Facultad de Economía y Negocios
Auditoría e Ingeniería en Control de Gestión

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO EN CONTROL DE GESTIÓN

Análisis al gasto en combustible de las empresas manufactureras de Chile utilizando herramientas de minería de datos

Autores : José Vásquez Valderrama
Obed Mora Villar

Profesor Guía : Hugo Nuñez Delafuente

Diciembre 2022

CONSTANCIA

La Dirección del Sistema de Bibliotecas a través de su unidad de procesos técnicos certifica que el autor del siguiente trabajo de titulación ha firmado su autorización para la reproducción en forma total o parcial e ilimitada del mismo.

Talca, 2022

Agradecimientos

Al finalizar esta investigación y trabajo de tesis, es inevitable no agradecer a todos los seres queridos que de alguna manera fueron partícipe de este proceso. Quiero agradecer de todo corazón a todas esas personas que quiero y que me escucharon con detención para darme ánimos o para soportar esos momentos de dificultad donde uno se queja del cansancio y de la ansiedad de esta etapa. Quiero agradecer a mi compañero de tesis, sin lugar a duda me sentí cómodo y con un gran apoyo constante para tomar distintas decisiones a la hora de desarrollar todo este trabajo. Por otra parte, y no menos importante, la disposición y paciencia del profesor Hugo Núñez fue una experiencia gratificante si tomo en cuenta el desafío que abordamos para este proceso, fue tolerante y la forma en cómo trabajamos en equipo fue innovadora, transformando el curso de esta investigación en algo sumamente enriquecedor para nuestro futuro.

(José Ignacio Vásquez Valderrama)

En primer lugar, agradezco a Dios por acompañarme en toda mi etapa universitaria y darme las fuerzas para terminar esta carrera, creo que todo viene de él, la Biblia dice "Porque Jehová da la sabiduría, y de su boca viene el conocimiento y la inteligencia" (Proverbios 2:6). Quiero agradecer también a mis padres Daniel Mora Sandoval y Fanny Villar Escobar por el sacrificio, esfuerzo y apoyo incondicional que me han entregado en mi formación personal y académica, acompañándome en todo momento para que yo pueda lograr las metas que me propuse. A mis hermanos Abner y David, como también tatas que siempre han estado a mi lado, a mis tíos, primos y amigos, por ayudarme y apoyarme en todo momento; quiero dar las gracias de forma especial mi tía Andrea Villar Escobar, quien fue mi mentora en matemáticas y a ella le debo todo lo que sé en esa materia. También, quiero agradecer a mi profesor guía Hugo Núñez Delafuente por apoyarnos y guiarnos en todo momento junto a mi compañero, ya que era un tema nuevo para nosotros pero se tomó el tiempo y la dedicación para enseñarnos entregarnos las herramientas necesarias para desarrollar un buen proceso de tesis. Finalmente quiero agradecer a mi compañero en esta investigación, ya que sin su apoyo esto no seria posible, fue capaz de seguirme en el tema de este trabajo que se escapa de nuestra zona de conocimiento, nos conocemos desde el inicio de este proceso universitario y ha sido fundamental su apoyo a lo largo de toda esta carrera.

(Obed Edom Mora Villar)

Índice

Índice de figuras

Índice de tablas

Resumen

Cuidar el planeta es de vital importancia para las personas ya que es el lugar que habitamos. Debemos cuidarlo de cara a la preservación de la vida. Esta investigación buscó caracterizar a las empresas que utilizan combustibles perjudiciales para el medio ambiente, esto se realizó mediante un estudio del gasto en combustible que tienen las empresas manufactureras en Chile y utilizando información generada por el INE a través de la ENIA.

Para hacer nuestros análisis, utilizamos algoritmos de clúster; uno de los métodos de minería de datos más importantes. Se utiliza para descubrir conocimiento en distintos conjuntos de datos multidimensionales. A través de esta metodología se identificaron dos grupos de empresas, de los cuales uno incorpora a las empresas que utilizan combustibles dañinos para el medio ambiente y el otro posee empresas cuyo gasto en combustible fue bajo.

Los análisis de esta tesis arrojaron que donde se concentra el mayor gasto en combustible es en la región de O´Higgins. Esta región se caracteriza por la presencia de empresas relacionadas con el rubro de la producción de alimentos, que generan un alto gasto en combustible a través de la utilización de gas en sus tres versiones para sus procesos productivos

En lo referido a los combustibles dañinos para el ecosistema, los resultados arrojaron que el petróleo en sus dos versiones es el más utilizado por las empresas y por tanto, lo convierte en el que genera más impacto negativo al medio ambiente.

Palabras clave: Combustibles, Clústeres, Empresas.

Abstract

Taking care of the planet is of vital importance for people since it is the place we inhabit. We must take care of it to preserve life. This research sought to characterize the companies that use fuels that harm the environment. This was done through a study of the fuel costs of manufacturing companies in Chile and using information generated by the INE through the ENIA.

To do our analysis, we use cluster algorithms, one of the most important data mining methods. It is used to discover knowledge in different multidimensional data sets. Using this methodology, two groups of companies were identified, one of which incorporates companies that use fuels that are harmful to the environment, and the other has companies whose fuel costs are low.

The analysis of this thesis showed that where the highest fuel spending is concentrated in the region of O'Higgins. This region is characterized by companies related to the food production sector, which generate high fuel costs due to using the three types of gases in their production processes.

Regarding fuels that are harmful to the ecosystem, the results showed that oil in its two forms is the most commonly used by companies, which is why it generates the most significant negative impact.

Keywords: Fuels, Clusters, Companies.

1. Introducción

El segundo gobierno de Sebastián Piñera buscaba mejorar la calidad de vida de los chilenos promoviendo el cuidado del medio ambiente, en aquella época el Estado se propuso como "El principal objetivo de la agenda ambiental del Gobierno para el período 2018-2021 en materia ambiental es mejorar el bienestar y calidad de vida de las personas a través de un desarrollo sustentable, lo cual reconoce que el desarrollo pleno de la sociedad descansa en tres pilares: crecimiento económico, cuidado del medio ambiente y equidad social." (REMA, 2019).

Uno de los desafíos con los que contaba ese gobierno era generar un plan de acción que tuviera un impacto positivo y de esta manera concientizar a la población chilena. Por esto se hacía necesario contar con más y mejor información, en la que se declarara cuáles eran los desafíos que tenía el gobierno de cara al estado del medio ambiente de nuestro país y el planeta; dichos planteamientos requieren acciones urgentes, coordinadas y que fueran respaldadas por el conocimiento científico para así lograr un impacto positivo que es lo que se buscaba a priori (REMA, 2019).

"En Chile, se reconocen tres grandes fuentes de contaminación del aire: los medios de transporte, las actividades industriales y la calefacción de las viviendas mediante la combustión de leña" (REMA, 2019). Por ese motivo nace la oportunidad de hacer una investigación acerca de esta temática, ya que en el reporte el Sistema Nacional de Información Ambiental (SINIA) que entregan año a año, no se encuentra un detalle relacionado con el consumo de combustible que tienen las empresas del rubro manufacturero, siendo este uno de los principales detonantes de la contaminación del aire.

La calidad del aire en Chile es un aspecto que preocupa a los gobiernos, en este sentido el Reporte del Estado del Medio Ambiente nos dice "por ejemplo, el año 2018, más de 9 millones de habitantes del país se encontraban bajo exposición de concentraciones promedio de material particulado fino (MP2,5) superiores a la norma, estimándose alrededor de 3.640 casos de mortalidad prematura por enfermedades cardiopulmonares, asociadas a la exposición crónica a este contaminante, entre otros impactos." (2019)

Actualmente los principales contaminantes del aire provienen del uso de combustibles como el carbón, utilizado en los procesos productivos de las industrias a lo largo de nuestro país "Bajo este contexto, se puede destacar la publicación del Plan de Prevención y Descontaminación Atmosférica para las comunas de Concón, Quintero y Puchuncaví (D.S. N°105/2018 del Ministerio de Medio Ambiente), que establece una serie de medidas para las principales fuentes de emisión identificadas en la zona" (REMA, 2019).

"Aunque Chile no es un actor relevante en el total de emisiones de gases de efecto invernadero (GEI) a nivel mundial (0,27 %), sí evidencia un aumento acelerado en el tiempo debido principalmente al incremento del consumo de combustibles fósiles, y por ello, se han propuesto compromisos ambiciosos de reducción de estos gases" (REMA, 2019). A pesar de esto, el IEMA y REMA no entregan un mayor detalle a las personas sobre qué empresas o rubros son aquellas que más gastan en combustibles dañinos para el medio ambiente. Para que de este modo, el gobierno pueda llevar a cabo acciones que permitan producir cambios sustanciales en Chile, dichas transformaciones podrían ser: crear impuestos verdes específicos de acuerdo al rubro de la empresa y al impacto negativo que tenga en el medio ambiente junto con considerar la salud de las persona, todo esto con el fin de promover la utilización

de combustibles amigables con el ecosistema.

Nuestro país no cuenta con un marco jurídico que permita asignar responsabilidades en lo relacionado a la contaminación que generan las empresas en diferentes áreas tales como: el aire, el suelo, el agua, entre otros (REMA, 2019).

A partir de lo señalado en los párrafos anteriores, el presente trabajo busca realizar un análisis exploratorio del gasto en combustibles que las empresas utilizan en sus procesos productivos. Para ello se utilizó una base de datos del Instituto Nacional de Estadísticas (INE) generada con datos de la Encuesta Nacional Industrial Anual (ENIA).

Todo este análisis se hace con el propósito de identificar y caracterizar el tipo de empresa que más contaminación genera en el aire a través del alto consumo de combustible, también se buscará identificar la región que más gasto genera en combustible y qué impacto tiene lo anterior en el medioambiente.

Esta investigación pretende mostrar al lector, un análisis del gasto en combustible que tienen las empresas a través de la utilización de los algoritmos de clústering de minería de datos, todo esto con el propósito de agrupar entidades y descubrir que características poseen en común unas con otras según el clúster donde se agrupen. De esta manera, se aportarán conclusiones y juicios de valor respecto al gasto elevado que las empresas tengan en el consumo combustibles y su relación al rubro que pertenecen según sus procesos productivos internos.

2. Problematización y Objetivos

2.1. Planteamiento del problema

El problema que abarcó y visibilizó este estudio, se relaciona con el alto impacto ambiental que tiene la utilización de combustibles. Actualmente, son cada vez más los países que adquieren políticas internas que buscan el cuidado del medio ambiente (United States Statistics Division, 2019), es por esto, que las autoridades del mundo se encuentran en vías de búsqueda de soluciones que lleven a un desarrollo sustentable y sostenible que permita al ser humano tener un menor impacto en el ecosistema, haciendo que los recursos naturales perduren por más tiempo y de alguna manera amortizar el efecto que genera la mala utilización de los bienes naturales de nuestro planeta (Tercera, 2022).

Por otro lado, en nuestro país, no existe un detalle más profundo de los gastos relacionados por concepto de combustible en el rubro manufacturero, por lo tanto, se volvió una necesidad prioritaria el hecho de tener esta información para efectos de análisis y para una futura toma de decisiones. Actualmente, los datos del Instituto Nacional de Estadísticas permiten conocer cómo está el país, en relación a su economía, población, territorio entre otros temas; esta institución nos entrega antecedentes acerca de hechos o fenómenos de interés para todas las personas y sirven de apoyo para que el Estado realice acciones que van en beneficio de toda la población, incluyendo también a empresas, académicos, estudiantes y público en general. Con las estadísticas podemos conocer, por ejemplo, cómo ha evolucionado la población en un tiempo determinado (Instituto Nacional de Estadísticas, 2022); de esta manera, nos percatamos que el INE no entrega conclusiones a partir de sus distintas bases de datos y no hay organismos que los depuren para un uso en la toma de decisiones en distintos aspectos del país, como lo es el medio ambiente.

Para ilustrar lo indicado anteriormente, existen diversos informes que hacen referencia a las consecuencias de los tipos de emisión de contaminantes, uno de estos es el informe de Chile Sustentable que abarca problemáticas relacionadas a la salud infantil provocadas por la exposición de agentes contaminantes del aire, entregando una escaza evidencia de los daños provocados a la sociedad (Tellerías, 2018). Por este motivo la información recopilada por el INE debiese usarse para analizar exhaustivamente cada uno de los parámetros de interés, en este caso, el impacto del uso de combustibles en la industria manufacturera de Chile.

Es relevante mencionar que en la actualidad se está viviendo un proceso extraordinario que promueve renovación de la Constitución chilena, la cual incluye aspectos primordiales en temáticas medioambientales que buscan la creación de una legislación moderna y vanguardista, que pondría a Chile en línea con países considerados del primer mundo en asuntos ecológicos, como bien lo explica la revista Crítica Urbana (Águila, 2021).

2.2. Pregunta de investigación

En este sentido, ante la creciente preocupación que muestran las autoridades de los países sobre el impacto negativo que tiene el uso de ciertos combustibles en el medio ambiente, se planteó la siguiente pregunta que será el foco de la investigación planteada en esta tesis:

"¿Qué características tienen las empresas que utilizan en sus procesos productivos combustibles que generan un alto impacto al medio ambiente?"

De esta forma la investigación clasificó las empresas manufactureras utilizando la metodología de clustering y de esta forma se realizó un análisis que permita responder la pregunta planteada.

El carácter de este trabajo de investigación será mixto, es por esto que el análisis será de forma cuantitativa y cualitativa a partir de los datos entregados por el Instituto Nacional de Estadísticas a través de la Encuesta Nacional Industrial Anual del 2019, también se analizarán diversos papers de investigación, así como artículos y estudios que potencien el proyecto.

La razón por la cual este proceso de investigación se trabajó en la modalidad mixta, se debe a los siguientes factores: en primer lugar, se desarrolló un análisis de datos numéricos como el gasto que tienen las empresas del rubro manufacturero de Chile en combustible; en segundo lugar, a través del estudio de estos datos se encontraron las características de las empresas que tienen un mayor impacto ambiental por concepto de este gasto. En síntesis, se planteó una investigación de este tipo, aprovechando las ventajas que ofrecen ambos métodos con el fin de realizar un análisis exhaustivo y bien fundamentado (Santander-Universidades, 2019).

2.3. Justificación de la investigación

La relevancia de esta investigación surge por la necesidad de conocer y mostrar al lector dónde se encuentra concentrado el mayor gasto de combustible dañino para el medio ambiente, ante lo cual analizaron tres focos para responder a la pregunta antes señalada. Para realizar este estudio se analizará la Encuesta Nacional Industrial Anual (ENIA) realizada por el Instituto Nacional de Estadísticas (INE) en el año 2019, la cual corresponde a la última, disponible al momento de realizar este proyecto.

El primer foco que se abordó en la investigación es conocer el tipo de empresa manufacturera que tiene el mayor gasto de combustible dañino para el medio ambiente, es decir, a través de las particularidades de cada compañía se encontrarán semejanzas que nos lleven a una conclusión más exacta.

El segundo enfoque que se trabajó corresponde al análisis de la ubicación geográfica a nivel regional donde se concentra el mayor gasto en combustible dañino para el medio ambiente, tratando de analizar los diferentes aspectos que caracterizan a dicha región.

El tercer foco de análisis buscó identificar cuál es el tipo de combustible más y menos utilizado en Chile por las empresas manufactureras y a partir de ello indagar si el impacto es perjudicial o beneficioso para el medio ambiente.

2.4. Delimitación de la investigación

A raíz de que los datos analizados son entregados por el Instituto Nacional de Estadísticas de Chile, este país fue el primer límite geográfico a considerar; el segundo límite propuesto corresponde a las empresas manufactureras de la nación que respondieron la encuesta ENIA; el tercer límite solo incluyó el gasto de combustible que tienen las empresas manufactureras de Chile y que registraron en la encuesta del año 2019; no se midió el impacto que generen otras prácticas que no sean las descritas a lo largo de este trabajo y que tienen relación con el uso de combustibles.

Para realizar esta investigación se contó con un plazo estimado de 4 meses, en los cuales se recopiló y analizó las cifras disponibles en la base de datos de la ENIA que proporciona la página web del INE.

2.5. Alcance de la investigación

En este sentido y para el desarrollo esta investigación se utilizó la base de datos que entrega el INE con los datos recopilados por la ENIA, cabe señalar que esta encuesta solo es respondida por aquellas empresas del rubro manufacturero de Chile. De esta encuesta se consideraron solo las preguntas que tienen relación con los datos de empresas tales como: su clasificación, ubicación geográfica o datos relacionados que permitan caracterizar la entidad, con el fin de realizar una investigación más concreta y precisa; además, se incluyó preguntas relacionadas con el gasto y uso de combustible que tuvieron en el año. De esta forma se efectuó un análisis cuyo alcance se concentra en las empresas manufactureras de Chile que respondieron la encuesta y que a la vez utilizaron combustible para su funcionamiento.

2.6. Objetivos de la investigación

2.6.1. Objetivo general

Determinar las características que poseen las empresas según el tipo de combustible utilizado para generar energía en sus procesos productivos.

2.6.2. Objetivos específicos

I. Identificar los tipos de combustibles más utilizados por las empresas y su efecto en el ecosistema.

II. Determinar las características comunes de las empresas que utilizan combustibles amigables con el medio ambiente.

III. Determinar las características comunes de las empresas que utilizan combustibles no amigables con el medio ambiente.

IV. Identificar las diferencias que puedan existir por industria, región y tamaño de empresa.

3. Marco Teórico

3.1. Historia del combustible

El origen del combustible se remonta al momento en el que el hombre aprendió a utilizar el fuego, ya que necesitaba de este recurso para mantenerlo en funcionamiento y así tener una fuente de calor además de un lugar donde cocinar los alimentos que cazaban. La materia prima que descubrieron los primeros hombres y que podían utilizar para mantener el fuego en funcionamiento eran las ramas y troncos de árboles que encontraban en los bosques que habitaban o recorrían cazando o recolectando sus alimentos, dicho material era empleado

como leña para que el fuego perdurara más tiempo. Posteriormente, conforme el hombre iba evolucionando, descubrió que la madera que se recolectaba de los troncos se podía convertir en un combustible mucho más eficiente y es así como se descubrió el carbón de leña, un combustible con que el hombre logro un avance, pudiendo trabajar el hierro desde hace unos 1500 años antes de Cristo (Fernández-Betancur, 2005).

Con este nuevo descubrimiento básico, en lo relacionado a la utilización de los combustibles, el hombre pudo desenvolverse cientos de años hasta lograr nuevos avances significativos en esta materia. Durante ese período de estancamiento, los antiguos seres humanos utilizaban la leña y el petróleo que brotaba de algunos yacimientos de forma natural y que les servían al ser humano como combustible para mantener el fuego vivo.

Conforme avanza la historia de la humanidad, se van desarrollando nuevos trabajos que conllevan la extracción de materias prima, en ese sentido se necesitaba de forma imperiosa encontrar un combustible que fuera eficiente y que potenciara ese trabajo; ya que durante mucho tiempo el hombre solo ocupaba la fuerza física como medio para obtener dicho material, la cual en sus inicios era proporcionada por los esclavos, pero con el paso del tiempo pasó a ser una tarea de los animales e incluso se buscaron nuevas formas para hacer esta labor más efectiva como lo fue el uso de molinos de viento (Aquiles Gay, 1996).

Pero todas estas formas de generar energía y combustible eran muy deficientes ya que implicaban grandes sacrificios físicos y era indispensable lograr un avance significativo, el cual se consiguió cuando en la revolución industrial el hombre creó una máquina a vapor potenciada por la utilización del carbón como combustible, pues venía a simplificar todas las necesidades que tenía la industria en esa época (Aquiles Gay, 1996).

Este descubrimiento trajo consigo ciertas consecuencias que no se tenían en cuenta, siendo la principal de estas, el alto requerimiento de combustible que necesitaba la maquinaria para funcionar, por tanto, la demanda de carbón fue en aumento conforme crecía la cantidad de máquinas en funcionamiento; para esa fecha, el carbón y la madera eran las principales fuentes de combustible, pero nadie era consciente de las funestas consecuencias que traía la alta demanda de este material, ya que la elevada explotación de árboles trajo consigo la erosión de los terrenos, la desaparición de fuentes naturales de agua y quebradas; sumado a ello, la producción de carbón pasó a ser un poco más técnica y generando la desaparición de zonas que eran utilizadas para la agricultura. Cabe señalar que el carbón como material, es peligroso para el ser humano debido a sus propiedades química, lo que en aquella época provocó mayor cantidad de accidentes letales. (Oviedo-Salazar et al., 2015).

En este orden cronológico y a raíz de los problemas generados por la extracción desmedida de carbón, en el año 1859 el hombre alcanza el siguiente avance en materia de combustible cuando se realiza la perforación del primer pozo de petróleo en Estados Unidos de América, de modo que, a raíz de este descubrimiento comienzan a surgir una serie de inventos que utilizaban este nuevo elemento como fuente de combustible, un ejemplo de ello es: el generador eléctrico, el motor de combustión interna, la luz eléctrica y el automóvil.

Del mismo modo que ocurrió con el carbón, el petróleo se posicionó como una gran amenaza para el ecosistema ya que en las zonas donde existen yacimientos había poca vegetación, debido a que los requerimientos en materia de terreno para la explotación de este recurso es alta, es decir, se requiere de mucho espacio para la instalación de las maquinarias necesarias para producir este combustible.

También es destructivo en lugares como lagunas o el mar, los cuales se ven altamente

afectados debido a la formación de una capa aceitosa en la superficie donde se extrae el petróleo que impide la oxigenación del agua y en consecuencia, inhibe la existencia de plantas marinas. Adicionalmente existe el riesgo que tiene el transporte de este combustible debido a que se pueden producir accidentes cuyas consecuencias son fatales para el medio ambiente. Además, la utilización del petróleo como combustible principal trae consigo la emisión de gases que son perjudiciales para la atmósfera tales como el monóxido de carbono o el cloro-fluoro-carbono que son los elementos causantes del llamado agujero en la capa de ozono que cada vez genera más daño a la tierra debido a las altas temperaturas (Fernández-Betancur, 2005).

Ante esta crisis que empezó a ser cada vez más evidente en la época, aparece otro combustible que es el gas, un combustible fósil que, a pesar de encontrarse muchas veces en los mismos yacimientos de petróleo, algunos estudios científicos demuestran que no es tan dañino para el medio ambiente su extracción y posterior utilización. Con el propósito de generar menos impacto al medio ambiente con la utilización de combustibles, los científicos de la época comenzaron a probar utilizando alcohol como fuente de energía, ya que es menor la contaminación que produce porque su extracción se hace mediante procesos fisicoquímicos realizados al bioma, lo que también trae el aprovechamiento de ciertos desechos orgánicos. Pero actualmente este combustible aún es un complemento a la gasolina debido a que tiene un bajo poder calorífico y por ende, su efectividad es reducida, por lo que aún se está trabajando en investigar sus propiedades y usos para así potenciar su utilización, aunque existen países que lo emplean como un combustible base(Fernández-Betancur, 2005).

Con el paso de los años el hombre realiza otro descubrimiento en materia energética y es la energía nuclear, que se concretó en la construcción del primer reactor nuclear en Estados Unidos en el año 1942, pero a pesar de las esperanzas que se pusieron en este nuevo método de obtener energía, el alto crecimiento en la población mundial lleva a que en el año 1973 solo ocupara una pequeña porción de la producción mundial que excedía las 6000 toneladas equivalentes de petróleo. De ahí en adelante, el ser humano ha ido investigando cada vez más formas de tener combustible para generar energía y se ha enfocado en obtener combustibles menos perjudiciales para el medio ambiente (Oviedo-Salazar et al., 2015).

3.1.1. Uso de combustibles en el mundo

Debido a la evolución que ha tenido la sociedad en términos tecnológicos e industriales, el uso de combustibles se ha vuelto una obligación para contribuir con el desarrollo de dichas actividades; los efectos secundarios de este tipo de tareas traen consecuencias tanto positivas cómo negativas para el medio ambiente y la salud humana.

El desarrollo tecnológico siempre ha sido un sinónimo de mejora en la calidad de vida de las personas, en lo relacionado a las necesidades básicas como en otros aspectos secundarios. Un claro ejemplo de esto son los medios de transportes que el ser humano ha implementado en los últimos dos siglos; esto ha permitido a las personas moverse a través del mundo por distintos medios para viajar, también ha permitido que el traslado de materiales para distintos usos sea más eficiente y rápido. Es aquí donde la sociedad evoluciona abruptamente en términos medicinales, educativos, tecnológicos e industriales .

Por otro lado, el uso de combustibles nos entrega consecuencias negativas como la contaminación o el aumento en la huella de carbono de las personas; el problema de esto es que

a medida que pasan las décadas, la extracción de estos recursos aumenta significativamente, según cifras que entrega el Programa de las Naciones Unidas para el Medio Ambiente, hay un incremento del 45 % en el uso de combustibles fósiles en el mundo, más del tripe de hace poco más de 50 años (United Nations, 2019).

Desde el punto de vista de las empresas, se cree que la utilización de combustibles fósiles es de las principales fuentes que emiten gases de invernadero a través de la fabricación, transporte o cualquier tipo de consumo de energía dentro de la organización. Es por esto que en los últimos años, las empresas se han comprometido de manera gradual para reducir este impacto en el medio ambiente, ya que debido al desarrollo nuevas tecnologías y distintas formas de abordar el funcionamiento de la empresa, en la mayoría de los casos se puede reducir este impacto, lo que genera una eficiencia energética en sus líneas de producción.

Otro factor a considerar es la brecha que existe entre los países desarrollados y subdesarrollados o naciones con peor o mejor calidad de vida; para ilustrar esto, de acuerdo con la División de Estadísticas de las Naciones Unidas, en países que poseen mayores ingresos, esta huella de carbono per cápita, o sea, la cantidad de materias primas para satisfacer nuestras necesidades es 10 veces mayor que en países de menor ingreso (United States Statistics Division, 2019).

Finalmente, el último tópico que se aborda, es el tipo de combustibles que usa para la producción de energía, desafortunadamente la mayoría de los combustibles que se utilizan son del tipo fósil, esto es negativo para la sociedad debido a que estos poseen el mayor efecto secundario que recae en la emisión de gases de invernadero que son perjudiciales tanto para el ecosistema como para la salud de las personas. A pesar de que se ha visto una disminución en la utilización de combustibles para producir energía a lo largo de los últimos 60 años, aún tiene más de tres cuartos del uso de estos, pasando desde 1960 con un 94 %, hasta la actualidad donde se utiliza un 81 % a nivel mundial (Banco Mundial de Estadísticas, 2019).

3.2. Historia de la industria manufacturera en Chile

La historia de la industria manufacturera de Chile se remonta principalmente a este siglo, sus inicios pueden ser establecidos en los gobiernos de Manuel Montt o José Joaquín Pérez; pero hay estudios y análisis históricos que mencionan al gobierno de Federico Errázuriz como uno de los actores principales en lo relacionado al desarrollo de la industria manufacturera en nuestro país a partir del año 1871. (De Ramón, 1988).

Tal y como pasa en la actualidad, el desarrollo de la industria manufacturera se llevó a cabo principalmente en las zonas de Valparaíso y Santiago; pero también tiene un importancia bastante preponderante el norte del país en todo lo relacionado con la minería, que durante los años posteriores tomó más realce histórico a nivel de país; desde sus inicios la industria ha estado centralizada principalmente alrededor de Santiago, la capital y Valparaíso por la existencia del puerto tan importante y famoso a nivel mundial durante esa época. (De Ramón, 1988).

El desarrollo de la industria manufacturera en el país trajo consecuencias negativas, una de estas es el alto impacto medioambiental que se genera a través de la intervención del terreno para la extracción de materias primas, otra es por la utilización de combustibles altamente contaminantes para el planeta que impactan negativamente en la calidad del aire; de hecho en sus inicios, la minería solo utilizaba carbón para su funcionamiento y si bien en

la actualidad, los avances tecnológicos han permitido un avance en esta materia, se puede observar como las empresas relacionadas con la minería registran gastos elevados en relación a la utilización de combustibles fósiles como el petróleo. (De Ramón, 1988).

Este desarrollo de la industria no solo trajo consecuencias negativas para el país, sino también positivas, un ejemplo de ello es que al establecerse las fundiciones en lugares como Coquimbo o Copiapó, llevo a la creación de escuelas especializadas en instruir profesionales con un conocimiento técnico que propiciara y mejorara la industria; además hubo un aumento en los puestos de trabajo ya que la minería requería de altas concentraciones de capital humano para su funcionamiento, lo cual trajo grandes beneficios para el país(De Ramón, 1988).

Para comprender aún más el desarrollo histórico que ha tenido la industria manufacturera en Chile, las Tablas 1 y 2 presentan la realidad de este rubro en el año 1979.

Tabla 1: Participación de los distintos tipos de establecimientos en el número total de establecimientos, el empleo y el valor agregado en 1979

Indicadores	Establecimientos menos de 10 trabajadores %	Establecimientos entre 10 y 49 trabajadores	Establecimientos con más de 50 trabajadores
Establecimientos	41.7	45.4	12.9
Empleo	7.7	27.0	65.3
Valor Agregado	1.8	12.2	86.0

Nota: Extraído de (Huelva and Núñez, 2010)

Como se puede observar en la Tabla 1, nos muestra una realidad específica del año 1979, donde la gran empresa era quien dominaba la industria y es la pequeña de estas la que presenta la participación más baja en el valor agregado aportado, esto muestra como la industria avanzó a lo largo de los años y se fue convirtiendo cada vez más en un potenciador económico para país(Huelva and Núñez, 2010).

La Tabla 2 nos mostrará un desglose que apunta al valor agregado aportado según empresa y número de trabajadores de estas.

Tabla 2: Participación en el Empleo de los Distintos Tipos de Establecimientos en los Diversos Sectores Productivos en %).

Sector	Participación empleo establecimientos menos de 10 trabajadores	Participación empleo establecimientos entre 10 y 50 trabajadores y 49 trabajadores	Participación empleo establecimientos más de 50 trabajadores
Productos alimenticios, bebidas y tabaco	9.2	32.5	58.5
Textiles, prendas de vestir e industrias del cuero	7.6	27.0	65.4
Industria de la madera y productos de la madera incluidos muebles	12.3	36.5	51.2
Fabricación de papel y sus productos. Imprentas y editoriales	8.3	22.9	68.7
Industria química de caucho y plástico, derivadas del petróleo y carbón	3.5	20.7	74.8
Fabricación de productos minerales no metálicos	9.7	19.0	71.3
Industrias metálicas básicas	1.1	7.5	91.4
Fabricación de productos metálicos maquinaria y equipos	5.9	24.8	69.3
Otras industrias manufactureras	22.4	41.7	36.0
Total, Industria Manufacturera	7.7	27.0	65.3

Nota: Extraído de (Huelva and Núñez, 2010)

En el Cuadro 2 se puede observar el alto impacto en el valor agregado que generar las empresas grandes, ya que su participación es bastante amplia en todos los sectores, el único que es relativamente bajo es en el rubro de otras industrias manufactureras que se encuentran en crecimiento y desarrollo en ese año.

Actualmente la industria ha crecido y se encuentra en constante desarrollo, a continuación, en las Tablas 3 y 4 se muestran algunos datos que fueron confeccionados en base a las estadísticas entregadas por la ENIA del año 2019.

Tabla 3: Participación en el Empleo de los Distintos Tipos de Establecimientos en los Diversos Sectores Productivos en %).

Sector	Valor Bruto de Producción en %	Consumo Intermedio en %	Valor Agregado en %
Elaboración de productos alimenticios	34.80	34.86	34.68
Elaboración de bebidas	7.14	6.60	8.31
Fabricación de productos textiles	0.41	0.40	0.42
Fabricación de productos de cuero y productos conexos	0.31	0.21	0.51
Fabricación de papel y de productos de papel	7.20	7.15	7.31
Impresión y reproducción de grabaciones	0.70	0.63	0.83
Fabricación de sustancias y productos químicos	16.37	17.95	12.98
Fabricación de productos farmacéuticos, sustancias químicas medicinales y productos botánicos de uso farmacéutico	1.90	1.45	2.88
Fabricación de productos de caucho y de plástico	3.98	3.96	4.01
Fabricación de otros productos minerales no metálicos	3.91	3.82	4.12
Fabricación de metales comunes	2.27	2.96	0.77
Fabricación de productos elaborados de metal, excepto maquinaria y equipo	3.81	3.63	4.19
Fabricación de productos de informática, de electrónica y de óptica	0.21	0.11	0.43
Fabricación de equipo eléctrico	0.96	1.02	0.85
Fabricación de maquinaria y equipo n.c.p.	1.54	1.38	1.88
Fabricación de vehículos automotores, remolques y semi remolques	0.14	0.14	0.13
Fabricación de otro equipo de transporte	0.12	0.09	0.19
Fabricación de muebles	0.86	0.75	1.11
Otras industrias manufactureras	0.11	0.10	0.13
Reparación e instalación de maquinaria y equipo	2.15	1.49	3.59
Industrias manufactureras	5.81	6.17	5.03
TOTAL	100	100	100

Nota: Extraído de Instituto Nacional de Estadísticas (2019)

Tabla 4: Valor Bruto, Consumo Intermedio y Valor Agregado Según Tramo por Número de Trabajadores.

Tramo por nº de trabajadores	Valor bruto de producción	Consumo intermedio en %	Valor agregado en %
Menos de 10 ocupados	2.27	1.51	3.91
10-19 Ocupados	1.67	1.72	1.56
20-49 Ocupados	6.69	6.17	7.81
50-99 Ocupados	9.85	10.23	9.03
100-199 Ocupados alimenticios	12.79	12.79	12.79
200-499 Ocupados	26.23	26.72	25.18
500 ocupados y más alimenticios	40.50	40.86	39.73
TOTAL	100	100	100

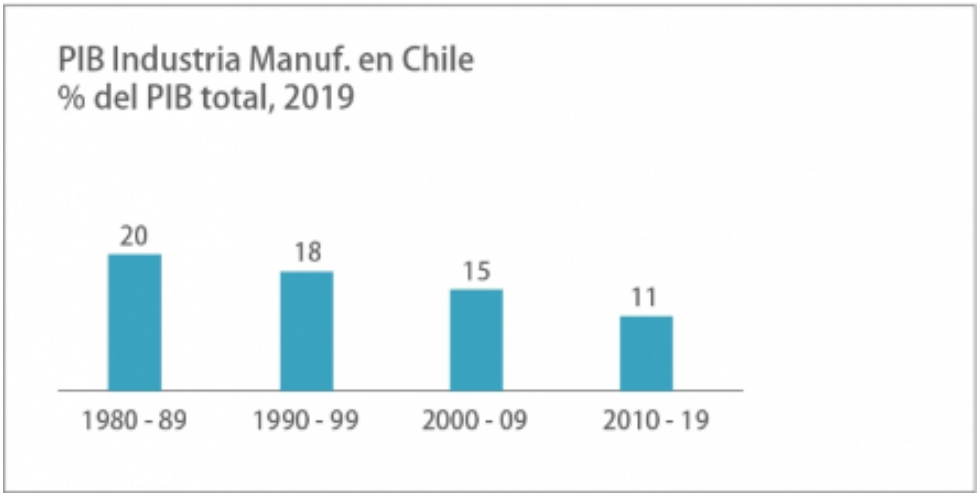
Nota: Extraído de Instituto Nacional de Estadísticas (2019)

Al establecer una comparativa de desarrollo respecto del año 1979, se ve como la industria chilena ha presentado un amplio desarrollo, abarcando aún más sectores que en ese año como es normal dado el crecimiento y desarrollo histórico que ha tenido Chile con el paso de los años; además en la actualidad la clasificación por número de trabajadores es más amplia debido al impacto generado por el crecimiento, llevando al INE a idear nuevas formas para clasificar a las empresas.

El crecimiento y desarrollo que ha tenido la industria manufacturera se observa en diferentes aspectos como el número de empresas, ya que en el año 1979 según los datos entregados por el banco nacional eran cerca de 1.200 empresas y actualmente en Chile existen más de 4.000 empresas dedicadas a este rubro, esto representa un importante crecimiento y desarrollo con el paso de los años.

Otro dato a destacar es el impacto que tiene la industria manufacturera en la participación del PIB nacional y como esta ha ido en declive con el paso de los años.

Figura 1: Producto Interno Bruto de la industria manufacturera en Chile en %. 2019



Nota: La figura muestra la evolución del PIB. Extraído del articulo "Plan de reactivación industrial para Chile después del covid". Escrito por Agosin and De Gregorio (2020)

En la Figura 1 junto con los cuadros 3 y 4, se puede analizar como la industria manufacturera dentro del país ha ido en constante crecimiento y desarrollo en cuanto a la variedad de rubros industriales presentes en el país, lo que impacta directamente en materias relacionadas a la empleabilidad nacional; pero también se puede ver un claro declive en la participación en el PIB nacional, esto debido a la entrada en la economía de otras áreas que están teniendo cada vez más presencia dentro del país y provoca que la participación del rubro manufacturero de Chile disminuya (Agosin and De Gregorio, 2020).

3.3. Estado del Medio Ambiente

La contaminación del medio ambiente es una temática que toma cada vez más relevancia y preocupación en los gobiernos del mundo, a raíz de esto, en Chile el gobierno ha implementado un sistema que permita el flujo de información, el cual consiste en reportes que se realizan de manera anual y un informe que se realiza cada 4 años; todo esto con el fin de mostrar a la población el estado del medio ambiente para su conocimiento y para que así las autoridades tomen acciones preventivas ante las consecuencias negativas que tiene la contaminación en la calidad de vida de las personas.

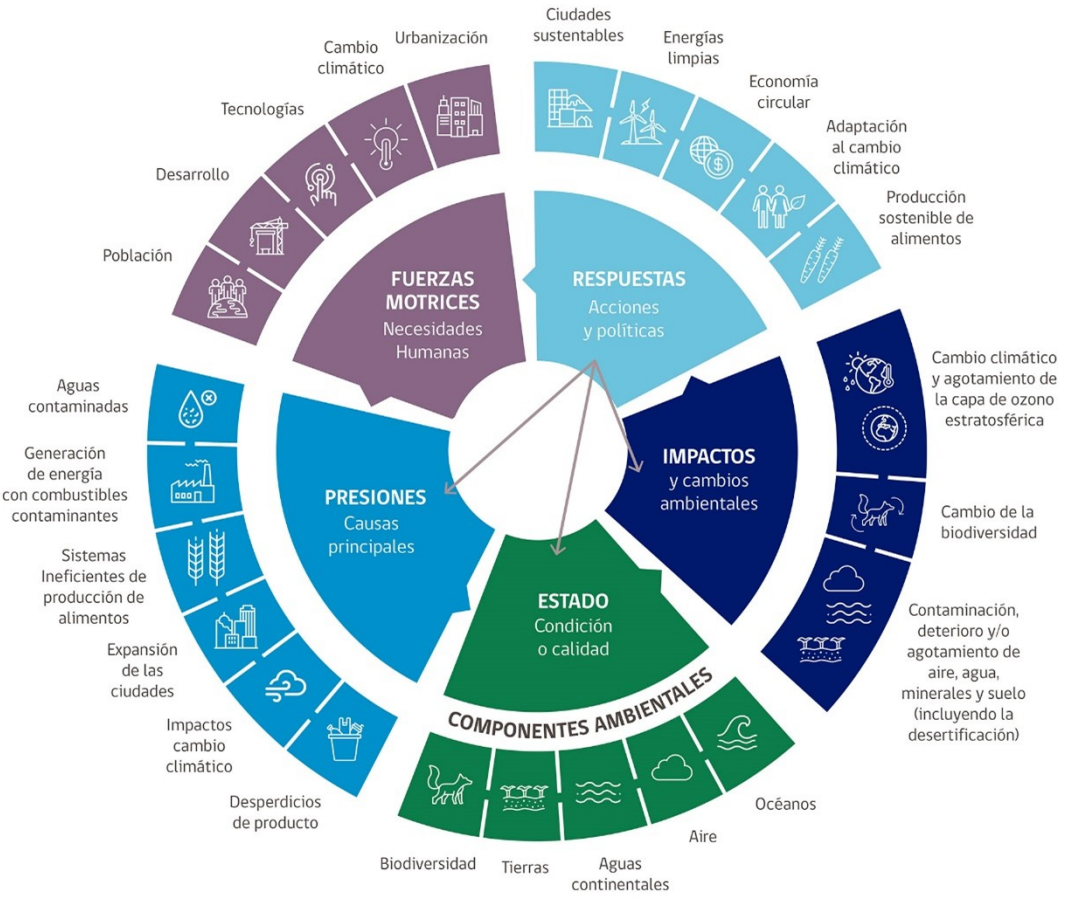
Así nace el Estado del Medio ambiente, que "es un instrumento que permite evaluar y monitorear el estado de los distintos componentes del medio ambiente, así como de las acciones y políticas públicas que se implementan para abordar los problemas que los afectan." (SINIA, 2021) y consiste en dos tipos de publicaciones, uno es un reporte que se publica año a año y el otro es un informe más completo que se entrega cada 4 años.

El objetivo de estos informes como lo detalla la página web del organismo SINIA es el siguiente: "Los informes y reportes indagan en las presiones que se ejercen sobre los ecosistemas, su estado de conservación y las respuestas de la institucionalidad ambiental, orientadas a armonizar las distintas actividades humanas con la protección del medio ambiente y la salud de las personas." (SINIA, 2021); por tanto, lo que se busca con este tipo de publica-

ciones es mantener a las autoridades y la población en general informada acerca de cómo se encuentra el medio ambiente de Chile.

La metodología que utiliza este informe es la Global Environment Outlook (GEO) o Perspectivas del Medio Ambiente Mundial de la Organización de las Naciones Unidas para el Medio Ambiente, esta considera un proceso participativo y consultivo." (SINIA, 2021). En la Figura 2 se observa la aplicación de esta metodología en los informes confeccionados por el ministerio del medio ambiente a través del SINIA.

Figura 2: Metodología del Estado del Medio Ambiente



Nota: La figura muestra la metodología utilizada por la SINIA para confeccionar los reportes e informes del estado del medio ambiente. Extraído de la página web SINIA (2021)

Estos reportes buscan presentar la realidad del país en todo lo relacionado al medio ambiente, dividiendo el estado del medio ambiente en 4 aristas: una de ellas detalla el impacto y cambios en el mismo que son provocados por la contaminación, otro detalla el estado de los componentes ambientales, con el fin de mostrar la realidad a las personas que lean las publicaciones, otra arista que muestra son las presiones, donde se detallan las causas

que provocan la contaminación, también se detallan las fuerzas motrices, arista que hace referencia a las necesidades que tiene la población respecto a la calidad de vida y por ultimo cada publicación nos muestra las respuestas que tiene cada gobierno ante esta temática.

Estas publicaciones son elaboradas por el Departamento de Información Ambiental de la División de Información y Economía Ambiental del Ministerio del Medio Ambiente de Chile. "La información ambiental es generada por los servicios públicos miembros del Comité Interinstitucional de Información Ambiental (CIIA), integrado por 25 instituciones con sus respectivas 60 subdivisiones administrativas y coordinado por el Ministerio del Medio Ambiente." (SINIA, 2021)

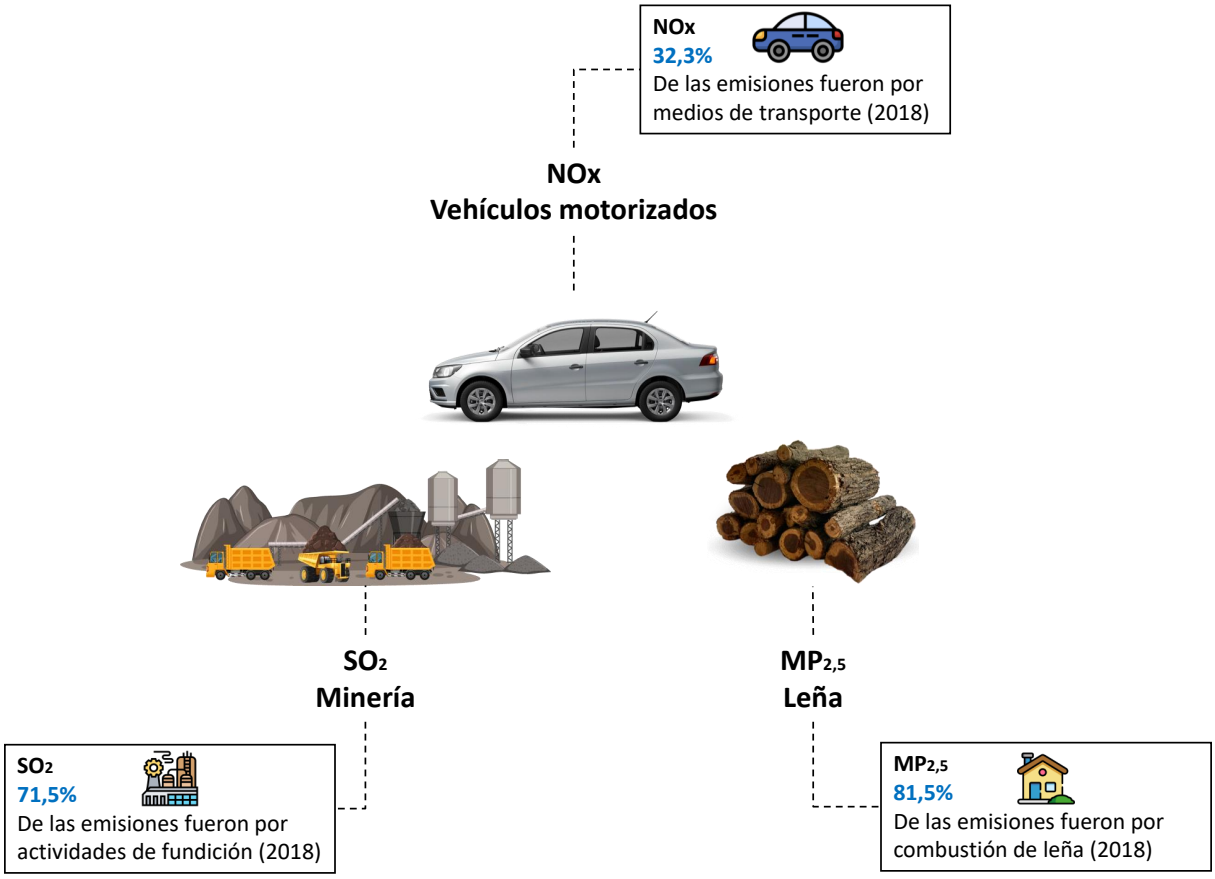
3.3.1. Contaminación del aire

La contaminación del aire es un tema que cada vez toma más auge en el mundo debido a su implicancia en la calidad de vida de las personas, sin ir más lejos, el REMA (Reporte del Estado del Medio Ambiente) del año 2019, expone que en el año 2018, alrededor de 9 millones de habitantes se encontraban con una exposición por sobre lo que exige la normal a material particulado fino, de hecho, se estima que alrededor de 3.640 casos donde existe una mortalidad prematura debido a enfermedades cardiopulmonares producto a una exposición constante a este tipo de contaminación. (REMA, 2019).

El IEMA (Informe del Estado del Medio Ambiente) que se entrega cada cuatro años y que el último fue en el año 2020 nos dice que "La contaminación del aire afecta la salud de personas y animales, daña la vegetación y el suelo, deteriora materiales, reduce la visibilidad y contribuye significativamente al cambio climático. Por ello, la calidad del aire sigue siendo una de las prioridades en materia de gestión ambiental. En Chile, las tres grandes fuentes de contaminación del aire son los medios de transporte, las actividades industriales y la calefacción de las viviendas por combustión de leña." (IEMA, 2020).; a raíz de esto el foco de las autoridades es crear políticas que contribuyan a bajar los indicadores que muestra este informe.

El IEMA 2020 presenta un resumen general de la contaminación del aire que se observa en la Figura 3.

Figura 3: Resumen de la Contaminación del Aire



Nota: La figura muestra un resumen de cómo se distribuye la emisión de partículas que son altamente perjudiciales para el ser humano y el medio ambiente . Elaboración propia a partir de IEMA (2020)

La Figura 3 muestra los tres agentes contaminantes más predominantes en el aire de nuestro país; donde el SO2 es el dióxido de azufre emanado por la utilización de combustibles dañinos en el sector industrial, el MP2,5 es el material particulado que posee un diámetro menor a 2,5 micras y que es perjudicial para el sistema respiratorio humano y el NOx que es un conjunto de gases muy reactivos como el óxido nítrico (NO) en conjunto con el dióxido de nitrógeno (NO2), ambos gases tienen en su composición nitrógeno y oxigeno e proporciones distintas y que regularmente se asocian a la emisión de gases por parte de los vehículos motorizados (IEMA, 2020).

Como se observa en la Figura 3, el sector industrial tiene una amplia participación en la emisión de uno de los gases que son tóxicos y perjudiciales para el medio ambiente como también para la calidad humana afectando de forma negativa al sistema respiratorio; la emisión de dióxido de azufre se concentra principalmente en el sector de la minería a través de las fundiciones según el reporte de 2018 concentrando el 71,5 % de las emisiones (IEMA, 2020).

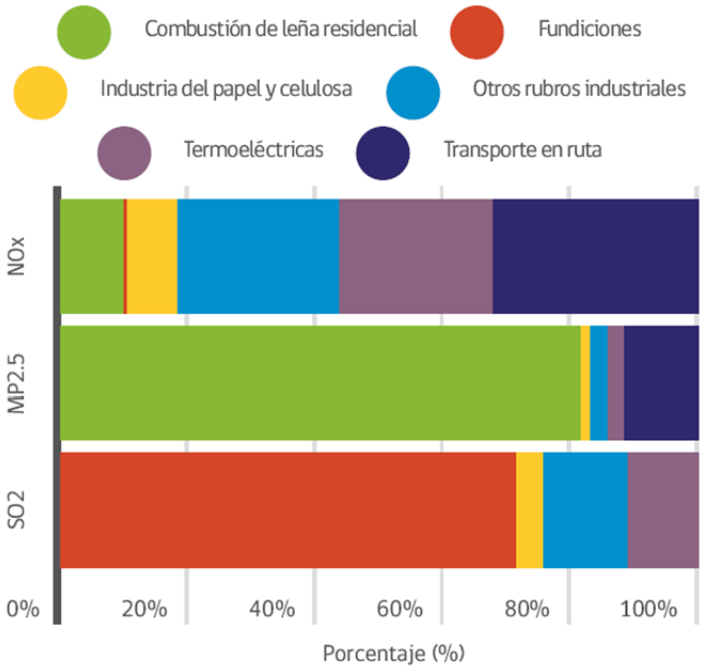
La preocupación de las autoridades viene dada porque "A nivel nacional se ha identificado a la combustión de leña residencial, las fundiciones de cobre, la industria del papel y

celulosa, las centrales termoeléctricas, el transporte en ruta y otros rubros industriales como las principales fuentes de emisiones de contaminantes locales: material particulado fino (MP2,5), dióxido de azufre (SO2) y óxidos de nitrógeno (NOX)"(IEMA, 2020).

Un dato preocupante, es que "Para el año 2018, el rubro de fundiciones de cobre constituye la fuente más relevante de emisión de SO2 con 191.000 toneladas, equivalentes a 71,54 % de las emisiones totales." (IEMA, 2020), estos datos son preocupantes ya que, si bien se muestra el impacto de este gas y el origen principal, no existe un detalle de que otro tipo de empresas tienen una alta emisión de estos gases por el consumo de combustible a nivel industrial, y que este trabajo pretende mostrar a través del gasto que reflejan los establecimientos en la ENIA.

La Figura 4, muestra la concentración de los 3 tipos de gases más contaminantes y perjudiciales para el ser humano, así como su origen principal.

Figura 4: Concentración de la emisión de gases dañinos para el ser humano y el medio ambiente así como su principal origen



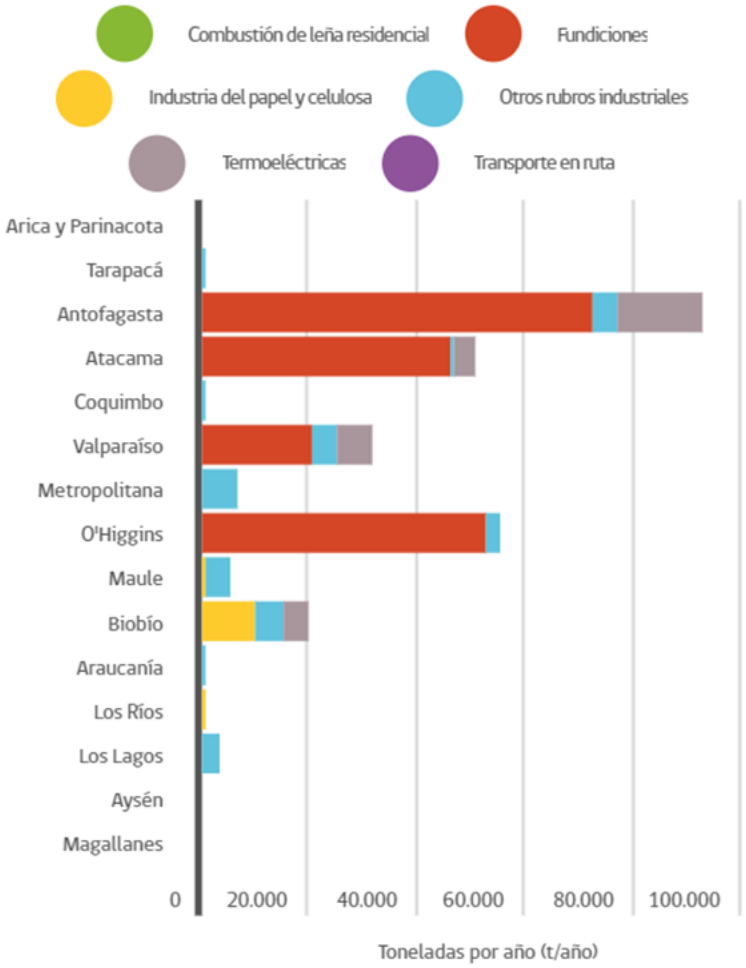
Nota: La figura muestra un resumen de cómo se distribuye la emisión de partículas que son altamente perjudiciales para el ser humano y el medio ambiente así como su origen. Extraído de IEMA (2020)

En la Figura 4 se presenta de forma más concreta los datos antes expuestos y nos da constancia sobre la alta concentración de estos tres tipo de gases negativos para el medio ambiente, se observa que la fuente principal de emisión de material particulado fino viene dado por la combustión de leña a nivel residencial con un 81,50 % de las emisiones, aproximadamente 83.528 toneladas anuales; también se observa que la emisión de óxidos de nitrógeno proviene mayoritariamente de dos fuentes que son el transporte en ruta el cual concentra alrededor del 32,32 % del total nacional con una cantidad de 44.714 toneladas emitidas y también el rubro de las termoeléctricas con el 23,95 % de las emisiones de este gas a nivel nacional, un aproximado de 33.136 toneladas.(IEMA, 2020).

Existen otras tres Figuras (5,6 y 7) que se presentan a continuación, que muestran la realidad de las emisiones de estos tres gases a nivel regional, además se clasifican por tipos de actividades e industrias con sus respectivas fuentes de emanación variando su intensidad según la naturaleza del sector y su ubicación.

La Figura 5 muestra las emisiones al aire de SO2 por región y según el tipo de fuente del año 2018:

Figura 5: Emisiones al aire de SO2 por región y según el tipo de fuente del año 2018

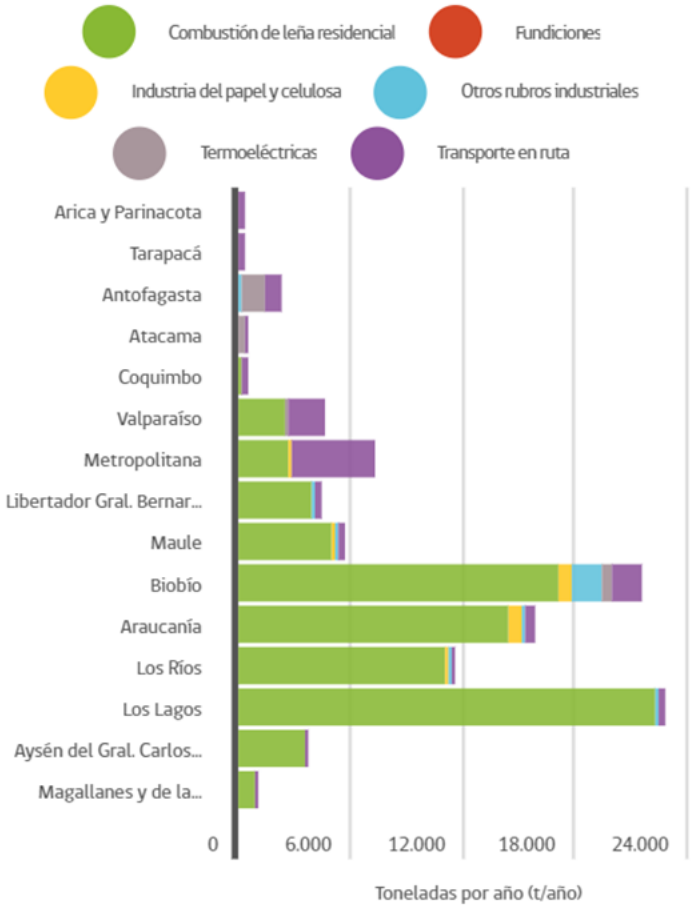


Nota: La figura muestra un resumen de cómo se distribuye la emisión de partículas SO2 por región y su origen. Extraído de IEMA (2020)

La Figura 5 presenta un resumen más detallado con la emisión de SO2 por parte de las actividades realizadas a nivel país, "En cuanto a óxido de azufre (SO2), las regiones de Antofagasta, Atacama, Valparaíso y O'Higgins son las que concentran los mayores niveles de emisión de este contaminante, debido a que en ellas se localizan las siete fundiciones de cobre que operan en el país." (IEMA, 2020).

La siguiente figura nos proporciona una visión de las emisiones al aire de MP25 por región y según el tipo de fuente del año 2018:

Figura 6: Emisiones al aire de MP25 por región y tipo de fuente del año 2018.

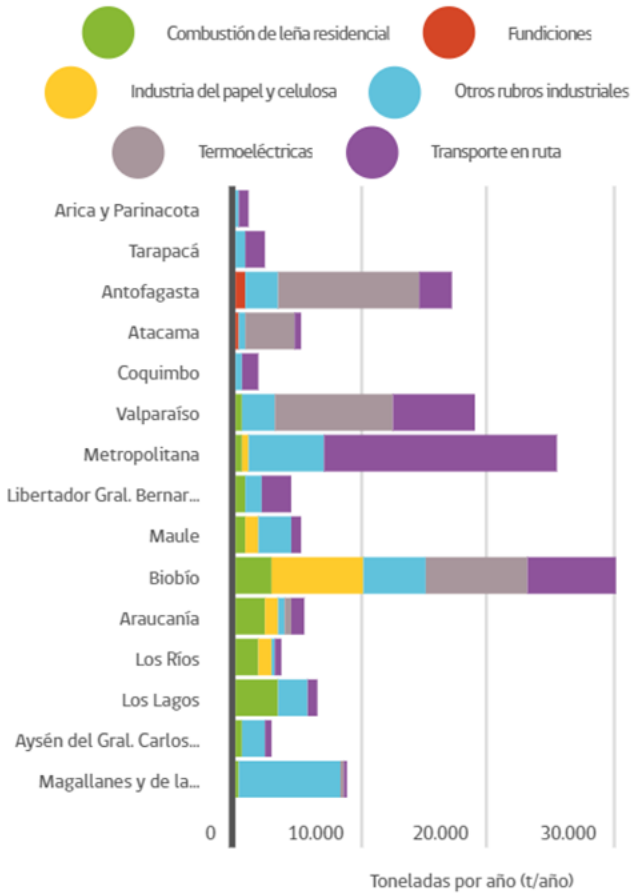


Nota: La figura muestra un resumen de cómo se distribuye la emisión de partículas MP25 por región y su origen. Extraído de IEMA (2020)

En la Figura 6 se puede ver con más detalle la emisión de MP 2,5 y se destaca de cómo varia la concentración de la emisión de MP 2,5 de norte a sur, y esto se da de forma natural ya que en nuestro país contamos con un clima bastante diverso que varía según la zona geográfica donde tengas tu residencia, ya que el norte del país se caracteriza por sus altas temperaturas, mientras que en el sur de Chile, la temperatura es cada vez más baja conforme al extremo sur del planeta; esto se ve reflejado en la Figura 6 ya que vemos como la concentración de MP 2,5 que se relaciona principalmente con la emisión producto de la combustión de la leña, aumenta en las regiones que se encuentran al sur de nuestro país debido a la alta utilización de la leña en los hogares de forma usual, como por ejemplo en cocinas a leña o estufas para dar calor en los hogares (IEMA, 2020).

En la Figura 7 y al igual que las anteriores, encontraremos un detalle por región sobre la emisión del gas NO que data del año 2018.

Figura 7: Emisión del gas NO año 2018.



Nota: La figura muestra un resumen de cómo se distribuye la emisión de partículas NO por región y su origen. Extraído de IEMA (2020)

En la Figura 7, se puede observar a detalle la emisión de óxidos de nitrógeno. En el norte sobresale la región de Antofagasta debido al rubro de las termoeléctricas. También destacan las regiones de Biobío y Metropolitana debido a que presentan una concentración alta de este gas producto de los transportes en ruta que viajan por estas zonas, y que generalmente está relacionado al uso de camiones para esta actividad(IEMA, 2020).

Actualmente el país se encuentra en vías de crear normas que controlen aún más la emisión de este tipo de gases implementando impuestos verdes que gravan las emisiones de material particulado (MP), los óxidos de nitrógeno (NOx), el dióxido de azufre (SO2) y dióxido de carbono (CO2) que provienen de fuentes fijas como industrias y de fuentes móviles como lo son los vehículos, este último depende de su rendimiento urbano y de las emisiones de NOx que tengan. (REMA, 2019).

3.4. Clúster como concepto general

El término clúster tiene diversas definiciones y aplicaciones, Michael Porter señaló "Un clúster es un grupo geográficamente próximo de empresas e instituciones asociadas, interco-

nectadas en un campo específico, ligadas por actividades e intereses comunes y complementarios" (Porter et al., 1998).

Los clústeres realizan asociaciones, para ello deben existir puntos donde las empresas analizadas converjan, ese lugar de encuentro corresponde a la ubicación geográfica "Todos los clúster comprenden una dimensión geográfica. La actividad productiva tiene lugar siempre en un espacio y en tal sentido, todo clúster industrial está geográficamente determinado." (Navarro, 2003). Para el contexto de análisis de este trabajo y considerando la cita anterior, el espacio geográfico que tienen en común las empresas es Chile.

Pero no en todos los casos la ubicación geográfica es el determinante principal para definir el marco espacial de los clústeres, de hecho Porter en su libro Çlusters and the new economics of competition"de 1980, menciona que los clústeres pueden aparecer en lugares geográficos diferentes; incluso en ciudades, pueblos o regiones de diferentes países, pero un clúster no puede ser de carácter global; hay autores que mencionan que un clúster no se puede realizar a nivel de un país porque falla en cuanto a la concentración geográfica debido a que todos pertenecen a un mismo lugar geográfico y se busca que exista distinción para los análisis posteriores (Navarro, 2003).

Para realizar un análisis utilizando la metodología de Clustering, se deben tomar en cuenta los siguientes aspectos:

- Límites espaciales del clúster: Como se mencionó con anterioridad, este aspecto es muy relativo y muchos autores difieren en sus definiciones respecto al espacio donde se debe desarrollar un clúster. Pero se debe tomar la realidad país, es decir no se puede comparar un país como Chile con Estados Unidos ya que la densidad demográfica es completamente distinta y que por lo tanto afectaría aplicar un clúster a nivel país (Espinoza Benedetti, 2003). Para este caso se aplicará un clúster a nivel de empresas del rubro manufacturero de Chile y para ello se tomarán todas las del país.

- Relación entre empresas que lo componen: el núcleo o core de un clúster se puede dar en empresas de similares características o del mismo ámbito pero de igual manera en empresas distintas pero que están claramente relacionadas e interconectadas a través del comercio y buscan en conjunto soluciones a diversas problemáticas que las atañen a ellas (Espinoza Benedetti, 2003).

- Longitud: Este concepto hace referencia a que tan disperso y diversificado está el clúster, puesto que uno muy disperso alteraría las condiciones para buscar similitudes dentro del conjunto (Espinoza Benedetti, 2003).

- Nivel de análisis: El enfoque planteado en un clúster depende de la finalidad que se pretende con el mismo, ya que las conclusiones obtenidas pueden ser a nivel micro, meso o macroeconómico. Por esta razón la información recopilada debe facilitar la creación de relaciones dentro de un sector del clúster a fin de que el análisis sea efectivo (Espinoza Benedetti, 2003).

Porter en su libro Çlusters and the new economics of competition"de 1980, muestra un ejemplo de un clúster sobre la industria del cuero más premium en Italia.

Figura 8: Clúster del cuero Italiano para el mercado de la moda.

Nota: La figura muestra el clúster del cuero Italiano utilizado para artículo relacionados con la moda. Extraído del libro escrito por Porter et al. (1998)

En la Figura 8 se puede ver de forma clara como se construye un clúster con empresas de similares características, en este caso se construye un clúster que parte por la industria del cuero pero que se va abriendo a nuevas relaciones dependiendo del tipo de producto que utilicen el cuero como materia prima pero a la vez se tiende a relacionar con el material sintético. A esto se refiere que un clúster se construye basándose en la geografía pero que después se van descubriendo relaciones entre sí hasta conformar un mapeo más completo con empresas de similares características.

3.4.1. Aplicación de la metodología de clustering en el País Vasco

Un caso muy particular sobre la utilización de la metodología clustering para buscar un beneficio mutuo entre empresas que cumplen con las condiciones para crear un clúster es lo echo en el País Vasco, España, por María José Aranguren Querejeta.

El País Vasco en la década de los ochentas, pasaba por una situación de recesión producto de la pérdida de ventajas competitivas económicas a nivel mundial; por tanto, los diferentes sectores industriales de ese entonces competían con sus precios, a raíz de esto, era necesario buscar ventajas competitivas que fueran sostenibles dado el contexto internacional en desarrollo (Aranguren Querejeta, 2010).

Todo este contexto provocó que las autoridades pertinentes del País Vasco contrataran los servicios de Monitor Company para realizar un estudio que analizara la situación competitiva que presentaba el País Vasco y la potencial competitividad que podía presentar este en el futuro; es así como esta consultora que era liderada por Michael Porter elaboró un reporte que en su primera fase identifico 50 sectores de la industria que presentaban un potencial

para competir en los mercados mundiales, y mediante este reporte fue capaz de evaluar cual era el potencial de estos sectores para competir de forma internacional (Aranguren Querejeta, 2010).

En la segunda fase del trabajo realizado por Porter, los 50 sectores que se identificaron fueron agrupados usando la metodología de clustering en nueve clústeres y esto además llevo hacia una reflexión sobre las posibilidades de mejorar su potencial competitivo. "El Gobierno Vasco asumió el papel dinamizador del proceso e impulsor de los clústeres finalmente seleccionados" (Aranguren Querejeta, 2010) . Es así como basado en las recomendaciones de Monitor Company, el gobierno Vasco estableció el programa de competitividad en el marco de la política industrial presente entre los años 1991 y 1995. (Aranguren Querejeta, 2010).

Este programa impulsado por el gobierno Vasco consistía en una financiación de grupos de trabajo formados por las empresas e instituciones que componían los nueve clústeres prioritarios planteados por Porter; estos clústeres eran: "electrodomésticos, máquina-herramienta, acero de valor añadido, Puerto de Bilbao, aeronáutica, papel, componentes de automoción, turismo y alimentos" (Aranguren Querejeta, 2010). Es así como los diferentes grupos de trabajos de estos clústeres trabajaron en propuestas de mejora prioritaria y propuestas de acción para cada clúster según correspondiera.

Esta dinámica planteada por el gobierno Vasco llevo a la conformación de múltiples asociaciones de clústeres entre los años 1995 y 2000 que buscaban la mejora constante en la competitividad de las organizaciones y asumían un papel dinamizador dentro del clúster (Aranguren Querejeta, 2010).

En el Cuadro 5 se pueden observar las Asociaciones de Clústeres que existen a la fecha de publicación de este articulo

Tabla 5: Participación en el Empleo de los Distintos Tipos de Establecimientos en los Diversos Sectores Productivos %).

Clúster	Año de creación	Asociación de clúster	Nº de socios
Máquina - Herramientas	1992	AFM	94
Electrodomésticos	1992	ACEDE	11
Automoción	1993	ACICAE	90
Medio ambiente	1995	ACLIMA	93
Puerto de Bilbao	1995	UNIPORT BILBAO	151
Telecomunicaciones	1996	GAIA	238
Energía	1996	Clúster de energía	76
Aeronáutica	1997	HEGAN	36
Sector marítimo	1997	Foro marítimo	192
Papel	1998	Clúster de papel	20
Audiovisual	2004	EIKEN	54
Transporte y logística	2005	Clústeril	88

Posteriormente en el año 2000 el gobierno Vasco inicio un proceso de reflexión sobre la política de clúster, a raíz de este proceso se evaluó la efectividad de los mismos y se tomó la decisión continuar impulsando esta política. Luego en el año 2009 el País Vasco realizo un

nuevo impulso a la política clúster, indicando que aun existían algunas actividades económicas importantes que podían integrarse a los clústeres; a raíz de este impuso se llevó a la creación de cuatro nuevas asociaciones de clústeres que fueron la de fundición, artes gráficas, hábitat y equipamiento de interiores y como ultimo la de biociencias (Aranguren Querejeta, 2010).

Como se pudo observar en este planteamiento, se puede observar como la metodología de clustering se puede asociar a diferentes sectores, pero en este proyecto de memoria se utilizará la aplicación de estos en la minería de datos que funciona de manera similar que en los otros sectores.

3.5. Minería de datos

La minería de datos como concepto "puede definirse inicialmente como un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al examinar grandes cantidades de datos" (Pérez López and Santin González, 2007). Por tanto, en el mundo actual donde contar con información precisa es relevante a la hora de tomar decisiones, esta temática va tomando más relevancia conforme avanza el tiempo.

"La revolución digital ha hecho posible que la información digitalizada sea fácil de capturar, procesar, almacenar, distribuir, y transmitir (Riquelme Santos et al., 2006), actualmente el mundo se encuentra en una época donde todo esta digitalizado y la información es cada vez más voluminosa, ya que la cantidad de datos cada vez crece de forma progresiva y se hace difícil el análisis de estos.

Es así como en la actualidad se van creando bases de datos enormes y "descubrir conocimiento de este enorme volumen de datos es un reto en sí mismo" (Riquelme Santos et al., 2006), ahí radica la importancia de la minería de datos, ya que esta busca darle sentido a ese gran volumen de datos para transformarlos en información más acotada y precisa para quien haga uso de ésta.

Hoy en día los datos almacenados en bases de datos no solo están limitados a números o caracteres, "el avance de la tecnología para la gestión de bases de datos hace posible integrar diferentes tipos de datos, tales como imagen, video, texto, y otros datos numéricos, en una base de datos sencilla, facilitando el procesamiento multimedia." (Riquelme Santos et al., 2006); por eso la minería de datos es un área que cobra cada vez más relevancia en el mundo actual ya que busca darle un sentido lógico a este volumen de datos y tipo de información estructurada en las bases de datos por medio de técnicas estadísticas o herramientas de gestión que existen para realizar análisis de datos numéricos o caracteres.

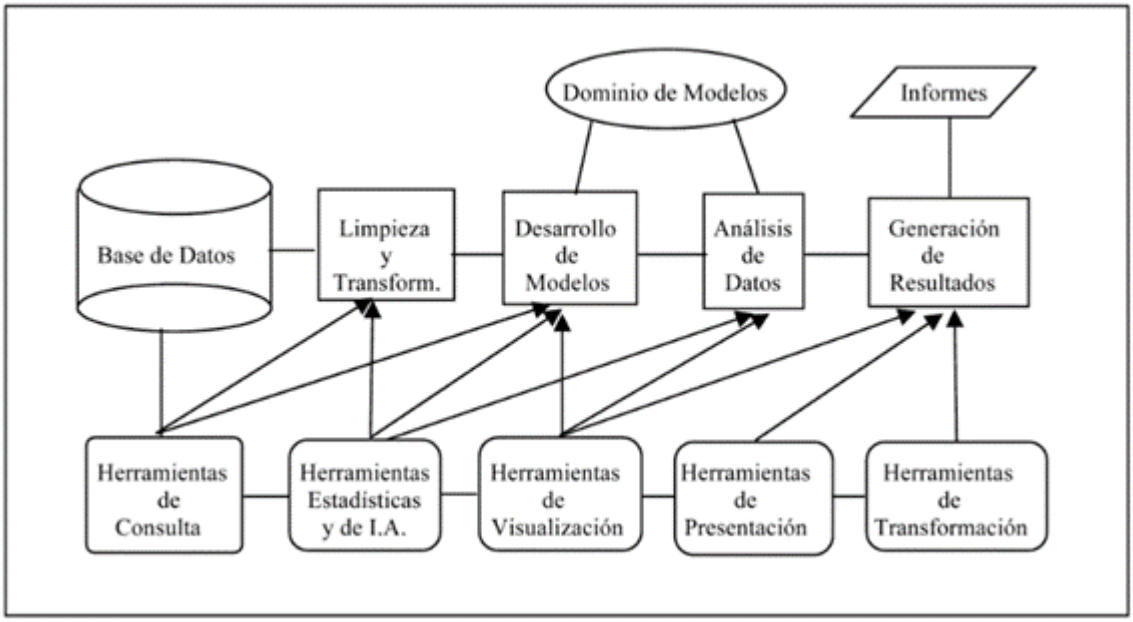
De hecho, la invención de internet como herramienta de uso común y su creciente demanda en las personas ha provocado que se desarrollen técnicas de minería de datos cada vez más sofisticadas que sean capaces de analizar datos cada vez más heterogéneos, puesto que las bases de datos actualmente pueden contener imágenes, vídeos e incluso GIFs. "El desarrollo de la tecnología de minería de datos avanzada continuará siendo una importante área de estudio, y en consecuencia se espera gastar muchos recursos en esta área de desarrollo en los próximos años" (Riquelme Santos et al., 2006); todo esto debido a que la demanda por obtener una interpretación de estos datos irá en aumento en todo el mundo.

Uno de los métodos más utilizados por la minería de datos es KDD (Knowledge Discovery in Databases), que nace en el año 1989 y se refiere al proceso donde se extrae conocimiento

de una base de datos para ser utilizado con fin o propósito establecido previamente (Riquelme Santos et al., 2006).

En la Figura 9 el proceso de KDD.

Figura 9: Proceso KDD.

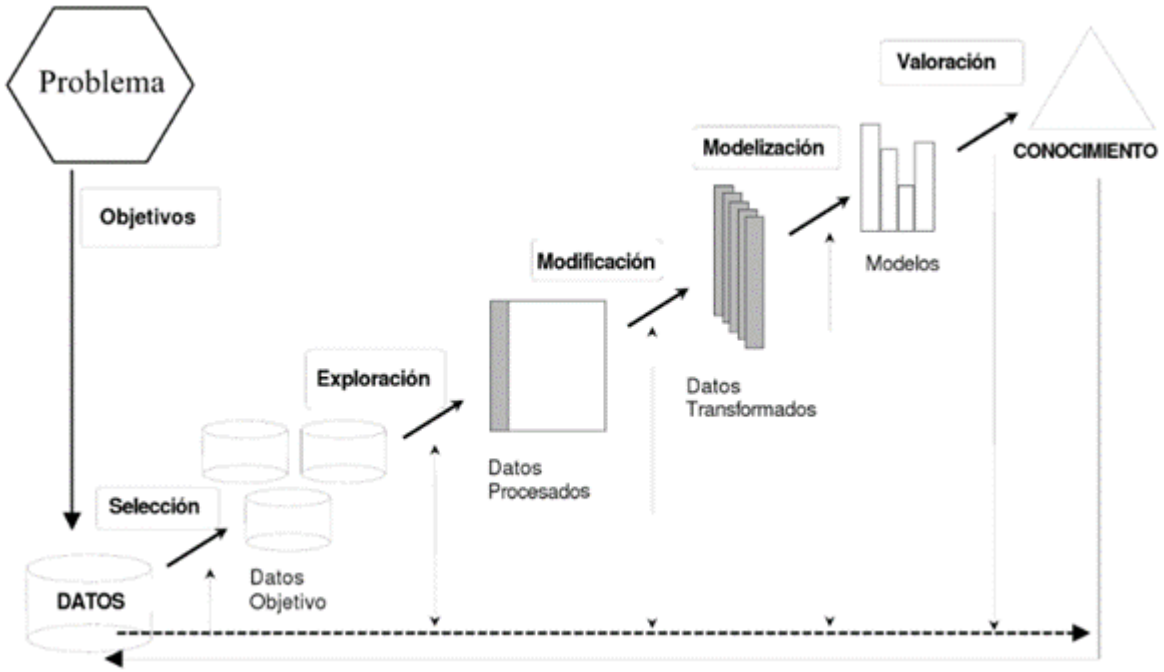


Nota: Adaptado del libro "Minería de Datos: Técnicas y Herramientas". Escrito por (Riquelme Santos et al., 2006)

Como se observa en la Figura 9, este proceso consta de varias etapas que se deben seguir, una de ellas es la preparación de datos donde se selecciona, limpia y se transforman los datos según el requerimiento, luego se sigue con el procesos de exploración y auditoría donde se realiza un análisis de datos y un desarrollo de modelos para trabajar los datos seleccionados con anterioridad, una vez los datos fueron seleccionados y analizado se puede realizar la generación de resultados para así presentar un informe a quien lo requiera (Riquelme Santos et al., 2006).

Otra forma de hacer minería de datos es aplicando el concepto que propone SAS Institute que define el concepto de data mining como el proceso de seleccionar, explorar, modificar, modelizar y valorar los grandes volúmenes de datos que son analizados con el objetivo de obtener ventajas comparativas a través del descubrimiento de patrones que permitan diferenciarse de los competidores(Riquelme Santos et al., 2006). Este proceso es resumido con las siglas SEMMA y se muestra en la Figura 10.

Figura 10: Proceso SEMMA.



Nota: Adaptado del libro "Minería de Datos: Técnicas y Herramientas". Escrito por (Riquelme Santos et al., 2006)

Como se puede observar en la Figura 10, tanto el proceso presentado por SAS Institute como el de las fases del KDD de la Figura 9 presentan múltiples semejanzas, pero son utilizados en distintos softwares de minería de datos, ya que SEMMA es utilizado en el software "Enterprise Miner" (Riquelme Santos et al., 2006).

Otra metodología para hacer minería de datos es la planteada por SPSS y considera que las seis fases que conforman el proceso conocido como minería de datos comienzan con la comprensión del negocio, la comprensión de los datos analizados, la preparación de los datos para ser utilizados, el modelado, la evaluación y la posterior utilización del modelo antes planteado; este modelo planteado por SPSS se utiliza en el software "Clementine", un software relacionado con la minería de datos (Riquelme Santos et al., 2006).

En materia relacionada con nuestra área de estudio durante la carrera, las finanzas, el mayor desafío que enfrentan hoy las empresas es mantener una cartera de clientes que se mantenga y sea lucrativa de cara a lo que toda empresa desea; es mediante la minería de datos que las empresas pueden tener un conocimiento adquirido sobre los clientes, pueden ser capaz de interpretar sus objetivos, expectativas y deseos (Braga et al., 2009).

Como dice el articulo "minería de datos: conceptos y tendencias" (Riquelme Santos et al., 2006),en la actualidad, la minería de datos tiene diversas aplicaciones. Se puede aplicar en prácticamente todas las actividades humanas que generen datos:

- Comercio y Banca: segmentación de clientes, previsión de ventas y análisis de riesgo.
- Medicina y Farmacia: diagnóstico de enfermedades y la efectividad de los tratamientos.

- Seguridad y Detección de Fraude: reconocimiento facial, identificaciones biométricas, accesos a redes permitidos, etc.
- Recuperación de Información no Numérica: minería de texto, minería web, búsqueda e identificación de imagen, vídeo, voz y texto de bases de datos multimedia.
- Astronomía: identificación de nuevas estrellas y galaxias.
- Ciencias Ambientales: identificación de modelos de funcionamiento de ecosistemas naturales y/o artificiales (ejemplo, plantas depuradoras de aguas residuales) para mejorar su observación, gestión y/o control.
- Geología, Minería, Agricultura y Pesca: identificación de áreas de uso para distintos cultivos o de pesca o de explotación minera en bases de datos de imágenes de satélites.
- Ciencias Sociales: Estudio de los flujos de la opinión pública. Planificación de las Ciudades: identificar barrios con conflicto en función de los valores socio demográficos.

3.6. Clustering en minería de datos

El clustering es una de las principales técnicas de modelado de la minería de datos la cual consiste en dividir la información en grupos diferentes, internamente los miembros de cada grupo son muy similares unos de otros y disimiles respecto a los miembros de los otros grupos. Los grupos o clústeres pueden ser usados para clasificar nuevos datos (Mamani Rodríguez et al., 2017).

Otro aspecto que debemos tener en cuenta son las distintas aplicaciones donde se utilizan los análisis de clúster, los cuales pueden ser:

- Big data.
- Clústering empresarial.
- Data science
- Clúster geográfico
- Clúster sectorial
- Clúster horizontal
- Clúster vertical
- Clústers que evalúan el sistema de valor
- Clústers de capital humano
- Robótica
- Biología

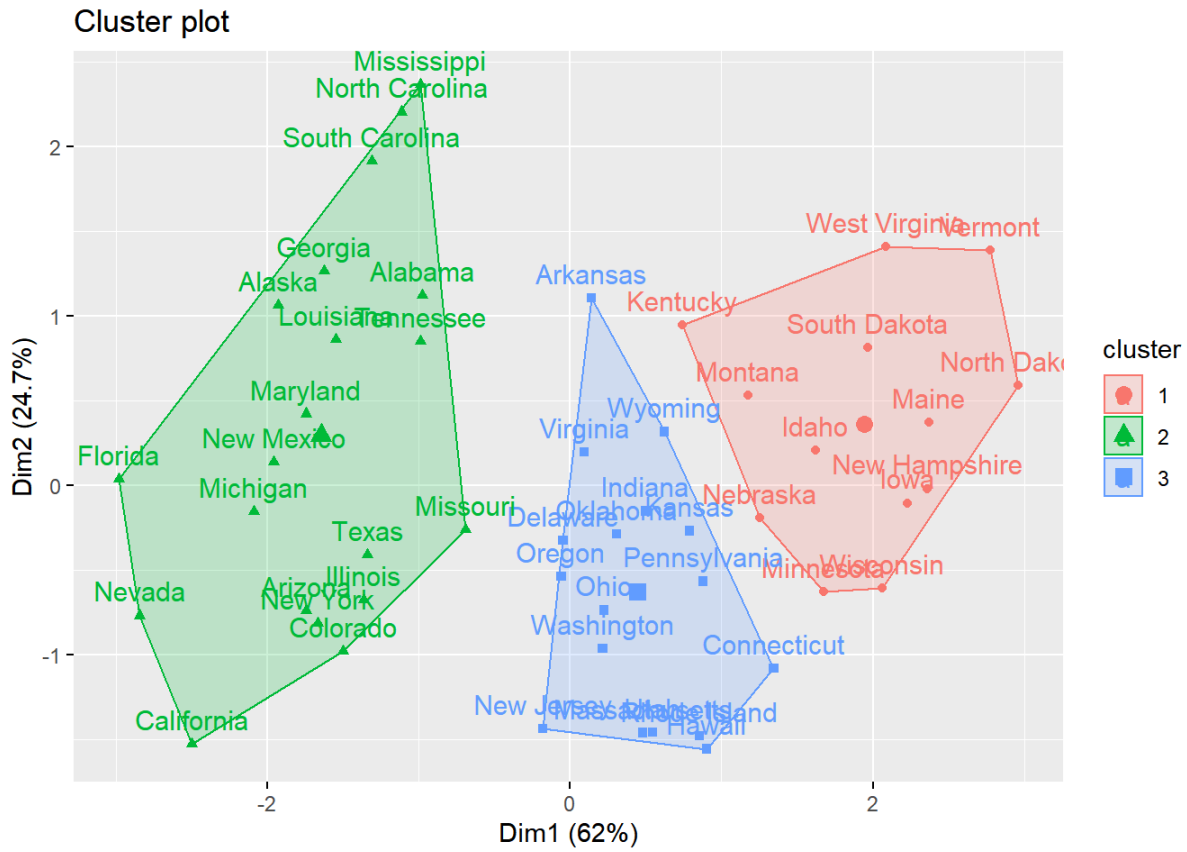
- Medicina

Los métodos de clustering están enmarcados dentro de las técnicas de machine learning y de aprendizaje no supervisado; aunque existen multitud de métodos, los dos más conocidos son el K means y el clúster jerárquico, pero también se deben considerar otros parámetros que serán considerados en el desarrollo de este informe, tales como la distancia entre clusters, técnicas de normalización, formas gráficas, entre otros.

3.6.1. Algoritmos de clústering

Método K-Means: K-means es una metodología que tiene como objetivo crear una partición de un conjunto de un cierto número de observaciones en k grupos, cada grupo está representado por el promedio de los puntos que lo componen, el representante de cada grupo se denomina centroide; la cantidad de grupos a descubrir, k, es un parámetro que se debe fijar de forma previa. Este método de clústering comienza con k centroides ubicados de forma aleatoria, y asigna cada observación al centroide más cercano. Después de asignarlos, los centroides se mueven a la ubicación promedio de todos los datos asignados a él, y se vuelven a reasignar los puntos de acuerdo a las nuevas posiciones de los centroides (MacQueen, 1967).

Figura 11: Ejemplo del método KMeans.



Nota: La figura muestra un ejemplo del método KMeans utilizando 3 clústeres para agrupar ciudades de Estados Unidos con la información de delitos.. Extraído de Alboukadel (2017)

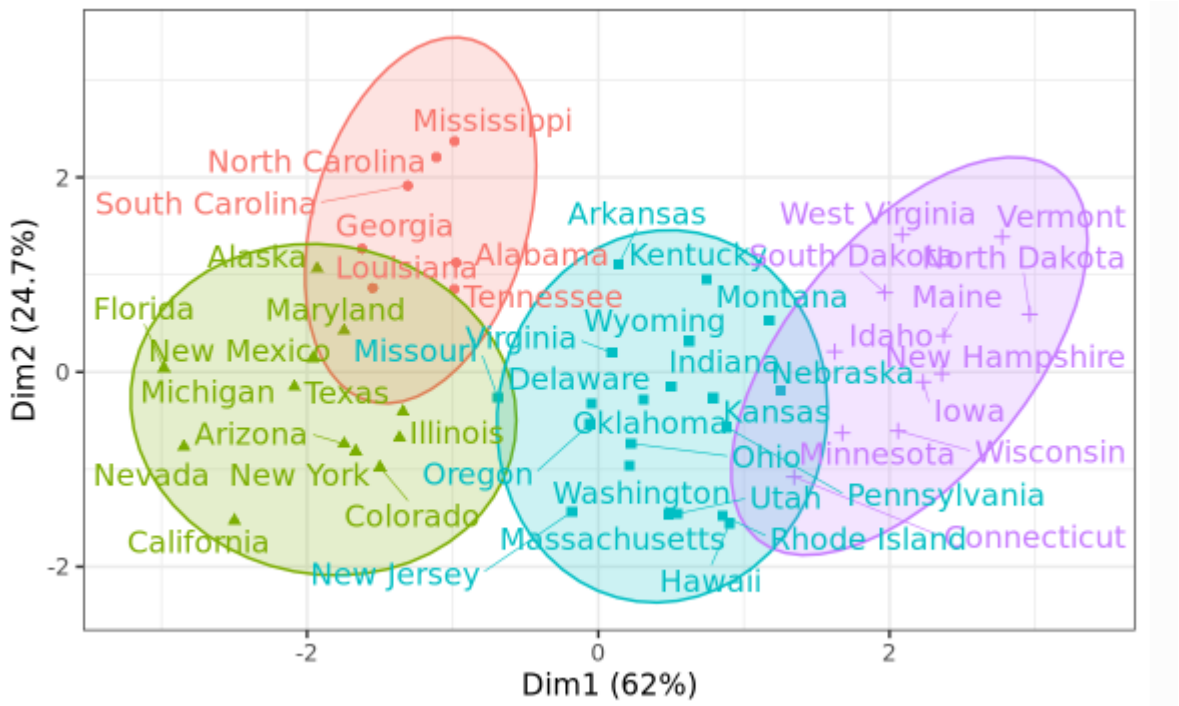
En la Figura 11 observamos información sobre clusters que representan los delitos (asaltos, asesinatos y secuestros) junto a los 50 estados de USA. Con esta representación gráfica se pretende estudiar si existe una agrupación subyacente de los estados empleando k-medias.

Método PAM (Partitioning Around Medoids): K-medoids o PAM, es muy similar a K-means ya que ambos agrupan las observaciones en K clústeres, donde K es un valor que fue preestablecido por quien realiza el análisis, en este caso utilizando las diferentes opciones que existen para obtener la cantidad de clústeres que se deben ejecutar de acuerdo con el análisis de los datos; pero PAM utiliza un medoide como referencia central del clúster. (Amat, 2017).

La definición más exacta del término medoide es un "Elemento dentro de un clúster cuya distancia (diferencia) promedio entre él y todos los demás elementos del mismo clúster es lo menor posible." (Amat, 2017). Por tanto, este dato puede considerarse como el más representativo dentro del clúster. Este método resulta mejor que K-means debido a que al utilizar centroides se ve menos afectado por aquellos datos que desestabilizan la varianza interna de los clústeres, los datos outliers que son datos atípicos observados dentro de los datos analizados.

Una vez aplicado el método de clustering PAM el cluster queda de la siguiente forma como se observa en la Figura 12:

Figura 12: Ejemplo del método PAM.

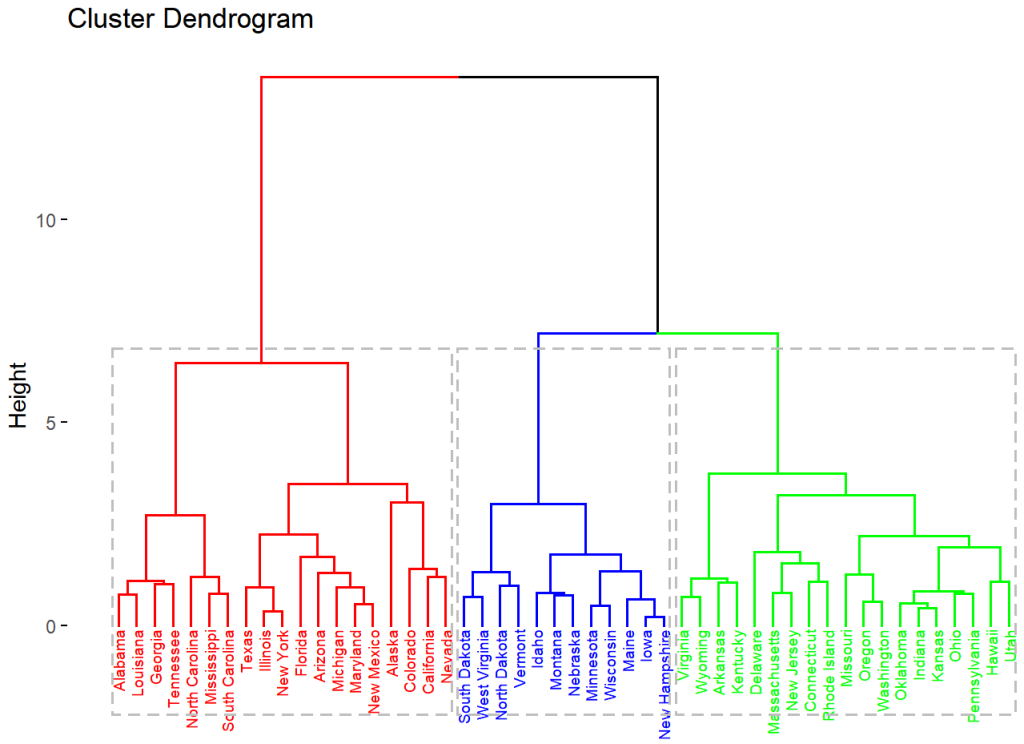


Nota: La figura muestra un ejemplo gráfico de cómo se representa el método de PAM, donde cada color indica un clúster distinto. Extraído de Amat (2017)

Método Jerárquico *Ward.d2*: Es un enfoque alternativo a la agrupación de *kmeans* que sirve para identificar grupos en el conjunto de datos donde el resultado de la agrupación es

una representación que llamamos dendrograma (Alboukadel, 2017); las observaciones pueden subdividirse en grupos cortando el dendograma a un nivel de similitud deseado (*ward.d2* , esto apunta a minimizar la varianza total dentro del grupo. En cada paso, se fusionan el par de clústeres con una distancia mínima entre los mismos; en otras palabras, se forman grupos de una manera que minimiza la pérdida asociada con cada uno, también se considera la unión de cada par de clústeres posible y se combinan los dos cuya fusión da como resultado un aumento mínimo en la pérdida de información (Amat, 2017).

Figura 13: Ejemplo del método ward.d2.



Nota: La figura muestra un ejemplo del método jerárquico, estas figuras también se conocen como dendograma. Extraído de Alboukadel (2017).

En la Figura 13 se utiliza el método jerárquico para representar los delitos (asaltos, asesinatos y secuestros) de todos las ciudades de Estados Unidos con el fin de caracterizar los grupos que crea este algoritmo.

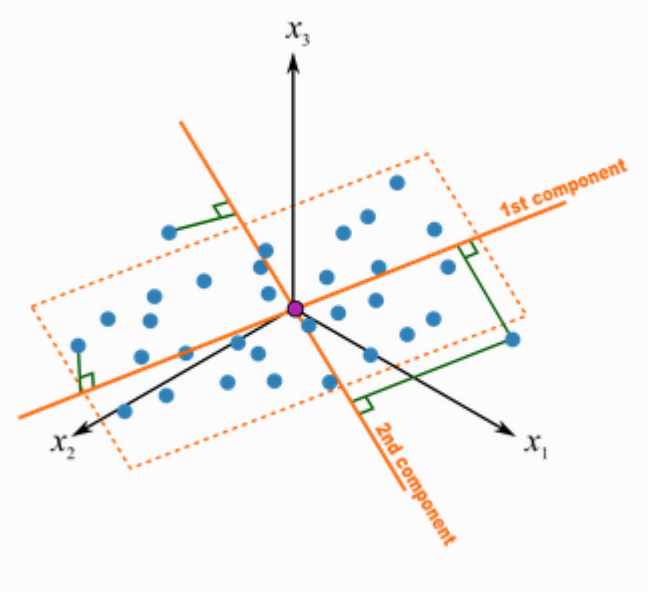
3.6.2. Parámetros de clusters

Método estadístico PCA: PCA(Principal Component Analysis) o análisis de componentes principales es una de las técnicas de aprendizaje no supervisado, las cuales suelen aplicarse como parte del análisis exploratorio de los datos; solo contamos con un número de variables de las cuales nos interesa conocer o de las que queremos extraer información, por ejemplo, sobre la existencia de subgrupos entre las variables u observaciones; PCA permite reducir los datos a un número menor de variables transformadas (Gil, 2018).

Antes de todo, se deben centrar los datos de tal manera que obtengan una media de 0, el primer componente estará representado por un vector de cargas, el segundo constará de un vector unitario de dirección; la distancia perpendicular de cada observación al plano (2 componentes) o hiperplano (3 o más componentes) será el error residual, es decir, los se escalan a una varianza unitaria para eliminar el efecto de las distintas unidades en las que puedan estar medidos los datos. Se traza la línea que mejor se ajusta a los datos centrados y escalados.

La Figura 14, muestra gráficamente el resultado de la aplicación del algoritmo PCA a un conjunto de datos:

Figura 14: Ejemplo de la técnica PCA.



Nota: La figura muestra un ejemplo gráfico de cómo se representa el método de PCA, donde el primer y segundo componente definen un plano que constituye una mejor representación de los datos analizados. Extraído de Gil (2018).

Promedio Silhouette: El método para validar clústeres por medio del promedio silhouette mide la calidad de un agrupamiento, es decir, nos dice si los objetos dentro de cada grupo están bien posicionados, estimando las distancias promedio entre los agrupamientos; si este promedio es alto indica que existe un buena, generalmente estos valores están entre -1 y 1 agrupación. (Alboukadel, 2017).

Para calcular el coeficiente de silueta (Si) para cada observación i se utiliza la siguiente fórmula:

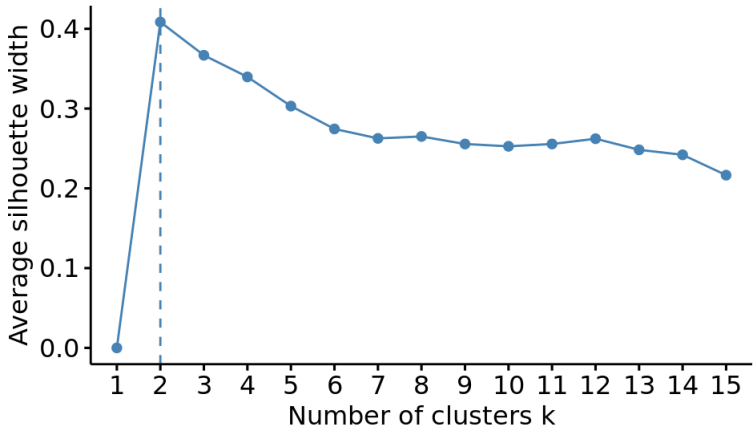
$$S_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{1}$$

Donde:

- ai es el promedio de las distancias entre la observación i y el resto de las observaciones que pertenecen a un mismo clúster; mientras más bajo sea este factor, mejor es la asignación de i al mismo.

- bi es la menor de las distancias promedio entre la observación i con el resto de los clústeres del análisis, es decir, es la distancia con el clúster más próximo.

Figura 15: Ejemplo del método Silhouette.



Nota: La figura muestra un ejemplo gráfico de cómo se establece el método del promedio Silhouette . Extraído de Amat (2017)

En la Figura 15, y tal como se explicó, la cantidad óptima de clústeres es 2, esto se puede observar ya que la media del coeficiente de silouette es la más alta (0,4) y además cumple con la condición de ser cercano a 1.

En síntesis, el método del promedio silhouette considera como cantidad efectiva de clústeres aquel donde la media del coeficiente de silhouette se maximiza entre -1 y 1.

Estadístico Hopkins: Es un método para validar clústeres que permite evaluar la tendencia a formar clúster que tienen el conjunto de datos que se está analizando, este método utiliza el cálculo de probabilidad con el fin de ver si estos datos proceden de una distribución uniforme, es decir, analiza la distribución espacial que tienen los datos que se están analizando con el fin de determinar si es posible hacer o no el o los clústeres. (Alboukadel, 2017). La forma de calcular este estadístico es con la siguiente fórmula:

$$H = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i} \tag{2}$$

Y se utiliza de la siguiente forma:

- Primero se extrae una muestra uniforme de n (p1,. . . ,pn) observaciones de los datos analizados.
- Para cada observación pi seleccionada, encontrar la observación más cercana pj y se debe calcular la distancia entre ambas, xi= dist(pi,pj)

- Simular un conjunto de datos cuyo tamaño sea n (q1,. . . ,qn) extraídos de una distribución uniforme que tengan la misma variación que los datos originales
- Para cada observación simulada qi, se debe encontrar la observación más cercana qj y se debe calcular una distancia entre ambas, yi = dist(qi,qj)
- Luego se debe calcular el estadístico de Hopkins (H) con la fórmula anterior y teniendo en cuenta lo calculado para xi e yi respectivamente ya que se debe calcular la media de las distancias de las observaciones más cercanas en el set de datos que se simuló, y se debe dividir por la suma de las medias de las distancias más cercanas del set de datos original (xi) y el simulado (yi)

Si los datos obtenidos en H son en torno a 0,5 o mayor, nos indican que los datos están distribuidos uniformemente y por tanto no tiene sentido aplicar clustering. En cambio si los datos tienden a 0, existen evidencias de que los datos tienen agrupaciones entre sí en caso de aplicar clustering correctamente y los grupos resultantes serán reales.(Alboukadel, 2017).

Para efectos de esta investigación, se utilizó la librería factoextra del lenguaje R, la cual utiliza los parámetros de forma diferente a lo planteado por Alboukadel Kassambara en su libro "Practical Guide To Cluster Anaysis in R". Esta librería calcula el estadístico de Hopkins de igual manera, pero en forma posterior normaliza los datos entre 0 y 1, los valores cercanos a 1 indicarán que el dataset presenta una alta tendencia a formar clusters.

Distancia Euclídea La distancia euclídea entre dos puntos p y q se define como la longitud del segmento que une ambos puntos. En coordenadas cartesianas, la distancia euclídea se calcula empleando el teorema de Pitágoras.(Alboukadel, 2017) Por ejemplo, en un espacio de dos dimensiones, en el que cada punto está definido por las coordenadas (x,y), la distancia euclídea entre p y q viene dada por la ecuación:

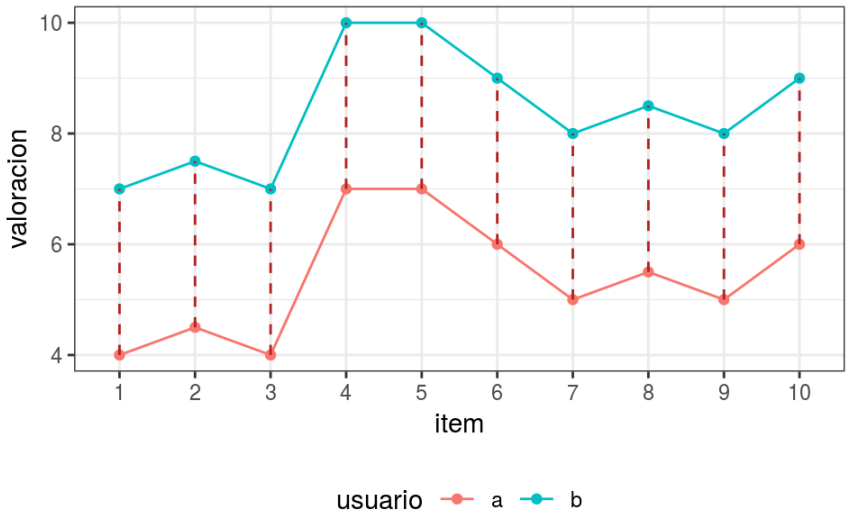
$$d_{euc}(p,q) = \sqrt{(x_p - x_q)^2 + (y_p - y_q)^2} \tag{3}$$

Esta ecuación puede generalizarse para un espacio euclídeo n-dimensional donde cada punto está definido por un vector de n coordenadas:

$$d_{euc}(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + ... + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2} \tag{4}$$

La Figura 16 muestra el perfil de dos usuarios definidos por las valoraciones que han hecho de 10 ítems (espacio con 10 dimensiones).

Figura 16: Ejemplo de Distancia Euclídea.



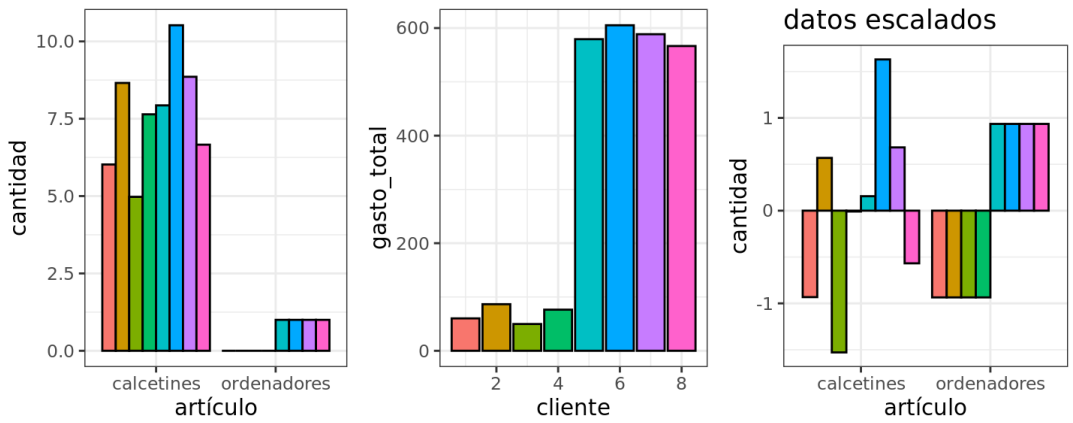
Nota: La figura muestra un ejemplo perfil de dos usuarios definidos por las valoraciones que han hecho de 10 ítems (espacio con 10 dimensiones). Extraído de Amat (2017)

La distancia euclídea entre las dos observaciones equivale a la raíz cuadrada de la suma de las longitudes de los segmentos rojos que unen cada par de puntos. Tiene en cuenta, por lo tanto, el desplazamiento individual de cada una de las variables (Alboukadel, 2017).

3.6.3. Normalización de datos

Es un proceso de procesamiento previo en los datos con el objetivo de ajustar la escala o características para tener una medida estándar, esto en machine learning también se conoce como escalamiento. Para ilustrar esto con un ejemplo, una tienda online quiere clasificar a los compradores en función de los artículos que adquieren, para esta caso, calcetines y ordenadores. La Figura 17 muestra el número de artículos comprados por 8 clientes a lo largo de un año, junto con el gasto total de cada uno.

Figura 17: Ejemplo de aplicaciones de clústeres en minería de datos.



Nota: La figura muestra un ejemplo de cómo funciona la normalización de datos en el proceso de clustering de datos.Escalando y centrando las variables se consigue igualar la influencia de calcetines y ordenadores. Extraído de Amat (2017)

Si se intenta agrupar a los clientes por el número de artículos comprados, dado que los calcetines se compran con mucha más frecuencia que los ordenadores, van a tener más peso al crear los clusters. Por el contrario, si la agrupación se hace en base al gasto total de los clientes, como los ordenadores son mucho más caros, van a determinar en gran medida la clasificación (Amat, 2017).

Escalar datos (*scale*) es una función que centra y escala las columnas de una matriz de datos; el parámetro del centro toma un vector numérico similar o un valor lógico, si se proporciona un vector numérico, entonces a cada columna de la matriz se le resta el valor correspondiente desde el centro, si el valor es verdadero, las medias de las columnas de la matriz se restan de sus columnas correspondientes; esta función considera un vector numérico similar o un valor lógico, cuando se le proporciona un vector de tipo numérico, cada columna de la matriz se divide por el valor correspondiente de *scale*. Si el valor lógico se proporciona parámetro validados, luego las columnas centradas de la matriz se dividen por sus desviaciones estándar, y la raíz cuadrática media de lo contrario. Si es falso, no se escala en la matriz.

3.7. Metodología CRISP-DM

Para el enfoque en cómo se abordarán los datos, se decidió emplear la metodología de CRISP-DM (Cross Industry Standard Process for Data Mining) que brinda información normalizada de un proyecta que requiera análisis de datos para aplicarlo en ámbitos de negocios; este modelo considera las fases del proyecto, las respectivas tareas de este, y las relaciones existentes entre las tareas, además, desde un punto de vista de modelo de procesos, CRISP-DM ofrece un resumen del ciclo de vida de los datos que es importante para cualquier trabajo orientado en este campo.

El ciclo de vida de este modelo contempla 6 fases que muestran las dependencias más importantes y frecuentes entre sí (Hren, 2020):

- Fase 1: Business Understanding es la fase inicial donde la prioridad es obtener una comprensión clara de los objetivos del proyecto para luego definir el problema de la minería de datos, y finalmente diseñar un plan de forma preliminar para alcanzar los objetivos.

- Fase 2: Data Understanding, comienza con la colección de datos inicial para luego seguir con actividades para familiarizarse con los datos, identificar problemas en la calidad de los datos a utilizar y/o descubrir conocimientos preliminares sobre los datos o identificar subconjuntos importantes para formar una hipótesis.

- Fase 3: Data Preparation, en esta etapa se cubren todas las actividades para elaborar el conjunto definitivo de los datos (los datos que sí se utilizarán) a partir de los datos iniciales en bruto. Aquí se incluirán las tareas que inclu6yen la selección de las tablas, registros, atributos, así como cambios, transformaciones de datos o limpieza de estos.

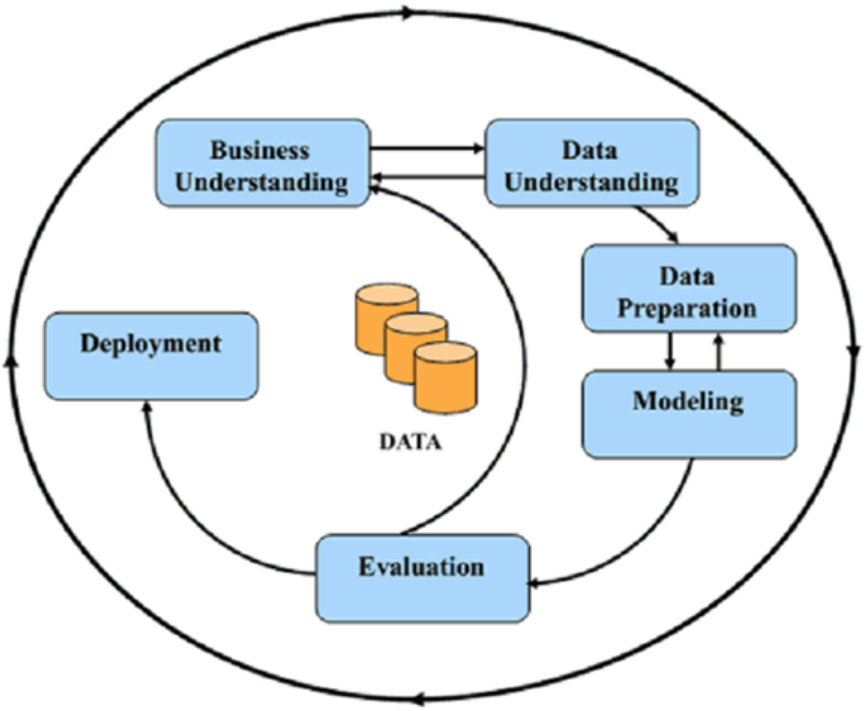
- Fase 4: Modeling, aquí se seleccionarán y aplicarán las técnicas para el modelado más pertinente en función del problema, además se calibrarán los parámetros de los datos para obtener valores óptimos. Esto dependerá de la técnica que requieran los datos, por lo tanto, es de esperar que esta fase se relacione de forma constante con la preparación de datos.

- Fase 5: Evaluation, en esta etapa ya deberán haberse construido uno o más modelos que demuestren una calidad considerable en los datos desde una perspectiva de análisis. Antes de desplegar el modelo definitivo es importante que sea evaluado a fondo y revisar los pasos previamente ejecutados. Al finalizar esta fase, se debería decidir la aplicación de los resultados del proceso de análisis de datos.

- Fase 6: Deployment, cuando el modelo está creado, no quiere decir que ese sea el final. Dependiendo de los requerimientos del caso, el objetivo deberá organizarse y presentarse de acuerdo con el cliente. Por ejemplo, puede ser desplegado un informe final o crear una automatización de procesos en análisis de datos para una compañía.

La Figura 18 muestra el funcionamiento de esta metodología.

Figura 18: Mapa conceptual CRISP-DM.



Nota: Extraído de *What is the CRISP-DM cycle?* (Hren, 2020).

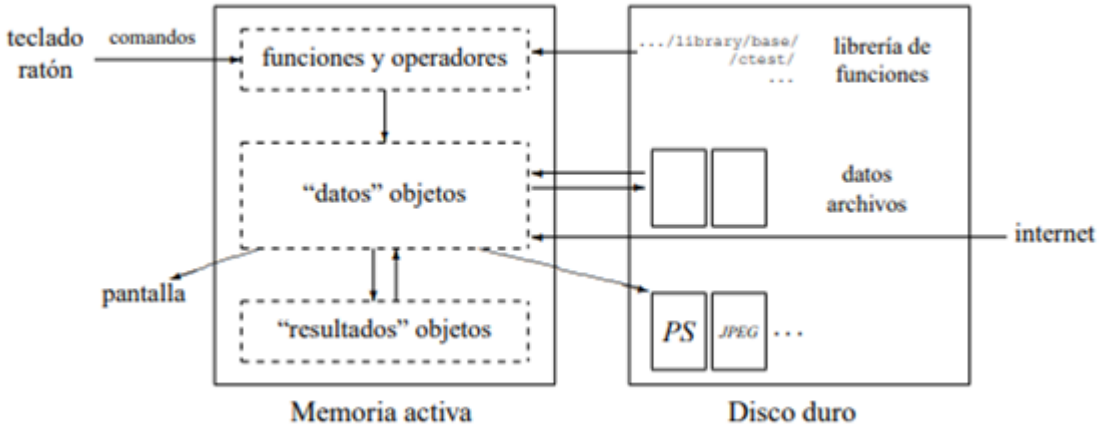
3.8. R como lenguaje de programación

R es un sistema para análisis estadístico y gráficos creado por Ross Ihaka y Robert Gentleman; R se distribuye gratuitamente bajo los términos de la GNU General Public Licence, su desarrollo y distribución son llevados a cabo por varios estadísticos conocidos como el Grupo Nuclear de Desarrollo de R; este lenguaje posee muchas funciones para análisis estadísticos y gráficos; estos últimos pueden ser visualizados de manera inmediata en su propia ventana y ser guardados en varios formatos (jpg, png, bmp, ps, pdf, emf, pictex, xfig; los formatos disponibles dependen del sistema operativo) (Paradis, 2003).

En este análisis se decidió utilizar este lenguaje gracias a la relación directa con el tema a investigar y dadas las flexibles características del mismo, se puede permitir combinar distintas funciones estadísticas que respaldan los resultados que se analizaron al termino de esta. Al mismo tiempo, se guardarán los resultados como un objeto, lo cual permite analizar sin mostrar un resultado, y con esto, podemos descomponer distintas ramas de la investigación a medida que la investigación avanza.

Para clarificar la idea se presenta la Figura 19 a continuación con un esquema del funcionamiento del lenguaje R.

Figura 19: Visión esquemática del funcionamiento de R.



Nota: La figura muestra un esquema de cómo funciona el lenguaje R desde el punto de vista técnico. Extraído de Paradis (2003)

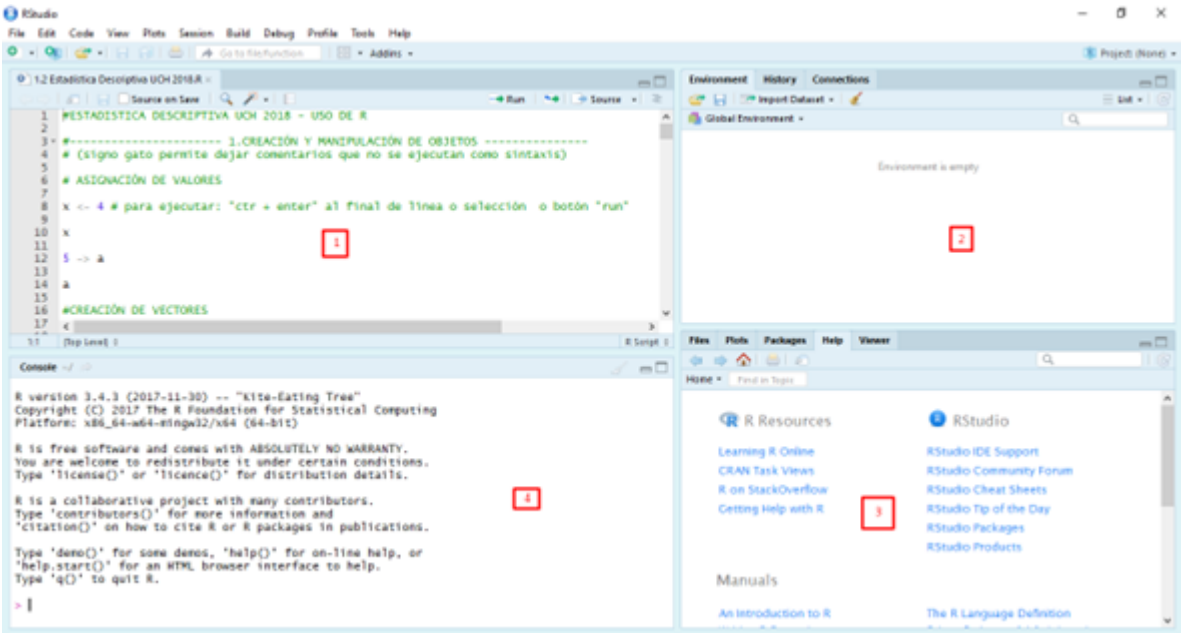
Los resultados que se obtienen desde los datos analizados son de excelente calidad en función de este análisis; cabe señalar que este lenguaje funciona sin inconvenientes en Linux, macOS y Windows.

Los inconvenientes que tiene este lenguaje no afectan al desarrollo de este proyecto, ya que las desventajas como la poca compatibilidad con otros lenguajes, no poseer soporte para gráficos en 3D o que se actualice de manera constante y que el usuario deba aprender nuevas funciones, están fuera del marco en que se trabajará con R; es decir, las principales características de este lenguaje están identificadas para desplegar la indagación (Paradis, 2003).

3.8.1. RStudio como Software

Para aprovechar al máximo las herramientas del lenguaje de R, se necesitará un software/interface que permita crear interacciones sencillas para poseer un entendimiento efectivo de los datos analizados; básicamente se trata de una máscara para visualizar el software que tiene como principales ventajas el orden y la visualización de los procesos que son llevados a cabo con R, todo de manera simultánea (Boccardo, 2019).

Figura 20: Interface de RStudio.



Nota: La figura muestra la interface general de RStudios. Extraído de Boccardo (2019).

En la Figura 20 se muestran las siguientes ventanas del software:

- Ventana (1): es el editor de sintaxis: se trata del lugar donde editamos la sintaxis para posteriormente ejecutarla. Al escribir allí no sucederá nada, a no ser que se apriete algún botón para ejecutar los comandos o la tecla ctrl+enter.
- Ventana (2): es el "entorno de trabajo" del programa: en este lugar se muestra el conjunto de datos y los "objetos" (resultados, variables, gráficos, etc.) que se almacenan al ejecutar diferentes análisis.
- Ventana (3) tiene varias sub pestañas: (i) la pestaña files permite ver el historial de archivos trabajados con el programa; (ii) la pestaña plots permite visualizar los gráficos que se generen; (iii) la pestaña packages permite ver los paquetes descargados y guardados en el disco duro así como gestionar su instalación o actualización; (iv) la ventana help permite acceder al CRAN - Comprehensive R Archive Network (siempre que se cuente con conexión a Internet), página oficial del software que ofrece diferentes recursos para el programa: manuales para el usuario, cursos on line, información general, descarga de paquetes, información de los paquetes instalados, etc. Esta última pestaña es bastante útil: empleando el motor de búsqueda se accede de manera rápida a manuales de uso de los diferentes paquetes (y sus funciones) instalados en el computador (esto no requiere conexión a Internet).7; (v) la ventana viewer muestra los resultados al construir reportes mediante funcionalidades tipo rmarkdown.
- Ventana (4): es la consola. Corresponde a lo que sería el software R en su versión básica. Allí el software ejecuta las operaciones realizadas desde el editor de sintaxis.

4. Marco Metodológico

4.1. Elaboración de hipótesis

Debido a que la investigación no busca pronosticar datos ni hechos concretos y que se trató de un estudio descriptivo, no se realizó una hipótesis. "Para un estudio descriptivo, en raras ocasiones se establece una hipótesis antes de ingresar en el ambiente o contexto y comenzar la recolección de los datos"(Grinnell Jr and Unrau, 2005). En esta investigación no se creó una hipótesis, ya que en este caso no se empleó ningún tipo de pronóstico acerca del análisis final de los datos analizados a través de la metodología clustering.

4.2. Diseño de la investigación

El tipo de investigación fue un diseño no experimental, esta se define como "estudios que se realizan sin la manipulación deliberada de variables y en los que sólo se observan los fenómenos en su ambiente natural para después analizarlos" (Hernandez Sampieri Roberto, 2010). En esta investigación no se experimentó con los datos obtenidos en la ENIA del año 2019, sino que se utilizó esta información para obtener una clasificación que llevó a determinar las similitudes que tienen las empresas utilizando clústers.

La investigación no experimental posee dos caminos a seguir según el tipo de análisis que se realice, para este trabajo se tomó el rumbo que nos entrega el diseño transeccional, estas son "investigaciones que recopilan datos en un momento único" (Hernandez Sampieri Roberto, 2010). El propósito de este análisis con estas variables en el diseño transeccional fue ver el impacto o la relación que puedan tener estas variables analizadas en el periodo o momento que se está analizando, en simples palabras es similar a lo que sucede cuando se toma una fotografía en un lugar donde ocurrió un hecho impactante como un incendio o catástrofe natural.

El carácter de esta investigación no experimental con un diseño transeccional es del tipo descriptivo que se define de la siguiente manera, "indagan la incidencia de las modalidades, categorías o niveles de una o más variables en una población, son estudios puramente descriptivos" (Hernandez Sampieri Roberto, 2010). Para este caso, en la investigación se analizaron los datos entregados por la ENIA del año 2019 y luego estos fueron clasificados utilizando clústers, y finalmente fueron clasificados según sus características que nos llevó a describir la similitud que existía entre empresas del mismo conjunto de la agrupación; de esta manera abordamos y encontramos características similares entre las empresas que poseían un elevado gasto en combustible y que como consecuencia tenían un alto impacto negativo para el medio ambiente.

4.3. Muestreo

En este proyecto no se utilizó muestreo, ya que el estudio se realizó sobre toda la población; este corresponde al total de empresas de Chile que respondieron la Encuesta Nacional Industrial Anual (ENIA) realizada Instituto Nacional de Estadísticas (INE) en el año 2019. Esta encuesta se realiza a todas las empresas del rubro manufacturero de Chile dividida en

cuestionarios que sirven para clasificar los datos que deben aportar las empresas para la confección de estadísticas; para el desarrollo de este trabajo, se utilizaron tres cuestionarios que clasifican a las empresas de forma inicial por su volumen, el primer cuestionario lo responden aquellas empresas que solo poseen una sucursal en funcionamiento, el segundo lo contestan aquellas entidades que poseen más de una sucursal en funcionamiento, el tercer formulario lo rellenan todas las empresas que posean múltiples entidades asociadas a una marca, a través de esta clasificación inicial que determina la misma encuesta se puede seccionar el tamaño de la empresa que se analizará

Los datos recolectados por la ENIA hacen referencia al periodo que va desde el 1 de enero al 31 de diciembre del año en que se aplica; Y su periodo de recolección va desde mayo del año siguiente hasta fines de febrero del año subsiguiente, por tanto, la periodicidad de la etapa de recolección es anual (Instituto Nacional de Estadísticas, 2019).

4.4. Recolección de datos

La recolección de datos se realizó mediante el análisis de una base de datos que se genera por una encuesta que realiza el INE y que está disponible en su página web oficial en formato de tipo acces (.accdb), la función de este organismo en el país es brindar estadísticas oficiales a quienes deseen utilizarlas con fines positivos, otro dato importante es que este organismo es el encargado de realizar los censos en nuestro país, por tanto el nivel de confianza que nos entrega para este trabajo es alto; la labor que realiza el INE dentro de nuestro país es de suma relevancia puesto que diferentes entidades sobre todo gubernamentales pueden implementar acciones basadas en estas estadísticas que traigan beneficios para la población, los datos obtenidos a través de los diferentes resultados que entrega este organismo no solo son utilizados por el ámbito público sino también por entidades del sector privado que mediante estos datos pueden tomar diferentes decisiones que traigan estabilidad a la misma.(Instituto Nacional de Estadísticas, 2022)

El INE como organismo entrega a través de sus encuestas muchos datos ricos en información útil para quienes la utilizan, pero para el desarrollo de esta investigación se utilizó una de ellas llamada encuesta nacional industrial anual, la cual es un instrumento de carácter estadístico censal que se realiza año a año y que tiene como principal objetivo reflejar información detallada acerca de diferentes aspectos internos de las empresas que pertenecen al rubro manufacturero de Chile. “La encuesta recaba datos de insumos utilizados, productos elaborados y factores productivos que participan en el proceso de transformación, información que permite calcular variables sectoriales de importancia, tales como Valor Agregado (VA), Consumo Intermedio (CI), producción física y valorada.” (Instituto Nacional de Estadísticas, 2022).

Como esta encuesta va dirigida al rubro manufacturero de Chile, estas estadísticas son útiles para analizar la capacidad manufacturera que tiene Chile para de esta forma establecer políticas públicas que potencien este rubro, otro punto importante que agrega valor a las estadísticas que se encuentran en esta encuesta es que estos resultados sirven para elaborar las cuentas nacionales con información de mejor calidad así como construir ciertos indicadores estadísticos-económicos que son útiles dentro de este rubro sobre todo en la toma de decisiones por parte de las gerencias de cada entidad.

4.5. Descripción de la base de datos

La base de datos obtenida de la página web del INE está disponible en formato access y está abierta a ser descargada por cualquier persona que desee utilizarla. Este archivo contiene 281 columnas y 4.255 filas.

Esta base de datos agrupa todos los datos obtenidos de las 3 encuestas que realiza el INE a las empresas que pertenecen al rubro manufacturero de Chile, que en total son 4.254; los datos se agrupan conforme a la pregunta que se hizo en la encuesta, se identifican con las letras del abecedario y un número, donde cada letra es una agrupación de datos y la acompaña un código Ejemplo: A001, B011, I067, entre otros; de la siguiente manera:

- CIIU-4: Esta columna contiene datos con una codificación asociada al rubro donde se desarrolla la empresa, por ejemplo: elaboración de productos alimenticios, elaboración y conservación de carne, etc.
- A: Estas 12 columnas aportan datos relacionados con información general de la entidad como la moneda en que llevan su contabilidad, el tipo de organización jurídica, si están conformadas por capital nacional o extranjero, entre otras.
- B: Estas 30 columnas aportan datos relacionados con los ingresos que tiene la entidad que respondió la encuesta.
- C: Estas 7 columnas entregan información relacionada con las materias primas y los materiales que son comprados para la reventa.
- D: Estas 47 columnas muestran los diferentes gastos en que incurre la entidad para su funcionamiento algunos de estos son el impuesto renta, patentes y derechos municipales, entre otros.
- E: Estas 4 columnas reflejan los traspasos que tiene cada empresa.
- F: Estas 12 columnas aportan información respecto a la valorización de inventario inicial y final que tienen las entidades.
- G: Estas 21 columnas reflejan el gasto neto en combustible que tienen las empresas y los clasifican por tipo (Leña, Parafina, Bencina, Petróleo, etc.).
- H: Estas 20 columnas entregan la información relacionada a las ventas de activos fijos e intangibles de la empresa, también se puede encontrar datos relacionado a la depreciación que registra cada entidad.
- I: estas 110 columnas aportan toda la información que se relaciona con los trabajadores de cada entidad, como remuneraciones, honorarios, entre otros.
- K: Estas 11 Columnas muestran los totales de todos los datos clasificados con anterioridad.

La base de datos resultante de la aplicación de la ENIA a las empresas del rubro manufacturero de Chile es muy grande y posee datos que son irrelevantes para nuestros análisis, por lo tanto, fue necesario depurar esta base de datos para así quitar todos aquellos datos que el equipo consideró como no útiles para la aplicación de la metodología de clustering.

Para efectos de este trabajo se utilizaron las columnas: K, G, algunas D, una sumatoria de las B, algunos datos de las A y otras que entregan información que ayuda a caracterizar la empresa, ya que la ENIA no menciona nombres ni datos específicos de la entidad como lo son: RUT, dirección, para resguardar el anonimato de las empresas.

También el INE entrega algunos datos estadísticos confeccionados a través de la base de datos de la ENIA y que permiten observar el volumen de datos con que se pretende trabajar, una muestra de esto es el Cuadro 6 que nos entrega los totales de la columna CIIU-4.

Tabla 6: Cantidad de establecimientos según columna CIIU-4.

CIIU REV.4	Nº de establecimientos
Elaboración de productos alimenticios	1105
Elaboración de bebidas	163
Fabricación de productos textiles	111
Fabricación de prendas de vestir	153
Fabricación de productos de cuero y productos conexos	34
Producción de madera y fabricación de productos de madera y corcho*	240
Fabricación de papel y de productos de papel	155
Impresión y reproducción de grabaciones	120
Fabricación de sustancias y productos químicos	220
Fabricación de productos farmacéuticos, sustancias químicas medicinales y productos botánicos de uso farmacéutico	37
Fabricación de productos de caucho y de plástico	334
Fabricación de otros productos minerales no metálicos	221
Fabricación de metales comunes	62
Fabricación de productos elaborados de metal, excepto maquinaria y equipo	381
Fabricación de productos de informática, de electrónica y de óptica	17
Fabricación de equipo eléctrico	90
Fabricación de maquinaria y equipo n.c.p.	152
Fabricación de vehículos automotores, remolques y semirremolques	36
Fabricación de otro equipo de transporte	13
Fabricación de muebles	130
Otras industrias manufactureras	29
Reparación e instalación de maquinaria y equipo	256
C** - Industrias manufactureras	195
Total	4254

Nota: * Excepto muebles; fabricación de artículos de paja y de materiales trenzables. ** Establecimientos que por secreto estadístico han sido clasificadas.

Como la ENIA la responden 3 tipos de empresa (uni establecimiento, consolidados, multi establecimientos) estos datos (CIIU-4) se encuentran codificados para cada tipo de empresa que contestó el formulario (cada rubro está 3 veces), pero el Cuadro 6 muestra una agrupación total de estos datos.

Otro dato estadístico que aporta el INE a partir de la base de datos de la ENIA es la cantidad de establecimientos manufactureros por región. El cuadro 7 aporta esta información:

Tabla 7: Número de Establecimientos por región.

Región	Total
Tarapacá	53
Antofagasta	124
Atacama	34
Coquimbo	82
Valparaíso	291
Libertador Bernardo O.	154
Maule	220
Biobío	313
Araucanía	130
Los Lagos	186
Aysén	12
Magallanes	52
Metropolitana	2439
De los Ríos	49
Arica y Parinacota	36
Ñuble	79
Total	4254

La Enia 2019, posee una columna (A001) que hace referencia al tipo de organización que llena el formulario; la encuesta proporciona 12 respuestas enumeradas para tipificar la empresa que responde de acuerdo con su composición jurídica. (Diccionario Básico Tributario Contable, 2022).

A continuación, se presentan las 12 alternativas junto a la definición que proporciona una perspectiva clara de los tipos de organizaciones jurídicas que responden y participan en la encuesta

I. Persona Natural: Es todo individuo de la especie humana, cualquiera sea su edad, sexo, estirpe o condición;

II. Sociedad de responsabilidad limitada: Es aquella en que todos los socios administran por sí mismos o por mandatarios elegidos de común acuerdo, y en que la responsabilidad de los socios está limitada al monto de sus aportes según lo determinan los estatutos;

III. Sociedad colectiva: Es aquella que se celebra entre una o más personas que prometen llevar a la caja social un determinado aporte, y una o más personas que se obligan

a administrar exclusivamente la sociedad por sí mismos o por sus delegados y en su nombre particular;

IV. Sociedad anónima cerrada: Las sociedades anónimas cerradas no pueden hacer oferta pública de sus acciones, salvo que se sometan voluntariamente a la fiscalización de la SVS;

V. Sociedad anónima abierta: son aquellas que pueden ofrecer públicamente sus acciones, para lo cual deben inscribirse en el Registro de Valores dentro de los 60 días desde su formación, quedando sujetas a la fiscalización de la Superintendencia de Valores y Seguros (SVS). Tratándose de entidades bancarias, éstas son fiscalizadas por la Superintendencia de Bancos e Instituciones Financieras;

VI. Cooperativas: tienen tratamientos tributarios especiales. En el caso de las cooperativas, éstas tienen beneficios tributarios consistentes en la exención del 100 % o 50 % de los impuestos, dependiendo del caso;

VII. Empresa pública: Es aquella en que tanto la propiedad del capital como su gestión y toma de decisiones está bajo control estatal. Uno de los principales objetivos de la empresa pública es el bien común o la producción de bienes esenciales. Hay que señalar que las organizaciones estatales que tienen autonomía financiera no constituyen empresas públicas;

VIII. Sociedad contractual minera: sociedades de capital, con un objeto especial minero y cuyo capital se divide en acciones, más similares a aquellos tipos de sociedades más comúnmente conocidos;

IX. Estatales: Es aquella en que tanto la propiedad del capital como su gestión y toma de decisiones está bajo control estatal. Uno de los principales objetivos de la empresa pública es el bien común o la producción de bienes esenciales. Hay que señalar que las organizaciones estatales que tienen autonomía financiera no constituyen empresas públicas;

X. Empresa Individual de Responsabilidad Limitada (EIRL): Persona jurídica formada exclusivamente por una persona natural, con patrimonio propio y distinto al del titular, que realiza actividades de carácter netamente comercial y están sometidas a las normas del Código de Comercio, cualquiera sea su objeto, y pudiendo realizar toda clase de operaciones civiles y comerciales, excepto las reservadas por la ley a las Sociedades Anónimas;

XI. Sociedad por acciones: es una persona jurídica creada por una o más personas mediante un acto de constitución perfeccionado de acuerdo con los preceptos siguientes, cuya participación en el capital es representada por acciones;

XII. Otra: cualquier otro tipo de constitución jurídica no incluida en las opciones anteriores.

.

Todos estos datos nos ayudan a caracterizar la base de datos con la que se llevará a cabo el proceso de clusterización.

5. Resultados

5.1. Estadístico de Hopkins

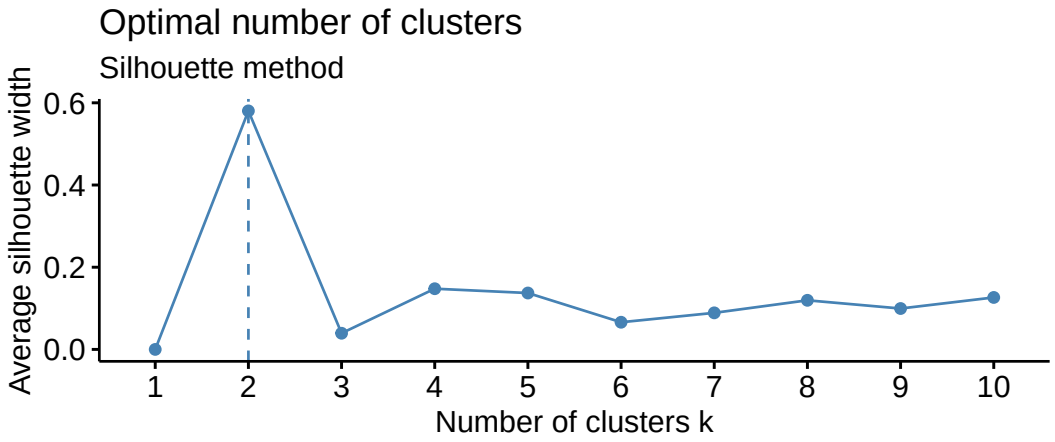
Al ejecutar el código utilizado para calcular el estadístico de Hopkins en la base de datos, el resultado fue 0.9885529, lo que indica que el dataset es altamente clusterizable. Ya que como el parámetro tiende a 1; y como se explicó en el marco teórico, según la librería de R utilizada, valores cercanos a 1 indican que los datos analizados son clusterizables.

5.2. Número óptimo de clusters

Una vez se validó la utilización de los procesos de clustering usando el estadístico de Hopkins, se empleó el promedio silhouette para determinar la cantidad de clústeres que se requerían en el análisis.

En la Figura 21 se muestra el gráfico resultante de la aplicación del promedio silhouette en nuestra base de datos; la ejecución de este método arroja que el número de óptimo clústeres que debemos utilizar en los futuros análisis es 2.

Figura 21: Número óptimo de clústers según promedio silhouette.



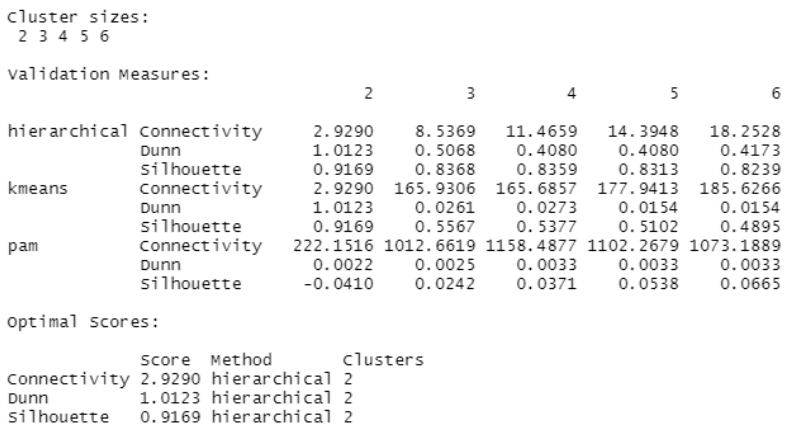
Nota: La imagen muestra un gráfico extraído del software RStudios.

5.3. Algoritmo de Clustering

Antes de ejecutar el proceso de clusterización que nos lleve a los resultados esperados, se determinó el método de clustering a utilizar. Para esto se evaluaron los algoritmos de K-means, PAM y Jerárquico, considerando un rango de 2 a 6 clusters.

El resultado del proceso de validación lo muestra la Figura 22 y es el siguiente:

Figura 22: Determinación de algoritmo de clustering a utilizar.



```
Cluster sizes:
 2 3 4 5 6

Validation Measures:
                                   2         3         4         5         6

hierarchical Connectivity     2.9290    8.5369   11.4659   14.3948   18.2528
             Dunn             1.0123    0.5068    0.4080    0.4080    0.4173
             Silhouette       0.9169    0.8368    0.8359    0.8313    0.8239
kmeans       Connectivity     2.9290  165.9306  165.6857  177.9413  185.6266
             Dunn             1.0123    0.0261    0.0273    0.0154    0.0154
             Silhouette       0.9169    0.5567    0.5377    0.5102    0.4895
pam          Connectivity   222.1516 1012.6619 1158.4877 1102.2679 1073.1889
             Dunn             0.0022    0.0025    0.0033    0.0033    0.0033
             Silhouette      -0.0410    0.0242    0.0371    0.0538    0.0665

Optimal Scores:

             Score  Method       Clusters
Connectivity 2.9290 hierarchical 2
Dunn         1.0123 hierarchical 2
Silhouette   0.9169 hierarchical 2
```

Nota: La figura fue extraída de la consola de RStudio.

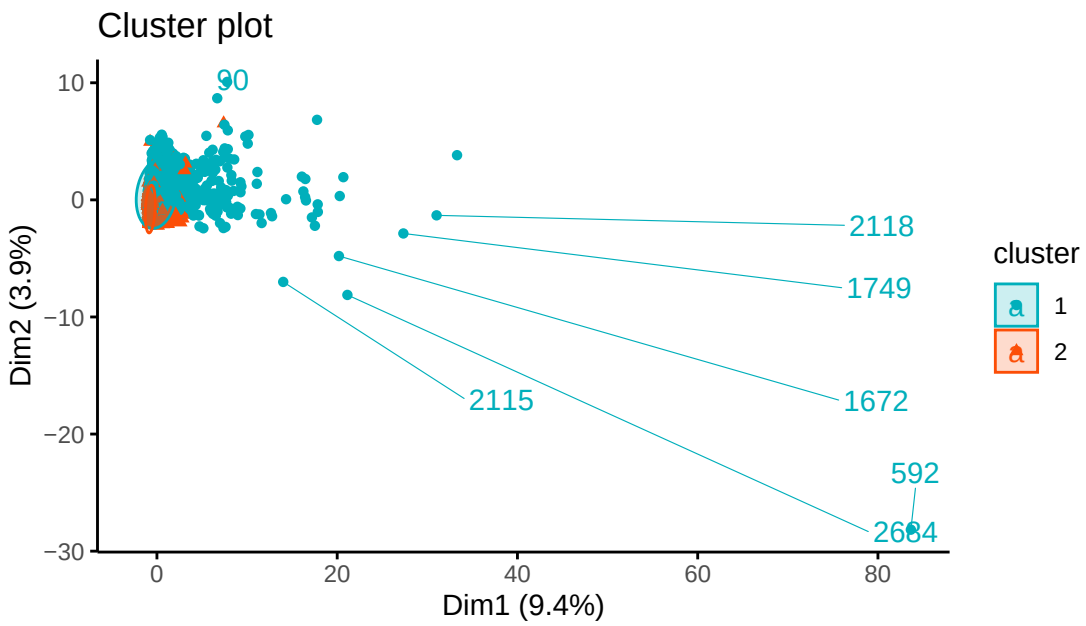
AL evaluar las distintas combinaciones, se determinó que el número óptimo de clústeres es 2, resultado congruente con lo obtenido en la Figura 22; además, al considerar los resultados desde 2 a 6 cluster, se determina que el algoritmo óptimo es el Jerárquico. Cómo el número de cluster a utilizar es 2 y para este número de cluster K-means posee los mismos resultados que cluster jerárquico, se decide utilizar en esta tesis K-means como algoritmo de clustering.

5.4. Método PAM

Pese a que el método elegido para este proyecto fue k-means, de igual manera se ejecutó el algoritmo PAM en la consola de rstudios con el fin de comparar resultados.

La Figura 28 muestra el resultado obtenido al ejecutar este método y es el siguiente:

Figura 23: Clusters PAM.



Nota: La figura el resultado de la ejecución del algoritmo de PAM.

Como se observa en la Figura 28, el resultado de aplicar el método PAM no resulta clarificador para los análisis posteriores de este proyecto, ya que los componentes de los dos clústers generados por este, se encuentran muy superpuestos para las dos dimensiones analizadas, por tanto, resulta complejo establecer diferencias y caracterizar cada clúster.

La Figura 24 muestra los medoides que arrojo la aplicación del método PAM a la base de datos:

Figura 24: Medoides de los clústers PAM.

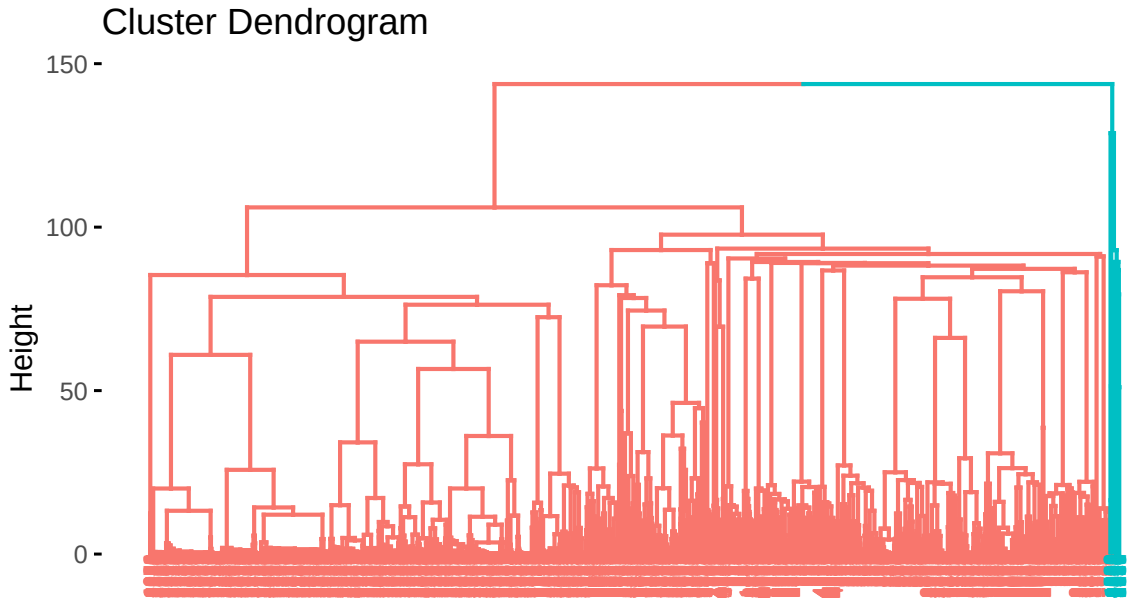
```
       REGION1    REGION2     REGION3    REGION4    REGION5    REGION6    REGION7   REGION8    REGION9   REGION10
3137 -0.112308 -0.1732546 -0.08974953 -0.1401792 -0.2709464 -0.1937838 -0.2335029 -0.281785 -0.1775257 -0.2138036
2208 -0.112308 -0.1732546 -0.08974953 -0.1401792 -0.2709464 -0.1937838 -0.2335029 -0.281785 -0.1775257 -0.2138036
       REGION11   REGION12  REGION13   REGION14   REGION15   REGION16      A0011     A0012       A0013      A0014
3137 -0.05318072 -0.1112302 0.8625442 -0.1079354 -0.0923734 -0.1375417 -0.2503762 -0.905588 -0.03429972  1.4223145
2208 -0.05318072 -0.1112302 0.8625442 -0.1079354 -0.0923734 -0.1375417 -0.2503762  1.103995 -0.03429972 -0.7029141
         A0015       A0016       A0017       A0018       A0019     A00110     A00111     A00112      A0055       D003
3137 -0.2008549 -0.03067499 -0.05091067 -0.02168539 -0.0265622 -0.159064 -0.1216002 -0.2851719 -0.2238287  0.1383607
2208 -0.2008549 -0.03067499 -0.05091067 -0.02168539 -0.0265622 -0.159064 -0.1216002 -0.2851719 -0.2238287 -0.1143837
          D004       G017       G020       G023       G026       G029       G032       G035       G038       G041
3137 -0.05193228 -0.05647561 -0.03077238 -0.1356062 -0.1331001 -0.04845501 -0.05158937 -0.1146424 -0.1032803 -0.07157238
2208 -0.05193228 -0.05647561 -0.03077238 -0.1356062 -0.1097407 -0.04845501 -0.05158937 -0.1068964 -0.1032803 -0.07157238
          G044       G047       G050       G051 SUMAINGRESOS       K001       K009       K010       K011
3137 -0.04230224 -0.0469458 -0.09983842 -0.09363189  -0.03213638  0.09199322 -0.01819859  0.06237245 -0.1524900
2208 -0.04230224 -0.0469458 -0.09983842 -0.09363189  -0.15361426 -0.21738102 -0.23397139 -0.22582873 -0.1896795
       TAMANO1    TAMANO2    TAMANO3    TAMANO4    TAMANO5    TAMANO6    TAMANO7    TAMANO8 IMPORTACION1 EXPORTACION1
3137 -0.2107808 -0.502951 -0.7006758 -0.432317 -0.3417618 -0.3104337 -0.1644193 -0.1264838   -0.5640722    0.5629861
2208 -0.2107808 -0.502951  1.4268580 -0.432317 -0.3417618 -0.3104337 -0.1644193 -0.1264838   -0.5640722    0.5629861
```

Nota: La figura fue extraída de la consola de rstudios una vez ejecutado el algoritmo PAM. Fue elaborada por el equipo.

5.5. Método Jerárquico

En el dendograma obtenido a través del software de RStudio, se observaran los clusters determinados por el algoritmo, cada color representa un cluster.

Figura 25: Clusters representados en un dendograma.



Nota: La figura muestra el resultado de la ejecución del algoritmo de *ward.d2*.

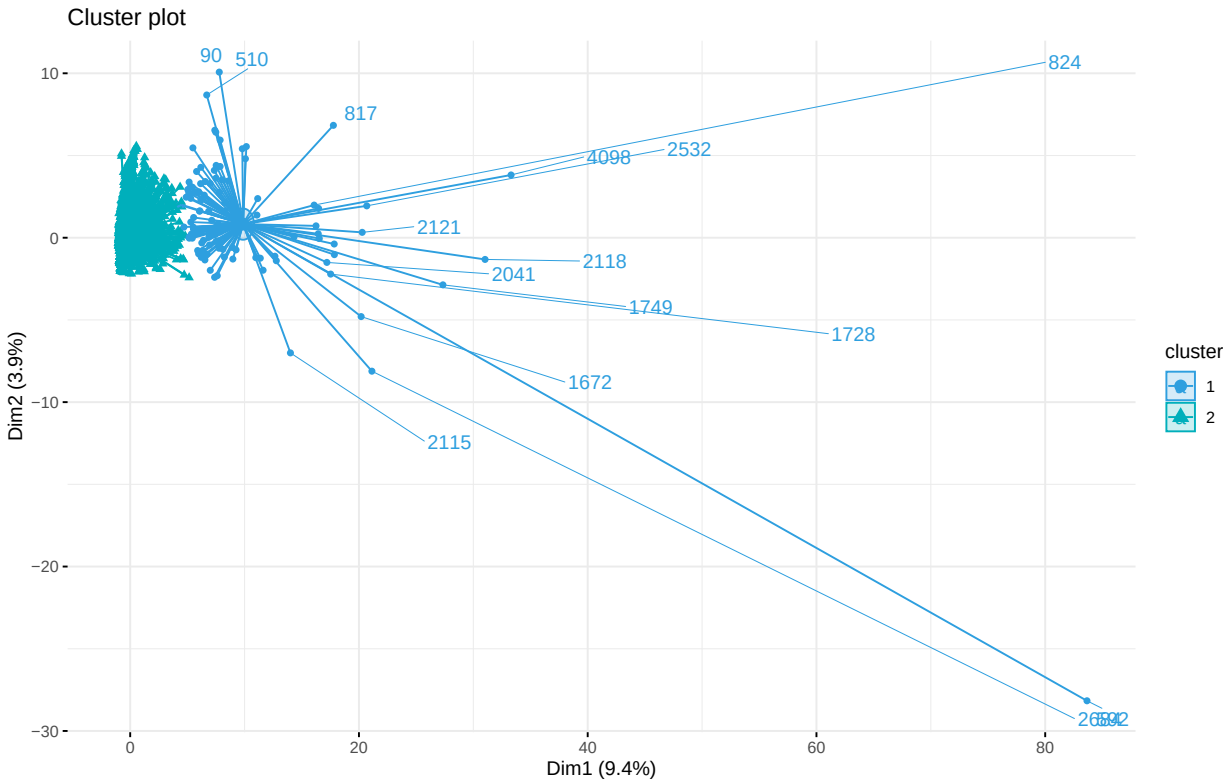
En la Figura 25 se ve cómo el dendograma comienza a dividirse de arriba a abajo en dos grupos parecidos, para luego repetir el proceso a través de la ejecución del algoritmo, volviendo a dividir en nuevos grupos; esto es, calculando constantemente las distancias entre las observaciones dentro de los grupos.

Por lo tanto, el algoritmo creó dos conglomerados donde el de color naranjo agrupa la mayor cantidad de observaciones, mientras que los datos con una caracterización más particular componen un conglomerado más pequeño, este caso el color calipso.

5.6. Método K-Means

Al ejecutar el código de k-means en el software utilizado para esta tesis, el resultado fue el siguiente:

Figura 26: Cluster K-means.



Nota: La figura el resultado de la ejecución del algoritmo de k-means.

La Figura 26 muestra el resultado de la ejecución del método k-means, donde se observan claramente los dos clústers en dos dimensiones, el primero (color celeste) posee en su interior 117 datos, mientras que el segundo (color calipso) tiene en total de 4137 datos en su interior, siendo este último el que concentra la mayoría de los datos de la base analizada.

La Figura 27 muestra los centroides arrojados por k-means para cada clúster.

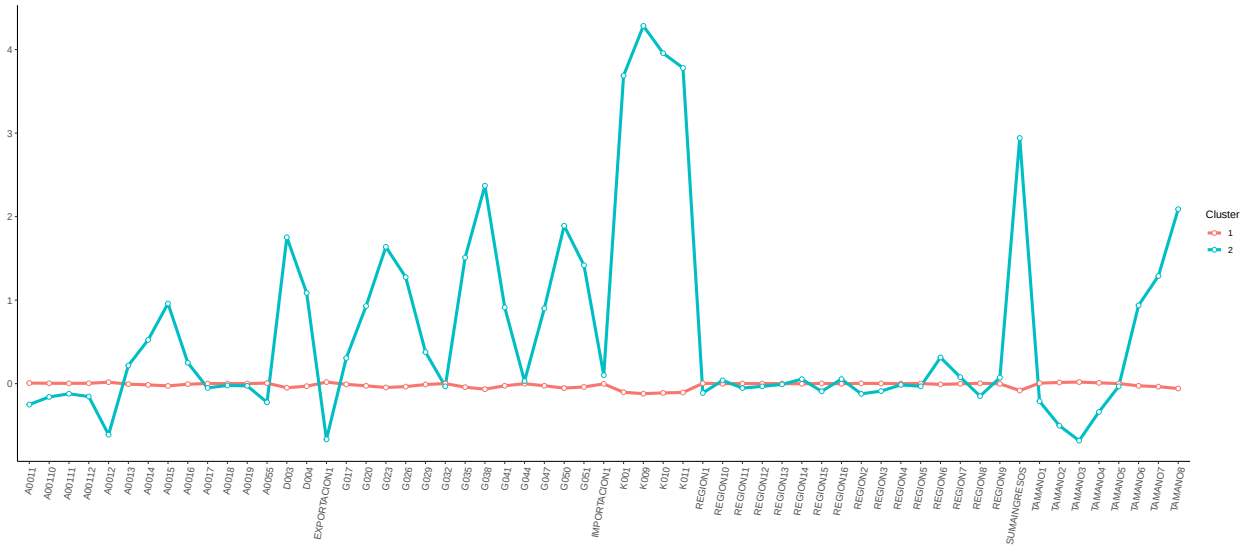
Figura 27: Centroides de los clústers k-means.

```
        PC1         PC2          PC3          PC4         PC5          PC6          PC7          PC8          PC9         PC10          PC11         PC12
1  9.8706898  0.8269898  0.317224680 -2.01152295  0.39524138  0.253495632  0.45475042 -0.70205362 -0.26147715  0.48832957  0.0268487611  0.037721470
2 -0.2791566 -0.0233884 -0.008971546  0.05688861 -0.01117796 -0.007169202 -0.01286096  0.01985503  0.00739493 -0.01381063 -0.0007593196 -0.001066815
         PC13         PC14         PC15         PC16         PC17         PC18         PC19         PC20         PC21         PC22         PC23          PC24
1 -0.129854734 -0.129844768  0.150083349  0.047973892 -0.292752722 -0.078350830 -0.38776852  0.185710939 -0.048608391  0.054876632  0.014425614  1.261231e-03
2  0.003672469  0.003672187 -0.004244562 -0.001356767  0.008279446  0.002215868  0.01096662 -0.005252159  0.001374712 -0.001551986 -0.000407976 -3.566933e-05
         PC25         PC26         PC27         PC28         PC29         PC30         PC31         PC32         PC33         PC34         PC35
1 -0.0117254117 -0.081920147 -0.0339666743 -0.144963822 -0.118635581 -0.1342757 -0.189087606 -0.039768616  0.0059591385  0.116826224 -0.28995349
2  0.0003316106  0.002316813  0.0009606239  0.004099775  0.003355176  0.0037975  0.005347655  0.001124711 -0.0001685326 -0.003304005  0.00820028
         PC36         PC37         PC38         PC39         PC40
1 -0.0254238054  0.213535239 -0.158503531 -0.113425019 -0.36058878
2  0.0007190199 -0.006039068  0.004482696  0.003207814  0.01019794
```

Nota: La figura fue extraída de la consola de rstudios una vez ejecutado el algoritmo k-means. Fue elaborada por el equipo.

La Figura 28 muestra el gráfico de los centroides determinados por k-means para cada clúster.

Figura 28: Gráfico de centroides k-means.



Nota: La figura fue extraída de la consola de RStudios una vez ejecutado el algoritmo k-means.

El gráfico de centroides entregado permitirá caracterizar a las empresas que pertenecen a cada cluster.

6. Discusión y conclusión

6.1. Discusión

Utilizando k-means como algoritmo de clustering, se determinaron 2 cluster; al analizar los centroides que originò la aplicación de este código se pudieron determinar las características de cada agrupación.

En clúster 1 se encuentran empresas de pequeño tamaño y que prácticamente no registran gastos elevados en combustible, esta agrupación de datos contiene 4.137 datos (97,24 %) siendo la mayor agrupación de datos de esta investigación.

De este clúster, en relación con la ubicación geográfica, la característica que más destaca, ubica a las empresas en la segunda (actual) región de Tarapacá junto con la tercera (actual) región de Antofagasta, ambas caracterizadas por la presencia de actividades relacionadas con la minería y termoeléctricas, por tanto, no es una sorpresa ya que estas actividades requieren de transporte en camiones y este tipo de transporte concentra una alta emanación de partículas contaminantes para el aire que por ende son dañinas para el ser humano.

Otra característica que tiene este clúster es en lo que se relaciona al tamaño, puesto que son empresas de entre 10 y 49 trabajadores, es decir, hablamos de empresas de pequeño tamaño, según lo describe el estatuto para pymes.

Otra particularidad que posee este clúster, tiene relación con el tipo de organización jurídica, siendo la de sociedades de responsabilidad limitada (A00110) una de las características principales de las empresas de este clúster, lo cual confirma la teoría de que se trata de empresas de pequeño tamaño.

En síntesis, el clúster 1 es el más grande de esta investigación y agrupa empresas que no poseen una relevancia alta en lo relacionado al gasto de combustible, es decir, las entidades dentro de este clúster no registraron altos gastos en combustibles y por tanto su impacto a nivel individual en el medio ambiente podríamos definirlo como bajo, además son empresas pequeñas o tal vez nacientes que poseen bajos ingresos y gastos elevados.

En relación con el cluster 2, este agrupa a las empresas que son contaminantes para el medio ambiente, se puede observar que posee 117 empresas (2,75 %), las características principales son las siguientes:

En relación con la ubicación geográfica, podemos concluir que existe una región que destaca por sobre el resto y esta es la región de O'Higgins, la cual posee proporcionalmente más empresas del cluster 2 que del 1; esta es una región que es caracterizada por sus actividades industriales ligadas a la minería, fabricación de alimentos y las actividades agrícolas; por tanto, estamos hablando de una región donde la utilización de combustibles es alta y por tanto el impacto generado a través de ello al medio ambiente es mayor al de otras ubicaciones geográficas del país. Otras regiones que se destacan en esta característica son las regiones del Maule, Araucanía y Los Ríos, que poseen similitudes con la región de O'Higgins en lo referido a las actividades económicas presentes en la región.

En lo referido al tipo de organización jurídica, se destacan las sociedades anónimas abiertas (A0015), esto nos da luces acerca del tamaño de las empresas que componen este clúster, ya que las sociedades anónimas abiertas generalmente son empresas de gran tamaño y presencia a nivel nacional; otro tipo de organización que destaca en esta característica analizada, son las sociedades colectivas (A0013), las sociedades anónimas cerradas (A0014) y las

cooperativas (A0016); es decir, las empresas que caracterizan este clúster son en su mayoría grandes.

Al analizar las variables (TAMANO 7 y 8 expuestos en la Tabla 8) que mayor tamaño tienen, observamos que las empresas de este clúster cuentan con más de 500 trabajadores, por tanto, se habla de empresas de gran tamaño, cuyos procesos productivos requieren de grandes volúmenes de capital humano; otro punto que apoya esta idea es que en este cluster se encuentran las empresas con el mayor valor de la variable K009, que refleja la sumatoria de los ingresos, inventarios, inversiones, entre otros; además, al contrastarlo con sumaingresos, nos indica que en este cluster se encuentran empresas con ingresos altos y que poseen gran cantidad de inventario de cara a enfrentar el nuevo año que comienza.

Respecto a los datos relacionados al gasto en combustible que reflejaron las entidades que contestaron la ENIA 2019, se destacan los siguientes: el gas (natural, licuado y natural licuado), el petróleo (combustible y diésel), la leña y el coque.

Las tres columnas relacionadas con el gas son: natural (G035), licuado (G038) y natural licuado (G041); siendo el gas natural el combustible más utilizado dentro de las empresas de este clúster y en este trabajo; este combustible verde, es considerado así, porque es un tipo de materia que no es de alto impacto medioambiental; al contrastarlo con la ubicación geográfica y el tipo de actividad económica que en ella se destaca, podemos decir que existe una relación estrecha, puesto que, la fabricación de alimentos es un rubro productivo que requiere de grandes cantidades de gas para sus procesos internos.

Con respecto al petróleo, en el clúster hay 2 tipos y son: petróleo combustible (G023) y diésel (G026); el valor que muestran estos, arroja que este tipo de combustible dañino para el medio ambiente, es el segundo más utilizado por las empresas del rubro manufacturero de chile; la relación con el impacto medioambiental que tiene este gasto, se debe a que la combustión de petróleo genera gases que son perjudiciales para el ecosistema y la salud humana, según se explicó con anterioridad en el estado del medio ambiente del marco teórico.

Otro dato de los combustibles que destaca, es el de valor neto de coque (G020), el cual también es considerado un combustible dañino para el medio ambiente puesto que es un derivado del carbón y es utilizado para ciertos procesos productivos dentro las empresas manufactureras de nuestro país.

La leña (G047), es otra variable que destaca dentro de los combustibles y que es considerado como un alto contaminante al medio ambiente, puesto que al igual que el coque y el petróleo, la combustión de este, genera partículas que contaminan el aire y que son perjudiciales sobre todo para la salud humana pero también para el ecosistema.

En síntesis, este clúster agrupa a las empresas que más alto tienen su gasto en combustible y arroja que las entidades que más gastan en estos, están en la región de O'Higgins, tienen más de 500 trabajadores, son sociedades anónimas abiertas y su principal gasto en combustible lo generan a través del gas natural; pero también agrupa a las empresas que gastan en combustibles dañinos como lo es: el petroleo, coque o la leña, por tanto es considerado el cluster contaminante de esta investigación.

Cabe señalar que los resultados obtenidos de la clusterización de los datos, nos arroja una realidad nacional, puesto que en Chile existen menos empresas grandes respecto de las micro, pequeñas y medianas empresas, cuya cantidad es mucho mayor que las entidades grandes.

Cuando se inició esta investigación, el Instituto Nacional de Estadísticas tenía a disposición la Encuesta Nacional Industrial Anual del año 2019 como la más actual, y si considera-

mos que en ese año ocurrió un estallido social seguido de casi dos años de ardua pandemia de covid-19 o incluso un posible período de receso en nuestro país, se podría adicionar un sesgo en futuras versiones de los datos que son reflejados con esta encuesta. Por consecuencia, cientos de empresas realizaron un cierre abrupto de sus operaciones, esto trae consecuencias negativas si realizamos comparaciones con los datos sin depurar de una eventual ENIA 2020 o 2021, ya que no solo se genera una dificultad comparativa en términos de cantidades de empresas, sino que también, las externalidades a considerar en el momento que se realizó la encuesta podría reflejar datos menos fidedignos; sin embargo, para efectos de esta investigación se usó la encuesta más actual hasta la fecha en que se inició esta misma y era la ENIA 2019.

6.2. Conclusión

El problema abordado a lo largo de esta investigación, se relaciona con el impacto generado al medio ambiente a través del uso de combustible por parte de las industrias manufactureras en nuestro país, para ello analizamos el gasto reflejado por concepto de combustibles en la Encuesta Nacional Industrial Anual del año 2019, que está disponible en la página web del Instituto Nacional de Estadísticas.

El medio ambiente es un tema muy importante a analizar en estos días, ya que actualmente vivimos en el mundo un proceso de cambio de mentalidad, que busca cuidar el medio ambiente, pues es el lugar que habitamos y que debemos cuidar de cara al futuro y preservación de la vida sobre él; ante esto, las autoridades tienen un mecanismo de flujo de información por medio de reporte (REMA) cada año e informes (IEMA) cada cuatro años, pero dichos documentos no incluyen en su interior un desglose sobre el rubro industrial; de hecho, la clasificación que este le asigna es de "industria manufacturera", pero como se vio a lo largo del desarrollo de esta investigación, no todas las empresas de este rubro tienen el mismo impacto en el ecosistema.

Para realizar los análisis, se decidió utilizar técnicas de clustering en minería de datos, ya que el volumen de datos al que se tuvo acceso es alto, por ende hacer estudios de estos datos era complejo, esta metodología nos permitió agrupar datos con el fin de descubrir patrones y de esta forma cumplir con los objetivos de esta investigación; para ello fue necesario tomar la base de datos que inicialmente contiene un gran volumen de datos (281 columnas y 4.255 filas) y se procedió a depurar los mismos, con la finalidad de solo utilizar datos que nos permitieran caracterizar las empresas, algunos de estos fueron: sus ingresos, su ubicación geográfica, entre otros aspectos relevantes.

En la metodología clustering se pueden utilizar distintos algoritmos de agrupamiento, en nuestra investigación utilizamos k-means como se mostró en los resultados; este tipo de cluster utiliza como referencia los centroides, estos son un reflejo de la media de cada agrupación de datos analizada en la agrupación, es decir, es un dato que no necesariamente existe dentro de la base de datos puesto que es el promedio.

Al aplicar los algoritmos que nos indican los parámetros necesarios para realizar la clusterización, se obtuvo como resultado que los datos eran clusterizables y que se debían utilizar dos clusters con el fin de agrupar todos los datos que se seleccionaron anteriormente; el tamaño de cada agrupación también es un dato relevante puesto que el cluster 1 agrupó el 97,24 % (4.137) de las empresas , mientras que el 2 contiene el 2,75 % (117) de las entidades

analizadas, siendo este último, el cluster contaminante, puesto que contiene a las empresas que poseen un alto impacto en el medio ambiente a través del elevado gasto en combustible.

Recordando nuestro objetivo general de la tesis: **"Determinar las características que poseen las empresas según el tipo de combustible utilizado para generar energía en sus procesos productivos"**, la investigación nos brindó la información suficiente para determinar que gran parte de la concentración de las empresas poseen características que involucran al rubro minero, teniendo gastos considerablemente más altos por concepto de petróleo, o por otro lado, empresas del rubro de la fabricación de alimentos concentrando gastos en el uso del gas.

En primera instancia con el objetivo específico N° 1 pretendemos **"Identificar los tipos de combustibles más utilizados por las empresas y su efecto en el ecosistema."** Para ello utilizando la metodología clustering, se obtuvo como resultado que los combustibles más utilizados por las empresas son: el gas en sus tres versiones, el petróleo en sus dos versiones, la leña y el coque. Siendo el gas, el combustible más utilizado por las empresas del rubro manufacturero de nuestro país, este tipo de materia es considerada de poco impacto medioambiental pues su utilización no desemboca en problemas relacionados con la contaminación del aire; este es el caso del segundo combustible más utilizado, el petróleo, ya que como se analizó en el marco teórico, este combustible es altamente contaminante, tiene perjuicios sobre el ecosistema y la salud de las personas; dada su utilización en los procesos productivos acarrea gran contaminación al aire.

Luego, con el objetivo específico N°2 se busca **"Determinar las características comunes de las empresas que utilizan combustibles amigables con el medio ambiente."** Al finalizar con la investigación y utilizando la metodología de clustering, resulto que el gas era el combustible más utilizado por las empresas del rubro manufacturero de Chile y como se mencionó en el marco teórico, este es un combustible verde; la característica principal que une a todas las entidades que utilizan este combustible en sus procesos productivos, es el rubro de la fabricación de productos alimenticios, por ende esta sería la singularidad en común que poseen las empresas que menos impacto generan al medio ambiente a través de la contaminación del aire.

Siguiendo con nuestros objetivos específicos, el N°3 hace referencia a **"Determinar las características comunes de las empresas que utilizan combustibles no amigables con el medio ambiente."** Esta investigación, arrojó que uno de los combustibles más utilizados por las empresas manufactureras es el petróleo, como se analizó en profundidad en nuestro marco teórico, este combatible es uno de los principales contaminantes del aire, esto lo convierte en uno de los principales combustibles no amigables para el medio ambiente; la característica que tienen en común las empresas que utilizan este combustible en sus procesos productivos es que pertenecen al rubro de la minería, dado que este tipo de industria requiere de grandes cantidades de petróleo para sus procesos productivos, pero no solo esto, también requieren de este combustible para mover sus camiones que son la fuente principal de transporte de materia prima.

Finalmente, con el objetivo específico N°4 nos propusimos **"Analizar las diferencias que puedan existir por industria, región y tamaño de empresa."** En respuesta a esto, se puede notar que en Chile tenemos un variado conglomerado de empresas, de hecho los resultados obtenidos muestran que principalmente son dos industrias las que generan un alto gasto en combustible, estas son la empresas con fabricación de alimentos a través del

gas y las industrias relacionadas con la minería; ambos rubros completamente diferentes, ya que uno se dedica a la extracción de materia prima para su posterior fundición con el fin de ser exportada para la de fabricación de objetos; Otra característica es la diversidad de regiones donde se produce el gasto, ya que las empresas relacionadas con la fabricación de alimentos están concentradas en la zona central para luego extenderse hacia el sur del país, mientras que hacia el norte es más común con las industrias relacionadas con la minería, esta característica tiene directa relación con la variedad zonal que posee nuestro país; Otra cualidad que se destaca es el tamaño de las entidades, pues tiene directa relación con el nivel de gasto que estas organizaciones poseen, así como sus ingresos, lo que no es una sorpresa, ya que como es normal, mientras más grande la empresa, concentra un gasto mayor en las diferentes áreas de sus procesos productivos ya que trabajan con más altos volúmenes de producción que las entidades pequeñas.

Una de las limitaciones que debemos destacar de nuestra investigación, es que no existe manera de comparar los datos obtenidos con otros estudios relacionados a la ENIA, puesto que no existían indicios de la utilización de esta metodología con un fin igual a lo propuesto en esta tesis, al menos en Chile; además, debemos tener en cuenta la variada cantidad de empresas que participan al momento de responder la encuesta, ya sea, empresas grandes con una figura jurídica donde participan muchas personas, o por otro lado con empresas pequeñas que son formadas con la participación de solo una persona; por lo tanto, no es de extrañar que al momento de aplicar la metodología de clustering se destaque fácilmente este aspecto en forma gráfica, por otra parte, la cantidad de clusters se relaciona con lo anterior, ya que la metodología usada en esta investigación determina una cantidad igual a 2 de estas agrupaciones, es decir, empresas muy grandes se quedan reunidas en una agrupación (la que contiene menos empresas) y empresas más pequeñas quedan agrupadas en otro clúster, que como se analizó previamente, no poseen gastos destacables en combustible.

Como se pudo analizar en nuestro marco teórico, la metodología clustering tiene múltiples aplicaciones, por tanto, se vuelve factible la utilización de esta investigación en futuros proyectos relacionados al análisis de la base de datos que entrega la ENIA; puesto que utilizando esta metodología se podría obtener resultados positivos para el caso de estudiar el impacto de otras variables de esta encuesta, tal y como fue el caso de esta tesis.

7. Bibliografía

Manuel Agosin and José De Gregorio. Plan de reactivación industrial para Chile después del covid. December 2020. URL https://mirada.fen.uchile.cl/articulo/ver/plan-de-reactivacion-industrial-para-chile-despues-del-covid.

Kassambara Alboukadel. *Practical Guide To Cluster Analysis in R*. Number 3. 2017.

Rodrigo Amat. Clustering y heatmaps: aprendizaje no supervisado, September 2017. URL https://www.cienciadedatos.net/documentos/37_clustering_y_heatmaps.

Aquiles Gay. Educación Tecnológica, 1996. URL https://es.slideshare.net/AliumCalderon/educacin-tecnolgica-aquiles-gay.

María José Aranguren Querejeta. Política clúster del país vasco: lecciones aprendidas y retos. *Revista EAN*, (68):86–99, 2010.

Banco Mundial de Estadísticas. Consumo de energía procedente de combustibles fósiles (total), 2019. URL https://datos.bancomundial.org/indicator/EG.USE.COMM.FO.ZS.

Felipe Boccardo, Giorgio R̃uiz. Rstudio para estadística descriptiva en ciencias sociales. manual de apoyo docente para la asignatura estadística descriptiva. carrera de sociología, universidad de chile (segunda edición). 07 2019. doi: 10.13140/RG.2.2.18323.22564.

Luis Paulo Vieira Braga, Luis Iván Ortiz Valencia, and Santiago Segundo Ramírez Carvajal. *Introducción a la Minería de Datos*. Editora E-papers, 2009.

Armando De Ramón. Historia del sector industrial en chile. *Ambiente y Desarrollo*, 4(1), 1988.

Hugo Espinoza Benedetti. Clusters: Teoría y desarrollo. 2003.

León D. Fernández-Betancur. Energías alternativas. *TecnoLógicas*, (14):105, July 2005. ISSN 2256-5337, 0123-7799. doi: 10.22430/22565337.538.

Cristina Gil. Análisis de Componentes Principales., June 2018. URL https://github.com/CristinaGil/Ciencia-de-Datos-R/blob/master/PDF/Analisis_de_Componentes_Principales_(PCA).pdf.

Richard M Grinnell Jr and Yvonne Unrau. *Social work research and evaluation: Quantitative and qualitative approaches*. Cengage Learning, 2005.

Baptista Lucio Pilar Hernandez Sampieri Roberto, Fernandez Collado Carlos. *Metodología de la Investigación*. Number 5. 2010.

Erik Hren. CRISP-DM, October 2020. URL https://github.com/erikhren/CRISP-DM. original-date: 2020-02-23T17:08:20Z.

Daniel Coq Huelva and Sandra Ríos Núñez. Cambio estructural de la industria manufacturera en chile: 1979-2004. *rEviSta dE Economía mundial*, (26):27–51, 2010.

IEMA. Informe del estado del medio ambiente. Technical report, Gobierno de Chile., 2020. URL https://sinia.mma.gob.cl/wp-content/uploads/2021/04/14-calidad-del-aire.pdf.

Instituto Nacional de Estadísticas. Industria manufacturera. http://www.ine.cl/estadisticas/economia/industria-manufacturera/industria-manufacturera, 2019.

Instituto Nacional de Estadísticas. Sabes qué es el INE, 2022. URL http://www.ine.cl/ine-ciudadano/conoce-el-ine/sabes-que-es-el-ine.

J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, 5.1:281–298, January 1967. URL https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992. Publisher: University of California Press.

Zoraida Mamani Rodríguez, Luz Del Pino Rodríguez, and Augusto Cortez Vasquez. Minería de datos distribuida usando clustering k-means en la predictibilidad del proceso petitorio en una organización pública. *Industrial Data*, 20(2):123, December 2017. ISSN 1810-9993, 1560-9146. doi: 10.15381/idata.v20i2.13949. URL http://revistasinvestigacion.unmsm.edu.pe/index.php/idata/article/view/13949.

Mikel Navarro. Análisis y políticas de clusters: teoría y realidad. *Ekonomiaz: Revista vasca de economía*, (53):14–49, 2003.

JL Oviedo-Salazar, MH Badii, A Guillen, and O Lugo Serrato. Historia y uso de energías renovables history and use of renewable energies. *Daena Int. J. Good Conscience*, 10(1): 1–18, 2015.

Manuel Paradis. R para Principiantes, March 2003. URL https://cran.r-project.org/doc/contrib/rdebuts_es.pdf.

César Pérez López and Daniel Santin González. *Minería de datos. Técnicas y herramientas: técnicas y herramientas*. Editorial Paraninfo, 2007.

Michael E Porter et al. *Clusters and the new economics of competition*, volume 76. Harvard Business Review Boston, 1998.

REMA. Quinto reporte del estado del medio ambiente. Technical report, Gobierno de Chile., 2019. URL https://sinia.mma.gob.cl/wp-content/uploads/2022/06/REMA2019.pdf.

José Cristóbal Riquelme Santos, Roberto Ruiz, and Karina Gilbert. Minería de datos: Conceptos y tendencias. *Inteligencia Artificial: Revista Iberoamericana de Inteligencia Artificial, 10 (29), 11-18.*, 2006.

Santander-Universidades. Investigación cualitativa y cuantitativa: características y ventajas. https://www.becas-santander.com/es/blog/cualitativa-y-cuantitativa.html, 2019.

SINIA. Ministerio del Medio Ambiente, 2021. URL https://sinia.mma.gob.cl/estado-del-medio-ambiente/.

Lidya Tellerías. *INFORME PARA CHILE SUSTENTABLE*. PhD thesis, Pontificia Universidad Católica de Chile, 2018.

La Tercera. Ministra del Medio Ambiente y su presentación en la Conferencia sobre los Océanos: "Estamos en un momento crítico" - La Tercera, June 2022. URL https://www.latercera.com/que-pasa/noticia/ministra-del-medio-ambiente-y-su-presentacion-en-la-conferencia-sobre-los-oceanos-est WAINUHE2XNFMZE7JFBG7MJJEN4/.

United Nations. United nations environment programme-sdg indicators, 2019. URL https://unstats.un.org/sdgs/report/2020/goal-12/.

United States Statistics Division. Ensure sustainable consumption and production patterns, 2019. URL https://unstats.un.org/sdgs/report/2020/goal-12/.

Raquel Águila. La protección del medio ambiente en la constitución de chile. *Crítica urbana: revista de estudios urbanos y territoriales.*, 4(20):6, 2021.

Anexos

A. Anexo I: Tabla de tamaños según el número de trabajadores

Tabla 8: Asignación de intervalo de trabajadores ENIA 2019

Tamaño	Número de trabajadores
0	0 <= EMPTOT >= 04
1	05 <= EMPTOT >= 09
2	10 <= EMPTOT >= 19
3	20 <= EMPTOT >= 49
4	50 <= EMPTOT >= 99
5	100 <= EMPTOT >= 199
6	200 <= EMPTOT >= 499
7	500 <= EMPTOT >= 999
8	EMPTOT >= 1000

B. Anexo II: Tabla de centroides

Tabla 9: Tabla de centroides, elaborada con RStudio.

Categoría	Valor	Cluster
REGION1	0,00314831	1
REGION2	0,00342044	1
REGION3	0,00251594	1
REGION4	0,00041478	1
REGION5	0,00089509	1
REGION6	-0,00879741	1
REGION7	-0,00218329	1
REGION8	0,00419721	1
REGION9	-0,00204276	1
REGION10	-0,0010967	1
REGION11	0,00149081	1
REGION12	0,0009191	1
REGION13	0,00024807	1
REGION14	-0,00150329	1
REGION15	0,00258949	1
REGION16	-0,00151383	1
A0011	0,00701876	1
A0012	0,01713033	1

Continúa en la página siguiente

Tabla 9 – Continuación de la página anterior

Categoría	Valor	Cluster
A0013	-0,00609073	1
A0014	-0,01470572	1
A0015	-0,02690656	1
A0016	-0,00702382	1
A0017	0,00142717	1
A0018	0,0006079	1
A0019	0,00074461	1
A00110	0,00445902	1
A00111	0,0034088	1
A00112	0,00432961	1
A0055	0,00627456	1
D003	-0,0491147	1
D004	-0,03044356	1
G017	-0,00858361	1
G020	-0,02603997	1
G023	-0,04591815	1
G026	-0,03573583	1
G029	-0,01058304	1
G032	0,00091704	1
G035	-0,04231043	1
G038	-0,06643036	1
G041	-0,0255882	1
G044	-0,00085417	1
G047	-0,0252528	1
G050	-0,05298779	1
G051	-0,03974126	1
SUMAINGRESOS	-0,08245427	1
K001	-0,10340143	1
K009	-0,12008761	1
K010	-0,11090451	1
K011	-0,10597629	1
TAMANO1	0,00590879	1
TAMANO2	0,01409916	1
TAMANO3	0,01912781	1
TAMANO4	0,00946574	1
TAMANO5	0,0008957	1
TAMANO6	-0,02628316	1
TAMANO7	-0,03613879	1
TAMANO8	-0,05855794	1
IMPORTACION1	-0,00282054	1
EXPORTACION1	0,01869529	1
REGION1	-0,11230796	2

Continúa en la página siguiente

Tabla 9 – Continuación de la página anterior

Categoría	Valor	Cluster
REGION2	-0,12201537	2
REGION3	-0,08974953	2
REGION4	-0,01479604	2
REGION5	-0,03192987	2
REGION6	0,31382469	2
REGION7	0,07788311	2
REGION8	-0,14972448	2
REGION9	0,07287031	2
REGION10	0,03912201	2
REGION11	-0,05318072	2
REGION12	-0,03278635	2
REGION13	-0,00884913	2
REGION14	0,0536259	2
REGION15	-0,0923734	2
REGION16	0,05400206	2
A0011	-0,25037618	2
A0012	-0,61108008	2
A0013	0,21727099	2
A0014	0,52458861	2
A0015	0,95982183	2
A0016	0,25055652	2
A0017	-0,05091067	2
A0018	-0,02168539	2
A0019	-0,0265622	2
A00110	-0,15906399	2
A00111	-0,12160016	2
A00112	-0,15444764	2
A0055	-0,22382874	2
D003	1,75203997	2
D004	1,08599517	2
G017	0,30619797	2
G020	0,92890851	2
G023	1,63801138	2
G026	1,27478347	2
G029	0,3775227	2
G032	-0,03271295	2
G035	1,50931507	2
G038	2,36973129	2
G041	0,91279301	2
G044	0,03047019	2
G047	0,90082845	2
G050	1,89020221	2

Continúa en la página siguiente

Tabla 9 – Continuación de la página anterior

Categoría	Valor	Cluster
G051	1,41766681	2
SUMAINGRESOS	2,94134286	2
K001	3,68857858	2
K009	4,28381493	2
K010	3,95623143	2
K011	3,78042998	2
TAMANO1	-0,21078079	2
TAMANO2	-0,50295096	2
TAMANO3	-0,68233501	2
TAMANO4	-0,33766564	2
TAMANO5	-0,03195161	2
TAMANO6	0,93758367	2
TAMANO7	1,28915802	2
TAMANO8	2,0889029	2
IMPORTACION1	0,10061557	2
EXPORTACION1	-0,66690609	2