

|  |
| --- |
| **Universidad Internacional de La Rioja**  **Escuela Superior de Ingeniería y Tecnología**  **Máster Universitario en Análisis y Visualización de Datos Masivos** |
|  |
| Desarrollo de Modelos Analíticos para medir atención a Jóvenes con sanciones penales |

**Trabajo Fin de Máster**

**Tipo de trabajo:** Piloto Experimental

**Presentado por:** Celis Pineda, Diana Paola

**Director/a:** Garay Gallastegui, Luis Miguel

**Resumen**

**Palabras Clave:**

**Abstract**

**Keywords:**

**Índice de contenidos**

[1. Introducción 6](#_Toc87984348)

[1.1 Justificación 6](#_Toc87984349)

[1.2 Planteamiento del trabajo 7](#_Toc87984350)

[1.3 Estructura de la memoria 8](#_Toc87984351)

[2. Contexto y estado del arte 8](#_Toc87984352)

[3. Objetivos concretos y metodología de trabajo 14](#_Toc87984353)

[3.1. Objetivo general 14](#_Toc87984354)

[3.2. Objetivos específicos 14](#_Toc87984355)

[3.3. Metodología del trabajo 15](#_Toc87984356)

[4. Conclusiones y trabajo futuro 30](#_Toc87984357)

[4.1. Conclusiones 30](#_Toc87984358)

[4.2. Líneas de trabajo futuro 30](#_Toc87984359)

[5. Bibliografía 30](#_Toc87984360)

[Anexos 32](#_Toc87984361)

[Anexo I. Título anexo I 32](#_Toc87984362)

[Anexo II. Título anexo II 32](#_Toc87984363)

**Índice de tablas**

[Tabla 1 Diccionario de Datos 18](#_Toc88047139)

[Tabla 2 Cargar Información 20](#_Toc88047140)

[Tabla 3 ETL Dimensiones 22](#_Toc88047141)

[Tabla 4 STG1\_HECHOS 23](#_Toc88047142)

[Tabla 6 STG2\_Hechos 23](#_Toc88047143)

[Tabla 7 STG3\_Hechos 24](#_Toc88047144)

[Tabla 8 STG4\_Hechos 25](#_Toc88047145)

[Tabla 9 STG5\_Hechos 26](#_Toc88047146)

[Tabla 10 STG6\_Hechos 26](#_Toc88047147)

[Tabla 11 STG7\_Hechos 27](#_Toc88047148)

[Tabla 12 Fact\_corp 28](#_Toc88047149)

[Tabla 13 Estructura Final tabla FACT\_CORP 29](#_Toc88047150)

[Tabla 14 Variables de Salida. 29](#_Toc88047151)

**Índice de figuras**

[Ilustración 1 Cronograma de actividades 16](#_Toc88047155)

[Ilustración 2 Consentimiento tratamiento de datos 17](#_Toc88047156)

[Ilustración 3 Job Inicial. 19](#_Toc88047157)

[Ilustración 4 DWH Carga Información 19](#_Toc88047158)

[Ilustración 5 Validación Archivos 20](#_Toc88047159)

[Ilustración 6 ETL Validación Archivos 21](#_Toc88047160)

[Ilustración 7 ETL Dimensiones 22](#_Toc88047161)

[Ilustración 8 ETL STG1\_HECHOS 23](#_Toc88047162)

[Ilustración 9 ETL STG2\_HECHOS 24](#_Toc88047163)

[Ilustración 10 ETL STG3\_HECHOS 24](#_Toc88047164)

[Ilustración 11 ETL STG4\_HECHOS 25](#_Toc88047165)

[Ilustración 12 ETL STG5\_HECHOS 26](#_Toc88047166)

[Ilustración 13 ETL STG6\_HECHOS 27](#_Toc88047167)

[Ilustración 14 ETL STG7\_HECHOS 27](#_Toc88047168)

[Ilustración 15 ETL FACT\_CORP 28](#_Toc88047169)

# Introducción

El siguiente TFM (Trabajo Fin de Máster) pretende implementar modelos descriptivos, predictivos y la generación de visualizaciones que permitan medir la atención en los adolescentes y jóvenes que ingresan al programa "Atención de la sanción Semicerrado, Externado Media Jornada de Adolescentes y Jóvenes en Conflicto con la Ley Penal (Santander - Colombia)”.

Este proyectó se realizará como un plan Piloto en la Corporación Corpoadases “Alianza para el Desarrollo Ambiental Social y Económico Sostenible”, con los resultados que se obtengan en el análisis de la información se podrán desarrollar estrategias entre las áreas transversales, (Terapia Ocupacional, Psicología, Nutrición, Trabajo Social, Artes, Cultura y Deportes), lo cual permitirá agregar a la institución valor al trabajo que realizan con los jóvenes y adolescente mejorando así la calidad del servicio.

## Justificación

Actualmente la Corporación no tiene como medir el efecto en la atención en los Adolescentes y Jóvenes que ingresan al programa "Atención de la sanción Semicerrado, Externado media Jornada de Adolescentes y Jóvenes en Conflicto con la Ley Penal (Santander – Colombia)”, por qué la información no la tienen almacenada en un repositorio de datos, esta se encuentra en archivos de Excel y documentos físicos que están expuestos a daños y a pesar que se deben entregar reportes a los entes de control colombianos (ICBF Instituto colombiano del Bienestar Familiar – Procuraduría, Contraloría, Personería) para sus informes mensuales, trimestrales y anuales, toma tiempo y sobreesfuerzos en el equipo de trabajo para recompilar la información total de un Adolescente o Joven que se encuentran tratando.

La corporación quiere impactar en la vida de los adolescentes y jóvenes que ingresan al programa y con este proyecto además de almacenar la información en un repositorio de datos busca mediante el análisis de los datos un insumo que permita detectar puntos de mejora donde pueda definir: que paso, como paso, porque paso y que podría pasar con los adolescentes y jóvenes que se ingresa al programa.

El resultado del análisis de la información permitirá diseñar e implementación estrategias en la atención individual y grupal de los adolescentes y jóvenes que ingresan a la corporación, para que estos cambien de estilo de vida y puedan tener mejores oportunidades, disminuyendo de esta manera las posibilidades de reincidencia en el delito; también, Corpoadases quiere destacarse como una institución que se preocupa y ataca los problemas que pueda tener un Adolescente o Joven.

## 1.2 Planteamiento del trabajo

Lo que se quiere es aprender de los datos, por eso se tomara la información histórica del programa para procesarla y mediante modelos de datos estructurados e implementando técnicas para análisis de datos poder describir, predecir y generar visualizaciones que responsan las siguientes preguntas:

- Cuál es el Motivo de Ingreso de Jóvenes con sanciones penales.

- Cuál es la Edad de ingreso de Jóvenes con sanciones penales

- Cuánto es el Numero de reincidencias por periodo.

- Cuál es la tipología Familiar de los Jóvenes con sanciones penales

- Nivel de Consumo de los Jóvenes con sanciones penales.

- Nivel de Escolaridad de los Jóvenes con sanciones penales.

- Comparación el estrato Sociofamiliar Vs Delito.

- Comparación el nivel escolaridad Ingreso Vs nivel escolaridad Egreso.

- Porcentaje de Jóvenes que pasan de trabajo Informal a trabajo Formal.

- Cual es la relación del Motivo de Ingreso al programa con el Consumo de drogas.

- Cual es la relación del Nivel de escolaridad con Motivo de ingreso.

- Cual es la relación entre la tipología familiar y el consumo de drogas.

- Cual es la probabilidad de reincidencia de un Joven que tiene sanciones penales.

- Clasificacion del tipo de tratamiento que requiere un Joven con sanciones penales por el nivel de consumo de drogas.

- Definicion de áreas de refuerzo que requiera un joven que ha ingresado a la corporación.

¿Qué es lo que se propone?

**Lo cual le permitirá a la Corporación:**

Detectar el mayor motivo de ingreso al programa de Responsabilidad penal.

Cuantificar el número de reincidencias.

Conocer la tipología familiar de los Adolescentes y Jóvenes que ingresan al programa.

Detectar el nivel de consumo de los adolescentes y jóvenes que ingresan al programa.

Conocer el nivel de escolaridad.

Comparar el estrato sociofamiliar y el delito

Evaluar el nivel de escolaridad de los adolescentes y jóvenes que ingresan egresan del programa.

Detectar si un adolescente o joven puede reincidir en un delito.

Definir áreas de refuerzo que requiera un joven que ha ingresado a la corporación

Que le permitirá a la corporación poder reestructurar su plan de acción buscando la mejora constante en la calidad del servicio.

## 1.3 Estructura de la memoria

El resto de la Memoria se estructura de la siguiente manera:

En el capítulo 2 Se sitúa el lector en el contexto y estado del arte que se ha tomado como base para el desarrollo de este proyecto el cual abarca temas técnicos relacionados con la definición de análisis de datos, técnicas de machine Learning, proyectos y aplicaciones basadas en IA para la prevención y predicciones de delitos y consumo de drogas, además una definición de visualización de datos y las herramientas mas populares en el mercado.

En el capitulo 3 se abarca los Objetivos generales y específicos del alcance del proyecto y la metodología del trabajo usada para el desarrollo del proyecto.

En el Capítulo 4 se describe en detalle el análisis y planeación la definición de la necesidad de la corporación, análisis del negocio, las técnicas para el tratamiento de la información, limpieza de datos, desarrollo de los algoritmos y las visualizaciones.

En el Capítulo 5 Conclusiones y trabajo futuro se medirá si el proyecto Responde a la totalidad de las preguntas planteadas, se puedan generar visualización de información en Línea y si los modelos de predicción tienen una precisión del 70%, además se dejará planteado posibles líneas de trabajo a futuro para complementar el proyecto.

En el Capítulo 6 Bibliografía se listan las diferentes fuentes tomadas como referencia para el desarrollo de este proyecto.

Finalizado se encuentra el apartado de anexos donde se referencia anexos e información adicional el proyecto.

# Contexto y estado del arte

Descifrar el comportamiento de una persona ha sido un tema de trabajo en el trascurso del tiempo, de hecho existe la pseudociencia conocida como: La Frenología, que se conocía como la doctrina del cerebro la cual fue desarrollada por Franz Joseph Gall en el siglo XVIII que buscaban mediante la forma del cráneo descifrar como era una persona y sus comportamientos[1], basado en esto Cesare Lombroso que se considera el padre de la Criminología desarrolla teorías para definir los tipos de delincuentes que de acuerdo a su fisonomía, genero, entre otros factores, permitiera clasificar a una persona en los 6 tipos de criminales (Criminal nato, Delincuente loco moral, delincuente epiléptico, delincuente pasional, delincuente loco, delincuente ocasional ). [2]

Sin embargo, la frenología ha dejado de ser una disciplina de estudio pues para el desarrollo de las teorías no se usó un método científico, cuando se realizó pruebas experimentales se validaron que algunas de sus afirmaciones no podrían ser respaldadas ya que se basaba en observaciones y en evidencias anecdótica. [3]

En la actualidad se ha pedido no realizar publicaciones sobre algoritmos que mediante el uso de reconocimiento facial pueda detectar si una persona es criminal por la forma de su cara,[4] pues se consideran que llegan a ser racistas y generar injusticias sociales.[5].

Actualmente Existen aplicaciones que usan análisis de datos para predecir crímenes, consumo de drogas, reincidencia de un delito, a continuación, se mencionaran las más relevantes:

**Key Crime, “**la tecnología incursiona en la lucha contra los delitos en serie” [6].

Es una aplicación que mediante la recopilación de información de los delitos que se han cometido permite a la policía de Milán (Italia) realizar perfiles de criminales y realizando cruce de información implementar modelos para predecir, prevenir y compartir delitos.

“Este sistema de “predicción delictiva”, ha sido replicado a otros países, se inspiró en un modelo utilizado para los análisis de riesgo sísmico, el cual pretende predecir réplicas de temblores. En el ámbito criminal funciona igual, pues analiza las conductas reincidentes del delincuente” [6]

Este modelo sirvió de inspiración para “**El PredPol** (Predict Crime in Real Time) creado en el año 2011 por George Mohler. “Analiza lugares y horarios de crímenes cometidos recientemente y los combina con información sociológica y con datos de los patrones de comportamiento criminal, para prevenir delitos cometidos en tiempo real”

Este aplicativo tienen detractores quienes indican que el uso de estas tecnologías podría ocasionar “detenciones arbitrarias” [6].

Basado en esto en la ciudad de Bogotá (Colombia) existe un proyecto que busca predecir los crímenes con inteligencia artificial, “Según indicadores de la Secretaría de Seguridad, Convivencia y Justicia, quienes viven en Bogotá se sienten inseguros y esta percepción puede estar influenciada por los 14 hurtos que se producen cada hora, los más de 60 casos de lesiones personales denunciados y los 2,7 homicidios registrados a diario. “[7]**.**

Inspirados en las aplicaciones Key Crime y El PredPol, “Un equipo de matemáticos de la Universidad Nacional desarrolla un modelo de predicción de crímenes en Bogotá. Un sistema pionero en América Latina que impulsará la formación de talento humano en este campo.” [8]

Se recompilará la información que suministra la Policía Nacional, la Fiscalía, Línea de atención de emergencias (123), cámaras de vigilancia y buscará crear patrones para diseñar modelos predictivos “no ira detrás de las personas” [8].

La buena calidad y correcto procesamiento de la información será el éxito de este aplicativo, evitar sesgos es el mayor reto, pues en las aplicaciones Key Crime y El PredPol el uso de la variable Raza incidió en los resultados del análisis de la información por tal motivo se acusó falsamente a delincuentes de raza negra con mayor probabilidad de reincidir en un delito. [8].

EL uso de técnicas de IA (inteligencia Artificial) también se está implementando para atacar otros problemas de salud pública como lo es el consumo de drogas para esto en Ecuador se está desarrollando “algoritmos emergentes de inteligencia artificial para la predicción del consumo problemático y controlado de drogas a partir de información sociodemográfica, psicosocial y de salud; y la creación de un “observatorio para el estudio del alcohol y otras drogas” que garantice el acceso a información actualizada y la difusión periódica de los resultados” esto para desarrollar programas preventivos y generar políticas publicas para el consumo de drogas. [9]

**Análisis de información para medir atención a Jóvenes con sanciones penales**

En la ciudad de Bucaramanga (Santander – Colombia), las instituciones que trabajan con Jóvenes Infractores no cuentan con herramientas tecnológicas para el análisis de su información por tal motivo este Piloto permitirá a Corpoadases convertirse en una institución pionera en uso de análisis de datos para medir el servicio en la atención de los Jóvenes y Adolescentes que ingresan al programa de "Atención de la sanción semicerrado, Externado media Jornada de Jóvenes en conflicto con la Ley Penal.

La Corporación suministrara la data mediante la carga histórica de archivos en Excel y Se cargara en base de datos mediante procesos ETL.

**ETL: “**Como sus siglas indican, consiste en la extracción, transformación y carga de los datos en el **datawarehouse** (almacén de datos), de modo que se puede afirmar que es una parte fundamental de este. Antes de guardar los datos, deben ser transformados, limpiados, filtrados y redefinidos.”[9]

No se busca Juzgar a los jóvenes, lo que se quiere es analizar la información y con los resultados cuantificar las relaciones entre cada una de las variables y entender comportamientos de los jóvenes implementando con modelos preventivos patrones que ayuden a determinar si un joven reincidir en delitos, clasificar por los niveles de consumo el tipo de atención que requiere y ayudando a detectar las áreas de refuerzo.

Para evitar caer en algún tipo de sesgo social como ha sucedido en las aplicaciones mencionadas anteriormente se usarán en los algoritmos predictivos variables que describen características psicosociales descartando variables como: genero, etnia y raza para el entrenamiento de los modelos.

Procesada ya la información se realizará **análisis descriptivo** de los datos y basados en las preguntas realizadas por el equipo de trabajo de la corporación se usarán algunos modelos de Aprendizaje supervisado y no supervisados como:

**Random Forest:** “El bosque aleatorio es un algoritmo de aprendizaje supervisado que se utiliza tanto para la clasificacion y regresión. Sin embargo, se utiliza principalmente para cuestiones de clasificacion. Como sabemos, un bosque está formado por árboles y más arboles significa un bosque más robusto. Asimismo, el algoritmo de bosque aleatorio crea arboles de decisión sobre muestras de data, luego obtiene la predicción de cada uno de ellos y finalmente selecciona la mejor solución mediante votación. Es un método de conjunto que es mejor que un árbol de decisión único porque reduce el sobreajuste al promediar el resultado".[10]

Este modelo se selecciona ya que el equipo de trabajo quiere conocer si un joven puede o no reincidir en un delito, este algoritmo se acopla a lo que se requiere ya que es un modelo de clasificacion que a diferencia de un árbol de decisión reduce el sobreajuste en los resultados que pueda generar algún tipo de sesgo.

**Regresiones:** “Dos variables están relacionadas de una forma determinista, es decir, dado un valor de una variable, el valor de la otra variable se determina automáticamente sin error.

Mediante las regresiones se podrá no solo predecir el valor de una variable, también encontrar relación entre X (llamada variable explicativa, variable de predicción o variable independiente) y Y (llamada variable respuesta o variable dependiente). “[11]

para este TFM usáremos regresiones para validar si existe relación entre las variables que se han recolectado con la información brindada por la corporación

Hay diferentes tipos de modelos para hacer regresiones:

a. Regresión lineal simple: Variable dependiente continúa e independiente continúa.

b. Regresión lineal múltiple: Variable dependiente continua y variables múltiples independientes.

c. Regresión logística nominal: Variable dependiente binaria, independientes continuas o categóricas.

d. Regresión logística multinomial: Variable dependiente categórica, independientes continuas o categóricas.

e. Regresión logística ordinal: Variable dependiente categórica ordinal, independientes continuas o categóricas.

f. Regresión de poisson: Variable dependiente discreta, independientes continuas o categóricas.

g. Regresión de cox: Variable dependiente nominal y el tiempo como independiente

Dependiendo de los tipos de datos se podrá aplicar algunos de los modelos de regresión mencionados.

**Clúster:** “El clustering es uno de los métodos de aprendizaje no supervisado más importantes y, como cualquier otro método de aprendizaje no supervisado, busca caracterizar conceptos desconocidos a partir de instancias de estos. En este tipo de problemas de aprendizaje no supervisado la clase es desconocida y es, precisamente el descubrimiento de esta clase, el objetivo, a través de la agrupación de instancias en base a un esquema de similitud”. [12]

Esta técnica se aplicará en el proyecto ya que permitirá agrupar a los jóvenes para identificar patrones que ayudaran por ejemplo para definir los tipos de atención que requiere un joven dependiendo el nivel de consumo.

**Lógica Difusa:** “La lógica difusa o borrosa, también conocida como lógica fuzzy, es una lógica que nos permite procesar datos inciertos. Al contrario que la lógica booleana que sigue el principio bivalente o binario en el que la lógica solo permite dos estados posibles (verdadero y falso), la lógica difusa consigue tener infinitas degradaciones entre el valor verdadero y falso.” [13]

Este algoritmo se aplicará con el objetivo de detectar las áreas de refuerzo de los jóvenes,

Ya que esta enfocado a la toma de decisiones, teniendo en cuenta que algunos de los patrones que se detecte puedes ser ambiguos.

Con los resultados se creará una capa de visualizaciónmediante el uso de la herramienta Power BI

**Visualización de Datos**

“La visualización de datos es la presentación de datos en un formato pictórico o gráfico. Permite a los tomadores de decisiones ver los análisis presentados visualmente, para que puedan captar conceptos difíciles o identificar nuevos patrones.” [14]

Mediante gráficos se permitirá a las áreas trasversales de la corporación entender mejor los resultados del análisis de información que se está realizando, además de tomar las mejores decisiones para mejorar la prestación del servicio a los jóvenes con sanciones penales que ingresan a la corporación.

**Herramientas para la Visualizacion de Datos**

Las 7 herramientas de visualización de datos más populares son:

1. Tableu
2. Qlik
3. Ploty
4. CartoDB
5. Power BI
6. DataWrapper
7. DataStudio
8. Power BI

Para este piloto la Herramienta seleccionada es Power BI:

**Power BI:** “Microsoft Power BI, [o por sus siglas en inglés “Business Intelligence”](https://inbest.solutions/que-es-power-bi-de-office-365/), es un servicio de análisis de negocio basado en la nube que proporciona una vista única de los datos más críticos del negocio. [15]

Dentro de las ventajas de realizar las visualizaciones con esta herramienta tenemos las siguientes:

Información masiva, pero consolidada dentro de un mismo libro Excel. Se esta recompilando toda la información dentro de un modelo de datos y si en algún momento alguien del equipo de trabajo de la corporación quisiera realizar otras visualizaciones las podrían realizar y les agilizaría el trabajo que tienen que hacer.

Interfaz de lenguaje natural. Esto es poder realizar consultas a Power BI del mismo modo que solemos interactuar con otras personas al modo pregunta-respuesta.

Con capacidad de presentar de forma atractiva múltiples datos extremadamente complejos, en un único panel, realizar dashboard que podrán refrescar frente a una nueva carga de información y que podrán monitorear.

Puedes tener acceso a todos los datos donde quiera que vayas, desde Windows, iOS y Android. [15]

Se debe tener en cuenta que las personas que integran el equipo de trabajo no son técnicas sin embargo Power Bi es una herramienta intuitiva, pero se debe considerar por parte de los directores de la corporación capacitaciones sobre esta herramienta para que puedan obtener el mejor provecho, se debe considerar la compra de licencias de Power BI Pro, la empresa cuenta afiliación con Microsoft 365 que podría en un futuro mejorar el precio en la compra de licencias de esta herramienta.

## 3. Objetivos concretos y metodología de trabajo

## 3.1. Objetivo general

Desarrollar visualizaciones, modelos descriptivos y Predictivos que permitan medir la atención de los Jóvenes y Adolescentes que ingresan al programa de "Atención de la sanción semicerrado, Externado media Jornada de Jóvenes en conflicto con la Ley Penal (Santander - Colombia)”.

## 3.2. Objetivos específicos

- Recopilar la información histórica de la “Corporación (Alianza para el Desarrollo Ambiental Social y Económico Sostenible)” (Corpoadases) para ser disponibilizada para el insumo de datos para los modelos.

- Explicar mediante técnicas de Análisis de datos y Generación de visualizaciones los comportamientos de las variables recolectadas y sus características que permitan detectar tendencias.

- Relacionar el tipo de delito, situación sociofamiliar, nivel de estudio entre otras variables que se encuentren con la reincidencia y el consumo de drogas de los Jóvenes.

- Implementar modelos que permita predecir de acuerdo con las variables situación sociofamiliar, nivel de estudio, consumo de drogas si un Joven puede reincidir en un delito.

## 3.3. Metodología del trabajo

**3.3.1 CRISP-DM**

Cross-Industry Standard Process for Data Mining (CRISP-DM), es una metodología para dar organización a los proyectos analíticos, consta de los siguientes 6 pasos los cuales se espera que el desarrollo del proyecto sea efectivo:

**Comprensión del Negocio**

**Comprensión de datos**

**Preparación de datos**

**Modelado**

**Evaluación**

**Despliegue**

A continuación, se expondrá cada una de las fases de la metodología CRISP-DM aplicada al desarrollo del TFM.

**3.3.1.1 Comprensión del Negocio:** En esta fase se exploran las necesidades que tienen en la corporación, recopilando la información sobre la situación actual y como se puede disponibilizar para el desarrollo de los modelos de datos, permitiendo establecer los objetivos y el alcance para este Piloto.

Se realizaron sesiones de trabajo con las áreas transversales de la corporación y se obtiene realiza el siguiente cronograma de actividades.

Tabla

Descripción generada automáticamente

Ilustración 1 Cronograma de actividades

**3.3.1.2 Comprensión de datos:** Durante esta fase se le solicita a la corporación la información que tengan del programa Atención de la sanción semicerrado, Externado media Jornada de Jóvenes en conflicto con la Ley Penal (Santander - Colombia)”, la información entregada corresponde a archivos de Excel, para tratar los datos se cuenta con el vbo del director.

No se aplicarán técnicas de anonimización y disociaciones de los datos personales ya que la información que se recompile quedara disponible para que el personal del equipo de trabajo pueda generar informes, sin embargo, para los entrenamientos de los modelos no se usaran variables como el Nombre, Historia y Numero Documento de los jóvenes.

Analizado los datos más relevantes, se define las tablas de dimensiones y Hechos para la creación del DWH(DataWareHouse) para que la información quede disponibilizada en los modelos analíticos para que se pueden implementar de acuerdo con la información obtenida.

****

Ilustración 2 Consentimiento tratamiento de datos

A continuación, se describe cada uno los campos del archivo de Excel BASE:

|  |  |
| --- | --- |
| CAMPO | DEFINICIÓN |
| Nombre | Nombre del joven |
| Edad | Edad con la que el Joven ingresa al programa. |
| Historia | Número que el ICBF[[1]](#footnote-1) da al caso del Joven. |
| Defensor | Nombre defensor del Joven. |
| Equipo Psicosocial | Equipo asignado por el ICBF para la atención para el caso del Joven. |
| Juzgado | Nombre del Juzgado donde está el caso del Joven. |
| Asistente Social | Profesional que delega el juzgado para hacerle seguimiento a los casos de los chicos sancionados. |
| Delito | Nombre del delito con el que el Joven ingresa al programa. |
| Tiempo de sanción | Tiempo de sanción del Joven de acuerdo con el delito. |
| Documento | Numero de documento del Joven. |
| Ocupación | Ocupación actual del Joven. |
| Escolaridad Aprobada | Nivel escolaridad con la que ingresa el Joven. |
| Colegio | Nombre institución educativa. |
| Barrio | Nombre del barrio donde se encuentra el domicilio del Joven. |
| Sector | Nombre del Sector donde se encuentra el domicilio del Joven. |
| Jornada Programa | Jornada en la que el joven debe asistir a la corporación. (Mañana y/o Tarde) |
| Fecha Ingreso | Fecha de ingreso al programa. |
| Dirección | Dirección del acudiente. |
| Teléfono | Teléfono del acudiente |
| Responsable | Nombre del acudiente. |
| Numero Documento | Numero de documento del acudiente |
| Salud | EPS en la que se encuentra afiliado el joven. |
| Tutorías |  |
| Empleo | Indicador si el joven trabaja o no trabaja. |

Tabla 1 Diccionario de Datos

**3.3.1.3 Preparación de datos**

Para el desarrollo del procesamiento de la información se realizará usando el programa PENTAHO y para el almacenamiento el motor de base de datos POSTGRESQL.

**Inicio:** Job Inicial con el que arranca el procesamiento de información el cual Tiene como primer paso el arranque del proceso Star, continuando con el procesamiento de la información y finalmente la generación de los modelos Analíticos.

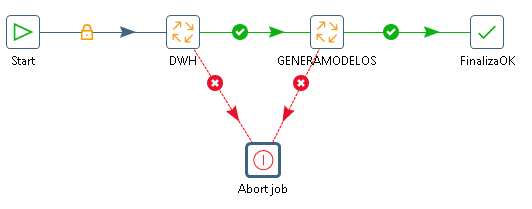


Ilustración 3 Job Inicial.

**TFM\_DWH:** Este paso se incluye el procesamiento de la información, se inicia con la validación de los archivos, continuando con la carga de las dimensiones y continua con la limpieza de datos.

Gráfico

Descripción generada automáticamente

Ilustración 4 DWH Carga Información

**Validación Archivos:** En este paso se realiza la validación del archivo de Excel entregado por la corporación.

Se busca un archivo con el Nombre BASE\_[enumerado].

Interfaz de usuario gráfica, Texto, Aplicación, Correo electrónico

Descripción generada automáticamente

Ilustración 5 Validación Archivos

**Cargar Información**

|  |  |
| --- | --- |
| **Proceso:** CargaInformacion.ktr | |
| **Origen** | Archivo Excel Base\_[Enumerado] |
| **Destino** | Public.COR\_IN\_BASE  Public.log\_estrcuturaarchivos |
| **Descripción** | Carga la información del archivo plano Base validando lo siguiente: **Numero Documento que sea numérico.**  **Tipo de sanción contenga los valores Meses, Días.**  Si alguno de los dos (2) campos no cumple con las reglas se enviará la tabla log\_estrcuturaarchivos, de lo contrario pasaran los registros en la tabla **Public. COR\_IN\_BASE.** |

Tabla 2 Cargar Información

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

Ilustración 6 ETL Validación Archivos

**Dimensiones**

|  |  |
| --- | --- |
| **Proceso:** Dimensiones.ktr | |
| **Origen** | dim\_delito.csv  dim\_tip\_familiar.csv  dim\_niv\_consumo.csv  dim\_escolaridad.csv  dim\_sit\_joven.csv  dim\_genero.csv  dim\_ocupacion.csv |
| **Destino** | Public.DIM\_DELITOS  Public. DIM\_TIP\_FAMILIAR  Public.DIM\_NIV\_CONSUMO  Public.DIM\_ESCOLARIDAD  Public.DIM\_SIT\_JOVEN  Public.DIM\_GENERO  Public.DIM\_OCUPACION  Public.LOG\_DIM\_DELITOS  Public.LOG\_DIM\_TIPOfAMILIAR  Public.LOG\_ DIM\_NIVELCONSUMO  Public.LOG\_DIM\_ESCOLAPR  Public.LOG\_DIM\_SITJOVEN  Public.LOG\_DIM\_GENERO  Public.LOG\_DIM\_OCUPACION |
| **Descripción** | En este proceso se crean archivos paramétricos de extensión .CSV, que contiene el campo id y descripción de las variables: Delitos, Tipo\_Familia, Nivel\_Consumo, Escolaridad, Situación\_Joven, Genero, Ocupación. Esto usado para los procesos de homologación y limpieza de los datos que se cargan desde los archivos Excel que suministra la corporación, finalizado el paso los registros se insertaran en las tablas Destino mencionados en este apartado. |

Tabla 3 ETL Dimensiones

Imagen que contiene Calendario

Descripción generada automáticamente

Ilustración 7 ETL Dimensiones

**STG1\_HECHOS**

|  |  |
| --- | --- |
| **Proceso:** STG1\_Hechos.ktr | |
| **Origen** | Public.cor\_in\_base  public.dim\_delitos |
| **Destino** | Public.STG1\_Hechos  Public.log\_stg1\_hechos |
| **Descripción** | Este proceso cruza la información de la tabla **COR\_IN\_BASE**, limpiando y validando el campo Nombre del Delito con la tabla **DIM\_DELITO,** los registros que no crucen se envían a una tabla(**log\_stg1\_hechos**) para que el equipo de trabajo pueda revisarlos y de ser necesario arreglarlos para volver a procesar la información, (Se tiene un control para evitar la inserción de duplicados), de lo contrato se insertan los registros en la tabla **STG1\_Hechos.** |

Tabla 4 STG1\_HECHOS

Escala de tiempo

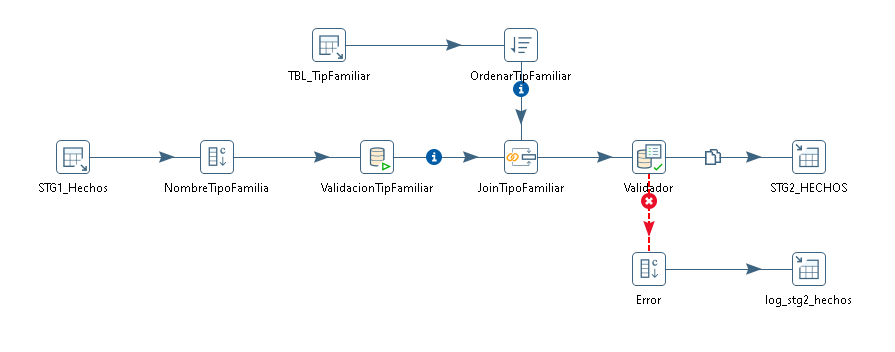
Descripción generada automáticamente

Ilustración 8 ETL STG1\_HECHOS

**STG2\_HECHOS**

|  |  |
| --- | --- |
| **Proceso:** STG2\_Hechos.ktr | |
| **Origen** | Public.STG1\_Hechos  public.dim\_tip\_familiar |
| **Destino** | Public.STG2\_HECHOS  Public.log\_stg2\_hechos |
| **Descripción** | Este proceso cruza la información de la tabla **STG1\_Hechos**, limpiando y validando el campo Nombre Tipo Familia con la tabla **DIM\_TIP\_FAMILIAR,** los registros que no crucen se envían a una tabla(**log\_stg2\_hechos**) que el equipo de trabajo pueda revisarlos y de ser necesario arreglarlos para volver a procesar la información, de lo contrato se insertan los registros en la tabla **STG2\_HECHOS.** |

Tabla 6 STG2\_Hechos

Ilustración 9 ETL STG2\_HECHOS

**STG3\_HECHOS**

|  |  |
| --- | --- |
| **Proceso:** STG3\_Hechos.ktr | |
| **Origen** | Public.STG2\_HECHOS  public.dim\_niv\_consumo |
| **Destino** | Public.STG3\_HECHOS  Public.log\_stg3\_hechos |
| **Descripción** | Este proceso cruza la información de la tabla **STG2\_Hechos**, limpiando y validando el campo nivel Consumo con la tabla **DIM\_NIV\_CONSUMO,** los registros que no crucen se envían a una tabla(**log\_stg3\_hechos**) que el equipo de trabajo pueda revisarlos y de ser necesario arreglarlos para volver a procesar la información, de lo contrato se insertan los registros en la tabla **STG3\_HECHOS.** |

Tabla 7 STG3\_Hechos

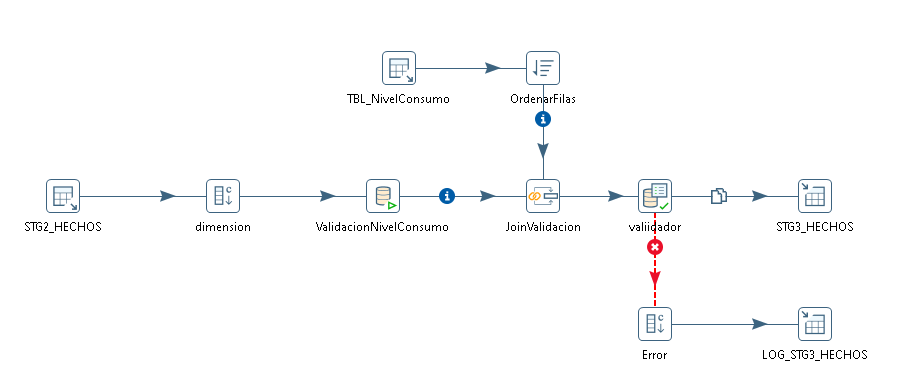


Ilustración 10 ETL STG3\_HECHOS

**STG4\_HECHOS**

|  |  |
| --- | --- |
| **Proceso:** STG4\_Hechos.ktr | |
| **Origen** | Public.dim\_sit\_joven  Public.STG3\_HECHOS |
| **Destino** | Public.STG4\_HECHOS  Public.log\_stg4\_hechos |
| **Descripción** | Este proceso cruza la información de la tabla **STG3\_Hechos**, limpiando y validando el campo situación Joven con la tabla **DIM\_SIT\_JOVEN,** los registros que no crucen se envían a una tabla(**log\_stg4\_hechos**) que el equipo de trabajo pueda revisarlos y de ser necesario arreglarlos para volver a procesar la información, de lo contrato se insertan los registros en la tabla **STG4\_HECHOS.** |

Tabla 8 STG4\_Hechos

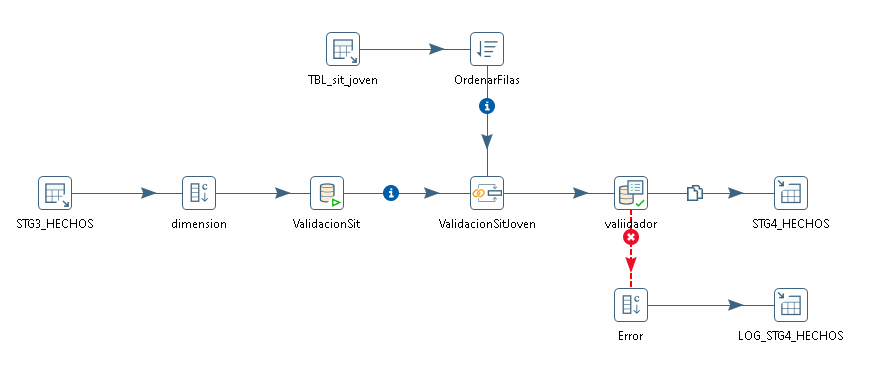


Ilustración 11 ETL STG4\_HECHOS

**STG5\_HECHOS**

|  |  |
| --- | --- |
| **Proceso:** STG5\_Hechos.ktr | |
| **Origen** | Public.dim\_genero  Public.STG4\_HECHOS |
| **Destino** | Public.STG5\_HECHOS  Public.log\_stg5\_hechos |
| **Descripción** | Este proceso cruza la información de la tabla **STG4\_Hechos**, limpiando y validando el campo genero con la tabla **DIM\_GENERO,** los registros que no crucen se envían a una tabla(**log\_stg5\_hechos**) que el equipo de trabajo pueda revisarlos y de ser necesario arreglarlos para volver a procesar la información, de lo contrato se insertan los registros en la tabla **STG5\_HECHOS.** |

Tabla 9 STG5\_Hechos

Diagrama

Descripción generada automáticamente

Ilustración 12 ETL STG5\_HECHOS

**STG6\_HECHOS**

|  |  |
| --- | --- |
| **Proceso:** STG6\_Hechos.ktr | |
| **Origen** | Public.dim\_escolaridad  Public.STG5\_HECHOS |
| **Destino** | Public.STG6\_HECHOS  Public.log\_stg6\_hechos |
| **Descripción** | Este proceso cruza la información de la tabla **STG5\_Hechos**, limpiando y validando el campo escolaridad con la tabla **DIM\_ESCOLARIDAD,** los registros que no crucen se envían a una tabla(**log\_stg6\_hechos**) que el equipo de trabajo pueda revisarlos y de ser necesario arreglarlos para volver a procesar la información, de lo contrato se insertan los registros en la tabla **STG6\_HECHOS.** |

Tabla 10 STG6\_Hechos

Diagrama

Descripción generada automáticamente

Ilustración 13 ETL STG6\_HECHOS

**STG7\_HECHOS**

|  |  |
| --- | --- |
| **Proceso:** STG7\_Hechos.ktr | |
| **Origen** | Public.dim\_ocupacion  Public.STG6\_HECHOS |
| **Destino** | Public.STG7\_HECHOS  Public.log\_stg7\_hechos |
| **Descripción** | Este proceso cruza la información de la tabla **STG6\_Hechos**, limpiando y validando el campo ocupación con la tabla **DIM\_OCUPACION,** los registros que no crucen se envían a una tabla(**log\_stg7\_hechos**) que el equipo de trabajo pueda revisarlos y de ser necesario arreglarlos para volver a procesar la información, de lo contrato se insertan los registros en la tabla **STG7\_HECHOS.** |

Tabla 11 STG7\_Hechos

Diagrama

Descripción generada automáticamente

Ilustración 14 ETL STG7\_HECHOS

**FACT\_CORP**

|  |  |
| --- | --- |
| **Proceso:** FACT\_CORP.ktr | |
| **Origen** | Public.STG7\_HECHOS  Public.fact\_corp |
| **Destino** | Public.fact\_corp |
| **Descripción** | Este proceso inserta los registros con la información transformada y homologada en la tabla **FACT\_CORP**, antes se realiza una validación para controlar los duplicados. |

Tabla 12 Fact\_corp

Diagrama

Descripción generada automáticamente

Ilustración 15 ETL FACT\_CORP

**Tabla Final**

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **Nombre Columna** | **Tipo** | **Tabla Asociada** |
| 1 | ID | Double | Número de registro |
| 2 | DELITO | Text | public.dim\_delitos |
| 3 | NOM\_APE | Text |  |
| 4 | NUM\_DOC\_ID | Text |  |
| 5 | PROGRAMA | Text |  |
| 6 | ANNO\_INGRESO | Text |  |
| 7 | MES\_INGRESO | Text |  |
| 8 | GENERO | Text | public.dim\_genero |
| 9 | EDAD | Text |  |
| 10 | ESCO\_APRO | Text | public.dim\_escolaridad |
| 11 | ESCO\_ACT | Text | public.dim\_escolaridad |
| 12 | OCUPACION | Text | public.dim\_ocupacion |
| 13 | AUTORIDAD | Text |  |
| 14 | NUM\_JUEZ | Text |  |
| 15 | TIEMPO\_SAN | Text |  |
| 16 | SALUD | Text |  |
| 17 | REGIMEN | Text |  |
| 18 | TIP\_FAMILIAR | Text | public.dim\_tip\_familiar |
| 19 | BARRIO | Text |  |
| 20 | MUNICIPIO | Text |  |
| 21 | NIV\_CONSU | Text | public.dim\_niv\_consumo |
| 22 | SIT\_ATUAL | Text | public.dim\_sit\_joven |
| 23 | ARCHIVO | Text | Nombre Del Archivo para control. |
| 24 | FEC\_DATA | Text | Fecha Procesamiento |

Tabla 13 Estructura Final tabla FACT\_CORP

**3.3.1.4 Modelado:** En estaFase se seleccionarán y aplicarán las técnicas de modelado mencionados (métodos estadísticos, aplicación de técnicas de Machine Learning).

Se identifican las siguientes variables como salidas para los modelos:

|  |  |
| --- | --- |
| TIPO\_TRATAMIENTO | Text |
| POR\_ATEN\_PEDA | Int |
| POR\_ATEN\_TOCU | Int |
| POR\_ATEN\_PSICO | Int |
| POR\_ATEN\_TRAB\_SOCI | Int |
| REIN\_DELITO | Int |

Tabla 14 Variables de Salida.

**TIPO\_TRATAMIENTO**

Se define un modelo para realizar la clasificacion del tipo de tratamiento que requiere el joven basados por el nivel de consumo.

Su clasificacion es la siguiente:

* Consumo abusivo - requiere tratamiento intramural
* Consumo disfuncional - tratamiento ambulatorio
* Consumo experimental/social - citas de seguimiento con psicología
* No consume - participación en actividades de prevención
* Abstinencia - participación en grupo de apoyo

**REIN\_DELITO**

Este modelo se desarrollará para clasificar si un joven podrá incidir en un delito basados en las siguientes variables: delito, tipo familia, nivel consumo, escolaridad aprobada y Ocupación.

Las salidas de modelo será una clasificacion:

* 1: Si
* 0: No

**AREAS DE REFUERZO**

Se requiere desarrollar un modelo el cual se pueda por Joven definir de acuerdo con las variables: Delito, Programa, Edad, Escolaridad Aprobada, Escolaridad Actual, Ocupación, Tiempo Sanción, Tipo Familia, Nivel de consumo los porcentajes de atención que requiera de cada una de las áreas de trabajo (esto basado en las reglas que definen en el equipo de trabajo de la corporación):

**PEDAGOGICA**

**TERAPIA OCUPACIONAL**

**PSICOLOGICA**

**TRABAJO SOCIAL**

**3.3.1.5 Evaluación:** Durante esta etapaya se han construido los modelos por lo cual nos va a permitir evaluar los resultados revisando de los datos para trabajos futuros.

**3.3.1.6 Despliegue:** En esta fase se entregarán los resultados del análisis de datos mediante la generación de las visualizaciones en POWER BI, adicional se entregará una -BAT llamado EjecutarEtls.bat, con el cual se podrán ejecutar los procesos de ETL y generación de los modelos y visualización de los resultados.

# 4. Conclusiones y trabajo futuro

## 4.1. Conclusiones

## 4.2. Líneas de trabajo futuro

# 5. Bibliografía

[1] Walter L. Arias G, La frenología y sus implicancias: un poco de historia sobre un tema olvidado

[2]<< Tipos de delincuentes, según Lombroso>> [En Línea].

Available: <https://www.unir.net/derecho/revista/tipos-de-delincuentes-criminales/>

[3] << Frenología: Historia y Frenólogos Destacados.>> [En Línea].

Available: <https://www.lifeder.com/frenologia/>.

[4] << El polémico programa de ordenador que predice si eres un criminal por tu aspecto: "Tienes cara de malo">>[EnLínea]

Available: <https://www.elmundo.es/tecnologia/2020/06/29/5ef9243721efa04e418b459e.html>

[5] <<1.700 expertos condenan la IA que predice crímenes basándose en el rostro por ser socialmente injusta>>[EnLínea]

Available: <https://www.europapress.es/portaltic/sector/noticia-1700-expertos-condenan-ia-predice-crimenes-basandose-rostro-ser-socialmente-injusta-20200624135813.html>

[6] << ‘Key Crime’, un software que predice los crímenes>>[EnLínea]

Available:<https://www.semana.com/mundo/articulo/key-crime-un-software-que-predice-previene-crimenes/364708-3/>

[7] << En Bogotá se predecirán los crímenes con inteligencia artificial>>[EnLínea]

Available:<https://bogota.gov.co/mi-ciudad/seguridad/modelo-de-prediccion-de-crimenes-en-bogota>

[8] << La Universidad Nacional ayudará a predecir crímenes en Bogotá>>[EnLínea]

Available: <https://www.semana.com/educacion/articulo/la-universidad-nacional-y-la-alcaldia-de-bogota-trabajan-en-un-sistema-de-prediccion-de-crimenes/609920/>

[9] Toan C. Ong, 2017

[10] <<tecnicas-de-procesamiento-automatico-aplicadas-al-analisis-y-prediccion-del-consumo-de-drogas>>[EnLínea]

Available: <https://www.cedia.edu.ec/es/proyectos-ganadores/cepra-xii/tecnicas-de-procesamiento-automatico-aplicadas-al-analisis-y-prediccion-del-consumo-de-drogas>

[11] Mario F. Triola, Correlación y Regresión.

[12] Universidad Internacional de La Rioja, Clustering: Agrupamiento o clasificación no supervisada

[13] << Fundamentos de la lógica difusa>>[EnLínea].

Available: <https://shakaran.net/blog/2008/12/fundamentos-de-la-logica-difusa-fuzzy-logic/>

[14] <<AI en 3 minutos: Tipos de Machine Learning>>[EnLínea].

Available: <https://www.sas.com/es_mx/insights/big-data/data-visualization.html>

[15] << 10 ventajas de Microsoft Power BI que unificarán datos de tu empresa

>>[EnLínea].

Available <https://inbest.solutions/10-ventajas-de-microsoft-power-bi-que-unificaran-datos-de-tu-empresa/>

# Anexos

## Anexo I. Título anexo I

## Anexo II. Título anexo II

1. ICBF (Instituto colombiano de Bienestar Familiar) [↑](#footnote-ref-1)