**Proyecto Integrador**

Agrupación y clasificación de las características de calidad del algodón basados en la producción por regiones y sitios de recibo con respecto a las solicitudes por cliente, para optimizar el almacenamiento y la preparación de pedidos en la logística de la empresa Diagonal Colombia.

**Integrantes**:

Mauricio Jiménez Sánchez - mjjimenezs@eafit.edu.co

Jefferson Quiroz Fino - jaquirozf@eafit.edu.co

Carlos Enrique Salazar- csalazar@eafit.edu.co

César Augusto Cardona Ortiz - cacardonao@eafit.edu.co

Jesús Alberto Arcia Hernández - jaarciah@eafit.edu.co

**ESCUELA DE CIENCIAS**

**MAESTRÍA EN CIENCIA DE LOS DATOS Y ANALÍTICA.**

**UNIVERSIDAD EAFIT.**

**MAYO DE 2020.**

**Tabla de contenido**

[**1.** **Descripción del Problema:** 3](#_Toc40113225)

[**2.** **Metodología.** 4](#_Toc40113226)

[a. **Definición de ciclo de vida de ciencia de datos:** 4](#_Toc40113227)

[**b.** **Estructura de proyecto estandarizada.** 6](#_Toc40113228)

[**c.** **Infraestructura y recursos para proyectos de ciencia de datos.** 7](#_Toc40113229)

[**d.** **Herramientas y utilidades para la ejecución de proyectos.** 8](#_Toc40113230)

[**3.** **Aplicando la metodología “*TDSP*” al proyecto integrador.** 10](#_Toc40113231)

[**a.** **Ciclo de vida** 10](#_Toc40113232)

[**i.** **Conocimiento del negocio** 10](#_Toc40113233)

# **Descripción del Problema:**

Actualmente, el algodón involucra cerca de 17 características de calidad que son tomadas en cuenta en el proceso de hilandería, a fin de ejecutar lotes muy homogéneos en la actividad de alimentación de las plantas, que permitan una producción de tela todos los días que esté dentro de los estrechos márgenes de variabilidad que los clientes exigen.

Este proceso genera un gran desafío logístico, en razón a que las hilanderías solicitan despachos de vehículos con cerca de 130 pacas de algodón específicas, las cuales deben ser escogidas muchas veces de entre un almacenamiento a campo abierto que contiene más de 20,000 pacas.

El mayor inconveniente se presenta en que las características de calidad como longitud, resistencia, basura, color etc, son medidas de unas muestras tomadas de cada producto que son enviadas por los vendedores a un laboratorio de análisis en Medellín.

Desafortunadamente el algodón es recibido en los sitios de recepción repartidos en todo el país antes de que lleguen dichos resultados, lo que no permite un almacenamiento programado por grupos de algodones homogéneos, que disminuya el movimiento de montacargas en la actividad de escogencia de las pacas.

El objeto de este trabajo es utilizar una amplia base de datos con el histórico de las características paca a paca compradas por Diagonal en la última década que, contrastadas con el histórico de solicitudes por vehículo y cliente, permitan sugerir un mecanismo de almacenamiento por grupos que disminuya los movimientos internos al momento de escogencia de pacas.

La idea es que, basados en algún criterio específico como proveedor, zona o desmotadora, se puedan preestablecer un tipo de almacenaje con grupos de calidades esperadas homogéneas, ya que normalmente los vehículos se solicitan con unos grupos de calidades muy similares.

# **Metodología.**

La normalización del proceso de hallar conocimiento en los datos ha convocado esfuerzo desde finales de los años 90, durante este tiempo se han planteado metodologías que buscan alinear la ejecución de los proyectos analíticos con los objetivos del negocio sin descuidar los elementos técnicos que deben tenerse en cuenta. Dentro de estas metodologías se encuentra “***Team Data Science Process***” conocida como “***TDSP***” planteada por Microsoft donde indica que ayuda a mejorar la colaboración y el aprendizaje en equipo al sugerir cómo los roles de equipo funcionan mejor juntos. TDSP incluye procedimientos recomendados y estructuras de Microsoft y otros líderes del sector para ayudar a implementar correctamente iniciativas de ciencia de datos. El objetivo es ayudar a las empresas a que se den cuenta de las ventajas de su programa de análisis. Los principales componentes de esta metodología son:

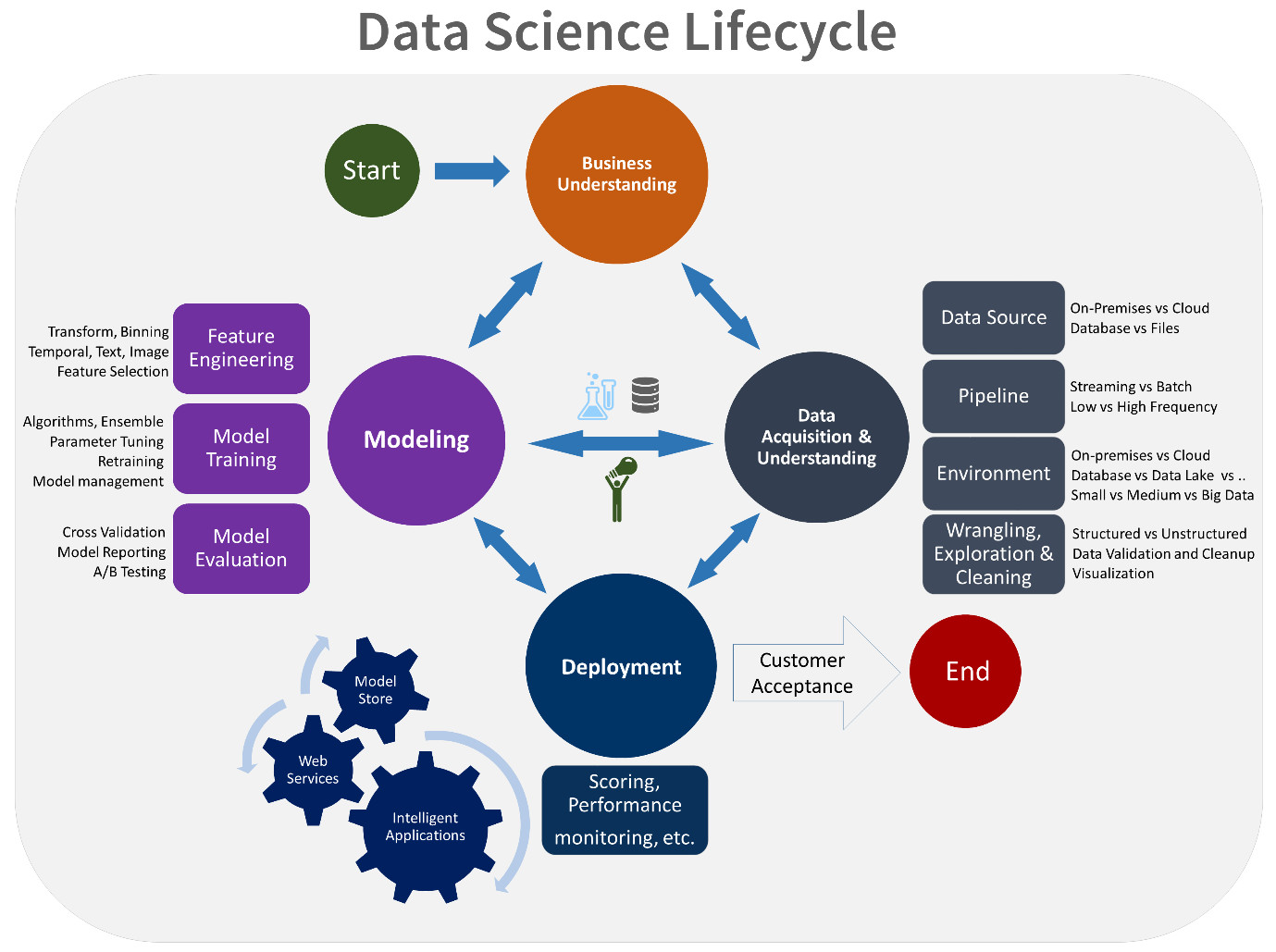
## **Definición de ciclo de vida de ciencia de datos:**

El proceso de ciencia de datos para la metodología (TDSP) proporciona un ciclo de vida para estructurar el desarrollo de los proyectos de ciencia de datos. En el ciclo de vida se describen todos los pasos que siguen los proyectos correctos. Este ciclo de vida se ha diseñado para proyectos de ciencia de datos que se enviarán como parte de aplicaciones inteligentes. Estas aplicaciones implementan modelos de aprendizaje o inteligencia artificial de máquina para realizar un análisis predictivo. Los proyectos de ciencia de datos exploratorios o proyectos de análisis improvisados también se pueden beneficiar del uso de este proceso. Pero, en estos casos, puede que algunos de los pasos descritos no sean necesarios.

El ciclo de vida describe las fases principales por las que pasan normalmente los proyectos, a menudo de forma iterativa:

* Conocimiento del negocio.
* Adquisición y comprensión de los datos.
* Modelado.
* Implementación.
* Aceptación del cliente.

Esta es una representación visual del ciclo de vida del proceso de ciencia de datos en equipo.



En la metodología **“*TDSP”*** se describen los objetivos, las tareas y los artefactos de documentación de cada fase del ciclo de vida de ***TDSP***. Estas tareas y artefactos están asociados con roles de proyecto:

* Arquitecto de soluciones.
* Jefe de proyecto.
* Científico de datos.
* Responsable de proyecto.

En el siguiente diagrama se proporciona una vista de cuadrícula de las tareas (en azul) y los artefactos (en verde) asociados con cada fase del ciclo de vida (eje horizontal) de estos roles (eje vertical).

Captura de pantalla de un celular

Descripción generada automáticamente

## **Estructura de proyecto estandarizada.**

Cuando todos los proyectos comparten una estructura de directorio y usan plantillas para los documentos de proyecto, resulta fácil para los miembros del equipo encontrar información sobre sus proyectos. Todo el código y los documentos se almacenan en un sistema de control de versiones (VCS), como Git, TFS o Subversión para permitir la colaboración en equipo. El seguimiento de las tareas y las características en un sistema de seguimiento de proyectos ágil, como Jira, Rally y Azure DevOps permite seguir más de cerca el código para conocer sus características individuales. Este seguimiento también permite a los equipos obtener mejores estimaciones de los costos. TDSP recomienda crear un repositorio independiente para cada proyecto en el VCS de cara al control de versiones, la seguridad de la información y la colaboración. La estructura estandarizada para todos los proyectos ayuda a crear conocimiento institucional en toda la organización.

Se proporcionan plantillas para la estructura de carpetas y los documentos necesarios en ubicaciones estándar. Esta estructura de carpetas organiza los archivos que contienen código para la exploración de datos y la extracción de características, y los que registran las iteraciones de los modelos. Estas plantillas permiten a los miembros del equipo comprender el trabajo que otros realizan, y agregar nuevos miembros a los equipos de forma fácil. Las plantillas de documento se pueden ver y actualizar fácilmente en formato de marcado. Use plantillas para proporcionar listas de comprobación con preguntas clave en cada proyecto y de esta forma garantizar que el problema esté bien definido y que los resultados entregados satisfagan la calidad esperada. Algunos ejemplos son:

* Una carta de constitución de proyecto para documentar los problemas empresariales y el ámbito del proyecto.
* Informes de datos para documentar la estructura y las estadísticas de los datos sin procesar.
* Informes de modelo para documentar las características derivadas.
* Métricas de rendimiento de modelo, como curvas ROC o MSE.

Captura de pantalla de un celular

Descripción generada automáticamente

## **Infraestructura y recursos para proyectos de ciencia de datos.**

TDSP proporciona recomendaciones para administrar análisis compartido e infraestructura de almacenamiento, por ejemplo:

* Sistemas de archivos en la nube para almacenar conjuntos de datos.
* Databases.
* Clústeres de macrodatos (Hadoop o Spark).
* Servicio de aprendizaje automático.

La infraestructura de análisis y almacenamiento, donde se almacenan los conjuntos de datos sin procesar y los procesados, puede estar en la nube o en un entorno local. Esta infraestructura permite un análisis reproducible. También evita la duplicación, lo que puede llevar a incoherencias y costos de infraestructura innecesarios. Se proporcionan herramientas para aprovisionar los recursos compartidos, realizar un seguimiento de ellos y permitir que cada miembro del equipo se conecte a dichos recursos de forma segura. También es una buena práctica pedir a los miembros del proyecto que creen un entorno de proceso coherente. Luego, diferentes miembros del equipo pueden replicar y validar los experimentos.

Este es un ejemplo de un equipo que trabaja en varios proyectos y que comparte diversos componentes de la infraestructura de análisis.

Imagen que contiene texto, mapa

Descripción generada automáticamente

## **Herramientas y utilidades para la ejecución de proyectos.**

En la mayoría de las organizaciones la introducción de procesos presenta ciertos desafíos. Las herramientas proporcionadas para implementar el proceso y el ciclo de vida de ciencia de datos ayudan a reducir las barreras a su adopción y la normalizan. TDSP proporciona un conjunto inicial de herramientas y scripts para impulsar la adopción de TDSP dentro de un equipo. También ayuda a automatizar algunas de las tareas comunes del ciclo de vida de ciencia de datos, como la exploración de datos y el modelado de línea de base. Existe una estructura bien definida que se proporciona a los individuos para que contribuyan con herramientas y utilidades compartidas al repositorio de código compartido de su equipo. Estos recursos se pueden aprovechar luego en otros proyectos dentro del equipo o en la organización. TDSP también tiene previsto habilitar las contribuciones de herramientas y utilidades a toda la comunidad. Las utilidades de TDSP se pueden clonar desde GitHub.

# **Aplicando la metodología “*TDSP*” al proyecto integrador.**

## **Ciclo de vida**

El ciclo de vida se compone de cinco faces principales que se ejecutan en forma iterativa y estas están compuestas por:

### **Conocimiento del negocio**

#### **Definición de objetivos.**

* **Objetivos del modelado**

1. clusterización de las calidades del algodón basados en sus características.
2. Clasificación de las pacas de algodón según la calidad, basados en las solicitudes de los clientes, zonas de producción y sitios de recibo.

* **Métricas asociadas**

Aquí las métricas

#### **Identificación de los orígenes de datos.**

* Los datos son suministrados por la empresa Diagonal Colombia en formato .csv descritos de la siguiente manera a un tiempo de 10 años:
  1. Características de recibos de mercancías.
  2. Características de despachos de mercancía.
  3. Movimientos de pacas entre sitios de almacenamientos.
  4. Históricos de las condiciones meteorológicas de las zonas de producción.
  5. Localización histórica de las pacas en sitios de distribución.

**Artefactos:**

* Carta del proyecto:

