Ciencia de Datos

Asignatura

MINERIA DE DATOS

Mto. José C Roberto Olvera López
Data Science Consultant

jroberto.olveral@gmail.com

Agosto 2024

Minería de Datos
Objetivos del curso

- Analizar **grandes volúmenes de datos** de una organización permitiendo extraer patrones de comportamiento que generen valor a una organización
- Seleccionar y aplicar técnicas de Minería de Datos para **construir modelos predictivos**, que permitan extraer conocimiento del negocio**.**
- Elaborar una **arquitectura** que asocie los componentes involucrados en el proceso de Minería de Datos
- Describir el **Proceso de Minería de Datos** en todas sus etapas detallando Actividades, Técnicas, Participantes, dando énfasis en los requerimientos de negocio para su aplicación.
- Mostrar Experiencias y **Casos de Éxito** aplicado a Minería de Datos

MINERIA DE DATOS

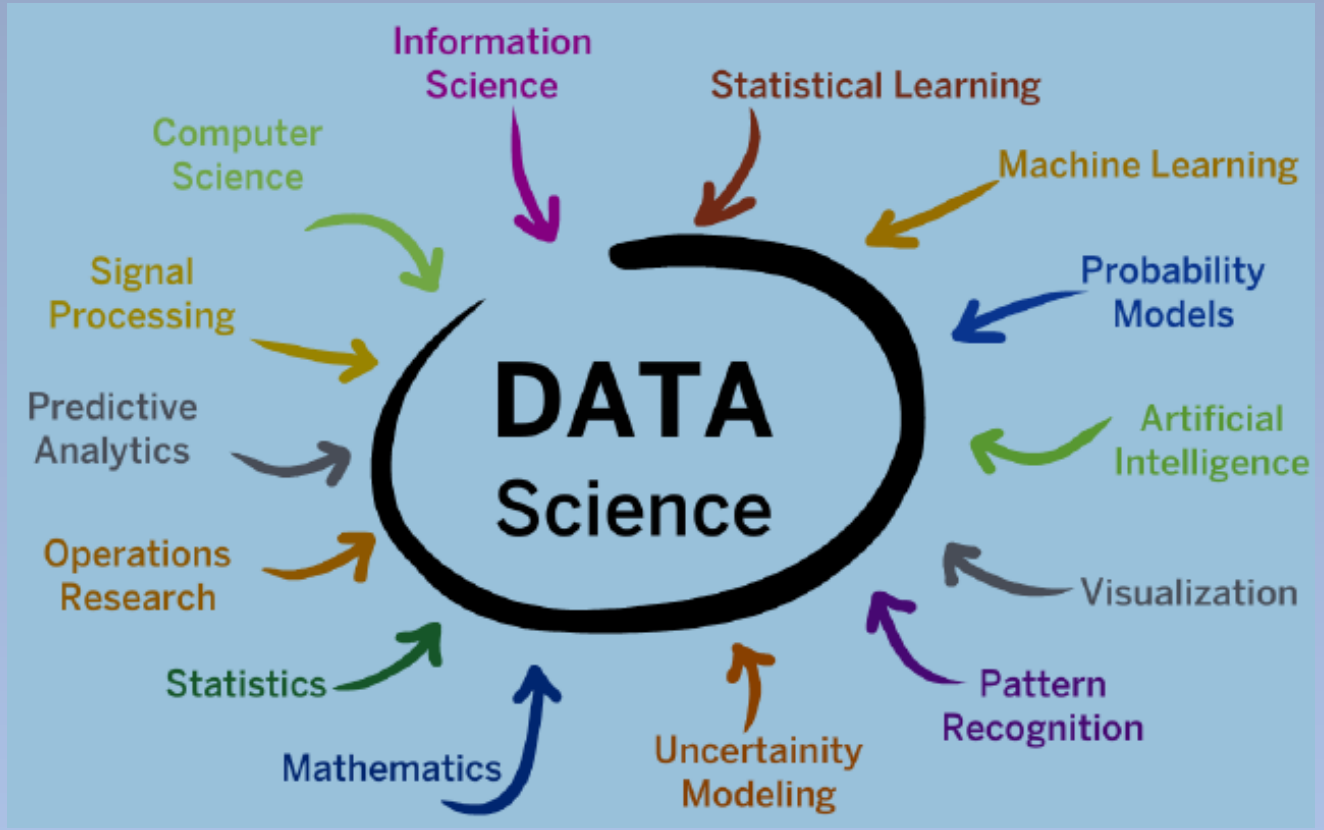
Contenido Temático

1. Introducción al Descubrimiento del Conocimiento en Datos (KDD)
2. Arquitectura de Minería de Datos
3. Proceso de Minería de Datos
4. Técnicas de Evaluación de Modelos
5. Aplicación en la Minería de Datos

Ciencia de Datos

Ciencia de Datos es una interdisciplina acerca de los procesos y sistemas que habilitan la obtención del conocimiento a partir de los datos.

- Recopilar la información, hacer un análisis y aplicar a la toma de decisiones
- La información se toma de cualquier tipo de datos, tanto estructurados como no estructurados.
- Ciencia de Datos emplea técnicas y teorías en un amplio rango de disciplinas.



Ciencia de Datos

La Ciencia de Datos es una de las disciplinas con mayor demanda profesional, dado que la cantidad de usos y aplicaciones en el mundo real es cada vez mayor.

- el reconocimiento facial,
- las órdenes por voz a los **asistentes virtuales**, los chatbots,
- la **prevención** del fraude bancario,
- cuando **Netflix** nos recomienda una película o **LinkedIn** nos sugiere una búsqueda laboral,
- el otorgamiento de **préstamos**
- En la optimización de **procesos de negocio**
- el **diagnóstico médico … Etc.**

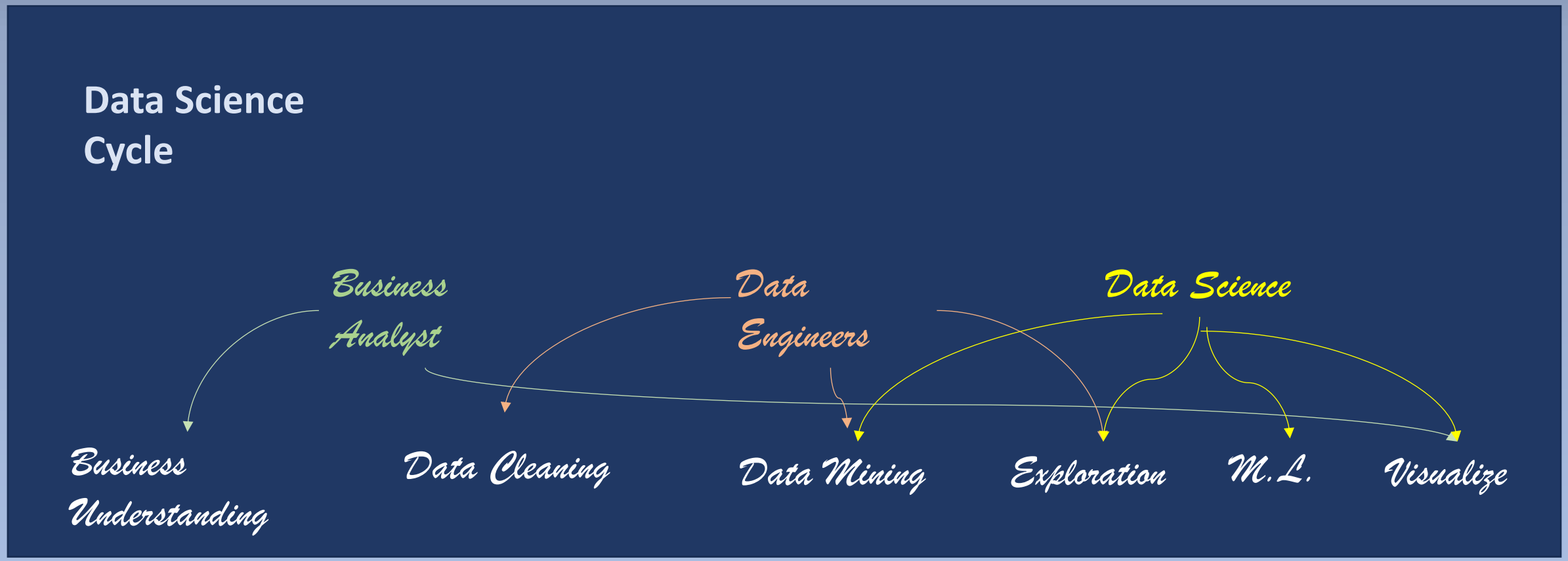
nos dan una idea de las oportunidades que se presentan en diferentes industrias.

Los profesionales en Ciencia de Datos que cuenten con las herramientas adecuadas, visión de negocios, creatividad y espíritu emprendedor estarán frente a una gran oportunidad para tener una ventaja competitiva.

Data Science

Business Data → Knowledge Insights → Actions

Data Science Cycle



Ciencia de Datos

Temas de importantes

Democratización de los Datos - **C**reando nuevas formas de trabajo, donde las herramientas, aplicaciones y dispositivos empujan los conocimientos inteligentes a las manos de todos para permitirles hacer su trabajo de manera más efectiva y eficiente

Inteligencia Artificial - Su efecto en el análisis de negocios será permitir predicciones más precisas, reducir la cantidad de tiempo que dedicamos al trabajo repetitivo para que actúen sobre información basada en datos, independientemente de su función y nivel de experiencia técnica.

Cloud y Data as a Service - Significa que las empresas pueden acceder a fuentes de datos que han sido recopiladas y seleccionadas a través de servicios en la nube en un modelo de pago por uso. Esto reduce la necesidad de recopilación y almacenamiento de datos costosos.

Real Time Data - Saber qué está pasando en este momento tener los datos en tiempo real se está convirtiendo cada vez más en la fuente de información más valiosa para las empresas para la toma de decisiones.

Ciencia de Datos

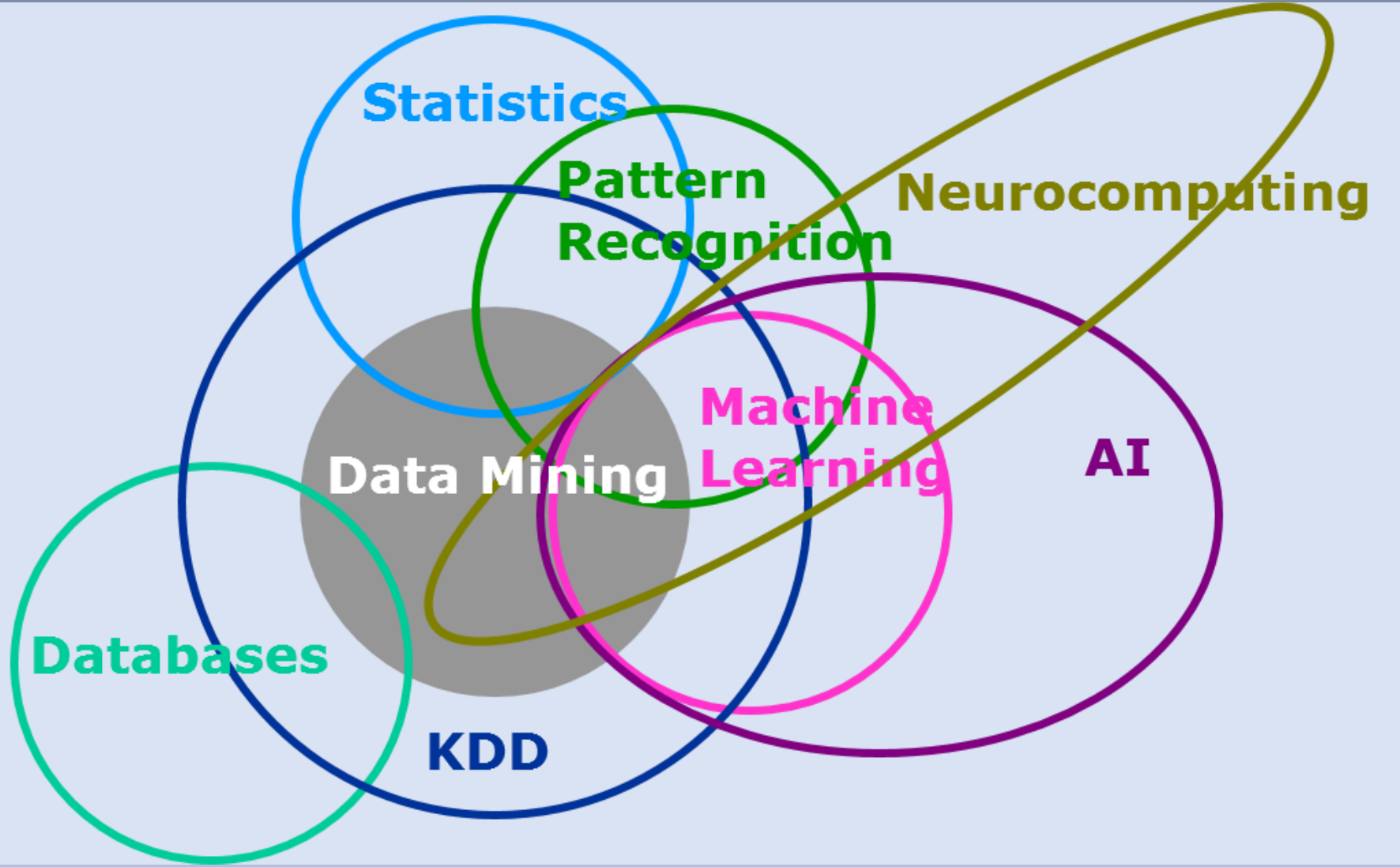
MINERIA DE DATOS

Introducción al Descubrimiento del conocimiento en datos (KDD)

Mto. José C Roberto Olvera López

Agosto, 2024

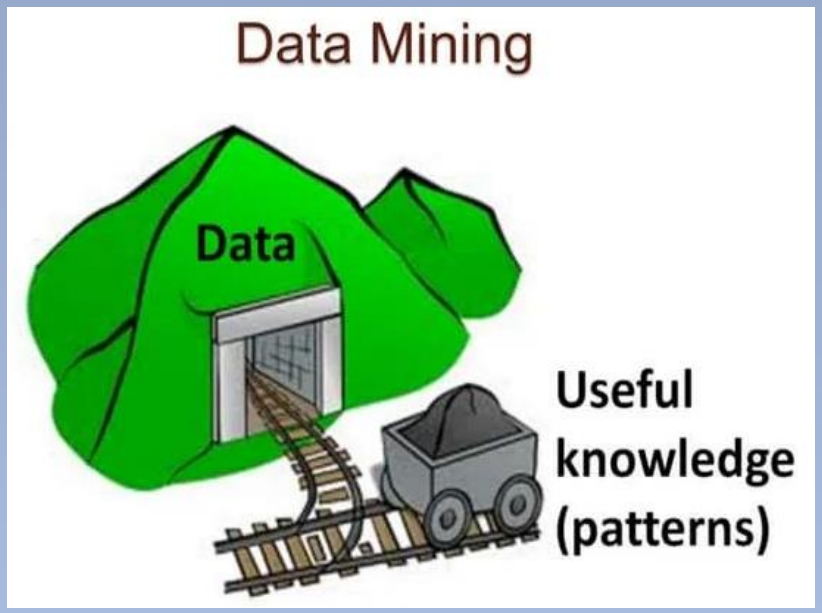
Tecnologías aplicadas en la Información



Minería de Datos

Minería de Datos es el proceso de análisis de grandes conjuntos de datos para encontrar información relevante que se pueda usar con un fin específico.

- Permite hallar anomalías, patrones y correlaciones para predecir resultados.
- Descubrir patrones y tendencias de datos importantes y transformar el Big Data en soluciones de gestión eficaces
- La Minería de Datos es fundamental para la Ciencia de los Datos y la Inteligencia de Negocio
- Mediante su uso las empresas pueden aprender más sobre sus clientes y proveedores para desarrollar estrategias más efectivas.

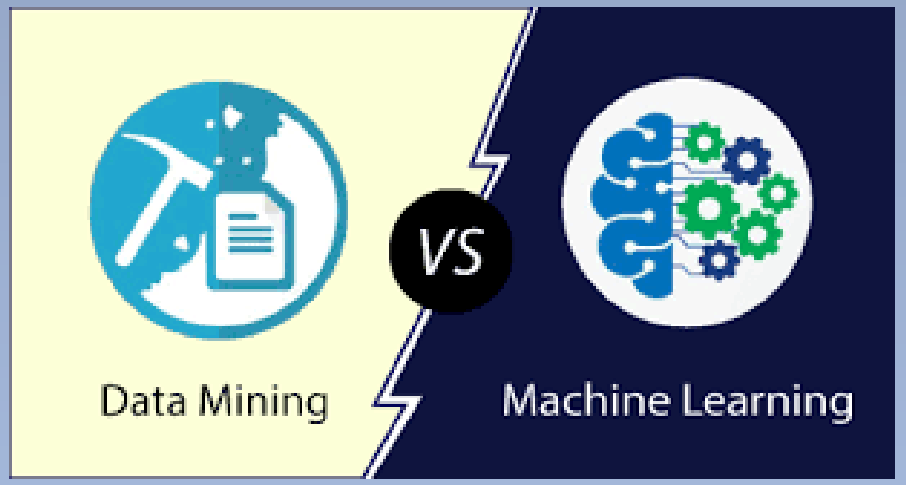


Minería de datos vs. KDD (Knowledge Discovery in Databases)

- La **Minería de Datos** es una parte del procedimiento denominado Knowledge Discovery in Databases (KDD).
- Mientras que **KDD** es un proceso general de extracción de conocimiento de datos.
- La Minería de Datos se ocupa específicamente del reconocimiento de patrones en los datos.
- En otras palabras, la Minería de Datos es la aplicación de un algoritmo particular para el propósito general del proceso KDD.
- KDD es iterativo, y durante este proceso, se pueden hacer varios ajustes, incluido el refinamiento de la evaluación y en la Minería.

Data Mining and Machine Learning?

- Ambos analizan conjuntos de datos para hacer predicciones y obtener información.
- Sin embargo, se basan en principios diferentes.
- En la **Minería de Datos** los patrones no se conocen de antemano y deben establecerse, y utiliza algoritmos para descubrir correlaciones e interdependencias en los datos y descifrar su significado.
- En **Machine Learning**, el análisis está precedido por el establecimiento de criterios para la categorización de datos. Permite descartar datos inadecuados del análisis y aprender de datos que son generados.



Minería de Datos

Minería de Datos Procesos y técnicas - con los que hacer ese análisis y extracción de datos para detectar patrones concretos.

- **Técnica de asociación o relación.** para detectar un patrón de comportamiento como identificar productos combinados que los usuarios suelen adquirir juntos habitualmente.
- **Técnica de clasificación**. clasifica elementos o variables en grupos predefinidos asociados a una variable objetivo. Utilizan estadísticas, árboles de decisión, redes neuronales…
- **Técnica de agrupación**. Agrupar elementos u objetos con características similares
- **Técnica de predicción**. se basa en el uso de los datos históricos, para hacer predicciones de comportamiento.
- **Técnica de patrones secuenciales**. Se usan los datos de transacciones para identificar patrones o tendencias similares, logrando que se pueden comparar periodos de un año a otro e identificar posibles oportunidades de negocio.

¿Quién lo utiliza?

La minería de datos participa en diversas industrias y disciplinas.

Seguros

Pueden resolver problemas complejos concernientes a:

- **Fraude,**
- **Cumplimiento,**
- **Gestión de riesgo**
- **Clasificación de clientes,**
- **Asignar precios a productos con mayor eficacia**
- **Ofrecer productos competitivos a sus clientes existentes.**

¿Quién lo utiliza?

La minería de datos participa en diversas industrias y disciplinas.

Educación

Permite mejorar el sistema educativo :

- Progreso de los estudiantes,
- P**redecir el desempeño** de sus alumnos
- **Desarrollar estrategias de intervenció**n para mantenerlos en curso,
- Predecir niveles de logro de objetivos
- Detectar estudiantes que necesitan atención extra

¿Quién lo utiliza?

La minería de datos participa en diversas industrias y disciplinas.

Manufactura

Permite mejorar a las organizaciones en:

- Alineación de planes de suministro con **pronósticos de demanda**,
- Detección temprana de problemas,
- Garantía de calidad
- Inversión en equidad de marca**,**
- **Predecir el desgaste de activos** de producción
- Anticipar su mantenimiento

¿Quién lo utiliza?

La minería de datos participa en diversas industrias y disciplinas.

Bancos

Permite mejorar a las organizaciones Financieras en:

- Entender a su base de clientes,
- Tener una mejor vista de los **riesgos del mercado,**
- **Detectar el fraude** de manera inmediata,
- Gestionar las obligaciones de **cumplimiento** de las regulaciones
- Obtener **retornos óptimos** de sus inversiones en marketing

¿Quién lo utiliza?

La minería de datos participa en diversas industrias y disciplinas.

Retail

Grandes bases de datos de clientes contienen insights ocultos que le pueden ayudar a:

- M**ejorar las relaciones con clientes,**
- **Optimizar campañas de marketing**
- **Pronosticar ventas**
- **E**ncontrar la oferta que tenga el mayor impacto en el cliente

Minería de Datos

Proceso de Minería de Datos

Minería de Datos - Proceso

Business Responsible

Caso de Negocio

Business Processes

Data Manager

Business Analyst

Data Scientist

Data Scientist

Data Manager

Business Rep.

Entendimiento del Dominio de Negocio

Identificación de los Datos Relevantes

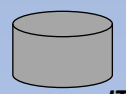
Preparación de Datos
- Limpieza
-Transformación

Identificación de Tareas de Minería de Datos

Implementación de Algoritmos

Interpretación y Evaluación de Datos

Datos Históricos



IT / Data manager

Datos Actuales

MINERIA DE DATOS

Descubrimiento del conocimiento en datos

Entendimiento del Dominio del Negocio

CASOS DE NEGOCIO

Clientes y Marketing

- Lanzamiento de nuevos productos
- Promociones correctas y oportunas
- Sentimiento de marca y ventas
- Análisis de compra
- Valor de tiempo de vida
- Optimización de precios

Cadena de Suministro

- Predicción de stock
- Previsión de la demanda
- Planificación de inventario y logística

Retos de Negocio

Operaciones

- Detección de anomalías
- Segmentación de entregas
- Previsión de uso
- Pronóstico de KPI
- Optimización de tamaño y zona
- Predicción de cuota de mercado

Fraude / Riesgos

- Calificación crediticia
- Gestión y Prevención del Fraude
- Optimización de la calidad

Finanzas

- Cash Flow y Forecasting
- Simulación de presupuestos
- Análisis de rentabilidad y margen

Entendimiento del Dominio de Negocio

Propósito de Negocio establece objetivos en terminología empresarial.

Propósito de Minería de Datos establece los objetivos del proyecto en términos técnicos

Crear un plan que permita lograr las metas de Minería de Datos y lograr los objetivos de Negocio

Entendimiento del Dominio de Negocio

Objetivo

- Comprender los objetivos y requerimientos del proyecto desde la perspectiva de negocio.
- Identificar el caso de Minería de Datos y crear un plan que permita cumplir los objetivos.

Actividades principales

- Definir los objetivos de negocio.
- Determinar la situación actual.
- Determinar las metas de Minería de Datos.
- Crear el plan de proyecto
- Criterios de éxito del proceso de Minería de Datos

Preguntas sobre Negocio

Iniciar preguntando:
¿Cuáles son los principales objetivos y metas de su negocio en este momento?
¿Cuáles son los desafíos más importantes que enfrenta actualmente su negocio?
Información del Cliente:
¿Quiénes son sus clientes principales/ Proveedores y sus comportamientos de compra?
¿Cómo obtienen información de sus clientes?
Datos Disponibles:
¿Qué tipo de datos tienen actualmente disponibles (ventas, inventario, datos de clientes, etc.)?
¿Cómo se utilizan estos datos?

Preguntas técnicas

Infraestructura:
¿Qué tipo de infraestructura tecnológica tienen implementada para la gestión de datos (servidores, bases de datos, sistemas de gestión de inventario, etc.)?
¿Tienen personal capacitado en análisis de datos o minería de datos?

Frecuencia y Volumen de Datos:
¿Con qué frecuencia se actualizan y se generan nuevos datos?
¿Cuál es el volumen de datos que manejan en su operación diaria?

Integración y Herramientas:
¿Qué herramientas y sistemas utilizan actualmente que necesitarían integrarse con nuevas soluciones de minería de datos?
¿Están dispuestos a invertir en nuevas tecnologías o software si es necesario?

Expectativas y Resultados

1. **Expectativas del Proyecto**:
    1. ¿Qué resultados esperan obtener al implementar minería de datos en su negocio?
    2. ¿Cuáles son sus expectativas en términos de tiempo y recursos dedicados al proyecto?
2. **Medición del Éxito**:
    1. ¿Cómo medirán el éxito del proyecto de minería de datos?
    2. ¿Tienen algún KPI o métrica específica que les gustaría mejorar?
3. **Presupuesto y Recursos**:
    1. ¿Cuál es el presupuesto disponible para este proyecto?
    2. ¿Tienen algún personal o recursos dedicados que podrían colaborar en este proyecto?

Puntos clave a validar

1. **Sobre la Información de negocio**

- **Calidad de Datos**: Asegúrate de que los datos disponibles son precisos, completos, consistentes y actualizados.
- **Disponibilidad de Datos**: Verifica que los datos necesarios estén disponibles y accesibles. Esto incluye datos históricos, transaccionales, de inventario, y de clientes.

2. **Infraestructura Tecnológica**

- **Capacidad de Almacenamiento y Procesamiento**: Asegúrate de que la infraestructura tecnológica actual puede manejar el volumen y la velocidad de los datos. Esto puede implicar la necesidad de actualizar sistemas o invertir en nuevas tecnologías.
- **Herramientas y Software**: Evalúa las herramientas y software de minería de datos disponibles. La elección de herramientas adecuadas es crucial para el éxito del análisis.

Puntos clave a validar

3. Definición Clara de Objetivos del Proyecto de minería de datos. ¿Qué problemas específicos intentan resolver? ¿Qué resultados esperan obtener? Establece indicadores clave de rendimiento (KPI) y métricas para medir el éxito del proyecto. Esto permitirá evaluar el impacto de las soluciones implementadas.

4. Capacitación y Competencias

Habilidades del Personal: Asegúrar que el personal clave tenga las habilidades necesarias para utilizar herramientas de minería de datos y analizar los resultados.

Colaboración entre diferentes departamentos (TI, marketing, ventas, etc.) para asegurar una implementación efectiva y una interpretación correcta de los datos.

5. Consideraciones Éticas y de Privacidad

•**Protección de Datos**: Cumple con las regulaciones de protección de datos y privacidad,

•**Transparencia y Consentimiento** con los clientes sobre cómo se utilizan sus datos.

MINERIA DE DATOS

Introducción al Descubrimiento del conocimiento en datos

Identificación de Datos Relevantes

PROCESO DE MINERIA DE DATOS

Identificación de los datos relevantes

Entendimiento de Datos

Objetivo

Recolectar, seleccionar y entender los datos para identificar hallazgos iniciales a partir de los datos, identificar hipótesis candidatas y problemas de calidad de datos

Actividades principales

- Recolectar datos iniciales.
- Describir los datos.
- Explorar los datos.
- Valorar la calidad de los datos.

Preparación de Datos

Objetivo

Decidir sobre los datos que se utilizarán para el análisis, criterios que incluyan la relevancia para los objetivos de Minería de Datos y las limitaciones técnicas como el volumen o los tipos de datos.

Actividades principales

- Creación de atributos derivados
- Valores transformados para atributos existentes combinados a partir de varias fuentes de datos
- Crear datasets de entrenamiento y prueba

Identificación de los datos relevantes

Recolectar Datos Iniciales

- Cargar los datos en distintos entornos de almacenamiento para ser procesados o analizados efectivamente en cada etapa.
- Los destinos de datos pueden incluir:
  - Contenedores de archivos locales o en la nube.
  - Bases de datos relacionales.
  - Repositorios NoSQL.
  - Almacenes Hadoop.
- Los factores decidir el destino de datos incluyen:
  - Repositorio origen de los datos.
  - Tamaño del dataset.
  - Requisitos de muestreo (downsampling) del dataset

Identificación de los datos relevantes

¿Cuáles son las fuentes y el destino de datos?

¿Cómo se van a mover los datos entre la fuente y el destino, con qué frecuencia, se requiere de actualizaciones?

¿Cuáles son las características de los datos?

- Tipos de datos.
- Formato de los archivos de datos.
- Volumen de datos, Historia que se tiene de los datos.

¿Cuál es la situación actual en cuanto a calidad de datos?

- Patrones, tendencias, *outliers*, valores faltantes.

¿Las herramientas previstas soportan los formatos de datos origen o se requieren conversiones?

¿Se requiere de preprocesamiento o limpieza de los datos origen?

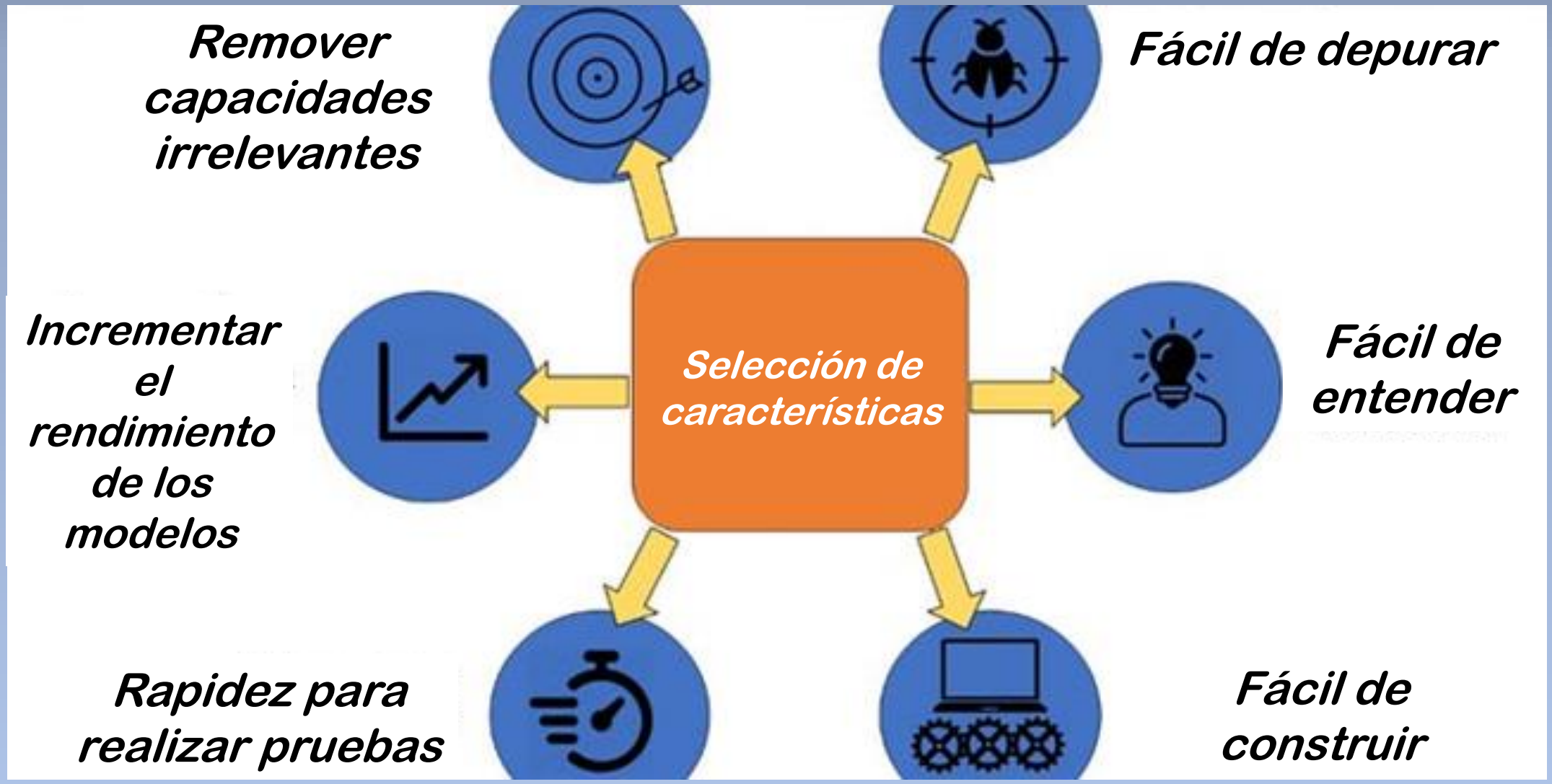
Identificación de los datos relevantes

Muestreo : Si el dataset que se planea analizar es grande, se recomienda generar una versión reducida y representativa que sea más manejable. Esto facilita los procesos de entendimiento de datos, exploración e ingeniería de características.

Tipo de muestreo

- **Muestreo uniforme aleatorio**: Cada renglón del dataset tiene la misma probabilidad de formar parte de la muestra.
- **Muestreo aleatorio por grupos**: Permite incluir o excluir todas las instancias de un valor categórico.
- **Muestreo estratificado**: El muestreo aleatorio es estratificado con respecto a una variable categórica cuando las muestras tienen valores de dicho categórico que se encuentran en la misma proporción que la población.

Identificación de los datos relevantes – Capacidades



MINERIA DE DATOS

Introducción al Descubrimiento del conocimiento en datos

Limpieza de Datos

Preparación de Datos - Limpieza de datos - Calidad de los Datos

La limpieza de datos es un paso importante en el proceso de minería de datos para garantizar que los análisis y modelos resultantes sean precisos y fiables.

Factores a considerar

- Número de registros.
- Número de atributos o características.
- Tipos de datos.
- Número de valores faltantes.
- Presencia de *outliers* (no son necesariamente un problema).
- Datos bien formateados.
- Consistencia de los registros de datos.
- Datos duplicados.

Consideraciones

- Generar reporte de calidad de datos.
- Planificar la remediación durante la etapa de preparación de datos.
- Verificar la calidad de datos en relación a si cubre los casos requeridos.

Limpieza de datos - Recomendaciones importantes

- **Comprender los Datos:**
  - Familiarízate con el contexto y origen de los datos. Esto te ayudará a identificar valores anómalos o inconsistencias.
  - Realiza un análisis exploratorio para identificar patrones, distribuciones, y valores atípicos.
- **Manejo de Datos Faltantes:**
  - Detecta valores faltantes y decide cómo manejarlos.
  - Considera técnicas de imputación como media, mediana, moda o métodos más avanzados como k-NN, regresión, o imputación múltiple.
  - Si la cantidad de datos faltantes es significativa o si ciertos registros no son útiles, puede ser mejor eliminarlos.
- **Detección y Corrección de Valores Atípicos**:
  - Usa métodos estadísticos (como IQR o z-score) o visualizaciones (boxplots) para detectar outliers.
  - Decide si los outliers son errores o si representan información valiosa. Puedes eliminarlos, corregirlos, o tratarlos por separado.

Limpieza de datos - Recomendaciones importantes

- **Estandarización y Normalización:**
  - Asegurarse de que todas las unidades de medida y formatos sean consistentes a lo largo del dataset.
  - Si es necesario, aplica técnicas de normalización (min-max scaling) o estandarización (z-score) para que los datos sean comparables.
- **Corrección de Errores:**
  - Detecta y corrige errores tipográficos, de formato o duplicados en los datos.
  - Verifica la coherencia de los datos comparando diferentes fuentes o utilizando reglas de negocio.
- **Manejo de Duplicados:**
  - Busca registros duplicados que pueden estar afectando la calidad del análisis.
  - Elimina o combina registros duplicados de manera adecuada.

Limpieza de datos - Recomendaciones importantes

- **Transformación y Reducción de Dimensionalidad:**
  - Transforma variables categóricas (por ejemplo, usando one-hot encoding).
  - Considera técnicas de reducción de dimensionalidad (como PCA) si tienes un número muy alto de variables, para simplificar el modelo.
- **Documentación y Automatización:**
  - Documenta cada paso del proceso de limpieza para poder replicarlo o revisarlo.
  - Usa scripts o herramientas para automatizar partes del proceso de limpieza, garantizando consistencia y ahorrando tiempo.
- **Revisar Regularmente:**
  - La limpieza no es un proceso único; revisa y ajusta periódicamente tus datos conforme cambien o se amplíen.

Preparación de Datos - Limpieza de datos - Valores Faltantes

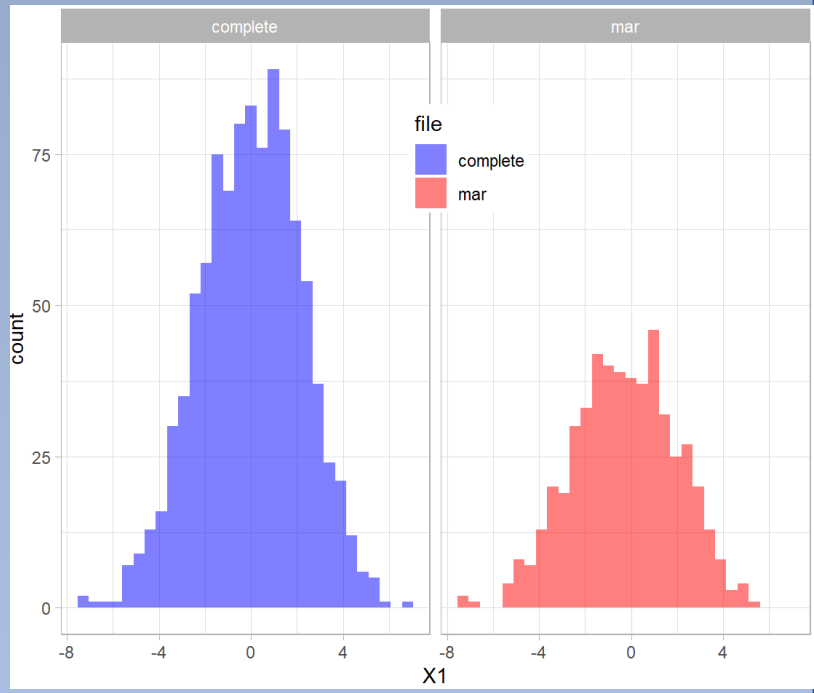
- La estrategia de manejo de los datos faltantes (*missing data*) afecta dramáticamente los resultados de los modelos
- Se recomienda intentar diversos métodos y considerar tanto la justificación para el uso de un método en particular y la calidad de los resultados.
- Conocer el significado del dato es imperativo para determinar la estrategia.

Estrategias de manejo:

– Eliminar los objetos de datos (renglones o columnas).
– Estimar los valores faltantes y sustituirlos
– Ignorar los valores faltantes durante el análisis.
– Sustituir por los valores posibles (ponderados por su probabilidad).

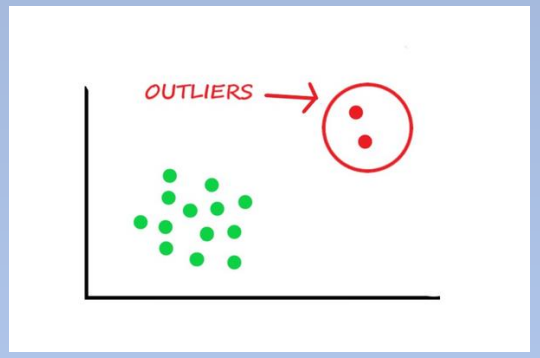
Preparación de Datos - Limpieza de datos - Valores Faltantes

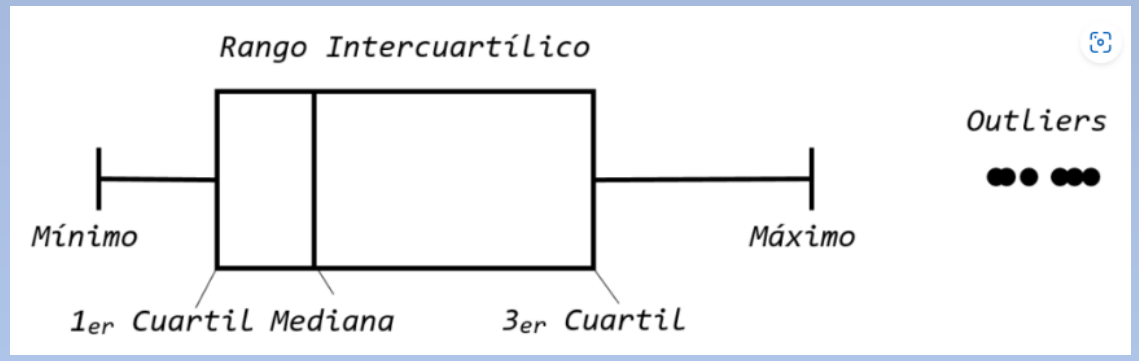
- Si la estimación depende de la suposición de que el patrón de datos faltantes está relacionado únicamente con los datos observados, se denomina falta al azar (MAR).
- Esta hipótesis permite ajustar las estimaciones utilizando la información disponible.
- Por ejemplo, en un estudio de educación e ingresos, los estudiantes con bajo nivel educativo pueden tener más valores de ingresos faltantes.
- En este caso, la probabilidad de que se registren los ingresos depende del nivel de educación del estudiante.
- La probabilidad puede variar según la educación, pero no según los ingresos dentro de ese nivel de educación.

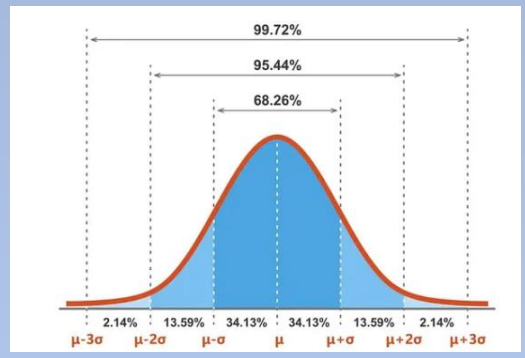


Preparación de Datos - Limpieza de datos - Outliers

- *En estadística, un valor atípico es una observación que está numéricamente distante del resto de los datos.*
- *Los valores atípicos pueden producirse debido a errores de medición y pueden eliminarse del conjunto de datos o corregirse.*
- *Pueden ocurrir naturalmente y, por lo tanto, deben tratarse con cuidado.*
- *Algunas estadísticas / algoritmos pueden estar muy sesgados por valores atípicos.*
- *Los valores atípicos se pueden detectar visualmente, por ejemplo, diagramas de dispersión y diagramas de caja*. -







Preparación de Datos - Limpieza de datos - *Outliers*

***Outlier*:** Elemento de datos que no corresponde con el comportamiento general de los datos. Puede ser considerado como ruido o como una excepción, sin embargo son de gran utilidad en la detección de fraudes y el análisis de eventos atípicos

Causas:

–Datos erróneos: Errores de lectura en sensores, error humano, error de software.

–Datos no representativos : Datos reales de no interés para el estudio.

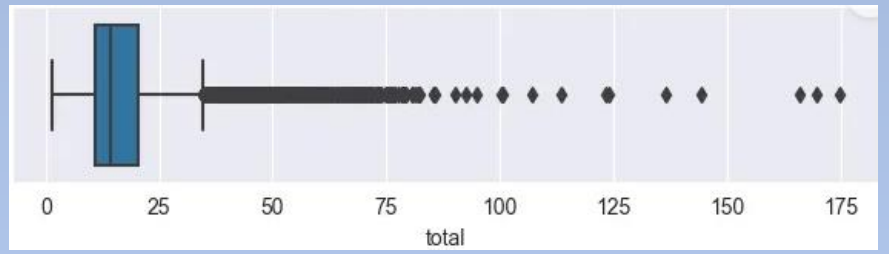
- Se tiene que proveer de un argumento legítimo para considerar un *outlier*.

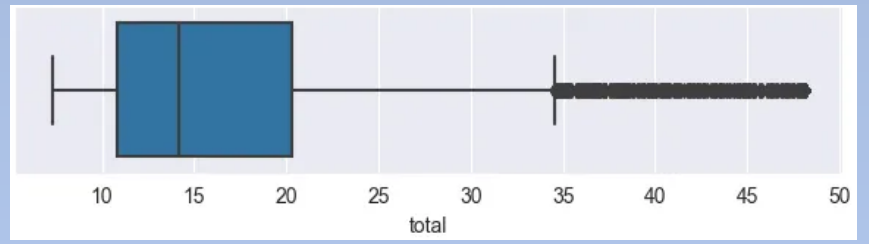
Consideraciones:

–Identificación: Probar si una observación es *outlier* o no.

–Gestión: Emplear técnicas estadísticas robustas para lidiar con los *outliers*.

–Resolución: Corregir los datos para que no tenga *outlier y no in*fluya en las conclusiones estadísticas.





Preparación de Datos - Limpieza de datos **- Datos Duplicados**

El dataset puede incluir objetos de datos que se estén duplicados, o sus valores sean muy cercanos a otros objetos de datos. Esto es un problema frecuente cuando se combinan datos de múltiples fuentes

Escenarios:
–Datos redundantes en la fuente.
–Datos redundantes enviados recurrentemente.
–Replicación de información como estrategia de prevención en manejo de desastres.

Consideraciones:
–Remediación (des duplicación) en la fuente.

Preparación de Datos - Limpieza de datos **- Selección de Datos**

Elegir los datos que van a ser empleados para el análisis. Los criterios incluyen relevancia para las metas del proyecto de ciencia de datos, calidad y restricciones técnicas tales como límites en el volumen de datos o tipos de datos.

Escenarios:
–Eliminar predictores que tengan varianza cercana a cero.
–Eliminación de predictores que tengan una alta correlación.
–Centrar y escalar los predictores.
–Datos redundantes en la fuente.
–Datos redundantes enviados recurrentemente.

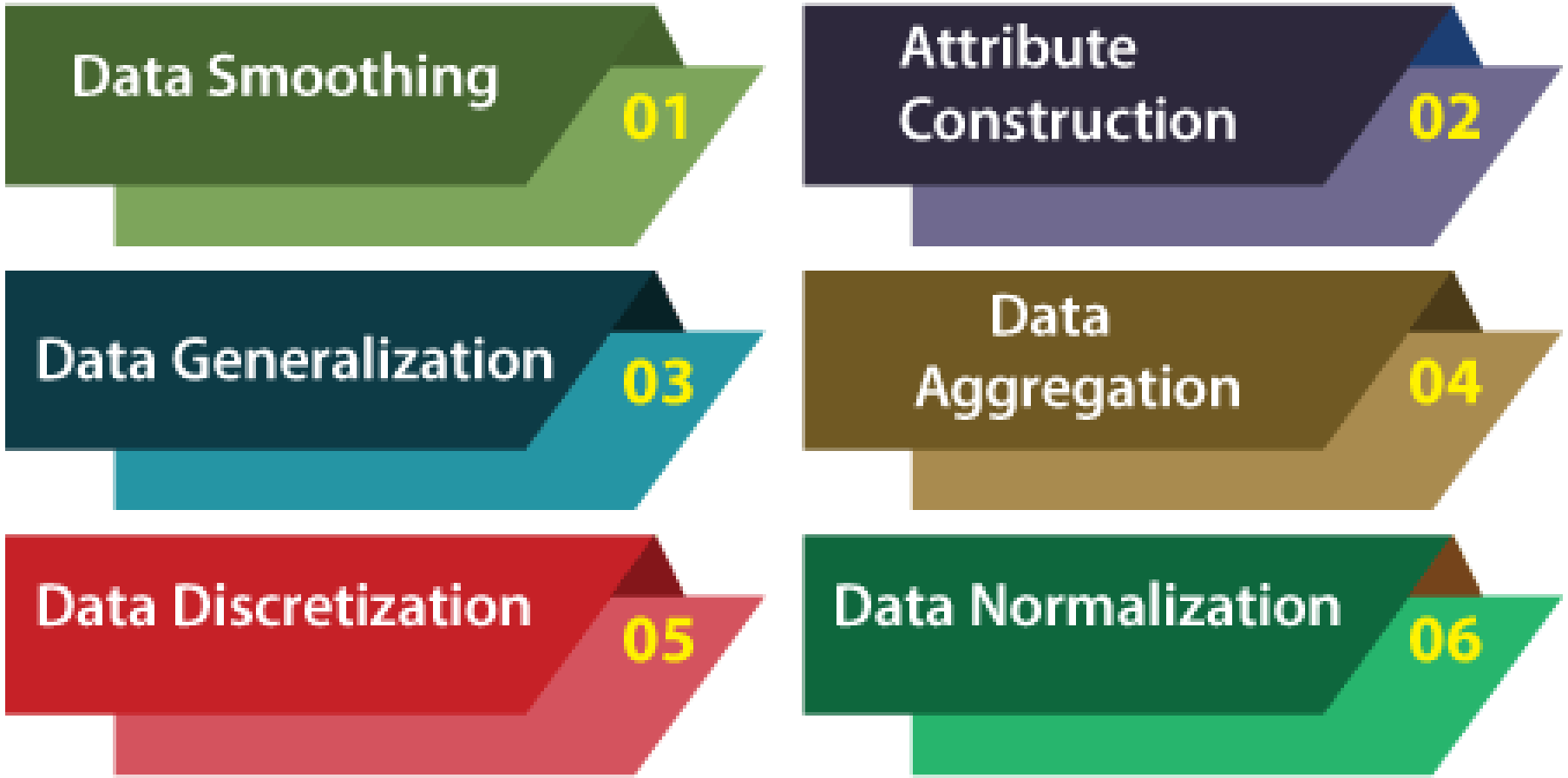
Consideraciones:
–Remediación (des duplicación) en la fuente.
–Remediación en el destino
--Sincronización con la fuente de datos origen.

MINERIA DE DATOS

Introducción al Descubrimiento del conocimiento en datos

Transformación de Datos

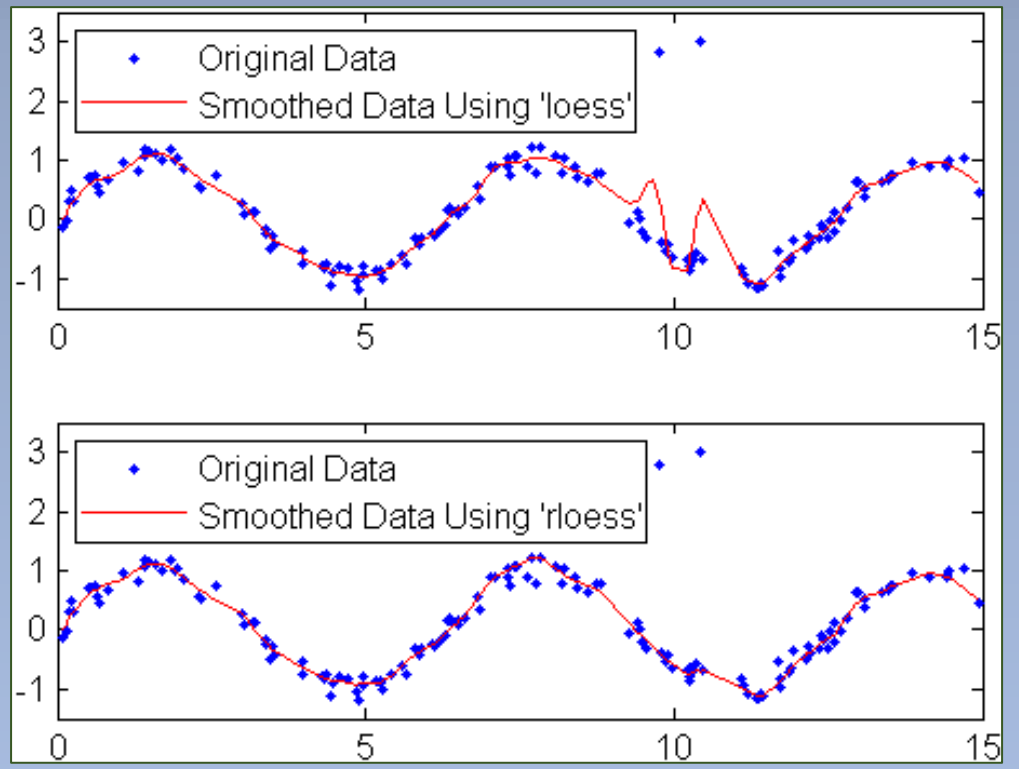
Data Transformation Techniques



Preparación de Datos - Transformación de datos

Data Smoothing

El Suavizado de Datos se realiza mediante el uso de un algoritmo para eliminar el ruido de un conjunto de datos. Esto permite que los patrones importantes se destaquen más claramente.



Preparación de Datos - Transformación de datos

Data Smoothing

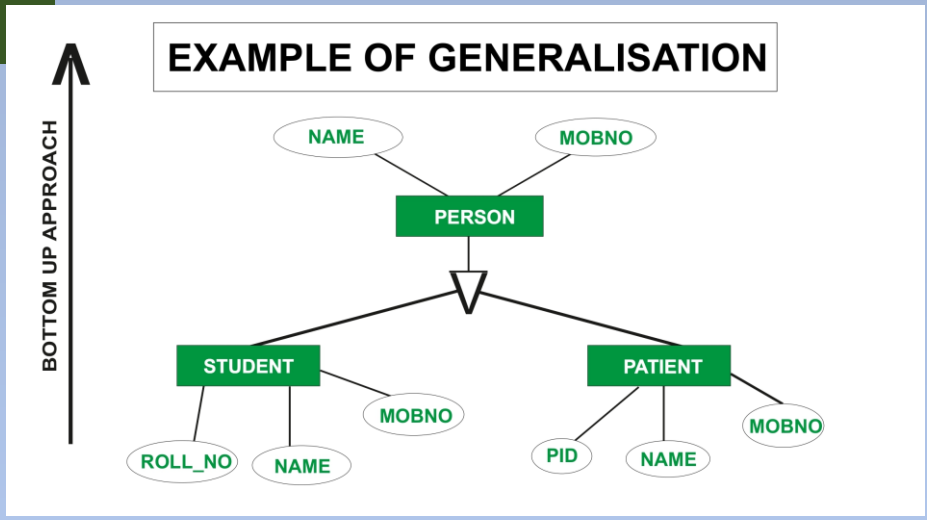
Attribute Construction

Construcción de Atributos es construir nuevos atributos a partir de los originales, transformando la representación de datos original en una nueva donde las regularidades en los datos se detectan más fácilmente mediante el algoritmo de clasificación.

Preparación de Datos - Transformación de datos

- Data Smoothing
- Attribute Construction
- Data Generalization

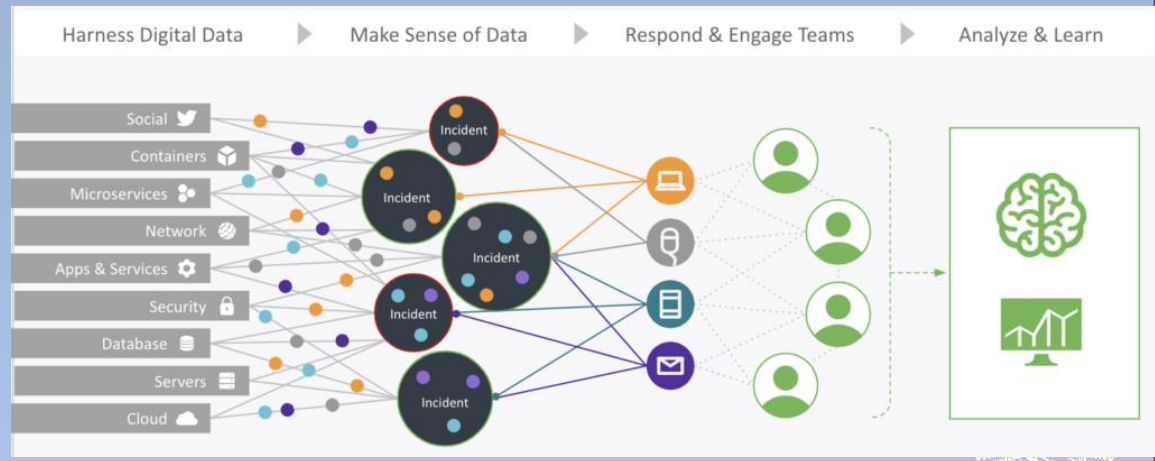
Generalización de Datos, también conocida como resumen de datos o compresión de datos, es el proceso de reducir la complejidad de grandes conjuntos de datos al identificar y representar patrones en los datos de una forma más simplificada.



Preparación de Datos - Transformación de datos

- Data Smoothing
- Attribute Construction
- Data Generalization
- Data Aggregation

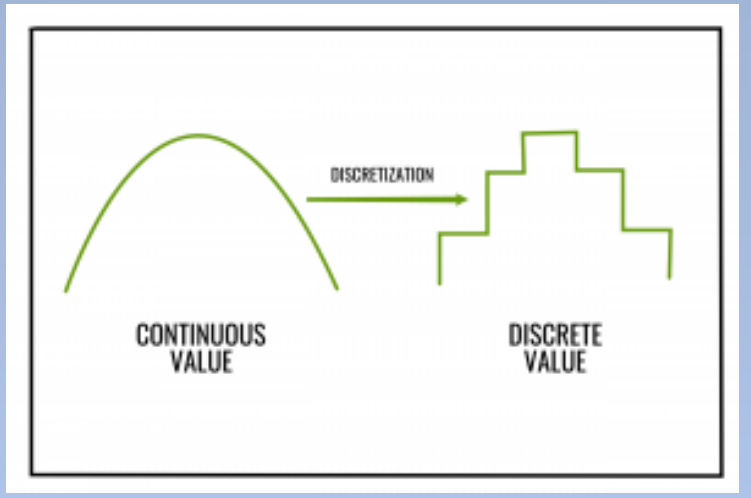
Agregación de Datos es el proceso en el que los datos se recopilan y presentan en un formato resumido para el análisis estadístico y para lograr los objetivos comerciales de manera efectiva.



Preparación de Datos - Transformación de datos

- Data Smoothing
- Attribute Construction
- Data Generalization
- Data Aggregation
- Data Discretization

La Discretización de Datos es un método que convierte los valores de atributos de datos continuos en una colección discreta de intervalos mientras minimiza la cantidad de datos que se pierden en el proceso.



Preparación de Datos - Transformación de datos

- Data Smoothing
- Attribute Construction
- Data Generalization
- Data Aggregation
- Data Discretization
- Data Normalization

La Normalización de Datos es una técnica utilizada en la minería de datos para transformar los valores de un conjunto de datos en una escala común.

person_name	Salary	Year_of_ experience	Expected Position Level
Aman	100000	10	2
Abhinav	78000	7	4
Ashutosh	32000	5	8
Dishi	55000	6	7
Abhishek	92000	8	3
Avantika	120000	15	1
Ayushi	65750	7	5

The attributes salary and year_of_experience are on different scale and hence attribute salary can take high priority over attribute year_of_experience in the model.

Preparación de Datos - Transformación de datos

Predictores con Varianza Cero (o cercana a)

Concepto

La mayoría de los modelos requieren que cada predictor tenga al menos dos valores únicos.

¿Por qué?

–Un predictor con solo un valor único tiene una varianza de cero y no tiene información acerca de la respuesta.

– En general es recomendado eliminarlos.

Adicionalmente, si las distribuciones de los predictores es muy dispersa

– Puede tener un efecto negativo en la estabilidad de la solución del modelo.

–Podrían incluirse descriptores con varianza cercana a cero durante el muestreo

Preparación de Datos - Transformación de datos

- Combinación de Variables: Crea nuevas variables combinando las existentes. Por ejemplo, si tienes peso y altura, crear el Índice de Masa Corporal (IMC).
- Agregaciones Temporales: Si tienes datos temporales, puedes crear agregaciones como periodos específicos (diario, semanal, mensual).
- Reducción de Dimensionalidad:.
- Transformación de Variables Categóricas:
- Transformaciones de Tiempo: hora del día, día de la semana, o estacionalidad de los datos temporales. estacional.
- Clusterización: Agrupa datos en clusters usando algoritmos
- Transformaciones Textuales:, convierte el texto en una representación numérica Análisis de Sentimiento
- Integración de Datos Externos como datos demográficos, económicos, o del clima que puedan proporcionar contexto adicional.
- Transformaciones Espaciales: crea variables adicionales como distancia a ciertos puntos de interés, densidad de población,

Preparación de Datos - Transformación de datos

Modelado

Objetivo

Aplicar diversas técnicas de modelado, calibrando sus parámetros a valores óptimos. Ajustar la estructura del dataset de acuerdo a los requisitos de la técnica.

Actividades principales

- Seleccionar las técnicas de modelado.
- Crear el diseño de pruebas.
- Construir el modelo.
- Validar el modelo.

Preparación de Datos - Transformación de datos

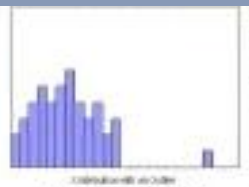
Incluir Variables Derivadas

El proceso de incluir variables derivadas está basado en la combinación de otras variables que están incluidas en los datos, por ejemplo:

- Total de las transacciones y suma de montos
- Numero de meses cuando hay cambios
- Crecimiento en un cierto periodo de tiempo
- Promedios de tiempo atribuidos por el numero de eventos
- Numero de eventos sobre un incidente
- Identificar si un cliente se dio de baja, se añade un indicador de cero / uno
- Una transacción basada en un umbral se identifica como posiblemente fraudulenta

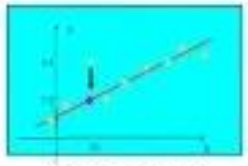
Preparación de Datos - Proceso

Limpieza de datos

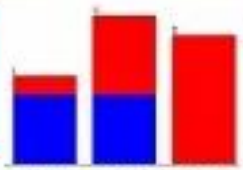


Valores extremos

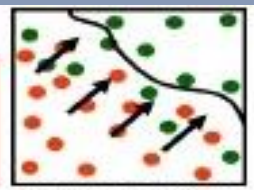
Valores null



Valores con ruido



Valores inconsistentes



Valores mal clasificados

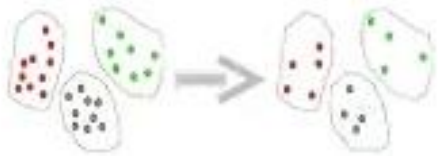
Integración de Datos

Problemas de semántica ¿significa lo mismo?
Problemas de temporalidad ¿cuándo fue registrado?
Problemas de codificación y formato.
Problemas de idioma.
Problemas de duplicidad ¿cuál vale?
Problema de medio de almacenamiento.

Transformación de datos

Filas: suma y agregación de datos
Columnas: log, sin, exp, tan, etc…
Funciones entre columnas.

Reducción de datos



MINERIA DE DATOS

Introducción al Descubrimiento del conocimiento en datos

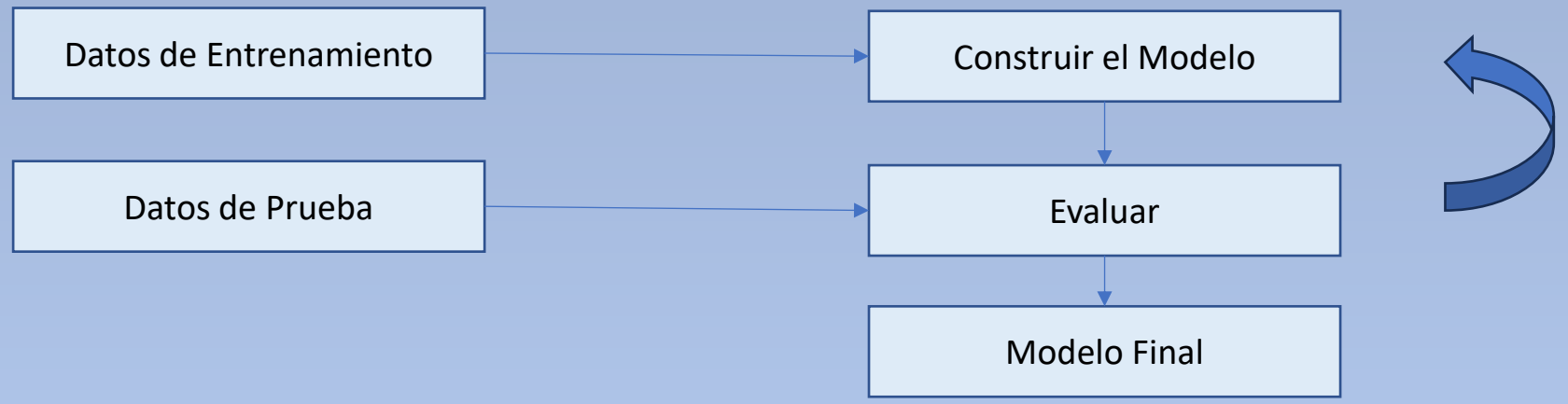
Identificación de Tareas de Minería de Datos

Identificación de tareas de Minería de datos

Crear un conjunto de datos para probar el modelo ⟷ Escoger el modelo predictivo

- Considerar la historia de los datos que estarán participando
- Frecuencia de los resultados
- Actualizaciones
- Nivel de detalle de los datos

- Elegir el algoritmo predictivo *de Clasificación / Redes Neuronales*
- Analizar resultados
- Validar el nivel de confianza resultante del modelo
- Identificar ventajas de la técnica empleada



Estadísticas de destino

Partición de datos	Categoría de destino	Frecuencia
Entrenamiento	1	18,57 %
Entrenamiento	0	81,43 %
Validación	1	17,93 %
Validación	0	82,07 %

Fluctuaciones

Las observaciones pasadas no tienen ningún impacto en la serie cronológica.

Estadísticas de destino

Entrenamiento

Mínimo:
80.804

Máximo:
177.543

Media:
128.150,13

Desviación estándar:
19.760,66

Fluctuaciones

Las observaciones pasadas no tienen ningún impacto en la serie cronológica.

Estadísticas de destino

Validación

Mínimo:
94.149

Máximo:
173.591

Media:
133.291,93

Desviación estándar:
18.454,34

Identificación de tareas de Minería de datos

Ejemplo *Detección de Fraudes*

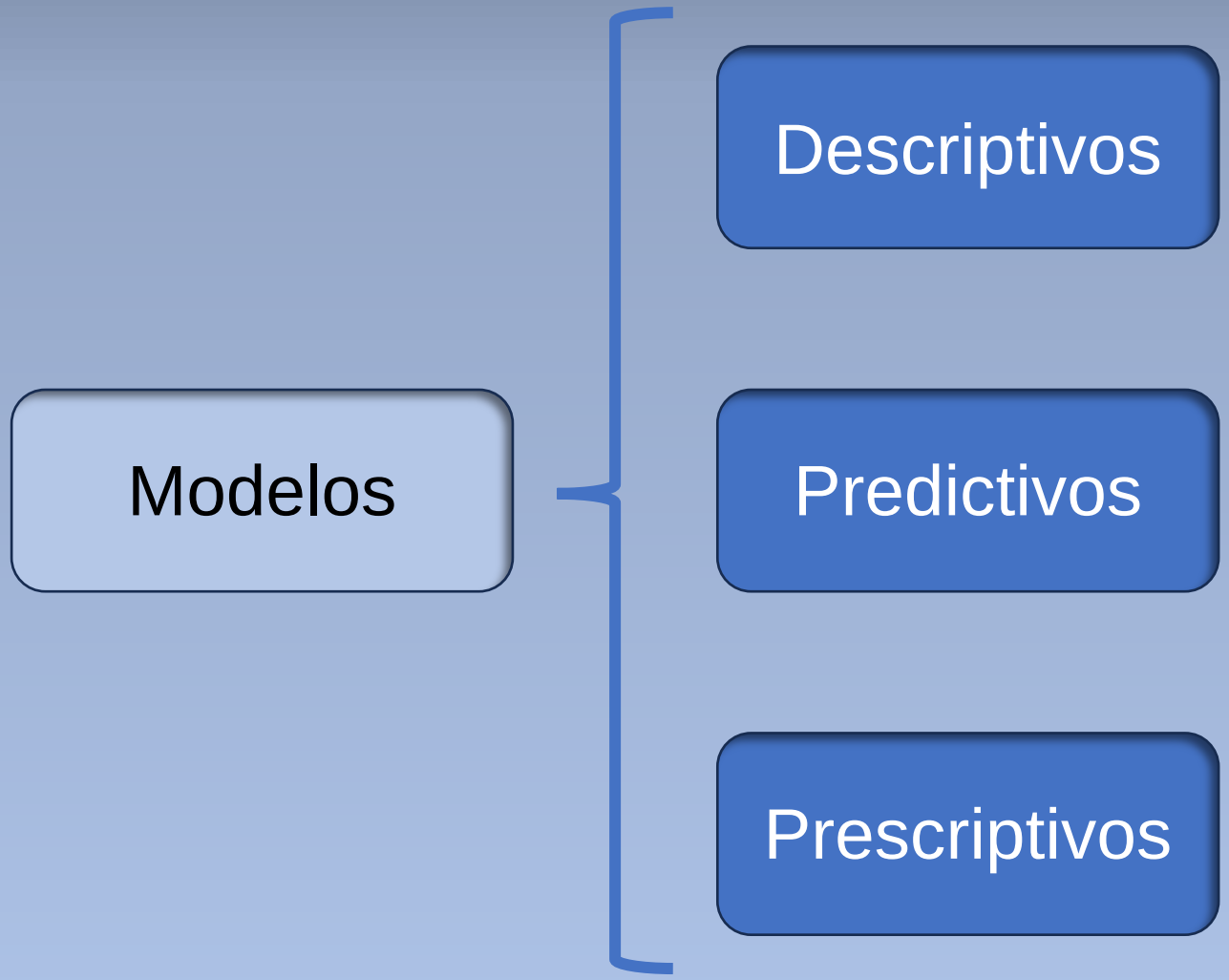
- Identificar el requerimiento de Datos
  - *Transacciones financieras por cliente*
- Obtener los Datos
  - *Historia de operaciones / Transacción realizada*
- Validar, Explorar, y Limpiar los Datos
  - *Seleccionar datos por caso y periodos de tiempo*
- Transformar los Datos
  - *Asociar las operaciones por Cliente / Tipo Operación/….*
- Incluir Variables Derivadas
  - *Incluir nivel de riesgo p.e. Indicador*
- Crear un conjunto de datos para probar el modelo
  - *Seleccionar una Ciudad y tipo de operación*
- Escoger el modelo predictivo
  - *Modelo de Clasificación / Redes Neuronales*

MINERIA DE DATOS

Introducción al Descubrimiento del conocimiento en datos

Implementación de Algoritmos

Implementación de algoritmos de minería de datos



Describir y comprender lo que está pasando.

Big Data

Predecir lo que puede pasar.

Pronostico y Simulación

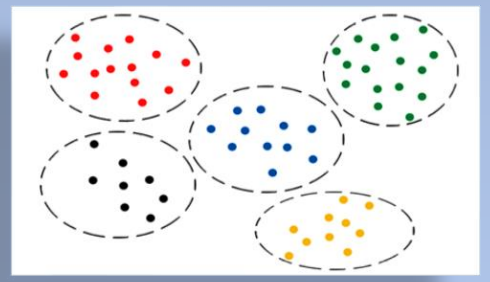
Lo que se debe hacer en el futuro.

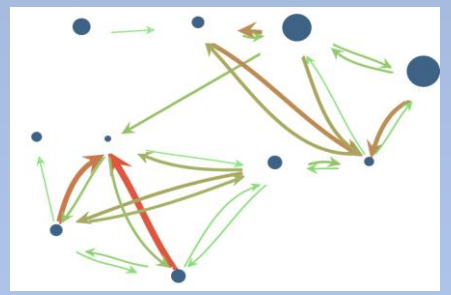
Técnicas de Optimización

Implementación de algoritmos de minería de datos

En Modelos Descriptivos

- El análisis descriptivo describe o resume los datos sin procesar y lo hace más interpretable. Describe los esquemas de asociación y relación entre sus variables.
- Los análisis descriptivos son útiles porque nos permiten aprender de comportamientos pasados y entender cómo estos podrían influir en los resultados futuros.
- Los modelos analíticos descriptivos incluyen modelos de: **Agrupación, Reglas de asociación y Análisis de redes.**

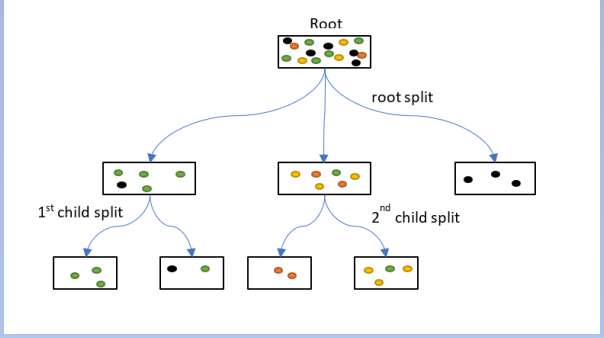


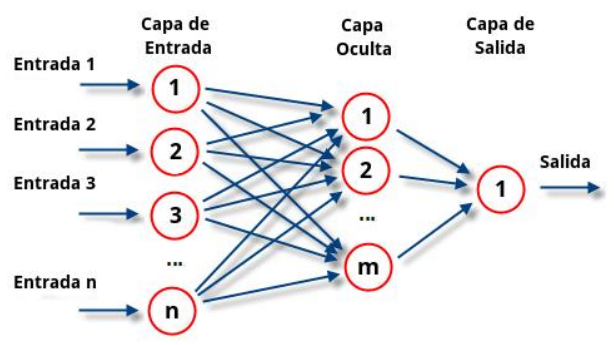


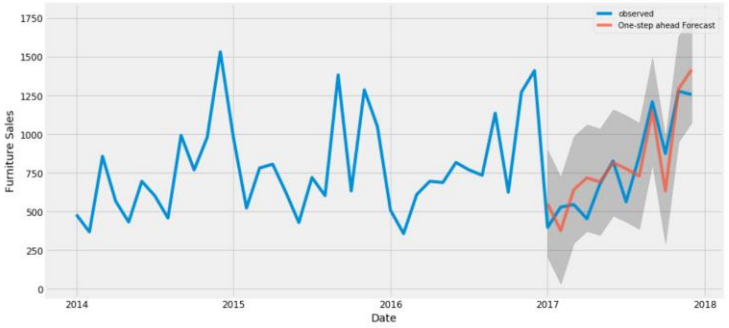
Implementación de algoritmos de minería de datos

En Modelos Predictivos

- El análisis predictivo predice lo que podría suceder en el futuro, proporcionando estimaciones sobre la probabilidad de un resultado futuro.
- Una aplicación común es el uso de análisis para producir un puntaje de crédito. Estas puntuaciones son utilizadas por los servicios financieros para determinar la probabilidad de clientes que realizan pagos de crédito futuros a tiempo.
- Predecir qué artículos compraran los clientes en una misma compra, o pronosticar los niveles de inventario basados en ciertas variables.
- Los modelos analíticos predictivos incluyen: **Modelos de Clasificación, Modelos de Regresión y Modelos de Redes Neuronales, Series de Tiempo**







Implementación de algoritmos de minería de datos

Criterios de éxito para elegir el modelo

Anomalías

¿Qué anomalías o valores inusuales podrían existir?
¿Son errores o cambios reales en el comportamiento?

Tendencias

¿Cuáles son las tendencias, tanto históricas como emergentes, y cómo podrían continuar?

Asociación

¿Cuáles son las correlaciones en los datos? ¿Cuáles son las oportunidades de venta cruzada?

Relaciones

¿Cuáles son los principales influenciadores?
Por ejemplo, la rotación de empleado, abandono de Clientes

Agrupación

¿Existen agrupaciones claras de los datos, Por ejemplo, segmentos de clientes para Campañas de marketing específicas?

Implementación de algoritmos de minería de datos

Selección del Modelo

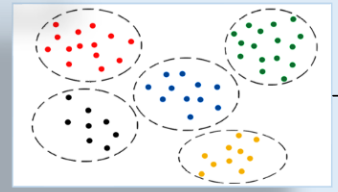
Análisis de Comportamiento

Caso de Negocio

Clientes

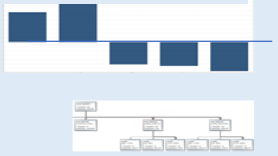
- **Transacciones**
- **Historia**

Cluster Model

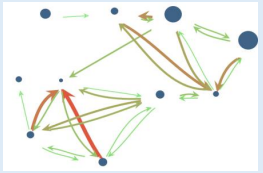


- **Segmentación de Clientes**
- **Perfilamiento de Segmentos**

Classification Model



Association Model



Time Series Model

Social Network Model

Funciones Participativas

- **Variables de Impacto**
- **Árbol de decisión / Correlación**
- **Simulación**

- **Asociación de Productos**
- **Recomendación de Productos**
- **Comportamiento de Productos especiales**

- **Pronostico de Evolución**
- **Identificación de Outliers**
- **Tendencias**

- **Relación Producto “A” <-> Producto “B”**
- **Relación Clientes <-> Productos**
- **Recomendaciones**
- **Identificación de Comunidades**

KDD Descubrimiento del Conocimiento en los Datos

EJEMPLOS

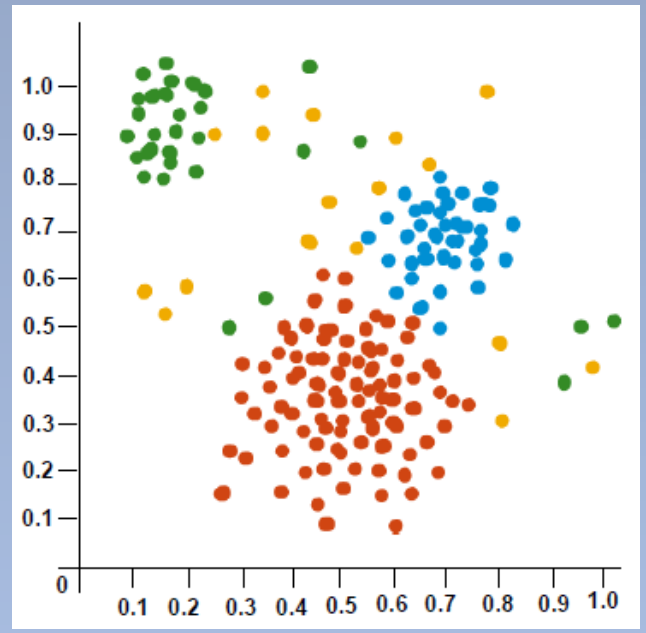
- Análisis de bases de datos de **clientes** para identificar grupos y predecir su comportamiento.
- Análisis de la **canasta de compra** para determinar probabilidades de compra de producto.
- **Administración de Capitales** se empleó sistemas expertos, redes neuronales y algoritmos genéticos para administrar portafolios de inversión
- Sistemas para **monitorear fraude** en tarjetas de crédito en millones de cuentas.
- Sistema para identificar transacciones financieras relacionadas con **lavado de dinero**.
- Sistema de **gestión de errores** para diagnosticar y predecir problemas con el Boeing 737.

Implementación de algoritmos de minería de datos

Segmentación

Propósito
Separar los datos en subgrupos o clases de interés donde todos los integrantes tengan características en común.

Técnicas
- Clustering.(K-Means)
- Redes Neuronales.
- Visualizaciones.

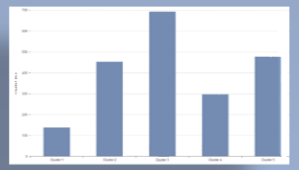


Un modelo de Segmentación requiere que sea:
- Homogéneo en los de miembros dentro de los segmentos
- Heterogéneo entre segmentos
- Estable para que se puede implementar una actividad comercial
- Los segmentos deben tener sentido para el negocio
- Manejable el número y la complejidad de los segmentos

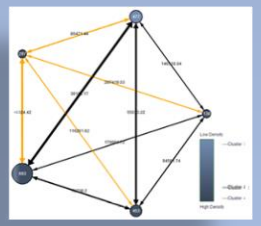
Implementación de algoritmos de minería de datos

Resultados de Segmentación

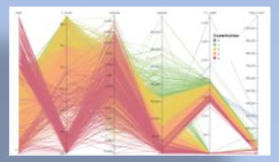
Tamaño de los Segmentos



Densidad de cada segmento y distancia entre ellos



Asociación de las Variables por Segmento



Ejemplo

Marketing del Banco desea crear segmentos de clientes en base a su capacidad potencial o LTV, y desarrollar una campaña que permita potenciar a los Clientes objetivo en base a nuevos productos o servicios

Beneficios:

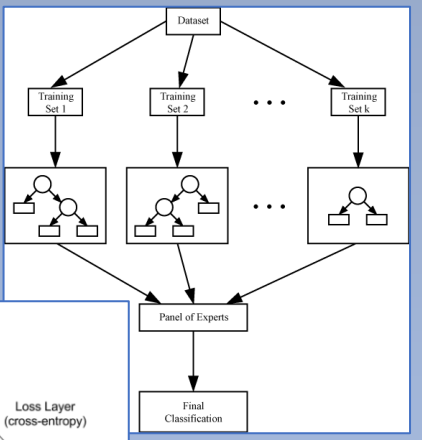
- *Identificar potencial de cada Segmento*
- *Promover nuevos Productos según perfil de cada Segmento*

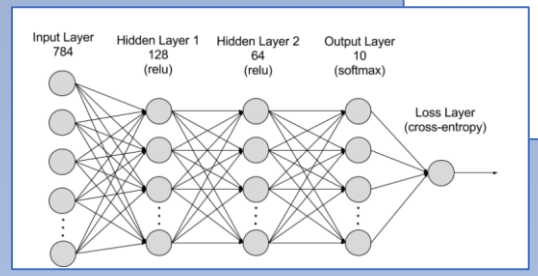
Implementación de algoritmos de minería de datos

Modelo de Clasificación

Propósito

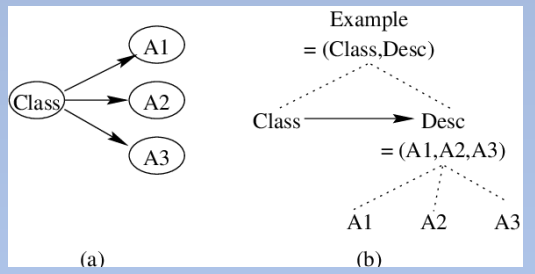
Clasificación se considera una instancia de aprendizaje donde un conjunto de observaciones correctamente identificadas y se tiene definida una variable objetivo (dependiente) , que permite identificar el comportamiento de las variables independientes y su impacto asociado a la variable objetivo.





Técnicas

- Regresión logística.
- Análisis discriminatorio.
- Métodos de inducción de reglas.
- Árboles de decisión.
- Árboles de clasificación.
- Redes neuronales.
- KNN.
- NaiveBayes.
- Support Vector Machines.



Implementación de algoritmos de minería de datos

Modelo de Clasificación – Ejemplo de Churn Analysis

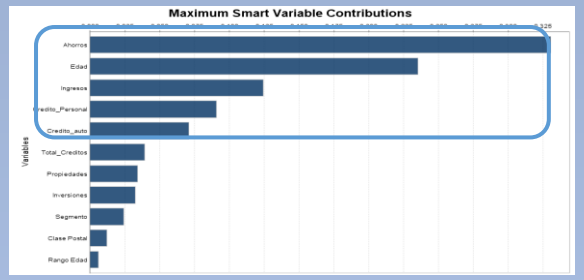
Objetivo:

- Identificar las variables de negocio de mayor impacto en los Clientes que han abandonado el Banco
- Valorar que Clientes con posibilidad de abandono

- Variable objetivo: Clientes que han abandonado
- Historia: 2 años

Variables de Impacto:

- *Ahorros (32%)*
- *Edad (22%)*
- *Ingresos (12%)*
- *Crédito Personal (9%)*
- *Crédito Auto (7%)*



El abandono se da a clientes cuando:su

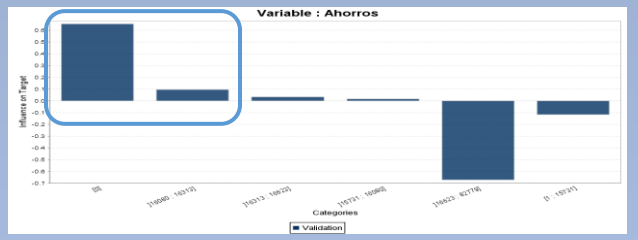
***Saldo en AHORROS** era de:*

- *Sin Saldo ($0) o < 3,000*

***Edad** del Cliente entre:*

18 a 25 años

***Ingresos** entre 20,000 y 40,000*

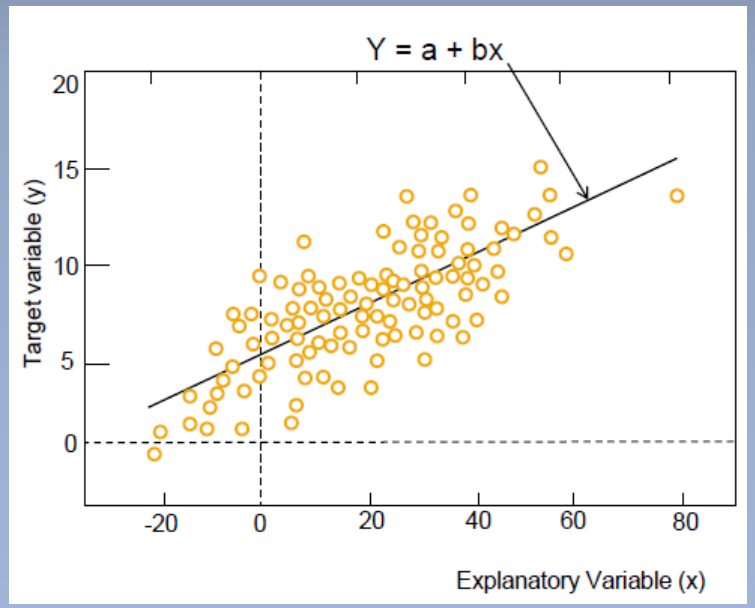


Beneficios de Negocio esperados

- Anticiparse al abandono de Clientes
- Mejorar las propuestas de productos financieros

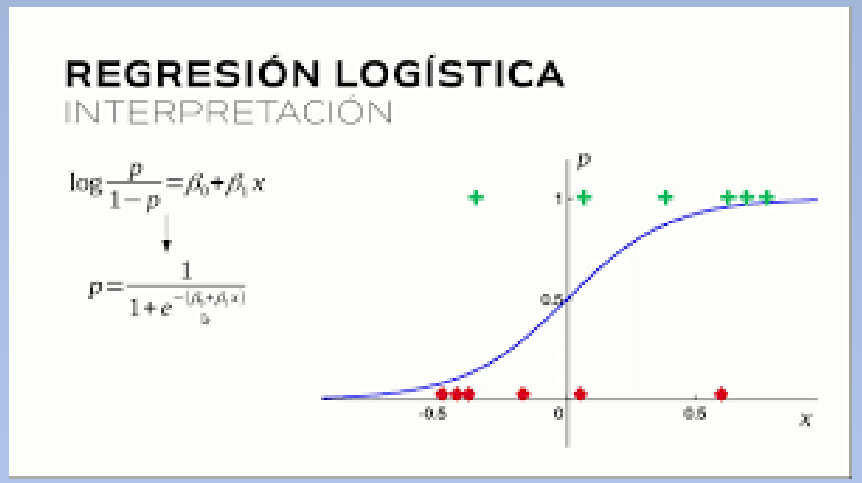
Implementación de algoritmos de minería de datos

Regresión - Predicción



Propósito

- Elemento de la inferencia estadística en el que se recolectan datos acerca de la variable a ser predicha y un conjunto de variables que se cree que tienen influencia sobre ella.
- Se formula la hipótesis de una función que expresa la relación entre la variable dependiente y las variables independientes.

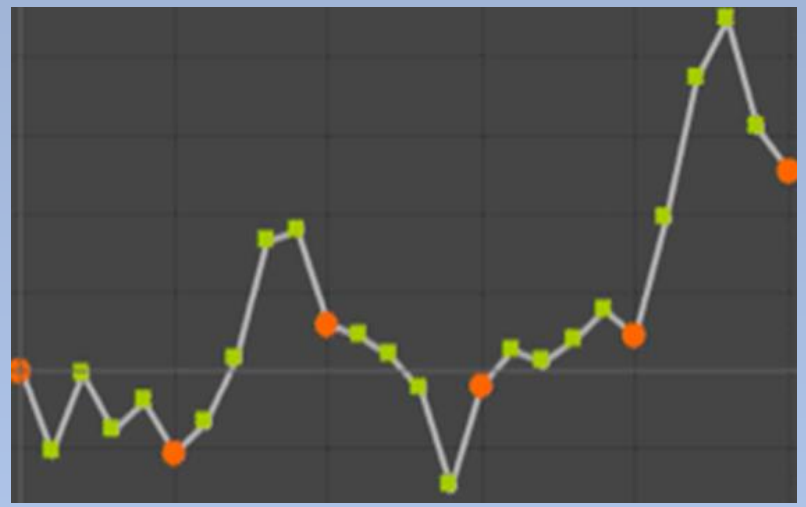


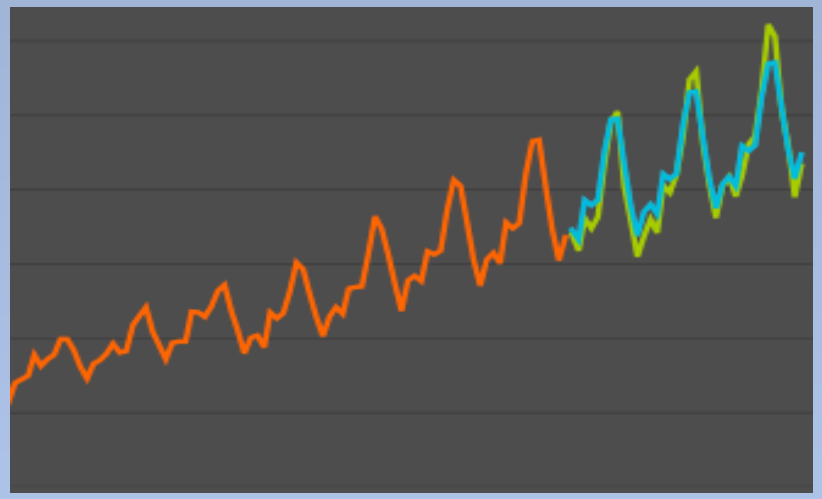
Técnicas

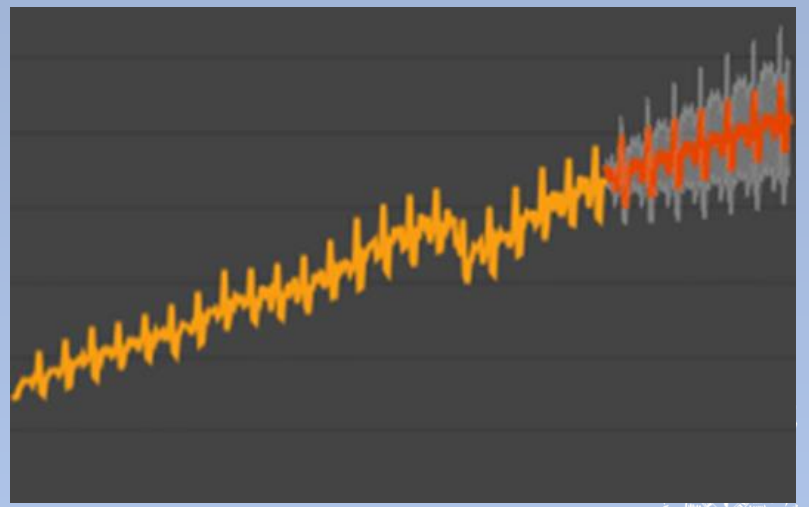
- Bi-Variate Regression Variations
- Polynomial Regression
- Logistic Regression
- Linear Regression

Time Series

- Serie temporal: un conjunto secuencial de puntos de datos medidos en un conjunto de intervalos espaciados regularmente (por ejemplo, semanal, mensual, cada hora)
- Análisis de series temporales: el proceso de ajustar una serie temporal a un modelo con el fin de pronosticar.
- Se utiliza para apoyar a las decisiones empresariales, por ejemplo: plan de Producción, Situación de Inventario, situación de Ventas



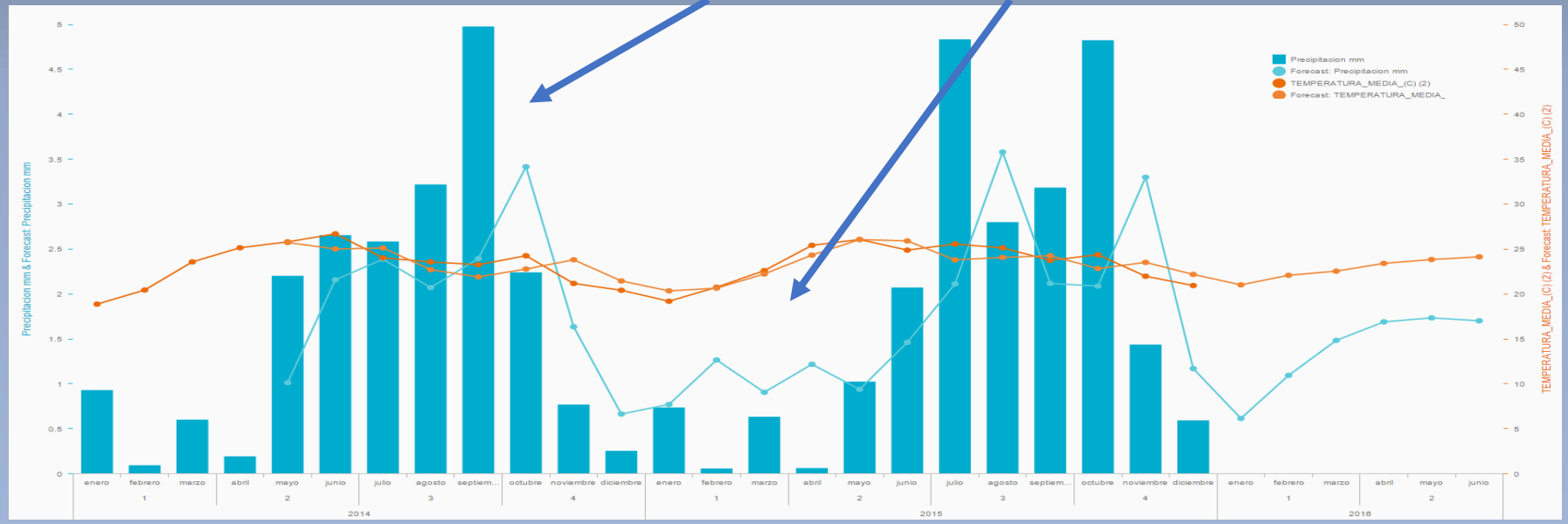




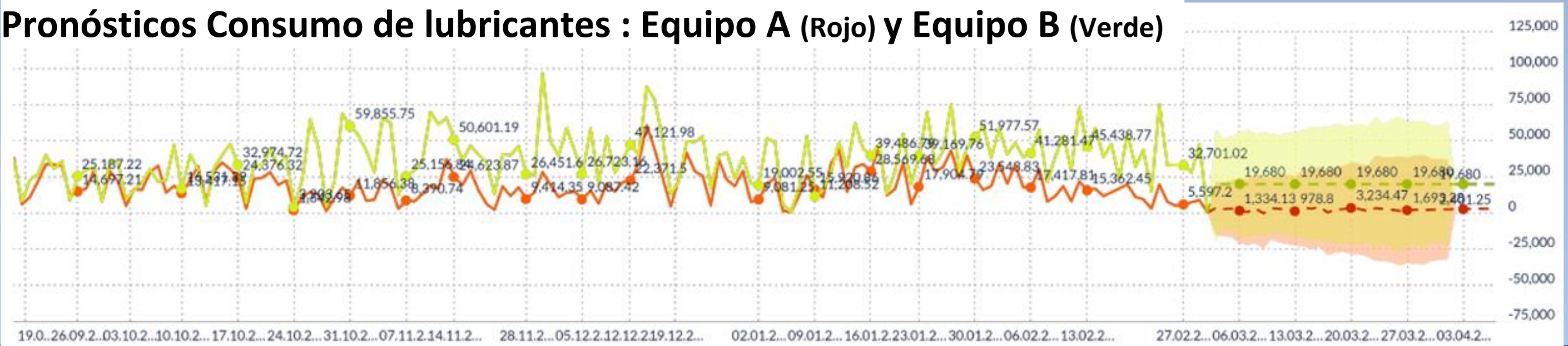
PROCESO DE CIENCIA DE DATOS

Time Series

Precipitación (mm) / Temperatura



Pronósticos Consumo de lubricantes : Equipo A (Rojo) y Equipo B (Verde)



Elección del modelo

Cuáles son los grupos de Clientes con características similares en Comportamiento, Perfil socioeconómico, etc.

Segmentación

Provee recomendaciones sobre sitios Web o para almacenes análisis de canasta (Basket analysis)

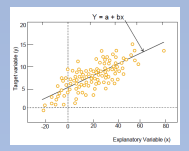
Asociación

Identifica los principales influenciadores en casos como abandono de Clientes, detección de fraude o riesgo crediticio

Clasificación

Se desea realizar un pronóstico para estimar los costos o ventas en los próximos meses.

Series de Tiempo



Relaciona una variable de Negocio (dependiente) con una o más variables de Negocio (independientes).

Regresión

Analiza las interacciones para identificar comunidades, influenciadores, etc.

Análisis de Redes

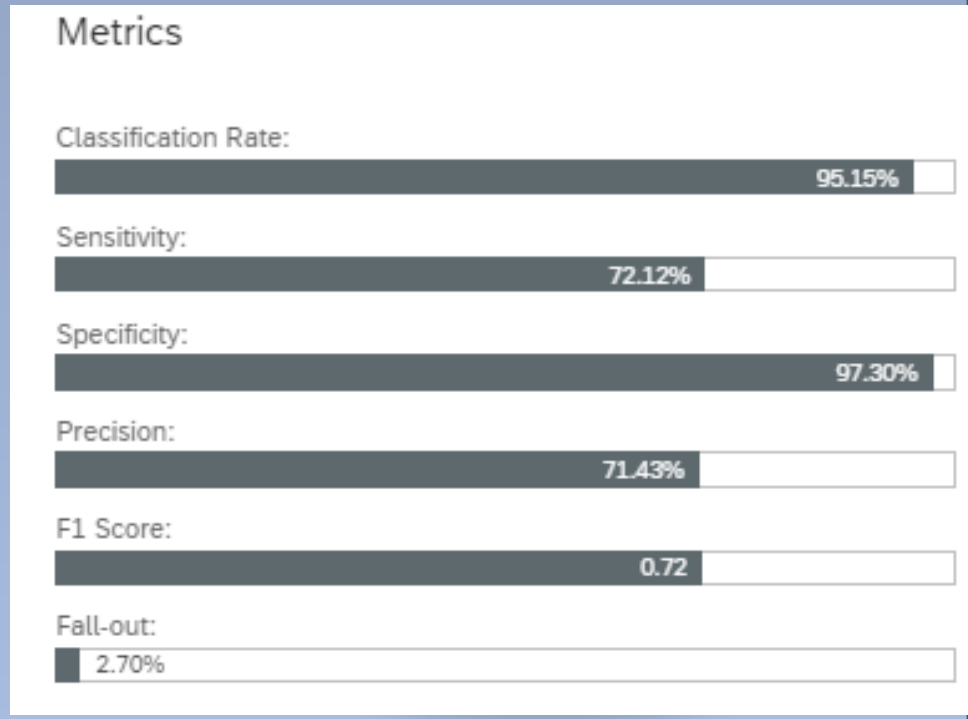
Implementación de algoritmos de minería de datos

Selección del Mejor Modelo

Seleccionar el modelo que se tenga precisión y robustez, son factores para determinar la calidad de la predicción, que refleja cómo es un modelo exitoso.

- La **precisión** analiza la calidad de un modelo predictivo, La precisión mide la proporción de correctas predicciones del número total de casos evaluado.

- La **robustez** de un modelo predictivo se refiere a qué tan bien funciona un modelo con datos alternativos. Estos podrían ser datos de retención o nuevos datos que el modelo se va a aplicar.

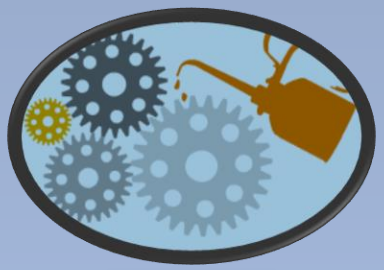


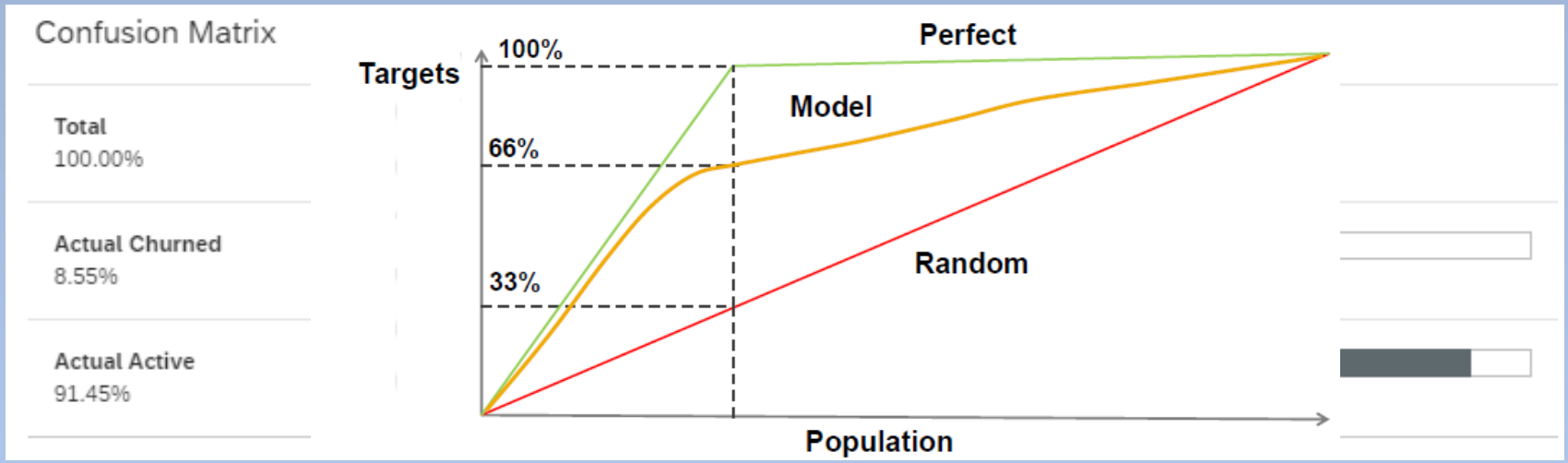
Antes de construir un modelo, necesitamos generar un procedimiento o mecanismo para probar la calidad y validez del modelo

Interpretación y evaluación de datos

Métricas de rendimiento para evaluar el éxito del modelo de clasificación:

- La **Matriz de Confusión** depende de los valores de los errores Tipo I y Tipo II
- Utilizando gráficos con Lift, Gains, ROC y área bajo la curva (AUC)
- Valores en las métricas del Poder predictivo y la Confianza de Predicción





PROCESO DE MINERIA DE DATOS

Selección del Mejor Modelo

Implementación de algoritmos de Minería de Datos

Model Overview

Overview

Model: Potencial_VentasPotencial_Farm	
Data Set:	**VentasPotencial_Farm.csv**
Initial Number of Variables:	**5**
Number of Selected Variables:	**3**
Number of Records:	**256**
Building Date:	**2023-08-14 14:56:50**
Learning Time:	**0 s**
Engine Name:	**Kxen.RobustRegression**
Author:	**Lenovo**

Modeling Warnings	
Auto-selection	**Not enough informative variables to perform auto-selection**

Suspicious Variables

Variable	Target	KI	KR
Num_Compras	Potencial	1.0000	0.934

Monotonic Variables

Variable	Value	Storage	Role	Monotonicity
Clave_Farm	continuous	integer	input	increase

Nominal Targets

Potencial	
Target Key	**NoPotenc**
NoPotenc - Frequency	**49.46%**
Potencial - Frequency	**50.54%**

Performance Indicators

Target: Potencial

rr_Potencial	
Predictive Power (KI)	**1.0000**
Prediction Confidence (KR)	**0.9982**

MINERIA DE DATOS

Introducción al Descubrimiento del conocimiento en datos

Interpretación y Evaluación de Datos

Interpretación y evaluación de datos

Interpretación y evaluación de los datos son etapas cruciales en el proceso de minería de datos, ya que determinan la utilidad y la aplicabilidad de los resultados obtenidos.

Interpretación de Datos: en minería de datos implica darle sentido a los patrones, relaciones y modelos descubiertos durante el análisis. Esto incluye:

- **Identificación de Patrones Significativos**: Analizar los patrones descubiertos para determinar si son relevantes y tienen sentido en el contexto del problema que se está resolviendo.
- **Relacionar los resultados con el dominio de negocio**, interpretando los hallazgos de una manera que sea comprensible y útil para los expertos en el área.
- **Verificar si los resultados obtenidos cumplen** con los supuestos establecidos al inicio del análisis, como correlaciones esperadas o comportamientos predichos.

Interpretación y evaluación de datos

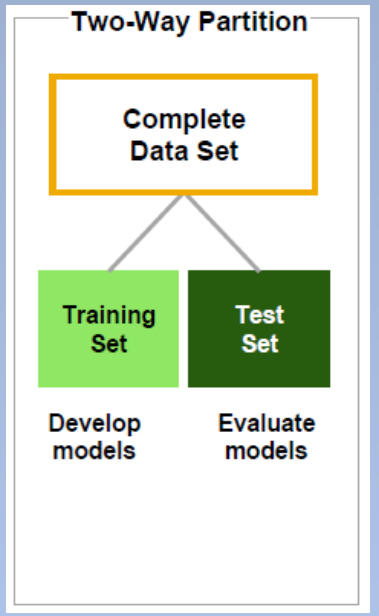
Evaluación de Resultados: es el proceso de medir la eficacia y precisión de los modelos y patrones encontrados. Esto puede incluir:

- **Evaluar la exactitud de los modelos** utilizando métricas como precisión, recall, F1-score, ROC-AUC, entre otras, dependiendo del tipo de análisis realizado.
- **Utilizar técnicas** como la validación cruzada para asegurarse que los resultados no están sobre ajustados y que el modelo generaliza bien con datos no vistos.
- **Evaluar cómo los cambios en las variables** de entrada afectan los resultados del modelo, ayudar a identificar variables clave y entender la robustez del modelo.
- **Comparar los resultados obtenidos** con modelos básicos o métodos tradicionales para determinar la mejora que ofrece el enfoque de minería de datos.
- **Evaluación de Impacto:** Considerar cómo los resultados pueden influir en decisiones prácticas o políticas dentro del contexto del negocio o la investigación.

Interpretación y evaluación de datos

Entrenamiento y Pruebas

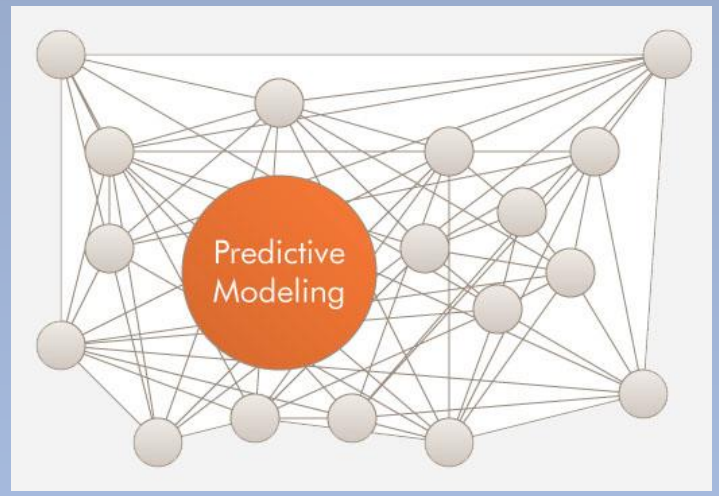
- Fundamental para desarrollar modelos predictivos y evaluar si tienen éxito es un régimen de entrenamiento y prueba.
- Los datos se dividen en entrenamiento y subconjuntos de prueba. Hay una variedad de estrategias de corte por ejemplo: Aleatorio, Secuencial, Periódico
- Facilidad de desarrollo de modelos para que el cliente pueda construir nuevos modelos y actualizar los modelos existentes rápidamente.
- Facilidad de implementación de modelos para que el cliente pueda crear y aplicar conjuntos de datos fácilmente e implementar modelos
- Capacidad de integración con otros sistemas



Interpretación y evaluación de datos

Pruebas Criterios de aceptación

- Facilidad de desarrollo de modelos para que el cliente pueda construir nuevos modelos y actualizar los modelos existentes rápidamente.
- Facilidad de implementación de modelos para que el cliente pueda aplicar nuevos conjuntos de datos fácilmente e implementar modelos rápidamente con los resultados requeridos
- Facilidad de mantenimiento del modelo: modelos requieren actualización / reconstrucción y llevarlo a cabo de forma rápida y sencilla



Interpretación y evaluación de datos

Pruebas Criterios de aceptación

Generalmente, el término **'Calidad**' en Minería de Datos corresponde a las siguientes cuestiones:

- **Representación del conocimiento 'real' incluido en los datos analizados.** Los datos analizados esconden información interesante que los métodos de Data Mining están llamados a revelar para que sea explotable por los expertos del dominio es más fuerte que nunca.
- **Ajuste de algoritmos**. La selección de un método adecuado para una tarea de análisis de datos se basa en su rendimiento y la calidad de sus resultados es uno de los principales problemas de la Minería de Datos.
- **Selección de los patrones más interesantes y representativos para los datos**. pero es probable que solo algunos de estos patrones sean de interés para el experto en el dominio que analiza los datos.

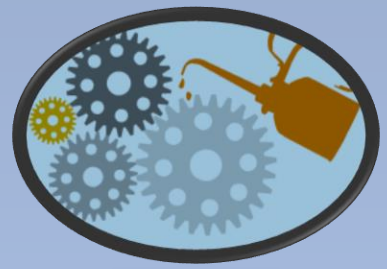
Interpretación y evaluación de datos

Monitoreo y Mantenimiento

Por lo tanto, es importante que los procesos de Monitoreo y Mantenimiento de modelos se puedan automatizar, para que pueda ser lo más eficiente posible y liberar los recursos de los científicos de datos.

Hay una serie de procesos básicos que requieren automatización:

- Programación de modelos
- Aplicar los conjuntos de datos
- Desviación de datos
- Análisis de desviación del modelo
- Identificación de Anomalías
- Actualización o reconstrucción automatizada del modelo



Interpretación y evaluación de datos

Monitoreo y Mantenimiento

- Los modelos de producción casi siempre tienen acuerdos de **nivel de servicio** que tienen que ver con la rapidez con la que deben producir resultados y cómo a menudo se les permite fallar.

- Estas consideraciones operativas pueden ser tan importantes como la precisión del modelo … los resultados correctos devueltos tarde son mucho peores que los resultados correctos devueltos a tiempo.

UNAM Facultad de Ingeniería

MINERIA DE DATOS

José C Roberto Olvera López
Data Science Consultant

jroberto.olveral@gmail.com