Entendiendo la Analítica Predictiva

Esta presentación describe, bajo un marco común, los conceptos fundamentales de Analítica Predictiva.

Descargue la última versión de este documento de: https://github.com/jdvelasq/Lecture-notes-on-analytics/blob/master/intersoftware-predictive-analytics.pdf

JUAN DAVID VELÁSQUEZ HENAO, MSc, PhD Profesor Titular

Departamento de Ciencias de la Computación y la Decisión Facultad de Minas Universidad Nacional de Colombia, Sede Medellín ♠ jdvelasq@unal.edu.co

y @jdvelasquezh

Thttps://github.com/jdvelasq

https://goo.gl/prkjAq

RG https://goo.gl/vXH8jy

Big Data &
Data Analytics
Categoría A,
Convocatoria 781 de 2017

	RDBMS Bases de datos relacionales. SQL Desarrollado por Oracle.			Análisis RDBMS decision Data N Proceso de patro	de inf para es de linin de de nes y grands. Exp Wel De u gube com	negocio. g escubrimientendencias des conjunto	to os la	Web Services SaaS Infraestructura, mantenimiento, operación y soporte dado por un tercero y accesado por internet PaaS Plataformas de desarrollo, Middleware, sistemas operativos IaaS Pago por uso de memoria, disco, procesador, red.		Bases de Datos NoSQL Open Data Science Modern Analytics Explosión IoT			
1970 1976 1985		1989 1995		1996	1999	2002 2005 Internet of Things (IoT) Red de dispositivos físicos con sensores y conectividad que les permiten recolectar e		2009	2011	Big Data Analytics Apache Mah	2016 out		
Data Warehousing OLAP Almacenes de datos que garantizan la permanencia de la información				Business Intelligence 2.0 Plataformas de BI Data Science Área relacionada con los procesos y sistemas para la extracción de conocimiento de datos almacenados electrónicamente				Traditional Analytics Proceso científico de transformación de datos en conocimiento para mejorar el proceso de toma de decisiones de negocio Hadoop / MapReduce Primera implementación de un sistema Big Data		Masifica Big	ación Data		

Data Mining & Business Intelligence

una organización. Data Marts y reportes automáticos. Data BI warehouse DM **Data Mining**

Inteligencia de Negocios

Conjunto de herramientas, productos y tecnologías para la creación y gestión del conocimiento del medio a partir de los datos disponibles en

Tareas típicas: visualización de datos, cálculo de indicadores, dashboards

Enterprise Resource Planning (ERP) Executive information systems (EIS)

Proceso computacional de descubrimiento de patrones y tendencias útiles en grandes conjuntos de datos usando métodos provenientes de la Estadística, el Aprendizaje de Máquinas y los sistemas de bases de datos.

Tareas típicas: detección de anomalías, modelado de dependencias agrupamiento, clasificación, regresión.

CRISP-DM (1996)

Cross-industry standard process for data mining

Comprensión del negocio

Comprensión de los datos

Preparación de los datos

Modelado

Evaluación

Distribución

Resultados del proyecto

Establecer objetivos Formulación del plan preliminar Criterios de éxito

Valoración de la situación actual

Inventario de recursos Requerimientos, hipótesis y restricciones Riesgos y contingencias Terminología Costos y beneficios

Objetivos de la minería de datos

Objetivos de la DM Criterios de éxito desde el negocio Criterios de éxito desde la DM

Plan del proyecto

Desarrollo del plan de proyecto Selección de herramientas y técnicas

Recopilación de datos iniciales

Datos existentes Datos adquiridos Datos adicionales

Descripción de los datos

Cantidad de los datos Tipos de valores Esquema de codificación

Exploración de datos

Análisis exploratorio

Verificación de calidad de datos

Datos perdidos Errores de datos Errores de medición Incoherencias de codificación Metadatos erróneos

Selección de datos

Selección de elementos (filas) Selección de atributos (columnas)

Limpieza de datos

Datos perdidos Errores de datos Errores de medición Incoherencias de codificación Metadatos erróneos

Construcción de nuevos datos

Derivación de atributos (Columnas) Generación de registros (filas)

Integración de datos

Fusión (columnas) Adición (filas)

Formato de datos

CRISP-DM (1996)

Cross-industry standard process for data mining

Comprensión del negocio

Comprensión de los datos

Preparación de los datos

Modelado

Evaluación

Distribución

Selección de la técnica (factores)

Datos disponibles Objetivos de la DM Requisitos específicos

Generación de un diseño de comprobación

Criterios de bondad Definición de datos de testing

Generación de modelos

Configuración de parámetros Listado de modelos producidos Descripción de resultados

Evaluación del modelo

Evaluación global Ajuste de parametros

Evaluación de resultados

Se cumplen los criterios?

Proceso de revisión

Valor de los resultados de cada fase Mejoras posibles Errores o fallos cometidos Callejones sin salida Sorpresas Decisiones alternativas

Determinación del siguiente paso

Continuar con la fase de desarrollo Refinar o sustituir modelos

Planificación de la distribución

Resultados, modelos y descubrimientos Selección de modelos a distribuir Plan de difusión de los descubrimientos Plan de distribución de resultados

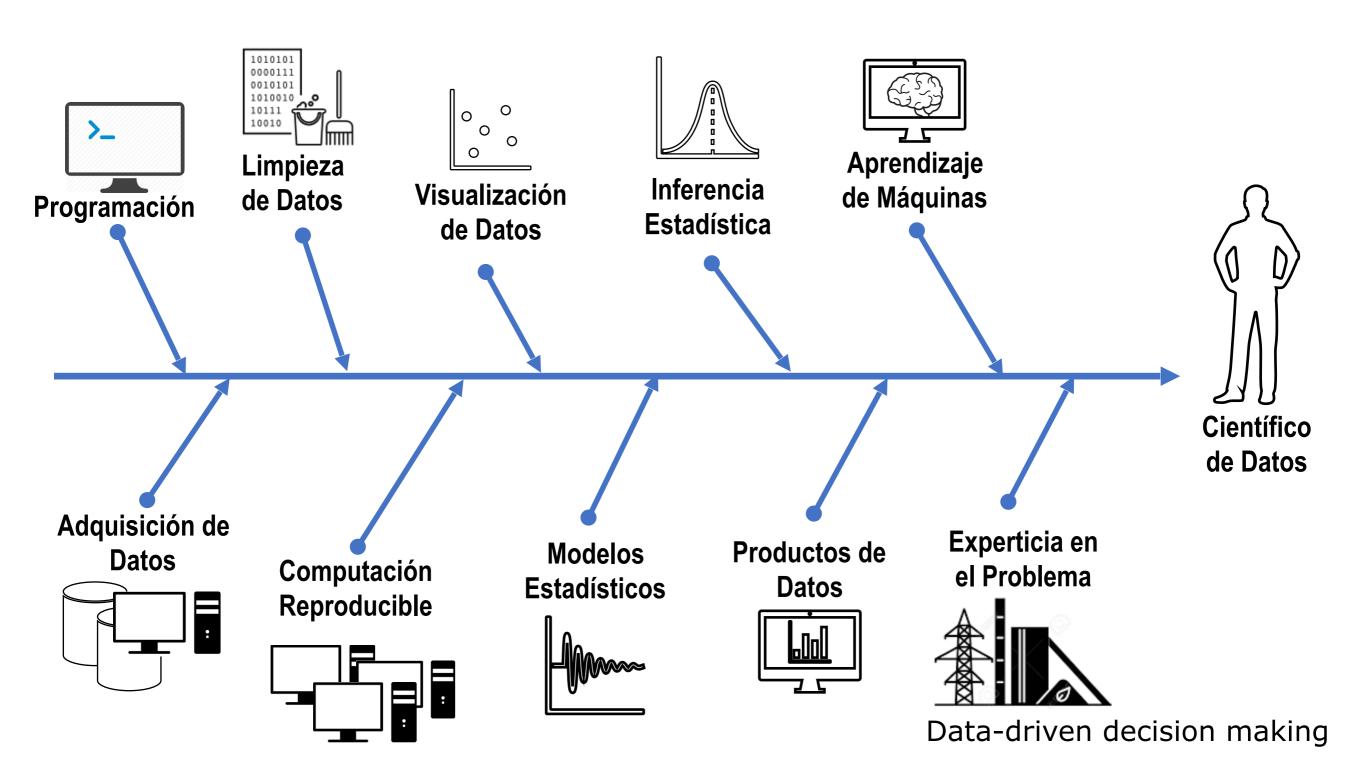
Planificación del control y mantenimiento

Factores o influencias a controlar Medición y control de la validez y precisión Desecho de modelos Reuso de modelos en problemas similares

Creación del informe final

Revisión final del proyecto

Data Science (1996)



Traditional Analytics (2005)

Estadística Descriptiva

Aprendizaje Estadístico

Conjunto de herramientas fundamentadas en conceptos estadísticos para la comprensión de conjuntos de datos con el fin de modelar y predecir.

Aprendizaje de Máquinas•

Área de las ciencias de la computación orientada al desarrollo de sistemas inteligentes (solución de problemas como lo haría un experto humano).

Investigación de operaciones

Inteligencia de negocios -

Área enfocada al uso de técnicas estadísticas descriptivas y KPI para analizar la situación actual

Modelado Predictivo

Área enfocada al uso de técnicas estadísticas utilizadas para pronosticar resultados (de un proceso).

Minería de Datos

Descubrimiento de patrones y conocimiento de grandes conjuntos de datos usando técnicas estadísticas, aprendizaje de máquinas y herramientas de bases de datos.

Analítica Descriptiva / Diagnostica

Análisis de la situación actual para la toma de decisiones operativas

Analítica Predictiva

Área enfocada al uso de técnicas de modelado predictivo, aprendizaje de máquinas y minería de datos para pronosticar los resultados de un proceso en un contexto organizacional.

Analítica Prescriptiva

Uso de técnicas de simulación, optimización y análisis de riesgo e incertidumbre para la toma de decisiones organizacionales

- Compresión de la historía.
- Pronóstico del futuro.
- Los datos están listos.
- Sólida fundamentación matemática.

Aprendizaje Estadístico

Conjunto de herramientas fundamentadas en conceptos estadísticos para la comprensión de conjuntos de datos con el fin de modelar y predecir.

Aprendizaje de Máquinas

Área de las ciencias de la computación orientada al desarrollo de sistemas inteligentes (solución de problemas como lo haría un experto humano).

- No se quiere comprender que pasó.
- Se desea mimificar la inteligencia.
- Pronóstico del futuro.
- · Los datos están listos.
- Fundamentación matemática, pero sin el rigor de la estadística.

Traditional Analytics (2005)

- No se quiere comprender que pasó.
- Pronóstico del futuro.
- · Los datos están listos.
- Sólida fundamentación matemática.

Modelado Predictivo •

Área enfocada al uso de técnicas estadísticas utilizadas para pronosticar resultados (de un proceso).

Minería de Datos

Descubrimiento de patrones y conocimiento de grandes conjuntos de datos usando técnicas estadísticas, aprendizaje de máquinas y herramientas de bases de datos.

- No se quiere comprender que pasó.
- Pronóstico del futuro.
- Los datos están NO están listos.
- Sólida fundamentación matemática.

Analítica Predictiva

Área enfocada al uso de técnicas de modelado predictivo, aprendizaje de máquinas y minería de datos para pronosticar los resultados de un proceso en un contexto organizacional.

- · Pronóstico del futuro.
- Los datos están NO están listos.
- Agrupa todas las anteriores.

Open Data Science & Optimización Simulación Modern Analytics (2009) Análisis de **Decisiones DataOps** 0000111 0010101 Inteligencia 1010010 Colectiva >_ **Aprendizaje** Limpieza Inferencia Visualización de Máquinas de Datos Programación **Estadística** de Datos **Analytics Practitioner** Adquisición de Experticia en **Modelos Datos** el Problema Computación Analítica de Productos de **Estadísticos** Reproducible **Redes Datos Sociales**

Modern Analytics (2009)

FASES / DIMENSIONES

Definición del problema de negocio.

La habildad de entender un problema de negocios y determinar cuando el problema es solucionable mediante la analítica.

- Obtener o recibir la definición del problema y los requisitos de usabilidad.
- Identificar stakeholders.
- Determinar si el problema es solucionable mediante analítica.
- Refinar la definición del problema y definir restricciones.
- Definir el conjunto inicial de beneficios y costos para el negocio.
- *Obtener concenso entre los stakeholders sobre la definicón del problema.

Definición del problema de analítica.

Habilidad para reformular un problema de negocio en un problema de datos con una solución analítica potencial.

- •Reformular el problema en términos de los datos.
- Desarrollar un conjunto propuesto de variables explicativas y relaciones con las salidas.
- Establecer el conjunto de supuestos relacionados con el problema.
- •Definir los criterios de éxito.
- •Realizar el inventario de recursos para la ejecución del proyecto.
- *Obtener consenso entre los stakeholders.

Datos

Habilidad de trabajar efectivamente con datos para identificar relaciones potenciales entre variables que ayudarán a refinar la formulación del problema en términos del negocio y de la analítica.

- Identificar, priorizar los datos necesarios y sus fuentes.
- Adquirir los datos.
- Armonizar, escalar, limpiar y compartir datos.
- Construir, integrar y formatear los datos.
- Análizar los datos e identificar relaciones entre los datos.
- Verificar la calidad de los datos.
- Documentar y reportar hallazgos.
- Refinar la formulación del problema en términios del negocio y de la analítica.

Modern Analytics (2009)

FASES / DIMENSIONES

Selección de la metodología.

Habilidad para identificar y seleccionar aproximaciones potenciales para resolver el problema de negocio

- Identificar las aproximaciones disponibles para la solución del problema.
- Seleccionar las herramientas de software.
- Definir los métodos para probar los modelos.
- Refinar el análisis de costos y beneficios
- Desarrollar un plan inicial del proyecto.
- Seleccionar las aproximaciones a utilizar.

Desarrollo del modelo

Habilidad para identificar y construir modelos efectivos para ayudar a resolver el problema de negocio.

- Identificar las estructuras del o los modelos.
- Correr y evaluar modelos.
- Calibrar modelos y datos.
- Integrar modelos.
- Documentar y reportar hallazgos.

Despliegue del modelo

Habilidad para realizar el despliegue del modelo que ayuda a solucionar el problema de negocio.

- Evaluación del modelo en términos del negocio.
- Publicar reportes con hallazgos.
- Desarrollar los requerimientos del modelo, del sistema y de usabilidad para producción.
- Despliegue del modelo/ sistema en producción.
- Soportar el desarrollo

Gestión del ciclo de vida del modelo

Habilidad para gestionar el ciclo de vida con el fin de evaluar los beneficios del modelo para el negocio sobre el tiempo.

- Documentar la estructura del modelo.
- Evaluar permanentemente la calidad del modelo
- Recalibrar y mantener el modelo.
- Desarrollar actividades de entrenamiento.
- Evaluar periodicamente los beneficios del modelo.

Problemas abordados en la analítica predictiva

- Clasificación de patrones
- Agrupamiento
- Asociación
- Regresión
- Series de tiempo.

Problemas detectados en el medio

- Hay fuertes debilidades en los curriculums de pregrado y postgrado.
- No hay pregrados que den formación integral.
- Fuertes barreras de entrada a nivel conceptual y de herramientas.
- Falta de reconocimiento de la profesión.
- Las organizaciones son usuarios de herramientas y no desarrolladores de algoritmos.
- Se terminan adaptando productos genéricos.
- Falta de reconocimiento del producto de datos.
- Las empresas deben formar personal sobre la marcha.
- No es claro como medir el ROI en productos de analítica.
- Vendedores y Clientes sin sociedad estratégica.

Entendiendo la Analítica Predictiva

Esta presentación describe, bajo un marco común, los conceptos fundamentales de Analítica Predictiva.

Descargue la última versión de este documento de: https://github.com/jdvelasq/Lecture-notes-on-analytics/blob/master/intersoftware-predictive-analytics.pdf

JUAN DAVID VELÁSQUEZ HENAO, MSc, PhD Profesor Titular

Departamento de Ciencias de la Computación y la Decisión Facultad de Minas Universidad Nacional de Colombia, Sede Medellín ♠ jdvelasq@unal.edu.co

y @jdvelasquezh

Thttps://github.com/jdvelasq

https://goo.gl/prkjAq

RG https://goo.gl/vXH8jy

Big Data &
Data Analytics
Categoría A,
Convocatoria 781 de 2017