patrones descubiertos debería ayudar a tomar decisiones más seguras que reporten, por tanto, algún beneficio a la organización.

Por lo tanto, dos son los retos de la minería de datos: por un lado, trabajar con grandes volúmenes de datos, procedentes mayoritariamente de sistemas de información, con los problemas que ello conlleva (ruido, datos ausentes, intratabilidad, volatilidad de los datos…), y por el otro usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil. En muchos casos la utilidad del conocimiento minado está íntimamente relacionada con la comprensibilidad del modelo inferido. No debemos olvidar que, generalmente, el usuario final no tiene por qué ser un experto en las técnicas de minería de datos, ni tampoco puede perder mucho tiempo interpretando los resultados.

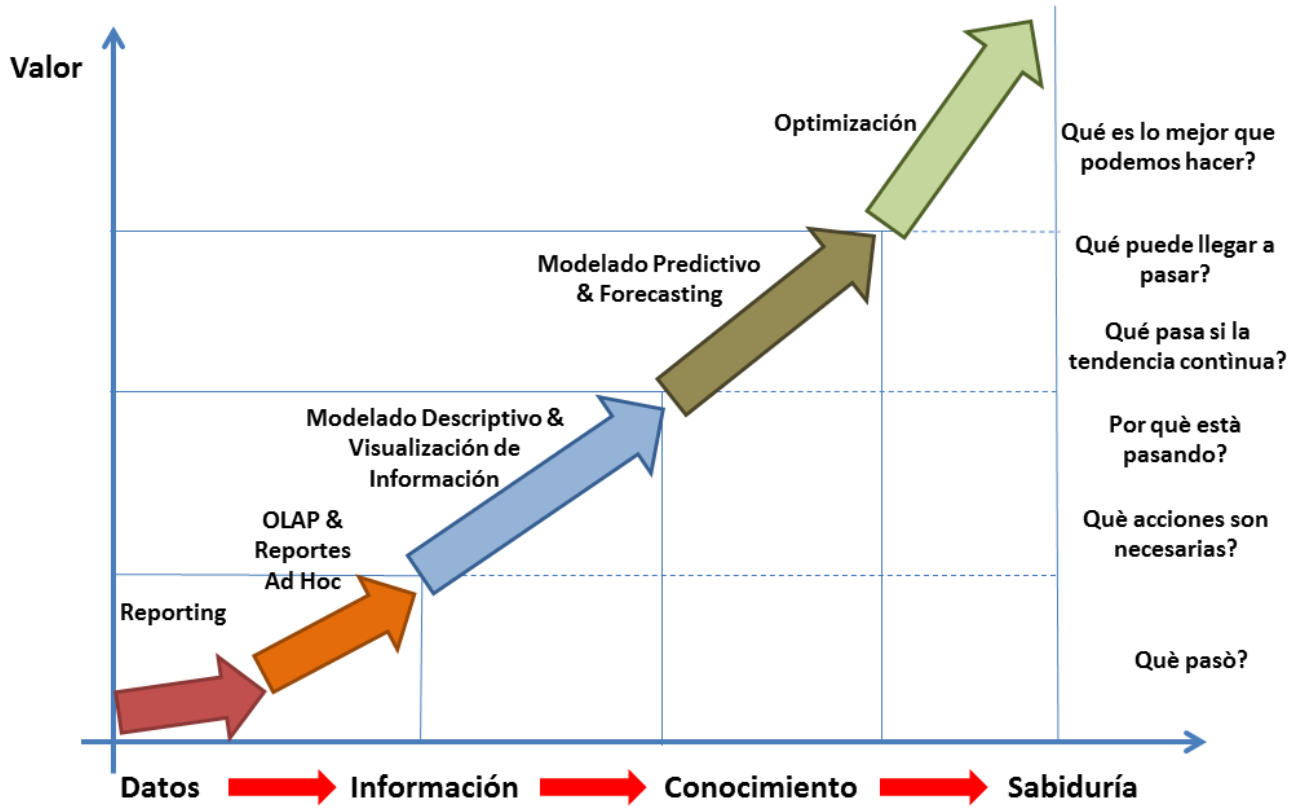


Fig. 2: Evolución del Valor Aportado por el análisis de datos.

A.3.1 Descubrimiento del Conocimiento

Fayyad define el Knowledge Discovery in Databases (KDD) como “el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y, en última instancia, comprensibles a partir de los datos”.[3]

[3] 1996, Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview", in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, *Advances in Knowledge Discovery and Data Mining,* AAAI Press / The MIT Press, Menlo Park, CA, pp.1-34

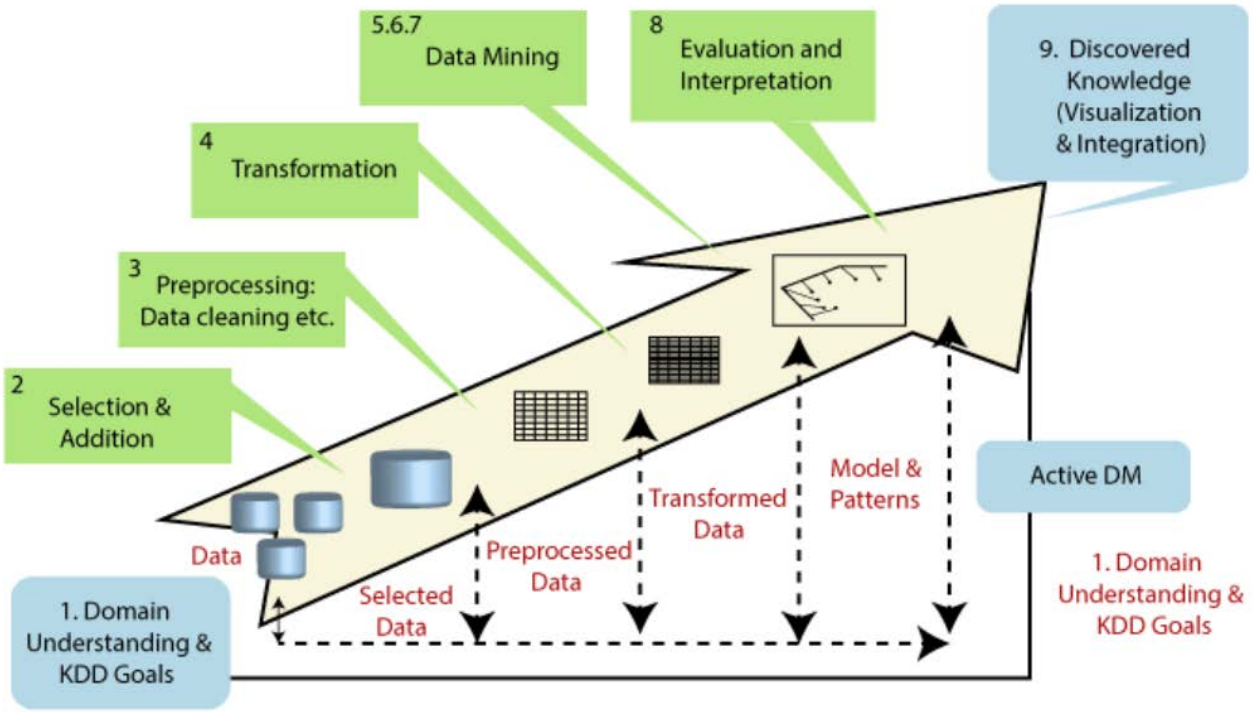


Fig. 3: KDD Model.

En esta definición se resumen cuáles deben ser las propiedades deseables del conocimiento extraído:

- **Válido:** hace referencia a que los patrones deben seguir siendo precisos para datos nuevos (con un cierto grado de certidumbre), y no sólo para aquellos que han sido usados en su obtención.
- **Novedoso:** que aporte algo desconocido tanto para el sistema y preferiblemente para el usuario.
- **Potencialmente útil:** la información debe conducir a acciones que reporten algún tipo de beneficio para el usuario.

- **Comprensible:** la extracción de patrones no comprensibles dificulta o imposibilita su interpretación, revisión, validación y uso en la toma de decisiones. De hecho, los datos crudos no proporcionan conocimiento (al menos desde el punto de vista de su utilidad).

Como se deduce de la anterior definición, el KDD es un proceso complejo que incluye no sólo la obtención de los modelos o patrones (el objetivo de la minería de datos), sino también la evaluación y posible interpretación de los mismos que se representa en la siguiente figura:

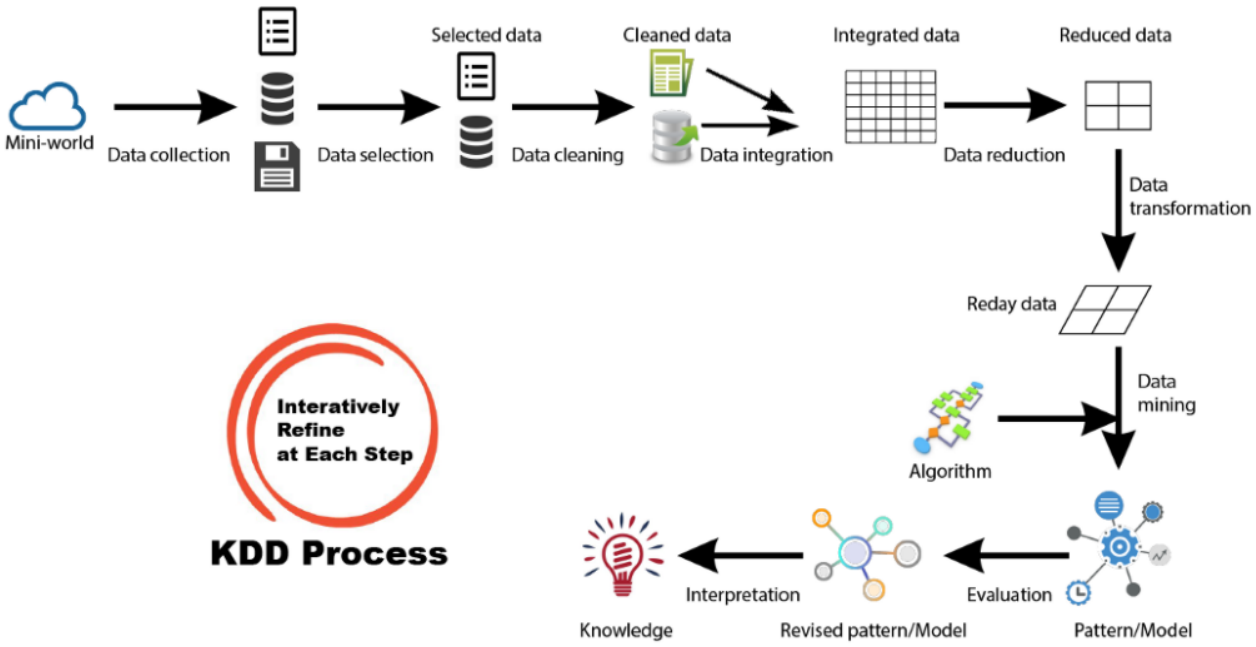


Fig. 4: Proceso de datos KDD.

A.3.2 Técnicas

Las técnicas empleadas en esta especialidad son comúnmente técnicas de predicción y descripción de datos, tales como regresiones lineales y logísticas, árboles de clasificación y regresión, redes neuronales, algoritmos bayesianos, técnicas de segmentación y/o de reducción de dimensionalidad, asociación, pronósticos, etc.

Estas técnicas son empleadas para resolver problemas enmarcables en las siguientes tareas:

- Segmentación.
- Clasificación.
- Estimación.
- Asociación y secuenciación.
- Pronóstico.
- Visualización.

A.4.0 Modelos Metodológicos

Las metodologías posibles para explotar en forma eficiente los datos son CRISP, SEMMA; DMAIC. Las mismas las describiremos a continuación.

A.4.1 CRISP-DM

Acrónimo de *Cross Industry Standard Process for Data Mining*. Se trata de un modelo de proceso de minería de datos que describe los enfoques comunes que utilizan los expertos en esta materia.[4]

La metodología se compone de las siguientes instancias a saber:

[4] 2000, Shearer C., *el modelo CRISP-DM: el nuevo plan para la minería de datos,* almacenamiento de los datos J; 5:13-22.

- **Comprensión del Negocio:** Entender el objetivo y problema a resolver.
- **Comprensión de los datos:** Conformar un equipo interdisciplinario que nos permita entender las variables que intervienen en la solución del problema.
- **Preparación de Datos:** El arte y la clave del éxito en el desarrollo de modelos analíticos.
- **Modelado:** Es una representación simplificada de la realidad que nos permite comprender algunos aspectos de un determinado problema. (Ej.: Ecuaciones, Reglas generales o Puntuales, Diagramas, etc.)
- **Evaluación:** La validación de que el modelo funciona
- **Implementación:** La puesta en producción – El punto donde se genera valor.
- **Monitoreo:** El proceso de control (¿El modelo, funciona aún?).

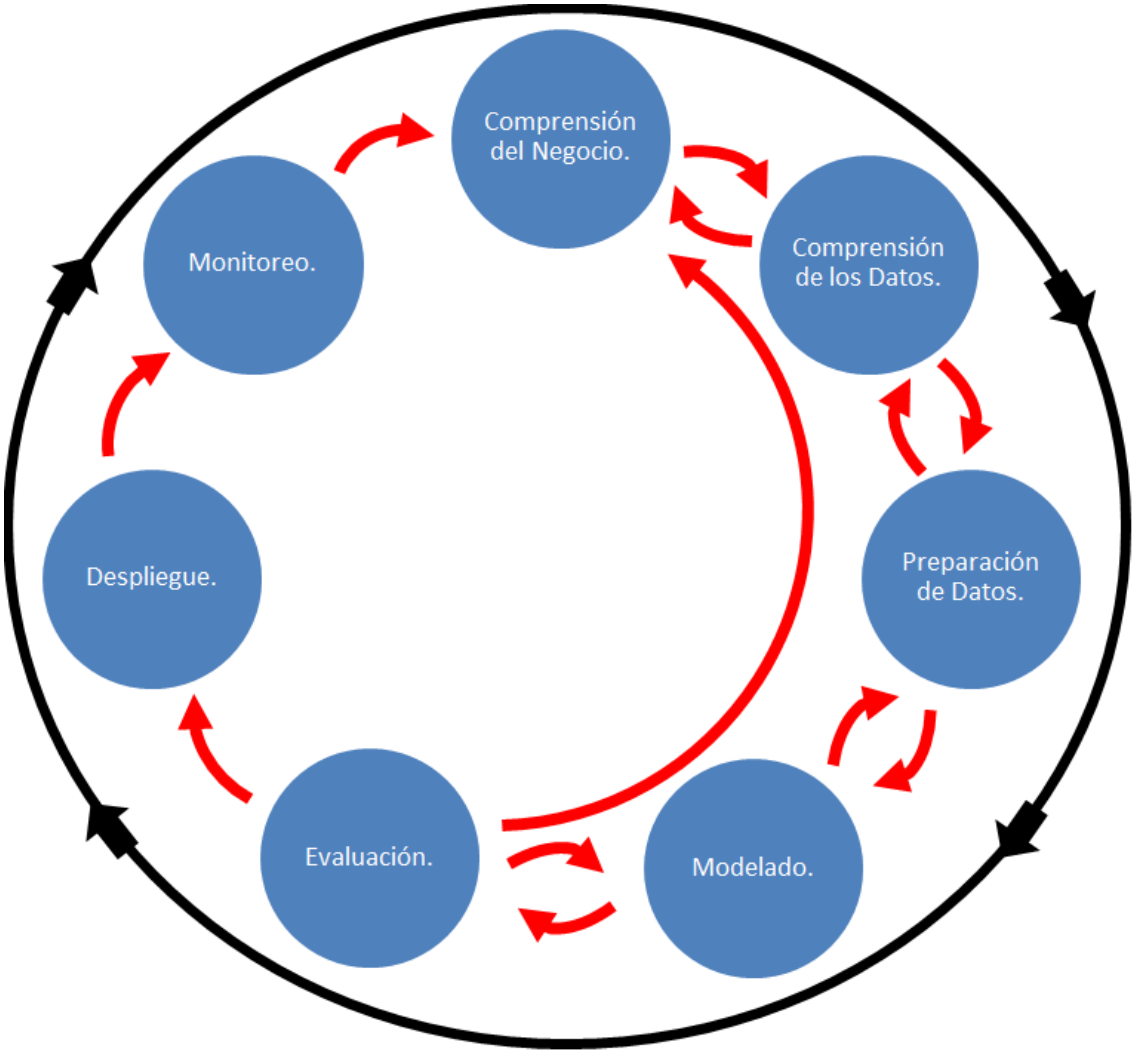


Fig. 5: Metodología CRISP-DM adaptada.

A.4.2 DMAIC

Esta metodología tradicional Six Sigma fue diseñada para resolver procesos problemáticos y/o controlar la oferta de producto y servicio. Está orientada a mejorar la productividad (que cantidad), el estado financiero (cuánto cuesta), la calidad (qué tan bien se realiza) y el tiempo (la rapidez) - PFQT. Originalmente los costos dominaban los aspectos financieros, pero últimamente el enfoque

de los proyectos se ha desplazado a los ingresos y el crecimiento. Es común que se utilice DMAIC como una metodología de mejora de procesos.

El enfoque DMAIC[5] Está diseñado para permitir flexibilidad y el trabajo iterativo, si es necesario. A medida que se aprende a través del proceso de 5 pasos, los supuestos o hipótesis iniciales sobre la causa raíz del problema puede perder vigencia, lo que requiere el equipo del proyecto para volver a ellos para modificarlos o explorar otras posibles alternativas. Por ejemplo, en la causa raíz de un problema de eficacia de una fuerza de ventas, la hipótesis de que se trata de un problema de capacitación en ventas en una región geográfica específica, puede llevar a sacar conclusiones apresuradas mediante la implementación de un nuevo programa de capacitación en ventas. En su lugar, el equipo Six-Sigma, decide sabiamente recolectar datos sobre el problema en primer lugar y después de un poco de investigación y análisis, descubre que la raíz apunta a un problema con la dirección de la gestión de ventas y no la falta de conocimientos y habilidades de los representantes de ventas.

DMAIC se basa en el control estadístico de procesos y en tres principios fundamentales:

- Centrada en los resultados, impulsados por datos, hechos y métricas.
- Estructura de Proyecto.
- Combinación de herramientas-tareas-entregables que varía según la etapa de la metodología.

[5] 1980, Creador Bill Smith, metodología Six Sigma, Motorola Inc.

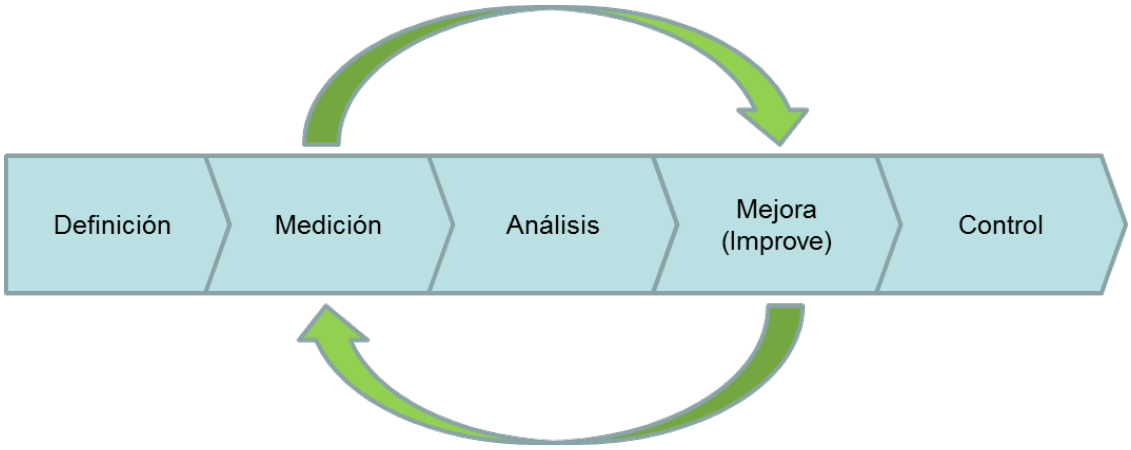


Fig. 6: Proceso DMAIC

Los proyectos DMAIC normalmente son de corta duración (de 3 a 9 meses)

Definición: Definir el problema y el alcance del esfuerzo de trabajo del equipo de proyecto. La descripción del problema debe incluir la sensación que tiene respecto del cliente y/o del negocio, así como el tiempo que el problema ha existido. Se necesitan identificar los clientes, los beneficios, metas del proyecto y los plazos de ejecución. Los diferentes tipos de problemas tienen alcance y escala diferentes, desde problemas de los empleados a problemas con el proceso de producción. Independientemente del tipo de problema, el mismo debe ser sistémico y recurrente de un proceso existente, detectado en más de un ciclo del proceso.

OBJETIVO: Que define el Cliente como problema

TAREA	ENTREGABLE	HERRAMIENTA
Identificar objetivos, problemas y oportunidades	Definición de Objetivos Definiciones de alcance Etapas del Proyecto	Project Documento de Alcance Mapa de relaciones
Mapear proceso	Mapa de Procesos de alto nivel (As-Is y To-Be) Mapa de Responsabilidades	Mapa de Procesos Matriz RACI
Relevar requerimientos y necesidades de clientes	Documento de Requerimientos de Negocio y parámetros críticos Minutas de reunión	Técnicas de manejo de Reuniones Gráficos de control de procesos Costos-Tiempos-Calidad
Elaborar plan de comunicación	Plan de proyecto comunicado y publicado	Template de plan de comunicaciones
Finalizar el Plan de proyecto	Plan de Proyecto detallado y aprobado	Project Matriz RACI Mapas de procesos Matriz DAFO

Medición: Medir el proceso actual o su rendimiento. Identificar los datos que están disponibles y cuáles son sus fuentes. Desarrollar un plan para recopilación, recopilarlos y resumirlos, realizando una descripción del problema. Esto implica generalmente la utilización de herramientas gráficas.

OBJETIVO: Cuales son las características del problema y cómo fue cambiando a lo largo del tiempo

TAREA	ENTREGABLE	HERRAMIENTA
Recolectar fuentes de datos, tiempos base, performance actual, controles, verificar que el proceso actual, eliminar causas especiales que no impactan en el problema.	Datos relevados	Mapa de relaciones Modelo de datos Gráficos de Control Muestras estadísticas Métodos gráficos
Desarrollar mapas de procesos	Mapa de procesos detallado Métricas	Mapa de Procesos Matriz RACI
Validar sistemas de medición y capacidades del proceso	Evaluación de capacidad y medición	Análisis de sistemas de medición (MSA) Análisis de capacidades de proceso
Revisar plan de proyecto, objetivos, necesidades y oportunidades	Plan revisado	Project Matriz RACI

Análisis: Analizar el rendimiento actual para aislar el problema. A través del análisis (tanto estadístico como cualitativo), empezar a formular y probar hipótesis acerca de la causa raíz del problema.

OBJETIVO: Cuales son las causas raíz

TAREA	ENTREGABLE	HERRAMIENTA
Validar gaps de requerimientos vs mediciones, establecer relaciones y cuantificar oportunidades de solución	Datos analizados	Mapas de relaciones Análisis crítico de Gaps Análisis estadístico Correlación y regresión
Desarrollar mapas de procesos detallados, establecer relaciones y cuantificar oportunidades	Procesos analizados	Mapa de Procesos Matriz RACI Mapas de relaciones Gráficos de Pareto Análisis de capacidades de procesos Mapa de parámetros críticos y puntos de control
Analizar conductas de las causas raíz, priorizarlas y cuantificar oportunidades	Causas Raíz analizadas	Técnicas de Brainstorming Diagramas de Causa-Efecto Diagramas de afinidad

		Test de Hipótesis sobre causas raíz Estadísticas de inferencia
Revisar plan de proyecto, objetivos, necesidades y oportunidades	Plan revisado	Project Matriz RACI

Mejora: Mejorar el problema mediante la selección de una solución. Basado en las causas identificadas en el paso anterior, atacar directamente la causa con una mejora. Las posibles soluciones son priorizadas a partir de los requerimientos del cliente, se realiza una selección y se la prueba para ver si resuelve el problema.

OBJETIVO: Que acciones de mejora corrigen las causas raíz de los requerimientos establecidos por el Cliente

TAREA	ENTREGABLE	HERRAMIENTA
Desarrollar mejoras potenciales o soluciones para causas raíz	Solución potencial generada	Técnicas de Brainstorming Desvíos positivos
Desarrollar criterios de evaluación, medir resultados, evaluar los	Solución potencial evaluada	Pilotos y pruebas Análisis de costos-beneficios

objetivos alcanzados y evaluar riesgos		Diseño de experimentación básicos Modo de falla y análisis de efectos (FMEA)
Recomendar e implementar soluciones y métricas	Solución recomendada	Matriz de selección Diagrama de fuerzas Análisis de sistemas de medición (MSA) Análisis de capacidades de proceso
Desarrollar mapas del proceso futuro detallados	Mapas de procesos futuros detallados Plan de Implementación y transición	Mapa de Procesos detallados Matriz RACI futura Manual de procedimientos Plan de implementación y de transición
Revisar plan de proyecto, objetivos, necesidades y oportunidades	Plan revisado	Project Matriz RACI

Control: Controlar el proceso mejorado y/o el rendimiento del producto para asegurar el cumplimiento de los objetivos del proyecto. Una vez resuelto el problema, la solución debe ser estandarizada y sustentable en el tiempo. Los procedimientos operativos pueden requerir

revisiones y un plan de control debe ser implementado. El equipo de proyecto transmite la solución sustentable, los procedimientos y los mecanismos de control a los actores del proceso y se cierra el proyecto.

OBJETIVO: Que controles deben ser implementados para hacer sustentable la solución

TAREA	ENTREGABLE	HERRAMIENTA
Documentar proceso de control y definir un plan de control	Plan de control definido	Diseño de plan de Control Gráficos de Control Análisis de Riesgos Plan de Comunicación Análisis de Partes Interesadas
Validar cumplimiento de métricas de implementación	Mejora implementada	Análisis de sistemas de medición (MSA) Análisis de capacidades de proceso Análisis de Costo-Beneficio
Capacitar	Capacitación realizada	Plan de Capacitación y Transición
Documentar recomendaciones y	Procesos documentados	Mapa de Procesos Matriz RACI Manual de procedimientos

principales puntos de la mejora		
Establecer un sistema de seguimiento	Sistemas de seguimiento desarrollados	Tablero de control Tablero de datos
Registrar lecciones aprendidas y revisar problemas ocurridos durante el proyecto.	Lecciones aprendidas Cierre de proyecto	Project Matriz RACI Matriz DAFO (Nueva)

A.4.3 SEMMA

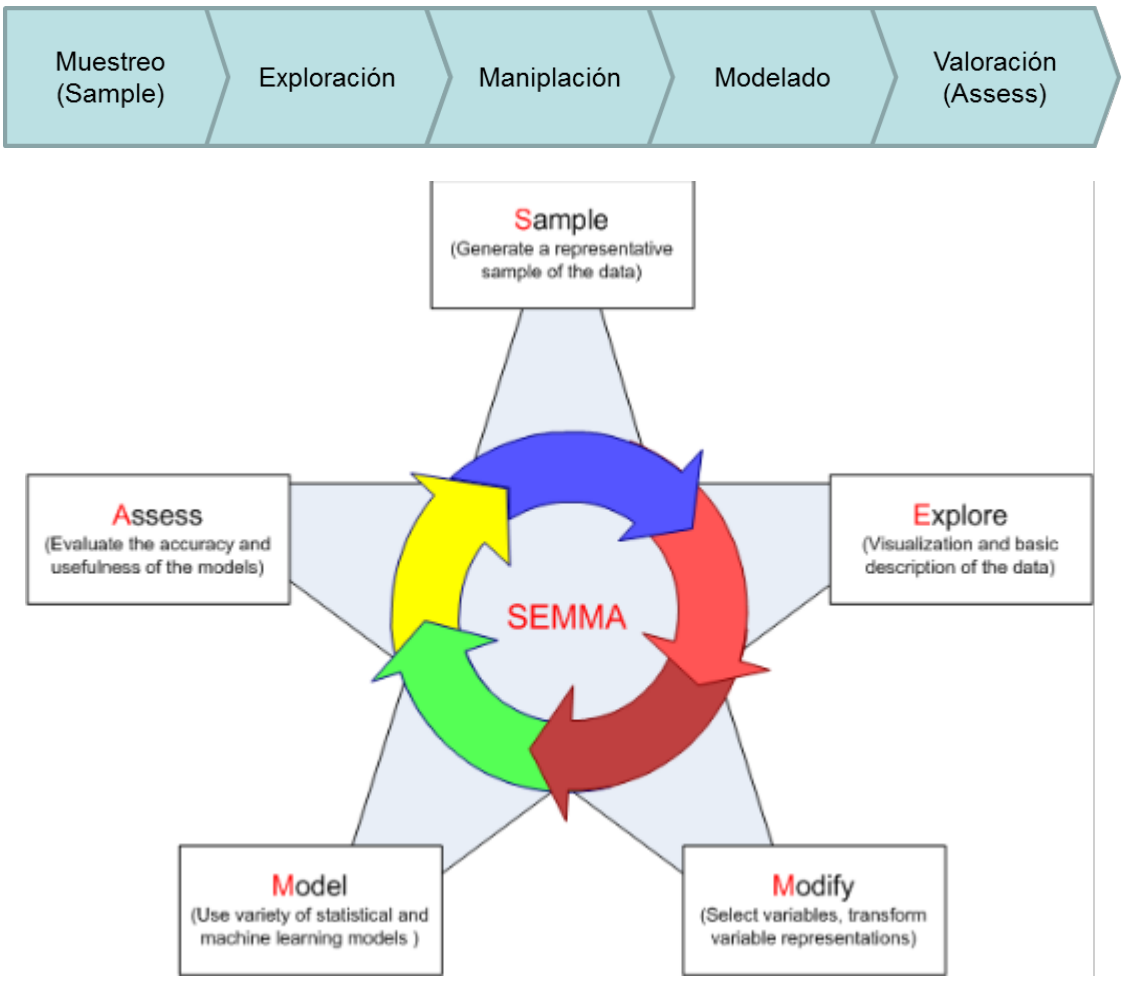


Fig. 7: Proceso SEMMA.

SEMMA[6] Es una metodología más corta y menos extensa que el CRISP-DM porque se centra más en el desarrollo del proceso de Minería de datos y no se orienta a objetivos empresariales.

Sample/Muestreo: Extracción de una muestra representativa. En esta primera fase de la metodología, se realiza la extracción de un conjunto de datos que sean una buena representación de la población a analizar, esto se hace con el objetivo de facilitar los procesos de minado sobre los datos, reduciendo los tiempos que se necesita para determinar la información valiosa para el negocio.

Explore/Exploración: Exploración de los datos en la muestra. En esta fase, se hace un recorrido a través de los datos extraídos en la muestra para detectar, identificar y eliminar datos anómalos, ayudando a refinar los procesos de descubrimiento de información en fases siguientes del proceso. En este punto del proceso, la exploración se puede realizar a través de medios visuales, aunque muchas veces no es suficiente este método, es por eso, que además de la visualización se pueden manejar diferentes técnicas estadísticas como análisis de factores, análisis de correspondencias, entre otros.

Modify/Manipulación: Modificación de los datos. Esta modificación de los datos se puede realizar creando, seleccionando y transformando las variables en las cuales se va a enfocar el proceso de selección del modelo. Muchas veces se tendrá la necesidad de realizar modificaciones cuando los datos que se están analizando cambien. Esto se debe a que el entorno en el que se trabaja la minería de datos es dinámico e iterativo.

Model/Modelado: Modelación de los datos. En esta fase, las herramientas de software se encargan de realizar una búsqueda completa de combinaciones de datos que juntos predecirán de una

[6] 2012, SAS Institute Inc autor

manera confiable los resultados buscados. Es en esta parte donde las técnicas y métodos de minería de datos entran a jugar un papel importante para la solución de los problemas que fueron identificados al iniciar el proyecto de minería de datos.

Assess/Valoración: Evaluación de los datos obtenidos. Después de que la fase de modelación presente los resultados obtenidos de la aplicación de los métodos de minería de datos al conjunto de datos. Se deberá realizar un análisis de los resultados para ver si estos fueron exitosos de acuerdo a las entradas que se tuvieron para analizar el problema. Una buena práctica para identificar si los

SEMMA	CRISP
Orientado al desarrollo del proceso de MD	Orientado a los objetivos empresariales
Se inicia analizando los datos	Se inicia analizando los objetivos del negocio
Ligada a productos SAS	Metodología abierta y gratuita
	Orientado a una metodología de **gestión de proyectos**

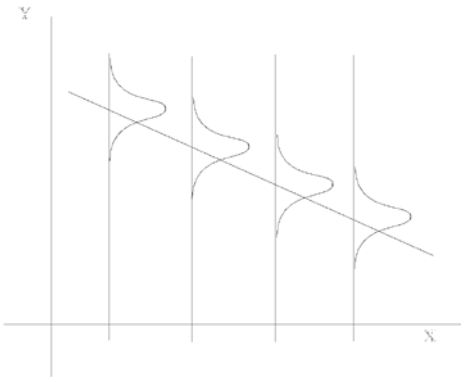
resultados con el modelo creado son los esperados, es aplicar este modelo a una porción de datos diferente. Si el modelo funciona correctamente para esta muestra y para la muestra utilizada para el proceso de creación del modelo, se tiene una buena probabilidad de tener un modelo válido.

A.5.0 Algoritmos

A.5.1 Regresión Lineal

La Regresión lineal se refiere a la predicción del valor de una variable a partir de una o más variables. En ocasiones se denomina a la variable dependiente (*y)* variable de respuesta y a la variable independiente (*x)* variable de predicción. En muchos problemas hay dos o más variables inherentemente relacionadas, y es necesario explorar la naturaleza de esta relación. El análisis de regresión puede emplearse por ejemplo para construir un modelo que exprese el rendimiento como una función de la temperatura. Este modelo puede utilizarse luego para predecir el rendimiento en un nivel determinado de temperatura. También puede emplearse con propósitos de optimización o control del proceso. Comenzaremos con el caso más sencillo, la predicción de una variable *(y)* a partir de otra variable *(x).*

Supuestos para inferencia[7] en el modelo de regresión lineal[1]



[7] Estos supuestos son para poder hacer inferencia una vez estimados los parámetros. Para poder estimar el modelo de regresión se necesita una muestra donde los predictores no estén perfectamente correlacionados.

1. Para cada valor de *x*, la variable aleatoria ε se distribuye normalmente.

2. Para cada valor de *x*, la media o valor esperado de ε es 0; esto es, $E(\varepsilon) = \mu_\varepsilon = 0$.

3. Para cada valor de *x,* la varianza de ε es la constante σ^2 (llamada varianza del error).

4. Los valores del término de error ε son independientes.

5. Para un valor fijo de *x*, la distribución muestral de *y* es normal, porque sus valores dependen de los de ε.

6. Para un valor fijo *x*, es posible predecir el valor de *y*.

7. Para un valor fijo *x*, es posible estimar el valor promedio de *y*

Ecuación Canónica de regresión Lineal

Modelo de regresión lineal

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Donde

y = variable dependiente

β_0 = ordenada al origen

β_1 = pendiente

x = variable independiente

ε = Error aleatorio

La expresión $\beta_0 + \beta_1 x$ se denomina componente **recta de regresión poblacional** del modelo de regresión lineal. La muestra de pares de datos se usará para estimar los parámetros $\beta_0 y \beta_1$ de la recta de regresión poblacional. En el ejemplo los diferentes rendimientos para un mismo tamaño de motor se atribuyen al término de error en el modelo de regresión.

Cálculo de la ecuación de regresión

También es llamada ecuación de predicción de mínimos cuadrados. La ecuación de regresión estimada es: $\hat{y} = b_0 + b_1 x.$

Donde:

$\hat{y} =$ Valor predicho de $\hat{y}$ para un valor particular de x.

b_0 = Estimador puntual de β_0 .(ordenada al origen)

b_1= Estimador puntual de β_1. (pendiente)

Para el cálculo de b_0 y b_1 se utilizamos las siguientes fórmulas:

$$SS_x = \sum x^2 - \frac{\left(\sum x\right)^2}{n}$$

$$SS_y = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

$$SS_{xy} = \sum xy - \frac{\left(\sum x\right)\left(\sum y\right)}{n}$$

$$b_1 = \frac{SS_{xy}}{SS_x}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

Donde:

SS = suma de cuadrados

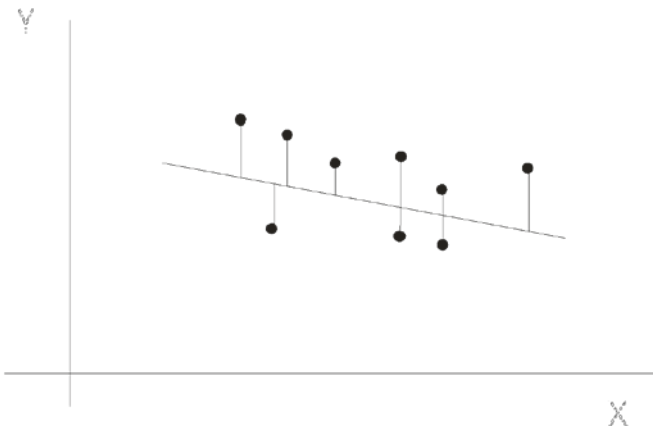
b_1 = pendiente

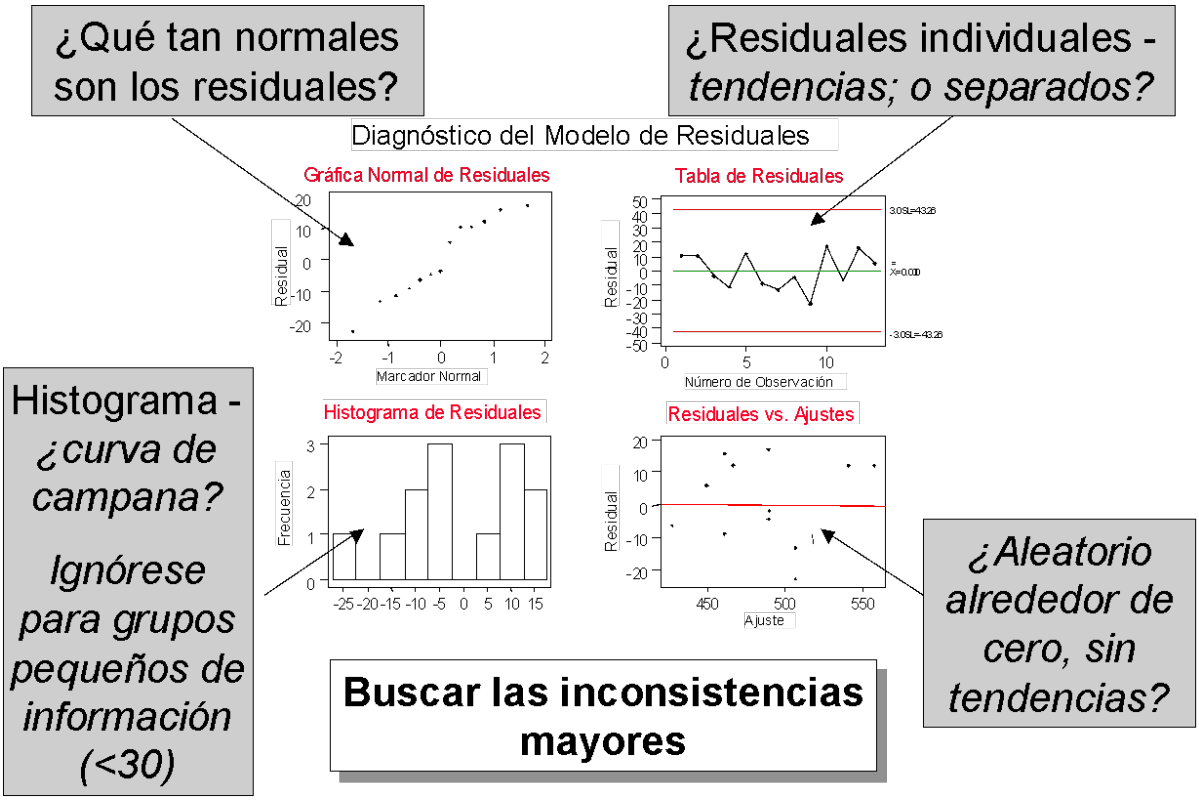
b_0 = ordenada al origen

n = número de pares de datos

Error

Los errores se denominan frecuentemente **residuales**. Podemos observar en la gráfica de regresión los errores indicados por segmentos verticales.





Al usar el criterio de mínimos cuadrados para obtener la recta que mejor se ajuste a nuestros datos, podemos obtener el valor mínimo para la suma de cuadrados del error (SSE)

$$SSE = SS_y - b_1 SS_{xy}$$

A la varianza de los errores *e* se le llama **varianza residual** siendo denotada por s_e^2, se encuentra dividiendo SSE entre n-2

$$S_e^2 = \frac{SSE}{n-2}$$

La raíz cuadrada positiva de la varianza residual se llama **error estándar de estimación** y se denota por S_e.

Análisis de correlación

Establece si existe una relación entre las variables y responde a la pregunta,"¿Qué tan evidente es esta relación?".

La correlación es una prueba fácil y rápida para eliminar factores que no influyen en la predicción, para una respuesta dada.

Coeficiente de Correlación de Pearson

- Es una medida de la fuerza de la relación lineal entre dos variables *x e y*.
- Es un número entre -1 y 1
- Un valor positivo indica que cuando una variable **aumenta,** la otra variable **aumenta**
- Un valor negativo indica que cuando una variable **aumenta,** la otra **disminuye**
- Si las dos variables no están relacionadas, el coeficiente de correlación se aproxima a 0.

El coeficiente de correlación *r* se calcula mediante la siguiente fórmula:

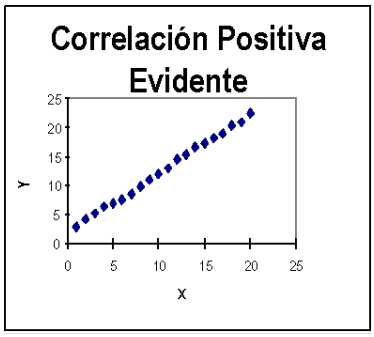
$$r = \frac{SSxy}{\sqrt{SSxSSy}}$$

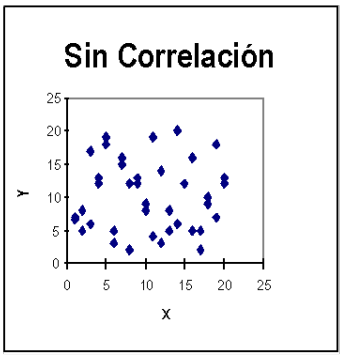
Tabla de Correlación

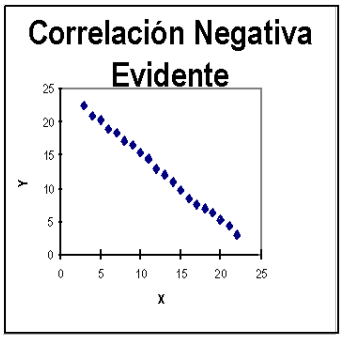
Por su importancia, ¿cuál es el coeficiente mínimo de correlación?

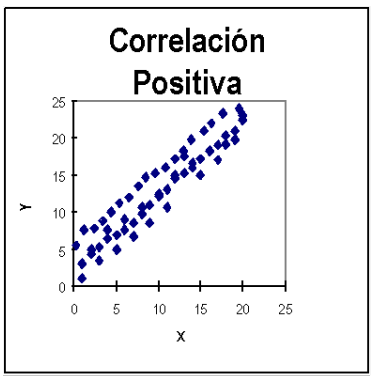
n	95% de confianza	99% de confianza	n	95% de confianza	99% de confianza
3	1.00	1.00	15	0.51	0.64
4	0.95	0.99	16	0.50	0.61
5	0.88	0.96	17	0.48	0.61
6	0.81	0.92	18	0.47	0.59
7	0.75	0.87	19	0.46	0.58
8	0.71	0.83	20	0.44	0.56
9	0.67	0.80	22	0.42	0.54
10	0.63	0.76	24	0.40	0.52
11	0.60	0.73	26	0.39	0.50
12	0.58	0.71	28	0.37	0.48
13	0.53	0.68	30	0.36	0.46
14	0.53	0.66			

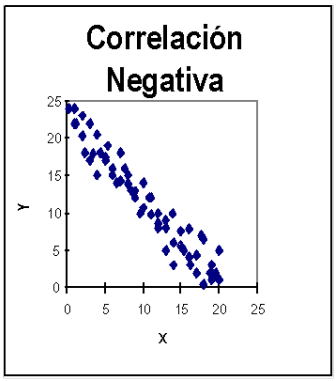
Para un 95% de confianza, con una muestra de 10, el coeficiente (r) debe ser al menos .63











A.5.2 Regresión Logística

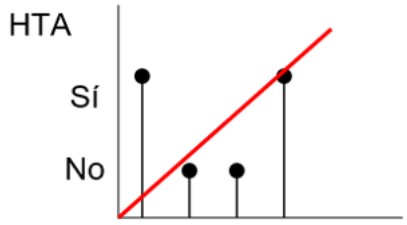
Un ejemplo de modelo no lineal (en los coeficientes) es el **modelo de regresión logística** que se llama así porque la función f(x) que la define es una curva logística. Constituye un planteamiento especial que busca un modelo o ecuación capaz de predecir el valor que tomará una variable dependiente (y) en función de los valores que presenten diversas variables independientes ($x_1..x_p$); pero ahora con tres importantes características y que tienen una traducción práctica muy frecuente y útil en situaciones de investigación biológica.:

1. La variable dependiente es cualitativa, generalmente dicotómica (0=no, 1=si)
2. Las independientes pueden ser cuantitativas o cualitativas, en el segundo caso recodificadas como binarias si fuera necesario.
3. La relación que se busca no es una ecuación lineal (pocos procesos en medicina guardan este tipo de relación), sino exponencial de tipo sigmoideo.

Modelo de regresión múltiple no lineal

↓

Modelo de regresión logística multivariante



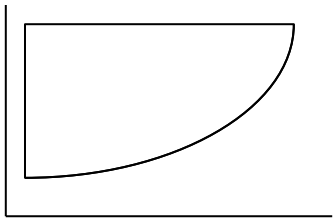
Como vemos en la figura, Parece claro que de momento no puedo establecer una ecuación del tipo y = a.x (línea roja), X [col], por ello lo que hacemos con vars dicotómicas es ver la probabilidad de que se dé una circunstancia (HTA sí/no).

Asignando probabilidades (valor numérico) podré conseguir un modelo de probabilidad y obtendremos una función logística que permitirá clasificar a los individuos en uno de los dos grupos.

CARACTERÍSTICAS DE ESTE MODELO

1. Regresión: porque tiene variables dependientes e independientes
2. Múltiple: hay más de una variable independiente.
3. No lineal: es una ecuación exponencial.

Una ecuación exponencial indica crecimiento. Y=a exp (bx,) crecimiento infinito.



Pero existen planteamientos en biología en los que varía; ej.: Bioensayos, cuando analizamos la dosis-respuesta de un fármaco, llega un momento en que la respuesta no sube más, aunque se aumente la dosis, porque los receptores están saturados (es la máxima probabilidad); es un modelo exponencial pero la ecuación es algo diferente y se llama modelo o **ecuación logística**. En este

ejemplo del fármaco-dosis sería un ejemplo de regresión UNIVARIANTE (hay una variable independiente, la dosis) que nos sirve de ayuda para entender el tema, aunque la clase sea de modelo multivariante.

ECUACIÓN

(ejemplo de **bivariante**, para que sea más fácil entenderlo)

$$P(Enf/A) = \frac{1}{1 + e^{-(b_0 + b_1 A)}}$$

P= variable dependiente (respuesta al tto)

X= variable independiente (dosis)

Donde b_0 = constante y b_1 multiplica a la variable A (o a x si la representásemos como $b_0 + ax$).

Las que mejor funcionan son las variables dicotómicas, pero si no lo son, pueden valer las ordinales. Las cuantitativas funcionan peor.

Una vez sé b_0 y b_1, el modelo me sirve para saber una respuesta a una dosis determinada sin necesidad de medir por ejemplo las concentraciones plasmáticas a todos los pacientes (de ahí la utilidad de crear ecuaciones, gracias a una variable independiente fácil de medir obtenemos otra dependiente que es complicada de medir o haría falta métodos cruentos y/o costoso para ello). El cálculo de b_0 y b_1 es complicado y dijo que no venía a la clase su explicación.

En **multivariante**: P ($E/A_1, A_2, A_3$) sería como el de bivariante pero con b_1, b_2 y b_3 y para interpretarlo intentaremos usar el OR en lugar de b_1, b_2 y b_3.

5.3 Árbol de decisión

Los árboles de decisión son uno de los modelos de minería de datos más intuitivos, debido a que muestran una estructura “si-entonces-otro”, fácil de comprender. Los árboles de decisión se pueden utilizar tanto para clasificación como para análisis de regresión Los árboles de decisión separan un conjunto de datos en distintos subconjuntos de datos. En lugar de utilizar un enfoque de aprendizaje sin guía (como en el análisis clúster), el árbol de decisión utiliza un algoritmo guiado, de forma que los subconjuntos de datos creados comparten una característica de destino determinada, que proporciona una variable dependiente. Las demás características las proporcionan las variables independientes, que se utilizan para dividir los conjuntos de datos originales en subconjuntos de datos. Normalmente, se utiliza la variable independiente con mayor poder previsible y así sucesivamente. Muchas herramientas implementan el algoritmo de árbol de clasificación y regresión (CART).

Los árboles de clasificación también son llamados de decisión o de identificación, constituyen una aproximación radicalmente distinta a todas las estudiadas hasta el momento.. Como forma de representación del conocimiento, los árboles de clasificación destacan por su sencillez. A pesar de que carecen de la expresividad de las redes semánticas o de la lógica de primer orden, su dominio de aplicación no está restringido a un ámbito concreto, sino que pueden utilizarse en diversas áreas: diagnóstico médico, juegos, predicción meteorológica, control de calidad, etc.

Un árbol de clasificación es una forma de representar el conocimiento obtenido en el proceso de aprendizaje inductivo. Puede verse como la estructura resultante de la partición recursiva del espacio de representación a partir del conjunto (numeroso) de prototipos. Esta partición recursiva

se traduce en una *organización jerárquica* del espacio de representación que puede modelarse mediante una estructura de tipo árbol. Cada *nodo interior* contiene una pregunta sobre un atributo concreto (con un hijo por cada posible respuesta) y cada *nodo hoja* se refiere a una decisión (clasificación).

La clasificación de patrones se realiza en base a una serie de preguntas sobre los valores de sus atributos, empezando por el nodo raíz y siguiendo el camino determinado por las respuestas a las preguntas de los nodos internos, hasta llegar a un nodo hoja. La etiqueta asignada a esta hoja es la que se asignará al patrón a clasificar.

Cada nodo incluye la información siguiente:

- **Puntuación (node´s score)**: el resultado más común (o dominante) de los registros de datos para el nodo.
- **Predicado (predicate)**: una sentencia lógica que se usa para separar registros de datos del origen de un noto. Los registros pertenecen a un nodo si se evalúa como verdadero (pueden incluir una o varias sentencias lógicas combinadas con los operadores AND, OR, XOR, etc).
- **Gráfico de ojo (eye chart)**: representación gráfica de la distribución de puntuaciones.
- **Distribución de puntuaciones (score distribution)**: tabla que muestra el desglose de registros de datos de formación asociados con el nodo. Sirve de leyenda para el gráfico de ojo. La distribución de puntuaciones contiene el conteo real de los registros de datos de formación del nodo que se representa mediante cada clase de destino. La proporción de

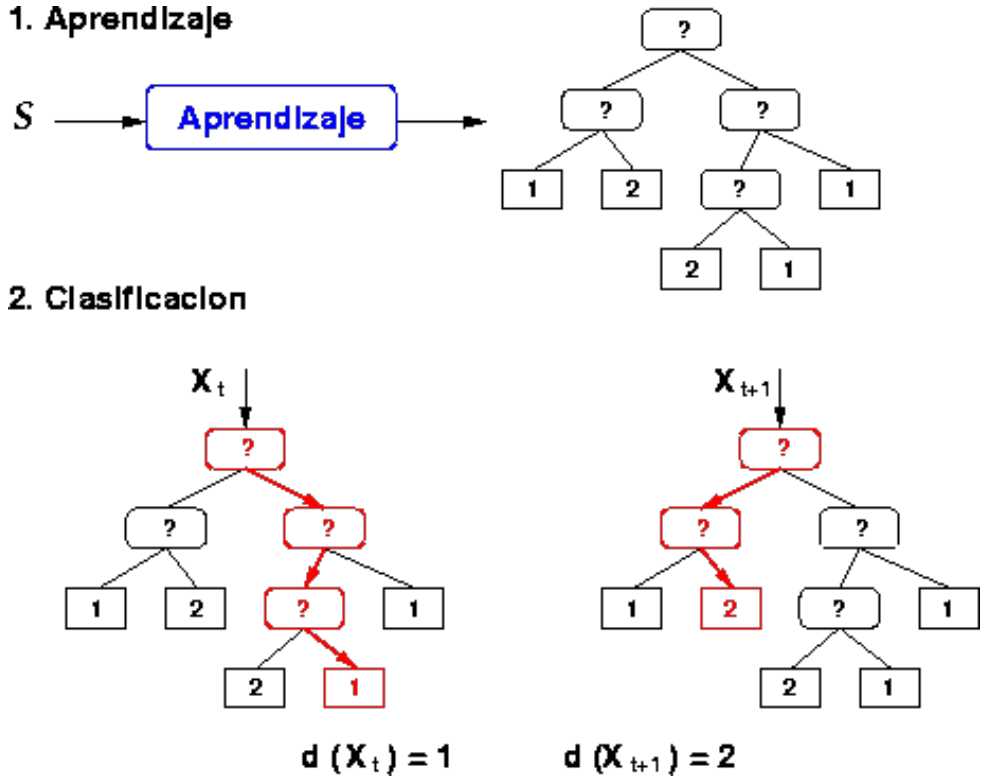
cada clase de registros de datos se muestra como porcentaje de los conteos totales de este nodo y como porcentaje de la población total.

- **Resumen de nodo (node summary)**: porcentaje de todos los registros de datos de formación asociados con el nodo.
- **ID de nodo (node id)**: identificador único o referencia para cada nodo. Se usa un formato de profundidad en nivel (el id es una serie de números).

Desafortunadamente, los árboles generalmente no tienen el mismo nivel de performance predictiva en comparación con otros algoritmos. Además, los árboles pueden ser muy poco robustos. En otras palabras, un pequeño cambio en los datos puede causar un gran cambio en la estimación final del árbol.

Sin embargo, al agregar muchos árboles de decisión, utilizando métodos como bagging, random forest y boosting el rendimiento predictivo de los árboles puede ser sustancialmente mejorado.

Entre los clasificadores basados en árboles descritos en la literatura (ID3, C4, C4.5, árboles Bayesianos, etc.) estudiaremos **CART** (acrónimo de **C**lassification **A**nd **R**egression **T**rees o árboles de clasificación y regresión), propuesto por Breiman y otros en [B.1]. Las diferencias principales entre los distintos algoritmos de construcción de árboles de decisión radican en las estrategias de poda y en la regla adoptada para particionar nodos. Así, CART se caracteriza, fundamentalmente, por realizar particiones binarias y por utilizar una estrategia de poda basada en el criterio de coste-complejidad. Entre las dos aplicaciones de CART (clasificación y regresión) nos centraremos exclusivamente en la primera. La metodología para seguir puede resumirse en dos pasos, y se esquematiza en la siguiente figura.



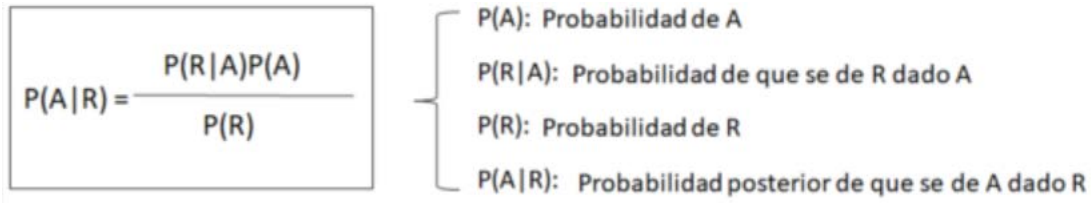
Aprendizaje y clasificación con un árbol de decisión

1. **Aprendizaje**. Consiste en la *construcción del árbol* a partir de un conjunto de prototipos, *S*. Constituye la fase más compleja y la que determina el resultado final. A esta fase dedicamos la mayor parte de nuestra atención.

2. **Clasificación**. Consiste en el etiquetado de un patrón, *X*, independiente del conjunto de aprendizaje. Se trata de responder a las preguntas asociadas a los nodos interiores utilizando los valores de los atributos del patrón *X*. Este proceso se repite desde el nodo raíz hasta alcanzar una hoja, siguiendo el camino impuesto por el resultado de cada evaluación.

A.5.4 Naive Bayes

Los modelos de Naive Bayes son una clase especial de algoritmos de clasificación de Machine Learning. Se basan en una técnica de clasificación estadística llamada “teorema de Bayes”. Estos modelos son llamados algoritmos “Naive”, o “Inocentes” en español. En ellos se asume que las variables predictoras son independientes entre sí dentro de cada clase de la variable de target. En otras palabras, que la presencia de una cierta característica en un conjunto de datos no está en absoluto relacionada con la presencia de cualquier otra característica. Proporcionan una manera fácil de construir modelos con un comportamiento muy bueno debido a su simplicidad. Lo consiguen proporcionando una forma de calcular la probabilidad ‘posterior’ de que ocurra un cierto evento A, dadas algunas probabilidades de eventos ‘anteriores’.



VENTAJAS

- Una manera fácil y rápida de predecir clases, para problemas de clasificación binarios y multiclase.
- En los casos en que sea apropiada una presunción de independencia, el algoritmo se comporta mejor que otros modelos de clasificación, incluso con menos datos de entrenamiento.
- El desacoplamiento de las distribuciones de características condicionales de clase significa que cada distribución puede ser estimada independientemente como si tuviera una sola

dimensión. Esto ayuda con problemas derivados de la dimensionalidad y mejora el rendimiento.

DESVENTAJAS

- Aunque son unos clasificadores bastante buenos, los algoritmos Naive Bayes son conocidos por ser pobres estimadores. Por ello, no se deben tomar muy en serio las probabilidades que se obtienen.

- La presunción de independencia Naive muy probablemente no refleja cómo son los datos en el mundo real.

- Cuando el conjunto de datos de prueba tiene una característica que no ha sido observada en el conjunto de entrenamiento, el modelo le asignará una probabilidad de cero y será inútil realizar predicciones. Uno de los principales métodos para evitar esto, es la técnica de suavizado, siendo la estimación de Laplace una de las más populares.

A.5.5 Arquitecturas basadas en redes Neuronales

Las Redes Neuronales constituyen la base de una familia muy variada de arquitecturas. El sistema se compone de un número elevado de unidades muy simples (neuronas) altamente interconectadas. Durante el proceso de entrenamiento se modifican los “pesos” de las conexiones entre las unidades al igual que la arquitectura de la red.. Básicamente el conocimiento es el resultado buscado en estos sistemas que se obtiene por un complejo proceso de aprendizaje que

luego es almacenado en las sinapsis interneuronales caracterizando patrones que luego pueden ser reconocidos analizando dichas sinapsis.

A.5.5.1 Estructura de red neuronal

Podemos entender una red neuronal compuesta por lo que se denomina la entrada formada por una red de perceptrones multicapa que actúan como predictores de las variables de salida.

Este tipo de funcionamiento se lo conoce como feedforward ya que las conexiones de la red fluyen unidimensionalmente desde la entrada hasta la capa salida sin que exista una retroalimentación de la información.

También existe el funcionamiento retroalimentado o feedback para aquellos sistemas donde las respuestas de las capas N vuelven a alimentar las capas N-1-

Podemos resumir la estructura y las características de las redes neuronales de la siguiente manera:

- Capa de entrada**:** Es la que posee los predictores o variables de entrada.
- Capa oculta: es la capa que contiene los nodos que no podemos ver ni analizar y es función de los predictores.
- Capa de salida: es la capa que tiene las variables respuesta o de salida.
- Reglas de aprendizaje
- Topología de interconexión

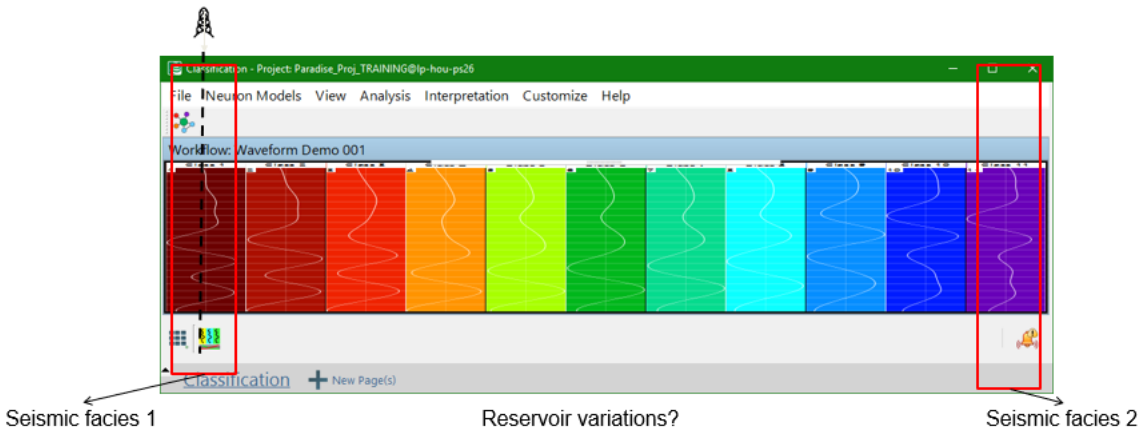
Cuando hablamos de redes multicapa son aquellas que permiten una o más capas ocultas por lo que la segunda capa es función de la primera oculta y las n capas función de la n-1 anterior. Por lo dicho cada respuesta será función de la última capa oculta.

Entrenamiento de la Red Neuronal:

- Supervisado: es aquel en el que se controla como la red resuelve el modelo y se le añade información para que coteje los resultados que queremos alcanzar.
- No supervisado: la red neuronal no se alimenta con información de los resultados que queremos alcanzar por lo que no tiene contra que cotejar sus resultados.

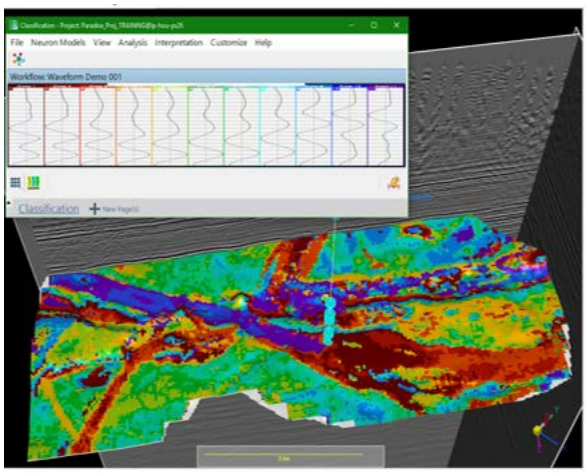
A.5.5.2 Aplicaciones en la industria de hidrocarburos:

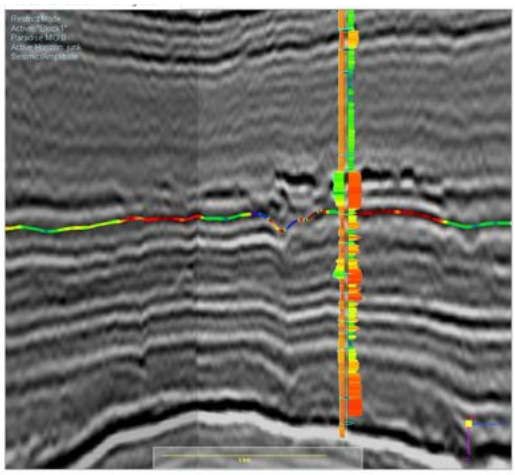
Básicamente las redes neuronales se utilizan para predecir series temporales y sobre todo para reconocimiento de patrones tanto de imágenes como de voz. En el caso de la industria del oil & gas son muy usadas para la búsqueda de patrones de imágenes sísmicas. Las mismas pueden ser caracterizadas por tipo de ondícula como se vé en la imagen adjunta y se alimenta a la red con ondículas exitosas por pertenecer al tipo donde hay gran cantidad de gas o petróleo conformando una red neuronal supervisada.



Para este tipo de aplicaciones predictivas del tipo de roca reservorio que se quiere hallar se utiliza una red neuronal basada en un algoritmo desarrollado por la empresa Paradigm que es asociativo neuronal. Este algoritmo primero tiene que clasificar las ondas sísmicas y compararlas con modelos

sísmicos de campos exitosos. La siguiente figura nos muestra una clasificación de cada ondícula y como las mismas se aparecen dependiendo del tipo de formación a la que pertenecen y determinando el tipo de roca.





A.5.6 Genéticos

Los Algoritmos Genéticos (AGs) son heurísticas que pueden usarse para resolver problemas de búsqueda y optimización. En nuestro problema los algoritmos genéticos pueden usarse para la optimización de hiper parámetros de los modelos. Estos permiten evitar el sobreajuste al no considerar todas las configuraciones posibles de hiper parámetros al tiempo que permiten obtener soluciones en menor tiempo computacional comparadas con esquemas de optimización como grid search.

Están basados en el proceso genético de los organismos vivos. A lo largo de las generaciones, las poblaciones evolucionan en la naturaleza de acorde con los principios de la selección natural y la supervivencia de los más fuertes, postulados por Darwin (1859). Por imitación de este proceso, los Algoritmos Genéticos son capaces de ir creando soluciones para problemas del mundo real. La

evolución de dichas soluciones hacia valores óptimos del problema depende en buena medida de una adecuada codificación de las mismas.

Los principios básicos de los Algoritmos Genéticos fueron establecidos por Holland (1975), y se encuentran bien descritos en varios textos – Goldberg (1989), Davis (1991), Michalewicz (1992), Reeves (1993).

En la naturaleza los individuos de una población compiten entre sí en la búsqueda de recursos tales como comida, agua y refugio. Incluso los miembros de una misma especie compiten a menudo en la búsqueda de un compañero. Aquellos individuos que tienen más éxito en sobrevivir y en atraer compañeros tienen mayor probabilidad de generar un gran número de descendientes. Por el contrario, individuos poco dotados producirán un menor número de descendientes. Esto significa que los genes de los individuos mejor adaptados se propagarán en sucesivas generaciones hacia un número de individuos creciente. La combinación de buenas características provenientes de diferentes ancestros puede a veces producir descendientes "super individuos", cuya adaptación es mucho mayor que la de cualquiera de sus ancestros. De esta manera, las especies evolucionan logrando unas características cada vez mejor adaptadas al entorno en el que viven.

Los Algoritmos Genéticos usan una analogía directa con el comportamiento natural. Trabajan con una población de individuos, cada uno de los cuales representa una solución factible a un problema dado. A cada individuo se le asigna un valor o puntuación, relacionado con la bondad de dicha solución. En la naturaleza esto equivaldría al grado de efectividad de un organismo para competir por unos determinados recursos. Cuanto mayor sea la adaptación de un individuo al problema, mayor será la probabilidad de que el mismo sea seleccionado para reproducirse, cruzando su

material genético con otro individuo seleccionado de igual forma. Este cruce producirá nuevos individuos – descendientes de los anteriores – los cuales comparten algunas de las características de sus padres. Cuanto menor sea la adaptación de un individuo, menor será la probabilidad de que dicho individuo sea seleccionado para la reproducción, y por tanto de que su material genético se propague en sucesivas generaciones.

De esta manera se produce una nueva población de posibles soluciones, la cual reemplaza a la anterior y verifica la interesante propiedad de que contiene una mayor proporción de buenas características en comparación con la población anterior. Así a lo largo de las generaciones las buenas características se propagan a través de la población. Favoreciendo el cruce de los individuos mejor adaptados, van siendo exploradas las áreas más prometedoras del espacio de búsqueda. Si el Algoritmo Genético ha sido bien diseñado, la población convergerá hacia una solución óptima del problema.

Si bien no se garantiza que el Algoritmo Genético encuentre la solución óptima del problema, existe evidencia empírica de que se encuentran soluciones de un nivel aceptable, en un tiempo competitivo con el resto de los algoritmos de optimización combinatoria. En el caso de que existan técnicas especializadas para resolver un determinado problema, lo más probable es que superen al Algoritmo Genético, tanto en rapidez como en eficacia. El gran campo de aplicación de los Algoritmos Genéticos se relaciona con aquellos problemas para los cuales no existen técnicas especializadas.

SECCIÓN B- Desarrollo Aplicado de Modelo- MVP

B.1.0 Capital Humano

B.1.1 Liderazgo

Antes de embarcarnos en el desarrollo del modelo se debe lograr el apoya e las máximas autoridades de la organización porque de otra manera el modelo que realicemos perecerá con el tiempo. De esta manera se acordó con los máximos responsables del negocio los objetivos que abarcaría el modelo para asignar partida presupuestaria y evaluar el caso de negocio.

- STAKEHOLDER: Socios y Directores de Negocio.
- SPONSOR: Director Ejecutivo
- DIRECTOR DE PROYECTO : Líder Transformación Digital

B 1.2 Especialistas

Conformamos un equipo multidisciplinario integrado por personas con distintos perfiles, entre ellos, ingenieros de reservorios y producción, gestores de datos, consultores funcionales, ingenieros en sistemas y especialistas en inteligencia analítica.

El mismo quedó compuesto de la siguiente manera:

LÍDERES: Gerente de Reservas y Reservorios / Business Partners de TI para E&P

EQUIPO NEGOCIO:

- Especialista de Ingeniería y procesos
- Especialista de geofísica (Sísmica)

- Especialista geología SCh
- Referente de Ingeniería

EQUIPO IT:

- Especialista de Analytics
- Especialista en integración & desarrollo
- Analista Funcional y de soporte

EQUIPO CHANGE MANAGEMENT & PROCESS

- Especialista en gestión de Cambios
- Analista de procesos (modelado de roles y responsabilidades)

EQUIPO EXTERNOS

- Schlumberger / Paradigm: Especialistas en procesamiento. modelado de imágenes y machine learning
- TIBCO: Especialistas en Analytics / Machine Learning

B.2.0 Objetivo de Negocio

El punto de partida de este estudio es la determinación de las necesidades del negocio. Por tal motivo el equipo de trabajo se abocó a la búsqueda de una solución que se adecue a la satisfacción de las necesidades relevadas del negocio.

Las mismas podemos resumirlas en los siguientes puntos:

- Contar con una metodología de explotación de datos.
- Establecer la gobernanza de datos.
- Disponer de una solución técnica y funcional para el problema de calidad de datos.

- Contar con un repositorio de datos único y centralizado.
- Unificar el lenguaje con una nomenclatura estándar.
- Dar soporte al proceso de toma de decisiones.
- Disponer de los *Key Performance Indicators* (KPIs) en tiempo y forma.

De manera puntual y con el fin de evaluar tanto herramientas como metodología se eligió para el proyecto piloto un set de datos de casos de fractura pertenecientes a un yacimiento de la cuenca Neuquina para lograr la optimización de la producción a través de la explotación de los datos.

Los datos pertenecen a las formaciones de Choiyoi, Petrolífera y Quintuco.

Nuestros especialistas de Ingeniería de Reservorios proveyeron de las características de esta zona:

- Muy buenas propiedades petrofísicas en dolomías gruesas y finas de Quintuco Superior que aseguran una buena conectividad entre pozos vecinos.
- Las capas Complejo Superior y Capa 2 (Qco. Sup.) contienen el 77% del POIS del proyecto y ahí centramos la selección.
- Una secundaria bien desarrollada asegura tener medianamente presurizados los reservorios.
- Fracturas con una edad promedio de 14 años que pueden mejorarse con una nueva estimulación.

B.3.0 Infraestructura Técnica

B.3.1 Base integrada

La base integrada centraliza los datos de las diferentes fuentes de una manera estandarizada y consolidada de manera de cumplir con dos objetivos básicos: satisfacer las necesidades del negocio de E & P y ser un repositorio único de datos.

Es importante destacar, que, si bien tenemos manera de analizar los datos en tiempo real, para este estudio trabajamos con datos que se almacenan en forma diaria por lo que prescindimos de series temporales.

B.3.2 PPDM

PPDM[8] (*Professional Petroleum Data Management*) es un estándar definido y creado por la PPDM *Association* para ayudar a las empresas de gas y petróleo a administrar sus datos de exploración y producción.

PPDM *Data Model* es un modelo de datos relacional robusto diseñado por expertos en distintas áreas de aplicación de la industria del petróleo, profesionales del *data management*, desarrolladores de soluciones, entes reguladores, etc. que atiende 53 temáticas diferentes de la industria, cubriendo un espectro muy amplio.

[8] 2017, Modelo de Datos de profesionales de Petróleo, PPDM3.8.

B.3.3 ETL

El término ETL (*Extract, Transform & Load*) se refiere al proceso de extracción de datos de sus fuentes originales, transformación de los mismos con el fin de adecuarlos al formato definitivo, y finalmente cargarlos en la base de datos destino.

Fig. 8: Diagrama del proceso de ETL.

B.3.4 Selección de modelo estándar de almacenamiento.

El equipo de trabajo estudió el estándar definido por PPDM concluyendo que este modelo seleccionado corporativamente por empresas tales como TOTAL, Petrobras y las principales empresas de servicios cubría mayormente las necesidades del negocio local de E & P.

El proceso de análisis detallado de este estándar determinó que el modelo PPDM representa adecuadamente nuestro modelo de producción, en tanto que debió extenderse para satisfacer los requerimientos de datos del modelo de reservorios de la empresa.

Del modelo estándar de PPDM se utilizaron los siguientes componentes:

- Áreas
- Campos
- Pozos (estados y clases)
- Punzados
- Completaciones
- Estratigrafías
- Producción
    - *Oil*
    - Gas
    - Por formación
    - Mensual
- *Well Tests*
- Fracturas y tratamientos
- Instalaciones

La extensión del modelo, propuesta por el equipo de trabajo sigue los lineamientos del estándar PPDM en cuanto al diseño de tablas y relaciones entre ellas.

Las adaptaciones se hicieron en torno a la información de:

- Propiedades petrofísicas
- Permeabilidades relativas
- PVT
- PLT
- Diseño y resultado de estimulaciones

- Pretratamientos en fracturas

B.3.5 Selección de las plataformas tecnológicas

La base de datos de PPDM se instaló sobre un servidor Oracle. Cada una de las interfaces de transformación de datos se implementó en *Sql Server Integration Service* (SSIS), la solución de Microsoft para la integración y transformación de datos.

Este conjunto de herramientas aportó las capacidades de almacenamiento, automatización y *scheduling* de los datos e interfaces.

El desarrollo de cada una de estas interfaces fue hecho con Microsoft Visual Studio y Microsoft Integration.

SSIS provee una gran cantidad de conectores a diferentes tipos de bases de datos y archivos, esto nos permite poder tomar diferentes fuentes de datos.

Estas tecnologías se seleccionaron en función del *know-how* técnico del equipo y disponibilidad tecnológica de la compañía.

Como plataforma de *analytics* se seleccionó TIBCO *Spotfire* (*Spotfire Analyst* y *Spotfire Miner*) a la cual el equipo consideró como la más apropiada y completa suite de *analytics*.

B.3.6 Interfaces & Procesamiento

Establecidos los datos a incorporar al estándar de PPDM se determinaron las fuentes de origen de los datos para poblar el modelo PPDM implementado.

En el proceso de integración intervienen personas con diferentes roles y perfiles, ya que es un proceso continuo e iterativo en el que se realizan las diferentes acciones para sanear los datos y asegurar su calidad.

Fig. 9: Proceso de saneamiento y carga de datos

La siguiente imagen muestra la arquitectura de técnica para la toma de datos desde los sensores en boca de pozo y como ese dato es obtenido para posterior explotación en el aplicativo específico.

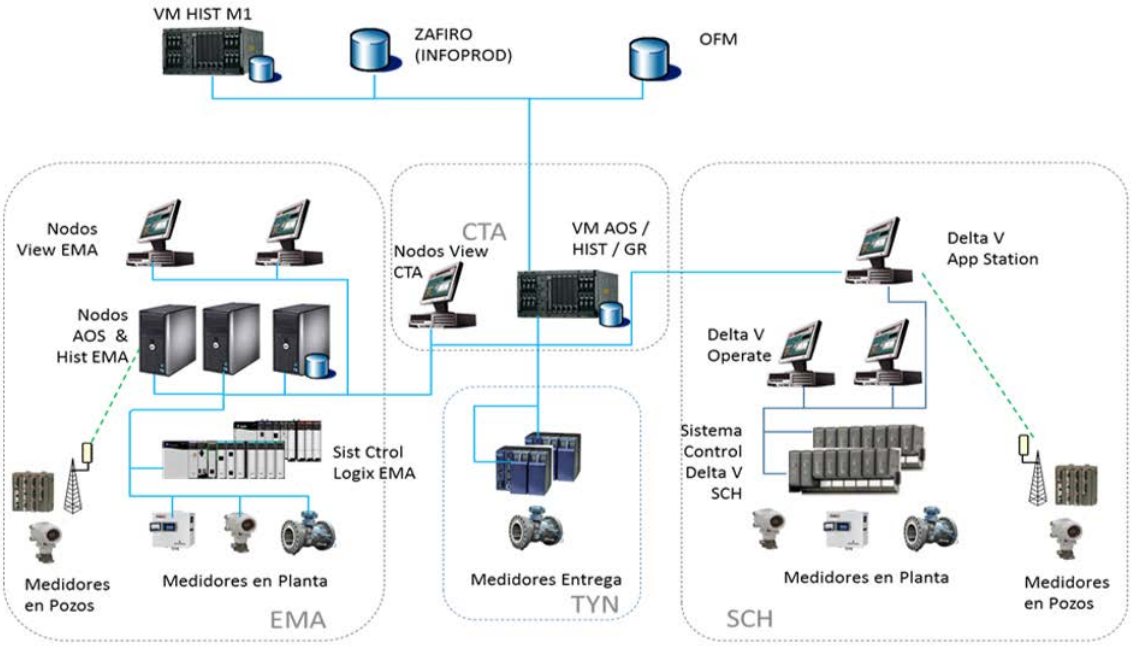


Fig. 10: Esquema funcional de origen del dato desde su fuente de datos original.

Los datos que llegan a los aplicativos son integrados para poder ser explotados y poder tomar decisiones en el menor tiempo posible.

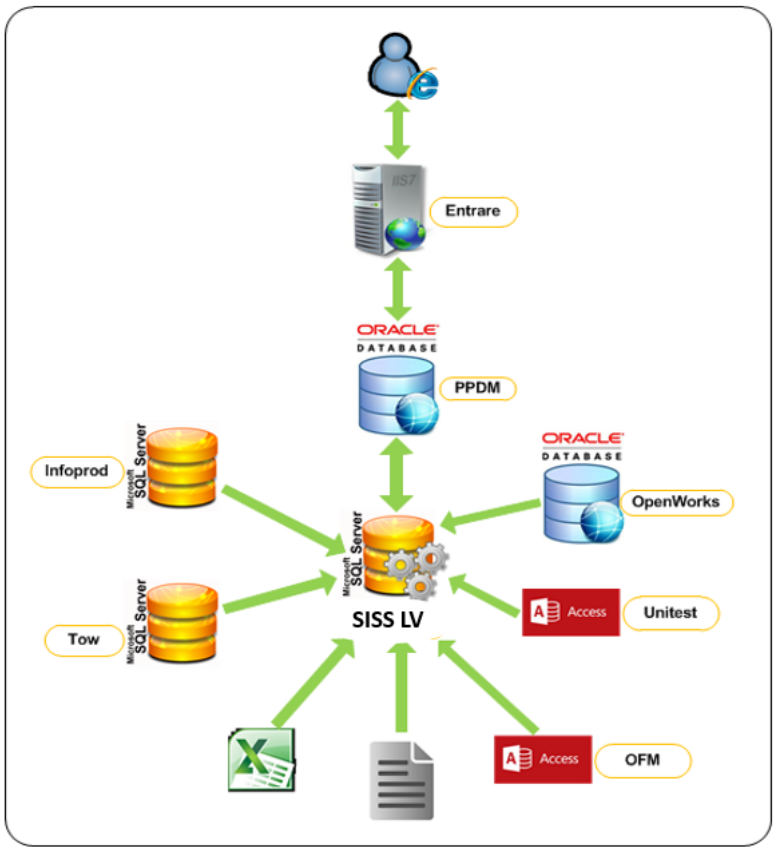


Fig. 11: Esquema funcional de base de datos integrada y sus fuentes de datos

Con SSIS se implementan las interfaces de migración de datos a PPDM desde los siguientes orígenes:

- SQL Server: TOW e Infoprod
- Oracle: Open Works
- Access: Unitest, OFM
- Excel: información de fracturas, PVT, Propiedades petrofísicas, Permeabilidades relativas, etc.
- Archivos de texto

Se desarrolló Entrare (Ensayos y Tratamientos de Reservorios) como la aplicación a ser utilizada como *front-end* de la base integrada. Esta aplicación tiene como objetivo principal ayudar a los

usuarios a registrar de una manera sencilla y consistente los datos que originalmente eran registrados en planillas.

Entre las funcionalidades más destacadas de la aplicación se encuentran:

- Registro de fracturas y tratamientos: la información de las fracturas se registra en diferentes grupos de datos; se registran las relaciones entre fracturas y punzados, las propiedades y propantes definidos en el diseño de la fractura, como así también los resultados. De esta manera se puede comparar lo planificado de lo realizado en campo. Otro grupo de datos se refiere a las propiedades petrofísicas que se encuentran en cada capa del pozo donde se realiza la fractura. También se registran las diferentes presiones y aditivos usados en la fractura. Por último, se registran los fluidos inyectados en los pretratamientos.
- PLTs: se registra para cada punzado asociado el aporte en porcentaje de agua, oil y gas, utilizados luego para calcular la producción de gas.
- Propiedades petrofísicas: la aplicación permite registrar las propiedades petrofísicas de las formaciones y vincular estos datos con pozos, áreas y campos.
- *Well tests*: es el registro de los diferentes ensayos realizados sobre un pozo para poder determinar su potencial.

Técnicamente Entrare está desarrollada como una aplicación web utilizando el *Framework .NET* de Microsoft, implementada en un servidor IIS, y consumida por los usuarios dentro de la intranet con Microsoft EDGE.

Los software específicos utilizados en las disciplinas de geología, geofísica y reservorio siguen un flujo de trabajo que se detalla en la figura adjunta. Los datos resultados son los que serán usados para la explotación del modelo.

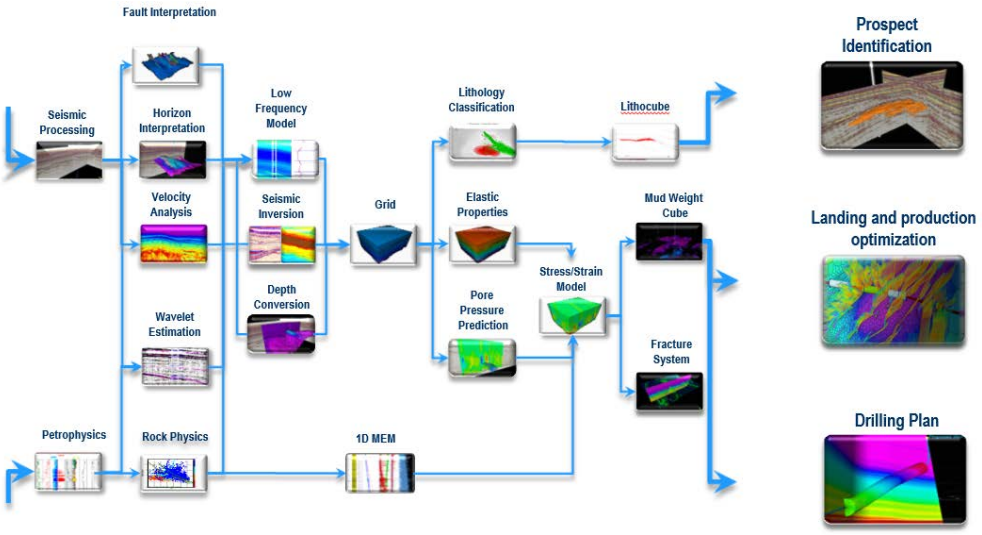


Fig. 12: Flujo de datos en aplicativos específicos de G&G[9]

B.4.0 Definición de Metodología para explotación de datos

Tras un análisis detallado de la información disponible el equipo concluyó que la metodología a seguir más se adecuaba a los procesos de nuestra empresa es la denominada CRISP-DM. Esta metodología fue ampliada agregándose de manera explícita una etapa de monitoreo de resultados considerada como imprescindible en el seguimiento permanente de los modelos analíticos desarrollados. Esta instancia se incorporó para evitar problemas de pérdida de performance por envejecimiento de los modelos desarrollados puestos en producción.

[9] 2021 – Schlumberger Inc. Flujo de Datos para aplicaciones de software específico de E & P.

B.5.0 Desarrollo de modelos analíticos

B.5.1 Comprensión del Negocio:

Del análisis del problema surge la recomendación de particionar el modelado en dos etapas. Una primera etapa en la que se seleccionará el par POZO-CAPA a ser fracturado y una segunda etapa en la que se aportará la información de los tratamientos realizados a fin de mejorar la calidad de la fractura obtenida y por ende su rendimiento.

Una primera etapa, a la que llamamos MODELO PRE, en la cual se tiene en cuenta solamente la información conocida con antelación a la ejecución de los tratamientos que llevaron a la realización de la fractura hidráulica.

Esta información consta básicamente de las variables petrofísicas medidas y/o extraídas de los perfiles de los pozos luego de perforados.

La segunda etapa, a la que denominamos MODELO POST, incorpora las variables relacionadas al tratamiento realizado sobre la capa para realizar la fractura hidráulica en la capa seleccionada. El objetivo de esta segunda etapa es esencialmente aportar al diseño de una configuración óptima de los estímulos de fractura, dadas las variables petrofísicas y la puntuación PRE del par POZO-CAPA a ser tratado.

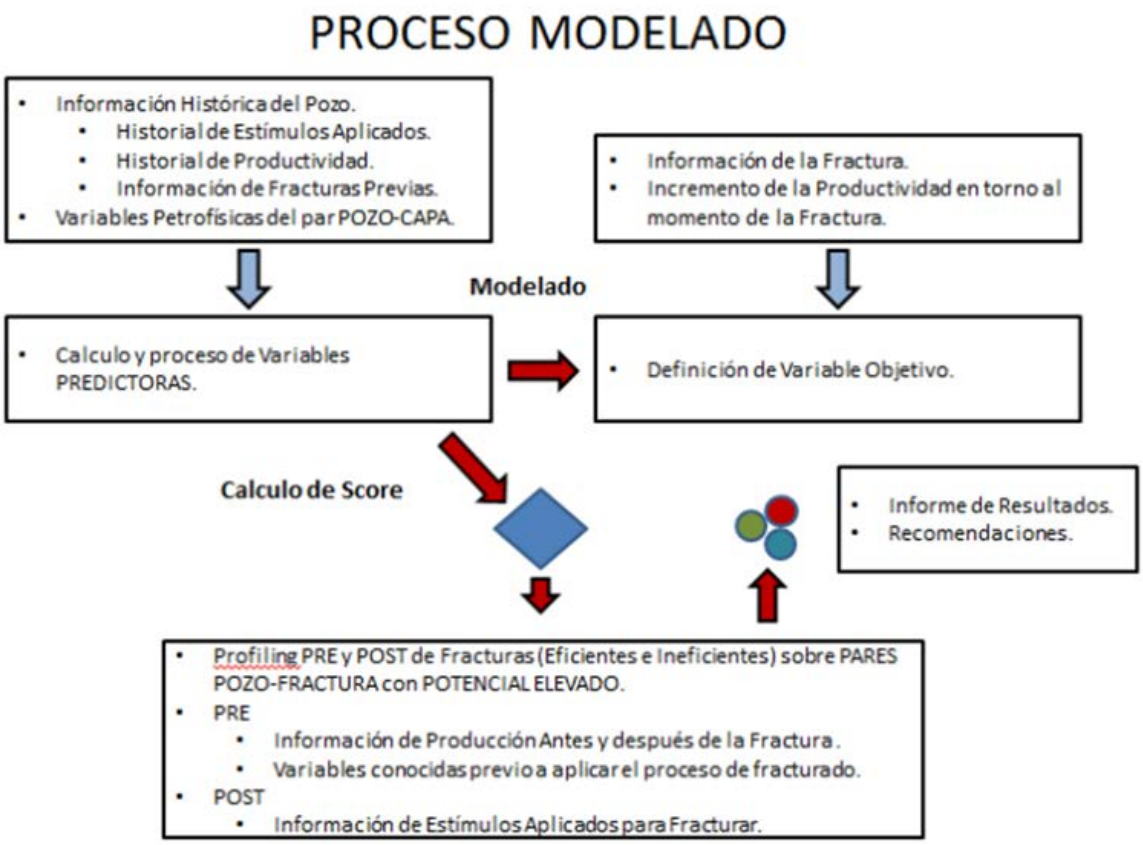


Fig. 13: Diagrama del proceso de Modelado, de los inputs, de sus etapas y resultados esperables.

En todos los casos el objetivo consiste en generar las mejores recomendaciones posibles de pares POZO-CAPA a ser fracturados que permitan maximizar el nivel de producción de la capa (i.e. del pozo) luego de realizada la fractura, evitando y/o reduciendo las pérdidas económicas devenidas de una decisión errónea en la selección de la capa a ser fracturada en un pozo determinado.

Las fuentes de datos utilizadas contienen información de productividad antes y después del proceso de fracturado, variables de completación (datos de punzado), variables petrofísicas (de reservorios) e información de los estímulos aplicados (tipo de agentes, cantidad de bolsas, forma de la colocación en la formación, volúmenes inyectados, fluidos, presiones, etc.) para producir cada una de las fracturas en los pozos del reservorio.

En la figura 13 se muestra un esquema de cómo serán las etapas del proceso de modelado.

B.5.2 Comprensión de los Datos:

El proceso de comprensión de los datos se desarrolló con *Spotfire Analyst*. Con esta herramienta se caracterizaron las fuentes de datos en cuanto a su completitud, distribución, duplicidad de registros, concentración de valores y calidad de datos de manera general.

En la figura 14 se muestran capturas de pantalla de la herramienta en este proceso, conocido como *PROFILING* de la fuente de datos.

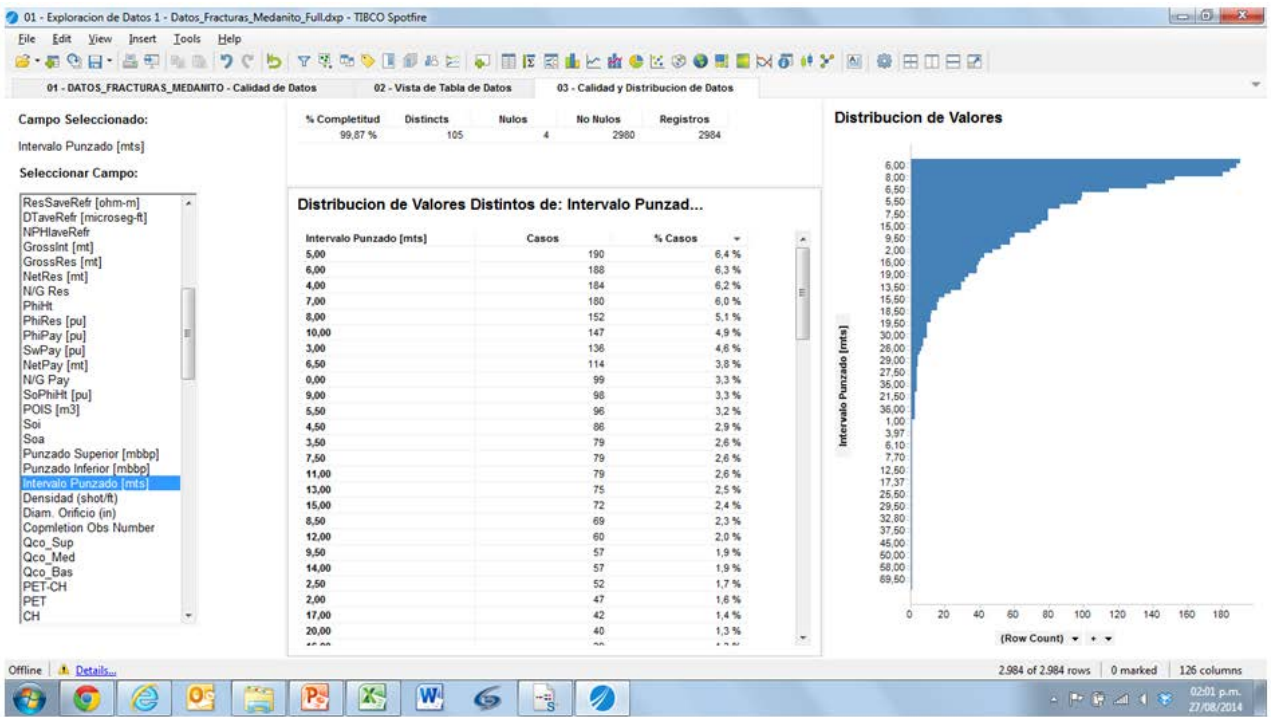
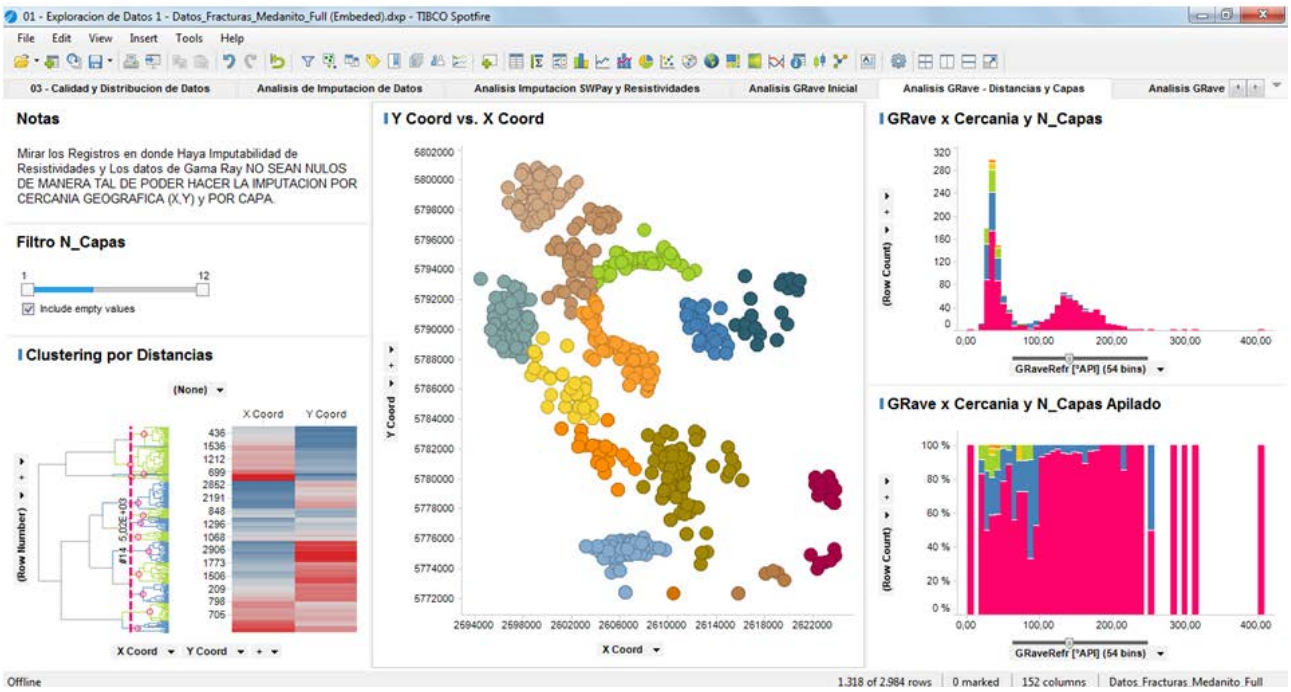


Fig. 14: Spotfire Analyst en el proceso de PROFILING (QA) de las fuentes de datos.

El resultado de esta primera instancia permite hacer una primera selección de variables, las que por su completitud y variabilidad puedan ser incorporadas en un análisis posterior para evaluar su poder predictivo.

Realizada esta preselección de campos se procedió a la realización de un análisis exploratorio univariado y multivariado a fin de caracterizar adecuadamente a cada una de las variables a ser empleadas.

A continuación, en la figura 15, se muestran capturas de pantalla de *Spotfire Analyst* utilizándose con este fin exploratorio.



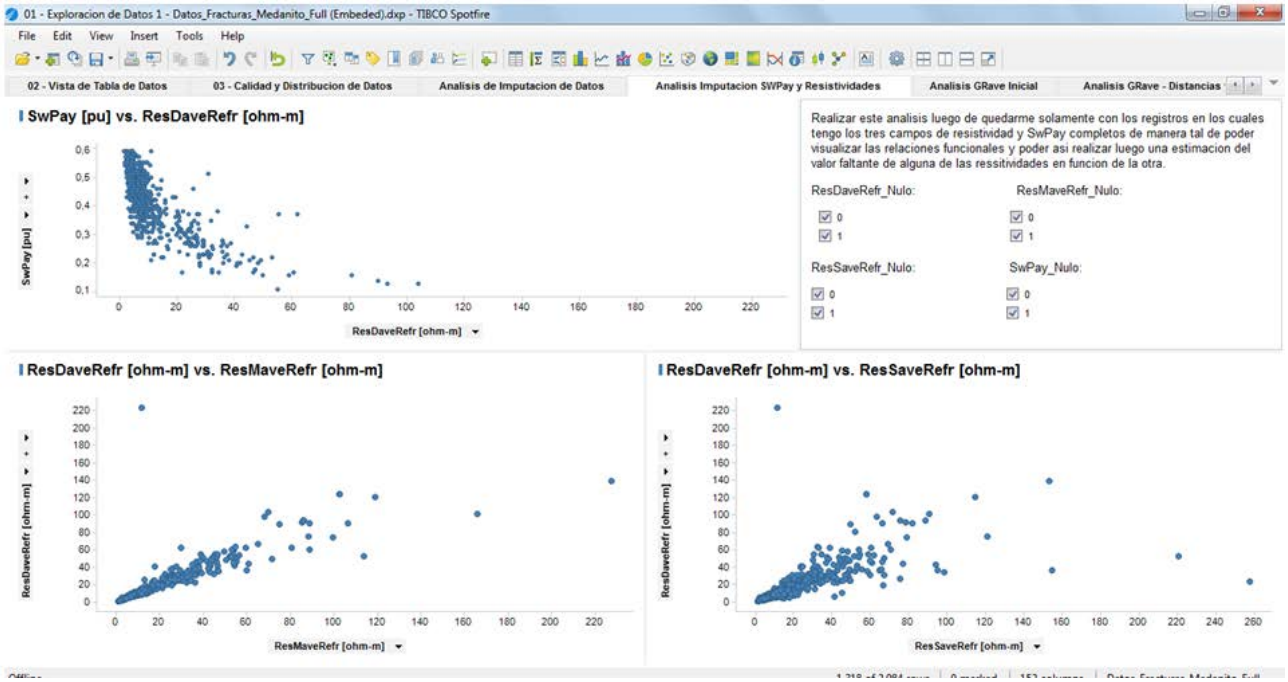


Fig. 15: Spotfire Analyst utilizado en el proceso de Exploración de Datos

En esta etapa exploratoria y de comprensión de los datos se definió también el OBJETIVO TÉCNICO del análisis. Este objetivo técnico consiste en la marca a ser utilizada como variable objetivo por los algoritmos de aprendizaje.

El mismo fue definido utilizando en *BOX Plot*, observando el cambio en la productividad diaria acumulada, calculada desde datos de la producción de la capa 6 meses antes y después del proceso de fractura.

En la figura 16 se muestra una captura de pantalla de *Spotfire Analyst*, en la cual se visualiza este proceso.

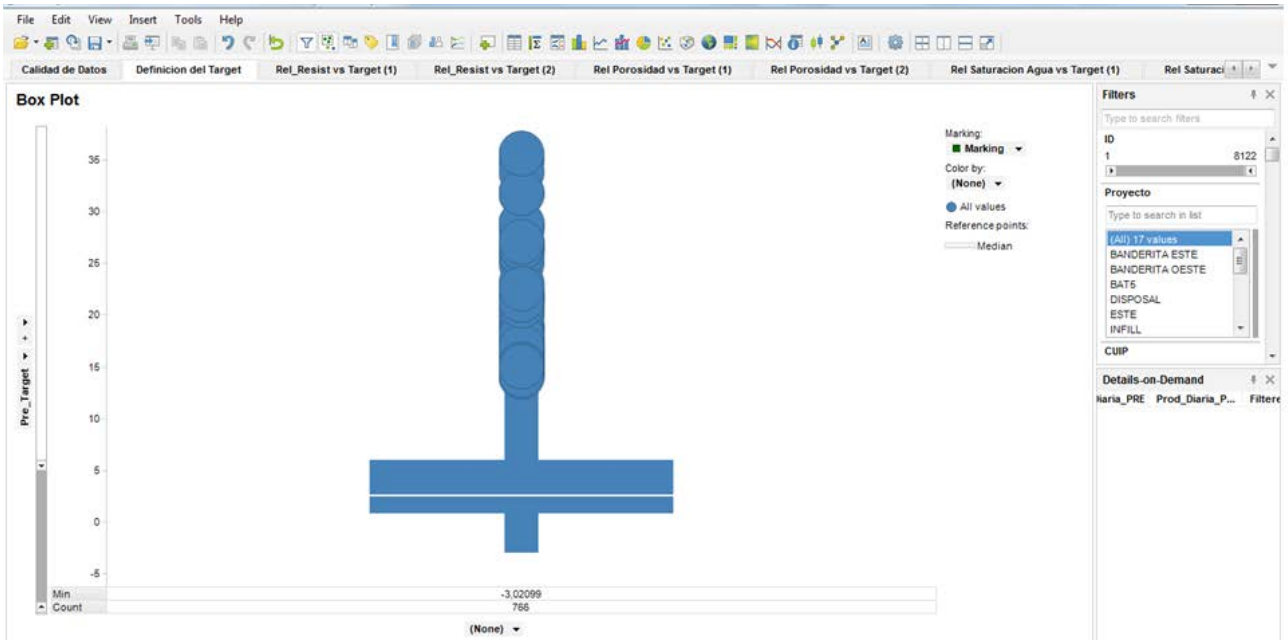


Fig. 16: Definición de Objetivo Técnico de análisis utilizando un Box Plot de Spotfire Analyst.

Como parte de este proceso exploratorio, se realizó en proceso de preselección de variables teniendo en cuenta el aporte de poder de discriminación "APARENTE" de cada una de ellas en relación al objetivo técnico de análisis.

De este análisis se seleccionaron para el modelo PRE las siguientes variables:

- ResDaveRefr - Resistividad *Depth* o Profunda, promedio
- ResSaveRefr - Resistividad *Sallow*, superficial o poco profunda, promedio.
- Se calculó, además, la diferencia entre estas dos resistividades.
- GraveRefr – Emisión de Rayos Gamma promedio – Indicador de Litología.
- Porosidad Media
- SWPay – Saturación de Agua promedio.
- RHOBaveRefr – Densidad Promedio.
- Flg_Mas_de_Una_Capa –Indicador distingue entre los casos en que se fracturó una o más de una capa.

Y las siguientes se adicionaron a las anteriores para el Modelos POST:

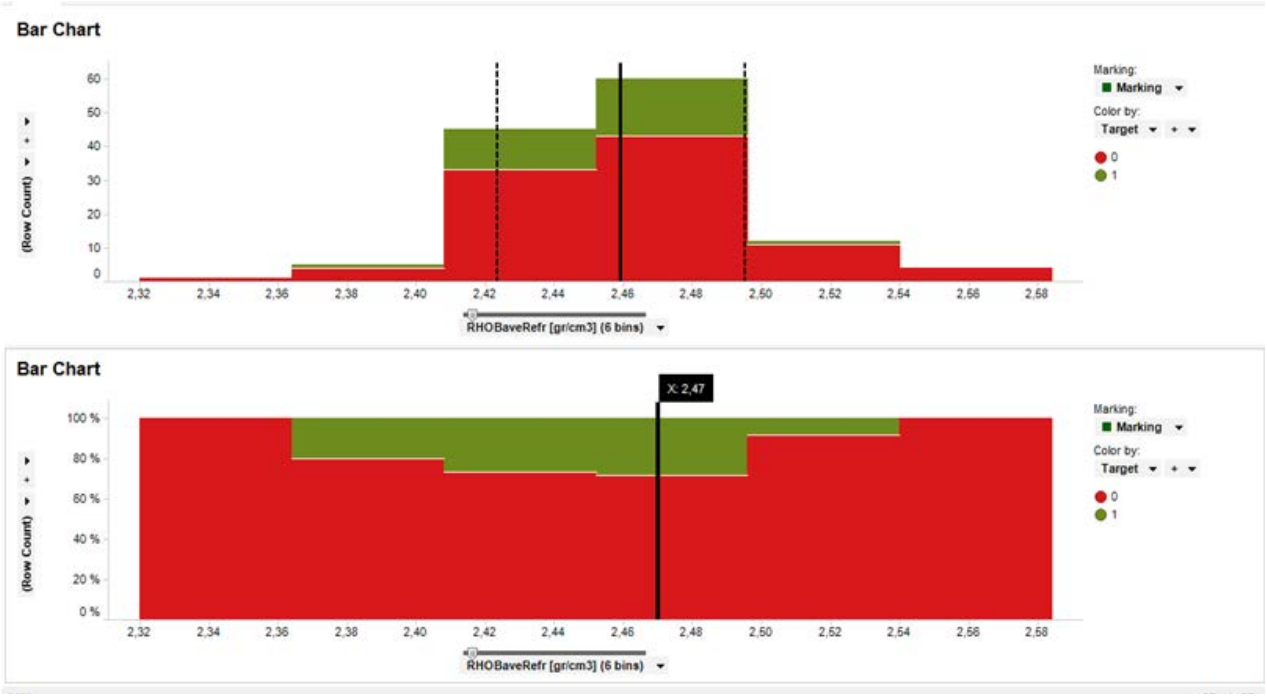
- Intervalo_Punzado
- Tipo_Arena_Marca
- Arena_Bombeada
- Arena_en_Formacion
- Caudal_x_mt_Punzado
- Caudal
- Colchon
- HHP
- Vol_Iny_x_mt_Capa
- Vol_Iny
- Vol_TT_x_mt_Punzado
- Vol_Tratamiento
- Fluido_Fractura

En esta etapa se definieron además los procesos a realizarse sobre estas variables con la finalidad de obtener una representación de los datos que permita resolver la tarea objetivo con mayor facilidad.

B.5.3 Preparación de Datos:

En esta instancia se realizó el procesado de las variables definidas en la última etapa del paso anterior. Las variables se procesaron aplicando procesos de *capping, ranking, centering* y *flagging* según resultó necesario de manera tal de lograr esta expresividad.

En la figura 17 se ven partes de este proceso, el cual fue implementado posteriormente en Spotfire *Miner.*



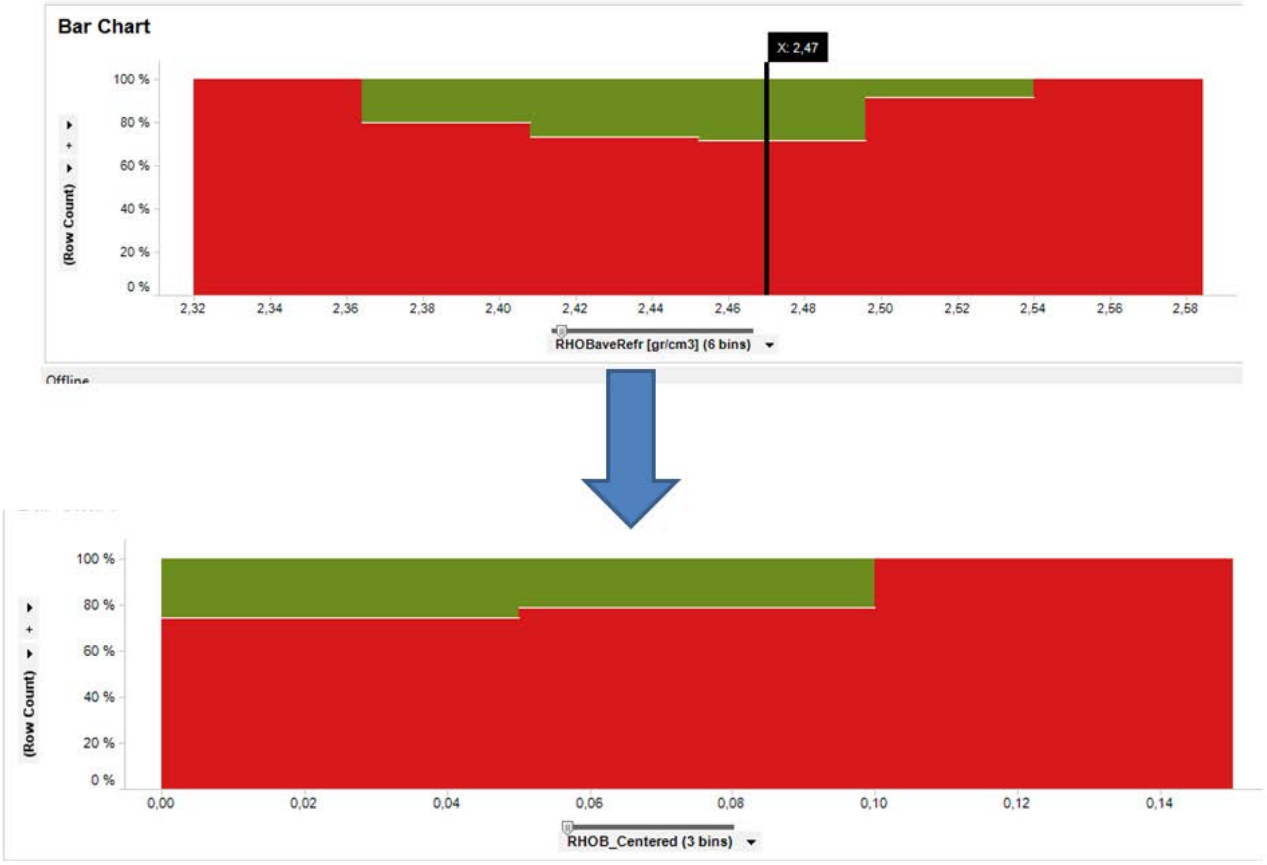


Fig.17: Procesamiento de variables

El paso posterior consistió en determinar las muestras sobre la que se desarrollarán los modelos y la muestra sobre la cual será testeado el correcto aprendizaje de los modelos.

Las proporciones de estas fueron configuradas en 70% para entrenamiento y 30% para validación. Esta selección fue hecha a fin de contar con los registros suficientes en la muestra de entrenamiento para lograr el mejor ajuste posible de los modelos.

En la figura 18 se muestra una ruta de *Spotfire Miner* en la que se ajustan diversos modelos para luego comprarse y seleccionarse.

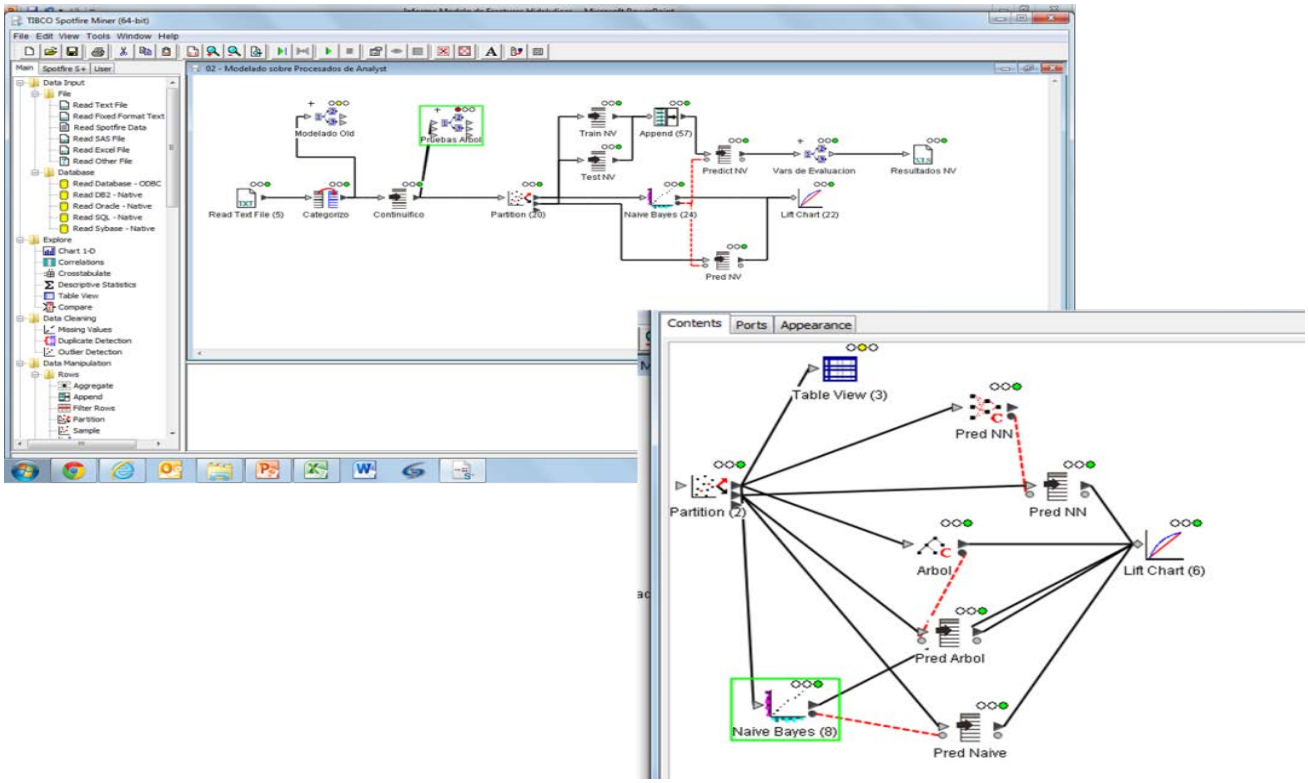


Fig. 18: Ejemplo de rutas de análisis en Spotfire Miner en la que se ven el procesado de datos, el ajuste de modelos y su posterior evaluación y comparación.

B.5.4 Modelado:

En esta etapa se desarrollaron los dos modelos antes mencionados (modelos PRE y POST respectivamente).

Se probaron distintas técnicas de modelado tales como Árboles de Clasificación, Naive Bayes y Redes Neuronales, dejándose de lado a las Regresiones Logísticas debido a que la cantidad de registros en las muestras no era suficiente como para garantizar un ajuste estable de este tipo de modelos.

B.6.0 Evaluación de Modelos:

Los distintos modelos fueron evaluados con los gráficos de Ganancias. En las figuras 19 y 20 se muestran estos gráficos para los modelos PRE y POST respectivamente.

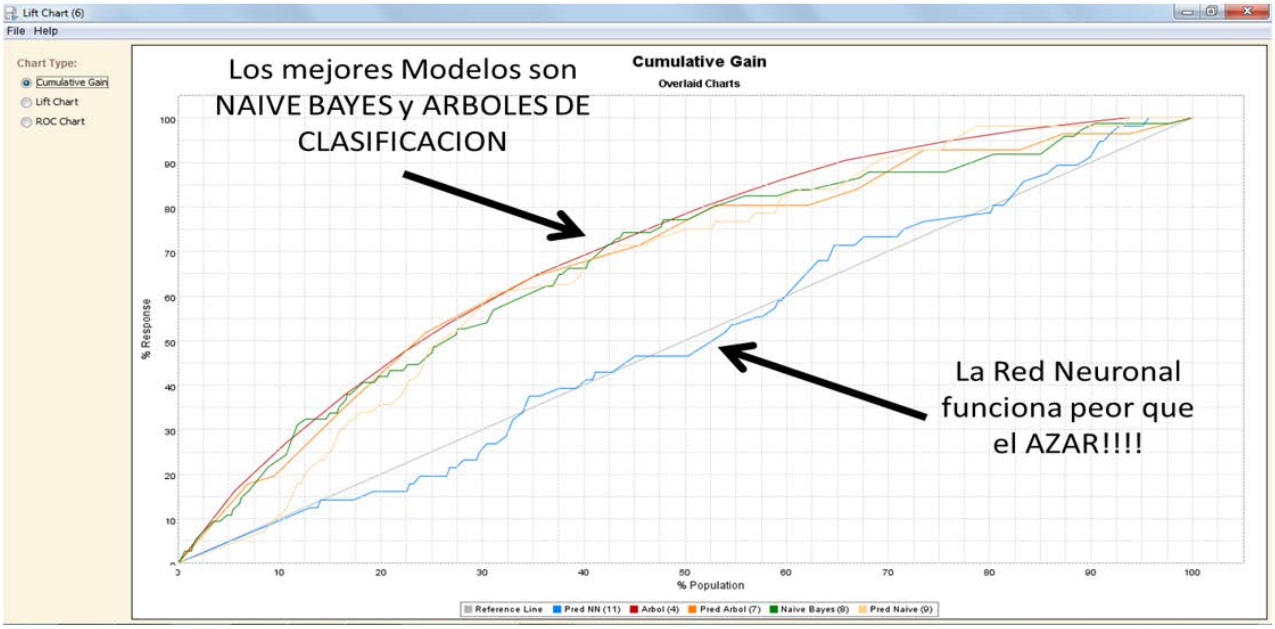


Fig. 19: Evaluación de Modelo Pre.

En la figura 19 se muestra por qué se descartaron a las redes neuronales. El modelo PRE finalmente seleccionado fue el NAIVE BAYES, debido a que, por la cantidad de registros disponibles, los árboles de decisión no resultaban fuertemente estables.

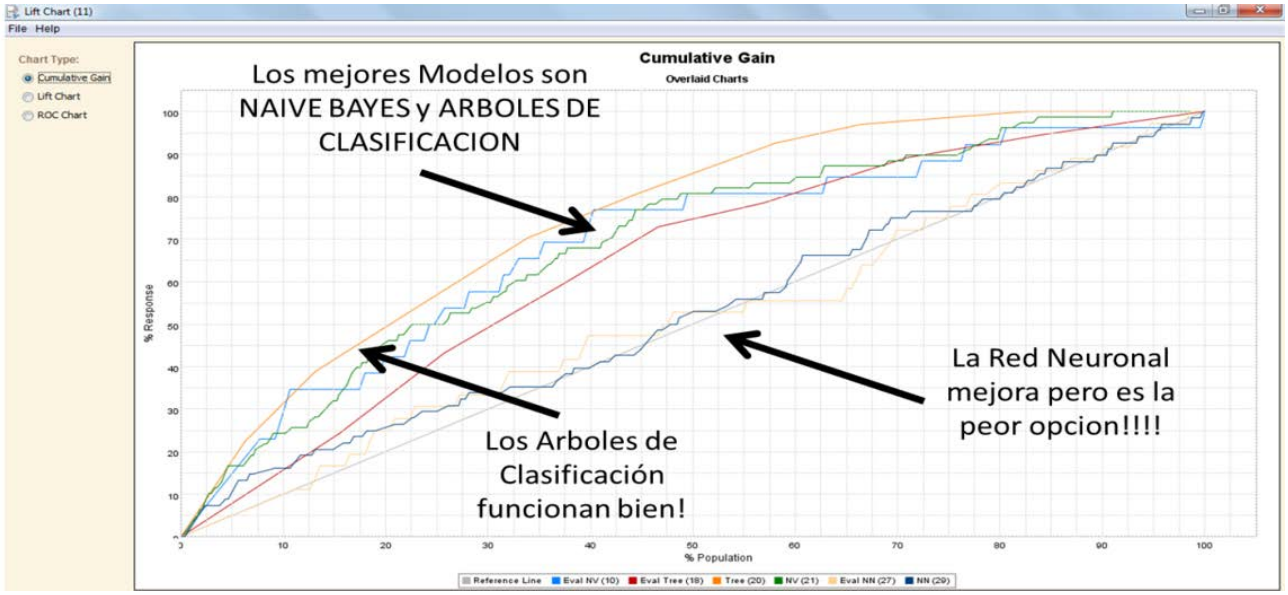


Fig.20: Evaluación de Modelos POST

En la figura 20, se visualiza la evaluación para los modelos POST. Nuevamente la cantidad de registros disponibles nos llevó a seleccionar a NAIVE Bayes como el modelo a ser implementado.

En la figura 21 se muestra la curva de ganancias para el modelo POST con un par de conclusiones:

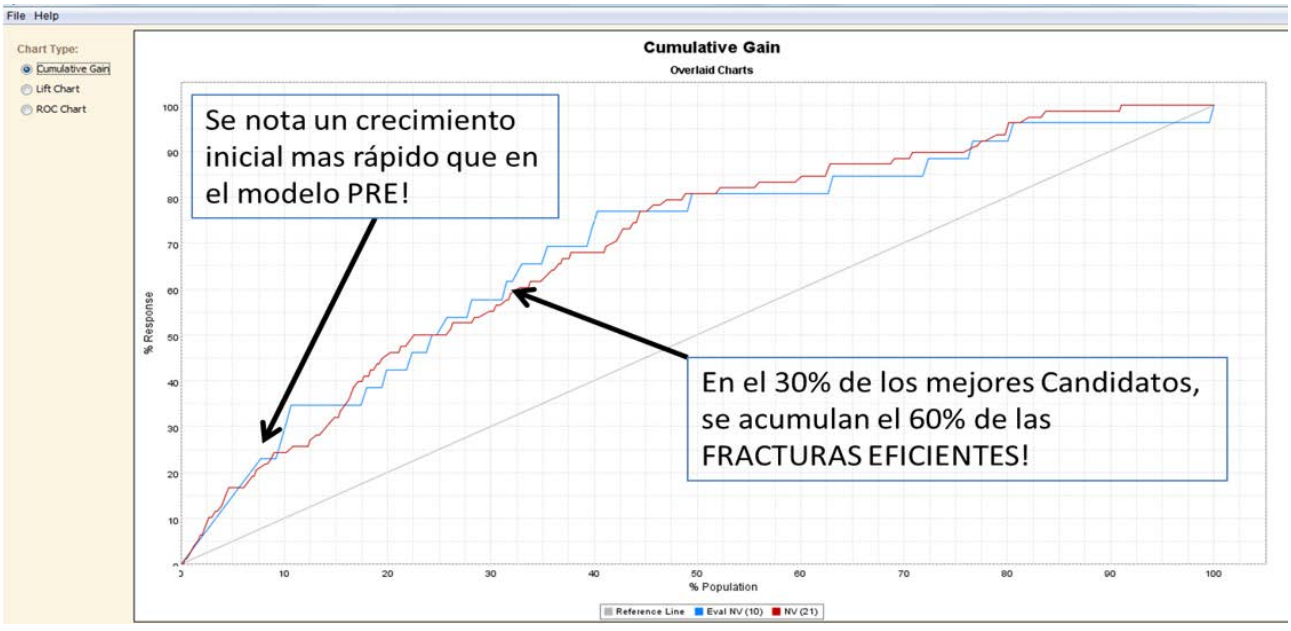


Fig. 21: Curva de Ganancias para Evaluación del Modelo POST seleccionado.

B.6.1 Implementación (Generación de Recomendaciones):

Una vez seleccionados los modelos analíticos a ser implementados, los mismos se incorporaron en una ruta de Spotfire Miner a fin de iniciar el proceso de generación de recomendaciones. En la figura 22 se muestra esta ruta impactando sobre un archivo Excel como salida.

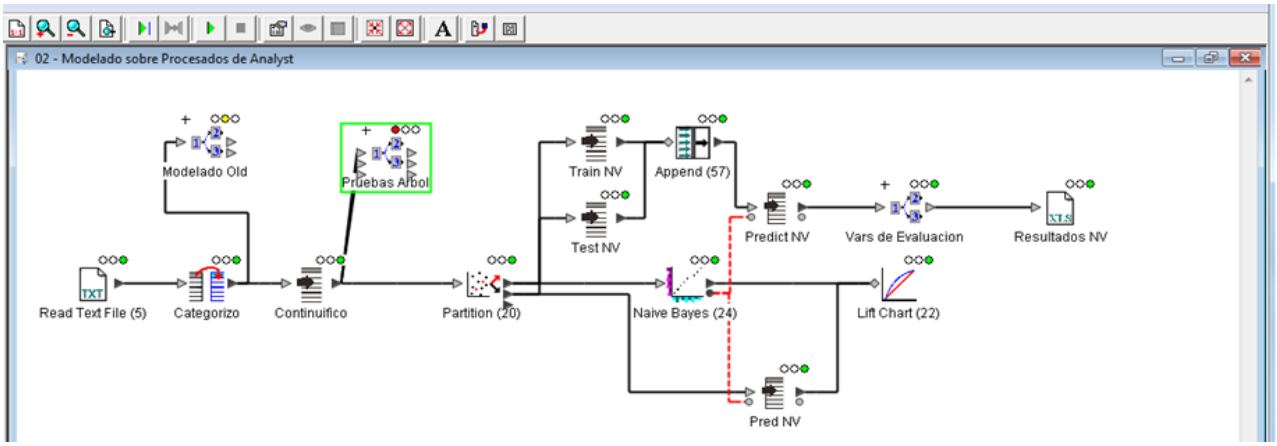


Fig. 22: Ruta en Spotfire Miner con la implementación de los modelos para la generación de recomendaciones.

B.6.2 Integración del producto en el proceso de toma de decisiones.

La información proporcionada por el modelo será estudiada en detalle por los especialistas del negocio quienes evaluarán la factibilidad física y económica de continuar con las recomendaciones generadas en el campo.

Esta metodología cambia la forma de trabajo a la que están habituados los ingenieros del sector, agregando una nueva serie de indicadores desarrollada desde los datos y produciendo la necesidad de generar un cambio en el método de trabajo actual.

Es importante señalar que este modelo puede ser utilizado repetidamente por ingenieros sin conocimientos de *Data Mining* ya que solamente deben suministrar datos en la entrada y estudiar la salida aportando una manera sencilla y precisa para la toma de decisiones.

7.0 Conclusiones

La primera conclusión en este trabajo es que el camino hacia la explotación de datos sostenible conlleva una gobernanza de datos mandatoria, así como un examen continuo de las soluciones que existen en el mercado tanto para predicción como para análisis.

El conocimiento detallado de los datos y una adecuada selección de variables acorde al problema a estudiar es otra condición necesaria para el éxito de los proyectos.

Alcanzar este objetivo implicó disponibilizar la infraestructura tecnológica necesaria, generar los procesos de trabajo y sobre todo trabajar en la capacitación y el cambio de la forma de trabajo de todos los intervinientes. Este trabajo es continuo y requiere de gran compromiso de la alta gerencia y sobre todo de equipos multidisciplinarios integrados en cada una de las etapas. Una de las claves en la conformación del equipo fue la inclusión de un especialista de analytics perteneciente a una empresa de servicios que permitió acelerar la curva de aprendizaje y el proceso de desarrollo en sí.

Establecer el objetivo del trabajo es fundamental para entender el problema y el tipo de herramienta a utilizar. Entender qué necesitamos resolver posibilita la solución.

De manera general las conclusiones obtenidas en el desarrollo del presente trabajo se encuadran en el hecho de que la implementación de técnicas de *analytics,* no solo son posibles en la industria del Petróleo y gas, sino que además aportan un gran valor al negocio, optimizando el proceso de

toma de decisiones en la selección de pares POZO-CAPA a ser fracturados con el fin de maximizar la capacidad de producción acumulada de un pozo.

La calidad de dicha selección, basada en un *SCORE* obtenido por técnicas de aprendizaje automático, produce un gran valor económico para la empresa debido a que posibilita la reducción de costos de fracturas que no han producido los efectos deseados en la productividad del pozo.

De manera puntual podemos concluir que los mejores modelos han sido obtenidos con los algoritmos NAIVE Bayes y árboles de clasificación. Dejando finalmente fuera a estos últimos, y a las regresiones logísticas debido a que la cantidad de registros no fue la suficiente para obtener resultados estables en los modelos.

Por otro lado, pudimos ver como los modelos basados en redes neuronales para clasificación no funcionaron de manera adecuada, pudiendo inferir que quizá, este tipo de problemas no sean del tipo adecuado para ser tratados con dichas técnicas.

En cuanto a la evaluación de la plataforma analítica *Tibco Spotfire,* podemos concluir que ha resultado buena en cuanto a su flexibilidad ya que contiene todas las funcionalidades y algoritmos necesarios para el desarrollo y evaluación de modelos analíticos, alcanzando el grado de óptima al conjugar simpleza y velocidad en el desarrollo además de aportar información y conocimiento instantáneo en el proceso.

Podemos destacar que la preparación de datos es crítica para asegurar el éxito del proyecto y al que más tiempo recomendamos dar en la planificación de un proyecto de estas características.

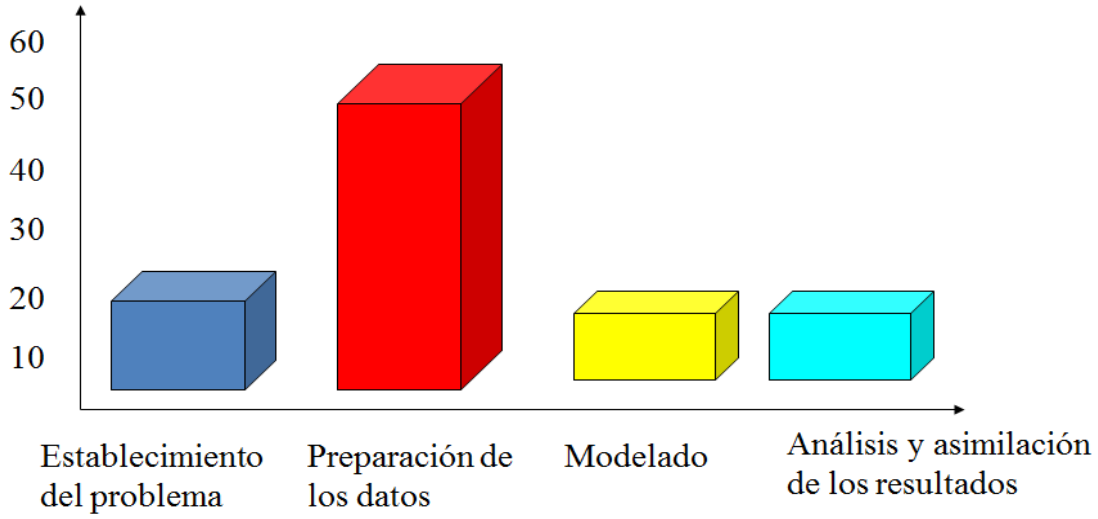


Fig.23. Distribución de tiempos de proyecto.

Como corolario de este trabajo se ofrece un conjunto de recomendaciones clave para lograr una implementación exitosa:

- Comenzar con un fin en mente (enfocar el problema).
- Seguir una metodología probada. Entender que Inteligencia Analítica es un proceso iterativo e interactivo.
- Adecuar las expectativas a las posibilidades reales de la Inteligencia Analítica.
- Delimitar claramente el problema y las métricas de evaluación.
- Participación de gente con conocimiento del negocio y el suficiente nivel de autoridad en cada etapa del proceso.
- No es posible hacer inteligencia analítica solo con consultores externos.
- Disponer de los datos en tiempo y forma, con soporte de todas las áreas involucradas en su registro, almacenamiento, extracción y manejo.
- Que el usuario pueda interpretar los resultados (conocimiento, indicador, pronóstico, etc.)

8.0 Bibliografía

- Keith R. Holdaway (2014). Harness Oil and Gas BIG DATA with Analytics.
- Gareth James, Daniela Witten, Trevor Hastie, Robert TibshiraniAn (2013) Introduction to Statistical Learning: With Applications in R.
- José Hernández Orallo, María José Ramírez Quintana, César Ferri Ramírez (2004). Introducción a la minería de datos.
- Tim Crocker, 2014, *Spotfire Oil* & *Gas Production Optimization*, *TIBCO Spotfire Conference*, USA.
- Taylor, L., Schroeder, R., Meyer, E. (2014). Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same? Big Data Society.
- Fayyad, Piatetsky-Shapiro, Smyth, "From Data Mining to Knowledge Discovery: An Overview", in Fayyad, Piatetsky-Shapiro, Smyth, Uthurusamy, *Advances in Knowledge Discovery and Data Mining,* AAAI Press / The MIT Press, Menlo Park, CA, 1996, pp.1-34
- Shearer C., *el modelo CRISP-DM: el nuevo plan para la minería de datos,* almacenamiento de los datos J (2000); 5:13-22.