**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.**

**JUNIO DE 2020.**

**Tabla de contenido**

[**1.** **Descripción del Problema:** 4](#_Toc44066468)

[**2.** **Metodología.** 6](#_Toc44066469)

[a. **Definición de ciclo de vida de ciencia de datos:** 6](#_Toc44066470)

[**b.** **Estructura de proyecto estandarizada.** 8](#_Toc44066471)

[**c.** **Infraestructura y recursos para proyectos de ciencia de datos.** 9](#_Toc44066472)

[**d.** **Herramientas y utilidades para la ejecución de proyectos.** 10](#_Toc44066473)

[**3.** **Aplicando la metodología “*TDSP*” al proyecto integrador.** 12](#_Toc44066474)

[**a.** **Ciclo de vida** 12](#_Toc44066475)

[**i.** **Conocimiento del negocio.** 12](#_Toc44066476)

[**ii.** **Adquisición y comprensión de datos** 13](#_Toc44066477)

[**iii.** **Modelado** 14](#_Toc44066478)

[**iv.** **Conclusión** 14](#_Toc44066479)

**4. Anexos**

# **Descripción del Problema:**

Actualmente, el algodón involucra cerca de 16 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.

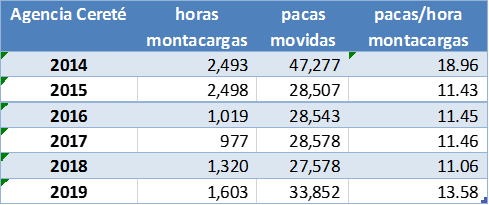
El algodón es recibido en los sitios de recepción repartidos en todo el país con los resultados de los análisis de las características por paca entregados por el laboratorio. No obstante, el proceso de almacenamiento se realiza de forma aleatoria sin tener en cuenta las características de calidad de 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. Para esto se escogerá la información de los años 2017 y 2018 para la predecir los grupos que deberían hacerse para el año 2019.

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.

**ESTIMACIÓN DE LOS BENEFICIOS**

* El siguiente es el cuadro de horas montacargas de una de las agencias de Diagonal​



* Teniendo en cuenta que la tarifa de montacargas es de $ 50,000/hora, que un montacarga mueve en promedio 13 pacas/hora y que se mueven en todo Diagonal unas 100,000 pacas año, el egreso por montacargas es aproximadamente de $ 384 millones por año.
* Adicionalmente, la carga de un vehículo está tomándose en ocasiones hasta medio día, se estima que se puede conseguir un descuento del 5% en transportes si este guarismo se puede reducir a la mitad, porque el tiempo muerto de la mula le representa dinero a la compañía de transporte.
* Los cálculos que se han hecho con la compañía Diagonal, es que el tiempo de montacargas se gasta un 20% en almacenaje y un 80% en preparación de pedidos para despachos. Así mismo, este 80% se reparte en 50% para desplazamientos y un 50% para manipulación de pacas. ​
* Es así como se estima que una reducción de un 50% en el desplazamiento de preparación de despachos, podrá significar un menor egreso por montacargas del orden del 25%, que significa aproximadamente 96 millones al año. ​
* Así mismo, si se logra reducir a la mitad el tiempo de carga de un vehículo, se puede conseguir una rebaja del 5% en fletes. El transporte terrestre de una paca cuesta aproximadamente $31,818 pesos por paca. Un 5% de este valor en 100,000 pacas anuales, representaría un ahorro de aproximadamente $ 159 millones anuales.

# **Metodología.**

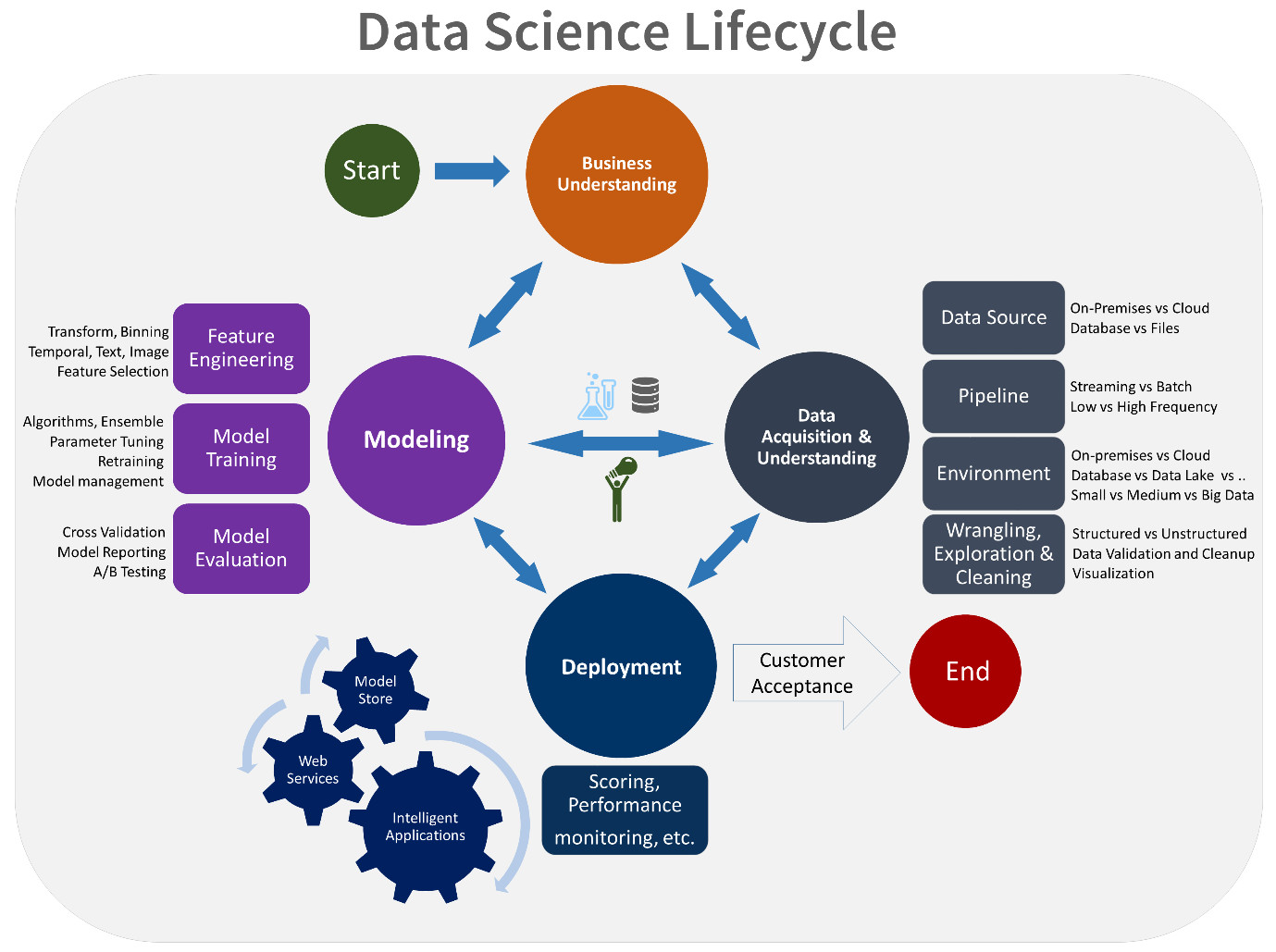
La normalización del proceso de hallar conocimiento en los datos ha convocado esfuerzo desde finales de los años 90, durante este tiempo se han planteado metodologías que buscan alinear la ejecución de los proyectos analíticos con los objetivos del negocio sin descuidar los elementos técnicos que deben 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.



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.

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.
* Captura de pantalla de un celular

  Descripción generada automáticamenteMétricas de rendimiento de modelo, como curvas ROC o MSE.

## **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.

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.

##### **Objetivo general del modelado.**

1. Reducir los tiempos en la preparación de pedidos del sitio de almacenaje.

##### **Objetivos específicos del modelado**

1. Clusterización de las calidades del algodón basados en sus características utilizando la información o los datos recolectados en los años 2017 y 2018.
2. Clasificación de las pacas de algodón según los resultados de laboratorio para predecir la ubicación de cada paca para el año 2019.

#### Métricas asociadas

* Basados en el objetivo general, se requiere reducir los tiempos en la preparación de pedidos en al menos un 40% con respecto al año 2019.

#### Identificación de los orígenes de datos.

##### Los datos son suministrados por la empresa Diagonal Colombia en formato .csv descritos de la siguiente manera para los años 2017 y 2018:

* 1. Características de recibos de mercancías.
  2. Características de despachos de mercancía.
  3. Movimientos de pacas entre sitios de almacenamientos.
  4. Localización histórica de las pacas en sitios de distribución.

#### Artefactos.

##### **Carta del proyecto:** (ver anexo A)

##### **Origen de los datos:** (ver anexo B)

##### **Diccionario de datos:** (ver anexo C)

### **Adquisición y comprensión de datos**

#### Introducción de los datos

##### Los datos ubicados en una base de datos SQL se descargar en archivo tipo csv, una vez los datos son descargados, se procede a cargarse en el bucket de datos crudos en AWS S3. Los datos se cargar en un jupyter-notebook para hacer el procesamiento correspondiente.

#### *Exploración de los datos*

##### En la etapa de preparación y exploración de datos se logra limpiar la data quitando los datos nulos, datos no invalido y outliers, para cada una de las variables influyentes en el proyecto. Una vez los datos se encuentran limpio se procede a realizar la exploración de datos en la cual comparamos correlación entre variables utilizando distintas métricas y análisis de componentes principales.

#### Configuración de una canalización de datos

##### Una vez los datos se encuentran limpios y analizados, se procede a exportar estos datos limpios a un bucket en el cual se introducen los datos procesados llamado processeddiagonalcluster.

#### Artefactos:

##### **Descripción de los datos:** (veranexo D)

### **Modelado**

#### Diseño de características

##### En este proyecto se implementaron dos algoritmos, el primero se trata del algoritmo K-medoids, el cual se encarga de agrupar el conjunto de datos proveniente de las calidades de cada paca introducidas en cada despacho realizado entre los años 2017 y 2018.

##### El segundo algoritmo implementado en este proyecto fue el RandomForest encargado de la clasificación de las pacas.

#### Entrenamiento del modelo

##### En el modelo de K- medoids se tomó la información de los años 2017 y 2018 para la creación de los grupos.

##### Para el algoritmo RandomForest se dividieron los datos en una proporción 80-20, tomando el 80% para el entrenamiento y el 20% para el test. Los datos de entrenamiento y test a su vez están compuestos por 2 set de datos, los datos X de entrenamiento y test, lo cuales contienen las características de las pacas y los datos Y de entrenamiento y test, los cuales contiene la información acerca de que grupo pertenece cada paca, la cual proviene de la implementación del algoritmo k-medoids. Toda la información tanto de test y entrenamiento proviene de un rango de fechas ubicados entre los años 2017 y 2018.

##### 

#### Evaluación del modelo

##### El modelo de K.medoids es evaluado por medio del algoritmo silhouette para medir la calidad de los agrupamientos (clusters) encontrados. Este método va a permitir determinar qué tan bien cada objeto se encuentra dentro de su agrupación. El número óptimo de clusters k es el que maximiza la silueta promedio en un rango de posibles valores de k.

##### Para el modelo RandomForest se utiliza métrica accuracy para evaluar la eficiencia del modelo.

#### Artefactos

##### **Informe de modelo:** (ver anexo E)

### **Conclusión**

Se seleccionó el modelo de RandomForest debido a una mejor distribución en los grupos de almacenamiento obteniendo un 31.6% de reducción en las distancias recorridas por lo montacargas, el mejor valor de reducción que es del 37.3% se obtuvo con una regresión logística, esta reducción representa un ahorro de 72.000.000 millones de pesos al año para la empresa.

# **4.Anexos**

**Anexo A. Carta del proyecto**

**Conocimiento del negocio**

* **¿Quién es el cliente, en qué dominio comercial se encuentra el cliente?**

El cliente es Diagonal Colombia, es una corporación de derecho privado, sin ánimo de lucro, constituida en abril de 1950 para facilitar el aprovisionamiento de materias primas a la industria textil colombiana. En la actualidad Diagonal comercializa el 100% de la fibra de algodón consumida por la industria textil nacional e importa otros insumos como hilazas, fibras sintéticas y telas.

* **¿Qué problemas comerciales estamos tratando de abordar?**

El problema que estamos abordando es logístico ya que se quiere agrupar y clasificar el algodón según sus características basados en la producción por regiones y sitios de recibo con respecto a las solicitudes del cliente, de esta forma optimizar el almacenamiento y preparación de pedidos en la logística.

**Alcance**

* ¿Qué soluciones de ciencia de datos estamos tratando de construir?
  1. En primera instancia se pretende realizar una clusterización de las calidades del algodón basadas las salidas de despachos.
  2. En segunda instancia se pretende clasificar las pacas de algodón según los grupos obtenidos en la primera instancia.
* ¿Qué haremos?
  1. Definición del problema.
  2. Adquisición de los datos.
  3. Origen de los datos.
  4. Entendimiento de los datos.
  5. Ambiente tecnológico.
  6. Almacenamiento de los datos.
  7. Frameworks de procesamiento de datos.
  8. Tratamiento de los datos.
  9. Ingesta de los datos.
  10. Limpieza y filtrado de los datos a través de métricas y distancias adecuadas.
  11. Identificación de outliers.
  12. Aplicación de valores y vectores propios en la identificación de proyecciones y reducción de dimensión “**PCA**”.
  13. Normas y determinantes para calcular varianzas globales.
  14. Productos internos para la identificación de asociaciones lineales entre variables.
  15. Número condición de las matrices de covarianzas.
  16. Creación de modelos
  17. Modelos de clusterización.
  18. Modelos de calificación.

1. Validación de los modelos.
2. Despliegue de modelos.

* ¿Cómo va a ser consumido por el cliente?

El cliente va a consumir los resultados de los modelos por medio de un dashboard en la plataforma de su elección.

**Personal**

* Quiénes están en este proyecto:
  + Diagonal-cluster:
    - **Líder del Proyecto**.
      * César Augusto Cardona Ortiz.
    - **PM**.
      * Carlos Enrique Salazar Escobar.
    - **Científico de datos**
      * Jesús Alberto Arcia Hernández.
      * Mauricio José Jiménez Sánchez.
    - **Gerente de cuentas.**
      * Jefferson Adolfo Quiroz Fino.
  + **Cliente**:
    - **Administrador de datos**, personal de Diagonal Colombia.
    - **Contacto de negocio,** personal de Diagonal Colombia.

**Métrica**

* ¿Cuáles son los objetivos cualitativos?
  1. Reducir los recorridos de los montacargas en la preparación de pedidos.
  2. Incrementar la capacidad de despachos diarios de vehículos.
  3. Aumentar la experiencia del cliente garantizando pedidos rápidos y oportunos.
* ¿Cuáles son las métricas cuantificables del proyecto?
  1. Reducir el 40% los recorridos de los montacargas en la preparación de pedidos.
* Cuantifique las mejoras de los valores de las métricas que son útiles para el escenario del cliente.
  1. Disminución del tiempo de uso de los montacargas.
  2. Reducción en la preparación de los pedidos.
* ¿Cuál es el valor de referencia (actual) de la métrica?
  1. Tiempo de uso del montacarga.
  2. Tiempo en la preparación de los pedidos.
* ¿Cómo mediremos la métrica? (por ejemplo, prueba A / B en un subconjunto especificado durante un período específico; o comparación del rendimiento después de la implementación con la línea de base)

**Plan**

* Fases (hitos), línea de tiempo, breve descripción de lo que haremos en cada fase.



**Arquitectura**

* Datos
  + ¿Qué datos esperamos?

Esperamos archivos .csv descritos de la siguiente manera para las fechas 2017 y 2018:

* + - * Características de recibos de mercancías.
      * Características de despachos de mercancía.
      * Movimientos de pacas entre sitios de almacenamientos.
      * Localización histórica de las pacas en sitios de distribución.
* Movimiento de datos de local a AWS usando el componente S3 mediante boto3 y sqoop, donde vamos a trabajar sobre el siguiente esquema:
  + **Raw**, bucket de S3 donde van a estar alojados los datos crudos.
  + **Processed,** bucket de S3 donde van a estar alojados los datos procesados.
  + **Modeling,** bucket de S3 donde van a estar alojados los datos para realizar el modelado.
* **Herramientas y recursos de análisis**
  + Microsoft project 2019 para el modelado del proyecto.
  + Simple Storage Service (Amazon S3) para el almacenamiento de la informacion.
  + Python para construcción de las ETL y modelado de la información.
  + Amazon Elastic MapReduce (Amazon EMR).
* **Consumo de los modelos creados en el proyecto**
  + ¿Por medio de que herramienta se va a consumir la información suministrada por los modelos?

Por medio de la fuente de datos ODBC de POWER BI utilizando AWS S3.

* + ¿Cómo utilizará el cliente los resultados del modelo para tomar decisiones?

Se montará en la página web “***https://www.diagonal.com.co/#”*** en la sección de “***HVI***” un dashboard con los filtros de agencias de recibo, agremiación y fecha. El aplicativo debe sugerir el almacenamiento de las mercancías recibidas.

* + Tubería de movimiento de datos en producción
    - Se realiza la ingesta de datos a bucket raw por medio de sqoop.
    - Luego se realizan transformaciones a los datos y esta información queda alojada en el bucket processed.
    - Se realizan transformaciones de los datos al esquema definido en producción, esta información queda alojada en el bucket modeling.

Imagen que contiene camiseta, negro

Descripción generada automáticamente

**Comunicación**

* ¿Cómo nos mantendremos en contacto?

Se tienen proyectadas reuniones diarias para verificar el progreso de cada una de las tareas programadas.

* ¿Quiénes son las personas de contacto en ambos lados?

Por parte de Diagonal Colombia el gerente financiero Carlos Enrique Salazar Escobar que es el **PM.**

**Anexo B. Origen de datos**

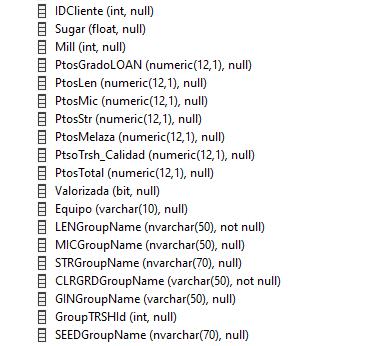
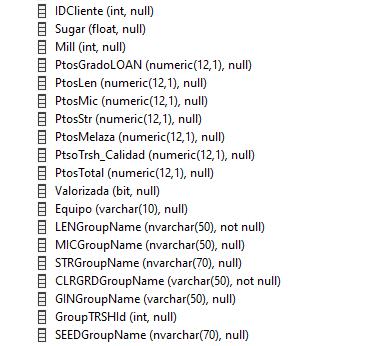
**DEFINICIONES DE DATOS Y CARACTERÍSTICAS**

**Fuentes de datos sin procesar:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Nombre del conjunto de datos** | **Ubicación original** | **Ubicación de destino** | **Herramientas de movimiento de datos / scripts** |
| Características de recibos de mercancías. | Diagonal\_DW en MSSQL 2016 Diagonal Colombia. | https://rawcsalazar.s3.us-east-2.amazonaws.com/ProyectoIntegrador/ | Script en Python |
| Características de despachos de mercancía. | Diagonal\_DW en MSSQL Diagonal Colombia. | https://rawcsalazar.s3.us-east-2.amazonaws.com/ProyectoIntegrador/ | Script en Python |
| Movimientos de pacas entre sitios de almacenamientos. | Diagonal\_DB en MSSQL Diagonal Colombia. | https://rawcsalazar.s3.us-east-2.amazonaws.com/ProyectoIntegrador/ | Script en Python |
| Localización histórica de las pacas en sitios de distribución. | Diagonal\_DB en MSSQL Diagonal Colombia. | https://rawcsalazar.s3.us-east-2.amazonaws.com/ProyectoIntegrador/ | Script en Python |

**Anexo C. Diccionario de datos**





**Anexo D. Descripción de los datos**

**Características de recibos de mercancías.**

|  |  |  |  |
| --- | --- | --- | --- |
| **Campo** | **Descripción.** | **Tipo de dato.** | **Tipo de dato cuantitativo.** |
| IdMuestra | Indicador propio de la fábrica que corresponde al id de la muestra, este dato es incremental. | Numérico (Entero). | Discreto |
| CodPaca | Identificación única del producto, está compuesto por la siguiente estructura:   * Los tres primeros dígitos corresponden al código del proveedor. * Los dos dígitos siguientes corresponden a un consecutivo anual de contratos, 01 y 02 identifica la primera y segunda cosecha nacional y 03 en adelante contratos internacionales. * Los cuatro dígitos siguientes corresponden al número del lote. * Los cuatro dígitos siguientes corresponden al número de la paca. * El digito siguientes corresponde al último número del año de la cosecha. | Numérico (Entero). | Discreto. |
| Lote | Identificador del lote o grupo de pacas que fueron desmotadas al mismo tiempo, este número está compuesto por cuatro dígitos. | Numérico (Entero). | Discreto. |
| Paca | Identificador del consecutivo de la paca en el lote desmotado, este número está compuesto por 4 dígitos. | Numérico (Entero). | Discreto. |
| IDEntrada | Identificador de la entrada al análisis de laboratorio. Llave primaria de la tabla “***Entradas***”. |  | Discreto. |
| IDCosecha | Identificador interno del laboratorio de la muestra que ingreso al laboratorio. Llave primaria de la tabla “***Cosechas***”. |  | Discreto. |
| CodProveedor | Identificador del proveedor que hace referencia a la asociación que envía el algodón, este número está compuesto por 3 dígitos y es la llave primaria de la tabla “***Proveedores***”. | Numérico (Entero). | Discreto. |
| FechaAmbientacion | Fecha en la que se mete a preparar la muestra en ambientación antes de entrar al análisis, este campo está compuesto por la siguiente estructura (AAAAMMDD). | Alfanumérico. | Discreto |
| FechaAnalisis | Fecha en la que se analiza la muestra, este campo está compuesto por la siguiente estructura (AAAAMMDD). | Alfanumérico. | Discreto. |
| Temperature | Condiciones de temperatura en grados centígrados que presenta el laboratorio al momento del análisis de la muestra. |  |  |
| Humidity | Condiciones de humedad relativa que presenta el laboratorio al momento del análisis de la muestra.  <https://www.diagonal.com.co/ClasificacionHVI/Humedad> | Numérico (Entero). | Discreto |
| Amount | Número de puntos no blancos que encuentra la máquina al tomar una foto en blanco y negro de la muestra. | Numérico (Entero). | Discreto |
| Length | Longitud de la muestra en pulgadas.  <https://www.diagonal.com.co/ClasificacionHVI/Longitud> | Numérico (Entero). | Discreto |
| EquivalenciaId | **Campo redundante** | Numérico (Entero). | Discreto |
| Uniformity | Medida de la uniformidad de la longitud de la muestra.  <https://www.diagonal.com.co/ClasificacionHVI/Uniformidad> | Numérico (Entero). | Discreto |
| Strength | Resistencia de la muestra en gramos por tex (medida textil).  <https://www.diagonal.com.co/ClasificacionHVI/Resistencia> | Numérico (Entero). | Discreto |
| Elongation | Medida del estiramiento de la muestra antes de romperse cuando es sometida a tracción.  <https://www.diagonal.com.co/ClasificacionHVI/Elongacion> | Numérico (Entero). | Discreto |
| SFI | Contenido de fibras cortas de longitud inferior a 0.5 pulgadas, en porcentaje del total de la muestra.  <https://www.diagonal.com.co/ClasificacionHVI/FibrasCortas> | Numérico (Entero). | Discreto |
| Grade | Grado es una valoración discreta que daban los clasificadores, o personas encargadas de determinar la calidad cuando esta determinación se hacía de manera manual y no por máquina. | Numérico (Entero). | Discreto |
| Moisture | Cantidad de humedad de la muestra al momento del análisis. | Numérico (Entero). | Discreto |
| Micronaire | Medida de la finura del algodón.  <https://www.diagonal.com.co/ClasificacionHVI/Micronaire> | Numérico (Entero). | Discreto |
| Rd | Grado de reflactancia del algodón, medida por la reflección de luz de la muestra. |  |  |
| Plusb | Grado de amarillamiento de la muestra, medida por la interacción con un haz de luz. |  |  |
| ColorGrade | Valor comercial establecido de una combinación del valor de Rd y PlusB.  <https://www.diagonal.com.co/ClasificacionHVI/Color> | Numérico (Entero). | Categórica |
| TrashCount | Medida de la cantidad de puntos de basura encontrados en una foto en blanco y negro de la muestra, por unidad de area. | Numérico (Entero). | Discreto |
| TrashArea | Medida del area de basura encontrados en una foto en blanco y negro de la muestra. | Numérico (Entero). | Discreto |
| TrashCode | Metrica que indica la cantidad de basura que lleva el algodón. Es una fórmula basada en el TrashCount y el TrashArea. Es un número de 1 a 7, donde indica el má lismpio y 7 indica el ms á sucio.  <https://www.diagonal.com.co/ClasificacionHVI/Basura> | Numérico (Entero). | Categórica |
| SCI | Métrica que indica el grado de hilabilidad estimado del hilo elaborado con el algodón testeado. Es una fórmula que incluye 7 características, pero en Colombia no es muy usado.  <https://www.diagonal.com.co/ClasificacionHVI/Hilabilidad> | Numérico (Entero). | Categórica |
| Maturity | Indica el grado de Madurez de la muestra (algoritmo propio).  <https://www.diagonal.com.co/ClasificacionHVI/Madurez> | Numérico (Entero). | Discreto |

**Anexo E. Informe de modelo**

Informe de modelo

## Para este proyecto se realizaron dos tipos de modelos, el primero es un modelo de agrupamiento no supervisado y el segundo es un modelo de clasificación supervisado. El modelo de agrupación se realizados con el fin de crear grupos basados en los despachos realizados hacia los clientes y el modelo supervisado se realizó con el fin de clasificar las pacas entrantes a la bodega de almacenamiento según los grupos formados previamente por el modelo k-medoids.

## Para este proyecto se realizaron dos tipos de modelos, el primero es un modelo de agrupamiento no supervisado y el segundo es un modelo de clasificación supervisado. El modelo de agrupación se realizados con el fin de crear grupos basados en los despachos realizados hacia los clientes y el modelo supervisado se realizó con el fin de clasificar las pacas entrantes a la bodega de almacenamiento según los grupos formados previamente por el modelo k-medoids.

# **Descripción del modelo**

# **K-Medoids**

## En el algoritmo k-medoides se encarga de realizar agrupaciones dividiendo el conjunto datos en k grupos, cada grupo tiene como representante uno de los puntos de datos en el grupo. Estos puntos se denominan medoides. La palabra medoide hace referencia a un objeto dentro de un grupo para el cual la diferencia promedio en términos de distancia entre este y todos los demás miembros del grupo es mínima. Corresponde al punto más centralmente ubicado en el grupo. Y este punto a su vez, es considerado como el elemento representativo de los miembros de ese grupo que pueden ser útiles en algunas situaciones.

En este proyecto, el modelo tiene como entrada un data set que contiene las calidades de cada paca por despacho.

## **RandomForest**

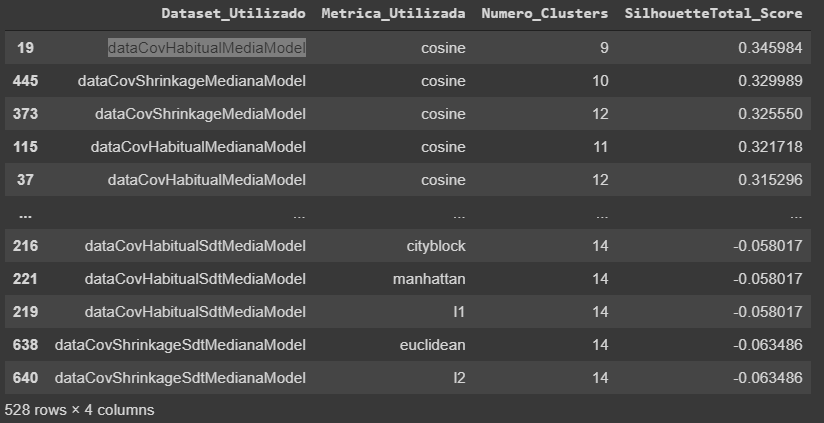
## Este es un algoritmo de aprendizaje supervisado que se encarga de crear y combinar aleatoriamente diversos árboles de decisión en un "bosque". La finalidad del algoritmo no es confiar en un solo modelo de aprendizaje, sino más bien en un conjunto de modelos de decisión para mejorar la precisión. La principal diferencia entre este enfoque y los algoritmos de árbol de decisión estándar es que los nodos raíz presentan nodos de división que se generan aleatoriamente.

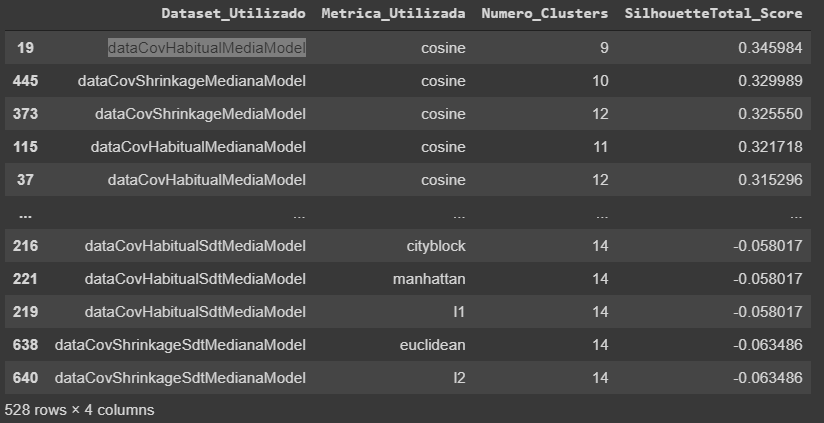
## En este proyecto, el algoritmo tiene como entrada en los datos X de entrenamiento las calidades de cada paca y como Y de entrenamiento el grupo al que pertenece cada paca proveniente del algoritmo k-medoids implementado con anterioridad. Toda la información de los datos tanto de entrenamiento como de test provienen del rango de fechas entre el 2017 y 2018.

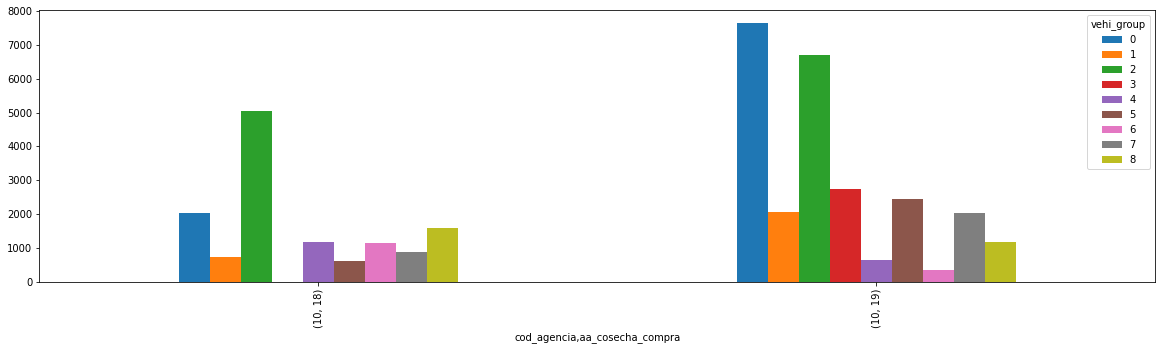
## **Resultado de rendimiento del modelo**

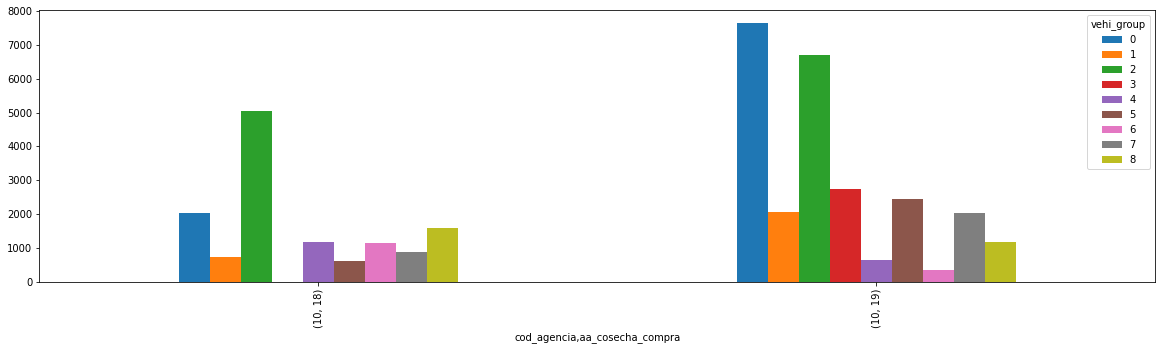
## **K-Medoids**

## En la siguiente tabla podemos observar los parámetros utilizados para este algoritmo.





cluster, el cual obtuvo un score de 0.3459 generado por el modelo de evaluación silhouette, el cual realiza una evaluación para determinar que tan bueno es el parámetro k.



En el anterior grafico se puede observar la distribución de los grupos que se generó al implementar el modelo para los datos de los años 2017 y 2018.

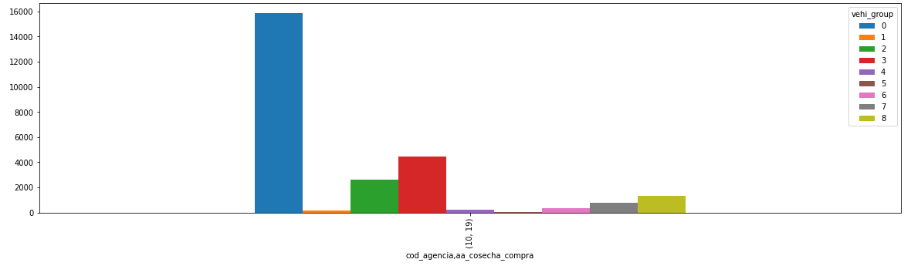
## **Random Forest**

El modelo RandomForest fue seleccionado debido a que presenta una distribución de grupo más homogénea a pesar de no ser el modelo con mayor reducción de distancias.

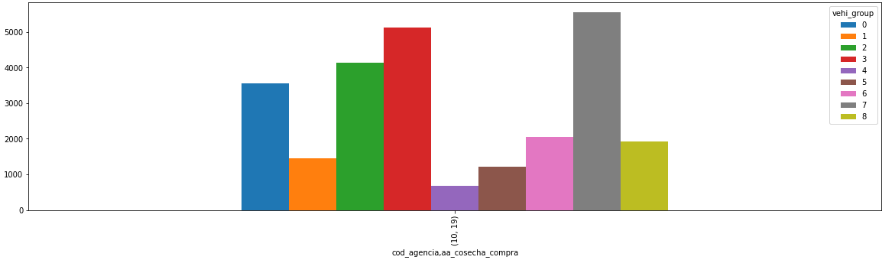
|  |  |
| --- | --- |
| Método | Métrica |
| data\_predicted | 32.7% |
| data\_predicted\_balanced | 30.7% |
| data\_predicted\_random\_forest | 31.6% |
| data\_predicted\_logistic\_regresion | 37.3% |
| data\_predicted\_decision\_tree | 31.4% |
| data\_predicted\_knn\_sin\_balanceo | 30% |

**Predicciones**

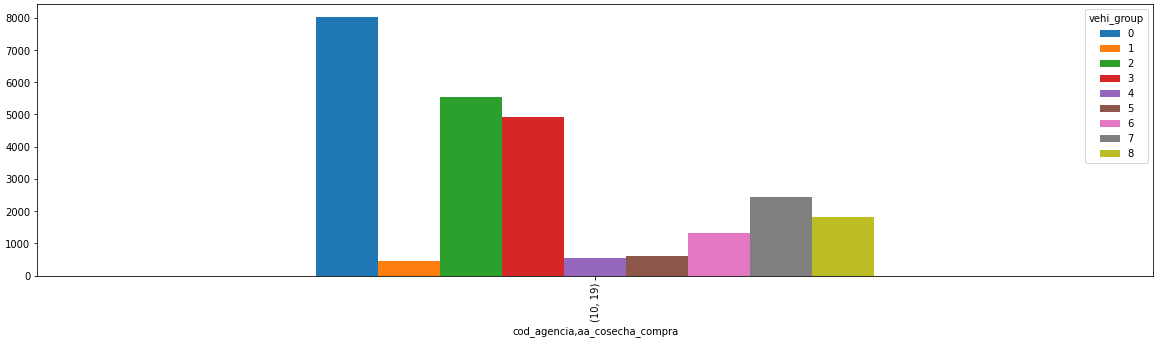
* **data\_predicted**



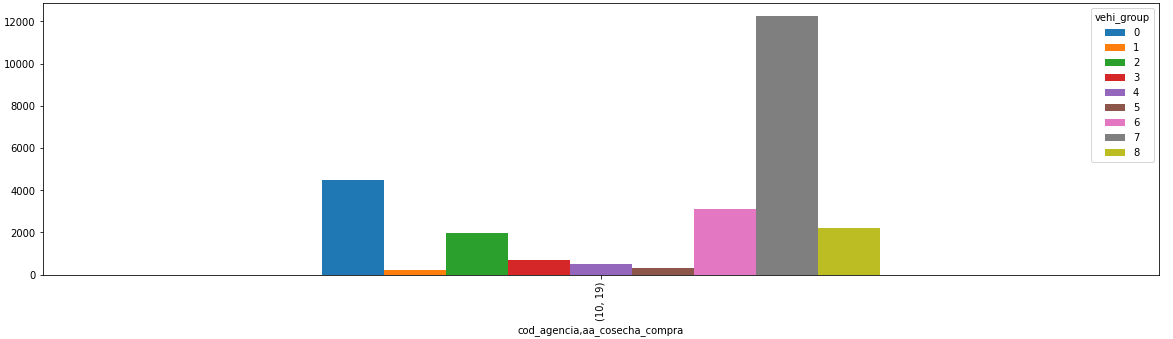


* **data\_predicted\_balanced**

**• data\_predicted\_random\_forest**

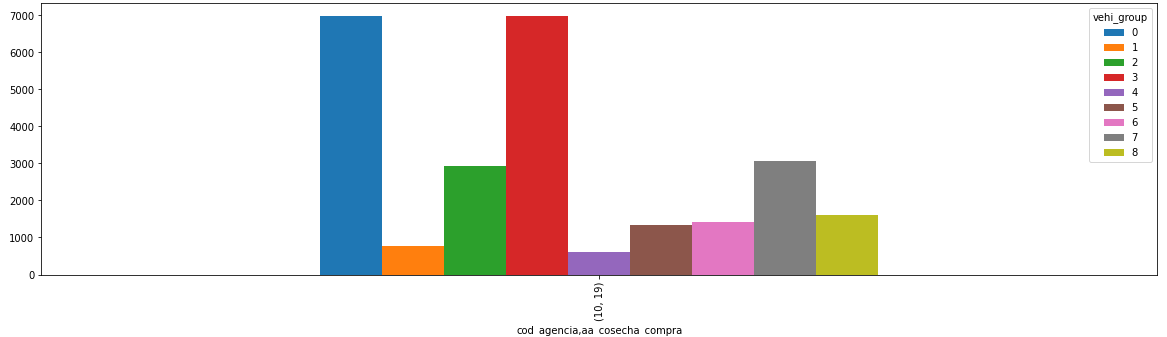
****

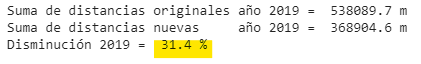
**• data\_predicted\_logistic\_regresion**

****

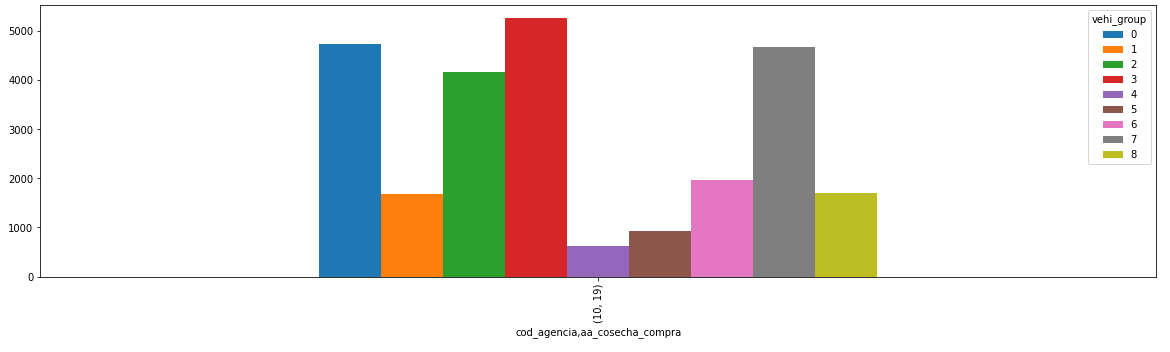
****

**• data\_predicted\_decision\_tree**

****

****

**• data\_predicted\_knn\_sin\_balanceo**

****

****