MAESTRÍA INGENIERÍA COMPUTACIONAL

Modelo de minería de datos en la empresa INTEGRA S.A., operador de transporte masivo del Área Metropolitana Centro Occidente para definir patrones de conducción de los operadores de los buses articulados y alimentadores

Ing. Jorge Augusto Sánchez Jácome
Univesidad de Caldas
Facultad de Ingenierías, Departamento de Sistemas
Manizales
2022

MAESTRÍA INGENIERÍA COMPUTACIONAL

Modelo de minería de datos en la empresa INTEGRA S.A., operador de transporte masivo del Área Metropolitana Centro Occidente para definir patrones de conducción de los operadores de los buses articulados y alimentadores

Ing. Jorge Augusto Sánchez Jácome

Director
M.Sc. Jairo Iván Vélez
CoDirector
Ph.D. Luis Miguel Escobar Falcón

2022

Resumen

Agradecimientos

Agradezco a mi esposa Sonia y a mi hija Celeste por todo el apoyo en la elaboración de esta investigación. Su entendimiento y sacrificio me inspiraron a ser un mejor profesional y obtener más conocimiento.

De manera profunda agradezco a la empresa INTEGRA S.A. en cabeza de Ramón Antonio Toro Pulgarín, al Gerente de I+D+i, César Augusto Marín Moreno y a la Gerente Administrativa y Financiera, María Elena Vélez Taborda ya que su confianza permitió que pudiera realizar esta maestría y obtener estos resultados.

Tabla de contenido

Indice de figuras

Indice de Tablas

Capítulo 1

Introducción

La ciudad, día a día va cambiando, transformándose de un pueblo pequeño a una gran metrópoli, donde el transporte viene desempeñando un papel estructurante en las mismas, determinando los sitios donde se puede poblar la ciudad más densamente a través de la concurrencia de rutas y la cobertura de las necesidades de la sociedad, este dinamismo hace cada vez más imperante que haya una movilidad sostenible en las grandes y medianas urbes, buscando proteger el medio ambiente, reduciendo los tiempos de viaje y dignificando a la persona en su necesidad por transportarse. En Colombia esto no es ajeno, por esta razón desde el gobierno nacional se asumió la necesidad de desarrollar un documento construido por el Consejo Nacional de Política Ecónomica y Social (CONPES), donde se introduce, reglamenta y se traza la ruta de implementación de sistemas de transporte público masivo de pasajeros de buses de tránsito rápido (BRT por su nombre en inglés *Bus Rapid Transit System*), la principal característica de los sistemas BRT es que cuentan con un carril exclusivo para la operación de buses con alta capacidad de integración física y tarifaria, contando con rutas alimentadoras, las cuales acercan a las personas a las estaciones troncales, todo con el objetivo de dar solución a las necesidades de transporte eficiente y sostenible en las ciudades que tienen una cantidad considerada de habitantes.

El auge tecnológico que arrancó en la época de los 90 ha transformado poco a poco cada una de las actividades diarias de la sociedad. Tecnologías como el wifi, bluetooth, GPS, red móvil de datos han permitido interconectar al mundo y a las personas, pero las personas no son las únicas que han sido relacionadas, los sistemas informáticos también lo han sido y se han integrado con la ciudad. Las tecnologías de la información y Comunicación (TIC) en los últimos años han ayudado a tecnificar la industria del transporte, por medio de la inclusión de dispositivos GPS que permiten monitorear y controlar en tiempo real la flota lo cual ayuda para mejorar la eficiencia y cumplimiento en los servicios programados, plataformas de comunicación bidireccional por medio de tabletas electrónicas conectadas a los GPS, permiten a las empresas operadoras de los

sistemas BRT conocer la realidad de la ciudad en temas de movilidad y tomar decisiones que estén en pro de prestar un servicio ideal.

En el Área Metropolitana Centro Occidente (AMCO), de la cual hacen parte las ciudades de Pereira, Dosquebradas y la Virginia se cuenta con Megabús, un BRT controlado por un ente gestor en cuya estructura existe un operador de transporte (INTEGRA S.A.) que se ha encargado durante los últimos 10 años de incluir tecnologías de software y de hardware que permitan prestar un mejor servicio, disminuir los tiempos de viajes, optimizar los recursos finitos tales como combustible, llantas entre otros, por medio del almacenamiento de información de operación de la flota cada tres (3) segundos en un sistema de información conocido como InnoBUS Masivo.

InnoBUS Masivo va en búsqueda de simular todo el comportamiento de los procesos pilares de la organización, talento humano, operaciones y mantenimiento, para que, por medio del mismo; tareas que se hacían de manera repetitiva, acudiendo a la mecánica y apoyo de múltiples herramientas que dispersan la información y generan reprocesos y retrabajos con una técnica que puede ser manipulable o de forma no muy tecnificada se pueda en un futuro realizar de forma automática y que permita a la empresa seguir en la mejora continua de la operación.
En este orden de ideas la inclusión de tecnologías o metodologías que están en auge en el mundo empresarial son necesarias para sacar el máximo provecho de los datos que están alojados en la base de datos y transformarlos en información de valor para la toma de decisiones de la compañía, por eso la minería de datos, la analítica de datos o como en este caso servirá para premiar o castigar al operador según su parámetro de conducción.

1.1 Planteamiento del Problema

¿De que forma se puede caracterizar la población de Operadores de INTEGRA S.A. usando la minería de datos para definir patrones de conducción en los buses articulados y alimentadores del Área Metropolitana Centro Occidente?

INTEGRA S.A. es una de las empresas operadoras del sistema de transporte público masivo de pasajeros del Área Metropolitana Centro Occidente AMCO en los municipios de Pereira, La Virginia y Dosquebradas (MEGABUS, 2005), desde su visión, la compañía se ha trazado el objetivo de ser una empresa de clase mundial, reconocida por la innovación en modelos de negocio aplicados a la cadena de valor en la industria del transporte (Colciencias, 2016), para ello desde su área de Investigación, Desarrollo e Innovación (I+D+i) se ha dedicado, desde hace más de 10 años al interior de la empresa, a desarrollar proyectos de investigación, desarrollo aplicado e innovación, creando una red de conocimiento con diferentes actores de la comunidad académica y empresarial tanto en el ámbito nacional como internacional.

Es así como en el año 2012, a través del convenio 373 de 2011 (INTEGRA-SENA, 2011) firmado entre Integra S.A. y el Servicio Nacional de Aprendizaje SENA se desarrolló e implementó un software denominado InnoBUS Masivo, el cual facilita la gestión de las áreas de planeación, operación, mantenimiento y talento humano de la empresa.

InnoBUS Masivo permite controlar de manera centralizada la operación diaria de la compañía por medio de un monitoreo constante de la flota de vehículos con los que cuenta actualmente (37 buses articulados y 42 buses alimentadores); cada vehículo está equipado con un dispositivo GPS que reporta cada 3 segundos datos de posicionamiento de latitud y longitud, velocidad, aceleración, odómetro, entre otros, por medio de una conexión de plan de datos verticales usando la red móvil.

Dado que la información es enviada cada 3 segundos durante un periodo de 20 horas diarias de operación, se requiere almacenar un total de 24.000 tramas por vehículo para un total de 1.896.000 tramas al día por todos los buses de la empresa, en este sentido, la base de datos de la compañía crece a un ritmo de 300MB por día, lo que es una cantidad considerable de información. Sin embargo, este gran volumen de datos no genera ningún valor comercial u operacional, ya que actualmente la compañía carece de un sistema integral soportado en técnicas de *Big Data* o *Data Mining* que le permitan realizar el tratamiento adecuado de los datos para convertirlos en información que genere valor para la empresa.

Así las cosas, la plataforma InnoBUS Masivo se ha convertido en un eje fundamental al interior de la organización para recolectar y centralizar información digital en grandes volúmenes, optimizando tareas diarias de la empresa, pero esa misma recolección de información se ha convertido de manera simultánea en un nuevo problema que impone nuevos retos, ya que la búsqueda de información concreta, de resultados o de mediciones por medio del actual proceso de modelación se basa en la elaboración de consultas a la base de datos del sistema, que al realizarlas reduce el rendimiento del servidor afectando la operación diaria.

Como se menciona en [Rolim et al., 2017b] los patrones de conducción están determinados por 4 grandes grupos (las características del conductor, la tipología del vehículo, los patrones de movilización tales como tiempos de viaje y condiciones de tráfico entre otros, y por último los indicadores del medio ambiente).

Si en el transporte colectivo se evaluaba al conductor por el valor del producido diario [Castaño Vanessa, 2013], en el transporte masivo las mejoras sustanciales en la tipología del vehículo obliga a las empresas no solo a evaluar la eficiencia del operador desde el cumplimiento de servicio, sino también desde el rendimiento de combustible, tiempos de servicio (efectividad), conducciones

agresivas, excesos de velocidad, niveles de accidentalidad.

Actualmente InnoBUS muestra informes gerenciales que no se acomodan a las necesidades actuales de la compañía, basándose en todos los datos contenidos en las más de 256 GB de información con la que se cuenta en este momento acumulada entre los años 2015 y 2016.

InnoBUS Masivo necesita informes gerenciales que se construyan a partir de los datos alojados, sirviendo como un insumo importante para el actual modelo de bonificación de los Operadores de INTEGRA S.A., de tal manera que se describa un promedio estándar de puntajes a distintas variables basados en datos históricos de la operación de INTEGRA S.A. almacenado en la base de datos y que este promedio permita conocer el comportamiento de conducción para un operador en un mes específico Tal cual fue diseñado, la falta de aplicación de conceptos como *Big Data* o *Data Mining* en la base de datos de InnoBUS Masivo no permite en la actualidad maximizar al 100% su uso, lo cual ayudaría a la empresa convertir sus datos en información realmente relevante, y a convertir a InnoBUS MASIVO de una gran base de datos a una base de información al interior de la organización que permita aplicar nuevas metodologías, no solo para situaciones como las mencionadas sino para nuevos requerimientos que se generen de la aplicación de Inteligencia de Negocios y descubrimiento de información. Vale la pena resaltar que es precisamente en la búsqueda de la innovación de sus procesos y de generación de valor agregado que se aúnan los esfuerzos de Integra S.A. a fin de orientar de mejor manera sus esfuerzos optimizando sus recursos.

1.2 Justificación

En Integra S.A se cuenta con una base de datos que recolecta información relacionada con los vehículos que se encuentran en operación, pero no se encuentra realizando dentro de la empresa algún tipo de aprovechamiento de estos datos utilizando el *Data mining* o el *BIG DATA*. Estos datos se recolectan a través InnoBUS Masivo, pero no se está obteniendo ninguna información. Donde revisando la literatura especializada, se puede encontrar que en empresas de transporte similares a Integra S.A, se han aplicado modelos en arquitecturas similares, donde se aprovecha La aplicabilidad de estos modelos en arquitecturas similares a las descritas anteriormente han comprobado tal cual como se relaciona en [Viswanathan, 2013] donde se realiza un análisis de uso y los parámetros de conducción para medir las variables que afectan el consumo de combustible de una forma directa en buses con chasis similares con los que cuenta actualmente INTEGRA S.A. la principal diferencia de este estudio con los buses de la empresa es la marca.

También en [Nagar et al., 2016] se habla de la importancia que tiene la In-

teligencia de Negocio por medio del *Data Mining* para fortalecer las capacidades de respuesta empresariales, ofrecer más información a los niveles de gerencia con tiempos de respuesta menores. Por mandato de la alta dirección de INTEGRA S.A. para la planeación estratégica 2017 – 2021 InnoBUS Masivo se debe convertir en un ente centralizador de todas las actividades diarias de la organización realizándose una mejora continua sobre el mismo (INTEGRA S.A., 2013). La implementación de *Data Mining* es un objetivo dentro de la planeación estratégica del Proceso de TI buscando fortalecer debilidades en el manejo de la información y la gestión del conocimiento cerrando brechas tecnológicas y realizando conexiones con otras investigaciones realizadas anteriormente como lo es por ejemplo la tesis de maestría "Evaluación de eficiencias relativas en el desempeño sostenible de los operadores en un sistema BRT" desarrollada por la Magister en gerencia de Operación y Gerente del proceso de Operaciones de INTEGRA S.A. Vanessa Castaño Bañol donde se busca dar los insumos necesarios a través de la aplicabilidad de un modelo de Minería de Datos o Data Mining por sus siglas en Inglés y poder realizar un tratamiento de datos que generen una información significativa a la empresa.

Es por esto que este trabajo toma relevancia, donde se busca utilizar los datos recolectados por el software InnoBus, utilizando técnicas de análisis de datos, construyendo metodologías, que puedan ser replicables en la organización y la ayude a tomar mejores decisiones desde la planeación operativa de la empresa.

1.3 Objetivos

1.3.1 Objetivo General

Implementar un modelo aplicable a la base de datos InnoBUS Masivo propiedad de la operadora de Transporte Masivo INTEGRA S.A. en el Área Metropolitana Centro Occidente que permita encontrar patrones de comportamiento de los operadores de los vehículos articulados y alimentadores.

1.3.2 Objetivos Específicos

- Determinar las variables que afectan los comportamientos de conducción normal de los operadores de los vehículos articulados y alimentadores a través del entendimiento de las necesidades de la empresa.
- Generar un modelo utilizando técnicas de *Clustering* que permita la obtención de los comportamientos de conducción normal de los operadores de los vehículos articulados y alimentadores usando las variables previamente definidas.

- Validar el modelo generado comprobando la confiabilidad del mismo por medio de un proceso estadístico.
- Interpretar los resultados obtenidos en conjunto con el grupo de expertos de INTEGRA S.A. basado en los conocimientos operacionales para definir las características de la población.

1.4 Estructura del Documento

El presente documento está organizado de la siguiente forma: En el capítulo 1 se habla de la introducción, el planteamiento del problema, la justificación del problema y los objetivos tanto el general como los específicos.

En el capítulo 2, se presenta la revisión bibliográfica, donde se incluye el marco teórico para el desarrollo del presente trabajo de grado, así como el estado del arte del mismo, se explicarán las técnicas existentes y cuales se aplicaron.

En el capítulo 3, se presentan los inicios de la investigación, la metodología de trabajo que se usó y la forma en como se obtuvieron los resultados, las problemáticas que se tuvieron y las soluciones aplicadas.

En el capítulo 4, se presentan todo el análisis de la estadística, la aplicación de los modelos de minería de datos y *clustering* y la interpretación de los mismos.

En el capitulo 5, se entregan las conclusiones del proyecto y los nuevos retos que surgen a partir de este proyecto.

Capítulo 2

Revisión Bibliográfica

2.1 Marco Teórico

Los comportamientos de conducción de los humanos son un concepto complejo, que en términos generales, describen la forma cómo el conductor opera el vehículo bajo el control del contexto de conducción y las condiciones externas [Martinez et al., 2017], por lo que para poder encontrar los patrones de conducción de los operadores de INTEGRA S.A. se requiere una gran cantidad de datos que permita conocer a partir de variables definidas las formas de conducir de cada operador y poder clasificarlo dentro de la población general, esto a través de técnicas de minería de datos y análisis estadístico.

El análisis de comportamiento de conducción consiste en una secuencia de procesos que incluye la limpieza de datos, la integración de los datos, la selección de los datos, la transformación de datos, la minería de datos y por último el conocimiento y evaluación. [Hwang et al., 2018].

En INTEGRA S.A. los operadores de los buses tienen la posibilidad de acceder como en cualquier industria a bonificaciones de productividad, esta productividad se mide respecto al comportamiento de conducción, estudios similares se han realizado no aplicado directamente al transporte público de pasajeros de un sistema BRT pero si al público general para el pago de tarifas aplicando métodos como ”Paga como conduces” [Martinelli et al., 2018].

Para el desarrollo de este proyecto se apoyarán en distintas herramientas, lenguajes de programación y en el uso de la estadística que se profundizará un poco en este momento.

2.1.1 Python

Es un interprete orientado a objetos y de programación a alto nivel con semánticas dinámicas. Permite la creación de estructura de datos de alto nivel combinado con una escritura dinámica, es muy usado por su rápido desarrollo de aplicaciones. Es uno de los lenguajes con la sintaxis más fácil de aprender y enfatizado en la lectura para reducir el coste del mantenimiento de programación. Python soporta módulos y paquetes que permite la programación modular y el reuso de código, sus extensas librerías de fuentes o binarios permite que sea fácilmente distribuido [varios Artists, 2020b].

2.1.2 Numpy

Numerical Python es el paquete fundamental para computación numérica en Python y contiene un poderoso arreglo de objetos multidimensionales. En propósitos generales Numpy es un paquete de procesamiento de arreglos que provee un alto rendimiento en objetos multidimensionales llamados arreglos y herramientas que permiten trabajar con ellos. Esta librería permite disminuir el problema de lentitud ofreciendo funciones y operadores que operan eficientemente en estos arreglos [Arora, 2020]. Como ventajas esta librería permite una mayor eficiencia por su computación orientada a arreglos, hacer más rápida las computaciones y compactas usando la vectorización. Esta librería tiene aplicaciones extensamente en la analítica de datos, sirve de base para otras librerías como scipy y scikit-learn.

2.1.3 Pandas

Python data analysis es una librería imprescindible en el ciclo de vida de la ciencia de datos por las herramientas que ofrece para el análisis y limpieza de datos. Pandas ofrece una estructura de datos, que permite trabajar con la estructura de datos de forma eficiente e intuitiva.

Entre las ventajas se tiene que ofrece una sintaxis y funcionalidades que permite el fácil manejo de datos perdidos. Permite crear funciones al usuario y testearlas en series de datos, alto nivel de abstracción. Las aplicaciones de Pandas permite la extracción, transformación y carga de trabajos de datos y tiene un excelente soporte para la carga de archivos CSV y convertirlos al formato *Data Frame*. Funciones especificas para las series de tiempo como generación de rango de edad, ventana de movimiento, regresiones lineales entre otras [Arora, 2020].

2.1.4 Web Service

Para la obtención de datos y facilitar el análisis de la misma se va usar web services. Un web service, es una colección de protocolos abiertos y estándares que permiten el intercambio de información entre aplicaciones, en este caso se va

conectar la Aplicación InnoBUS Masivo con un web service que se realizará para el estudio para almacenar en bases de datos no relacionales como MongoDB los datos, esto facilitando la recolección de información y obteniéndola en tiempo real y disminuirá enormemente el tiempo de computación para las relaciones requeridas.

2.1.5 RStudio

RStudio, es un entorno de desarrollo para el lenguaje de Programación R, el cual es un lenguaje para la computación estadística y gráficos, este es muy similar al lenguaje S el cuál fue desarrollado por Laboratorios BELL. RStudio provee una gran variedad de estadística (lineal, no lineal, test de estadísticas clásica, análisis de series de tiempo, clasificación, *clustering*) y técnicas de gráficas y altamente extensible. El ambiente de R permite un efectivo manejo y almacenamiento de los datos, una suite de operaciones de cálculos sobre arreglos y en particular de matrices, una larga. coherente e integrada colección de herramientas para el análisis de datos. [varios Artists, 2020a]

2.1.6 Estadística Descriptiva

En este tipo de estadística y que es la que vamos a usar dentro del presente proyecto, permite la presentación, el resumen y la organización de los datos (población) incluso con cálculos numéricos, gráficos y tablas.

Dentro de la ejecución de los modelos de la estadística descriptiva se pueden encontrar los datos extremos que variarían el comportamiento real del sistema, también facilita la búsqueda de datos faltantes o datos erróneos, por ejemplo medir la variabilidad por medio de los intercuartiles que dejan ver la varianza, la desviación estándar y encontrar el valor de diversidad de los datos o su variabilidad.

2.1.7 Estadística Inferencial

Este tipo de estadística es producida por cálculos matemáticos más complejos, facilitando encontrar tendencias y realizar supuestos y predicciones sobre unos datos (población) basados en el estudio de una muestra tomada de la misma. Este tipo de estadística es más usada cuando no hay posibilidad de recolectar todos los datos para ser analizados y se debe tomar una muestra y hacer proyecciones del comportamiento de la misma.

2.1.8 Minería de Datos

La evolución natural de la tecnología da como el resultado la minería de datos, la cual nace de las bases de datos y el manejo de datos de la propia industria,

en la actualidad los sistemas de información permiten realizar transacciones y consultas como una práctica común, lo natural sería que se tengan análisis de datos más avanzados de toda la información.

Por esto la minería de datos hace parte de las ciencias de la computación como un proceso que permite ahondar y encontrar características ocultas en los datos almacenados en los distintos sistemas de bases de datos, más conocidos como patrones, para alcanzar esta meta, la minería de datos se vale de herramientas como los sistemas de bases de datos, la inteligencia artificial, procesos automatizados de recolección de información, aprendizaje automático y procesos estadísticos. Por si sola la minería de datos no podría dar patrones correctos ya que previo a ella se debe realizar procesos de extracción, tratamiento y carga más conocido como el proceso ETL (extraction, transformation, load) [Thirumagal et al., 2014].

2.1.9 ETL *(Extraction, Transform, Load)*:

Este proceso dentro de la minería de datos y la *clusterización* es un proceso fundamental ya que permite que se realice la selección, limpieza y transformaciones necesarías de los datos necesarios para analizar y cargarlos en un repositorio de datos listos para poder ser usados [Kimball and Caserta, 2004].

- ***Extraction***: El proceso de *extraction* es el proceso por el cual se comienza a seleccionar los datos, su origen, se realiza las transformaciones necesarias para extraerlos, por ejemplo construcción de *web services*, descarga de archivos, unión de bases de datos [Trujillo, 2013]

- ***Transformation*:** Esta segunda etapa consiste en realizar la limpieza o la transformación de datos, eliminando datos incompletos o reemplazándolos a datos de la media, eliminando extremos o datos atípicos, procesos que se realizan en esta etapa esta depuración de valores nulos, distintas codificaciones para el mismo término (entrada, ingreso), descripción inconsistente de los campos entre otras.[González Echeverri, 2018]

- ***Load*:** Al finalizar el proceso de transformación se procede con la etapa de carga de datos, que no consiste en cargar los datos en el repositorio para que estén listos para usar y aplicar técnicas de minería de datos, importante conocer que se debe realizar nuevamente un proceso de depuración comprobando las reglas de integridad, ordenando los datos entre otras [Trujillo, 2013]

Como se ha indicado, la limpieza, gestión y transformación de los datos no hacen parte de la minería de datos, pero son las etapas previas que ayudan al éxito del mismo, por esto mismo. Un proceso de Minería generalmente realiza

estos procesos.

- **Selección del conjunto de datos:** Selección de las variables analizar como objetivo del modelo, como también los hechos establecidos de las dimensiones.

- **Análisis de la propiedad de datos:** Proceso estadístico de la información, se realizan un filtrado para comprobar la ausencia de datos (valores nulos), los histogramas y diagramas necesarios con el fin de que el modelo sea lo más cercano a la realidad.

- **Transformación del conjunto de datos de entrada:** Etapa de preprocesamiento de los datos.

Hasta este paso no se ha realizado minería de datos como tal y es lo que se ha mencionado que se conoce como (ETL), en la siguiente fase comienza como tal el proceso de minería de datos.

- Seleccionar y aplicar la técnica de minería de datos: Se construye el modelo basándose en los datos y se escoge la técnica más adecuada.

- Extracción de conocimiento: Como resultado del paso anterior, la aplicación del modelo genera un conocimiento antes desconocido o por el contrario permite corroborar hipótesis que se tenían previo al estudio, en esta fase se puede llegar a encontrar nuevas relaciones que antes del estudio eran nulas.

- Interpretación y evaluación de los datos: El último proceso que debe realizarse en la minería de datos es la de entender que nuevo conocimiento se presentó y darle una interpretación basándose generalmente en la opinión de un experto, como se indicará posteriormente, la minería de datos puede ser a través de un aprendizaje supervisado o por medio de un aprendizaje no supervisado. Si se realiza un proceso de extracción de conocimiento a través de distintas técnicas, los resultados de estas se deben confrontar y verificar si dan un resultado similar, en caso contrario buscar por qué no sucedió esto. En caso tal que los resultados no sean los esperados, se puede comenzar el proceso desde el paso que se consideré necesario e iterarlo cuantas veces sea necesario hasta lograr el objetivo.

La minería de datos consiste en arrojar información oculta en datos almacenados de distintas fuentes, es así por ejemplo se pueden encontrar relaciones que no tienen una lógica inicial o relacional oculta, así mismo también se usa para validar información que ya esta implícita, esto con el fin de ayudar en la toma de las decisiones de las organizaciones.

2.1.10 Tipos de Minería de Datos

Predicción

Este tipo de minería de datos, consiste en encontrar patrones a partir de datos que tienen relación que son evidentes o de probabilidad muy alta.

Agrupación

En este tipo de minería de datos, la extracción de conocimiento es a través de la identificación de grupos naturales en los datos.

2.1.11 Funciones de la Minería de Datos

La minería de datos se divide en 2 categorías, supervisadas y no supervisadas.

Minería de Datos Supervisada

Este tipo de minería tiene una alta relación con los tipos de minería de datos predictivos, la característica en una función supervisada es el entrenamiento, por medio de valores conocidos y donde el software realiza el análisis de distintos datos y que arrojan un mismo resultado o ya se tiene un valor conocido.

Minería de Datos No Supervisada

En las funciones no supervisadas no hay relación entre datos dependientes e independientes y generalmente sirven para propósitos descriptivos o encontrar patrones. Dentro de la minería de datos existen distintas técnicas, para el presente proyecto debido a los datos que se almacenan a lo largo del tiempo se usará la técnica de agrupamiento o *clustering*, buscando obtener grupos con características similares de los operadores de los buses, la clusterización es un procedimiento de agrupación de datos vectorizados según criterios de distancia. En la clusterización se busca ubicar los datos vectorizados de entrada de forma que estén más cercanos aquellos que tengan características comunes. Algunos ejemplos de estos son los Algoritmos K-medias. [Han et al., 2011]

2.2 Metodologías de Minería de Datos

La diferencia fundamental entre un modelo de proceso y una metodología radica en que el modelo de proceso establece el qué hacer, mientras que la metodología específica el cómo. Dentro de las distintas metodologías que existen en la

minería de datos hay 2 metodologías que son muy usadas la KDD (Knowledge Discovery in Database) y CRISP – DM.

2.2.1 KDD *(Knowledge Discovery in Database)*

KDD es el proceso por el cual se identifican procesos novedosos, útiles y comprensibles de los datos que son obtenidos de forma no trivial.[Fayyad and Stolorz, 1997] KDD, es considerado el modelo inicial de la minería de datos y es aceptado como el modelo de la comunidad científica y esta formado 9 etapas [Moine et al., 2011].

Etapas de *KDD*:

- **Compresión del negocio:** En esta etapa se trata de captar la necesidad de la organización al igual que las metas del proyecto, conocer las limitaciones, las reglas del escenario, ya que esto va ayudar mucho más a entender la problemática, es una etapa de concientización de la realidad y necesidad empresarial.

- **Seleccionar conjunto de datos:** En esta fase se busca seleccionar los datos y las distintas fuentes de los datos, si esta en una base de datos centralizada o esta en varias, entre otras cosas

- **Limpieza y pre-procesamiento de datos:** Como su nombre lo indica en esta fase se limpia los datos seleccionados en la fase anterior, retirando datos atípicos y haciendo un pre-procesamiento de los mismos. Esto se hace con el fin de garantizar la utilidad de los datos

- **Reducción y proyección de los datos:** Esta etapa busca minimizar la cantidad de variables y así facilitar el estudio, esto sin perder de foco el objetivo del negocio.

- **Relación del objetivo del proceso:** Se ajusta el objetivo con el método y algoritmo de minería de datos que más se ajuste para alcanzar la meta.

- **Análisis exploratorios y selección del modelo e hipótesis:** Se selecciona y usa el metodo y algoritmo que más se ajuste a las necesidades del objetivo

- **Implementar minería de datos:** Se ajusta el proceso y se realiza la minería de datos para obtener los mejores resultados

- **Interpretar los resultados:** Esta etapa consiste en el análisis de los resultados, entendiendo e interpretando los mismos, en ocasiones es necesario volver a iterar sobre el mismo proceso para obtener mejores resultados o confirmar una relación que se de como resultado.

- **Actuar sobre el conocimiento descubierto:** Con los resultados obtenidos en el paso anterior y la interpretación adecuada se procede a tomar decisiones o integrar estos resultados con conocimientos previos adquiridos.

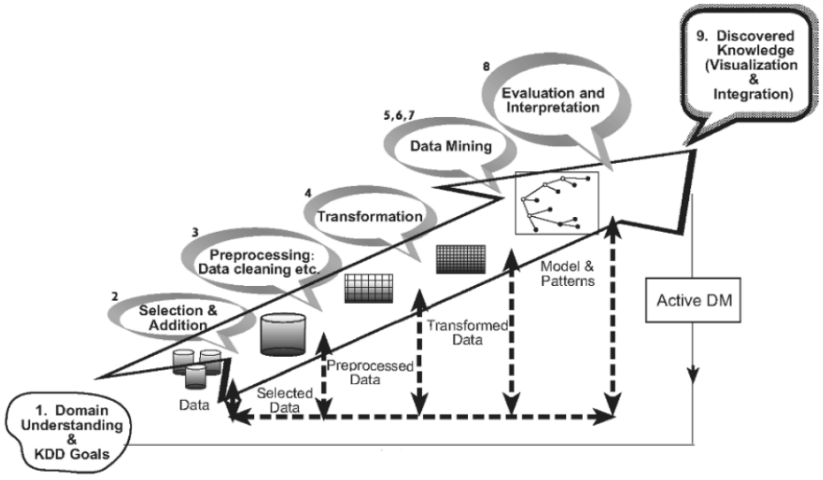

Figura 2.1: Proceso KDD, tomado de [Maimon and Rokach, 2005]

2.2.2 CRISP-DM *Cross-Industry Standard Process for Data Mining*

Desarrollado por *Daimler Chrysler, SPSS, y NCR* en 1999 como una mejora y dar un oportuno servicio a las necesidades de los clientes, hoy por hoy esta metodología es la más usada y aceptada a nivel empresarial a diferencia de *KDD* que esta más aceptada en el mundo cientifico. Esta metodología esta compuesta por 6 etapas las cuales están estructuradas y definidas en [Shearer, 2000]

Etapas de CRISP-DM

- **Entendimiento del negocio:** Al igual que en *KDD* en esta fase se trata de entender la necesidad empresarial, los criterios de negocio, todos los

requerimientos técnicos.

- **Entendimiento de los datos:** Esta segunda fase consiste en entender la colección de los datos, revisar su calidad, realizar exploraciones iniciales de los datos buscando *insights* de los datos de formar hipótesis para información no descubierta. En esta fase puede ocurrir que se tenga que iniciar de la fase 1 ya que los datos que se entendierón pueden no estar acorde con las necesidades empresariales haciendo que se itere a la primera fase.

- **Preparación de los datos:** En esta fase se selecciona y se preparan los datos finales con los cuales se van a trabajar en etapas posteriores, esta etapa se considera también la limpieza de datos y la transformación de los mismos, selección de tablas, selección de atributos entre otras.

- **Modelamiento:** En esta etapa se selecciona y modela con distintas técnicas y aplicaciones, con diferentes parámetros para el mismo problema de minería de datos buscando obtener la mejor respuesta.

- **Evaluación:** En esta etapa se revisan los datos obtenidos de los modelos y se decide como se pueden usar estos resultados. La interpretación de los modelos depende del algoritmo usado y se realiza la revisión para ver si cumple con los objetivos planteados inicialmente.

- **Implementación:** En esta última fase se determina el conocimiento obtenido y los resultados, en esta fase también se organiza, se reporta y se presenta el conocimiento ganado de lo que se necesitaba.

Es importante resaltar que el proceso CRISP-DM puede ser un proceso iterativo en cualquiera de sus etapas, es decir si por ejemplo al llegar al paso de la evaluación no se obtienen los resultados requeridos, se tiene que volver a empezar en la fase 1, o si en el modelamiento los resultados no son los esperados se puede devolver a la fase de preparación de los datos tal como se muestra en 2.2.

2.3 *Clustering*

El análisis de *cluster* o simplemente *clustering* es el proceso de particionar un conjunto de objeto de datos u observaciones en subconjuntos. Cada subconjunto es denominado como *cluster* y los elementos que hacen parte de este subconjunto tienen la característica que son similares unos a los otros, pero distintos de los elementos que estan contenidos en otro *cluster*.El conjunto de

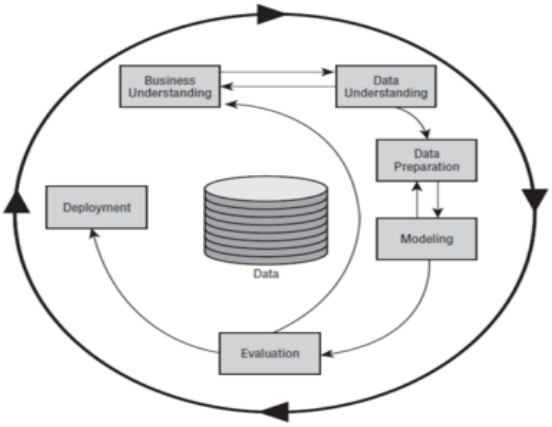

Figura 2.2: Proceso CRISP-DM, tomado de [Shafique and Qaiser, 2014]

Data Minig Process Model	KDD	CRISP-DM
Número de Pasos	9	6
9*Nombre de los Pasos	Compresión del negocio	Entendimiento del negocio
	Seleccionar conjunto de datos	2*Entendimiento de los datos
	Limpieza y pre-procesamiento de datos	
	Reducción y proyección de los datos	Preparación de los datos
	Relación del objetivo del proceso	3*Modelamiento
	Análisis exploratorios y selección del modelo e hipótesis	
	Implementar minería de datos	
	Interpretar los resultados	Evaluación
	Actuar sobre el conocimiento descubierto	Implementación

Tabla 2.1: Tabla Comparativa KDD vs CRISP-DM

cluster resultantes en el análisis de *cluster* es referido como *Clustering*. Diferentes métodos de *Clustering* puede generar subconjuntos distintos con el mismo conjunto de objetos de datos. La generación de los subconjuntos no es realizada por un proceso humano sino por los algoritmos de *Clustering* lo que permite que sea un proceso muy usable que puede liderar a descubrimiento de conocimiento previamente desconocido por los grupos a través del conjunto de datos u observaciones [Han et al., 2011].

Otra definición de *Clustering* esta dada por [Jain and Dubes, 1988] donde se agrupan los objetos que tienen características o atributos similares formando nuevos conjuntos o clases. Se diferencia del proceso de Clasificar ya que en esta sí existe una división previa entre las clases mientras que en el *Clustering* estás clases se desconocen. Otra diferencia entre clasificación y *Clustering* es que en este último no se agrupan los datos en conjunto sino que se clasifican uno por uno al usar sus algoritmos.

Las aplicaciones de la clusterización va desde el reconocimiento de imágenes

para reconocer patrones, como por ejemplo reconocer la escritura de una persona como el reconocimiento de los números en formularios escritos a mano. Otros usos se pueden encontrar ampliamente en las búsquedas en linea donde se puede generar una consulta por palabras claves que puede retornar una gran cantidad de resultado, la clusterización se puede usar para presentar el retorno de estos resultados en grupos de forma que puedan ser fácilmente accesibles y concisos respecto a lo que la persona busca. la inteligencia de negocio, la biología y la seguridad.

Una de las ventajas de los *cluster* es que permite encontrar los puntos *Outliner* que luego de una limpieza de datos hayan podido quedar o simplemente se puede usar como un proceso de limpieza para encontrar estos mismos puntos.

En general la mayoría de los distintos algoritmos de *Cluster* que se explicarán en el documento tienen el mismo comportamiento, es decir están basados en una función de optimización que es la suma ponderada de las distancias a los centros, eso no quiere decir que todos los algoritmos tienen la misma función objetivo a optimizar y estas pueden ser distintas entre cada una. Generalmente uno de los pasos es el de asignar a cada dato u objeto una medida de semejanza al patrón o centroide de cada *cluster*, esto con el fin de ”clasificar” ese objeto o dato a que *cluster* pertenece, esto se conoce como cálculo de función de distancia.

2.3.1 Medidas de Distancia entre Objetos

Lo primero que se debe decir antes de comenzar a explicar las distintas distancias, es que dentro de las funciones existen 2 tipos de índices, el primero de ellos se utiliza para obtener la cercanía o similitud de los objetos y será identificado por el índice i, mientras el segundo se usa para evidenciar la lejanía o disimilitud de los mismos y será identificado por el índice k, quedando la expresión d(i,k) y debe cumplir con las condiciones mostradas a continuación:

$$d(i,i) \begin{cases} = 0, \forall\, i & \text{(disimilitud)} \\ \geq max_k d(i,k), \forall\ (i,k) & \text{(similitud)} \end{cases} \tag{2.1}$$

$$d(i,k) = d(k,i), \forall (i,k) \tag{2.2}$$

$$d(i,k) \geq 0, \forall (i,k) \tag{2.3}$$

Los índices anteriormente definidos se pueden usar tanto en objetos cuantitativos como cualitativos y que se usen en los algoritmos de Clusterización no signfica que sean exclusivos de ellos y pueden ser usados en otras técnicas y otros objetos.

2.3.2 Distancia Minkowksi

Como se puede ver en la figura 2.3 existen 3 tipos de distancias para los datos cuantitativos, las cuales son la euclídea, la manhattan y del supremo que se explicarán en breve.

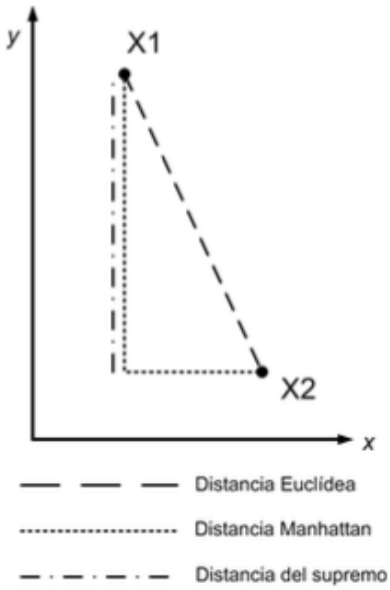

Figura 2.3: Distancias de Minkowski

Como se muestra en 2.4 se asignan distintos valores a r se obtienen las distancías anteriormente mencionadas y en la figura 2.3 se muestra gráficamente como funciona cada una de ellas.

$$d(i,k) = \left(\sum_{j=1}^{d} |x_{ij} - x_{kj}|^r \right)^{\frac{1}{r}}, donde\ r \geq 1 \tag{2.4}$$

Distancia Manhattan

Se conoce también con el nombre de *taxicab* ya que su recorrido es parecido al de un taxi que recorre las calles en cuadriculas, una particularidad es que si todas las dimensiones del conjunto son binarias pasa a llamarse **distancia Hamming**, la cual no se explicará en este documento pero su uso es particularmente alto en la codificación y transmisión de datos. La distancia manhattan se presenta cuando r=1.

$$d(i,k) = \sum_{j=1}^{d} |x_{ij} - x_{kj}| \tag{2.5}$$

Distancias Euclídea

Esta distancia esta basada en el teorema de pitágoras y es la distancia más corta entre 2 puntos, como se muestra en la figura 2.3 forma una diagonal entre los puntos analizando generalmente, cabe adicionar que esta distancia es la más usada en las técnicas de *Clustering*. La distancia euclídea esta relacionada directamente con la similitud de los casos y cuando se usa se aplica la formula de la ecuación 2.6

$$d(i,k) = \left(\sum_{j=1}^{d} |x_{ij} - x_{kj}|^2 \right)^{\frac{1}{2}} = \sqrt{(x_i - x_k)^T (x_i - x_k)} = ||x_i - xj|| \tag{2.6}$$

2.3.3 Distancia del Supremo

A diferencia de la euclídea, esta distancia se presente con las distancias más grandes entre los objetos, es decir esta relaciona con la disimilitud de los datos y esto sucede cuando r tiene a infinito (r $\rightarrow \infty$),

$$d(i,k) = max_{1 \leq j \leq d} |x_{ij} - xkj| \tag{2.7}$$

2.3.4 Distancia de Mahalanobis

Acá en vez de tener un valor r tenemos un valor que tiene el valor de la matriz de la covarianza. Esta distancia al igual que la distancia de Minkowski sirve para datos de tipo cuantitativo y esta definida por la expresión 2.8.

$$d(i,k) = (x_i - x_j^T \varphi^- 1 (x_i - xk) \tag{2.8}$$

Cuando el valor de phi esta elevado a la -1 este adquiere el valor de una matriz de identidad y el resultado termina siendo la distancia euclídea elevada al cuadrado.

$$\varphi = \frac{1}{N} \sum_{k=1}^{N} (x_k - \bar{x})^T \tag{2.9}$$

2.3.5 Coeficiente de Correlación

Este coeficiente permite medir la dependencia lineal entre 2 atributos o dimensiones y esta dada por la expresión 2.10

$$d(j,r) = \left| \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - m_j)(x_{ir} - m_r)}{S_j S_r} \right| \tag{2.10}$$

m representa la media de los datos y S la varianza, si *d(j,r)* es cero, significa que los datos son linealmente independientes es decir no existe una relación entre los mismos, si por el contrario el valor de *d(j,r)* es cercano a 1, esto significa que hay un alto grado de dependencia entre los atributos, esta distancia sirve para identificar las dependencias que existen entre uno y otro atributo.

2.3.6 *Matching Coefficients*

Es la primera de las distancias que se usa para objetos de tipo cualitativo, para entender un poco mejor esto, se pone el caso de 2 objetos x_i y x_j con valores binarios, es decir se pueden tener la relación de valores (0,0), (0,1), (1,0), (1,1), cómo se puede evidenciar el valor posible de combinaciones en un caso binario estaría dado por n^2, en caso que hubieran más objetos, como más resultados y no fuera binario estaría dado por la forma n^j donde n es la cantidad de objetos y j los atributos. Dentro de de esta distancia hay 3 tipos de cálculo, la distancia jaccard, la distancia simple y la proximidad única

Distancia Jaccard: Esta dada por la expresión 2.11 donde se elimina n las cuentas con los valores en 0 para ambos objetos.

$$d(i,k) = \frac{a_{11}}{a_{11} + a_{01} + a_{10}} = \frac{a_{11}}{d - a_{00}} \tag{2.11}$$

Simple matching coefficient: Dada por la expresión 2.12, a diferencia de la distancia jaccard, en esta distancia se le da la misma importancia a los valores en 0 que a los valores en 1

$$d(i,k) = \frac{a_{00} + a_{11}}{a_{00} + a_{11} + a_{01} + a_{10}} = \frac{a_{11} + a_{00}}{d} \tag{2.12}$$

Proximidad única: Es de las distancias más fáciles de usar y como las 2 anteriores es usada para datos cualitativos, también es conocida como *overlap metric* y se define como 2.13 y sólo se contabilizan los casos donde los valores no coincidan.

$$d(i,k) = \sum_{j=1}^{d} \delta_{x_{ik}x_{kj}} \tag{2.13}$$

en donde

$$\delta_{x_{ik}x_{kj}} = \begin{cases} 0 si x_{ij} = x_{kj} \\ 1 si x_{ij} \neq x_{kj} \end{cases} \tag{2.14}$$

2.3.7 Entropía

Como se menciona en [Barbará et al., 2002] la entropía nos permite conocer la diversidad o heterogeneidad de los objetos, esto en los *cluster* ya que permite a través de las diferencias de los objetos crear las homogeneidades de los mismos. Otra definición de entropía esta dada por [Andritsos, 2004] donde se define como la medida de información e incertidumbre de una variable aleatoria.

La entropía de una variable esta dada por la formula en 2.15, donde p(x) es la probabilidad de la variable, S(X) los distintos valores de la variable X y E(X) el valor de su entropía.

$$E(X) = -\sum_{x \in S(X)} p(x) log(p(x)) \tag{2.15}$$

2.3.8 Distancia de Kullback-Leibler

Cuando se tienen 2 bases de datos o datos distribuidos en el tiempo se usa la distancia Kullback-Leibler ya que permite conocer la medida de divergencia o convergencia entre las distribuciones de probabilidad, un ejemplo de esto se puede observar en [Lin et al., 2016] se aplica esta distancia con éxito para poder identificar las desigualdades entre las variables o dicho en otras palabras la divergencia. La formula que la define esta dada en 2.16

$$D_{KL}[p||q] = \sum_{t \in T} p(t) log \frac{(p(t)}{q(t)} \tag{2.16}$$

2.4 Algoritmos de *Cluster*

2.4.1 Factores a tener en cuenta en los Algoritmos de *Cluster*

Antes de hablarse de los distintos algoritmos de *Cluster* se deben tener en cuenta algunas situaciones para escoger el algoritmo que mejor se ajuste a las necesidades, es importante aclarar que escoger el algoritmo correcto permitirá tener un mayor ajuste del proceso. [Mamani Rodríguez, 2015] lista 4 factores a tener en cuenta para la escogencia de algoritmos de *Cluster* que son:

- **Objetivo de la Aplicación:** La elección de un algoritmo de *Cluster* debe estar basado en la pertinencia del mismo para el caso de estudio, por ejemplo si se quiere ubicar la disponibilidad de camas en un hospital según la complejidad clínica del caso se puede utilizar el algoritmo de particionamiento *k-means*, pero por ejemplo si lo que se desea es crear un reconocimiento de imágenes se puede usar los algoritmos basados en densidad.

- **Elección entre calidad y velocidad:** Como todo sistema, es ideal que existe una relación entre la velocidad de los datos a procesos para obtener el resultado y la calidad de los resultados, pero encontrar el equilibrio entre estas 2 situaciones es a veces bastante complejo ya que por ejemplo algoritmos que producen *cluster* de calidad por lo general son incapaces de manejar grandes volumenes de información.
- **Características y tipos de los datos:** Es importante tener la claridad de los tipos de atributos, si son cuantitativos o cualitativos, si son numéricos o texto, por eso evaluar las características de los datos es importante para elegir el algoritmo adecuado.
- **Dimensionalidad:** Este factor tiene que ver con la cantidad de atributos que tiene cada objeto, hay algoritmos de *cluster* que tienen mejores resultados cuando se manejan pocos atributos pero su calidad merma o tiene una reducción en la velocidad de los resultados al aumentar la cantidad de atributos.
- **Cantidad de ruido en los datos:** Es importante realizar un adecuado proceso de limpieza de datos ya que algunos algoritmos de *cluster* son sensibles a datos atípicos o incompletos y no funcionan correctamente.

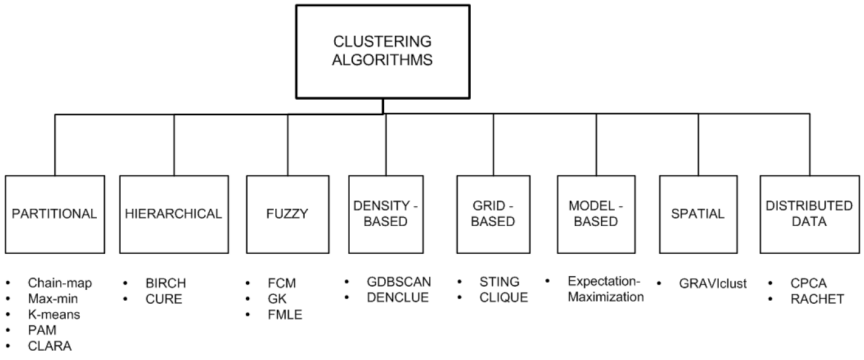

Figura 2.4: Algoritmos de *Cluster*, Tomado de [Sánchez, 2005]

Como se muestra en la 2.4, existen 8 categorías para los algoritmos de *cluster cluster*, estos son: algoritmos jerárquicos, algoritmos particionales, algoritmos fuzzy, algoritmos basados en cuadricula, algoritmos basados en modelo, algoritmos espaciales y algoritmos de datos distribuidos

2.4.2 Algoritmos Espaciales

Los *clustering* espaciales forman grupos de datos donde se busca maximizar las similaridades de los datos dentro del mismo *cluster* y minimizar la similaridad dentro de 2 *cluster* distintos [Han et al., 2000] Una de las características de los *clustering* espaciales es que no poseen un conocimiento de fondo y estas agrupaciones se forman a partir de los atributos de los objetos. Las leyes y modelos que se aplican al conjunto de objetos de datos esta muy basada en la medición de la distancia o en otras palabras de la similaridad y con esto es posible delinear las áreas de densidad que se conocerá como *cluster* y separá estos objetos de datos de otros por medio de su dsimilaridad y los clasificará en otro *cluster* [Murray and Shyy, 2000]. Los álgoritmos espaciales usan las distancias geométricas mencionadas en la sección anterior y dependiendo de la distancia que se use puede mostrar formaciones de *clustering* distintas [Ester et al., 2000]

El Algoritmo *GRAVIclust* está explicado en [Indulska and Orlowska, 2002] donde se definen 3 fases para este algoritmo, el primero es la fase de la función principal donde se define la distancia de la matriz y los *cluster* requeridos, en la segunda fase, conocida como de inicialización se selecciona una región R con un conjunto de puntos P, adicional de los *cluster* a formar y la distancia de la matriz. Con estos seleccionados se selecciona un radio R con la mayor cobertura de puntos a partir de los k seleccionados, luego de esto se repite el proceso tantas k veces sea necesario, cuando los puntos se encuentran dentro de un radio son eliminados para que no queden sobreseleccionados en otro punto. Por último viene la fase de optimización, en esta fase se toman los *cluster* obtenidos durante la fase de inicialización y para cada *cluster*, un nuevo centro del radio es calculado basado en el punto que este más cerca. Desde que el centro del radio haya cambiado los miembros pertenecientes a ese *cluster* también cambian por lo que es importante volver a calcular las distancias de cada uno de los puntos pertenecientes al *cluster*.

2.4.3 Algoritmos Basados en Cuadrícula

Los algoritmos basados en cuadricula particionan el espacio en cuadriculas y buscan o seleccionan los objetos que están dentro de estas cuadriculas, ejemplos de estos son STING (*Statical information grid*) y CLIQUE (*Clustering In QUEST*). Estos tipos de algoritmos o metodos son más que eficiente ya que sólo dependende del número de celdas y es independiente del número de objetos y requieriendo menor tiempo de procesamiento respecto a los algoritmos o metodos basados en densidad. [Chai and Ngai, 2020]

CLIQUE (*Clustering in Quest*

Los pasos de CLIQUE son particionar el espacio de datos y encontrar el número de puntos que están en cada una de las celdas de la partición, luego de esto se

identifica los subespacios que contengan *clustering* utilizando el principio anterior [Agrawal et al., 1998].

Luego de esto viene el paso de identificar los *cluster* y esto se hace determinando las unidades de densidad en cada uno de los subespacios de interés, paso seguido se determina la conexión de las unidades de densidad en los subespacios de interés.

Por último se genera el mínimo de las descripciones del *clustering*, esto se hace determinando el máximo de las regiones que puede cubrir el *cluster* conectando a través de unidades de densidad y por último se determina el mínimo de cubrimiento para cada *cluster*.

STING (*Statical Information Grid*) Lo primero que se hace es dividir el área en celdas rectangulares las cuales son representadas por una estructura jerárquica, siendo la zona de mayor jerarquía el nivel 1, sus "hijos serán el nivel 2 y así sucesivamente cuántos niveles hayan [Wang et al., 1997]. El número de niveles o capas puede ser obtenido cambiando el número de celdas rectangulares que se forman en un nivel superior, Una celda en un nivel i puede corresponder a la unión de las áreas de sus hijos en un nivel $i+1$. En el algoritmo STING cada celda rectangular tiene 4 hijos y cada hijo corresponde a un cuadrante dentro de la celda del padre.

2.4.4 Algoritmo de *Clustering* Jerárquico

Este algoritmo de *cluster* va partiendo el conjunto de datos en grupos jerárquicos recursivamente hasta quedar conjuntos cada vez más pequeños hasta quedar con un solo objeto en su hoja de la rama. En los métodos jerárquicos se puede formar de 2 formas de abajo hacia arriba o de arriba hacia abajo, en otras palabras puede ir de lo particular a lo general o de lo general a lo particular.

BIRCH (*Balanced Iterative Reducing and Clustering using Hierachies*)

Este algoritmo almacena una lista de 3 datos en los que se incluye, el número de objetos que pertenecen al *cluster*, la sumatoria de los valores de los objetos pertenecientes y la sumatoria de los cuadrados de los atributos que pertenecen al *cluster* [Zhang et al., 1997]. Esta información permite crear el *CF-Tree (Cluster Features Tree)*, cada nodo creado indica el grupo que pertenece en esa ramificación y que características posee, el algoritmo BIRCH funciona de la siguiente manera:

- Se crea un *CF-tree* inicial leyendo cada uno de los objetos y asignándose a cada una de las ramas, si la distancia de los objetos es mayor a un parámetro T se procede a crear una nueva rama

- Se revisa la estructura del árbol formado, si el mismo es demasiado grande, se procede a modificar el valor del parámetro *T*, si T es muy grande el árbol tiende a tener una sola rama, si T es más pequeña el árbol tiene más nodos.
- Se procede aplicar cualquier algoritmo de *clustering* sobre los nodos formados en el árbol.
- Se redistribuyen los centroides a partir del resultado del paso anterior logrando una mayor agrupación.

CURE *(*Clustering Using Representatives)

Es de tipo aglomerativo, es decir que comienza de abajo hacia arriba tomando a cada objeto del conjunto de datos como un grupo independiente, paso seguido a esto realiza un proceso de combinación y comenzando a formar los *cluster*, por último acerca los objetos extremos usando un factor de acercamiento que generalmente es la distancia media de todos los elementos que componen el grupo [Guha et al., 1998]. Una de sus ventajas es que es capaz de dtectar grupos con multiples formas y tamaños.

2.4.5 Algoritmos de *Clustering* Borroso o *Fuzzy*

Generalmente en un proceso de *clustering* un objeto pertenece a un sólo *cluster* y no puede pertenecer a otro como se denota en la 2.17, las formulas 2.18 2.19 indican que ningún objeto puede pertenecer al grupo vacio y siempre deben quedar en un *cluster*.

$$A_i \bigcap A_j = \varnothing, 1 \leq i \neq j \leq c \tag{2.17}$$

$$A_i \subset A_j \subset Z, 1 \leq i \leq c \tag{2.18}$$

$$\bigcup_{i=1}^{c} A_i = Z \tag{2.19}$$

En otras palabras si se hace una matriz de dimensiones c x N donde c es la cantidad de *cluster* y N el número de objetos resulta la expresión 2.20 que como se muestra el valor sólo puede pertenecer a 0 o a 1, estar o no estar en un *cluster* determinado

$$M_{hc} = \left\{ U \in R^{cxN} | \mu_{ik} \in \{0,1\}, \forall i,k; \sum_{i=1}^{c} \mu_{ik} = 1, \forall k; 0 < \sum_{k=1}^{N} \mu_{ik} < N, \forall i \right\} \tag{2.20}$$

En cambio en los algoritmos fuzzy todo esto cambia, los elementos de la matriz no tienen un valor determinado de 0 o 1 sino que puede tomar cualquier valor en el mismo rango. Lo que si debe ser obligatorio es que la suma de un único objeto debe sumar uno, independiente si esta en 1 o en varios *cluster*.

$$M_{fc} = \left\{U \in R^{cxN} | \mu_{ik} \in [0,1], \forall i,k; \sum_{i=1}^{c} \mu_{ik} = 1, \forall k; 0 < \sum_{k=1}^{N} \mu_{ik} < N, \forall i\right\} \tag{2.21}$$

Este es un requerimiento que se conoce como no-posibilista. Entre los algoritmos borrosos que usan el requerimiento no-posiblista se tiene los siguientes:

- **Algoritmo *c-means*:** El Algoritmo *c-mean* funciona de la siguiente manera, primero se inicializa la matriz de pertenencias *U* con valores aleatorios para que cumplan los requísitos de 2.21

  Luego se calcula los centros de los *cluster* con la expresión 2.22

  $$c_i = \frac{\sum_{k=1}^{N} (\mu_{ik})^m z_k}{\sum_{k=1}^{N} (\mu_{ik})^m}, 1 \leq i \leq c \tag{2.22}$$

  Se halla la distancia de los objetos a su centro con la expresion 2.23

  $$||z_k - c_i||_B^2 = (z_k - c_i)^T B (z_k - c_i) = D_{ik}^2 B \tag{2.23}$$

Se recalcula toda la matriz y aplicando los 2 casos posibles según sea el caso se verifica si se cumple con la condición, si no se cumple se debe volver a calcular los centros del *cluster* y calcular la distancia, esto se hace sólo si ε es mayor que 0.001 lo que indica que la matriz de pertenencias es muy similar a la anterior lo que para el algoritmo.

Algoritmo de *Gustafon-Kessel o GK* El algoritmo consiste en los siguientes pasos:

1. Se inicializa la matriz de pertenencia *U* con valores aleatorios.

2. Se calculan los centros de los grupos

3. Se calcula la matriz de covarianza de cada clase

4. Se calcula las distancias

5. Se hallan los nuevos valores de la matrixz de pertenencia, siguiendo el procedimiento del algirtmo C-mean y teniendo en cuenta que existe una norma distinta para cada *cluster*

6. Se verifica el patrón de parada tal cual se hace en el algoritmo de *c-mean*

2.4.6 Algoritmo de Datos Distribuidos

Este tipo de algoritmo es muy usado para poder reconocer patrones y generar agrupamiento de objetos por similitud en sus valores, en datos que no se encuentran en el mismo conjunto de datos, esto quiere decir que a partir de los comportamientos en un conjunto de datos se generaliza y estas mismas condiciones pueden aplicarse a otro conjunto de datos en otra zona. En un escenario entendible suponga que se tiene una locación de un supermercado en el lado norte de la ciudad, a partir de la data capturada en este lado norte, se crean los *clusters* de forma local y estos son enviados a través de internet a los servidores centrales del supermercado que a su vez envia esta información a almacenes al lado sur de la ciudad permitiendo crear *clusters* con la misma información que se capturo en el lado norte pero con los datos propios del lado sur, haciendose un sistema distribuido.

CPCA (*Collective Principal Components Analysis*

El CPCA utiliza los siguientes pasos según [Kargupta et al., 2001]

- Realizar localmente PCA

- Proyectar los datos localmente y aplicar el algoritmo de *clustering* en cada sitio

- Seleccionar un conjunto de datos representativos en cada nodo.

- Todos los sitios comunicarán al nodo central las filas de datos correspondiente y sus respectivos índices a la central.

- El sitio central realizara el PCA con los datos recolectados en cada sitio y enviará los resultados a cada sitio.

- Cada sitio recibe los datos globales y realiza nuevamente *clustering* usando nuevamente los algoritmos pero basados en los resultados obtenidos globalmente.

- La central recibe los diferentes grafos obtenidos de los resultados locales y los computa, los métodos de combinación pueden variar.

RACHET (*Recursive Agglomeration of Clustering Hierachies by Encircling Tactic*):

Es un algoritmo diseñado para objetos distribuidos homogéneamente en distintos nodos de la red y es de tipo jerárquico, el proceso comienza cuando cada nodo de la red genera su árbol jerárquico de agrupamiento usando los objetos contenidos en él, luego transmite estos datos al nodo central, donde es analizado con todos los demás árboles de los otros nodos de la red, con estos árboles, el nodo central, crea un árbol global el cual transmite a todos los nodos nuevamente, para esto el dendograma global crea ramificaciones para cada uno de los nodos y en la medida de lo posible los unifica o agrupa si cumplen ciertas condiciones.

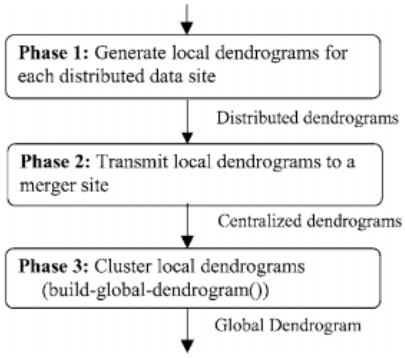

Figura 2.5: Flujo de control RACHET, Tomado de [Samatova et al., 2002]

2.4.7 Algoritmos de *Clustering* Basado en Modelos

Como su nombre indica, usa modelos matemáticos para los datos y busca los parámetros que mejor representan el comportamiento. Generalmente usan la función de distribución gausiana esto para que cada uno de los objetos se ubique o ajuste dentro de la distribución en cada *cluster*. El algoritmo de *clustering* basado en modelos tiene una representación que se conoce como Expectation-Maximization y una de sus desventajas es que tiene un alto consumo computacional al realizar todos los cálculos.

2.4.8 Algoritmos Particionales

Como su nombre lo indica estos algoritmos funcionan particionando los datos en espacio de conjuntos o *cluster* de datos tratando de optimizar un índice de coste [Kaufman and Rousseeuw, 2009].

Algoritmo *Chain-map*

Con un conjunto de objetos, los organiza en forma de vectores, donde cada elemento es el valor de una característica, el algoritmo selecciona aleatoriamente un objeto disponible y con esto ordena a los demás objetos del vector según la proximidad, cómo se ve en 2.24 en donde i, indica el objeto que se ha seleccionado de primero.

$$z_i(0), z_1(1), ..., z_i(N-1) \tag{2.24}$$

Usando las distancias euclidiana entre el objeto k y el objeto k-1 se dispone el orden en que estarán en la cadena. Una distancia euclidiana grande signficia que el objeto pertenece a otro grupo, mientras que una distncia pequeña indican que son del mismo grupo. Es importante determinar a partir de que valor de la distancia euclidiana se considera que el objeto pertenece a otro *cluster* o sigue en el mismo. Este algoritmo no se caracteriza por ser uno de los más optimos o más precisos, pero si se puede usar para estimar la cantidad de *cluster* a usar.

Algoritmo PAM *(Partition Around Medoids)*

Como lo explica [Kaufman and Rousseeuw, 2009] PAM se basa en la búsqueda de los objetos representativos de *k* a lo largo del *dataset*. Los objetos *k* deben representar los distintos aspetos de la estructura de datos y son llamados centrotipos en los *cluster* pero en el algoritmo PAM se llaman medoides. Luego de encontrar los *k* representativos de los objetos, se comienza con la construcción de los *k cluster*.

El algoritmo contempla 2 fases, en la primera fase se construye la colección de objetos *k* seleccionando de los objetos que no fuerón selecionados.

La segunda fase consiste en la fase de cambio, donde se mejora la calidad del *clustering* realizando cambios entre los objetos seleccionados inicialmente con otrs objetos no seleccionados.

El objetivo del algoritmo es la de minimizar el promedio de disimilaridad de los objetos con los objetos cercanos seleccionados.

Algoritmo K-Means

Es uno de los más usados actualmente más que todo por su robustez y eficacia, El procedimiento consiste en el siguiente:

- Se seleccionan *k* centroides al azar de las clases que se van a buscar.

- Se calcula la distancia euclidiana de cada uno de los objetos a los *k* centroides seleccionados en el paso anterior y se cuadra la pertenencia del objeto al *k* centroide más cercano por el valor de la distancia.
- Se recalcula el centroide del primer paso, como la media de todos los objetos que lo componen, buscando minimizar el valor de una función de coste, que es igual a la sumatoria de todas las sumatoria de la distancia euclídea de los objetos de cada clase al centroide de su respectiva clase, tal como se muestra en la figura 2.25.

$$J = \sum_{i=1}^{k} \left(\sum_{j, z_j \in A_i} ||z_j - ci|| \right) \tag{2.25}$$

El éxito en este algoritmo esta en la selección del valor *k*, ya que si es muy grande o muy pequeño va suceder que se creen grupos ficticios o se agrupan objetos que deben pertenencer a *cluster* distintos.

2.5 Estado del Arte

2.5.1 Industria 4.0

La nueva revolución industrial o como se ha conocido la industria 4.0 ha generado un proceso disruptivo en las empresas, que día por día incluyen elementos de los siguientes 3 factores que se encuentran mutuamente interconectados tal como se explica en [Zezulka et al., 2016].

1. La digitalización y la integración de cualquier técnica simple económica a la interconexión de redes técnicas más complejas económicamente.
2. Digitalización de ofertas de productos y servicios.
3. Nuevos modelos de negocios.

Otra definición de la industria 4.0 la podemos encontrar en [Rojko, 2017] y [Pereira and Romero, 2017] que la define como un nuevo modelo de organización y de control de la cadena de valor a través de sistemas de fabricación apoyados por tecnología. Es una aproximación basada en la integración de los procesos comerciales y de fabricación, así como de todos los actores de la cadena de valor de una empresa (proveedores y clientes), donde el sistema de ejecución se basa en la aplicación de sistemas ciber físicos y tecnologías como Internet de las Cosas, robótica, *Big Data* y realidad aumentada, para el desarrollo de procesos de fabricación más inteligentes, que incluyen dispositivos, máquinas, módulos de producción y productos que pueden intercambiar información de forma independiente y controlarse entre sí, permitiendo un entorno

de fabricación inteligente.

Los sistemas tecnológicos integrados con control descentralizado y conectividad avanzada que caracterizan a la industria 4.0 recopilan e intercambian información en tiempo real con el objetivo de identificar, rastrear, monitorear y optimizar los procesos de producción. Además, presentan un amplio soporte de software basado en versiones descentralizadas y adaptadas de sistemas de ejecución de fabricación y planificación de recursos empresariales para una integración perfecta de los procesos de fabricación y comerciales. Otro aspecto importante es el manejo de una gran cantidad de datos recopilados de los procesos, máquinas y productos. Por lo general, los datos se almacenan en un almacenamiento en la nube [Rojko, 2017]

2.5.2 Situación en Colombia

Por otra parte ante la inminente llegada de las tecnologías disruptivas en el país, tales como el *Big Data* ó el *Data Mining*, entre otras, el Gobierno Nacional no ha sido indiferente a estos retos y a través del Consejo Nacional de Política Económica y Social (CONPES) ha publicado políticas públicas acordes al Plan Nacional de Desarrollo (PND) 2018-2022 Pacto por Colombia, Pacto por la Equidad y creo los CONPES 3920 [Calderón et al.,] y CONPES 3975 [Márquez et al.,] con el fin de hacer frente y preparar las empresas y personas para estas tecnologías.

Los CONPES mencionados anteriormente son importantes ya que uno habla de la transformación digital e inteligencia artificial que tienen que afrontar las empresas y define un lineamiento base para la misma en Colombia, este documento muestra las debilidades y retos que tienen las empresas Colombianas para implementar una transformación digital exitosa tal cual como se muestra en la figura 2.6

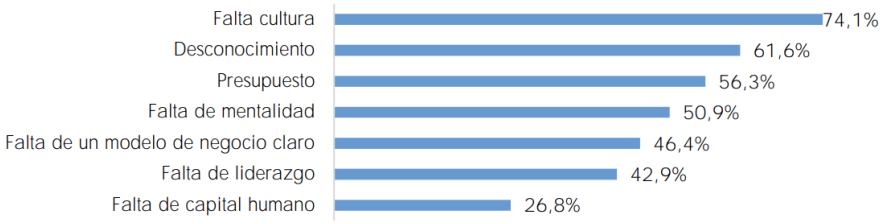

Figura 2.6: Barreras y Desafios que enfrentan las empresas para lograr una transformación digital exitosa. Tomado de [Márquez et al.,]

Esta encuesta muestra que aunque en la actualidad se hablé de Industria 4.0, *Big Data* y otras tecnologías, las empresas Colombianas no están preparadas para una transformación digital que contenga la aplicación de las tecnologías anteriormente mencionadas, adicionalmente las empresas pertenecientes a los sectores industrial, comercio y servicios que participaron en la Gran Encuesta TIC adelantada en el 2017 [Márquez et al.,] arroja que el 66% de las empresas no tienen un área, dependencia o persona encargada de los temas de TIC (Tecnologías de información y comunicación), una de los principales motivos para no contar con dicha área es que el negocio o entorno de negocio no lo exige.

Esta percepción de no exigencia del mercado en tener un área de TIC o una persona que este pendiente de las tecnologías emergentes y logre incluirlas en el sector privado hace que Colombia este muy lejos de países que sí implementan áreas de TIC e invierten en tecnologías emergentes como se muestra en 2.7

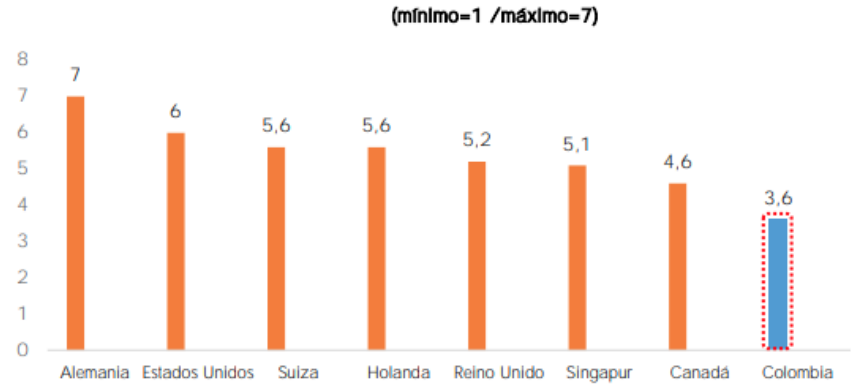

Figura 2.7: Inversión de las empresas en tecnologías emergentes. Tomado de [Kearney, 2018]

Ahora bien, el PND 2018-2022 *Pacto por Colombia, Pacto por la Equidad*, en el Pacto VII Pacto por la transformación digital de Colombia se plantean 2 vías complementarías para la transformación digital: en primer lugar se hace referencia a la masificación de Internet de banda ancha e inclusión digital para toda la población colombiana. Esto tiene como objetivo crear condiciones habilitantes para la masificación de las TIC a través del fortalecimiento del marco normativo del sector: en segundo lugar, la transformación digital se enfoca en la implementación de tecnologías digitales avanzadas (tales como *blockchain, IoT, IA*, entre otras) y en la búsqueda de una relación más eficiente, efectiva y transparente entre mercados, ciudadadano y Estado.

Igualmente el BID en [Antonio and Enrique, 2017] menciona que sectores como el TIC, financiero y comercio tienen alto potencial, tanto de apropiarse

de las TIC como de beneficiarse de la economía digital. En contraste hay otros sectores donde la apropiación de las TIC es más lenta, pero también podrían beneficiarse bastante de la economía digital (salud, agricultura, transporte y gobierno). El impacto de digitalización en Colombia es desigual en distintos sectores económicos, en general los medios tradicionales son preferidos que los digitales (sitio web de la entidad, correo electrónico y aplicación móvil) esto se puede deber a un analfabetismo digital de la sociedad, una poca capacidad de propagación y uso de los ecosistemas digitales o simplemente desinterés por el uso de nuevas tecnologías por parte de las mismas empresas, por ejemplo la Encuesta de Desarrollo e Innovación Tecnológica realizada por el DANE muestra que menos del 40 porciento de las empresas en sectores industriales (alimentos, textiles, metalúrgicos y de refinación de petróleo) usan internet para innovar. [DANE, 2017]

En el CONPES 3920 Política Nacional de Explotación de Datos (BIG DATA) se busca aumentar el aprovechamiento de los datos en Colombia, mediante el desarrollo de condiciones para que estos sean gestionados como activos generadores de valor social y económico en el país. Uno de los grandes limitantes para la explotación masiva de los datos en Colombia es que no se cuenta con el capital humano para la misma tarea, o esta en incipiente crecimiento.

En la linea de acción 8. Medición de la brecha de capital humano y actualización de compentencias del CONPES 3920, se indica que para la masificiación del aprovechamiento de datos en todos los sectores y la transición hacia una economía más intensiva en conocimiento aumenta la demanda de personal que pueda desempeñarse con solvencia en este contexto. Es decir, con educación y habilidades relacionadas con la explotación de este activo. Particularmente, matemáticas, estadística, aprendizaje de máquina y ciencia de datos. [Calderón et al.,]. Es así como el Ministerio de Tecnologías de la Información y las Comunicaciones no sólo ha realizado estudios para medir las diferencias entre los perfiles de cargo necesarios a cubrir y su demanda sino también los conocimientos y competencias específicas para el aprovechamiento de los datos, sino que también ha diseñado e implementado cursos de formación en materia de explotación de datos, al igual que ruedas de negocios donde se creen empresas dedicadas a la explotación de datos.

Es importante resaltar que la creación del CONPES 3920 permitio que se crearán sitios como www.datos.gov.co el cual permite la consulta de la base de datos de las distintas entidades territoriales, para llegar a esto en este mismo documento se planteo los documentos para evitar la duplicidad de datos entre las distintas bases de datos como a su vez la normalización de la base de datos, también se presenta el tema legal para la publicación de estos datos.

La gráfica 2.8 muestra la proyección de crecimiento de los datos para el año 2025 a nivel mundial mostrando que va haber un crecimiento mayo en los tipos de datos no estructurados, seguido de los datos estructurados y por último los datos semiestructurados que incluyen los datos que tienen que ver

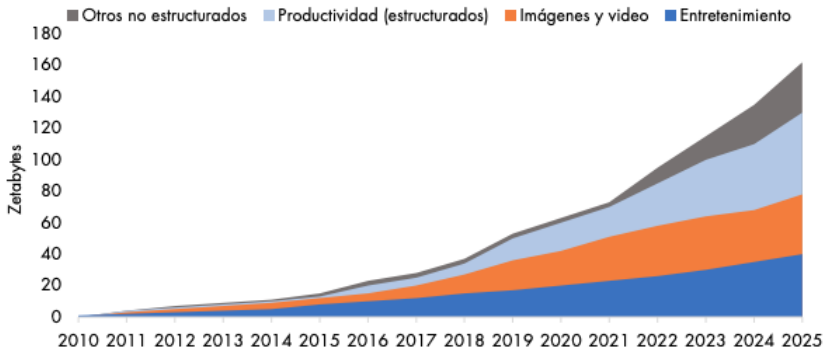

Figura 2.8: Generación de Datos Estructurados y no estructurados proyectados a 2025. Tomado de DNP con datos de [Reinsel et al., 2019]

Estructurado	Están organizados conforme a un modelo o esquema. Se almacenan en forma tabular y algunas veces su estructura también incluye la definición de las relaciones entre ellos. Típicamente están representados en bases de datos que hacen parte del funcionamiento de sistemas de información.
No estructurado	Su organización y presentación no está guiada por ningún modelo o esquema. En esta categoría se incluyen, por ejemplo, las imágenes, texto, audios, contenidos de redes sociales, vídeos.
Semiestructurado	Su organización y presentación tiene una estructura básica (etiquetas o marcadores), pero no tiene establecida una definición de relaciones en su contenido. En esta categoría se incluyen contenidos de e-mails, tweets, archivos XML.

Tabla 2.2: Organización y almacenamiento de los datos digitales, Tomado de [Calderón et al.,]

con imágenes, vídeos y entretenimiento.

Los datos como se menciono anteriormente pueden ser de 3 tipos tal cual como lo muestra la tabla 2.2.

El *Big Data* es la inmensa cantidad de datos que son generados día a día por diversas fuentes principalmente por la masificación de dispositivos móviles, esto hizo que las bases de datos tradicionales y sus consultas sean insuficientes para procesar y almacenar esta gran cantidad de datos, por lo que se debe remontar al año 2001 cuando [Laney, 2001] identifico el reto tecnológico que implicaba la generación de datos cuyo volumen, velocidad y variedad desafiaba los sistemas tradicionales. El crecimiento exponencial de los datos hace que cantidades que se consideraban enormes hace unos años hoy día sean considerados normales [Adolph, 2013]. Tomando esto como base el reto de los países y las empresas

no esta en solo almacenar datos sino darle valor a los datos almacenados siendo el insumo central de la economía digital.

Hoy es posible por ejemplo que a través de un *smartwatch*, el médico de una persona pueda monitorear en tiempo real el estado de salud de un paciente y generar una medicina preventiva en vez de una medicina correctiva, otro ejemplo se puede mencionar una persona que va en su carro pero no lo esta conduciendo sino que el carro es autónomo y permite que la persona pueda ir haciendo otras actividades y conduciendo de forma adecuada, también se puede ingresar a un restaurante donde no hay mesero, el pedido se hace a través de una tablet y es entregado a través de una banda transportadora no sin antes saber que la comida fue preparada por un robot [Oppenheimer, 2018].

La figura 2.9 muestra el proceso que expone el CONPES 3920 se debe realizar para la explotación de datos y sí se hace un simil con las metodologías *KDD ó CRISP-DM* tiene una estructura base muy parecida donde existe una etapa de generación y recolección, luego un momento de compartición y agregación, paso seguido el proceso de explotación y por último el proceso de innovación con la información adquirida o con los procesos realizados, transversal a esto, el documento permite ver que el capital humano transversal a estos procesos son la oferta en el mercado y la cultura de datos que cree el gobierno o las empresas es la demanda, de nada sirve contar con un capital humano capacitado si no se implementan procesos de cultura organizacional para la debida recolección y documentación de los datos, de igual forma tampoco sirve contar con una cultura de datos pero tener una ausencia de capital humano para su explotación

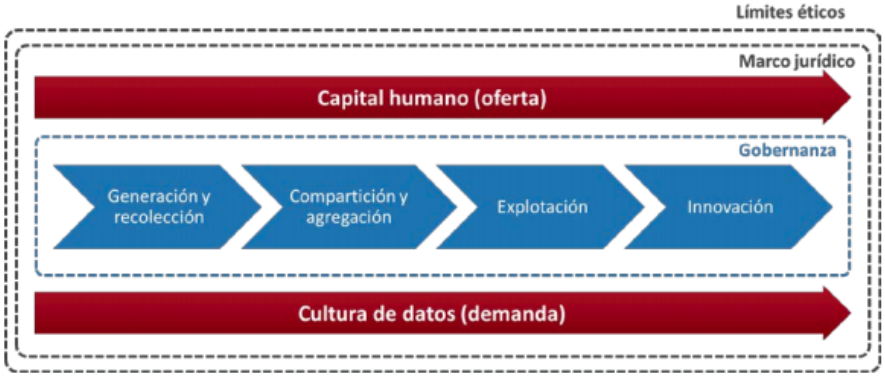

Figura 2.9: Generación de valor con datos. Tomado de DNP

Las políticas públicas mencionadas anteriormente, como también el proceso natural de la economía digital hace que hoy día sea posible por ejemplo que a través de un *smartwatch*, el médico de una persona pueda monitorear en tiempo

real el estado de salud de un paciente y generar una medicina preventiva en vez de una medicina correctiva, otro ejemplo se puede mencionar una persona que va en su carro pero no lo esta conduciendo sino que el carro es autónomo y permite que la persona pueda ir haciendo otras actividades y conduciendo de forma adecuada, también se puede ingresar a un restaurante donde no hay mesero, el pedido se hace a través de una tablet y es entregado a través de una banda transportadora no sin antes saber que la comida fue preparada por un robot [Oppenheimer, 2018].

Todas estas actividades humanas están interconectadas por un montón de sistemas de comunicación en este momento. Las tecnologías más prometedoras son *Internet of Things (IoT)*, *Internet of Services (IoS)* y *Internet of People (IoP)*.[Zezulka et al., 2016].

IDC, indica que todo esto ha generado unagran cantidad de datos en distintos sectores económicos, con mayor crecimientos en algunos de ellos, pero con una proyección de crecimiento en otros, como por ejemplo el séctor salud o el sector transporte.

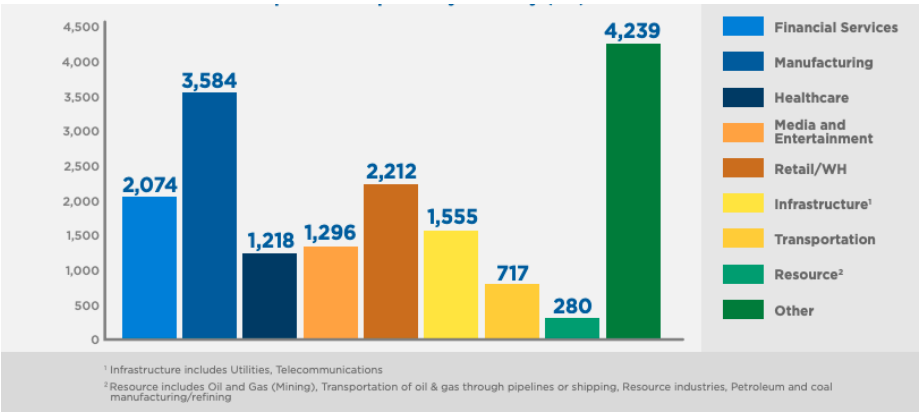

Figura 2.10: Datos de Empresas por tipo de Industria, tomado de: [Reinsel et al., 2019]

2.5.3 Situación en el Sector Transporte

Como se evidencia en la 2.10, la industria del transporte no es ajeno a este fenómeno, la cual ha servido para el caso de Colombia que las empresas se tecnifiquen y pasen de tener sus modelos construidos en la experiencia de su personal operativo a contar con ecosistemas digitales donde la tecnología es el pilar fundamental tal cual como se documenta en [Giraldo Martínez, 2014].

Estas tecnologías no están disponibles sólo para servir al ser humano sino también al mundo, es así como las tecnologías cada día buscan ser más limpias así como innovadoras, ahora no sólo basta con innovar a través de la tecnología sino medir el impacto que esto tendrá en el medio ambiente, es así por ejemplo como en [Kargupta et al., 2001] se habla no sólo del control sobre las emisiones del gas invernadero por parte de los sistemas de transporte sino también el rol de la minería de datos en los sistemas de transportes, los cuales son:

- Reducción de gases de efecto invernadero a través de medios tecnológicos y socio-económicos.
- Desarrollo de motores eficientes y limpios y trenes de fuerza incluyendo tecnologías híbridas.
- Usar combustibles alternativas para las aplicaciones de transporte, en especiales partículas de hidrógeno y celdas de combustible.
- Tener en cuenta la relación costo-eficiencia y eficiencia energética.
- Desarrollar estrategias del fin de ciclo de vida para los vehículos.

Las empresas de transporte tienen una canasta de costos fija, donde tiene 3 rubros que cobran una relevancia en cualquier estudio que son el combustible, llantas y el capital humano, por lo que buscan de todas las formas posibles reducir el costo generado por estos mismos. Como se menciono en el capitulo 1, INTEGRA S.A. tiene un modelo de bonificación al capital humano, primero buscando reducir el costo de combustible y segundo retener a las personas por más tiempo en la compañia para evitar procesos de inducción continuos.

Como se ve en [Rolim et al., 2017b] esta preocupación no es única en INTEGRA S.A. sino también por ejemplo del transporte público de la ciudad de Lisboa en Portugal, donde el transporte fue el responsable de cerca del 81 porciento del consumo de energía en este país. Esta claro que la inclusión de vehículos eléctricos y de tecnologías verdes como de eficiencia energética reducirá la huella de carbono que la operación del transporte de pasajeros deja, pero ninguna de estas tecnologías podrán ser bien aprovechadas si no se toman patrones adecuados de conducción por parte de los operadores de buses, el estudio de [Rolim et al., 2017b] permite ver el impacto que tiene un sistema de retroalimentación en tiempo real para mejorar el rendimiento de combustible.

Hay que resaltar que aunque este sector tiene críticas muy grandes por su alta dependencia a los combustibles fósiles, aumento del consumo de energía, niveles de congestión y ruido demasiado altos, por lo que se han generado distintos soluciones para reducir por ejemplo el consumo energético como se muestra en [Saboohi and Farzaneh, 2009]. La innovación en este sector no solo ha sido en temas de reducción de emisiones de carbono o menor consumo energético sino que también se han ido incluyendo tecnologías de información

y comunicación *(ICT, information and communication technologies)* o conocidas con la sigla en español TIC, estas tecnologías permitirán reducir la congestión de tráfico, reducir los tiempos de los servicios, reducir los costos de combustible e incrementar la eficiencia de los operadores [Mannering et al., 1995]

Las *ICT* permiten tener dispositivos de monitoreo abordo que recogen información de los patrones de conducción como por ejemplo (velocidad, tiempo de marcha en vacio, kilometraje, número de aceleraciones y desaceleraciones entre otras variables) [Rolim et al., 2017a] estos datos permiten realizar correciones en habitos o patrones de conducción a los operadores de los buses permitiendo analizar donde presenta las deficiencias y haciendo ejercicio de refuerzo en puntos críticos o situaciones determinadas, también permite crear conciencia en el comportamiento de conducción y sus consecuencias de manejar adecuadamente o de forma no adecuada, permite tener una reducción de cónsumo de combustible y por lo mismo una reducción de emisión de polución[af Wåhlberg, 2007].

Sea como sea esta búsqueda de reducción de costos nunca puede ir en contra de la calidad en la prestación de servicio y siempre se debe tomar las necesidades de los pasajeros como lo son la accesibilidad al servicio, mantenimiento de los vehículos, empleados amables y valor económico del servicio, tal como lo muestra la figura 2.11 en la India donde los pasajeros consideran la seguridad, el tiempo de viaje y el comfort como características esenciales de un servicio de transporte público.

.

Dentro del estudio que hace [Rolim et al., 2017b] se muestra que para definir patrones o técnicas de conducción se debe tener en cuenta distintas variables, no sólo se puede revisar las variables relacionadas al bus o las variables relacionadas al conductor, ya que se puede relacionar depronto una mala conducción cuando el verdadero problema puede estar relacionado con el estado de mantenimiento de la máquina, por lo que es muy importante delimitar cuales son las características que deben tenerse en cuenta para un estudio en particular o cuales no se pueden controlar para identificar cuando un patorn de conducción sea incorrecto o correcto tener la claridad que se tienen las variables correctas.

.

Como se evidencia en la figura 2.12 no sólo se tienen que analizar la información sociodemográfica del conductor, sino que también se tiene que evaluar, el vehículo, los patrones de movilidad y los indicadores del ambiente, en caso de no poder documentar toda esta información se debe limitar muy bien el estudio, ya que los patrones de conducción relacionan sólo datos cuantificables de la máquina y se debe encontrar su relación con las otras características anteriormente mencionadas.

Estudios alrededor de los patrones de conducción como los de [Rohani,

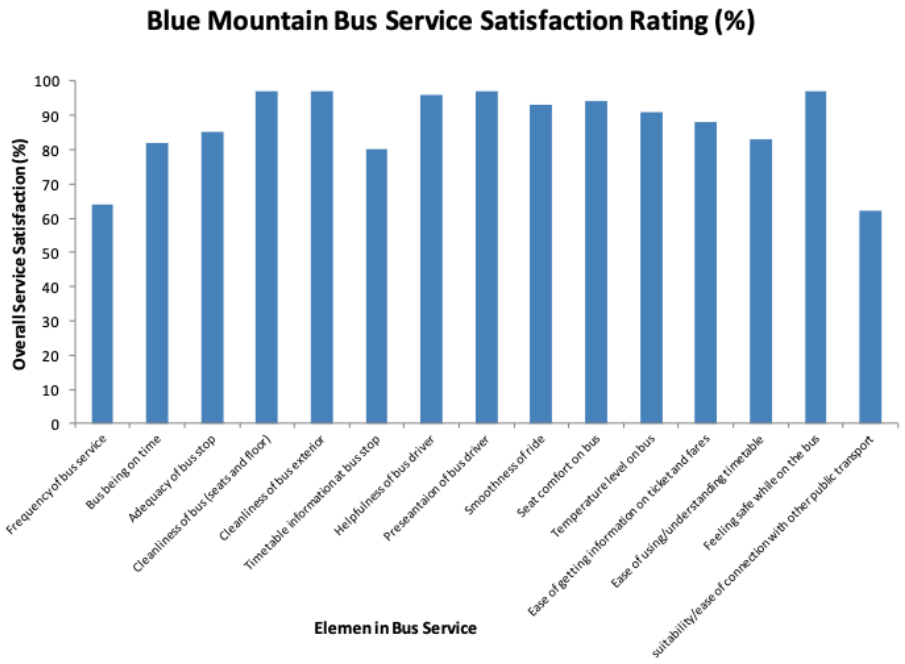

Figura 2.11: Satisfacción de los pasajeros en Blue Mountain Bus Service, tomado de: [Rohani et al., 2013]

2012], [Rohani et al., 2013], [af Wåhlberg, 2007], [Strömberg et al., 2015] hablan sobre la importancia no sólo de medir estos patrones, sino que haya un proceso de retro alimentación en la medida de lo posible buscan siempre que esta sea en tiempo real o realizando un nivel de comparación de antes de la capacitación y posterior a esta, pero lo más importante es que esta se mantenga en el tiempo.

En [Rolim et al., 2017a] se mencionan las características de comportamiento del operador, al igual que los comportamientos de conducción de un operador de bus de servicio público, por ejemplo para lo primero se da la definición de patrón usando como referencia a [Ouellette and Wood, 1998] que dice que la practica repetitiva de comportamientos en contextos constantes va crear automáticamente patrones de comportamiento sin la deliberación de la conciencia, es por lo anterior que se puede evidenciar que al conducir se generan distintos patrones entre las personas, por ejemplo algunas personas siempre mantienen el pie en el pedal del embrague, otras son más agresivas para conducir y otras simplemente conducen de una manera más lenta. La conducción es un proceso que esta basado en 2 variables, la primera es el rendimiento y la segunda es el comportamiento, el rendimiento con practica, técnica y entrenamiento puede ser mejorada pero el comportamiento es relacionado con

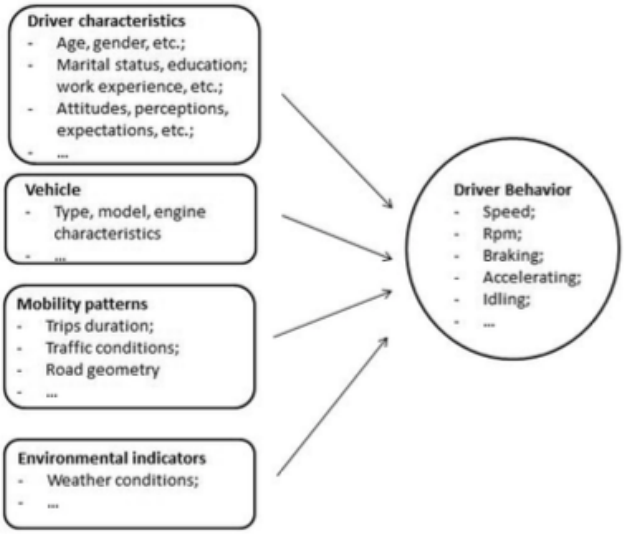

Figura 2.12: Algunas de las variables que son usadas para determinar los patrones de conduccion, tomado de: [Rolim et al., 2017b]

los hábitos propios de la persona y estos pueden estar formados por distintos factores como la edad, el genero, el stress que maneja la persona entre otras.

Cambiar los hábitos de conducción en las personas es posible pero para esto es necesario dar las herramientas que permitan que la persona se pueda ir reeducando [Murcotts, 2015], la retroalimentación positiva, la motivación y la actitud de la persona para mejorar cada día va ayudar a obtener mejores resultados. Los comportamientos mentales de las personas frente a las actitudes que se toman frente a evaluar los resultados son materia de otro tipo de estudio que no se considerará en el presente documento.

Dentro de los comportamientos de conducción se puede evidenciar que estos pueden ser muy distintos entre conductores experimentados o con conductores novatos, adicional a como los conductores los afecta el stress o la negatividad por parte de la demandas de los usuarios y el ambiente, adicional de las condiciones viales, no es lo mismo manejar en un sistema *BRT, Bus Rapid Transit*, donde la mayor parte del tiempo tiene un carril exclusivo y las exigencias pueden ser menores a realizar la prestación del servicio en rutas de carriles mixto y no contar con sistemas de pago electrónicos. De igual forma estos conductores dentro de sus habilidades deben tener las habilidades para cumplir con los servicios establecidos por el proceso operativo, manejar con una calidad de servicio adecuada y por último transportar a los usuarios con rapidez y seguridad, es importante recordar como se mencionaba en [Rohani et al., 2013] es importante tener el comfort, la seguridad y el cumplimiento de servicios como

componentes a tener en cuenta.

En [Zfnebi et al., 2017] hacen un análisis y clasificación de las variables cuantitativas que sirven para la construcción de un modelo para el comportamiento de conducción, se comienza con una revisión de la literatura, luego de la revisión de la literatura, realizan la extracción de las variables para el comportamiento de conducción y por último clasifican los parámetros del comportamiento de conducción, para la construcción del módelo evitan realizar proceso de encuestas e incluirlas en el análisis ya que el juzgamiento de las respuestas del cuestionario puede estar muy basada en las percepciones propias de las habilidades o de la percepción de terceros, en el segundo momento realizan una extracción de las variables que son usadas en los distintos documentos consultados en el estado del arte del paso 1, en el tercer y último paso califican las variables en 2 grandes categorías, la primera como prioritaria y la segunda como secundaria, la clasificación de la variable en cada categoría esta basado en la importancia o impacto dentro del articulo analizado, luego de esto le dan un *rate* a cada variable definiéndola como la cantidad de apariciones sobre la cantidad de artículos consultados.

Este estudio permite conocer como muestra la figura 2.13, una de la variable más usada para cualquier análisis es la velocidad, seguido por la aceleración o desaceleración y la acción de frenado. Estos resultados son importantes ya que la velocidad y la aceleración están estrechamente relacionados con la sensación de confort en un viaje.

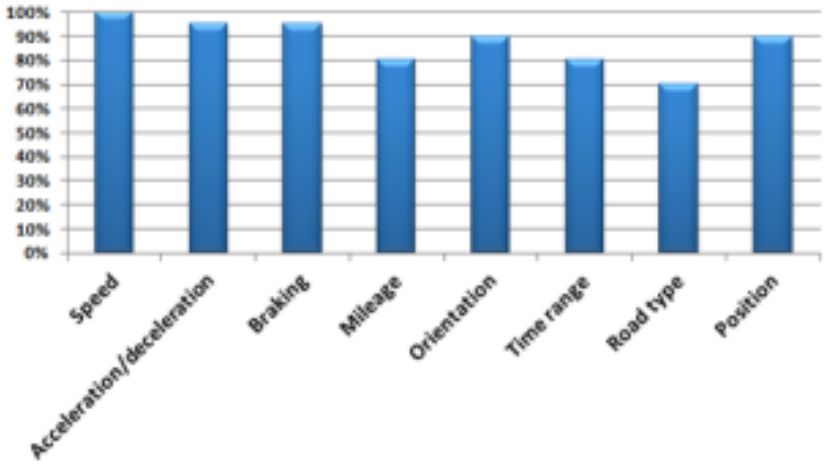

Figura 2.13: *Rate* de variables más usadas en distintos artículos, tomado de: [Zfnebi et al., 2017]

Orden de Importancia	Variable	Tipo
1	Velocidad	Prioritario
2	Aceleración Desaceleración	Prioritario
	Frenado	Prioritario
3	Orientación	Secundario
	Posición	Prioritario
4	Rango de tiempo	Prioritario
	Kilómetraje	Prioritario
5	Tipo de Carretera	Secundario

Tabla 2.3: Orden y Categoría de Importancia de las variables, fuente elaboración propia y apoyo en [Zfnebi et al., 2017]

La figura 2.14 permite conocer la categoría a la que pertenece cada una de las variables, siendo una variable prioritaria o secundaria, todo esto basado en distintos artículos publicados.

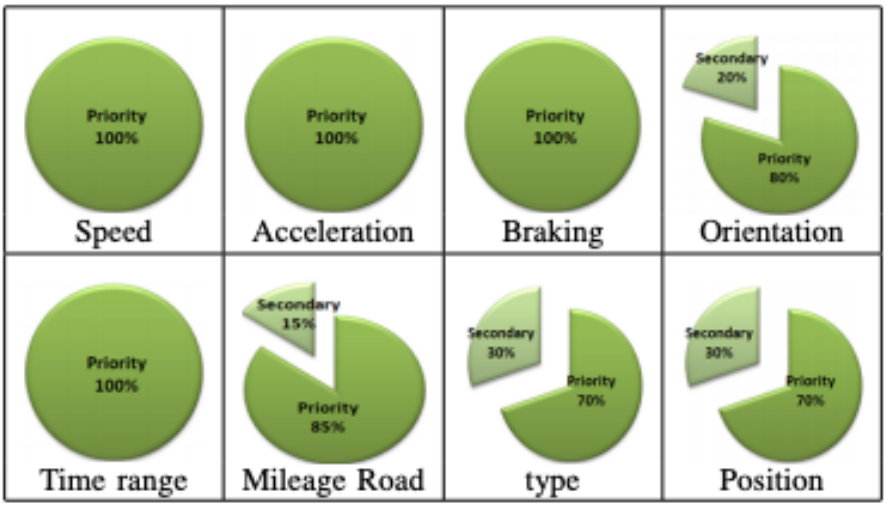

Figura 2.14: *Rate* Selección categoría primaria o secundaria, tomado de: [Zfnebi et al., 2017]

La tabla 2.5.3 muestra el orden de importancia y el tipo de variable que es, si es prioritaria o secundaria. Es importante resaltar que este orden no es único y que como muestra en el artículo.

En la Figura 2.15, se expresa lo que en [Martinez et al., 2017] se identifica que la eficiencia de combustible esta directamente influenciado pero no sólo exclusivo de las características de los vehículos, el tipo de carretera, las

condiciones del tráfico y los patrones de conducción, Por otro lado en el mismo artículo definen que el estilo de conducción es un concepto complejo influenciado por distintos factores que complica su descripción. El proceso de conducción entiende un evento de tareas como la aceleración, la desaceleración, los cambios de carril y todas estas tareas pueden ser usadas para identificar el estilo de conducción, como se puede observar las variables o factores que se usan son similares a las que se mencionan en la 2.5.3, en este artículo se hace una diferenciación entre los patrones de conducción y el comportamiento de conducción, el patrón de conducción está relacionado con el tipo de carretera, las condiciones del clima y es conocido también como condiciones de conducción, en cambio el comportamiento de conducción esta enfocado en el conductor u operador y las decisiones que toma y los factores externos incluyendo también las habilidades de conducción, esto sirve para diferenciar los conductores experimentados de los conductores promedios u ordinarios.

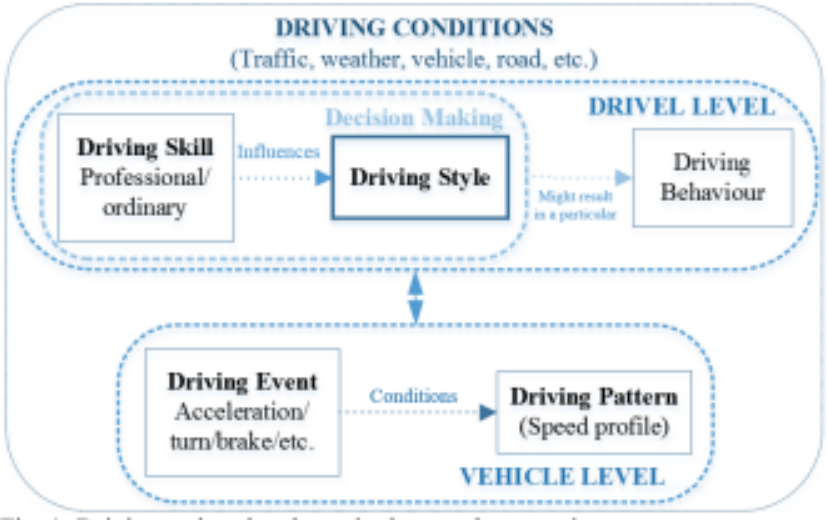

Figura 2.15: Condiciones de manejo, tomado de: [Martinez et al., 2017]

Al realizar una revisión de los factores que influencian el estilo de conducción se encuentra la definición de [Taubman-Ben-Ari et al., 2004] donde se analizan los factores humanos como la experiencia de conducción, el entrenamiento que tiene la persona y la familiaridad que tiene con el vehículo y el ambiente pueden definir los estilos de conducción, también otros factores como el genero, la edad, el nivel educativo, el tipo de personalidad de la persona, la autoestima, la paciencia, el manejo de ira, la imprudencia, entre otras, también es importante en el estudio del artículo el estado del conductor y si este se encuentra influenciado por drogas, fatiga, distracciones entre otras, esto respecto a los factores humanos, respecto a los factores externos, se tiene el tráfico, el tipo de clima, la estación en la que se encuentra, no es lo mismo manejar en

verano que en invierno, el tipo de carretera, si es carril exclusivo o carril mixto, la hora del día, no es lo mismo manejar en hora pico a manejar en hora valle, manejar con el vehículo lleno al vehículo vació. Es decir en el momento de realizar el análisis, no sólo se debe analizar las condiciones de los datos que se arrojan por la telemetría del *GPS Global Position System* sino también se debe dedicar un poco del análisis a como los factores anteriormente mencionados pueden afectar en la condición, o si existe algún patrón especifico a conductores con condiciones humanas similares.

En [Constantinescu et al., 2010] realizan un análisis de distintos conductores basados en variables básicas, donde nuevamente la velocidad, la aceleración, el frenado y el trabajo mecánico se utiliza para hacer un análisis del estilo de conducción usando minería de datos, lo importante de este análisis es que la fuente de información es generada a través de un *gps* que esta instalado en el vehículo y enviada la información a un servidor usando la red de comunicaciones *GPRS, General Packet Radio System*, el artículo analiza a 23 distintos conductores de la ciudad de Bucarest usando las variables anteriormente mencionadas y aplicando distintos métodos de análisis como PCA, HCA permiten encontrar que es posible clasificar a los conductores de los buses basando su estilo de conducción en las variables anteriormente mencionadas y relacionando los *clusters* resultantes en 6 Niveles que son: no agresivo, un poco agresivo, neutral, agresivo moderado y muy agresivo, este estudio permite que mes a mes se pueda volver a realizar el modelo y encontrar si cierto conductor salto de un *clusters* o si paso de un nivel a otro ya sea de forma positiva o de forma negativa.

Por su parte [Hwang et al., 2018] hace un análisis estadístico de comportamientos de conductores específicos, este estudio se basa en la información almacenada de los conductores en una plataforma *cloud* donde la información acumulada alcanza niveles de *Big Data* rápidamente, basándose paquetes como *numpy, pandasy scipy* puede generar una respuesta de los eventos por un árbol de decisión basado en conducción defensiva, defensivo débil, agresivo débil y agresivo. Para obtener estos resultados Wang realizo un proceso de limpieza de datos, integración de datos, selección de datos, transformación de datos, minería de datos y conocimiento o evaluación, en otras palabras uso la metodología *KDD*, en el paso 1, se realiza la limpieza de datos a través de la velocidad, usando la media y calculando la desviación estándar ya que se tenían muchos datos en 0, luego de eso se hace una integración de datos, donde se toma la velocidad, las revoluciones por minuto en el motor y otros parámetros. En la selección de los datos, da un ranking a las variables establecidas en el paso 2, dejando las RPM de primero, de segundo la velocidad y por último el consumo de combustible, el proceso de minería de datos se hace con un árbol de decisión que es entrenado para clasificar en los estados anteriormente mencionados.

Como se muestra en la figura 2.16 la arquitectura del *Web Server* recibe como variables de entrada la velocidad, la posición, y el tiempo que se reciben los datos, estos son obtenidos mediante el gps que se tiene en el vehículo, mien-

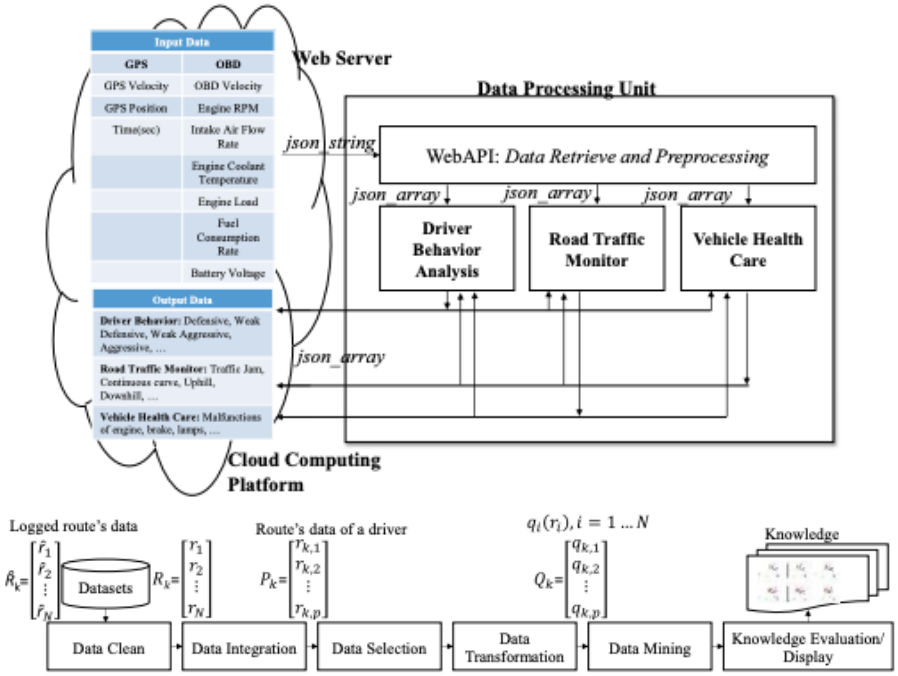

Figura 2.16: Arquitectura y proceso del sistema propuesto por [Hwang et al., 2018]

tras que por la conexión con el OBD se puede recibir la Velocidad del OBD, esta velocidad es distinta a la recibida del GPS ya que es dada por la computadora abordo del sistema, por lo que es un poco más precisa, también se reciben las revoluciones por minuto del motor, la rate del flujo de aire, temperatura del liquido refrigerante, la carga del motor, la rate del consumo de combustible y el voltaje de la batería. Lo ideal en estos estudios es tener la mayor cantidad de variables de forma cuantitativa, ya que como se menciono anteriormente los datos cualitativos son datos basados en una apreciación de persona a persona o de la misma persona. Estos datos tanto del OBD como del GPS son enviados a través de un string de json, es importante recalcar que esta conexión de escucha de los datos esta protegida por una conexión cifrada para que estos datos no sean robados o corruptos. La Web Api recive la información y hace la limpieza, selección y transformación de los datos y pasa 3 distintos *array* de *json* que son el Análisis del comportamiento del conductor, el monitoreo del tráfico en el camino y por último el cuidado de la salud del vehículo, esto entrega unas salidas que para el caso del conductor ya se sabe que son lso 4 estados de conducción, defensivo, poco defensivo, poco agresivo y agresivo. Así mismo para el monitoreo de tráfico se obtienen otros datos y para el seguimiento de la salud del vehículo se puede encontrar el mal funcionamiento del motor, de los frenos

entre otros.

La segunda parte de la gráfica muestra cada una de las fases que se han explicado y parte del proceso que se recibe. Para que un estilo de conducción sea exitoso no se debe realizar una selección ilimitada de variables ya que esto genera reglas innecesarias y complejas que en vez de ayudar complican la minería de datos y la creación de *cluster* ya que podría suceder que los mismos no alcancen a estar debidamente definidos y se mezclen con otros.

En [e Silva et al., 2015] hacen un estudio evaluando 3 distintos campos, la ruta, el conductor y el vehículo, por ejemplo para el análisis de los conductores u operadores de los buses se usa la edad, el tiempo laborando en la empresa, actividades de entrenamiento o re-entrenamiento y el número de eventos de conducción como por ejemplo, la cantidad de excesos de velocidad, puertas abiertas, excesos de aceleración entre otras), para el vehículo toman el peso, la edad del vehículo y el consumo de combustible. Para el caso de las rutas toman el tiempo de viaje, la distancia de viaje, velocidad comercial, el tipo de ruta, si es urbana, semi urbana o rural.

Con todos estos campos se procede a calcular a cada uno la mediana y la desviación estándar, antes de realizar estos cálculos se hace una limpieza de los datos lo que permite eliminar datos erróneos y reemplazarlos por otros o eliminarlos por completo. Con estos datos y con información cruzada facilitada por el Operador se podía relacionar el vehículo-ruta-conductor y encontrar los conductores que podían estar por encima o por debajo de la ruta, como resultados se obtuvo por ejemplo que los entrenamientos y re-entrenamientos son técnicas esenciales que sirven para reducir el consumo de combustible por rutas, esto quiere decir que a mayor cantidad de entrenamientos mejor rendimiento en el consumo de combustible se va obtener, de igual forma se encontró que el personal de conductores con una mayor edad tienen más resistencia al cambio de las habilidades de conducción por lo que deben ser un grupo focal al cual se le debe realizar demasiada insistencia o buscar procesos para agilizar su retiro, por último y a diferencia de lo que se cree el estudio demostró que las aceleración y desaceleraciones afectan el rendimiento del combustible pero no son valores determinantes por lo menos en las sesiones de monitoreo.
Volviendo a [Martinez et al., 2017] indican que existen 3 formas para reconocer los estilos de conducción:

- El primero consiste en sistemas basados en reglas, donde se seleccionan parámetros que se deben cumplir o monitorear, esto facilita la interpretación e implementación pero se limita a una cantidad de datos que pueden ser monitoreados, estos generalmente son basados en un sólo parámetro y posterior a esto lo robusto y la exactitud de los resultados puede ser considerablemente limitado.
- El segundo es basado en modelos, este consiste en la descripción del estilo de conducción o la habilidad de conducción a través de ecuaciones

o características predefinidas, este modelo es ajustado a cada uno de los parámetros de conducción para lograr que los datos obtenidos puedan ser usados en el método o modelo presentado.

- El tercero es el aprendizaje de máquina y es ampliamente usado cuando se deben procesar grandes cantidades de datos y obtener información valida ya sea de forma general o llevarlo al específico de un conductor, dentro de este aprendizaje de máquina podemos tener el aprendizaje supervisado y el aprendizaje no supervisado, en el primero se puede decir que es cuando la máquina aprende conociendo los resultados, por ejemplo si una persona que tiene dolores de cabeza, perdida de la visión y y congelación de un sector de su cuerpo tiene generalmente un ataque cerebro vascular, entonces a través de esto el algoritmo, revisa las condiciones donde estas condiciones se presentaron y aprende que cuando se presentan estas situaciones es que la persona va tener un ataque cerebro vascular, el aprendizaje no supervisado es cuando la máquina desconoce los resultados pero a través de patrones logra descifrar una conducta y es ahí donde se genera nuevo conocimiento a través de la máquina.

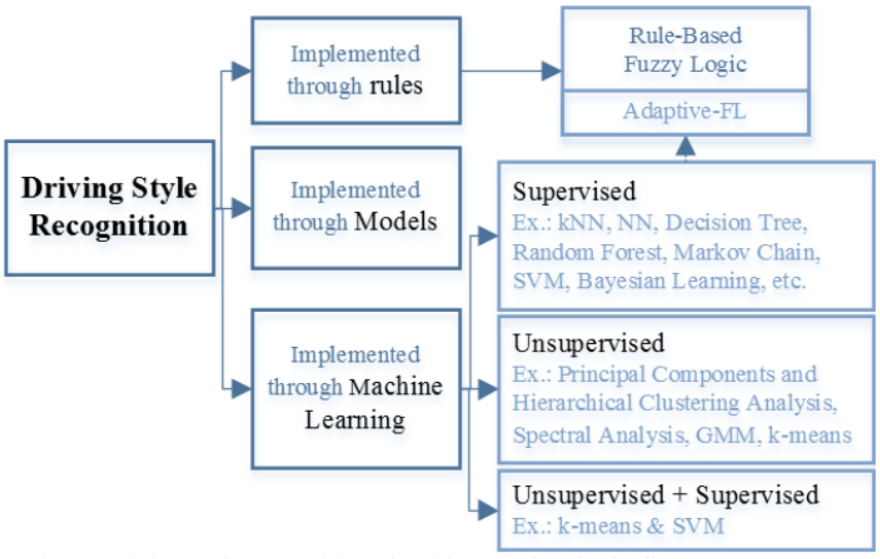

Figura 2.17: Tipos de algoritmos para reconocimiento de estilos de conducción tomado de [Martinez et al., 2017]

Capítulo 3

Descripción Detallada del Proceso

3.1 Selección del Método

Como se vio en 2.2 hay varias metodologías para el proceso de obtener información a través de la minería de datos, la *KDD* y *CRISP-DM*, la primera usada más que todo en el campo académico, mientras que la segunda se usa más para la inteligencia de negocios de las empresas.

CRISP-DM es una de las metodologías más usadas en la actualidad, financiada por la comunidad europea que se ha unido para desarrollar una plataforma para minería de datos, usada por grandes empresas tales como IBM, LLoyds Banking Group, ABB entre otras. Está basada en actividades ordenadas en 6 fases. Una de las ventajas más grandes de esta metodología es su neutralidad al uso de herramientas, está enfocada a los problemas de negocios, así como al análisis técnico entre otras.

En el primer objetivo de este documento se plantea determinar las variables que afectan los comportamientos de conducción normal de los operadores de los vehículos articulados y alimentadores a través del entendimiento de las necesidades de la empresa. Basándose en la metodología CRISP-DM en este objetivo se implementarán las 3 primeras etapas o fases, el primer paso a realizare es entender el requerimiento de la empresa INTEGRA S.A., que en este caso es poder caracterizar la población de conductores respecto a su comportamiento al conducir los vehículos articulados y alimentadores para premiar o castigar su bonificación mensual monetaria, debe quedar muy claro que el objetivo del modelo no considera simular cuanto debe recibir cada operador por esta bonificación, por el contrario este modelo sólo caracterizará la población operativa en grupos o clases, esto con el fin de poder realizar un seguimiento mes a mes.

El segundo paso dentro del objetivo es la selección de los datos del modelo, por lo que es necesario basarse en los datos con los que cuenta la empresa en este momento y que estén alineados con las necesidades de la empresa, por ejemplo, al quererse categorizar la población de operadores, información relacionado con los técnicos de mantenimiento no debe ser tenida en cuenta, por lo cual en esta fase se realizará un filtrado riguroso en la selección de la consulta a la base de datos. Inicia la construcción de una base de datos la cual contiene todas las características requeridas como candidatas para una variable que se espera predecir (selección de datos), también se realiza un proceso de limpieza de datos para obtener un nivel de calidad óptimo requerido por las técnicas de minería de datos seleccionadas. Como esta fase sirve de insumo para la fase siguiente, puede ocurrir más de una iteración sobre esta fase. Como se mencionó durante el Marco Teórico estos 3 primeros pasos son conocidos como el proceso *ETL*.

El segundo objetivo, es donde se generará un modelo utilizando técnicas de *Clustering* que permita la obtención de los comportamientos de conducción normal de los operadores de los vehículos articulados y alimentadores usando las variables previamente definidas, en esta etapa se hace la fase 4 de *CRISP-DM* y se realizará la prueba de *clustering* con distintos algoritmos buscando el que mejor se ajuste, adicionado o quitando datos esto con el fin de tener una mejor consistencia del mismo.

El tercer objetivo es la validación del modelo generado comprobando la confiabildiad del mismo por medio de un proceso estadístico. Con el modelo ya construido se validarán los resultados respecto al objetivo del entorno del negocio definido en el objetivo 1, donde se deberá seleccionar como resultado el algoritmo con mayor confiablidad, se realizará una retroalimentación en caso tal que se encuentren comportamientos que salgan de los parámetros normales del objetivo del negocio, si los niveles de confiabilidad en el objetivo 3 para cualquiera de las simulaciones del objetivo 2 con distintos algoritmos son demasiadas bajas se deberá proceder a realizar nuevamente todo el proceso desde el objetivo 1, esto con el fin de lograr una mayor confiabilidad en la evaluación de los resultados

Por último el cuarto objetivo que es la interpretación de los resultados obtenidos en conjunto con el grupo de expertos de INTEGRA S.A.m basado en los conocimientos operacionales para definir las características de la población, tendríamos la última fase del proceso *CRISP-DM* ya que con una interpretación y validación de los resultados el modelo quedaría listo para ser implementado y mes a mes hacerle seguimiento al mismo.

3.1.1 Entendimiento del Negocio:

INTEGRA S.A. es un grupo económico con presencia en 2 países de Latinoamérica, Colombia y Perú y cuenta en sus empresas con 2 empresas de transporte masivo, una en la ciudad de Pereira, Colombia y Arequipa, Perú, así mismo en esta última ciudad cuenta con una operación de taxis.

Integra S.A. operadora de transporte masivo cuenta con 37 buses articulados de marca VOLVO, 24 buses modelo B10 y 13 buses modelo B12 y tiene su operación en el área metropolitana Centro Occidente (AMCO), cuenta con más de 200 empleados, la mayor parte de ellos en la parte operativa. Una característica principal de INTEGRA S.A. es que la mayor parte de sus colaboradores llevan más de 5 años en la empresa, tal como lo muestra la Figura 3.1, con un 32 mientras que en el periodo de 1 año a 5 años se tiene el 49 por ciento, la suma de estos 2 suma el 81 por ciento. Esta permanencia de los empleados se debe en gran parte a la seguridad monetaria, al buen ambiente de trabajo y a la formación constante que tienen los empleados tal cual como muestra la tabla **??**

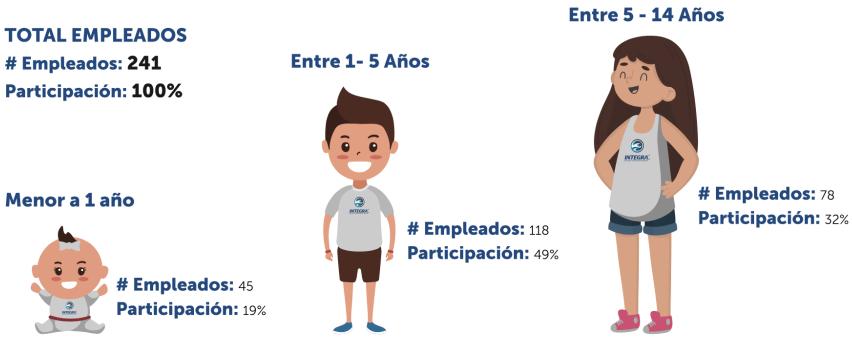

Figura 3.1: Tiempo de los empleados de INTEGRA S.A., tomado de Informe de Gestión 2019 INTEGRA S.A.

Como se puede observar en las Tablas 3.1, 3.2, 3.3 y **??**, la formación es un proceso constante en INTEGRA S.A. tanto que hace parte de su ADN como empresa y esto se puede notar ya que año tras año su personal operativo tiene que realizar o contar con más cursos es así como en el año 2019 los operadores de buses tanto de articulados como de alimentadores reciben formación alistamiento de vehículos, servicio de movilización de pasajeros, conducción e vehículos de transporte masivo, estos cursos de formación siempre están enfocados en reducir los grandes costos operativos mencionados en el Capitulo 2 que es el combustible, lubricantes, llantas y personal.

Un personal que se sostiene en el tiempo genera un menor impacto en el costo, ya que un alto índice de rotación acarrea con ello costos ocultos que a

Certificación en el periodo	2014	2015	2016	2017	2018	2019
Nivel Avanzado	16	3	0	0	37	16
Nivel Intermedio	55	48	6	0	87	21
Nivel Básico	10	45	28	0	27	5
Total Operaciones Certificados	81	96	34	0	151	42

Tabla 3.1: COMPETENCIA SENA EN ALISTAR EQUIPOS DE TRANSPORTE MASIVO DE PASAJEROS

Certificación en el periodo	2014	2015	2016	2017	2018	2019
Nivel Avanzado	0	0	0	14	0	32
Nivel Intermedio	0	0	0	145	0	12
Nivel Básico	0	0	0	1	0	0
Total Operaciones Certificados	0	0	0	160	0	44

Tabla 3.2: COMPETENCIA SENA EN SERVICIO DE MOVILIZACIÓN DE PASAJEROS

Certificación en el periodo	2014	2015	2016	2017	2018	2019
Nivel Avanzado	0	15	0	0	0	4
Nivel Intermedio	0	72	52	0	0	27
Nivel Básico	0	3	41	0	0	3
Total Operaciones Certificados	0	90	93	0	0	34

Tabla 3.3: COMPETENCIA SENA CONDUCIR LOS VEHÍCULOS DE TRANSPORTE AUTOMOTOR MASIVO

veces son difíciles de cuantificar para las empresas como lo son: entrega de dotación, plan padrino para aprender la cultura organizacional de la empresa y se aprovecha mejor las formaciones que se le da a cada uno de los empleados como se mostraba en la Tabla **??**. Por otra parte un empleado que siente que su empresa cree en él, termina generando un compromiso hacia la empresa y trata de hacer mejor su trabajo, es así como no sólo la empresa realiza inversión en formar a sus empleados en competencias técnicas sino que también genera espacios para que muchos de ellos se puedan terminar de formar ya sea en primaria, bachillerato o cuadrar sus turnos de trabajo para que pueda realizar estudios universitarios nocturnos. Otro valor agregado es que en el año 2019 INTEGRA S.A. fue reconocida por Happy At Work como una empresa feliz según sus trabajadores y como un excelente lugar para trabajar al superar el estandard 1.

INTEGRA S.A. cuenta con un área de Investigación, Desarrollo e Innovación la cual ha ideado y desarrollado con éxito distintos proyectos de Ciencia, Tecnología e Innovación, para lo cual, era indispensable en primera medida la creación de alianzas estratégicas con el estado, la academia.

Es así como en su plan estratégico 2017-2021 Integra S.A. definen cuatro (4) valores que abarcan y moldean la cultura corporativa.
[INTEGRASA, 2017]

- **Autenticidad:** Representa la veracidad, integridad y honradez de nuestra organización y su personal. Implica tener la iniciativa como empresa de ser coherentes, estables y sinceros en el desarrollo de las labores.
- **Efectividad:**Es la eficiencia y eficacia con las que desarrollamos nuestras labores y procesos, atendiendo de manera fluida y satisfactoria la prestación del servicio.
- **Compromiso:** Surge de la convicción personal en torno a los beneficios que trae el desempeño responsable de las tareas a nuestro cargo. El Compromiso permite pasar de las promesas a los hechos, generando resultados y beneficios tangibles.
- **Innovación:** Es la generación de valor para la organización por medio de la aplicación de técnicas y uso de herramientas de ideación y creatividad. La Innovación nos permite encontrar mayores beneficios de los existentes.

Todos estos referentes son importantes ya que como se puede notar la empresa tiene un alto grado de innovación y de procesos de investigación por lo que la inclusión de nuevas tecnologías son constantes, es por esto que los vehículos de la compañía cuentan con sistemas de comunicación a bordo y sistemas de posicionamiento global *GPS*, estos dispositivos envían información cada 3 segundos, sobre la latitud, longitud, velocidad, aceleración, dirección entre otros datos, así mismo aunque no se va abarcar en esta investigación la

adquisición de los 4 valores relacionados en la planeación estratégica hace que el personal que trabaje en INTEGRA S.A. se sienta feliz de trabajar en la empresa, genere un compromiso en la misma, sean auténticos, traten de ser más efectivos y generen procesos de innovación.

En la Tabla 3.4 se presentan las bonificaciones en términos de salarios mínimos mensuales legales vigentes para los operadores de INTEGRA S.A. esta es indistinta para un operador de vehículo alimentador o de articulado.

Categoría	Incentivo
Operador Alimentación	0,23 SMLV
Operador Articulado	0,23 SMLV

Tabla 3.4: Incentivo de bonificación, tomado de Procedimiento Incentivos Sistema de Gestión Integral

Esta bonificación está compuesta por distintos parámetros que se describen en la Tabla 3.5 y los ítems pueden ser acumulativos en el tiempo, por ejemplo un operador puede quedarse varios periodos sin bonificación así haya cumplido con los ítems, pero si incurrió en un daño, no recibirá bonificación hasta que haya cancelado por completo el daño, un operador puede incurrir en distintos ítems lo que hará que se vea afectado en un mayor o menor valor su bonificación hasta un 100 %, esto quiere decir que si la persona, incurre en faltas que suman más de este porcentaje no verá afectado su ingreso de bonificación el siguiente mes por el porcentaje excedido en el mes anterior.

Es de aclarar que los factores son acumulables en un semestre, esto quiere decir que si la persona llega tarde una primera vez se le descuenta un 10% de la bonificación, una segunda llegada tarde en un mes distinto pero en un periodo menor de 6 meses genera un descuento del 20% y así sucesivamente, cada tercer evento se generará una medida disciplinaria y perdida total de esa bonificación.

En la Tabla 3.5 se evidencia que gran parte de los factores son de tipo cuantitativo, aunque algunas son o pueden ser parte de la subjetividad, por ejemplo el ítem de mala presentación personal, aunque se pueda definir un estándar para la presentación personal de los empleados de la empresa, no se puede definir que es mala presentación para una persona X o para una persona Y y 2 personas distintas pueden tener conceptos distintos de una mala presentación personal y esto puede suceder con otros ítems.

Factores que afectan la productividad	Porcentaje de afectación
Golpe o daño (g)	Valor del daño
Llegada tarde (l.t)	10
Mala presentación personal (p.p)	10
Incumplimiento de tiempo de operación (o)	5
Incapacidad o ausentismo (i.a)	25
Mal Servicio al cliente	10
Sanción	20
Incumplimiento a los procedimientos de operación (i.p.o)	5
Incumplimiento normas de tránsito (i.n.t)	10
Abandono del puesto de trabajo	100
Accidente (a)	50
Mala actitud (ac)	20
Altercado (a)	100
Evasión (e)	50
Llegada en estado de embriaguez	100
Mala imagen corporativa	30
Técnicas de conducción (t.c)	10
No uso de herramientas TIC (Tablet, Micrófono)	50

Tabla 3.5: Parámetros de bonificación

3.1.2 Entendimiento de los Datos

INTEGRA S.A. tiene 2 grandes fuentes de información, la primera fuente es el sistema InnoBUS Masivo, el cual es un software de transporte ajustado a las necesidades empresariales de los transportes masivos y donde la empresa controla la realidad empresarial de los procesos pilares de la compañía que son: talento humano, mantenimiento y operaciones. En este software se recibe la información brindada por el *GPS* se computa y se presenta en informes visuales que muestran la información del momento, de igual forma esta información es almacenada y se conserva los datos cada 3 segundos.

La Figura 3.2 es la interfaz de monitoreo y control de vehículos en operación, es una de las interfaces más usadas por parte de este proceso, en la misma se muestra el movimiento de los vehículos acorde a los datos enviados por el *GPS* conectado en el bus, esto permite al proceso de Operaciones revisar espacialmente cómo están repartidos los vehículos y tomar decisiones en tiempo real. Integra S.A. al ser un operador de transporte masivo recibe unas programaciones que debe cumplir al ente gestor Megabús S.A., las programaciones pueden ser de 4 tipos, programación de día hábil, programación de día sábado, programación de día festivo, programación días especiales. Estas a su vez están formadas por servicios los cuales tienen una hora de inicio y una hora de fin y según lo dictaminado por contrato, el operador de transporte masivo tiene entre

2 y 4 minutos para que se tome el servicio como cumplido, fuera de este rango el servicio se tomará como incumplido, esta variable es importante ya que los cambios de velocidad y aceleración a veces están determinados porque los operadores al querer cumplir con el rango de estos tiempos conducen más rápido lo que puede ocasionar que se genere un mayor consumo de combustible, en ciertas ocasiones los operadores realizan una conducción un poco más lenta, en este caso no hay un aumento del consumo de combustible, pero sí se genera una pérdida del servicio definido por el ente gestor por llegar sin una justificación por encima del tiempo de llegada.

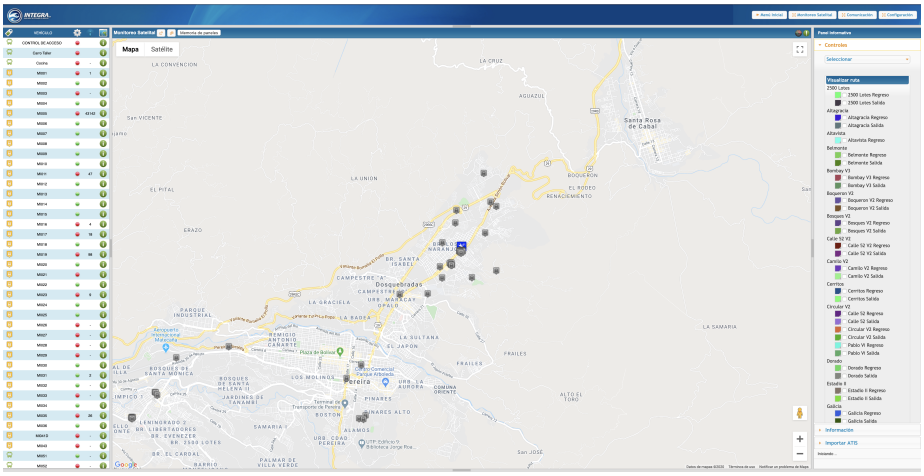

Figura 3.2: Interfaz de monitoreo del Software InnoBUS Masivo en operación. Fuente: Elaboración propia.

El sistema en esta interfaz cuenta también con un manejador de alertas que informa si hay un exceso de velocidad superior a 50 km/h, como también puede adicionar en el tablero de control rutas y estaciones. Esta interfaz aunque es muy completa permite un control en el momento, ya para revisiones posteriores se debe recurrir a la interfaz que se presenta en la Figura 3.3, donde toda la información que es visualizada en la interfaz presentada en la Figura 3.2, es almacenada para su posterior consulta, en esta interfaz donde se puede consultar por fecha, se puede consultar por vehículo, por ruta y por servicio y por conductor. Una de las situaciones que presentan problema es que al recibir datos cada 3 segundos, la cantidad de datos a computar se vuelve extremadamente grande y realizar comparativos entre vehículos es más complejo, ni hablar en las situaciones cuando se requiere analizar conductores o rutas que incluyen más datos.

Una de las grandes debilidades del sistema InnoBUS es que cuenta con un módulo de modelado de datos que permite realizar consultas a la base de datos

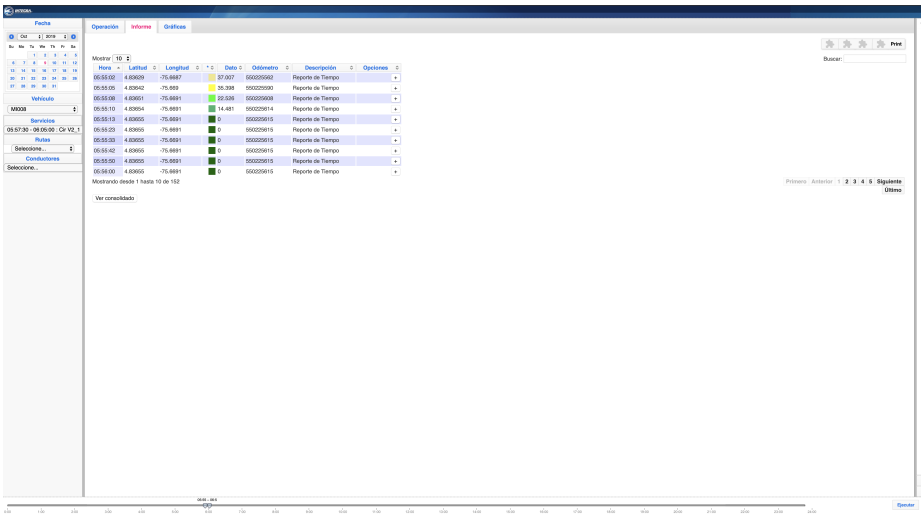

Figura 3.3: Interfaz del reporte de operación del software InnoBUS Masivo en operación. Fuente: Elaboración propia.

y a sus distintos esquemas, donde se requiere tener un conocimiento avanzado de las distintas relaciones entre las tablas existentes del sistema lo que para un usuario final vuelve esto un tema muy difícil, por otra parte para poder realizar las consultas se requiere tener conocimiento de SQL (Structured Query Language) y un poco de conocimiento en relaciones de bases de datos. Tal como se muestra en la Figura 3.4 se puede ir seleccionando las tablas que se requieran y los campos, adicional en este módulo solo se permite la consulta mas no permite funciones de inserción o de eliminación de los datos. En este módulo las consultas se hacen en en tiempo real, por lo que si se hacen consultas por parte del usuario en su desconocimiento que conlleve al tratamiento de grandes cantidades de datos al sistema, esto hace que que se vea afectada su operación normal, por lo que el uso de este módulo esta restringido al personal de Tecnologías de la información en compañía de la empresa desarrolladora, de igual forma las consultas realizadas en este módulo no pueden ser canceladas por lo que si el usuario realiza el envío de una consulta muy grande, se debe esperar hasta que este proceso termine dentro del servidor aunque el usuario haya realizado el cierre.

La Figura 3.5 presenta la interfaz del control de personas de Talento Humano, aunque es una interfaz bastante limpia, tiene datos muy básicos y no permite realizar extracción de los mismos para ser tratados externamente, en este módulo se incluye toda la información relacionada con el empleado, recolectando resultados de su evaluación desempeño realizada por sus jefes directos, distintos tipos de datos personales, su nivel educativo, la cantidad de capacitaciones que posee, las novedades que le han sido registradas, estos datos son

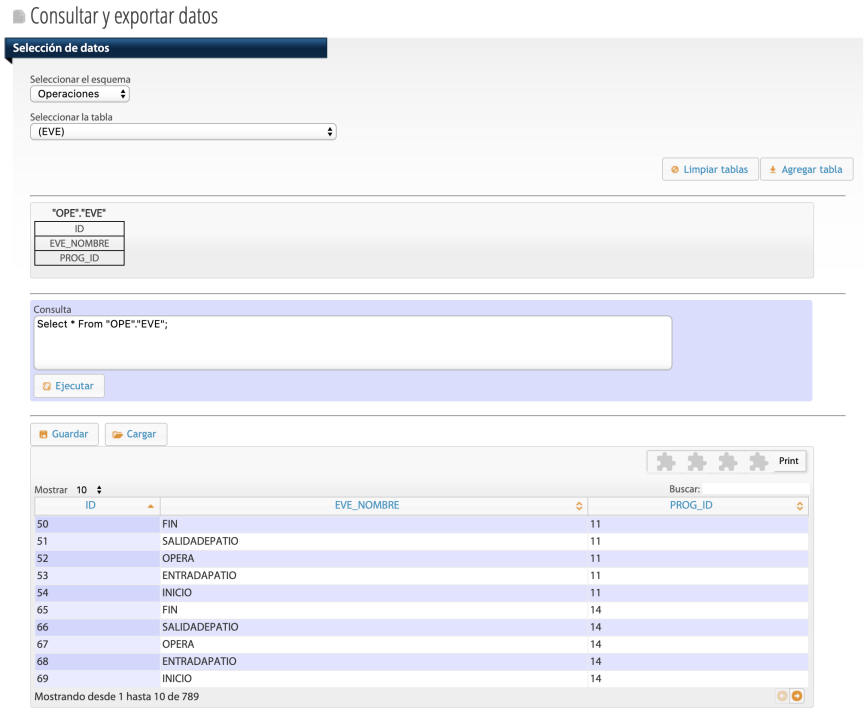

Figura 3.4: Interfaz de Consulta de Datos en InnoBUS Masivo. Fuente: Elaboración propia.

alimentados por el proceso de Talento Humano y generalmente no están actualizados ya que la mayoría de datos tienen que ser actualizados en distintos sistemas como la base de datos de Nómina que se usa para la liquidación quincenal de la misma, esta carencia de una fácil conexión entre distintos sistemas con el software InnoBUS dificulta tener información fiable en el módulo de Talento Humano, esto se debe a la alta dependencia de InnoBUS de recibir datos por parte de una interacción humana, para los casos de los módulos de Operaciones y Mantenimiento este efecto no es tan notorio ya que la recolección de la información es más automática.

Al quererse realizarse un análisis de datos en cualquier ejercicio no sólo se debe preocuparse por preparar los datos como se verá en la etapa posterior sino que también se tiene que tener la claridad de la fuente de los datos, la confiabilidad y la forma de su recolección, puede que un dato sea muy importante para el análisis pero su recolección sea compleja o no se cuente con los medios necesarios, por ejemplo para el caso de estudio, un dato muy importante sería las Revoluciones por minuto del vehículo o RPM, esto con el fin de encon-

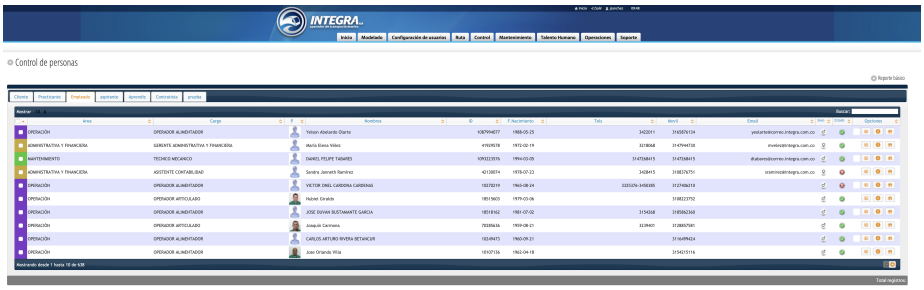

Figura 3.5: Interfaz del control de personal del software InnoBUS Masivo en operación. Fuente: Elaboración propia.

trar una relación de mayor consumo de combustible a una mayor cantidad de revoluciones, aunque este dato es importante, la recolección u obtención del mismo exige un proyecto o un proceso más grande como por ejemplo acceder en tiempo real a través del módulo CAN-BUS y crear un sistema de recolección de estos datos lo cual no aplicaría para el actual proyecto, de igual manera este valor se podría calcular por medio de formula, pero se tendría que analizar el trabajo computacional de hacer este análisis y revisar si realmente es un valor cercano o es un valor teórico.

Otro punto a valorar es la confiabilidad de los datos, esto puede suceder en ocasiones donde se manejan muchos eventos al mismo tiempo y el dato es mal tomado, por ejemplo la recolección se hace por una persona que en sus funciones no sólo tiene la toma de datos sino otras más, al tener un evento de 2 o más situaciones, realizará el evento de más prioridad y puede que relegue la toma del dato para ser escrito en un momento posterior, cuando llegue el momento de escribir el dato puede ocurrir, que la persona simplemente no recuerde el dato exacto y de un valor aproximado o decida no colocar el dato. De igual forma la fuente del dato es importante puede ocurrir el caso que el dato tenga que ser escrito a mano y no sea legible y la persona que transcribe este valor no lo entienda y escriba el valor que el cree que es más no el valor que exactamente fue.

Usando la interfaz de la figura 3.3 se recolecta los distintos datos que se requieren para el análisis y usando como respaldo los datos más usados en este tipo de investigación se procede a buscar recolectar datos de velocidad, aceleración, desaceleración, orientación, posición, rango de tiempo de trabajo, ruta, kilometraje recorrido. Es importante resaltar que no se realizará la recolección de variables categóricas. A continuación se definirán las variables a usar dentro del documento:

- **Velocidad:** Variable de tipo numérica que puede contener valores entre 0 a 90 Km/h, esta variable esta definida por los valores recolectados del

GPS, esta variable se recolecta cada 3 segundos.

- **Aceleración y Desaceleración:** Variable de tipo numérica que puede contener valores entre 20 y -20 metros sobre segundo al cuadrado, esta variable luego se tiene que crear una variable categórica para indicar el tipo de evento si es una aceleración se deberán tomar los valores positivos, en caso de ser una desaceleración deberá ser un valor negativo.
- **Posición**: Variable de tipo numérica. Esta compuesta por 2 variables, Latitud y Longitud que la unión de las 2 da la posición del reporte del vehículo en un momento determiando, al igual que la velocidad y la aceleración esta variable es recolectada del GPS cada 3 segundos.
- **Ruta:** Variable tipo texto. Indica la ruta por la cual estuvo operando el vehículo y el operador, se recolecta de la programación de servicios que es subida en InnoBUS y los cambios de programación que pudo haber ocurrido en el día.
- **Kilometraje recorrido:** Variable de tipo numérica que da la cantidad de kilómetros recorrido en un día por un vehículo o por un operador, este sale de la resta del primer dato del odómetro del día con el último dato del mismo filtrando estos datos ya sea por vehículo o por operador.
- **Nombre del Conductor:** Variable de tipo texto, indica el nombre completo del conductor
- **Vehículo**: Variable de tipo texto, Identifica el vehículo con el cual se esta operando o que reporta el dato de la posición del GPS.
- **Fecha**: Variable de tipo Date, Identifica la fecha y hora del evento.

3.1.3 Preparación de los datos

Esta fase es una de las fases más importantes ya que como se muestra en la Figura 2.2 iterará mucho con la etapa de modelamiento en caso de no conseguirse los resultados esperados o en caso que no se lleguen a tomar decisiones con el modelamiento que se realice, en esta fase también se hace la selección de las variables a usar en la etapa posterior, su limpieza y transformación.

Como se evidenció en la fase anterior con las variables seleccionadas para el estudio, estas provienen de distintas fuentes y pueden tener distintos formatos, además de contener complejidades para su recolección.

Al comenzar con la descarga de los datos a través del módulo de modelado se encontró que la cantidad de datos era demasiado grande aproximadamente más de 11 millones de datos, adicional a esto y como se ha descrito en el documento este proceso de consulta se hacía en el servicio de producción por lo que cada vez que se intentaba descargar este proceso se generaba un proceso que

hacía que el sistema se detuviera por varios minutos y no generaba la consulta.

En la Tabla 3.6 se presentan los códigos de los valores a consultar en la Red *GPS* que tienen los vehículos de Integra S.A.

Tabla de Red Gps
RepGPSCodigo
EquCodigo
TramGpsCodigo
EveGPSCodigo
FechaGps
HoraGps
LongGps
LatGps
VelGps
DirGps
AcelGps
OdoGps
AltGps

Tabla 3.6: Tabla de Red *GPS* con cada uno de sus campos

La consulta que se uso inicialmente para extraer es la siguiente:

*SELECT * FROM GPS.REPGPS WHERE (GPS.REPGPS.FECHAGPS BETWEEN '2018-10-01' AND '2018-10-30')*

Pero como se mencionó, esta tabla al ser demasiado pesada por todos los campos que se manejan y se ven en la tabla por lo que la anterior consulta se modifica y se procede a realizar una nueva consulta, en este caso se hace el proceso de grupos de 10. El tiempo aproximado de respuesta de la consulta es de entre 2 minutos a 3 minutos, con la complejidad que al realizarse este proceso de forma continua el rendimiento va mermando hasta alcanzar tiempos más altos entre 4 a 5 minutos o inhabilitar la función de extraer la consulta en un archivo plano de *.csv*.

Con la nueva consulta que se plantea a continuación, se pueden obtener los resultados deseados, pero se encuentra el inconveniente que se van generando datos en grupos de 10 vehículos por lo que para un sólo día en vehículos articulados se generan 4 archivos que posteriormente deben ser unificados en uno solo, este proceso inicialmente se hacía de manera manual, pero sumado al proceso de extraer los datos y ahora compilar todos los archivos en uno sólo se presentaban 2 situaciones, se adicionaba un proceso manual a la ejecución de la misma que podía generar errores y por ende llevar a una posible perdida de la información y la segunda situación se sumaba más tiempo a este proceso de

extracción de información que ya estaba alrededor de 15 minutos.

*SELECT * FROM GPS.REPGPS WHERE GPS.REPGPS.FECHAGPS" ='2019-10-01' AND (GPS.REPGPS.EQUCODIGO BETWEEN 'MI051' AND 'MI060')*

Como se ve en la Figura 3.6 es el resultado de la Tabla 3.6, para el caso que se muestra en la figura el archivo contiene más de 100 mil registros. Por lo anterior para optimizar este proceso se realizó un script en python que una vez teniendo los archivos ubicados todos en una carpeta al correr genera un nuevo archivo unificando todos los archivos que se encuentren en la carpeta, este código puede ser consultado en el Apéndice A de este documento. El resultado de este script es que se pasa de tener 131 elementos para el mes de Octubre a tener solo 1 elemento tal cual como se muestra en la figura 3.7.

20191001MI051MI060

REP_GPS_CODIGO	EQU_CODIGO	TRAM_GPS_CODIGO	EVE_GPS_CODIGO	FECHA_GPS	HORA_GPS	LON_GPS	LAT_GPS	VEL_GPS	DIR_GPS	ACL_GPS	ODO_GPS	ALT_GPS
MI051_1832614921789237	MI051	EV	3	2019-10-01	04:35:38	-75.7444	4.79857	0	0	0	28229762	
MI051_1832614964979379	MI051	EV	3	2019-10-01	04:35:46	-75.7444	4.79857	0	0	0	28229762	
MI051_1832616366109032	MI051	EV	3	2019-10-01	04:35:56	-75.7444	4.79857	0	0	0	28229762	
MI051_1832617794914447	MI051	EV	3	2019-10-01	04:36:06	-75.7442	4.79907	0	101	0	28229768	
MI051_1832619516509591	MI051	EV	3	2019-10-01	04:36:14	-75.7442	4.79908	0	101	0	28229770	
MI051_1832619543973570	MI051	EV	10	2019-10-01	04:36:18	-75.7442	4.79908	0	101			
MI051_1832619567713135	MI051	EV	3	2019-10-01	04:36:24	-75.7442	4.79908	0	101	0	28229770	
MI051_1832621219564093	MI051	EV	3	2019-10-01	04:36:33	-75.7442	4.79908	0	101	0	28229770	
MI051_1832621261913959	MI051	EV	3	2019-10-01	04:36:42	-75.7442	4.79908	0	101	0	28229770	
MI051_1832622485957549	MI051	EV	3	2019-10-01	04:36:51	-75.7442	4.79908	0	101	0	28229770	
MI051_1832624089714516	MI051	EV	3	2019-10-01	04:37:00	-75.7442	4.79908	0	101	0	28229770	
MI051_1832625425889108	MI051	EV	3	2019-10-01	04:37:09	-75.7442	4.79908	0	101	0	28229770	
MI051_1832629106016149	MI051	EV	3	2019-10-01	04:37:18	-75.7442	4.79908	0	101	0	28229770	
MI051_1832638566436340	MI051	EV	3	2019-10-01	04:37:27	-75.7442	4.79908	0	101	0	28229770	
MI051_1832647726472078	MI051	EV	3	2019-10-01	04:37:36	-75.7442	4.79908	0	101	0	28229770	
MI051_1832656366345604	MI051	EV	3	2019-10-01	04:37:45	-75.7442	4.79908	0	101	0	28229770	
MI051_1832664806694465	MI051	EV	3	2019-10-01	04:37:53	-75.7442	4.79908	0	101	0	28229770	
MI051_1832674246569902	MI051	EV	3	2019-10-01	04:38:03	-75.7442	4.79908	0	101	0	28229770	
MI051_1832683966509107	MI051	EV	3	2019-10-01	04:38:12	-75.7442	4.79908	0	101	0	28229770	
MI051_1832693446592857	MI051	EV	3	2019-10-01	04:38:22	-75.7442	4.79908	0	101	0	28229770	
MI051_1832702967157405	MI051	EV	3	2019-10-01	04:38:31	-75.7442	4.79908	0	101	0	28229770	
MI051_1832711186900377	MI051	EV	3	2019-10-01	04:38:37	-75.7442	4.79908	0	101	0	28229770	
MI051_1832720726968952	MI051	EV	3	2019-10-01	04:38:46	-75.7442	4.79908	0	101	0	28229770	
MI051_1832729027337296	MI051	EV	3	2019-10-01	04:38:55	-75.7442	4.79908	0	101	0	28229770	
MI051_1832738607220521	MI051	EV	3	2019-10-01	04:39:04	-75.7442	4.79908	0	101	0	28229770	
MI051_1832747067027753	MI051	EV	3	2019-10-01	04:39:13	-75.7442	4.79908	0	101	0	28229770	
MI051_1832756407176110	MI051	EV	3	2019-10-01	04:39:22	-75.7442	4.79908	0	101	0	28229770	
MI051_1832765047505100	MI051	EV	3	2019-10-01	04:39:30	-75.7442	4.79908	0	101	0	28229770	
MI051_1832774467781985	MI051	EV	3	2019-10-01	04:39:40	-75.7442	4.79908	0	101	0	28229770	
MI051_1832783947533451	MI051	EV	3	2019-10-01	04:39:49	-75.7442	4.79908	0	101	0	28229770	
MI051_1832792268242895	MI051	EV	3	2019-10-01	04:39:58	-75.7442	4.79908	0	101	0	28229770	
MI051_1832802108533929	MI051	EV	3	2019-10-01	04:40:08	-75.7442	4.79908	0	101	0	28229770	
MI051_1832806227464746	MI051	EV	8	2019-10-01	04:40:11	-75.7442	4.79908	0	101			
MI051_1832811568000615	MI051	EV	3	2019-10-01	04:40:17	-75.7442	4.79908	0	101	0	28229770	
MI051_1832821107685162	MI051	EV	3	2019-10-01	04:40:27	-75.7442	4.79908	0	101	0	28229770	
MI051_1832829647911816	MI051	EV	3	2019-10-01	04:40:35	-75.7442	4.79908	0	101	0	28229770	

Figura 3.6: Estructura de archivo de excel extraído del módulo de consulta para el día 2019-10-01 de los vehículos MI051 al MI060

Una de las primeras ventajas es que se automatiza gran parte del proceso y que cualquier persona con conocimientos básicos de copiar y pegar la consulta en el módulo de modelado puede generar los archivos para analizar un mes, es más, este proceso se puede hacer diariamente de tal forma que al final del mes sólo sea aplicar el archivo del Apéndice **??** y tener unificado todos los datos. Este proceso que se describe acá es más conocido como *Intermediate File Approach*

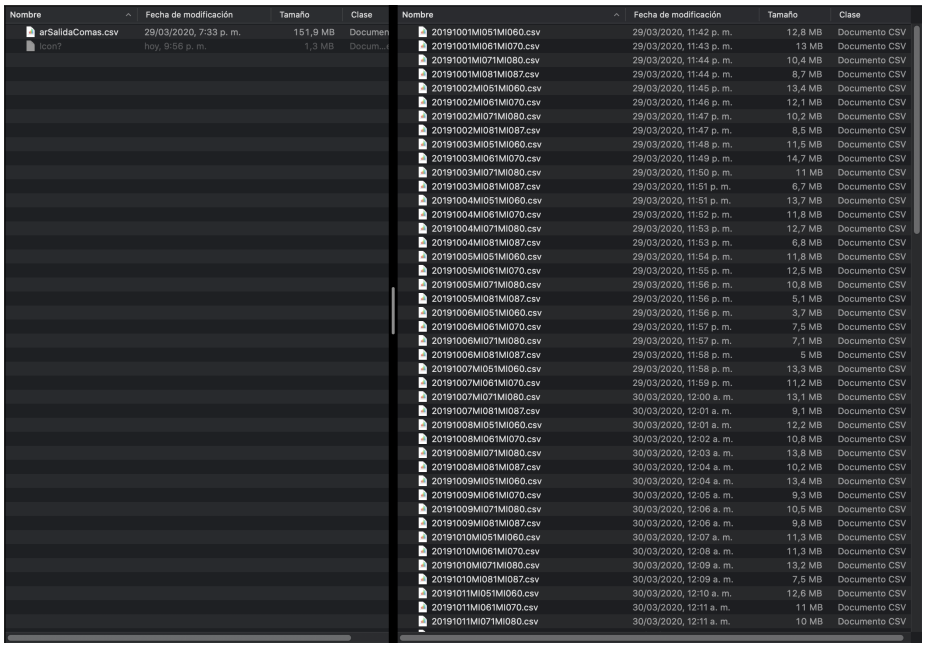

Figura 3.7: Comparación de generación de archivo unificado contra individualizado con el script.

que no es más que el proceso por el cuál se extrae información de una base de datos en un archivo ya sea *.csv*, *.xml* u otro formato para ser utilizado en otro sistema o sistemas, tal cual como lo muestra la Figura 3.8º.

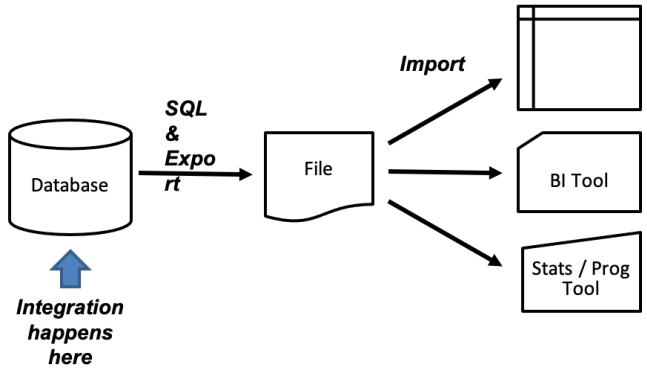

Figura 3.8: Representación de *Intermediate File Approach*

Aunque el proceso *Intermediate File Approach* es un proceso válido y muy us-

ado en muchas empresas, como la carga de archivo planos tiene una debilidad y es que al realizarse un proceso manual entre el proceso de exportar la información de la base de datos y el proceso de carga en los nuevos sistemas puede presentarse alteración o manipulación de los datos, adicional a esto generalmente el segundo paso que es la carga en los nuevos archivos es algo rígido ya que cada parte del documento o archivo resultante de exportar la información en la base de datos tiene su asignación dentro de los nuevos sistemas y al incluirse nuevos campos en la base de datos esto no permitirá que se tome la información de manera adecuada. Otra situación que se puede presentar con este proceso es que el operario o la persona encargada de realizar el proceso no lo realice por olvido o desconocimiento y esto retrase el proceso de la toma de datos.

En la Tabla 3.6 se presenta información relacionada con las variables de estudio, aún así es necesario realizar cruces de información con otras tablas del sistema o con archivos internos del sistema.

Para los demás datos se procede a realizar otras consultas dentro del módulo de InnoBUS Masivo, donde se relaciona el inicio y fin de tabla con el fin de a cada dato recibido de la red *GPS* asignarle un conductor, ya que como se vio anteriormente la estructura actual de la red *GPS* no contiene un dato donde se relacione directamente a los datos recibidos cada 3 segundos el operador del vehículo por lo que se debe proceder a realizar esa combinación para eso se usa la siguiente consulta:

SELECT cfecha, cequid, cconductor, cservicioid, coPERAID, shoraini, shorafin, srutaid, initcap(lower(pPERANOMBRE1 o pPERANOMBRE2 o pPERAAPELLIDO1 o pPERAAPELLIDO2)) AS nombre, pPERACEDULA, pPERAFNACIMIENTO, pPERASEXO, pPERACIVID, eEQUCODIGO FROM OPEconduce AS c INNER JOIN OPEservicio AS s ON (sid = cservicioid) INNER JOIN OPERUTNOM AS r ON (rID = srutaid) INNER JOIN MANEQU AS e ON (eequid = cequid) INNER JOIN OPECOND AS co ON (coCONDCODIG = cconductor) INNER JOIN THUPERA AS p ON (pID = coPERAID) WHERE (cfecha between '2019-10-01' and '2019-10-31') and (eEQUCODIGO between 'MI051' and 'MI087') ORDER BY shoraini ASC

La consulta anterior nos arroja una tabla cómo se muestra en la figura 3.9 se obtendrían los datos restantes de las variables escogidas anteriormente quedando pendiente algunas variables que se pueden adicionar posteriormente. Esta consulta relaciona un operador con un servicio, es decir tiene un inicio y un fin del servicio, pero un operador tiene varios servicios en el día, por lo que a esta tabla se le debe hacer un tratamiento con el fin de unificar todos los inicios y fin de servicios de un operador y depurar los servicios intermedios con el fin que solo quede el primer inicio del servicio y el último fin de servicio, esto se hace con el único objetivo de disminuir la carga computacional ya que los registros que se encuentran en la figura 3.6 suman más de 13 millones de registros al realizar un cruce con los 16522 mil registros que se tienen en la figura 3.9 nos

da un total de más de 214.786.000.000 computaciones lo cual daría un tiempo muy elevado de procesamiento.

Para esto lo primero que se hace es depurar esta tabla de registros de servicios con conductor realizando el proceso que se muestra en el Apendice A se rebaja en el caso de los articulados de 16522 a 2445 pasando el registro de cruces de 214.786.000.000 a 31.785.000.000 logrando una reducción del 85 porciento de las computaciones, igualmente sigue siendo un gran número de cruces a realizarse.

fecha	equ_id	conductor	servicio_id	PERA_ID	hora_ini	hora_fin	ruta_id	nombre	PERA_CEDULA	PERA_FNACIMIENTO	PERA_SEXO	PERA_CIV_ID	EQU_CODIGO
2019-10-01	62	0165	536015	606	08:32:00	08:48:00	51	Alberto Antonio Mejia Marulanda	9760389	1959-12-20	M	6	MI031
2019-10-01	62	0165	536016	606	08:48:00	09:04:00	51	Alberto Antonio Mejia Marulanda	9760389	1959-12-20	M	6	MI031
2019-10-01	62	0165	536017	606	09:04:00	09:20:00	51	Alberto Antonio Mejia Marulanda	9760389	1959-12-20	M	6	MI031
2019-10-01	62	0165	536018	606	09:20:00	09:36:00	51	Alberto Antonio Mejia Marulanda	9760389	1959-12-20	M	6	MI031
2019-10-01	62	0165	536019	606	09:36:00	09:52:00	51	Alberto Antonio Mejia Marulanda	9760389	1959-12-20	M	6	MI031
2019-10-01	62	0165	536020	606	09:52:00	10:08:00	51	Alberto Antonio Mejia Marulanda	9760389	1959-12-20	M	6	MI031
2019-10-01	62	0165	536021	606	10:08:00	10:24:00	51	Alberto Antonio Mejia Marulanda	9760389	1959-12-20	M	6	MI031
2019-10-01	62	0165	536022	606	10:24:00	10:40:00	51	Alberto Antonio Mejia Marulanda	9760389	1959-12-20	M	6	MI031
2019-10-01	62	0165	536023	606	10:40:00	10:56:00	51	Alberto Antonio Mejia Marulanda	9760389	1959-12-20	M	6	MI031
2019-10-01	62	0165	536024	606	10:56:00	11:12:00	51	Alberto Antonio Mejia Marulanda	9760389	1959-12-20	M	6	MI031
2019-10-01	62	0165	536025	606	11:12:00	11:28:00	51	Alberto Antonio Mejia Marulanda	9760389	1959-12-20	M	6	MI031
2019-10-01	62	0165	536026	606	11:28:00	11:44:00	51	Alberto Antonio Mejia Marulanda	9760389	1959-12-20	M	6	MI031
2019-10-01	62	0165	536027	606	11:44:00	12:00:00	51	Alberto Antonio Mejia Marulanda	9760389	1959-12-20	M	6	MI031
2019-10-01	62	0165	536028	606	12:00:00	12:16:00	51	Alberto Antonio Mejia Marulanda	9760389	1959-12-20	M	6	MI031

Figura 3.9: Representación de *Consulta de tabla de servicios con datos de conductor*

Para este proceso de cruce se realizó una primera prueba usando el lenguaje R y la interfaz de Rstudio, y usando el código que se encuentra en el Apéndice B, una de las desventajas de R es que no usa la totalidad de la disponibilidad de los recursos, por ejemplo para la ejecución del código en mención se uso un servidor IBM x3550 M5 con 64 Gb de memoria Ram, 1 procesador con 10 núcleos y 20 procesadores lógicos y el cruce de datos entre la tabla de conductores y la tabla de Red GPS se tomo aproximadamente 30 días de procesamiento ya que aunque se tenía a disposición distintos procesadores, R solo usa un procesador y su procesamiento es lineal, algo que sí se noto fue el consumo elevado de recursos de memoria RAM que se va acumulando en el tiempo y se va guardando

como basura, adicionalmente la eficiencia que tiene R para el manejo de datos muy grande no se nota en la eficiencia para ejecutar los 2 ciclos de los FOR que se tenían que realizar, esto lo que quiere decir es que cuando se deben de procesar grandes volúmenes de datos, Python tiende a ser mejor que R, pero para ser más rígidos estadísticamente, se usa R.

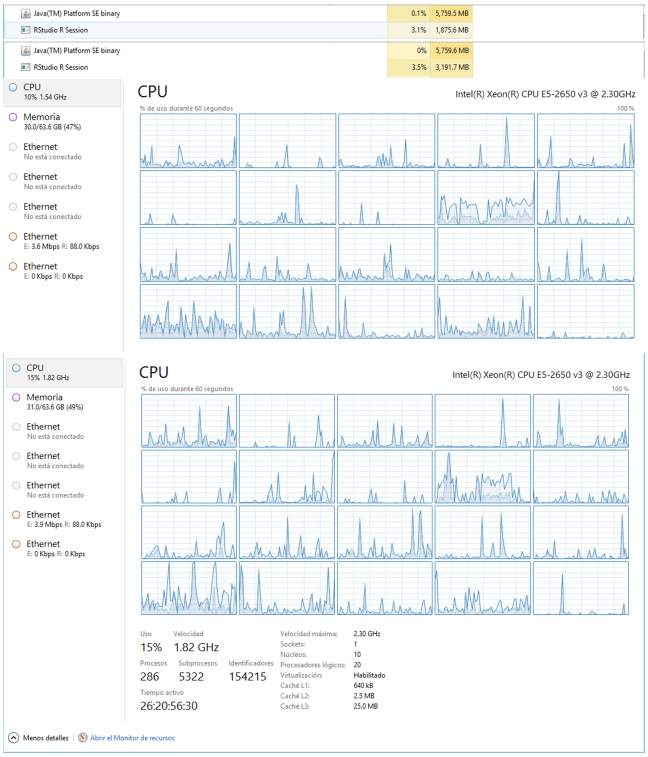

Figura 3.10: Representación de Comparación consumo de recursos en el tiempo

Como se muestra en la figura 3.10 el consumo de Memoria en R crece de forma exponencial y en menos de un minuto se pasa casi a duplicar el consumo de Ram pero con las velocidades iguales en su operacionalidad, esto se debe a que R no gestiona adecuadamente el garbage collector y no optimiza de manera adecuada la memoria RAM, por otro lado como se ven los consumos de procesadores son muy similares y no se aprovecha al máximo todos los procesadores, aunque se aclara que para el ejercicio en el código fuente no se coloco limitación alguna al uso de procesadores y el software tenía a disponibilidad todos los recursos, de igual forma se estaba corriendo en un Windows Server 2012 con permisos de Administrador lo que permitía que el mismo programa usará si así lo necesitará todos los recursos disponibles. Con todas estas claridades el proceso para relacionar las 13 millones de lineas con un Operador

tomo aproximadamente 20 días y el proceso apenas completaban el 75% de la tarea. Basado en esta realidad se procedió entonces a buscar una alternativa de solución o alternativas de solución ya que este modelo no se ajusta a la realidad empresarial, la toma de datos es de un mes más otro mes de procesamiento sería un proceso que no permitiría la bonificación empresarial adecuada o con 2 meses de retraso.

Paso seguido a esto se realizó el traslado del código en R y se migro al lenguaje de programación Python, tal cual como se muestra en el Apéndice A, y como se puede ver en 3.10 se presenta un mayor uso de procesador con la rutina corriendo en Python, los resultados en Python fueron muchos mejores se paso de 30 días a 15 días, aunque se consigue una eficiencia aproximadamente del 50% la alta computación requerida y el alto tiempo de demora para obtener la relación de un evento de la red *GPS* con un conductor hace que este no sea el proceso indicado para recolectar la data. Aunque es un gran avance y reduce la complejidad sigue siendo un proceso ineficiente.

Hay funciones y clases en Python que podrían apoyar con las funcionalidades tales como append y mejoraría el rendimiento de manera considerable pero de igual manera los resultados no serían óptimos además que como se ha recalcado se requeriría realizar el primer proceso ya mencionado acá de exportar por grupo de 10 la información de los vehículos alimentadores y articulados es decir generándose más archivos a unificar y aumentando considerablemente los datos analizar porque el dato de 100.000 datos por archivo es un aproximado por día por grupo de 10 vehículos y en total se tiene una flota de 87 vehículos por lo que se estaría hablando de 16 millones de registros aproximadamente con un cruce de relación de 2300 servicios lo que daría una iteración total 36.800.000.000 registros para relacionar y por lo que ya ha sido explicado anteriormente no es el proceso más eficiente.

Para suprimir este proceso poco eficiente se genera la tercera mejora al proceso la cual consiste en que por medio de un webservice se recolecta la información arrojada por el sistema InnoBUS, en esta versión se solicita al proveedor tecnológico que la información relacionada con el conductor, ruta y vehículo este relacionada ya en la trama de *GPS* y cuando se realice la captura no se tenga que hacer ningún procedimiento adicional, esto permite eliminar no sólo el proceso de consulta a la base de datos con su respectivo tiempo de creación sino que también elimina la rutina de relacionar el conductor con el dato del *GPS* que como se manifestó tomaba un tiempo aproximado de 15 días corriendo en Python.

Campo	Descripción	Tipo de dato
VEH.CODIGO	Código del vehículo	Carácter
codigo_conductor	Código del conductor	Carácter
LAT_GPS	Latitud entregada por el GPS.	Flotante
LON_GPS	Longitud entregada por el GPS.	Flotante
COMPASS	Número entregado por la brújula del GPS.	Entero
DISTANCE	Distancia recorrida hasta el momento en el vehículo por el conductor.	Flotante
ACE_GPS	Valor de la aceleración entregada por el GPS.	Entero
VEL_GPS	Velocidad tomada en el momento en el vehículo por el conductor.	Flotante
FECHA_GPS	Fecha en la que se toma la observación.	Fecha
VEH_ULT_REPORTE	Fecha en la que se realizó la última toma en el vehículo.	Fecha
ODO_GPS	Valor entregado por el ódometro del GPS.	Entero
ruta_id	id de la ruta que está tomando el vehículo.	Entero
PERA_NOMBRE1	Nombre del conductor.	Carácter
PERA_NOMBRE2	Nombre del conductor.	Carácter

Tabla 3.7: Campos de la Base de Datos construida a través de las diferentes consultas realizadas al servidor de InnoBUS

3.2 Algoritmo de Clusterización

3.2.1 Selección del Algoritmo

Antes de iniciar el modelamiento de los datos, se debe seleccionar el método de clusterización a aplicar, basados en la información previamente recolectada. Como en Integra S.A. los grupos o perfiles de conducción no se encuentran previamente definidos, se necesita utilizar un algoritmo de clusterización que nos permita encontrar estos grupos que nunca han sido caracterizados.

Revisando la información recolectada, se puede evidenciar que la mayoría es de clase numérica, lo que facilita la implementación de cualquier algoritmo de clusterización. Es así como se decide trabajar con el *algoritmo K-means* para identificar los perfiles de conducción dentro de Integra S.A. Este algoritmo se caracteriza por utilizar solamente variables númericas, por esto en el modelamiento de los datos y la construcción de la nueva base de datos con la cual se trabajará, se necesita que las variables sean todas númericas y que cada observación sea por cada conductor por cada vehículo manejado.

3.2.2 Modelamiento de los Datos

Primero se analizan los datos obtenidos por las consultas realizadas anteriormente, presentándose en la Tabla 3.7, donde la primer columna nos indica el nombre del campo, la segunda la descripción del campo y la tercera el tipo de dato obtenido.

Conociendo cada campo y su descripción se procede a realizar el modelamiento de los datos. Se necesita construir una nueva base de datos de manera ordenada (*TIDY DATA*), ya que la que construimos no lo está. Para poder organizar la información se deben cumplir las siguientes tres premisas [Wickham,

2014]:

- Cada variable debe formar una columna.
- Cada observación debe formar una fila.
- Cada unidad observacional tiene su tabla.

Para poder organizar la información y transformarla en *Tidy Data* se debe definir cuál será la observación en nuestro caso de estudio y las variables a tener en cuenta, así podremos construir la base de datos con la cual realizaremos el proceso de agrupamiento.

La base de datos que se construye contiene 5 variables, cada una representada en una columna y cada fila representa la observación de cada persona por cada vehículo. Las variables que se decidieron construir son las siguientes:

- **Aceleración Promedio:** Esta variable indica el valor de la aceleración promedio de cada conductor en cada vehículo que haya conducido durante el periodo de observación.
- **Velocidad Promedio:** Esta variable indica el valor de la velocidad promedio de cada conductor en cada vehículo que haya conducido durante el periodo de observación.
- **Número de Violaciones:** Esta variable indica el número de veces que cada conductor ha violado el límite de velocidad en cada vehículo durante el periodo de observación.
- **Tomas por conductor por vehículo:** Esta variable indica el número de tomas en cada vehículo durante el periodo de observación.
- **Distancia Recorrida:** Esta variable indica la distancia recorrida en cada vehículo durante el periodo de observación.

Adicional cada fila tiene su nombre que hace referencia al conductor y al vehículo al que pertenece la información. Por motivos de privacidad en el tratamiento de datos, se utiliza el código que tiene el conductor dentro de la empresa y el código del vehículo para construir esta referencia.

En la Tabla 3.8 se puede observar las primeras 5 observaciones y cómo queda construida la Base de datos.

Código	Aceleración Promedio	Velocidad Promedio	Número de Violaciones	Tomas Conductor Vehículo	Distancia Recorrida
100MI001	-3.0	31,261281	6	972	8.830.620.037
100MI002	-1.0	17,607057	0	595	2.192.574.966

100MI017	0.0	19,119311	1	938	1.438.141.624
100MI019	-2.0	20,404584	8	2534	46.721.132.488
100MI023	1.0	29,721341	0	481	33.281.832

Tabla 3.8: Base de datos construida a partir del procesamiento de los datos obtenidos en las consultas realizadas.

Análisis de los Datos Obtenidos

La Tabla 3.8 es la base de datos con la cual se trabajará el algoritmo de *K-means* para identificar los perfiles de conducción dentro de la empresa Integra S.A.

Lo primero que hacemos es escalar y normalizar los datos, esto con el fin de al momento de realizar el análisis de las variables, este se pueda realizar en las mismas unidades.

Luego, se realiza un diagrama de caja, el cual nos permite visualizar los datos de una manera globalizada y separada por variables. El diagrama lo podemos ver en la Figura 3.11. Donde se visualizan varios datos atípicos por variable, así que se debe de tratar los datos para eliminar estos datos atípicos, los cuales pueden producir una lectura errónea del comportamiento de los conductores, ya que el algoritmo *k-means*, el cual es el que se va a utilizar para realizar el agrupamiento de los individuos, es altamente susceptible a los datos atípicos, por esta razón se deben eliminar.

Para eliminar estas observaciones atípicas, se realizan dos metodologías de limpieza de los datos, la primera de ellas es utilizando la teoría que entre más o menos tres desviaciones estándar se encuentra el 97% de los datos según el teorema de Chebyshev. Luego procedemos a realizar otra vez el diagrama de cajas y bigotes de los datos, donde se presenta en la Figura 3.12.

Se realiza la metodología de limpieza interquartil, donde solo se toman los datos que se encuentran entre el quartil 1 y el 3.

3.2.3 Definición del Número de *Clusters*

Una vez realizadas los dos tipos de limpieza de datos atípicos, se procede a calcular el número *a priori* de grupos dentro del conjunto de datos, esto se realiza ya que para la metodología *k-means*, se debe de especificar previamente el número de conglomerados a identificar. Por esta razón existen diferentes indicadores, de tipo numérico y gráfico. En la Tabla 3.9 se presentan los indicados numéricos utilizados.

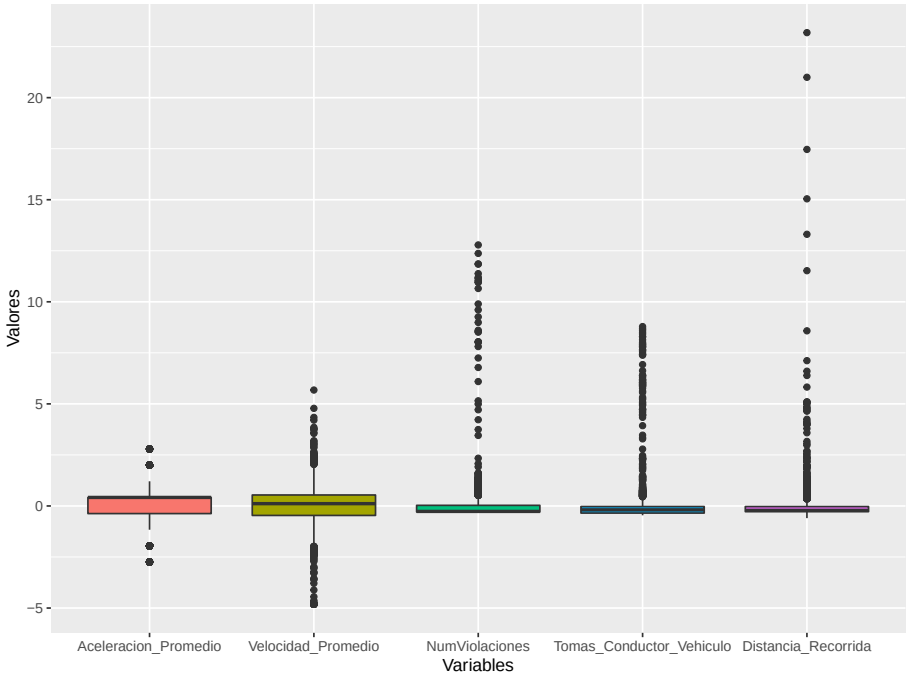

Figura 3.11: Diagrama de caja de las variables antes de realizar el proceso de limpieza de los datos

Definición de *Cluster* Metodología Interquartil

Primero se trabajará con los datos procesados por la metodología de limpieza interquartil.

En la Figura 3.14 se muestra el indicador gráfico Diagrama Silhoutte, el cual nos indica que el número de *clusters* para el conjunto de datos que estamos analizando es de dos (2) *clusters*.

En la Figura 3.15 se muestra el indicador gráfico Diagrama WSS, el cual nos indica que el número de *clusters* para el conjunto de datos que estamos analizando es de tres (3) *clusters*.

Por último tenemos la Figura 3.16 que nos permite visualizar los valores obtenidos en los 24 indicadores numéricos descritos en la Tabla 3.9, donde vemos que la mayoría de los indicadores nos aconsejan a trabajar con tres (3) conglomerados.

Luego de haber realizado todos estos procedimientos, se decide trabajar con 2 y con 3 *cluster*, para luego ser presentados a los expertos en Integra S.A y

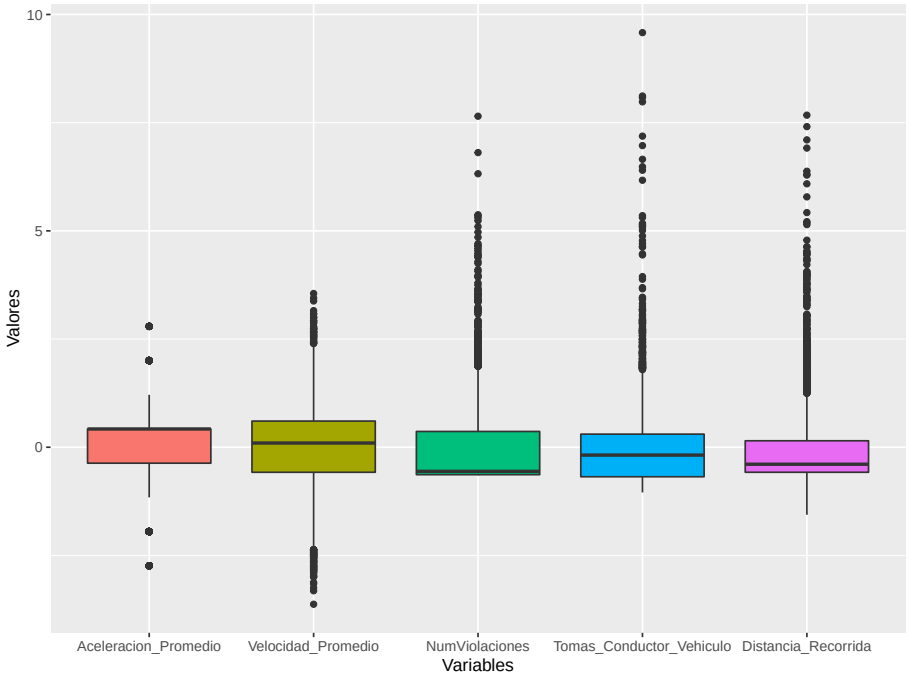

Figura 3.12: Diagrama de caja de las variables después de realizar el proceso de limpieza de los datos con la metodología de las desviaciones entandar

con estos tomar una decisión sobre cuál es el que más representa los perfiles de conducción dentro de la empresa.

Luego se trabaja con los datos que han sido procesados mediante la limpieza del teorema de Chebyshev.
En la Figura 3.17 se muestra el indicador gráfico Diagrama Silhoutte, el cual nos indica que el número de *clusters* para el conjunto de datos que estamos analizando es de cuatro (4) *clusters*.
En la Figura 3.18 se muestra el indicador gráfico Diagrama WSS, el cual nos indica que el número de *clusters* para el conjunto de datos que estamos analizando es de cuatro (4) *clusters*, el mismo número que en la Figura 3.17.
Por último tenemos la Figura 3.19 que nos permite visualizar los valores obtenidos en los 24 indicadores numéricos descritos en la Tabla 3.9, donde vemos que la mayoría de los indicadores nos aconsejan a trabajar con tres (3) conglomerados.

Luego de haber realizado todos estos procedimientos, se decide trabajar con tres (3) y (4) *cluster*, para luego ser presentados a los expertos en Integra S.A y con estos tomar una decisión sobre cuál es el que más representa los perfiles de

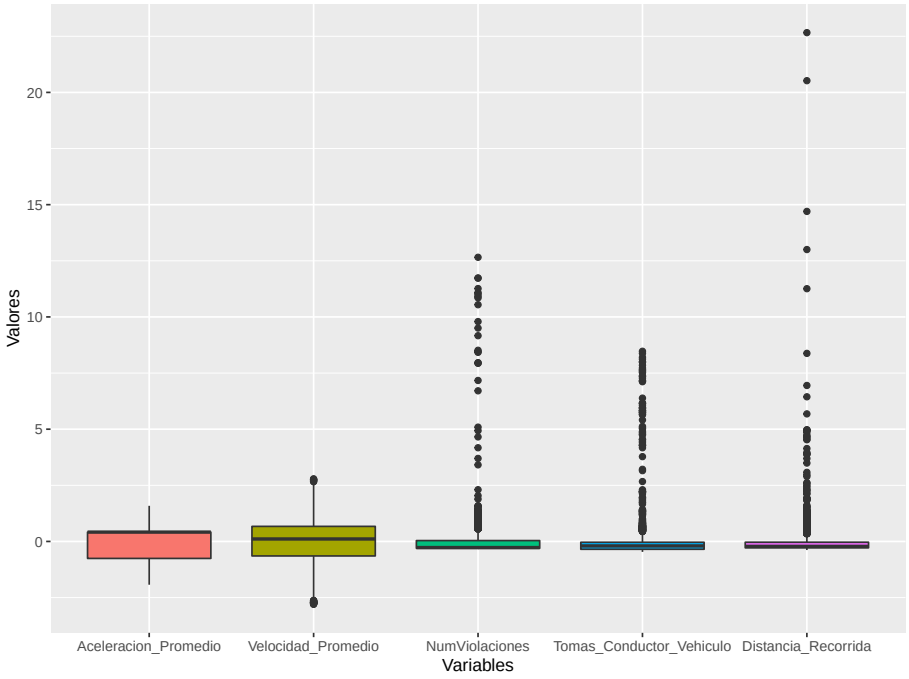

Figura 3.13: Diagrama de caja de las variables despues de realizar el proceso de limpieza de los datos con la metodologia de interquartil

conducción dentro de la empresa.

Índice	Descripción
Krzanowski-Lai [Krzanowski and Lai, 1988]	Usa una heurística para determinar el número de grupos de un dataset con la función objetivo es el mejor valor del determinante de la matriz de covarianza agrupada para cualquier grupo dado de la muestra.
Calinski-Harabasz [Caliński and Harabasz, 1974]	Se construye un gráfico de dendrita más corta usando tanto los métodos divisivos como aglomerativos, construido bajo un algoritmo de vecino más cercano y luego aplicando el criterio divisivo de la mínima suma de cuadrados entre *cluster*.
Hartigan [Hartigan, 1975]	Es un índice basado en la suma de cuadrados, calculado como el logaritmo negativo de la división de la suma de cuadrados intra *clusters* (SSW) por la suma de cuadrado entre *clusters* (SSB).
Criterio Cúbico de Clusterización [Sarle, 1983]	Índice basado en la suma de cuadrado de los errores y sus productos cuzados.
Scott-Symons [Scott and Symons, 1971]	Índice basado en los métodos de máxima verosimilitud de clasificación.
Marriot [Marriott, 1971]	Índice que utiliza el determinante de la matriz de covarianza para cada uno de los posibles grupos.
Índice de la traza de covarianza W [Milligan and Cooper, 1985]	Este índice representa la traza de la matriz de covarianza agrupada dentro de los conglomerados.
Índice de la traza de W [Milligan and Cooper, 1985]	Éste índice es muy utilizado en el contexto de la determinación de conglomerados, donde nota que la suma de cuadrados del error intraclusters es igual a la traza de la matriz del total de los suma de cuadrados de los productos cruzados W.
Índice Friedman[Friedman and Rubin, 1967]	Éste índice fue propuesto como base para el método no jerárquico de agrupamiento.
Índice de Rubin [Friedman and Rubin, 1967]	Este índice se basa en la razón entre determinante de la matriz de la suma total de cuadrados y productos cruzados, y el determinante de la matriz agrupada dentro de los conglomerados.
Hubert & Levin [Hubert and Levin, 1977]	Este índice utiliza las disimilitudes para encontrar el número de *cluster*.
Índice de Davies & Bouldin [Davies and Bouldin, 1979]	Este índice es una función de la relación de la suma de dispersión intra-*cluster* a la separación entre *clusters*.
Índice Silhouette [Kaufman and Rousseeuw, 2009]	Este índice indica qué tan parecido es un objeto con su propio conglomerado, así asignándolo a un grupo y entregando el número de *clusters* que se deberían tomar.
Índice Duda [Duda et al., 1973]	Índice que utiliza la suma de cuadrado del error intra *cluster*, donde el criterio de selección del mejor número de *cluster* es dónde el índice sea mayor al valor crítico establecido.
Índice *Pseudot2* [Duda et al., 1973]	Índice que utiliza la suma de cuadrador del error intra *cluster*, donde el criterio de selección del mejor número de *cluster* es dónde el índice sea menor al valor crítico establecido.
Índice Beale [Beale, 1969]	Índice que utiliza el contraste basado en la distribución F de Snedecor, usando la suma de cuadrados las desviaciones cuadráticas medias.
Índice Ratkowsky [Ratkowsky and Lance, 1978]	Índice que se basa en el promedio de la relación de la suma de cuadrados entre *cluster* y las suma total de cuadrados para cada una de las variables.
Índice Ball [Ball and Hall, 1965]	Índice basado en el promedio de la distancia de cada individuo con sus respectivos centroides.
Índice *Point-Biserial* ([Milligan, 1981],[Milligan, 1980])	Este índice es la medida de correlación entre una variable continua A y una variable binaria B.
Índice GAP [Tibshirani et al., 2001]	Índice que utiliza la matriz de dispersión definida en [Hartigan, 1975].
ïndice Frey [Frey and Van Groenewoud, 1972]	Índice basado en la relación de la diferencia entre dos níveles jerárquicos.
Índice McClain-Rao [McClain and Rao, 1975]	Índice basado en la relación entre el promedio de la distancia intra *cluster* y el número de distancias intra *cluster*.
Índice Dunn [Dunn, 1973]	Índice basado en la relación entre la distancia mínima intracluster y la distancia máxima entre *cluster*.
Índice SD [Halkidi et al., 2000]	Índice basado en el concepto de dispersión media de los *clusters* y la separación total entre *clusters*.

Tabla 3.9: Indicadores numéricos para determinar el número de *clusters*.

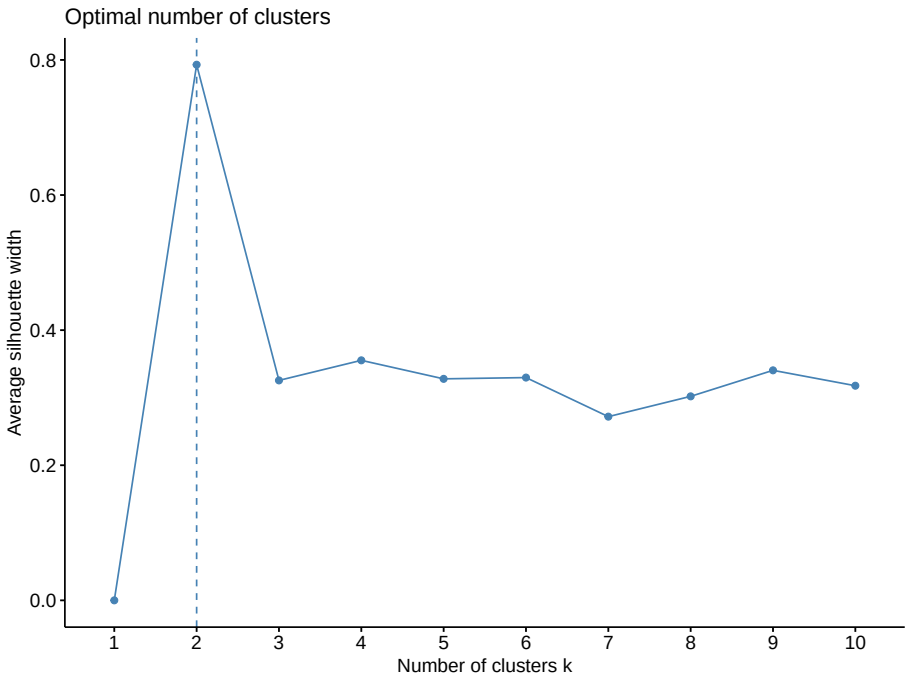

Figura 3.14: Diagrama *silhoutte* para determinar el número de *cluster* datos interquartil.

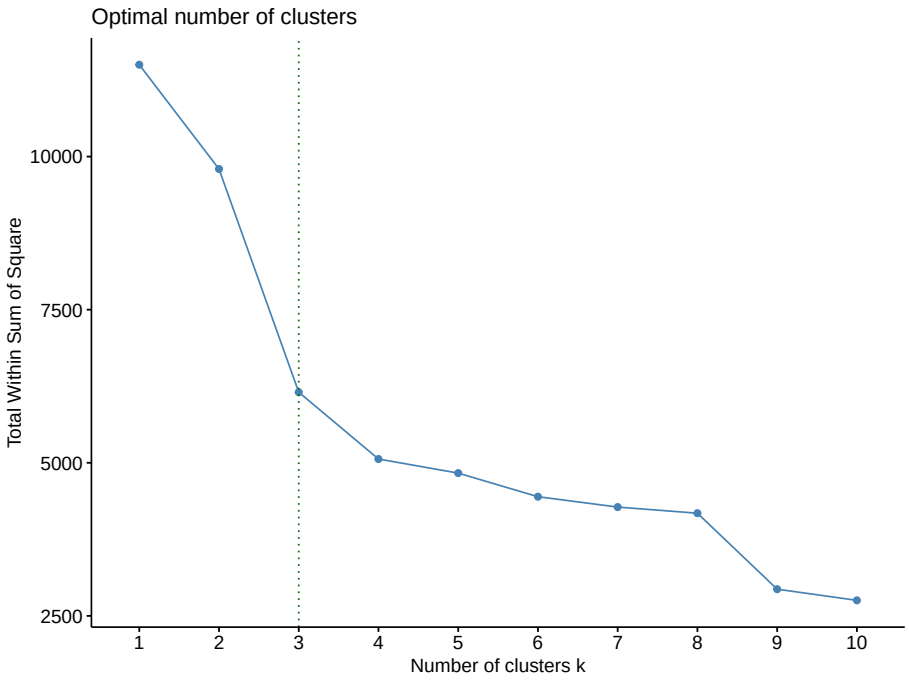

Figura 3.15: Diagrama *wss* para determinar el número de *cluster* datos interquartil.

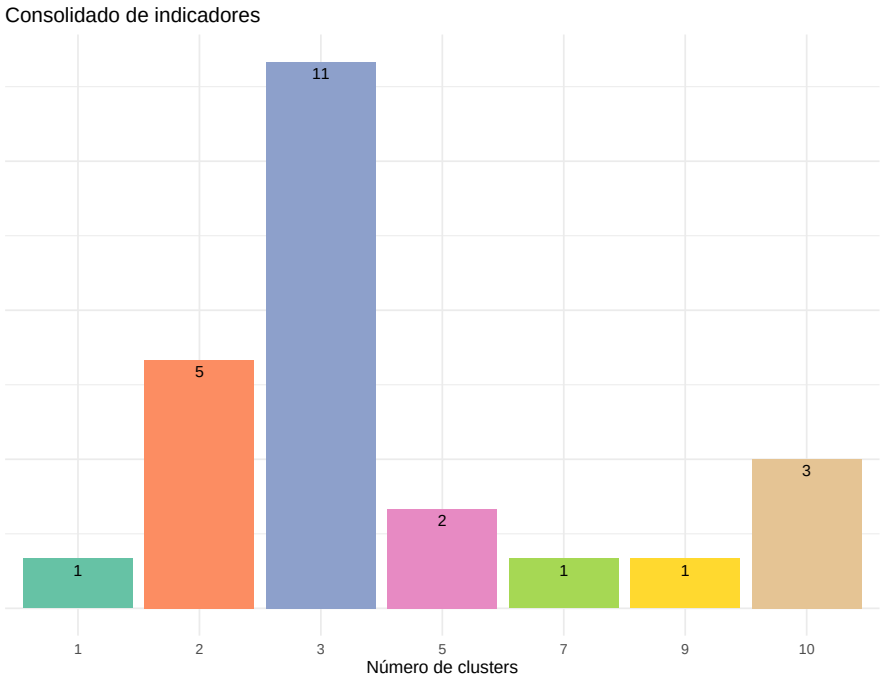

Figura 3.16: Consolidado indicadores numero de *cluster* para determinar el número de *cluster* datos interquartil

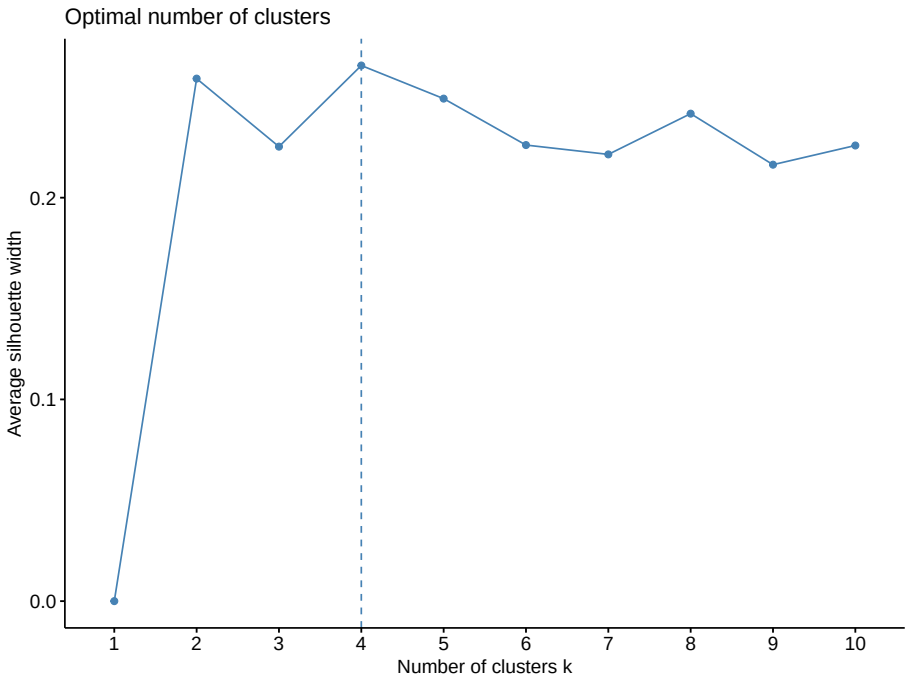

Figura 3.17: Diagrama *silhoutte* para determinar el número de *cluster* datos desviaciones.

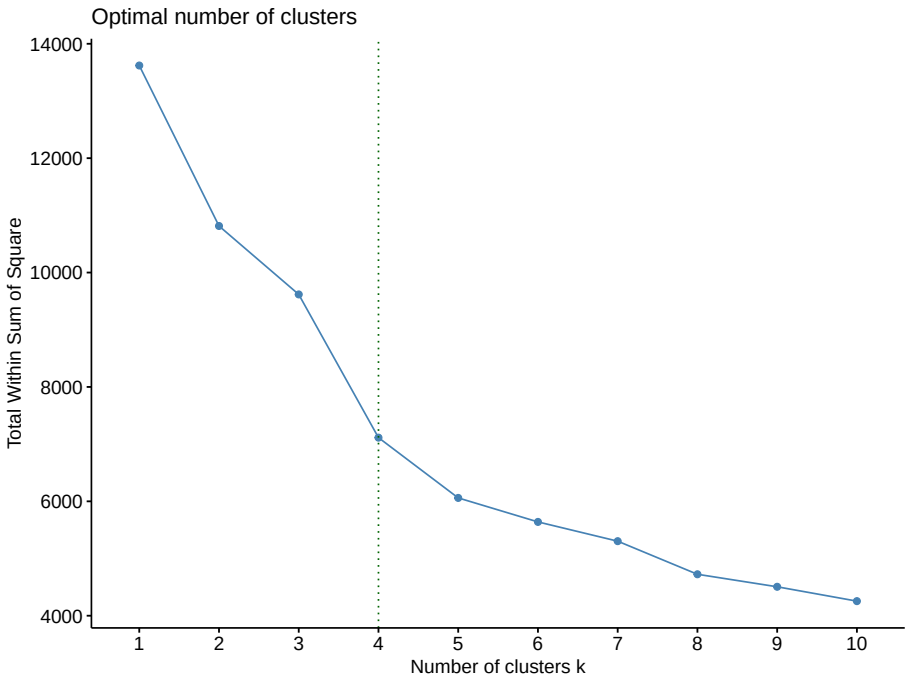

Figura 3.18: Diagrama *wss* para determinar el número de *cluster* datos desviaciones

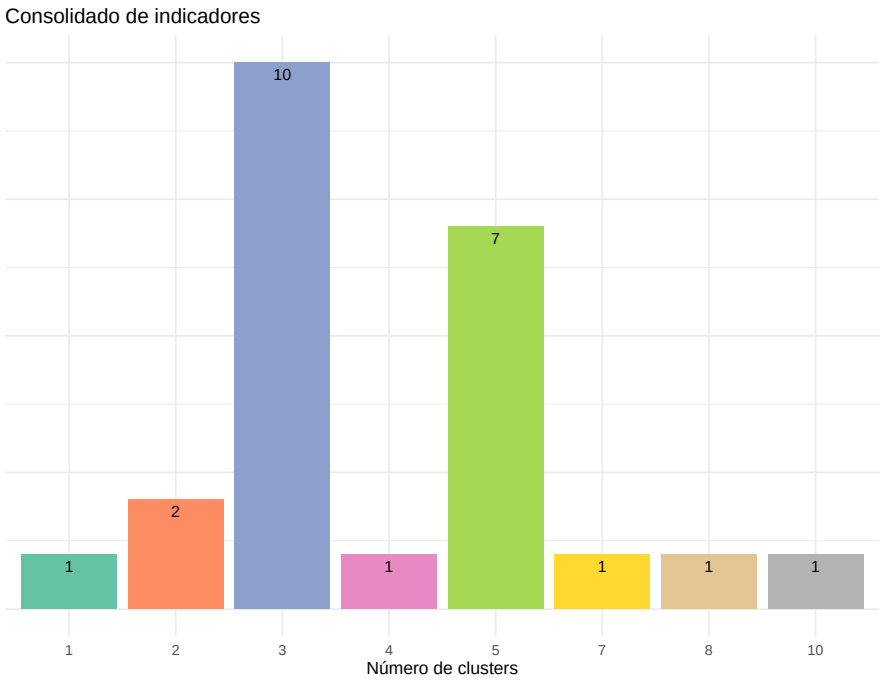

Figura 3.19: Consolidado indicadores numero de *cluster* para determinar el número de *cluster* datos desviaciones

Capítulo 4

Resultados

Una vez definidos los cuatro (4) experimentos que se van a realizar, se procede a implementar el algoritmo *K-means* con las dos bases de datos, procesada con la limpieza interquartil y la procesada con la limpieza del teorema de Chebyshev. Para realizar los experimentos se utiliza el paquete *stats* para realizar la clusterización, utilizando la función *k-means()*.

4.1 Datos Procesados con la Limpieza Interquartil

A continuación se presentan los resultados gráficos de la clusterización con dos (2) conglomerados, el gráfico se realiza utilizando la función *fviz_cluster* del paquete de R *factoextra*.

En la Figura 4.1 se evidencia los dos conglomerados separados, podemos ver que el conglomerado número 1 es mucho más grande que el conglomerado 2. Donde el conglomerado 2 se encuentra un poco más disperso que el 1.

En la Tabla 4.1 se presenta la información relacionada con la clusterización realizada. Se evidencia que efectivamente el *cluster* 1 es más grande que el *cluster* 2, teniendo el primero el 97.6 % de los datos. Igualmente podemos analizar los valores de la suma de cuadrados entre *cluster* e intra-*cluster*. La suma de cuadrados entre *cluster*, corresponde a la distancia de cada uno de los puntos con su respectivo centroide. La suma de cuadrados entre *cluster* se calcula, primero se halla la distancia media de los individuos y luego se calcula al distancia de cada individuo con la distancia media total, como si cada individuo fuese su respectivo centroide.

En la Figura 4.2 se presenta el la clusterización con 3 conglomerados, donde podemos ver que en este caso el *Cluster* 1 es el mismo *Cluster* 2 de la Figura 4.1, y el *cluster* 1 se divide en dos nuevos conglomerados.

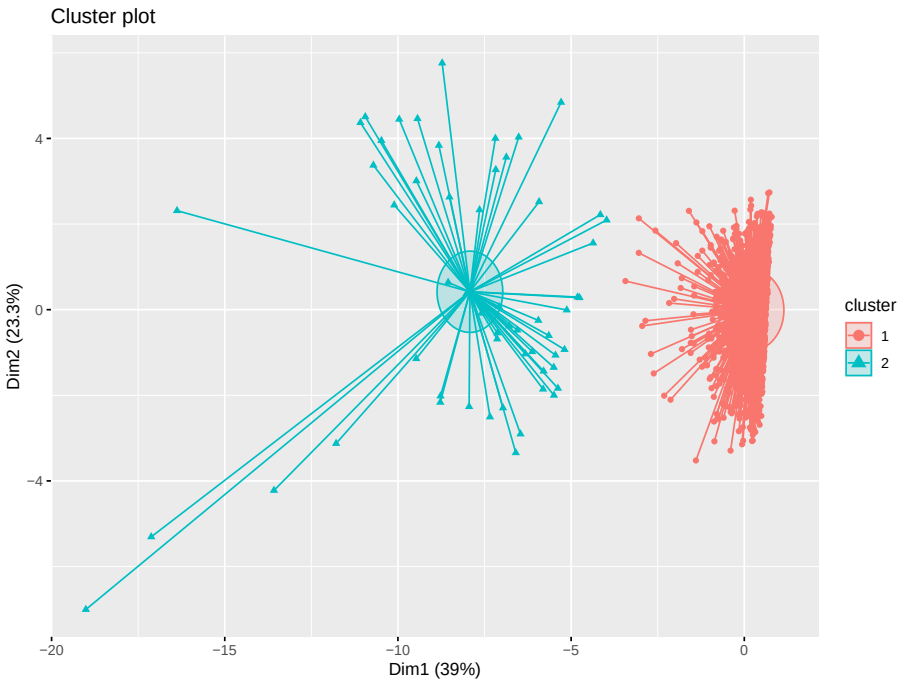

Figura 4.1: *Cluster plot* metodología interquartil 2 *cluster*

	Cluster 1	**Cluster 2**
Tamaño	2246	55
Suma de cuadrados entre *cluster*	5260	2645
Total suma de cuadrados intra *cluster*	7905	
Porcentaje de participación	68,74%	
Suma de cuadrados entre *cluster*	3595	
Porcentaje de participación	31,26%	
Suma de cuadrados total	11500	

Tabla 4.1: Suma de cuadrados 2 *Cluster* técnica interquartil

En la Tabla 4.2 se evidencia ahora que el *Cluster* 2 es el más grande, seguido del *Cluster* 3 y por último el *Cluster* 1. Adicional a esto, se evidencia que la suma de cuadrados total es la misma que la de la Tabla 4.1, pero la suma de cuadrados intra-*cluster* es menor, es decir que en este caso los individuos están más cerca a sus respectivos centroides.

Al final se decide presentar a Integra S.A los resultados con la clusterización de 3 conglomerados, es decir, 3 perfiles de conducción dentro de la empresa.

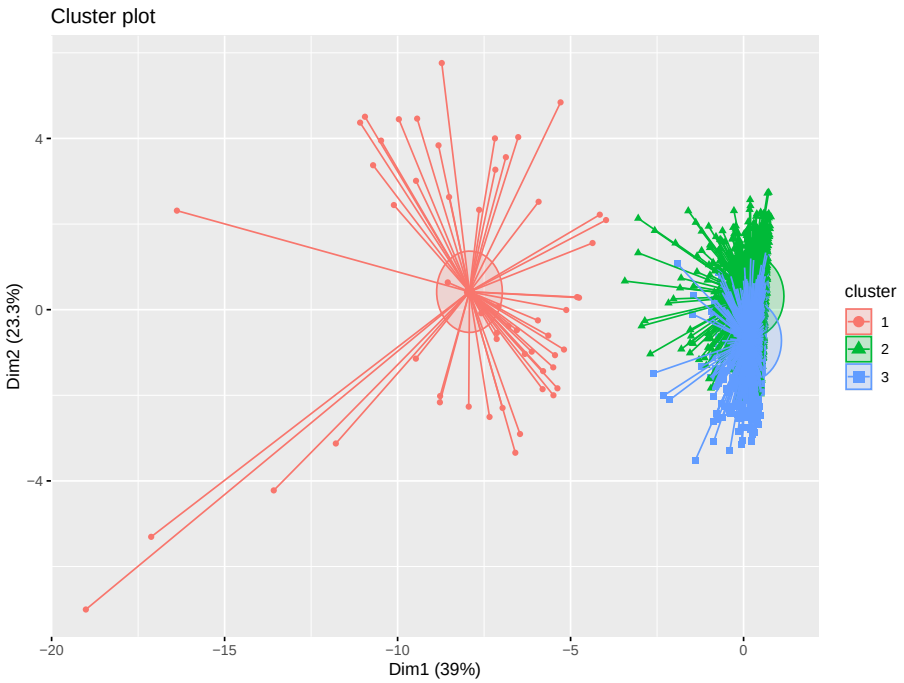

Figura 4.2: *Cluser plot* metodoloogia interquartil 3 *cluster*

	Cluster 1	**Cluster 2**	**Cluster 3**
Tamaño	55	1541	705
Suma de cuadrados entre cluster	2645	2355	1152
Total suma de cuadrados intra cluster	6152		
Porcentaje de participación	53,50%		
Suma de cuadrados entre cluster	5348		
Porcentaje de participación	46,50%		
Suma de cuadrados total	11500		

Tabla 4.2: Suma de cuadrados Clusterización de 3 conglomerados metodología interquartil

4.2 Datos Procesados con la Limpieza a través del Teorema de Chebyshev

En la Figura 4.3 se evidencia una clusterización parecida con los resultados obtenidos en la Figura 4.2, solo que en este caso cabe recordar que se encuen-

tran muchos más individuos presentados. Aquí podemos ver que el *Cluster* 3 se encuentra demasiado separado de los otros dos *cluster*, es decir que podemos ver que este grupo en los 3 resultados anteriores se mantiene. Los *cluster* 1 y 2 son parecidos a los evidenciados en la Figura 4.2.

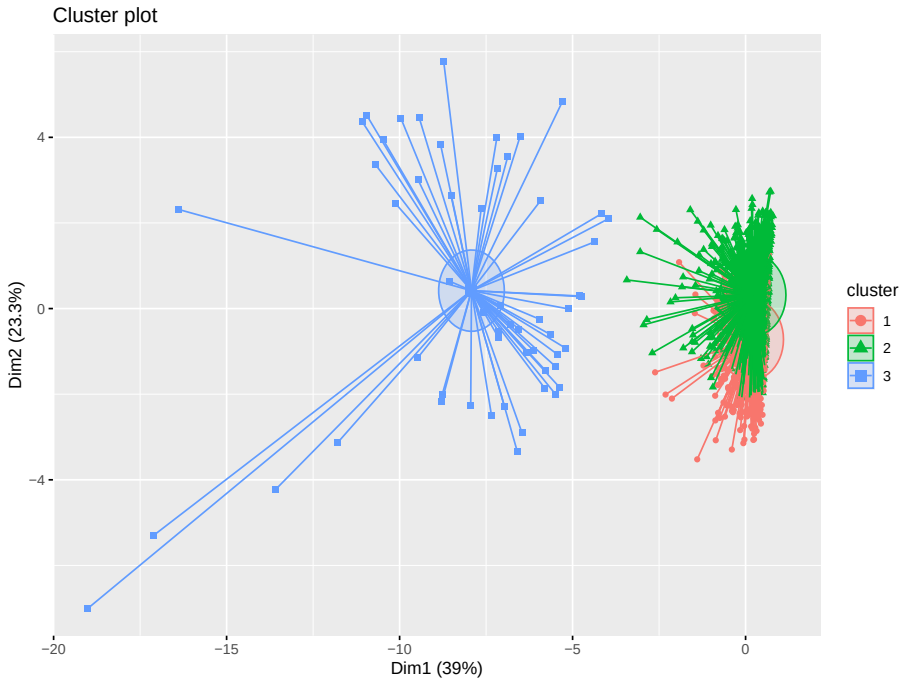

Figura 4.3: *Cluser plot* metodología desviaciones 3 *cluster*

En la Tabla4.3 se presenta la suma de cuadrados de los 3 conglomerados identificados con la metodología *k-means*. Aquí podemos evidenciar que el *Cluster* 1 y el *Cluster* 2 contienen casi la misma cantidad de individuos, mientras que el *cluster* 3 es mucho más pequeño que estos dos anteriores. Además, en este caso al ser más individuos la suma de cuadrados total es mayor a la de los dos anteriores experimentos, pero en este caso la suma intra *cluster* es mucho mayor en participación de la suma total de cuadrados de las anteriores.

En la Figura 4.4 se presentan los resultados de la clusterización con 4 conglomerados. Allí se observa una división de los conglomerados mucho más clara que la presentada en la Figura 4.3, se puede observar una divisón marcada en los individuos a analizar.

En la Tabla 4.4 se presenta la suma de cuadrados de los 4 conglomerados. Podemos observar que en este caso el *Cluster* más grande es el 1, con el 41,76 % de los individuos, el que sigue es el *Cluster* 4 con el 36,99 %, donde podemos observar que contienen la mayoría de los individuos. Ahora el *cluster* más pequeño es el 3, que es parecido al *Cluster* 3 de la Figura 4.3. También se puede

	Cluster 1	Cluster 2	Cluster 3
Tamaño	1224	1266	235
Suma de cuadrados entre *cluster*	3828	3090	1656
Total suma de cuadrados intra *cluster*	8574		
Porcentaje de participación	62,95%		
Suma de cuadrados entre *cluster*	5046		
Porcentaje de participación	37,05%		
Suma de cuadrados total	13620		

Tabla 4.3: Suma de cuadrados de 3 conglomerados utilizando la metodología de las desviaciones estándar.

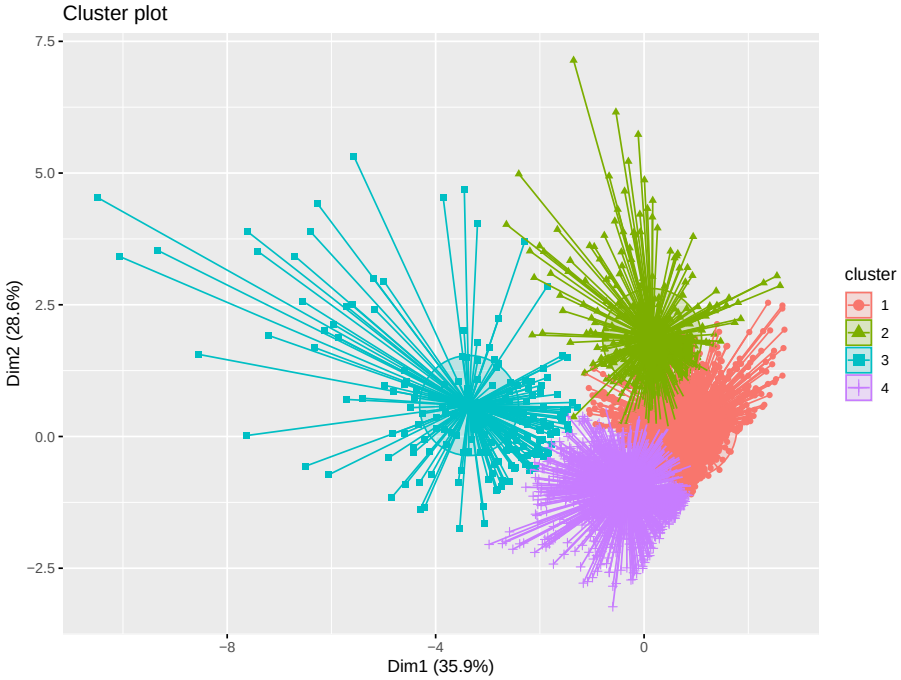

Figura 4.4: *Cluser plot* metodología desviaciones 4 *cluster*

observar que la suma de cuadrados es la misma que la presentada en el anterior experimento, pero en este caso la suma de cuadrados intra *cluster* tiene menor participación sobre la suma de cuadrados total.

Esta clusterización es la que se presenta a Integra S.A, para elegir entre esta y la presentada en la Figura 4.2. Es allí donde a cada una de estas clusterizaciones se les analizará sus respectivos centroides y se decidirá cuál se ajusta más a lo ocurrido dentro de la empresa Integra S.A.

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
Tamaño	1138	379	200	1008
Suma de cuadrados entre *cluster*	2039	1164	1339	2572
Total suma de cuadrados intra *cluster*	7114			
Porcentaje de particpación	52,23%			
Suma de cuadrados entre *cluster*	6506			
Porcentaje de particpación	47,77%			
Suma de cuadrados total	13620			

Tabla 4.4: Suma de cuadrados 4 conglomerados metodología de las desviaciones estándar.

4.3 Análisis de las Clusterizaciones Escogidas

Una vez escogidas las dos clusterizaciones que se van a presentar a la empresa Integra S.A., se debe de realizar en análisis a sus centroides, es decir, cada conglomerado cómo se define a partir de las variables de estudio. Es así como se presentan los valores en tablas de cada uno de los centrodies, pero se realiza una gráfica con cada uno superpuesto para así poder realizar un mejor análisis.

4.3.1 Clusterización con 3 Perfiles

En la Tabla 4.5 se presentan los valores para cada variable de los 3 centroides encontrados. Los valores se presentan en unidades escaladas y normalizadas, ya que así es más fácil poder interpretarlos.

Cluster	Aceleración Promedio	Velocidad Promedio	Número de Violaciones	Tomas por Conductor por Vehículo	Distancia Recorrida
1	0.59	0.07	-0.09	-0.14	-0.11
2	-0.01	-0.19	3.80	5.85	3.89
3	-1.30	-0.14	-0.10	-0.15	-0.07

Tabla 4.5: Valores de los 3 centroides

Para entender mejor la Tabla 4.5, se grafican los valores de cada una de las variables unidas, además de graficas los tres centroides en la misma. En la Figura 4.5 se puede observar que los conglomerados 1 y 2 son muy parecidos pero el conglomerado 3 es completamente diferente a estos dos anteriores.

Se identifica que el conglomerado 3 son los conductores que tienen mucho más turnos de trabajo dentro de la empresa, ya que tienen muchísimas más tomas que los otros dos conglomerados, por ende ya pasaron por periodo de capacitación y han estado mejorando su velocidad promedio y su aceleración promedio, es claro que al tener muchas más tomas, es más probable que cometan muchas más infracciones.

En los conglomerados 1 y 2, se diferencian en que, los conductores que se encuentran perfilados en el conglomerado 2, son aquellos que tienen un peor

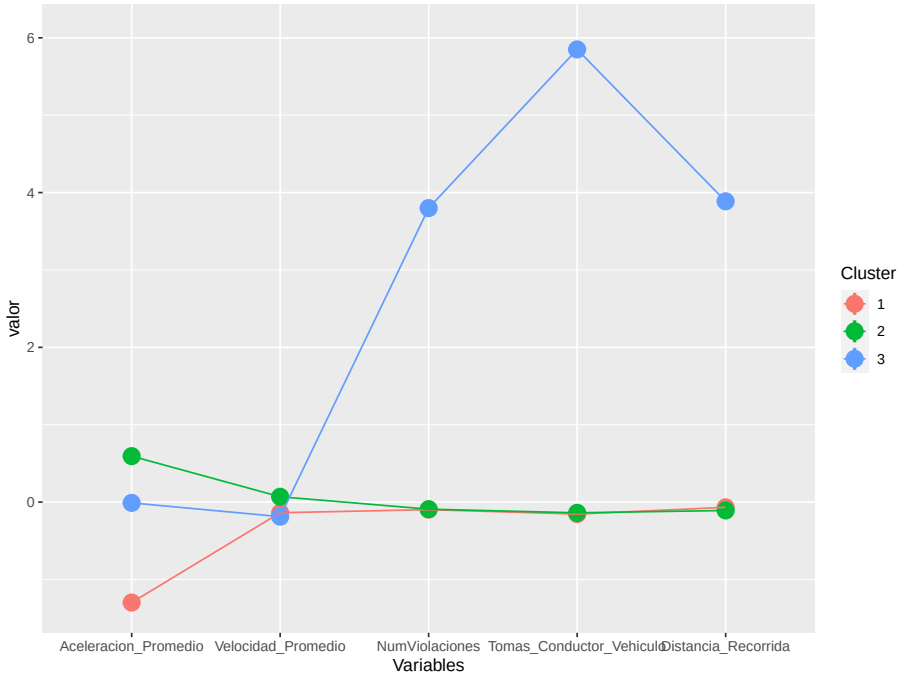

Figura 4.5: Análisis de variables de los *cluster* de 3 por medio de la metodología interquartiles

estilo de conducción, esto se puede deducir debido a que su valor de aceleración y velocidad promedio es más alta que la media de conductores de la empresa. En el caso de los del conglomerado 2, se puede identificar que tienen unos cambios más bruscos en el frenado, ya que su aceleración promedio está por debajo de la media, pudiendo catalogarse como conductores imprudentes.

4.3.2 Clusterización con 4 perfiles

En la Tabla 4.6 se presentan los valores para cada variable de los 4 centroides encontrados. Los valores se presentan en unidades escaladas y normalizadas, ya que así es más fácil poder interpretarlos.

Clúster	Aceleración Promedio	Velocidad Promedio	Número de Violaciones	Tomas por conductor por Vehículo	Distancia Recorrida
1	-0.23	-0.41	-0.45	2.22	2.65
2	0.51	0.57	-0.13	-0.40	-0.42
3	-0.55	-0.78	-0.52	-0.16	-0.02
4	0.04	0.58	2.01	0.47	-0.07

Tabla 4.6: Valores de los centroides

Gráfica de la tabla anterior.

Para entender mejor la Tabla 4.6, se grafican los puntos de cada uno de los conglmerados y sus valores en las variables de estudio, presentados en la Figura 4.6. En esta figura, se puede observar que cada uno de los conglomerados representa un perfil de conducción marcado dentro de la empresa, esto por la diferencia entre los valores de los centroides en la variable de estudio.

Se puede evidenciar como cada uno de los perfiles de conducción identificados en esta clusterización, son diferentes en las variables de estudio, así como se presenta en su gráfica de resultado presentada en la Figura 4.4.

Se presentan estas dos clusterizaciones a la empresa Integra S.A, para decidir cuál se ajusta más al entendimiento del negocio de la empresa y sus necesidades.

4.4 Elección de Clusterización

Una vez presentados las dos clusterizaciones a la empresa Integra S.A, se decide utilizar la clusterización de los datos procesados bajo el teorema de Chebyshev y los 4 perfiles de conducción. Esto ya que por un lado se quedan menos individuos por fuera del análisis, de como quedan a través del procesamiento por la ley de los cuartiles.

Adicional a la cantidad de individuos analizados, los perfiles de conducción en esta clusterización son mucho más marcados. Lo que facilita establecerlos dentro de la empresa de una manera más rígida para mantenerlos en el tiempo. Además, que se puede explicar con mejor detalle a los empleados, siendo estos los más interesados en conocer estos perfiles, debido a que de estos perfiles se establecerá después las respectivas bonificaciones.

4.5 Perfiles de Conduccion.

A continuación se describen los perfiles de conducción de la clusterización seleccionada.

4.5.1 Perfil de Conducción 1

Este perfil corresponde al color salmón de la Figura 4.6, donde se puede inferir que este perfil son los conductores que son los que menos trabajan, pero aún así tienen un número de violaciones considerado, además de ser los que tienen una conducción agresiva, esto debido al valor de su aceleración promedio, siendo la más alta de los 4 perfiles y su velocidad promedio la más alta, empatada con los conductores del perfil de conducción 2. Estos conductores serán los que se deberán capacitar sobre conducción agresiva y sobre las políticas de velocidad

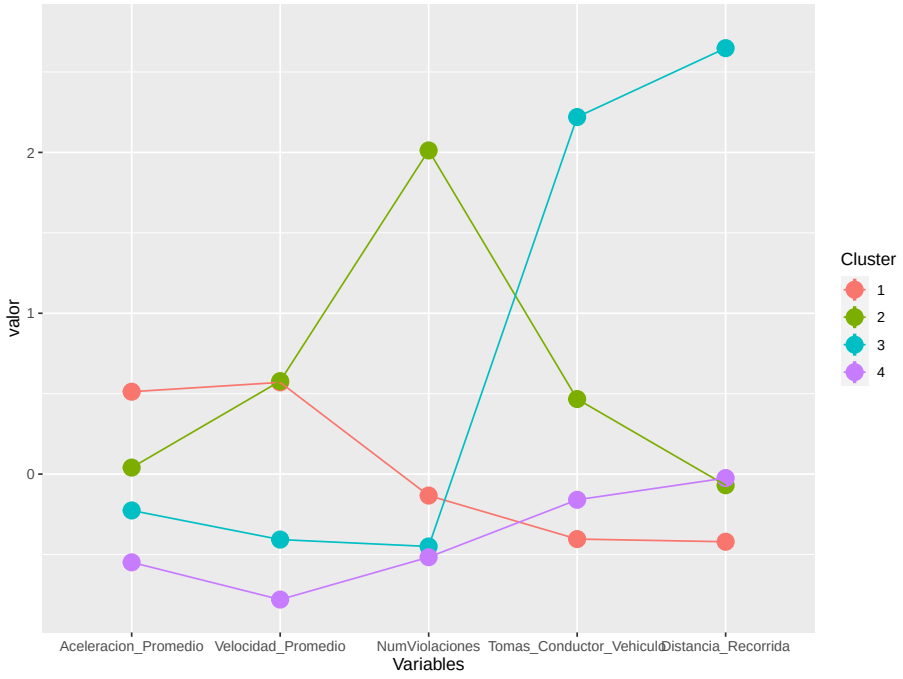

Figura 4.6: Análisis de variables de los *cluster* de 4 por medio de la metodología desviación estándar

dentro de la empresa.

4.5.2 Perfil de Conducción 2

Este perfil corresponde a los conductores que tienen peor conducción dentro de la empresa pero llevan mucho tiempo en ella o trabajan más turnos. Esto por ser los segundos con más tomas y los terceros con más distancia recorrida. Adicional son los que mayor número de violaciones de velocidad tienen y esto también se refleja en que son los que más alta tiene su velocidad promedio. Estos conductores no clasifican en conducción agresiva, debido a que su aceleración promedio no es tan alta comparada con los otros perfiles de conducción, tendiendo un poco a acelerar que a desacelerar los vehículos. Estos conductores no deberían recibir bonificación debido a su perfil de conducción.

4.5.3 Perfil de Conducción 3

Este perfil corresponde a aquellos que mejor conducen, ya que a pesar de ser los que más distancia han recorrido y más tomas tienen, el número de violaciones, su velocidad promedio y su aceleración está por debajo de la media, que es alta debido a los valores de los conductores de los perfiles 1 y 2. Este tipo de conductores, podrían ser aquellos que más tiempo lleven en la empresa, ya que su número de tomas es mucho mayor así como su distancia recorrida, teniendo mayor probabilidad de realizar más violaciones, manejar el vehículo con mayor velocidad o acelerar y frenar de manera agresiva, pero los resultados son todo lo contrario. Estos conductores deberían de recibir la totalidad de la bonificación.

4.5.4 Perfil de Conducción 4

Este perfil de conducción corresponde a aquellos conductores que a pesar de tener pocas violaciones a la velocidad y tener un poco más de tiempo conduciendo los vehículos, deben de mejorar su desaceleración, es decir, tienen un poco de conducción agresiva al momento de frenar los vehículos. Podemos observar que en velocidad promedio y número de violaciones son los que tienen los menores valores, pero en su aceleración promedio logran estar muy por debajo de la media. Este tipo de conductores pueden recibir parcialmente la bonificación, hasta que mejores su conducción agresiva.

Capítulo 5

Conclusiones y Trabajos Futuros

- Se puede identificar que en la empresa Integra S.A. sí era necesario realizar un análisis de conglomerados de sus conductores. Ya que es notoria la existencia de diferentes perfiles de conducción dentro de sus colaboradores. Esto permite que la empresa Integra S.A pueda establecer su escalafón de bonificaciones.

- Aunque las dos clusterizaciones pueden representar las necesidades y la perfilación de los conductores dentro de la empresa Integra S.A. La clusterización utilizando cuatro (4) conglomerados es la que mejor representa los individuos analizados en este caso de estudio. Esto al evidenciarse notoriamente las divisiones entre perfiles que ayudan a establecer estas categorías que no existían antes en la empresa.

- Uno de los posibles trabajos futuros partiendo de esta investigación, es tomar muchas más variables, de carácter categórico que permitan analizar rasgos individuales del conductor, como lo puede ser la edad, tiempo que lleva en la empresa, estado civil, enfermedades previas, género, tipo de licencia, entre otras. Donde se puede plantear una clusterización jerárquica que es la que mejor resultados tiene frente a este tipo de datos.

- Un posible trabajo futuro es establecer un algoritmo de clasificación con los perfiles de conducción establecidos en esta investigación, así cada mes ir clasificando nuevamente a los conductores ya sean nuevos o antiguos, para seleccionar quienes obtienen el beneficio de la bonificación total, parcial o ningún reconocimiento por su estilo de conducción.

Apéndice A

Código Fuente

En el script A.1 se presenta la exploración del código fuente para la unificación de los archivos generados por día y por mes.

Listing A.1: Rutina unificación

```
#Importar librerias para manejo de archivos
y directorios (gestion de sistema operativo)

from os import listdir
from os.path import isfile, join

#Directorio donde se encuentran los archivos
mypath = './archivos/'

#Filtrar directorios y archivos ocultos
onlyfiles = [f for f in listdir(mypath)

if isfile(join(mypath, f)) and f[0] != '.']

#Extraer los numeros de los vehiculos para ordenar los
archivos y reflejar esto en los registros de salida
diccionarioCodigoArchivos = {}
listadoCodigos = []
for nombreArchivo in onlyfiles:
    listadoCodigos.append(nombreArchivo[11:13]+
    nombreArchivo[6:8])
    diccionarioCodigoArchivos[nombreArchivo[11:13]
    +nombreArchivo[6:8]] = nombreArchivo

#Ordenar listado de codigos
listadoCodigos.sort(key=int)

#Actualizar listado de nombres de archivos con el orden
obtenido
```

```
30
31 onlyfiles.clear()
32 for codigo in listadoCodigos:
33     onlyfiles.append(
34     diccionarioCodigoArchivos[codigo])
35
36 #Salidas de prueba
37 print(listadoCodigos)
38 print(diccionarioCodigoArchivos)
39 print(onlyfiles)
40 print(type(onlyfiles))
41
42 #Parar ejecucion para diagnosticar
43 #input()
44
45 #Crear el archivo de salida que recibira todos los
46 registros
47 archivoComas = open('arSalidaComas.csv', 'w+')
48
49 #Bandera primera vez para los nombres de los campos
50 banderaPrimeraVez = True
51
52 #Recorrer el listado de archivos
53 for nombreArchivo in onlyfiles:
54
55     #Abrir cada archivo
56     f = open(mypath+nombreArchivo, 'r')
57
58     #Contador de lineas del archivo de la iteracion
59     actual
60     contadorLineas = 0
61
62     #Recorrer linea a linea el
63     archivo sobre el que se esta iterando
64     y pasar cada linea al archivo
65     general de salida\\
66
67     for line in f:
68         if contadorLineas == 0 and
69         banderaPrimeraVez == False:
70             contadorLineas += 1
71             continue
72         else:
73             archivoComas.write(line)
74             contadorLineas += 1
75
76     #Del segundo archivo en adelante desactivar
77     bandera
78     banderaPrimeraVez = False
79
```

```
80      #Cerrar cada archivo
81      f.close()
82
83  #Cerrar el archivo de salida general
84  archivoComas.close()
```

En el script A.2 se presenta la exploración del código fuente para la unificación de los archivos generados por día y por mes.

Listing A.2: Análisis Datos R

```
1  #Carga de Datos Servicios con Conductor y Bus, se adecuan
2  las variables en formato de fecha y no como factor
3
4  datos_conductores <- read.csv("I:/TESIS/DatosOctubreGPS/OperadorBusHora.csv")
5  datos_conductores$fecha <- as.POSIXct(datos_conductores$fecha,
6  format = "%Y-%m-%d")
7  datos_conductores$hora_ini <- as.POSIXct
8  (datos_conductores$hora_ini, format = "%H:%M:%S")
9  datos_conductores$hora_fin <-
10 as.POSIXct(datos_conductores$hora_fin, format = "%H:%M:%S")
11 datos_conductores$PERA_FNACIMIENTO <-as.POSIXct(datos_conductores$PERA_FNACIMIENTO, format= "%Y-%m-%d")
12 datos_conductores$PERA_FNACIMIENTO <- as.integer(now() - datos_conductores$PERA_FNACIMIENTO)/365
13 datos_conductores$NOMBRE <- as.character(datos_conductores$NOMBRE)
14 datos_conductores$EQU_CODIGO <- as.character(datos_conductores$EQU_CODIGO)
15 diasemanaconductor <- wday(datos_conductores$fecha, label= FALSE)
16 datos_conductores <- mutate(datos_conductores, dia = diasemanaconductor)
17 longitud_conductores <<- length(datos_conductores$fecha)
18 str(datos_conductores)
19 head(datos_octubre)
20
21 # se reduce la tabla de servicios de conductores tomando la hora inicio y la hora final
22 #se reduce el dataset de 16522 a 2445
23 fecha_conductor <- datos_conductores_habil$fecha[1]
24 nombre_conductor <- datos_conductores_habil$NOMBRE[1]
25 inicio_servicio_conductor <- datos_conductores_habil$hora_ini[1]
26 fin_servicio_conductor <- datos_conductores_habil$hora_fin[1]
27 codigo_equipo <- datos_conductores_habil$EQU_CODIGO[1]
28
29
30 conductores_reducido <- data.frame(fecha_conductor, nombre_conductor, inicio_servicio_conductor, fin_
       servicio_conductor, codigo_equipo)
31 j <- 1
32
33
34 for (i in 1:longitud_conductores)
35   {
36     if((datos_conductores_habil$fecha[i] == conductores_reducido$fecha_conductor[j]) && (datos_conductores_
           habil$NOMBRE[i] == conductores_reducido$nombre_conductor[j]))
37
38       {
39         if (datos_conductores_habil$hora_fin[i] == datos_conductores_habil$hora_ini[i+1]){
40
41           conductores_reducido$fin_servicio_conductor[j] <- datos_conductores_habil$hora_fin[i+1]
42         }
43         else{
44
45           tablatemporal<- data.frame(fecha_conductor=datos_conductores_habil$fecha[i+1],
46                                      nombre_conductor= datos_conductores_habil$NOMBRE[i+1],
47                                      inicio_servicio_conductor = datos_conductores_habil$hora_ini[i+1],
48                                      fin_servicio_conductor = datos_conductores_habil$hora_fin[i+1],
49                                      codigo_equipo = datos_conductores_habil$EQU_CODIGO[i+1])
50           conductores_reducido <- rbind(conductores_reducido, tablatemporal)
51           j <- j+1
52         }
53 }
54
55 }
56
57 write.csv(conductores_reducido, file="datosconducereducidos.csv")
```

Exploración del código fuente para la unificación de los archivos generados por día y por mes.

```
1
2  #Carga de Datos de Red GPS
3  datos_octubre <- read.csv("I:/TESIS/Datos Octubre GPS/arSalidaComas.csv")
4  #conversi  n de la fecha en tipo Date
5
6
```

```
#conversion  a formato POSIXct y a tipo caracter
datos_octubre$FECHA_GPS <- as.POSIXct(datos_octubre$FECHA_GPS, format = "%Y-%m-%d")
datos_octubre$HORA_GPS <- as.POSIXct(datos_octubre$HORA_GPS, format = "%H:%M:%S")
datos_octubre$EQU_CODIGO <- as.character(datos_octubre$EQU_CODIGO)

#creacion de las columnas Operador, Edad, Estado_civil
datos_octubre$Operador <- ""
datos_octubre$Edad <- ""
datos_octubre$Estado_civil <- ""
str(datos_octubre)
head(datos_octubre)
#creaci  n de variable d  ???a de la semana y mutaci  n del dataset para incluir el d  ???a de la semana
diasemana <- wday(datos_octubre$FECHA_GPS, label= FALSE)
datos_octubre <- mutate(datos_octubre, dia = diasemana)
#creacion de la longitud del arreglo para el posterior for
longitud_octubre <<- length(datos_octubre$FECHA_GPS)

datos_octubre <- filter(datos_octubre, datos_octubre$EQU_CODIGO != 'EV')

#Filtro para separar los d as sabados, domingos y festivos
datos_octubre_domingo <- filter(datos_octubre, datos_octubre$dia==1)
datos_octubre_sabado <- filter(datos_octubre, datos_octubre$dia==7)
#se filtra por la fecha de festivo que es el Lunes 14
datos_octubre_festivo <- filter(datos_octubre, datos_octubre$FECHA_GPS == '2019-10-14')

#se juntan en un dataframe, el d a domingo y el d a lunes festivo del 14
datos_octubre_festivo <- rbind(datos_octubre_festivo, datos_octubre_domingo)

#se construyen los datos del d a h bil
datos_octubre_habil <- filter(datos_octubre, datos_octubre$dia > 1)
datos_octubre_habil <- filter(datos_octubre_habil, datos_octubre_habil$dia < 7)
datos_octubre_habil <- filter(datos_octubre_habil, datos_octubre_habil$FECHA_GPS != '2019-10-14')

#lineas de c digo para saber si el filtro esta correctamente aplicado
#datos_octubre_habil_nolabora1 <- filter(datos_octubre_habil, datos_octubre_habil$HORA_GPS <= '2019-12-26
      00:15:00')
#datos_octubre_habil_nolabora2 <- filter(datos_octubre_habil, datos_octubre_habil$HORA_GPS <= '2019-12-26
      04:40:00')

#se filtran los datos del d a h bil que est n entre las 12:00 y las 12:15 y desde las 4:40 de la ma ana
      en adelante.
datos_octubre_habil <- filter(datos_octubre_habil, datos_octubre_habil$HORA_GPS <= '2019-12-26 00:15:00' |
      datos_octubre_habil$HORA_GPS >'2019-12-26 04:40:00')

#filtro de los conductores para los d as h biles
datos_conductores_habil <- filter(datos_conductores, datos_conductores$dia > 1)
datos_conductores_habil <- filter(datos_conductores_habil, datos_conductores_habil$dia < 7)
datos_conductores_habil <- filter(datos_conductores_habil, datos_conductores_habil$fecha != '2019-10-14')
datos_conductores_habil <- arrange(datos_conductores_habil, fecha, hora_ini)


#se filtran los datos del d a s bado que est n entre las 12:00 y las 12:15 y desde las 4:40 de la ma ana
       en adelante.
datos_octubre_sabado <- filter(datos_octubre_sabado, datos_octubre_sabado$HORA_GPS <= '2019-12-26 00:15:00'
      | datos_octubre_sabado$HORA_GPS >'2019-12-26 04:40:00')

#se filtran los datos del d a festivo que est n entre las 12:00 y las 12:15 y desde las 4:40 de la ma ana
       en adelante.
datos_octubre_festivo <- filter(datos_octubre_festivo, datos_octubre_festivo$HORA_GPS <= '2019-12-26
      00:15:00' | datos_octubre_festivo$HORA_GPS >'2019-12-26 04:40:00')


datos_octubre_habil <- select(datos_octubre_habil, FECHA_GPS:VEL_GPS, ACL_GPS, EQU_CODIGO, dia)

datos_octubre_habil <- arrange(datos_octubre_habil, FECHA_GPS, EQU_CODIGO, HORA_GPS)

datos_octubre_habil$Operador <- ""
datos_octubre_habil$Edad <- ""
datos_octubre_habil$Estado_civil <- ""

datos_conductores <- select(datos_conductores, fecha, hora_ini:EQU_CODIGO, -ruta_id)
datos_conductores_habil <- arrange(datos_conductores, fecha, NOMBRE, hora_ini)


# se reduce la tabla de servicios de conductores tomando la hora inicio y la hora final
#se reduce el dataset de 16522 a 2445
fecha_conductor <- datos_conductores_habil$fecha[1]
nombre_conductor <- datos_conductores_habil$NOMBRE[1]
inicio_servicio_conductor <- datos_conductores_habil$hora_ini[1]
fin_servicio_conductor <- datos_conductores_habil$hora_fin[1]
codigo_equipo <- datos_conductores_habil$EQU_CODIGO[1]

system.time({
for(j in 1:longitud_octubre){ #arrancamos a las 21/02/2020 3,18 #3349574
    for(i in 1:2445){
      if ( (datos_octubre_habil$FECHA_GPS[j]== conductores_reducido$fecha_conductor[i])){
        if (datos_octubre_habil$EQU_CODIGO[j]== conductores_reducido$codigo_equipo[i]){
          if((conductores_reducido$inicio_servicio_conductor[i] <= datos_octubre_habil$HORA_GPS[j]) & (datos
                _octubre_habil$HORA_GPS[j] <= conductores_reducido$fin_servicio_conductor[i])){
            datos_octubre_habil$Operador[j] <- datos_conductores$NOMBRE[i]
            datos_octubre_habil$Edad[j] <- datos_conductores$PERA_FNACIMIENTO[i]
```

```
            datos_octubre_habil$Estado_civil[j] <- datos_conductores$PERA_CIV_ID[i]
            print(i)
            print(j)
          }
        }
      }
    }
  }
}
)
```

Bibliografía

[Adolph, 2013] Adolph, M. (2013). Big data: Big today, normal tomorrow. *ITU-T Technology Watch Report*, 28:28.

[af Wåhlberg, 2007] af Wåhlberg, A. E. (2007). Long-term effects of training in economical driving: Fuel consumption, accidents, driver acceleration behavior and technical feedback. *International journal of industrial ergonomics*, 37(4):333–343.

[Agrawal et al., 1998] Agrawal, R., Gehrke, J., Gunopulos, D., and Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pages 94–105.

[Andritsos, 2004] Andritsos, P. (2004). *Scalable clustering of categorical data and applications*. University of Toronto.

[Antonio and Enrique, 2017] Antonio, G. Z. and Enrique, I. R. (2017). Economía digital en américa latina y el caribe.

[Arora, 2020] Arora, S. (22 de Enero de 2020). Top 5 python libraries for data science. urlhttps://www.simplilearn.com/top-python-libraries-for-data-science-article.

[Ball and Hall, 1965] Ball, G. H. and Hall, D. J. (1965). Isodata, a novel method of data analysis and pattern classification. Technical report, Stanford research inst Menlo Park CA.

[Barbará et al., 2002] Barbará, D., Li, Y., and Couto, J. (2002). Coolcat: an entropy-based algorithm for categorical clustering. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 582–589.

[Beale, 1969] Beale, E. (1969). *Euclidean cluster analysis*. Scientific Control Systems Limited.

[Calderón et al.,] Calderón, J. M. S., Trujillo, G. R. Ó. A. N., Flórez, G. A. R., del Interior, M., Santamaría, M. C., de Hacienda, M., Público, C., Echeverri, L. C. V., Uribe, A. G., de Salud, M., et al. Documento conpes 3920 dnp de 2018.

[Caliński and Harabasz, 1974] Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27.

[Castaño Vanessa, 2013] Castaño Vanessa, M. J. (2013). Evaluación de eficiencias relativas en el desempeño sostenible de los operadores en un sistema brt. *Universidad de la Sabana*, 1(1):10–20.

[Chai and Ngai, 2020] Chai, J. and Ngai, E. W. (2020). Decision-making techniques in supplier selection: Recent accomplishments and what lies ahead. *Expert Systems with Applications*, 140:112903.

[Constantinescu et al., 2010] Constantinescu, Z., Marinoiu, C., and Vladoiu, M. (2010). Driving style analysis using data mining techniques. *International Journal of Computers Communications & Control*, 5(5):654–663.

[DANE, 2017] DANE (2017). Encuesta de desarrollo e innovación tecnológica (edit). urlhttps://www.dane.gov.co/index.php/estadisticas-por-tema/tecnologia-e-innovacion/encuesta-de-desarrollo-e-innovacion-tecnologica-edit.

[Davies and Bouldin, 1979] Davies, D. and Bouldin, D. (1979). A cluster separation measure, ieee transactions on patter analysis and machine intelligence. vol.

[Duda et al., 1973] Duda, R. O., Hart, P. E., and Stork, D. G. (1973). *Pattern classification and scene analysis*, volume 3. Wiley New York.

[Dunn, 1973] Dunn, J. C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters.

[e Silva et al., 2015] e Silva, J. d. A., Moura, F., Garcia, B., and Vargas, R. (2015). Influential vectors in fuel consumption by an urban bus operator: Bus route, driver behavior or vehicle type? *Transportation Research Part D: Transport and Environment*, 38:94–104.

[Ester et al., 2000] Ester, M., Frommelt, A., Kriegel, H.-P., and Sander, J. (2000). Spatial data mining: database primitives, algorithms and efficient dbms support. *Data Mining and Knowledge Discovery*, 4(2-3):193–216.

[Fayyad and Stolorz, 1997] Fayyad, U. and Stolorz, P. (1997). Data mining and kdd: Promise and challenges. *Future generation computer systems*, 13(2-3):99–115.

[Frey and Van Groenewoud, 1972] Frey, T. and Van Groenewoud, H. (1972). A cluster analysis of the d2 matrix of white spruce stands in saskatchewan based on the maximum-minimum principle. *The Journal of Ecology*, pages 873–886.

[Friedman and Rubin, 1967] Friedman, H. P. and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62(320):1159–1178.

[Giraldo Martínez, 2014] Giraldo Martínez, L. N. (2014). Diseño de una arquitectura ti para el sistema avanzado de transporte público en el area metropolitana centro occidente.

[González Echeverri, 2018] González Echeverri, J. S. (2018). Identificación de los factores sociales discriminantes de personas expuestas a violencia intrafamiliar y conflicto armado en la ciudad de bogotá.

[Guha et al., 1998] Guha, S., Rastogi, R., and Shim, K. (1998). Cure: an efficient clustering algorithm for large databases. *ACM Sigmod record*, 27(2):73–84.

[Halkidi et al., 2000] Halkidi, M., Vazirgiannis, M., and Batistakis, Y. (2000). Quality scheme assessment in the clustering process. In *European conference on principles of data mining and knowledge discovery*, pages 265–276. Springer.

[Han et al., 2000] Han, E.-H., Karypis, G., and Kumar, V. (2000). Scalable parallel data mining for association rules. *IEEE transactions on knowledge and data engineering*, 12(3):337–352.

[Han et al., 2011] Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[Hartigan, 1975] Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, Inc.

[Hubert and Levin, 1977] Hubert, L. J. and Levin, J. R. (1977). Inference models for categorical clustering. *Psychological Bulletin*, 84(5):878.

[Hwang et al., 2018] Hwang, C.-p., Chen, M.-S., Shih, C.-M., Chen, H.-Y., and Liu, W. K. (2018). Apply scikit-learn in python to analyze driver behavior based on obd data. In *2018 32nd International Conference on Advanced Information Networking and Applications Workshops (WAINA)*, pages 636–639. IEEE.

[Indulska and Orlowska, 2002] Indulska, M. and Orlowska, M. E. (2002). Gravity based spatial clustering. In *Proceedings of the 10th ACM international symposium on Advances in geographic information systems*, pages 125–130.

[INTEGRASA, 2017] INTEGRASA (2017). Plan estratégico 2017-2021.

[Jain and Dubes, 1988] Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.

[Kargupta et al., 2001] Kargupta, H., Huang, W., Sivakumar, K., and Johnson, E. (2001). Distributed clustering using collective principal component analysis. *Knowledge and Information Systems*, 3(4):422–448.

[Kaufman and Rousseeuw, 2009] Kaufman, L. and Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

[Kearney, 2018] Kearney, A. (2018). Readiness for the future of production report 2018. In *World Economic Forum*.

[Kimball and Caserta, 2004] Kimball, R. and Caserta, J. (2004). Practical techniques for extracting, cleaning, conforming, and delivering data.

[Krzanowski and Lai, 1988] Krzanowski, W. J. and Lai, Y. (1988). A criterion for determining the number of groups in a data set using sum-of-squares clustering. *Biometrics*, pages 23–34.

[Laney, 2001] Laney, D. (2001). 3d data management: Controlling data volume, velocity and variety. *META group research note*, 6(70):1.

[Lin et al., 2016] Lin, W., Philip, S. Y., Zhao, Y., and Deng, B. (2016). Multitype clustering in heterogeneous information networks. *Knowledge and Information Systems*, 48(1):143–178.

[Maimon and Rokach, 2005] Maimon, O. and Rokach, L. (2005). Data mining and knowledge discovery handbook.

[Mamani Rodríguez, 2015] Mamani Rodríguez, Z. E. (2015). Aplicación de la minería de datos distribuida usando algoritmo de clustering k-means para mejorar la calidad de servicios de las organizaciones modernas caso: Poder judicial.

[Mannering et al., 1995] Mannering, F., Kim, S.-G., Ng, L., and Barfield, W. (1995). Travelers' preferences for in-vehicle information systems: An exploratory analysis. *Transportation Research Part C: Emerging Technologies*, 3(6):339–351.

[Márquez et al.,] Márquez, I. D., Blanco, M. L. R., Castañeda, N. P. G., del Interior, M., García, C. H. T., de Relaciones Exteriores, M., Barrera, A. C., de Hacienda, M., Público, C., Blanco, M. C., et al. Documento conpes 3975 dnp de 2019 (bogotá, noviembre 8 de 2019)¡ fuente: Archivo interno entidad emisora¿ consejo nacional de política económica y social república de colombia.

[Marriott, 1971] Marriott, F. H. C. (1971). Practical problems in a method of cluster analysis. *Biometrics*, pages 501–514.

[Martinelli et al., 2018] Martinelli, F., Mercaldo, F., Nardone, V., Orlando, A., and Santone, A. (2018). Cluster analysis for driver aggressiveness identification. In *ICISSP*, pages 562–569.

[Martinez et al., 2017] Martinez, C. M., Heucke, M., Wang, F.-Y., Gao, B., and Cao, D. (2017). Driving style recognition for intelligent vehicle control and advanced driver assistance: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):666–676.

[McClain and Rao, 1975] McClain, J. O. and Rao, V. R. (1975). Clustisz: A program to test for the quality of clustering of a set of objects. *Journal of Marketing Research*, pages 456–460.

[Milligan, 1980] Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *psychometrika*, 45(3):325–342.

[Milligan, 1981] Milligan, G. W. (1981). A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, 46(2):187–199.

[Milligan and Cooper, 1985] Milligan, G. W. and Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50(2):159–179.

[Moine et al., 2011] Moine, J. M., Gordillo, S. E., and Haedo, A. S. (2011). Análisis comparativo de metodologías para la gestión de proyectos de minería de datos. In *Congreso Argentino de Ciencias de la Computación*, volume 17.

[Murcotts, 2015] Murcotts (2015). Driver behaviour training. urlhttp://www.murcotts.edu.au/resources/driver-behaviour-training/.

[Murray and Shyy, 2000] Murray, A. T. and Shyy, T.-K. (2000). Integrating attribute and space characteristics in choropleth display and spatial data mining. *International Journal of Geographical Information Science*, 14(7):649–667.

[Nagar et al., 2016] Nagar, P., Atriwal, L., Mehra, H., and Tayal, S. (2016). Comparison of generalized and big data business intelligence tools. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 3585–3588. IEEE.

[Oppenheimer, 2018] Oppenheimer, A. (2018). *¡ Sálvese quien pueda!: El futuro del trabajo en la era de la automatización*. Vintage Espanol.

[Ouellette and Wood, 1998] Ouellette, J. A. and Wood, W. (1998). Habit and intention in everyday life: The multiple processes by which past behavior predicts future behavior. *Psychological bulletin*, 124(1):54.

[Pereira and Romero, 2017] Pereira, A. and Romero, F. (2017). A review of the meanings and the implications of the industry 4.0 concept. *Procedia Manufacturing*, 13:1206–1214.

[Ratkowsky and Lance, 1978] Ratkowsky, D. and Lance, G. (1978). Criterion for determining the number of groups in a classification.

[Reinsel et al., 2019] Reinsel, D., Gantz, J., and Rydning, J. (2019). The digitization of the world: From edge to core, 2018.

[Rohani, 2012] Rohani, M. (2012). *Bus driving behaviour and fuel consumption*. PhD thesis, University of Southampton.

[Rohani et al., 2013] Rohani, M. M., Wijeyesekera, D. C., and Karim, A. T. A. (2013). Bus operation, quality service and the role of bus provider and driver. *Procedia Engineering*, 53:167–178.

[Rojko, 2017] Rojko, A. (2017). Industry 4.0 concept: background and overview. *International Journal of Interactive Mobile Technologies (iJIM)*, 11(5):77–90.

[Rolim et al., 2017a] Rolim, C., Baptista, P., Duarte, G., Farias, T., and Pereira, J. (2017a). Impacts of real-time feedback on driving behaviour: a casestudy of bus passenger drivers. *European Journal of Transport and Infrastructure Research*, 17(3).

[Rolim et al., 2017b] Rolim, C., Baptista, P., Duarte, G., Farias, T., and Pereira, J. (2017b). Real-time feedback impacts on eco-driving behavior and influential variables in fuel consumption in a lisbon urban bus operator. *IEEE Transactions on Intelligent Transportation Systems*, 18(11):3061–3071.

[Saboohi and Farzaneh, 2009] Saboohi, Y. and Farzaneh, H. (2009). Model for developing an eco-driving strategy of a passenger vehicle based on the least fuel consumption. *Applied Energy*, 86(10):1925–1932.

[Samatova et al., 2002] Samatova, N. F., Ostrouchov, G., Geist, A., and Melechko, A. V. (2002). Rachet: an efficient cover-based merging of clustering hierarchies from distributed datasets. *Distributed and Parallel Databases*, 11(2):157–180.

[Sánchez, 2005] Sánchez, I. J. B. (2005). Técnicas de agrupamiento para el análisis de datos cuantitativos y cualitativos.

[Sarle, 1983] Sarle, W. S. (1983). *Cubic clustering criterion*. SAS Institute.

[Scott and Symons, 1971] Scott, A. J. and Symons, M. J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics*, pages 387–397.

[Shafique and Qaiser, 2014] Shafique, U. and Qaiser, H. (2014). A comparative study of data mining process models (kdd, crisp-dm and semma). *International Journal of Innovation and Scientific Research*, 12(1):217–222.

[Shearer, 2000] Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22.

[Strömberg et al., 2015] Strömberg, H., Karlsson, I. M., and Rexfelt, O. (2015). Eco-driving: Drivers' understanding of the concept and implications for future interventions. *Transport policy*, 39:48–54.

[Taubman-Ben-Ari et al., 2004] Taubman-Ben-Ari, O., Mikulincer, M., and Gillath, O. (2004). The multidimensional driving style inventory—scale construct and validation. *Accident Analysis & Prevention*, 36(3):323–332.

[Thirumagal et al., 2014] Thirumagal, R., Suganthy, R., Mahima, S., and Kesavaraj, G. (2014). Etl tools in data mining—a review. *International Journal of Research in Computer Applications & Robotics*, 2(1):62–69.

[Tibshirani et al., 2001] Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423.

[Trujillo, 2013] Trujillo, J. C. (2013). *Diseño y explotación de almacenes de datos: conceptos básicos de modelado multidimensional*. Ecu.

[varios Artists, 2020a] varios Artists (2020a). Rstudio. urlhttps://rstudio.com/products/rstudio/.

[varios Artists, 2020b] varios Artists (2020b). What is python? executive summary. urlhttps://www.python.org/doc/essays/blurb/.

[Viswanathan, 2013] Viswanathan, A. (2013). Data driven analysis of usage and driving parameters that affect fuel consumption of heavy vehicles.

[Wang et al., 1997] Wang, W., Yang, J., Muntz, R., et al. (1997). Sting: A statistical information grid approach to spatial data mining. In *VLDB*, volume 97, pages 186–195.

[Wickham, 2014] Wickham, H. (2014). Tidy data.

[Zezulka et al., 2016] Zezulka, F., Marcon, P., Vesely, I., and Sajdl, O. (2016). Industry 4.0–an introduction in the phenomenon. *IFAC-PapersOnLine*, 49(25):8–12.

[Zfnebi et al., 2017] Zfnebi, K., Souissi, N., and Tikito, K. (2017). Driver behavior quantitative models: Identification and classification of variables. In *2017 International Symposium on Networks, Computers and Communications (ISNCC)*, pages 1–6. IEEE.

[Zhang et al., 1997] Zhang, T., Ramakrishnan, R., and Livny, M. (1997). Birch: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, 1(2):141–182.