PROYECTO DE PRACTICA EN EQUIPO NO. 1
Módulo IV- Ciencia de Datos e Inteligencia Artificial
PROF. RENATO R. GONZALEZ DISLA

CONTENIDO

1. Introducción

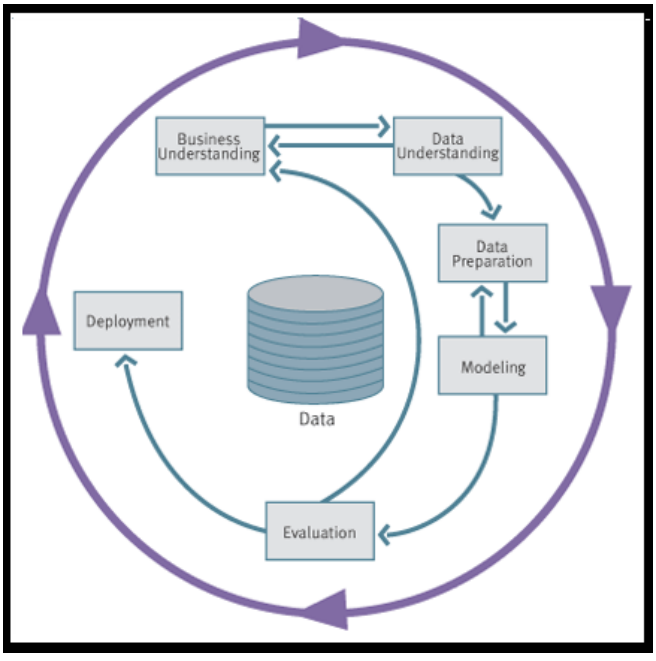
Identificar un proyecto de Ciencia de Datos, Maching Learning o Data Mining desarrollado en una rama de la industria, negocios o gobierno con el objeto de producir una documentación del mismo basado en la metodología de gestión de proyectos de Ciencia de Datos, tal como CRISP-DM, de Ciclo de Vida o de Epiciclos o un hibrido de ellas.

El proyecto ha de consistir en describir cada etapa del proyecto de acuerdo a la documentación obtenida como a continuación describimos usando el modelo CRISP-DM como prototipo metodológico.

2. Metodología de Gestión del Proyecto

2.1 Metodología CRISP-DM

La ciencia de datos sigue una metodología de ciclo de vida de proyectos con características particulares por el tipo de herramienta y objetivos de proyectos definidos en el alcance del proyecto. Se ha establecido un ciclo de proyecto basado en la metodología CRISP-DM, de acuerdo a la naturaleza del proyecto. Cada etapa desarrollada esta seguida por un entregable como producto de las actividades de la metodología.



El estándar incluye un modelo y una guía, estructurados en **seis fases (ver Anexo Documento Guía de CRISP-DM o del libro The Art of Data Science de Pen)**, algunas de estas fases son bidireccionales, lo que significa que algunas fases permitirán revisar parcial o totalmente las fases anteriores. Existe una VII fase que es la de gestión y evaluación del modelo, que dependerá de los planes de mercadeo que empresa desarrolle a partir de los resultados del modelo predictivo. Estas las definiremos a continuación:

I. Comprensión del Negocio

En esta primera etapa se definen los objetivos y requerimientos desde una perspectiva no técnica, sino de negocios. Se han de definir los alcances del proyecto en términos de la visión creada en cuanto a Política de la Empresa respecto ala problema de Negocios a resolver, definiendo al mismo tiempo los alcances y límites de la misma. Se han de definir las áreas involucradas y su relación, el staff, actores y organización del proyecto con sus roles, perfiles y responsabilidades respectivos de acceso a la información generada durante el proceso. Se establece el procedimiento de aprobación de las etapas y entregas del proyecto. Los cambios y actualizaciones deben ser contemplados en este proceso. El Equipo de Proyecto y el Comité de Dirección son los encargados de estas revisiones y aprobaciones.
Nota: Puede crear escenarios ficticios a partir del problema elegido.

Comprende:

- Entendimiento de las áreas del negocio
- Políticas de la empresa
- Establecimiento de los objetivos de la organización y criterios de éxito.
- Evaluación de la situación (Inventario de recursos de datos, requerimientos, supuestos, terminologías propias del proyecto)
- Establecimiento de los objetivos de la ciencia de datos (objetivos y criterios de éxito)
- Generación del plan detallado del proyecto (plan, herramientas, equipo y técnicas)

Entregable: Documento de Modelo de Negocio y Definición del problema

II. Comprensión y Análisis de Calidad de los datos

Consiste en familiarizarse con los datos teniendo presente los objetivos del proyecto. Los datos del proyecto se almacenan en diversas bases de datos por los sistemas institucionales, tanto los datos básicos demográficos y personales de las

entidades envueltas como la información transaccional y la Big Data. Consiste en extraer los datos relevantes para crear un repositorio usado posteriormente por el analizador de datos R u otra herramienta.

Comprende:

- Recopilación inicial de datos
- Descripción de los datos
- Exploración de los datos (Visualización)
- Verificación de calidad de datos

Entregable: Documento con la descripción de los datos y análisis de consistencia y calidad de la información periódica seleccionada

III. Preparación de los datos

El objetivo es obtener la vista o data set de la minería de datos o de la big data para el desarrollo de los modelos. La limpieza y transformación de datos es el más intensivo de los recursos del proceso de proyecto de minería de datos. El propósito de la limpieza de datos es eliminar el ruido e información irrelevante del conjunto de datos, y modificar la fuente de datos en diferentes formatos en función de los tipos de datos y valores.

Comprende las siguientes actividades:

- Selección de los datos
- Limpieza de datos (Data Clean)
- Construcción de datos
- Integración de datos
- Formateo de datos

Entregable: Documento informe con la estructura de datos y los archivos de datos depurados que servirán de entrada al Modelo

IV. Modelado

En esta etapa se aplican las técnicas de ciencia de datos a los dataset. Una vez que los datos se limpian y las variables son transformadas, podemos empezar a construir modelos basados en los recursos algorítmicos provistos por el lenguaje R u otra herramienta. Antes de construir cualquier modelo, tenemos que entender el objetivo del proyecto de ciencia de datos y el tipo de la tarea de minería de datos, como se ha definido en la sección anterior.

Una vez que entienda el tipo de tarea de ciencia de datos, se seleccionan los algoritmos de análisis de datos correctos. Para cada tarea de minería, hay algunos algoritmos adecuados como lo hemos definido en la sección anterior. Se seleccionan un conjunto de datos preliminares de prueba. En muchos casos, no sabemos que es el mejor algoritmo de ajuste para los datos antes del entrenamiento del modelo.

La precisión y desempeño del algoritmo depende de la naturaleza de los datos como el número de los estados del atributo de predicción, la distribución del valor de cada atributo, las relaciones entre los atributos, y así sucesivamente.

Comprende:

- Selección de la técnica de modelado y del algoritmo
- Diseño de la evaluación
- Construcción del modelo
- Evaluación del modelo

Entregable: Modelo y documento descriptivo del modelo y sus componentes: Rutas, algoritmo empleado y los parámetros de precisión del mismo y las variables explicativas y objetivo usadas y evaluadas

V. Predicción y Evaluación del Modelo

En la etapa de construcción del modelo, se construye un conjunto de modelos que utilizan diferentes algoritmos y ajustes de parámetros. Entonces, ¿cuál es el mejor modelo en términos de precisión? ¿Cómo evalúa estos modelos? El Lenguaje R u otro proveen herramientas para evaluar la calidad de un modelo. Se utiliza un modelo de formación para predecir los valores del conjunto de datos de prueba, sobre la base del valor de predicción y la probabilidad.

Comprende:

- Evaluación de resultados mediante matriz de confusión o contingencia
- Revisar el proceso
- Establecimiento de los siguientes pasos o acciones

Entregable: Documento evaluativo del resultado del modelo por parte del equipo de proyecto

VI. Despliegue e implementación.

Consiste en explotar la utilidad de los modelos, integrándolos en las tareas de toma de decisiones de la organización. La Implementación de un Modelo de Machine Learning cubre los siguientes componentes:

Comprende las siguientes actividades:

- Planificación de despliegue
- Planificación de la monitorización y del mantenimiento
- Capacitación de las áreas de usuarios
- Generación de informe final
- Revisión del proyecto

Entregable: Manuales de cpacitacion y Documento evaluativo de Implementación

VII. Gestión del Modelo

Es difícil mantener el estado de los modelos de minería de datos. Cada modelo de minería tiene un ciclo de vida. En algunas instituciones, los patrones de datos son relativamente estables y los modelos no requieren de reciclaje con frecuencia. Sin embargo, en los patrones de muchas instituciones varían con frecuencia, esto significa que las nuevas reglas de asociación aparecen cada día. Es un proceso dinámico y una nueva versión del modelo se debe crear con frecuencia a partir de los resultados de la aplicación de las medidas y campañas de retención de clientes desarrolladas (si es un proyecto de clientes). En última instancia, determinar la exactitud del modelo y la creación de nuevas versiones de este debe llevarse a cabo mediante el uso de procesos automatizados, y una herramienta versátil de gestión de contenidos y versiones de los modelos.

Comprende:

- Evaluación de la campaña implementada
- Evaluación del modelo
- Actualización o adaptación del modelo

Entregable: Documento de evaluación de modelo